Tratamiento Estadístico de Datos con Aplicaciones en R

Luis Alberto Díaz Chávez
Jairo Rafael Rosado Vega

UNIVERSIDAD DE LA GUAJIRA | SHIKII EKIRAJIA PULEE WAJIIRA

Tratamiento Estadístico de Datos con Aplicaciones en R

Luis Alberto Díaz Chávez
Jairo Rafael Rosado Vega

UNIVERSIDAD DE LA GUAJIRA | SHIKII EKIRAJIA PULEE WAJIIRA

Tratamiento Estadístico de Datos con Aplicaciones en R

Primera edición, 2019

ISBN: 978-958-8942-58-2

Carlos Arturo Robles Julio
Rector

Hilda María Choles Almazo
Vicerrectora Académica

Víctor Pinedo Guerra
Vicerrector de Investigación y Extensión

Sulmira Patricia Medina
Directora Centro de Investigaciones

Impresión*:*
Editorial Gente Nueva

Depósito legal

Impreso en Colombia
Printed in Colombia

A mi familia,
principales testigos y creyentes
de mis esfuerzos y pasión
por la academia.

Contenido

Presentación

Como rector de la universidad de La Guajira me complace poner a disposición de la comunidad estudiantil, docentes e investigadores de nuestra alma mater la presente obra, que se constituye en un primer esfuerzo dentro de nuestra institución para contar con un documento de consulta en donde se referencien los procedimientos estadísticos de mayor relevancia, para apoyar las actividades en nuestras aulas durante el desarrollo de los cursos de Estadística Descriptiva, Estadística Inferencial y Diseño de Experimentos, y así mismo, otros de singular importancia en el tratamiento de datos productos de las investigaciones desarrolladas por los grupos de investigación.

Por otro lado, esta obra se suma a los esfuerzos a nivel global en la adopción y masificación del uso de softwares libres, como es el caso de R, un lenguaje de programación de gran versatilidad y potencia, que en los últimos años ha tomado un reconocimiento exponencial y favoritismo de investigadores y analistas de datos, gracias al amplio espectro de herramientas que ofrece y la excelente presentación gráfica de reportes estadísticos que a través de él se pueden lograr, y sobre el cual, se sustentó la aplicación de los métodos y procedimientos descritos en este texto.

Hago llegar mis felicitaciones a los autores de esta obra por entregarnos esta valiosa herramienta que fortalecerá el desarrollo académico de nuestra institución y catapultará la proyección social de nuestra alma mater, en la comunidad académica y científica.

Carlos Arturo Robles Julio
Rector

Agradecimientos

Los autores de esta obra expresamos nuestro sincero agradecimiento a todas las personas que hicieron posible la realización de la misma, especialmente a:

Carlos Robles Julio, Rector de la Universidad de La Guajira, por su apoyo y valiosa gestión que hizo posible la publicación de la obra y ofrecer la oportunidad de capacitación de los autores en diferentes cursos de formación continuada que permitieron ampliar los conocimientos que hicieron posible la realización de la misma.

Víctor Pinedo Guerra, Vicerrector de Investigación y Extensión y Sulmira Medina Payares, Directora del Centro de Investigaciones, por el apoyo, estímulo y gestión logística para su publicación.

A los jóvenes integrantes del Grupo de Investigación Pichihuel, por su dedicada participación en los diferentes proyectos realizados por el grupo, que permitieron obtener datos, a través de los cuales se modelaron las diferentes técnicas y procedimientos estadísticos que en el libro se describen.

Prólogo

El contenido de esta obra se condensa en diez capítulos ordenados de forma sistemática para permitir al lector el aprendizaje paso a paso de cada una de las técnicas estadísticas que en ella se exponen, acorde al contenido programático que se imparte en la gran mayoría de cursos de Estadística Descriptiva, Estadística Inferencial y Diseño de Experimentos, tratando de hacer énfasis en la interpretación de resultados, más que al desarrollo matemático de las mismas.

En el capítulo 1, se presenta una conceptualización y definición de algunos términos básicos de la estadística necesarios para una mejor comprensión de los capítulos posteriores.

El capítulo 2, enfatiza las temáticas más relevantes relacionadas con la estadística descriptiva, especialmente en el conocimiento e interpretación de medidas de resumen (tendencia central y variabilidad), que no faltan en cualquier reporte científico, medidas de forma, para entender la distribución de los datos y construcción de gráficos estadísticos como herramienta de soporte en la presentación de resultados.

El capítulo 3, se adentra en el campo de la probabilidad, específicamente en el estudio de variables aleatorias de naturaleza discreta y continua, en el estudio de las principales distribuciones de probabilidad y su campo de aplicación en el desarrollo de experimentos científicos reales.

En el capítulo 4, se inicia el estudio de la inferencia estadística, a través de la estimación puntual por intervalos de parámetros poblacionales con el propósito de emitir conclusiones validas de los mismos bajo el concepto de confiabilidad.

En capítulo 5, expone los procedimientos de prueba de hipótesis de parámetros poblacionales a partir de una o dos muestras, donde se aprenderá a validar o verificar conjeturas que se realicen sobre uno o dos parámetros poblacionales.

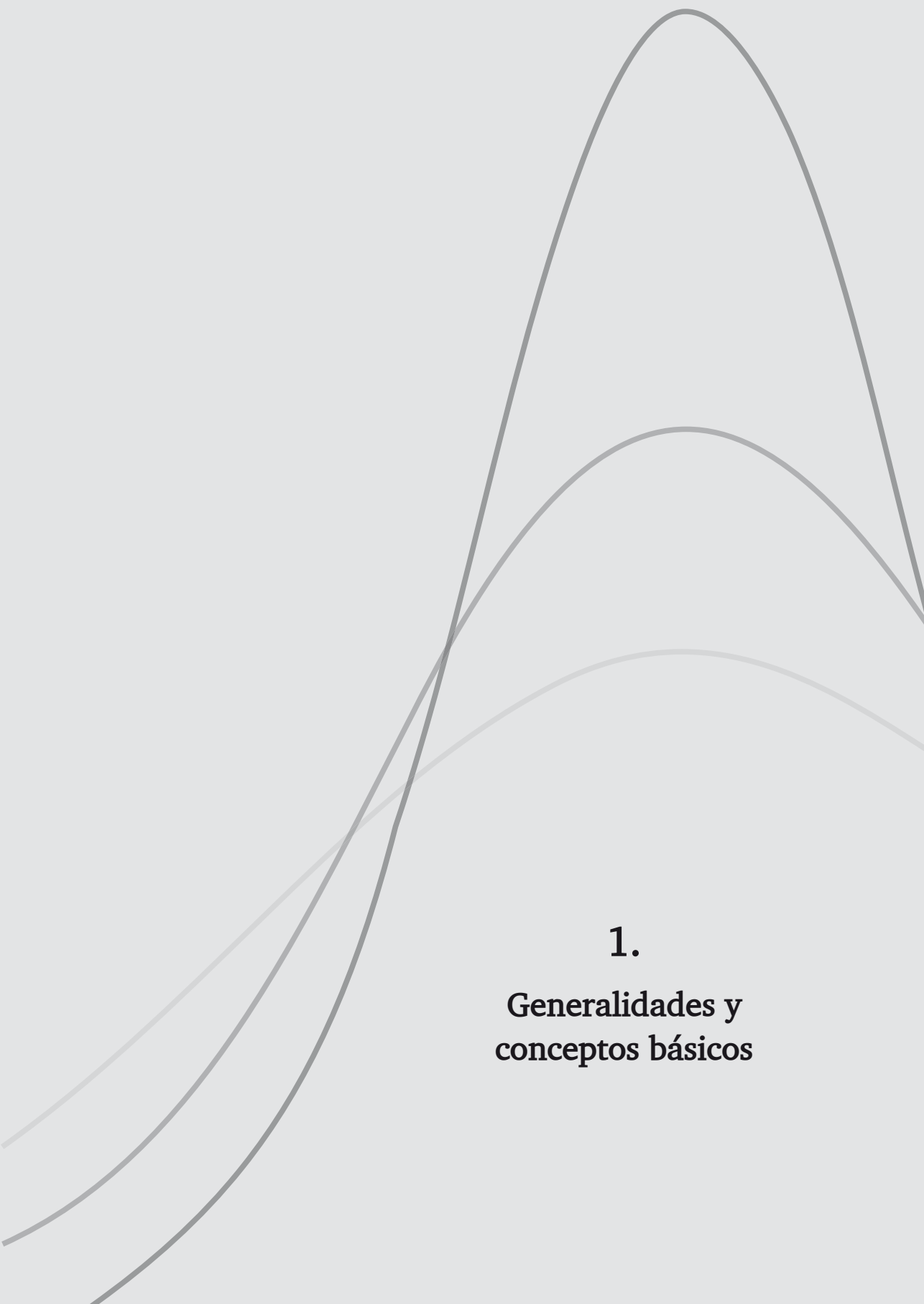
El capítulo 6, presenta las pruebas de bondad de ajuste de uso más concurrido en la práctica para la diagnosis de los datos a ser utilizados en procedimientos de prueba de hipótesis estadísticas

El capítulo 7, se dedica ampliamente al estudio de diseño de experimentos, específicamente en el tratamiento profundo del análisis de varianza en diferentes situaciones experimentales, generalidades, diagnosis de los datos para su aplicación y la exposición de un conjunto de pruebas a usarse posteriormente al análisis de varianza (pruebas de comparaciones múltiples).

El capítulo 8, contiene el estudio de los modelos de regresión, desde las generalidades de los mismos y la exposición de escenarios donde es pertinente su utilización, construcción de modelos simples, múltiples, no lineales y logísticos, la evaluación del cumplimiento de los supuestos que cada modelo debe cumplir y la presentación de diferentes pruebas que permiten determinar la eficiencia de los modelos construidos.

En el capítulo 9, se exponen los principales procedimientos enmarcados dentro de la estadística no paramétrica o de distribución libre, como alternativa a los procedimientos de prueba de hipótesis y análisis de varianza expuestos en los capítulos 5 y 7, cuando en estos no se cumplan los supuestos para su aplicación.

Por último, el capítulo 10, se dedica al estudio de las principales técnicas de análisis multivariante de datos, para el tratamiento simultáneo de tablas de datos con un elevado número de variables, y de las cuales se desean extraer conclusiones con el examen conjunto de dichas variables.



1.

Generalidades y conceptos básicos

1.1. Panorama general: Tratamiento de datos

Los procesos de recolección, organización, presentación, procesamiento, análisis e interpretación de datos numéricos son aspectos fundamentales en el desarrollo de un estudio o una investigación científica en general (Vargas, 2007); esto último, es la esencia de la rama de las matemáticas denominada estadística y que dependiendo del campo de aplicación, toma distintas denominaciones, por lo que un término más generalizado para referirnos a esta ciencia sin discriminar a que disciplina es aplicada es "tratamiento de datos".

Así, los datos se constituyen en la materia prima a través de los cuales se busca sacar conclusiones validas por parte de los investigadores que hacen uso de métodos estadísticos, brindando validez, confiabilidad y por supuesto objetividad a los resultados de sus investigaciones.

Los datos provienen de un proceso de medición u observación que debe hacerse de manera regular, organizada y sistemática, de tal forma que permita obtener un sistema confiable de observaciones con el fin de acercarse a la respuesta de los interrogantes específicos de una investigación (Vargas, 2007). Sin embargo, los procesos de recolección y calidad de los datos es un aspecto poco conocido por los investigadores, ignorando que es una de las fases de la experimentación que debe planearse con mucho cuidado, pues como ya se ha mencionado, de ello depende la calidad de las conclusiones que se generen de una investigación.

1.2. Definiciones básicas

Como cualquier ciencia, la estadística también tiene una terminología (lenguaje) propia que facilita su comprensión. A continuación, se presentan algunos conceptos y definiciones que son básicos para la comprensión de muchos de los temas que serán tratados en las demás secciones de este libro.

Población estadística. Población se refiere al grupo completo de interés que deseamos describir o del cual deseamos obtener conclusiones (Ferrer,

2007). Una población puede ser definida como un grupo de individuos, como por ejemplo, personas, animales, objetos o mediciones.

Parámetro. Se refiere a un indicador estadístico que es calculado a través del total de observaciones o datos de la población. El valor del parámetro es constante y generalmente desconocido, el cual se estima a través de los datos de una muestra.

Muestra. Una muestra no es más que una parte de la población o universo. Esto es, salvo en las raras excepciones que podemos obtener información para la población o universo entero, lo que se obtiene siempre que vamos al campo a muestrear, es decir, a obtener muestras. Ejemplo de esto, es cada vez que vamos a un rio determinado (por ejemplo el Ranchería) y obtenemos valores de oxígeno disuelto, esos valores representan una muestra de la población total del oxígeno disuelto del rio, cada vez que colectamos individuos de *Tubificidae* en la Laguna Salada, el grupo de individuos colectados representan una muestra de la población total de *Tubificidae* presentes en la Laguna Salada.

Es importante enfatizar que el éxito y confiabilidad de las inferencias estadísticas que hagamos de cualquier atributo o proceso, dependerá de la calidad de la muestra y no de la población completa (Ferrer, 2007).

Estadístico. Un estadístico, o también llamado *estadígrafo,* se refiere a un indicador que es calculado de las observaciones o datos de la muestra. En general estos indicadores son los que pretenden generalizar a la población a través del proceso de inferencia estadística (Vargas, 2007).

Variable. Una variable es una característica o fenómeno que puede tomar diferentes valores, por ejemplo, peso, longitud, sexo, salinidad, concentración de oxígeno disuelto, abundancia, riqueza de especies, diversidad, etc. La clave es que todas estas características o atributos difieren de individuo a individuo, de ecosistema a ecosistema, etc. En resumen, cualquier objeto o evento que puede variar en sucesivas observaciones, bien sea cualitativa o cuantitativamente, se denomina variable (Ferrer, 2007).

En términos estadísticos, las variables pueden ser clasificadas de dos formas generales: como **variables cuantitativas** (intervalares), que son aquellas de naturaleza numérica, por ejemplo, la estatura de una persona, la longitud de un pez, el número de colonias de coliformes fecales en un

cultivo microbiológico, etc. Estas a su vez se clasifican en **variables discretas**, cuyos valores son enumerables y solo toman valores enteros (número de hijos, número de personas, número de quironómidos en una laguna, etc.), y **variables continúas**, que pueden tomar infinitos valores dentro de un intervalo (peso, temperatura, concentración de oxígeno disuelto, etc.).

Asimismo, las variables pueden ser clasificadas como **cualitativas** (categóricas), las cuales representan una característica o atributo no sometido a cuantificación, ejemplo de este tipo de variables, es el color de ojos de una persona, la presencia o ausencia de un grupo bacteriano determinado en un cultivo, el estrato socioeconómico, etc.

A cada tipo de variable, le corresponde una escala de medición. En este sentido, las variables cualitativas pueden ser **nominales**, si sus valores no están ordenados de modo natural (lugar de nacimiento, especie), u **ordinales**, si sus valores tienen un orden (por ejemplo, una variable "toxicidad" que toma los valores, nada, poco, bastante y muy toxico) (Guisande *et al.*, 2011).

Igualmente, las variables cuantitativas pueden estar enmarcadas dentro de dos escalas de medición: la **Escala de intervalo**, cuya propiedad característica es la igualdad de la distancia entre puntos de escala de la misma amplitud. Aquí puede establecerse orden entre sus valores, hacerse comparaciones de igualdad, y medir la distancia existente entre cada valor de la escala. El valor cero de la escala no es absoluto, sino un cero arbitrario: no refleja ausencia de la magnitud medida, por lo que las operaciones aritméticas de multiplicación y división no son apropiadas. Cumple con las propiedades de identidad, magnitud e igual distancia. La igual distancia entre puntos de la escala significa que puede saberse cuántas unidades de más tiene una comparada con otra, con relación a cierta característica analizada. Por ejemplo, en la escala de temperatura centígrada puede decirse que la distancia entre 25° y 30°C es la misma que la existente entre 20° y 25° C, pero no puede afirmarse que una temperatura de 40° C equivale al doble de 20° C en cuanto a intensidad de calor se refiere, debido a la ausencia de cero absoluto (Orlandoni, 2010).

La **Escala de razón**, corresponde al nivel de medición más completo para variables cuantitativas. Tiene las mismas propiedades que la escala intervalos, y además posee el cero absoluto. Aquí el valor cero no es arbitrario, pues representa la ausencia total de la magnitud que se está

midiendo. Con esta escala se puede realizar cualquier operación lógica (ordenamiento, comparación) y aritmética. A iguales diferencias entre los números asignados corresponden iguales diferencias en el grado de atributo presente en el objeto de estudio. Ejemplos: longitud, peso, distancia, ingresos, precios.

Es importante conocer el tipo de variable con la que estemos tratando durante el desarrollo de una investigación, así como su escala de medición, pues de esto depende, el tipo de tratamiento estadístico que se le va a dar para extraer las mejores inferencias de las mismas, la Figura 1.1 muestra un diagrama donde se presenta la clasificación de las variables estadísticas y su consecuente escala de medición.

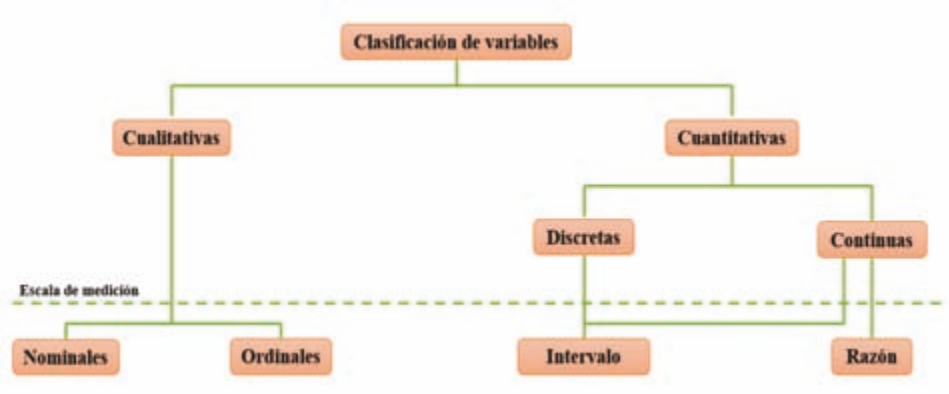
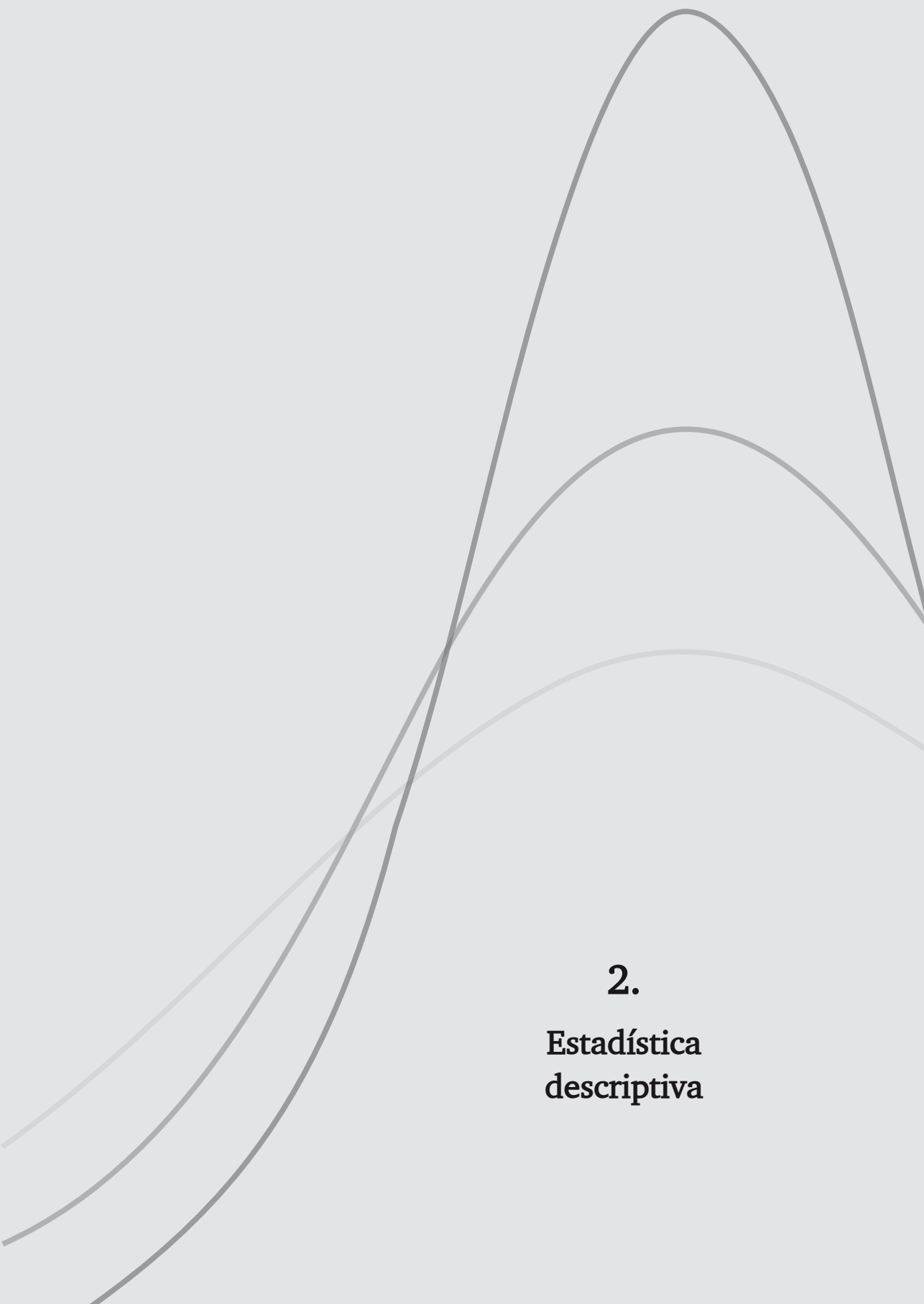


Figura 1.1. Clasificación de las variables estadísticas

1.3. Estadística descriptiva e inferencial

La estadística para su estudio se divide en dos tipos: la Estadística Descriptiva y la Estadística Inferencial. La primera representa y describe un conjunto de datos utilizando métodos numéricos y gráficos de resumen, y presenta la información contenida en ellos, mientras que la segunda se apoya en el cálculo de probabilidades y datos muéstrales para efectuar estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos (Población).



2.

Estadística descriptiva

2.1. Organización y presentación de datos

El proceso subsecuente a la recolección de datos se refiere a la organización de los mismos para realizar una presentación que permitan realizar cierto nivel de análisis y de descripción del fenómeno que se está estudiando en el desarrollo de una investigación. Una de las formas más simples y usuales de organización de la información es a través de **tablas de distribución de frecuencias**, que son la representación conjunta de los datos en forma tabular correspondientes a un fenómeno en estudio y su ordenamiento en base al número de observaciones que corresponden a cada dato o a cada grupo de datos, adecuados según cronología, geografía, análisis cuantitativo o cualitativo (Córdova & Cortés, 2010).

La construcción de tablas de distribución de frecuencias, tiene un tratamiento matemático distinto dependiendo de la tipología de las variables (discretas o continuas). A continuación se describirá las dos metodologías y su posterior aplicación en el entorno de R.

Variables discretas

La construcción de tablas de distribución de frecuencias para variables discretas es bastante sencillo, basta con la organización de la información en un formato tabular que contenga en su primera columna los valores de la variable que se está estudiando (y_i), en la segunda columna, se ubican las frecuencias absolutas (n_i), es decir, el número de veces que se repite cada valor de la variable de estudio, y la tercera columna contendrá las frecuencias relativas (h_i), correspondiente al cociente entre las frecuencias absolutas y al número de observaciones de la muestra (n); otra información de interés que puede anexarse a las tablas de distribución de frecuencias son las frecuencias absolutas acumuladas (N_i) y las frecuencias relativas acumuladas (H_i), esto último con el fin de facilitar algunas inferencias, dependiendo de las necesidades de cada caso en particular. En general, la tabla de distribución de frecuencias resultante deberá verse así:

y_i	n_i	h_i	N_i	H_i
y_1	n_1	h_1	N_1	H_1
y_2	n_2	h_2	N_2	H_2
y_3	n_3	h_3	N_3	H_3
$\sum$	n	1,00	-	-

A continuación, veremos a través de un ejemplo práctico, la metodología para elaborar tablas de distribución de frecuencias de forma mecánica y a través de programación en el entorno de R.

Ejemplo 2.1. Un experimento agrícola consistió en contar el número de flores por planta de una muestra de 50 plantas ($n = 50$). Los valores resultantes del conteo son los siguientes:

10	8	6	3	9	7	5	4	6	9
8	10	7	9	10	6	8	6	3	2
4	3	2	7	5	5	4	3	7	6
6	7	8	8	6	7	7	9	8	6
5	3	2	1	4	3	6	7	7	0

Para una mejor interpretación de los datos se requiere resumirlos a través de una tabla de distribución de frecuencias

Solución

Para cada número de flores, la distribución de frecuencias se muestra en la siguiente tabla

y_i	n_i	h_i	N_i	H_i
0	1	0.02	1	0.02
1	1	0.02	2	0.04
2	3	0.06	5	0.10
3	6	0.12	11	0.22
4	4	0.08	15	0.30
5	4	0.08	19	0.38
6	9	0.18	28	0.56
7	9	0.18	37	0.74
8	6	0.12	43	0.86
9	4	0.08	47	0.94
10	3	0.06	50	1.00
$\sum$	**50**	**1.00**	-	-

A continuación, mostraremos la ruta de programación utilizada en R para obtener la tabla anterior. Inicialmente tabulamos los datos en la hoja de cálculo de Excel y los guardamos en un archivo de extensión *.csv* (Delimitado por comas), para importarlos con mayor facilidad y eficiencia en el ambiente de programación de R. Luego cargamos en R los datos originales a través de la función ***read.csv2***, indicando en su argumento el nombre del archivo entre comillas, seguido a esto se especifica si la primera fila de los datos corresponde al nombre de la variable (encabezado) con la orden *header = TRUE* y la identificación del acento (tildes) con la orden *encoding = "latin1"*, tal como se muestra a continuación

```
> Flores<-read.csv2("Numero de flores.csv",header=TRUE, encoding=
"latin1")
```

Una vez cargados los datos en R, proseguimos con determinar las frecuencias absolutas a través de la función ***table***, las frecuencias relativas con la función ***prop.table*** y las frecuencias absolutas acumuladas y frecuencias relativas acumuladas con la orden ***cumsum***, colocando en el argumento los comandos para determinar frecuencias absolutas y relativas, respectivamente

```
> Frec.abs<-table(Flores)
> Frec.abs
Flores
 0  1  2  3  4  5  6  7  8  9 10
 1  1  3  6  4  4  9  9  6  4  3
> Frec.rel<-round(prop.table(Frec.abs),2)
> Frec.rel
Flores
   0    1    2    3    4    5    6    7    8    9   10
0.02 0.02 0.06 0.12 0.08 0.08 0.18 0.18 0.12 0.08 0.06
> Frec.abs.acum<-cumsum(Frec.abs)
> Frec.abs.acum
 0  1  2  3  4  5  6  7  8  9 10
 1  2  5 11 15 19 28 37 43 47 50
> Frec.rel.acum<-cumsum(Frec.rel)
> Frec.rel.acum
   0    1    2    3    4    5    6    7    8    9   10
0.02 0.04 0.10 0.22 0.30 0.38 0.56 0.74 0.86 0.94 1.00
```

Note que al momento de determinar las frecuencias relativas se usó de manera simultánea una orden llamada ***round***, esta nos permite hacer un redondeo de las unidades decimales al objeto puesto en el argumento de esta función (***prop.table(Frec.abs)***), en nuestro caso este redondeo se hizo a nivel de las centésimas, es decir, dos unidades decimales.

Interpretación: Algunas inferencias que se pueden extraer de la tabla de frecuencia construida anteriormente es que los valores 6, 7 y 8 de la variable número de flores por planta, fueron los que se observaron con mayor frecuencia, 9 plantas (18%) presentaron 6 flores, 9 plantas (18%) tuvieron 7 flores, 6 plantas tuvieron 8 flores; pocas fueron las plantas sin flores (2%); el 10% de las plantas tuvieron 2 o menos flores; el número máximo de flores por planta en esta experiencia fue de 10 y sólo en el 6% de la muestra se registró este valor máximo.

Estas afirmaciones, como algunas otras, pueden obtenerse de la lectura de una tabla de frecuencias, y no son fáciles de formular a partir de los datos sin procesar, sobre todo cuando *n* es grande (Di Rienzo *et al.*, 2005).

Variables continuas

La elaboración de tablas de distribución de frecuencias para variables continuas difiere un poco a la de variables discretas, principalmente porque asume mayores cálculos matemático, esto se debe a la naturaleza de este tipo de variables de asumir infinitos valores dentro de un intervalo, por lo que el recuento de las frecuencias absolutas se hace sobre la base del número de intervalos creados que se repiten y no sobre el número de cada caso en particular, pues existe el riesgo de que ninguno de los valores se repitan dificultando observar la tendencia de los datos a través de un gráfico. A continuación listaremos la metodología paso a paso para la construcción de tablas de frecuencia con variables continuas.

- Inicialmente determinamos el valor mínimo y máximo del conjunto de datos: $x_{\min}$, $x_{\max}$
- Hallamos el *rango* o *recorrido* del conjunto de datos, definido como la diferencia entre el valor máximo y mínimo de los mismos: $R = x_{\max} - x_{\min}$ (2.1)
- Calculamos el número de intervalos (*m*) que se utilizaran para agrupar los datos. Una de las formas de hacer esto es aplicando la regla de *Sturges*, definida matemáticamente como

$$m = 1 + 3.3\log(n) \tag{2.2}$$

- Luego de determinar el número de intervalos, procedemos a calcular la amplitud (A) que tendrá cada uno de los intervalos. Así:

$$A = \frac{x_{\max} - x_{\min}}{m} = \frac{R}{m} \tag{2.3}$$

- Como paso siguiente definimos los límites de cada intervalo, se inicia con el valor inicial que puede ser definido como el valor mínimo del conjunto de datos, o como el menor valor entero al valor mínimo, con el fin de que los límites de los intervalos tengan valores enteros y esto facilite la interpretación de la distribución de frecuencias.
- El último paso es determinar cuántos de los datos (frecuencias) se encuentran dentro de cada uno de los intervalos de agrupación formados.

Veamos un ejemplo donde se ilustre lo anterior

Ejemplo 2.2. A continuación se muestran los niveles de presión sonora de ruido medida en decibeles (dB) en diferentes estaciones de muestreo, en un estudio de ruido ambiental en la ciudad de Cali en horario diurno (Vargas, 2007). Agrúpelos a través de una tabla de frecuencias.

63.7	75.0	70.5	72.1	67.2	65.1	59.6	64.1	61.1	62.0
66.9	76.3	73.7	74.1	62.3	55.3	70.6	53.3	65.9	64.0
66.8	71.4	71.0	76.5	69.4	71.3	65.3	62.5	62.6	58.7
75.3	77.4	56.1	57.3	60.5	72.3	74.0	62.3	50.2	68.2
70.8	71.6	69.0	71.6	75.0	64.6	74.9	75.4	50.9	61.6

Solución

Inicialmente identificamos el valor máximo y mínimo del conjunto de datos, correspondientes a las magnitudes $x_{\min} = 50.2\text{dB}$ y $x_{\max} = 77.4\text{dB}$

Calculamos el rango del conjunto de datos $R = 77.4 - 50.2 = 27.2\,dB$

Se determina el número de intervalos en los cuales se agruparán los datos a través de la ecuación 2.2, redondeando las cifras decimales siempre al mayor entero más cercano

$$m = 1 + 3.3\log(50) = 6.6 \approx 7$$

Esto indica que debemos construir 7 intervalos. Siguiendo con lo anterior, hallamos la amplitud de los intervalos con la ecuación 2.3, en esta se utiliza el valor no aproximado (redondeado) del cálculo del número de intervalos

$$A = \frac{27.2}{6.6} = 4.1 \approx 5$$

Esto es, cada intervalo tendrá una amplitud de 5 dB. Ahora definimos los límites de los intervalos; considerando que el valor mínimo del conjunto de datos es 50.2 dB, se selecciona el menor entero al valor mínimo como valor inicial, este valor corresponde a 50 dB.

A partir de este valor se construyen los límites superiores de cada intervalo sumando de manera sucesiva la amplitud a los límites inferiores de cada intervalo. Por último, se identifican cuantos datos se encuentran contenido en los intervalos, es decir sus frecuencias absolutas, y a partir de estas, se calculan las frecuencias relativas, las frecuencias absolutas acumuladas y las frecuencias relativas acumuladas, como se indica en la siguiente tabla.

Intervalos de ruido	n_i	h_i	N_i	H_i
50 – 55	3	0.06	3	0.06
55 – 60	5	0.10	8	0.16
60 – 65	12	0.24	20	0.40
65 – 70	9	0.18	29	0.58
70 – 75	16	0.32	45	0.90
75 – 80	5	0.10	50	1.00
$\sum$	**50**	**1.00**	-	-

Note que el séptimo intervalo (80 – 85), se suprimió de la tabla, dado que en este no se encuentra contenido ninguno de los datos (frecuencias iguales a cero), lo que desmejora la presentación de la tabla y puede dificultar su interpretación.

Para la solución de este problema en R, debemos hacerle “trampa”, es decir, utilizar una ruta de programación diferente a la empleada cuando se tratan variables discretas. Lo primero que debemos hacer es tabular los datos, guardarlos bajo la extensión *.csv* y cargarlo en el ambiente de R como se indicó en el ejemplo anterior; luego creamos un objeto al que

denominaremos *Tabla*, al que se le asignará un histograma de frecuencias, construido con la función ***hist***, luego ordenamos a R que nos muestre las frecuencias absolutas, invocando a un objeto de esta función llamado ***count***, es decir, el conteo de los valores contenidos en cada uno de los intervalos, las frecuencias relativas se obtienen indicando a R que nos muestre el resultado de dividir las frecuencias absolutas por el número total de observaciones (***length***) de nuestro conjunto de datos; por último, obtenemos las frecuencias absolutas y relativas acumuladas, con la orden ***cumsum***, antes utilizada. La salida de resultados de R para este procedimiento es la siguiente

```
> Presión<-read.csv2("Presión Sonora.csv",header=TRUE,encoding=
"latin1")
> attach(Presión)
> Tabla<-hist(Presión.Sonora)
> Frec.abs<-Tabla$count
> Frec.abs
[1]  3  5 12  9 16  5
> Frec.rel<-Frec.abs/length(Presión.Sonora)
> Frec.rel
[1] 0.06 0.10 0.24 0.18 0.32 0.10
> Frec.abs.acum<-cumsum(Tabla$counts)
> Frec.abs.acum
[1]  3  8 20 29 45 50
> Frec.rel.acum<-cumsum(Frec.rel)
> Frec.rel.acum
[1] 0.06 0.16 0.40 0.58 0.90 1.00
```

Interpretación: Algunas de las variadas interpretaciones que se pueden realizar de la tabla de frecuencias anterior es que de todos los puntos muestreados, 16 de ello presentaron niveles de presión sonora entre 70 y 75 dB (32%), 12 presentaron valores entre 60 y 65 dB, abarcando el 29% de los puntos de muestreo, y que el 16% de los puntos mostraron valores de presión sonora entre 50 y 60 dB.

2.2. Gráficos

Los gráficos estadísticos, consisten en la utilización de puntos, líneas y figuras que sirven para mostrar magnitudes, asociadas a una escala de medición, de manera que se facilita la comparación e interpretación de los datos estadísticos, sin que necesariamente se incluyan los valores numéricos. No obstante, la diversidad de gráficos y sus modalidades de ilustración es muy variada, y a pesar de constituir opciones útiles para

una consulta ágil de la información, existe también el riesgo de no usar el tipo de gráfico adecuado o enfatizar el atractivo visual, dejándose en un segundo plano el objetivo esencial del gráfico en cuanto a facilitar la consulta de los datos. Por ello, en la presentación de datos es necesario considerar criterios básicos sobre el uso de recursos gráficos y lineamientos sobre su diseño, de tal manera que las distintas opciones sean eficientemente aprovechadas (INEI, 2006).

El gráfico estadístico debe estructurarse teniendo en cuenta la utilidad que preste al usuario común; es decir, que quien lo diseña deba colocarse en el lugar del que utilizará la información. La construcción del mismo es una labor aparentemente sencilla, sin embargo en la práctica es necesario tener en cuenta elementos que faciliten su comprensión e interpretación.

Así, las representaciones gráficas deben conseguir que un simple análisis visual ofrezca la mayor información posible, y dependiendo del tipo de variables que estemos estudiando, usaremos una representación gráfica u otra. En las secciones siguientes estudiaremos las representaciones graficas más utilizadas en el tratamiento de datos, ilustradas a través de ejemplos prácticos modelados en el ambiente de programación de R, y se brindaran interpretaciones de cada uno de ellos.

2.2.1. Gráficos para variables cualitativas o cuantitativas discretas

Existen dos tipos de gráficos usados con mayor frecuencia en el tratamiento de datos de variables nominales o cualitativas, o cuantitativas discretas: el gráfico de barras y el gráfico de sectores. Ambos se construyen en función de las frecuencias con las que se presentan cada uno de los datos, en breve daremos una descripción más detallada de estos y mostraremos algunos ejemplos.

Gráficos de barras

Un gráfico de barras es aquella representación gráfica bidimensional en que los objetos gráficos elementales son un conjunto de rectángulos dispuestos paralelamente de manera que la extensión de los mismos es proporcional a la magnitud que se quiere representar.

En este tipo de gráficos se puede representar la información contenida en una tabla de frecuencias, colocando en el eje de las ordenas las frecuencias absolutas o relativas y en el eje de las abscisas los distintos valores que toma la variable. Veamos un ejemplo de esto.

Ejemplo 2.3. Los siguientes datos muestran un estudio sobre la composición vegetal en el sector oeste de la Laguna salada (Rosado, 2009). A partir de estos datos construya un gráfico de barras en función de sus frecuencias absolutas y sus frecuencias relativas.

L.racemosa	*L. racemosa*	*L. racemosa*	*T. cattapa*	*E. caribeae*	*C. luzulae*	*C. luzulae*	*C. ferax*
L. racemosa	*L. racemosa*	*L. racemosa*	*P. adunca*	*E. caribeae*	*C. luzulae*	*C. flavus*	*C. ferax*
L. racemosa	*L. racemosa*	*C. erectus*	*P. adunca*	*E. caribeae*	*C. luzulae*	*C. flavus*	*C. ferax*
L. racemosa	*L. racemosa*	*T. cattapa*	*P. adunca*	*E. caribeae*	*C. luzulae*	*C. flavus*	*C. ferax*
L. racemosa	*L. racemosa*	*T. cattapa*	*R. mangle*	*E. caribeae*	*C. luzulae*	*C. ferax*	*C. ferax*
L. racemosa	*L. racemosa*	*T. cattapa*	*E. caribeae*	*C. luzulae*	*C. luzulae*	*C. ferax*	*C. ferax*
L. racemosa	*L. racemosa*	*T. cattapa*	*E. caribeae*	*C. luzulae*	*C. luzulae*	*C. ferax*	*C. ferax*

Solución

Como ya se ha indicado anteriormente lo primero con lo que debemos iniciar es tabulando los datos y cargarlos en el entorno de R.

Luego, como insumo para construir el gráfico de barras en función de las frecuencias absolutas, debemos crear la tabla de frecuencia del conjunto de datos a través de la orden ***table***, de esta forma, se hace sencillo elaborar el grafico, solo basta con utilizar el comando ***barplot,*** y con el propósito de darle más presentación al grafico se le añaden ciertos atributos, a través de argumentos generales, tales como ***col***, que nos permite insertar un vector constituido por los colores de cada una de las barras del gráfico, ***main*** que inserta un título al gráfico, ***xlab*** y ***ylab***, permiten insertar etiquetas a los ejes, el argumento ***font***, define la fuente de los textos a través de un valor numérico, donde 1 es normal, 2 es negrita, 3 es cursiva y 4 es cursiva y negrita, este argumento puede ser aplicado a los ejes, etiquetas, títulos y subtítulos (***font.axis***, ***font.lab***, ***font.main***, ***font.sub***, respectivamente).

```
> Vegetación<-read.csv2("Vegetación laguna.csv",header=TRUE,
encoding="latin1")
> attach(Vegetación)
> Tabla.frec<-table(Especie)
> Diagrama<-barplot(Tabla.frec,col=rainbow(9),main="Composición
vegetal Laguna salada",xlab="Especies",ylab="Frecuencias
absolutas",ylim=c(0,17),cex.main=1.5,
cex.lab=1.5,font.lab=2,font.axis=2)
```

El gráfico construido con las órdenes anteriores se muestra en la figura 2.1. Es posible insertar otros atributos de utilidad al gráfico, tales como la frecuencia que representa cada una de las barras, esto se consigue a través de la función ***text***.

En los argumentos de esta función, indicamos las coordenadas *x* e *y*, donde se ubicarán las etiquetas que representan las frecuencias de cada barra (***x = Diagrama, y = Tabla.frec + 1***), se debe indicar a su vez cuales son la etiquetas, en este caso, son las contenidas en el objeto ***Tabla.frec***, por último, utilizamos el argumento lógico ***xpd*** e indicándole el valor ***TRUE***, que permite que las etiquetas sean dispuestas fuera de las barras. Los nuevos atributos del grafico se muestran en la Figura 2.2.

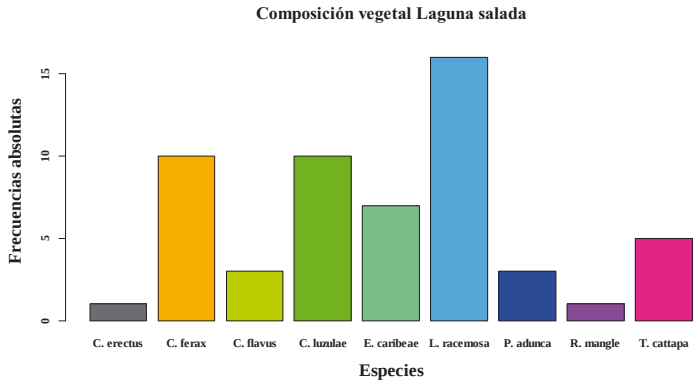


Figura 2.1. Gráfico de barras en función de las frecuencias absolutas

```
> text(Diagrama,Tabla.frec+1,Tabla.frec,xpd=TRUE)
```

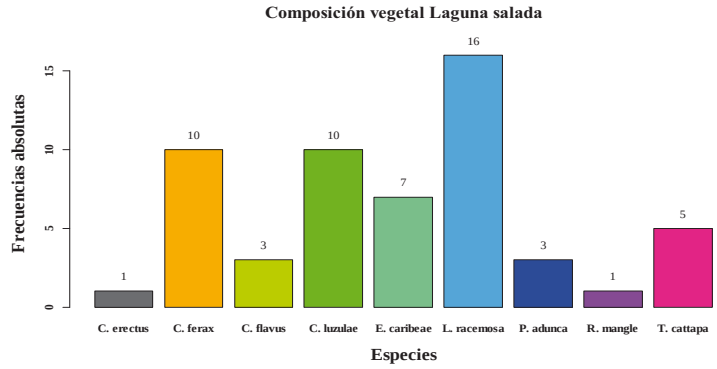


Figura 2.2. Gráfico de barras con sus etiquetas de valor

El gráfico de barras sobre las frecuencias relativas (Figura 2.3) se construye siguiendo las mismas órdenes anteriores, con la leve modificación de trabajar con la tabla de frecuencias relativas, determinadas con la orden ***prop.table***, como se muestra en la siguiente salida de resultados de R

```
> Vegetación<-read.csv2("Vegetación laguna.csv",header=TRUE,
encoding="latin1")
> attach(Vegetación)
> Tabla.frec.rel<-round(prop.table(table(Especie))*100,2)
> Diagrama1<-
barplot(Tabla.frec.rel,col=rainbow(9),main="Composición
vegetal  Laguna salada",xlab="Especies",ylab="Frecuencias
relativas (%)",ylim=c(0,30),
font.lab=2,font.axis=2)
> text(Diagrama1,Tabla.frec.rel+1,Tabla.frec.rel,xpd=TRUE)
```

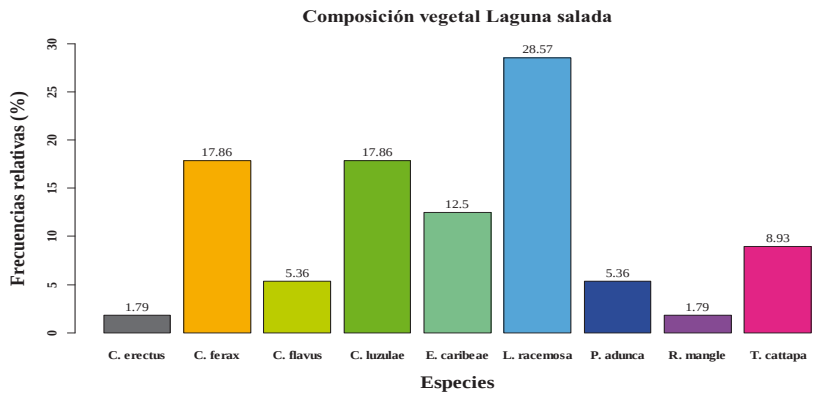


Figura 2.3. Gráfico de barras en función de las frecuencias relativas

Existen situaciones en las que un investigador le interesa realizar agrupaciones de una variable categórica en función de algunas variables cuantitativas, esto se puede realizar a través de una modificación de los diagramas de barras convencionales, denominado **diagrama de barras agrupado**. Observemos un ejemplo de esto.

Ejemplo 2.4. A continuación se muestran los datos de algunas variables fisicoquímicas medidas en un estudio de calidad de agua en la Laguna Navío Quebrado (Rosado, 2009), en ellos se registran los valores de potencial de hidrogeno (pH), Oxígeno disuelto (OD) en mg/L y demanda biológica de oxigeno (DBO) en mg/L. A partir de estos datos construya

un gráfico de barras agrupado sobre los valores medios de las variables mencionadas para cada estación de muestreo (E1, E2 y E3).

Estación	pH	OD	DBO	Estación	pH	OD	DBO
E1	7.96	5.2	4.27	E1	8.0	5.4	3.9
E1	7.85	5.6	3.38	E1	7.8	5.5	5.2
E2	7.90	5.4	5.01	E2	7.9	5.5	4.9
E2	8.51	5.2	5.00	E2	8.4	5.4	5.2
E3	8.11	5.1	5.20	E3	8.2	5.2	4.8

Solución

Iniciamos cargando los datos en la consola de R, luego los adjuntamos a la memoria con la función ***attach***, una vez realizado esto, hacemos una agrupación de las variables pH, OD y DBO a través de una lista de las medias de estas variables por estación, a través de la función ***agreggate***, en cuyo argumento se debe especificar la lista a través de la cual se realizará la agrupación con la orden ***by***. Luego en la agrupación realizada, suprimimos la primera columna para dejar en el data frame solo las columnas de interés para el grafico (***Agrupación[,-1]***), asignamos las etiquetas a las filas con la orden ***rownames***, transformamos el data frame a formato matricial, que es la tipología de objetos que mejor admite la función ***barplot***; construimos el grafico de barras (Figura 2.4) cuidando que el argumento ***beside*** tenga el valor lógico *TRUE* para obtener barras agrupadas y no apiladas, y por último insertamos una leyenda al grafico de las variables representadas por cada barra con la función ***legend***.

```
> Fisicoquimica<-read.csv2("Variables fisicoquimicas LNQ.csv",
header=TRUE,encoding="latin1")
> attach(Fisicoquimica)
> Agrupación<-aggregate(Fisicoquimica[,c("pH","OD","DBO")],
by=list(Estación),mean)
> Agrupación<-Agrupación[,-1]
> rownames(Agrupación)<-c("E1","E2","E3")
> Agrupación<-as.matrix(Agrupación)
> barplot(t(Agrupación),beside=TRUE,main="Variables
fisicoquímicas LNQ",
ylim=c(0,10),font.main=2,font.axis=2,font.lab=2,col=1:3)
>legend(locator(1),colnames(Agrupación),bty="n",fill=c("black",
"red","green"))
```

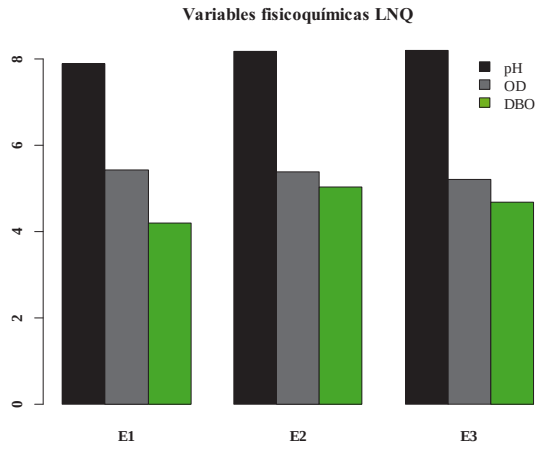


Figura 2.4. Diagrama de barras agrupadas del ejemplo 2.4

Existen situaciones en que nos gustaría elaborar gráficos de barras un poco más atractivos en cuanto a apariencia. R permite elaborar estos gráficos con barras cilíndricas. Para lograr este cometido es necesario instalar el paquete "*plotrix*" (Lemon *et al.*, 2015) y de este utilizar la función ***barp***, cuyos resultados son análogos a los de la función ***barplot***.

A continuación elaboraremos un diagrama de barras cilíndricas (Figura 2.5) para los datos del ejemplo 2.4. Los primeros pasos son exactamente igual a los mostrados en el desarrollo del ejemplo 2.4, las órdenes cambian cuando se le pida a R elaborar el gráfico. En los argumentos de esta función ***barp***, ***name.arg*** define el nombre de las etiquetas del eje *x*, con ***cylindrical=TRUE***, se especifica que las barras sean cilíndricas, con ***shadow=TRUE*** se ordena que las barras tengan sombra y con ***staxx=TRUE*** y ***staxy=TRUE***, se busca que las etiquetas del eje *x* e *y* no se pongan al mismo nivel vertical, lo que es útil cuando las etiquetas se solapan.

```
> library(plotrix)
> Fisicoquimica<-read.csv2("Variables fisicoquimicas LNQ.csv",
header=TRUE,encoding="latin1")
> attach(Fisicoquimica)
> Agrupación<-aggregate(Fisicoquimica[,c("pH","OD","DBO")],
by=list(Estación),mean)
> Agrupación1<-Agrupación[,-1]
> Agrupación2<-t(Agrupación1)
> colnames(Agrupación2)<-c("E1","E2","E3")
>barp(Agrupación2,names.arg=colnames(Agrupación2),cex.axis=1.2,col
```

```
=rainbow(3),cylindrical=TRUE,shadow=TRUE,staxx=FALSE,staxy=FALSE,l
egend.lab=c("pH","OD","DBO"),legend.pos=locator(1),xlab="",ylab=""
,border=TRUE)
> title("variables fisicoquímicas LNQ",cex=1.5,font=2)
> mtext("Estación",1,line=2.8,font=2,cex=1.5)
```

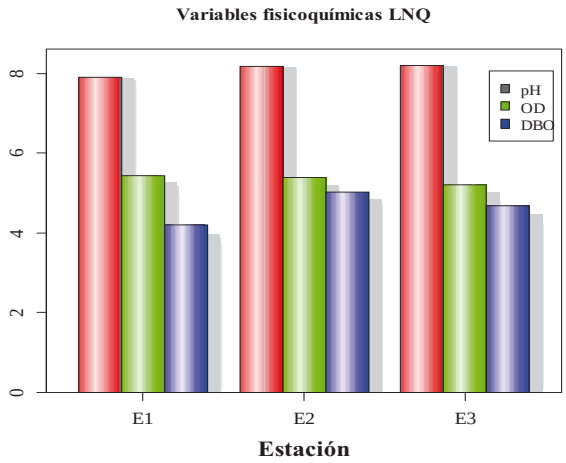


Figura 2.5. Gráfico de barras cilíndricas del ejemplo 2.4.

Gráfico de sectores

En muchas situaciones interesa presentar gráficamente la composición relativa o porcentual en vez de los valores absolutos (Villegas, 2002). El gráfico que logra esto con mucha eficiencia es el gráfico de sectores o circular. Éste representa las distribuciones de frecuencias relativas (es decir, porcentajes o proporciones) haciendo corresponder la medida de la frecuencia relativa con la medida del ángulo en grados, es decir, si el 100% de los datos son 360º de la circunferencia, a cada 1% le corresponden 3.6º; así, para obtener la medida del ángulo del sector, multiplicamos la frecuencia correspondiente por 3.6º.

El entorno de programación de R, posee herramientas (funciones) muy versátiles que nos permite construir gráficos de este tipo, en el plano y en tres dimensiones (3D), tal como ilustraremos en el siguiente ejemplo práctico.

Ejemplo 2.5. A partir de los datos de composición vegetal del sector Oeste de la Laguna salada (Ejemplo 2.3), constrúyase un gráfico de sectores de los mismos.

Solución

Iniciaremos construyendo el grafico de sectores en dos dimensiones haciendo uso de la función ***pie*** del paquete de instalación básica de R, en cuyos argumentos, ***clockwise*** indica si los sectores del grafico cambian en el sentido de las manecillas del reloj (*TRUE*) o a la inversa (*FALSE*), con ***init.angle*** se indica en que ángulo empieza el primer sector (valor por defecto de cero) y ***radius*** define el radio de la circunferencia del gráfico, es decir, el tamaño de éste.

```
> Vegetación<-read.csv2("Vegetación laguna.csv",header=TRUE,
encoding="latin1")
> attach(Vegetación)
> Tabla.frec.rel<-round(prop.table(table(Especie))*100,2)
>pie(Tabla.frec.rel,labels=paste(names(Tabla.frec.rel),Table.frec.
rel, "%"),col=rainbow(9),main=paste("Composición vegetal",
sep="\n","sector oeste Laguna salada"),cex.main=1.5,
clockwise=FALSE,init.angle=0,radius=1)
```

El gráfico de sectores resultante se muestra en la Figura 2.6.

Ahora, construyamos el grafico de sectores en 3D (Figura 2.7), usando la función ***pie3D*** del paquete "*plotrix*" (Lemon *et al.*, 2015), donde el argumentos ***explode***, indica que tan alejados se encontrarán los sectores, ***radius*** define el radio de la circunferencia del gráfico, ***height*** establece la altura del gráfico (o profundidad), y ***theta*** define el ángulo de inclinación del gráfico.

```
> library(plotrix)
> Vegetación<-read.csv2("Vegetación laguna.csv",header=TRUE,
encoding="latin1")
> attach(Vegetación)
> Tabla.frec.rel<-round(prop.table(table(Especie))*100,2)
> pie3D(Tabla.frec.rel,labels=paste(Tabla.frec.rel,"%"),
col=rainbow(9),main=paste("Composición vegetal",sep="\n","sector
Oeste Laguna salada"),cex.main=1.5,explode=0.1,radius=1,
height=0.15,theta=0.7, start=0,labelcol="black",labelcex=1.2)
>legend(locator(1),legend=names(Tabla.frec.rel),bty="n",fill=rainb
ow(9),xpd=TRUE)
```

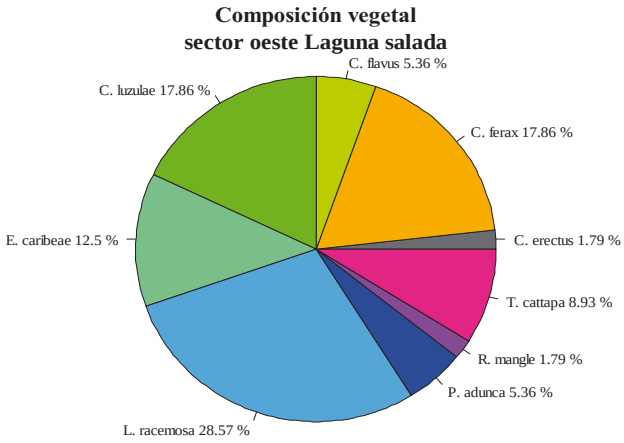


Figura 2.6. Gráfico de sectores ejemplo 2.5

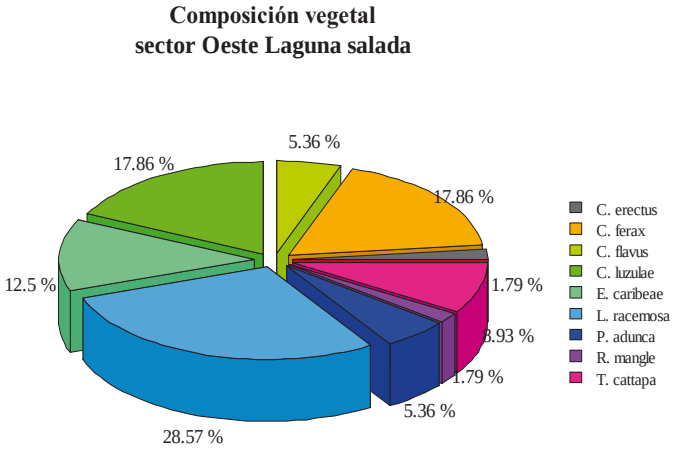


Figura 2.7. Gráfico de barras 3D del ejemplo 2.5

2.2.2. Gráficos para variables cuantitativas

Para el tratamiento gráfico de variables cuantitativas se pueden aplicar los gráficos vistos en la sección anterior, si y solo si, las variables sean de naturaleza discreta o se establezcan categorías a través sus valores. Por otro lado, cuando nos encontramos manipulando variables de naturaleza continúa, el grafico de uso más difundido para evaluar el comportamiento de la variación de un conjunto de datos, distribución y simetría, es el **histograma de frecuencias**, que consiste en una gráfica en

forma de barras que consta de dos ejes, uno horizontal, llamado eje de la variables en observación, en donde situamos la base de una serie de rectángulos o barras contiguas, es decir, que no van separadas, y que se rotula con los limites inferiores de cada clase o intervalo, excepto el último que deberá llevar también el límite superior, centradas en la marca de clase, y un eje vertical llamado eje de las frecuencias, en donde se miden las alturas que vienen dadas por las frecuencia del intervalo que representa (Córdova & Cortez, 2010).

En breve observaremos un ejemplo, donde se ilustre la construcción de un histograma de frecuencias en el entorno de programación de R.

Ejemplo 2.6. Usando los datos del nivel de presión sonora (dB) del ejemplo 2.2, tomados en un estudio de ruido ambiental en la ciudad de Cali. Construir un histograma de frecuencias de los mismos.

Solución

En R, la construcción de histogramas de frecuencias (Figura 2.8) se realiza haciendo uso de la función ***hist***, vista anteriormente, en cuyos argumentos ***breaks*** establece el número de barras que tendrá el histograma, es decir, el número de intervalos de clase, por defecto, este argumento tomara el número de clases determinadas con la *regla de Sturges*. El argumento ***freq***, que por defecto toma el valor lógico *TRUE*, indica a R que se representen las frecuencias absolutas; al asignarle el valor lógico *FLASE*, se representarán los valores de densidad de probabilidad (Frecuencias relativas). La función ***lines***, permite añadir al histograma la curva de densidad de probabilidad, y el argumento ***lwd*** de esta última función define el grosor de la línea. En la función ***hist***, también es posible utilizar los demás argumentos gráficos generales, para agregar títulos, etiquetas, definir límites, agregar color, etc. Como se observa en la siguiente salida de resultados

```
> Presión<-read.csv2("Presión Sonora.csv",header=TRUE, encoding=
"latin1")
> attach(Presión)
>  hist(Presión.Sonora,freq=TRUE,col="gold",main="Ruido ambiental
en la ciudad de Cali",xlab="Niveles de presión sonora
(dB)",ylab="Frecuencias absolutas",cex.main=1.5,font.lab=2)
```

Ahora observemos el mismo histograma, pero construido sobre las densidades de probabilidad y adjuntado la respectiva curva de densidad (Figura 2.9).

```
> Presión<-read.csv2("Presión Sonora.csv",header=TRUE, encoding=
"latin1")
> attach(Presión)
> hist(Presión.Sonora,freq=FALSE,col="gold",main="Ruido ambiental
en la ciudad de Cali",xlab="Niveles de presión sonora
(dB)",ylab="Densidades",cex.main=1.5,font.lab=2)
> lines(density(Presión.Sonora),lwd=2,col="red")
```

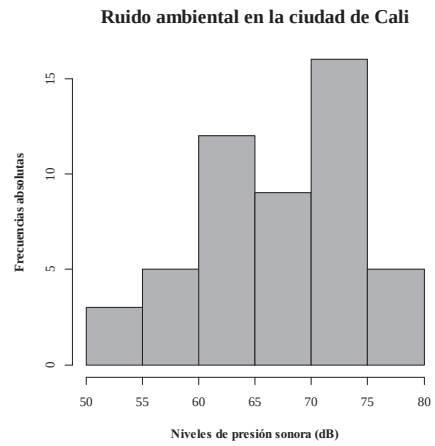


Figura 2.8. Histograma de frecuencias ejemplo 2.6

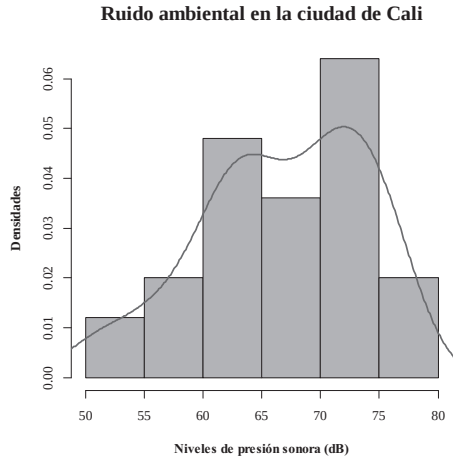


Figura 2.9. Histograma de frecuencias en función de las densidades de probabilidad

2.2.3. Gráficos de líneas y gráficos de dispersión

Los gráficos de líneas se emplean cuando es necesario representar las tendencias de una serie de datos, y éstos son numerosos o continuos; generalmente este tipo de gráficos se utilizan para cubrir períodos de minutos, horas, días, semanas, meses o años (IMEI, 2009).

Por otra parte, los gráficos de dispersión son usados con el objeto de representar la relación existente entre dos variables cuantitativas.

En R, la función ***plot*** es la función básica para la construcción de estos dos tipos de gráficos. A continuación mostraremos dos ejemplos en donde se muestra el proceso de construcción de estos gráficos en R.

Ejemplo 2.7. A continuación se muestra una porción de los datos colectados sobre las concentraciones de Amonio (mg/L) y Coliformes fecales (UFC/100mL), correspondiente a los muestreos realizados entre los meses de febrero y julio del 2012 en el marco del programa de Calidad Ambiental de Playas Turísticas, CAPT. A partir de estos datos construir un gráfico de línea para las concentraciones de amonio durante los seis meses de muestreo y un diagrama de dispersión entre los Coliformes fecales y las concentraciones de Amonio.

Mes	Amonio	Col. fecal
Febrero	43.78	339
Marzo	38.22	1901
Abril	23.58	494
Mayo	1.49	1244
Junio	88.94	761
Julio	108.2	2241

Solución

La construcción del gráfico de líneas para los datos de concentración de Amonio (Figura 2.10), se realiza siguiendo las siguientes líneas de programación, previa tabulación de los datos bajo la extensión *.csv*.

```
> Calidad.Amb<-read.csv2("Calidad Ambiental.csv",header=TRUE,
encoding="latin1")
> attach(Calidad.Amb)
> plot(Amonio,type="l",main=paste("Comportamiento concentración de
Amonio", sep="\n","durante los seis meses de
muestreo"),xlab="Meses",ylab="Concentración(mg/L)",font.lab=2,lwd=
3,col="red")
```

Observe que las concentraciones de Amonio, mostraron el valor más alto en el sexto mes de monitoreo (Julio), correspondiente a 108.2 mg/L. El argumento ***type = "l"*** usado, especifica que se trace la línea de tendencia de los datos y no solo los puntos. Para mayor información consúltese la ayuda de la función ***plot***.

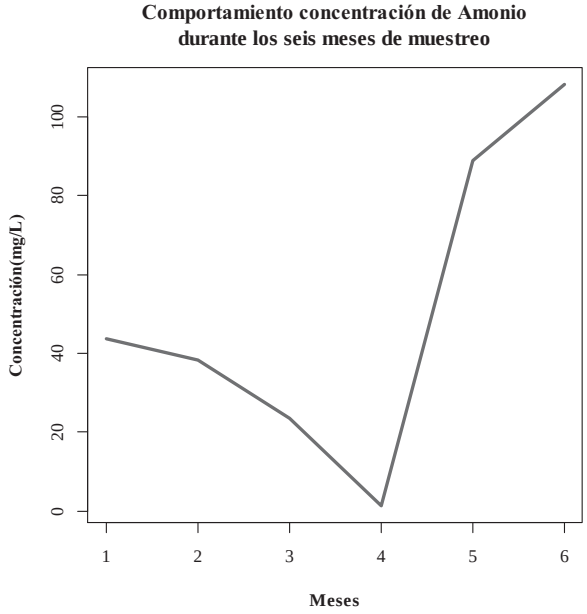


Figura 2.10. Gráfico de línea: Comportamiento del amonio en los meses de monitoreo.

El gráfico de dispersión entre los Coliformes fecales y las concentraciones de amonio se muestra en la Figura 2.11, el cual se construyó siguiendo las siguientes órdenes de programación de R

```
> Calidad.Amb<-read.csv2("Calidad Ambiental.csv",header=TRUE,
encoding="latin1")
> attach(Calidad.Amb)
> plot(Amonio,Col.fecal,main=paste("Relación entre los Coliformes
fecales y",sep="\n","las concentraciones de
Amonio"),xlab="Concentración de Amonio (mg/L)",ylab="Coliformes
fecales (UFC/100 mL)",font.lab=2,pch=19)
```

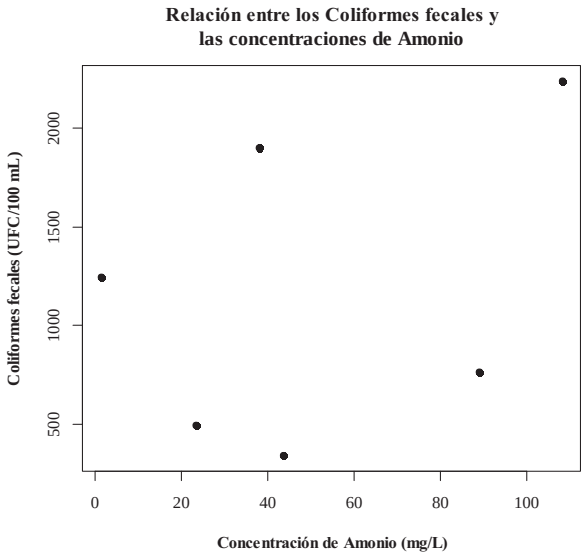


Figura 2.11. Diagrama de dispersión entre Col. Fecales y Conc. Amonio

Cabe comentar de las líneas de código anteriores, que el argumento ***pch***, permite definir la forma de los puntos, en este caso el 19 corresponde a puntos circulares con relleno sólido (consúltese la ayuda de la función ***plot***). De este gráfico, se observa que no existe una relación aparente entre las dos variables estudiadas, dada la aleatoriedad en que están dispuestos los puntos en el gráfico sin mostrar ninguna tendencia (lineal, exponencial, logarítmica, etc.).

2.3. Medidas de tendencia central

Las medidas de tendencia central son medidas de un conjunto de datos que proporcionan un valor simple y representativo que resume un gran volumen de información. Estas medidas corresponden a un valor que tiende a ubicarse en el centro del conjunto de observaciones, alrededor del cual los demás datos muéstrales se distribuyen. Las medidas de tendencia central más usualmente utilizadas en el tratamiento de datos son: La media aritmética, la mediana y la moda.

2.3.1. Media aritmética

La media aritmética ($\bar{x}$) muestra el valor central de los datos, constituyéndose en la medida de ubicación que más se utiliza. En general, es calculada sumando los valores de interés y dividiendo entre el número total de observaciones.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.4}$$

Esta medida presenta la ventaja de ser única y estar expresada en las mimas unidades en que se encuentran expresados los datos. Sin embargo, la media aritmética es muy sensible a los valores extremos, por lo que esto último representa su más importante desventaja.

2.3.2. Mediana

La mediana ($\tilde{x}$), es el valor de en medio de un grupo de números u observaciones (ordenados en forma ascendente) o el promedio aritmético de los dos valores de en medio. Geométricamente hablando, la mediana es el valor de *x* (abscisa) correspondiente a esa línea vertical que divide a un histograma en dos partes teniendo áreas iguales (Quevedo, 2006).

Visto de otra manera, la mediana se encarga de dividir la distribución de los datos en el punto medio, donde el 50% de los datos están por encima de la mediana y el otro 50% está por debajo de la misma. No existe una ecuación matemática única para el cálculo de la mediana, sin embargo esta medida se puede calcular a través de las siguientes ecuaciones, una vez los datos se hayan ordenado de menor a mayor

$$\tilde{x} = \begin{cases} \dfrac{x_{n+1}}{2} & \text{si } n \text{ es impar} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{si } n \text{ es par} \end{cases} \tag{2.5}$$

Entre las ventajas del uso de la mediana se encuentra que es una medida fácil de calcular y de entender, sus unidades se expresan en las mismas unidades de la variable que se esté estudiando y no se ve afectada por la presencia de valores extremos. Entre sus ventajas o limitaciones encontramos la poca estabilidad que muestra respecto a la media aritmética por lo tanto es poco útil en la estadística inferencial, y la necesidad de ordenar los datos previamente vuelve un poco engorroso su cálculo cuando se trata se conjunto de datos muy grandes.

2.3.3. Moda

La moda ($\hat{x}$) es un estadístico que demuestra el valor que ocurre con más frecuencia en una muestra (poniendo los datos en forma ascendente). Una distribución puede tener una moda, puede ser bimodal, o multimodal. Sin embargo, existen algunas ocasiones en que la moda no existe.

Las principales ventajas de la moda radican en que no se ve afectada por los valores extremos, no se requieren cálculos para su determinación, y sus unidades se expresan en las mismas que la variable de estudio. Por otro lado, las principales desventajas de esta medida es que no necesariamente ocurrirá como un valor central, no siempre existe y posee muy poca utilidad en la estadística inferencial.

En el ambiente de programación de R es muy sencillo determinar las medidas de tendencia central mencionadas anteriormente, todas las funciones para tal fin, se encuentran dentro del paquete de instalación de instalación básico de este software. Así, ***mean***, ***median*** son las funciones programadas en R para calcular la media aritmética y la mediana, respectivamente.

Ejemplo 2.8. Se saca una muestra aleatoria de análisis químicos de compuestos de cloruros (Cl^-) expresados en unidades de mg/L procedentes de una muestra de agua residuales. Estos análisis se hicieron usando el método de nitrato de mercurio descrito en el *Standard Methods*. Los valores de concentración de cloruros determinados, se muestran a continuación

17.2	17.1	17.0	17.1	16.9	17.0	17.1	17.0	17.3	17.2
16.9	17.0	17.1	17.3	17.2	17.4	17.1	17.1	17.0	17.1

A partir de estos datos calcular: a) la media aritmética; b) la mediana, y c) la moda.

Solución

Iniciamos el con el cálculo de la media aritmética ($\overline{x}$), como se muestra a continuación

$$\bar{x} = \frac{1}{2}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{17.2 + 17.1 + 17.0 + \cdots + 17.1}{20}$$

$$\bar{x} = 17.105 \text{ mg/L}$$

Para calcular la mediana ($\tilde{x}$), primeramente ordenamos los datos de forma ascendente (de menor a mayor)

16.9 16.9 17.0 17.0 17.0 17.0 17.0 17.1 17.1 17.1 17.1 17.1 17.1 17.2 17.2 17.2 17.2 17.2 17.3 17.4

Como *n* es par, se deben promediar los valores de las posiciones $x_{\frac{n}{2}}$ y $x_{\frac{n}{2}+1}$,

es decir, los valores de las posiciones $x_{10} = 17.1$ mg/L y $x_{11} = 17.1$ mg/L, así,

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{10} + x_{11}}{2}$$

$$\tilde{x} = \frac{17.1 + 17.1}{2}$$

$$\tilde{x} = 17.1 \text{ mg/L}$$

La moda, al no requerir ninguna especie de cálculo, se determina ordenando los datos de forma ascendente y observando cuál de ellos es el que presenta mayor repitencia o frecuencia, en nuestro caso en particular, este valor corresponde a 17.1 mg/L.

Ahora veamos la solución del este ejercicio, a través de ordenes en el entorno de programación de R, utilizando las funciones mencionadas anteriormente.

Por facilidad metodológica, al tratarse de pocos datos, estos se cargarán en R a través de la creación de un vector de datos, empleando la función *c* (concatenación), así,

```
Cloruro<c(17.2,17.1,17.0,17.1,16.9,17.0,17.1,17.0,17.3,17.2,1
6.9, 17.0,17.1,17.2,17.2,17.4,17.1,17.2,17.0,17.1)
```

Como paso siguiente se calculan las medidas pedidas

```
> mean(Cloruro)
[1] 17.105
```

Este valor muestra que el valor medio del conjunto de observaciones de las concentraciones de Cl^- en muestras de aguas residuales es de 17.105 mg/L.

```
> median(Cloruro)
[1] 17.1
```

La mediana nos muestra que el 50% de las concentraciones de Cl^- son menores que 17.01 mg/l y que el 50% restante son mayores que este valor

En el caso de la moda, en la literatura revisada, no se encontró una función específica que permita el cálculo de este indicador en R. Generalmente, se sugieren o exponen procedimientos para la construcción de funciones definidas por el usuario para el cálculo de este, es decir, construir funciones que no están programadas en el paquete de instalación básica de R.

2.4. Medidas de posición: Cuantiles

Estas medidas de posición o localización, permiten dividir en partes iguales un conjunto ordenado de datos (de manera ascendente), de manera que los nuevos grupos formados posean el mismo número de observaciones. De esta forma, se definen los cuantiles de orden *k* como los valores de la variable estudiada, que previo ordenamiento de la misma de forma ascendente (de menor a mayor), la dividen con la misma frecuencia de observaciones, por lo tanto, existirán $k - 1$ cuantiles de orden *k* (Guisande *et al.*, 2011).

De acuerdo a lo anterior, el primer cuantil de orden *k* deja a su izquierda la fracción $1/k$ de frecuencias de observaciones. El segundo cuantil de orden *k* deja a su izquierda la fracción $2/k$ de frecuencia de observaciones, y de forma general el *j*-ésimo cuantil de orden *k* deja a su izquierda la fracción j/k de frecuencia de observaciones (Guisande *et al.*, 2011).

Los cuantiles surgen por las limitaciones que presentan las medidas de tendencia central cuando se desean realizar análisis respecto a la posición

de un valor específico en relación con el resto de los datos, y la importancia de estas medidas radica en que sintetizan las distribuciones de frecuencias e indican que porcentaje de datos, dentro de una distribución, hay antes o después de un valor determinado. Los cuantiles más utilizados en estadística son los **cuartiles**, **deciles** y **percentiles**.

2.4.1. Cuartiles

Dividen el conjunto de datos en cuatro partes iguales, de manera que las observaciones del tercer cuartil (Q_3) constituyen el cuarto superior del conjunto de datos, el segundo cuartil (Q_2) es idéntico a la mediana y el primer cuartil (Q_1) separa el cuarto inferior de los tres cuartos superiores (Devore, 2008).

2.4.2. Deciles

De manera similar a los cuartiles, los deciles permiten hacer una división de un conjunto ordenado de datos en diez partes iguales con el mismo número de observaciones, de manera que el quinto decil (D_5) resulta ser igual al segundo cuartil, y a su vez igual a la mediana del conjunto de datos.

2.4.3. Percentiles

Al igual que en las medidas anteriores, si deseamos dividir más finamente un conjunto de datos lo podemos hacer a través de percentiles, es decir, en cien partes iguales, donde el percentil cincuenta (P_{50}), es igual a la mediana, al segundo cuartil y al quinto decil.

En general, existen muchos métodos para el cálculo de los cuantiles. Sin embargo, los softwares estadísticos de más amplio uso se basan generalmente en nueve de estos (Hyndman & Fan, 1996), donde R, como muestra de su gran versatilidad, brinda la posibilidad de trabajar con cualquiera de ellos de acuerdo a nuestras exigencias como usuarios. No obstante, para no extendernos demasiado en la discusión de estos métodos de cálculo, solo trataremos el método con el que se encuentra configurado R por defecto, el cual se basa en la siguiente expresión

$$\hat{Q}_p = x_{\lfloor h \rfloor} + \left(h - \lfloor h \rfloor\right)\left(x_{\lfloor h \rfloor+1} - x_{\lfloor h \rfloor}\right) \qquad (2.6$$

con

$$h = (n-1)\,p + 1$$

y

$$p = \frac{j}{k}$$

Donde $\lfloor h \rfloor$ (denominada función piso de *h*), denota el mayor entero no mayor que *h*.

En R, los cuantiles, se pueden determinar a través de la función ***quantile***, en cuyo argumento se especifica el objeto ***x*** al cual se le hallaran estas medidas, seguidas del argumento ***probs***, donde se define la fracción de frecuencias de observaciones que se dejara a la izquierda.

```
quantile(x, probs = c(0.01,...,1))
```

Ejemplo 2.9. Se tienen los datos de Demanda Química de Oxigeno (DQO) en mg/L, en el efluente de una planta de tratamiento anaeróbico para agua residual tipo UASB (*Upflow Anaerobic Sludges Blanket*). La serie de datos ordenados en forma creciente se presenta a continuación (Vargas, 2007):

110 126 135 154 152 155 160 181 191 191 200 208 216 257 260 312 315 320 320 (mg/L).

A partir de los datos anteriores, determinar: a) Los cuartiles; b) D_1 y D_8 ; c) P_{35} y P_{70} .

Solución

Teniendo en cuenta que los datos ya se encuentran dispuestos en orden creciente, donde *n = 19*, calculamos las medidas pedidas a través de la ecuación (2.6).

El orden de los cuartiles es $k = 4$, por lo tanto, el valor de *p* para Q_1, Q_2 y Q_3 es respectivamente, 0.25, 0.50, 0.75. De esta forma, para el primer cuartil se tiene que

$$h = (19-1)0.25 + 1 = 5.5$$

y

$$\hat{Q}_{0.25} = 152 + (5.5 - 5)(155 - 152)$$

$$\hat{Q}_{0.25} = 153.5 = Q_1$$

Es decir, el 25% de los datos sobre la concentración de DQO, del efluente de una planta de tratamiento anaerobio de aguas residuales tipo UASB, son menores que 153.5 mg/L, y el 75% de los datos restantes superan este valor.

Para el segundo cuartil tenemos que

$$h = (19 - 1)0.50 + 1 = 10$$

y

$$\hat{Q}_{0.50} = 191 + (10 - 10)(200 - 191)$$

$$\hat{Q}_{0.50} = 191 = Q_2$$

Es decir, el 50% de los datos sobre la concentración de DQO, del efluente de una planta de tratamiento anaerobio de aguas residuales tipo UASB, son menores que 191 mg/L, y el 50% de los datos restantes superan este valor.

Para el tercer cuartil tenemos que

$$h = (19 - 1)0.75 + 1 = 14.5$$

y

$$\hat{Q}_{0.75} = 257 + (14.5 - 14)(260 - 257)$$

$$\hat{Q}_{0.75} = 258.5 = Q_3$$

Es decir, el 75% de los datos sobre la concentración de DQO, del efluente de una planta de tratamiento anaerobio de aguas residuales tipo UASB, son menores que 285.5 mg/L, y el 25% de los datos restantes superan este valor.

El orden *k* de los deciles es 10, de allí que el valor de *p* para los deciles que se desean hallar sean respectivamente 0.10 y 0.80. Así, para D_1, tenemos que

$$h=(19-1)0.10+1=2.8$$

y

$$\hat{Q}_{0.10}=126+(2.8-2)(135-126)$$

$$\hat{Q}_{0.10}=133.2=D_1$$

Concluyéndose que el 10% de los datos sobre la concentración de DQO, del efluente de una planta de tratamiento anaerobio de aguas residuales tipo UASB, son menores que 133.2 mg/L, y el 90% de los datos restantes superan este valor.

De forma análoga, los cálculos para hallar D_8, serian

$$h=(19-1)0.80+1=15.4$$

y

$$\hat{Q}_{0.80}=260+(15.4-15)(312-260)$$

$$\hat{Q}_{0.80}=280.8=D_8$$

Aquí, el 80% de los datos sobre la concentración de DQO, del efluente de una planta de tratamiento anaerobio de aguas residuales tipo UASB, son menores que 280.8 mg/L, y el 20% de los datos restantes superan este valor.

Por último, para el cálculo de los percentiles, debemos tener en cuenta que el orden *k* de estos es 100, por lo tanto, los valores de *p* para P_{35} y P_{70} son, respectivamente, 0.35 y 0.70. El cálculo de estos percentiles, iniciando por P_{35} , se muestra a continuación

$$h=(19-1)0.35+1=7.3$$

y

$$\hat{Q}_{0.35}=160+(7.3-7)(181-160)$$

$$\hat{Q}_{0.35}=166.3=P_{35}$$

Es decir, que el 35% de los datos sobre la concentración de DQO, del efluente de una planta de tratamiento anaerobio de aguas residuales tipo UASB, son menores que 166.3 mg/L, y el 65% de los datos restantes superan este valor.

Para el cálculo de P_{70} , tenemos que

$$h = (19-1)0.70+1 = 13.6$$

y

$$\hat{Q}_{0.70} = 216 + (13.6-13)(257-216)$$
$$\hat{Q}_{0.70} = 240.6 = P_{70}$$

Aquí, el 70% de los datos sobre la concentración de DQO, del efluente de una planta de tratamiento anaerobio de aguas residuales tipo UASB, son menores que 240.6 mg/L, y el 30% de los datos restantes superan este valor.

Ahora veamos la salida de resultados de R para el cálculo de cuartiles, deciles y percentiles, luego de la aplicación de la función ***quantile***

```
> DQO<-c(110,126,135,145,152,155,160,181,191,191,200,208,216, 257,
260,312,315,320,320)
> Cuartiles<-quantile(DQO,probs=c(0.25,0.50,0.75))
> Cuartiles
  25%   50%   75%
153.5 191.0 258.5
> Deciles<-quantile(DQO,probs=c(0.10,0.50,0.80))
> Deciles
  10%   50%   80%
133.2 191.0 280.8
> Percentiles<-quantile(DQO,probs=c(0.35,0.50,0.70))
> Percentiles
  35%   50%   70%
166.3 191.0 240.6
```

2.4.4. Gráficos basados en los cuartiles: Gráfico de caja

Un método común para la detección de valores atípicos es el llamado diagrama de caja o boxplot. Este método es válido para cualquier

conjunto de datos, independientemente de la forma de su distribución de frecuencias. Este gráfico, se ha utilizado en años recientes con éxito para describir varias de las características más prominentes de un conjunto de datos, entre las que se incluyen: 1) el centro, 2) la dispersión de los datos, 3) el grado y naturaleza de cualquier alejamiento de la simetría y 4) la identificación de observaciones "extremas" inusualmente apartadas del cuerpo principal de los datos (Devore, 2008). La construcción de estos gráficos se basa en el primer, segundo y tercer cuartil, y las observaciones mínima y máxima del conjunto de datos, por ser estas medidas robustas a la presencia de valores extremos.

La construcción de gráficos de caja simple, se realiza ordenando las observaciones de la más pequeña a la más grande, se ubica el primer, segundo y tercer cuartil de los datos, se dibuja una caja rectangular desde la posición del primer al tercer cuartil, cuya longitud corresponde a la diferencia de estos, comúnmente denominada **Rango intercuartílico** (IQR, por sus siglas en ingles), finalmente se dibujan líneas desde los extremos de la caja hacia los valores mínimo y máximo, estas líneas representan los bigotes o alambres (Figura 2.12).

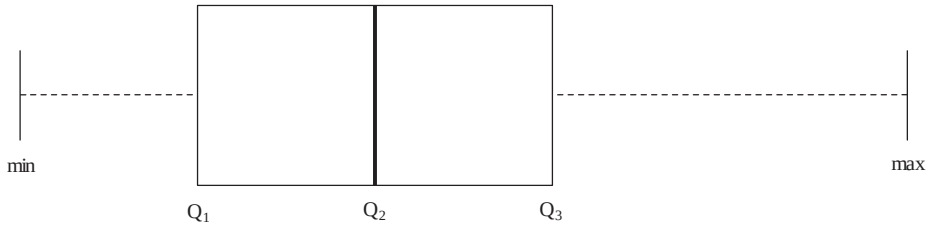


Figura 2.12. Gráfico de caja y bigotes (boxplot)

Una gráfica de caja puede ser embellecida para indicar explícitamente la presencia de valores apartados, que pueden dar evidencia de que los datos no se encuentran normalmente distribuidos (una curva en forma de campana), impidiendo la aplicación de ciertas técnicas de estadística inferencial. Cuando este es el objetivo, las cajas se construyen como en el caso del gráfico de caja simple, adicionando los límites generados por cercos internos y externos, utilizando las siguientes expresiones (Vargas, 2007):

Cerco interno inferior $\rightarrow Q_1 - 1.5 * IQR$
Cerco interno superior $\rightarrow Q_3 + 1.5 * IQR$
Cerco extremo inferior $\rightarrow Q_1 - 3 * IQR$
Cerco extremo superior $\rightarrow Q_3 + 3 * IQR$

Los valores que se encuentran entre las distancias $1.5*IQR$ y $3*IQR$, son denominados **valores atípicos**; y aquellos que se encuentren por encima de una distancia de $3*IQR$, son denominados **valores atípicos extremos** u **"*outlier*".** En caso de no existir datos en esta región se considera que no hay datos outlier en el conjunto de datos (Figura 2.13)

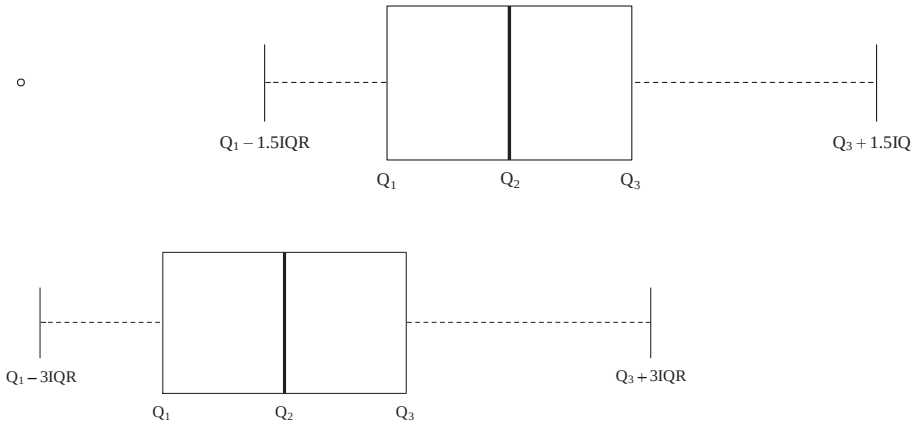


Figura 2.13. Gráficos de caja mostrando valores atípicos y atípicos extremos.

Según Vargas (2007), entre las interpretaciones que se le pueden hacer a un gráfico de cajas sobresalen las siguientes:

- La longitud de la caja refleja el grado de dispersión de los datos. A mayor longitud, mayor dispersión.
- La línea que divide la caja principal es el valor de la mediana. Si esta se encuentra en el punto medio de la caja o cercano a este, indica simetría de los datos con relación a la mediana. También indica homogeneidad en la distribución de los datos.
- La dispersión está dada tanto por la longitud de la caja, como por la distancia entre los extremos de los bigotes.
- El sesgo se observa en la desviación que exista entre la línea de la mediana en relación con el centro de la caja, y también la relación entre las longitudes de los bigotes.
- Las colas de la distribución se pueden apreciar por la longitud de los bigotes, y también por las observaciones que se marcan explícitamente.

Otra notable aplicación que proporciona la construcción de gráficos de caja, es mostrar similitudes y diferencias entre grupos sobre la medición

de la misma variable, esto se hace con la construcción simultánea de gráficos de caja para cada grupo, donde la evidencia de solapamiento entre las cajas, sugieren que los grupos estudiados son iguales, y posiblemente provenientes de la misma población (Figura 2.14).

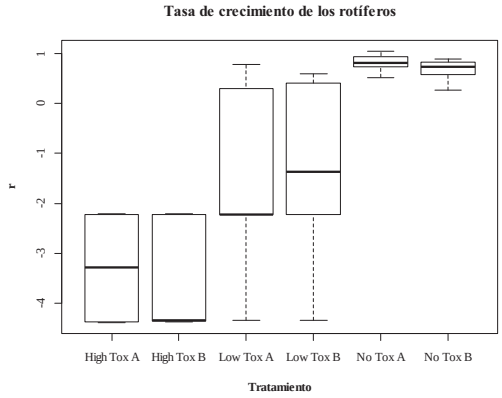


Figura 2.14. Gráficos de caja comparativos.

A los gráficos de caja, se le pueden agregar muescas en la posición de la mediana, con el propósito de facilitar las comparaciones entre grupos. La anchura de las muescas representa un intervalo de confianza aproximado para la mediana a una confiabilidad del 95% (Pérez, 2004). Así, con el objeto de comparar los grupos sobre la medición de una misma variable, la atención del investigador se debe centra en observar el solapamiento de las muescas de cada una de las cajas (Figura 2.15)

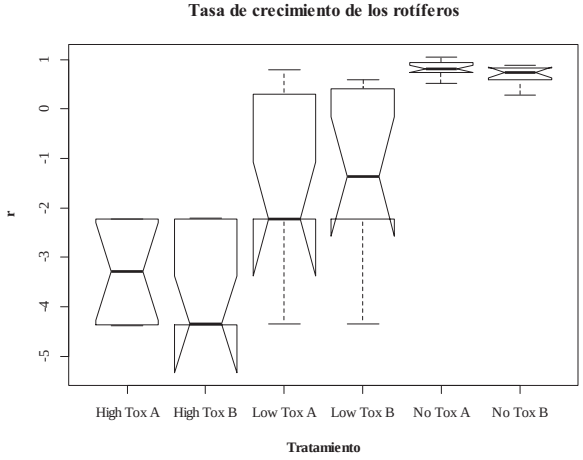


Figura 2.15. Gráficos de caja comparativos con muescas

En R, la función ***boxplot*** permite construir gráficos de caja, en su primer argumento se puede colocar el nombre de un objeto (*x*), cuando el objetivo es solo representar un solo grupo, o una fórmula de la forma ***x~gorup*** cuando se quiere realizar comparaciones entre grupos, donde ***x*** es la variable sobre la cual se efectúan las mediciones y ***group,*** son los diferentes grupos que se desean representar. En el caso en que deseemos disponer los gráficos de manera horizontal, debemos agregar el argumento ***horizontal*** y asignarle el valor lógico ***TRUE***, con ***col*** se puede asignar color a los gráficos y con ***notch,*** se define si se grafican las muescas, asignándole el valor ***TRUE***.

A continuación, se mostrará un ejemplo donde se ejemplifique mejor lo anterior.

Ejemplo 2.10. Se coleccionaron las concentraciones atmosféricas de SO_2 (en ppm) provenientes de 5 muestreadores localizados a diferentes distancias (aleatoriamente asignadas), de una fuente industrial emisora. Con base en estos datos, construir un gráfico de caja comparativo y realizar algunas inferencias

A	B	C	D	E
500	550	648	720	890
510	540	630	700	900
490	500	620	710	920
530	520	600	736	880

Solución

En el ambiente de programación de R, resulta muy sencillo construir el grafico que se está pidiendo, solo basta con cargar los datos y aplicar la función ***boxplot*** con sus respectivos argumentos. Las líneas de código introducidas en R y el grafico generado se muestran a continuación (Figura 2.16).

```
> Conc<-read.csv2("Concentraciones de SO2.csv",header=TRUE,
encoding="latin1")
> attach(Conc)
> boxplot(Concentración~Muestreador,notch=TRUE,
main="Concentraciones de
SO2",xlab="Muestreadores",ylab="ppm")
```

En la gráfica, se puede apreciar que las concentraciones medias de dióxido de azufre (ppm), son iguales en los muestreadores A y B (solapamiento de las cajas) y estas a su vez, son muy distintas respecto a los muestreadores C, D y E, pues no se observa coincidencia en la posición de las cajas, especialmente de la mediana o las muescas.

Por lo anterior, es evidente la importancia potencial de este tipo de gráficos en el tratamiento de datos, que para el caso de este ejemplo, ayuda a identificar qué puntos están siendo mayormente afectados por las emisiones de SO_2 provenientes de una fuente industrial, lo que ayudaría al investigador e incluso a la misma fuente a identificar donde se deben tomar los respectivos correctivos de control ambiental.

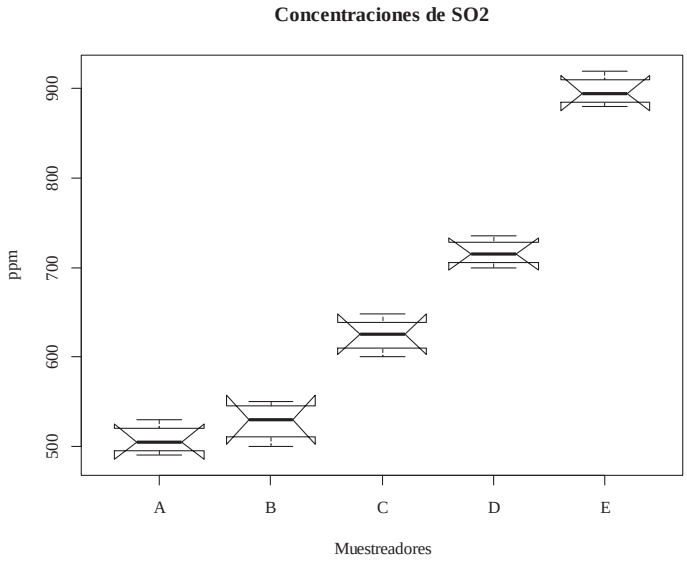


Figura 2.16. Gráfico de caja ejemplo 2.10.

2.5. Medidas de variabilidad o dispersión

Para realizar una descripción de un conjunto de datos no es suficiente contar solo con las medidas de tendencia central, es necesario también obtener información acerca de la dispersión o variabilidad de los datos. Por ello, es importante obtener estas medidas al tiempo que se persigue determinar la centralidad de las observaciones.

Las medidas de variabilidad o dispersión permiten generar criterios sobre el grado de homogeneidad o heterogeneidad del conjunto de datos que se

está analizando, en relación con una medida de centralidad (Vargas, 2007), como la media aritmética. En el tratamiento de datos las medidas de variabilidad que muestran mayor utilidad y aceptación por la mayoría de los investigadores son: la varianza, la desviación estándar y el coeficiente de variación. A continuación se brindará una breve descripción de cada una de estas medidas

2.5.1. Varianza

La varianza de un conjunto de datos es una medida estadística que permite medir las variaciones del conjunto de datos respecto a su media aritmética y es definida como la media aritmética de los cuadrados de las desviaciones de cada dato a la media aritmética (Vargas, 2007; Martínez, 2011), matemáticamente la varianza se calcula como se indica en (2.7):

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1} \tag{2.7}$$

El denominador $n - 1$, se utiliza cuando la varianza se calcula a partir de los datos muéstrales; en caso de contar con los datos de la población, se sustituye $n - 1$ por n, y de ese modo se calcula la varianza poblacional (σ^2).

En materia de interpretación, cuanto menor sea el valor de la varianza, menor es el grado de variación o heterogeneidad del conjunto de datos respecto a su media aritmética. Sin embargo, una desventaja muy marcada de esta medida estadística es la expresión de los resultados como unidades al cuadrado, lo que dificulta su interpretación.

En R, la varianza se determina a través de la función ***var***, en cuyo argumento se debe especificar la variable a la cual se desea calcular este indicador

2.5.2. Desviación estándar

La definición de la desviación estándar es simple, no es más que la raíz cuadrada de la varianza, y nació como una forma de superar la limitación

que presentaba esta de expresar sus unidades al cuadrado. Matemáticamente la desviación estándar muestral se expresa como sigue

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2.8}$$

y la desviación estándar poblacional

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{2.9}$$

En R, la desviación estándar se determina a partir de la función ***sd***, especificando en el argumento la variable estudiada

2.5.3. Coeficiente de variación

Es una medida estadística que permite valorar en forma relativa (porcentual) la forma en que se dispersan los datos respecto a su media aritmética (Morales, 2008; Sáez, 2012; Di Rienzo *et al.,* 2005). La principal ventaja del coeficiente de variación es que no tiene unidades de medida, lo que hace más fácil su interpretación.

Otra de las ventajas de este indicador es que permite comparar la homogeneidad de dos o más conjunto de datos así se encuentren expresados en diferentes unidades. Por ejemplo, si se comparan dos índices para medir nivel de ataque de pulgones y ambos están basados en técnicas completamente diferentes, que dan puntajes cuyas unidades de medida son distintas, se dirá que el índice que tenga menor coeficiente de variación es el menos variable.

Matemáticamente el coeficiente de variación se calcula a través de la siguiente expresión

$$CV = \frac{s}{\bar{x}} * 100 \tag{2.10}$$

Como una guía para la interpretación de esta medida se pueden tomar los siguientes criterios propuestos por Vargas (2007):

- Si $CV \leq 30\%$, entonces el conjunto de datos es poco variable u homogéneo con relación a la media.
- Si $30\% < CV \leq 70\%$, entonces el conjunto de datos es variable o heterogéneo con relación a la media.
- Si $CV > 70\%$, entonces el conjunto de datos es muy variable o muy heterogéneo con relación a la media.

En el ambiente de programación de R, para calcular el coeficiente de variación se usa este como una calculadora o cualquier hoja de cálculo, es decir, utilizando la expresión ***(sd()/mean())*100***, ubicando dentro de los paréntesis la variable a la cual se le está realizando el análisis.

Ejemplo 2.11. A partir de los datos consignados en el Ejemplo 2.8, calcular las medidas de variabilidad vistas anteriormente e interpretarlas.

Solución

Como se indicó anteriormente la varianza (s^2), se calcula a partir de (2.7)

$$s^2 = \frac{(17.2-17.105)^2 + (17.1-17.105)^2 + \cdots + (17.1-17.105)^2}{20-1}$$

$$s^2 = 0.17 \text{ mg}^2/\text{L}^2$$

$$s = \sqrt{0.17}$$

$$s = 0.13 \text{ mg/L}$$

$$CV = \frac{0.13}{17.105} * 100$$

$$CV = 0.75\%$$

En R, construimos el vector de datos y se aplican cada una de las funciones descritas anteriormente

```
> Cloruro<-c(17.2,17.1,17.0,17.1,16.9,17.0,17.1,17.0,17.3,17.2,
16.9,17.0,17.1,17.2,17.2,17.4,17.1,17.2,17.0,17.1)
> var(Cloruro)
[1] 0.01628947
> sd(Cloruro)
[1] 0.1276302
> (sd(Cloruro)/mean(Cloruro))*100
[1] 0.7461574
```

Estos resultados muestran que en promedio el conjunto de datos presenta un alejamiento de su media aritmética de 0.13 mg/L, en decir, presenta una variabilidad muy pequeña. Este resultado lo corrobora el coeficiente de variación que muestra que las observaciones presentan una variabilidad del 0.75 %, es decir, los datos son poco variables u homogéneos.

2.6. Medidas de forma

Las medidas de forma comparan la forma que tiene la representación gráfica de los datos, bien sea el histograma o el diagrama de barras de la distribución, con una situación ideal en la que los datos se reparten en igual medida a la derecha y a la izquierda de la media (Sáez, 2012). Existen dos medidas estadísticas para determinar la forma de una distribución: El coeficiente asimetría y la kurtosis

2.6.1. Coeficiente de asimetría

Esa situación en la que los datos están repartidos de igual forma a uno y otro lado de la media se conoce como **simetría**, y se dice en ese caso que la distribución de los datos es simétrica. Además, su mediana, su moda y su media coinciden. Por otra parte, se dice que una distribución es **asimétrica a la derecha** si las frecuencias (absolutas o relativas) descienden más lentamente por la derecha que por la izquierda. Si las frecuencias descienden más lentamente por la izquierda que por la derecha diremos que la distribución **es asimétrica a la izquierda** (Figura 2.16).

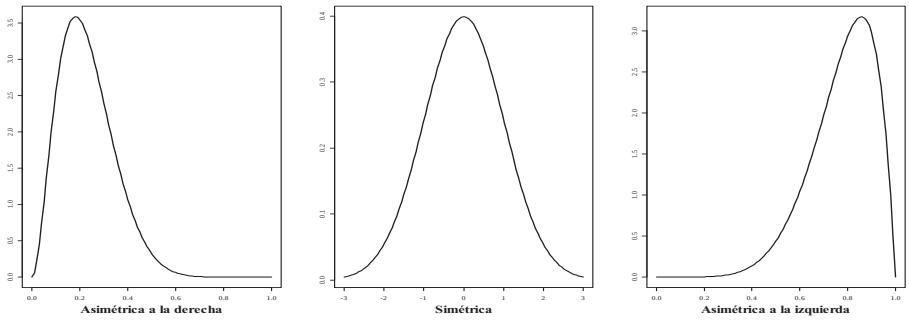


Figura 2.16. Diferentes tipos de asimetría de una distribución de observaciones.

Para calcular la simetría de un conjunto de observaciones, la medida estadística de uso más difundido es el coeficiente de asimetría de Fisher, el cual es calculado a través de la siguiente ecuación

$$As = \frac{m_3}{s^3} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{ns^3} \tag{2.11}$$

Obsérvese que para evitar problemas de discrepancias de unidades y hacer que la medida sea escalar, y por lo tanto relativa, dividimos por el cubo de su desviación típica. De esta forma podemos valorar si unos datos son más o menos simétricos que otros, aunque no estén medidos en la misma unidad de medida (Sáez, 2012). La interpretación de este coeficiente de asimetría es la siguiente (Martínez, 2010; Molina & Rodrigo, 2010):

- Si $As > 0$, los datos presentan asimetría a la derecha, asimetría positiva.
- Si $As = 0$, los datos son simétricos o presentan una distribución normal.
- Si $As < 0$, los datos presentan asimetría a la izquierda o asimetría negativa.

En R, el cálculo del coeficiente de asimetría es simple, solo basta con aplicar a nuestra variable de estudio la función ***skewness***. Para la aplicación de esta función es necesario el llamado previo de los paquetes "*class"* (Ripley & Venables, 2015) y "*e1071"* (Meyer *et al.*, 2014).

2.6.2. Apuntamiento o kurtosis

El apuntamiento o kurtosis de una distribución de frecuencias no tiene un referente natural como en el caso de la simetría, sino que se sustenta en la comparación respecto a una distribución de referencia, en concreto, la distribución normal o campana de Gauss. En consecuencia, su obtención sólo tendrá sentido en variables cuya distribución de frecuencias sea similar a la de la curva normal, en la práctica ello se reduce, básicamente, a que sea unimodal y más o menos simétrica. Existen diversas formas de estimar la kurtosis de una distribución de datos, pero una de las más usadas es el coeficiente de apuntamiento de Fisher que se expresa matemáticamente como se muestra a continuación

$$Ap = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{ns^4} - 3 \qquad (2.12)$$

El valor de este coeficiente para la distribución normal será igual a 0, o sea que cualquier distribución para la que se obtenga un valor de *Ap* igual o próximo a 0 significará que su nivel de apuntamiento es como el de la distribución normal (mesocúrtica). Valores mayores que 0, expresan que la distribución es leptocúrtica, mientras que si son menores que 0 ponen de manifiesto que la distribución es platicúrtica (Figura 2.17).

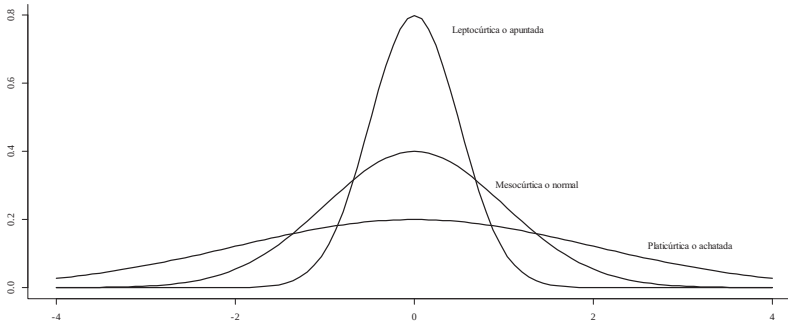


Figura 2.17. Tipos de apuntamiento o kurtosis.

En R, el coeficiente de apuntamiento de *Fisher* o kurtosis se determina con la orden ***kurtosis***, incluyendo dentro de sus paréntesis la variable estudiada. Al igual que en la función ***skewness***, es necesario cargar previamente los paquetes "*class*" (Ripley & Venables, 2015) y "*e1071*" (Meyer *et al.*, 2015).

Ejemplo 2.12. A partir de los datos consignados en el Ejemplo 2.8, calcular las medidas de forma (Asimetría y kurtosis) vistas anteriormente e interpretarlas.

Solución

En coeficiente de asimetría (*As*), se determina con la ecuación (2.11) como se indicó anteriormente

$$As = \frac{(17.2-17.105)^3 + (17.1-17.105)^3 + \cdots + (17.1-17.105)^3}{(20)(0.13)^3}$$

$$As = 0.345$$

Por su parte la kurtosis es determinada con la ecuación (2.12)

$$Ap = \frac{(17.2-17.105)^4 + (17.1-17.105)^4 + \cdots + (17.1-17.105)^4}{(20)(0.13)^4} - 3$$

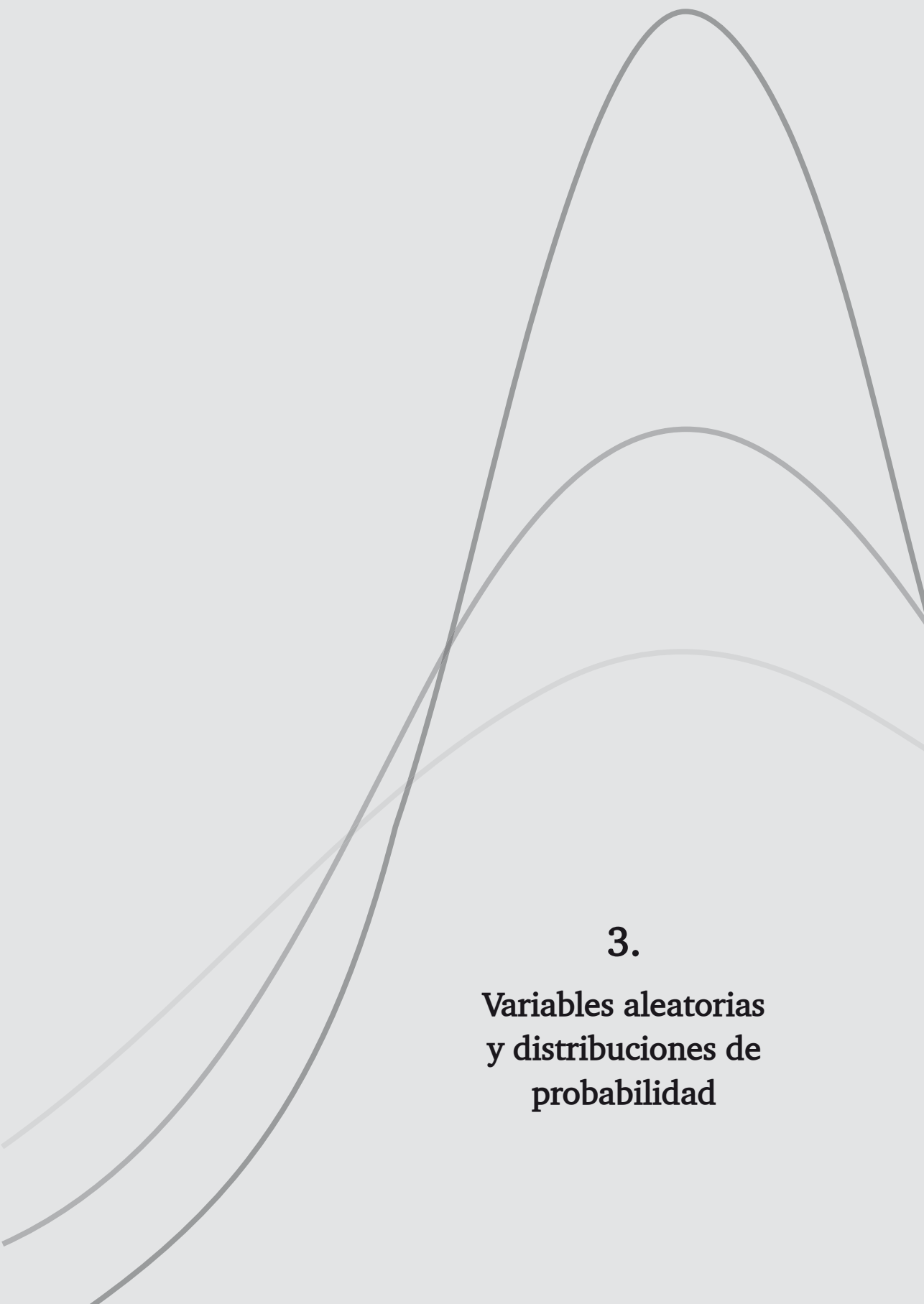
$$Ap = -0.444$$

Dado que $As > 0$, el conjunto de observaciones de las medidas de concentración del ion cloruro (mg/L) efectuadas en muestras de cierto tipo de aguas residuales presentan una asimetría a la derecha o positiva, es decir, las observaciones tienden a agruparse en la parte izquierda de la distribución.

En el caso de la kurtosis, por tomar un valor menor que cero ($Ap < 0$), muestra evidencia de que los datos presentan una distribución platicúrtica o achatada, es decir, presenta menos apuntamiento que la curva de distribución normal, o hay poca concentración de las observaciones alrededor de la media aritmética.

Los códigos para calcular las medidas o indicadores de forma en R se muestran a continuación. Adviértase que puede existir una leve diferencia en las unidades decimales respecto a las medidas calculadas manualmente, esto se atribuye principalmente al redondeo de las unidades.

```
> Cloruro<-
c(17.2,17.1,17.0,17.1,16.9,17.0,17.1,17.0,17.3,17.2,
16.9,17.0,17.1,17.2,17.2,17.4,17.1,17.2,17.0,17.1)
> library(class)
> library(e1071)
> skewness(Cloruro)
[1] 0.3452333
> kurtosis(Cloruro)
[1] -0.443637
```



3.

Variables aleatorias y distribuciones de probabilidad

3.1. Concepto de variable aleatoria

En las secciones anteriores hemos trabajado con conjuntos de datos ya proporcionados de alguna manera sin interrogarnos sobre la forma en que esos datos fueron obtenidos. Estadísticamente hablando, al proceso mediante el cual se recolecta datos u observaciones se denomina **experimento**, y este puede clasificarse en dos tipos: los experimentos determinísticos, definidos como aquellos que al ser repetidos bajo las mismas condiciones se obtienen los mismos resultados, es decir, los resultados son predecibles y, **los experimentos aleatorios**, definidos como aquellos que al ser repetidos bajo condiciones análogas se obtienen diferentes resultados, es decir, dentro de los posibles resultados, el resultado del experimento es impredecible (Febrero *et al.*, 2008).

En materia de inferencia estadística, son de mayor utilidad los experimentos aleatorios, dada la naturaleza aleatoria de la mayoría de los fenómenos que se estudian en la realidad, por ejemplo, el número de bacterias coliformes que se pueden encontrar en una muestra de agua superficial; este número de bacterias puede variar dependiendo del método de análisis que se utilice o el volumen de agua utilizado para el análisis, incluso con la utilización del mismo método, se pueden obtener resultados ligeramente diferentes. En virtud de lo anterior, se define a una variable aleatoria, como el resultado que se obtiene luego de realizar un experimento aleatorio. En el ejemplo, podemos definir la variable aleatoria X = "*Numero de bacterias coliformes encontradas en una muestra de agua superficial*". Este experimento se puede repetir tantas veces como se quiera. Si realizamos este experimento 100 veces y tomamos el registro del número de coliformes encontrados en la muestra de agua, obtenemos una muestra de valores de la variable aleatoria de tamaño muestral 100.

Las variables aleatorias, como cualquier variable estadística pueden ser discretas, si los valores que asume son siempre números enteros, por ejemplo, el número de flores de cierta especie de árbol, cantidad de larvas de *tubifex* encontradas durante un muestreo de

macroinvertebrados acuáticos, entre otros, o continuas, si los valores que asume la variable aleatoria se encuentran dentro de un intervalo, por ejemplo, la temperatura del agua de un rio determinado, los niveles de presión sonora en decibeles durante un estudio de ruido ambiental, la producción de residuos sólidos urbanos en kilogramos en cierta ciudad, etc.

Predecir cuál será el valor que asuma una variable aleatoria durante un experimento, es una tarea muy importante dentro de la inferencia estadística, lo cual se consigue gracias al estudio de las probabilidades y las distribuciones de probabilidad teórica que asumen las variables de estudio. El desarrollo de estas dos disciplinas ha permitido el desarrollo de los métodos estadísticos que permitan extrapolar las decisiones que se toman sobre una muestra a la población.

En las secciones siguientes abordaremos todo lo relacionado con los modelos de distribución de probabilidad, tanto para variables aleatorias discretas como continuas. Se ilustrarán los conceptos con algunos ejemplos y luego se procederá a la aplicación de los mismos en el entorno de programación de R.

3.2. Distribuciones discretas de probabilidad

La forma en que se asigna la probabilidad a los resultados de una variable aleatoria discreta está dada por una función de los valores numéricos que toma la variable aleatoria y que denotaremos como $p(x)$, de manera que $p(x) = P(X = x)$. Así, al conjunto de pares ordenados $(x, p(x))$ se le llama **función de probabilidad** o **distribución de probabilidad** de la variable aleatoria *X*. Esta función debe satisfacer las siguientes condiciones:

1. $p(x) \geq 0$
2. $\sum_x p(x) = 1$
3. $P(X = x) = p(x)$

En términos simples, la distribución de probabilidad de la variable aleatoria *X*, indica cómo está distribuida (asignada) la probabilidad total de 1 entre los varios posibles valores que puede tomar *X* (Devore, 2008).

Ejemplo 3.1. Considérese un experimento que consta de la observación de 4 semillas en un cierto orden, cada una de las cuales puede estar sana

(situación que se representará con el signo "+") o bien enferma (situación que se representará con el signo "-"), tal como se muestra a continuación

$$S = (++++, +++-, ++-+, ++--, +-++, +-+-, +--+, +---,$$
$$-+++, -++-, -+-+, -+--, --++, --+-, ---+, ----)$$

Así, las probabilidades de encontrar 0, 1, 2 o 3 semillas sanas con sus respectivas probabilidades se muestran en la siguiente tabla

X	0	1	2	3	4
$P(X = x)$	1/16	1/4	3/8	1/4	1/16

Esta representa la función de probabilidad de la variable aleatoria *número de semillas sanas*. A partir de ella y haciendo uso de algunos conceptos de probabilidad elemental se pueden calcular otras probabilidades de interés, por ejemplo, la probabilidad de encontrar por lo menos una semilla sana sería

$$P(X \geq 1) = 1 - P(X \leq 0) = 1 - \frac{1}{16} = \frac{15}{16}$$

La probabilidad de encontrar como máximo 2 semillas sanas es,

$$P(X \leq 2) = p(0) + (1) + p(2)$$
$$P(X \leq 2) = \frac{1}{16} + \frac{1}{4} + \frac{3}{8} = \frac{11}{16}$$

La probabilidad de encontrar por lo menos 3 semillas sanas sería

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - \frac{11}{16} = \frac{5}{16}$$

A partir de lo anterior se define el concepto de **función de la distribución acumulada**, como aquella función que nos permite calcular la probabilidad de que el valor observado de la variable aleatoria X sea menor o igual que algún número real x, es decir $p(x) = p(X \leq x) = \sum_{t \leq x} p(t)$,

para $-\infty < x < \infty$.

En términos generales la función acumulada de probabilidad de este ejemplo sería

$$P(x)=\begin{cases}0, \text{ para } x<0\\ \frac{1}{16}, \text{ para } 0\le x<1\\ \frac{5}{16}, \text{para } 1\le x<2\\ \frac{11}{16}, \text{ para } 2\le x<3\\ \frac{15}{16}, \text{para } 3\le x<4\\ 1, \text{ para } x\ge 4\end{cases}$$

La distribución de probabilidad y la función de la distribución acumulada se pueden representar gráficamente como se muestra en la Figura 3.1.

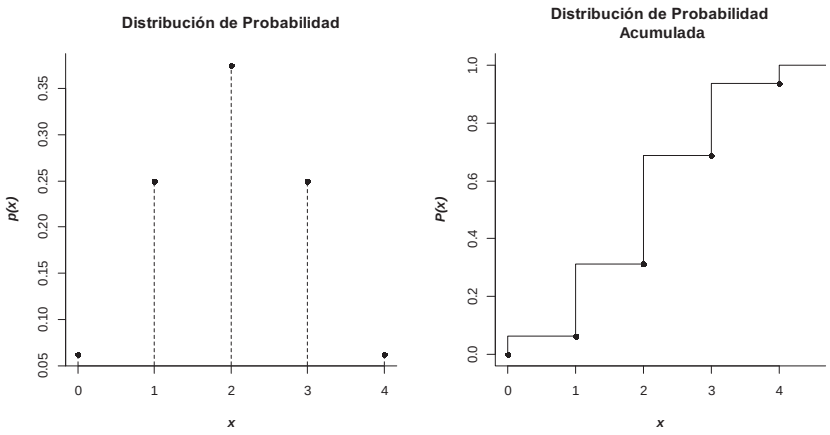


Figura 3.1. Representación gráfica de la distribución de probabilidad y la distribución de probabilidad acumulada

A continuación se hará una exposición de las principales distribuciones de probabilidad para variables aleatorias discretas de mayor interés en el tratamiento de datos, específicamente en el campo de la inferencia estadística.

3.2.1. Distribución de probabilidad binomial

Una distribución de probabilidad de una variable aleatoria discreta utilizada ampliamente es la distribución binomial. Esta distribución es

apropiada para una variedad de procesos que describe datos discretos, que son resultado de un experimento conocido como proceso de Bernoulli, en honor al matemático Suizo Jacob Bernoulli (1654 - 1705), el cual nos llevará a uno de sólo dos resultados posibles que son mutuamente excluyentes, tales como muerto o vivo, enfermo o saludable, etc., en donde la obtención del resultado deseado se considera como éxito "p" y el resultado no deseado como fracaso "q", donde, $q = 1 - p$.

Todo proceso de Bernoulli (y por ende un experimento binomial) debe ajustarse a cada una de las siguientes propiedades:

1. El experimento consta de una secuencia de n experimentos más pequeños llamados ensayos, donde n se fija antes del experimento.
2. Cada ensayo puede dar por resultado uno dos resultados posibles (ensayos dicotómicos), los cuales se denotan por éxito o fracaso.
3. Los ensayos son independientes, de modo que el resultado de cualquier ensayo particular no influye en el resultado de cualquier otro.
4. La probabilidad de éxito, denotada por p, permanece constante de un ensayo a otro.

Ejemplos de experimentos binomiales pueden ser, el número de organismos (por ejemplo, *E. coli*) que permanecen vivos o muertos luego de un bioensayo a donde se exponen a los individuos a cierta concentración de un desinfectante; que las descargas de aguas residuales domesticas de cierta ciudad a un cuerpo de agua cumpla o no con la legislación nacional de límites permisibles de vertimiento.

En virtud de lo anteriormente mencionado, la distribución de probabilidad de una variable aleatoria binomial X, con una probabilidad p de éxitos en n ensayos independientes se define de la siguiente forma

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \quad x = 1, 2, \ldots, n. \tag{3.1}$$

Como comentario adicional se puede decir que la gráfica de la distribución binomial puede ser simétrica o no dependiendo de los valores de p. Así, cuando $p = 0.5$ la distribución toma una forma simétrica, cuando $p < 0.5$ la distribución toma una forma asimétrica con cola hacia la derecha, por ultimo cuando $p > 0.5$ la distribución toma una forma asimétrica con cola hacia la izquierda (Figura 3.2).

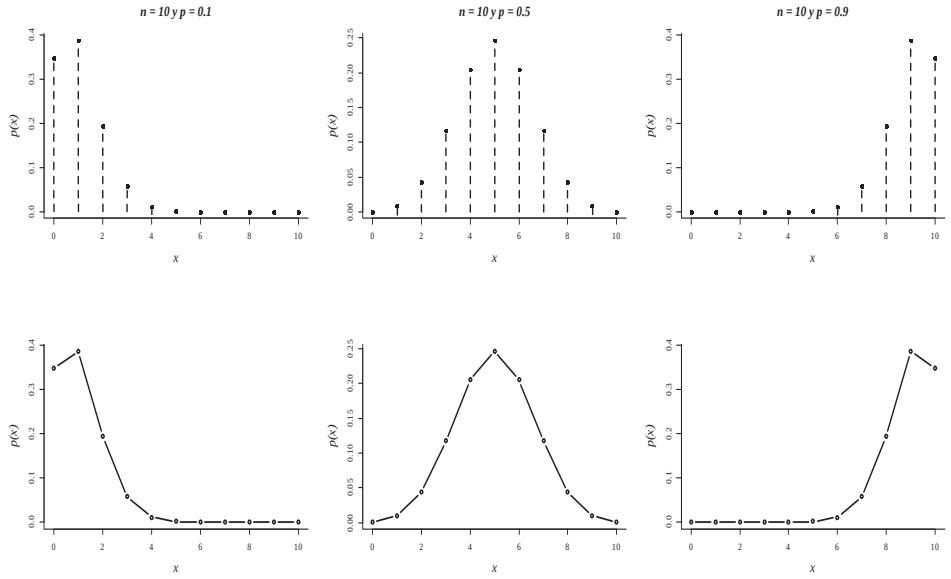


Figura 3.2. Gráficas de la función binomial de probabilidad

Ejemplo 3.1. Supóngase que el 40% de los ríos de cierta región industrial de México están contaminados con benceno. Si tomamos una muestra aleatoria de tamaño $n = 30$, calcular las siguientes probabilidades:

a) Que exactamente 15 ríos estarán contaminados con benceno

b) Cuando menos 15 ríos estarán contaminados con este compuesto orgánico cancerígeno, de una muestra de $n = 25$.

c) No más de 10 ríos, pero cuando menos de 5 ríos estarán contaminados de una muestra aleatoria de $n = 25$.

Solución

a) La probabilidad de que exactamente 15 ríos se encuentren contaminados con benceno se define como $P(X = 15) = b(15, 30, 0.4)$, cuyo cálculo respectivo sería

$$b(15,30,0.4) = \binom{30}{15}(0.4)^{15}(0.6)^{15} = \frac{30!}{15!15!}(0.4)^{15}(0.6)^{15}$$

$$b\,(15,30,0.4) = 0.078$$

b) La probabilidad de que cuando menos 15 ríos estén contaminados con benceno de una muestra de 25 está dada por:

$$P(X \geq 15) = 1 - P(X < 15)$$

En este caso es preciso realizar la suma de todas las probabilidades de los valores que asume *x* desde 0 hasta 14. Para simplificar esto, se encuentran disponibles arreglos tabulares para las sumas binomiales para diferentes valores *n* y probabilidades *p* entre 0.1 y 0.9 (Tabla A.1 del Apéndice). Así,

$$P(X \leq 15) = 1 - \sum_{x=0}^{14} b(x; 25, 0.4)$$

$$P(X \leq 15) = 1 - 0.966 = 0.034$$

c) Aquí la $P(5 \leq X \leq 10)$ se calcula así:

$$P(5 \leq X \leq 10) = P(X \leq 10) - P(X \leq 4)$$

$$P(5 \leq X \leq 10) = \sum_{x=0}^{10} b(x; 25, 0.4) - \sum_{x=0}^{4} b(x; 25, 0.4)$$

$$P(5 \leq X \leq 10) = 0.586 - 0.009$$

$$P(5 \leq X \leq 10) = 0.577$$

En R, la distribución de probabilidad binomial se modela a través de las funciones:

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

Donde las letras que anteceden a la expresión ***binom*** ejecutan un proceso diferente, por ejemplo, la ***d*** ordena a R que se calculen densidades (probabilidades), la letra ***p*** permite el cálculo de la distribución de probabilidad acumulada, la letra ***q*** calcula cuantiles y la letra ***r***, genera números aleatorios que siguen una distribución de probabilidad binomial.

Los argumentos involucrados con esta función expresan lo siguiente:

x, q: vector de cuantiles.

p: vector de probabilidades.

size: número de observaciones *n*.

prob: probabilidad de éxito en cada ensayo.

log.p: valor lógico, si es TRUE, las probabilidades *p* se ofrecen como *log* (*p*).

lower.tail: valor lógico, si es TRUE (por defecto), las probabilidades son $P(X \leq x)$ (Cola de la izquierda) de lo contrario $P(X > x)$ (cola de la derecha).

Veamos el uso de estas funciones en el ejemplo abordado anteriormente.

```
> a<-dbinom(15,30,0.4)
> a
[1] 0.07831221
> b<-(1-pbinom(14,25,0.4))
> b
[1] 0.03439152
> c<-(pbinom(10,25,0.4)-pbinom(4,25,0.4))
> c
[1] 0.5763041
```

3.2.2. Distribución de probabilidad hipergeométrica

La función hipergeométrica es una distribución discreta de probabilidad, la cual está estrechamente ligada a la distribución binomial. La manera más simple de ver la diferencia entre las dos distribuciones radica en la forma que se realiza el muestreo. La diferencia entre estas dos distribuciones radica en que, en la distribución binomial, los intentos son independientes, porque hay reemplazo en la selección de la muestra. Sin embargo, en el caso de la distribución hipergeométrica, hay dependencia, porque la selección de la muestra se hace sin reemplazo y la probabilidad de éxito cambia de un intento a otro (Quevedo, 2006).

La distribución hipergeométrica se apoya en las siguientes suposiciones:

1. La población o conjunto que se va a muestrear se compone de *N* individuos, objetos o elementos (población finita).

2. Cada individuo puede ser caracterizado como un *éxito* o una *falla*, existiendo k éxitos y $N - k$ fallas en la población.
3. Se selecciona una muestra de n individuos sin remplazo, de tal modo que cada subconjunto de tamaño n es igualmente probable de ser seleccionado.

La distribución de probabilidad hipergeométrica tiene las mismas aplicaciones que la distribución binomial, con la diferencia de que en la distribución hipergeométrica el muestreo se realiza sin remplazo.

En virtud de lo anterior, la distribución de probabilidad de la variable aleatoria hipergeométrica X, el número de éxitos en una muestra aleatoria de tamaño n que se selecciona de N artículos, en los que existen k éxitos y $N - k$ fallas, está dada por

$$h(x;N,n,k)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad máx\{0,n-(N-k)\}\le x\le mín\{n,k\}. \tag{3.2}$$

La Figura 3.3, muestra diferentes gráficos de la función de distribución hipergeométrica a diferentes valores de los parámetros N, n y k.

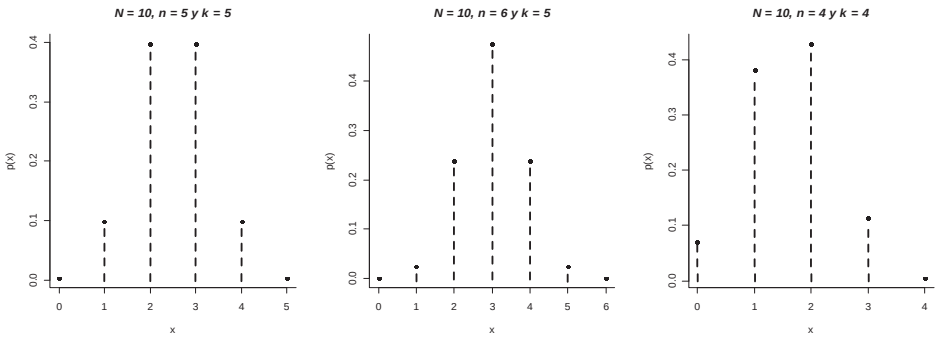


Figura 3.3. Graficas de la función de probabilidad hipergeométrica.

En R, el modelamiento de la distribución de probabilidad hipergeométrica se hace a través de las siguientes órdenes,

```
dhyper(x, k, N - k, n, log = FALSE)
phyper(q, k, N - k, n, lower.tail = TRUE, log.p = FALSE)
qhyper(p, k, N - k, n, lower.tail = TRUE, log.p = FALSE)
rhyper(n, k, N - k, n)
```

Ejemplo 3.2. Asúmase que de 10 ríos del departamento de La Guajira se tiene certeza que 5 están contaminados con organoclorados. Si se selecciona al azar una muestra de 4 ríos, encontrar las siguientes probabilidades:

a) Exactamente un río se encuentre contaminado con organoclorados.
b) Dos ríos estén contaminados.
c) A lo sumo dos ríos estén contaminados.
d) Entre uno y tres ríos se encuentren contaminados

Solución

Como datos de importancia, se tiene que los parámetros de la distribución dados corresponden a $N=10$, $n=4$, $k=5$. Así,

a) La probabilidad de encontrar exactamente un río contaminado es

$$P(X=1)=h(1;10,4,5)=\frac{\binom{5}{1}\binom{5}{3}}{\binom{10}{4}}=\frac{\left(\frac{5!}{1!4!}\right)\left(\frac{5!}{3!2!}\right)}{\left(\frac{10!}{4!5!}\right)}$$

$$P(X=1)=0.238$$

b) La probabilidad de encontrar dos ríos contaminados se determina por

$$P(X=2)=h(2,10,4,5)=\frac{\binom{5}{2}\binom{5}{2}}{\binom{10}{4}}$$

$$P(X=2)=0.476$$

c) La probabilidad que máximo dos ríos estén contaminados se determina por

$$P(X\leq 2)=\sum_{x=0}^{2}h(x;10,4,5)=h(0;10,4,5)+h(1;10,4,5)+h(2;10,4,5)$$

$$P(X\leq 2)=0.738$$

d) Por último, la probabilidad de encontrar entre 1 y 3 ríos contaminados por organoclorados se establece a través del siguiente procedimiento

$$P(1 \le X \le 3) = \sum_{x=0}^{3} h(x;10,4,5) - \sum_{x=0}^{0} h(x;10,4,5) = \sum_{x=0}^{3} h(x;10,4,5) - h(0;10,4,5)$$

$$P(1 \le X \le 3) = 0.952$$

Los cálculos anteriores se resumen en R con solo aplicar las líneas de comando que se mostraron anteriormente, como se muestra en la siguiente salida de resultados

```
> a<-dhyper(1,5,5,4)
> a
[1] 0.2380952
> b<-dhyper(2,5,5,4)
> b
[1] 0.4761905
> c<-phyper(2,5,5,4)
> c
[1] 0.7380952
> d<-(phyper(3,5,5,4)-phyper(0,5,5,4))
> d
[1] 0.952381
```

Es de notar que los cálculos de las probabilidades en R son mucho más preciso gracias al número de cifras significativas que esta toma en consideración para los cálculos respectivos.

3.2.3. Distribución de probabilidad de Poisson

Esta distribución es una de las más importantes distribuciones de probabilidad para variables aleatorias discretas, desarrollada por el francés Siméon Denis Poisson, 1837. Sus principales aplicaciones hacen referencia a la modelación de situaciones en las que nos interesa determinar el número de hechos de cierto tipo que se pueden producir en un intervalo de tiempo o de espacio, bajo supuestos de aleatoriedad y ciertas circunstancias restrictivas. En otras palabras, la distribución de Poisson es un modelo que puede usarse para calcular la probabilidad correspondiente al número de *éxitos* que ocurren en una región o en un

intervalo de tiempo especificados, si se conoce el número promedio de *éxitos* que ocurren.

Este modelo de distribución requiere que se cumplan las siguientes suposiciones (Rodríguez, 2007):

1. El número de *éxitos* que ocurre en la región o intervalo es independiente de lo que ocurre en otra región o intervalo.
2. La probabilidad de que un resultado ocurra en una región o intervalo muy pequeño, es igual para todos los intervalos o regiones de igual tamaño y es proporcional al tamaño de la región o intervalo.
3. La probabilidad de que más de un resultado ocurra en una región o intervalo muy pequeño no es significativa.

Las diferencias más importantes de la distribución de Poisson respecto a la distribución binomial son que la distribución de Poisson se aplica a sucesos que pueden tener una probabilidad muy baja y, además, el tamaño de n es infinito. En algunos casos la distribución de Poisson se utiliza como aproximación a la binomial cuando n es muy grande y, por tanto, es difícil el cálculo de la binomial. Además, cuando la probabilidad de algunos de los eventos es muy baja (Guisande *et al.*, 2011).

La aproximación de la distribución de Poisson a la distribución binomial es apropiada cuando $p \leq 0.05$ y $n \geq 20$. En realidad, el porcentaje de error de los resultados obtenidos usando la distribución de Poisson, como una aproximación a la distribución binomial, es de 1 en 270 o cerca de 0.4% (Quevedo, 2006).

La función de probabilidad de la distribución de Poisson se expresa como,

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 1,2,... \tag{3.3}$$

Donde λ es el promedio de resultados por unidad de tiempo, distancia, área o volumen.

En la Figura 3.4, se proporcionan distintas graficas de la función de probabilidad de Poisson para diferentes valores de λ.

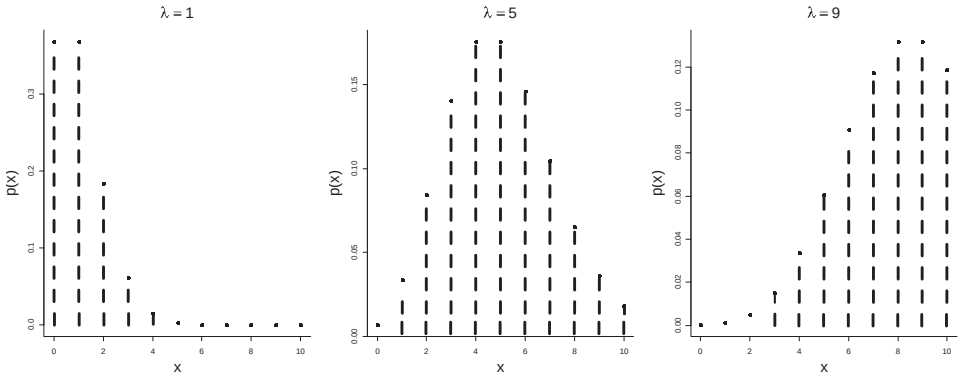


Figura 3.4. Gráficas de la función de probabilidad de Poisson.

En R el modelamiento de la distribución de probabilidad de Poisson se realiza a través de las funciones

```
dpois (x, lambda, log = FALSO)
ppois (q, lambda, lower.tail = TRUE, log.p = FALSO)
qpois (p, lambda, lower.tail = TRUE, log.p = FALSO)
rpois (n, lambda)
```

Ejemplo 3.3. Supóngase que en un estudio de contaminación ambiental se instala una red de 3,840 sensores de alto volumen para medir las concentraciones de partículas atmosféricas, menores que 10 micras (PM_{10}). Si en promedio cuatro muestreadores fallan al año, encontrar la probabilidad de:

a) Que fallen exactamente 5 muestreadores
b) A lo sumo 4 muestreadores fallen
c) Entre 1 y 5 muestreadores fallen

Solución

a) La probabilidad de que fallen exactamente 5 muestreadores está dada por

$$P(X=5) = p(5;4) = \frac{e^{-4}\lambda^5}{x!}$$

$$P(X=5) = 0.156$$

b) La probabilidad de que máximo 5 muestreadores fallen se determina por

$$p(X \leq 4) = \sum_{x=0}^{4} p(x;4) = p(0;4) + p(1;4) + p(2;4) + p(3;4)$$

$$P(X \leq 4) = 0.688$$

Al igual que para la distribución binomial, la distribución de Poisson posee unos formatos tabulares para el cálculo de probabilidades acumulativas. Estas se encuentran disponibles en la Tabla A.2 (del apéndice) y permiten solucionar este y el enciso siguiente.

c) Por último, la probabilidad de fallar entre 1 y 5 muestreadores se calcula a través de

$$P(1 \leq X \leq 5) = \sum_{x=0}^{5} p(x;4) - \sum_{x=0}^{0} p(x;4)$$

Haciendo uso de la Tabla A2, para hallar las probabilidades acumuladas en cuestión, se obtiene

$$P(1 \leq X \leq 5) = 0.7851 - 0.0183$$

$$P(1 \leq X \leq 5) = 0.767$$

La salida de resultados de R para los cálculos mecánicos realizados para este ejemplo se muestra a continuación

```
> a<-dpois(5,4)
> a
[1] 0.1562935
> b<-ppois(4,4)
> b
[1] 0.6288369
> c<-(ppois(5,4)-ppois(0,4))
> c
[1] 0.7668147
```

3.3. Distribuciones continúas de probabilidad

Una distribución de probabilidad es continua cuando los resultados posibles del experimento son obtenidos de variables aleatorias continuas,

es decir, de variables cuantitativas que pueden tomar cualquier valor dentro de un intervalo, y que resultan principalmente de un proceso de medición. Esta característica de las variables continuas, permiten definir a priori que la probabilidad de que este tipo de variables tome exactamente cualquiera de sus valores es igual a cero, no siendo el caso, si nos referimos a la probabilidad de que la variable aleatoria continúa tomo valores dentro de un intervalo.

Así mismo, esta característica impide que la distribución de probabilidad de una variable aleatoria continua se represente en forma tabular, sin embargo, si es posible establecer una ecuación matemática que dependa de los valores de la variable aleatoria, la cual denotaremos de ahora en adelante $f(x)$. Al tratar con variables aleatorias continuas $f(x)$, recibe el nombre de **función de densidad de probabilidad**, o simplemente **función de densidad**. Esta función de densidad de probabilidad se construye de manera que el área bajo la curva limitada por el eje *x* sea igual a 1, cuando se calcula en el rango de *X* para el que se define $f(x)$. Así, para dos números cualesquiera *a* y *b* con $a \le b$, la probabilidad de que *X* tome un valor dentro $[a,b]$ es igual al área bajo la curva de la función de densidad entre las ordenadas $x = a$ y $x = b$ (Figura 3.5), y matemáticamente está dada por

$$P(a \le X \le b) = \int_a^b f(x)dx$$

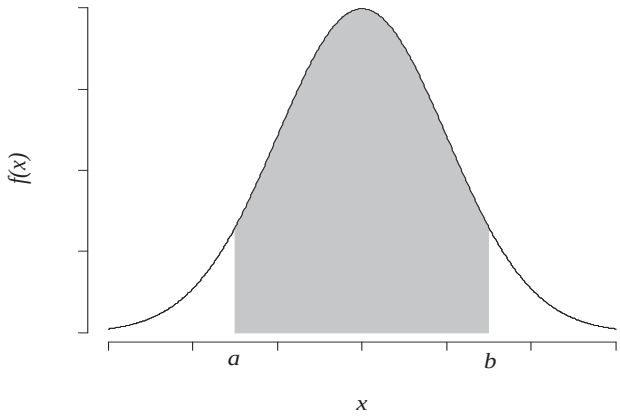


Figura 3.5. $P(a \le X \le b)$

Toda función de densidad de probabilidad para variables aleatorias continuas, debe satisfacer las siguientes condiciones:

1. $f(x) \geq 0$, para todo $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.
3. $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Ejemplo 3.4: Al simbolizar con X la cantidad de tiempo de incubación de bacterias en un plato de agar de prueba durante 2 horas. Supóngase que la variable aleatoria X tiene función de densidad de $f(x) = 0.5x$, para el conjunto posible de valores de X en el intervalo $(0 \leq X \leq 2)$. Siendo así, calcular las siguientes probabilidades:

a) $P(X \leq 1)$.
b) $P(0.5 \leq X \leq 1.5)$.
c) $P(1.5 < X)$.

Solución

a) Teniendo en cuenta que X se encuentra definida en el intervalo $(0 \leq X \leq 2)$, se tiene que

$$f(x) = \begin{cases} 0.5x, & 0 \leq x \leq 2 \\ 0, \text{en cualquier otro caso} \end{cases}$$

Así,

$$P(X \leq 1) = \int_0^1 0.5dx = \left[0.5\left(\frac{x^2}{2} \right) \right]_0^1$$

$$P(X \leq 1) = \left[0.5\left(\frac{1}{2} \right) \right] - \left[0.5\left(\frac{0}{2} \right) \right]$$

$$P(X \leq 1) = 0.25$$

b) La probabilidad de que el tiempo de incubación tome un valor entre 0.5 y 1.5 horas se determina como

$$P(0.5 \leq X \leq 1.5) = \int_{0.5}^{1.5} 0.5xdx = \left[0.5\left(\frac{x^2}{2} \right) \right]_{0.5}^{1.5}$$

$$P(0.5 \leq X \leq 1.5) = \left[0.5\left(\frac{2.25}{2} \right) \right] - \left[0.5\left(\frac{0.25}{2} \right) \right]$$

$$P(0.5 \leq X \leq 1.5) = 0.5$$

c) Por último la probabilidad de que el tiempo de incubación sea mayor a 1.5 horas es

$$P(1.5 < X) = \int_{1.5}^{2} 0.5x dx = \left[0.5\left(\frac{x^2}{2} \right) \right]_{1.5}^{2}$$

$$P(1.5 < X) = \left[0.5\left(\frac{4}{2} \right) \right] - \left[0.5\left(\frac{2.25}{2} \right) \right]$$

$$P(1.5 < X) = 0.44$$

Análogo a las distribuciones de probabilidad para variables discretas, se puede definir una función de distribución acumulada $F(x)$, que nos proporcione la probabilidad de que la variable aleatoria adopte un valor menor o igual al valor especificado, es decir, $F(x) = P(X \leq x)$. Esta función de probabilidad, para el caso de variables aleatorias continuas, se halla calculando el área delimitada por la gráfica de la función de densidad de probabilidad, a la izquierda del punto especificado (Figura 3.6).

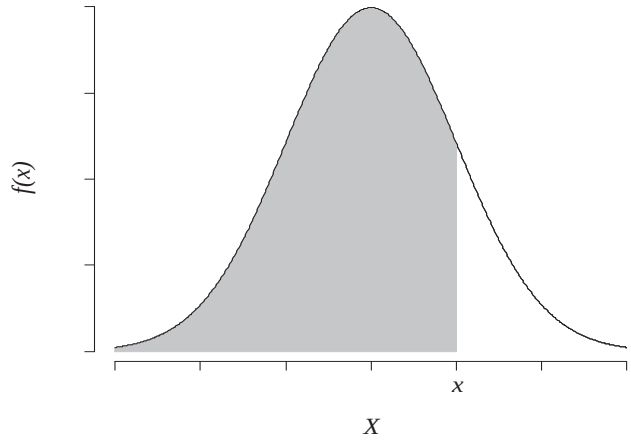


Figura 3.6. $P(X \leq x)$.

En términos matemáticos, la función de distribución acumulada $F(x)$ de una variable aleatoria continua X con función de densidad $f(x)$ se define como

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \text{, para } -\infty < x < \infty .$$

Cuya importancia, al igual que variables aleatorias discretas, está en que las probabilidades de varios intervalos pueden ser calculadas como una ecuación de $F(x)$, así

$$P(X > a) = 1 - F(a)$$

$$P(a \leq X \leq b) = F(b) - F(a)$$

Ejemplo 3.5. Para la función de densidad del ejemplo 3.4 encuentre $F(x)$, y utilícela para evaluar $P(0.5 \leq X \leq 1.5)$.

Solución

Dado que X se encuentra definida en el intervalo $(0 \leq X \leq 2)$, entonces

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} 0.5tdt = \left[\frac{t^2}{4}\right]_{0}^{x} = \frac{x^2}{4}$$

Por lo tanto,

$$F(x) = \begin{cases} 0, & x < 0, \\ \dfrac{x^2}{4}, & 0 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

La distribución acumulada $F(x)$ se ilustra en la Figura 3.7. Así,

$$P(0.5 \leq X \leq 1.5) = F(1.5) - F(0.5)$$

$$P(0.5 \leq X \leq 1.5) = \frac{9}{16} - \frac{1}{16}$$

$$P(0.5 \leq X \leq 1.5) = 0.5$$

Que concuerda con el resultado obtenido al utilizar la función de densidad de probabilidad en el ejemplo 3.4 (área sombreada del gráfico).

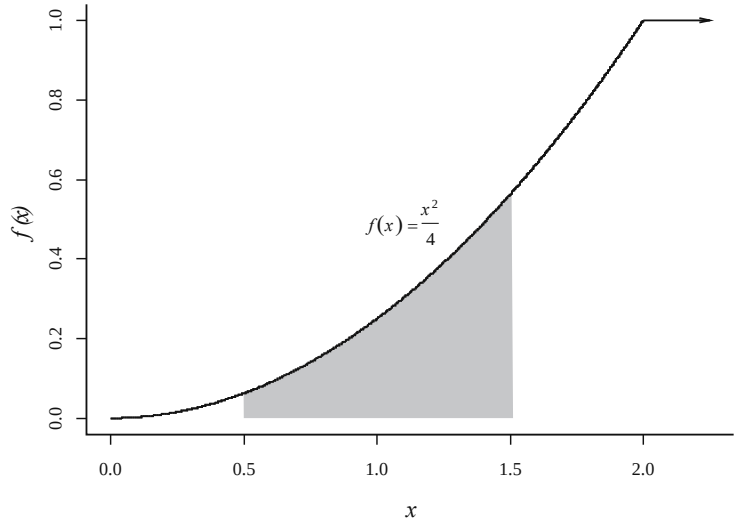


Figura 3.7. Función de distribución acumulada.

A continuación se hará una exposición de las principales distribuciones de probabilidad para variables aleatorias continuas, y de mayor interés en el tratamiento de datos, específicamente en el campo de la inferencia estadística.

3.3.1. Distribución de probabilidad normal

La distribución normal es indudablemente la más importante y la de mayor uso de todas las distribuciones continuas de probabilidad (Yakir, 2011). Es la piedra angular en la aplicación de la inferencia estadística en el análisis de datos puesto que las distribuciones de muchas estadísticas muéstrales tienden hacia la distribución normal conforme crece el tamaño de la muestra ($n \geq 30$). La apariencia grafica de la distribución normal es una curva simétrica con forma de campana, que se extiende sin límite tanto en la dirección positiva como en la negativa (Figura 3.8) (Conavos, 1988). Un gran número de estudios indica que la distribución normal proporciona una adecuada representación de datos provenientes de mediciones meteorológicas como temperatura y precipitación, mediciones efectuadas en organismos, mediciones de concentración de contaminantes atmosféricos, etc.

El desarrollo de esta distribución tuvo su inicio en 1733 por Abraham DeMoivre quien desarrollo la ecuación matemática de la curva normal. Sin embargo, este descubrimiento no llamo mucho la atención, perdiéndose el trabajo de DeMoivre por casi medio siglo cuando Pierre-Simon Laplace y

Carl Friedrich Gauss redescubrieron esta distribución de manera independiente, en sus estudios del comportamiento de los errores de las medidas astronómicas, de allí que en muchas ocasiones a esta distribución se le denomine **distribución gaussiana**.

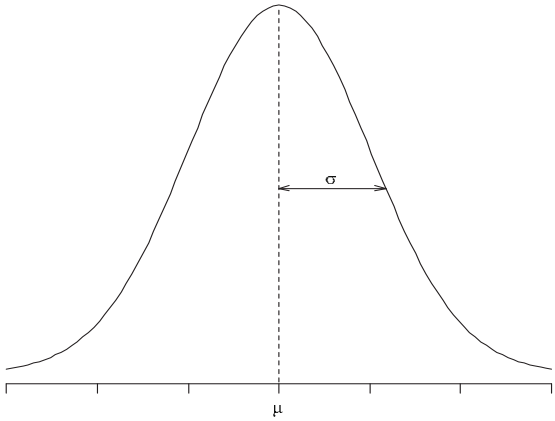


Figura 3.8. Curva de una distribución normal con media μ y desviación estándar σ.

La ecuación matemática descubierta por DeMoivre para modelar la función de densidad de probabilidad de una variable aleatoria X que sigue una distribución normal está dada por

$$n(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty, \tag{3.4}$$

Donde $\pi = 3.14159...$, y $e = 2.71828...$.

Así, la forma de la distribución normal queda completamente determinada una vez se encuentra especificado el valor de la media (μ) y la desviación estándar (σ). En las Figuras 3.9, 3.10 y 3.11, se muestran diferentes curvas normales con diferentes valores de μ y σ. Así mismo, a partir de una inspección de estas graficas se pueden establecer las siguientes características de la distribución normal citadas por Quevedo (2006):

1. Es simétrica alrededor de su promedio μ y en forma de campana.
2. El promedio, la mediana y la moda son iguales.
3. El área total bajo la curva es igual a uno. El 50% de las observaciones están a la derecha del promedio y el otro 50% de las observaciones están a la izquierda del promedio.

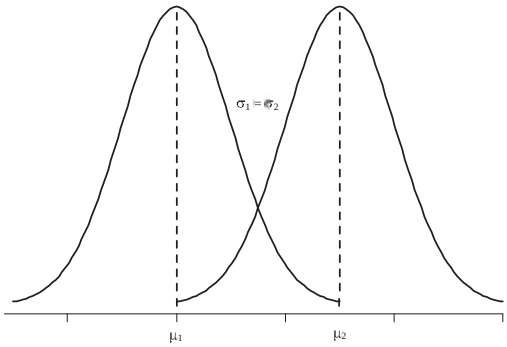


Figura 3.9. Curvas normales con $\mu_1 < \mu_2$ y $\sigma_1 = \sigma_2$.

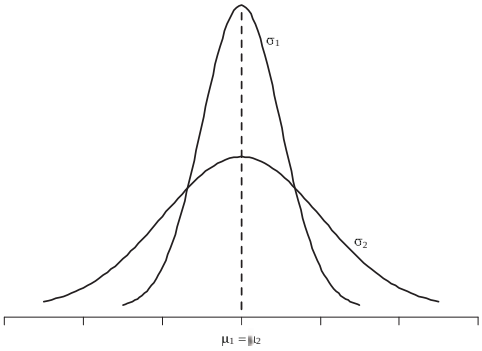


Figura 3.10. Curvas normales con$\mu_1 = \mu_2$ y $\sigma_1 < \sigma_2$.

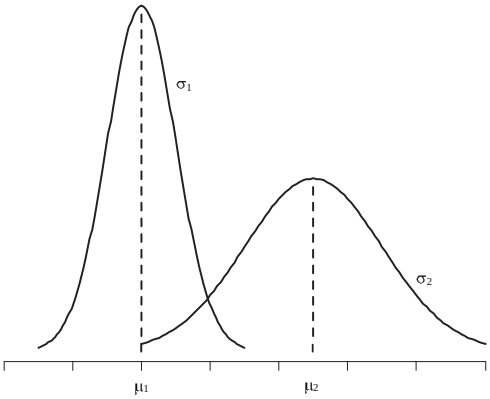


Figura 3.11. Curvas normales con $\mu_1 < \mu_2$ y $\sigma_1 < \sigma_2$.

Como cualquier distribución de probabilidad el interés de investigar distribuciones normales se centra en determinar su función de densidad de probabilidad o función de distribución acumulada, para el cálculo de

probabilidades como $P(a \leq X \leq b)$, que queda determinada matemáticamente por

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} n(x;\mu,\sigma)dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \qquad (3.5)$$

Gráficamente el cálculo de $P(x_1 \leq X \leq x_2)$ se representa por el área sombreada en la Figura 3.12.

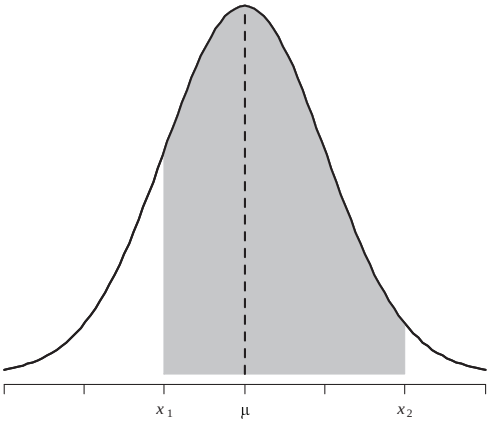


Figura 3.12. $P(x_1 \leq X \leq x_2)$.

Sin embargo, ningún método de integración conocido puede usarse para evaluar (3.5), por ello se han recurrido a métodos tabulares cuando $\mu = 0$ y $\sigma = 1$ para ciertos valores de x_1 y x_2, siendo posible la utilización de la tabla obtenida para calcular probabilidades cuando se tiene otros valores para μ y σ. Así, se origina la **distribución normal estándar**, definida como aquella distribución normal con $\mu = 0$ y $\sigma = 1$. Cualquier variable aleatoria que tiene una distribución normal estándar se denomina **variable aleatoria normal estándar**, denotada por *Z*, y su función de densidad de probabilidad es

$$n(z;0,1) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2}}, \qquad -\infty < z < \infty \qquad (3.6)$$

Así mismo, la función de probabilidad acumulada de *Z* es

$$P(Z \leq z) = \int_{-\infty}^{z} n(t;0,1)dt \qquad (3.7)$$

El proceso de transformación de todas las observaciones de cualquier variable aleatoria normal X a un nuevo conjunto de observaciones de una variable aleatoria normal estándar Z, se denomina **estandarización** o **tipificación**, lo cual se consigue a través de la expresión

$$Z = \frac{X - \mu}{\sigma}$$

Cuyo estimador muestral es

$$z = \frac{X - \bar{x}}{s}$$

En la Tabla A.3 del apéndice, se da $P(Z \leq z)$, para diferentes valores de z, es decir, el área bajo la curva de densidad normal estándar a la izquierda de z.

Ejemplo 3.6. En Colombia los ministerios de protección social y ambiente, vivienda y desarrollo territorial, con la formulación de la resolución 2115 de 2007 establecen un índice para evaluar el riesgo sanitario del agua destinada al consumo humano, denominado Índice de Riesgo de Calidad de Agua (IRCA), el cual entre sus diferentes categorías establece como agua sanitariamente inviables a aquellas que toman un puntaje entre 80.1% y 100%. De una muestra de 100 viviendas del municipio de Riohacha, La Guajira cuyos puntajes del IRCA mensual se ajustan a una distribución normal con $\bar{x} = 54.9\%$ y $s = 15.2\%$, calcular las siguientes probabilidades:

a) Que el IRCA mensual sea menor que 80.1%.
b) Que el IRCA mensual se encuentre entre 80.1 y 100%.

Solución

a) Como paso previo al cálculo de las probabilidades pedidas, debemos estandarizar la variable X (IRCA mensual).

$$z_1 = \frac{80.1 - 54.9}{15.2} \therefore z_1 = 1.66$$

$$z_2 = \frac{100 - 54.9}{15.2} \therefore z_2 = 2.97$$

Así, la probabilidad de que el IRCA mensual sea menor que 80.1%, se determina haciendo uso de la Tabla A.3 como

$$P(X \leq 80.1) = P(Z \leq 1.66) = 0.9515$$

b) Por definición sabemos que para variables aleatorias continuas $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) = P(X \leq x_2) - P(X \leq x_1)$, y haciendo uso de la Tabla A.3, tenemos que la probabilidad de que el IRCA mensual se encuentre entre 80.1 y 100% es

$$P(80.1 \leq X \leq 100) = P(1.66 \leq Z \leq 2.97) = P(Z \leq 2.97) - P(Z \leq 1.66)$$

$$P(80.1 \leq X \leq 100) = 0.9985 - 0.9515$$

$$P(80.1 \leq X \leq 100) = 0.047$$

Análogo a las distribuciones discretas de probabilidad, es posible el modelamiento de las distribuciones continuas de probabilidad en R, y en el caso específico de la distribución normal, este modelamiento se realiza a través de las siguientes funciones

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Aquí, las letras que anteceden a la expresión ***norm*** cumplen las mismas funciones descritas cuando se trataron las distribuciones discretas de probabilidad. Los argumentos ***mean*** y ***sd***, especifican la media aritmética y la desviación estándar, respectivamente; los argumentos, ***log***, ***lower.tail*** y ***log.p***, ejecutan las mismas ordenes vistas en la sección de distribuciones discretas de probabilidad.

Así, la solución del ejercicio anterior en R, corresponde a la siguiente salida de resultados

```
> a<-pnorm(80.1,mean=54.9,sd=15.2,lower.tail=TRUE)
> a
[1] 0.9513306
> b<-(pnorm(100,mean=54.9,sd=15.2,lower.tail=TRUE)-
pnorm(80.1,mean=54.9,sd=15.2,lower.tail=TRUE))
> b
[1] 0.04716627
```

La representación gráfica de estos resultados se muestra en la Figura 3.13 donde el grafico de la izquierda representa $P(X \leq 80.1)$ y el de la derecha muestra $P(80.1 \leq X \leq 100)$.

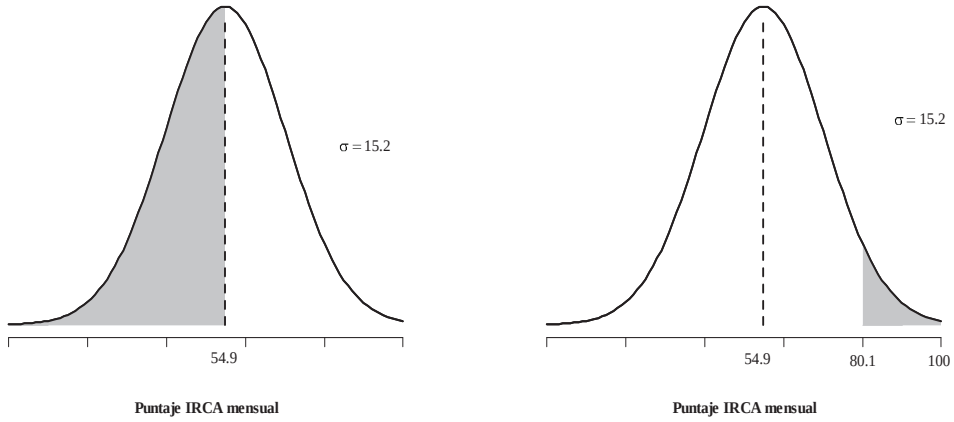


Figura 3.13. $P(X \leq 80.1)$ y $P(80.1 \leq X \leq 100)$ respectivamente.

3.3.2. Distribución de probabilidad t de student.

La distribución *t* de *student* es una distribución de densidad de probabilidad simétrica en forma de campana muy parecida a la distribución normal. En la practica la gran mayoría de los experimentos que se puedan realizar implican desconocimiento absoluto de la desviación estándar poblacional σ, dado que solo se trabaja con una pequeña muestra representativa de la población, y además, puede ocurrir que el número de observaciones sea pequeño ($n \leq 30$). En estos casos se puede utilizar la cuasidesviación estándar (*s*), y la distribución *t* de *student* para realizar estimaciones acerca de la media poblacional μ, siempre que se tenga certeza que la variable estudiada tiene una distribución normal. La variable aleatoria *T* se define matemáticamente como:

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

Cuya función de densidad de probabilidad viene dada por

$$h(t;\upsilon) = \frac{1}{\sqrt{\upsilon\pi}} \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)} \left(1+\frac{t^2}{\upsilon}\right)^{-\left(\frac{\upsilon+1}{2}\right)}, \quad -\infty < t < \infty, \qquad \upsilon > 0 \tag{3.8}$$

Donde $\upsilon = n - 1$ representan los grados de libertad

La distribución *t* de *student* se publicó por primera vez en 1908 en un artículo de W. S. Gosset. En esa época, Gosset era empleado de una cervecería irlandesa que no autorizaba la publicación de investigaciones de sus empleados. Para evadir tal prohibición, publico su trabajo en secreto bajo el nombre de *student*, de allí que deba su nombre (Walpole *et al.*, 2007).

La distribución *t* de *student* puede tener diferentes formas dependiendo de los grados de libertad υ (Figura 3.14). Como se mencionó antes, la apariencia general de la distribución *t* es similar a la distribución normal estándar. Sin embargo, la distribución *t* tiene colas más amplias que la normal, es decir, la probabilidad de las colas es mayor que en la distribución normal. La distribución *t* se transforma en una distribución normal cuando el tamaño de la muestra crece hacia el infinito (Guisande *et al.*, 2011).

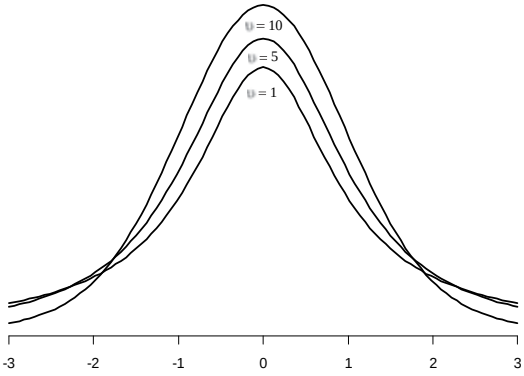


Figura 3.14. Curvas de distribución *t* de *student* para $\upsilon = 1, 5 \text{ y } 10$ grados de libertad.

Tal como ocurre con todas las distribuciones de probabilidad estudiadas, muchas veces se está interesado en determinar la función de distribución acumulada. Para ello, se establece un numero $t_{\alpha;\upsilon}$ como aquel sobre el eje de medición con el cual el área bajo la curva *t* con υ grados de libertad a la derecha de $t_{\alpha;\upsilon}$ es α; este número se denomina **valor critico *t*** (Figura 3.15a). Así,

$$P(T \geq t_{\alpha;\upsilon}) = \alpha \tag{3.9}$$

Como la distribución *t* es simétrica alrededor de cero, tenemos que $t_{1-\alpha;\upsilon} = -t_{\alpha;\upsilon}$; es decir, el valor *t* que deja un área de $1-\alpha$ a la derecha y,

por lo tanto un área de α a la izquierda de $t_{1-\alpha;\upsilon}$ (Figura 3.15b). Por ejemplo, $t_{0.95} = -t_{0.05}$, $t_{0.99} = -t_{0.01}$, etc.

A partir de lo anterior, tenemos que la función de distribución acumulada, queda matemáticamente determinada por

$$P(T \leq t_{1-\alpha,\upsilon}) = \int_{-\infty}^{t_{1-\alpha,\upsilon}} h(t;\upsilon)dt = 1-\alpha, \qquad 0 \leq \alpha \leq 1 \tag{3.10}$$

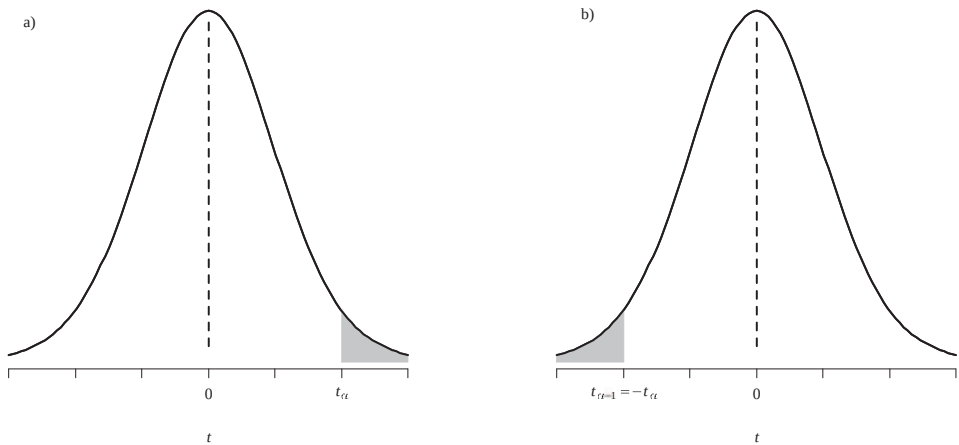


Figura 3.15. Propiedades de simetría de la curva de la distribución *t*.

Ejemplo 3.7. El ministerio de ambiente, vivienda y desarrollo territorial (hoy ministerio de ambiente y desarrollo sostenible), a través de la resolución 610 de 2010, estableció como nivel máximo permisible a condiciones de referencia para el dióxido de nitrógeno (NO_2), una concentración media diaria de 150 $\mu g/m^3$. Supóngase que a la chimenea de cierta industria se le realizó una evaluación del nivel de NO_2 emitido a la atmosfera, obteniéndose los siguientes resultados

Concentración diaria de NO_2 ($\mu g/m^3$)				
148	159	150	153	153
155	159	152	153	150

A partir de estos datos determinar la probabilidad de

a) Que la industria en promedio emita menos de 150 $\mu g/m^3$ de NO_2.

b) Que la concentración de media NO_2 emitida se encuentre entre 150 y 160 $\mu g/m^3$.
c) Que la concentración de media NO_2 sea mayor que 150 $\mu g/m^3$.

Asuma que los datos siguen una distribución normal

Solución

A partir de los datos se tiene que $s = 3.65\ \mu g/m^3$ y $\upsilon = 9$. Así, y a partir de 3.9 y 3.10, tenemos que

a) La probabilidad de que la emisión promedio de NO_2 en la chimenea de la industria sea menor que 150 $\mu g/m^3$, es

$$P(X \leq 150) = P(T \leq t) = 1 - \alpha$$

Del cálculo de t se tiene

$$t = \frac{150 - 150}{3.65/\sqrt{10}} \therefore t = 0$$

Igual que las demás distribuciones vistas anteriormente, la distribución t de *student*, también cuenta con arreglos tabulares para diferentes valores críticos de esta distribución (Tabla A.4 del apéndice). Haciendo uso de esta tabla para 9 grados de libertad, se tiene que $\alpha = 0.493$. Entonces

$$P(T \leq 0) = 1 - 0.493 = 0.507$$

b) La probabilidad de que la emisión media de NO_2 se encuentre entre 150 y 160 $\mu g/m^3$, se determina por

$$P(150 \leq X \leq 160) = P(t_1 \leq T \leq t_2) = P(T \leq t_2) - P(T \leq t_1)$$

$$P(0 \leq T \leq 3.664) = P(T \leq 3.664) - (P \leq 0) = (1 - \alpha_2) - (1 - \alpha_1)$$

De la Tabla A.4, tenemos que $\alpha_1 = 0.493$ y $\alpha_2 = 0.0026$. Así,

$$P(0 \leq T \leq 3.664) = (1 - 0.0026) - (1 - 0.493)$$

$$P(0 \leq T \leq 3.664) = 0.490$$

c) Por último, la probabilidad de que la concentración de media dióxido de nitrógeno sea mayor que 150 $\mu g/m^3$ es

$$P(X \geq 150) = P(T \geq 0) = 0.493$$

El modelamiento en R de la distribución *t* de *student* se realiza a través de las siguientes funciones y sus argumentos

```
dt(x, df, log = FALSE)
pt(q, df, lower.tail = TRUE, log.p = FALSE)
qt(p, df, lower.tail = TRUE, log.p = FALSE)
rt(n, df)
```

Donde los argumentos de estas funciones, indican lo mismo que las vistas en secciones anteriores en el tratamiento de otras distribuciones de probabilidad.

De esta forma, la solución del ejercicio anterior en R, se consigue ejecutando las siguientes líneas de comando

```
> a<-pt(0,df=9,lower.tail=TRUE)
> a
[1] 0.5
> b<-(pt(3.664,df=9,lower.tail=TRUE)-pt(0,9,lower.tail=TRUE))
> b
[1] 0.4973985
> c<-pt(0,df=9,lower.tail=FALSE)
> c
[1] 0.5
```

Estos resultados, corresponden a una buena aproximación de los cálculos realizados mecánicamente para la resolución de este ejercicio.

3.3.3. Distribución chi-cuadrado (χ^2)

La distribución chi-cuadrado, descubierta por Karl Pearson (1900), es considerada una distribución muestral de la varianza (s^2); es decir, aquella que obtiene si se extraen todas las muestras posibles de una población normal y a cada muestra se le calcula su varianza. Este hecho, explica la importancia de esta distribución en problemas de muestreo de poblaciones con distribución normal.

De lo anterior, si se tiene una muestra aleatoria de una población con distribución normal. Entonces la variable

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma^2}$$

Sigue una distribución chi-cuadrado (χ^2) con $\upsilon = n-1$ grados de libertad, cuya función de densidad de probabilidad está dada por la expresión

$$f(x;\upsilon) = \frac{1}{2^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon}{2}\right)} e^{-\frac{x}{2}} x^{\left(\frac{\upsilon}{2}-1\right)} \tag{3.11}$$

La curva de esta distribución tiene forma de campana con asimetría positiva, que al igual que la distribución *t* de *student*, depende de los grados de libertad (Figura 3.16).

De un análisis de estas curvas, se pueden resaltar las siguientes características de la distribución χ^2.

1. La distribución es asimétrica positiva.
2. A medida que aumenta el tamaño de la muestra la curva es menos asimétrica, aproximándose a una curva normal.
3. Para cada tamaño muestral, se tendrá una distribución χ^2 diferente.
4. El parámetro que caracteriza a una distribución χ^2 son sus grados de libertad (*n* - 1), originado una distribución diferente para cada grado de libertad.

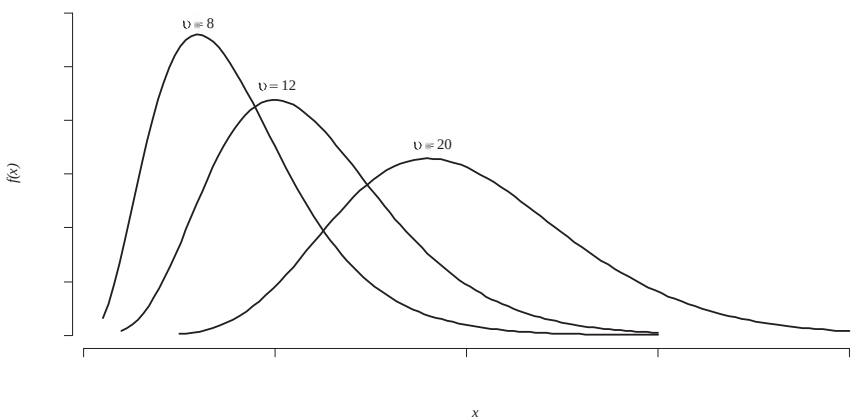


Figura 3.16. Curvas de funciones de densidad de probabilidad χ^2.

Para especificar procedimientos inferenciales que utilizan la distribución χ^2, se requiere una notación análoga a aquella para un valor t crítico $t_{\alpha,\upsilon}$. De esta forma se define $\chi^2_{\alpha,\upsilon}$, **llamado valor critico chi cuadrado**, como el número sobre el eje de medición de forma que el área bajo la curva de la distribución χ^2 con υ grados de libertad y a la derecha de $\chi^2_{\alpha,\upsilon}$, sea igual a α Figura (3.17).

Una vez definidos los valores críticos de la distribución χ^2, proporcionados en la Tabla A.5 del apéndice, se puede establecer la función de distribución acumulada dada mediante la siguiente expresión

$$P(X^2 \leq \chi^2_{1-\alpha,\upsilon}) = \int_0^{\chi^2_{1-\alpha,\upsilon}} f(x;\upsilon)dx \tag{3.12}$$

Sustituir 3.11 en 3.12 y luego realizar la integración, resulta ser una tarea dispendiosa y poco práctica, por tal motivo se hace uso de la Tabla A.5, para realizar el cálculo de probabilidades.

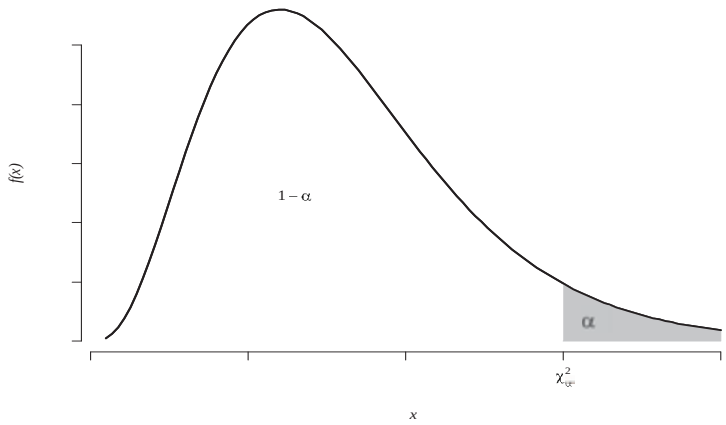


Figura 3.17. Ilustración de la notación adoptada para $\chi^2_{\alpha,\upsilon}$.

Ejemplo 3.8. Los registros históricos de precipitación anual en el municipio de Riohacha, La Guajira, sugieren que dichos datos se distribuyen normalmente con $\sigma = 1.07\,\text{mm}$. Si se selecciona una muestra aleatoria de 6 registros, encuentre la probabilidad de

a) Que la varianza muestral se menor que 2.0 mm.
b) Que la varianza muestral se encuentre entre 0.5 y 1.5 mm.

Solución

a) Como paso inicial, calculamos el valor de la variable aleatoria χ^2.

$$\chi^2 = \frac{(6-1)(2.0)}{(1.07)^2} \therefore \chi^2 = 8.734$$

A partir del valor calculado de χ^2, se halla el valor de α con $\upsilon = 5$ grados de libertad, haciendo uso de la Tabla A.5, encontrado que $\alpha = 0.126$.

Entonces la probabilidad de que la varianza sea menor que 2.0 mm es

$$P(X^2 \leq 8.734) = 1-\alpha$$

$$P(X^2 \leq 8.734) = 1-0.126$$

$$P(X^2 \leq 8.734) = 0.874$$

b) La probabilidad de que la varianza muestral se encuentre entre 0.5 y 1.5 mm, se calcula por

$$P(0.5 \leq s^2 \leq 1.5) = P(\chi_1^2 \leq X^2 \leq \chi_2^2) = P(X^2 \leq \chi_2^2) - P(X^2 \leq \chi_1^2)$$

$$P(2.184 \leq X^2 \leq 6.551) = P(X^2 \leq 6.551) - P(X^2 \leq 2.184) = (1-\alpha_2) - (1-\alpha_1)$$

De la Tabla A.5, tenemos que $\alpha_1 = 0.822$ y $\alpha_2 = 0.257$. Así,

$$P(2.184 \leq X^2 \leq 6.551) = (1-0.257) - (1-0.822)$$

$$P(2.184 \leq X^2 \leq 6.551) = 0.565$$

El modelamiento en R de la distribución χ^2 se realiza a través de las siguientes funciones y sus argumentos

```
dchisq(x, df, log = FALSE)
pchisq(q, df, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df)
```

Donde los argumentos de estas funciones, indican lo mismo que las vistas en secciones anteriores en el tratamiento de otras distribuciones de probabilidad.

De esta forma, la solución del ejercicio anterior en R, se consigue ejecutando códigos de la salida de resultados siguiente

```
> a<-pchisq(8.734,df=5,lower.tail=TRUE)
> a
[1] 0.8798439
>b<-(pchisq(6.551,df=5,lower.tail=TRUE)-
pchisq(2.184,df=5,lower.tail=TRUE))
> b
[1] 0.5669132
```

3.3.4. Distribución F de Fisher-Snedecor

La distribución *F*, es una distribución de probabilidad de gran aplicación en la inferencia estadística, fundamentalmente en el contraste de la igualdad de varianzas de dos poblaciones normales. Estadísticamente, esta distribución se define como el cociente de dos variables aleatorias que siguen una distribución χ^2 con $\upsilon_1 = n_1 - 1$ y $\upsilon_2 = n_2 - 1$ grados de libertad. Así, siendo dos muestras aleatorias independientes de tamaño n_1 y n_2, con varianzas muéstrales s_1^2 y s_2^2, tomadas de poblaciones normales con varianzas poblacionales σ_1^2 y σ_2^2, respectivamente. Entonces, la variable aleatoria

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2}$$

Tiene una distribución *F* con $\upsilon_1 = n_1 - 1$ y $\upsilon_2 = n_2 - 1$ grados de libertad. La función de densidad de probabilidad de la distribución *F*, está dada por la expresión

$$h(x,\upsilon_1,\upsilon_2) = \frac{\Gamma\left(\frac{\upsilon_1+\upsilon_2}{2}\right)\left(\frac{\upsilon_1}{\upsilon_2}\right)^{\frac{\upsilon_1}{2}}}{\Gamma\left(\frac{\upsilon_1}{2}\right)\Gamma\left(\frac{\upsilon_2}{2}\right)} \frac{x^{\frac{\upsilon_1}{2}-1}}{\left(1+\frac{\upsilon_1}{\upsilon_2}x\right)^{\frac{\upsilon_1+\upsilon_2}{2}}} \tag{3.13}$$

Del mismo modo que la distribución χ^2, la distribución *F* presenta asimetría positiva, donde la forma de su curva característica depende de los grados de libertad y del orden en que estos se establezcan (Figura 3.18). Así mismo, la función de distribución acumulada de la distribución *F*, se determina a partir de la siguiente expresión

$$P(F \leq f_{1-\alpha[\upsilon_1,\upsilon_2]}) = \int_0^{f_{1-\alpha[\upsilon_1,\upsilon_2]}} h(x,\upsilon_1,\upsilon_2)dx = 1-\alpha \qquad 0 \leq \alpha \leq 1 \qquad (3.14)$$

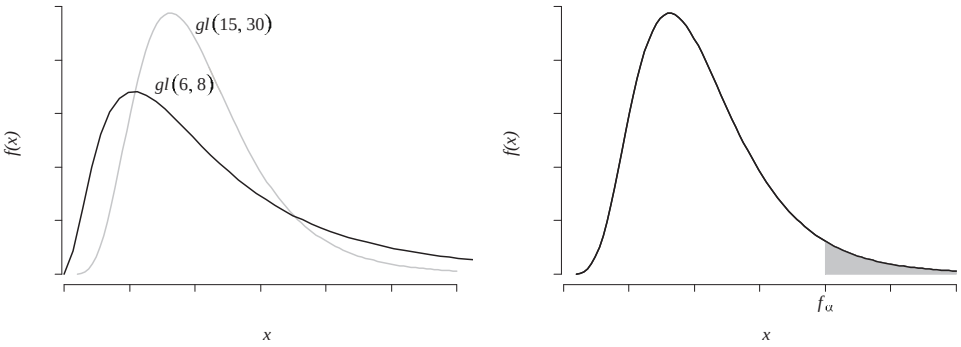
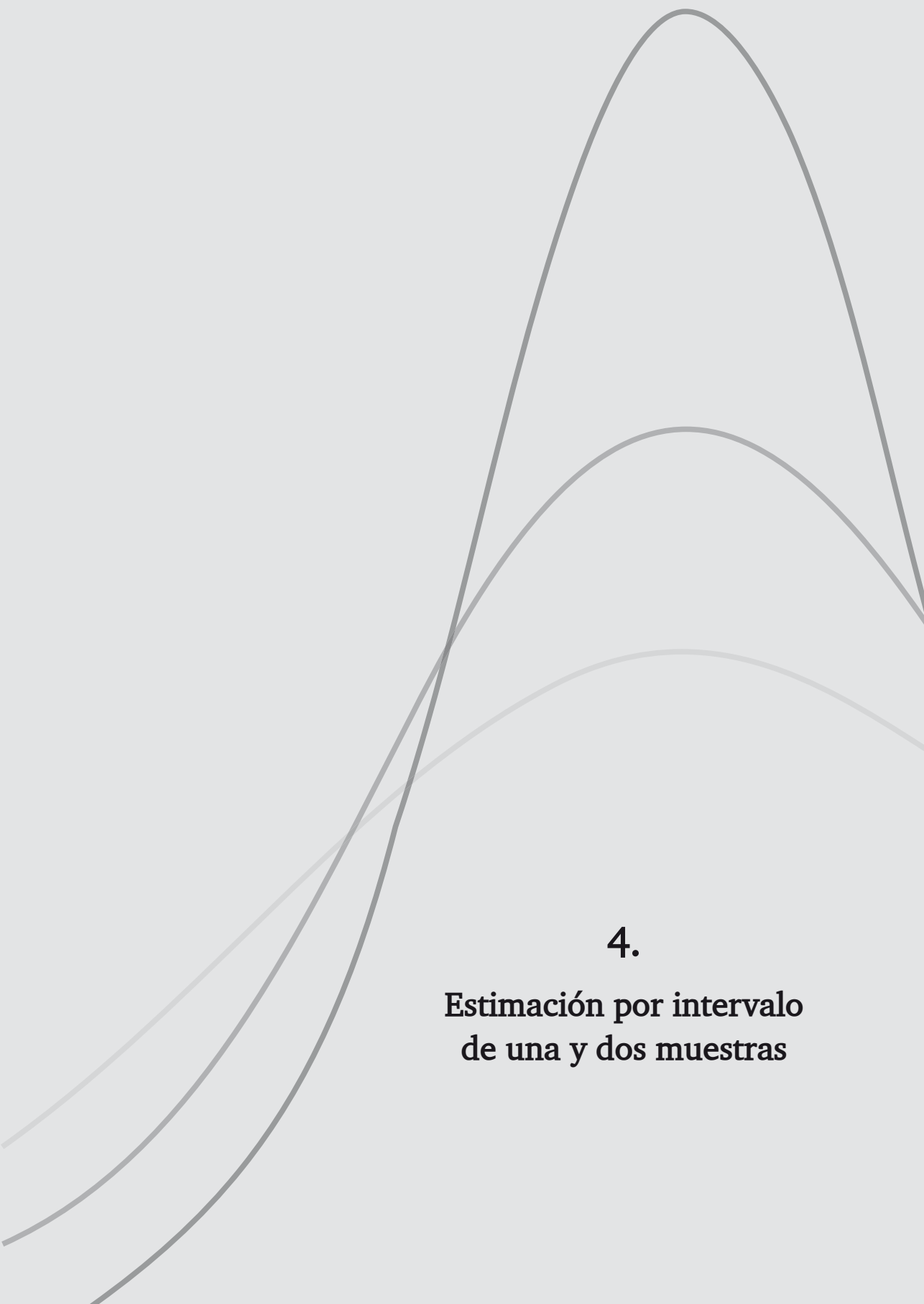


Figura 3.18. Representación gráfica de la distribución *F*.

De la Figura 3.18. f_α es el valor de *f* a la derecha del cual se encuentra un área bajo la curva igual a α (región sombreada). Este valor, se encuentra en tablas elaboradas para la distribución *F* con valores de α de 0.05 y 0.01 y diferentes grados de libertad (Tabla A.6 del apéndice). Sin embargo, es posible encontrar valores de $f_{0.95}$ y $f_{0.99}$, a través de la siguiente relación

$$f_{1-\alpha[\upsilon_1,\upsilon_2]} = \frac{1}{f_{\alpha[\upsilon_2,\upsilon_1]}}$$

En esta sección para la distribución *F* no se darán ejemplos sobre el uso de la misma, pues como se mencionó antes su principal objetivo es el de comprobar la igualdad de varianza de dos poblaciones normales, y en capítulos posteriores cuando se aborde esta temática se mostraran los ejemplos pertinentes.



4.

Estimación por intervalo de una y dos muestras

4.1. Generalidades

En el capítulo 1 se mencionó que el objetivo principal de la estadística inferencial es realizar generalizaciones acerca de ciertas características (parámetros) de la población a partir de los datos muéstrales que se obtengan como resultado de un experimento, es decir, es un proceso de inducción en la que se busca realizar inferencias validas sobre los parámetros poblacionales a partir de los datos de una muestra representativa de la población, basándose en el cálculo de probabilidades.

Genéricamente, algunos métodos de inferencia estadística buscan que una **estimación puntual** de algún parámetro poblacional θ sea un único valor $\hat{\theta}$ de un estadístico $\hat{\Theta}$, denominado **estimador puntual**. Así, la media muestral $\bar{x}$ del estadístico $\bar{X}$, que se calcula de una muestra de tamaño *n*, es una estimación puntual de la media poblacional μ.

Dado que todo experimento o proceso de medición está sometido a errores, de los datos muéstrales no se espera obtener un estimador que realice una estimación exacta del parámetro poblacional que se estudia. Sin embargo, si se espera que el valor de este estimador se encuentre los más cercano posible al verdadero valor del parámetro poblacional. De esta forma se introduce el término de **estimador insesgado**, definido como aquel estimador cuya media es igual al parámetro estimado (Figura 4.1).

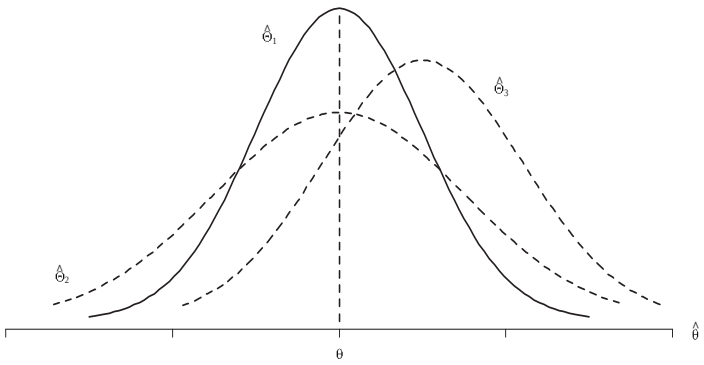


Figura 4.1. Distribuciones muéstrales de estimadores diferentes de θ.

Del gráfico se puede observar que $\hat{\Theta}_1$ y $\hat{\Theta}_2$ se encuentran centrados en θ, por lo tanto son los únicos estimadores que cumplen con la característica de insesgamiento. No obstante $\hat{\Theta}_1$, por tener menor varianza que $\hat{\Theta}_2$, resulta ser más eficaz; de esta forma se puede generalizar que de todos los posibles estimadores insesgados de un parámetro poblacional θ, resulta más eficaz aquel cuya varianza sea más pequeña.

Es improbable que incluso el estimador insesgado más eficaz estime con exactitud el parámetro poblacional (Walpole *et al.*, 2007), aun cuando se aumente el tamaño de la muestra con el objeto de aumentar la precisión. Dado lo anterior en la mayoría de las situaciones prácticas es preferible realizar la estimación a través de la construcción de un intervalo dentro del cual se espera encontrar con mucha certeza el valor del parámetro poblacional, tal intervalo se denomina **intervalo de predicción**.

La estructura de un intervalo de predicción toma la forma $\hat{\theta}_L < \theta < \hat{\theta}_U$, donde $\hat{\theta}_L$ y $\hat{\theta}_U$ son los limites inferior y superior del intervalo, y dependen del valor del estadístico Θ para una muestra específica, y de la distribución de muestreo de Θ. De aquí, queda claro que generalmente muestras distintas, darán valores distintos de Θ y, por tanto, valores diferentes de los extremos $\hat{\theta}_L$ y $\hat{\theta}_U$, que serían valores de las variables aleatorias $\hat{\Theta}_L$ y $\hat{\Theta}_U$. A partir del cálculo de probabilidades, y a través de la distribución muestral de Θ, se podría determinar $\hat{\Theta}_L$ y $\hat{\Theta}_U$, de manera que,

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1-\alpha$$

para $0<\alpha<1$, tenemos entonces una probabilidad de $1-\alpha$ de seleccionar una variable aleatoria que produzca un intervalo que contenga θ (Walpole *et al.*, 2007). El intervalo $\hat{\theta}_L < \theta < \hat{\theta}_U$, que se calcula a partir de los datos de la muestra seleccionada, se denomina **intervalo de confianza** de $(1-\alpha)100\%$, y la fracción $1-\alpha$, es llamada **grado de confianza**. De esta forma cuando $\alpha = 0.05$, tenemos un intervalo de confianza del 95%, y cuando $\alpha = 0.01$ obtenemos un intervalo de confianza más amplio del 99%. Cuanto más amplio sea el intervalo de confianza, tendremos mayor confianza de que el intervalo construido contenga al parámetro desconocido. Sin embargo, es preferible un intervalo corto con un alto grado de confianza.

A continuación, trataremos las técnicas estadísticas usadas para la construcción de intervalos de confianza para la media, proporción y varianza de una población.

4.2. Intervalo de confianza para μ de una población normal con σ conocida

Esta situación en la que se desea construir un intervalo de confianza para la media de una población distribuida normalmente con varianza o desviación estándar conocida es un poco irreal, pues aun cuando la suposición de normalidad es razonable, el desconocimiento de la media poblacional implica, a su vez, el desconocimiento de la variabilidad de la población. Sin embargo, con fines de explicación de un método para la construcción de intervalos de confianza tomaremos este escenario como posible, y más adelante se desarrollaran metodologías para construir intervalos de confianza en situación que requieran de menos restricciones y más adaptadas a los experimentos de la realidad.

De acuerdo con el teorema del límite central (se recomienda consultar a Walpole *et al.*, 2007), la distribución muestral de $\bar{X}$, se encuentra distribuida casi normalmente con media $\mu_{\bar{X}} = \mu$ y deviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Así, al denotar por $z_{\alpha/2}$ al valor de z por encima del cual se encuentra un área de $\alpha/2$, de la figura 4.2 podemos observar que

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha$$

donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Así,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

Luego, multiplicando cada término de la desigualdad por $\sigma/\sqrt{n}$, después restar $\bar{X}$ de cada término y multiplicar toda la desigualdad por -1, para invertir el sentido de la desigualdad, se obtiene

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Ahora, si se selecciona una muestra aleatoria de tamaño *n* de una población distribuida normalmente con varianza conocida σ^2, un intervalo de confianza de para μ se determina a través de

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (4.1)$$

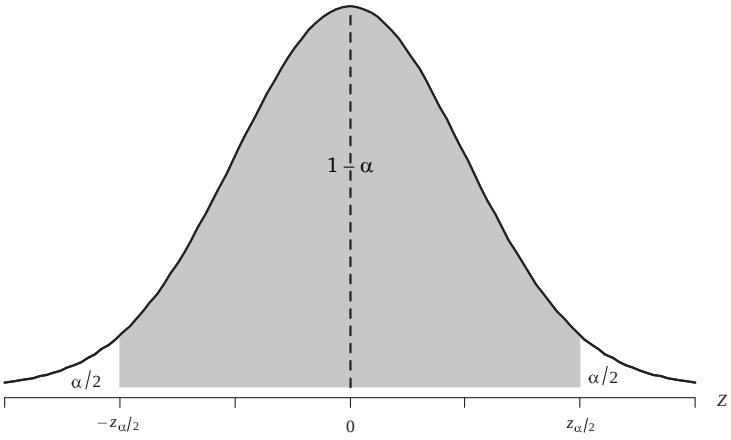


Figura 4.2. $P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha$.

Ejemplo 4.1. En un estudio de contaminación atmosférica se determinó la concentración de partículas suspendidas totales (PST) en el casco urbano del municipio de Guamal, Magdalena, expresada como $\mu g/m^3$ en un tiempo de muestreo de 24 horas. Los datos obtenidos en 14 estaciones de muestreo son los siguientes:

110	120	105	100	140	115	110	121	130	145
110	150	130	120	105	150	140	115	110	120

A partir de estos datos construya un intervalo de confianza del 95% de confiabilidad para la media poblacional de concentración PST. Asúmase

que los datos provienen de una población distribuida normalmente con $\sigma = 18 \mu g/m^3$.

Solución

La estimación puntual de μ, es decir, la media muestral de las concentraciones de PST es $\bar{x} = 123.21 \mu g/m^3$, y el valor de *z*, que deja un área a la derecha de 0.025 y, por lo tanto, un área de 0.975 a la izquierda es $z_{0.025} = 1.96$ (Tabla A.3 de apéndice). De lo anterior y haciendo uso de (4.1), el intervalo de confianza del 95% la para la media poblacional de las concentraciones de PST es

$$123.21 - 1.96\frac{18}{\sqrt{20}} < \mu < 123.21 + 1.96\frac{18}{\sqrt{20}}$$

$$115.32 < \mu < 131.10$$

Es decir, que el verdadero valor medio de las concentraciones de PST en el casco urbano del municipio de Guamal Magdalena, se encuentran entre 115.32 y 131.10 $\mu g/m^3$, con una confiabilidad del 95%.

En R, no existe una función específica para la construcción de intervalos de confianza, puesto que las funciones utilizadas para las pruebas hipótesis que se verán más adelante, arrojan dichos intervalos en sus salidas de resultados. Sin embargo, se pueden utilizar líneas de código para determinar los intervalos de confianza, considerando la plataforma de R como una calculadora, como veremos a continuación, con los datos del ejemplo anterior.

```
> PST<-
c(110,120,105,100,140,115,110,121,130,145,110,150,130,120,
115,150,140,110,120)
> Lim.Inf<-mean(PST)-qnorm(0.975)*(18/sqrt(length(PST)))
> Lim.Sup<-mean(PST)+qnorm(0.975)*(18/sqrt(length(PST)))
> Inter.Conf<-c(Lim.Inf,Lim.Sup)
> Inter.Conf
[1] 115.1169 131.3042
```

Note que los resultados solo varían un poco por efectos del redondeo de las cifras cuando se realizan los cálculos mecánicamente.

4.3. Intervalo de confianza para μ de una población normal con σ desconocida a través de una muestra pequeña ($n < 30$)

Una situación más real a la planteada en la sección anterior es la de realizar estimaciones acerca de la media de una población cuando se desconoce su varianza y del mismo modo su desviación estándar. Este problema se soluciona fácilmente sustituyendo la distribución normal por la distribución *t* de *student* con $n - 1$ grados de libertad y reemplazando σ por *s*, la desviación estándar de la muestra, en adelante se sigue el mismo procedimiento visto anteriormente. De acuerdo a la Figura 4.3, se tiene que

$$P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1-\alpha$$

Donde $t_{\alpha/2}$ es el valor de *t* con $n - 1$ grados de libertad, a la derecha del cual se encuentra un área de $\alpha/2$ y, debido a la simetría de la distribución *t* de *student*, un área igual a $\alpha/2$ se encuentra a la izquierda de $-t_{\alpha/2}$.

Como

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Entonces

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$$

Multiplicando cada término de la desigualdad por $S/\sqrt{n}$, y después de restar $\bar{X}$ de cada termino y multiplicar con -1, obtenemos

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Así, para una muestra aleatoria de tamaño *n*, se calcula la media muestral $\bar{x}$, la desviación estándar s y se obtiene un intervalo de confianza de $(1-\alpha)100\%$ para μ a través de la expresión

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{4.2}$$

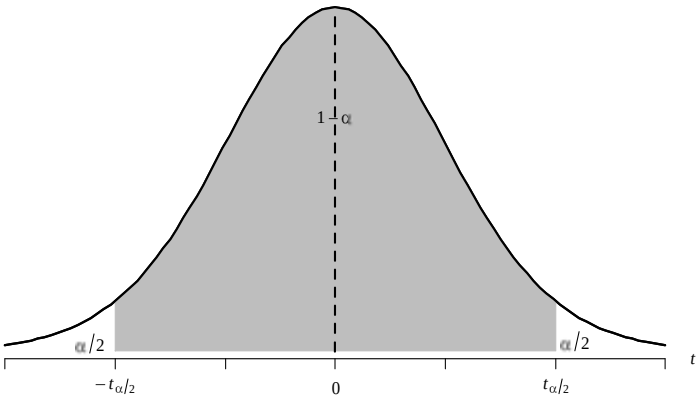


Figura 4.3. $P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1-\alpha$

Ejemplo 4.2. Con el objetivo de evaluar la eficiencia u operación de un rector anaerobio de flujo ascendente (UASB, por sus siglas en ingles), que trata las aguas residuales domesticas de un sector residencial de la ciudad de Barranquilla, se está llevando a cabo el análisis de ciertas variables fisicoquímicas entre las que se encuentra la DQO, parámetro importante en la determinación de la eficiencia del reactor. Los resultados del porcentaje de remoción de DQO se muestran a continuación

61.5	50	25	44.4	25	25	50	57.1	50
50	50	16.6	50	50	66.6	75	75	66.6

Asumiendo que los datos provienen de una población con distribución normal, construya un intervalo de confianza del 99%.

Solución

Para el presente ejercicio, se tiene que la estimación puntal para μ y σ, son respectivamente $\bar{x} = 49.32$ y $s = 17.15$. Así mismo, el valor de *t*, que deja un área de 0.005 a la derecha de él, es $t_{0.005(17)} = 2{,}898$. De aquí que el intervalo de confianza de 99% solicitado sea

$$49.32 - 2{,}898\frac{17.15}{\sqrt{18}} < \mu < 49.32 + 2{,}898\frac{17.15}{\sqrt{18}}$$

$$37.61 < \mu < 61.03$$

El intervalo de confianza calculado nos muestra que el verdadero valor medio del porcentaje remoción de DQO en el reactor UASB estudiado, se encuentra entre 37.61 y 61.03%, con una confiabilidad del 99%.

El modelado en R de intervalos de confianza para la media cuando se desconoce su desviación estándar, se puede realizar a través de las siguientes órdenes

```
> REM.DQO<-
c(61.5,50,25,44.4,25,25,50,57.1,50,50,50,16.6,50,50,
66.6,75,75,66.6)
> Lim.Inf<-mean(REM.DQO)-qt(0.995,length(REM.DQO)-
1)*(sd(REM.DQO)/ sqrt(length(REM.DQO)))
> Lim.Sup<-mean(REM.DQO)+qt(0.995,length(REM.DQO)-
1)*(sd(REM.DQO)/ sqrt(length(REM.DQO)))
> Inter.Cof<-c(Lim.Inf,Lim.Sup)
> Inter.Cof
[1] 37.60983 61.03461
```

4.4. Intervalo de confianza para μ de una muestra grande ($n \geq 30$)

Con mucha frecuencia se recomienda que aun cuando no es razonable suponer la normalidad de los datos, con desconocimiento de la varianza poblacional σ, y con un número de observaciones mayor o igual a 30 ($n \geq 30$), s puede reemplazar a σ, y la construcción de un intervalo de confianza de $(1-\alpha)100\%$ para la media de la población queda dada por la expresión

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{4.3}$$

La justificación a lo anterior, se sustenta en la presunción de que al tener una muestra tan grande como 30 y se cuente con una distribución de la población no sesgada, s estará muy cerca del verdadero valor de σ. Es de

destacar que conforme se haga más grande el tamaño de la muestra, aumentara la calidad de los resultados en la estimación de μ.

Ejemplo 4.3. Las siguientes puntuaciones representan la calificación en el examen final para un curso de estadística elemental. Con base en ellos construya un intervalo de confianza de 95% para la media μ.

23	60	79	32	57	74	52	70	82
36	80	77	81	95	41	65	92	85
55	76	52	10	64	75	78	25	80
98	81	67	41	71	83	54	64	72
88	62	74	43	60	78	89	76	84
48	84	90	15	79	34	67	17	82
69	74	63	80	85	61			

Solución

El cálculo previo de la media y desviación estándar muestral, como estimadores puntuales de μ y σ, arroja como resultado que $\bar{x} = 65.48$ y $s = 21.13$. Del mismo modo $z_{\alpha/2} = 1.96$. Así el intervalo de confianza solicitado se determina aplicando la ecuación brindada en (4.3)

$$65.48 - 1.96\frac{21.13}{\sqrt{60}} < \mu < 65.48 + 1.96\frac{21.13}{\sqrt{60}}$$

$$60.13 < \mu < 70.83$$

Es decir, que el verdadero valor medio de las calificaciones del examen final del curso de estadística elemental, se encuentra entre 60.13 y 70.83, con una confiabilidad del 95%.

Para modelar este ejemplo en R, dada la cantidad de datos y la dificultad que representa condensarlos en un vector de datos, crearemos un archivo en Excel que llamaremos "*Estadística*", guardado bajo la extensión ***.csv***, cuyo encabezado de la variable se denominara "*Notas*", para luego cargarlo en R a través de la función ***read.csv2***, como se ha visto en apartados anteriores.

```
> Estadistica<-read.csv2("Estadistica.csv",header=TRUE,
encoding="latin1")
> Lim.Inf<-mean(Estadistica$Notas)-qnorm(0.975)*
(sd(Estadistica$Notas)/sqrt(length(Estadistica$Notas)))
> Lim.Sup<-mean(Estadistica$Notas)+qnorm(0.975)*
(sd(Estadistica$Notas)/sqrt(length(Estadistica$Notas)))
> Inter.Conf<-c(Lim.Inf,Lim.Sup)
> Inter.Conf
[1] 60.13591 70.83076
```

4.5. Intervalo de confianza para $\mu_1 - \mu_2$; con σ_1^2 y σ_2^2 conocidas

En muchos fenómenos de la realidad se está interesado en el estudio de la diferencia de cierta característica medida en dos poblaciones con el objeto de realizar comparaciones e inferencias sobre dicha diferencia. Un ejemplo muy frecuente de lo anterior, es estar interesado en realizar estimaciones sobre las diferencias de las medias de dos poblaciones.

Una estimación puntual de $\mu_1 - \mu_2$, se obtiene seleccionando dos muestras aleatorias independientes, una de cada población, de tamaños n_1 y n_2, y calculando la diferencia $\bar{x}_1 - \bar{x}_2$ de las medias muéstrales. Ahora, análogo a los procedimientos que involucran una sola muestra, la distribución muestral de la diferencia $\bar{X}_1 - \bar{X}_2$, se espera que se encuentre distribuida de forma aproximadamente normal con media $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ y desviación estándar $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Por lo tanto, se puede aseguran con una probabilidad de $1-\alpha$ que la variable aleatoria normal estándar

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Caerá entre $-z_{\alpha/2}$ y $z_{\alpha/2}$ (Walpole *et al.*, 2007). De manera que

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha$$

Al sustituir Z y realizar las operaciones algebraicas correspondientes, similar a las expuestas en el apartado de estimaciones para una muestra,

se llega a la siguiente expresión para la construcción de intervalos de confianza de $(1-\alpha)100\%$ para $\mu_1-\mu_2$ cuando σ_1^2 y σ_2^2 conocen.

$$(\bar{x}_1-\bar{x}_2)-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}<\mu_1-\mu_2<(\bar{x}_1-\bar{x}_2)+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \qquad (4.4)$$

Es preciso comentar que la calidad de la estimación es más precisa cuando las muestras se seleccionan de poblaciones con distribuciones normales. Para poblaciones no normales se pueden tener buenas aproximaciones cuando se aumenta el tamaño de la muestra.

Ejemplo 4.4. Durante los años 2004 a 2005 el grupo de investigación Pichihuel de la Universidad de La Guajira, realizó un estudio sobre la dinámica fisicoquímica del ecosistema estuarino el Riito. Los datos referentes a las concentraciones de ortofosfatos (mg/L), desde noviembre de 2004 a septiembre de 2005, en dos estaciones de muestreo se muestran a continuación

	Concentración de ortofosfatos (mg/L)	
Meses	E_1	E_2
Nov	2.43	1.73
Dic	2.87	0.69
Ene	2.31	1.66
Feb	2.12	1.10
Mar	2.42	0.68
Abr	2.04	0.55
May	2.49	1.43
Jun	2.90	2.06
Jul	2,48	2.59
Ago	2.10	1.50
Sep	2.65	2.10

A partir de estos datos, se desea construir un intervalo de confianza de 95% para la diferencia de las medias de las concentraciones de PO_4 en las dos estaciones de muestreo. Asúmase que los datos provienen de poblaciones con distribución normal con $\sigma_{E_1}=0.31\text{mg/L}$ y $\sigma_{E_2}=0.66\text{mg/L}$.

Solución

Las determinaciones de las medias muéstrales de las concentraciones de ortofosfatos en las estaciones E_1 y E_2, son respectivamente $\bar{x}_{E_1}=2.44$ y

$\bar{x}_{E_2} = 1.46\,\text{mg/L}$. Así, el intervalo de confianza para la diferencia de las medias de las concentraciones de ortofosfatos, se determina de la siguiente forma

$$(2.44-.146)-1.96\sqrt{\frac{(0.31)^2}{11}+\frac{(0.66)^2}{11}} < \mu_1-\mu_2 < (2.44-.146)+1.96\sqrt{\frac{(0.31)^2}{11}+\frac{(0.66)^2}{11}}$$

$$0.55 < \mu_1-\mu_2 < 1.41$$

Cuya interpretación literal sería que la verdadera diferencia entre las medias de las concentraciones de ortofosfatos en las dos estaciones de muestreo ubicadas en el Riito se encuentra entre 0.55 y 1.41 mg/L, con una confiabilidad del 95%.

En R, las líneas de comando para la resolución de este tipo de problemas se muestran a continuación

```
> E1<-
c(2.43,2.87,2.31,2.12,2.42,2.04,2.49,2.90,2.48,2.10,2.65)
> E2<-
c(1.73,0.69,1.66,1.10,0.68,0.55,1.43,2.06,2.59,1.50,2.10)
> Lim.Inf<-(mean(E1)-mean(E2))-qnorm(0.975)*
sqrt((0.31)^2/11+(0.66)^2/11)
> Lim.Sup<-(mean(E1)-mean(E2))+qnorm(0.975)*
sqrt((0.31)^2/11+(0.66)^2/11)
> Inter.Conf<-c(Lim.Inf,Lim.Sup)
> Inter.Conf
[1] 0.5436369 1.4054540
```

4.6. Intervalo de confianza para $\mu_1 - \mu_2$; con $\sigma_1^2 = \sigma_2^2$ pero desconocidas.

Generalmente en los fenómenos estudiados de la realidad, no es posible conocer con exactitud σ_1^2 y σ_2^2. Sin embargo, si las dos muestras implicadas provienen de poblaciones con distribuciones aproximadamente normales, de nuevo la distribución *t* de *Student*, resultará ser de utilidad para el modelado de este tipo de problemas.

Cuando σ_1^2 y σ_2^2 se desconocen, pero $\sigma_1^2 = \sigma_2^2 = \sigma^2$, demostrado a través de un proceso de prueba de hipótesis sobre la homogeneidad de varianzas de dos poblaciones, que discutiremos en capítulos siguientes, la construcción de

intervalos de confianza de $(1-\alpha)100\%$ para la diferencia de las medias de las dos poblaciones estudiadas, se realiza a través de la siguientes expresión

$$(\bar{x}_1-\bar{x}_2)-t_{\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}<\mu_1-\mu_2<(\bar{x}_1-\bar{x}_2)+t_{\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \quad (4.5)$$

Donde $s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$, es la estimación de la unión de las desviaciones estándar poblacionales y $t_{\alpha/2}$ es el valor de t con $\upsilon=n_1+n_2-2$ grados de libertad que deja un área de $\alpha/2$ a la derecha del mismo.

Ejemplo 4.5. Se reportaron las concentraciones atmosféricas de SO_2 (en ppm) provenientes de dos muestreadores localizados a diferentes distancias de una fuente industrial emisora. A partir de los datos que se muestran a continuación construya un intervalo de confianza del 95% para la diferencia de las concentraciones medias de SO_2 en los dos muestreadores. Asúmase que los datos provienen de poblaciones normales con varianzas iguales, aunque desconocidas.

Concentración de SO_2 (ppm)	
Muestreador A	**Muestreador B**
680	500
630	510
620	490
600	530

Solución

Las estimaciones para la media y la varianza de las emisiones de SO2 determinadas en los muestreadores A y B, respectivamente son $\bar{x}_A=632.5$, $\bar{x}_B=507.5$, $s_A^2=1158.33$ y $s_B^2=291.67$. Así mismo, $t_{\alpha/2}=2.447$, es el valor de t con $\upsilon=4+4-6=6$ grados de libertad, que deja una área de 0.025 a la derecha. De esta forma el valor de la estimación de la varianza común queda dado por

$$s_p=\sqrt{\frac{(4-1)*1158.33+(4-1)*621.67}{4+4-2}}=26.93$$

y el intervalo de confianza del 95% para la diferencia de las medias de las concentraciones de SO2 en los dos muestreadores, queda determinado por

$$(632.5-507.5)-(2.447)(26.93)\sqrt{\frac{1}{4}+\frac{1}{4}}<\mu_1-\mu_2<(632.5-507.5)+(2.447)(26.93)\sqrt{\frac{1}{4}+\frac{1}{4}}$$

$$78.40<\mu_1-\mu_2<171.59$$

Es decir, que la verdadera diferencia entre las medias de las concentraciones de SO_2 en los dos muestreadores se encuentra entre 78.40 y 171.59 ppm, con una confiabilidad del 95%.

El modelamiento en R de este tipo de problemas de realiza a través de las siguientes sentencias de comando.

```
> A<-c(680,630,620,600)
> B<-c(500,510,490,530)
> sp<-sqrt(((length(A)-1)*var(A)+(length(B)-
1)*var(B))/(length(A)+ length(B)-2))
> Lim.Inf<-(mean(A)-mean(B))-qt(0.975,length(A)+length(B)-
2)*sp* sqrt((1/length(A))+(1/length(B)))
> Lim.Sup<-(mean(A)-mean(B))+qt(0.975,length(A)+length(B)-
2)*sp* sqrt((1/length(A))+(1/length(B)))
> Conf.Inter<-c(Lim.Inf,Lim.Sup)
> Conf.Inter
[1]  78.41219 171.58781
```

4.7. Intervalo de confianza para $\mu_1-\mu_2$; con $\sigma_1^2\neq\sigma_2^2$ y desconocidas

Si ahora consideramos el caso en que se quiera construir intervalos de confianza para $\mu_1-\mu_2$, de poblaciones con distribución aproximadamente normal, con desconocimiento de sus varianzas poblacionales, y sumado a esto, que estas no sean iguales como ocurrió en la sección anterior, la expresión matemática para construir dichos intervalos de $(1-\alpha)100\%$, a partir de muestras de tamaño n_1 y n_2, está dada por

$$(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (4.7)$$

Donde $t_{\alpha/2}$ es el valor de t con

$$\upsilon = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left[\left(s_1^2/n_1\right)^2/(n_1 - 1)\right] + \left[\left(s_2^2/n_2\right)^2/(n_2 - 1)\right]}$$

grados de libertad, que deja un área de $\alpha/2$ a la derecha. Como rara vez el cálculo de υ dará como resultado un entero, es preciso hacer un redondeo de las cifras decimales al menor entero más cercano.

Ejemplo 4.6. Dentro del Programa de Calidad Ambiental de Playas Turísticas- CAPT, del cual la universidad de La Guajira, a través del grupo de investigación Pichihuel, tiene participación, se han desarrollado monitoreos con el objeto de evaluar la calidad de las playas del municipio de Riohacha, donde la humedad es una de las variables tenidas en cuenta como factor que puede condicionar la presencia de ciertos grupos bacterianos y fúngicos en la arena de playa. Para el estudio de este parámetro de realizaron mediciones *in situ* a través de un sensor en muestras de arena húmeda y arena seca. Los resultados de estas mediciones se consignan en la siguiente tabla.

% de humedad	
Arena húmeda	Arena seca
70	73
53	57
42	48
80	76
79	83
81	64
77	77
79	60
76	44
66	41

A partir de estos datos construir un intervalo de confianza del 99% para la diferencia de las medias del porcentaje de humedad, en los dos tipos de arena. Asuma que los datos provienen de poblaciones con distribuciones normales y varianzas desconocidas y diferentes.

Solución

Las determinaciones de las medias y las desviaciones muéstrales son respectivamente $\overline{x}_{AH} = 78.3$, $\overline{x}_{AS} = 62.3$, $s^2_{AH} = 36.46$ y $s^2_{AS} = 219.57$. Entonces

$$\upsilon = \frac{\left(\frac{36.46}{10} + \frac{219.57}{10}\right)}{\left[(36.46/10)^2/(10-1)\right] + \left[(219.57/10)^2/(10-1)\right]}$$

$$\upsilon = 11.9 \approx 12$$

Así, $t_{0.005\,(12)} = 3.055$, es decir, el valor de *t* que deja un área de 0.005 a la derecha. Por último, el intervalo de confianza para la diferencia de las medias del porcentaje de humedad en muestras de arena húmeda y arena seca en las playas del municipio de Riohacha, está dado por

$$(78.3-62.3) - 3.055\sqrt{\frac{36.46}{10} + \frac{219.57}{10}} < \mu_1 - \mu_2 < (78.3-62.3) + 3.055\sqrt{\frac{36.46}{10} + \frac{219.57}{10}}$$

$$0.54 < \mu_1 - \mu_2 < 31.46$$

Es decir, que la verdadera diferencia entre las medias del porcentaje de humedad en muestras de arena húmeda y arena seca en las playas del municipio de Riohacha, se encuentra entre 0.54 y 31.46%, con una confiabilidad del 99%.

Los comandos en R, para modelar intervalos de confianza para la diferencia de las medias de dos poblaciones normales cuando se desconocen sus varianzas y se tiene certeza que son diferentes, se muestran a continuación

```
> AH<-c(70,73,72,80,89,81,77,79,76,86)
> AS<-c(73,57,48,76,83,64,77,60,44,41)
> v<-
round(((var(AH)/length(AH))+(var(AS)/length(AS)))^2/(((var(AH)
/length(AH))^2/(length(AH)-1))+((var(AS)/length(AS))^2)/
(length(AS)-1)),0)
> v
[1] 12
> Lim.Inf<-(mean(AH)-mean(AS))-
qt(0.995,v)*sqrt((var(AH)/length(AH) )+(var(AS)/length(AS)))
> Lim.Sup<-(mean(AH)-
mean(AS))+qt(0.995,v)*sqrt((var(AH)/length(AH)
)+(var(AS)/length(AS)))
> Inter.Conf<-c(Lim.Inf,Lim.Sup)
> Inter.Conf
[1]  0.5444455 31.4555545
```

Note que nuevamente se usó la función ***round***, descrita en capítulos anteriores, para realizar el redondeo de las cifras decimales del cálculo de los grados de libertad para la determinación del valor de *t*, en este caso a cero cifras decimales, es decir, al valor entero más cercano.

4.8. Intervalo de confianza para una proporción *p* de una muestra grande

En esta sección abordaremos lo referente a la estimación por intervalos de la verdadera proporción *p*. Para ello, iniciaremos con dejar por sentado que una estimación puntual de la proporción *p* en un experimento binomial está dado por el estadístico $\hat{P} = X/n$, donde *X* representa el número de éxitos en *n* intentos. Por lo tanto, la proporción muestral $\hat{p} = x/n$ se utiliza como la estimación puntual del parámetro *p* (Walpole *et al.*, 2007).

De esta forma y sin el ánimo de profundizar en demostraciones matemáticas, si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño *n*, y $\hat{q} = 1 - \hat{p}$, un intervalo de confianza aproximado de $(1-\alpha)100\%$ para el parámetro binomial *p* es:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad (4.8)$$

Donde $z_{\alpha/2}$ es el valor de *z* con un área de $\alpha/2$ a la derecha.

Adviértase que cuando n es pequeña y se cree que la proporción desconocida p se acerca a 0 o a 1, la expresión anterior para el intervalo de confianza no es confiable, y por lo tanto, no debe ser utilizado. Para tener seguridad de cuando utilizar este procedimiento, se requiere que ambos $n\hat{p}$ y $n\hat{q}$ sean mayores o iguales a 5 (Walpole *et al.*, 2007).

Ejemplo 4.7. Supóngase que en cierta ciudad solo 54 de las 178 industrias que existen cumplen con la normativa ambiental vigente sobre concentraciones máximas de material particulado (PST) emitidas a la atmosfera. Con base en lo anterior, encuentre un intervalo de confianza del 95% para la verdadera proporción de industrias que cumplen con el marco normativo ambiental vigente de la ciudad en cuestión.

Solución

Con $n = 178$ y $x = 54$, la estimación puntual de p, sería $\hat{p} = 54/178 = 0.303$, $\hat{q} = 1 - 0.303 = 0.675$ y $z_{0.025} = 1.96$. Así, el intervalo de confianza que se pide encontrar está dado por

$$0.303 - 1.96\sqrt{\frac{(0.303)(0.697)}{178}} < p < 0.303 - 1.96\sqrt{\frac{(0.303)(0.697)}{178}}$$

$$0.235 < p < 0.371$$

Es decir, que la verdadera proporción de industrias que cumplen con la normativa ambiental vigente que regula la emisión de PST a la atmosfera se encuentra entre 23.5 y 37.1%, con una confiabilidad del 95%.

La resolución mediante código en la plataforma de R, para problemas que involucra la estimación por intervalos de la verdadera proporción se muestra a continuación.

```
> x<-54
> n<-178
> p<-x/n
> q<-(1-p)
> Lim.Inf<-p-qnorm(0.975)*sqrt((p*q)/n)
> Lim.Sup<-p+qnorm(0.975)*sqrt((p*q)/n)
> Inter.Conf<-c(Lim.Inf,Lim.Sup)
> Inter.Conf
[1] 0.2358363 0.3709053
```

4.9. Intervalo de confianza para la diferencia entre dos proporciones para muestras grandes

Al igual que en el estudio de medias poblacionales, existen situaciones en las ciencias experimentales en las que se desea realizar estimaciones de las proporciones de dos poblaciones en las que a través de un proceso de muestreo se midió una característica en particular. Una de estas estimaciones es la construcción de intervalos de confianza del $(1-\alpha)100\%$ para la diferencias entre dos proporciones p_1 y p_2. Para ello, se seleccionan muestras aleatorias independientes de tamaños n_1 y n_2, a partir de las dos poblaciones con medias $n_1 p_1$ y $n_2 p_2$ y varianzas $n_1 p_1 q_1$ y $n_2 p_2 q_2$, respectivamente. Luego se determina el número de éxitos (posee o cumple con la característica o atributo medido) x_1 y x_2, y se forman las proporciones $\hat{p}_1 = x_1/n_1$ y $\hat{p}_2 = x_2/n_2$.

Con base en lo anterior, un estimador puntual de la diferencia entre las verdaderas proporciones $p_1 - p_2$, está dado por el estadístico $\hat{P}_1 - \hat{P}_2$. Por lo tanto, la diferencia de las proporciones muéstrales, $\hat{p}_1 - \hat{p}_2$, se utilizan como la estimación puntual de $p_1 - p_2$. De esta forma la estimación por intervalo del $(1-\alpha)100\%$ para $p_1 - p_2$, está dada por la expresión

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \quad (4.9)$$

Donde $\hat{p}_1$ y $\hat{p}_2$ son las proporciones de éxitos en las muestras aleatorias de tamaño n_1 y n_2, respectivamente, y $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$, las proporciones de incumplimiento de la característica medida. $z_{\alpha/2}$ es el valor de *z* que deja un área de $\alpha/2$ a la derecha.

Ejemplo 4.8. Una planta de tratamiento de agua potable desea conocer la diferencia entre el método de desinfección con cloro gaseoso y luz ultravioleta en función de la presencia de *E. coli* en muestras de aguas a la salida de esta etapa del sistema de tratamiento. Para tal fin, se seleccionaron aleatoriamente 80 muestras a la salida del sistema luego de utilizar los dos métodos de desinfección mencionados, donde se encontró de para el sistema de desinfección con cloro gaseoso 25 muestras fueron positivas al examen de *E. coli* y para el método de desinfección con luz ultravioleta 18 resultaron ser positivas. Con base en lo anterior, encuentre

un intervalo de confianza del 95% para la diferencia entre las proporciones de la presencia de *E. coli* del agua a la salida de la planta de tratamiento.

Solución

Con $n_1 = n_2 = 80$, $x_1 = 25$ y $x_2 = 14$, las estimaciones puntuales de p_1 y p_2, son respectivamente $\hat{p}_1 = 0.313$ y $\hat{p}_2 = 0.175$, $\hat{q}_1 = 0.687$, $\hat{q}_2 = 0.825$ y $z_{0.025} = 1.96$. Así, el intervalo de confianza para la diferencia de las proporciones de presencia de *E. coli* en muestras de aguas a la salida de la etapa de desinfección de la planta de tratamiento usando los métodos de desinfección con cloro gaseoso y luz ultravioleta está dado por,

$$(0.313-0.175)-1.96\sqrt{\frac{(0.313)(0.687)}{80}+\frac{(0.175)(0.825)}{80}} < p_1 - p_2 < (0.313-0.175)-1.96\sqrt{\frac{(0.313)(0.687)}{80}+\frac{(0.175)(0.825)}{80}}$$

$$0.06 < p_1 - p_2 < 0.269$$

Como se observa, la diferencia entre la verdadera proporción de la presencia de *E. coli* en muestras de agua a la salida de la etapa de desinfección de la planta de tratamiento, utilizando cloro gaseoso y luz ultravioleta se encuentra entre 6 y 26.9%, con una confiabilidad del 95%.

El conjunto de códigos en R para la construcción de intervalos de confianza para la diferencia entre dos proporciones, se muestras en la salida de resultados siguiente

```
> x1=25
> n1=80
> x2=14
> n2=80
> p1<-x1/n1
> q1<-1-p1
> p2<-x2/n2
> q2<-1-p2
> Lim.Inf<-(p1-p2)-
qnorm(0.975)*sqrt(((p1*q1)/n1)+((p2*q2)/n2))
> Lim.Sup<-(p1-
p2)+qnorm(0.975)*sqrt(((p1*q1)/n1)+((p2*q2)/n2))
> Inter.Conf<-c(Lim.Inf,Lim.Sup)
> Inter.Conf
[1] 0.00616436 0.26883564
```

4.3. Intervalo de confianza para la varianza de una población normal

Aun cuando los métodos inferenciales sobre la varianza o la desviación estándar de una población, son de menor interés que aquellos respecto a la media o proporción, hay ocasiones en las que se requieren dichos procedimientos para conocer la concentración de valores alrededor de la media (Devore, 2008; Guisande *et al.,* 2011). A continuación, desarrollaremos intuitivamente una expresión para la construcción de intervalos de confianza para la varianza de una población normal.

Si se extrae una muestra de tamaño *n* de una población normal con varianza σ^2 y se calcula la varianza muestral s^2, se obtiene un valor del estadístico S^2, que representa una estimación puntual de σ^2. De esta forma y basándonos en lo comentado en secciones anteriores, se puede establecer una estimación por intervalos de σ^2 utilizando el estadístico

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

De manera que X^2 tiene una distribución chi-cuadrado con $\upsilon = n-1$ grados de libertad, cuando las muestras se eligen de una población con distribución normal. De esta forma tenemos que

$$P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$$

Donde $\chi^2_{1-\alpha/2}$ y $\chi^2_{\alpha/2}$ son valores de la distribución chi cuadrado con $\upsilon = n-1$ grados de libertad, que dejan un área de $1-\alpha/2$ y $\alpha/2$, respectivamente, a la derecha (Figura 4.4). Al sustituir para X^2, tenemos

$$P\left[\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right] = 1-\alpha\,.$$

Ahora, al dividir cada término de la desigualdad entre $(n-1)S^2$ y, después, invertir cada termino (ocasionando un cambio en el sentido de las desigualdades), obtenemos

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha$$

De esta forma, para una muestra aleatoria de tamaño *n*, se calcula s^2 y se obtiene el intervalo de confianza de $(1-\alpha)100\%$ para σ^2 con la expresión

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \tag{4.10}$$

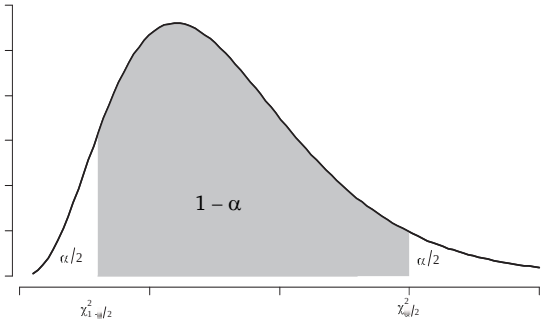


Figura 4.4. $P\left(\chi^2_{1-\alpha/2} < \mathrm{X}^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$.

Ejemplo 4.9. Supóngase que se ha realizado un estudio para estimar las características del terreno inundable de una parte de cierto rio. Una de las variables es la anchura del terreno inundable. De una muestra de las mediciones obtenidas en 61 lugares seleccionados aleatoriamente se obtuvo como resultado $s = 100$ m. Construya un intervalo de confianza del 99% para la varianza de anchura del terreno inundable en la región del rio estudiado. Asúmase que las mediciones de la variable estudiada siguen una distribución normal (Milton, 2004).

Solución

Para este ejemplo, se tiene que una estimación puntual de σ^2 es $s^2 = 10000\ \mathrm{m}^2$ y, de la Tabla A.5 del apéndice, con $\upsilon = 60$ grados de libertad, tenemos que $\chi^2_{0.05} = 79.082$ y $\chi^2_{0.95} = 43.188$. Así, el intervalo pedido, queda determinado por

$$\frac{(61-1)(10000)}{79.082} < \sigma^2 < \frac{(61-1)(10000)}{43.188}$$

$$7587.06 < \sigma^2 < 13892.75$$

De esta manera se concluye con un 90% de confiabilidad que la verdadera varianza de anchura del terreno inundable en la región del rio estudiado se encuentra entre 7587.06 y 13892.75 m^2, o que su desviación estándar poblacional se encuentra entre 87.10 y 117.87 m, para facilitar la compresión al expresar los resultados en las mismas unidades de medida de la variable de estudio.

El conjunto de comandos utilizados en R para la construcción de intervalos de confianza para la varianza de una población con distribución normal, se muestran a continuación

```
> s=100
> s2=s^2
> n=61
> Lim.Inf<-((n-1)*s2)/qchisq(0.95,n-1)
> Lim.Sup<-((n-1)*s2)/qchisq(0.05,n-1)
> Inter.Conf<-c(Lim.Inf,Lim.Sup)
> Inter.Conf
[1]  7587.067 13892.761
```

4.4. Intervalo de confianza para la razón de dos varianzas de poblaciones normales

Como se hizo mención en la sección anterior, existen situaciones experimentales en las que es necesario realizar inferencias sobre la varianza o la desviación estándar de poblaciones normales donde se realiza el estudio de una variable en particular, con el objeto de realizar comparaciones entre la variabilidad de las dos poblaciones o debido a que existen métodos inferenciales comparativos en los que es necesario conocer de antemano que las varianzas de las poblaciones son iguales o diferentes, ejemplo de ello lo constituyen las pruebas de hipótesis para la diferencias de dos medias poblacionales o el análisis de varianza (Anova), que se discutirán más adelante. En esta sección, desarrollaremos una metodología, para la construcción de intervalos de confianza sobre la razón (cociente) de las varianzas de dos poblaciones con distribución normal σ_1^2/σ_2^2. Una estimación puntual de esta razón está dada por s_1^2/s_2^2 de las varianzas muéstrales, de tal forma que el estadístico S_1^2/S_2^2 se denomina estimador de σ_1^2/σ_2^2.

Como se mencionó antes, si σ_1^2 y σ_2^2 son varianzas de poblaciones normales, podemos establecer una estimación por intervalos para σ_1^2/σ_2^2 usando el estadístico

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

De acuerdo a lo que se estableció en el capítulo 3, la variable aleatoria *F* tiene una distribución *F* con $\upsilon_1 = n_1 - 1$ y $\upsilon_2 = n_2 - 1$ grados de libertad. Por lo tanto podemos afirmar que

$$P\left[f_{1-\alpha/2(\upsilon_1,\upsilon_2)} < F < f_{\alpha/2(\upsilon_1,\upsilon_2)} \right] = 1-\alpha$$

Donde $f_{1-\alpha/2(\upsilon_1,\upsilon_2)}$ y $f_{\alpha/2(\upsilon_1,\upsilon_2)}$ son valores de la distribución *F* con υ_1 y υ_2 grados de libertad, que dejan áreas de $1-\alpha/2$ y $\alpha/2$, respectivamente a la derecha (Figura 4.5). Al sustituir en *F*, tenemos que

$$P\left[f_{1-\alpha/2(\upsilon_1,\upsilon_2)} < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2(\upsilon_1,\upsilon_2)} \right] = 1-\alpha$$

Ahora, al multiplicar cada término de la desigualdad por S_2^2/S_1^2, y después invertir cada termino (nuevamente para cambiar el sentido de las desigualdades), obtenemos

$$P\left[\frac{S_1^2}{S_2^2} \frac{1}{f_{\alpha/2(\upsilon_1,\upsilon_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \frac{1}{f_{1-\alpha/2(\upsilon_1,\upsilon_2)}} \right] = 1-\alpha$$

En el capítulo 3, se estableció que $f_{1-\alpha[\upsilon_1,\upsilon_2]} = \frac{1}{f_{\alpha[\upsilon_2,\upsilon_1]}}$, por tanto,

$$P\left[\frac{S_1^2}{S_2^2} \frac{1}{f_{\alpha/2(\upsilon_1,\upsilon_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2(\upsilon_2,\upsilon_1)} \right] = 1-\alpha$$

Así, para cualquiera dos muestras aleatorias independientes de tamaño n_1 y n_2 que se selecciones de poblaciones normales, un intervalo de confianza de $(1-\alpha)100\%$, para σ_1^2/σ_2^2 esta dado por

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2(\upsilon_1,\upsilon_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2(\upsilon_2,\upsilon_1)} \tag{4.11}$$

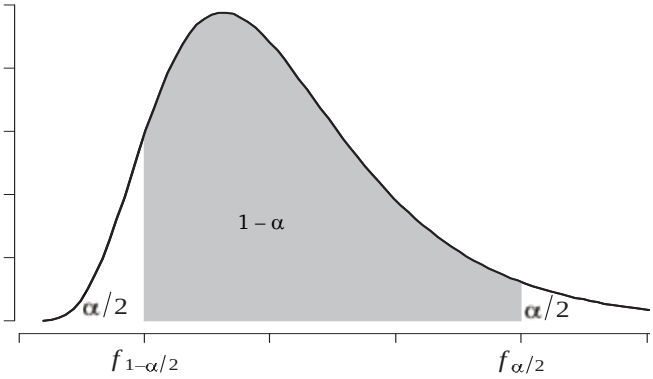


Figura 4.5. $P\left[f_{1-\alpha/2(\upsilon_1,\upsilon_2)} < F < f_{\alpha/2(\upsilon_1,\upsilon_2)} \right] = 1-\alpha$.

Ejemplo 4.10. En el ejemplo 4.4, se construyó un intervalo de confianza para la diferencia de la concentración media de SO_2 (ppm) de dos muestreadores localizados a diferentes distancias de una fuente emisora, bajo el supuesto que los datos provenían de poblaciones normales y que sus varianzas eran iguales. Justifique esta suposición mediante la construcción de un intervalo de confianza del 99% para σ_1^2/σ_2^2 y para σ_1/σ_2, donde σ_1^2 y σ_2^2 son las varianzas poblacionales de la concentración de SO_2 en los muestreadores A y B, respectivamente. Los datos de las concentraciones en cada muestreador se exponen a continuación

Concentración de SO_2 (ppm)	
Muestreador A	**Muestreador B**
680	500
630	510
620	490
600	530

Solución

De estos datos tenemos que $n_1 = n_2 = 4$, $s_1 = 34.03$ y $s_2 = 17.08$. Para un intervalo de confianza del 99%, $\alpha = 0.01$. De la Tabla A.6 del apéndice tenemos que $f_{\alpha/2(\upsilon_1,\upsilon_2)} = f_{\alpha/2(\upsilon_2,\upsilon_1)} = f_{0.05(3,3)} = 9.28$. Por lo tanto, el intervalo de confianza del 95% para σ_1^2/σ_2^2 es

$$\frac{34.03^2}{17.08^2}\frac{1}{9.28} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{34.03^2}{17.08^2}9.28$$

$$0.428 < \frac{\sigma_1^2}{\sigma_2^2} < 36.838$$

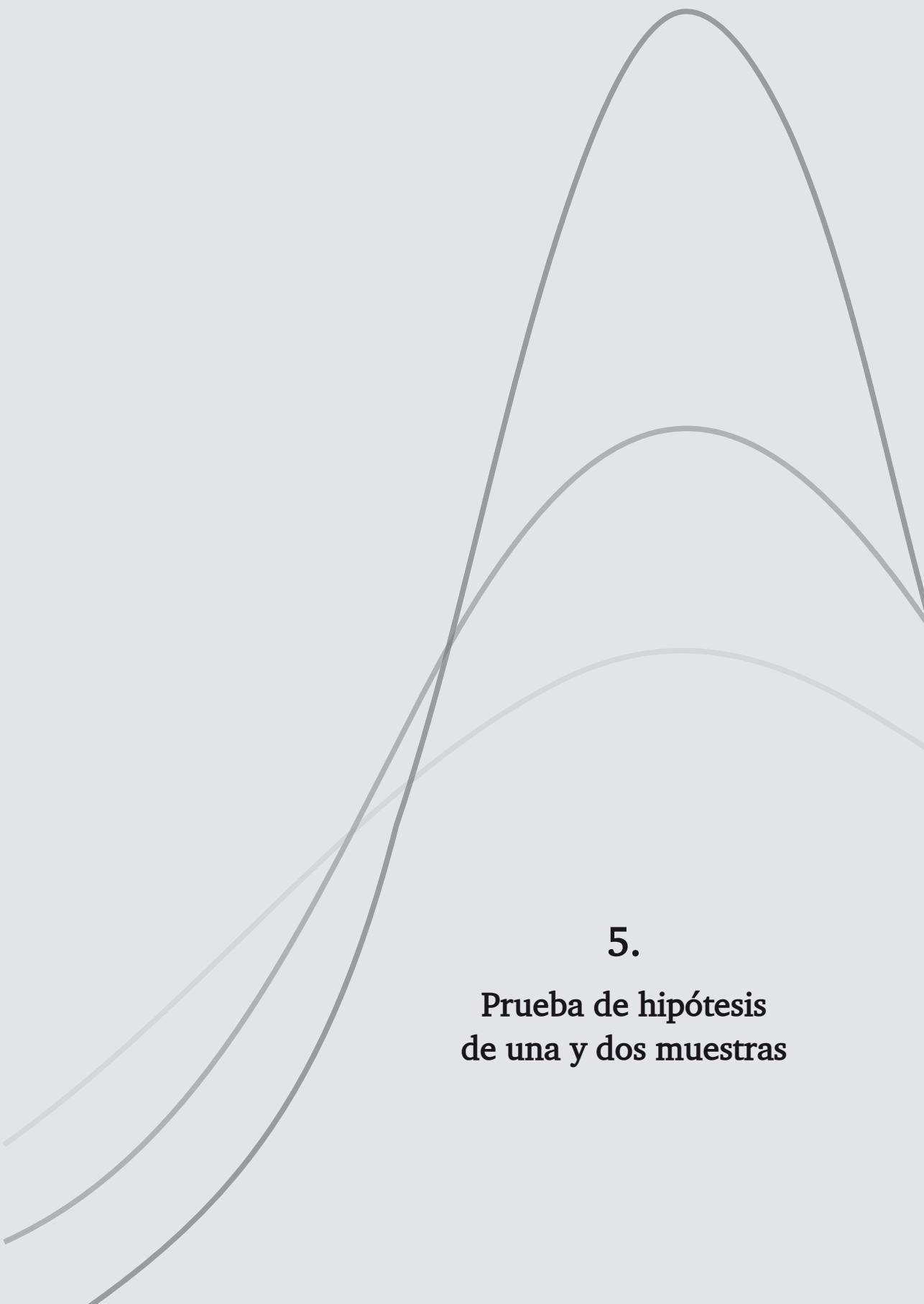
Al calcular las raíces cuadradas de los límites de confianza del intervalo, encontramos que un intervalo de confianza para σ_1/σ_2 es

$$0.654 < \frac{\sigma_1}{\sigma_2} < 6.069$$

En decir, la razón de la verdadera varianza (varianza poblacional) de las concentraciones de SO2 en los muestreadores de A y B se encuentra entre 0.428 y 36.838 ppm2, o lo que es lo mismo que la verdadera desviación estándar se encuentra entre 0.654 y 6.069 ppm. Como se puede observar en estos intervalos incluye la posibilidad de que σ_1^2/σ_2^2 o σ_1/σ_2 sea igual a 1. Por tanto, es acertado decir que $\sigma_1^2 = \sigma_2^2$ o $\sigma_1 = \sigma_2$.

La aplicación en la plataforma de R para la construcción de intervalos de confianza para la razón de la varianza de dos poblaciones normales, se realiza a través de las siguientes líneas de código

```
> A<-c(680,630,620,600)
> B<-c(500,510,490,530)
> v1<-length(A)-1
> v2<-length(B)-1
> Lim.Inf<-(var(A)/var(B))*1/qf(0.95,v1,v2)
> Lim.Sup<-(var(A)/var(B))*qf(0.95,v1,v2)
> Inter.Conf<-c(Lim.Inf,Lim.Sup)
> Inter.Conf
[1]  0.4281112 36.8414661
```



5.

Prueba de hipótesis de una y dos muestras

5.1. Generalidades

En el campo de la experimentación, es frecuente que el científico se enfrente a situaciones donde su objetivo no se limita a realizar estimaciones de los parámetros de la población que estudia a partir de las técnicas que se expusieron en el capítulo anterior, sino, por el contrario, tomar decisiones acerca de esos parámetros poblacionales basado en los datos muéstrales de los que se disponga que permitan producir una conclusión acerca de los objetivos planteados antes de la experimentación. Para tal fin, el científico debe formular *conjeturas* o *hipótesis* acerca de las variables estudiadas en su experimento, y estas deben corresponder con la formulación de **hipótesis estadísticas**, definidas como las afirmaciones que se realizan acerca de los valores de uno o más parámetros poblacionales (Devore, 2008). La veracidad o falsedad de estas hipótesis, se establecerá sobre la base de la información experimental (datos) que se tenga, a través de procedimientos de amplio uso en el campo de la inferencia estadística denominados **pruebas de hipótesis** o **contrastes de hipótesis**.

Conceptualmente la prueba de hipótesis es sencilla: se examina un conjunto de datos muéstrales y a partir de ellos se calcula un estadístico cuya distribución depende de la hipótesis planteada. Sobre la base de la distribución especificada para el estadístico y de su valor observado en la muestra, se decide el rechazo o no de la hipótesis estadística y en consecuencia de la hipótesis científica (Di Rienzo *et al.*, 2006). Un ejemplo relacionado con procedimientos de prueba de hipótesis, es el de determinar bajo la evidencia de datos muéstrales si la precipitación media mensual en el municipio de Riohacha es mayor a 150 mm, durante el mes de Octubre.

Es obvio que la hipótesis estadística debe ser equivalente a la hipótesis científica postulada, de lo contrario, aceptar o rechazar la hipótesis estadística no implicará necesariamente lo propio para la hipótesis científica.

5.2. Hipótesis nula e hipótesis alterna

Para poder construir una prueba estadística se debe especificar una hipótesis que se supone, provisoriamente como verdadera, llamada

hipótesis nula y es simbolizada con H_0. Esta hipótesis especifica los valores de uno o varios parámetros de la distribución de la variable aleatoria observada en el experimento. Cuando la hipótesis nula se somete a prueba, el resultado es su aceptación o rechazo. En este último caso se aceptará una hipótesis especificada de antemano que se llama hipótesis alterna, que se simboliza por H_1 y que propone como posibles valores del o los parámetros en cuestión al conjunto de valores complementarios al postulado bajo H_0.

Siguiendo con el ejemplo de la precipitación media mensual, se podría establecer las siguientes hipótesis:

$H_0 : \mu = 150$; "el verdadero valor medio de precipitaciones mensuales durante el mes de octubre en el municipio de Riohacha no difiere significativamente de 150 mm".

$H_1 : \mu \neq 150$; "el verdadero valor medio de precipitaciones mensuales durante el mes de octubre en el municipio de Riohacha difiere significativamente de 150 mm".

$H_1 : \mu < 150$; "el verdadero valor medio de precipitaciones mensuales durante el mes de octubre en el municipio de Riohacha es significativamente menor que 150 mm".

$H_1 : \mu > 150$; "el verdadero valor medio de precipitaciones mensuales durante el mes de octubre en el municipio de Riohacha es significativamente mayor que 150 mm".

Con base en lo anterior y dependiendo de la estructura de sus hipótesis, se distingue entre los siguientes tipos de contrastes (Arianza *et al.*, 2008):

Contrastes bilaterales. En ellos se propone un valor puntual para el parámetro bajo estudio, de forma que se rechazará bien porque la evidencia muestral lleve a decidir que el valor es mayor que el propuesto o bien que es menor. Formalmente

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

1. **Contrastes unilaterales.** En ellos se propone que el valor del parámetro se encuentre por debajo (o por encima) de un cierto valor. Las dos situaciones se plantearían de la siguiente forma:

$$H_0 : \theta = \theta_0 \qquad H_0 : \theta = \theta_0$$
$$H_1 : \theta < \theta_0 \qquad H_1 : \theta > \theta_0$$

5.3. Estadístico de prueba, región crítica y región de aceptación

Hasta el momento se ha establecido que una prueba de hipótesis estadística con respecto a alguna característica desconocida de interés es cualquier regla para decidir si se rechaza la hipótesis nula con base en una muestra aleatoria de la población. Ahora, esa decisión se basa en algún estadístico apropiado que recibe el nombre de **estadístico de prueba**. Para ciertos valores del estadístico de prueba, la decisión será de rechazar la hipótesis nula. Estos valores constituyen lo que se conoce como la **región crítica** o **región de rechazo** de la prueba (Conavos, 1988). Por ejemplo, para el problema relacionado con la precipitación media mensual en el municipio de Riohacha durante el mes de octubre, se estableció como hipótesis nula $H_0 : \mu = 150$. Para un tamaño de muestra *n*, supóngase que se decide rechazar H_0 si se observa un valor de la media muestral $\bar{X}$ que sea mayor que 160. De esa manera, $\bar{X}$ es el estadístico de prueba, el valor $\bar{X} = 160$ es el *valor crítico*, y el conjunto de valores mayores que 160 constituyen la región crítica de la prueba. En contraparte, el conjunto de valores menores que 160 constituyen la de aceptación. La representación gráfica de la región crítica y la región de aceptación se muestra en la Figura 5.1.

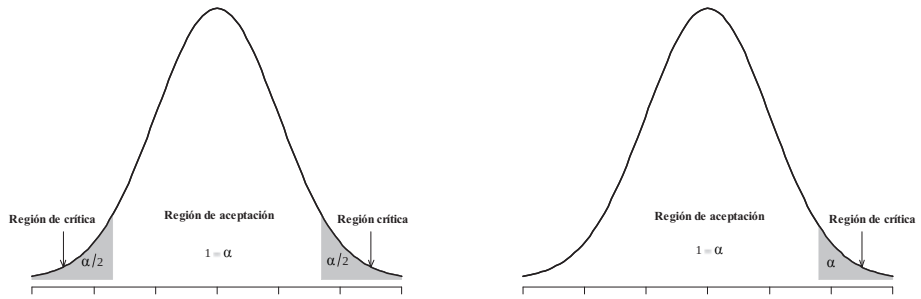


Figura 5.1. Regiones críticas y de aceptación para contrastes bilaterales y unilaterales.

5.4. Tipos de errores

Cuando realizamos contrastes de hipótesis, si no se realizan correctamente los protocolos de muestreo y establecimiento de las hipótesis a probar, somos susceptibles a incurrir en cualquiera de los dos tipos de errores que existen:

1. **Error tipo I.** Es aquel que cometemos cuando rechazamos la hipótesis nula siendo esta verdadera. Su probabilidad se representa generalmente como α y se conoce como *nivel de significación*. Usualmente a este error se le dan valores entre 0.01 y 0.05 (del 1 al 5%). Sin embargo, fijar una tasa de error tipo I demasiado pequeña aumenta el riesgo de incurrir en el siguiente tipo de error (Guisande *et al.*, 2011).

2. **Error tipo II.** Es aquel que se comete cuando se acepta la hipótesis nula cuando esta es falsa. Su probabilidad se representa generalmente por la letra griega β. Este error se tiene en cuenta determinando el tamaño de muestra necesario para garantizar el valor de β prefijado. En el supuesto de que no se tenga en cuenta el tamaño de muestra necesario y, por lo tanto, se omita el error tipo II, entonces el procedimiento se suele denominar *contraste de significación*, ya que solo tiene en cuenta el error tipo I. El error tipo II no se suele tener en cuenta porque generalmente se desconoce la información necesaria para ello (Guisande *et al.*, 2011).

De forma similar a lo visto cuando tratamos intervalos de confianza, la cantidad $1-\alpha$ se llama *nivel de confiabilidad*. Análogamente, existe una cantidad $1-\beta$ llamada *potencia de la prueba*. Ambos errores son contrapuestos entre sí, y fijado un tamaño muestral, cuando uno de los dos crece el otro decrece. En la Tabla 5.1 se muestra un resumen de las cuatro posibles situaciones que determinan si la decisión es correcta o equivocada, cuando se prueba cualquier hipótesis (Walpole *et al.*, 2007).

Tabla 5.1. Posibles situaciones al probar hipótesis estadísticas.

	H_0 es verdadera	H_0 es falsa
Se acepta H_0	Decisión correcta	Error tipo II
Se rechaza H_0	Error tipo I	Decisión correcta

5.5. Uso del p-valor como herramienta para la toma de decisiones en un procedimiento de prueba de hipótesis

Con base en lo comentado anteriormente, la preselección del nivel de significancia α, se realiza bajo el criterio de arbitrariedad, aunque por generaciones, en el análisis estadístico se ha estandarizado elegir un α igual a 0.05 o 0.01. Esta preselección del nivel de significancia α, tiene su apoyo en el intento de controlar el riesgo de cometer error tipo I. Sin embargo, no explica los valores del estadístico de prueba que están "Cercanos" a la región crítica (Walpole *et al.*, 2007), es decir, no brinda información acerca de que la afirmación que se realice sea dudosa o bastante precisa (Devore, 2008).

Como consecuencia de lo anterior, una vez se haya obtenido la muestra, se puede calcular una cantidad que permita resumir el resultado del experimento de una forma objetiva. Esta cantidad es el **p-valor**, que corresponde al nivel de significación más pequeño posible que puede escogerse con el cual H_0 sería rechazada cuando se utiliza un procedimiento de prueba especificado por el conjunto de datos de la muestra dada.

El uso del p-valor, ha sido adoptado extensamente por los usuarios de la estadística aplicada como una forma de expresar los resultados en términos de probabilidad a la simple conclusión de "rechazo" o "no rechazo". Así mismo, el cálculo del p-valor también da al usuario información importante cuando el valor del estadístico de prueba cae, o no, por completo dentro de la región critica. Siendo de esta forma, el estadístico de prueba, significativo (difiere) o no significativo (no difiere), cuando:

1. p-valor $\leq 0.05 \Rightarrow$ Rechazar H_0 al nivel preseleccionado α, es decir, difiere significativamente.
2. p-valor $>0.05 \Rightarrow$ No rechazar H_0 al nivel preseleccionado α, es decir, no difiere significativamente.

Tal vez una definición más asertiva y explicita del p-valor es la propuesta por Ferrer (2007), que establece que cuando se realiza el cálculo del p-valor, se está realmente estimando α. Así, define al p-valor más precisamente como a probabilidad de que se cometa error tipo I al rechazar erróneamente la hipótesis nula cuando esta es verdadera.

Obviamente al ser el p-valor una probabilidad, se espera que su cálculo se encuentre entre

$$0 \leq \text{p-valor} \leq 1.$$

En materia de cálculo, existen tres formas de calcular el p-valor, dependiendo del tipo de contraste que estemos realizando, a saber:

$$\text{p-valor} = P\begin{cases} 1-\Phi(x); \text{para una prueba unilateral de cola superior} \\ \Phi(x); \text{para una prueba unilateral de cola inferior} \\ 2[1-\Phi(|x|)]; \text{para una prueba bilateral} \end{cases} \tag{5.1}$$

Sin embargo, el avance en los programas de cómputo y paquetes (software) estadísticos, calculan y muestran automáticamente un p-valor cuando se realiza un procedimiento de prueba de hipótesis, por lo que se puede sacar una conclusión directamente de los datos de salida, sin referencia a una tabla de valores críticos.

A continuación se expondrán diferentes procedimientos de prueba de hipótesis para una y dos muestras para parámetros poblacionales como la media, proporción y la varianza, y su consecuente aplicación en R.

5.6. Prueba de hipótesis para la media de una población con varianza desconocida

Contrario a lo planteado en el capítulo anterior, donde se inició la discusión contemplando el escenario en donde se requería la construcción de intervalos de confianza para la media μ de una población con distribución normal, con previo conocimiento de su varianza σ^2, en esta sección iniciaremos la discusión con el desarrollo de un procedimiento de prueba sobre la media poblacional μ, pero con desconocimiento de σ^2, dado a que representa una situación más acorde con el desarrollo de experimentos reales. Nuevamente, se hace uso de la distribución *t* de s*tudent*, en remplazo de la distribución normal y la varianza poblacional se remplaza por la varianza muestral s^2. Teniendo el siguiente estadístico de prueba

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Que sigue una distribución *t* de *student* con $n-1$ grados de libertad, para una muestra aleatoria $X_1, X_2, ..., X_n$.

Para la hipótesis nula $H_0 : \mu = \mu_0$, el rechazo de esta a un nivel de significancia α, a favor de la alternativas bilateral $H_1 : \mu \neq \mu_0$, resulta cuando un valor del estadístico calculado

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Excede a $t_{\alpha/2,n-1}$ o es menor que $-t_{\alpha/2,n-1}$, debido a la naturaleza simétrica de la distribución *t*. Para $H_1 : \mu > \mu_0$ y $H_1 : \mu < \mu_0$, el rechazo de H_0 resulta cuando $t_{\alpha,n-1}$ y $-t_{\alpha,n-1}$, respectivamente. A continuación mostramos un resumen del procedimiento de prueba para la media de una población normal con varianza desconocida.

Hipótesis nula: $H_0 : \mu = \mu_0$

Valor calculado del estadístico de prueba: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

Hipótesis alternativa	Región de rechazo a un nivel de significancia α
$H_1 : \mu \neq \mu_0$	$t \geq t_{\alpha/2,n-1}$ o $t \leq -t_{\alpha/2,n-1}$
$H_1 : \mu > \mu_0$	$t \geq t_{\alpha,n-1}$
$H_1 : \mu < \mu_0$	$t \leq -t_{\alpha,n-1}$

En R, existe una función denominada ***t.test*** que permite realizar procedimientos de prueba de hipótesis para la media solo con el ingreso de los argumentos necesarios para la ejecución de esta función, como se muestra a continuación

```
t.test(x, alternative = c("two.sided", "less", "greater"),mu = 0,
conf.level = 0.95)
```

En los argumentos de esta función ***x***, representa un vector de los datos muéstrales a partir de los cuales se realizara el procedimiento de prueba, ***alternativa***, establece cual es la hipótesis alternativa, en función del tipo de contraste, pudiéndose elegir entre ***two.side*** (dos colas o bilateral), ***less*** (menor que o unilateral de cola inferior) y ***greater*** (mayor que o unilateral

de cola superior), el argumento ***mu***, establece el valor de la media poblacional que se desea probar y ***conf.level***, indica el nivel de confiabilidad de la prueba.

Ejemplo 5.1. En un estudio de contaminación atmosférica se determinó la concentración de dióxido de nitrógeno (NO_2) en el casco urbano del municipio de Guamal, Magdalena, expresada como $\mu g/m^3$ en un tiempo de muestreo de 24 horas. Los datos obtenidos en 14 estaciones de muestreo son los siguientes:

110	120	105	100	140	115	110	121	130	145
110	150	130	120	105	150	140	115	110	120

Para que el municipio cumpla con la normativa vigente de emisión de este contaminante atmosférico (Res. 610 de 2010, MAVDT), su concentración media no debe exceder los 150 $\mu g/m^3$. ¿A un nivel de significancia de 0.05, la información muestral disponible sugiere que el municipio de Guamal, Magdalena incumple la normativa de calidad de aire vigente, en cuanto a la emisión de NO_2 en la atmosfera? Suponga que los datos provienen de una población con distribución normal.

Solución

De acuerdo al planteamiento del problema, se busca establecer si $\mu_{NO_2} > 150\,\mu g/m^3$ en el municipio de Guamal, Magdalena, de manera que el sistema de hipótesis que se desea probar es

$$H_0 : \mu_{NO_2} = 150$$
$$H_1 : \mu_{NO_2} > 150$$

Así, con $\bar{x}_{NO_2} = 122.3\,\mu g/m^3$ y $s_{NO_2} = 15.59\,\mu g/m^3$ tenemos que el valor del estadístico de prueba es

$$t = \frac{122.3 - 150}{15.59/\sqrt{20}} \therefore t = -7.946$$

El valor crítico, según Tabla 4.A del apéndice es $t_{\alpha,n-1} = t_{0.05,19} = 1.729$.

De esta forma como $t = -7.946 < t_{0.05,19} = 1.729$, se tiene evidencia para no rechazar H_0 a favor de H_1 a un nivel de significancia del 0.05, es decir,

los resultados son concluyentes en que la concentración media de NO_2 en el municipio de Guamal, Magdalena no difiere significativamente de 150 $\mu g/m^3$ con una confiabilidad de 95%. Por lo tanto, cumple con la normativa ambiental vigente.

En R, la resolución del problema anterior se realiza a través de la función ***t.test***, mencionada antes, obteniéndose la siguiente salida de resultados

```
> NO2<-
c(110,120,105,100,140,115,110,121,130,145,110,150,130,120,
105,150,140,115,110,120)
> t.test(NO2,alternative="greater",mu=150,conf.level=0.95)

        One Sample t-test

data:  NO2
t = -7.944, df = 19, p-value = 1
alternative hypothesis: true mean is greater than 150
95 percent confidence interval:
 116.2707      Inf
sample estimates:
mean of x
    122.3
```

Aquí la decisión la podemos basar en términos de probabilidad, haciendo uso del p-valor. Nótese que el p-valor = 1, es mayor al nivel de significancia preseleccionado, por lo tanto, la decisión es no rechazar H_0 con una confiabilidad del 95%. Así mismo, es de notar que las estimaciones de la media y el estadístico de prueba *t* (valores sombreados), corresponden a las realizadas mecánicamente.

5.7. Prueba de hipótesis sobre la diferencia de dos medias poblacionales: comparación de dos medias

Para el caso de comparar las medias de dos poblaciones, además de comprobar las hipótesis sobre normalidad y aleatoriedad, se plantean distintas situaciones. En primer lugar habrá que determinar si se tienen muestras independientes o pareadas (relacionadas). La diferencia entre uno y otro caso es que en el segundo, se dan dos mediciones de la misma o similar característica para cada individuo o para dos individuos de idénticas, respecto de los restantes, características relevantes de la muestra (Arriaza *et al.*, 2008).

Cuando se tratan muestras que son independientes, es necesario determinar si las varianzas de las poblaciones se pueden considerar iguales o no, y según el caso, tomar cualquiera de los procedimientos que se describen a continuación.

5.7.1. Varianzas desconocidas pero iguales

Las situaciones experimentales que más prevalecen que implican pruebas sobre dos medias son aquellas con varianzas desconocidas. Si el científico está dispuesto a suponer que ambas distribuciones son normales y que $\sigma_1^2 = \sigma_2^2 = \sigma^2$, luego un análisis de homogeneidad de varianzas, que se discutirá en secciones siguientes, se puede utilizar la *prueba t combinada*, para probar la hipótesis nula $H_0 : \mu_1 - \mu_2 = d_0$, contra la alternativa bilateral $H_1 : \mu_1 - \mu_2 \neq d_0$, o las alternativas unilaterales $H_1 : \mu_1 - \mu_2 < d_0$ y $H_1 : \mu_1 - \mu_2 > d_0$ de cola inferior y superior, respectivamente.

Bajo el supuesto de que las dos muestran de tamaño n_1 y n_2, respectivamente, provienen de poblaciones normales, el estadístico

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{1/n_1 + 1/n_2}}$$

usado en remplazo de la distribución normal, sigue una distribución *t* de *student* con $\upsilon = n_1 + n_2 - 2$ grados de libertad, donde el termino s_p^2, se denomina varianza común y se calcula a través de la expresión

$$s_p^2 = \frac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{n_1 + n_2 - 2}$$

El rechazo de la hipótesis nula $H_0 : \mu_1 - \mu_2 = d_0$, a favor de la hipótesis alternativa bilateral $H_1 : \mu_1 - \mu_2 \neq d_0$ se produce cuando $t \leq -t_{\alpha/2, n_1+n_2-2}$ o $t \geq t_{\alpha/2, n_1+n_2-2}$. Para las alternativas unilaterales $H_1 : \mu_1 - \mu_2 < d_0$ y $H_1 : \mu_1 - \mu_2 > d_0$, se propone el rechazo de $H_0 : \mu_1 - \mu_2 = d_0$ cuando $t \leq -t_{\alpha, n_1+n_2-2}$ y $t \geq t_{\alpha, n_1+n_2-2}$, respectivamente. A continuación mostramos un resumen de lo anterior.

Hipótesis nula: $H_0 : \mu_1 - \mu_2 = d_0$

Valor calculado del estadístico de prueba: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{1/n_1 + 1/n_2}}$

Varianza común: $s_p^2 = \dfrac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{n_1 + n_2 - 2}$

Hipótesis alternativa	Región de rechazo a un nivel de significancia α
$H_1 : \mu_1 - \mu_2 \neq d_0$	$t \leq -t_{\alpha/2, n_1+n_2-2}$ o $t \geq t_{\alpha/2, n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 < d_0$	$t \leq -t_{\alpha, n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 > d_0$	$t \geq t_{\alpha, n_1+n_2-2}$

La aplicación en el ambiente de programación de R de pruebas de hipótesis para la diferencia de dos medias poblacionales independientes, cuando se asume que sus varianzas poblacionales son iguales, aunque se desconozcan, al igual que en la sección anterior se realiza a través de la función ***t.test*** ciñéndose a la siguiente sintaxis de programación

```
t.test(x,y, alternative = c("two.sided", "less", "greater"),mu
= 0,  var.equal = TRUE, conf.level = 0.95)
```

Aquí, los argumentos ***x*** e ***y***, corresponden a las dos muestras que se desean contrastar, ***mu*** es el valor de la diferencia entre las medias, ***var.equal***, se fija como ***TRUE*** (verdadero), para ordenar al software que se trata de dos poblaciones que tienen igual varianza. Los demás argumentos, realizan las mismas funciones que se describieron en la sección anterior.

Ejemplo 5.2. En un estudio de la aplicación del pH (potencial hidrógeno que tiene una escala de 0 a 14, donde 7 es neutral y abajo de 7 es ácido y arriba de 7 es alcalino) para medir la alcalinidad y la acidez de soluciones, un científico, dedicado al estudio de la contaminación ambiental, asegura que dos muestras de agua (A y B) provienen del mismo lugar de un río, donde supuestamente hubo un descarga industrial de ácido clorhídrico (HCl). Si esto fuera cierto, entonces el pH de las dos muestras de agua serían iguales. Asumiendo que las observaciones provienen de poblacionales normales con varianzas iguales, probar la hipótesis nula de igualdad de los promedios de pH. Asumir $\alpha = 0.05$.

Mediciones de pH

Muestra A					Muestra B				
6.24	6.31	6.28	6.30	6.22	6.27	6.25	6.33	6.27	6.34
6.25	6.26	6.24	6.29	6.28	6.24	6.31	6.28	6.29	6.27

Solución

De acuerdo a lo planteado en el enunciando, el sistema de hipótesis que se desea probar es el siguiente

$$H_0 : \mu_A - \mu_B = 0$$
$$H_1 : \mu_A - \mu_B \neq 0$$

De esta forma, con $\bar{x}_A = 6.267$, $\bar{x}_B = 6.285$, $s_A^2 = 0.00087$ y $s_B^2 = 0.00107$, tenemos que la varianza común queda determinada por

$$s_p^2 = \frac{0.00087(10-1) + 0.00107(10-1)}{10+10-2} \therefore s_p^2 = 0.0097$$

$$\Rightarrow s_p = 0.0311$$

Así, el valor del estadístico de prueba es

$$t = \frac{6.267 - 6.285}{0.0311\sqrt{1/10 + 1/10}} = -1.294$$

El valor crítico, según la Tabla 4.A del apéndice es

$$t_{\alpha/2, n_1+n_2-2} = t_{0.025,18} = 2.101$$

De esta forma como $t = -1.294 < t_{0.025,18} = 2.101$, no se rechaza H_0 a un nivel de significancia de 0.05, es decir, existe evidencia suficiente para concluir que no existen diferencias significativas entre los valores medios de pH de las muestras de agua A y B del rio en cuestión con una confiabilidad del 95%, por lo que es posible afirmar con una alta confiabilidad que las muestras de agua A y B, proviene del mismo lugar del rio estudiado.

En el ambiente de programación de R, la salida de resultados a la solución de este ejercicio sería

```
> Muestra.A<-
c(6.24,6.31,6.28,6.30,6.22,6.25,6.26,6.24,6.29,6.28)
> Muestra.B<-
c(6.27,6.25,6.33,6.27,6.34,6.24,6.31,6.28,6.29,6.27)
> t.test(Muestra.A,Muestra.B,alternative="two.sided",mu=0,
var.equal=TRUE,conf.level=0.95)
```

```
        Two Sample t-test

data:  Muestra.A and Muestra.B
t = -1.2923, df = 18, p-value = 0.2126
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -0.04726246  0.01126246
sample estimates:
mean of x mean of y
    6.267     6.285
```

La decisión basada en términos de probabilidad, haciendo uso de p-valor, muestra que el p-valor = 0.2126 es mayor al nivel de significancia preseleccionado y, por lo tanto, se decide no rechazar H_0 con una confiabilidad del 95%.

En la sección 2.4.4, se discutió la utilidad de los gráficos de cajas y bigotes para la evaluar, de una forma descriptiva, las similitudes o diferencias entre grupos sobre la medición de una misma variable. En los procedimientos de comparación de dos medias poblacionales, estos gráficos ofrecen una representación reveladora al analista de datos, pues a través de estos, puede sacar inferencias acerca de la existencia (o no) de diferencias significativas entre las medias de los grupos que se estudien, solo con observar el grado de solapamiento de las cajas.

En general no existe reglas definitivas respecto a cuándo las gráficas de cajas brindan evidencia de diferencias significativas entre las medias. Sin embargo, una pauta aproximada es que si la línea del percentil 25 (primer cuartil) para una muestra excede la línea de la mediada para la otra muestra, hay evidencia sólida de una diferencia entre las medias (Walpole *et al.*, 2007).

Para el ejemplo anterior los gráficos de cajas (con y sin muescas) para las medidas de pH en las muestras de agua A y B, respectivamente, se muestran en la Figura 5.2. Los comandos en R para la construcción de estos gráficos, se proporcionan en el siguiente recuadro

```
> Muestra.A<-
c(6.24,6.31,6.28,6.30,6.22,6.25,6.26,6.24,6.29,6.28)
```

```
> Muestra.B<-
c(6.27,6.25,6.33,6.27,6.34,6.24,6.31,6.28,6.29,6.27)
> par(mfrow=c(1,2))
> boxplot(Muestra.A,Muestra.B,names=c("Muestra A", "Muestra B"))
> boxplot(Muestra.A,Muestra.B,notch=TRUE,names=c("Muestra A",
"Muestra B"))
```

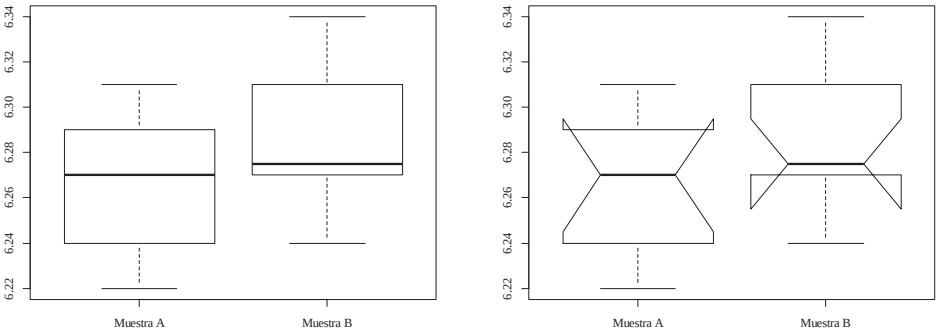


Figura 5.2. Gráficos de cajas para as medidas de pH en las muestras de agua A y B.

5.7.2. Varianzas desconocidas pero diferentes

Contrario al procedimiento descrito anteriormente, hay situaciones en que el analista o científico no es capaz de suponer que $\sigma_1^2 = \sigma_2^2 = \sigma^2$; bajo este escenario, si se puede suponer que las poblaciones se distribuyen normalmente, el estadístico

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Sigue una distribución *t* de *student*, en remplazo de la distribución normal con

$$\upsilon = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

grados de libertad.

Así, el procedimiento de prueba y las regiones críticas o de rechazo se resume de la siguiente forma

Hipótesis nula: $H_0 : \mu_1 - \mu_2 = d_0$

Valor calculado del estadístico de prueba: $t = \dfrac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$

Grados de libertad: $\upsilon = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$

Hipótesis alternativa	Región de rechazo a un nivel de significancia α
$H_1 : \mu_1 - \mu_2 \neq d_0$	$t \leq -t_{\alpha/2,\upsilon}$ o $t \geq t_{\alpha/2,\upsilon}$
$H_1 : \mu_1 - \mu_2 < d_0$	$t \leq -t_{\alpha,\upsilon}$
$H_1 : \mu_1 - \mu_2 > d_0$	$t \geq t_{\alpha,\upsilon}$

Como rara vez el cálculo de υ, denominados grados de libertad de Welch-Satterthwaite (Guisande *et al.*, 2011), dará como resultado un entero, es preciso hacer un redondeo de las cifras decimales al menor entero más cercano.

Los procedimientos de pruebas de hipótesis para la diferencia de dos medias poblacionales independientes, cuando se asume que sus varianzas poblacionales son diferentes, aunque se desconozcan, al igual que en las secciones anteriores se realiza a través de la función ***t.test*** ciñéndose a la siguiente sintaxis de programación

```
t.test(x,y, alternative = c("two.sided", "less", "greater"),mu
= 0,  var.equal = TRUE, conf.level = 0.95)
```

Aquí, solo se requiere fijar el valor de ***var.equal*** como ***FALSE*** (falso), para ordenar al software que se trata de dos poblaciones que no tienen igual varianza. Los demás argumentos, realizan las mismas funciones antes descritas.

Ejemplo 5.3. Dentro del Programa de Calidad Ambiental de Playas Turísticas- CAPT, del cual la universidad de La Guajira, a través del grupo de investigación Pichihuel, tiene participación, se han desarrollado monitoreos con el objeto de evaluar la calidad de las playas del municipio de Riohacha, donde la humedad es una de las variables tenidas en cuenta como factor que puede condicionar la presencia de ciertos grupos bacterianos y fúngicos en la arena de playa. Para el estudio de este parámetro de realizaron mediciones *in situ* a través de un sensor en

muestras de arena húmeda y arena seca. Una porción de los resultados de estas mediciones se consignan en la siguiente tabla.

% de humedad	
Arena húmeda	Arena seca
84	83
75	69
82	82
83	80
87	84
78	53
85	65
86	69
77	67
77	73

A partir de estas observaciones se desea saber, con nivel de significancia de 0.05, si los porcentajes medios de humedad en arena húmeda son significativamente mayores a la humedad media obtenida en muestras de arena seca. Se supone que los datos provienen de poblaciones normales con varianzas diferentes.

Solución

De acuerdo a lo planteado, el sistema de hipótesis que se desea probar es el siguiente

$$H_0 : \mu_{AH} - \mu_{AS} = 0$$

$$H_1 : \mu_{AH} - \mu_{AS} > 0$$

Un análisis gráfico de los datos para la verificación de estas hipótesis, a través de un gráfico de cajas, se muestra en la Figura 5.3. Allí, se puede ver que el percentil 25, del porcentaje de humedad en las muestras de arena húmeda, excede a la mediada (percentil 50) del porcentaje de humedad en las muestras de arena seca, por lo que se podría concluir que el porcentaje de humedad es significativamente mayor en las muestras de arena húmeda. Sin embargo, veamos qué resultados nos arroja el procedimiento de prueba de nuestro sistema de hipótesis

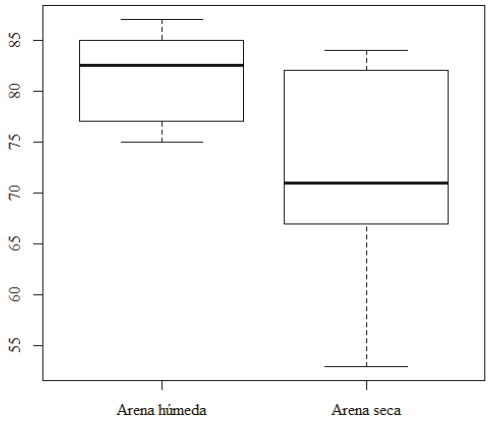


Figura 5.3. Porcentaje de humedad en muestras de arena húmeda y arena seca (Ejemplo 5.3)

Con $\overline{x}_{AH} = 81.4$, $\overline{x}_{AS} = 72.5$, $s^2_{AH} = 18.49$ y $s^2_{AS} = 97.83$, el valor del estadístico de prueba es

$$t = \frac{81.4 - 72.5}{\sqrt{18.49/10 + 97.83/10}} \therefore t = 2.6095$$

Así mismo, los grados de libertad de Welch-Satttherthwaite, son iguales a

$$\upsilon = \frac{(18.49/10 + 97.83/10)^2}{(18.49/10)^2/(10-1) + (97.83/10)^2/(10-1)}$$

$$\upsilon = 12.284 \approx 12$$

El valor crítico, según la Tabla 4.A del apéndice es $t_{\alpha,\upsilon} = t_{0.05,12} = 1.782$.

De esta forma como $t = 2.6095 > t_{0.05,12} = 1.782$, se rechaza H_0 a favor de H_1 con un nivel de significancia de 0.05, es decir, existe evidencia suficiente para concluir que los valores medios del porcentaje de humedad en muestras de arena húmeda son significativamente mayor a los valores medios del de este parámetro en muestras de arena seca con una confiabilidad del 95%.

La salida de resultados y las órdenes utilizadas en la plataforma de R, para la resolución de este tipo de problemas se muestra a continuación

```
> AH<-c(84,75,82,83,87,78,85,86,77,77)
> AS<-c(83,69,82,80,84,53,65,69,67,73)
> t.test(AH,AS,alternative="greater",mu=0,var.equal=FALSE,
conf.level=0.95)

        Welch Two Sample t-test

data:  AH and AS
t = 2.6095, df = 12.284, p-value = 0.01123
alternative hypothesis: true difference in means is greater
than 0
95 percent confidence interval:
 2.833027      Inf
sample estimates:
mean of x mean of y
     81.4      72.5
```

Observe que en estos resultados, el p-valor = 0.011 es menor que el nivel de significancia fijado de 0.05, evidencia suficiente para rechazar H_0 a favor de H_1 con una confiabilidad del 95%.

5.7.3. Observaciones pareadas (emparejadas)

Habitualmente las pruebas estadísticas de comparación de medias poblacionales se fundamentan en que las observaciones pertenecientes a cada una de las muestras son independientes entre sí o no guardan relación; siendo precisamente ese uno de los objetivos de la aleatorización (elección aleatoria de los sujetos o unidades de observación). Sin embargo, existen situaciones en las que las condiciones de las dos poblaciones no se asignan de manera aleatoria a las unidades experimentales, más bien, cada unidad experimental reciba ambas condiciones experimentales (Walpole *et al.*, 2007), es decir, se tiene un conjunto de *n* individuos o unidades experimentales y se realizan dos observaciones de cada una (Devore, 2008). Este tipo especial de situaciones, recibe el nombre de *pruebas con observaciones pareadas*.

Aquí, el estadístico de prueba queda definido por

$$T = \frac{\bar{D} - \mu_D}{S_d / \sqrt{n}}$$

Donde $\bar{D}$ y S_d son las media y la varianza de los valores de una muestra aleatoria $D_1, D_2, ..., D_n$ de una población de las diferencias de las dos observaciones en cada una de las unidades experimentales, que supondremos se encuentra distribuida normalmente. De esta forma el

problema de dos muestras se reduce en esencia a un problema de una muestra utilizando las diferencias calculadas $d_1, d_2, ..., d_n$. Donde el sistema de hipótesis que se desea probar puede ser cualquiera de los siguientes

$$H_0 : \mu_D = d_0 \qquad H_0 : \mu_D = d_0 \qquad H_0 : \mu_D = d_0$$
$$H_1 : \mu_D \neq d_0 \qquad H_1 : \mu_D < d_0 \qquad H_1 : \mu_D > d_0$$

El estadístico de prueba calculado, queda determinado por

$$t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$$

Que sigue una distribución *t* de *student* con $n-1$ grados de libertad.

El resumen del procedimiento de prueba para observaciones pareadas con las diferentes regiones críticas se muestra a continuación

Hipótesis nula: $H_0 : \mu_D = d_0$	
Valor calculado del estadístico de prueba: $t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$	
Hipótesis alternativa	Región de rechazo a un nivel de significancia α
$H_1 : \mu_D \neq d_0$	$t \leq -t_{\alpha/2,n-1}$ o $t \geq t_{\alpha/2,n-1}$
$H_1 : \mu_D < d_0$	$t \leq -t_{\alpha,n-1}$
$H_1 : \mu_D > d_0$	$t \geq t_{\alpha,n-1}$

Nuevamente, la función ***t.test*** se utiliza para realizar el procedimiento de prueba de hipótesis para observaciones pareadas en el ambiente de programación de R, con los mismos argumentos relacionados. Solo se destaca que se debe prescindir del argumento ***var.equal*** dado que en este tipo de contrastes el que las poblaciones tengan o no varianzas iguales no representa ninguna importancia. Sin embargo, hay que incluir un nuevo argumento, denominado ***paired***, al cual se le asigna el valor lógico ***TRUE***, para indicar que se están tratando observaciones pareadas, quedando la estructura de la función como sigue

```
t.test(x,y, alternative = c("two.sided", "less", "greater"),mu =
0,  paired = TRUE, conf.level = 0.95)
```

Ejemplo 5.4. Para probar la eficiencia de una planta de tratamiento de lodos activados se midió la concentración del DBO_5 en el afluente (entrada) y efluente (salida) de la misma. Se requiere saber, a un nivel de significancia de 0.05, si el sistema de tratamiento de aguas residuales es eficiente en la remoción de este parámetro. Las concentraciones de DBO5 se muestran en la tabla siguiente:

Concentraciones DBO en el afluente (mg/L)	Concentraciones DBO en el efluente (mg/L)	Diferencia en las concentraciones (mg/L)
170.5	140.4	30.1
207.4	174.7	32.7
215.9	170.2	45.7
209.0	174.6	34.4
171.6	154.6	17.0
201.2	185.0	16.2
209.9	118.9	91.0
213.3	169.8	43.5
184.1	174.7	9.4
220.4	176.7	43.7

Solución

Se incluyeron las diferencias de las observaciones en la tabla de datos con el objeto de facilitar los cálculos. De esta forma $\bar{d} = 36.37$, $s_p = 22.954$ y el valor del estadístico prueba sería

$$t = \frac{36.37}{22.954/\sqrt{10}} \therefore t = 5.0105$$

El valor crítico para esta ocasión es $t_{\alpha,n-1} = t_{0.05,9} = 9$.

Así, como $t = 5.0105 > t_{0.05,9} = 1.383$, se rechaza H_0 a favor de H_1 con un nivel de significancia de 0.05, es decir, existe evidencia suficiente para concluir que las concentraciones medias de DBO_5 en el afluente son significativamente mayor a las concentraciones medias de este parámetro en el efluente con una confiabilidad del 95%.
La salida de resultados y las órdenes utilizadas en la plataforma de R, para la resolución de este tipo de problemas se muestra a continuación

```
> DBO.Afluente<-
c(170.5,207.4,215.9,209.0,171.6,201.2,209.9,213.3, 184.1,220.4)
> DBO.Efluente<-
c(140.4,174.7,170.2,174.6,154.6,185.0,118.9,169.8, 174.7,176.7)
> t.test(DBO.Afluente,DBO.Efluente,alternative="greater",
paired=TRUE,con.level=0.95)

        Paired t-test

data:  DBO.Afluente and DBO.Efluente
t = 5.0105, df = 9, p-value = 0.0003642
alternative hypothesis: true difference in means is greater than
0
95 percent confidence interval:
 23.06379      Inf
sample estimates:
mean of the differences
                 36.37
```

Observe que en estos resultados, el $\text{p-valor} = 0.0003642 < 0.05$, evidencia suficiente para rechazar H_0 a favor de H_1 con una confiabilidad del 95%.

5.8. Prueba de hipótesis para una proporción

Las pruebas de hipótesis relacionadas con proporciones (porcentajes) se requieren en muchas áreas de la ingeniería. En el campo de la ingeniería ambiental estamos interesados en saber qué fracción de las industrias están cumpliendo con las legislaciones ambientales. Igualmente, es de interés saber qué fracción o proporción de personas que puedan estar conscientes de la magnitud del problema de la contaminación ambiental, etc. Las pruebas de hipótesis con la estadística $\hat{p} = x/n$ (que estima a *p*) están basadas en una muestra aleatoria de tamaño *n* de la población muestreada (Quevedo, 2006). Igual a como se discutió en la sección 4.8, los procedimientos de prueba de hipótesis para una proporción, se basan en el uso de la distribución normal como una aproximación a la binomial cuando el tamaño de la muestra *n* es grande.

El sistema de hipótesis que se desea probar puede tomar cualquiera de las siguientes formas

$$H_0 : p = p_0 \qquad H_0 : p = p_0 \qquad H_0 : p = p_0$$

$$H_1 : p \neq p_0 \qquad H_1 : p < p_0 \qquad H_1 : p > p_0$$

Cuyo valor del estadístico de prueba está dado por

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \quad ,$$

El cual es un valor de la variable normal estándar Z. De aquí que para una prueba bilateral a un nivel de significancia α, la región critica es $z < -z_{\alpha/2}$ y $z > z_{\alpha/2}$. Para la alternativa unilateral $p < p_0$, la región de crítica es $z < -z_{\alpha}$ y para la alternativa $p > p_0$, la región de rechazo es $z > z_{\alpha}$ (Walpole *et al.*, 2007).

De forma resumida el procedimiento de prueba de hipótesis para una proporción es el siguiente

Hipótesis nula: $H_0 : p = p_0$

Valor calculado del estadístico de prueba: $z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$

Hipótesis alternativa	Región de rechazo a un nivel de significancia α
$H_1 : p \neq p_0$	$z \leq -z_{\alpha/2}$ o $z \geq z_{\alpha/2}$
$H_1 : p < p_0$	$z \leq -z_{\alpha}$
$H_1 : p > p_0$	$z \geq z_{\alpha}$

Sin embrago, como uno de los objetivos de este texto es hacer corresponder los resultados de los cálculos mecánicos con los resultados de la salida de la consola de R, es preciso comentar otra metodología utilizada para probar hipótesis sobre proporciones, basado en la distribución de probabilidad chi-cuadrado (χ^2), en la cual el software R basa una de sus funciones para la ejecución de esta metodología.

En este procedimiento lo que se busca es comparar las frecuencias de casos favorables (éxitos) en la muestra estudiada (frecuencias observadas, o_i) con la frecuencia de casos favorables que habría en una muestra con el mismo número de datos si la hipótesis nula fuera verdadera (frecuencias esperadas, e_i) (Sáenz, 2010). El estadístico de prueba utilizado para este procedimiento sería

$$\chi^2 = \sum_{i=1}^{c} \frac{(o_i - e_i)^2}{e_i}$$

que, bajo el supuesto de que ninguna frecuencia esperada e_i es inferior a 5, se distribuye según una distribución chi-cuadrado con $c-1$ grados de

libertad (número de categorías menos uno). En el caso de que alguna frecuencia esperada sea inferior a 5 se suele utilizar la corrección por continuidad de Yates, en la que el estadístico es

$$\chi^2 = \sum_{i=1}^{c} \frac{\left(|o_i - e_i| - 0.5\right)^2}{e_i}$$

La región crítica para una prueba bilateral a un nivel de significancia α está dada por $\chi^2 \leq \chi^2_{1-\alpha/2,c-1}$ o $\chi^2 \geq \chi^2_{\alpha/2,c-1}$, para el caso de alternativas unilaterales de cola inferior y superior, las regiones críticas serían $\chi^2 \leq \chi^2_{1-\alpha,c-1}$ y $\chi^2 \geq \chi^2_{\alpha,c-1}$, respectivamente. Note que al existir solo dos categorías para el criterio de clasificación (favorable y no favorable), la relación $c-1$ siempre será igual a 1. A continuación se presenta un resumen para este procedimiento

Hipótesis nula: $H_0 : p = p_0$

Valor calculado del estadístico de prueba: $\chi^2 = \sum_{i=1}^{c} \frac{(o_i - e_i)^2}{e_i}$

Valor calculado del estadístico de prueba con corrección: $\chi^2 = \sum_{i=1}^{c} \frac{\left(|o_i - e_i| - 0.5\right)^2}{e_i}$

Hipótesis alternativa	**Región de rechazo a un nivel de significancia α**
$H_1 : p \neq p_0$	$\chi^2 \leq \chi^2_{1-\alpha/2,[1]}$ o $\chi^2 \geq \chi^2_{\alpha/2,[1]}$
$H_1 : p < p_0$	$\chi^2 \leq \chi^2_{1-\alpha,[1]}$
$H_1 : p > p_0$	$\chi^2 \geq \chi^2_{\alpha,[1]}$

En R, la aplicación de este procedimiento de prueba se realiza a través de la función ***prop.test***, en cuyos argumentos ***x*** representa el número de aciertos ("éxitos"), ***n*** el número de observaciones (tamaño de la muestra), ***p*** el valor de la proporción a probar, ***alternative*** fija la hipótesis alternativa, bilateral o unilateral, que se escoja, ***conf.level***, establece el nivel de confiabilidad de la prueba y ***correct***, indica la aplicación de la corrección de Yates, tomando los valores lógicos ***TRUE*** o ***FALSE***, según sea el caso.

```
prop.test(x, n, p = NULL ,alternative = c("two.sided", "less",
"greater"), conf.level = 0.95, correct = TRUE)
```

Ejemplo 5.4. Supóngase que los criterios de control de calidad del proceso de purificación de agua de una planta embotelladora de este líquido,

establecen que para que el producto sea sanitariamente seguro, menos del 95% de las unidades de cada lote no deben registrar presencia de indicadores microbianos de patógenos. De un total de 250 muestras de agua analizadas, 17 resultaron ser positivos al análisis de estos indicadores. A un nivel de significancia de 0.05, ¿se puede afirmar que la planta embotelladora cumple con este criterio de control sanitario de calidad?

Solución

La resolución de este problema la realizaremos a través de los dos métodos comentados anteriormente para luego realizar la aplicación en la interfaz de R. Tomaremos como éxito al evento en que el examen de la presencia de indicadores microbianos patógenos resulte negativo. De esta forma, la hipótesis que deseamos probar sería

$$H_0 : p = 0.95$$

$$H_1 : p < 0.95$$

Con $x = 233$, $\hat{p} = 233/250 = 0.932$ y $q_0 = 1 - p_0 = 1 - 0.95 = 0.05$, el valor del estadístico de prueba para el primer método es

$$z = \frac{0.932 - 0.95}{\sqrt{(0.95)(0.05)/250}} \therefore z = -1.306$$

El valor critico según la Tabla 3.A del apéndice es $z_\alpha = z_{0.05} = 1.645$

De esta forma como $z = -1.306 > -z_{0.05} = -1.645$, no se rechaza H_0 a un nivel de significancia de 0.05, siendo los resultados concluyentes en que la verdadera proporción de muestras de agua en la planta embotelladora en el que no se registra presencia de indicadores microbianos de patógenos no difiere significativamente de 95%, con una confiabilidad del 95%, es decir, la planta si cumple con sus criterios de control sanitario de calidad.

Para la realización de este ejercicio a través del otro procedimiento de prueba, es de notar que tenemos un solo criterio de clasificación (Ausencia o no de indicadores microbianos de patógenos) que se puede resumir mediante el siguiente formato tabular, donde se muestran las frecuencias observadas para cada categoría.

Ausencia de indicadores microbianos de patógenos		
Si	No	Total
233	17	250

Como en nuestro sistema de hipótesis se estableció a priori que $p_0 = 0.95$ y $q_0 = 0.05$, de esta manera, las frecuencias esperadas serían

$$e_1 = np_0 = (250)(0.95) \therefore e_1 = 237.5 \text{ y}$$

$$e_2 = nq_0 = (250)(0.05) \therefore e_2 = 12.5$$

Así, el valor del estadístico de prueba es

$$\chi^2 = \frac{(233-237.5)^2}{237.5} + \frac{(17-12.5)^2}{12.5} \therefore \chi^2 = 1.705$$

El valor crítico según la Tabla 5.A del apéndice es $\chi^2_{1-\alpha,c-1} = \chi^2_{0.95,1} = 0.00393$.

A partir de lo anterior, como $\chi^2 = 1.705 > \chi^2_{0.95,1} = 0.00393$, no se rechaza H_0 a un nivel de significancia de 0.05, llegando a las mimas conclusiones mencionadas anteriormente.

El modelado de este problema con la respectiva salida de resultados en la consola de R, se muestra a continuación

```
> prop.test(x=233,n=250,p=0.95,alternative="less",conf.level=0.95,
correct=FALSE)

        1-sample proportions test without continuity correction

data:  233 out of 250, null probability 0.95
X-squared = 1.7053, df = 1, p-value = 0.0958
alternative hypothesis: true p is less than 0.95
95 percent confidence interval:
 0.0000000 0.9538308
sample estimates:
    p
0.932
```

De esta salida de resultados se observa que el p-valor = 0.0958 es mucho mayor que el nivel de significancia seleccionado, de allí que la conclusión es no rechazar la hipótesis nula.

5.9. Prueba de hipótesis sobre la diferencia entre dos proporciones

Es frecuente que en diferentes experimentos surjan situaciones en donde se desea realizar inferencias acerca de la diferencia entre dos proporciones poblacionales a través de muestras de tamaño n_1 y n_2, respectivamente, deseándose probar cualquiera de los siguientes sistemas de hipótesis que implican la igualdad de las proporciones estudiadas.

$$H_0 : p_1 - p_2 = 0 \qquad H_0 : p_1 - p_2 = 0 \qquad H_0 : p_1 - p_2 = 0$$
$$H_1 : p_1 - p_2 \neq 0 \qquad H_1 : p_1 - p_2 < 0 \qquad H_1 : p_1 - p_2 > 0$$

Cuando H_0 es verdadera, puede sustituirse $p_1 = p_2 = p$ y $q_1 = q_2 = q$, donde p y q, se denominan valores comunes, de tal forma que las regiones de aceptación y rechazo se establecen con la variable de distribución normal

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{pq\left[\left(1/n_1\right) + \left(1/n_2\right)\right]}}$$

La estimación de los parámetros p y q que aparecen en el radical para el cálculo del valor de Z se realiza al combinar los datos de ambas muestras, de tal forma que la **estimación combinada de la proporción p** es:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Donde x_1 y x_2 son el número de éxitos en cada una de los dos muestras. Al sustituir $\hat{p}$ para p y $\hat{q} = 1 - \hat{p}$ para q, el valor de z para probar $H_0 : p_1 - p_2 = 0$ queda determinado por la expresión:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left[\left(1/n_1\right) + \left(1/n_2\right)\right]}}$$

Las regiones críticas para las hipótesis alternativas que se establecen como antes utilizando los puntos críticos de la curva normal estándar. Así, a un nivel de significancia α, las regiones críticas para la alternativa bilateral son $z \leq -z_{\alpha/2}$ y $z \geq z_{\alpha/2}$. Para las alternativas unilaterales de cola inferior y

superior, las regiones críticas son $z \leq -z_{\alpha}$ y $z \geq z_{\alpha}$, respectivamente. A continuación mostramos un resumen de este procedimiento de prueba

Hipótesis nula: $H_0 : p_1 - p_2 = 0$

Valor calculado del estadístico de prueba: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left[\left(1/n_1\right)+\left(1/n_2\right)\right]}}$

Hipótesis alternativa	Región de rechazo a un nivel de significancia α
$H_1 : p_1 - p_2 \neq 0$	$z \leq -z_{\alpha/2}$ o $z \geq z_{\alpha/2}$
$H_1 : p_1 - p_2 < 0$	$z \leq -z_{\alpha}$
$H_1 : p_1 - p_2 > 0$	$z \geq z_{\alpha}$

Como se comentó en la sección anterior, R utiliza una función que basa sus cálculos en la distribución chi-cuadrado, debido a que las proporciones son en esencia la frecuencia en que se presenta cierta característica (casos favorables) de la población estudiada. Para el caso de dos proporciones, se tendrían dos criterios de clasificación, de esta forma el valor del estadístico de prueba se calcula por la expresión

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$$

Que tiene una distribución chi-cuadrado con $(r-1)(c-1)$ grados de libertad. Los casos favorables y no favorables para cada una de las muestras estudiadas se pueden disponer en un arreglo tabular como el que se muestra a continuación denominado tabla de contingencia 2x2.

	Casos favorables	Casos no favorables	Total
Muestra 1	$o_{11} = x_1$	$o_{12} = n_1 - x_1$	n_1
Muestra 2	$o_{21} = x_2$	$o_{22} = n_2 - x_2$	n_2
Total	$x_1 + x_2$	$\left(n_1 + n_2\right) - \left(x_1 + x_2\right)$	$n_1 + n_2$

Para este tipo de situaciones las frecuencias esperadas, igual que en el apartado anterior se hace bajo el supuesto de que la hipótesis nula es verdadera y por lo tanto $p_1 = p_2 = p$. Por ello una expresión intuitiva para el cálculo de las frecuencias esperadas para los casos favorables y no favorables sería

Frecuencias esperadas casos favorables $= e_i = n_i p$

Frecuencias esperadas casos no favorables $= e_i = n_i q$

Sabiendo que si la hipótesis nula es cierta, la estimación de p a través de la evidencia muestral es $\hat{p} = \frac{x_1 + x_2}{n_1 + n_1}$ y la estimación de q sería $\hat{q} = 1 - \hat{p}$. Si tabuláramos las frecuencias esperadas en una tabla de doble entrada como la mostrada antes, luciría como la que se muestra a continuación

	Frecuencias esperadas casos favorables	Frecuencias esperadas casos no favorables	Total
Muestra 1	$e_{11} = n_1 p$	$e_{12} = n_1 q$	n_1
Muestra 2	$e_{21} = n_2 p$	$e_{22} = n_2 p$	n_2
Total	$p(n_1 + n_2)$	$q(n_1 + n_2)$	$n_1 + n_2$

Por último, la región crítica para una prueba bilateral a un nivel de significancia α está dada por $\chi^2 \leq \chi^2_{1-\alpha/2,[(r-1)(c-1)]}$ o $\chi^2 \geq \chi^2_{\alpha/2,[(r-1)(c-1)]}$, para el caso de alternativas unilaterales de cola inferior y superior, las regiones críticas serían $\chi^2 \leq \chi^2_{1-\alpha,[(r-1)(c-1)]}$ y $\chi^2 \geq \chi^2_{\alpha,[(r-1)(c-1)]}$, respectivamente. Como siempre se trata de dos criterios de clasificación (dos renglones y dos columnas en la tabla), el producto $(r-1)(c-1)$ siempre dará 1 grados de libertad. A continuación se presenta un resumen para este procedimiento de prueba.

Hipótesis nula: $H_0 : p_1 - p_2 = 0$

Valor calculado del estadístico de prueba: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

Hipótesis alternativa	**Región de rechazo a un nivel de significancia α**
$H_1 : p_1 - p_2 \neq 0$	$\chi^2 \leq \chi^2_{1-\alpha/2,[1]}$ o $\chi^2 \geq \chi^2_{\alpha/2,[1]}$
$H_1 : p_1 - p_2 < 0$	$\chi^2 \leq \chi^2_{1-\alpha,[1]}$
$H_1 : p_1 - p_2 > 0$	$\chi^2 \geq \chi^2_{\alpha,[1]}$

La aplicación en R del procedimiento de prueba para la diferencia de dos proporciones poblacionales, al igual que en la sección anterior, se realiza a través de la función ***prop.test***, sin embrago, antes de la aplicación de esta

función es necesario crear una matriz de datos donde se especifique el número de casos favorables y no favorables en cada muestra con la siguiente configuración

Nº casos favorables en la muestra 1 (O_{11})	Nº casos no favorables en la muestra 1 (O_{12})
Nº casos favorables en la muestra 2 (O_{21})	Nº casos no favorables en la muestra 2 (O_{22})

En consecuencia, el procedimiento en R para probar las hipótesis especificadas sigue la siguiente estructura de programación

```
Tabla<-matrix(c( O11, O12, O21, O22 ),nrow=2,ncol=2,byrow=TRUE)
prop.test(Tabla, n, p = 0, alternative=c("two.sided","less",
"greater"), conf.level = 0.95, correct = TRUE)
```

Los argumentos cumplen la misma función que se describió antes cuando se trató el procedimiento de prueba de hipótesis para una sola proporción, solo hay que destacar que el valor de ***p*** siempre será cero.

Ejemplo 5.5. La administración de una planta de tratamiento de agua potable desea probar si el método de desinfección con luz ultravioleta resulta ser más eficiente que la desinfección con cloro gaseoso en función de la presencia de *E. coli* en muestras de aguas a la salida de esta etapa del sistema de tratamiento. Para tal fin, se seleccionaron aleatoriamente 80 muestras a la salida del sistema luego de utilizar los dos métodos de desinfección mencionados, donde se encontró que el sistema de desinfección con cloro gaseoso 25 muestras fueron positivas al examen de *E. coli* y para el método de desinfección con luz ultravioleta 18 resultaron ser positivas. Con base en lo anterior, a un nivel de significancia de 0.05, ¿se puede establecer como cierta esta afirmación?

Solución

Tomando como $x =$ resultado positivo al análisis de *E. coli*, el interés se centra en probar el siguiente sistema de hipótesis

$$H_0 : p_{UV} - p_{Cl} = 0$$

$$H_1 : p_{UV} - p_{Cl} < 0$$

Al tabular la evidencia muestral de casos favorables (resultado positivo al análisis de *E. coli*) y casos no favorables (Resultados negativos al análisis de *E. coli*) en una tabla 2x2 tenemos el siguiente arreglo

	Frecuencias observadas análisis de *E. coli*		
	Positivo	**Negativo**	**Total**
Luz UV	18	62	80
Cloro gaseoso	25	55	80
Total	43	117	160

Así, con $\hat{p}_1 = 0.225$, $\hat{p}_2 = 0.313$, $\hat{p} = 0.269$, y $\hat{q} = 0.731$, el valor del estadístico de prueba sería

$$z = \frac{0.225 - 0.313}{\sqrt{(0.269)(0.731)\left[(1/80) + (1/80)\right]}} \therefore z = -1.225$$

De la Tabla A.3 del apéndice, tenemos que el valor crítico es $z_\alpha = z_{0.05} = 1.645$.

De esta forma como $z = -1.225 > -z_{0.05} = -1.645$, no se rechaza H_0 a un nivel de significancia de 0.05, es decir, no existen diferencias significativas entre la verdadera proporción de los resultados positivos al examen de *E coli*, después de un proceso de desinfección con luz UV o con cloro gaseoso, con una confiabilidad del 95%. Por lo tanto, la administración de la planta de tratamiento de agua potable puede inclinarse por cualquiera de los métodos de desinfección, dependiendo de cuál les resulta más factible económicamente.

Ahora veamos la resolución de este mismo ejercicio con la metodología basada en la distribución chi-cuadrado. Para ello, realizamos el cálculo de las frecuencias esperadas, las cuales se consignas en la siguiente tabla

	Frecuencias esperadas análisis de *E. coli*	
	Positivo	**Negativo**
Luz UV	21.52	58.48
Cloro gaseoso	21.52	58.48

A partir de estos resultados el valor del estadístico de prueba es

$$\chi^2 = \frac{(18-21.52)^2}{21.52} + \frac{(62-58.48)^2}{58.48} + \frac{(25-21.52)^2}{21.52} + \frac{(55-58.48)^2}{58.48}$$

$$\chi^2 = 1.557$$

De la Tabla A.5 del apéndice, el valor crítico para este caso es $\chi^2_{1-\alpha} = \chi^2_{0.95} = 0.00393$. Dado que $\chi^2 = 1.557 > \chi^2_{0.95} = 0.00393$, se llega a la misma de decisión de no rechazar H_0 a un nivel de significancia de 0.05, llegando a las misma conclusión antes mencionada.

Ahora veamos la salida de resultados de la consola de R, después de la aplicación de la función ***prop.test***

```
> Tabla<-matrix(c(18,62,25,55),nrow=2,ncol=2,byrow=TRUE)
>
prop.test(Tabla,alternative="less",conf.level=0.95,correct=FALSE)

        2-sample test for equality of proportions without
continuity
        correction

data:  Tabla
X-squared = 1.5583, df = 1, p-value = 0.106
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.02723057
sample estimates:
prop 1 prop 2
0.2250 0.3125
```

Nótese de esta salida de resultados que el p-valor = 0.106, es mayor que el nivel de significancia seleccionado, por lo tanto la evidencia muestral, señala el no rechazo de la hipótesis nula.

5.10. Prueba de hipótesis sobre dos varianzas poblacionales: Prueba de homogeneidad de varianzas

En capítulos anteriores, se hizo hincapié en que muchos procedimientos inferenciales, como la prueba *t* para dos medias de poblaciones normales, para poder ser aplicados requieren de un conocimiento anticipado de la igualdad (homogeneidad) o diferencia (heterogeneidad) de las varianzas de las dos poblaciones que se desean contrastar, a partir de la evidencia muestral con que se disponga. Así mismo, se mencionó que la distribución de probabilidad utilizada para este propósito es la distribución *F* de Fisher-Snedecor. En esta sección presentaremos un procedimiento de prueba para

probar la hipótesis de igualdad de varianzas de dos poblaciones normales, a través de los siguientes sistemas de hipótesis posibles

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2 \qquad H_1 : \sigma_1^2 < \sigma_2^2 \qquad H_1 : \sigma_1^2 > \sigma_2^2$$

Siendo la alternativa bilateral la de uso más recurrente, pues generalmente el interés no se centra en establecer que población tiene mayor o menor varianza que la otra, sino establecer si en efecto difieren significativamente, o no.

El valor *f* para probar la hipótesis nula está dado por la siguiente expresión

$$f = \frac{s_1^2}{s_2^2}$$

Que sigue una distribución F de Fisher-Snedecor con $\upsilon_1 = n_1 - 1$ y $\upsilon_2 = n_2 - 1$ grados de libertad. En esta expresión, s_1^2 y s_2^2 son las varianzas calculadas de las dos muestras independientes de tamaño n_1 y n_2, respectivamente.

Así, a un nivel de significancia α, las regiones críticas para la alternativa bilateral son $f \leq -f_{1-\alpha/2(\upsilon_1,\upsilon_2)}$ y $f \geq f_{\alpha/2(\upsilon_1,\upsilon_2)}$. Para las alternativas unilaterales de cola inferior y superior, las regiones críticas son, respectivamente, $f \leq -f_{1-\alpha(\upsilon_1,\upsilon_2)}$ y $f \geq f_{\alpha(\upsilon_1,\upsilon_2)}$.

Un resumen del procedimiento descrito para probar la igualdad de las varianzas de dos poblaciones normales e independientes se muestra en el siguiente recuadro

Hipótesis nula: $H_0 : \sigma_1^2 = \sigma_2^2$

Valor calculado del estadístico de prueba: $f = \frac{s_1^2}{s_2^2}$ con $\upsilon_1 = n_1 - 1$ y $\upsilon_2 = n_2 - 1$ grados de libertad

Hipótesis alternativa	**Región de rechazo a un nivel de significancia α**
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$f \leq -f_{1-\alpha/2(\upsilon_1,\upsilon_2)}$ o $f \geq f_{\alpha/2(\upsilon_1,\upsilon_2)}$
$H_1 : \sigma_1^2 < \sigma_2^2$	$f \leq -f_{1-\alpha(\upsilon_1,\upsilon_2)}$
$H_1 : \sigma_1^2 > \sigma_2^2$	$f \geq f_{\alpha(\upsilon_1,\upsilon_2)}$

En R, para la ejecución de este procedimiento de prueba se utiliza la función ***var.test***, donde se deben especificar los argumentos ***x*** e ***y***, que corresponden a los vectores de datos de las muestras a partir de las cuales se hará la inferencia, ***ratio*** indica el valor de la razón entre las varianzas a probar, al tratarse de una prueba de igualdad de las varianzas, se dejará su valor por defecto de 1; ***alternative***, como antes especifica el sentido de la hipótesis alternativa, y ***conf.level***, fija el nivel de confiabilidad de la prueba

```
var.test(x, y, ratio = 1, alternative = c("two.sided", "less",
"greater"),conf.level = 0.95)
```

Ejemplo 5.6. En el ejemplo 5.3, se realizó un procedimiento de prueba de hipótesis sobre la diferencias de las medias del porcentaje de humedad relativa en muestras de arena húmeda y arena seca, tomadas dentro del marco del programa CAPT, bajo el supuesto de que los datos provenían de poblaciones normales con varianzas diferentes. A un nivel de significancia de 0.05, verifíquese el supuesto de homogeneidad de varianza.

% de humedad	
Arena húmeda	Arena seca
84	83
75	69
82	82
83	80
87	84
78	53
85	65
86	69
77	67
77	73

Solución

A partir de lo planteado, es evidente que se tiene por como objetivo probar la siguiente hipótesis

$$H_0 : \sigma_{AH}^2 = \sigma_{AS}^2$$

$$H_1 : \sigma_{AH}^2 \neq \sigma_{AS}^2$$

De esta forma con $s_{AH}^2 = 18.49$ y $s_{AS}^2 = 97.83$, el valor del estadístico de prueba es

$$f = \frac{18.49}{97.83} \therefore f = 0.189$$

El valor critico según la Tabla A.6 del apéndice después de un proceso de interpolación lineal simple es $f_{\alpha/2(\upsilon_1,\upsilon_2)} = f_{0.025(9,9)} = 4.536$, como se mencionó en la sección referente al modelado de la distribución *F* de Fisher-Snedecor,

$$f_{0.975(9,9)} = \frac{1}{f_{0.025(9,9)}} = \frac{1}{4.536} = 0.220$$

Así, como $f = 0.189 < f_{0.975(9,9)} = 0.220$, se rechaza H_0 a favor de H_1, a un nivel de significancia de α. Por lo tanto, los resultados del procedimiento de prueba son concluyentes en que las varianzas de las dos poblaciones del porcentaje de humedad en arena húmeda y arena seca difieren significativamente con una confiabilidad del 95%.

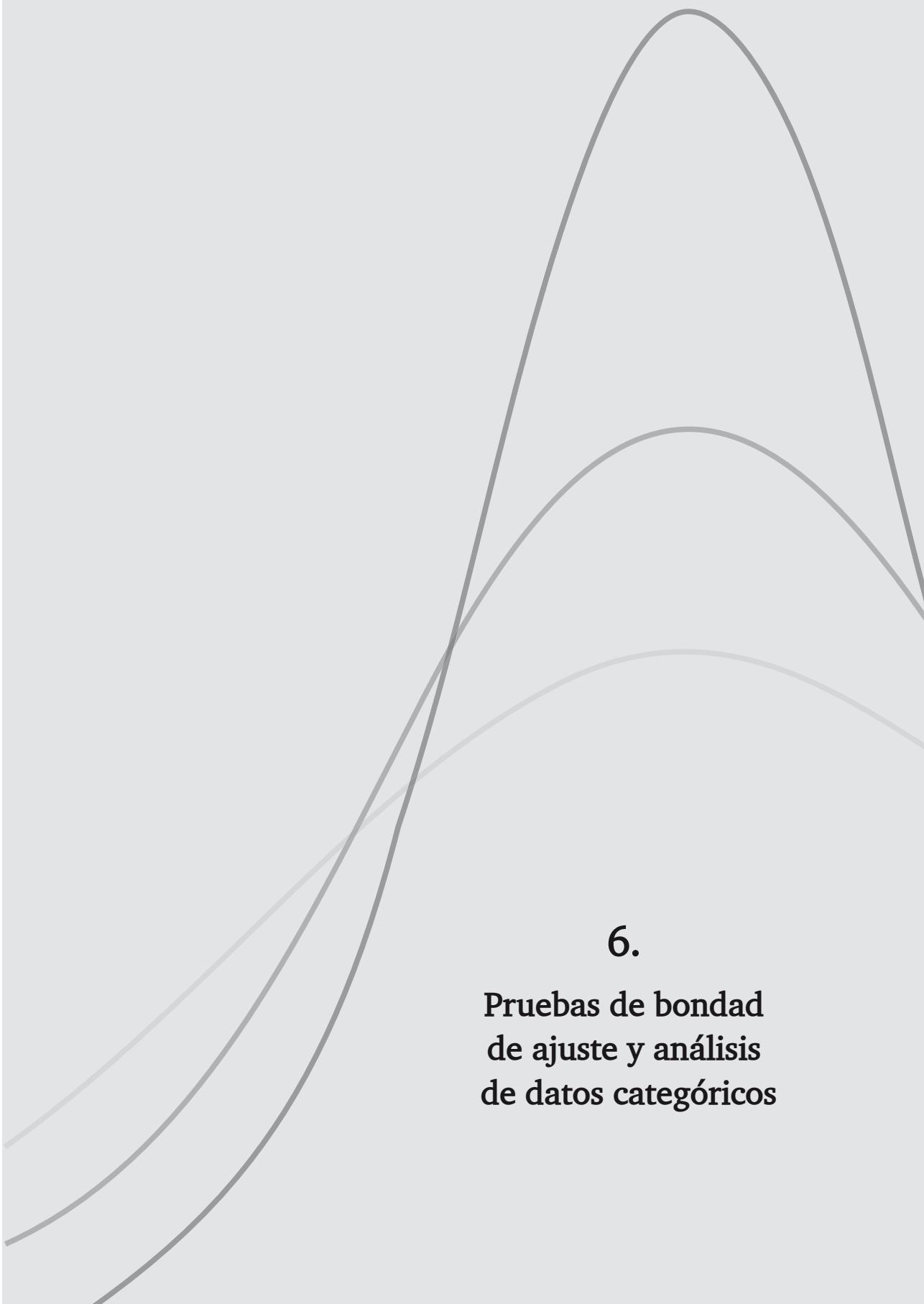
La salida de resultados en la consola de R para este tipo de procedimiento, se muestra a continuación

```
> AH<-c(84,75,82,83,87,78,85,86,77,77)
> AS<-c(83,69,82,80,84,53,65,69,67,73)
>
var.test(AH,AS,ratio=1,alternative="two.sided",conf.level=0.95)

        F test to compare two variances

data:  AH and AS
F = 0.189, num df = 9, denom df = 9, p-value = 0.02077
alternative hypothesis: true ratio of variances is not equal to
1
95 percent confidence interval:
 0.04694084 0.76084660
sample estimates:
ratio of variances
         0.1889835
```

Nótese que en estos resultados, el p-valor es menor que el nivel de significancia elegido de 0.05, por lo tanto, se toma la misma decisión de rechazar la hipótesis nula a favor de la hipótesis alternativa.



6.

Pruebas de bondad de ajuste y análisis de datos categóricos

6.1. Generalidades

Recuérdese que una hipótesis estadística es una afirmación con respecto a una o más característica que se desconocen de una población de interés. En el capítulo anterior, se discutieron procedimientos exclusivos para probar el valor θ, que asumía algún parámetro en una o dos poblaciones. En este capítulo, se examinaran las pruebas de hipótesis estadísticas en las que la característica que se desconoce es alguna propiedad de la forma funcional de la distribución que se muestrea (Conavos, 1988), esto es, determinar si una población tiene una distribución teórica específica, a través del análisis de que tan buen ajuste existe entre la frecuencia de ocurrencia de las observaciones en una muestra y las frecuencias esperadas que se obtienen a partir de la distribución hipotética (Walpole *et al.*, 2007). Este tipo de procedimientos son conocidos como **pruebas de bondad de ajuste**, y son de gran utilidad, puesto que muchos procedimientos estadísticos, requieren tener un conocimiento anticipado de que los datos tengan una distribución teórica específica, usualmente una distribución normal, como se ha mencionado en los capítulos anteriores, en donde se expusieron los métodos usados para la construcción de intervalos de confianza y pruebas de hipótesis para la media, varianza o proporción de una o dos poblaciones.

Existen distintas pruebas de bondad de ajuste que se utilizan en función del tipo de datos y la distribución teórica esperada. Una clasificación de los ajustes más empleados son (Guisande *et al.*, 2011):

1. Prueba chi-cuadrado, cuando tratamos muestras categorizadas (los datos son códigos asignados a los valores de una variables cualitativa o a las clases en las que se agrupan los valores de una variables cuantitativa). Útil para evaluar el ajuste de los datos a cualquier distribución.
2. Prueba de Kolmogov-Smirnov (test K-S), cuando tratamos muestras no categorizadas (variables cuantitativas, continuas o discretas, no agrupadas en intervalos o clases). Útil para evaluar el ajuste de los datos a cualquier distribución.

3. Prueba de Shapiro-Wilk, cuando tratamos muestras no categorizadas (variables cuantitativas, continuas o discretas, no agrupadas en intervalos o clases). Útil para solo para evaluar el ajuste de los datos a la distribución normal.

Además, se discutirán **las pruebas de independencia** y **homogeneidad** entre dos variables aleatorias en las que la evidencia muestral se obtiene mediante la clasificación de cada variable en un cierto número de categorías, es decir, aquellos procedimientos en los que se busca establecer si dos variables aleatorias cualitativas, difieren significativamente o no hay asociación entre las mismas, para el caso de las pruebas de independencia, o por su parte, si las proporciones para cada categoría de dos o más variables categóricas, se pueden asumir como iguales (homogéneas).

6.2. Prueba de bondad de ajuste chi-cuadrado

Este contraste se puede aplicar tanto a distribuciones continuas (con los datos previamente agrupados en clases o categorías) como a distribuciones discretas o variables cualitativas. Su uso se basa en cuantificar las diferencias entre las frecuencias observadas en cada clase y las frecuencias esperadas, partiendo de la hipótesis nula de que los datos se ajustan a una distribución teórica específica, que puede ser la normal, Poisson, etc. Para su aplicación, en las clases existentes, se contabiliza el número de casos observados (o_i) y, a través de la función de distribución teórica que se desea testar, se calcula el número de casos esperados (e_i). A partir de estas frecuencias se calcula el valor del estadístico χ^2, con la siguiente expresión (Guisande *et al.*, 2011; Pérez, 2004; Walpole *et al.*, 2007)

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Donde χ^2 es un valor de una variable aleatoria cuya distribución muestral se aproxima muy de cerca con la distribución chi-cuadrado con $\upsilon = k - 1$ grados de libertad.

Si la diferencia entre las frecuencias observadas y las esperadas es muy pequeña, el valor de χ^2 también lo será, lo que indicaría un buen ajuste de los datos, que conduciría al no rechazo de la hipótesis nula. En caso contrario, si las diferencias entre las frecuencias observadas y las esperadas es grande, el valor de χ^2 también lo será, e indicará un ajuste deficiente de

los datos, evidencia suficiente para rechazar la hipótesis nula y concluir que los datos no siguen la distribución teórica que se analiza.
La región critica para este tipo de contraste, cae en la cola derecha de la distribución chi-cuadrado. Así, para un nivel de significancia α, encontramos el valor crítico de la Tabla A.5 del apéndice y, entonces, $\chi^2 > \chi^2_\alpha$ constituye la región critica o de rechazo.

Se aconseja que si el número de casos (frecuencias) esperados dentro de cada una de las categorías es menor de 5 se combinen varias categorías en una, hasta conseguir que todas tengan una frecuencia esperada mayor o igual de 5, lo que conlleva a una reducción del número de grados de libertad (Guisande *et al.*, 2011; Pérez, 2004; Walpole *et al.*, 2007, Conavos, 1988).

Ejemplo 6.1. A continuación se muestran los niveles de presión sonora de ruido medida en decibeles (dB) en diferentes estaciones de muestreo, en un estudio de ruido ambiental en la ciudad de Cali en horario diurno (Vargas, 2007), utilizados en el ejemplo 2.2.

63.7	75.0	70.5	72.1	67.2	65.1	59.6	64.1	61.1	62.0
66.9	76.3	73.7	74.1	62.3	55.3	70.6	53.3	65.9	64.0
66.8	71.4	71.0	76.5	69.4	71.3	65.3	62.5	62.6	58.7
75.3	77.4	56.1	57.3	60.5	72.3	74.0	62.3	50.2	68.2
70.8	71.6	69.0	71.6	75.0	64.6	74.9	75.4	50.9	61.6

A un nivel de significancia de 0.05, probar la hipótesis de que estos datos siguen una distribución normal con media $\mu = 68$ y desviación estándar $\sigma = 5$.

Solución

Los datos del enunciado, sugieren probar la siguiente hipótesis

H_0 : Los datos del nivel de presión sonora se distribuyen normalmente
H_1 : Los datos del nivel de presión sonora no se distribuyen normalmente

Antes de hallar el valor de χ^2, organizamos los datos en una tabla de distribución de frecuencias, siguiendo el procedimiento estudiado en el capítulo 2, donde se obtiene el siguiente arreglo

Intervalos de clase	o_i	e_i
50 – 55	3	0.2
55 – 60	5	2.5
60 – 65	12	11
65 – 70	9	19.1
70 – 75	16	13.2
75 – 80	5	3.6

Dado que la distribución teórica escogida es la normal, las frecuencias esperadas se determinan calculando el área bajo la curva normal hipotética que cae entre los diversos límites de clase (Walpole *et al.*, 2007). Al ser esta una probabilidad (frecuencia relativa) se utiliza la expresión

$$e_i = np_i$$

para calcular la frecuencia esperada de cada clase. Ilustraremos esto para la primera clase que se muestra en el arreglo tabular de los datos del ejemplo. Los valore de *z* que corresponden a los límites de clase de la primera clases son

$$z_1 = \frac{50-68}{5} = -3.6 \qquad z_2 = \frac{55-68}{5} = -2.6$$

De la Tabla A.3 del apéndice, encontramos que el área entre $z_1 = -3.6$ y $z_2 = -2.6$ es

$$\begin{aligned} P(-3.6 \le Z \le -2.6) &= P(Z \le -2.6) - P(Z \le -3.6) \\ &= 0.0047 - 0.0002 = 0.0045 \end{aligned}$$

De aquí que la frecuencia esperada para la primera clase es

$$e_1 = (50)(0.0045) = 0.2$$

Acostumbrándose a redondear las frecuencias esperadas que se calculan a una sola cifra decimal (Walpole *et al.*, 2007).

Ya calculadas todas las frecuencias esperadas, se determina el valor del estadístico de prueba, teniendo la precaución de combinar clases (valores sombreados) cuando el valor de las frecuencias esperadas sean inferiores a 5. Así,

$$\chi^2 = \frac{(20-13.7)^2}{13.7} + \frac{(9-19.1)^2}{19.1} + \frac{(21-16.8)^2}{16.8} \therefore \chi^2 = 9.29$$

Según la Tabla A.5 del apéndice el valor cítrico es $\chi^2_{\alpha,k-1} = \chi^2_{0.05,2} = 5.991$.

De esta forma, como $\chi^2 = 9.29 > \chi^2_{0.05,2} = 5.991$, se rechaza H_0 a favor de H_1 a un nivel de significancia de 0.05; concluyéndose que los datos del nivel de presión sonora de ruido ambiental, en horario diurno en la ciudad de Cali no presentan un buen ajuste con la curva de distribución normal, o en términos más simples, estos datos no siguen una distribución normal con media $\mu = 68$ y desviación estándar $\sigma = 5$, con una confiabilidad del 95%. Esto se puede visualizar gráficamente al comparar la curva de la distribución normal, con los parámetros dados, y la curva de densidad del conjunto de datos (Figura 6.1).

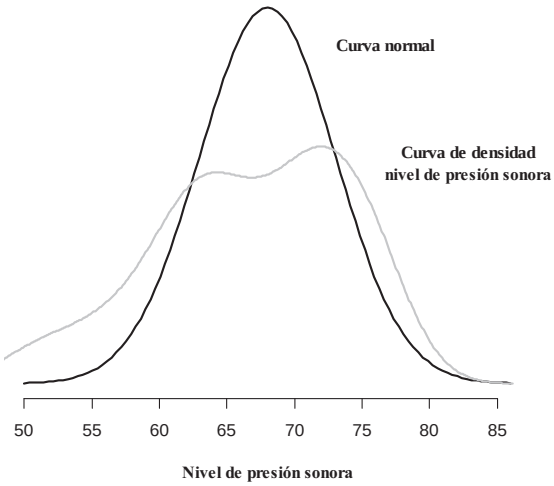


Figura 6.1. Comparación gráfica de la curva normal y la curva de densidad de los datos

En R, se puede utilizar la función ***chisq.test***, vista anteriormente para realizar pruebas de bondad de ajuste de los datos a cualquier tipo de distribución. Sin embargo, antes de la aplicación de esta función se debe hacer una preparación de los datos, como en nuestro caso particular, donde nuestro objetivo se centra en determinar si los datos dados se ajustan a una distribución normal. A continuación se muestra la salida de resultados de R, del procedimiento programático que se realizó para tal fin y se harán comentarios al respecto, donde se considere necesario.

```
> Ruido<-c(63.7,75.0,70.5,72.1,67.2,65.1,59.6,64.1,61.1,62.0,66.9,
76.3,73.7,74.1,62.3,55.3,70.6,53.3,65.9,64.0,66.8,71.4,71.0,76.5,69.
4,71.3,65.3,62.5,62.6,58.7,75.3,77.5,56.1,57.3,60.5,72.3,74.0,
62.3,50.2,68.2,70.8,71.6,69.0,71.6,75.0,64.6,74.9,75.4,50.9,61.6)
> Tabla<-hist(Ruido,plot=FALSE)
> mu=68
> sigma=5
> x<-Tabla$breaks
> Frec.obs<-Tabla$count
> n<-sum(Frec.obs)
> z<-(x-mu)/sigma
> Frec.esp<-round((pnorm(z[2:7])-pnorm(z[1:6]))*n,1)
> Frec.esp>=5
[1] FALSE FALSE  TRUE  TRUE  TRUE FALSE
> Frec.esp
[1]  0.2  2.5 11.0 19.1 13.2  3.6
> Frec.obs<-c(sum(Frec.obs[1:3]),Frec.obs[4],sum(Frec.obs[5:6]))
> Frec.esp<-c(sum(Frec.esp[1:3]),Frec.esp[4],sum(Frec.esp[5:6]))
> Frec.esp>=5
[1] TRUE TRUE TRUE
> chisq.test(Frec.obs,p=Frec.esp,correct=FALSE,rescale.p=TRUE)

        Chi-squared test for given probabilities

data:  Frec.obs
X-squared = 9.2104, df = 2, p-value = 0.01
```

Nótese que basando la decisión en el p-valor = 0.01, también se llegaría a la decisión de rechazar de H_0 por ser este menor que el nivel de significancia preseleccionado.

La aplicación de este procedimiento en R se realizó inicialmente con construir un vector de datos que contiene las medidas de nivel de presión sonora. Dado el volumen considerado de observaciones, habría sido más práctico tabular los datos en una hoja de cálculo de Excel, guardarlos bajo la extensión *.csv* y luego cargarlos en R a través de la función ***read.csv2***, como se ha visto en ejemplo anteriores. Paso seguido se construyó la tabla de distribución de frecuencias (*Tabla*), siguiendo el procedimiento descrito en el capítulo 2, para el tratamiento de variables continuas, donde básicamente se hace "trampa" al software, ordenándole que construya un histograma de frecuencias con la función ***hist***, pero que no lo visualice, dándole a su argumento ***plot***, el valor lógico ***FALSE***. Luego se asignaron a dos objetos diferentes (*mu* y *sigma*) el valor de la media y desviación

estándar poblacional, respectivamente, que sigue la curva normal teórica con la que se desea comparar los datos. A continuación, se creó un objeto llamado ***x*** al que se le asignaron los límites de clase (*breaks*) de cada intervalo del objeto *Tabla* (*Tabla$breaks*). Se asignaron a diferentes objetos (*Frec.obs*, *n* y *z*), los valores de las frecuencias absolutas (observadas), el número total de observaciones y los límites de clase estandarizados, respectivamente. Luego se realizó el cálculo de las frecuencias esperadas, que se consignaron en un objeto denominado *Frec.esp*. Seguido de esto, se consultó si todas las frecuencias esperadas eran cuanto menos iguales a 5 (*Frec.esp>=5*), al darnos cuenta que no todas cumplían este requerimiento, se procedió a realizar las agrupaciones pertinentes tanto en las frecuencias esperadas como en las observadas, para terminar con la aplicación de la función ***chisq.test***, en cuyos argumentos ***p*** establece cuales con las frecuencias esperadas de la muestra, ***correct***, comentado anteriormente indica si se realiza la corrección de continuidad de Yates y el argumento ***rescale.p***, al darle el valor lógico ***TRUE***, le ordena a R realizar un reajuste (en caso de ser necesario) de las probabilidades, para que estas sumen 1.

Como en los casos prácticos de la vida real, se tiene más interés en determinar si un conjunto de observaciones se ajustan a una distribución normal, el modelado de las demás distribuciones no se tocará en este texto para no extendernos al respecto. Sin embargo, se invita al lector que quiera profundizar más al respecto, la consulta de otros recursos bibliográficos al respecto; en la web, se encuentran muchos documentos en castellano que realizan incluso la modelación en R.

6.3. Test de Kolmogorov-Smirnov

En la sección anterior, vimos que un requerimiento para la aplicación del test de bondad de ajuste chi-cuadrado era el agrupamiento de los datos observados en un número finito de intervalos de clase, sobre todo cuando el modelo propuesto es continuo, como lo es el caso de la distribución normal. Este requisito para el test chi-cuadrado, puede ser una limitación del mismo, pues implica disponer de una muestra más o menos grande. Una prueba de bondad de ajuste más apropiada que la chi-cuadrado y que en la actualidad tiene un amplio uso es el test de Kolmogorov-Smirnov, que no requiere el agrupamiento de los datos y es de más fácil aplicación.

Esta prueba es adecuada para testar la normalidad de una muestra si el número de datos es grande ($n > 30$), aunque se puede usar tanto para

muestras grandes como pequeñas. También se puede usar para testar otras distribuciones como la Binomial o de Poisson. Es un test muy conservador que se aplica a variables continuas, y se basa en la determinación de la máxima diferencia (D) entre las frecuencias acumuladas observadas (Ao_i) y las frecuencias acumuladas esperadas (Ae_i), partiendo de la hipótesis nula de que los datos se ajustan a una distribución determinada (Guisande *et al.*, 2011). El cálculo del estadístico de prueba se realiza a través de la siguiente expresión

$$D = \max \left| Ao_i - Ae_i \right|$$

Una vez calculado el estadístico D se contrasta con el valor critico ($D_{\alpha,n}$) que se encuentra tabulado, y que podemos encontrar en la Tabla A.7 del apéndice, para un nivel de significancia preseleccionado. De esta forma, la hipótesis nula es rechazada cuando $D > D_{\alpha,n}$, o cuando el p-valor es menor que el nivel de significancia elegido.

En R, se dispone de una función que realiza el test de Kolmogorov-Smirnov de forma rápida, llamada ***ks.test***, en cuyos argumentos se debe indicar el vector de datos (***x***) a quien se le aplicará el test, entre comillas (caracter) se indica el modelo teórico de probabilidad propuesto (por ejemplo, "*pnorm*"), y por último se especifican los parámetros del modelo, para el caso de la distribución normal, estos son la media (***mean***) y la desviación estándar (***sd***).

```
ks.test(x, "Distribution", mean = mean(x), sd = sd(x))
```

Ejemplo 6.2. En el ejemplo 5.1, se supuso que la concentración de dióxido de nitrógeno (NO_2) en el casco urbano del municipio de Guamal, Magdalena, expresada como $\mu g/m^3$ en un tiempo de muestreo de 24 horas, obtenidos en un estudio de contaminación atmosférica, seguían una distribución normal. Dichas concentraciones de NO_2 se muestran a continuación:

110	120	105	100	140	115	110	121	130	145
110	150	130	120	105	150	140	115	110	120

A un nivel de significancia de 0.05, comprobar que se cumple este supuesto de normalidad.

Solución

El interés en esta situación se centra probar la siguiente hipótesis

H_0: Los datos del nivel de presión sonora se distribuyen normalmente
H_1: Los datos del nivel de presión sonora no se distribuyen normalmente

Para el cálculo del estadístico de prueba nos basaremos en el siguiente arreglo tabular de los datos ordenados en forma creciente, para facilitar las operaciones aritméticas

NO_2	Ao_i	z	Ae_i	$\lvert Ao_i - Ae_i \rvert$
100	0.05	-1.43	0.0764	0.0264
105	0.10	-1.11	0.1335	0.0335
105	0.15	-1.11	0.1335	0.0165
110	0.20	-0.79	0.2148	0.0148
110	0.25	-0.79	0.2148	0.0352
110	0.30	-0.79	0.2148	0.0852
110	0.35	-0.79	0.2148	0.1352
115	0.40	-0.47	0.3192	0.0808
115	0.45	-0.47	0.3192	0.1308
120	0.50	-0.15	0.4404	0.0596
120	0.55	-0.15	0.4404	0.1096
120	0.60	-0.15	0.4404	0.1596
121	0.65	-0.08	0.4681	0.1819
130	0.70	0.49	0.6879	0.0121
130	0.75	0.49	0.6879	0.0621
140	0.80	1.14	0.8729	0.0729
140	0.85	1.14	0.8729	0.0229
145	0.90	1.46	0.9279	0.0279
150	0.95	1.78	0.9625	0.0125
150	1.00	1.78	0.9625	0.0375

De esta manera $D = \max \lvert Ao_i - Ae_i \rvert = 0.1819$ (valor sombreado en la tabla). El valor crítico según la Tabla A.7 del apéndice es $D_{\alpha,n} = D_{0.05,20} = 0.294$. Así, como $D = 0.1819 < D_{0.05} = 0.294$, no se rechaza H_0 a un nivel de significancia de 0.05, concluyéndose que los datos de concentración NO_2 en el casco urbano del municipio de Guamal, Magdalena, presentan un buen ajuste a

la distribución normal, o se encuentran normalmente distribuidos, con una confiabilidad del 95%.
La salida de resultados de R, después de la aplicación de la función ***ks.test*** es la siguiente

```
> NO2<-
c(110,120,105,100,140,115,110,121,130,145,110,150,130,120,
105,150,140,115,110,120)
> ks.test(NO2,"pnorm",mean=mean(NO2),sd=sd(NO2))

        One-sample Kolmogorov-Smirnov test

data:  NO2
D = 0.1832, p-value = 0.513
alternative hypothesis: two-sided
```

Nótese que el valor del estadístico de prueba (*D*), es ligeramente mayor al que se obtuvo mediante cálculos, esto se debe a la perdida de cifras decimales al realizar redondeos en el cálculo de los valor de *z*. Sin embargo, al basar la decisión en el p-valor, al ser este mayor que el nivel de significancia preseleccionado, es evidencia suficiente para no rechazar la hipótesis nula. No obstante, no es correcto que en esta función se utilice el p-valor para la hipótesis de normalidad, ya que la distribución de la estadística de prueba es diferente cuando se estiman los parámetros (Gross, 2015), que podría hacer que se incurra en errores a la hora de rechazar (o no) la hipótesis nula. En virtud de esto, este test fue recalculado para la distribución normal, al estudiar y tabular las frecuencias esperadas a partir de la media y la varianza muestral (Lillifors, 1967) y se conoce como test *K-S-L* (Guisande *et al.*, 2011), ampliamente usado, a pesar de tener la limitación de ser aplicable más eficientemente a muestras grandes. El cálculo del estadístico de prueba y la región critica de definen igual al descrito antes, pero los valores críticos se encuentran disponibles en la Tabla A.8 del apéndice.

Para la aplicación del test *K-S-L* en R, se utiliza la función ***lillie.test***, en cuyo argumento se especifica el vector de datos en donde se encuentran las observaciones a partir de las cuales se realizará la aplicación del test. Esta función requiere la instalación previa del paquete "*nortest*" (Gross, 2015).

```
lillie.test(x)
```

Para nuestro caso de estudio con $D = 0.1819$ y el valor crítico según la Tabla A.8 del apéndice igual a $D_{0.05,\,20} = 0.190$, como $D = 0.1819 < D_{0.05,\,20} = 0.190$, se decide no rechazar H_0 con una confiabilidad del 95%.

La salida de resultados en R para la aplicación del test *K-S-L*, se muestra a continuación

```
> NO2<-
c(110,120,105,100,140,115,110,121,130,145,110,150,130,120,
105,150,140,115,110,120)
> library(nortest)
> lillie.test(NO2)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  NO2
D = 0.1832, p-value = 0.07707
```

Obsérvese, que dado que el p-valor = 0.08 es mayor que el nivel de significancia seleccionado, se decide no rechazar la hipótesis nula, concluyéndose que existe un ajuste normal de los datos con una confiabilidad del 95%.

6.4. Test de Shpairo-Wilk

El test de Shapiro-Wilk, es la prueba más recomendable para testar la normalidad de una muestra, sobre todo si se trabaja con un número pequeño de observaciones ($n \leq 30$). Téngase en cuenta que es una prueba estadística orientada solo a testar el ajuste de un conjunto de observaciones a la distribución normal (Shapiro & Wilk, 1965). Su principio de aplicación se basa en medir el ajuste de los datos a una recta probabilística normal (Figura 6.2); si el ajuste fuera perfecto los puntos formarían una recta de 45º, es decir, las frecuencias observadas y esperadas, serían iguales (Guisande *et al.*, 2011).

El cálculo del estadístico de prueba *W*, dada una muestra aleatoria de tamaño *n*, $x_1, x_2, ..., x_n$, se obtiene al aplicar el siguiente procedimiento (Shapiro & Wilk, 1965):

1. Ordenar las observaciones para obtener una muestra ordenada de forma ascendente $y_1 \leq y_2 \leq ... \leq y_n$.

2. Calcular

$$s^2 = \sum_1^n \left(x_i - \bar{x}\right)^2 = \sum_1^n \left(y_i - \bar{y}\right)^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2$$

3.1. Si n es par, $n = 2k$, calcular

$$b = \sum_{i=1}^{n} a_{n-i+1}\left(y_{n-i+1} - y_i\right)$$

Donde los valores de a_{n-i+1} se encuentran tabulados en la Tabla A.9 del apéndice.

3.2. Si n es impar, $n = 2k+1$, el cálculo de b es igual que el descrito en 3.1, ya que a_{k+1} cuando $n = 2k+1$. por lo tanto se encuentra que

$$b = a_n\left(y_n - y_1\right) + \ldots + a_{k+2}\left(y_{k+2} - y_k\right)$$

Donde el valor de y_{k+1}, la mediana de la muestra, no es tenida en cuenta en el cálculo de b.

4. Calcular

$$W = \frac{b^2}{s^2}$$

5. Se busca el valor crítico de la distribución W al nivel de significancia elegido. Dicho valor crítico se halla en la Tabla A.10 del apéndice.

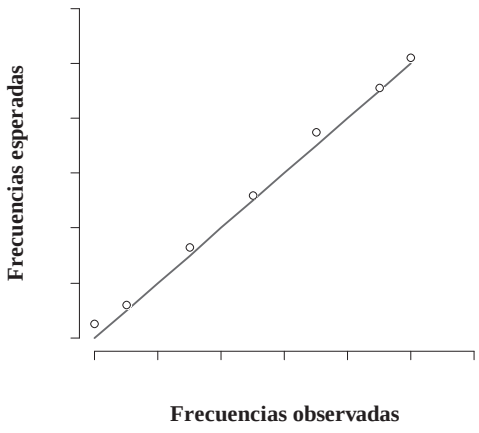


Figura 6.2. Ajuste de las frecuencias (probabilidades) observadas y esperadas a una recta de ajuste perfecto a una distribución normal.

Una vez calculado el estadístico W se contrasta con el valor crítico tabulado $W_{\alpha,n}$. Como esta prueba mide el ajuste a una recta y no la distancia a la distribución normal (como en las pruebas descritas anteriormente), y se puede interpretar de forma aproximada como un coeficiente de correlación entre los valores observados y esperados (un valor próximo a 1 indica buen ajuste, y próximo a 0, mal ajuste), la hipótesis nula se rechaza cuando $W \leq W_{\alpha,n}$.

En R, la aplicación del test de Shapiro-Wilk, se ejecuta de una forma fácil y sencilla a través de la función ***shapiro.test***, en cuyo argumento solo se especifica el vector de datos al cual se aplicará el test.

```
shapiro.test(x)
```

Ejemplo 6.3. Retomando los datos del Ejemplo 4.2, donde se asumió que los datos del porcentaje de remoción de DQO de un rector anaerobio de flujo ascendente (UASB, por sus siglas en ingles), que trata las aguas residuales domesticas de un sector residencial de la ciudad de Barranquilla, sigue una distribución normal. Con los resultados del porcentaje de remoción de DQO se muestran a continuación, probar la hipótesis nula de que estos datos se encuentran normalmente distribuidos a un nivel de significancia de 0.05.

61.5	50	25	44.4	25	25	50	57.1	50
50	50	16.6	50	50	66.6	75	75	66.6

Solución

Al disponer de un tamaño de muestra pequeño, el test de Shapiro-Wilk, es ideal para evaluar el supuesto de normalidad de estas observaciones. Los cálculos necesarios se resumen en la siguiente tabla donde se inicia el ordenamiento de los datos en orden ascendente

i	y_i	$\bar{y}$	y_{n-i+1}	$y_{n-i+1}-y_i$	a_{n-i+1}	$b=a_{n-i+1}\left(y_{n-i+1}-y_i\right)$	$s^2=\left(y_i-\bar{y}\right)^2$
1	16.6	49.32	75.0	58.4	0.4886	28.5342	1070.5984
2	25.0		75.0	50.0	0.3253	16.2650	591.4624
3	25.0		66.6	41.6	0.2553	10.6205	591.4624
4	25.0		66.6	41.6	0.2027	8.4323	591.4624
5	44.4		61.5	17.1	0.1587	2.7138	24.2064

i	y_i	$\bar{y}$	y_{n-i+1}	$y_{n-i+1}-y_i$	a_{n-i+1}	$b=a_{n-i+1}\left(y_{n-i+1}-y_i\right)$	$s^2=\left(y_i-\bar{y}\right)^2$
6	50.0		57.1	7.1	0.1197	0.8499	0.4624
7	50.0		50.0	0.0	0.0837	0.0000	0.4624
8	50.0		50.0	0.0	0.0496	0.0000	0.4624
9	50.0		50.0	0.0	0.0163	0.0000	0.4624
10	50.0		50.0	0.0	0.0000	0.0000	0.4624
11	50.0		50.0	0.0	0.0000	0.0000	0.4624
12	50.0		50.0	0	0.0000	0.0000	0.4624
13	57.1		50.0	-7.1	0.0000	0.0000	60.5284
14	61.5		44.4	-17.1	0.0000	0.0000	148.3524
15	66.6		25.0	-41.6	0.0000	0.0000	298.5984
16	66.6		25.0	-41.6	0.0000	0.0000	298.5984
17	75.0		25.0	-50.0	0.0000	0.0000	659.4624
18	75.0		16.6	-58.4	0.0000	0.0000	659.4624
Total	-	-	-	-	-	67.4157	4997.4312

De esta forma, el valor del estadístico de prueba queda dado por

$$W=\frac{\left(67.4157\right)^2}{4997.4312}\therefore W=0.9094$$

Según la Tabla A.9 del apéndice, el valor crítico sería $W_{\alpha,n}=W_{0.05,18}=0.897$.

Así, como $W=0.9094>W_{0.05,18}=0.897$, no rechaza la hipótesis nula, y se concluye que el porcentaje de remoción de DQO del reactor UASB en cuestión tiene un buen ajuste a la distribución normal, con una confiabilidad del 95%.

La salida de resultados en la consola de R para este caso se muestra a continuación

```
> shapiro.test(REM.DQO)

        Shapiro-Wilk normality test

data:  REM.DQO
W = 0.9091, p-value = 0.08296
```

De esta salida de resultados, obsérvese, que el p-valor = 0.0829 es mayor que el nivel de significancia elegido, evidencia suficiente para no rechazar la hipótesis nula y concluir que los datos de porcentaje de remoción de DQO de reactor UASB, se encuentran normalmente distribuidos.

6.5. Prueba de independencia: Tablas de contingencia *r* x *c*

Muchas veces surge la necesidad de determinar si existe alguna relación entre dos rasgos diferentes en los que una población ha sido clasificada y en los que cada rasgo se encuentra subdividido en cierto número de categorías (Conavos, 1988), es decir, se busca establecer un procedimiento de prueba para valorar si existe relación significativa entre dos variables cualitativas o categóricas, decidir si existe o no asociación entre dos variables categóricas, o en términos más simples si estas variables son independientes entre sí. El estudio de la independencia entre las dos variables estudiadas se realiza a través de una tabla de contingencia de *r* renglones (filas) por *c* columnas, dando lugar a *r* x *c* celdas o categorías. Así, el análisis de los datos se basa en comparar el número de observaciones que caen dentro de cada categoría (frecuencias marginales u observadas) y las que se esperaría obtener en cada celda bajo el supuesto de que la hipótesis nula es verdadera.

En virtud de lo anterior, el procedimiento de prueba chi-cuadrado, presentado en secciones anteriores, resulta ser adecuadamente utilizado para este tipo de situaciones, donde se busca probar la hipótesis nula de independencia de dos variables de clasificación (Walpole *et al.*, 2007). La forma general de una tabla de contingencia de *r* x *c*, se muestra a continuación

	Categorías	Característica B				Totales
		1	2	…	c	
Característica A	1	O_{11}	O_{12}	…	O_{1c}	$O_{1\cdot}$
	2	O_{21}	O_{22}	…	O_{2c}	$O_{2\cdot}$
	⋮	⋮	⋮	⋮	⋮	⋮
	r	O_{r1}	O_{r2}	…	O_{rc}	$O_{r\cdot}$
	Total	$O_{\cdot 1}$	$O_{\cdot 2}$	…	$O_{\cdot c}$	n

Para deducir un estadístico para la prueba de independencia, denotemos por *A* y *B* las dos variables (características) en las que centramos nuestro estudio. Así, la hipótesis nula de no asociación se establece como:

H_0: A y B son independientes

La hipótesis alterna es que hay una asociación entre *A* y *B* o que *A* y *B* no son independientes.

Para este tipo de situaciones el valor del estadístico de prueba chi-cuadrado se define por la siguiente expresión

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

Donde cada una de las frecuencias esperadas se calcula de la siguiente forma

$$e_{ij} = \frac{(total\ de\ la\ columna)\,\mathrm{x}\,(total\ del\ renglón)}{gran\ total}$$

Las frecuencias esperadas para cada celda, se acostumbra a registrarse entre paréntesis a un lado del valor observado real (Walpole *et al.*, 2007), tomando la tabla de contingencia la siguiente forma.

Característica A	Categorías	Característica B 1	2	…	c	Totales
	1	$o_{11}(e_{11})$	$o_{12}(e_{12})$	…	$o_{1c}(e_{1c})$	$o_{1\cdot}(e_{1\cdot})$
	2	$o_{21}(e_{21})$	$o_{22}(e_{22})$	…	$o_{2c}(e_{2c})$	$o_{2\cdot}(e_{2\cdot})$
	⋮	⋮	⋮	⋮	⋮	⋮
	r	$o_{r1}(e_{r1})$	$o_{r2}(e_{r2})$	…	$o_{rc}(e_{rc})$	$o_{r\cdot}(e_{r\cdot})$
	Total	$o_{\cdot 1}(e_{\cdot 1})$	$o_{\cdot 2}(e_{\cdot 2})$	…	$o_{\cdot c}(e_{\cdot c})$	n

Nótese que la suma de las frecuencias esperadas en cualquier renglón o columna da el total marginal apropiado. Para este procedimiento el rechazo de H_0 a un nivel de significancia α, se da cuando $\chi_0^2 > \chi^2_{\alpha[(r-1)(c-1)]}$.

Ejemplo 6.4. Supóngase que se realizó un estudio para determinar si el aumento de los cuadros clínicos de cáncer está asociado a la escala de explotación del carbón. Para ello se seleccionó una muestra aleatoria 560 personas en diferentes de regiones de explotación de carbón a diferentes escalas, para evaluar si padecían o no de cáncer. En una región donde se desarrolla minería del carbón a cielo abierto a pequeña escala 86 padecían

la enfermedad; en una región donde la explotación del carbón se realiza a mediana escala, se encontró que 126 personas se encontraban enfermas de cáncer, y en una zona donde la explotación de este mineral se desarrolla a gran escala, 175 personas padecían la enfermedad. Con base en lo anterior, pruebe si efectivamente el padecimiento de cáncer en zonas de explotación mineras, está asociado a la escala de explotación de la misma, a un nivel de significancia de 0.05.

Solución

A partir de la información muestral, el interés en este caso se centra en probar la siguiente hipótesis

H_0: El padecimiento de cáncer en zonas de explotación del carbón es independiente a la escala de explotación.

H_1: El padecimiento de cáncer en zonas de explotación del carbón no es independiente a la escala de explotación.

Para iniciar con la ejecución del procedimiento de prueba, tabularemos los datos en una tabla de contingencia de 2x3, como se muestra a continuación

		Escala de explotación de carbón			Total
		Pequeña	Mediana	Grande	
Padecimiento de cáncer	Si	86 (103.7)	126 (124.4)	175 (158.9)	387
	No	64 (46.3)	54 (55.6)	55 (71.1)	173
	Total	150	180	230	560

Luego de construida la tabla de contingencia, procedemos con realizar el cálculo de las frecuencias esperadas, las cuales se colocarán entre paréntesis en la tabla antes construida, al lado de cada una de las frecuencias esperadas.

$$e_{11}=\frac{(387)(150)}{560}=103.7 \qquad e_{12}=\frac{(387)(180)}{560}=124.4 \qquad e_{13}=\frac{(387)(230)}{560}=158.9$$

$$e_{21}=\frac{(173)(150)}{560}=46.3 \qquad e_{22}=\frac{(173)(180)}{560}=55.6 \qquad e_{23}=\frac{(173)(230)}{560}=71.1$$

De esta forma, el valor del estadístico de prueba χ^2 es

$$\chi^2 = \frac{(86-103.7)^2}{103.7} + \frac{(126-124.4)^2}{124.4} + \frac{(175-158.9)^2}{158.9} + \frac{(64-46.3)^2}{46.3} + \frac{(54-55.6)^2}{55.6} + \frac{(55-71.1)^2}{71.1}$$

$$\chi^2 = 15.1$$

Usando la Tabla A.5 del apéndice, encontramos que

$$\chi^2_{\alpha,[(r-1)(c-1)]} = \chi^2_{0.05,2} = 5.991$$

Por lo tanto, como $\chi^2 = 15.1 > \chi^2_{0.05,2} = 5.991$, se rechaza H_0 a favor de H_1, a un nivel de significancia de 0.05, es decir, los resultados son concluyentes en que el padecimiento de cáncer no es independiente a la escala de explotación de carbón con una confiabilidad del 95%, o sea, que el padecimiento de cáncer en zonas de explotación de este mineral está asociado con la escala de explotación del mismo, siendo mayor a medida que la escala de explotación sea más grande.

Para la ejecución de esta prueba en R, es necesario realizar previamente el ordenamiento de las frecuencias observadas en una matriz de datos y luego aplicar la función ***chisq.test***, como se observa en la siguiente salida de resultados de la consola de R.

```
> Cancer<-
matrix(c(86,126,175,64,54,55),ncol=3,nrow=2,byrow=TRUE)
> Cancer
     [,1] [,2] [,3]
[1,]   86  126  175
[2,]   64   54   55
> chisq.test(Cancer,correct=FALSE)

        Pearson's Chi-squared test

data:  Cancer
X-squared = 15.0554, df = 2, p-value = 0.000538
```

Nótese, que al ser el p-valor = 0.0005, mucho menor que el nivel de significancia seleccionado, hay evidencia para rechazar la hipótesis nula a favor de la alternativa.

6.6. Prueba de homogeneidad

En la sección anterior, al probar la independencia de dos variables cualitativas, se seleccionó una muestra aleatoria *n*, y los totales de cada

renglón y columna para la tabla de contingencia construida se determinaron al azar. Otro tipo de problema para el que se aplica el test chi-cuadrado, es aquel donde se predeterminan los totales de los renglones y las columnas, con el objetivo de probar la hipótesis de que las frecuencias dentro de cada renglón son las mismas, es decir, determinar si cada una de las categorías de una variable son *homogéneas* respecto a las categorías que asume la otra variable estudiada. Este tipo de contrastes son denominados, **pruebas de homogeneidad para variables cualitativas**.

El procedimiento de prueba para este tipo de contrastes es igual al descrito en la sección anterior, por lo que no se discutirá nuevamente, sino que se ilustrará con el siguiente ejemplo, tomado de Milton (2004)

Ejemplo 6.5. Un gran número de personas que viven en una sección determinada de una comunidad han estado expuestas durante los últimos diez años a la radiactividad procedente de un vertedero en el que se almacenan deshechos atómicos. Se realiza una investigación para descubrir si hay asociación aparente entre la exposición y el desarrollo de una cierta enfermedad en la sangre. Para llevar a cabo el experimento se eligen muestras aleatorias de 300 personas de la comunidad que han estado expuestas al peligro y 320 no expuestas. Se estudia cada sujeto para determinar si tiene la enfermedad. El experimento genera una tabla de contingencia de 2x2 que se presentan a continuación, donde se muestran las frecuencias observadas y esperadas para cada categoría

		Tiene la enfermedad		Total
		Si	No	
Expuesto a la radiactividad	Si	52 (48.39)	248 (251.61)	300
	No	48 (51.61)	272 (268.39)	320
	Total	100	520	620

A un nivel de significancia de 0.05, evaluar si la proporción de personas con la enfermedad que se encuentran expuestos la radiactividad es igual a la proporción de personas con la enfermedad entre los no expuestos.

Solución

Obviamente la hipótesis que se desea probar es la siguiente

H_0 : La proporción de personas con la enfermedad que se encuentran expuestos a radiactividad es igual a la proporción de personas con la enfermedad entre los no expuestos.

H_1: La proporción de personas con la enfermedad que se encuentran expuestos a radiactividad es diferente a la proporción de personas con la enfermedad entre los no expuestos.

Con base en la tabla de contingencia suministrada, el valor del estadístico de prueba es

$$\chi^2 = \frac{(52-48.39)^2}{48.39} + \frac{(248-251.61)^2}{251.61} + \frac{(48-51.61)^2}{51.61} + \frac{(272-268.39)^2}{268.39}$$

$$\chi^2 = 0.622$$

De acuerdo a la Tabla A.5 del apéndice tenemos que $\chi^2_{\alpha,[(r-1)(c-1)]} = \chi^2_{0.05,1} = 3.841$

De esta forma, como $\chi^2 = 0.622 < \chi^2_{0.05,1} = 3.841$, no se rechaza la hipótesis nula a un nivel de significancia de 0.05, es decir, que no hay evidencia de asociación entre la enfermedad sanguínea y la exposición a la radiactividad a un nivel de significancia del 95%, dado que la proporción de personas con la enfermedad expuestos a radiactividad no difiere significativamente a la proporción de personas con la enfermedad entre los no expuestos.

La salida de resultados de R, para este ejemplo sería la siguiente

```
> chisq.test(Enfermedad,correct=FALSE)

        Pearson's Chi-squared test

data:  Enfermedad
X-squared = 0.6232, df = 1, p-value = 0.4299
```

Note que no se rechaza H_0, dado que el p-valor es menor que el nivel de significancia elegido.

7.

Análisis de varianza (ANOVA)

7.1. Generalidades

En capítulos anteriores se discutieron diferentes procedimiento estadísticos para estudiar uno o dos parámetros poblacionales, como por ejemplo, la prueba *t* para comparar dos medias poblacionales. Sin embargo, en muchas situaciones experimentales que se presentan con frecuencia se disponen de más de dos poblaciones, por lo que es necesario contar con herramientas estadísticas que permitan realizar este tipo de inferencias. El **Análisis de varianza (ANOVA)**, permite generalizar el contraste de igualdad de medias de dos a *k* poblaciones; este procedimiento estadístico se puede utilizar en las situaciones en las que nos interesa analizar una respuesta cuantitativa, llamada habitualmente variable dependiente, medida bajo ciertas condiciones experimentales identificadas por una o más variables categóricas (por ejemplo tratamiento, sexo), llamadas variables independientes.

El ANOVA, es una de las técnicas de tratamiento estadístico de datos de más amplio uso en el campo de la investigación científica, pues es común que los objetivos de diferentes investigaciones estén orientados a evaluar el comportamiento de las unidades experimentales, bajo la influencia de diferentes tratamientos o grupos de clasificación, por ejemplo, se puede tener interés en la dinámica de las concentraciones de ciertas variables fisicoquímicas espacialmente, es decir, conocer si estas presentan la misma concentración medias en diferentes estaciones de muestreo, o determinar si la aplicación de diferentes tipos de biofertilizantes sobre un cultivo específico, muestra diferencias significativas en la producción del mismo, etc.

7.2. Análisis de varianza (ANOVA) de un factor: Diseño completamente al azar

Cuando solo hay una sola variable que proporciona condiciones experimentales distintas, el análisis recibe el nombre de ANOVA de un factor. En este tipo de situaciones de *k* poblaciones se seleccionan muestras aleatorias de tamaño *n* y se clasifican con base en un criterio único, como tratamientos o grupos diferenciales. En la actualidad, el término

tratamiento se utiliza, por lo general, para designar las diferentes clasificaciones (Walpole *et al.*, 2007), por ejemplo, estaciones de muestreo, tipo de biofertilizante utilizado, regiones, etc.

7.2.1. Diagnosis e hipótesis del modelo del ANOVA de un factor

Para la aplicación del ANOVA de un factor, es necesario que se evalúen las hipótesis previas relativas a la calidad de la muestra (aleatoriedad e independencia), a la estructura de probabilidad normal o no de la población y así las distintas poblaciones tienen varianzas iguales o distintas, esta última propiedad conocida como *homocedasticidad* (Conavos, 1988; Walpole *et al.*, 2007; Devore, 2008; Guisande *et al.*, 2011).

El supuesto de normalidad de las muestras, como se discutió en capítulos anteriores, se evalúa a través de los test de **Shapiro-Wilk** ($n \leq 30$) o el test de **Kolmogorov-Smirnov** ($n > 30$). A continuación, si la muestra no está contaminada y no hay desviaciones importantes de normalidad, se comprobará la hipótesis de homocedasticidad a través del ***test de homogeneidad de varianzas de Bartlett,*** el ***test de Levene*** o el ***test de Cochran*** (Arriaza *et al.*, 2008; Hoff, 2005; Walpole *et al.*, 2007). A la vista del cumplimiento de ambos supuestos se procede a realizar el procedimiento de prueba del ANOVA o de lo contrario, se decide entre transformar los datos para dar cumplimientos a los supuestos del modelo, o buscar una prueba estadística alternativa de distribución libre (no paramétrica), que se discutirán en capítulos posteriores.

Dado que el ANOVA busca generalizar el contraste de igualdad de medias de dos a *k* poblaciones, el objetivo del procedimiento de prueba que se presentará en esta sección se encuentra orientado a probar las hipótesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

H_1 : Al menos dos de las medias no son iguales

7.2.2. Procedimiento de prueba del ANOVA de un factor

Antes de emprender el desarrollo matemático de los datos, se aconseja realizar un acomodo de los datos a través de un formato tabular como el que se muestra en la Tabla 7.1, donde $\bar{y}_{ij}$ representa la *j*-ésima observación del *i*-ésimo tratamiento, $T_{i.}$ es el total de todas las observaciones en la

muestra correspondiente al *i*-ésimo tratamiento, $\bar{y}_{i\cdot}$ es la media de todas las observaciones en la muestra en el *i*-ésimo tratamiento, $T_{\cdot\cdot}$ es el total de las *nk* observaciones (Gran total) y $\bar{y}_{\cdot\cdot}$ la media de todas las *nk* observaciones (Gran media).

Tabla 7.1. Arreglo de la *k* muestras aleatorias para los cálculos del ANOVA.

	Tratamiento						
	1	**2**	**…**	**i**	**…**	**k**	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	.	.		.		.	
	.	.		.		.	
	.	.		.		.	
Total	$T_{1\cdot}$	$T_{2\cdot}$	$\cdots$	$T_{i\cdot}$	$\cdots$	$T_{k\cdot}$	$T_{\cdot\cdot}$
Media	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	$\cdots$	$\bar{y}_{i\cdot}$	$\cdots$	$\bar{y}_{k\cdot}$	$\bar{y}_{\cdot\cdot}$

Desde el comienzo de este capítulo, se ha hecho hincapié en que la técnica del análisis de varianza tiene como objetivo realizar un análisis de la variación de las medias de *k* poblaciones. Sin embargo, esta variación entre las *k* medias se logra mediante la partición de la variación total de las observaciones en dos componentes; una componente que expresa la variación de las unidades experimentales debido al efecto que ejerce cada uno de los tratamientos, y otra componentes que mide la variación de las observaciones debida a la aleatoriedad, es decir, al error experimental o intravarianza. La partición de la variabilidad total se encuentra especificada por la siguiente expresión

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{\cdot\cdot}\right)^2 = n\sum_{i=1}^{k}\left(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2$$

Cuya notación para cada uno de los términos de esta identidad es la siguiente

$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{\cdot\cdot}\right)^2 =$ Suma total de cuadrados,

$SSA = n\sum_{i=1}^{k}\left(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2 =$ Suma de cuadrados de los tratamientos,

$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2 =$ Suma de los cuadrados de los errores.

Así, la identidad puede expresarse más fácilmente como

$$SST = SSA + SSE$$

No obstante, es fácil observar que las expresiones anteriores involucran ciertos cálculos bastante tediosos, por lo que se han establecido expresiones algebraicamente equivalentes y de cálculos mucho más convenientes para determinar cada uno de los términos de la identidad (Milton, 2004)

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{T_{..}^2}{nk}$$

$$SSA = n\sum_{i=1}^{k}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = \frac{\sum_{i=1}^{k} T_{i.}^2}{n} - \frac{T_{..}^2}{nk}$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2 = SST - SSA$$

Durante el análisis también son de interés el cálculo de dos cantidades conocidas como media cuadrática del tratamiento (s_1^2) y el error cuadrático medio (s^2), calculados a través de las siguientes expresiones

$$s_1^2 = \frac{SSA}{k-1}$$

$$s^2 = \frac{SSE}{k(n-1)}$$

Donde los denominadores de cada una de estas medidas representan los grados de libertad, con base en los cuales se realizan las estimaciones de la variabilidad.

El estadístico de prueba, a través del cual se decide rechazar o no la hipótesis nula se basa en la distribución *F* de Fisher-Snedecor con $k - 1$ y $k(n - 1)$ grados de libertad, definido por la razón

$$f = \frac{s_1^2}{s^2}$$

Así, la hipótesis nula H_0 se rechaza con un nivel de significancia α cuando

$$f > f_{\alpha[k-1,k(n-1)]}$$

Comúnmente, los cálculos en un problema de análisis de varianza se resumen en forma tabular, tal como se muestra a continuación en la Tabla 7.2.

Tabla 7.2. Resumen tabular del análisis de varianza de un factor.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*f* calculada
Tratamientos	*SSA*	$k-1$	$s_1^2 = \frac{SSA}{k-1}$	$f = \frac{s_1^2}{s^2}$
Error	*SSE*	$k(n-1)$	$s^2 = \frac{SSE}{k(n-1)}$	
Total	*SST*	$nk-1$	-	-

Las expresiones presentadas anteriormente para el cálculo de las componentes de la variabilidad de las observaciones, son utilizadas bajo el supuesto de tamaño igual de las muestras. Sin embargo, puede ocurrir el caso en que se tenga diferentes número de observaciones para cada uno de los tratamientos, es decir, se tengan k muestras aleatorias de tamaños $n_1, n_2, ..., n_k$, respectivamente. En este tipo de situaciones los cálculos de las medidas de variabilidad presentadas antes, sufren ligeras modificaciones para ajustarse al tamaño desigual de las muestras, tomando las siguientes formas

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{T_{..}^2}{N}$$

$$SSA = \sum_{i=1}^{k} n_i\left(\bar{y}_{i\cdot} - \bar{y}_{..}\right)^2 = \frac{\sum_{i=1}^{k} T_{i\cdot}^2}{n_i} - \frac{T_{..}^2}{N}$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2 = SST - SSA$$

Donde $N = \sum_{i=1}^{k} n_i$. Luego, la partición de los grados de libertad similar al mostrado anteriormente: $N-1$ para *SST*, $k-1$ para *SSA*, y $N-1-(k-1)$ para *SSE* (Walpole *et al.*, 2007).

Cabe anotar que este estadístico de prueba se aplica cuando se da cumplimiento a todos los requerimientos exigidos por el modelo del ANOVA, es decir, cuando se verifiquen todas las hipótesis exigidas, la alternativa preferida será el test *F*. Cuando se dé la normalidad pero no la homocedasticidad, se recomienda el uso del *test de Kruskal Wallis*, que trataremos en capítulos posteriores. Si falla, aunque no de forma drástica la normalidad (con valores de *p* entre 0,01 y 0,05), la robustez del test *F* le hace seguir siendo una buena opción. Por último, si fallara fuertemente la normalidad, se recomienda el uso del test de *Kruskal Wallis* (Arriaza *et al.*, 2008). Otro procedimiento alternativo a transgresiones de los supuestos del modelo, sería cambiar la escala de medición de las observaciones a través de transformaciones que se discutirán en secciones siguientes.

En R, el modelado del análisis de varianza resulta ser muy simple, pues existe una función llamada ***aov***, que permite realizar este tipo de procedimientos y muestra en su salida de resultados la tabla resumen del análisis, similar a la presentada en la Tabla 7.2. En los argumentos de esta función solo se especificar la fórmula del modelo, es decir, cuál es nuestra variable dependiente y cuál es nuestro grupo diferencial o factor (tratamiento), separados por el símbolo "~".

```
aov(y ~ factor)
```

7.3. Pruebas sobre homogeneidad de diversas varianzas (homocedasticidad)

Anteriormente, se comentó que el análisis de varianza (ANOVA) es la técnica más poderosa para probar hipótesis acerca de la variación media de *k* poblaciones, o evaluar la influencia de *k* tratamientos o grupos diferenciales en un conjunto de unidades experimentales, siempre que se cumplan los supuestos de normalidad, homogeneidad de varianza e independencia. La falta de cualquiera de estos supuestos deterioraría la utilidad de la prueba y conllevaría a formular conclusiones erróneas y no válidas (Vorapongsathorn *et al.*, 2004). Por lo tanto, es necesario probar estas hipótesis antes de usar el análisis de la varianza.

La literatura actual recomienda el uso de varios procedimientos estadísticos para probar la hipótesis de homogeneidad de varianza. Entre estas pruebas, el test de Bartlett, test de Levene y el test de Cochran son ampliamente utilizados para comprobar este supuesto, (Box *et al.*, 2008;

Montgomery, 2001; Kuehl, 2001; Gutiérrez & de la Vara, 2008; Walpole *et al.*, 2007). En esta sección, discutiremos el procedimiento de prueba de cada uno de estos test y las situaciones en las que estos pueden ser utilizados.

7.3.1. Test de Bartlett

El test de Bartlett, es el test de más amplio uso para probar la homogeneidad de las varianzas de *k* poblaciones siempre que se tenga previa certeza del cumplimiento del supuesto de normalidad de las observaciones (Bartlett, 1937). Este test, involucra el cálculo de un estadístico cuya distribución de muestreo está estrechamente aproximada a la distribución de chi-cuadrado con $k-1$ grados de libertad, cuando las *k* muestras aleatorias son de poblaciones normales e independientes (Montgomery, 2001). Obviamente, la hipótesis que se desea probar es

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

H_1 : Al menos dos varianzas son diferentes.

El estadístico de prueba para este test cuando se tienen *k* muestras aleatorias de tamaño $n_1, n_2, ..., n_k$, está definido por la siguiente expresión (Montgomery, 2001; Gutiérrez & de la Vara, 2008)

$$\chi^2 = 2.3026 \frac{q}{c}$$

donde

$$q = \left(N-k\right)\log s_p^2 - \sum_{i=1}^{k}\left(n_i - 1\right)\log s_i^2$$

y

$$c = 1 + \frac{1}{3\left(k-1\right)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right)$$

con

$$s_p^2 = \frac{\sum_{i=1}^{k}\left(n_i - 1\right)s_i^2}{N-k}$$

Recuérdese que el término s_p^2, es una estimación de la varianza común.

Para este procedimiento de prueba, se rechaza la hipótesis nula de igualdad de las varianzas cuando $\chi^2 > \chi^2_{\alpha,k-1}$.

En R, existe una función incorporada en su paquete básico que permite realizar la aplicación del test de Bartlett de una forma fácil y rápida. La orden al software se realiza a través del código ***bartlett.test***, en cuyo argumento se indica la fórmula de cálculo, especificando la variable dependiente y el factor, a través de la siguiente estructura

```
bartlett.test(x, factor)
```

Ejemplo 7.1. Durante los años 2004 a 2005 el grupo de investigación Pichihuel de la Universidad de La Guajira, realizó un estudio sobre la dinámica fisicoquímica del ecosistema estuarino el Riito. Los datos referentes a las concentraciones de DBO (mg/L), desde noviembre de 2004 a septiembre de 2005, en cuatro estaciones de muestreo se muestran a continuación

	Estaciones			
Meses	**E1**	**E2**	**E3**	**E4**
Nov	5.65	1.48	1.64	2.38
Dic	2.40	2.61	2.71	2.70
Ene	1.21	2.00	2.31	1.95
Feb	2.45	1.84	0.35	1.52
Mar	1.25	3.71	1.92	1.08
Abr	1.90	1.75	2.28	2.11
May	0.27	0.44	0.11	0.20
Jun	1.20	1.80	0.80	0.78
Jul	1.25	1.42	1.10	1.17
Ago	1.25	1.47	1.10	1.17
Sep	3.17	4.01	2.15	1.50

A partir de estos datos evaluar si los valores de DBO presentan homogeneidad de varianza en las cuatro estaciones de muestreo. Asúmase que los datos provienen de poblaciones normales.

Solución

Con $N = 44$ y las varianzas de los datos de DBO en cada una de las estaciones de muestreo iguales a $s_{E1}^2 = 2.10$, $s_{E2}^2 = 1.08$, $s_{E3}^2 = 0.74$ y $s_{E4}^2 = 0.54$, respectivamente. Así, tenemos que

$$s_p^2 = \frac{10\left(2.10 + 1.08 + 0.74 + 0.54\right)}{40} \therefore s_p^2 = 1.115$$

de esta manera

$$q = \left(40\right)\log\left(1.115\right) - 10\left[\log(2.10) + \log(1.08) + \log(0.74) + \log(0.54)\right]$$

$$q = 2.318$$

y

$$c = 1 + \frac{1}{3(3)}\left(\frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} - \frac{1}{40}\right)$$

$$c = 1.042$$

Así, el valor del estadístico de prueba es

$$\chi^2 = 2.3026\frac{2.318}{1.042} \therefore \chi^2 = 5.122$$

El valor crítico para la distribución chi-cuadrado, según la Tabla A.5 sería $\chi_{0.05,3}^2 = 7.815$.

Como $\chi^2 = 5.122 < \chi_{0.05,3}^2 = 7.815$, no se rechaza la hipótesis nula de igualdad de las varianzas a un nivel de significancia de 0.05, concluyéndose que las concentraciones de DBO en las diferentes estaciones de muestreo, presenta varianzas homogéneas, con una confiabilidad del 95%.

A continuación, mostraremos la salida de resultados en R para la aplicación del test de Bartlett a través de dos metodologías diferentes, la primera de ellas con la creación de un vector de datos para la variable respuesta y un vector de datos para los diferentes niveles del factor, y otro método, y el que más se aconseja por comodidad de trabajo y porque generalmente en el desarrollo de nuestros experimentos los datos se encuentran tabulados en una hoja de cálculo, que guardaremos bajo la extensión *.csv*, y luego cargaremos en la consola de R siguiendo las instrucciones de la función ***read.csv2***, discutida en secciones anteriores. Enseguida veremos y discutiremos la primera metodología

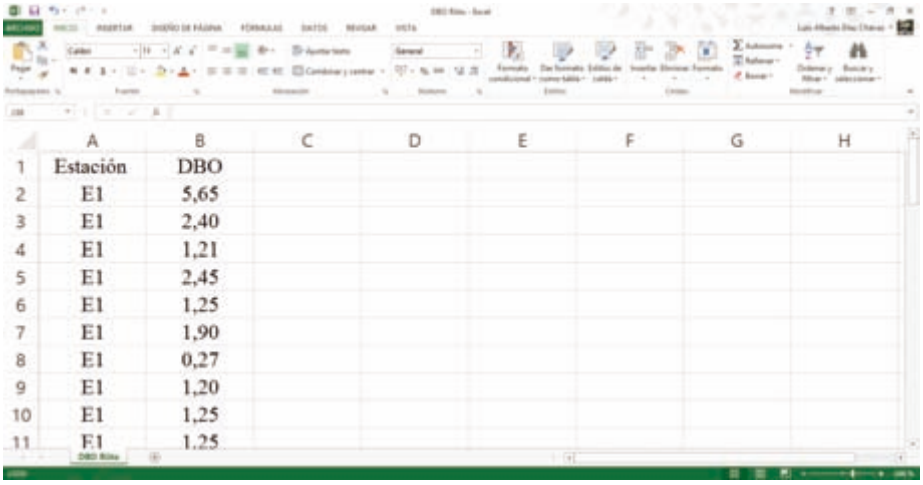
```
> DBO.E1<-
c(5.65,2.40,1.21,2.45,1.25,1.90,0.27,1.20,1.25,1.25,3.17)
> DBO.E2<-
c(1.48,2.61,2.00,1.84,3.71,1.75,0.44,1.80,1.42,1.47,4.01)
> DBO.E3<-
c(1.64,2.71,2.31,0.35,1.92,2.28,0.11,0.80,1.10,1.10,2.15)
> DBO.E4<-
c(2.38,2.70,1.95,1.52,1.08,2.11,0.20,0.78,1.17,1.17,1.50)
> DBO<-c(DBO.E1,DBO.E2,DBO.E3,DBO.E4)
> Estación<-rep(1:4,each=11)
> Estación<-factor(Estación,labels=c("E1","E2","E3","E4"))
> bartlett.test(DBO,Estación)

        Bartlett test of homogeneity of variances

data:  DBO and Estación
Bartlett's K-squared = 5.1317, df = 3, p-value = 0.1624
```

Nótese que el p-valor = 0.1624, al ser mayor que el nivel de significancia seleccionado, brinda evidencia suficiente para no rechazar la hipótesis nula y concluir que las concentraciones de DBO posee varianzas homogéneas a lo largo de las cuatro estaciones de muestro.

Para la ejecución de la segunda metodología, como se mencionó antes, se debe tabular los datos en una hoja de cálculo de Excel y se guardan bajo la extensión *.csv*, siguiendo la estructura que se muestra en la Figura 7.1



	A	B	C	D	E	F	G	H
1	Estación	DBO						
2	E1	5,65						
3	E1	2,40						
4	E1	1,21						
5	E1	2,45						
6	E1	1,25						
7	E1	1,90						
8	E1	0,27						
9	E1	1,20						
10	E1	1,25						
11	E1	1,25						

Figura 7.1. Tabulación de los datos en Excel para importarlos a R.

En adelante se sigue el siguiente procedimiento en R

```
> Datos<-read.csv2("DBO
Riito.csv",header=TRUE,encoding="latin1")
> attach(Datos)
> bartlett.test(DBO,Estación)

        Bartlett test of homogeneity of variances

data:  DBO and Estación
Bartlett's K-squared = 5.1317, df = 3, p-value = 0.1624
```

7.3.2. Test de Levene

La prueba estándar para verificar la homogeneidad de varianzas es el test de Bartlett, el cual es una efectiva herramienta sólo si la población se encuentra distribuida aproximadamente normal, o hay cumplimiento del supuesto de normalidad. Cuando el supuesto de normalidad es violado fuertemente, un procedimiento que es relativamente insensible (robusto) al incumplimiento de este supuesto es el test de Levene (De la Huerta, 2012).

Al igual que el test de Bartlett, el test de Levene es aplicable aún cuando el tamaño muestral de los *k* grupos o tratamientos es desigual. De esta forma, el estadístico de prueba de este test está dado por la siguiente expresión (Levene, 1960)

$$W_0 = \frac{(N-k)\sum_{i=1}^{k} n_i \left(\overline{z}_{i\cdot} - \overline{z}_{\cdot\cdot}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N}\left(z_{ij} - \overline{z}_{i\cdot}\right)^2}$$

donde

$$\overline{z}_{i\cdot} = \sum_{j=1}^{n_i} z_{ij}/n_i \quad \text{y} \quad \overline{z}_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}/N \,,$$

con

$$z_{ij} = \left|x_{ij} - \overline{x}_i\right| \quad \text{y} \quad N = \sum_{i=1}^{k} n_i \,.$$

La expresión anterior para el cálculo del estadístico de prueba del test de Levene, fue replanteado, para aumentar su robustez cuando las poblaciones involucradas presentaban sesgo y por lo tanto no normalidad.

De esta forma se consideró la mediana y la media recortada al 10% como mejores estimadores del centro de los datos, sustituyendo a la media en el cálculo de las desviaciones absolutas (z_{ij}) (Brown & Forsythe, 1974). Esta media recortada al 10%, es la media de las observaciones después de haber eliminado los valores del 10% superior y 10 inferior en cada grupo. De esta forma la mediana, también puede ser considerada como la media recortada al 50%, y la media, como una media recortada al 0%. La elección del porcentaje de recorte es arbitrario.

En virtud de los anterior, cuando se estima el estadístico de prueba del test de Levene al remplazar la media $\bar{x}_i$ por la mediana $\tilde{x}_i$ para formar las deviaciones absolutas z_{ij}, el estadístico de Levene, se define como W_{50}. Cuando el remplazo de $\bar{x}_i$, se realiza por x'_i, donde x'_i es la media recortada al 10% en el *i*-enésimo grupo, el estadístico se define como W_{10} (Vorapongsathorn *et al.*, 2004; Brown & Forsythe, 1974).

En todos los casos anteriores la distribución muestral del estadístico de Levene se aproxima a una distribución *F* de Fisher-Snedecor con $k-1$ y $N-k$ grados de libertad. Por lo que el rechazo de la hipótesis nula de igualdad de las varianzas en todos los grupos $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, se rechaza cuando $W > f_{0.05[(k-1),(N-k)]}$ (Brown & Forsythe, 1974; Conover *et* al., 1981; Kuehl, 2001; Vorapongsathorn *et al.*, 2004).

La aplicación del test de Levene en R se puede realizar a través de dos rutas de análisis: la primera a través de la función ***levene.test*** del paquete "*lawstat*" (Gastwirth *et al.*, 2015), siguiendo la siguiente sintaxis

```
levene.test(y, group, location=c("median", "mean", "trim.mean"),
trim.alpha=0.25)
```

Donde ***y***, especifica la variable independiente, ***group***, indica el factor del modelo (grupo diferencial), ***location***, permite especificar si el cálculo del estadístico de prueba se realiza a través del método establecido por Levene (1960) o por el modificado por Brown & Forsythe (1974), solo basta con indicar si se usa como medida central la media (mean), para el primer método, o la mediana (***median***) o media recortada (***trim.mean***), para el segundo método. En caso de usarse la media recortada para el cálculo de las desviaciones absolutas, en el argumento ***trim.alpha***, se debe establecer cuál será el porcentaje de recorte, que puede ser cualquier valor entre 0-50% (0-0.50).

La otra alternativa para la aplicación del test de Levene en R, es a través de la función leveneTest del paquete "car" (Fox *et al.*, 2015), cuya ejecución se realiza a través de la siguiente línea de comando

```
leveneTest(y, group, center=median, trim=0.1)
```

Donde los argumentos ***y***, ***group***, ***center*** y ***trim***, cumplen respectivamente el mismo propósito que los argumentos descritos para la función ***levene.test***.

Ejemplo 7.2. Durante los años 2004 a 2005 el grupo de investigación Pichihuel de la Universidad de La Guajira, realizó un estudio sobre la dinámica fisicoquímica del ecosistema estuarino el Riito. Los datos referentes a las concentraciones de NO_2 (mg/L), desde noviembre de 2004 a septiembre de 2005, en cuatro diferentes estaciones de muestreo se muestran a continuación

	Estaciones			
Meses	**E1**	**E2**	**E3**	**E4**
Nov	0.43	0.53	0.43	0.44
Dic	0.30	0.28	0.28	0.22
Ene	0.26	0.28	0.14	0.22
Feb	0.20	0.18	0.16	0.19
Mar	0.05	0.14	0.04	0.04
Abr	0.13	0.22	0.13	0.13
May	0.25	0.21	0.18	0.21
Jun	0.38	0.39	0.33	0.39
Jul	0.52	0.57	0.49	0.58
Ago	5.58	5.32	2.60	3.38
Sep	0.34	0.47	0.31	0.37

A partir de estos datos evaluar si los valores de NO_2 presentan homogeneidad de varianza en las cuatro estaciones de muestreo. Asúmase que supuesto de normalidad es violado fuertemente.

Solución

Con $N = 44$, las varianzas de las concentraciones de NO_2 en cada una de las estaciones de muestreo iguales a $s^2_{E1} = 2.57$, $s^2_{E2} = 2.29$, $s^2_{E3} = 0.52$ y $s^2_{E4} = 0.90$, respectivamente, y las medianas iguales a $\tilde{x}_{E1} = 0.30$, $\tilde{x}_{E2} = 0.28$, $\tilde{x}_{E3} = 0.28$ y $\tilde{x}_{E4} = 0.22$, respectivamente, tenemos que las desviaciones absolutas z_{ij}, son iguales a

	Estaciones			
Meses	**E1**	**E2**	**E3**	**E4**
Nov	0.13	0.25	0.15	0.22
Dic	0.00	0.00	0.00	0.00
Ene	0.04	0.00	0.14	0.00
Feb	0.10	0.10	0.12	0.03
Mar	0.25	0.14	0.24	0.18
Abr	0.17	0.06	0.15	0.09
May	0.05	0.07	0.10	0.01
Jun	0.08	0.11	0.05	0.17
Jul	0.22	0.29	0.21	0.36
Ago	5.28	5.04	2.32	3.16
Sep	0.04	0.19	0.03	0.15
$\overline{Z}_{i\cdot}$	0.578	0.568	0.319	0.397

$$\overline{z}_{\cdot\cdot} = \frac{0.13 + 0.00 + \ldots + 0.15}{44} \therefore \overline{z}_{\cdot\cdot} = 0.466$$

y

De esta forma, el valor del estadístico de prueba es

$$W_{50} = \frac{(40)(11)\left[(0.578 - 0.466)^2 + (0.568 - 0.466)^2 + (0.319 - 0.466)^2 + (0.397 - 0.466)^2\right]}{(3)\left[(0.13 - 0.466)^2 + (0.00 - 0.466)^2 + \ldots + (0.15 - 0.466)^2\right]}$$

$$W_{50} = 0.121$$

y el valor crítico según la Tabla A.6 es $f_{0.05[3,40]} = 2.84$.

Ahora, como $W_{50} = 0.121 < f_{0.05[3,40]} = 2.84$, no se rechaza a hipótesis nula de igualdad de varianzas a un nivel de significancia de 0.05, concluyendo que

las concentraciones de NO_2 en las cuatro estaciones de muestreo poseen varianzas homogéneas a un nivel de confiabilidad del 95%.

La aplicación del test de Levene en R, se puede realizar siguiendo las dos metodologías explicadas en la sección anterior. Sin embargo, optaremos por utilizar la primera de ellas, con el objetivo de que el lector se afiance más con el lenguaje de programación del R.

```
> NO2.E1<-
c(0.43,0.30,0.26,0.20,0.05,0.13,0.25,0.38,0.52,5.58,0.34)
> NO2.E2<-
c(0.53,0.28,0.28,0.18,0.14,0.22,0.21,0.39,0.57,5.32,0.47)
> NO2.E3<-
c(0.43,0.28,0.14,0.16,0.04,0.13,0.18,0.33,0.49,2.60,0.31)
> NO2.E4<-
c(0.44,0.22,0.22,0.19,0.04,0.13,0.21,0.39,0.58,3.38,0.37)
> NO2<-c(NO2.E1,NO2.E2,NO2.E3,NO2.E4)
> Estación<-rep(1:4,each=11)
> Estación<-factor(Estación,labels=c("E1","E2","E3","E4"))
> library(lawstat)
> levene.test(NO2,Estación,location="median")

        modified robust Brown-Forsythe Levene-type test based on
the absolute
        deviations from the median

data:  NO2
Test Statistic = 0.1217, p-value = 0.9468
```

Note que el p-valor = 0.9468 es mayor que el nivel de significancia seleccionado, evidencia suficiente para retener la hipótesis nula y aceptar la homogeneidad de varianza de las concentraciones de NO_2 en las cuatro estaciones de muestreo.

7.3.3. Test de Cochran

Otro test alternativo, para probar la homogeneidad de varianza de k poblaciones o grupos, cuando el supuesto de normalidad de los datos es violado es el test de Cochran (Vorapongsathorn et al., 2004), que proporciona un procedimiento de prueba con cálculos simples, pero que se restringe a situaciones en las cuales los tamaños muéstrales de los k grupos son iguales (Conover et al., 1981; Walpole & Myers, 2007). El estadístico de prueba para este test está dado por la siguiente expresión (Cochran, 1941).

$$C = \frac{\max s_i^2}{\sum_{i=1}^{k} s_i^2}$$

y la hipótesis de igualdad de las varianzas en los *k* grupos se rechaza cuando $C > C_{\alpha,n,k}$, donde el valor que se obtiene de la Tabla A.10 del apéndice.

El cálculo de este test en R se realiza a través de la función ***C.test*** del paquete "GAD" (Sandrini & Camargo, 2015), siguiendo la siguiente sintaxis de programación

```
C.test(lm(y ~ group))
```

En cuyos argumentos, ***lm*** específica la construcción de un modelo lineal con las variables a considerar en el análisis de varianza, esta función se tratará, con más detalles cuando abordemos el capítulo de análisis de regresión. Los argumentos ***y*** y ***group***, cumplen las mismas funciones descritas antes.

Ejemplo 7.3. Durante los años 2004 a 2005 el grupo de investigación Pichihuel de la Universidad de La Guajira, realizó un estudio sobre la dinámica fisicoquímica del ecosistema estuarino el Riito. Los datos referentes a las concentraciones de DQO (mg/L), desde noviembre de 2004 a septiembre de 2005, en cuatro diferentes estaciones de muestreo se muestran a continuación

	Estaciones			
Meses	**E1**	**E2**	**E3**	**E4**
Nov	42	55	40	48
Dic	190	199	148	139
Ene	58	61	55	97
Feb	136	147	133	135
Mar	91	321	515	516
Abr	136	147	133	130
May	119	91	62	112
Jun	123	118	110	237
Jul	128	145	160	363
Ago	44	159	287	133
Sep	91	83	77	78

A partir de estos datos evaluar si los valores de DQO presentan homogeneidad de varianza en las cuatro estaciones de muestreo. Asúmase que supuesto de normalidad es violado fuertemente.

Solución

Con $n = 11$ y las varianzas de las concentraciones de DQO en las cuatro estaciones de muestro iguales a $s^2_{E1} = 2048.62$, $s^2_{E2} = 5630.82$, $s^2_{E3} = 18802.85$ y $s^2_{E4} = 19738.42$, respectivamente, el valor del estadístico de Cochran es igual a

$$C = \frac{19738.42}{2048.62 + 5630.82 + 18802.85 + 19738.42}$$

$$C = 0.427$$

y el valor crítico para este test, según la Tabla A.10 es $C_{0.05,11,4} = 0.4884$.

De esta forma, como $C = 0.427 < C_{0.05,11,4} = 0.4884$, no se rechaza la hipótesis nula de igualdad de las varianzas a un nivel de significancia de 0.05, es decir, los resultados son concluyentes en que existe homogeneidad de varianzas entre las concentraciones de DQO en las cuatro estaciones de muestreo con una confiabilidad del 95%.

A continuación mostraremos la salida de resultados de R luego de la aplicación del test de Cochran. Nuevamente seguiremos la metodología donde se requiere construir un vector de datos y un factor, para luego aplicar la función que permitirá el cálculo de este test.

```
> DQO.E1<-c(42,190,58,136,91,136,119,123,128,44,91)
> DQO.E2<-c(55,199,61,147,321,147,91,118,145,159,83)
> DQO.E3<-c(40,148,55,133,515,133,62,110,160,287,77)
> DQO.E4<-c(48,139,97,135,516,130,112,237,363,133,78)
> DQO<-c(DQO.E1,DQO.E2,DQO.E3,DQO.E4)
> Estación<-rep(1:4,each=11)
> Estación<-factor(Estación,labels=c("E1","E2","E3","E4"))
> library(GAD)
> C.test(lm(DQO~Estación))

        Cochran test of homogeneity of variances

data:  lm(DQO ~ Estación)
C = 0.427, n = 11, k = 4, p-value = 0.1733
alternative hypothesis: Group E4 has outlying variance
sample estimates:
       E1        E2        E3        E4
 2048.618  5630.818 18802.855 19738.418
```

Nótese que el p-valor = 0.1733 es mayor que el nivel de significancia elegido, por lo tanto, se tiene evidencia suficiente para no rechazar la hipótesis nula y concluir que la concentraciones de DQO presentan homogeneidad de varianzas en las cuatro estaciones de muestreo.

A continuación, veremos un ejemplo de aplicación del análisis de varianza de un factor, donde la evaluación de los requisitos previos para aplicación de este procedimiento se obviará en los cálculos mecánicos y se partirá bajo ciertas asunciones. Sin embargo, cuando se realice la aplicación en R, nos ceñiremos al método y ejecutaremos todos los pasos del procedimiento, tal como ocurriría en la aplicación de este test en el tratamiento de nuestros datos, resultados de nuestros experimentos.

Ejemplo 7.4. La Universidad de La Guajira a través del Grupo de Investigación Pichihuel en asociación con el Instituto de Estudios Ambientales y Aprovechamiento de Agua – INESAG, llevaron a cabo monitoreos de algunas variables ambientales en las playas del municipio de Riohacha, en el marco del programa de Calidad Ambiental de Playas Turísticas – CAPT, en cuatro diferentes estaciones de monitoreo. Dentro de sus variables de estudio se encuentran el grupo de los enterococos fecales en agua, un grupo bacteriano de interés sanitario por ser indicador de contaminación de origen fecal y representar riesgo para los usuarios de las playas por ser precursor de algunas enfermedades gastrointestinales. A continuación se muestran las densidades de enterococos fecales en UFC/100 mL en las diferentes estaciones, obtenidas en una jornada de muestreo. Se desea determinar: (a) si existen diferencias significativas en la concentración media de enterococos en las diferentes estaciones de muestreo; (b) en caso de existir diferencias significativas, ¿en cuales estaciones de monitoreo se presentan dichas diferencias? Asúmase que los datos de densidad de enterococos se encuentra normalmente distribuida en cada una de las estaciones de monitoreo con varianzas homogéneas.

	Estaciones de monitoreo				
	RT_1	RT_2	RT_3	RT_4	
	1000	300	3300	1600	
	2100	100	2000	2500	
	700	1000	2200	3400	
	500	600	3200	300	
	100	600	2700	1000	
	200	700	2500	800	
Total	4600	3300	15900	9600	33400
Media	766.67	550	2650	1600	1391.67

Solución

A partir de lo planteado en el enunciado es evidente que se tiene interés en probar las siguientes hipótesis

$$H_0 : \mu_{RT_1} = \mu_{RT_2} = \mu_{RT_3} = \mu_{RT_4}$$

H_1 : Al menos dos medias son distintas

Bajo este sistema de hipótesis iniciamos el análisis con determinar las sumas de cuadrados

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{T_{..}^2}{nk}$$

$$SST = 1000^2 + 2100^2 + 700^2 + ... + 800^2 - \frac{33400^2}{24}$$

$$SST = 27638333$$

$$SSA = \frac{\sum_{i=1}^{k} T_{i.}^2}{n} - \frac{T_{..}^2}{nk}$$

$$SSA = \frac{4600^2 + 3300^2 + 15900^2 + 9600^2}{6} - \frac{33400^2}{24}$$

$$SSA = 16355000$$

$$SSE = SST - SSA$$

$$SSE = 27638333 - 16355000$$

$$SSE = 11283333$$

Una vez halladas las sumas de cuadrados, continuamos con los cálculos de los cuadrados medios

$$s_1^2 = \frac{SSA}{k-1}$$

$$s_1^2 = \frac{16355000}{4-1} \Rightarrow s_1^2 = 5451667$$

$$s^2 = \frac{SSE}{k(n-1)}$$

$$s^2 = \frac{11283333}{4(6-1)} \Rightarrow s^2 = 564167$$

ya obtenidos estas medidas, el valor del estadístico de prueba sería

$$f = \frac{s_1^2}{s^2}$$

$$f = \frac{5451667}{564167} \Rightarrow f = 9.663$$

y el valor crítico de la distribución F de Fisher-Snedecor, según la Tabla A.5 es $f_{0.05[3,20]} = 3.10$

De esta forma, como $f = 9.663 > f_{0.05[3,20]} = 3.10$, se rechaza H_0 a favor de H_1 a un nivel de significancia de 0.05, concluyéndose que existen diferencias significativas en el verdadero valor medio de la concentración de enterococos fecales (UFC/100 mL) en al menos dos estaciones de monitoreo con una confiabilidad del 95%.

El resumen del análisis de varianza se resume en la siguiente tabla ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	*f* calculada
Enterococos	16355000	3	5451667	9.663
Error	11283333	20	564167	
Total	27638333	23	-	-

La salida de resultados para cada una de las etapas del análisis de varianza de un factor, se muestra a continuación. Iniciamos la ejecución del procedimiento, cargando los datos en R, a través de la construcción de vectores de datos para la densidad de enterococos en cada estación de monitoreo, y el grupo factor

```
> RT1<-c(1000,2100,700,500,100,200)
> RT2<-c(300,100,1000,600,600,700)
> RT3<-c(3300,2000,2200,3200,2700,2500)
> RT4<-c(1600,2500,3400,300,1000,800)
> Entero<-c(RT1,RT2,RT3,RT4)
> Estación<-rep(1:4,each=6)
> Estación<-factor(Estación,labels=c("RT1","RT2","RT3","RT4"))
```

Luego evaluamos el supuesto de normalidad de las densidades de enterococos en cada una de las estaciones de monitoreo a través del test de Shapiro-Wilk, dado a que nos enfrentamos a un conjunto pequeño de observaciones. Este test se aplica a través de la función ***shapiro.test***, discutida en secciones anteriores, y en este caso particular, para aplicar este test simultáneamente a las densidades de enterococos en las cuatro estaciones de monitoreo, utilizaremos la función ***by***, que permite ejecutar una función a un conjunto de datos categorizado por un factor (o grupo diferencial).

```
> by(Entero,Estación,shapiro.test)
Estación: RT1

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.8735, p-value = 0.2407

------------------------------------------------------------
Estación: RT2

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9645, p-value = 0.8536

------------------------------------------------------------
Estación: RT3

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9356, p-value = 0.6241

------------------------------------------------------------
Estación: RT4

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9447, p-value = 0.6974
```

Nótese que todos los p-valor > 0.05, por lo tanto, lo resultados son concluyentes en que las densidades de enterococos están normalmente distribuidas en las cuatro estaciones de monitoreo.

Como paso siguiente, evaluamos el supuesto de homogeneidad de las varianzas.

```
> bartlett.test(Entero,Estación)

        Bartlett test of homogeneity of variances

data:  Entero and Estación
Bartlett's K-squared = 7.5761, df = 3, p-value = 0.05564
```

Observe que al ser el p-valor mayor que el nivel de significancia, es evidencia suficiente para asumir que las densidades de enterococos presentan varianzas homogéneas en las cuatro estaciones de monitoreo.

Por último, dado el cumplimiento del modelo ANOVA, proseguimos a aplicar este procedimiento para evaluar la existencia, o no, de diferencias significativas en las densidades de enterococos entre las estaciones de monitoreo.

```
> Anova<-aov(Entero~Estación)
> summary(Anova)
            Df   Sum Sq Mean Sq F value   Pr(>F)
Estación     3 16355000 5451667   9.663 0.000377 ***
Residuals   20 11283333  564167
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Aquí, el p-valor = 0.000377 es mucho menor que el nivel de significancia preseleccionado, por lo tanto, se rechaza la hipótesis nula a favor de la alternativa y se concluye que existen diferencias significativas en las densidades de enterococos entre las cuatro estaciones de monitoreo en el municipio de Riohacha, La Guajira, con una confiabilidad del 95%.

7.4. Pruebas de comparaciones múltiples (post-hoc)

Un aspecto importante a considerar del análisis de varianza, es que si la conclusión del test aplicado fuera el rechazo de la hipótesis nula, no ocurriría como en el caso de dos poblaciones en el que claramente una de ellas tendría media superior a la otra, sino que habría que evaluar las relaciones entre las *k* poblaciones, bien dos a dos o a través de

combinaciones entre ellas, mediante los denominados *test de comparaciones múltiples*.

Existe una gran cantidad de test que realizan las comparaciones múltiples, tratando cada uno de ellos de adaptarse mejor a determinadas circunstancias. Cabe destacar, por ser de uso más extendido, los test de LSD de Fisher, Scheffé, HSD de Tukey, Student-Newman-Keuls, Duncan y Dunnett. Dependiendo de que las comparaciones sean entre parejas de medias o más generales, combinaciones de las mismas, será más aconsejable el test de Tukey o el de Scheffé. En el caso de comparaciones de parejas de medias, puesto que el de Tukey proporciona intervalos de confianza de menor longitud, se preferirá al de Scheffé. Por otro lado, si el interés se centra en realizar comparaciones de las medias de todos los tratamientos con un grupo control, el test de Dunnett, se constituye en el más adecuado para este tipo de situaciones (Gutiérrez & De la Vara, 2008; Montgomery, 2001; Zimmermann, 2004).

A continuación se discutirán los procedimientos de pruebas de cada uno de estos test y se realizarán comentarios de los escenarios en los que es más factible la aplicación de cada uno de ellos.

7.4.1. Test de la mínima diferencia significativa o LSD de Fisher

El test de mínima diferencia significativa (LSD, del inglés *least significant difference*), es un procedimiento que fue sugerido por Fisher (1935). Es el procedimiento para la comparación de las medias de *k* tratamientos más antiguo, y su aplicación consiste en una prueba de hipótesis por parejas basada en la distribución *t* de *student*, con el objetivo de probar la hipótesis nula $H_0 : \mu_i = \mu_j$, contra la alternativa $H_1 : \mu_i \neq \mu_j$ (Montgomery, 2001; Kuehl, 2001; Gutiérrez & De la Vara, 2008).

El estadístico de prueba para este test, se basa en el cálculo de la **mínima diferencia significativa (*LSD*)** que se puede obtener entre las medias de dos tratamientos para considerar que estos son significativamente diferentes entre sí. Bajo el supuesto de que este procedimiento sigue una distribución *t* de *student*, la diferencia mínima significativa se determina a través de la siguiente expresión

$$LSD = t_{\alpha/2, N-k} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Donde $t_{\alpha/2,N-k}$, se encuentra en la tablas de la distribución *t* de *student* (Tabla A.4) con $N - k$ grados de libertad que corresponde a la suma de los cuadrados de los errores (*SSE*), s^2 es el error cuadrático medio calculado en el análisis de varianza, n_i y n_j, son el número de observaciones para los tratamientos *i* y *j*.

El rechazo de la hipótesis nula para este test ocurre cuando

$$\left| \overline{y}_{i\cdot} - \overline{y}_{j\cdot} \right| > LSD$$

Cuando nuestro diseño experimental está balanceado, es decir, si $n_1 = n_2 = \cdots = n_k = n$, la expresión de cálculo de la diferencia mínima significativa se reduce a

$$LSD = t_{\alpha/2,N-k}\sqrt{\frac{2s^2}{n}}$$

Ésta última propiedad de este test, de poder ser aplicado aún cuando el diseño se encuentra desbalanceado, ya sea por la pérdida de unidades experimentales o por la falta de información para algunos tratamientos, se constituye en la principal ventaja de este procedimiento. Sin embargo tiene como desventaja que no controla la tasa de error tipo I, y para un número relativamente grande de tratamientos, el número de posibles falsos rechazos de la hipótesis nula puede ser elevado aunque no existan diferencias reales, es decir, es un test muy liberal o poco conservador (Cardoso & Veitía, 2008).

La aplicación del test *LSD* de Fisher en R se realiza a través de la función ***LSD.test*** del paquete "*agricolae*" (de Mendiburu, 2015), siguiendo la siguiente sintaxis de programación

```
LSD.test(y, trt, DFerror, MSerror, alpha = 0.05,p.adj=c("none",
"holm", "hochberg", "bonferroni", "BH", "BY", "fdr"),
group=TRUE)
```

Donde ***y*** representa a la variable dependiente o respuesta, ***trt*** indica el vector de los tratamientos, ***DFerror***, establece los grados de libertad de *SSE*, ***MSerror*** representa al error cuadrático medio (s^2), ***alpha*** indica la tasa de error tipo I o nivel de significancia, ***p.adj*** establece el método a usarse para el ajuste del p-valor, que en caso de ser "*none*", es el *t* de *student*, ***group***

indica si se establecen comparaciones entre grupos de medias de los tratamientos, dependiendo de si asume el valor lógico ***TRUE*** o ***FALSE***.

Ejemplo 7.5. Dado que el análisis de varianza de los datos de densidad de enterococos en cuatro estaciones de monitoreo de las playas del municipio de Riohacha resultó significativo, encontrar en que estaciones se presentaron dichas diferencias, con base en el test LSD de Fisher.

Solución

Con un valor del error cuadrático medio, $s^2 = 564167$, $n = 6$ y $t_{0.025,20} = 1.725$, el valor de la mínima diferencia significativa es

$$LSD = 1.725\sqrt{\frac{(2)(564167)}{6}} \therefore LSD = 748.053$$

El valor de las medias de las densidades de enterococos en las cuatro diferentes estaciones de monitoreo es

$\bar{y}_{RT_1}$	$\bar{y}_{RT_2}$	$\bar{y}_{RT_3}$	$\bar{y}_{RT_4}$
766.67	550	2650	1600

De esta forma los resultados del examen de diferencias entre las medias de las densidades de enterococos entre las estaciones de monitoreo sería

$\left|\bar{y}_{RT_1} - \bar{y}_{RT_2}\right| = 216.67 < 748.053$, no existen diferencias significativas
$\left|\bar{y}_{RT_1} - \bar{y}_{RT_3}\right| = 1883.33 > 748.053$, existen diferencias significativas
$\left|\bar{y}_{RT_1} - \bar{y}_{RT_4}\right| = 833.33 > 748.053$, existen diferencias significativas
$\left|\bar{y}_{RT_2} - \bar{y}_{RT_3}\right| = 2100 > 748.053$, existen diferencias significativas
$\left|\bar{y}_{RT_2} - \bar{y}_{RT_4}\right| = 1050 > 748.053$, existen diferencias significativas
$\left|\bar{y}_{RT_3} - \bar{y}_{RT_4}\right| = 1050 > 748.053$, existen diferencias significativas

Con base en lo anterior, vemos que existen diferencias significativas entre las medias de densidad de enterococos en todas las estaciones de monitoreo, a excepción de $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_2}$, a un nivel de significancia de 0.05. Un examen gráfico de estos resultados se puede realizar a través de la construcción de un gráfico de cajas, o a través de un gráfico de barras de las medias de cada grupo (Figura 7.2).

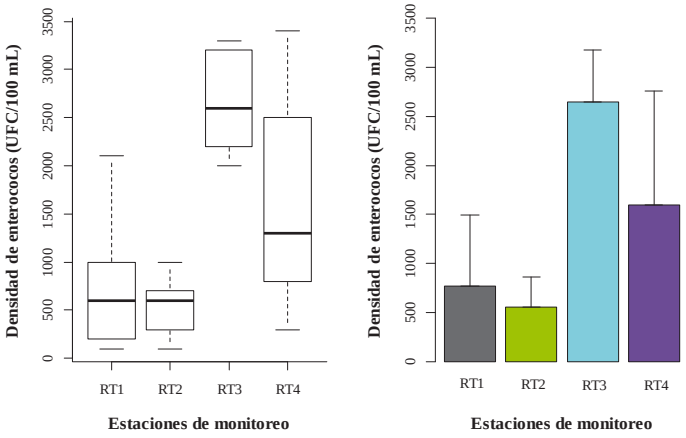


Figura 7.1. Densidad media de enterococos.

La salida de resultados para el test *LSD* de Fisher se muestra a continuación

```
> LSD.test(Entero,Estación,DFerror=20,MSerror=564167,alpha=0.05,
p.adj="none",group=FALSE,console=TRUE)

Study: Entero ~ Estación

LSD t Test for Entero

Mean Square Error:  564167

Estación,  means and individual ( 95 %) CI

       Entero       std r        LCL      UCL  Min  Max
RT1  766.6667  731.2090 6  127.02771 1406.306  100 2100
RT2  550.0000  314.6427 6  -89.63896 1189.639  100 1000
RT3 2650.0000  524.4044 6 2010.36104 3289.639 2000 3300
RT4 1600.0000 1161.0340 6  960.36104 2239.639  300 3400
alpha: 0.05 ; Df Error: 20
Critical Value of t: 2.085963

Comparison between treatments means

          Difference       pvalue sig.        LCL         UCL
RT1 - RT2   216.6667 6.227875e-01       -687.9194  1121.25276
RT1 - RT3 -1883.3333 3.155712e-04   *** -2787.9194  -978.74724
RT1 - RT4  -833.3333 6.901452e-02     . -1737.9194    71.25276
RT2 - RT3 -2100.0000 9.878957e-05   *** -3004.5861 -1195.41391
RT2 - RT4 -1050.0000 2.509684e-02     * -1954.5861  -145.41391
RT3 - RT4  1050.0000 2.509684e-02     *   145.4139  1954.58609
```

Obsérvese que todos los p-valores resultaron ser significativos (< 0.05), a excepción de primero y tercero de ellos, que evidencian diferencias significativas entre $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_2}$, y $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_4}$, a un nivel de significancia de 0.05.

7.4.2. Test de Student-Newman-Keuls (SNK) de rangos múltiples

Esta prueba fue desarrollada independientemente por Newman (1939) y Keuls (1952) y se clasifica como una prueba de intervalos (rangos) múltiples, ya que se usan dos o más intervalos entre medias como criterio de prueba (Kuehl, 2001). Este test, se basa en una distribución de *rango estudentizado*, al igual que los procedimientos de Tukey y Duncan que se discutirán más adelante, y desde el punto de vista operativo, el test *SNK* es similar al test de Duncan (Montgomery, 2001).

El criterio de prueba *SNK* en un diseño balanceado está dado por la siguiente expresión

$$SNK_{(\alpha,k)} = q_{\alpha,k,\upsilon}\sqrt{\frac{s^2}{n}} \text{ para } k = 2,3,...$$

donde $q_{\alpha,k,\upsilon}$, es el intervalo de rango estudentizado que se obtienen de la Tabla A.12 del apéndice, p es el número de medias de los tratamientos y υ los grados de libertad de *SSE*.

Para la aplicación del test *SNK*, se deben ordenar las medias de los tratamientos de forma ascendente, es decir $\bar{y}_{[1]} \leq \bar{y}_{[2]} \leq \cdots \leq \bar{y}_{[t]}$, donde $\bar{y}_{[1]}$ es la media del tratamiento con el valor más pequeño y $\bar{y}_{[t]}$ es la media del tratamiento con el valor más grande.

Así, para las medias mayor y menor de un conjunto de p medias, digamos $\bar{y}_{i\cdot}$ y $\bar{y}_{j\cdot}$, la hipótesis nula $H_0 : \mu_{i\cdot} = \mu_{j\cdot}$ se rechaza si

$$\left|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}\right| > SNK_{(\alpha,k)}$$

Para la aplicación del test SNK, una vez las medias de los tratamientos estén ordenadas en orden creciente, las diferencias observadas entre las medias se comparan con los rangos $SNK_{(\alpha,k)}$ de la siguiente manera: primero se compara la diferencia de la media más grande y la más pequeña con el rango $SNK_{(\alpha,k)}$. Luego, la diferencia entre la media más grande y la

segunda más pequeña se compara con el rango $SNK_{(\alpha,k-1)}$. Estas comparaciones continúan hasta que la media mayor se haya comparado con todas las demás. Enseguida, se compara la diferencia entre la segunda media más grande y la media menor con el rango $SNK_{(\alpha,k-1)}$. Después, la diferencia entre la segunda media más grande y la segunda más pequeña se comparan con el valor de $SNK_{(\alpha,k-2)}$, y así sucesivamente hasta que se comparen los $k(k-1)$ pares de medias posibles con el rango que les corresponda.

Este test es más conservador que el test *LSD* de Fisher e incluso que el de Duncan, en el sentido de que la tasa de erros tipo I es menor para este test, siendo más difícil que en los test *LSD* de Fisher y de Duncan rechazar la hipótesis nula (Montgomery, 2001, Morales, 2011). La desventaja más marcada del test *SNK*, radica en que su aplicación se limita a situaciones en las que el diseño se encuentra balanceado (Morales, 2011).

En R, el test de Student-Newman-Keuls se aplica a través de la función ***SNK.test*** del paquete "*agricolae*" (de Mendiburu, 2015), siguiente la siguiente sintaxis de programación

```
SNK.test(y, trt, DFerror, MSerror, alpha = 0.05,
group=TRUE,console=TRUE)
```

Donde ***y*** representa a la variable dependiente o respuesta, ***trt*** indica el vector de los tratamientos, ***DFerror***, establece los grados de libertad de *SSE*, ***MSerror*** representa al error cuadrático medio (s^2), ***alpha*** indica la tasa de error tipo I o nivel de significancia, y ***group*** indica si se establecen comparaciones entre grupos de medias de los tratamientos, dependiendo de si asume el valor lógico ***TRUE*** o ***FALSE***.

Ejemplo 7.6. Dado que el análisis de varianza de los datos de densidad de enterococos en cuatros estaciones de monitoreo de las playas del municipio de Riohacha resultó significativo, encontrar en qué estaciones se presentaron dichas diferencias, con base en el test *SNK* de rangos múltiples.

Solución

Con un valor del error cuadrático medio, $s^2 = 564167$ y $n = 6$, los valores del criterio de prueba para los diferentes grupos de tratamientos son, respectivamente

$$SNK_{(0.05,2)} = 2.95\sqrt{\frac{564167}{6}} = 904.587$$

$$SNK_{(0.05,3)} = 3.58\sqrt{\frac{564167}{6}} = 1097.770$$

$$SNK_{(0.05,4)} = 3.96\sqrt{\frac{564167}{6}} = 1214.292$$

Donde los valores de $q_{0.05,2,20} = 2.95$, $q_{0.05,23,20} = 3.58$ y $q_{0.05,4,20} = 3.96$, se encuentran en la Tabla A.12 del apéndice.

El valor de las medias de las densidades de enterococos en orden ascendente es

$\bar{y}_{RT_2}$	$\bar{y}_{RT_1}$	$\bar{y}_{RT_4}$	$\bar{y}_{RT_3}$
550	766.67	1600	2650

De esta forma los resultados del examen de diferencias significativas entre las medias de las densidades de enterococos entre las estaciones de monitoreo sería

$\bar{y}_{RT_3} - \bar{y}_{RT_2} = 2100 > SNK_{(0.05,4)} = 1214.292$, existen diferencias significativas
$\bar{y}_{RT_3} - \bar{y}_{RT_1} = 1883.33 > SNK_{(0.05,3)} = 1097.770$, existen diferencias significativas
$\bar{y}_{RT_3} - \bar{y}_{RT_4} = 1050.00 > SNK_{(0.05,2)} = 904.587$, existen diferencias significativas
$\bar{y}_{RT_4} - \bar{y}_{RT_2} = 1050 < SNK_{(0.05,3)} = 1097.770$, no existen diferencias significativas
$\bar{y}_{RT_4} - \bar{y}_{RT_1} = 833.33 < SNK_{(0.05,2)} = 904.587$, no existen diferencias significativas
$\bar{y}_{RT_1} - \bar{y}_{RT_2} = 216.67 < SNK_{(0.05,2)} = 904.587$, no existen diferencias significativas

Los resultados anteriores, al presentarlos en el orden en que se muestran los datos crudos, desde $\bar{y}_{RT_1}$ hasta $\bar{y}_{RT_4}$, y como presenta R los resultados sería

$\left|\bar{y}_{RT_1} - \bar{y}_{RT_2}\right| = 216.67 < SNK_{(0.05,2)} = 904.587$, no existen diferencias significativas
$\left|\bar{y}_{RT_1} - \bar{y}_{RT_3}\right| = 1883.33 > SNK_{(0.05,3)} = 1097.770$, existen diferencias significativas
$\left|\bar{y}_{RT_1} - \bar{y}_{RT_4}\right| = 833.33 < SNK_{(0.05,2)} = 904.587$, no existen diferencias significativas
$\left|\bar{y}_{RT_2} - \bar{y}_{RT_3}\right| = 2100 > SNK_{(0.05,4)} = 1214.292$, existen diferencias significativas
$\left|\bar{y}_{RT_2} - \bar{y}_{RT_4}\right| = 1050 < SNK_{(0.05,3)} = 1097.770$, no existen diferencias significativas
$\left|\bar{y}_{RT_3} - \bar{y}_{RT_4}\right| = 1050 > SNK_{(0.05,2)} = 904.587$, existen diferencias significativas

Con base en lo anterior, vemos que existen diferencias significativas entre las medias de densidad de enterococos en las estaciones de monitoreo $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_2}$, $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_4}$ y $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_4}$ a un nivel de significancia de 0.05.

A continuación mostraremos la salida de resultados de R para la aplicación del test *SNK*, a través de la función ***SNK.test*** del paquete "*agricoale*" (de Mendiburu, 2014).

```
> RT1<-c(1000,2100,700,500,100,200)
> RT2<-c(300,100,1000,600,600,700)
> RT3<-c(3300,2000,2200,3200,2700,2500)
> RT4<-c(1600,2500,3400,300,1000,800)
> Entero<-c(RT1,RT2,RT3,RT4)
> Estación<-rep(1:4,each=6)
> Estación<-factor(Estación,labels=c("RT1","RT2","RT3","RT4"))
> library(agricolae)
> SNK.test(Entero,Estación,DFerror=20,MSerror=564167,alpha=
0.05, group=FALSE,console=TRUE)

Study: Entero ~ Estación

Student Newman Keuls Test
for Entero

Mean Square Error:  564167

Estación,  means

       Entero       std r  Min  Max
RT1  766.6667  731.2090 6  100 2100
RT2  550.0000  314.6427 6  100 1000
RT3 2650.0000  524.4044 6 2000 3300
RT4 1600.0000 1161.0340 6  300 3400

alpha: 0.05 ; Df Error: 20

Critical Range
       2        3        4
 904.586 1097.136 1213.769

Comparison between treatments means

        Difference   pvalue sig.        LCL        UCL
RT1-RT2   216.6667 0.622787       -687.9194 1121.25271
RT1-RT3 -1883.3333 0.000882  *** -2980.4697 -786.19697
RT1-RT4  -833.3333 0.069015    . -1737.9194   71.25271
RT2-RT3 -2100.0000 0.000530  *** -3313.7695 -886.23055
RT2-RT4 -1050.0000 0.062215    . -2147.1364   47.13636
RT3-RT4  1050.0000 0.025097    *   145.4140 1954.58604
```

Un examen de los p-valores del test, evidencia que existen diferencias significativas (p-valor < 0.05) entre las medias $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_2}$, $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_4}$ y $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_4}$, a un nivel de significancia de 0.05.

7.4.3. Test de Scheffé

Es uno de los contrastes más utilizado y no suele faltar en los textos de estadística. Este método, desarrollado por Scheffé (1953), está diseñado para probar todos los posibles contrastes de medias de los tratamientos que puedan interesar al investigador, sin el inconveniente de inflar por ello el error tipo I (detección de diferencias que no existen) (Gutiérrez & De la Vara, 2008; Kuehl, 2001), como ocurre con el test LSD de Fisher.

Se trata de un método seguro, que puede ser siempre utilizado, y permite comparar las medias de dos en dos, o una media con una combinación lineal de otras, etc., y las muestras pueden ser tanto de idéntico tamaño (diseño equilibrado) como de tamaño desigual (diseño desequilibrado). Así mismo, es el método más fiable y más seguro de utilizar cuando se violan los supuestos de normalidad y de homogeneidad de varianzas. Esta es una buena razón para utilizarlo cuando dudamos de la legitimidad del análisis de varianza porque no se cumplen los requisitos previos (Morales, 2011).

A pesar de ser posible la aplicación del test de Scheffé, en todas las situaciones, no siempre resulta ser el test más recomendable, dado a que es un test muy conservador y se puede tender a no rechazar la hipótesis nula, cuando en realidad si se pueden encontrar diferencias significativas con otros test, principalmente cuando las comparaciones de las medias de los tratamientos se hace por pares. Sin embargo, su poder aumentar por encima del de otras pruebas de comparaciones múltiples cuando las comparaciones se hacen entre grupos de medias o entre una media y combinaciones lineales de las otras, de la forma $C = \sum_{i=1}^{k} a_i \mu_i$, por ejemplo, se puede estar interesado en probar la hipótesis

$$H_0 : 2\mu_h = \mu_i + \mu_j \qquad H_0 : 2\mu_h - \mu_i - \mu_j = 0$$
$$H_1 : 2\mu_h \neq \mu_i + \mu_j \quad \text{o} \quad H_1 : 2\mu_h - \mu_i - \mu_j \neq 0$$

Que implica que la hipótesis está definida por el contraste $C = 2\mu_h - \mu_i - \mu_j$, cuya estimación a partir de los datos de la muestra está dado por

$$c = 2\bar{y}_{h\cdot} - \bar{y}_{i\cdot} - \bar{y}_{j\cdot}$$

Con una desviación estándar estimada dada por la expresión

$$s_c = \sqrt{s^2 \left(\sum_{i=1}^{k} \frac{a_i^2}{n_i} \right)}$$

Así, el criterio de prueba para el procedimiento de Scheffé se define como

$$S_\alpha = s_c \sqrt{(k-1)\, f_{\alpha, k-1, N-k}}$$

Donde $f_{\alpha,k-1,N-k}$ es el valor que se encuentra en las tablas de la distribución *F* de Fisher-Snedecor con *k* - 1 grados de libertad en el numerador y *N* - *k* grados de libertad en el denominador (Tabla A.6) que corresponden a los grados de libertad de la suma de cuadrados de los tratamientos (*SSA*) y la suma de cuadrados de los errores (*SSE*) del ANOVA, respectivamente.

En virtud de los anterior la hipótesis nula se puede expresar alternativamente como $H_0 : C = 0$ y ésta se rechaza cuando

$$|c| > S_\alpha$$

La principal ventaja del test de Scheffé es que es un test blindado contra el error tipo I (aceptar la hipótesis nula cuando ésta en realidad es falsa), y además permite realizar el contraste entre las medias de tratamientos, o combinaciones de ellas, que tengan diferente número de observaciones. Sin embargo, al ser un test muy conservador, es posible que no se encuentren diferencias significativas que objetivamente se puedan encontrar con otros procedimientos, por ello para la aplicación de este test, es recomendable trabajar con tasas de error o nivel de significancia más grande (0.10), o cuando el objetivo sea la comparación de los tratamientos por parejas y no combinaciones de ellos, es recomendable la utilización de otros procedimientos como el de Tukey.

En el ambiente de programación de R, la aplicación de este test se realiza a través de la función ***scheffe.test*** del paquete "*agricolae*" (de Mendiburu, 2015), siguiendo la siguiente línea de comando

```
scheffe.test(y, trt, DFerror, MSerror, alpha = 0.05, group=
FALSE,console= TRUE)
```

Donde ***y*** representa a la variable dependiente o respuesta, ***trt*** indica el vector de los tratamientos, ***DFerror***, establece los grados de libertad de *SSE*,

MSerror representa al error cuadrático medio (s^2) que se obtiene en el ANOVA, ***alpha*** indica la tasa de error tipo I o nivel de significancia, y ***group*** indica si se establecen comparaciones entre grupos de medias de los tratamientos, dependiendo de si asume el valor lógico ***TRUE*** o ***FALSE***.

Ejemplo 7.7. Dado que el análisis de varianza de los datos de densidad de enterococos en cuatros estaciones de monitoreo de las playas del municipio de Riohacha resultó significativo, encontrar en que estaciones se presentaron dichas diferencias, con base en el test Scheffé.

Solución

Con un valor del error cuadrático medio, $s^2 = 564167$ y $n = 6$, dado que el contraste se realizará por pares de tratamiento y no combinaciones de ellos, el valor de S_c y S_α, será el mismo para todas las parejas, cuyos valores respectivos son

$$S_c = \sqrt{564167\left(\frac{1+1}{6}\right)} = 433.654$$

y

$$S_\alpha = 433.654\sqrt{(3)(3.10)} = 1322.467$$

Donde el valor 3.10 dentro del radical, corresponde al valor tabulado $f_{0.05,3,20}$, que se encuentran en la Tabla A.6 del apéndice.

De esta forma los resultados del examen de diferencias significativas entre las medias de las densidades de enterococos en las estaciones de monitoreo, se puede evaluar de una forma resumida a través de la siguiente tabla

Pares de tratamiento	$\lvert c \rvert$	S_c	S_α	Decisión
$RT_1 - RT_2$	$\lvert \bar{y}_{RT_1} - \bar{y}_{RT_2} \rvert = 216.67$	433.654	1322.467	No existen diferencias significativas
$RT_1 - RT_3$	$\lvert \bar{y}_{RT_1} - \bar{y}_{RT_3} \rvert = 1883.33$			Existen diferencias significativas
$RT_1 - RT_4$	$\lvert \bar{y}_{RT_1} - \bar{y}_{RT_4} \rvert = 833.33$			No existen diferencias significativas
$RT_2 - RT_3$	$\lvert \bar{y}_{RT_2} - \bar{y}_{RT_3} \rvert = 2100$			Existen diferencias significativas
$RT_2 - RT_4$	$\lvert \bar{y}_{RT_2} - \bar{y}_{RT_4} \rvert = 1050$			No existen diferencias significativas
$RT_3 - RT_4$	$\lvert \bar{y}_{RT_3} - \bar{y}_{RT_4} \rvert = 1050$			No existen diferencias significativas

Con base en lo anterior, vemos que no existen diferencias significativas entre las medias de densidad de enterococos en todas las estaciones de monitoreo, a excepción de $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_3}$ y $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_3}$, a un nivel de significancia de 0.05.

A continuación se muestra la salida de resultados de R para la aplicación del test Scheffé, a través de la función ***scheffe.test*** del paquete "*agricoale*" (de Mendiburu, 2014).

```
> RT1<-c(1000,2100,700,500,100,200)
> RT2<-c(300,100,1000,600,600,700)
> RT3<-c(3300,2000,2200,3200,2700,2500)
> RT4<-c(1600,2500,3400,300,1000,800)
> Entero<-c(RT1,RT2,RT3,RT4)
> Estación<-rep(1:4,each=6)
> Estación<-factor(Estación,labels=c("RT1","RT2","RT3","RT4"))
> library(agricolae)
>
scheffe.test(Entero,Estación,DFerror=20,MSerror=564167,alpha=
0.05,group=FALSE,console=TRUE)

Study: Entero ~ Estación

Scheffe Test for Entero

Mean Square Error  : 564167

Estación,  means

       Entero       std r  Min  Max
RT1  766.6667  731.2090 6  100 2100
RT2  550.0000  314.6427 6  100 1000
RT3 2650.0000  524.4044 6 2000 3300
RT4 1600.0000 1161.0340 6  300 3400

alpha: 0.05 ; Df Error: 20
Critical Value of F: 3.098391

Comparison between treatments means

          Difference   pvalue sig        LCL       UCL
RT1 - RT2   216.6667 0.968407     -1094.8331 1528.1664
RT1 - RT3 -1883.3333 0.003516  ** -3194.8331 -571.8336
RT1 - RT4  -833.3333 0.324602     -2144.8331  478.1664
RT2 - RT3 -2100.0000 0.001204  ** -3411.4998 -788.5002
RT2 - RT4 -1050.0000 0.153422     -2361.4998  261.4998
RT3 - RT4  1050.0000 0.153422      -261.4998 2361.4998
```

El examen de los p-valores del test, evidencia que existen diferencias significativas (p-valor < 0.05) entre las medias $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_3}$ y $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_3}$, a un nivel de significancia de 0.05.

7.4.3.1. Test de Tukey o prueba de la Diferencia Honestamente Significativa (HSD)

Un método bastante conservador y de amplio uso para la comparación de pares de medias es el test desarrollado por Tukey (1953), también conocido como la prueba de la Diferencia Honestamente Significativa (*HSD*, del inglés *Honestly Significant Difference*). Este test es considerado más eficaz que el test de Scheffé discutido anteriormente, cuando las comparaciones de las medias de los tratamientos se realiza por pares (Zimmermann, 2004).

La aplicación del test de Tukey, siempre que se pueda garantizar el cumplimiento de los supuestos de homogeneidad de varianzas y normalidad de los datos, se basa en una función de α, k y υ, y las diferencias significativas entre las medias i y j $(i \neq j)$, se encuentran cuando

$$\left|\bar{y}_{i.} - \bar{y}_{j.}\right| > T = q_{\alpha,k,\upsilon}\sqrt{\frac{s^2}{n}}$$

Donde α es el nivel de significancia seleccionado, k es el número de tratamientos, υ, son los grados de libertad para *SSE*, s^2 es el error cuadrático medio, n es el número de observaciones en los k tratamientos y $q_{\alpha,k,\upsilon}$ es el intervalo de rango estudentizado encontrados en la Tabla A.12 del apéndice. Cabe destacar que el test de Tukey requiere para su aplicación que el diseño se encuentre balanceado, por lo que para algunas situaciones experimentales podría significar una desventaja, de otra forma, es el test preferido pues al mantener fija la tasa de error tipo I, identifica verdaderas diferencias significativas entre las medias de los tratamientos.

En R el test HSD de Tukey se realiza a través de la función ***HSD.test*** del paquete "*agricolae*" (de Mendiburu, 2015), siguiendo la sintaxis que se muestra a continuación

```
HSD.test(y, trt, DFerror, MSerror, alpha = 0.05, group=TRUE,
console = TRUE)
```

Donde ***y*** representa a la variable dependiente o respuesta, ***trt*** indica el vector de los tratamientos, ***DFerror***, establece los grados de libertad de *SSE*, ***MSerror*** representa al error cuadrático medio (s^2) que se obtiene en el ANOVA, ***alpha*** indica la tasa de error tipo I o nivel de significancia, y ***group*** indica si se establecen comparaciones entre grupos de medias de los tratamientos, dependiendo de si asume el valor lógico ***TRUE*** o ***FALSE***.

Ejemplo 7.8. Dado que el análisis de varianza de los datos de densidad de enterococos dé en cuatro estaciones de monitoreo de las playas del municipio de Riohacha resultó significativo, encontrar en qué estaciones se presentaron dichas diferencias, con base en el test *HSD* de Tukey.

Solución

Con un valor del error cuadrático medio, $s^2 = 564167$, $n = 6$ y $q_{0.05,4,20} = 3.96$, el valor del criterio de prueba de Tukey es

$$T = 3.96\sqrt{\frac{(564167)}{6}} \therefore T = 1214.293$$

El valor de las medias de las densidades de enterococos en las cuatro diferentes estaciones de monitoreo es

$\overline{y}_{RT_1}$	$\overline{y}_{RT_2}$	$\overline{y}_{RT_3}$	$\overline{y}_{RT_4}$
766.67	550	2650	1600

De esta forma los resultados del examen de diferencias entre las medias de las densidades de enterococos entre las estaciones de monitoreo sería

$\left|\overline{y}_{RT_1} - \overline{y}_{RT_2}\right| = 216.67 < 1214.293$, no existen diferencias significativas

$\left|\overline{y}_{RT_1} - \overline{y}_{RT_3}\right| = 1883.33 > 1214.293$, existen diferencias significativas

$\left|\overline{y}_{RT_1} - \overline{y}_{RT_4}\right| = 833.33 < 1214.293$, no existen diferencias significativas

$\left|\overline{y}_{RT_2} - \overline{y}_{RT_3}\right| = 2100 > 1214.293$, existen diferencias significativas

$\left|\overline{y}_{RT_2} - \overline{y}_{RT_4}\right| = 1050 < 1214.293$, no existen diferencias significativas

$\left|\overline{y}_{RT_3} - \overline{y}_{RT_4}\right| = 1050 < 1214.293$, no existen diferencias significativas

Con base en lo anterior, vemos que existen diferencias significativas entre las medias de densidad de enterococos de las estaciones de monitoreo $\overline{y}_{RT_1}$ *vs* $\overline{y}_{RT_3}$, y $\overline{y}_{RT_2}$ *vs* $\overline{y}_{RT_3}$, a un nivel de significancia de 0.05.

A continuación se muestra la salida de resultados de R para la aplicación del test de Tukey, a través de la función ***HSD.test*** del paquete "*agricoale*" (de Mendiburu, 2014).

```
> RT1<-c(1000,2100,700,500,100,200)
> RT2<-c(300,100,1000,600,600,700)
> RT3<-c(3300,2000,2200,3200,2700,2500)
> RT4<-c(1600,2500,3400,300,1000,800)
> Entero<-c(RT1,RT2,RT3,RT4)
> Estación<-rep(1:4,each=6)
> Estación<-factor(Estación,labels=c("RT1","RT2","RT3","RT4"))
> library(agricolae)
> HSD.test(Entero,Estación,DFerror=20,MSerror=564167,alpha=
0.05, group=FALSE,console=TRUE)

Study: Entero ~ Estación

HSD Test for Entero

Mean Square Error:  564167

Estación,  means

       Entero       std r  Min  Max
RT1  766.6667  731.2090 6  100 2100
RT2  550.0000  314.6427 6  100 1000
RT3 2650.0000  524.4044 6 2000 3300
RT4 1600.0000 1161.0340 6  300 3400

alpha: 0.05 ; Df Error: 20
Critical Value of Studentized Range: 3.958293

Comparison between treatments means

          Difference   pvalue sig.        LCL       UCL
RT1 - RT2   216.6667 0.958212       -997.1028 1430.4361
RT1 - RT3 -1883.3333 0.001655   ** -3097.1028 -669.5639
RT1 - RT4  -833.3333 0.250872      -2047.1028  380.4361
RT2 - RT3 -2100.0000 0.000530  *** -3313.7695 -886.2305
RT2 - RT4 -1050.0000 0.105113      -2263.7695  163.7695
RT3 - RT4  1050.0000 0.105113       -163.7695 2263.7695
```

El examen de los p-valores del test, evidencia que existen diferencias significativas (p-valor < 0.05) entre las medias $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_3}$ y $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_3}$, a un nivel de significancia de 0.05.

7.4.4. Test de Duncan de rangos múltiples

Este procedimiento introducido por Duncan (1955) para la comparación de medias si los *k* tratamientos poseen el mismo número de observaciones.

Adopta el principio de rango estudentizado al igual el test de Tukey y *SNK*. La aplicación de este test sigue la misma dinámica operativa del test *SNK*, como se había comentado antes, implicando el ordenamiento de las medias de los tratamientos en orden ascendente, y al ser un test iterativo, al comparar dos medias cualesquiera, deben ser consideradas no solamente aquellas medias, sino todas las demás que se sitúen entre ellas.

El criterio de prueba de este test se basa en el cálculo de cierto valor que recibe el nombre de **rango mínimo significativo** (Walpole *et al.*, 2007), el cual, cuando se consideran *p* medias a un nivel de significancia α (tasa de error tipo I) se encuentra determinado por la expresión

$$R_k = r_{\alpha,k,\upsilon}\sqrt{\frac{s^2}{n}} \text{ para } k = 2,3,...$$

Donde υ, son los grados de libertad para *SSE*, s^2 es el error cuadrático medio y *n* es el número de observaciones en los *k* tratamientos. Los valores de $r_{\alpha,k,\upsilon}$, llamados **rango mínimo significativos estudentizado**, se encuentran tabulados en la Tabla A.13 del apéndice.

Las diferencias significativas entre las medias i y j $(i \neq j)$, se encuentran cuando

$$\left|\bar{y}_{i.} - \bar{y}_{j.}\right| > R_p$$

Para la aplicación del test de rangos múltiples de Duncan, una vez las medias de los tratamientos estén ordenadas en orden creciente, las diferencias observadas entre las medias se comparan con los rangos R_k de la siguiente manera: primero se compara la diferencia de la media más grande y la más pequeña con el rango R_k. Luego, la diferencia entre la media más grande y la segunda más pequeña se compara con el rango R_{k-1}. Estas comparaciones continúan hasta que la media mayor se haya comparado con todas las demás. Enseguida, se compara la diferencia entre la segunda media más grande y la media menor con el rango R_{k-1}. Después, la diferencia entre la segunda media más grande y la segunda más pequeña se comparan con el valor de R_{k-2}, y así sucesivamente hasta que se comparen los $k(k-1)$ pares de medias posibles con el rango que les corresponda (Gutiérrez & De la Vara, 2008).

Entre los test *LSD* de Fisher y el test *HSD* de Tukey, el test de Duncan es un procedimiento intermedio en función de su característica

conservacionista, pues proporciona menos diferencias significativas entre las medias que el test *LSD*, pero más que el test *HSD* de Tukey, por ello se puede elegir entre cualquiera de ello, dependiendo del riesgo que queramos tomar acerca de rechazar o no, la hipótesis nula de igualdad de las medias.

En R la aplicación del test de Duncan de rangos múltiples es simple, y se realiza a través de la función ***duncan.test*** del paquete "*agricolae*" (de Mendiburu, 2015), ciñéndose a la siguiente línea de código

```
duncan.test(y, trt, DFerror, MSerror, alpha = 0.05,
group=TRUE, console = TRUE)
```

Donde nuevamente ***y*** representa a la variable dependiente o respuesta, ***trt*** indica el vector de los tratamientos, ***DFerror***, establece los grados de libertad de *SSE*, ***MSerror*** representa al error cuadrático medio (s^2) que se obtiene en el ANOVA, ***alpha*** indica la tasa de error tipo I o nivel de significancia, y ***group*** indica si se establecen comparaciones entre grupos de medias de los tratamientos, dependiendo de si asume el valor lógico ***TRUE*** o ***FALSE***.

Ejemplo 7.9. Dado que el análisis de varianza de los datos de densidad de enterococos de en cuatros estaciones de monitoreo de las playas del municipio de Riohacha resultó significativo, encontrar en que estaciones se presentaron dichas diferencias, a través del test Duncan de rango múltiple.

Solución

Con un valor del error cuadrático medio, $s^2 = 564167$ y $n = 6$, los valores del criterio de prueba para los diferentes grupos de tratamientos son, respectivamente

$$R_2 = 2.95\sqrt{\frac{564167}{6}} = 904.587$$

$$R_3 = 3.097\sqrt{\frac{564167}{6}} = 949.663$$

$$R_4 = 3.190\sqrt{\frac{564167}{6}} = 978.180$$

Donde los valores de $r_{0.05,2,20} = 2.950$, $r_{0.05,23,20} = 3.097$ y $r_{0.05,4,20} = 3.190$, se encuentran en la Tabla A.13 del apéndice.

El valor de las medias de las densidades de enterococos en orden ascendente

$\bar{y}_{RT_2}$	$\bar{y}_{RT_1}$	$\bar{y}_{RT_4}$	$\bar{y}_{RT_3}$
550	766.67	1600	2650

De esta forma los resultados del examen de diferencias significativas entre las medias de las densidades de enterococos entre las estaciones de monitoreo sería

$\bar{y}_{RT_3} - \bar{y}_{RT_2} = 2100 > R_4 = 978.180$, existen diferencias significativas
$\bar{y}_{RT_3} - \bar{y}_{RT_1} = 1883.33 > R_3 = 949.663$, existen diferencias significativas
$\bar{y}_{RT_3} - \bar{y}_{RT_4} = 1050.00 > R_2 = 904.587$, existen diferencias significativas
$\bar{y}_{RT_4} - \bar{y}_{RT_2} = 1050 > R_3 = 949.663$, existen diferencias significativas
$\bar{y}_{RT_4} - \bar{y}_{RT_1} = 833.33 < R_2 = 904.687$, no existen diferencias significativas
$\bar{y}_{RT_1} - \bar{y}_{RT_2} = 216.67 < R_2 = 904.587$, no existen diferencias significativas

Los resultados anteriores, al presentarlos en el orden en que se muestran los datos crudos, desde $\bar{y}_{RT_1}$ hasta $\bar{y}_{RT_4}$, y como presenta R los resultados sería

$\left|\bar{y}_{RT_1} - \bar{y}_{RT_2}\right| = 216.67 < R_2 = 904.587$, no existen diferencias significativas
$\left|\bar{y}_{RT_1} - \bar{y}_{RT_3}\right| = 1883.33 > R_3 = 949.663$, existen diferencias significativas
$\left|\bar{y}_{RT_1} - \bar{y}_{RT_4}\right| = 833.33 < R_2 = 904.587$, no existen diferencias significativas
$\left|\bar{y}_{RT_2} - \bar{y}_{RT_3}\right| = 2100 > R_4 = 978.180$, existen diferencias significativas
$\left|\bar{y}_{RT_2} - \bar{y}_{RT_4}\right| = 1050 > R_3 = 949.663$, existen diferencias significativas
$\left|\bar{y}_{RT_3} - \bar{y}_{RT_4}\right| = 1050 > R_2 = 949.663$, existen diferencias significativas

Con base en lo anterior, vemos que existen diferencias significativas entre las medias de densidad de enterococos en las estaciones de monitoreo $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_3}$, $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_3}$, $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_4}$ y $\bar{y}_{RT_3}$ *vs* $\bar{y}_{RT_4}$ a un nivel de significancia de 0.05.

A continuación mostraremos la salida de resultados de R para la aplicación del test Duncan, a través de la función ***duncan.test*** del paquete "*agricoale*" (de Mendiburu, 2014).

```
> RT1<-c(1000,2100,700,500,100,200)
> RT2<-c(300,100,1000,600,600,700)
> RT3<-c(3300,2000,2200,3200,2700,2500)
> RT4<-c(1600,2500,3400,300,1000,800)
> Entero<-c(RT1,RT2,RT3,RT4)
> Estación<-rep(1:4,each=6)
> Estación<-factor(Estación,labels=c("RT1","RT2","RT3","RT4"))
> library(agricolae)
> duncan.test(Entero,Estación,DFerror=20,MSerror=564167,alpha
= 0.05,group=FALSE,console=TRUE)

Study: Entero ~ Estación

Duncan's new multiple range test
for Entero

Mean Square Error:  564167

Estación,  means

       Entero       std r  Min  Max
RT1  766.6667  731.2090 6  100 2100
RT2  550.0000  314.6427 6  100 1000
RT3 2650.0000  524.4044 6 2000 3300
RT4 1600.0000 1161.0340 6  300 3400

alpha: 0.05 ; Df Error: 20

Critical Range
       2        3        4
904.5860 949.5114 978.0625

Comparison between treatments means

          Difference   pvalue sig.        LCL         UCL
RT1 - RT2   216.6667 0.622787       -687.9194  1121.25271
RT1 - RT3 -1883.3333 0.000441  *** -2832.8448  -933.82190
RT1 - RT4  -833.3333 0.069015    . -1737.9194    71.25271
RT2 - RT3 -2100.0000 0.000177  *** -3078.0625 -1121.93750
RT2 - RT4 -1050.0000 0.031108    * -1999.5114  -100.48857
RT3 - RT4  1050.0000 0.025097    *   145.4140  1954.58604
```

El examen de los p-valores del test, evidencia que existen diferencias significativas (p-valor < 0.05) entre las medias $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_3}$, $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_3}$, $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_4}$ y $\bar{y}_{RT_3}$ *vs* $\bar{y}_{RT_4}$, a un nivel de significancia de 0.05.

7.4.5. Comparación de los tratamientos con un control: Test de Dunnett

Existen muchas situaciones en las que una vez se haya rechazado la hipótesis nula en el análisis de varianza, el interés recae en encontrar si existen diferencias significativas entre las medias de cada tratamiento y un control, es decir, la investigación está dirigida en comparar cada una de las otras k - 1 medias de los tratamientos contra el control, por lo tanto, existen k - 1 comparaciones. Un procedimiento para realizar estas comparaciones es la prueba introducida por Dunnett (1955), que en su forma bilateral busca probar el siguiente sistema de hipótesis

$$H_0 : \mu_i = \mu_c$$

$$H_1 : \mu_i \neq \mu_c$$

Donde μ_c representa el rendimiento promedio para la población de mediciones en la cual se utiliza el control. El criterio de Dunnett para comparar los $k - 1$ tratamientos con el control es:

$$D_{\alpha,k-1} = d_{\alpha,k-1,\upsilon}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_c}\right)}$$

Donde la varianza muestral s^2 se obtiene de la media cuadrática de los errores en el análisis de varianza y υ representa los grados de libertad para el cálculo de s^2. Así, se rechaza H_0 a un nivel de significancia α, cuando $|\bar{y}_{i.} - \bar{y}_{c.}| > D_{\alpha,k-1}$, haciendo uso de los valores tabulados de $d_{\alpha,k-1,\upsilon}$ para el método de Dunnett a dos colas (alternativa bilateral) (Tabla A.14 y A.16)

Si se desea probar la hipótesis nula $H_0 : \mu_i = \mu_c$, contra la alternativa unilateral $H_1 : \mu_i > \mu_c$ o $H_1 : \mu_i < \mu_c$, se rechaza H_0 a un nivel de significancia α, cuando $|\bar{y}_{i.} - \bar{y}_{c.}| > D_{\alpha,k-1}$, haciendo uso de los valores tabulados de $d_{\alpha,k-1,\upsilon}$ para el método de Dunnett a una cola (alternativa unilateral) (Tabla A.15 y A.17).

En R, la aplicación del test de Dunnett se ejecuta a través de la función ***SimTestDiff*** del paquete "*SimComp*" (Hasler, 2015), siguiendo la siguiente sintaxis de programación

```
SimTestDiff(data, grp, resp = NULL, type = "Dunnett", base = 1,
alternative = "two.sided")
```

Donde ***data***, se refiere al conjunto de datos (data frame), que contiene a las variables del estudio, ***grp*** corresponde al nombre del grupo diferencial o tratamiento, ***resp*** indica el nombre de la variable respuesta, ***type*** siempre se dejará con su valor por defecto (Dunnett), pues esta función también aplica otros test de comparaciones múltiples, ***base*** indica a través de un numero entero, cual es el grupo que se tomara como referencia para la realización de las comparaciones con los $k - 1$ tratamientos restantes, y ***alternative*** como se ha comentado antes establece la hipótesis alternativa que deseamos probar.

A continuación ejemplificaremos el uso de este test en una situación en la que los resultados del ANOVA resultan significativos.

Ejemplo 7.10. Se tiene la sospecha que las redes de distribución de agua de la universidad de La Guajira por su vejez pueden estar sufriendo de infiltraciones de contaminantes orgánicos producto del lavado de los suelos después de eventos de precipitación. Para probar si dicha conjetura es cierta se realizó un estudio por un grupo de estudiantes de Ingeniería Ambiental con el objeto de determinar la calidad fisicoquímica y microbiológica del agua de las redes de distribución, en el que se pretende comparar las concentraciones de las variables de estudio en diferentes puntos dentro de la universidad (Bloque Administrativo, Bloque I y Bloque IV) y la fuente de abastecimiento de agua (Pozo Aujero). A continuación se muestra una porción de los datos del estudio, en donde se considera solo las concentraciones de fosfatos (PO_4) en μg/L en los diferentes puntos de monitoreo. A partir de estos datos probar si la sospecha es cierta o falsa.

	Estaciones de monitoreo				
	Bloque Administrativo	Bloque I	Bloque IV	Pozo Aujero	
	87.9	36.7	37.5	35.9	
	75.05	62.7	47.9	27.5	
	62.2	24.8	19.4	30.2	
	55.9	24.9	24.9	24.9	
	56.7	45.05	60.2	29.9	
	69.9	35.7	46.2	25.2	
Total	407.65	229.85	236.1	173.6	1047.2
Media	67.94	38.31	39.35	28.93	43.63

Solución

Obviaremos todos los cálculos del ANOVA y a continuación mostraremos la tabla resultante

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	*f* calculada
Fosfatos	5122.13	3	1707.38	11.317
Error	3017.40	20	150.87	
Total	8139.53	23	-	-

Como $f = 11.317 > f_{0.05[3,20]} = 3.10$, se rechaza la hipótesis nula $H_0 = \mu_{BA} = \mu_{BI} = \mu_{BIV} = \mu_{PA}$, a un nivel de significancia de 0.05, concluyéndose que existen diferencias significativas en el verdadero valor medio de la concentración de PO_4 (μg/L) en al menos dos puntos de monitoreo con una confiabilidad del 95%.

Ahora, determinaremos que puntos de monitoreo presentan diferencias significativas en la concentración media de PO_4 (μg/L) con el Pozo Aujero (Grupo Control). Dada la existencia de un control, el test adecuado para este fin es el test de Dunnett, donde se tiene interés de probar las siguientes hipótesis

$$H_0 : \mu_i = \mu_c$$

$$H_1 : \mu_i > \mu_c$$

Con $k - 1 = 3$, $n_i = 6$, $n_c = 6$ y $d_{0.05,3,20} = 2.19$, el valor del criterio de prueba del test de Dunnett es

$$D_{0.05,3} = 2.19\sqrt{150.87\left(\frac{1}{6}+\frac{1}{6}\right)} \therefore D_{0.05,3} = 15.530$$

Los resultados del examen de las tres (k – 1) comparaciones por pares de cada tratamiento con el control se muestran en la siguiente tabla

Punto de monitoreo	$\bar{y}_{i.}$	$\lvert\bar{y}_{i.} - \bar{y}_{c.}\rvert$	Decisión
Pozo Aujero	28.93	-	-
Bloque Administrativo	67.94	39.01	> 15.530; hay diferencias significativas
Bloque I	38.31	9.38	< 15.530; no hay diferencias significativas
Bloque IV	39.35	10.42	< 15.530; no hay diferencias significativas

Como se puede observar solo la concentración media de PO_4 en el Bloque Administrativo es significativamente mayor que la del Pozo Aujero, siendo

ciertas las sospechas solo en este tramo de la red de distribución de agua, a un nivel de significancia de 0.05.

A continuación mostramos la salida de resultados de R para el test de Dunnett, a través de la función SimTestDiff del paquete "SimComp" (Hasler, 2015).

```
> BA<-c(87.9,75.05,62.2,55.9,56.7,69.9)
> BI<-c(36.7,62.7,24.8,24.9,45.05,35.7)
> BIV<-c(37.5,47.9,19.4,24.9,60.2,46.2)
> PA<-c(35.9,27.5,30.2,24.9,29.9,25.2)
> PO4<-c(BA,BI,BIV,PA)
> Punto<-rep(1:4,each=6)
> Punto<-factor(Punto,labels=c("BA","BI","BIV","PA"))
> PO4<-data.frame(Punto,PO4)
> library(SimComp)
> SimTestDiff(data=PO4,grp="Punto",resp="PO4",type="Dunnett",
base=4,alternative="greater")

Test for differences of means of multiple endpoints
Assumption: Heterogeneous covariance matrices for the groups
Alternative hypotheses: True differences greater than the
margins

  comparison endpoint margin estimate statistic p.value.raw
p.value.adj
1    BA - PA      PO4      0   39.008     7.370      0.0001
0.0006
2    BI - PA      PO4      0    9.375     1.552      0.0867
0.2295
3   BIV - PA      PO4      0   10.417     1.615      0.0800
0.2129
```

El examen de los p-valores del test de Dunnett, muestra la existencia de que las concentraciones de PO_4 es significativamente mayor (p-valor < 0.05) en el Bloque administrativo (BA) que en el Pozo Aujero (PA), a un nivel de significancia de 0.05.

7.5. Análisis de varianza para diseño de bloques completos aleatorios (BCA)

En secciones anteriores se discutió el diseño de experimentos completamente al azar para situaciones en las que las diferencias significativas entre las medias se debe a los efectos ejercidos por los

tratamientos, es decir, cuando la variabilidad de estos claramente sobrepasa la variabilidad debida al error experimental. En este tipo de diseño se asume que no existen efectos externos que podrían contribuir a la variabilidad de las observaciones, de tal forma que se distorsione la decisión de rechazar o no la hipótesis nula de igualdad de las medias de los *k* tratamientos, debido a que todas las unidades experimentales se suponen relativamente homogéneas.

Sin embargo, en las situaciones reales de cualquier experimento, puede existir efectos externos que puede aumentar la variabilidad de las observaciones y afectar los resultados del análisis, tales efectos externos son denominados **factores perturbadores**, definidos como factores del diseño que probablemente tengan efecto sobre la respuesta, pero que en ellos no existe un interés específico (Montgomery, 2001). Un factor perturbador puede **ser desconocido y no controlable**, **conocido pero no controlable**, y **conocido y controlable**; en el primero de los casos la técnica de diseño estadístico comúnmente utilizada para librarse de este tipo de factores es la *aleatorización*; en el segundo de los casos, siempre que se conozca el efecto del factor perturbador y esta pueda ser medida, la compensación de la variabilidad añadida por el mismo se puede realizar a través de una técnica de diseño estadístico llamada **análisis de covarianza**, que no será discutida en este texto. Por último, si la fuente de variabilidad extra se debe a un factor perturbador conocido y controlable, la técnica de diseño experimental utilizada para compensar esta variabilidad adicional se llama **formación de bloques**, la cual se revisará en el desarrollo de esta sección.

Para comprender lo anterior, considere el ejemplo en el que se determinan las emisiones de material particulado de una fuente industrial a través de muestreadores de alto volumen ubicados a diferentes distancias de la fuente de emisión. Evidentemente, se busca establecer si existen diferencias significativas entre las concentraciones de material particulado en función de las diferentes distancias a las que se encuentran instalados los muestreadores de la fuente de emisión, es decir, las concentraciones de particulado corresponde a la variable respuesta y las distancias de instalación de los muestreadores, corresponden a los tratamientos. Es de esperarse que a medida que los muestreadores se alejen de la fuente de emisión, las concentraciones de material particulado disminuyan, sin embargo, puede existir un efecto externo, ocasionado por la altura a la que se instalen los muestreadores o a la no uniformidad de las condiciones meteorológicas de la zona donde se realiza el experimento, lo que conllevaría a un aumento del error en las observaciones, reflejado a través

de un aumento del error experimental. De allí, que para eliminar del error experimental la variabilidad atribuida por las diferencias entre las observaciones, se podrían agrupar los muestreadores en conjuntos más homogéneos de unidades experimentales con el propósito de disminuir esta variabilidad extra, agrupando a los muestreadores por similitud de sus alturas de instalación o aquellos que se encuentren en zonas con condiciones climáticas homogéneas, sin perder de vista que el interés recae en encontrar diferencias significativas atribuidas al efecto de la distancia a la se encuentran dispuestos los muestreadores.

Este tipo de diseño mostrado en el ejemplo anterior, se denomina **diseño de bloques completos aleatorios (BCA)**, y es la técnica estadística más usada para reducir del diseño la variabilidad añadida por el efecto de factores perturbadores al error experimental. Una buena elección de criterios de formación de bloques (bloquización) es: 1) proximidad (parcelas de cultivo adyacentes), 2) características físicas (edad o peso), 3) tiempo, 4) administración de tareas en el experimento y 5) condiciones climatológicas homogéneas (Kuehl, 2001). En un diseño BCA la palabra "completos" indica que cada uno de los bloques formados contiene todos los *k* tratamientos estudiados, para el caso del ejemplo que se expuso antes, si el criterio de bloquización son las alturas de instalación de los muestreadores, cada categoría de ello, debe contener todas las distancias utilizadas para la medición del material particulado. Así mismo, el diseño BCA es aleatorio porque cada uno de los *k* tratamientos se asignan aleatoriamente dentro de cada bloque, esto último es frecuentemente considerado como una **restricción sobre la aleatorización**, es decir, que una réplica individual de cada uno de los tratamientos se asigna a cada bloque, de manera que cada bloque contenga exactamente una réplica de todos los tratamientos considerados en el experimento. Una forma útil de representar el diseño BCA se muestra en la Figura 7.3.

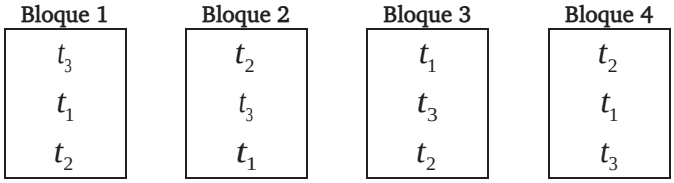


Figura 7.3. Diseño de bloques completos aleatorios.

Al generalizar nuestro diseño a *k* tratamientos asignados a *b* bloques, los datos se pueden resumir en un formato tabular *k* x *b*, como el que se muestra en la Tabla 7.1. Al igual que el diseño completamente al azar,

revisado en secciones anteriores, el diseño BCA para su aplicación requiere que las observaciones cumplan con ciertos requisitos, a saber: 1) que las muestras de donde se obtiene las observaciones para cada tratamiento y cada bloque sean independientes, 2) que la respuesta al *i*-ésimo tratamiento en el *j*-ésimo bloque provenga de una población distribuida normalmente, 3) que exista homocedasticidad en las *kb* poblaciones, y 4) que no exista interacción entre los bloques y los tratamientos, esto último, será tema de discusión en secciones siguientes.

Tabla 7.3. Arreglo *k* x *b* para un diseño de bloques completos aleatorios.

	Bloque							
Tratamiento	**1**	**2**	⋯	***j***	⋯	***b***	**Total**	**Media**
1	y_{11}	y_{12}	⋯	y_{1j}	⋯	y_{1b}	$T_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{12}	⋯	y_{2j}	⋯	y_{2b}	$T_{2\cdot}$	$\bar{y}_{2\cdot}$
⋮	⋮	⋮		⋮		⋮	⋮	⋮
i	y_{i1}	y_{i2}	⋯	y_{ij}	⋯	y_{ib}	$T_{i\cdot}$	$\bar{y}_{i\cdot}$
⋮	⋮	⋮	⋯		⋯	⋮	⋮	⋮
k	y_{k1}	y_{k2}	⋯	y_{kj}	⋯	y_{kb}	$T_{k\cdot}$	$\bar{y}_{k\cdot}$
Total	$T_{\cdot 1}$	$T_{\cdot 2}$	⋯	$T_{\cdot i}$	⋯	$T_{\cdot b}$	$T_{\cdot\cdot}$	
Media	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	⋯	$\bar{y}_{\cdot j}$	⋯	$\bar{y}_{\cdot b}$		$\bar{y}_{\cdot\cdot}$

De la Tabla 7.3, se definen los siguientes términos:

$T_{i\cdot}$ = suma de las observaciones para el *i*-ésimo tratamiento

$T_{\cdot i}$ = suma de las observaciones para el *j*-ésimo bloque

$T_{\cdot\cdot}$ = suma de todas las *bk* observaciones

$\bar{y}_{i\cdot}$ = medias de las observaciones para el *i*-ésimo tratamiento

$\bar{y}_{\cdot j}$ = media de las observaciones para el j-ésimo bloque

$\bar{y}_{\cdot\cdot}$ = media de todas las *bk* observaciones.

Para el modelado estadístico del diseño BCA, puesto que el interés se centra en determinar la existencia de diferencias significativas entre las medias de los *k* tratamientos y la formación de bloques solo constituye un mecanismo para reducir el efecto de factores perturbadores del diseño sobre el error experimental, el objetivo del diseño está orientado en probar las siguientes hipótesis

$$H_0 : \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{k\cdot}$$

H_1: al menos dos medias son diferentes.

Luego, para establecer un procedimiento de prueba, la variabilidad total de las observaciones es particionada en componentes, igual a como se procede en el diseño completamente al azar, una componente que expresa la variabilidad debida a los tratamientos, otra debida al error experimental y una tercera atribuida al efecto que ejercen los bloques sobre la variable respuesta. Lo anterior puede ser expresado a través de la siguiente identidad

$$SST = SSA + SSB + SSE$$

donde

$SST = \sum_{i=1}^{k}\sum_{j=1}^{b}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2 =$ Suma total de cuadrados

$SSA = b\sum_{i=1}^{k}\left(y_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2 =$ Suma de cuadrados de los tratamientos

$SSB = k\sum_{j=i}^{b}\left(y_{\cdot j} - \bar{y}_{\cdot\cdot}\right)^2 =$ Suma de cuadrados de los bloques

$SSE = \sum_{i=1}^{k}\sum_{j=1}^{b}\left(y_{ij} - y_{i\cdot} - y_{\cdot j} - \bar{y}_{\cdot\cdot}\right)^2 =$ Suma de cuadrados de los errores.

Sin embargo, en la práctica para agilizar los cálculos se hace uso de expresiones alternativas que son algebraicamente equivalentes a las anteriores, pero que simplifican enormemente los cálculos, las cuales se presentan a continuación

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{b} y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{bk}$$

$$SSA = \frac{\sum_{i=1}^{k} T_{i\cdot}^2}{b} - \frac{T_{\cdot\cdot}^2}{bk}$$

$$SSB = \frac{\sum_{j=1}^{b} T_{\cdot j}^2}{k} - \frac{T_{\cdot\cdot}^2}{bk}$$

$$SSE = SST - SSA - SSB$$

Posteriormente, se calculan estimaciones de σ^2, para los tratamientos, bloques y error, a través de sus medias cuadráticas s_1^2, s_2^2, y s^2, respectivamente, a través de las siguientes expresiones

$$s_1^2 = \frac{SSA}{k-1}$$

$$s_2^2 = \frac{SSB}{b-1}$$

$$s^2 = \frac{SSE}{(b-1)(k-1)}$$

El estadístico de prueba, a través del cual se decide rechazar o no la hipótesis nula se basa en la distribución *F* de Fisher-Snedecor con $k - 1$ y $(b - 1)(k - 1)$ grados de libertad, definido por la razón

$$f_1 = \frac{s_1^2}{s^2}$$

y la hipótesis nula H_0 se rechaza con un nivel de significancia α cuando $f_1 > f_{\alpha[k-1,(b-1)(k-1)]}$

También podría haber interés en comparar las medias de los bloques porque, en caso de que la diferencia entre las medias no sea considerable, quizá no sea necesaria la formación de bloques en experimentos futuros. En tal caso, se tendría interés en probar las siguientes hipótesis

$$H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot b}$$

H_1: Al menos dos medias entre los bloques son diferentes.

El estadístico de prueba, a través del cual se decide rechazar o no la hipótesis nula anterior, se basa en la distribución *F* de Fisher-Snedecor con $b - 1$ y $(b - 1)(k - 1)$ grados de libertad, definido por la razón

$$f_2 = \frac{s_2^2}{s^2}$$

y la hipótesis nula H_0 se rechaza con un nivel de significancia α cuando $f_2 > f_{\alpha[b-1,(b-1)(k-1)]}$

Sin embargo, la aleatorización sólo se ha aplicado a los tratamientos dentro de los bloques, es decir, los bloques representan una restricción sobre la aleatorización, y debido a que con frecuencia el supuesto de normalidad es

cuestionable, considerar f_2 como una prueba F exacta para la igualdad de las medias de los bloques no es una buena práctica general. No obstante, como un procedimiento aproximado para investigar el efecto de la variable formación de bloques, el estadístico f_2 es muy razonable. Si este estadístico es muy grande, implica que el factor de formación de bloques tiene un efecto considerable y que la reducción del ruido obtenida por la formación de bloques probablemente fue útil para mejorar la precisión de la comparación de las medias de los tratamientos (Montgomery, 2001).

Finalmente los resultados del análisis de varianza para el diseño BCA, se resumen en una tabla ANOVA como la mostrada en la Tabla 7.4.

Tabla 7.4. Análisis de varianza para el diseño de bloques completos aleatorios.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	f
Tratamientos	SSA	$k-1$	$s_1^2 = \frac{SSA}{k-1}$	$f_1 = \frac{s_1^2}{s^2}$
Bloques	SSB	$b-1$	$s_2^2 = \frac{SSB}{b-1}$	$f_2 = \frac{s_2^2}{s^2}$
Error	SSE	$(b-1)(k-1)$	$s^2 = \frac{SSE}{(b-1)(k-1)}$	
Total	SST	$bk-1$	-	

En R, la aplicación del ANOVA para un diseño de bloques completos aleatorios se realiza como en el caso del diseño completamente al azar a través de la función ***aov*** del paquete base, solo basta con adicionar en los argumentos el factor de bloqueo, como se muestra en la siguiente línea de código

```
aov(y~factor + bloque)
```

7.5.1. Interacción entre bloques y tratamientos

Cuando se revisaron los requisitos que deben cumplir las observaciones para poder aplicar el diseño BCA, se mencionó que un requisito de indispensable cumplimiento es que los efectos de los tratamientos y los bloques no interactúen, o lo que es los mismo que estos sean aditivos. Esto es, que las diferencias entre las medias poblacionales de dos bloques sea la misma para cada tratamiento, y la diferencia entre las medias poblacionales para dos tratamientos, sea la misma para cada bloque (Walpole *et al.*, 2008). Lo anterior se puede evaluar a través de un análisis

gráfico de las medias poblacionales de cada tratamiento en cada uno de los bloques, si las líneas resultantes son paralelas, se dice que los efectos de los tratamientos y los bloques son aditivos o no interactúan (Figura 7.4 a), en caso contrario, si las líneas se cruzan entre sí, se dice que hay interacción entre dichos efectos o no existe aditividad (Figura 7.4 b).

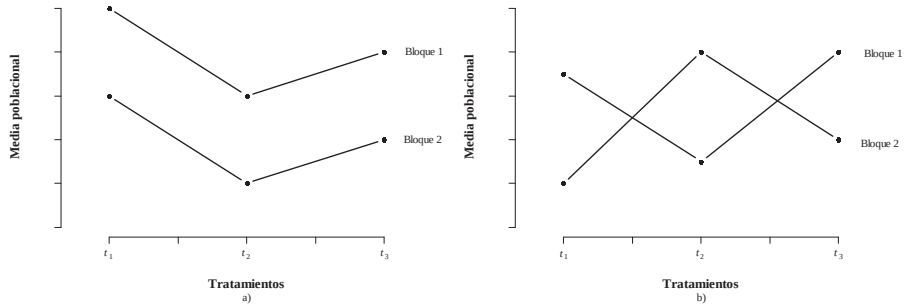


Figura 7.4. Medias poblacionales para a) efectos aditivos y b) efectos que interactúan.

Los gráficos de interacción, se construyen en R a través de la función ***interaction.plot***, donde se debe especificar la siguiente secuencia de argumentos

```
interaction.plot(factor, bloques, y, fun = mean,type = c("l",
"p", "b", "o", "c"), legend = TRUE, trace.label =
deparse(substitute (trace.factor)),...)
```

Donde ***factor***, ***bloques*** y ***y***, es evidente lo que indican, ***fun=mean***, especifica que el gráfico se construya en función de las medias de los tratamientos, ***type***, indica con qué tipo de trazo se construye el gráfico, ***legend***, establece si se muestra o no una leyenda para cada uno de los trazos de los bloques, ***trace.label***, especifica las etiquetas de los trazos, lo puntos suspensivos indican que se pueden añadir otros argumentos válidos para la construcción de gráficos en R. se recomienda consultar la ayuda de R para estudiar los demás argumentos.

7.5.2. Comparaciones múltiples para el diseño de bloques completos aleatorios

Si el análisis de varianza en el diseño BCA, indica diferencias significativas en las medias de los tratamientos, es razonable que el experimentador se interese en realizar comparaciones múltiples para descubrir cuales son los tratamientos cuyas medias difieren. Para ello puede utilizarse cualquiera de los procedimientos de comparaciones multiples revisados en la sección 7.4. Simplemente se sustituye en cada una de las ecuaciones para el cálculo

de los criterios de prueba, el número de replicas (n) usado en las pruebas del ANOVA de un factor por el número de bloques (b). Así mismo, se deben modificar el número de grados de libertad para la media cuadrática del error, que para el diseño BCA es $[(b-1)(k-1)]$ (Montgomery, 2001).

Ejemplo 7.11. La Universidad de La Guajira a través del Grupo de Investigación Pichihuel en asociación con el Instituto de Estudios Ambientales y Aprovechamiento de Agua – INESAG, llevaron a cabo monitoreos de algunas variables ambientales en las playas del municipio de Riohacha, en el marco del programa de Calidad Ambiental de Playas Turísticas – CAPT, en cuatro diferentes estaciones de monitoreo. Dentro de sus variables de estudio se encuentran las concentraciones de nitritos (NO_2). A continuación se muestran las concentraciones de NO_2 en mg/L en las diferentes estaciones, obtenidas durante los meses de junio a julio del 2013. Se desea determinar: (a) si existen diferencias significativas en la concentración media de NO_2 en las diferentes estaciones de monitoreo a través de un diseño de bloques completos aleatorios para compensar los efectos que pueden tener los meses de monitoreo en los datos de concentraciones de NO_2; (b) en caso de existir diferencias significativas determinar ¿cuáles estaciones de monitoreo presentan dichas diferencias? Asúmase que los datos de densidad de enterococos se encuentra normalmente distribuida en cada una de las estaciones de monitoreo con varianzas homogéneas.

	Estaciones de Monitoreo							
Meses	**Jun**	**Jul**	**Ago**	**Sep**	**Oct**	**Nov**	**Total**	**Media**
RT1	1.90	1.47	2.57	7.17	9.60	2.26	24.97	4.16
RT2	1.27	1.30	2.97	0.93	5.60	2.21	14.28	2.38
RT3	3.73	3.50	3.90	2.44	7.00	5.20	25.77	4.30
RT4	0.63	1.50	2.73	0.80	5.00	2.84	13.50	2.25
Total	7.53	7.77	12.17	11.34	27.20	12.51	78.52	
Media	1.88	1.94	3.04	2.84	6.80	3.13		13.09

Solución

Para determinar la existencia de diferencias significativas en las concentraciones medias de NO_2, el interés de este experimento se centra en probar las siguientes hipótesis

$$H_0 : \mu_{RT1} = \mu_{RT2} = \mu_{RT3} = \mu_{RT4}$$

H_1 : al menos dos medias son diferentes.

De esta manera iniciamos el desarrollo del procedimiento de prueba calculando las sumas de cuadrados

$$SST = 1.90^2 + 1.47^2 + \cdots + 2.84^2 - \frac{78.52^2}{(6)(4)}$$

$$SST = 120.566$$

$$SSA = \frac{24.97^2 + 14.28^2 + 25.77^2 + 13.50^2}{6} - \frac{65.42^2}{(6)(4)}$$

$$SSA = 22.069$$

$$SSB = \frac{7.53^2 + 5.77^2 + \cdots + 12.51^2}{4} - \frac{65.42^2}{(6)(4)}$$

$$SSB = 65.638$$

$$SSE = 120.566 - 22.069 - 65.638$$

$$SSE = 32.858$$

Luego se realizan las estimaciones de las medias cuadráticas

$$s_1^2 = \frac{22.069}{3} = 7.356$$

$$s_2^2 = \frac{65.638}{5} = 13.128$$

$$s^2 = \frac{32.858}{(3)(5)} = 2.191$$

Ahora, calculamos el valor de estadístico de prueba, con el cual basaremos nuestra decisión de rechazar no la hipótesis nula

$$f_1 = \frac{7.356}{2.191} = 3.357$$

Según la Tabla A.5 el valor crítico de la distribución F de Fisher-Snedecor es $f_{0.05[3,15]} = 3.29$.

Como $f = 3.357 > f_{0.05[3,15]} = 3.29$, se rechaza la hipótesis nula de igualdad de las medias de nitritos a un nivel de significancia de 0.05, llegando a la conclusión de que si existen diferencias significativas entre las medias de las concentraciones de nitritos en las cuatro estaciones de monitoreo de las playas del municipio de Riohacha, La Guajira, con una confiabilidad del 95%.

Para evaluar el efecto de los meses de monitoreo (bloques), se calcula el valor de $f_2 = 13.128/2.191 = 5.992$, y dado que es mayor que $f_{0.05[5,15]} = 2.90$, muestra evidencias de que los meses de monitoreo, ejercer un efecto considerable sobre la respuesta de las concentraciones de NO_2, por lo que la realización del bloqueo probablemente resulta ser una buena estrategia para reducir este efecto. La tabla ANOVA para este experimento se muestra a continuación

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*f*
Estaciones	22.069	3	7.356	3.357
Meses	65.638	5	13.128	5.992
Error	32.858	15	2.191	
Total	120.566	23	-	-

Ahora, dado a que se hallaron diferencias significativas entre las concentraciones medias de NO_2 en las cuatro diferentes estaciones de monitoreo, nos interesa saber en qué estaciones se presentan dichas diferencias. Para ello aplicaremos el test de comparaciones múltiples *LSD* de Fisher, cuyo valor del criterio de prueba con $t_{0.25,15} = 2.131$, es

$$LSD = 2.131\sqrt{\frac{(2)(2.191)}{6}} \therefore LSD = 1.821$$

De esta forma los resultados del examen de diferencias entre las concentraciones medias de NO2 entre las estaciones de monitoreo sería

$\left|\overline{y}_{RT_1} - \overline{y}_{RT_2}\right| = 1.782 < 1.821$, no existen diferencias significativas
$\left|\overline{y}_{RT_1} - \overline{y}_{RT_3}\right| = 0.133 < 1.821$, no existen diferencias significativas
$\left|\overline{y}_{RT_1} - \overline{y}_{RT_4}\right| = 1.912 > 1.821$, existen diferencias significativas
$\left|\overline{y}_{RT_2} - \overline{y}_{RT_3}\right| = 1.915 > 1.821$, existen diferencias significativas
$\left|\overline{y}_{RT_2} - \overline{y}_{RT_4}\right| = 0.130 < 1.821$, no existen diferencias significativas
$\left|\overline{y}_{RT_3} - \overline{y}_{RT_4}\right| = 2.045 > 1.821$, existen diferencias significativas

Con base en lo anterior, vemos que existen diferencias significativas en las concentraciones medias NO_2, para los pares de estaciones $\overline{y}_{RT_1}$ *vs* $\overline{y}_{RT_4}$, $\overline{y}_{RT_2}$ *vs* $\overline{y}_{RT_3}$, y $\overline{y}_{RT_3}$ *vs* $\overline{y}_{RT_4}$, a un nivel de significancia de 0.05. Un examen gráfico de estos resultados se puede realizar a través de la construcción de un gráfico de cajas como el que se muestra en la Figura 7.5.

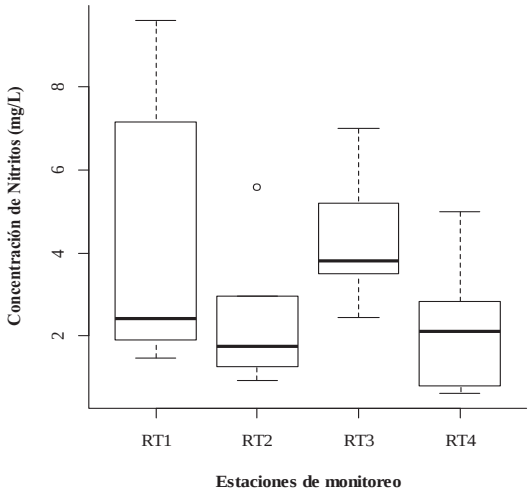


Figura 7.5. Concentración de NO_2 en las diferentes estaciones de monitoreo.

A continuación mostraremos las salidas de resultados de R, para la aplicación de un análisis de varianza para el diseño de bloques completos aleatorios. Iniciaremos con la construcción de los vectores de datos para las concentraciones de NO_2, y los factores que corresponden a las estaciones (tratamientos) y los meses de monitoreo (bloques).

```
> RT1<-c(1.90,1.47,2.57,7.17,9.60,2.26)
> RT2<-c(1.27,1.30,2.97,0.93,5.60,2.21)
> RT3<-c(3.73,3.50,3.90,2.44,7.00,5.20)
> RT4<-c(0.63,1.50,2.73,0.80,5.00,2.84)
> NO2<-c(RT1,RT2,RT3,RT4)
> Estación<-rep(1:4,each=6)
> Estación<-factor(Estación,labels=c("RT1","RT2","RT3","RT4"))
> Meses<-rep(1:6,each=1,times=4)
> Meses<-factor(Meses,labels=c("Jun","Jul","Ago","Sep","Oct",
"Nov"))
```

Luego realizamos la evaluación del cumplimiento del supuesto de normalidad, homogeneidad de varianzas y no interacción de los tratamientos y los bloques

```
> by(NO2,Estación,shapiro.test)
Estación: RT1

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.7937, p-value = 0.05153

------------------------------------------------------------
Estación: RT2

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.8313, p-value = 0.1103

------------------------------------------------------------
Estación: RT3

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9244, p-value = 0.5373

------------------------------------------------------------
Estación: RT4

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9074, p-value = 0.4196

> bartlett.test(NO2,Estación)

        Bartlett test of homogeneity of variances

data:  NO2 and Estación
Bartlett's K-squared = 4.1873, df = 3, p-value = 0.2419
```

Un examen de los p-valores para los test de normalidad y homogeneidad de varianzas, indica que las concentraciones de NO_2 siguen una distribución normal en todas las estaciones de monitoreo, con varianzas homogéneas (p-valor > 0.05).

El examen de interacción entre los tratamientos en los bloques se realiza a través de un gráfico de interacción (Figura 7.6) construido con la función ***interaction.plot***, como se muestra a continuación

```
interaction.plot(Estación,Meses,NO2,fun=mean,legend=TRUE,col=1
:6,xlab="Estaciones monitoreo",ylab="Concentración media de
nitritos (mg/L)",lwd=2,fixed=TRUE)
```

El gráfico muestra claramente cruces entre las líneas que representan a los meses de monitoreo, por lo tanto es evidente que existe fuerte interacción entre las estaciones y lo meses de monitoreo, por lo tanto, ante el incumplimiento de este requisito, la formación de bloques no resulta ser la mejor alternativa para compensar el efecto que ejercen los meses de monitoreo sobre las respuestas de la concentración de NO_2. Sin embargo para efectos prácticos continuaremos con el desarrollo de este análisis.

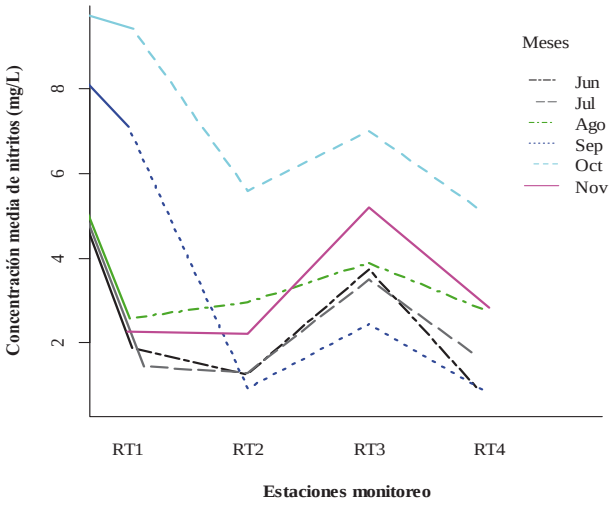


Figura 7.6. Gráfico de interacción entre las estaciones y meses de monitoreo.

Como paso siguiente, realizamos la construcción de la tabla ANOVA a través de la función ***aov***

```
> Anova<-aov(NO2~Estación+Meses)
> summary(Anova)
            Df Sum Sq Mean Sq F value  Pr(>F)
Estación     3  22.07   7.356   3.358 0.04714 *
Meses        5  65.64  13.128   5.993 0.00306 **
Residuals   15  32.86   2.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Los resultados del ANOVA muestran que existen diferencias significativas entre las concentraciones medias de NO_2 (p-valor < 0.05) en las cuatro estaciones de monitoreo, a un nivel de significancia de 0.05. Así mismo, revela que los meses de monitoreo tienen efecto en la variabilidad de los datos de concentración de NO_2 (p-valor < 0.05).

Por último, ante la consecuente existencia de diferencias significativas entre las medias de las concentraciones de NO_2 en las diferentes estaciones de monitoreo, se aplicará el test *LSD* de Fisher para determinar en cuales estaciones se presentan las diferencias. Recuérdese que la aplicación de este test en el entorno de R se realiza a través de la función ***LSD.test*** del paquete "agricolae" (de Mendiburu, 2015).

```
> library(agricolae)
>
LSD.test(NO2,Estación,DFerror=15,MSerror=2.191,alpha=0.05,p.adj=
"none",group=FALSE,console=TRUE)

Study: NO2 ~ Estación

LSD t Test for NO2

Mean Square Error:  2.191

Estación,  means and individual (95 %) CI

         NO2      std r       LCL      UCL  Min Max
RT1 4.161667 3.380399 6 2.8736526 5.449681 1.47 9.6
RT2 2.380000 1.746219 6 1.0919859 3.668014 0.93 5.6
RT3 4.295000 1.592755 6 3.0069859 5.583014 2.44 7.0
RT4 2.250000 1.638926 6 0.9619859 3.538014 0.63 5.0
alpha: 0.05 ; Df Error: 15
Critical Value of t: 2.13145

Comparison between treatments means

          Difference     pvalue sig.         LCL         UCL
RT1 - RT2  1.7816667 0.05459195    . -0.03986035  3.60319369
RT1 - RT3 -0.1333333 0.87809821      -1.95486035  1.68819369
RT1 - RT4  1.9116667 0.04090161    *  0.09013965  3.73319369
RT2 - RT3 -1.9150000 0.04059672    * -3.73652702 -0.09347298
RT2 - RT4  0.1300000 0.88112046      -1.69152702  1.95152702
RT3 - RT4  2.0450000 0.03023887    *  0.22347298  3.86652702
```

Los resultados del test *LSD* de Fisher muestran que existen diferencias significativas en las concentraciones medias NO_2 (p-valor < 0.05), para los pares de estaciones $\bar{y}_{RT_1}$ *vs* $\bar{y}_{RT_4}$, $\bar{y}_{RT_2}$ *vs* $\bar{y}_{RT_3}$, y $\bar{y}_{RT_3}$ *vs* $\bar{y}_{RT_4}$, a un nivel de significancia de 0.05.

7.6. Análisis de varianza de dos factores para diseños completamente aleatorios

En las secciones anteriores hemos revisado el uso del análisis de varianza para diseños experimentales completamente aleatorios y con formación de bloques completos aleatorios, donde se estudió el efecto de un solo factor sobre la variable respuesta del experimento. Sin embargo, existen otras situaciones en donde el experimento se lleva a cabo con la evaluación simultánea de dos factores en sus diferentes niveles sobre el comportamiento de la variable respuesta. A este tipo de situaciones se les denomina experimentos de dos factores.

Por ejemplo, se puede evaluar el comportamiento de la riqueza de especies de macroinvertebrados acuáticos en diferentes estaciones de muestreo y diferentes tipos de hábitats: bentos, neuston y necton. Aquí, se busca evaluar el efecto de los dos factores del diseño (Estaciones de monitoreo y hábitats), sobre la variable respuesta, riqueza de especies de macroinvertebrados acuáticos, de esta forma se analiza aleatoriamente todas las combinaciones posibles que pueden formarse con los niveles de los dos factores a investigar, con el fin de determinar cuáles de estas combinaciones tiene mayor efecto sobre la variable respuesta.

Un aspecto importante a tener en cuenta cuando nos enfrentamos a un diseño experimental completamente aleatorio de dos factores, es evaluar no solo el **efecto principal** de cada uno de los factores, es decir, el cambio en la media de la variable respuesta que se debe a la acción individual de cada factor, por un cambio de nivel en este último; sino evaluar también los **efectos de interacción** entre los factores, es decir, aquella situación que se presenta cuando el efecto de un factor depende del nivel en el que se encuentra el otro (Gutiérrez & de la Vara 2008; Ferrer, 2007; Conavos, 1988). Este efecto de interacción, ya se había revisado anteriormente cuando se trataron los diseños de bloques completos aleatorios, cuyo análisis se realizó a través de un gráfico de interacción.

No obstante, a pesar de la utilidad de los gráficos de interacción, este tipo de análisis es de naturaleza descriptiva y se sustenta en estimaciones muéstrales, por lo que es necesario establecer un procedimiento de prueba que nos permita saber si los efectos de interacción son estadísticamente significativos a nivel poblacional. Allí, es donde toma importancia el análisis de varianza (ANOVA) de dos factores.

Cuando la interacción entre los factores es probada, el análisis siguiente es la evaluación de los efectos principales de cada factor (independien-

temente de si estos resultan significativos o no), y determinar si el efecto de cada factor sobre la respuesta es **positivo**, cuando ocurre que la respuesta se incrementa conforme los niveles de un factor aumentan de acuerdo a un orden definido, para un nivel fijo del otro factor; o si el efecto es **negativo**, cuando ocurre una disminución de la respuesta a niveles crecientes del factor, para un nivel fijo del otro (Walpole *et* al., 2007).

Otro aspecto a tener en cuenta en el diseño de experimentos de dos factores es que la variabilidad atribuida a la interacción y el error experimental solo se separan si se hacen observaciones múltiples con las distintas combinaciones de tratamiento, es decir, si se utilizan réplicas en cada una de las combinaciones de tratamientos, teniendo en cuenta que se obtengan el mismo número de réplicas por combinación.

7.6.1. Procedimiento de prueba del ANOVA de dos factores

Antes de presentar el desarrollo matemático del ANOVA de dos factores, considérese un diseño de dos factores *A* y *B* con *a* y *b* niveles de cada factor, respectivamente, de los cuales se toman *n* réplicas de las combinaciones de cada tratamiento. Las observaciones recolectadas durante el experimento pueden clasificarse, como se ha acostumbrado, en un arreglo rectangular, donde las filas correspondan a los niveles del factor *A* y las columnas correspondan a los niveles del factor *B*. Cada combinación de niveles de los factores corresponde a un tratamiento, y dentro de los *ab* tratamientos se tendrían *n* observaciones (réplicas). Con este arreglo, se denota por y_{ijk} a la *k*-ésima observación tomada en el *i*-ésimo nivel del factor *A* y el *j*-ésimo nivel del factor *B*. De esta forma, las *abn* observaciones del experimento, quedarían organizadas de manera similar a como se muestra en la Tabla 7.5.

Tabla 7.5. Experimento de dos factores con *n* réplicas.

	Factor *B*					
Factor *A*	1	2	…	b	Total	Media
1	y_{111}	y_{121}	…	y_{1b1}	$T_{1..}$	$\bar{y}_{1..}$
	y_{112}	y_{122}	…	y_{1b2}		
	⋮	⋮		⋮		
	y_{11n}	y_{12n}	…	y_{1bn}		
2	y_{211}	y_{221}	…	y_{2b1}	$T_{2..}$	$\bar{y}_{2..}$

Factor *A*	Factor *B* 1	2	…	b	Total	Media
	y_{212}	y_{222}	$\cdots$	y_{2b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{21n}	y_{22n}	$\cdots$	y_{2bn}		
$\vdots$	$\vdots$	$\vdots$		$\vdots$		
	y_{a11}	y_{a21}	$\cdots$	y_{ab1}		
	y_{a12}	y_{a22}	$\cdots$	y_{ab2}	$T_{a\cdot\cdot}$	$\bar{y}_{a\cdot\cdot}$
a	$\vdots$	$\vdots$		$\vdots$		
	y_{a1n}	y_{a2n}	$\cdots$	y_{abn}		
Total	$T_{\cdot 1\cdot}$	$T_{\cdot 2\cdot}$	$\cdots$	$T_{\cdot b\cdot}$	$T_{\cdots}$	
Media	$\bar{y}_{\cdot 1\cdot}$	$\bar{y}_{\cdot 2\cdot}$	$\cdots$	$\bar{y}_{\cdot b\cdot}$		$\bar{y}_{\cdots}$

Igual que los diseños presentados en secciones anteriores requerían del cumplimiento de ciertas condiciones, el diseño completamente aleatorio de dos factores, requiere que las observaciones en la celda *ij*-ésima constituya una muestra aleatoria de tamaño *n* de una población que se supone tiene una distribución normal e igual varianza que las demás celdas (Walpole *et al.*, 2007).

De la Tabla 7.5, como en secciones anteriores, es importante definir los siguientes términos:

$T_{ij\cdot}$ = suma de las observaciones en la *ij*-ésima celda,

$T_{i\cdot\cdot}$ = suma de las observaciones para el *i*-ésimo nivel del factor *A*,

$T_{\cdot j\cdot}$ = suma de las observaciones para el *j*-ésimo nivel del factor *B*,

$T_{\cdots}$ = suma de todas las *abn* observaciones,

$\bar{y}_{ij\cdot}$ = media de las observaciones en la *ij*-ésima celda,

$\bar{y}_{i\cdot\cdot}$ = media de las observaciones para el *i*-ésimo nivel del factor *A*,

$\bar{y}_{\cdot j\cdot}$ = media de las observaciones para el *j*-ésimo nivel del factor *B*,

$\bar{y}_{\cdots}$ = media de todas las *abn* observaciones.

Para el modelado estadístico del diseño de dos factores, puesto que el interés se centra en determinar los efectos principales ejercidos por los factores *A* y *B*, así como la interacción entre ellos, el objetivo del diseño está orientado en probar las siguientes hipótesis

$$H_0^{'}: \mu_{1..} = \mu_{2..} = \cdots = \mu_{a..}$$

$H_1^{'}$: al menos dos medias del factor *A* son diferentes.

$$H_0^{''}: \mu_{.1.} = \mu_{.2.} = \cdots = \mu_{.b.}$$

$H_1^{''}$: Al menos dos medias del factor *B* son diferentes.

$H_0^{'''}$: no existe interacción entre los tratamientos *A* y *B*

$H_0^{'''}$: existe interacción entre los tratamientos *A* y *B*.

En adelante el efecto global de la interacción entre dos factores *A* y *B*, se denotara como *A:B*.

Luego, para establecer un procedimiento de prueba, la variabilidad total de las observaciones es particionada en componentes, igual como se ha discutido en secciones anteriores. Para el caso del diseño de dos factores, una componente expresa la variabilidad debida a los efectos del factor *A*, otra debida al efecto del factor *B*, una tercera atribuida a la interacción entre los factores *A:B*, y por último, una componente que expresa la variabilidad debida al error experimental. Lo anterior puede ser expresado a través de la siguiente identidad

$$SST = SSA + SSB + SS(AB) + SSE$$

donde

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \overline{y}_{...}\right)^2 = \text{ suma total de cuadrados}$$

$$SSA = bn\sum_{i=1}^{a}\left(y_{i..} - \overline{y}_{...}\right)^2 = \text{ suma de cuadrados para el efecto del factor } A$$

$SSB = an\sum_{j=i}^{b}\left(y_{\cdot j\cdot} - \bar{y}_{\cdots}\right)^2 =$ suma de cuadrados para el efecto del factor *B*

$SS(AB) = n\sum_{i=1}^{a}\sum_{i=1}^{b}\left(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}\right)^2 =$ suma de cuadrados de la interacción *A:B*

$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \bar{y}_{ij\cdot}\right)^2 =$ suma de cuadrados de los errores.

Aquí, también existen expresiones alternativas que ayudan a simplificar los cálculos de las sumas de cuadrados, basadas en la notación de puntos de la Tabla 7.5.

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{T_{\cdots}^2}{abn}$$

$$SSA = \frac{\sum_{i=1}^{a} T_{i\cdot\cdot}^2}{bn} - \frac{T_{\cdots}^2}{abn}$$

$$SSB = \frac{\sum_{j=1}^{b} T_{\cdot j\cdot}^2}{an} - \frac{T_{\cdots}^2}{abn}$$

$$SS(AB) = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{T_{ij\cdot}^2}{n} - \frac{T_{\cdots}^2}{abn} - SSA - SSB$$

$$SSE = SST - SSA - SSB - SS(AB)$$

Posteriormente, se calculan estimaciones de σ^2, para los efectos del factor *A*, factor *B*, interacción *A:B* y el error, a través de sus medias cuadráticas s_1^2, s_2^2, s_3^2 y s^2, respectivamente, a través de las siguientes expresiones

$$s_1^2 = \frac{SSA}{a-1}$$

$$s_2^2 = \frac{SSB}{b-1}$$

$$s_3^2 = \frac{SS(AB)}{(a-1)(b-1)}$$

$$s^2 = \frac{SSE}{ab(n-1)}$$

Para probar la hipótesis nula H_0', de no existencia de efectos significativos del factor *A*, basamos nuestra decisión en el cálculo de un estadístico de prueba que sigue una distribución *F* de Fisher-Snedecor con $a - 1$ y $(ab)(n - 1)$ grados de libertad, definido por la razón

$$f_1 = \frac{s_1^2}{s^2}$$

y la hipótesis nula H_0' se rechaza con un nivel de significancia α cuando $f_1 > f_{\alpha[a-1,\, ab(n-1)]}$

Asimismo, para probar la hipótesis nula H_0'', de no existencia de efectos significativos del factor *B*, basamos nuestra decisión en el cálculo de un estadístico de prueba que sigue una distribución *F* de Fisher-Snedecor con $b - 1$ y $(ab)(n - 1)$ grados de libertad, definido por la razón

$$f_2 = \frac{s_1^2}{s^2}$$

y la hipótesis nula H_0'' se rechaza con un nivel de significancia α cuando $f_2 > f_{\alpha[b-1,\, ab(n-1)]}$.

Por último, para probar la hipótesis nula H_0''', de no existencia de interacción significativa entre los factores *A* y *B*, también basamos nuestra decisión en el cálculo de un estadístico de prueba que sigue una distribución *F* de Fisher-Snedecor con $(a - 1)(b - 1)$ y $(ab)(n - 1)$ grados de libertad, definido por la razón

$$f_3 = \frac{s_3^2}{s^2}$$

y la hipótesis nula H_0''' se rechaza con un nivel de significancia α cuando $f_3 > f_{\alpha[(a-1)(b-1),\, ab(n-1)]}$.

Los resultados del análisis de varianza de dos factores con *n* réplicas, se resumen en una tabla ANOVA como la mostrada en la Tabla 7.6.

Tabla 7.6. Análisis de varianza para el diseño de dos factores con *n* réplicas.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*f*
Factor *A*	SSA	$a-1$	$s_1^2 = \frac{SSA}{a-1}$	$f_1 = \frac{s_1^2}{s^2}$
Factor *B*	SSB	$b-1$	$s_2^2 = \frac{SSB}{b-1}$	$f_2 = \frac{s_2^2}{s^2}$
Interacción *A:B*	SS(AB)	$(a-1)(b-1)$	$s_3^2 = \frac{SS(AB)}{(a-1)(b-1)}$	$f_3 = \frac{s_3^2}{s^2}$
Error	SSE	$ab(n-1)$	$s^2 = \frac{SSE}{ab(n-1)}$	
Total	SST	$abn-1$	-	

En R, la aplicación del ANOVA para un diseño completamente aleatorio de dos factores se realiza de forma similar a los diseños vistos anteriormente, solo basta con seguir cualquiera de las dos siguientes líneas de código

```
aov(y ~ factor A + factor A + factor A: factor B)
aov(y ~ factor A*factor B)
```

7.6.2. Comparaciones múltiples

En la sección 7.4 se introdujeron los test de comparaciones múltiple (pruebas post hoc), utilizados cuando el ANOVA resulta significativo para identificar entre que pares de medias de cada tratamiento se producen las diferencias. Estos procedimientos también son de utilidad cuando nos enfrentamos a diseños experimentales de dos factores. Sin embargo, cuando en el diseño el efecto de interacción entre los factores es significativo, las comparaciones entre las medias de uno de los factores (por ejemplo, el factor *A*) pueden ser oscurecidas por los efectos de la interacción *A:B,* por lo que los resultados del análisis pueden ser engañosos (Montgomery, 2001; Gutiérrez & de la Vara 2008).

Cuando los efectos de interacción son significativos, una forma de solucionar la limitación de la aplicación de los test de comparaciones múltiples consiste en fijar el factor *B* en un nivel específico y aplicar cualquiera de los procedimientos estudiados a las medias del factor *A* con ese nivel. De otro modo, cuando los efectos de la interacción *A:B* no son

significativos, se pueden calcular las diferencias entre las medias del factor *A* y del factor *B* de la forma en que tradicionalmente lo hemos hecho.

Un aspecto importante a tener en cuenta al realizar comparaciones múltiples en un diseño de dos factores, es que si las comparaciones se llevaran a cabo sin tener en cuenta los efectos de interacción, el cálculo del criterio de prueba del test elegido se debe realizar teniendo en cuenta el número total de observaciones de cada nivel del factor considerado, por ejemplo, si deseáramos realizar las comparaciones entre las medias del factor *A*, a través del test *LSD* de Fisher, su expresión de cálculo sería

$$LSD_A = t_{\alpha/2,ab(n-1)}\sqrt{\frac{2SSE}{n_A}}$$

donde n_A, es el número de observaciones en cada nivel del factor *A*.

Por otra parte, para realizar comparaciones múltiples entre las medias de un factor, tomando en cuenta el efecto de interacción, el cálculo del criterio de prueba del test elegido se lleva a cabo teniendo en cuenta el número de observaciones *n* de cada celda, para el nivel en que se fijó el otro factor. Nuevamente utilizando la expresión de cálculo del test *LSD* de Fisher, esta quedaría dada por

$$LSD_{B_i(A)} = t_{\alpha/2,ab(n-1)}\sqrt{\frac{2SSE}{n}}$$

Donde B_j denota que se harán comparaciones entre las medias del factor *A*, fijando al factor *B* en su nivel *j*, y *n*, representa el número de observaciones en cada celda.

Ejemplo 7.12. La universidad de La Guajira, a través del grupo de investigación Pichihuel, realizó un estudio sobre el efecto del humus sólido y líquido de lombriz roja californiana (*Eisenia foetida*) sobre el cultivo de pepino (*Cucumis sativus L.*), donde una de las variables de interés fue el contenido de clorofila foliar, expresado en mg/g-hojas. El diseño del experimento consistió en el establecimiento de tres grupos experimentales a los que se aplicaron los diferentes tipos de biofertilizantes y una combinación de estos. Determinar si existen diferencias significativas en el contenido clorofila foliar en los diferentes grupos experimentales, bajo la aplicación de los diferentes biofertilizantes.

Biofertilizantes	Grupos experimentales			Total	Media
	G1	G2	G3		
	1.1220	0.7698	0.6657		
HS	1.0306	0.7673	0,7510	7.7398	0.8600
	1.2214	0.6447	0.7673		
	2.9099	3.6659	2.7510		
HS+HL	1.9174	2.6476	3.6447	26.0125	2.9803
	2.1521	2.5341	3.7898		
Total	10.3534	11.0294	12.3695	33.7523	
Media	1.7256	1.8382	2.0616		1.8751

Solución

Para determinar la existencia de diferencias significativas entre las medias de cada combinación de humus utilizado y las medias de cada grupo experimental, en función de la existencia o no de efectos de interacción entre estos factores, centramos nuestro interés en probar las siguientes hipótesis:

$$H_0^{'} = \mu_{HS} = \mu_{HS+HL}$$

$H_1^{'}$ = Las medias son diferentes.

$$H_0^{''} = \mu_{G1} = \mu_{G2} = \mu_{G3}$$

$H_1^{''}$ = Al menos dos medias son diferentes.

$H_0^{'''}$ = Existe interacción significativa entre el tipo de biofertilizante y los grupos experimentales.

$H_1^{'''}$ = No existe interacción significativa entre el tipo de biofertilizante y los grupos experimentales.

Luego iniciamos los cálculos respectivos, para determinar las sumas de cuadrados de cada una de las componentes en que se divide la variabilidad total del diseño

$$SST = 1.220^2 + 1.0306^2 + \cdots + 3.7898^2 - \frac{33.7523^2}{(2)(3)(3)}$$

$$SST = 22.576$$

$$SSA = \frac{7.7398^2 + 26.0125^2}{(3)(3)} - \frac{33.7523^2}{(2)(3)(3)}$$

$$SSA = 18.550$$

$$SSB = \frac{10.3534^2 + 11.0294^2 + 12.3695^2}{(2)(3)} - \frac{33.7523^2}{(2)(3)(3)}$$

$$SSB = 0.351$$

$$SS(AB) = \frac{3.3740^2 + 2.1818^2 + 2.1840^2 + 6.9794^2 + 8.8476^2 + 10.1855^2}{3}$$
$$- \frac{33.7523^2}{(2)(3)(3)} - 18.550 - 0.351$$

$$SS(AB) = 1.6931$$

$$SSE = 22.576 - 18.550 - 0.351 - 1.6931$$

$$SSE = 1.9824$$

A partir de los cálculos de las sumas de cuadrados, continuamos el análisis con el cálculo de las medias cuadráticas

$$s_1^2 = \frac{18.550}{2-1} = 18.550$$

$$s_2^2 = \frac{0.351}{3-1} = 0.176$$

$$s_3^2 = \frac{1.6931}{(2-1)(3-1)} = 0.847$$

$$s^2 = \frac{1.9824}{(2)(3)(3-1)} = 0.165$$

Calculamos el valor del estadístico de prueba para determinar la existencia de efectos significativos en cada uno de los factores estudiados, y por supuesto determinar la existencia o no de efectos de interacción

significativa entre los factores. De esta forma, para determinar la existencia de diferencias significativas el contenido medio de clorofila foliar, después de la aplicación de las dos combinaciones de humus sólido y líquido, tenemos que el valor del estadístico de prueba es

$$f_1 = \frac{18.550}{0.165} = 112.42$$

El valor crítico la distribución *F* de Fisher-Snedecor es $f_{0,05[1,12]} = 4.75$.

Del mismo modo, el valor del estadístico de prueba para encontrar diferencias entre el contenido medio de clorofila foliar en los diferentes grupos experimentales del estudio, y el valor crítico de la distribución *F* de Fisher-Snedecor, es calculado como sigue

$$f_2 = \frac{0.176}{0.165} = 1.07$$

Con un valor crítico igual a $f_{0.05[2,12]} = 3.89$.

Por ultimo calculamos el efecto de interacción entre las combinaciones de humus utilizados y los grupos experimentales del estudio

$$f_3 = \frac{0.847}{0.165} = 5.13$$

Con un valor crítico de la distribución *F* de Fisher-Snedecor igual a $f_{0.05[2,12]} = 3.89$.

A continuación mostramos el resumen del análisis a través de una tabla ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*f*
Combinaciones	18.550	1	18.550	112.42
Grupos	0.351	2	0.176	1.07
Combinaciones:Grupos	1.693	2	0.847	5.13
Error	1.982	12	0.165	
Total		17	-	

Del análisis anterior, pueden realizarse las siguientes conclusiones

- Como $f_1 = 112.42 > f_{0,05[1,12]} = 4.75$, se rechaza la hipótesis nula a favor de la alternativa a un nivel de significancia de 0.05, y se concluye que existen diferencias significativas en el contenido medio de clorofila foliar entre las combinaciones de humus utilizadas.

- Como $f_2 = 1.07 < f_{0.05[2,12]} = 3.89$, no se rechaza la hipótesis nula a un nivel de significancia de 0.05, lo que nos lleva a afirmar que no existen diferencias estadísticamente significativas en el contenido medio de clorofila foliar entre los diferentes grupos experimentales estudiados.

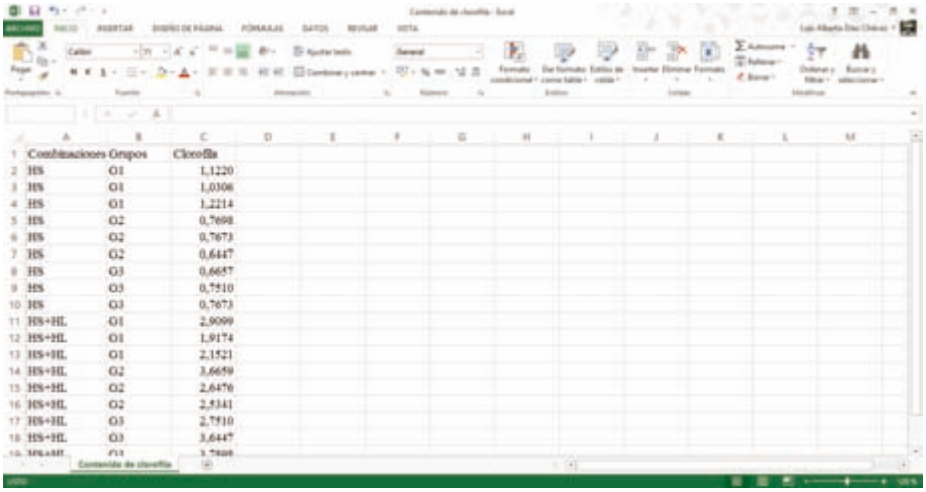
- Como $f_3 = 5.13 > f_{0,05[2,12]} = 3.89$, se rechaza la hipótesis nula a favor de la hipótesis alternativa a un nivel de significancia de 0.05, y se concluye que existen interacción significativa entre las combinaciones de humus utilizadas y los grupos experimentales con una confiabilidad del 95%. Más adelante, realizaremos la construcción del gráfico de interacción y realizaremos comentarios del mismo para una mejor interpretación de su utilidad.

Cabe destacar que a pesar de la existencia de diferencias significativas en las medias del contenido de clorofila foliar luego de la aplicación de las dos combinaciones de humus, no es necesario realizar test de comparaciones múltiples, dado que solo existen dos niveles para este factor, y es suficiente con evaluar cuál de ellos ofrece un mayor rendimiento en el cultivo en función de su contenido medio de clorofila foliar y el grupo experimental donde fue aplicado. Con la construcción del gráfico de interacción se puede llegar a estas conclusiones a través de su análisis, como veremos más adelante.

A continuación, describiremos la aplicación del ANOVA de dos factores en el entorno de programación de R, iniciando con la diagnosis del modelo, es decir, evaluando si las observaciones de cada celda provienen de distribuciones normales con varianzas homogéneas. Este análisis, lo podemos realizar en función de los grupos experimentales o a través de las combinaciones de humus utilizadas, para efectos prácticos, tomaremos la primera ruta.

Primeramente, realizamos la carga de los datos en R, en esta ocasión los datos serán tabulados en Excel y guardados bajo la extensión *.csv,* luego

se importarán en R a través de la función read.csv2 descrita en secciones anteriores. La forma de tabulación de los datos, deben seguir la estructura que se muestra en la Figura 7.7, donde cada una de la primera y segunda columnas corresponden a los dos factores del experimento, y la tercera columna contiene a la variable respuesta.



	A	B	C
1	Combinaciones	Grupos	Clorofila
2	HS	G1	1,1220
3	HS	G1	1,0308
4	HS	G1	1,2214
5	HS	G2	0,7698
6	HS	G2	0,7673
7	HS	G2	0,6447
8	HS	G3	0,6657
9	HS	G3	0,7510
10	HS	G3	0,7673
11	HS+HL	G1	2,9099
12	HS+HL	G1	1,9174
13	HS+HL	G1	2,1521
14	HS+HL	G2	3,6659
15	HS+HL	G2	2,6476
16	HS+HL	G2	2,5341
17	HS+HL	G3	2,7510
18	HS+HL	G3	3,6447

Figura 7.7. Tabulación de los datos en Excel.

```
> Clorofila.foliar<-read.csv2("Contenido de
clorofila.csv",header= TRUE,encoding="latin1")
> attach(Clorofila.foliar)
> by(Clorofila,Grupos,shapiro.test)
Grupos: G1

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.8925, p-value = 0.3317

------------------------------------------------------------
Grupos: G2

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.8418, p-value = 0.1348

------------------------------------------------------------
Grupos: G3

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.7937, p-value = 0.05155
```

El test de Shapiro-Wilk para evaluar el cumplimiento del supuesto de normalidad, muestra que las observaciones de contenido de clorofila foliar en todos los grupos estudiados siguen una distribución normal (p-valor > 0.05).

Ahora mostramos la salida de resultados de R para la evaluación del supuesto de homogeneidad de varianzas

```
> bartlett.test(Clorofila,Grupos)

        Bartlett test of homogeneity of variances

data:  Clorofila and Grupos
Bartlett's K-squared = 2.1915, df = 2, p-value = 0.3343
```

Los resultados del test de Bartlett para homogeneidad de varianzas, muestra que los datos de contenido de clorofila foliar en todos los grupos tienen varianzas homogéneas (p-valor > 0.05).

Ya probado el cumplimiento de los supuestos del modelo, podemos aplicar el análisis de varianza a los datos, obteniéndose la siguiente salida de resultados de R

```
> Anova<-aov(Clorofila~Combinaciones*Grupos)
> summary(Anova)
                    Df Sum Sq Mean Sq F value   Pr(>F)
Combinaciones        1 18.550  18.550 112.285 1.91e-07 ***
Grupos               2  0.351   0.175   1.062   0.3761
Combinaciones:Grupos 2  1.693   0.847   5.124   0.0246 *
Residuals           12  1.982   0.165
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Donde se observa evidencia suficiente de que existe interacción significativa entre las combinaciones de humus utilizados y los diferentes grupos experimentales (p-valor<0.05). Así mismo, se evidencia la existencia de diferencias estadísticamente significativas en el contenido medio de clorofila foliar entre las combinaciones de humus utilizadas (p-valor < 0.05).

Dada la existencia de efectos significativos de interacción, construimos un gráfico de interacción entre los factores, para observar el comportamiento de los mismos sobre el contenido medio de clorofila foliar (Figura 7.8).

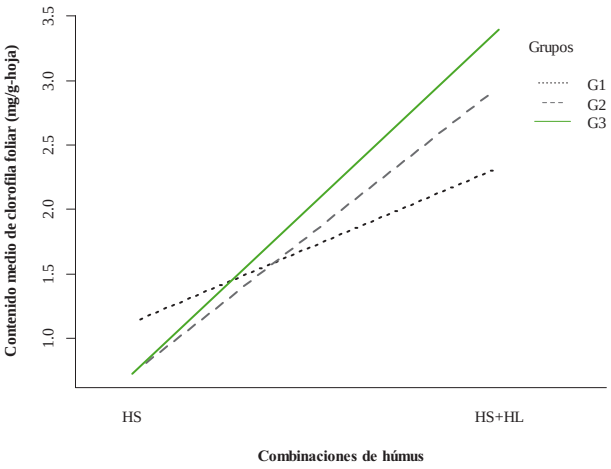


Figura 7.8. Gráfico de interacción entre las combinaciones de humus y los grupos experimentales sobre el contenido medio de clorofila foliar.

De la Figura 7.8, se observa el efecto marcado de interacción entre los factores del diseño. Así mismo, ante la consecuente existencia de diferencias significativas en las medias del contenido de clorofila foliar entre las combinaciones de humus utilizadas, se observa que el mayor rendimiento en el cultivo de pepino, se obtuvo en el grupo experimental G3, bajo la aplicación de una combinación de humus sólido y humus líquido.

7.7. Transformación de variables

En secciones anteriores se ha comentado que para que el análisis de varianza provea resultados confiables, es necesario que las observaciones del diseño cumplan con los requerimientos del modelo ANOVA, es decir, que exista aleatorización e independencia en los datos y que estos se distribuyan en forma normal con varianzas homogéneas (Box & Cox, 1964; Gaete, 1978).

Sin embargo, el cumplimiento de todos estos requerimientos supone una situación idealizada que raramente se cumple en los experimentos reales, donde eventualmente se puede presentar el incumplimiento de unos o más de los criterios antes citados. Ante esta situación el experimentador para analizar sus datos puede tomar dos caminos: 1) Transformar las variables, o 2) elegir otro modelo estadístico que no requiera del cumplimiento de estas restricciones, es decir, elegir métodos no paramétricos o de distribución libre que discutiremos posteriormente.

En esta sección, revisaremos diferentes métodos de transformación de las variables, en los que el propósito de cada técnica está orientada a cambiar la escala de medición en la que se encuentran los datos originales a otra donde se pueda dar cumplimiento a los supuestos del modelo ANOVA, especialmente al de varianzas homogéneas, dada la robustez que presenta el test al incumplimiento sutil del supuesto de normalidad; además de que los supuestos de independencia y aleatoriedad generalmente se satisfacen durante el desarrollo del experimento cuando se aplican los tratamientos al azar a las unidades experimentales.

La elección de la transformación específica a utilizar no es una tarea fácil, y depende en gran medida del conocimiento que el investigador tenga de su experimento. Sin embargo, dependiendo de la naturaleza y la tipología de las variables de estudio existen pautas que pueden servir al investigador inexperto para lograr una buena elección del método de transformación. A continuación, discutiremos los métodos de transformación de variables de uso más recurrente en la literatura revisada.

7.7.1. Transformación de variables con distribuciones conocidas

7.7.1.1. *Transformación raíz cuadrada* $x = \sqrt{y + k}$.

Este tipo de transformación es conveniente aplicarla cuando las observaciones de la variable que se pretende cambiar de escala siguen una distribución de Poisson o binomial. A menudo estas observaciones corresponden a investigaciones donde los resultados se expresan mediante un simple conteo o enumeración, en vez de tratarse de mediciones en una escala de intervalo (Gaete, 1978; Kuehl, 2001; Zimmermann, 2004).

El valor de k que se presenta dentro del radical de la expresión utilizada en este tipo de transformación puede asumir, en general, tres valores diferentes, que pueden ser 0, 0.5 o 1 (Zimmermann, 2004). Cuando el número de observaciones es grande, suele usarse una valor de $k = 0$, y con frecuencia es suficiente para estabilizar la varianza de los datos. Sin embargo, cuando el número de observaciones es bastante pequeño (de tres a diez observaciones solamente) y existen valores muy pequeños o ceros, la transformación anterior es menos estable y eficiente que aquella cuando se utilizan valores de k de 0.5 o 1 (Gaete, 1978; Zimmermann, 2004).

La transformación raíz cuadrada fue extensamente estudiada por Bartlett (1936), quien descubrió que con el uso de esta, aparte de la consecuente

estabilización que se logra de la varianza de cada una de las muestras o tratamientos, se consigue a su vez un beneficio en aquellas distribuciones sesgadas hacia la derecha o hacia la izquierda, pues en tales casos, se acorta la cola larga de las distribuciones.

7.7.1.2. Transformación angular

También conocida como ***transformación arco seno***, se aplica a variables que se encuentran expresadas en porcentajes o proporciones, más específicamente, a aquellos resultados de investigaciones en los cuales las anotaciones proporcionan el número de individuos o elementos *x* que poseen una determinada característica dentro de un grupo mayor *n* de ellos, es decir, la probabilidad de ocurrencia de un evento. Cuando los resultados se expresan así, la frecuencia de las observaciones tiende a seguir una distribución binomial; por lo tanto, para la aplicación de la transformación angular, es preciso que las observaciones del experimento se encuentren binomialmente distribuidas.

Si designamos por $\hat{p}$ a la proporción o fracción de los *n* individuos que poseen la característica en estudio y por otra parte $\hat{q} = 1 - \hat{p}$, representa la proporción de los individuos restantes que no la poseen, tenemos que $\hat{p} + \hat{q} = 1$, y si relacionamos esta expresión con la siguiente identidad pitagórica de la trigonometría

$$\text{sen}^2\theta + \cos^2\theta = 1$$

Donde θ es un ángulo, se puede considerar que $\hat{p} = \text{sen}^2\theta$, de donde se obtiene que $\theta = arcsen\sqrt{\hat{p}}$, con la cual se calcula a partir del valor de una proporción el valor de un ángulo, de allí el termino de *transformación angular*.

7.7.2. Transformación de variables con exponentes para estabilizar la varianza

Muchas veces tiene lugar la imposibilidad de determinar la distribución de probabilidad que siguen las observaciones de nuestro estudio con base en los estimadores de los parámetros que de ellas obtenemos. En estas circunstancias, es posible determinar una función de transformación a través de la suposición de que la desviación estándar de las observaciones es proporcional a alguna potencia de su media aritmética (Box & Cox, 1964), que se relacionan a través de

$$\sigma_y \propto \mu^{\beta}$$

Si empleamos una transformación con exponentes sobre las observaciones que tome la siguiente forma

$$x = y^{\lambda}$$

Tendríamos como resultado una relación proporcional de la desviación estándar con la media aritmética de las observaciones transformadas, dada por

$$\sigma_x \propto \mu^{\lambda+\beta-1}.$$

Si ocurre que $\lambda = 1-\beta$, entonces la desviación estándar de la variable transformada será constante, ya que $\lambda+\beta-1=0$ y $\sigma_x \propto \mu^0$ (Kuehl, 2001). De allí, que el problema se resume a hallar el valor de λ con el cual se efectuará la transformación de las observaciones. Sin embargo, el uso de extendido de algunas transformaciones, han permitido establecer cuáles son los valores de lambda, para los cuales se obtienen las transformaciones más útiles, a través de la denominada *escalera de exponentes* (Tukey, 1977) que se muestra en la Tabla 7.7.

En ella, los valores de λ inferiores a 1 acortaran la cola superior y alargaran la cola izquierda de las observaciones. De manera inversa, si λ es mayor que 1, una distribución sesgada a la izquierda cambia a una más simétrica al mover las observaciones que están en el lado izquierdo. La transformación logaritmo se coloca en la posición cero de la escalera porque su efecto sobre las observaciones cae de manera natural en esa posición (Kuehl, 2001). Siempre que se desee aplicar la transformación logaritmo, es recomendable examinar las observaciones en búsquedas de observaciones iguales a cero, en caso de existir algunas, es aconsejable agregar una pequeña constante *k*, tal que $0<k\leq 1$ a cada una de las observaciones para evitar el cálculo de un logaritmo de cero, dado que este no existe. Con frecuencia se usan 1 y $\frac{1}{2}$ para *k* (Gaete, 1978; Yamamura, 1999), pero Mosteller & Tukey (1977), citados por Kuehl (2001), sugieren un valor de $k=\frac{1}{6}$.

Tabla 7.7. Transformaciones en la escalera de exponentes

λ	y^{λ}	Nombre	Observaciones
2	y^2	Cuadrada	La más usada
1	y^1	Datos originales	No hay transformación
$\frac{1}{2}$	$\sqrt{y}$	Raíz cuadrada	Distribución de Poisson
0	$\ln y$	Logaritmo	"0" en la escalera
$-\frac{1}{2}$	$1/\sqrt{y}$	Raíz	El signo menos preserva el orden de las observaciones
-1	$1/y$	Recíproco	Reexpresa el tiempo como tasa

Análogo a lo anterior, Guisande *et al.*, (2011), propone una modificación de la escalera de exponentes realizada por Erickson & Nosanchuk (1977), que muestra el tipo de exponente recomendado para realizar la transformación en función de la asimetría o la dirección en la que van los casos extremos (Figura 7.9). Según este criterio, se deben elegir la transformación que haga más próximos a cero los coeficientes de asimetría y de kurtosis.

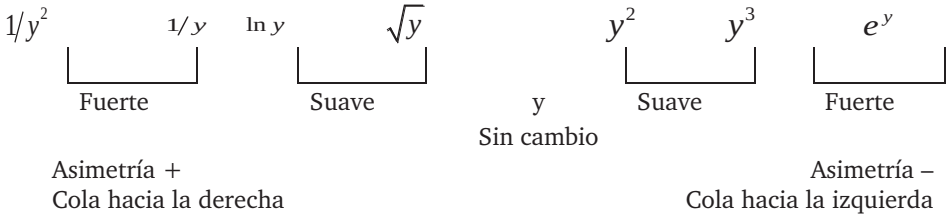


Figura 7.9. Escalera de exponentes adaptada por Guisande *et al.* (2011).

Otro enfoque para la elección de la transformación con exponentes más adecuada para nuestras variables, es la introducida por Box & Cox (1964), cuyo procedimiento de estimación se basa en el método de máxima verosimilitud (no discutido en este texto) para calcular el valor de λ, que define el tipo de transformación más apropiado para la variable respuesta y lograr que esta cumpla con los supuesto del análisis de varianza (Monteiro & Gómez, 2006). Este procedimiento recibe el nombre de **familia de transformaciones Box-Cox**, y se expresa matemáticamente por

$$x = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \ln y & (\lambda = 0) \end{cases}$$

El parámetro λ se calcula al maximizar la función de verosimilitud

$$L(\lambda) = -\frac{1}{2}\ln[s^2_{(\lambda)}]$$

Donde $S^2_{(\lambda)}$ es la media cuadrática del error calculado para el análisis de varianza, usando la transformación $x = (y^{\lambda} - 1)/\lambda$ para la elección dada de λ (Kuehl, 2001).

Una solución para esto se puede obtener si se determina $s^2_{(\lambda)}$ para un conjunto de valores elegidos de λ, y a partir de una gráfica de $L(\lambda)$ contra λ, se toma el valor de λ que hace que la función $L(\lambda)$ sea máxima. Este valor es el que se escogerá para aplicar la transformación respectiva.

En la práctica este procedimiento resulta ser bastante engorroso, dado que para el conjunto de valores seleccionados de λ, se debe realizar un ANOVA para hallar el valor de $s^2_{(\lambda)}$. Sin embargo, muchos de los paquetes estadísticos de la actualidad permiten realizar el procedimiento para la familia de transformaciones Box-Cox sin requerir mucho esfuerzo, y R no es la excepción. En su ambiente de programación se utiliza la función ***boxcox*** del paquete "*MASS*" (Ripley *et al.*, 2015), siguiendo la siguiente línea de código

```
boxcox(object, lambda = seq(-2, 2, 1/10),ylab = "log-
Likelihood", xlab = expression(lambda),...)
```

Donde ***object***, corresponde a la fórmula del modelo ANOVA, ***lambda*** especifica la secuencia de valores de λ a través de los cuales se construirá el gráfico, ***ylab*** y ***xlab***, establecen las etiquetas de los ejes *x* e *y* que serán insertadas en el gráfico.

A continuación, ilustraremos lo anteriormente discutido a través de un ejemplo.

Ejemplo 7.13. Durante los años 2004 a 2005 el grupo de investigación Pichihuel de la Universidad de La Guajira, realizó un estudio sobre la dinámica fisicoquímica del ecosistema estuarino el Riito. Los datos referentes a las concentraciones de NO_2 (mg/L), desde noviembre de 2004

a septiembre de 2005, en cuatro diferentes estaciones de muestreo se muestran a continuación

	Estaciones			
Meses	**E1**	**E2**	**E3**	**E4**
Nov	0.43	0.53	0.43	0.44
Dic	0.30	0.28	0.28	0.22
Ene	0.26	0.28	0.14	0.22
Feb	0.20	0.18	0.16	0.19
Mar	0.05	0.14	0.04	0.04
Abr	0.13	0.22	0.13	0.13
May	0.25	0.21	0.18	0.21
Jun	0.38	0.39	0.33	0.39
Jul	0.52	0.57	0.49	0.58
Ago	5.58	5.32	2.60	3.38
Sep	0.34	0.47	0.31	0.37

A partir de estas observaciones se desea saber si existen diferencias significativas en las concentraciones medias de NO_2 en las diferentes estaciones de muestreo.

Solución

Previamente al inicio de nuestro análisis, cargamos los datos en la consola de R a través de la construcción de vectores de datos, dada la pequeña cantidad de ellos.

```
> NO2.E1<-
c(0.43,0.30,0.26,0.20,0.05,0.13,0.25,0.38,0.52,5.58,0.34)
> NO2.E2<-
c(0.53,0.28,0.28,0.18,0.14,0.22,0.21,0.39,0.57,5.32,0.47)
> NO2.E3<-
c(0.43,0.28,0.14,0.16,0.04,0.13,0.18,0.33,0.49,2.60,0.31)
> NO2.E4<-
c(0.44,0.22,0.22,0.19,0.04,0.13,0.21,0.39,0.58,3.38,0.37)
> NO2<-c(NO2.E1,NO2.E2,NO2.E3,NO2.E4)
> Estación<-rep(1:4,each=11)
> Estación<-factor(Estación,labels=c("E1","E2","E3","E4"))
```

Ahora, evaluaremos el cumplimiento de los supuestos del modelo ANOVA, como se ha discutido en secciones anteriores. Iniciaremos con la evaluación del supuesto de normalidad.

```
> by(NO2,Estación,shapiro.test)
Estación: E1

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.4251, p-value = 1.937e-07

------------------------------------------------------------
Estación: E2

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.4327, p-value = 2.382e-07

------------------------------------------------------------
Estación: E3

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.5239, p-value = 2.95e-06

------------------------------------------------------------
Estación: E4

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.497, p-value = 1.396e-06
```

El examen de los p-valores, muestra que las concentraciones de NO_2 en todas las estaciones de muestreo no se ajustan a una distribución normal, a un nivel de significancia de 0.05 (p-valor < 0.05). Por lo tanto, el supuesto de normalidad del modelo no se satisface.

Ahora, veamos la salida de resultados para el examen de homocedasticidad de las concentraciones de NO_2.

```
> bartlett.test(NO2,Estación)

        Bartlett test of homogeneity of variances

data:  NO2 and Estación
Bartlett's K-squared = 7.578, df = 3, p-value = 0.05559
```

A través del p-valor, observamos que las concentraciones de NO_2 cumplen el criterio de homocedasticidad de las observaciones entre las estaciones de muestreo (p-valor > 0.05). Sin embargo, al incumplirse fuertemente el supuesto de normalidad, los resultados del ANOVA nos pueden llevar a conclusiones y decisiones erróneas, a pesar de la robustez de esta prueba a un sutil incumplimiento de la normalidad (p-valor entre 0.01 y 0.05). A partir de lo anterior, es recomendable utilizar alguna transformación en las observaciones, para cambiar su escala de medición y lograr que en esta nueva escala, las observaciones satisfagan los supuestos del modelo.

Para elegir la transformación más adecuada para estas observaciones, utilizaremos el procedimiento de la familia de transformaciones Box-Cox (Box & Cox, 1964; Kuehl, 2001; Monteiro & Gómez, 2006), por su fácil aplicabilidad metodológica a través de R.

Para ello realizaremos la construcción de un panel de dos gráficos con la función ***boxcox***, descrita anteriormente, apoyados con la función ***par***, como se muestra en las siguientes líneas de código

```
> par(mfrow=c(1,2),family="serif",bty="l",font.lab=2)
> boxcox(NO2~Estación,ylab="Log función de verosimilitud")
> boxcox(NO2~Estación,lambda=seq(-0.5,0,0.01),ylab="Log
función de verosimilitud")
```

Donde el argumento ***mfrow=c(1,2)***, indica que se construya un panel gráfico de una sola fila y dos columnas, ***family = "serif"*** establece que el tipo de letra de gráfico sea *Times New Roman*, ***bty = "l"***, especifica que solo se tracen los ejes inferior e izquierdo del gráfico, y ***fon.lab = 2***, ordena a R que las etiquetas de los ejes estén en negrita. El panel gráfico construido se muestra en la Figura 7.10, donde el gráfico de la izquierda corresponde a la primera línea de código ejecutada con la función ***boxcox***, para la secuencia de valores por defecto de λ, de -2 a 2, y las líneas punteadas delimitan la región de un intervalo de confianza del 95% donde se hace máxima la función de verosimilitud. De acuerdo a los límites de esa región (aproximadamente de -0.5 a 0.0), se construye otro gráfico con la función ***boxcox*** (gráfico de la derecha), para observar más detalladamente cual es el valor de λ que maximiza la función de verosimilitud; observamos en el gráfico que es un valor cercano a -0.2 ($-\frac{1}{5}$), de allí que una transformación adecuada para los datos de concentraciones de NO_2 es $1/\sqrt[5]{y}$.

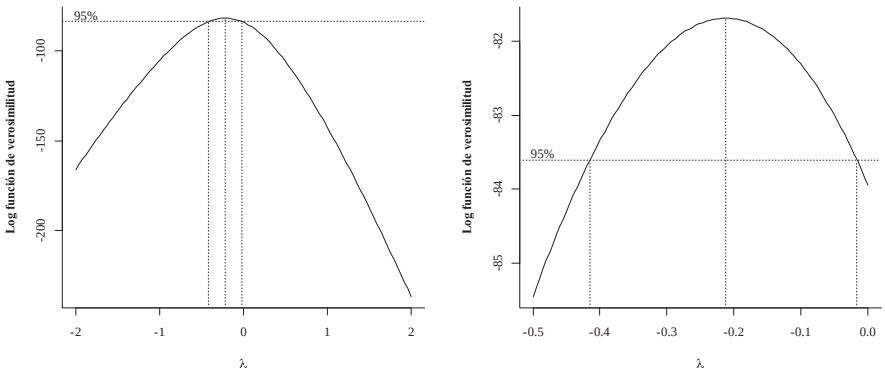


Figura 7.10. Estimación de λ a través del procedimiento de Box-Cox.

Ahora realizamos la transformación de los datos y evaluamos los supuestos de normalidad y de homogeneidad de varianzas a las concentraciones de NO_2 en las cuatro estaciones de muestreo como se muestra a continuación

```
> Trans.NO2<-1/(NO2^(1/5))
> by(Trans.NO2,Estación,shapiro.test)
Estación: E1

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.911, p-value = 0.2508

------------------------------------------------------------
Estación: E2

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.8722, p-value = 0.08275

------------------------------------------------------------
Estación: E3

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9454, p-value = 0.5857

------------------------------------------------------------
Estación: E4

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9278, p-value = 0.3893

> bartlett.test(Trans.NO2,Estación)

        Bartlett test of homogeneity of variances

data:  Trans.NO2 and Estación
Bartlett's K-squared = 0.9486, df = 3, p-value = 0.8137
```

De estos resultados se observa que luego de aplicar la transformación sugerida, las concentraciones de NO_2 en las cuatro estaciones de muestreo se ajustaron a una distribución normal (p-valor > 0.05), dándole cumplimiento a este supuesto del modelo. Así mismo, se logró una mayor estabilización de la varianza de las observaciones, reflejado en el aumento del p-valor del test de Bartlett, luego de aplicada la transformación.

Ahora, con los supuestos del modelo ANOVA satisfechos, podemos aplicar este procedimiento con los datos transformados, para evaluar la existencia de diferencias significativas de las concentraciones medias de NO_2, en las estaciones de muestreo, como se observa en la siguiente salida de resultados de R

```
> Anova<-aov(Trans.NO2~Estación)
> summary(Anova)
            Df Sum Sq Mean Sq F value Pr(>F)
Estación     3 0.0678 0.02260   0.345  0.793
Residuals   40 2.6175 0.06544
```

Los resultados del análisis de varianza, proporcionan evidencia estadística para concluir que no existe diferencias significativas en las concentraciones medias de NO2 en las cuatro estaciones con una confiabilidad del 95% (p-valor > 0.05). Una representación gráfica de los resultados del ANOVA para los datos antes y después de transformar se muestra en la Figura 7.11. Nótese como la transformación normaliza notablemente los datos.

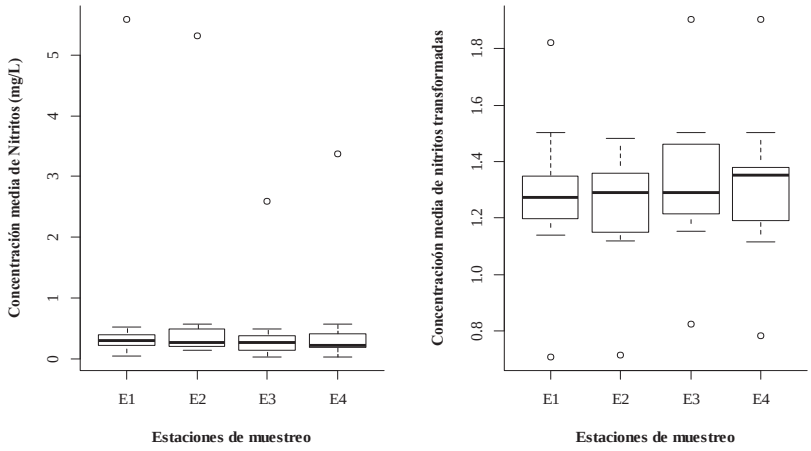


Figura 7.11. Concentraciones de NO_2 en las cuatro estaciones antes y después de transformar.

8.

Modelos de regresión

8.1. Generalidades

Durante el desarrollo de una investigación que involucra varias variables evaluadas simultáneamente, para cumplir un objetivo específico, es común intentar encontrar relaciones inherentes entre ellas, de tal forma que se pueda predecir con extrema precisión el valor de una variable en función del valor que toma otra u otras más. Lo anterior, se puede abordar a través de técnica estadísticas denominadas genéricamente como **modelos de regresión**.

En general, los modelos de regresión tiene como objetivo encontrar la "mejor" relación funcional entre una variable *y*, denominada variable *dependiente* o *respuesta*, y una o más variables $x_1, x_2, ..., x_k$, llamadas variables *independientes*, *regresoras* o *explicativas*, que son medidas con un error despreciable y se controlan durante el desarrollo del experimento (Walpole *et al.*, 2007; Vargas, 2007).

Para un conjunto de observaciones muéstrales, la relación funcional entre las variables respuesta y explicativa recibe el nombre de **ecuación de regresión**. Esta al ser usada para realizar predicciones de la variable respuesta, contiene un componente aleatorio que le da su naturaleza estocástica o probabilística, en contraposición con los modelos determinísticos, donde la relación entre las variables es exacta. De allí, que el problema del análisis de regresión, se reduzca a buscar la mejor relación entre las variables que permita realizar predicciones de la variable respuesta con un margen de error despreciable.

Los modelos de regresión, dependiendo de la tipología de las variables involucradas, especialmente de la variable respuesta, pueden tomar diferentes clasificaciones. Así, cuando la variable respuesta es de naturaleza cuantitativa se habla de un modelo de **regresión lineal**, el cual puede ser *simple* cuando existe solo una variable explicativa o *múltiple* cuando están involucradas dos o más variables regresoras. Sin embargo, existen situaciones en las que la relación inherente entre las variables de estudio no obedece un patrón lineal, por lo que es necesario explorar otras relaciones funcionales para evaluar el comportamiento entre las variables estudiadas, este tipo de modelos recibe el nombre de **regresión no lineal**.

Cuando se analiza la relación entre una variable respuesta de carácter cualitativo dicotómico (aquella que solo admite dos categorías que definen opciones o características mutuamente excluyentes, por ejemplo, si o no), frente a una variable explicativa cuantitativas, nos enfrentamos a un modelo de **regresión logística**. Todos estos modelos, serán revisados en el desarrollo de este capítulo, desde su fundamentación teórica, interpretación y aplicación práctica en el entorno de programación de R.

8.2. Regresión lineal simple

Antes, mencionamos que cuando intentamos encontrar la mejor relación funcional entre una variable respuesta de naturaleza cuantitativa frente a una variable explicativa, también de carácter cuantitativa, tratamos con un modelo de regresión lineal simple. También se hizo hincapié en que la característica fundamental de este modelo es que la relación entre las dos variables no es de naturaleza determinística, sino, que existe un componente aleatorio que le imprime características probabilísticas o estocásticas, es decir, la relación entre las dos variables no es exacta, y el valor de la variable respuesta puede ser distinto para una mismo valor de la variable regresora. En tales situaciones, el objetivo del análisis de regresión es encontrar una función que represente el mejor ajuste entre las dos variables y que tenga en cuenta el componente aleatorio del modelo. Para un modelo de regresión lineal simple la mejor y más simple forma de relacionar la variable respuesta *y*, y la variable regresora *x* de una forma determinística a través de la ecuación de la recta

$$y = \beta_0 + \beta_1 x$$

donde, por supuesto β_0, es la intersección de la recta y β_1 su pendiente.

Ahora, teniendo en cuenta el componente aleatorio del modelo de regresión, la ecuación de la recta anterior toma la forma

$$y = \beta_0 + \beta_1 x + \varepsilon$$

donde la cantidad ε de esta ecuación recibe el nombre de **error aleatorio** o **desviación aleatoria** (Walpole *et al.*, 2007; Devore, 2008). Esta ecuación, recibe el nombre de **recta de regresión verdadera** o **recta de regresión de la población**, y nunca es conocida con exactitud, por lo que es necesaria

estimarla a partir de la información muestral con que se disponga, es decir, una muestra de *n* parejas de valores (*x*, *y*).

8.2.1. Supuestos del modelo de regresión lineal simple

Cuando tratamos con variables dependientes cuantitativas, para aplicar un modelo de regresión entre las variables estudiadas, no se requiere que los datos presenten una distribución normal o que exista homogeneidad de varianzas, como lo exige el análisis de varianza. Sin embargo, para determinar si la función obtenida con el modelo de regresión es significativa, es necesario aplicar contrastes específicos que hace necesario se dé cumplimiento a requisitos que se enunciaran a continuación relacionados con los residuos (diferencia entre el valor observado de la variable respuesta y el valor ajustado por la función) (Guisande *et al.*, 2011). Estos requisitos, también aplican para los modelos de regresión lineal múltiple:

1. Los residuos obtenidos del modelo de regresión deben presentar una distribución normal.
2. Debe existir homocedasticidad en los residuos, es decir, la varianza de los mismos debe ser constante.
3. No debe existir autocorrelación en la serie de residuos (deben ser independientes). Si este requisito no se cumple, no es posible saber el grado de relación exacta entre a variable respuesta y la variable(s) explicativa(s), ya que parte de la predicción se debe a los propios valores de la variable respuesta.
4. En el caso del modelo de regresión múltiple, no debe existir relación entre las variables independientes, es decir, no debe existir multicolinealidad.

8.2.2. La recta de regresión ajustada

Anteriormente se dejó por sentado que la recta de regresión verdadera nunca es conocida, debido a que en la práctica se dispone de datos muéstrales, en lugar de los poblacionales, y es imposible determinar el valor exacto de los parámetros β_0 y β_1 del modelo, y el error aleatorio ε. Sin embargo, a partir de la información muestral de las *n* parejas de valores (*x*, *y*), es posible realizar estimaciones acerca de los parámetros del modelo, que en adelante los denominaremos $\hat{\beta}_0$ y $\hat{\beta}_1$, y con base en ellos hallar la recta de **regresión ajustada** o estimada, dada por

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x ,$$

donde $\hat{y}$ es el valor pronosticado o ajustado. Es evidente que la recta de regresión ajustada representa una estimación de la verdadera recta de regresión, y se espera que la recta de regresión ajustada esté tan cerca como sea posible de la verdadera recta de regresión cuando se disponga de una gran cantidad de datos.

En la Figura 8.1, se muestra un diagrama de dispersión entre una variable respuesta *y* y una variable regresora *x*. Así mismo, se muestra la recta de regresión verdadera $y = \beta_0 + \beta_1 x$, y la recta de regresión ajustada $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Además, es evidente que mientras el error aleatorio ε_i impide que la recta de regresión verdadera sea una ecuación determinista, los residuos e_i, indican un error de ajuste del modelo estimado $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

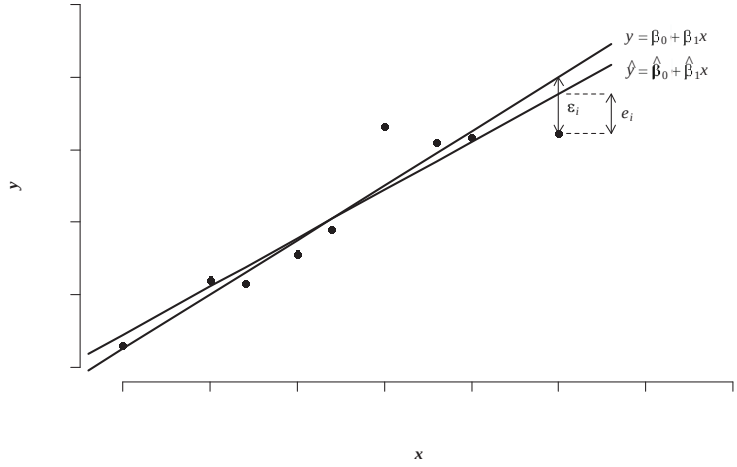


Figura 8.1. Comparación de $y = \beta_0 + \beta_1 x$ y $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

8.2.3. Estimación de los parámetros del modelo ajustado

El problema de estimar la recta de regresión ajustada, es equivalente a determinar las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ de los parámetros de la recta de regresión verdadera β_0 y β_1, respectivamente. Evidentemente, esto permite realizar el cálculo de valores ajustados o pronosticados a través de $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, y realizar inferencias acerca de la calidad del ajuste del modelo y la fuerza de la relación entre las variables, como se discutirá en secciones siguientes.

Antes de introducirnos en el procedimiento utilizado para la estimación de los parámetros del modelo, es necesario comprender el concepto de **residuo**, considerados como la diferencia entre los valores observados y los pronosticados por el modelo para un conjunto de *n* parejas de valores (*x*, *y*). Así, dado un conjunto de pares de datos $\left[(x_i, y_i); i = 1, 2, ..., n\right]$, y un modelo de regresión ajustado $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, el *i*-enésimo residuo e_i está dado matemáticamente por la expresión

$$e_i = y_i - \hat{y}_i, \qquad i = 1, 2, ..., n.$$

Cuanto más grande sea el valor de los residuos, mayor es la carencia de ajuste del modelo estimado. En oposición a lo anterior, cuanto más pequeño sea el valor de los residuos, mayor es el ajuste del modelo estimado.

La deducción de expresiones que permitan realizar estimaciones de los parámetros del modelo de regresión ajustado se realiza bajo la condición de que la suma de los cuadrados de los residuos sea mínima, comúnmente a esta suma de cuadrados de los residuos se les denomina *suma de cuadrados de los errores* respecto a la recta de regresión y se denota como *SSE* (Walpole *et al.*, 2007), debido a que estos expresan un error de ajuste del modelo. El procedimiento de minimización para estimar los parámetros del modelo se llama **método de los mínimos cuadrados**. Así, para encontrar las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$, se debe minimizar

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

Al diferenciar parcialmente *SSE* con respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$, se obtiene

$$\frac{\partial(SSE)}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right), \qquad \frac{\partial(SSE)}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i \ .$$

Al igualar a cero las derivadas parciales y reacomodar los términos, obtenemos las siguientes ecuaciones, llamadas **ecuaciones normales**

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \qquad \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \ ,$$

que al ser resueltas simultáneamente se obtienen las siguientes expresiones de cálculo de los parámetros estimados $\hat{\beta}_0$ y $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

y

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \overline{y} - \hat{\beta}_1 \overline{x}\ .$$

En R, la estimación de los parámetros del modelo de regresión lineal simple, se realiza a través de función ***lm*** del paquete básico de instalación del software, en cuyo argumento solo se debe especificar la variable respuesta y la variable explicativa de la siguiente forma

```
lm(y ~ x)
```

Ejemplo 8.1. A continuación se muestran las densidades de coliformes fecales (CF) expresadas como NMP/100 mL y las concentraciones de Amonio (NH_4) en mg/L, obtenidas en un estudio limnológico realizado en una laguna costera del departamento de La Guajira. A partir de estos datos se desea predecir el comportamiento del grupo bacteriano en función de los valores de concentración de NH_4.

CF	2100	2350	2500	2600	2750	2700	2700	2800	3000	3100	3400
NH_4	0.39	0.43	0.46	0.45	0.49	0.53	0.56	0.63	0.67	0.67	0.76

Solución

Lo expuesto en el enunciado, sugiere que el objetivo de análisis es encontrar un modelo que mejor represente la relación existente entre las dos variables estudiadas. Al observar el gráfico de dispersión de la Figura 8.2 se observa un buen ajuste de las variables a una línea recta, por ello construiremos un modelo de regresión lineal tomando las densidades de CF como variable respuesta y las concentraciones de NH_4 como variable regresora.

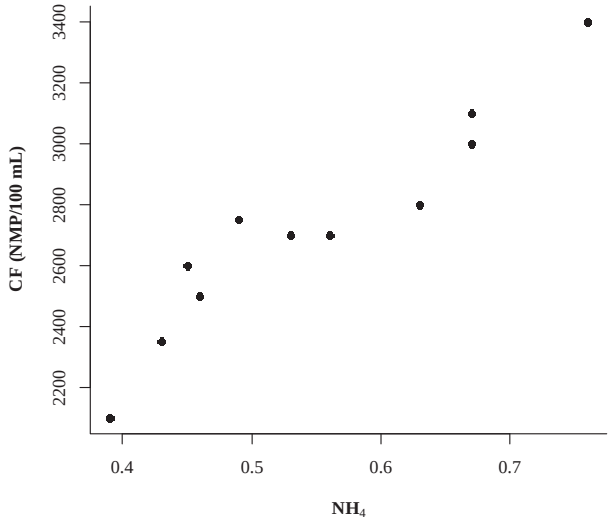


Figura 8.2. Densidad de coliformes fecales en función de las concentraciones de Amonio.

Para fines prácticos y para evitar la ejecución de cálculos tediosos, los resultados de las operaciones en que se incurren en el cálculo de los estimadores de los parámetros del modelo, se resumen en la Tabla 8.1.

Tabla 8.1. Resumen de cálculo de los términos para la estimación de los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$.

No	y_i	x_i	x_i^2	$x_i y_i$
1	2100	0.39	0.152	819
2	2350	0.43	0.185	1010.5
3	2500	0.46	0.212	1150
4	2600	0.45	0.203	1170
5	2750	0.49	0.240	1347.5
6	2700	0.53	0.281	1431
7	2700	0.56	0.314	1512
8	2800	0.63	0.397	1764
9	3000	0.67	0.449	2010
10	3100	0.67	0.449	2077
11	3400	0.76	0.578	2584
$\sum$	30000	6.04	3.458	16875

A partir de estos resultados, se determinan los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$, a través de las ecuaciones presentadas anteriormente, deducidas a través del método de los mínimos cuadrados

$$\hat{\beta}_1 = \frac{(11)(16875)-(6.04)(30000)}{(11)(3.458)-(6.04)^2}$$

$$\hat{\beta}_1 = 2843.1$$

y de esta forma el valor estimado $\hat{\beta}_0$ es

$$\hat{\beta}_0 = \frac{30000-(2843.1)(6.04)}{11}$$

$$\hat{\beta}_0 = 1166.2$$

Así, la recta de regresión ajustada para estas observaciones es $\hat{y} = 1166.2 + 2843.1x$, o en términos de las variables dadas $CF = 1166.2 + 2843.1NH_4$, lo que sugiere que por cada aumento de las concentraciones NH_4 en 1 mg/L, la densidad de CF aumenta 2843 NMP/ 100 mL. Esta recta de regresión estimada es posible superponerla al gráfico de dispersión mostrado en la Figura 8.2 para observar el ajuste de esta a los datos del problema (Figura 8.3).

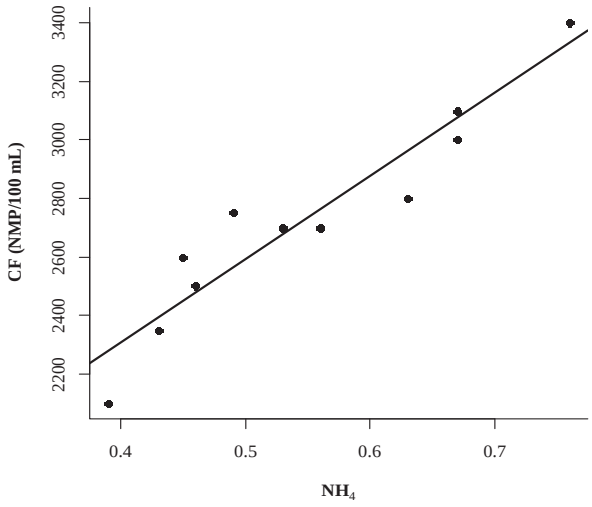


Figura 8.3. Recta de regresión estimada sobre el gráfico de dispersión de los datos.

De este gráfico se observa que la recta estimada se ajusta de forma aceptable a las observaciones, por ello puede ser usada con propósitos de pronóstico de las densidades de CF dado el conocimiento de las concentraciones de NH_4. No obstante, es necesario pruebas inferenciales de calidad de ajuste de la recta, que nos permitan determinar con un alto grado de confianza la utilidad del modelo como veremos más adelante.

La salida de resultados del entorno de R para el análisis de regresión se muestra a continuación. Inicialmente, realizamos el ingreso de los datos, que al ser pocos, lo hacemos a través de la construcción de vectores de datos, luego construimos el gráfico de dispersión de la Figura 8.2 con la función ***plot***, discutida en secciones anteriores

```
> CF <-
c(2100,2350,2500,2600,2750,2700,2700,2800,3000,3100,3400)
> NH4 <-
c(0.39,0.43,0.46,0.45,0.49,0.53,0.56,0.63,0.67,0.67,0.76)
> plot(NH4,CF,ylab="CF (NMP/100
mL)",xlab=expression(bold(NH[4])), pch=16,font.lab=2)
```

Después de la construcción del gráfico de dispersión, continuamos con la determinación de los parámetros del modelo a través de la función ***lm***, la cual se asigna a un objeto que llamaremos *Reg*

```
> Reg <- lm(CF~NH4)
> Reg

Call:
lm(formula = CF ~ NH4)

Coefficients:
(Intercept)          NH4
       1166         2843
```

Por último se inserta la recta de regresión ajustada al gráfico de dispersión a través de la función ***abline***, en cuyo argumento se indica el objeto que contiene la recta de regresión, y se añaden otros parámetros (argumentos) relativos a las propiedades que deseamos de la recta, como grosor (*lwd*), tipo (*lty*), color (*col*), como se muestra en la siguiente salida de R

```
> abline(Reg, lwd=2, col = "black")
```

8.2.4. Inferencias sobre la pendiente del modelo

Después de determinar los parámetros del modelo de regresión ajustado, es necesario realizar inferencias sobre la linealidad del modelo, más específicamente sobre la nulidad (o no) de la pendiente de la recta de regresión verdadera β_1. Antes de discutir este procedimiento de inferencia es necesario llegar a una estimación de la varianza del error del modelo σ^2, que refleja la variación aleatoria del error experimental alrededor de la recta de regresión.

Para resumir las expresiones que se introducirán a continuación se recomienda familiarizarse con la siguiente notación

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2; \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2; \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

A partir de lo anterior, y teniendo en cuenta que $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, la suma de cuadrados de los errores *SSE* puede ser expresada como

$$\begin{aligned} SSE &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x \right)^2 = \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right]^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx} = S_{yy} - \hat{\beta}_1 S_{xy} \end{aligned}$$

El paso final surge del hecho de que $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$.

De esta forma, un estimador insesgado para la varianza del error del modelo σ^2 es

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}.$$

Ahora, las inferencias sobre la nulidad de la pendiente de la recta de regresión verdadera a partir de las observaciones muéstrales, se expresa en forma de hipótesis estadística como sigue

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

y el estadístico de prueba, basado en una distribución *t* de *student* con *n* – 2 grados de libertad está dado por la expresión

$$t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}},$$

y el rechazo de H_0 se da cuando $t \geq t_{\alpha/2;n-2}$ o $t \leq -t_{\alpha/2;n-2}$.

Cuando la hipótesis nula es verdadera, *y* es independiente de *x*, así que el conocimiento de *x* no da información sobre el valor de la variable dependiente. La prueba de estas dos hipótesis planteadas anteriormente, a menudo se conoce como **prueba de utilidad del modelo** de regresión lineal simple. Con base en lo anterior, a menos que *n* sea demasiado pequeño, siempre se espera rechazar H_0. El modelo de regresión lineal simple no deberá ser utilizado para más inferencias (estimaciones del valor medio o predicciones de valores futuros) a menos que la prueba de utilidad del modelo dé por resultado el rechazo de H_0 con un α apropiadamente pequeño (Devore, 2008).

Para visualizar la prueba de utilidad del modelo en R, solo basta con aplicar la función ***summary*** al modelo lineal creado con la función ***lm***, como se muestra en la siguiente línea de código

```
Summary(lm(y ~ x))
```

8.2.5. Calidad del ajuste del modelo de regresión lineal simple

8.2.5.1. Coeficiente de determinación R^2.

En la sección anterior se estudió un procedimiento de prueba de hipótesis para verificar la utilidad del modelo de regresión, al evaluar la existencia de relación significativa entre la variable respuesta y la variable explicativa. Sin embargo, también es necesario contar con una medida que nos permita evaluar el ajuste de la recta de regresión estimada a las observaciones, es decir, evaluar la *calidad del ajuste de la recta de regresión estimada*. De

forma visual esto se puede distinguir al observar si los puntos del diagrama de dispersión entre las variables estudiadas tienden a ajustarse razonablemente bien a la recta de regresión ajustada. No obstante, un criterio de naturaleza cuantitativa es el que proporciona un indicador estadístico denominado **coeficiente de determinación**, que mide la calidad de ajuste de la recta de regresión ajustada en función de la proporción de la variabilidad que es explicada por el modelo ajustado (Gutiérrez & de la Vara, 2008). Lo anterior se puede expresar como

$$R^2 = \frac{\text{Variabilidad explicada por el modelo}}{\text{Variabilidad total}}$$

Nótese que al tratarse de componentes de la variabilidad total de las observaciones, es necesario introducir expresiones que denoten cada una de las componentes igual a como se discutió en el capítulo 7 del análisis de varianza. Para ello, partiremos de una de las expresiones que se dedujo para la estimación de la varianza del error σ^2, que expresaba que la suma de cuadrados de los errores estaba dada por

$$SSE = S_{yy} - \hat{\beta}_1 S_{xy} \therefore S_{yy} = \hat{\beta}_1 S_{xy} + SSE ,$$

teniendo en cuenta que $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, la expresión anterior se puede reexpresar como

$$S_{yy} = \frac{S_{xy}}{S_{xx}} S_{xy} + SSE \therefore \sum_{i=1}^{n} = (y_i - \bar{y})^2 = \frac{\sum_{i=1}^{n} (\hat{x}_i - \bar{x})(\hat{y}_i - \bar{y})}{\sum_{i=1}^{n} (\hat{x}_i - \bar{x})^2} \sum_{i=1}^{n} (\hat{x}_i - \bar{x})(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Realizando las simplificaciones algebraicas respectivas, la expresión anterior se reduce a

$$\sum_{i=1}^{n} = (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Esta expresión, representa una partición de la variabilidad total de la variable respuesta *y*, que en adelante simbolizaremos por *SST*, en dos componentes, la primera denominada suma de cuadrados de la regresión

SSR, que refleja la cantidad de variación de los valores de *y* que se explica con el modelo de regresión ajustado, y la segunda componente, ya conocida, denominada suma de cuadrados de los errores *SSE*, que refleja la variación alrededor de la recta de regresión (Walpole *et al.*, 2007). Esta partición de la variabilidad la podemos expresar simbólicamente como $SST = SSR + SSE$.

De esta forma la expresión dada para el cálculo de coeficiente de determinación, se puede reexpresar como

$$R^2 = \frac{SSR}{SST} \therefore R^2 = \frac{SST - SSE}{SST},$$

que es equivalente a decir que

$$R^2 = 1 - \frac{SSE}{SST}.$$

Es claro que los valores que toma el coeficiente de determinación se encuentran en el intervalo $0 \leq R^2 \leq 1$, debido que $0 \leq SSE \leq SST$. De forma ideal se desea tener un $R^2 \approx 1$, pues ello implica que $SSE \approx 0$, y toda la variación de la variable respuesta es explicada por el modelo lineal ajustado (Conavos, 1989; Gutiérrez & De la Vara, 2008).

En las publicaciones científicas es muy común que los investigadores reporten el valor de R^2 obtenido en su estudio, pues este ilustra al lector sobre la calidad del ajuste del modelo de regresión estimado para los datos del estudio, y es claro que entre más cercano de uno se encuentre el valor de R^2, mejor ajuste presenta la recta de regresión estimada. Sin embargo, ¿Qué valor de R^2 es considerado aceptable? Esta pregunta es muy difícil de contestar porque depende en primera instancia de cada experimento y la información que el investigador tenga del mismo, del tamaño del conjunto de datos y de la precisión que se requiere con los resultados del experimento. Por ejemplo, en estudios de calibración instrumental, es deseable que el valor de R^2 se encuentre por encima de 0.95 (incluso más); mientras que en estudios de las ciencias ambientales o biológicas valores de R^2 superiores a 0.70 son considerados adecuados, dado que los fenómenos científicos de estas áreas resultan ser muy variables. Es claro entonces que calificar la precisión o calidad de ajuste de un modelo de regresión, va a depender del fenómeno científico que se esté estudiando,

pues algunos requieren realizar el modelamiento con más precisión que otros.

En R, la salida de resultados al aplicar la función ***summary*** sobre el modelo lineal, como se describió en la sección anterior, también muestra el valor del coeficiente de determinación (R^2), a través de un objeto llamado *R-squared*, como se presentará más adelante cuando se muestre un ejemplo de aplicación del modelo de regresión lineal simple.

8.2.5.2. Coeficiente de correlación r.

En los problemas de análisis de regresión lineal es necesario determinar la fuerza o intensidad de la asociación entre la variable respuesta *y* y la variable explicativa *x*, para asegurarnos que una función lineal es la mejor alternativa para el modelado de nuestros datos. Una medida estadística que permite determinar esta fuerza de asociación a nivel poblacional es el conocido **coeficiente de correlación poblacional** ρ (ro), cuyo valor se encuentra comprendido en el intervalo de $-1 \leq \rho \leq 1$, donde valores de $\rho = \pm 1$ indican una relación lineal perfecta entre *y* y *x*; un valor de $\rho = 1$, indica una relación lineal positiva perfecta entre la variable respuesta y la variable explicativa, mientras que un valor de $\rho = -1$, es indicativo de una relación negativa perfecta entre *y* y *x*. Si $\rho = 0$, entonces no existe relación lineal entre *y* y *x*. De lo anterior también podría decirse que los estimadores muéstrales de ρ con magnitud cercana a la unidad implican una buena *correlación* o *asociación* lineal entre *y* y *x*, mientras que valores cercanos a cero indican poca o nula correlación (Walpole *et al.*, 2007).

Una estimación muestral de ρ, y que en adelante representaremos por *r* está dada por la expresión

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}},$$

Con frecuencia se acostumbra denominar al estimador *r* como coeficiente de correlación producto-momento de Pearson, o simplemente coeficiente de correlación muestral.

Es preciso realizar inferencias acerca de ρ, para evaluar si la asociación lineal entre la variable respuesta y la variable explicativa es significativa, pues como ya se mencionó antes, se desea obtener valores de $\rho = \pm 1$, pues cuando el valor de ρ es cero $\beta_1 = 0$, y la relación lineal entre la variable

dependiente y la variable regresora no sería significativa. Así, un procedimiento de prueba de hipótesis importante de contrastar es el relacionado con las siguientes hipótesis

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Cuyo estadístico de prueba que sigue una distribución *t* de *student* con *n* – 2 grados de libertad está dado por la expresión

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \ ,$$

y el rechazo de H_0 se da cuando $t \geq t_{\alpha/2;n-2}$ o $t \leq -t_{\alpha/2;n-2}$.

Lógicamente, se espera poder rechazar siempre la hipótesis nula para garantizar la significancia de la relación lineal entre nuestra variable respuesta y nuestra variable explicativa.

Se debe ser cuidadoso cuando se interpreta el coeficiente de correlación, pues es común que un valor cercano a cero en muchas ocasiones sea considerado como una falta de relación entre la variable dependiente y la variable regresora; sin embargo, esto no es del todo cierto, pues el coeficiente de correlación mide el *grado de relación lineal* entre las variables, de allí que un valor de *r* cercano a cero no es evidencia de la falta de una fuerte relación, sino solo de la ausencia de una relación estrictamente lineal (Devore, 2008), es decir, que puede existir otro tipo de relación funcional que se ajuste a los datos diferente a la lineal. En la Figura 8.4 se ilustra varios diagramas de dispersión con valores resultantes de *r* diferentes.

En la interfaz de programación de R, la determinación del coeficiente de correlación entre dos variables *x* y *y*, se realiza a través de la función ***cor*** del paquete básico de instalación del software, en cuyos argumentos solo basta con especificar las variables a quienes se les aplicará la función y el método de cálculo del coeficiente, siguiendo la siguiente instrucción de programación

```
cor(x, y = NULL, method = c("pearson", "kendall", "spearman"))
```

Donde *method*¸ corresponde al método de cálculo del coeficiente de correlación, que por defecto fija el método de Pearson.

Análogamente, el procedimiento de prueba de hipótesis para la correlación, se aplica en R haciendo uso de la función ***cor.test***, siguiendo la siguiente sintaxis de programación

```
cor.test(x, y, alternative = c("two.sided", "less",
"greater"), method = c("pearson", "kendall", "spearman"),
conf.level = 0.95)
```

Donde ***alternative*** establece la hipótesis alternativa, que para nuestros intereses siempre fijaremos como "*two.sided*" (bilateral) y *conf.level*, fija el nivel de confiabilidad de la prueba.

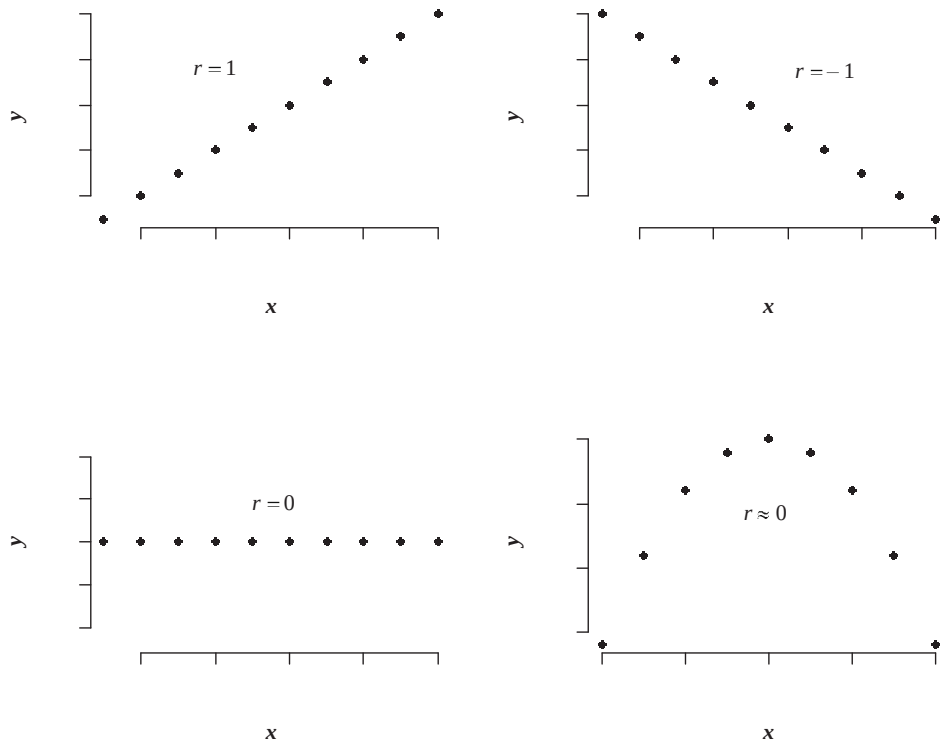


Figura 8.4. Diagramas de dispersión entre *x* y *y* con diferentes valores de *r*.

8.2.5.3. Análisis de varianza en los modelos de regresión lineal simple

Un procedimiento alternativo al estudiado en la sección 8.2.4 para analizar la calidad de la recta de regresión estimada o examinar la utilidad del modelo, se maneja a través del análisis de varianza que se introdujo en el capítulo 7. Es decir, el ANOVA es un procedimiento utilizado para probar las hipótesis

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0,$$

A partir de la partición de la variabilidad total de la variable respuesta en diferentes componentes, que para el modelo de regresión lineal simple, una componente refleja la cantidad de variación de la variable respuesta que es explicada por el modelo, y una segunda componente refleja la cantidad de variación de las observaciones alrededor a de la recta de regresión; esta partición se representó simbólicamente por

$$SST = SSR + SSE .$$

Para probar las hipótesis anteriores siguiendo el modelo ANOVA, se debe calcular y el valor del siguiente estadístico de prueba

$$f = \frac{SSR}{s^2} ,$$

donde

$$s^2 = \frac{SSE}{n-2} ,$$

y se rechaza H_0 al nivel de significancia α cuando $f \geq f_{\alpha(1,n-2)}$.

Como se ha acostumbrado, los resultados del ANOVA se presentan de manera resumida a través de un formato tabular como el que se muestra en la Tabla 8.1.

Tabla 8.1. Análisis de varianza para la prueba de $\beta_1 = 0$.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*f* calculada
Regresión	SSR	1	SSR	$f = \frac{SSR}{s^2}$
Error	SSE	n - 2	$s^2 = \frac{SSE}{n-2}$	
Total	SST	n - 1	-	

La prueba *f* del análisis de varianza proporciona exactamente el mismo resultado que la prueba *t* de utilidad del modelo, y la equivalencia entre ambas pruebas está dado por las siguientes expresiones

$$t^2 = f_{(1,n-2)} \; y \; t^2_{\alpha/2,n-2} = f_{\alpha(1,n-2)}.$$

En R, la tabla ANOVA del modelo de regresión lineal simple para validar la calidad del modelo ajustado, se construye a través de la función ***anova***, en cuyo argumento se indica el modelo de regresión lineal construido con la función ***lm***, siguiendo la siguiente línea de programación

```
anova(lm (y ~ x))
```

Ejemplo 8.2. Para el modelo de regresión ajustado determinado en el Ejemplo 8.1, aplicar la prueba de utilidad del modelo lineal y los contrastes de calidad de ajuste del mismo.

Solución

Iniciaremos contrastando la nulidad de la pendiente de la recta de regresión poblacional, expresado por las siguientes hipótesis

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Para determinar el estadístico de prueba de este procedimiento se requiere del cálculo de la varianza del error estimado, que su vez depende del cálculo de las expresiones S_{xx}, S_{xy} y S_{yy}.

Con $\bar{x} = 6.04/11 = 0.549$ y $\bar{y} = 30000/11 = 2727.273$, tenemos que

$$S_{xx} = (0.39 - 0.549)^2 + (0.43 - 0.549)^2 + \cdots + (0.76 - 0.549)^2$$

$$S_{xx} = 0.141$$

$$S_{xy} = (0.39 - 0.549)(2100 - 2727.273) + \cdots + (0.76 - 0.549)(3400 - 2727.273)$$

$$S_{xy} = 402.273$$

$$S_{yy} = (2100 - 2727.273)^2 + (2350 - 2727.273) + \cdots + (3400 - 2727.273)^2$$

$$S_{yy} = 1276818.18$$

De esta forma, la suma de cuadrados del error *SSE*, es igual a

$$SSE = 1276818.18 - (2843.1)(402.273)$$

$$SSE = 133115.814$$

y la varianza y desviación estándar del error estimadas corresponden a

$$s^2 = \frac{133115.814}{9} \therefore s^2 = 14790.646$$

$$\Rightarrow s = 121.617$$

Luego, el estadístico de prueba para nuestro contraste esta es igual a

$$t = \frac{2843.1}{121.617/\sqrt{0.141}} \therefore t = 8.778$$

De lo anterior, como $t = 8.778 > t_{0.025;9} = 2.262$, se rechaza H_0 con un nivel de significancia de 0.05. Esto nos permite concluir con una confianza del 95% que la pendiente de la recta de regresión poblacional es significativamente diferente de cero, lo que se traduce en una buena utilidad de nuestro modelo lineal ajustado, es decir, una buena relación lineal entre las variables estudiadas, permitiendo realizar predicciones con mucha precisión de la densidad media de CF para uno o varios valores de las concentraciones de NH_4, por ejemplo, para una concentración de NH_4

de 0.55, a través del modelo lineal estimado, las densidad media de CF es de 2730 NMP/100 mL.

Ahora determinaremos una medida de la calidad de ajuste de nuestro modelo lineal, más específicamente el coeficiente de determinación R^2, que como se discutió antes mide el ajuste de la recta de regresión estimada a las observaciones, a través de la cantidad de variabilidad total de las observaciones que es explicada por el modelo estimado, para nuestro ejemplo el valor de R^2 es

$$R^2 = 1 - \frac{133115.814}{1276818.18} \therefore R^2 = 0.8957$$

Este valor de R^2 indica que un 89.57% de la variabilidad de la densidad media de CF es explicada por el modelo de regresión ajustado, en otras palabras, nuestro modelo estimado se ajusta bastante bien a las observaciones. Esto se evidencia en la cercanía que muestra la recta de regresión a las observaciones (puntos del diagrama de dispersión) de la Figura 8.3.

La salida de resultados de R para la prueba de utilidad del modelo y el coeficiente de determinación R^2 se muestra a continuación

```
> summary(Reg)

Call:
lm(formula = CF ~ NH4)

Residuals:
    Min      1Q  Median      3Q     Max
-174.96  -64.66   26.02   51.03  190.73

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1166.2      181.3   6.433 0.000121 ***
NH4           2843.1      323.3   8.793 1.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 121.6 on 9 degrees of freedom
Multiple R-squared:  0.8957,    Adjusted R-squared:  0.8842
F-statistic: 77.33 on 1 and 9 DF,  p-value: 1.032e-05
```

De esta salida de resultados, cabe comentar que las diferencias entre el valor del estadístico *t* de la prueba de nulidad de la pendiente se debe a la

pérdida de decimales en las operaciones de redondeo de cifras decimales. No obstante, el examen del p-valor proporciona bases suficientes para concluir que la pendiente de la recta de regresión poblacional es significativamente distinta de cero. Por otra parte, la salida de resultados para R^2, corresponde con el hallado a través de cálculos mecánicos.

Ahora, determinaremos otra medida de calidad de ajuste de nuestro modelo estimado, esta vez examinaremos el grado de asociación (relación) lineal entre nuestra variables respuesta, densidad de CF, y nuestra variable explicativa, concentración de NH_4, a través del coeficiente de correlación *r*, que para nuestro ejemplo toma el valor

$$r = \frac{402.273}{\sqrt{(0.141)(1276818.18)}}$$

$$r = 0.948$$

Este valor de *r*, por su cercanía a la unidad, indica la existencia de una fuerte asociación lineal positiva entre la densidad media de CF y las concentraciones de NH_4, es decir, para valores crecientes de NH_4, se obtendrán también valores altos de densidad de CF.

A continuación mostramos la salida de R, para el cálculo del coeficiente de correlación muestral. Adviértase que las diferencias al nivel de las milésimas, obedece a la perdidas de cifras de decimales por causa del redondeo.

```
> cor(CF,NH4)
[1] 0.9464372
```

A nivel muestral la relación lineal entre las variables estudiadas es alta. Sin embargo, es prudente realizar un procedimiento de prueba de hipótesis sobre la significancia de esta relación a nivel poblacional, contrastando las siguientes hipótesis

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

El valor del estadístico de prueba para estas hipótesis es

$$t = \frac{0.948\sqrt{11-2}}{\sqrt{1-(0.948)^2}} \therefore t = 8.936$$

Como $t = 8.936 > t_{0.025;9} = 2.262$, se rechaza a un nivel de significancia de 0.05. Es decir, la verdadera correlación entre la densidad media de CF y las concentraciones de NH_4 es significativamente diferente de cero, lo que se traduce en una buena relación lineal entre estas variables.

La salida de resultados de R para esta prueba de hipótesis se muestra en el siguiente recuadro

```
> cor.test(CF,NH4)

        Pearson's product-moment correlation

data:  CF and NH4
t = 8.7935, df = 9, p-value = 1.032e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8017521 0.9863295
sample estimates:
      cor
0.9464372
```

De esta salida es de utilidad observar el valor del estadístico de prueba y el p-valor, este último aporta la evidencia necesaria para rechazar la hipótesis nula y concluir con una confiabilidad del 95% que la correlación entre las variables estudiadas es significativamente diferente de cero.

Por último, ilustraremos la aplicación de la tabla ANOVA para probar la hipótesis de nulidad de la pendiente, es decir, como procedimiento alternativo para probar la hipótesis de utilidad del modelo lineal, $H_0 : \beta = 0$, contra la alternativa bilateral $H_1 : \beta \neq 0$. Con $SST = 1276818.18$ y $SSE = 133115.814$, $SSR = 1143702.366$. Además, puesto que $s^2 = 14795.116$, la tabla ANOVA resultante es

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*f* calculada
Regresión	1143702.366	1	1143702.366	77.30
Error	133115.814	9	14795.116	
Total	1276818.18	10	-	

Ahora, como $f = 77.30 < f_{0.05\,(1,9)} = 5.12$, se rechaza H_0 con un nivel de significancia de 0.05 y se concluye que la pendiente de la recta de regresión población es significativamente diferente de cero, con una confiabilidad del 95%.

A continuación, se muestra la salida de resultados de R para esta tabla ANOVA

```
> anova(Reg)
Analysis of Variance Table

Response: CF
          Df  Sum Sq Mean Sq F value    Pr(>F)
NH4        1 1143701 1143701  77.325 1.032e-05 ***
Residuals  9  133117   14791
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obsérvese, que los resultados mostrados en esta salida (p-valor < 0.05) convergen con los realizados de forma mecánica, por ello se reafirman las conclusiones antes formuladas.

8.2.6. Intervalos de confianza y de predicción

Otra consecuente utilidad de la estimación de un modelo de regresión ajustado a partir de un conjunto de *n* pares de observaciones muéstrales (*x*, *y*), es realizar predicciones sobre la variable respuesta para uno o más valores de la variable explicativa.

En general, la ecuación de la recta del modelo de regresión ajustado $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, puede ser utilizada para estimar o predecir la respuesta media $\mu_{y|x_0}$ en $x = x_0$, a través de **intervalos de confianza**, o puede emplearse para pronosticar un valor único y_0 de y cuando $x = x_0$, a través de los llamados **intervalos de predicción**, que aunque similares en su función no deben confundirse y o malinterpretarse con los intervalos de confianza, pues estos realizan estimaciones por intervalo sobre el verdadero valor medio (o valor esperado) $\mu_{y|x_0}$ de la variable respuesta, cuando la variable explicativa toma un valor especifico x_0; mientras que los intervalos de predicción realizan estimaciones por intervalo para un valor particular y_0 que toma la variable respuesta, cuando se especifica un valor x_0 para la variable regresora.

Cuando se construyen intervalos de confianza, se usa la recta de regresión ajustada $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ como estimador puntual para construir intervalos de $(1-\alpha)100\%$ para $\mu_{y|x} = \beta_0 + \beta_1 x$ a través de la expresión

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{y|x_0} < \hat{y}_0 + t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución *t* de *student* con *n* – 2 grados de libertad.

Análogamente, cuando el objetivo es construir intervalos de predicción, la recta estimada $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ es usada como de costumbre, como un estimador puntual de $y = \beta_0 + \beta_1 x$. Así, un intervalos de predicción de $(1-\alpha)100\%$ para un valor único de la variable respuesta y_0 está dado por

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución *t* de *student* con *n* – 2 grados de libertad.

Al realizar consecutivamente los cálculos para los limites de los intervalos de confianza y de predicción, para diferentes valores x_0 de la muestra, se obtendrán múltiples límites de confianza y de predicción que se pueden graficar para ilustrar las regiones de estimación de los dos intervalos (de confianza y de predicción) alrededor de la recta de regresión ajustada (Figura 8.5).

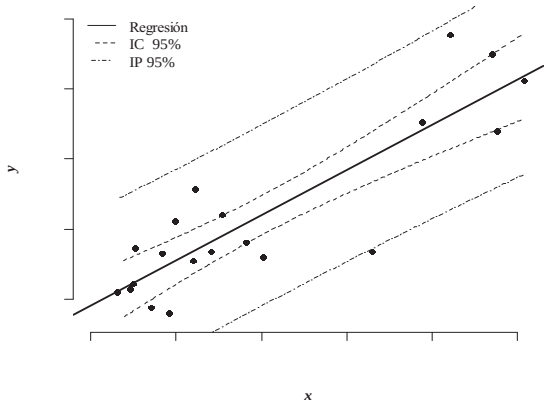


Figura 8.5. Límites de confianza para $\mu_{y|x}$ y el valor pronosticado y_0.

Cabe anotar que es aconsejable realizar las estimaciones de estos intervalos a partir de valores dentro del rango de valores muéstrales que toma la variable explicativa. El peligro de extrapolar a otros valores por encima o por debajo del rango de valores de *x*, se debe a que no se conoce el ajuste de la recta estimada por fuera de este rango.

En R, la construcción de intervalos de confianza o de predicción en los modelos de regresión se realiza a través de la función ***predict.lm*** del paquete base, adicionándole los argumentos que se muestran en la siguiente línea de código

```
Predict.lm(object, newdata, interval = c("none", "confidence",
"prediction"),level = 0.95)
```

Donde ***object***, corresponde al objeto asignado al modelo de regresión lineal (*lm (y ~ x)*), ***newdata*** es un data frame opcional que contenga una secuencia de valores de la variable explicativa, dentro del rango de valores dados para esta variable en el conjunto de *n* pares de observaciones, si no se especifica se utilizan los valores del modelo ajustado; ***interval***, especifica el tipo de intervalo que se construye para la secuencia de valores dados, los valores de este argumento pueden ser *none* si no se desea construir ningún intervalo, *confidence* si el objetivo construir intervalos de confianza y *prediction* para intervalos de predicción; por ultimo ***level*** indica el nivel de confiabilidad de los intervalos, que por defecto se asume son del 95% (0.95).

Ejemplo 8.3. Construir intervalos de confianza e intervalos de predicción del 95% para el modelo de regresión ajustado $\hat{y} = 1166.2 + 2843.1x$, que se estimó en el ejemplo 8.1.

Solución

El cálculo de las cotas superior e inferior de los intervalos de confianza y de predicción del modelo de regresión ajustado para los 9 pares de observaciones, puede resultar una tarea dispendiosa y algo extensa, por ello se ilustrará el cálculo intervalos de confianza y de predicción en el ambiente de programación de R, como se muestra en la siguiente salida de resultados, donde las cotas superior e interior de los intervalos de confianza y de predicción corresponden a las columnas llamadas *lwr* y *upr*, respectivamente.

```
> CF <-
c(2100,2350,2500,2600,2750,2700,2700,2800,3000,3100,3400)
> NH4 <-
c(0.39,0.43,0.46,0.45,0.49,0.53,0.56,0.63,0.67,0.67,0.76)
> Reg <- lm(CF~NH4)
> CI<-predict.lm(Reg,interval="confidence")
> CI
        fit      lwr      upr
1  2274.961 2132.062 2417.861
2  2388.685 2268.403 2508.967
3  2473.978 2368.495 2579.462
4  2445.547 2335.395 2555.699
5  2559.271 2465.737 2652.806
6  2672.995 2588.877 2757.113
7  2758.288 2674.955 2841.622
8  2957.305 2855.410 3059.201
9  3071.029 2949.781 3192.278
10 3071.029 2949.781 3192.278
11 3326.908 3151.761 3502.055
> IP<-predict.lm(Reg,interval="prediction")
> IP
        fit      lwr      upr
1  2274.961 1964.946 2584.977
2  2388.685 2088.423 2688.948
3  2473.978 2179.332 2768.625
4  2445.547 2149.198 2741.897
5  2559.271 2268.689 2849.854
6  2672.995 2385.306 2960.685
7  2758.288 2470.827 3045.750
8  2957.305 2663.924 3250.686
9  3071.029 2770.379 3371.680
10 3071.029 2770.379 3371.680
11 3326.908 3000.770 3653.046
```

Ahora, adjuntamos los intervalos construidos al gráfico de dispersión de la Figura 8.2, a través del trazado de líneas para los valores de concentración de NH_4 y los limites superior e inferior de los intervalos (Figura 8.6). Esto se ordena en R a través de la función ***lines***, como se muestra en las siguientes lines de comandos

```
> lines(NH4,IC[,2],lwd=2,lty=2)
> lines(NH4,IC[,3],lwd=2,lty=2)
> lines(NH4,IP[,2],lwd=2,lty=3)
> lines(NH4,IP[,3],lwd=2,lty=3)
```

Por último insertamos una leyenda que indica las líneas que representan a cada intervalo

```
> legend("topleft",legend=c("Recta de regresión","Intervalo de
confianza","Intervalo de predicción"),bty="n",lty=1:3)
```

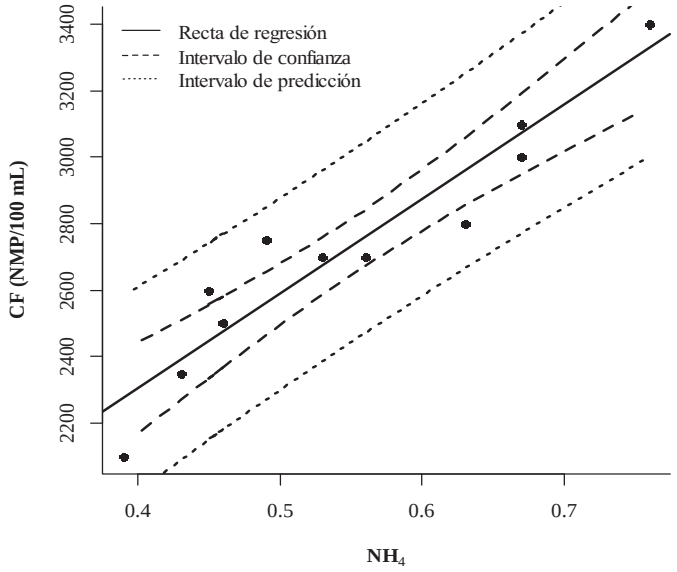


Figura 8.6. Intervalos de confianza y de predicción para los datos de densidad de CF.

8.2.7. Verificación de los supuestos del modelo de regresión

En la sección 8.2.1 se mencionaron lo requerimientos que deben ser cumplidos para garantizar la validez y confiabilidad de los resultados obtenidos en un análisis de regresión lineal. Para el caso del supuesto de normalidad de los errores, este puede ser verificado a través de cualquiera de los test que se revisaron en el capítulo 6; el supuesto de homocedasticidad e independencia de los errores, pueden ser evaluados por diversos test, pero en este texto, solo discutiremos los test de Breusch-Pagan y de Durbin-Watson, para verificar la homocedasticidad e independencia, respectivamente. El supuesto de multicolinealidad, se revisará cuando discutamos el modelo de regresión lineal múltiple.

8.2.7.1. Verificación de la homocedasticidad de los errores: Test de Breusch-Pagan

En los análisis de regresión, la verificación de igualdad de varianzas de los errores, es quizás el requisito más importante de cumplir para garantizar la confiabilidad de los resultados que se obtengan. De los múltiples test que se utilizan para la verificación de este supuesto, aquí discutiremos uno de

los más populares, propuesto por Breusch & Pagan (1979), basado en el test de los multiplicadores de Lagrange (Aitchison & Silvey, 1960), no discutido en este texto para no profundizar en cuestiones fuera de los objetivos del mismo.

La deducción de un procedimiento de prueba para la aplicación del test de Breusch-Pagan, es una tarea sencilla, aunque un poco engorrosa. Para ilustrarlo, considérese el siguiente modelo de regresión lineal, que generalizaremos a *k* variables regresoras

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

del mismo modo, supóngase que la varianza del error σ_i^2 se describe como

$$\sigma_i^2 = f\left(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_m z_{im}\right)$$

es decir, σ_i^2 es algún tipo de función de las variables *z* no estocásticas; las variables *x* pueden servir como *z*. Específicamente, supóngase que

$$\sigma_i^2 = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_m z_{im}$$

Lo anterior sugiere que σ_i^2 es una función lineal de las *z*. Si $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$, $\sigma_i^2 = \alpha_1$ que es una constante. Por consiguiente, para probar si σ_i^2 es homocedástica, se puede probar la hipótesis nula

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$$

que constituye la idea básica del test de Breusch-Pagan. El procedimiento de prueba de este test se puede resumir en los siguientes cinco pasos:

- Estimar el modelo de regresión lineal entre la variable respuesta y la variable regresora a través del método de los mínimos cuadrados y obténgase los residuales $e_1, e_2, ..., e_n$.
- Calcúlese $\hat{\sigma}_i^2 = \sum e_i^2 / n$, que es el estimador de máxima verosimilitud de σ_i^2.

- Constrúyase las variables p_i definidas como

$$p_i = e_i^2 / \hat{\sigma}_i^2$$

que es simplemente cada residual elevado al cuadrado dividido por $\hat{\sigma}_i^2$.

- Constrúyase un modelo de regresión entre las variables p_i sobre las variables *z*, dado por

$$p_i = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_m z_{im} + v_i$$

donde v_i es el término residual de esta regresión

- Obténgase la SSR_R (suma de cuadrados de la regresión del modelo ajustado entre p_i y *z*) y defínase el siguiente estadístico de prueba

$$BP = \frac{1}{2} SSR_R$$

que suponiendo que los e_i están normalmente distribuidos, se aproxima a una distribución chi-cuadrado con *m* - 1 grados de libertad. Por consiguiente, la hipótesis nula de homocedasticidad de los residuos se rechaza cuando $BP \geq \chi^2_{\alpha(m-1)}$.

El test de Breusch-Pagan, puede ser calculado en el ambiente de programación de R a través de la función ***bptest*** del paquete "*lmtest*" (Torsten *et al.*, 2015), utilizando la siguiente línea de programación

```
bptest(formula, studentize = FALSE)
```

8.2.7.2. Verificación de independencia de los errores: Test de Durbin Watson

El test conocido genéricamente como prueba ***d* de Durbin-Watson**, es la prueba mayormente conocida y utilizada para detectar correlación serial de los errores (autocorrelación), que es una estimación de falta de independencia de los mismos. Este test fue formulado por Durbin & Watson (1951), y estadísticamente se define a través de la expresión

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

que simplemente expresa la razón de la suma de las diferencias al cuadrado de residuos sucesivos sobre la *SSE* (Gujarati & Porter, 2010).

La principal ventaja que presenta este test es que se calcula a través de los residuos estimados que se determinan habitualmente en cualquier análisis de regresión. No obstante, como cualquier otro test requiere del cumplimiento de ciertos requerimientos para optimizar sus resultados, entre los que destacan que los errores estén normalmente distribuidos, el modelo de regresión incluya el término del intercepto y que no existan valores faltantes en los datos.

Derivar la distribución de probabilidad el estadístico *d*, resulta ser una tarea difícil debido a la dependencia que existe con los valores que toma la variable(s) regresoras. Por esta razón, a diferencias de otros test no existe un valor critico que defina cuando se rechaza o no la hipótesis nula de no autocorrelación de los residuos. Sin embargo, Durbin y Watson (1951), lograron encontrar un límite inferior d_L y un límite superior d_U tales que si el valor calculado de *d* cae por fuera de estos límites, pueden tomarse decisiones respecto a la presencia de autocorrelación. Además, estos límites tienen la principal característica que solo dependen del número de observaciones *n* y del número de variables explicativas (Gujarati & Porter, 2010). Estas tablas se encuentran disponibles en la Tabla A.18 del apéndice.

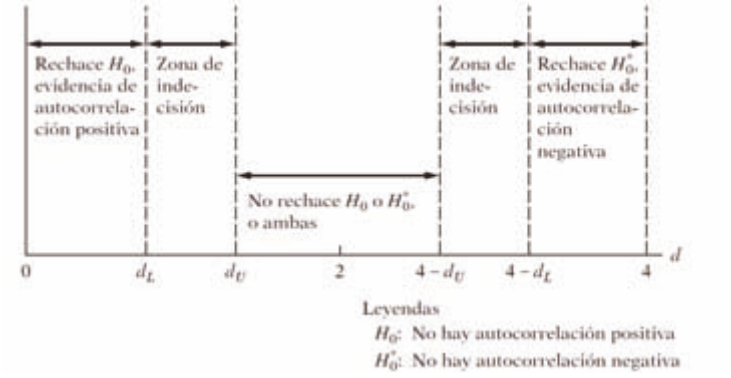


Figura 8.7. Resumen del estadístico *d* de Durbin-Watson.

La Figura 8.7 resume el procedimiento de prueba del test de Durbin-Watson, donde se observa que los límites de *d* son 0 y 4. Según Gujarati & Porter (2010), estos se determinan expandiendo *d*, para obtener

$$d = \frac{\sum e_i^2 - 2\sum e_i e_{i-1} + \sum e_{i-1}^2}{\sum e_i^2}$$

Como $\sum e_i^2$ y $\sum e_{i-1}^2$ difieren solo en una observación, son aproximadamente iguales. Por consiguiente, establecemos que $\sum e_{i-1}^2 \approx \sum e_i^2$ y la expresión anterior se puede escribir como

$$d \approx 2\left(1 - \frac{\sum e_i e_{i-1}}{\sum e_i^2}\right)$$

Ahora definimos

$$\hat{\rho} = \frac{\sum e_i e_{i-1}}{\sum e_i^2}$$

Como el coeficiente de autocorrelación muestral, un estimador de ρ. Con este estimador reexpresamos *d* como

$$d \approx 2\left(1 - \hat{\rho}\right)$$

Así, como $-1 \leq \hat{\rho} \leq 1$, implica que

$$0 \leq d \leq 4$$

De allí, que todo valor de *d* debe caer dentro de estos límites. De lo anterior es evidente que cuando $\hat{\rho} = 0$, $d = 2$; es decir, no hay autorrelación serial entre los residuos. Por lo que en la práctica siempre esperamos que *d* esté alrededor de 2. Si $\hat{\rho} = 1$, $d = 0$, lo que indica que existe autocorrelación positiva entre los residuos. Por último, si $\hat{\rho} = -1$, $d = 4$, indicativo de existencia de autocorrelación negativa perfecta entre los residuos.

Para ayudar a la toma de decisiones en la aplicación del test de Durbin-Watson, en la Tabla 8.2, se muestran unas reglas de decisión, que apoyan al resumen del test presentado en la Figura 8.4. En esta tabla sale a evidencia que el test de Durbin-Watson posee una desventaja y es que presenta zonas de indecisión que limitan su aplicabilidad. Por otro lado, según Gutiérrez & de la Vara (2008), otra desventaja que posee este test es tener el inconveniente de detectar sólo la estructura de correlación de residuos consecutivos. No detecta correlaciones entre residuos no consecutivos en el tiempo que también violan el supuesto de independencia. Este tipo de correlación ocurre en un experimento cuando la contaminación de una medición a otra no se refleja de inmediato, sino que actúa con retardo.

Tabla 8.2. Reglas de decisión del test *d* de Durbin-Watson.

Hipótesis	Regla de decisión	Decisión	Conclusión
$H_0: \rho = 0$ $H_1: \rho \neq 0$	Si $d < d_L$ o $4-d < d_L$	Rechazar H_0 al nivel 2α	Hay correlación entre los errores.
	Si $d > d_U$ o $4-d > d_U$	No rechazar H_0 al nivel 2α	No existe indicio de correlación.
	Si $d_L \leq d \leq d_U$	No hay decisión.	No se concluye.
$H_0: \rho = 0$ $H_1: \rho > 0$	Si $d < d_L$	Rechazar H_0 al nivel α	Hay correlación serial positiva entre los errores.
	Si $d > d_U$	No rechazar H_0 al nivel α	No existe indicio de correlación positiva.
	Si $d_L \leq d \leq d_U$	No hay decisión.	No se concluye.
$H_0: \rho = 0$ $H_1: \rho < 0$	Si $4-d < d_L$	Rechazar H_0 al nivel α	Hay correlación serial negativa entre los errores.
	$4-d > d_U$	No rechazar H_0 al nivel α	No existe indicio de correlación positiva.
	Si $d_L \leq d \leq d_U$	No hay decisión.	No se concluye.

En R, el test de *d* de Durbin-Watson se puede aplicar a través de la función ***dwtest*** del paquete "*lmtest*" (Torsten *et al.*, 2015), siguiendo la sencilla línea de código que se muestra a continuación

```
dwtest(formula, alternative = c("greater", "two.sided", "less"))
```

Donde el argumento ***formula*** toma la forma $y \sim x$, y ***alternative*** indica si se desea probar la existencia de autocorrelaciones seriales de los errores positivas (*greater*), negativas (*less*) o ambas (*two.sided*).

Ejemplo 8.4. Para los datos y el modelo de regresión lineal ajustado del ejemplo 8.1 aplicar las pruebas de verificación de los supuestos del modelo lineal.

Solución

La aplicación de los contrastes de verificación de los supuestos del modelo de regresión lineal, requiere del cálculo previo de los residuos del modelo, que según lo visto en secciones anteriores se determinan matemáticamente como $e_i = y_i - \hat{y}_i$. Así, para nuestras observaciones, los residuales serían

y_i	x_i	$\hat{y}_i$	e_i
2100	0,39	2275.009	-175.009
2350	0,43	2388.733	-38.733
2500	0,46	2474.026	25.974
2600	0,45	2445.595	154.405
2750	0,49	2559.319	190.681
2700	0,53	2673.043	26.957
2700	0,56	2758.336	-58.336
2800	0,63	2957.353	-157.353
3000	0,67	3071.077	-71.077
3100	0,67	3071.077	28.923
3400	0,76	3326.956	73.044

Estos se determinan en R, a través de indexación sobre el objeto que contiene nuestro modelo de regresión, como sigue

```
> CF <-
c(2100,2350,2500,2600,2750,2700,2700,2800,3000,3100,3400)
> NH4 <-
c(0.39,0.43,0.46,0.45,0.49,0.53,0.56,0.63,0.67,0.67,0.76)
> Reg <- lm(CF~NH4)
> Res<-Reg$residuals
> Res

     1          2          3          4          5          6
-174.96145  -38.68543   26.02159  154.45258  190.72860
27.00463
     7          8          9         10         11
-58.28836 -157.30532  -71.02930   28.97070   73.09175
```

La salida de resultados para este procedimiento es un vector de datos que contiene los residuos del modelo. Adviértase, que las discrepancias en los valores de la salida de R con los calculados obedecen a la perdida de cifras decimales en la estimación de los parámetros del modelo de regresión, por ello para hacer corresponder los resultados de las pruebas que aplicaremos a continuación, con las salidas de R, trabajaremos con los residuales proporcionados por R.

Ya discutimos que el examen de la normalidad de los residuos, se lleva a cabo con cualquiera de las pruebas que se introdujeron en el capítulo 6.

Para no extendernos en complejos y fatigantes cálculos que ya hemos visto antes, la verificación del supuesto de normalidad de los errores la ejemplificaremos solo en el entorno de R, inclinándonos por el uso del test de Shapiro-Wilk, dado el reducido número de observaciones que tenemos

```
> shapiro.test(Res)

        Shapiro-Wilk normality test

data:  Res
W = 0.9617, p-value = 0.7931
```

El p-valor de la salida de resultados de este test nos permite afirmar con un 95% de confiabilidad que los residuos de nuestro modelo lineal siguen una distribución normal, es decir, que nuestro modelo cumple con el primer requisito que garantiza la confiabilidad de nuestras predicciones.

El examen de la normalidad de los errores, usualmente es diagnosticado por los usuarios de la estadística de forma descriptiva, apoyándose en un gráfico cuantil-cuantil (QQ plot) de normalidad, el cual realiza una comparación de los valores ordenados de los errores estandarizados (e_i/s) con los cuantiles teóricos de la distribución normal. Si los errores presentan un buen ajuste a una distribución normal, los puntos de la gráfica resultante deben seguir un patrón lineal con un ángulo de inclinación de aproximadamente 45º. A continuación, mostramos el procedimiento de programación en R para la construcción del QQ plot de los errores del modelo lineal ajustado, a través de la función ***qqnorm*** y ***qqlines***. Esta última permite insertar al gráfico la línea de ajuste perfecto de los valores dados con lo cuantiles teóricos de la distribución normal.

```
> Res.Est<-Res/121.6
> qqnorm(Res.Est, xlab="Cuantiles teóricos", ylab= "Residuos
estandarizados", main="", font.lab=2,pch=16)
> qqline(Res.Est,lwd=2,lty=2)
```

Nótese que la función qqnorm, no realiza el estandarizado de los residuos, por ello este es un procedimiento que se debe realizar antes de ejecutar esta función, de no hacerlo se construiría el gráfico con base en los valores brutos de los errores. El gráfico generado, se muestra en la Figura 8.8, en él se observa que, salvo pocas observaciones, los residuos graficados se ajustan bien a la línea de ajuste perfecto de la distribución normal. De esta manera, el gráfico apoya los resultados obtenidos con el test de Shapiro-Wilk sobre los residuos.

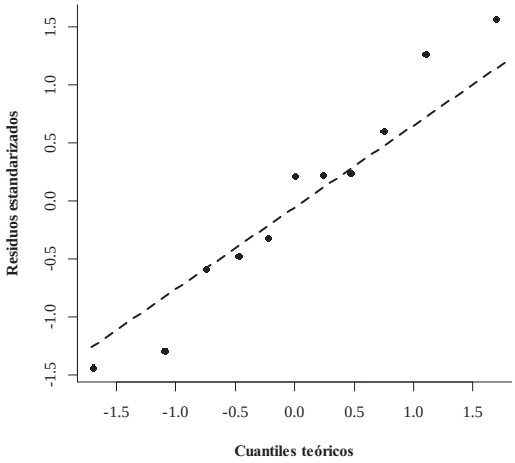


Figura 8.8. QQ plot para los errores del modelo ajustado de los datos de densidad de CF.

Ahora, realizaremos la verificación del supuesto de homocedasticidad de los errores, aplicando el test de Breusch-Pagan para tal propósito. Para ello, nos interesamos en probar las siguientes hipótesis

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_m$$

H_1 : Al menos dos α_m son diferentes.

Iniciamos calculando la estimación de máxima verosimilitud de la varianza del error como sigue

$$\hat{\sigma}_i^2 = \frac{1}{11}\left[(-174.961)^2 + (-38.685)^2 + \cdots + (73.092)^2\right]$$

$$\hat{\sigma}_i^2 = 12101.438$$

Como paso siguiente, definimos los valores de $p_i = e_i / \hat{\sigma}_i^2$, que se muestran enseguida

p_i :	2.53	0.12	0.56	1.97	3.00	0.06	0.28	2.04	0.41	0.06	0.44
	0	4	0	1	6	0	1	5	7	9	1

Luego, ajustamos un modelo de regresión entre las p_i contra los valores de nuestra variable explicativa, concentraciones de NH_4, y de su tabla ANOVA,

obtenemos que la suma de cuadrados de la regresión del modelo ajustado entre p_i y NH_4 es $SSR_R = 1.602$.

De lo anterior, el valor del estadístico de prueba es

$$BP = \frac{1}{2}(1.602) \therefore BP = 0.801$$

Ahora, como $BP = 0.801 < \chi^2_{0.05(9)} = 18.307$, no se rechaza H_0 a un nivel de significancia de 0.05. Es decir, se puede concluir a un nivel de confiabilidad del 95% que los errores del modelo de regresión ajustado entre las densidades de CF y las concentraciones de NH_4 cumplen con el requisito de homocedasticidad o igualdad de varianza de los mismos.

La salida de resultados de R para el test de Breusch-Pagan, a través de la función ***bptest*** se muestra a continuación

```
> library(lmtest)
> bptest(CF~NH4,studentize=FALSE)

        Breusch-Pagan test

data:  CF ~ NH4
BP = 0.8002, df = 9, p-value = 0.371
```

De esta salida de R, se observa que se llegan a los mismos resultados anteriores, es decir, con un nivel de confiabilidad del 95%, no se rechaza la hipótesis nula y se concluye que los residuos de nuestro modelo de regresión son homocedásticos.

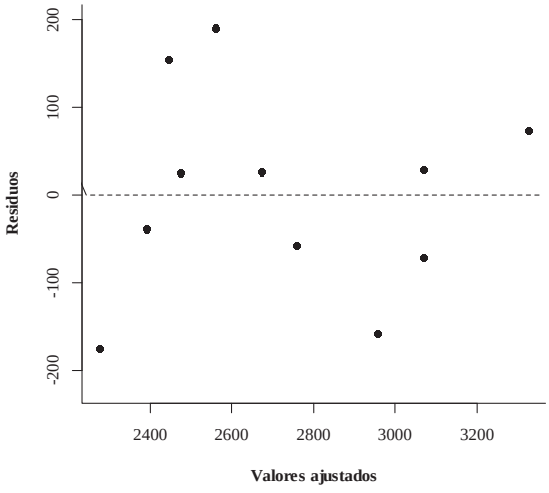


Figura 8.9. Gráfico de residuos contra valores ajustados.

Análogamente al examen de normalidad de los residuos, la homocedasticidad de los mismos se puede evaluar de manera informal o descriptiva a través de un gráfico de dispersión de los residuos contra los valores ajustados o predichos del modelo $\hat{y}_i$. En este gráfico, cuando los residuos no exhiben ningún patrón o relación sistemática con los valores ajustados y se distribuyen de forma equitativa alrededor de la recta $x = 0$, es indicativo de homocedasticidad de los errores o presencia de varianzas homogéneas. Por el contrario, cuando el gráfico muestra alguna relación entre los e_i y los $\hat{y}_i$, esto se toma como indicativo de heterocedasticidad de los errores. Para nuestro estudio de caso, el gráfico de e_i *vs* $\hat{y}_i$ se muestra en Figura 8.9, en ella se observa que los residuos no exhiben ningún patrón en particular, por tanto, se complementa los resultados del test de Breusch-Pagan, donde se concluyó homocedasticidad de los errores.

La construcción de este gráfico en R es simple, siguiendo las siguientes líneas de programación, donde el objeto *Ajus*, contiene los valores ajustados del modelo lineal, extraídos mediante indexación.

```
> CF <-
c(2100,2350,2500,2600,2750,2700,2700,2800,3000,3100,3400)
> NH4 <-
c(0.39,0.43,0.46,0.45,0.49,0.53,0.56,0.63,0.67,0.67,0.76)
> Reg <- lm(CF~NH4)
> Res<-Reg$residuals
> Ajus<-Reg$fitted
> plot(Ajus,Res,pch=16,xlab="Valores ajustados",
ylab="Residuos", font.lab=2)
> abline(h=0,lty=2)
```

Ahora, para terminar con la verificación del cumplimiento de los supuestos de nuestro modelo de regresión lineal ajustado, aplicaremos el test de Durbin-Watson para la comprobación de autocorrelación serial, o no independencia de los errores. Para ello, el valor del estadístico de prueba de *d* es

$$d = \frac{(-38.685+174.691)^2 + (26.022+38.685)^2 + \cdots + (73.092-28.971)^2}{133115.814}$$

$$d = 0.820$$

y los limites inferior y superior para el estadístico *d*, para un modelo lineal con una sola variable explicativa son $d_L = 0.927$ y $d_U = 1.324$, respectivamente. De esta forma, para las hipótesis

H_0 : no existe correlación positiva o negativa en los errores.
H_1 : existe correlación positiva o negativa en los errores.

Como $d = 0.820 < d_L = 0.927$, se rechaza H_0 a un nivel de significancia de 0.10, y se concluye la existencia de correlación serial positiva o negativa de los errores del modelo de regresión ajustado entre las densidades de CF y las concentraciones de NH_4.

La salida de resultados de R para el test de Durbin-Watson se muestra a continuación. Las discrepancias en los resultados obtenidos en esta salida, se deben a las operaciones de redondeo de cifras decimales al realizar los cálculos de forma mecánica.

```
> library(lmtest)
> dwtest(CF~NH4, alternative = "two.sided")

        Durbin-Watson test

data:  CF ~ NH4
DW = 0.7801, p-value = 0.003593
alternative hypothesis: true autocorrelation is not 0
```

El p-valor = 0.004 de esta salida, muestra suficiente evidencia para rechazar la hipótesis nula de existencia de autocorrelación.

Gráficamente, lo anterior es ilustrado a través de un gráfico de dispersión de los residuos estandarizados respecto el orden de ocurrencia de los mismos, el gráfico obtenido, se asemeja al gráfico de diagnóstico para la evaluación de la heterocedasticidad de los errores. Cuando no existe autocorrelación en los errores, y por lo tanto independencia de estos, el gráfico resultante muestra puntos distribuidos aleatoriamente alrededor del eje *x*, de otro modo, cuando los residuos se encuentran autocorrelacionados, el gráfico exhibirá patrones o tendencias de los residuos. Para nuestro modelo de regresión, el gráfico de residuos estandarizados en función del tiempo mostrado en la Figura 8.10, de él se observa que lo residuos siguen un patrón aproximadamente sinuidal alrededor del eje *x*, es decir, lo residuos del modelo de regresión ajustado

entre las densidades de CF y las concentraciones de NH_4 presentan autocorrelación, y por lo tanto, no son independientes.

Las líneas de programación usadas en R para la construcción de este gráfico se muestran enseguida

```
> Res<-Reg$residuals
> Res.Est<-Res/121.6
> par(family="serif",bty="l")
> plot(Res.Est,pch=16,xlab="Orden de los
residuos",ylab="Residuos estandarizados",font.lab=2)
> abline(h=0,lty=2)
```

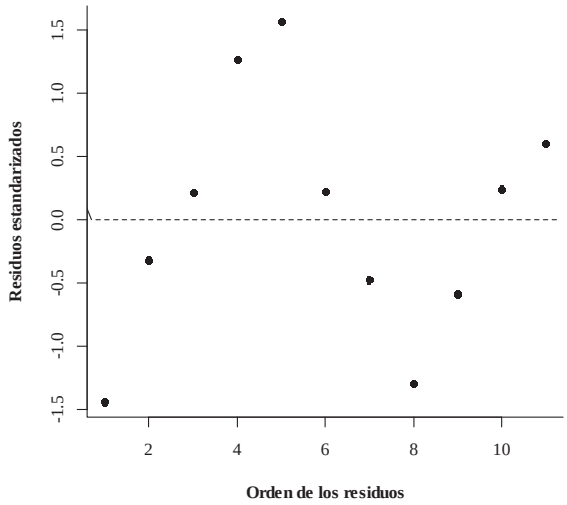


Figura 8.10. Gráfico de residuos estandarizados respecto al orden de ocurrencia.

8.2.8. Transformaciones

Cuando establecemos modelos de regresión entre dos variables *x* y *y*, siempre esperamos que estas intervengan en el modelo de forma lineal. Sin embargo, esto no siempre ocurre y se hace necesario trabajar con modelos alternativos en los que *x* o *y* (o ambas) intervengan en forma no lineal, debido a exigencias teóricas del fenómeno estudiado o especialmente cuando existe violación de las suposiciones del modelo lineal (Walpole *et al.*, 2007). Lo anterior sugiere el uso de transformaciones de los datos de *x* o *y* (o ambas) a través de ciertas relaciones funcionales conocidas que permitirán corregir la falta de ajuste del modelo ajustado o el cumplimiento

de los requisitos del modelo. La necesidad de llevar a cabo transformaciones de los datos en el caso de la regresión lineal simple es relativamente fácil de diagnosticar, debido a que las simples gráficas de dispersión entre *x* y *y* brindan un panorama verdadero de la forma en que se encuentran relacionadas funcionalmente estas variables.

Una vez realizada la transformación de los datos, el modelo ajustado resultante no debe verse como un modelo no lineal, dado que este seguirá siendo lineal en sus parámetros β_0 y β_1, y por ello conserva su carácter lineal, siendo válido aún, la aplicación de todos los test que se revisaron en las secciones anteriores para verificar la utilidad y calidad del modelo ajustado.

Como se discutió en el capítulo 7, son muchas las transformaciones disponibles para los datos. No obstante, existen algunas de ellas de uso más concurrido por ofrecer mayor facilidad de las interpretaciones de los resultados y por su fácil aplicabilidad. En la Tabla 8.3 se muestra un resumen de las principales transformaciones funcionales de los datos que producen una regresión lineal luego de su aplicación y correcciones en el cumplimiento de las premisas del modelo. En la Figura 8.11, se ilustra las funciones que se enlistan en la tabla 8.3, las cuales constituyen una buena herramienta a la hora de decidir que transformación utilizar.

Tabla 8.3. Algunas transformaciones útiles en la regresión lineal.

Relación funcional	Transformación	Forma de la regresión
Exponencial: $y = \beta_0 e^{\beta_1 x}$	$y' = \ln y$	$y' = \ln \beta_0 + \beta_1 x$
Potencia: $y = \beta_0 x^{\beta_1}$	$y' = \log y$; $x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y = \beta_0 + \beta_1 x'$
Recíproca: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x' = \frac{1}{x}$	$y = \beta_0 + \beta_1 x'$

Otro enfoque para realizar transformaciones y conseguir mayor calidad de ajuste, son las familias de transformaciones Box-Cox para la variable dependiente que se discutió en el capítulo 7. Su aplicación en el ambiente de programación de R se realiza exactamente igual a como se discutió antes. A continuación mostraremos un ejemplo de ajuste de un modelo de regresión lineal, donde sea necesario realizar transformaciones de los datos, este procedimiento se realizará de forma mecánica y luego se realizará la respectiva aplicación en el entorno de R.

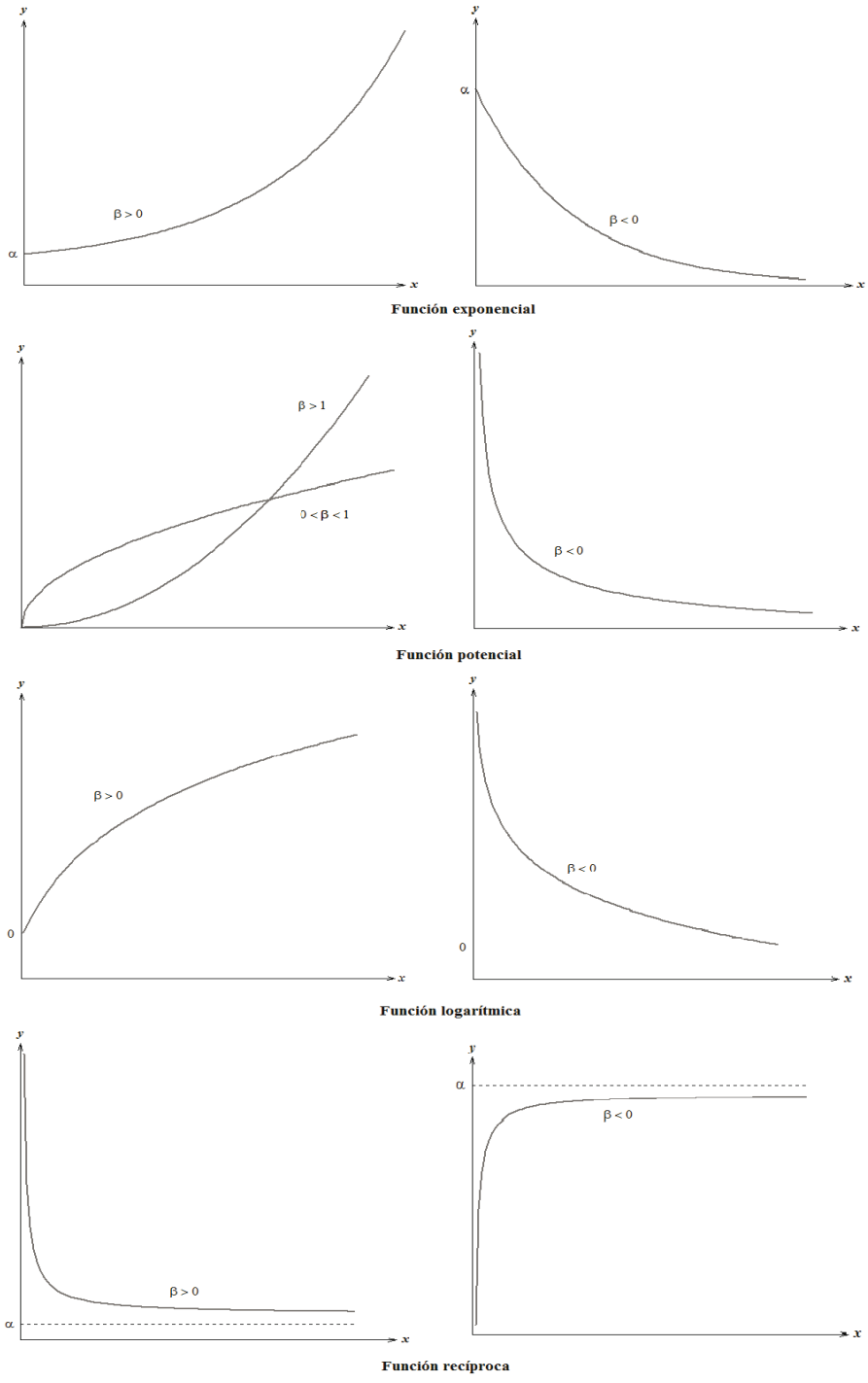


Figura 8.11. Graficas de las funciones enlistadas en la Tabla 8.3

Ejemplo 8.5. Considérese los datos que se presentan a continuación sobre la longitud (cm) de 20 individuos de cierta especie de peces y el contenido de mercurio (mg/kg) encontrado en sus tejidos. A partir de ellos ajuste un modelo de regresión lineal que permita realizar pronósticos del contenido de este metal en tejido, en función de la longitud alcanzada por los individuos.

x = Longitud (cm)	y = Mercurio (mg/kg)	x = Longitud (cm)	y = Mercurio (mg/kg)
89.2	0.78	70	0.53
83	2.19	57	0.16
95.5	1.20	84	0.54
63	0.18	74	0.52
104	3.46	87	1.01
105	1.55	91.8	0.90
101	2.16	52	0.22
105	1.99	91.5	1.33
89	1.08	94.5	1.05
100	1.81	97	1.26

Solución

Como en cualquier análisis de regresión, siempre es recomendable realizar un gráfico de dispersión entre las variables estudiadas con el objeto de observar la relación funcional que puede existir entre ellas y establecer si se requiere o no, la realización de transformaciones. El gráfico resultante para nuestro análisis se ilustra en la Figura 8.12, el cual obedece como se ha visto antes a las siguientes sentencias de programación.

```
>Long<-
c(89.2,83,95.5,63,104,105,101,105,89,100,70,57,84,74,87,
91.8,52,91.5,91.5,97)
>Merc<-
c(0.78,2.19,1.20,0.18,3.46,1.55,2.16,1.99,1.08,1.81,0.53,
0.16,0.54,0.52,1.01,0.90,0.22,1.33,1.05,1.26)
> par(family="serif",bty="l")
> plot(Long,Merc,pch=16,xlab="Longitud (cm)",ylab="Contenido
de Mercurio (mg/kg)",font.lab=2)
```

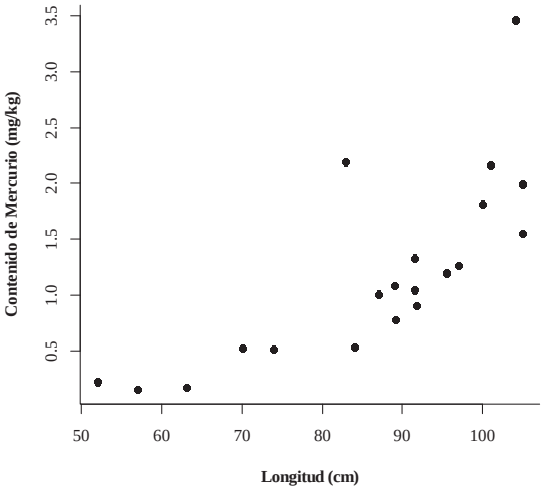


Figura 8.12. Diagrama de dispersión entre la longitud de los individuos y el contenido de mercurio en sus tejidos.

Del gráfico es posible observar que si ajustamos un modelo de regresión lineal entre las variables originales, este no se ajustaría bien a cada uno de los pares de observaciones, se violarían los supuestos del modelo y se reduciría su eficiencia a la hora de realizar pronósticos. Para ejemplificar lo anterior, obsérvese la salida de resultados de R que se muestra a continuación

```
> Reg<-lm(Merc~Long)
> summary(Reg)

Call:
lm(formula = Merc ~ Long)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55641 -0.34937 -0.13506  0.09402  1.57475

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.216728   0.710749  -3.119 0.005929 **
Long         0.039442   0.008088   4.877 0.000121 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.556 on 18 degrees of freedom
Multiple R-squared:  0.5692,    Adjusted R-squared:  0.5453
F-statistic: 23.78 on 1 and 18 DF,  p-value: 0.0001213
```

Nótese que la prueba de utilidad del modelo resulta ser significativa, a un nivel de significancia de 0.05 (p-valor < 0.05), no obstante la calidad de

ajuste del mismo, medido a través del coeficiente de determinación R^2, indica que el modelo ajustado está explicando el 56.92% de la variabilidad total de las observaciones, una cifra muy regular cuando se tienen propósitos de estimación o pronostico (Figura 8.13).

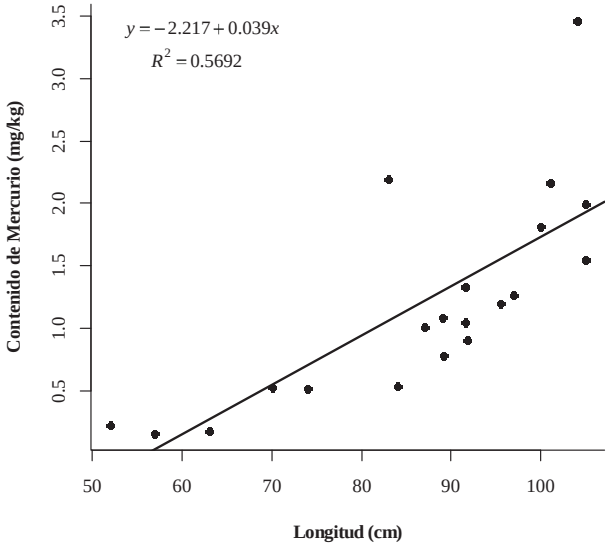


Figura 8.13. Modelo de regresión ajustado para los datos originales del Ejemplo 8.5.

Así mismo, el gráfico de dispersión entre las variables estudiadas sugiere que una transformación de tipo exponencial podría resolver los problemas de carencia de ajuste del modelo lineal con los datos originales. De allí, que en adelante emprenderemos el análisis del problema aplicando esta transformación. Recuérdese que este tipo de transformaciones solo es aplicada a la variable respuesta, en nuestro caso al contenido de mercurio en tejido de los peces estudiados, como se muestra en la siguiente tabla

x	y	$y' = \ln(y)$
89.2	0.78	-0.1079
83	2.19	0.3404
95.5	1.20	0.0792
63	0.18	-0.7447
104	3.46	0.5391
105	1.55	0.1903
101	2.16	0.3345
105	1.99	0.2989

x	y	$y' = \ln(y)$
89	1.08	0.0334
100	1.81	0.2577
70	0.53	-0.2757
57	0.16	-0.7959
84	0.54	-0.2676
74	0.52	-0.2840
87	1.01	0.0043
91.8	0.90	-0.0458
52	0.22	-0.6576
91.5	1.33	0.1239
94.5	1.05	0.0212
97	1.26	0.1004

Las estadísticas de resumen para ajustar la recta de regresión a los datos transformados son respectivamente $\sum_{i=1}^{n} x_i = 1733.5$, $\sum_{i=1}^{n} y_i^{'} = -1.9710$, $\sum_{i=1}^{n} x_i y_i^{'} = 60.1390$ y $\sum_{i=1}^{n} x_i^2 = 155015.63$, de modo que

$$\hat{\beta}_1 = \frac{(20)(60.1390) - (1733.5)(-1.9710)}{(20)(155015.63) - (1733.5)^2} = 0.048$$

Así, la recta de regresión ajustada con los datos trasformados es $\ln(y) = -4.259 + 0.048x$, que en la escala original en que están medidas las observaciones toma la forma

$$y = e^{(-4.259+0.048x)} \therefore y = 0.014e^{0.048x}$$

En la Figura 8.14 se muestra un gráfico de regresión con la recta de regresión ajustada a los datos transformados y otra en la escala original de las observaciones

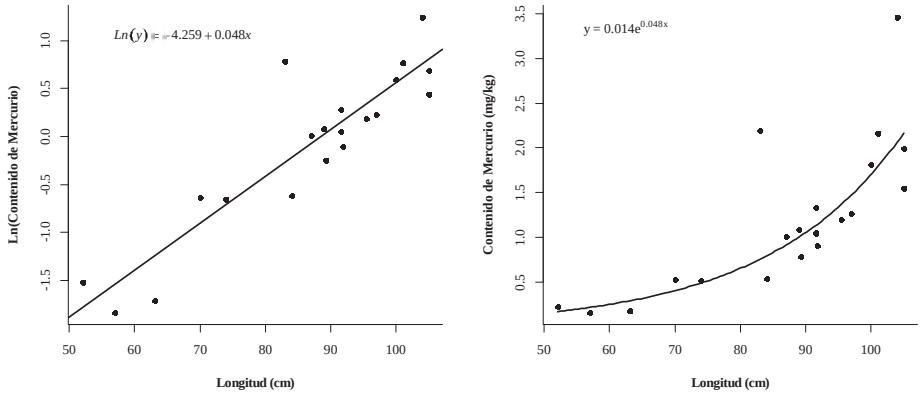


Figura 8.14. Diagrama de dispersión de las variables estudiadas con la recta de regresión ajustada en escala transformada y original.

En los gráficos anteriores se observa como mejora sustancialmente el ajuste de la recta estimada al comportamiento de las observaciones, por lo que los pronósticos que se realicen a través del modelo de regresión serán más confiables que cuando se ajustó el modelo de regresión con las observaciones sin transformar.

Ajustar el modelo de regresión ajustado en R es una tarea fácil y no muy diferente a modelos de regresión en escala original que hemos estudiado en secciones anteriores. Lo único necesario es emplear la función de transformación que se implementará dentro de los argumentos de la función ***lm***, como se observa en la siguiente salida de resultados de R para el ejemplo en cuestión

```
> Long<-
c(89.2,83,95.5,63,104,105,101,105,89,100,70,57,84,74,87,
91.8,52,91.5,91.5,97)
> Merc<-
c(0.78,2.19,1.20,0.18,3.46,1.55,2.16,1.99,1.08,1.81,0.53,
0.16,0.54,0.52,1.01,0.90,0.22,1.33,1.05,1.26)
> Reg1<-lm(log(Merc)~Long)
> Reg1

Call:
lm(formula = log(Merc) ~ Long)

Coefficients:
(Intercept)         Long
   -4.31914      0.04878
```

Ahora, observaremos cómo resultan algunos indicadores de calidad de ajuste del modelo luego de realizar la transformación, específicamente la prueba de utilidad del modelo y el coeficiente de determinación. Esto se observará en el siguiente resumen de la salida de resultados de R para el modelo de regresión ajustado

```
> summary(Reg1)

Call:
lm(formula = log(Merc) ~ Long)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46873 -0.26819 -0.03038  0.14644  1.05440

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.319137   0.464588  -9.297 2.71e-08 ***
Long         0.048779   0.005287   9.227 3.04e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3634 on 18 degrees of freedom
Multiple R-squared:  0.8255,    Adjusted R-squared:  0.8158
F-statistic: 85.13 on 1 and 18 DF,  p-value: 3.036e-08
```

De la salida anterior se evidencia que la prueba de utilidad del modelo de regresión ajustado resultó ser significativo (p-valor<0.05) y que el coeficiente de determinación mejoró sustancialmente respecto al modelo de regresión ajustado con las observaciones expresadas en su escala original, pasando a explicar del 56.92% al 82.55% de la variabilidad total de las observaciones.

8.3. Regresión lineal múltiple

En la práctica, muchos fenómenos científicos no solo involucran una variable respuesta predicha por una sola variable explicativa, sino que se encuentran involucradas dos o más variables explicativas que ayudan a predecir el comportamiento de una sola variable respuesta. Este tipo de situaciones se pueden modelar a través de una técnica de regresión denominada **regresión lineal múltiple**, cuando el modelo es lineal en sus coeficientes, igual a como se discutió cuando se abordó el modelo de regresión lineal simple. Para el caso de k variables independientes $x_1, x_2, ..., x_k$, la ecuación de regresión muestral está dada por

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Donde cada uno de los estimadores $\hat{\beta}_i$, son obtenidos de los datos muéstrales usando el método de los mínimos cuadrados discutido en secciones anteriores. Así, para cada uno de los puntos de los datos de la forma

$$\left\{\left(x_{1i}, x_{2i}, ..., x_{ki}, y_i\right), \qquad i = 1, 2, ..., n \text{ y } n > k\right\},$$

donde y_i, es la respuesta observada a los valores $x_{1i}, x_{2i}, ..., x_{ki}$ de las k variables independientes (explicativas) $x_1, x_2, ..., x_k$. Se supone que cada observación satisface la siguiente ecuación del modelo de regresión lineal múltiple

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

donde ε_i son los errores aleatorios asociados con la respuesta y_i.

Igual a los supuestos del modelo de regresión lineal simple, se supone que los ε_i son independientes y se encuentran distribuidos normalmente con media cero y varianza común. σ^2
Usando el método de los mínimos cuadrados para obtener los estimadores de los parámetros del modelo, minimizamos la suma cuadrática del error, *SSE*, que para el caso de la regresión lineal múltiple toma la forma

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_o - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki}\right)^2.$$

Esta expresión se deriva respecto a $\hat{\beta}_o, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$, se igualan a cero y se obtiene el siguiente **sistema de ecuaciones normales para estimación de los parámetros en la regresión lineal múltiple**, que debe ser resuelto a través de métodos algebraicos.

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} x_{2i} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{1i} x_{ki} = \sum_{i=1}^{n} x_{1i} y_i$$

$$\begin{array}{ccccc} . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{array}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ki} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ki} x_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ki} x_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} x_{ki} y_i$$

No obstante, en este sistema de ecuaciones cuando se encuentran involucradas muchas variables explicativas resulta bastante tedioso encontrar su solución, por ello, en esta sección sólo se expondrá el ajuste de modelos de regresión lineal múltiple a través del lenguaje de programación de R, con el fin de no hacer aburrido y dispendioso conocer la verdadera utilidad de esta técnica estadística. De acuerdo a lo anterior, en R un modelo de regresión lineal múltiple se ajusta de forma similar a una regresión lineal simple a través de la función ***lm***, con la única diferencia que al considerarse *k* regresores, estos se insertan en la línea de código separados por signos más "+", como se muestra a continuación

```
lm(y ~ x1 + x2 + … + xk)
```

8.3.1. Inferencias sobre el modelo de regresión lineal multiple

De forma análoga a lo expuesto en la sección donde se discutió el modelo de regresión lineal simple, para la regresión multiple, también existen ciertos procedimientos que permiten realizar inferencias y tomar decisiones respecto a la utilidad o buen ajuste del modelo ajustado, ejemplo de ello lo constituyen la construcción de intervalos confianza e intervalos de predicción, pruebas de utilidad del modelo, análisis de varianza coeficiente de determinación, entre otros. En esta sección, solo revisaremos los procedimientos más importantes a juicio de los autores y de uso más generalizado cuando se emprende el tratamiento estadístico de datos científicos.

8.3.1.1. Análisis de varianza en la regresión multiple

En el modelo de regresión lineal multiple, un análisis de varianza (ANOVA) se realiza con el objetivo de determinar la calidad de la ecuación de regresión ajustada. Esto se establece probando la siguiente hipótesis, que determina si el modelo explica una cantidad significativa de variación

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0.$$

Como es común el ANOVA implica una prueba *F*, cuyos resultados se resumen en una tabla como la siguiente

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*f* calculada
Regresión	SSR	k	$MSR = SSR/k$	$f = \frac{MSR}{MSE}$
Error	SSE	$n - (k + 1)$	$MSE = SSE/n-(k+1)$	
Total	SST	n - 1	-	

La hipótesis nula para esta prueba se rechaza a un nivel de significancia de α cuando $f \geq f_{\alpha[k,\, n-(k+1)]}$. Rechazar la hipótesis nula implica que uno de los coeficientes de la regresión difiere significativamente de cero, es decir, al menos una de las variables explicativas involucradas es importante.

En R, el análisis de varianza es aplicado como se expuso en secciones anteriores a través de la función ***anova***, aplicado al modelo lineal que se ajustó como se muestra a continuación

```
anova (lm(y ~ x1 + x2 + … + xk))
```

En la sección siguiente se dará una prueba que persigue el mismo objetivo, pero aplicado a cada coeficiente de forma individual

8.3.1.2. Pruebas t individuales para comparar variables

Las pruebas *t* que se consideran en la regresión lineal multiple es aquella donde se busca probar la nulidad de cada uno de los coeficientes de los regresores del modelo, es decir, se busca probar la hipótesis nula $H_0 : \beta_j = 0$, contra la alternativa bilateral $H_1 : \beta_j \neq 0$. Estas pruebas se denominan **pruebas *t* de comparación de variables**, y su objetivo es establecer cuáles regresores deben ser elegidos de modo que se aumente la utilidad y la capacidad de pronóstico del modelo ajustado. Esto último se establece con el rechazo o no rechazo de la hipótesis nula de cada prueba, pues si un coeficiente resulta no ser significativo (es decir, no se rechaza la hipótesis $H_0 : \beta_j = 0$), la conclusión que se obtiene es que la variables explicativas de este coeficiente es insignificante, es decir, explica una

cantidad insignificante de la variación de nuestra variable respuesta *y*, en la presencia de los demás regresores del modelo (Walpole *at al.*, 2007).

8.3.1.3. Coeficiente de determinación y coeficiente de determinación ajustado

En la regresión multiple también es común evaluar la calidad de ajuste del modelo de regresión ajustado a través del coeficiente de determinación. Recordemos que este mide la proporción de variabilidad de la variable respuesta *y*, que es explicada por el modelo ajustado y matemáticamente está definido por

$$R^2 = 1 - \frac{SSE}{SST}$$

No obstante, en la regresión multiple el considerar varios regresores no afecta considerablemente el valor del coeficiente de determinación, aún cuando estos regresores no sean importantes para el modelo ajustado, es decir, el R^2, no se ve afectado por problemas de un modelo sobreajustado. Ante esta desventaja del R^2, se ha propuesto en muchos textos de estadística el uso del coeficiente de determinación ajustado R^2_{Ajus}, definido matemáticamente como

$$R^2_{\text{Ajus}} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

Nótese que el ajuste de este último indicador se basa en dividir la suma de cuadrados del error y la suma de cuadrados totales por sus respectivos grados de libertad. Con ello se consigue obtener un estadístico que castiga la presencia de un modelo sobreajustado, de allí, que la gran mayoría de los software`s estadísticos de mayor uso en la actualidad en sus salidas de resultados proporcionen el valor de este estadístico.

En el entorno de R, las pruebas *t* individuales para comparación de las variables, el coeficiente de determinación y el correspondiente coeficiente de correlación ajustado se obtienen aplicando la función ***summary*** al modelo lineal ajustado, como se muestra en las siguientes líneas de comando

```
summary (lm(y ~ x1 + x2 + … + xk))
```

Ahora veamos un ejemplo donde podamos ilustrar lo conceptos dados sobre la regresión lineal multiple. El ejemplo se aplicará en R inmediatamente y nos saltaremos por completo los tediosos cálculos matemáticos, para no hacer extenuante la comprensión de esta técnica estadística.

Ejemplo 8.5. En la adsorción de tierra y sedimento, la magnitud de la acumulación en forma condensada de los productos químicos en la superficie es una característica importante que influye en la eficiencia de insecticidas y varios otros productos químicos. El artículo *"Adsorption of Phosphate, Arsenate, Methanearsonate and Cacodylate by Lake and Stream Sediments: Comparison with Soils*" (J. of Environ. Qual., 1984, pp. 499-504) presenta los siguientes datos en la tabla de abajo. Aquí se toma Y como la variable dependiente, la cual denota el índice de adsorción de fosfato, X_1 es una de las variables independientes denotando la cantidad de hierro extraíble y, X_2 es otra de las variables independientes denotando la cantidad de aluminio extraíble. (Devore, 2008).

Índice de Absorción	Hierro Extraíble	Aluminio Extraíble
4	61	13
18	175	21
14	111	24
18	124	23
26	130	64
26	173	38
21	169	33
30	169	61
28	160	39
36	244	71
65	257	112
62	333	88
40	199	54

Solución

Lo que se busca en este ejemplo es ajustar un modelo de regresión lineal entre el índice de absorción de fosfatos y el contenido de hierro y aluminio extraíbles. Para ello, iniciaremos con realizar un análisis gráfico para observar cómo se comporta nuestra variable respuesta, índice de absorción de fosfato, en función de las dos variables explicativas dadas, la cantidad de hierro y aluminio extraíbles, respectivamente. En la Figura 8.15 se

muestran los diagramas de dispersión que corresponden a las siguientes líneas de código

```
> Absorción<-c(4,18,14,18,26,26,21,30,28,36,65,42,40)
> Hierro<-
c(61,175,111,124,130,173,169,169,160,244,257,333,199)
> Aluminio<-c(13,21,24,23,64,38,33,61,39,71,112,88,54)
> par(mfrow=c(1,2),family="serif",bty="l")
> plot(Hierro,Absorción,pch=16,xlab="Cantidad de hierro
extraíble", ylab="Índice de absorción de fosfato",font.lab=2)
> plot(Aluminio,Absorción,pch=16,xlab="Cantidad de aluminio
extraíble",ylab="Índice de absorción de fosfato",font.lab=2)
```

De este gráfico que observa la existencia de una relativamente buena relación lineal entre la variable respuesta y las variables explicativas, lo que sugiere que un modelo de regresión lineal múltiple puede resultar un buen ajuste para las variables estudiadas.

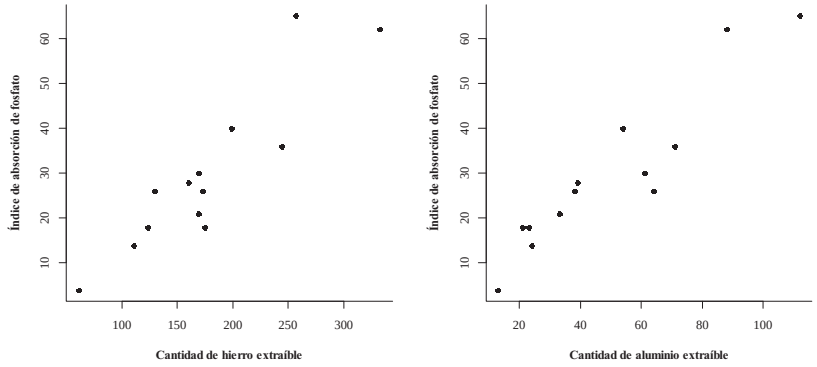


Figura 8.15. Diagramas de dispersión entre el índice de absorción de fosfatos y la cantidad de hierro y aluminio extraíbles, respectivamente.

Otra forma de mejor representación de nuestras variables de forma simultánea, lo constituye un gráfico de dispersión en el espacio (tercera dimensión), como se muestra en la Figura 8.16, en este gráfico se observa como la nube de puntos de las observaciones se ajustan bien a un plano recto. Este tipo de gráficos se construyen fácilmente en R a través de la función ***cloud*** del paquete "*lattice*" (Sarkar, 2015), cuya funcionalidad de los argumentos de esta función se dejan como material de consulta para el lector, para efectos de no extendernos demasiado en apartes que no constituyen el objetivo principal de este capítulo. No obstante, a

continuación mostramos las líneas de programación usadas para generar el gráfico de la Figura 8.16.

```
> library(lattice)
>  cloud(Absorción~Hierro*Aluminio,cex=0.8,pch=16,col="black",
zlab=list("Absorción",cex=1,font=2,fontfamily="serif",rot=90),
ylab=list("Aluminio",cex=1,font=2,fontfamily="serif",rot=15),x
lab=list("Hierro",cex=1,font=2,fontfamily="serif",rot=0),scale
s=list(arrows=FALSE,fontfamily="serif", distance=0.8),
screen=list(z=10, x=-90,y=-
35),zlim=c(4,65),ylim=c(13,112),xlim=c(61,333))
```

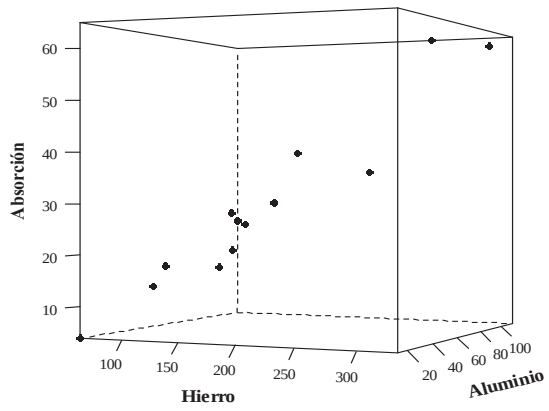


Figura 8.16. Gráfico de dispersión 3D para los datos del ejemplo 8.5.

Ahora, proseguiremos con ajustar el modelo de regresión multiple a las observaciones, como se muestra en la siguiente salida de resultados de R

```
> Reg<-lm(Absorción~Hierro+Aluminio)
> Reg

Call:
lm(formula = Absorción ~ Hierro + Aluminio)

Coefficients:
(Intercept)       Hierro     Aluminio
    -7.3507       0.1127       0.3490
```

Los resultados del ajuste de modelo indican que la ecuación representada por el modelo de regresión multiple es

$$\text{Índice de absorción} = -7.3507 + 0.1127\text{Hierro} + 0.3490\text{Aluminio}\,.$$

Ahora observemos la salida de resultados de R para las pruebas *t* individuales de comparación de variables y el valor del coeficiente de determinación múltiple y coeficiente de determinación ajustado

```
> summary(Reg)

Call:
lm(formula = Absorción ~ Hierro + Aluminio)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9352 -2.2182  0.4613  3.3448  6.0708

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.35066    3.48467  -2.109 0.061101 .
Hierro       0.11273    0.02969   3.797 0.003504 **
Aluminio     0.34900    0.07131   4.894 0.000628 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.379 on 10 degrees of freedom
Multiple R-squared:  0.9485,    Adjusted R-squared:  0.9382
F-statistic: 92.03 on 2 and 10 DF,  p-value: 3.634e-07
```

Obsérvese que las pruebas *t* para las dos variables explicativas involucradas resultan ser significativas a un nivel de significancia de 0.05 (p-valor < 0.05), esto sugiere que ambas variables explicativas son importantes para el modelo y explican una cantidad significativa de la variabilidad de la variable respuesta (índice de absorción de fosfatos). Esto se complementa con el valor del R^2 y R^2_{Ajus}, que toman valores de 0.9485 y 0.9382, respectivamente. Si interpretamos el R^2_{Ajus}, vemos que el modelo está explicando el 93.82% de la variabilidad total del índice de absorción de fosfatos, de aquí que se pueda concluir que nuestro modelo presenta un excelente ajuste.

A continuación, veremos un análisis gráfico sobre el cumplimiento de los supuestos de la regresión lineal. En la Figura 8.17 se muestran las gráficas de diagnóstico de los residuos de nuestro modelo ajustado, en él se observa que los residuos del modelo cumplen aceptablemente con el supuesto de normalidad, homogeneidad de varianzas y no autocorrelación.

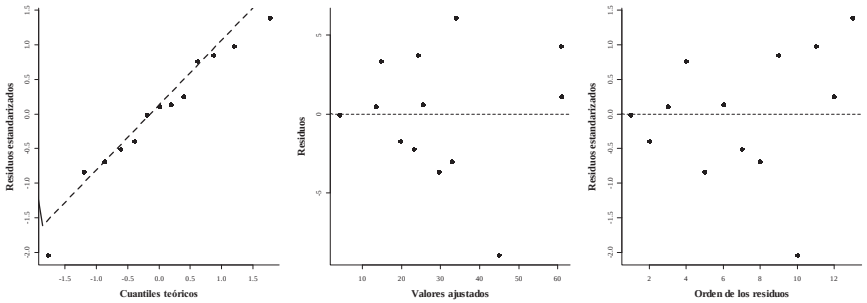


Figura 8.17. Gráficos de diagnóstico para el modelo de regresión del ejemplo 8.5.

8.3.2. Supuesto de multicolinealidad

La multicolinealidad es otro de los supuestos que deben ser evaluados cuando se ajustan modelos de regresión lineal. El termino multicolinealidad atribuido a Ragnar Frisch (1934), designa una relación lineal "perfecta" o exacta entre algunas o todas las variables explicativas de un modelo de regresión lineal multiple. Sin embargo, hoy día el termino incluye también aquellas situaciones en las que algunas variables independientes (regresoras o explicativas) están interrelacionadas pero no en forma perfecta, sino que contiene un término de error aleatorio (estocástico).

Suponer que no existe multicolinealidad entre los regresores del modelo, se debe a que ante la presencia de esta cuando se desea ajustar un modelo de regresión lineal, los coeficientes del mismo no pueden ser estimados con gran precisión o exactitud. Además, se tiene como consecuencia, que ante la presencia de multicolinealidad moderada o alta, las pruebas *t* de comparación de variables pueden resultar estadísticamente no significativa para algunos de los regresores, aun cuando el valor de R^2 y R^2_{Ajus}, asuman valores altos. No obstante a lo anteriormente comentado Achen citado por Gujarati & Porter (2010), afirma que la presencia de multicolinealidad no viola los demás supuestos básicos de la regresión, resultando inútil buscar una solución a la presencia de la misma, pues al igual que a la existencia de pocas observaciones en el experimento, no hay una solución estadística. Así mismo, ante la presencia de multicolinealidad el modelo de regresión ajustado sigue siendo útil para fines de pronostico (Gujarati & Porter, 2010).

Al evaluar el supuesto de multicolinealidad, no presume ninguna importancia determinar si esta está presente o no, sino en qué grado está

presente. Igualmente, como la multicolinealidad depende únicamente de las variables explicativas del modelo, que se suponen no son de naturaleza estocástica, esta es una característica de la muestra y no de la población. Por consiguiente, no es necesario llevar a cabo pruebas de multicolinealidad, pero si se desea, si es posible medir su grado en cualquier muestra determinada (Kmenta, 1986).

Al ser la multicolinealidad un fenómeno de tipo muestral, no existe un método o prueba estadística única para detectarla o medir su fuerza. Únicamente se cuenta con algunas reglas prácticas para medir el grado en que esta ocurre de las cuales solo discutiremos en este texto la del **factor de inflación de la varianza** (Gujarati & Porter, 2010), que se utiliza como una medida del grado en que la varianza de los estimadores del modelo de regresión se incrementa por la presencia de multicolinealidad entre los regresores del modelo. Su principio de aplicación se basa en efectuar regresiones auxiliares de la variable regresora x_i contra los *k* regresores restantes y se calcula el siguiente indicador

$$FIV_i = \frac{1}{1-R_i^2}$$

Donde R_i^2, es el coeficiente de determinación de cada una de las *i* regresiones auxiliares a que haya lugar. Entre mayor es el valor del FIV_i, mayor "problema" o colinealidad tiene la variable X_i. ¿Pero, cuánto debe ascender el *FIV* antes de que una regresora se convierta en un problema? Como regla práctica, si el *FIV* de una variable es superior a 10 (esto sucede si R_i^2 excede de 0.90), se dice que esa variable es muy colineal.

En el ambiente de programación de R, la multicolinealidad a través de la regla del factor de inflación de la varianza es fácilmente aplicado haciendo uso de la función ***VIF*** del paquete "*fmsb*" (Nakasawa, 2015), siguiendo la siguiente línea de código

```
VIF(X)
```

Donde *X* es un objeto de clase "*lm*" generado a través de la función ***lm*** para ajuste de modelos lineales, en donde se encuentra asignado nuestro modelo *j* de regresión lineal multiple auxiliar.

Ejemplo 8.6. Para los datos del Ejemplo 8.5, determínese el grado de la colinealidad existente entre las variables explicativas.

Solución

El conjunto de datos del Ejemplo 8.5, solo cuenta con dos variables regresoras, por ello solo se obtendría una sola regresión auxiliar, y a su vez solo es necesario el cálculo de un solo *FIV*. Con el objetivo de optimizar el procedimiento, correremos inmediatamente este análisis en el entorno de R, como se muestra a continuación

```
> Absorción<-c(4,18,14,18,26,26,21,30,28,36,65,62,40)
> Hierro<-
c(61,175,111,124,130,173,169,169,160,244,257,333,199)
> Aluminio<-c(13,21,24,23,64,38,33,61,39,71,112,88,54)
> library(fmsb)
> Reg<-lm(Absorción~Hierro+Aluminio)
> VIF(lm(Hierro~Aluminio))
[1] 2.710733
```

Siguiendo la regla practica de decisión, como el *FIV* = 2.711, es muy inferior a 10, se concluye que las variables regresoras contenido de hierro y aluminio extraíbles, no son significativamente colineales, es decir, que nuestro modelo de regresión multiple ajustado, no se ve afectado por problemas de multicolinealidad.

8.4. Regresión no lineal: Polinómica

En las secciones precedentes hemos discutido los apartes más importantes desde el punto de vista práctico de los modelos de regresión donde la relación inherente entre la variable respuesta y la o las variables regresoras es estrictamente lineal. Sin embargo, en muchas situaciones prácticas de la vida real, las variables de estudio no se encuentran relacionadas de forma lineal, sino que el gráfico de dispersión de la variable respuesta y la variable explicativa exhiben una relación funcional de tipo polinómico que puede dar una mejor aproximación de las observaciones a la verdadera función de regresión o línea de regresión poblacional.

En otras palabras, se intenta decir que el valor esperado de la variable respuesta *y* es una función con polinomios de *k*-ésimo grado de la variable explicativa *x*. Lo anterior se expresa matemáticamente como

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

Donde ε es una variable normalmente distribuida con media cero y varianza constante.

En este tipo de situación la obtención de los parámetros del modelo de regresión polinomial, se realiza de la forma tradicional a través del método de los mínimos cuadrados, que como sabemos su principio de aplicación se basa en minimizar la suma de cuadrados del error, *SSE*, que para el modelo de regresión polinomial ajustada $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_k x^k$ es

$$SSE = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \cdots + \hat{\beta}_k x_i^k \right) \right]^2$$

Para hallar los valores de los estimadores de los coeficientes del modelo debemos encontrar las $k + 1$ derivadas parciales $\partial(SSE)/\partial\hat{\beta}_0, \partial(SSE)/\partial\hat{\beta}_1, \ldots, \partial(SSE)/\partial\hat{\beta}_k$, igualarlas a cero y resolver el siguiente sistemas de ecuaciones normales

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_i^k = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_i^3 + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_i^{k+1} = \sum_{i=1}^{n} x_i y_i$$

$$\begin{array}{ccccc} . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{array}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i^k + \hat{\beta}_1 \sum_{i=1}^{n} x_i^{k+1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ki} x_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_i^{2k} = \sum_{i=1}^{n} x_i^k y_i$$

De forma mecánica, la resolución de este sistema de ecuaciones es una tarea dispendiosa, por ello a lo largo de este capítulo nos apoyaremos en el uso de herramientas tecnológicas. En nuestro caso particular, realizaremos los análisis a través de R, en donde el modelo polinomial se ajusta de forma análoga a los modelos de regresión tradicionales con la función ***lm***. Sin embargo, en los argumentos de esta función se debe especificar el comportamiento polinomial de la variable explicativa a través de la función ***poly*** del paquete base. La estructura de la línea de código para generar el modelo se muestra a continuación

```
lm (y ~ poly(x, k, raw = TRUE)
```

Donde *k* es el grado de la función polinómica.

Las pruebas de utilidad del modelo y de evaluación de la calidad del ajuste son las mismas a las discutidas en los modelos de regresión tradicionales. A continuación mostraremos un ejemplo para ilustrar la aplicación de los modelos de regresión polinomiales.

Ejemplo 8.7. A continuación se muestran datos tomados de Devore (2008) sobre un experimento realizado para estudiar la relación entre las concentraciones de aluminio (Al) y el pH del suelo.

Concentración de aluminio	Ph de suelo
1.20	4.01
0.78	4.07
0.83	4.08
0.98	4.10
0.65	4.18
0.76	4.20
0.40	4.23
0.45	4.27
0.39	4.30
0.30	4.41
0.20	4.45
0.24	4.50
0.10	4.58
0.13	4.68
0.07	4.70
0.04	4.77

A partir de esta información se desea establecer el modelo que se ajuste a la distribución de los datos.

Solución

Una vez cargados los datos en el entorno de R, iniciamos el análisis con la construcción de un gráfico de dispersión entre las variables de estudio para observar el patrón funcional que relaciona a dichas variables como se muestra en la Figura 8.18.

```
> Aluminio<-
c(1.20,1.78,0.83,0.98,0.65,0.76,0.40,0.45,0.39,0.30,
0.20,0.24,0.10,0.13,0.07,0.04)
```

```
> pH<-
c(4.01,4.07,4.08,4.10,4.18,4.20,4.23,4.27,4.30,4.41,4.45,
4.50,4.58,4.68,4.70,4.77)
> plot(pH,Aluminio,pch=16,ylab="Concentración de aluminio",
xlab= "pH del suelo",font.lab=2)
```

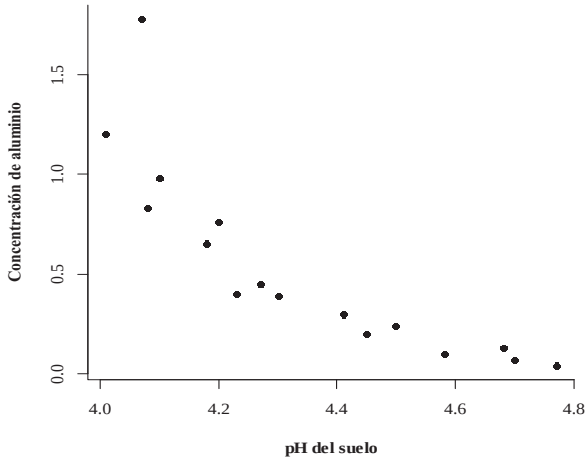


Figura 8.18. Gráfico de dispersión entre la concentración de aluminio y el pH de suelo.

Del gráfico es fácil notar que la relación entre las variables de estudio no es estrictamente lineal. Sin embargo, no es posible establecer por simple inspección que relación polinómica se ajusta mejor a las observaciones. Por ello, ajustaremos varias modelos de regresión polinomial con diferentes grados hasta quedarnos con aquel que nos presente mejor ajuste. A continuación mostramos la salida de resultados de R para este procedimiento de análisis

```
> Reg1<-lm(Aluminio~poly(pH,2,raw=TRUE))
> summary(Reg1)

Call:
lm(formula = Aluminio ~ poly(pH, 2, raw = TRUE))

Residuals:
     Min       1Q   Median       3Q      Max
-0.25953 -0.09014 -0.02375  0.05835  0.65345
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 70.835     22.016   3.217  0.00674 **
poly(pH, 2, raw = TRUE)1  -30.521     10.060  -3.034  0.00959 **
poly(pH, 2, raw = TRUE)2    3.291      1.146   2.871  0.01313 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2178 on 13 degrees of freedom
Multiple R-squared:  0.8216,    Adjusted R-squared:  0.7941
F-statistic: 29.93 on 2 and 13 DF,  p-value: 1.364e-05

> Reg2<-lm(Aluminio~poly(pH,3,raw=TRUE))
> summary(Reg2)

Call:
lm(formula = Aluminio ~ poly(pH, 3, raw = TRUE))

Residuals:
     Min       1Q   Median       3Q      Max
-0.26032 -0.04025 -0.02148  0.03972  0.64538

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               393.510    491.786   0.800    0.439
poly(pH, 3, raw = TRUE)1 -252.050    337.434  -0.747    0.469
poly(pH, 3, raw = TRUE)2   53.901     77.063   0.699    0.498
poly(pH, 3, raw = TRUE)3   -3.848      5.858  -0.657    0.524

Residual standard error: 0.2227 on 12 degrees of freedom
Multiple R-squared:  0.8277,    Adjusted R-squared:  0.7847
F-statistic: 19.22 on 3 and 12 DF,  p-value: 7.071e-05
```

De la salida anterior se puede observar que solo bastó con ajustar dos modelos polinomicos de grado dos ($k = 2$) y tres ($k = 3$), respectivamente. Siendo más útil para propósitos de predicción el modelo cuadrático, por el valor aceptable que arroja para el coeficiente de determinación, y así mismo, porque para este modelo las pruebas t de utilidad del modelo resultan significativas para todos los parámetros del modelo. En la Figura 8.19 se muestra las líneas de regresión de cada los dos modelos justados, allí se puede observar que a pesar de la de las dos curvas, el modelo cuadrático exhibe un mejor ajuste a las observaciones.

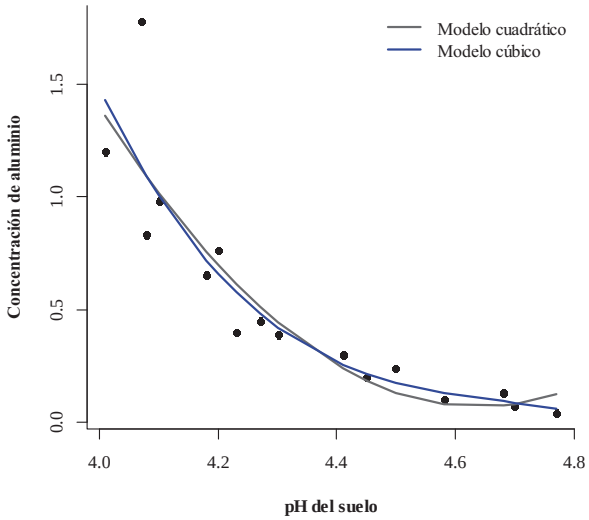


Figura 8.19. Modelos de regresión ajustados para los datos del Ejemplo 8.7.

Como acto siguiente al ajuste del modelo de regresión polinomial, se pueden evaluar los supuestos del modelo: Normalidad, homogeneidad de varianzas, autocorrelación y multicolinealidad de los residuos, siguiendo los procedimientos introducidos en secciones anteriores. Sin embargo, esto no se realizará en esta sección para no extendernos más en la discusión de temáticas, en las cuales el lector ya debe tener cierto conocimiento con base en lo visto en secciones precedentes.

8.5. Regresión logística

Anteriormente habíamos mencionado que existen situaciones prácticas en las que la variable respuesta es de naturaleza cualitativa dicotómica, y el objetivo de nuestro análisis es realizar pronósticos de esta a partir de la relación que existe con una o varias variables explicativas de naturaleza cuantitativa; siendo el modelo de regresión logística (o curva logística) el más adecuado para modelar este tipo de situaciones.

El modelo logístico es ampliamente usado en experimentos que involucran determinar la probabilidad en que una de las categorías de la variable dicotómica dependiente sucederá en función de los valores que toman una serie de variables explicativas. Por ejemplo, se podría establecer un modelo, con base en información muestral, que permita predecir la efectividad de cierta vacuna en función de la dosis aplicada de la misma, o sumar las variables sexo, edad etc. a este modelo.

El modelo logístico para una variable explicativa, se expresa matemáticamente como

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

En caso de existir varias variables pronosticadoras, el modelo logístico se expresa como

$$p(x) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}$$

Donde $p(x)$, es la probabilidad de que se presente el evento estudiado (éxito) en el caso *i*, *k* es el número de variables explicativas, β_0 es un coeficiente, β_j es el coeficiente de la variable predictora y x_{ij} es el valor de la variable explicativa *j* en el caso *i*. (Guisande *et al.*, 2011).

En la Figura 8.20 se muestra una gráfica del modelo logístico, donde se puede observar el comportamiento de éste para valores particulares de β_0 y β_1 con $\beta_1 > 0$ y $\beta_1 < 0$. Nótese que para $\beta_1 > 0$, cuando la variable explicativa *x* aumenta, la probabilidad de éxito se incrementa. Para $\beta_1 < 0$, la probabilidad de éxito es una función decreciente de *x*.

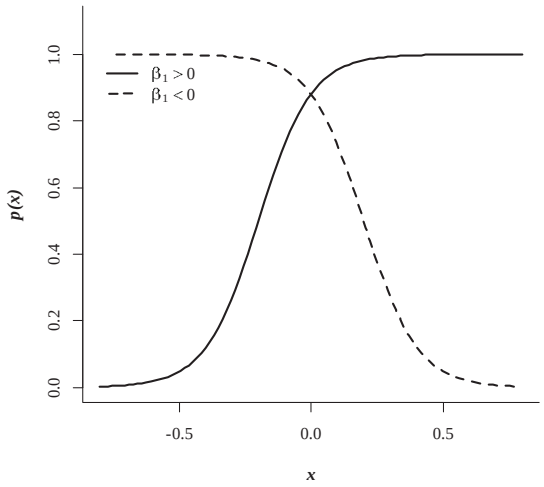


Figura 8.20. Curva logística con pendiente positiva y negativa.

En el modelo logístico, como en cualquier modelo de regresión, la importancia del análisis se centra en la estimación de los parámetros del modelo a partir de datos muéstrales. Por lo general en el modelo logístico, esto se hace usando técnicas de máxima verosimilitud que no han sido discutidas en este texto, dado que los detalles de las mismas son muy complejos y requieren por parte del lector profundizar en temas referentes a probabilidades, lo que haría muy dispendiosa las temáticas de la presente obra sin contar con que nos sesgaríamos fuertemente de los objetivos prácticos que se persiguen. Por fortuna, los paquetes estadísticos disponibles en el mercado, pueden realizar estos cálculos de forma automatizada, brindándonos la facilidad de solo centrarnos en la interpretación de las salidas de resultado. R no escapa a ello, de allí que en esta sección realizaremos el ajuste de un modelo logístico con la ayuda de este ambiente de programación y daremos una interpretación de los resultados más relevantes.

La interpretación de los parámetros del modelo logístico no es tan intuitiva a como ocurre en los modelos de regresión lineal, para facilitar las interpretaciones, al usar un poco de algebra elemental, el modelo logístico podemos rescribirlo de la siguiente forma

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} \therefore \frac{p(x)}{1-p(x)} = e^{\beta_0} \cdot e^{\beta_1 x}$$

Así, e^{β_0} expresa el número de veces en que es más probable que ocurra un éxito que un fracaso, cuando la variable explicativa toma un valor de cero, este coeficiente es denominado **probabilidad relativa u odds**.

Ahora, para buscar una interpretación de e^{β_1}, expresaremos el modelo logístico como

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} \quad \text{y} \quad \frac{p'(x)}{1-p'(x)} = e^{\beta_0+\beta_1 (x+1)}$$

Entonces

$$e^{\beta_1} = \frac{\dfrac{p(x)}{1-p(x)}}{\dfrac{p'(x)}{1-p'(x)}}$$

Es decir, e^{β} es el cambio en la probabilidad de ocurrencia de la variable respuesta asociadas con un aumento de una unidad de la variable explicativa, este coeficiente comúnmente es denominada **razón de ventaja** y **odds ratio**.

En R, el ajuste de modelos de regresión logístico se modela a través de la función ***glm*** (generalized lineal model), siguiendo la siguiente estructura de programación

```
glm(formula, family = binomial(logit), data, weights)
```

Donde el argumento ***formula*** indica la introducción de la estructura del modelo, igual que en los modelos lineales ordinarios ($y \sim x_1 + x_2 + \ldots + x_k$), ***family*** establece la familia de distribuciones a partir de la cual se generará el modelo lineal generalizado, en nuestro caso será un modelo binomial logístico, ***data*** especifica la matriz de datos (data frame) del cual se seleccionan las variables involucradas en el análisis, este argumento es útil cuando los datos son cargados en R como data frame, ***weigths*** establece un vector que indica los pesos a partir de los cuales se calculan las probabilidades de éxitos.

A continuación examinaremos un ejemplo de ajuste de modelos de regresión logística.

Ejemplo 8.8. A continuación se muestra un conjunto de datos de tomados de Walpole *et al.* (2007) sobre la respuesta a la dosis de estreptomicina, a partir de los cuales se desea desarrollar una relación entre la proporción de linfoblastos muestreados que contienes aberraciones y la dosis del medicamento. En total durante el estudio se aplicaron cinco niveles de dosis a los conejos utilizados como sujetos experimentales, obteniéndose los siguientes resultados

Dosis (mg/kg)	Número de linfoblastos	Número de aberraciones
0	600	15
30	500	96
60	600	187
75	300	100
90	300	145

Ajústese un modelo de regresión a partir de estos datos que permita predecir la probabilidad de ocurrencia de aberraciones en los linfoblastos muestreados en función de la dosis de estreptomicina aplicada.

Solución

Iniciaremos nuestro análisis con la introducción de los datos en el entorno de R. como los valores de la variable respuesta (linfoblastos muestreados con aberraciones) están datos como frecuencias absolutas, es necesario introducir una nueva variable que especifique las probabilidades de ocurrencia (frecuencias relativas) de esta variable, pues en el modelo logístico los valores de la variable respuesta deben estar comprendidos entre 0 y 1, por tratarse de probabilidades.

```
> Dosis<-c(0,30,60,75,90)
> Linfoblastos<-c(600,500,600,300,300)
> Aberraciones<-c(15,96,187,100,145)
> Prob.Aberraciones<-Aberraciones/Linfoblastos
```

Ahora, ya definidas las variables involucradas en el análisis, procedemos con ajustar el modelo logístico que nos permita establecer una relación funcional entre la proporción de linfoblastos con alteraciones y la dosis de estreptomicina aplicada. El modelaje en R de lo anterior, produce la siguiente salida de resultados

```
> Reg<-
glm(Prob.Aberraciones~Dosis,family=binomial(logit),weights=
Linfoblastos)
> summary(Reg)

Call:
glm(formula = Prob.Aberraciones ~ Dosis, family =
binomial(logit),
    weights = Linfoblastos)

Deviance Residuals:
      1        2        3        4        5
-4.0088   3.3902   1.3826  -1.9813  -0.6489

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.762042   0.127201  -21.71   <2e-16 ***
Dosis        0.030781   0.001973   15.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 342.651  on 4  degrees of freedom
```

De las interpretaciones más relevantes que salen a relucir de esta salida de resultados, nótese que β_1 resulta ser significativo a un nivel de significancia de 0.05, lo que muestra que la utilidad del modelo para fines predictivos es bastante sobresaliente. El valor de los coeficientes dados en esta salida está expresados como logaritmos por ello para fines de interpretación en las unidades de medidas de las variables es necesario aplicar trasformación exponencial a dichos coeficientes como se muestra a continuación

```
> exp(Reg$coefficients)
(Intercept)       Dosis
 0.06316263  1.03125975
```

De lo anterior, se pueden trazar las siguientes interpretaciones: por un lado el valor del odds, $e^{\beta_0} = 0.0632$, indica que es apenas 0.06 veces más probable encontrar aberraciones en los linfoblastos, cuando no se aplica una dosis de estreptomicina. Por otro lado, tenemos que el valor del odds ratios, $e^{\beta_1} = 1.0313$, establece que al aumentar la dosis 1 mg/kg, la probabilidad de encontrar aberraciones en los linfoblastos permanece aproximadamente constante. Esto sugiere que la dosis de estreptomicina aplicada a los sujetos experimentales no tiene un efecto significativo en la reducción de aberraciones de los linfoblastos.

9.

Estadística no paramétrica: procedimientos de distribución libre

9.1. Generalidades

En los capítulos anteriores se han discutidos diversas técnicas estadísticas de naturaleza inferencial en las que se ha hecho hincapié en que su eficiencia y utilidad depende un estricto cumplimiento de supuestos como la distribución de las observaciones de forma normal y cuando se contrasta varias poblaciones que estas posean varianzas iguales. Este tipo de procedimiento son genéricamente denominados **métodos paramétricos**. Sin embargo, en muchas ocasiones y situaciones practicas no es posible asumir los supuestos de normalidad e igualdad de varianzas en nuestros datos, por tanto, es necesario recurrir a procedimientos alternativos que no dependan de la forma en que los datos se encuentren distribuidos, este tipo de procedimientos son los llamados **métodos no paramétricos** o de **distribución libre** que serán discutidos a lo largo del desarrollo de este capítulo.

Es importante anotar que los métodos no paramétricos a pesar de su amplio uso en las ciencias e investigaciones científicas, son menos eficientes que los métodos paramétricos, por ello se prefiera el uso de los segundos sobre los primeros, siempre que sea posible. Además, se debe tener en cuenta que los métodos paramétricos usados tradicionalmente tienen una ligera robustez a transgresiones leves del supuesto de normalidad, como se ha comentado en secciones previas. Sin embargo, los métodos no paramétricos tiene como ventaja que son de aplicación más general que los métodos paramétricos, dada su naturaleza de no exigir ninguna condición sobre la forma en que están distribuidos los datos. Además, Guisande *et al.* (2011) afirma de forma general que la coincidencia entre los resultados obtenidos entre las pruebas paramétricas y no paramétricas es superior a un 90%.

En este capítulo revisaremos los métodos no paramétricos alternativos a las pruebas paramétricas que se han revisado a lo largo de esta obra, empezando por los test para una población, para la comparación dos poblaciones independientes y pareadas, hasta pruebas para la comparación de k poblaciones. Así mismo, se expondrá la forma en que cada una de estas pruebas son modeladas en el ambiente de programación de R.

9.2. Test no paramétricos para una población: test de rangos con signos de Wilcoxon

El lector debe recordar que en el Capítulo 5 se estudiaron procedimientos para probar la hipótesis nula de que la media μ de una población toma un valor especifico μ_0, cuando esta se distribuye normalmente o si la muestra es lo suficientemente grande. Cuando estas condiciones no son satisfechas y los datos de la muestra no provienen de una población normal y su vez pertenecen a una muestra de tamaño pequeño ($n < 30$), debemos recurrir a un procedimiento no paramétrico alternativo. En esta sección discutiremos uno de los test de más amplio uso, denominado **test de sumas de rango con Wilcoxon** (Wilcoxon, 1945).

El test de suma con rangos de Wilcoxon, como todos los test no paramétricos, se utiliza para realizar inferencias acerca del valor que toma la mediana $\tilde{\mu}$ de una población (en remplazo de la media), sin importar como se encuentra distribuida, pero se asume que la distribución de la cual provienen los datos es de naturaleza **continua** y **simétrica**. Bajo estas condiciones el procedimiento de prueba de este test para probar la hipótesis $H_0 : \tilde{\mu} = \tilde{\mu}_0$, es bastante sencillo e intuitivo. Consiste en restar el valor de $\tilde{\mu}_0$ a cada observación muestral y descartar todas las diferencias que sean iguales a cero. Luego se ordenan las diferencias resultantes de forma ascendente, prescindiendo del signo que estas posean. Se asigna un rango de 1 a la menor diferencia absoluta (es decir, sin signo), un rango de 2 a la siguiente más pequeña, y así sucesivamente. Cuando el valor absoluto de dos o más diferencias es el mismo, se asigna a cada una el promedio de los rangos que se asignarían si las diferencias fueran distinguibles (Walpole *et al.*, 2007).

Nótese que para el desarrollo metodológico anterior, si $H_0 : \tilde{\mu} = \tilde{\mu}_0$ es verdadera, el total de los rangos que corresponden a las diferencias positivas debería ser casi igual al total de los rangos que corresponden a las diferencias negativas. Estos totales los representaremos por W_+ y W_-, respectivamente. Al menor valor entre W_+ y W_-, lo designaremos como w. Así, $H_0 : \tilde{\mu} = \tilde{\mu}_0$ es rechazada para la alternativa bilateral $H_1 : \tilde{\mu} \neq \tilde{\mu}_0$, cuando $w < w_{\alpha,n}$. Para las alternativas unilaterales $H_1 : \tilde{\mu} < \tilde{\mu}_0$ y $H_1 : \tilde{\mu} > \tilde{\mu}_0$, es rechazada si $w_+ < w_{\alpha,n}$ y $w_- < w_{\alpha,n}$, respectivamente. Los valores de $w_{\alpha,n}$ se encuentran en la Tabla A.19 del apéndice para niveles de significancia de 0.01, 0.025 y 0.05 para una prueba unilateral y valores de $w_{\alpha,n}$ para niveles de significancia de 0.02, 0.05 y 0.10 para pruebas bilaterales.

El modelado de este test en R es bastante simple a través de la función ***wilcox.test***, solo basta con seguir la siguiente línea de códigos, una vez se tengan cargados los datos en la consola.

```
wilcox.test(x, alternative = c("two.sided", "less", "greater"),
mu = 0, correct = TRUE, conf.int = FALSE, conf.level = 0.95,
...)
```

De los argumentos de esta función tenemos que ***x*** es un vector de los datos a los que se desea aplicar el test, ***alternative*** especifica la hipótesis alternativa de la prueba (bilateral, unilateral a la izquierda y unilateral a la derecha, respectivamente), ***mu*** es el valor de la mediana que se desea contrastar, ***correct*** es un valor lógico (TRUE o FALSE) que establece si se realiza corrección de continuidad para el computo del p-valor, ***conf.int*** es un valor lógico que especifica la construcción de intervalos de confianza para la medina en la salida de resultados y ***conf.level*** especifica el nivel de confiabilidad del test.

Ejemplo 9.1. A continuación se presenta una serie de datos de concentraciones de nitritos (NO_2) registradas para un punto de monitoreo del ecosistema estuarino el Riito. A partir de estos datos se desea determinar si las concentraciones medias de NO_2 no rebasan los límites permisibles establecidos por el decreto 1594 de 1984 de 1.0 mg/L, para la destinación del agua para consumo humano y doméstico.

NO_2 (mg/L)	0.43	0.30	0.26	0.20	0.05	0.13	0.25	0.38	0.52	5.58	0.34

Solución

Según el planteamiento del problema, el objetivo del análisis se centra en probar el siguiente sistema de hipótesis, a partir de los datos muéstrales suministrados

$$H_0 : \tilde{\mu}_{NO_2} = 1.0$$

$$H_1 : \tilde{\mu}_{NO_2} > 1.0$$

A partir de la información muestral, tenemos que los resultados del test de Shapiro-Wilk, no mostrados aquí pues su aplicación ya fue expuesta en el capítulo 6, son concluyentes en que los datos de concentración de NO_2 no se ajustan a una distribución normal. Siguiendo el procedimiento de

prueba del test de rangos con signos de Wilcoxon, a continuación se muestran los resultados de restar a cada observación el valor de la mediana $\tilde{\mu}_0$ que se desea probar en el sistema de hipótesis planteado.

$x_i - \tilde{\mu}_0$	-0.57	-0.70	-0.74	-0.80	-0.95	-0.87	-0.75	-0.62	-0.48	4.48	-0.66

Ahora procedemos a ordenar de forma ascendente las magnitudes absolutas de las diferencias de $x_i - \tilde{\mu}_0$, para luego asignar los rangos respectivos y realizar el proceso de inferencia.

$\lvert x_i - \tilde{\mu}_0 \rvert$	0.48	0.57	0.52	0.66	0.70	0.74	0.75	0.80	0.87	0.95	4.48
Rango	1	2	3	4	5	6	7	8	9	10	11
Signo	-	-	-	-	-	-	-	-	-	-	+

Como $n = 11$, para un nivel de significancia de 0.05, la tabla A.17 del apéndice muestra que para rechazar la hipótesis nula es necesario que $w_- \leq 14$. Evidentemente esto no se cumple, pues $w_- = 55$, de allí que se decide no rechazar H_0 y concluir con una confiabilidad del 95% que las concentraciones medias de NO_2 en el ecosistema estuarino el Riito no rebasan los límites permisibles establecidos por la norma colombiana (decreto 1594 de 1984) para la destinación del agua para consumo humano y doméstico.

A continuación mostramos la salida de resultados de R para la aplicación del test de rangos con signos de Wilcoxon, incluyendo el test de Shapiro-Wilk donde se demuestra que datos muéstrales de concentraciones de NO_2 no se encuentran distribuidos normalmente.

```
> NO2<-
c(0.43,0.30,0.26,0.20,0.05,0.13,0.25,0.38,0.52,5.58,0.34)

> shapiro.test(NO2)

        Shapiro-Wilk normality test

data:  NO2
W = 0.4251, p-value = 1.937e-07
> wilcox.test(NO2,alternative="greater",mu=1.0,correct= TRUE,
conf.int=FALSE,conf.level = 0.95)

        Wilcoxon signed rank test

data:  NO2
V = 11, p-value = 0.979
alternative hypothesis: true location is greater than 1
```

Observe que de esta salida de resultados el valor V = 11, corresponde al valor de w definido anteriormente, es decir, el menor valor entre W_+ y W_-. Sin embargo, nuestras inferencias las basaremos en el uso del p-valor, y este al ser mucho mayor que el nivel de significancia elegido de 0.05, nos brinda soporte suficiente para retener la hipótesis nula.

9.3. Test no paramétricos para la comparación de dos poblaciones con base en muestras independientes: Test U de Mann-Whitney.

Durante el desarrollo del capítulo 5, se usó la prueba *t* de *student* para comparar dos medias poblacionales independientes que se suponían tenían distribuciones normales (al menos aproximadamente). Una alternativa no paramétrica de este test para situaciones experimentales en las que la que a partir de la información muestral no es posible suponer distribución normal de las observaciones es el comúnmente llamado **test U de Mann-Whitney** o **test de suma de rangos de Wilcoxon**, usado ampliamente siempre que se pueda garantizar que las observaciones son de naturaleza continua.

Con base en lo anterior, el objetivo del análisis es probar la hipótesis nula $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$, contra alguna alternativa adecuada. Como todo método no paramétrico, el procedimiento de prueba de este test es bastante simple e intuitivo. En Walpole *et al.* (2007) se encuentra muy bien descrito y a continuación les expondremos una cita resumida del mismo:

- Primero se seleccionan muestras aleatorias de cada una de las poblaciones, no necesariamente del mismo tamaño.
- Denotamos por n_1 al número de observaciones de la muestra más pequeña y por n_2 al número de observaciones de la muestra más grande. Cuando las muestras sean de igual tamaño es indistinto a cual se asigna n_1 y n_2, es decir, se pueden asignar de forma aleatoria.
- Luego se ordenan las $n_1 + n_2$ observaciones de las muestras combinadas en orden ascendente y se asignan los rangos $1, 2, ..., n_1 + n_2$ para cada observación. Recuérdese que en caso de empates (observaciones idénticas), asignamos la media de los rangos que tendrían si las observaciones fueran distinguibles.
- A la suma de los rangos que corresponden a las n_1 observaciones de la muestra más pequeña se denotan por W_1. De manera similar, el valor W_2 representa la suma de los n_2 rangos que corresponden a la muestra más grande.

- Es fácil notar que el total $w_1 + w_2$ depende solo del número de observaciones en las dos muestras y de ninguna manera resulta afectado por los resultados del experimento. Así, de forma general,

$$w_1 + w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2},$$

Una vez que se determine w_1, puede ser más fácil encontrar w_2 mediante la siguiente expresión

$$w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - w_1.$$

- Para basar nuestra decisión en algún estadístico de prueba, calculamos el valor de

$$u_1 = w_1 - \frac{n_1(n_1 + 1)}{2} \text{ o } \qquad u_2 = w_2 - \frac{n_2(n_2 + 1)}{2}$$

Del estadístico relacionado U_1 o U_2, o el valor de u del estadístico U, el mínimo valor entre U_1 y U_2.

A partir de lo anterior, $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$ es rechazada para la alternativa bilateral $H_1 : \tilde{\mu}_1 \neq \tilde{\mu}_2$, cuando $u \leq u_{\alpha(n_1, n_2)}$. Para las alternativas unilaterales $H_1 : \tilde{\mu}_1 < \tilde{\mu}_2$ y $H_1 : \tilde{\mu}_1 > \tilde{\mu}_2$, H_0 es rechazada cuando $u_1 \leq u_{\alpha(n_1, n_2)}$ y $u_2 \leq u_{\alpha(n_1, n_2)}$, respectivamente. En la Tabla A.20 del apéndice se dan los valores críticos de u_1 y u_2 para niveles de significancia iguales a 0.001, 0.002, 0.01, 0.02, 0.025 y 0.05 para una prueba unilateral, y valores críticos de u para niveles de significancia iguales a 0.002, 0.02, 0.05 y 0.10 para una prueba bilateral.

En entorno de R, la modelación del test U de Mann-Whitney se realiza haciendo uso de la función ***wilcox.test*** descrita en la sección precedente, solo basta con incluir en su sintaxis el vector de datos ***y*** correspondiente a la segunda población con la que se realiza el contraste.

```
wilcox.test(x, y, alternative = c("two.sided", "less",
"greater"), mu = 0, correct = TRUE, conf.int = FALSE,
conf.level = 0.95, ...)
```

Ejemplo 9.2. A continuación se presenta una serie de datos de demanda biológica de oxigeno (BDO) registradas para dos estaciones de monitoreo del ecosistema estuarino el Riito. La primera estación de monitoreo (E_1), se encuentra situada en las proximidades de una zona institucional donde se sospecha existen vertimientos de aguas residuales sin tratamiento previo, la segunda estación de monitoreo (E_2) se encuentra ubicada en la desembocadura del rio Ranchería al mar caribe, donde se suponen las condiciones del mismo son más homogéneas. A partir de estos datos se desea determinar si los medios de DBO en E_1, son significativamente mayores que en E_2, a causa del vertido de aguas servidas. Asúmase que no es posible suponer normalidad de las observaciones.

Demanda biológica de oxigeno (DBO) (mg/L)	
E_1	E_2
5.65	2.38
2.40	2.70
1.21	1.95
2.45	1.52
1.25	1.08
1.90	2.11
0.27	0.20
1.20	0.78
1.25	1.17
1.25	1.17
3.17	1.50

Solución

De acuerdo a lo que plantea el enunciado del problema, el sistema de hipótesis que se desea probar es el siguiente

$$H_0 : \tilde{\mu}_{E_1} = \tilde{\mu}_{E_2}$$

$$H_1 : \tilde{\mu}_{E_1} > \tilde{\mu}_{E_2}$$

Ahora, ordenamos las observaciones combinadas de las dos muestras, E_1 y E_2, respectivamente, de forma ascendente y asignamos sus rangos respectivos, distinguiendo cuales observaciones pertenecen a cuales muestras, como se muestra a continuación, donde los valores seguidos por un asterisco corresponden a los rangos de las observaciones de DBO de E_1.

Datos originales	Rango	Datos originales	Rango
0.20	1	1.50	12
0.27	2*	1.52	13
0.78	3	1.90	14*
1.08	4	1.95	15
1.17	5.5	2.11	16
1.17	5.5	2.38	17
1.20	7*	2.40	18*
1.21	8*	2.45	19*
1.25	10*	2.70	20
1.25	10*	3.17	21*
1.25	10*	5.65	22*

A partir de los anterior tenemos que

$$w_1 = 2+7+8+10+10+10+14+18+19+21+22 = 141,$$

y

$$w_2 = \frac{(22)(23)}{2} - 141 = 112$$

Por lo tanto,

$$u_1 = 141 - \frac{(11)(12)}{2} = 75, \; u_2 = 112 - \frac{(11)(12)}{2} = 46.$$

Según la Tabla A.20 del apéndice el valor crítico es $u_{0.05\,(11,11)} = 30$.

Ahora, como $u_2 = 75 > u_{0.05\,(11,11)} = 30$, se decide no rechazar H_0 con un nivel de significancia de 0.05. La evidencia estadística anterior, es concluyente en que no existen diferencias significativas en el valor medio de DBO entre las estaciones E_1 y E_2, con una confiabilidad del 95%. Es decir, el vertido de aguas servidas en E_1, no produce un aumento significativo de la DBO respecto a las condiciones del mismo parámetro en E_2.

En seguida mostramos la salida de resultados de R para este problema, para examinar si se llegan a las mismas conclusiones.

```
> E1<-
c(5.56,2.40,1.21,2.45,1.25,1.90,0.27,1.20,1.25,1.25,3.17)
> E2<-
c(2.38,2.70,1.95,1.52,1.08,2.11,0.20,0.78,1.17,1.17,1.50)
> wilcox.test(E1,E2,alternative="greater",mu=0,correct=TRUE,
conf.int=FALSE,conf.level= 0.95)

        Wilcoxon rank sum test with continuity correction

data:  E1 and E2
W = 75, p-value = 0.1786
alternative hypothesis: true location shift is greater than 0
```

Nótese que para esta salida de resultados el p-valor = 0.1786 es mayor al nivel de significancia elegido de 0.05, por lo tanto, se decide no rechazar la hipótesis nula y concluir que no existen diferencias significativas en los valores medios de DBO en las estaciones E_1 y E_2.

9.4. Test no paramétrico sobre observaciones pareadas

Recuérdese que en la sección 5.7.3 se introdujo un procedimiento de prueba basado en la distribución *t* de *student* para evaluar la existencia de diferencias significativas entre las medias de dos poblaciones que se encuentran emparejadas, es decir, mediciones de alguna variable de estudio tomadas en tiempos diferentes sobre una misma unidad experimental luego de aplicado un tratamiento. En esta sección discutiremos la alternativa no paramétrica para este tipo de situaciones experimentales, específicamente describiremos el uso del test de rangos con signos de Wilcoxon descrito en la sección 9.2 para este propósito.

El procedimiento de prueba de este test cuando tratamos observaciones (datos) pareados es en esencia igual al descrito anteriormente, la única diferencia que salta a relucir radica en que la asignación de los rangos se hacen sobre las diferencias de las observaciones pareadas prescindiendo del signo de las mismas. Los empates o la presencia de diferencias iguales, se manejan de la misma forma en que se ha comentado antes, del mismo modo cuando alguna diferencia sea igual a cero, esta debe ser omitida del análisis y se debe ajustar el valor de *n* (Walpole *et al.*, 2007; Conavos, 1989).

Así, la hipótesis nula $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$ es rechazada para la alternativa bilateral $H_1 : \tilde{\mu}_1 \neq \tilde{\mu}_2$, cuando $w \leq w_{\alpha,n}$. Para las alternativas unilaterales $H_1 : \tilde{\mu}_1 < \tilde{\mu}_2$ y $H_1 : \tilde{\mu}_1 > \tilde{\mu}_2$, H_0 es rechazada cuando $w_+ \leq w_{\alpha,n}$ y $w_- \leq w_{\alpha,n}$,

respectivamente. Téngase en cuenta que los valores de $w \leq w_{\alpha,n}$ se encuentran en la Tabla A.19 del apéndice.

En R el modelamiento de este test se realiza, como antes, a través de la función ***wilcox.test***. Sin embargo para indicar al software que tratamos con observaciones pareadas, incluimos el argumento ***paired*** y le asignamos el valor lógico ***TRUE***.

```
wilcox.test(x, y, alternative = c("two.sided", "less",
"greater"), mu = 0, paired = TRUE, correct = TRUE, conf.int =
FALSE, conf.level = 0.95, ...)
```

A continuación veamos un ejemplo de aplicación del test de rangos con signos de Wilcoxon para observaciones pareadas, con su respectiva aplicación en el ambiente de R.

Ejemplo 9.3. Se afirma que una nueva dieta reducirá el peso promedio (en kilogramos) de una persona en un periodo de dos semanas. Se registran los pesos de 10 mujeres que siguen esta dieta, antes y después de un periodo de dos semanas y se obtiene los siguientes datos (Adaptado de Walpole & Myers, 2007):

Mujer	Peso antes	Peso después
1	58.5	60.0
2	60.3	54.9
3	61.7	58.1
4	59.0	62.1
5	64.0	58.5
6	62.6	59.9
7	56.7	54.4
8	63.6	60.2
9	68.2	62.3
10	59.4	58.7

Demuéstrese estadísticamente, a un nivel de significancia de 0.05, si este tratamiento dietario reduce significativamente el peso de las mujeres estudiadas.

Solución

Representaremos por $\tilde{\mu}_{Ant}$ y $\tilde{\mu}_{Des}$ a las medianas del peso de las mujeres estudiadas antes y después de someterse al tratamiento dietario. Así, el sistema de hipótesis de prueba es

$$H_0 : \tilde{\mu}_{Ant} = \tilde{\mu}_{Des}$$

$$H_1 : \tilde{\mu}_{Ant} > \tilde{\mu}_{Des}$$

Ahora, las diferencias d_i entre las observaciones de peso en kilogramos de las mujeres antes y después de la dieta se muestran a continuación

d_i	-1.5	5.4	3.6	-3.1	5.5	2.7	2.3	3.4	5.9	0.7

En seguida procedemos con ordenar las magnitudes absolutas de las d_i para asignarles sus rangos respectivos y realizar el proceso de inferencia.

d_i	0.7	1.5	2.3	2.7	3.1	3.4	3.6	5.4	5.5	5.9
Rango	1	2	3	4	5	6	7	8	9	10
Signo	+	-	+	+	-	+	+	+	+	+

De lo anterior encontramos que $w_+ = 1+3+4+6+7+8+9+10 = 48$ y $w_- = 2+5 = 7$.

Según la Tabla A.19 del apéndice el valor crítico para esta prueba es $w_{0.05,10} = 11$.

Ahora, como $w_- = 7 < w_{0.05,10} = 11$, se rechaza H_0 a favor de H_1 con un nivel de significancia de 0.05. Es decir, los resultados son concluyentes en que el verdadero valor promedio del pesos de las mujeres antes de sometidas al tratamiento dietario es significativamente mayor al peso de las mismas después del tratamiento, a un 95% de confiabilidad. Lo anterior es prueba confiable de la efectividad de la dieta estudiada.

Examinemos la salida de resultados de R para la resolución de este problema

```
> Antes<-c(58.5,60.3,61.7,59.0,64.0,62.6,56.7,63.6,69.2,59.4)
> Después<-
c(60.0,54.9,58.1,62.1,58.5,59.9,54.4,60.2,62.3,58.7)
>
wilcox.test(Antes,Después,alternative="greater",mu=0,paired=TR
UE, correct=TRUE,conf.int=FALSE,conf.level=0.95)

        Wilcoxon signed rank test

data:  Antes and Después
V = 48, p-value = 0.01855
alternative hypothesis: true location shift is greater than 0
```

Nótese que al ser el p-valor = 0.0186 < 0.05, se llega a la misma decisión de rechazar H_0 a favor de H_1 con un nivel de significancia de 0.05 y llegar a las mimas conclusiones.

9.5. ANOVA de un factor no paramétrico: test de Kruskal-Wallis.

Durante el desarrollo del capítulo 7 se discutió el análisis de varianza de un factor para probar la hipótesis nula de igualdad de *k* medias de poblaciones independientes, siempre que se cumplieran estrictamente los supuestos de normalidad de las observaciones y homogeneidad de varianzas de los grupos experimentales o tratamientos. Sin embargo, es común que en la realidad las observaciones colectadas durante la experimentación infrinjan estos supuestos, y como consecuencia de ello existió la necesidad de establecer procedimientos alternativos no paramétricos que permitieran, en esencia, buscar el mismo propósito que el ANOVA de un factor, siempre que por lo menos se cuente con observaciones provenientes de distribuciones continuas. Uno de estos procedimientos es el **test de Kruskal-Wallis**, presentado por W. H. Kruskal y W. A. Wallis en 1952, como una extensión del test de suma de rangos con signos de Wilcoxon o test U de Mann-Whitney, utilizado para probar la hipótesis nula de igualdad de *k* medianas de poblaciones independientes (Guisande *et al.*, 2011; Conavos, 1989; Walpole *et al.*, 2007).

Al igual que la prueba U de Mann-Whitney, el test de Kruskal-Wallis se basa en la combinación de todas las muestras aleatorias involucradas en el análisis para formar un solo conjunto de $n = n_1 + n_2, ..., n_k$ observaciones; entonces, estas se ordenan en orden ascendente de magnitud y se asigna un rango a cada observación comenzando con un rango 1 hasta terminar con un rango *n*, en caso de empates (observaciones idénticas), seguimos el procedimiento acostumbrado de reemplazar las observaciones por las medias de los rangos que tendrían las observaciones si fueran distinguibles. Cuando el rango de todas las observaciones está completo, se determina la suma de los rangos para cada muestra n_i , que denotaremos como r_i . En esencia, el test de Kruskal-Wallis determina si las disparidades entre las r_i respecto a los tamaños n_i de las muestras es suficiente para garantizar el rechazo la hipótesis nula. Con base en lo anterior, el estadístico de prueba de Kruskal-Wallis es

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1),$$

que se aproxima muy bien a una distribución chi cuadrado con $k-1$ grados de libertad, cuando H_0 es verdadera y si cada muestra consiste en al menos 5 observaciones.

El rechazo de H_0 se da cuando *h* cae en la región crítica $\chi^2_{\alpha(k-1)}$, a un nivel de significancia α elegido, de otra manera no rechace H_0.

En el ambiente de programación de R, modelar el test de Kruskal-Wallis es una tarea bastante sencilla, solo basta usar la función ***kruskal.test***, siguiendo la sintaxis que se muestra a continuación

```
kruskal.test(x, g, ...)
```

En cuyos argumentos ***x*** es un vector de datos con las observaciones de los *k* grupos, en su orden, y ***g*** es un factor que especifica los niveles de cada uno de los grupos experimentales o muestras involucradas en el análisis.

Veamos a continuación un ejemplo de aplicación del test de Kruskal-Wallis.

Ejemplo 9.2. Durante los años 2004 a 2005 el grupo de investigación Pichihuel de la Universidad de La Guajira, realizó un estudio sobre la dinámica fisicoquímica del ecosistema estuarino el Riito. Los datos referentes a las concentraciones de DQO (mg/L), desde noviembre de 2004 a septiembre de 2005, en cuatro diferentes estaciones de muestreo se muestran a continuación

	Estaciones			
Meses	E1	E2	E3	E4
Nov	42	55	40	48
Dic	190	199	148	139
Ene	58	61	55	97
Feb	136	147	133	135
Mar	91	321	515	516
Abr	136	147	133	130
May	119	91	62	112
Jun	123	118	110	237
Jul	128	145	160	363
Ago	44	159	287	133
Sep	91	83	77	78

A partir de estos datos evaluar si los valores de DQO presentan diferencias significativas en las cuatro estaciones de monitoreo. Asúmase que supuesto de normalidad es violado fuertemente.

Solución

Es evidente que al tener 4 grupos experimentales representados por las cuatro estaciones de monitoreo, y dado el incumplimiento del supuesto de normalidad de las observaciones, utilizaremos el test de Kruskal-Wallis para probar las siguientes hipótesis nula $H_0 : \tilde{\mu}_{E1} = \tilde{\mu}_{E2} = \tilde{\mu}_{E3} = \tilde{\mu}_{E4}$, contra la alternativa H_1: Al menos dos medianas son diferentes.

Ahora, después de haber combinado y ordenado las observaciones, se asignan los rangos a cada una de las observaciones de cada muestra, como se muestra en la tabla 9.1.

Tabla 9.1. Rangos de concentraciones de DQO en las estaciones de monitoreo.

Estaciones de monitoreo			
E1	**E2**	**E3**	**E4**
2	5.5	1	4
37	38	34	30
7	8	5.5	16
28.5	32.5	25	27
14	41	43	44
28.5	32.5	25	23
20	14	9	18
21	19	17	39
22	31	36	42
3	35	40	25
14	12	10	11
$r_{E1}=197$	$r_{E2}=268.5$	$r_{E3}=245.5$	$r_{E4}=279$

Ahora, con $n=11+11+11+11=44$, el valor del estadístico de prueba es

$$h=\frac{12}{(44)(45)}\left(\frac{197^2}{11}+\frac{268.5^2}{11}+\frac{245.5^2}{11}+\frac{279^2}{11}\right)-(3)(45)=2.1970$$

Como $h=2.1970<\chi^2_{0.05(3)}=7.815$ (Tabla A.5 del apéndice), tenemos insuficiente evidencia para rechaza la hipótesis nula con un nivel de significancia de 0.05, es decir, los resultados son concluyentes a una confiabilidad del 95% que no existen diferencias significativas en el verdadero valor promedio de DQO en las cuatro estaciones de monitoreo del ecosistema estuarino el Riíto.

Ahora, examinemos la salida de resultados de para este ejercicio luego de aplicar la función ***kruskal.test*** y comparemos los resultados

```
> DQO.E1<-c(42,190,58,136,91,136,119,123,128,44,91)
> DQO.E2<-c(55,199,61,147,321,147,91,118,145,159,83)
> DQO.E3<-c(40,148,55,133,515,133,62,110,160,287,77)
> DQO.E4<-c(48,139,97,135,516,130,112,237,363,133,78)
> DQO<-c(DQO.E1,DQO.E2,DQO.E3,DQO.E4)
> Estación<-rep(1:4,each=11)
> Estación<-factor(Estación,labels=c("E1","E2","E3","E4"))
> kruskal.test(DQO,Estación)

        Kruskal-Wallis rank sum test

data:  DQO and Estación
Kruskal-Wallis chi-squared = 2.1987, df = 3, p-value = 0.5322
```

Observe que los resultados de esta salida aportar evidencia suficiente (p-valor = 0.5322 > 0.05) para retener la hipótesis nula y concluir que no existen diferencias significativas en el valor medio de DQO a en las cuatro estaciones de monitoreo. Las pequeñas diferencias en el valor del estadístico de prueba, se debe a que R utiliza un factor de corrección para los casos en que se cuenta con menos de 5 observaciones en los grupos, pero como se puede observar no hace discrepar significativamente los resultados de la salida con los computados mecánicamente.

9.6. ANOVA para diseños de bloques completamente aleatorios no paramétrica: Test de Friedman.

El test de Friedman es el equivalente no paramétrico del ANOVA en experimentos de bloques completamente aleatorios que se discutió en la sección 7.5, usado para evaluar la existencia de diferencias significativas en los efectos de k tratamientos de un factor en presencia de un factor externo (factor perturbador conocido y controlable). De manera similar al procedimiento paramétrico, se crea un bloque para cada una de las n condiciones de los factores externos de tal manera que cada bloque contiene una observación proveniente de cada uno de los k tratamientos. Además, se supone que los tratamientos se asignan en forma aleatoria y que no existe ninguna interacción entre los bloques y los tratamientos. Las nk observaciones se arreglan como se ilustra en la tabla 9.2, donde los bloques corresponden a los renglones de la tabla y los tratamientos a las columnas (Conavos, 1989).

Tabla 9.2. Arreglo de las observaciones para el test de Friedman.

	Tratamientos					
Bloques	**1**	**2**	**…**	**j**	**…**	**k**
1	y_{11}	y_{12}	…	y_{1j}	…	y_{1k}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2k}
⋮	⋮	⋮		⋮		⋮
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{nk}

Al igual que en los otros procedimientos no paramétricos, la prueba de Friedman se basa en los rangos. Para cada bloque (renglón) se asigna un rango a las observaciones comenzando con un rango 1 y terminando con un rango *k*; entonces se suman los rangos para cada tratamiento. Sea r_j la suma de los rangos del *j*-ésimo tratamiento (columna). Si los efectos de los tratamientos son idénticos, r_j deberá tener prácticamente el mismo valor para cada *j*. por lo tanto, el procedimiento de Friedman cuándo una disparidad observada entre los r_j es suficiente para rechazar la hipótesis nula de igualdad de las medianas en los *k* tratamientos. Los empates en la asignación de los rangos es tratada de forma similar a como se ha visto en las secciones anteriores.

De esta forma el valor del estadístico de prueba para el test de Friedman está dado por

$$s = \frac{12}{nk(k+1)} \sum_{j=1}^{k} r_j^2 - 3n(k+1)$$

que para valores de *n* y *k* no muy pequeños se aproxima a una distribución chi cuadrado con $k-1$grados de libertad. Así, la hipótesis nula se rechaza a un nivel de significancia de α cuando $s \geq \chi^2_{\alpha(k+1)}$.

La aplicación de este test en R se realiza a través de la función ***friedman.test*** de forma fácil y rápida, solo basta con seguir la siguiente línea de código

```
friedman.test(y, groups, blocks, ...)
```

Donde el argumento ***y*** denota la variable dependiente objeto de estudio, ***group*** es un factor que especifica los *k* grupos experimentales o

tratamientos, y ***blocks*** es un factor que especifica los diferentes bloques formados durante el experimento.

Ejemplo 9.5. Considérese los datos del ejemplo 9.4 que se vuelven a mostrar a continuación y tómense a los meses de muestreo como bloques para examinar la existencia de diferencias significativas en los verdaderos valores promedio de DQO en las cuatro estaciones de monitoreo del ecosistema estuarino el Riito.

	Estaciones			
Meses	**E1**	**E2**	**E3**	**E4**
Nov	42 (2)	55 (4)	40 (1)	48 (3)
Dic	190 (3)	199 (4)	148 (2)	139 (1)
Ene	58 (2)	61 (3)	55 (1)	97 (4)
Feb	136 (3)	147 (4)	133 (1)	135 (2)
Mar	91 (1)	321 (2)	515 (3)	516 (4)
Abr	136 (3)	147 (4)	133 (2)	130 (1)
May	119 (4)	91 (2)	62 (1)	112 (3)
Jun	123 (3)	118 (2)	110 (1)	237 (4)
Jul	128 (1)	145 (2)	160 (3)	363 (4)
Ago	44 (1)	159 (3)	287 (4)	133 (2)
Sep	91 (4)	83 (3)	77 (1)	78 (2)

Solución

Los valores que figuran entre paréntesis en la tabla de datos corresponden a los rangos asignados a las observaciones de cada mes de monitoreo (bloques). Así, para cada estación de monitoreo, la suma de sus rangos es respectivamente $r_{E1} = 27$, $r_{E2} = 33$, $r_{E3} = 20$ y $r_{E4} = 30$. Entonces, el valor del estadístico de prueba de Friedman es

$$s = \frac{12}{(44)(5)}\left(27^2 + 33^2 + 20^2 + 30^2\right) - (3)(11)(5) = 5.0727.$$

Ahora, como $s = 5.0727 < \chi^2_{0.05\,(3)} = 7.815$ (según Tabla A.5 del apéndice), no se rechaza la hipótesis nula de igualdad de las mediadas de los valores de

DQO en las cuatro estaciones de monitoreo con un nivel de significancia de 0.05. Es decir, se concluye que no existen diferencias significativas en los verdaderos valores promedios de DQO en las cuatro estaciones de monitoreo con una confiabilidad del 95%.

A continuación veamos la salida de resultados de R para la aplicación del test de Friedman

```
> DQO.E1<-c(42,190,58,136,91,136,119,123,128,44,91)
> DQO.E2<-c(55,199,61,147,321,147,91,118,145,159,83)
> DQO.E3<-c(40,148,55,133,515,133,62,110,160,287,77)
> DQO.E4<-c(48,139,97,135,516,130,112,237,363,133,78)
> DQO<-c(DQO.E1,DQO.E2,DQO.E3,DQO.E4)
> Estación<-rep(1:4,each=11)
> Estación<-factor(Estación,labels=c("E1","E2","E3","E4"))
> Meses<-rep(1:11,length=44)
>                                                    Meses<-
factor(Meses,labels=c("Nov","Dic","Ene","Feb","Mar","Abr",
"May","Jun","Jul","Ago","Sep"))
> friedman.test(DQO,Estación,Meses)

        Friedman rank sum test

data:  DQO, Estación and Meses
Friedman chi-squared = 5.0727, df = 3, p-value = 0.1665
```

Obsérvese la convergencia en el cálculo del estadístico de prueba, y según el p-valor, al ser mayor que el nivel de significancia elegido, se retiene la hipótesis nula y acepta la igualdad del verdadero valor medio de DQO en las cuatro estaciones de monitoreo con una confiabilidad del 95%.

10.

Análisis multivariante de datos

10.1. Generalidades

A lo largo del desarrollo de este texto se han introducido una variedad de procedimientos estadísticos que tiene como finalidad la descripción o el análisis inferencial de máximo dos variables -con excepción del modelo de regresión lineal múltiple que se revisó en el capítulo 8- de naturaleza cuantitativa (pruebas de hipótesis para comparación de medias, regresión lineal simple y polinomial, etc.), cualitativa (pruebas de independencia) o mixta (ANOVA, regresión logística, etc.). Sin embargo, la mayoría de situaciones experimentales de la vida real involucran la medición y análisis simultaneo de muchas variables (cuantitativas o cualitativas) sobre un conjunto de casos o unidades experimentales, por esta razón en las últimas décadas han surgido diferentes procedimientos de **análisis multivariado o multivariante**, que consisten en un conjunto de técnicas estadísticas cuya finalidad es el análisis de datos en los que se dispone de más de dos variables medidas simultáneamente para las mismas unidades experimentales, y donde el análisis uni y bidimensional resulta limitado debido a que impide considerar toda la información existente simultáneamente e ignora los efectos conjuntos o interacciones entre las variables (Guisande *et al*., 2011).

Estos procedimientos de análisis multivariado de datos han sido tardíamente utilizados en el mundo práctico debido entre otras razones a la exigencia de los mismos de cálculos tediosos que solo pueden ser optimizados a través de computadores con capacidad y velocidad para ejecutarlos y que hasta épocas muy recientes se encuentran accesibles a cualquier usuario, a la necesidad de desarrollar software's específicos que permitieran el modelamiento de las técnicas multivariantes surgentes, y sin duda alguna a la no disponibilidad de textos comprensibles orientados fundamentalmente al conocimiento conceptual de los métodos con el mínimo contenido matemático posible y dirigidos hacia su aplicación práctica (Pérez, 2004).

En general los métodos multivariados de datos se pueden clasificar en **métodos de interdependencia** y **métodos de dependencia**. En los métodos de interdependencia se considera que todas las variables estudiadas tiene una importancia equivalente y ninguna de ellas destaca como dependiente

de las demás; en este tipo de métodos pueden estar persiguiéndose dos propósitos: por una parte, se puede buscar una reducción de la dimensionalidad de nuestro conjunto de datos, en el que a partir del número elevado de variables del que se disponga sintetizar nuevas variables no observables, combinación lineal de las originales, que recojan la mayor proporción de la información contenida en nuestros datos. En este caso es importante tener en cuenta la tipología de las variables para la elección del procedimiento estadístico multivariante más adecuado. Así, si las variables son de naturaleza cuantitativa, las técnicas multivariadas que permiten una reducción de la dimensionalidad del problema pueden ser el Análisis de componentes principales y el Análisis factorial, si las variables estudiadas son de naturaleza cualitativa, se puede acudir al Análisis de correspondencias. Por otra parte, si el propósito no es la reducción de la dimensionalidad de los datos, sino realizar una clasificación de los individuos en grupos más o menos homogéneos, la técnica predilecta para este tipo de situaciones es el Análisis clúster.

En los métodos de dependencia, no es aceptable suponer que todas las variables estudiadas tiene una importancia equivalente, porque alguna de las variables destaca como dependiente de las demás, en este escenario es imperativo acudir al uso de métodos inferenciales de explicación de una variables en función de las demás dependiendo de su naturaleza, como los modelos de regresión y análisis de varianza estudiados en secciones anteriores de este texto, y un método no discutido llamado Análisis discriminante cuyo objetivo es buscar una función lineal de varias variables que permita clasificar nuevas observaciones en los grupos existentes y denotados por una variable dependiente categórica.

En el desarrollo de este capítulo, revisaremos los fundamentos conceptuales y la aplicación práctica de cada una de las técnicas multivariantes mencionadas con el mínimo desarrollo matemático posible y profundizando más en aspectos prácticos de la vida real. Así mismo, cada una de las técnicas que se revisarán en el presente capitulo serán modeladas en el entorno de programación de R, que cuenta con una amplia gama de funciones y librerías (paquetes) para análisis multivariado de datos de uso bastante fácil e intuitivo.

10.2. Análisis de componentes principales (ACP)

Un problema central en el análisis de datos multivariados es la reducción de la dimensionalidad: si es posible describir con precisión los valores de p variables por un pequeño subconjunto $k < p$ de ellas, se habrá reducido la

dimensión del problema a costa de una pequeña perdida de información (Peña, 2002). El análisis de componentes principales (ACP), es un procedimiento estadístico multivariado clasificado dentro de los métodos de interdependencia que busca la simplificación o reducción de la dimensionalidad de los datos, cuando se disponen de un elevado número de variables de naturaleza cuantitativa, persiguiendo obtener un menor número de variables sintéticas (no observadas), combinación lineal de las variables originales, que se denominan componentes principales o factores, cuya posterior interpretación permitirá una interpretación más simple del problema estudiado (Pérez, 2004). Las nuevas variables o componentes principales deben ser luego nombradas e interpretadas, con el propósito de darle sentido a la obtención de las mismas, pues de nada serviría obtener nuevas variables, que aunque puedan ser calculadas, no permitan realizar una mejor explicación del fenómeno que se estudia.

La característica principal de la obtención de las componentes principales es que estas se forman como combinaciones lineales de las variables observadas y las componentes resultantes serán incorreladas entre sí, es decir, las componentes principales que se obtienen a través de asociaciones de las variables observadas (variables muy correlacionadas) deben ser independientes u ortogonales entre sí.

Las interrelaciones entre las variables observadas se puede observar a través de la matriz de covarianzas o a través de la matriz de correlaciones, dependiendo de la escala de medida de las variables, cuando el conjunto de variables es homogéneo, es decir, están expresadas en las mismas unidades de medidas, es aconsejable la utilización de la matriz de covarianzas. De otro modo, si las variables son heterogéneas, es decir, se encuentran expresadas en unidades de medidas diferentes, el análisis o la obtención de las componentes principales debe realizarse a partir de la matriz de correlaciones, lo que es equivalente a realizar una estandarización de las variables de tal manera que tengan media cero y desviación estándar igual a la unidad. Esto representa una pequeña perdida de información por ello la matriz de correlaciones solo debe ser usada para la construcción de las componentes principales cuando la escala de medida de las variables sea diferente.

Ahora como lo que se busca es reducir la dimensionalidad de nuestros datos, lo más lógico es explicar la mayor cantidad (proporción) de información de los mismos a través en un espacio de dimensión más simple y de más fácil interpretación e ilustración gráfica. Esta cantidad de información se expresa como variabilidad de los datos, así, a pesar de

obtenerse tantas componentes como variables observadas se tengan, algunas explicarán una variabilidad tan reducida de nuestros datos que podrán ser ignoradas, de tal forma que al retener pocas componentes (generalmente dos o tres) se explique la mayor proporción de variabilidad de las observaciones, que en la práctica y dependiendo del fenómeno estudiado se considera suficiente un 70% de la variabilidad.

La obtención de las componentes principales involucra una serie de cálculos tediosos que se decidió no incluir en este texto para no desviarnos del objetivo principal de este capítulo que es entender la aplicación práctica de los métodos multivariados que se introducirán en el mismo. Por ello, para mayor comprensión del método se ilustraran los apartes teóricos a través de un ejemplo de aplicación práctica en el entorno de programación de R a través de la función ***princomp*** del paquete básico de instalación del software.

Ejemplo 10.1. A continuación se muestran los datos de densidad poblacional (UFC/100 mL) de enterococos fecales (Ent), coliformes fecales (CF) y coliformes totales (CT), grupos bacterianos indicadores de calidad ambiental en cuerpos de agua, y los valores de diferentes variables fisicoquímicas medidas en cuatro estaciones de monitoreo de las playas turísticas de la zona urbana del municipio de Riohacha, La Guajira. Se busca establecer relaciones entre las variables medidas que ayuden a representar las diferentes estaciones de monitoreo en función de las mismas.

Est	Ent	CF	CT	Temp	pH	Cond	Sal	OD	Turb	NO2	NO3	NH4	PO4
E1	70	120	170	27.5	8.16	58.4	37.2	7.59	26.1	2.02	2.72	36.00	2.8
E2	300	280	460	27.2	8.21	58.1	37.3	8.89	67.5	2.50	2.99	123.71	3.8
E3	60	530	1100	26.4	8.08	53.6	37.6	8.09	37.4	1.59	2.34	84.21	1.6
E4	120	540	720	27.0	8.10	53.6	37.6	8.16	44.3	1.72	4.12	21.39	3.4
E1	250	2480	6080	28.3	9.13	38.7	23.0	7.63	74.6	0.40	2.90	13.40	7.7
E2	40	2360	5200	29.0	8.37	57.0	43.8	6.70	33.3	0.30	0.50	93.50	48.1
E3	120	1840	5760	29.0	7.86	54.6	36.6	7.51	33.0	0.03	3.60	111.40	9.1
E4	80	960	5840	28.8	8.16	57.9	35.4	7.55	57.3	0.80	3.80	93.40	41.0
E1	127	62	147	31.2	8.60	44.9	26.0	7.23	36.5	1.90	23.90	17.50	31.9
E2	103	126	231	30.5	8.57	56.0	33.0	7.23	35.1	1.30	10.50	21.40	15.0
E3	134	40	113	30.2	8.50	50.8	37.5	7.43	16.4	1.20	0.70	14.40	10.3
E4	232	104	223	30.3	8.54	55.7	37.2	7.67	18.1	1.50	1.00	33.70	8.0

E1	124	21000	21000	28.8	8.51	14.32	99.2	5.82	73.8	20.60	138.3	20.60	206.2
E2	391	322	598	31.7	8.65	47.5	22.6	7.38	46.8	20.60	91.3	92.40	41.2
E3	134	299	21000	30.8	9.09	54.8	2.9	7.68	13.9	8.60	69.6	101.00	84.2
E4	1035	47	134	31.5	9.15	74.6	5.3	7.70	12.2	5.20	31.3	8.70	51.3
E1	8	13	72	28.4	8.15	55.4	37.0	7.20	21.3	0.98	4.20	20.60	9.1
E2	21	71	489	28.9	7.50	53.8	35.7	7.62	28.5	2.72	15.98	20.60	2.5
E3	37	49	276	29.6	8.60	50.1	29.8	5.06	79.7	3.24	19.59	20.60	13.5
E4	34	14	167	29.6	8.44	53.3	31.8	4.58	61.9	3.34	9.56	20.60	5.7

Solución

Inicialmente, digitaremos nuestra tabla de datos en una archivo de Excel y lo guardaremos bajo con la extensión *.csv* al que llamaremos *ACP Riohacha.csv*, este archivo será luego importado desde R como un objeto de tipo data frame, haciendo uso de la función ***read.csv2***, como se ha mostrado en secciones anteriores. Luego se evaluará si existen fuertes asociaciones entre las variables a través de la construcción de su respectiva matriz de correlaciones, como se muestra en la siguiente salida de resultados de R, haciendo uso de la función ***cor*** del paquete "*stats*" discutida en el capítulo 8.

```
> Datos<-read.csv2("ACP Riohacha.csv", header=TRUE, encoding =
"latin1")
> cor(Datos[,2:14],method="pearson")
        Ent     CF     CT   Temp     pH   Cond    Sal     OD   Turb    NO2    NO3    NH4    PO4
Ent   1.000 -0.064 -0.103  0.422  0.538  0.359 -0.394  0.275 -0.208  0.230  0.183 -0.076  0.118
CF   -0.064  1.000  0.682 -0.109  0.059 -0.811  0.825 -0.308  0.397  0.611  0.716 -0.104  0.877
CT   -0.103  0.682  1.000  0.052  0.289 -0.545  0.310 -0.126  0.132  0.500  0.677  0.246  0.818
Temp  0.422 -0.109  0.052  1.000  0.561  0.085 -0.391 -0.253 -0.306  0.336  0.359 -0.176  0.218
pH    0.538  0.059  0.289  0.561  1.000 -0.067 -0.386 -0.141  0.034  0.272  0.324 -0.199  0.308
Cond  0.359 -0.811 -0.545  0.085 -0.067  1.000 -0.708  0.355 -0.549 -0.552 -0.626  0.186 -0.632
Sal  -0.394  0.825  0.310 -0.391 -0.386 -0.708  1.000 -0.275  0.386  0.317  0.353 -0.097  0.539
OD    0.275 -0.308 -0.126 -0.253 -0.141  0.355 -0.275  1.000 -0.410 -0.224 -0.262  0.365 -0.285
Turb -0.208  0.397  0.132 -0.306  0.034 -0.549  0.386 -0.410  1.000  0.222  0.199  0.027  0.188
NO2   0.230  0.611  0.500  0.336  0.272 -0.552  0.317 -0.224  0.222  1.000  0.956  0.076  0.737
NO3   0.183  0.716  0.677  0.359  0.324 -0.626  0.353 -0.262  0.199  0.956  1.000  0.022  0.869
NH4  -0.076 -0.104  0.246 -0.176 -0.199  0.186 -0.097  0.365  0.027  0.076  0.022  1.000  0.007
PO4   0.118  0.877  0.818  0.218  0.308 -0.632  0.539 -0.285  0.188  0.737  0.869  0.007  1.000
```

Se observa que existen relativamente muchos valores de correlación que sobrepasan el 50% y que pueden interpretarse como una fuerte asociación o relación entre las variables estudiadas, aunque el ideal es que estos valores de correlación sea al menos todos superiores al 60 – 70%, para propósitos ilustrativos, continuaremos con nuestro análisis.

A continuación, aplicaremos el análisis de componentes principales y realizaremos la interpretación de los resultados más relevantes.

```
> ACP<-princomp(Datos[,2:14],cor=TRUE)
> summary(ACP)
Importance of components:
                        Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
Stardard deviation     2.3011   1.6656   1.2586   0.9620   0.9180
Proprtion of           0.4073   0.2134   0.1219   0.0712   0.0648
Variance
Comulative             0.4073   0.6207   0.7426   0.8138   0.8786
Proportion
                        Comp.6   Comp.7   Comp.8   Comp.9   Comp.10
Stardard deviation     0.8295   0.6513   0.4932   0.3681   0.2433
Proprtion of           0.0529   0.0.26   0.0187   0.0104   0.0046
Variance
Comulative             0.9315   0.9641   0.9829   0.9933   0.9978
Proportion
                        Comp.11  Comp.12  Comp.13
Stardard deviation     0.1441   0.0721   0.0465
Proprtion of           0.0016   0.0004   0.0002
Variance
Comulative             0.9994   0.9998   1.0000
Proportion
```

Nótese que dentro de los argumentos de la función ***princomp*** se tomó de nuestra matriz de datos, las mediciones de la segunda a la treceava columna, que corresponden a los datos de naturaleza cuantitativa, la primera columna es una variable categórica que especifica la estación de monitoreo en la que se realizaron las mediciones de las variables ambientales estudiadas. Asimismo, se incluye el argumento ***cor***, al que le dio el valor lógico ***TRUE***, esto indica que se tome la matriz de correlaciones para efectuar el análisis, pues no todas las variables están expresadas en las mismas unidades de medida (son heterogéneas). En el caso en que las variables estén medidas en la mismas unidades, se deja el valor por defecto de la función para este argumento (***cor = FALSE***).

Entrando en materia, el resumen de resultados del análisis de componentes principales, nos muestra la desviación estándar de cada una de las componentes (raíz cuadrada de sus eigenvalores o valores propios), el porcentaje de varianza o variabilidad de los datos que es explicada por cada componente y por último el porcentaje de variabilidad acumulada por las componentes. Se observa que las tres primeras componentes explican un 74.26% de la variabilidad de los datos, según el criterio de la cantidad de información que recogen las componentes, estas tres primeras

componentes son un subespacio suficiente de representación de los datos. Otro criterio, para decidir cuantas componentes retener es a través de sus valores propios, es decir, escoger aquellas componentes que posean valores propios mayores que uno. En la siguiente salida de resultados de R, se muestra la forma en que se obtienen los valores propios mencionados.

```
> Eigenvalores<-ACP$sd^2
> Eigenvalores

Comp.1  Comp.2   Comp.3   Comp.4   Comp.5   Comp.6  Comp.7  Comp.8
5.2950  2.7744   1.5841   0.9254   0.8427   0.6680  0.4242  0.2433
Comp.9  Comp.10  Comp.11  Comp.12  Comp.13
0.1355  0.0592   0.0207   0.0052
```

Convergiendo con el criterio de la variabilidad acumulada, las tres primeras componentes resultar ser las que poseen valores propios mayores que la unidad e indica que estas componentes son las que deben ser retenidas para representar los datos en un espacio de dimensión más reducida. Este análisis también se puede realizar a través de un gráfico de los valores propios en orden decreciente, conocido como **gráfico de sedimentación** (scree plot) que se muestra en la Figura 10.1, y cuyas órdenes de programación para su construcción se muestran a continuación.

```
> plot(Eigenvalores, type="b", pch=16, xlab = "", ylab =
"Valores propios",font.lab=2,xaxt="n")
> axis(1,at=1:13,labels=names(Eigenvalores),las=2)
```

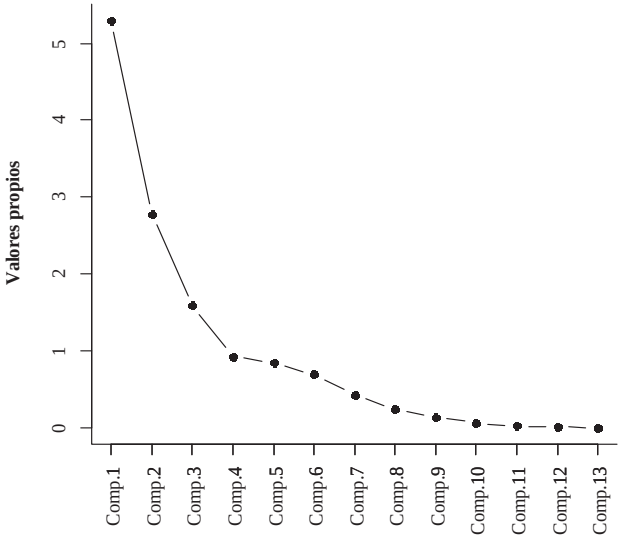


Figura 10.1. Gráfico de sedimentación ACP Ejemplo 10.1.

De este gráfico se retienen las componentes que se encuentran detrás del codo de la curva donde existe un cambio significativo en la pendiente y esta empieza a suavizarse, debido que después de este punto los valores propios tienden a estar tan cercanos de cero que pueden ser ignorados (Johnson, 2000). A partir de lo anterior, nuestro gráfico de sedimentación nos sugiere que deben ser retenidas las tres primeras componentes, decisión misma que se ha tomado con los otros criterios de extracción de las componentes principales.

Ahora, en la salida de resultados que se mostrará a continuación se muestras las instrucciones de programación para la obtención de las **cargas factoriales** o **saturaciones** de nuestro análisis de componentes principales.

```
> Cargas<-unclass(ACP$loadings)
> Cargas
        Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7
Ent     0,0285  0,4645  -0,0813 -0,1283 0,5849  -0,1413 0,3624
CF      -0,4038 -0,1108 -0,0378 -0,2034 0,1659  0,1088  0,2213
CT      -0,3266 0,0597  -0,2933 0,1177  -0,2537 0,5216  0,0760
Temp    -0,0482 0,4859  0,2112  -0,0204 -0,3857 -0,2081 -0,0695
pH      -0,0984 0,4587  0,1740  0,2842  0,1937  0,4745  -0,1010
Cond    0,3682  0,1802  -0,1251 -0,0170 -0,0507 -0,0977 0,5357
Sal     -0,2881 -0,3670 0,0148  -0,3191 0,1029  -0,1070 0,2210
OD      0,1838  0,0447  -0,5685 -0,2442 0,3627  0,1295  -0,4982
Turb    -0,1897 -0,2361 0,2143  0,6445  0,4386  -0,1224 -0,0070
NO2     -0,3458 0,2000  -0,1109 0,0185  0,0203  -0,5021 -0,2442
NO3     -0,3845 0,2023  -0,1075 -0,0268 -0,0630 -0,2531 -0,1787
NH4     0,0252  -0,0489 -0,6376 0,5022  -0,1891 -0,2037 0,2149
PO4     -0,4013 0,1168  -0,1257 -0,1293 -0,0507 0,1313  0,2776

        Comp.8  Comp.9  Comp.10 Comp.11 Comp.12 Comp.13
Ent     0,0251  0,1632  0,4035  0,0185  0,1199  -0,2493
CF      0,1476  0,0341  0,2044  -0,0359 -0,3664 0,7023
```

```
CT      -0,2184  0,3276   0,1675   -0,4647  0,1225   -0,1835
Temp    0,6045   0,3241   -0,0871  -0,1932  -0,0181  0,0628
pH      0,1185   -0,5857  -0,1505  -0,0544  -0,1155  0,0044
Cond    -0,2765  -0,0056  -0,5064  -0,3331  -0,2155  0,1632
Sal     0,3811   -0,3237  -0,1786  -0,3796  -0,0075  -0,4282
OD      0,2151   0,1659   -0,3083  -0,0950  -0,0323  0,0774
Turb    0,0770   0,3714   -0,2817  -0,1275  -0,0054  0,0078
NO2     -0,3114  -0,2738  -0,0217  -0,3221  0,4112   0,2652
NO3     -0,2818  0,0662   -0,1049  0,1795   -0,6719  -0,3471
NH4     0,3156   -0,2389  0,1809   0,1502   -0,0533  -0,0012
PO4     0,0222   0,1005   -0,4798  0,5507   0,3895   0,0090
```

Se observa que la primera componentes principal nos da una idea del grado de contaminación orgánica de las playas, pues se obtienen magnitudes absolutas altas de las cargas factoriales para las variables NO_2, NO_3, PO_4, CT, CF y Cond. La segunda componente está asociada al riesgo sanitario al que se exponen los usuarios de las playas de contraer enfermedades entéricas, dada la existencia de altas cargas factoriales en esta componente para las variables Ent, Temp, pH, Sal y Turb. Por último, no es fácil asignar una etiqueta a la tercera componente principal, sin embargo, se observa que está altamente explicada por las variables NH_4 y OD, lo que nos da una idea genérica de contaminación orgánica de reciente origen. En la Figura 10.2 se observa el gráfico de saturaciones en el plano de las dos primeras componentes retenidas, donde se ilustran las relaciones de las variables observadas con cada componente. Este gráfico corresponde a las siguientes órdenes de programación para su construcción.

```
> plot(Cargas[,1],Cargas[,2],type="n",xlab="Componente
1",ylab= "Componente 2")
> text(Cargas[,1],Cargas[,2],labels=rownames(Cargas),cex=0.7)
> abline(h=0,lty=2)
> abline(v=0,lty=2)
```

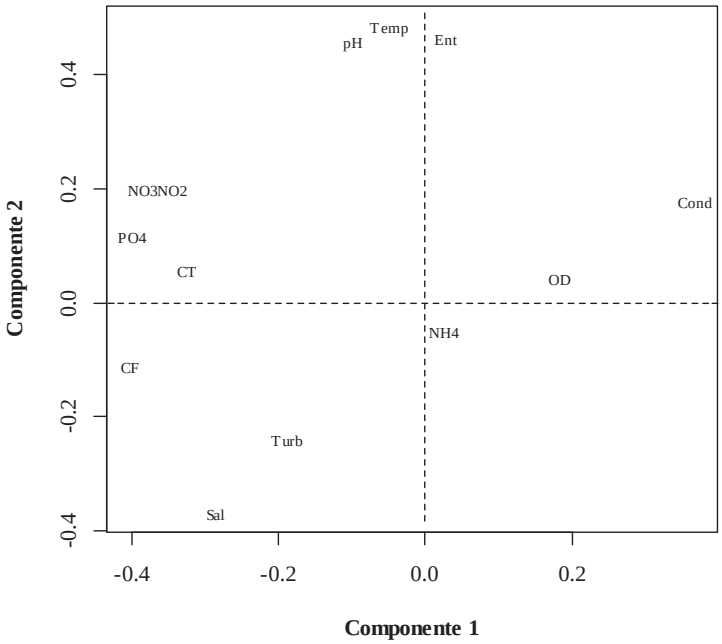


Figura 10.2. Gráfico de saturaciones en el plano de las dos primeras componentes.

Si realizáramos una análisis gráfico del problema, observamos que las variables que más se alejan en sentido horizontal del origen del plano, son las que poseen mayor valor absoluto de las cargas factoriales en la primera componente principal y por ello son las que mayor se encuentran relacionadas con la misma, asimismo, entre más agudo sea el ángulo de separación entre ellas, más correlación positiva existirá entre ellas, cuando el ángulo de separación entre las variables es de aproximadamente 180º, esto es indicativo de fuerte correlación negativa entre las variables. Por otro lado, cuando el ángulo de separación entre las variables es igual o aproximadamente igual a 90º, es indicativo de incorrelación entre las mismas o independencia. De la Figura 10.2, se observa además que las variables que más se alejan desde el origen en sentido vertical, son las que presenta mayores cargas factoriales absolutas para la segunda componente principal y son la que mejor la explican.

Para finalizar con nuestro análisis, analizaremos las relaciones entre las componentes retenidas y las estaciones de monitoreo, para ello debemos extraer las puntuaciones de cada una de las componentes y añadirlas nuestra matriz de datos, como se muestra en las siguientes ordenes de programación.

```
> Datos$CP1<-ACP$scores[,1]
> Datos$CP2<-ACP$scores[,2]
> Datos$CP3<-ACP$scores[,3]
> par(mfrow=c(1,2),family="serif",font.lab=2,bty="l")
>     plot(Datos[Est=="E1","CP1"],Datos[Est=="E1","CP2"],xlim=c(-
10,2),      ylim=c(-2,5),xlab="Componente      1",ylab="Componente
2",pch=19, col="blue")
>
points(Datos[Est=="E2","CP1"],Datos[Est=="E2","CP2"],pch=19,col=
"green")
>
points(Datos[Est=="E3","CP1"],Datos[Est=="E3","CP2"],pch=19,col=
"red")
>
points(Datos[Est=="E4","CP1"],Datos[Est=="E4","CP2"],pch=19,col=
"purple")
>  legend("topleft",legend=c("E1","E2","E3","E4"),pch=c(19,19,19,
19),col=c("blue","green","red","purple"),bty="n")
>     plot(Datos[Est=="E1","CP1"],Datos[Est=="E1","CP3"],xlim=c(-
10,2),   ylim=c(-2,3),xlab="Componente   1",ylab="Componente   3",
pch=19, col="blue")
>
points(Datos[Est=="E2","CP1"],Datos[Est=="E2","CP2"],pch=19,col=
"green")
>
points(Datos[Est=="E3","CP1"],Datos[Est=="E3","CP2"],pch=19,col=
"red")
>points(Datos[Est=="E4","CP1"],Datos[Est=="E4","CP2"],pch=19,col
= "purple")
>  legend("topleft",legend=c("E1","E2","E3","E4"),pch=c(19,19,19,
19),col=c("blue","green","red","purple"),bty="n")
```

En la Figura 10.3 se muestra el gráfico generado por las ordenes anteriores, en él se observa que en general, todas las estaciones de monitoreo muestran evidencia de contaminación orgánica, asociada con valores altos de densidad de coliformes y concentración de nutrientes (exceptuando el NH4). La estación E4 y en menor grado E3, exhiben el mayor riesgo sanitario para los usuarios de contraer enfermedades entéricas, asociados a la presencia de enterococos fecales. E1, E2 y en menor grado E4 son las estaciones más relacionadas con la tercera componente principal y se encuentran asociadas a altas concentraciones de NH4.

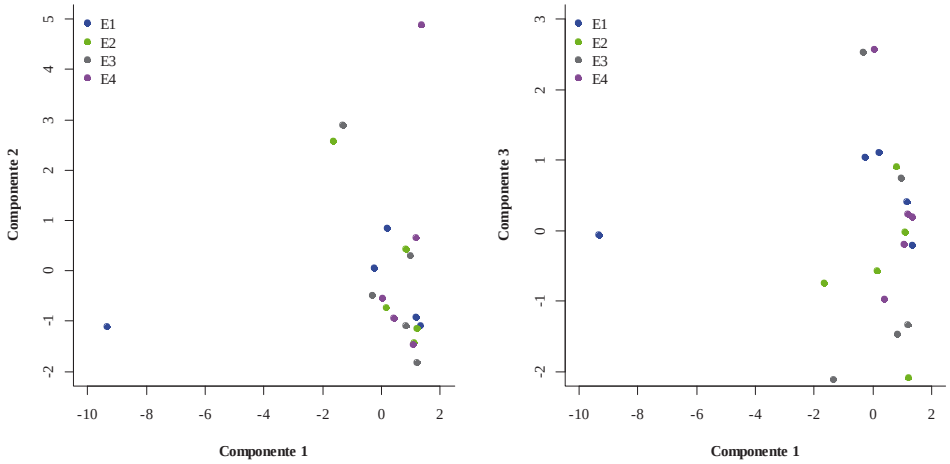


Figura 10.3. Gráfico de dispersión de las puntuaciones de las componentes principales del ejemplo 10.1 agrupadas por las estaciones de monitoreo.

10.3. Análisis factorial

10.3.1. Generalidades

El análisis factorial (AF) es una técnica de análisis multivariado, clasificada dentro de los métodos de interdependencia que busca básicamente la reducción de la dimensionalidad del problema planteado a través de la obtención de cierto número reducido de dimensiones comunes o **factores** no observables que explican las interrelaciones existentes entre las variables observadas (de naturaleza cuantitativa) y de las cuales tenemos datos u observaciones. Dicho de otro modo, el AF busca explicar las correlaciones entre p variables observadas, a través de $k < p$ factores no directamente observables, F_1, F_2,…, F_k, con la mínima perdida de información y que además tengan la características añadidas de ser independientes o incorrelados entre sí (ortogonales), sean fácilmente interpretables (principio de interpretabilidad) y sean en número los menores posibles (principio de parsimonia) (Pérez, 2004).

A simple vista podría pensarse que el AF es similar o busca el mismo propósito que el ACP, sin embargo, estos dos procedimientos son muy diferentes en su objetivo, características y grado de formalización, de hecho, lo único que poseen en común es que ambas son técnicas

multivariantes de interdependencia que permiten realizar una reducción de la dimensionalidad de nuestros datos.

En primer lugar, mientras el ACP busca la obtención de variables sintéticas que expliquen el mayor porcentaje de variabilidad de los datos, a través de meras relaciones matemáticas entre las variables, el AF se concentra en explicar la estructura de relación de las variables con la presunción de la existencia de factores que aunque no son observados, están latentes en nuestra tabla de datos y esperan ser encontrados. Ambas técnicas tienen propósitos inversos.

En segundo lugar, en ACP solo se establecen combinaciones lineales de las variables observadas para hallar las componentes principales que serán retenidas, y estas se explican a través de las saturaciones que cada variable tiene sobre las componentes. Muchas veces estas componentes carecen de sentido práctico y dificultan la interpretación y nombramiento de las mimas, pues solo existen a través de las relaciones matemáticas que exhiben nuestras variables pero que en la realidad no tienen ninguna utilidad práctica. Por el contrario, en AF la presunción de la existencia de factores comunes a las variables, siempre tienen utilidad práctica para la investigación y son establecidos por el conocimiento que el investigador posea sobre el fenómeno que estudia.

Por último, el ACP por sus características es situado por muchos autores entre los métodos multivariados de reducción de la dimensión, pertenecientes al dominio de la estadística descriptiva. En cambio, el AF implica la elaboración de un modelo que requiere la formulación de hipótesis y la aplicación de métodos de inferencia estadística (Pérez, 2004), como veremos más adelante.

10.3.2. Modelo factorial y obtención de los factores

Hasta el momento se ha hecho bastante énfasis en que el AF involucra un modelo matemático formal que establece a priori la existencia de ciertos factores $F_1, F_2,\ldots, F_k$, que ayudaran a explicar las complejas interrelaciones (correlaciones) de las variables observadas estandarizadas (tipificadas), $X_1, X_2,.., X_p$, con la mínima perdida de información y que el modelo con los factores –en número y significado- se corresponden con la realidad observada con nuestros datos (Pérez, 2004; Guisande *et al.*, 2011). El **modelo factorial** del que tanto se ha comentado es de la siguiente forma:

$$X_1 = \lambda_{11}F_1 + \lambda_{12}F_2, ..., \lambda_{1k}F_k + e_1$$
$$X_2 = \lambda_{21}F_1 + \lambda_{22}F_2, ..., \lambda_{2k}F_k + e_2$$
$$\vdots$$
$$X_p = \lambda_{p1}F_1 + \lambda_{p2}F_2, ..., \lambda_{pk}F_k + e_p$$

En este modelo, $F_1, F_2, \ldots, F_k$, son los **factores comunes** que explican las relaciones de todas las variables, $e_1, e_2, ..., e_p$, son los denominados **factores únicos** o **factores específicos**, adicionados al modelo para ayudar a explicar las relaciones de las variables que no logran ser explicadas por los factores comunes. Por último, λ_{ij} es el **peso** del factor *j* sobre la variable *i*, denominado también **carga factorial** o **saturación** de la variable *i* en el factor *j*. El modelo anterior expresa que cada una de las *p* variables es una combinación lineal de *k* factores comunes a todas las variables ($k < p$) y de un factor único para cada variable. Téngase siempre presente que tanto los factores comunes como los factores específicos son variables no observables, además ha de suponerse las hipótesis estadísticas que los factores comunes son variables estandarizadas con media cero, varianza igual a la unidad y se encuentran incorrelados entre sí. Por otra parte, los factores específicos se suponen incorrelados, con media cero y varianzas distintas para cada factor (Pérez, 2004).

A partir del modelo factorial planteado y de las hipótesis establecidas, se puede expresar la varianza de cada una de las variables estandarizadas X_i de la siguiente forma:

$$\sigma_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{ip}^2 + \varphi_i = 1$$

si establecemos que:

$$h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{ip}^2$$

Podemos reescribir la expresión para la descomposición de la varianza de la variable X_i como:

$$\sigma_i^2 = h_i^2 + \varphi_i = 1$$

Es decir, la varianza de cada una de las variables es igual a la suma de dos componentes: h_i^2, igual a la suma de los cuadrados de las cargas factoriales, que expresa la parte de la varianza de la variable X_i debida a los factores comunes, y que se denomina **comunalidad**, más una componente residual, φ_i, que denota la parte de la varianza de la variable X_i atribuida a los factores específicos, y que se denomina **especificidad**. Ambas cantidades se expresan a menudo en porcentaje o fracción de la varianza total de la variable. Así, cuanto más alta sean las comunalidades, mejor funciona el modelo factorial (Guisande *et al.*, 2011; Pérez, 2004; Johnson, 2000).

Las estimaciones de las cargas factoriales del modelo factorial se traducen en el problema fundamental del AF, pues a partir de estas se pueden determinar las comunalidades y especificidades de cada una de las variables, y a partir de las mismas realizar las inferencias pertinentes o a que haya lugar. Para las estimaciones de estas cargas factoriales son muchos los procedimientos existentes, no obstante, el denominado método de *máxima verosimilitud* es el predilecto por los estadísticos en la práctica, pues a diferencia de los demás métodos, resulta no ser un procedimiento directo, sino más bien un método inferencial basados en la teoría de probabilidad, además ofrece un test estadístico para evaluar la adecuación del modelo factorial con el número de factores seleccionados. El desarrollo matemático de este procedimiento se encuentra lejos del alcance de este texto, por ello no será tratado en el mismo. Sin embargo, se proporcionará un ejemplo de aplicación práctica del AF, donde se hará énfasis en la interpretación de cada uno de las estimaciones que se realicen. Este ejemplo se modelará en R a través de la función ***factanal*** del paquete "*stats*".

10.3.3. Contrastes en el modelo factorial

En el AF, aparte de la condición o requisito indispensable de que las variables observadas sean de naturaleza cuantitativa, se requiere que estas se encuentren altamente correlacionadas, por ello, es necesario la aplicación de ciertos test de adecuación de los datos para poder realizar el análisis. Así mismo, es importante evaluar la adecuación del modelo factorial planteado, es decir, que el mismo con sus factores tanto en número como en significado de los mismos se correspondan con la realidad observada. A continuación se describirán los test más usados para estos propósitos.

10.3.3.1. Test de esfericidad de Bartlett

Como se mencionó anteriormente, antes de realizar un AF debemos determinar si en esencia las variables observadas se encuentran correlacionadas, pues al no estarlo no existirían factores comunes y, por lo tanto, no tendría sentido emprender un AF. Una forma de determinar dicha correlación es a través de la inspección de la matriz de correlaciones. Sin embargo, cuando se cuenta con un elevado número de variables esta tarea resulta ser bastante dispendiosa y poco práctica. Otra alternativa es a través del denominado **test de esfericidad de Bartlett**. El principio de aplicación de este test se basa en la consideración de que al no existir correlación alguna entre las variables, su matriz de correlaciones (*R*) será en efecto una matriz identidad, compuestas por unos en su diagonal principal y ceros fuera de ella. Siendo así, el determinante de esta matriz ($|\boldsymbol{R}|$) será igual a uno, por lo tanto, la no correlación entre las variables se puede contrastar a través de las siguientes hipótesis:

$$H_0 : |R| = 1$$

$$H_1 : |R| \neq 1$$

Así, Bartlett introdujo un estadístico de prueba, que bajo tiene una distribución chi cuadrado con $p(p-1)/2$ grados de libertad.

$$B = -\left[n - 1 - \frac{1}{6}(2p+5) \right] \ln |R|$$

De esta forma, se rechaza H_0 cuando $B \geq \chi^2_{\alpha[p(p-1)/2]}$ a un nivel de significancia de α. Valores elevados de *B* que permitan el rechazo de H_0 indican que el determinante de la matriz de correlaciones es cercano a cero, lo que significa que tienen variables con altas correlaciones.

Cuando la decisión se basa en el uso del p-valor, se rechaza la hipótesis nula cuando este es inferior al nivel de significancia elegido.

El test de esfericidad de Bartlett en R es aplicado con el uso de la función ***cortest.bartlett*** del paquete "*psych*" (Revelle, 2015), a través de la siguiente línea de programación

```
cortest.bartlett(R, n = NULL)
```

Donde el argumento ***R***, especifica la matriz de correlación a partil de la cual se realizará el contraste y ***n*** es el tamaño de la muestra o número de individuos de nuestra tabla de datos, que no especificarse, el software toma por defecto un valor de 100.

10.3.3.2. Medida KMO de Kaiser, Meyer y Olkin de adecuación muestral

Otro mecanismo utilizado en el análisis factorial para evaluar la adecuación de nuestros datos para la realización del mismo, es a través de un indicador denominado *KMO* o índice de adecuación muestral, llamado asi por sus autores. Este indicador establece una relación entre los coeficientes de correlación observados de cada par de variables y sus coeficientes de correlación parcial mediante la siguiente expresión:

$$KMO = \frac{\sum\sum r_{ij}^2}{\sum\sum r_{ij}^2 + \sum\sum \rho_{ij}^2}$$

Donde r_{ij}^2 es el cuadrado del coeficiente de correlación lineal entre las variables *i* y *j*, y ρ_{ij}^2 es el cuadrado del coeficiente de correlación parcial entre dichas variables, excluyendo en ambas sumas los casos $i = j$. El coeficiente de correlación parcial entre pares de variables no ha sido discutido en este texto, se invita a los usuarios a consultar sobre el mismo. Sin embargo, para fines prácticos puede ser calculado en R a través de la función ***partial.cor*** del paquete "*RcmdrMisc*" (Fox *et al.*, 2015), que requiere la instalación previa del paquete "*Rcmdr*" (Fox *et al.*, 2015).

El indicador *KMO* toma valores entre valores 0 y 1, y su interpretación se basa en las siguientes reglas de decisión:

< 0.5	Inaceptable
0.5 – 0.6	Malo
0.6 – 0.7	Regular
0.7 – 0.8	Bueno
> 0.8	Excelente

En R, el indicador KMO es calculado de forma simple con el uso de la función ***KMO*** del paquete "*psych*" (Revelle, 2015), siguiendo la línea de código que se muestra a continuación

```
KMO(r)
```

Donde ***r***, denota la matriz de correlación a partir de la cual se realiza el cálculo del indicador.

10.3.3.3. Contraste de bondad de ajuste de máxima verosimilitud

Como se mencionó anteriormente la obtención de factores por el método de máxima verosimilitud ofrece una ventaja ante los otros procedimientos existentes en que proporciona un test que permite evaluar la adecuación del modelo factorial supuesto con el número de factores seleccionados. Es decir, se ofrece un test para contrastar las siguientes hipótesis:

H_0: *k* factores son suficientes para describir los datos.

H_1: *k* factores son insuficientes para describir los datos.

El estadístico de prueba para la hipótesis nula se basa en la función de máxima verosimilitud cuyo desarrollo matemático no ha sido ni será objeto de discusión de este texto, por su parte, el modelamiento de este test se encuentra incluida dentro de la función ***factanal*** del paquete "*stats*" de R. De este modo, las inferencias que se realicen para la adecuación del modelo factorial se explicarán cuando desarrollemos el ejemplo correspondiente. Lógicamente, se busca que el modelo factorial supuesto produzca una aceptación de la hipótesis nula, esta decisión de la salida de resultados de R se realizara a través de la inspección del p-valor, cuya aceptación de H_0 se conseguirá cuando este sea superior al nivel de significancia elegido.

10.3.4. Rotación de factores

Una de los mayores objetivos en el AF es que los factores tengan una interpretación clara, porque de esta forma se puede realizar una mejor evaluación de las interrelaciones que existen entre las variables observadas. Sin embargo en muy pocas ocasiones resulta fácil encontrar una interpretación clara de los factores. De allí, que han sido ideados los diversos métodos de rotación de factores para lograr que los mismos sean más fácilmente interpretables.

El principio de aplicación de los métodos de rotación de factores busca que las variables observadas tengan correlaciones altas (cercanas a uno) con

solo uno de los factores comunes y correlaciones bajas (cercanas a cero) con resto de factores. De esta forma y dado la existencia de más variables que factores comunes, cada factor estará fuertemente correlacionado con un grupo de variables e incorrelados con las restantes (Pérez, 2004).

Existen dos formas de realizar rotaciones de los factores del modelo factorial, a saber los métodos de **rotación ortogonal**, que son los más ampliamente utilizados por su característica distintiva de conservar la independencia de los factores luego de realizada la rotación, dentro de este enfoque existen diferentes métodos en los que el más utilizado en el método de rotación *Varimax*, que realiza una maximización de las varianzas de las cargas factoriales al cuadrado dentro de cada factor; otros métodos de rotación ortogonal son los métodos *Quartimax* y *Equamax*. El otro enfoque para la rotación de los factores en la **rotación oblicua**, de uso menor generalizado porque la rotación se realiza perdiendo la independencia o incorrelación de los factores, uno de los requisitos deseados del modelo factorial. Sin embargo, está perdida de independencia en ocasiones se ve compensada con una ganancia en la observación de la asociación de las variables con cada uno de los factores. Los métodos más destacados en este enfoque, son los procedimientos *Oblimin*, seguidos de otros menos utilizados como *Oblimax*, *Promax*, *Quartimin*, *Biquartimin* y *Covarimin*. Se invita a los lectores a la consulta más a fondo de estos procedimientos.

Para la función ***factanal*** del paquete "*stats*", solo se encuentra implementada la rotación *Varimax*, a través del argumento ***rotation***, por ser la de uso las recurrente. No obstante, la función ***fa*** del paquete "*psych*" (Revelle, 2015), para la ejecución de análisis factorial exploratorio, en su argumento ***rotate*** tiene implementado otros procedimientos para la realización de rotaciones tanto ortogonales como oblicuas. Se invita al lector a hacer uso de las mismas de acuerdo a sus necesidades.

10.3.5. Puntuación o medición de los factores

En la mayoría de las situaciones prácticas, el AF es un paso previo para la aplicación de otros procedimientos estadísticos en donde se sustituyen las variables originales por lo factores encontrados para resumir o simplificas las estructuras de correlación de dichas variables, tal es el caso en el que se emplean modelos de regresión usando como variables de respuesta a los factores encontrados. Por ello, es necesario conocer el valor que toman los factores en cada una de las observaciones. Estos valores son las llamadas **puntuaciones** o **mediciones factoriales**. Estas son estimadas por diversos

métodos, entre los que destacan por su uso más generalizado, e implementación en los diferentes paquetes estadísticos existentes los de *Mínimos cuadrados*, *Regresión*, *Anderson-Rubin* y *Bartlett*. No obstante, en la función ***factanal*** del software R, solo se encuentran implementados lo procedimientos de *Regresión* y *Bartlett*, por ser los de uso más recurrentes, a través de la inserción dentro de esta función del argumento ***scores*** asignándoles los valores tipo cadena "*none*", "*regression*" o "*Bartlett*", si no se desea realizar el cálculo de las puntuaciones, el método elegido sea Regresión o Bartlett, respectivamente.

Ejemplo 10.2. En Guisande *et al.* (2011), se proporcionan los siguientes datos de variables limnológicas en varios lagos neotropicales. Específicamente, se tiene datos de concentración de nutrientes y carbono orgánico disuelto que indican el grado de productividad de cada lago, y valores de profundidad, pH, conductividad, oxígeno disuelto y temperatura, que indican la adecuación del hábitat referente a sus condiciones para la vida. A partir de estas observaciones se desea comprobar si estos factores, productividad y hábitat, explican razonablemente la estructura de correlación entre las distintas variables y responde a la realidad observada.

Área	Lago	NO2	NO3	NH4	PO4	SiO2	Carb	Prof.	pH	Cond.	OD	Temp.
Amazonas	Correo	0.121	0.360	6.321	1.088	205.581	5.83	8.70	5.60	99.17	3.76	27.37
Amazonas	Correo	0.097	0.346	4.941	0.920	192.249	5.34	9.00	7.33	24.03	4.15	27.97
Amazonas	Correo	0.191	0.402	4.587	1.025	198.490	5.83	6.20	7.23	32.97	4.18	28.20
Amazonas	Tarapoto	0.086	0.430	3.344	0.473	219.947	6.16	6.10	7.28	31.50	3.86	27.83
Amazonas	Tarapoto	0.113	0.404	4.236	0.715	223.957	6.14	1.,50	7.08	37.97	2.37	27.43
Amazonas	Tarapoto	0.181	0.423	4.286	0.396	225.563	6.72	9.20	6.59	38.80	3.10	27.53
Amazonas	Tarapoto	0.135	0.470	9.079	0.644	198.085	6.15	6.50	6.41	70.60	2.70	27.60
Amazonas	Yahuarcaca	0.199	0.815	3.901	0.544	305.182	6.15	7.25	7.91	385.50	3.08	30.02
Amazonas	Yahuarcaca	0.141	0.578	3.133	0.301	306.138	6.49	6.90	7.88	223.67	2.45	30.27
Amazonas	Yahuarcaca	0.226	0.771	4.247	0.472	429.763	3.22	6.40	7.14	395.00	2.38	30.02
Amazonas	Yahuarcaca	0.268	0.620	4.232	0.385	325.059	5.21	6.30	6.90	349.00	1.15	28.00
Andes	Fuquene	0.038	0.292	3.657	0.017	22.918	3.35	1.65	7.37	121.20	5.53	18.68
Andes	Fuquene	0.044	1.336	2.773	0.018	22.459	3.71	2.20	7.31	123.33	3.15	17.98
Andes	Fuquene	0.048	1.264	2.354	0.000	25.627	3.90	4.30	7.41	121.75	3.48	18.78
Andes	Fuquene	0.020	0.865	2.647	0.000	19.905	3.53	0.93	8.08	131.25	5.96	18.45
Andes	Iguaque	0.052	0.183	0.994	0.000	0.128	2.80	4.50	6.66	10.53	5.10	12.90

Andes	Iguaque	0.034	0.457	0.816	0.000	0.080	2.70	4.60	6.67	957	5.20	12.23
Andes	Guatavita	0.010	0.595	7.957	0.000	45.500	3.21	9.60	7.09	16.60	4.63	16.65
Andes	Guatavita	0.007	0.470	6.521	0.000	22.436	1.86	19.75	7.33	14.75	4.48	16.63
Andes	Tota	0.069	0.872	2.317	0.024	18.250	1.59	50.00	7.35	95.60	5.45	16.65
Andes	Tota	0.077	0.932	1.749	0.000	19.766	1.60	46.00	7.40	94.95	5.51	16.60
Andes	Tota	0.085	0.990	4.440	0.000	19.691	2.26	28.10	7.25	97.35	5.06	16.75
Andes	Tota	0.043	1.438	1.216	0.000	21.435	1.70	27.70	7.30	96.60	5.49	16.40
Andes	Tota	0.053	1.528	1.776	0.015	22.005	1.97	23.50	7.35	96.20	5.67	16.35
Caribe	Momil	0.100	1.169	6.055	1.994	184.220	4.82	1.70	6.47	294.25	2.83	30.98
Caribe	S. Sebastian	0.119	1.926	6.392	1.846	177.000	4.70	2.60	6.95	261.60	0.76	29.32
Caribe	S. Sebastian	0.182	1.421	3.051	0.241	177.735	5.51	2.75	7,00	262.33	3.47	29.73
Caribe	Purisima	0.086	0.769	3.283	0.262	170.244	4.87	2.00	6.99	289.40	3.95	30.52
Caribe	Purisima	0.082	0.474	3.252	0.090	168.304	4.55	2.32	7.01	269.40	5.25	30.82

Solución

Para iniciar tabulamos los datos en un archivo de Excel y lo guardamos bajo la extensión *.csv* este lo cargamos en R y determinamos la matriz de correlación para realizar la inspección de la misma y establecer de forma descriptiva si nuestros datos se adecuan al modelo factorial.

```
> Lagos<-read.csv2("Lagos.csv",header=TRUE,encoding="latin1")
> r<-round(cor(Lagos[,3:13]),3)
> r
          NO2     NO3     NH4     PO4    SiO2    Carb    Prof      pH    Cond      OD    Temp
NO2         1  -0,065   0,194   0,398   0,861   0,608  -0,196  -0,159   0,577  -0,674   0,721
NO3    -0,065       1  -0,155   0,137  -0,187  -0,284   0,154   0,205   0,351   -0,15  -0,061
NH4     0,194  -0,155       1   0,527   0,342   0,387  -0,227   -0,37   0,022  -0,457   0,411
PO4     0,398   0,137   0,527       1   0,498   0,524  -0,312  -0,405   0,267  -0,609   0,642
SiO2    0,861  -0,187   0,342   0,498       1   0,729  -0,368  -0,125   0,599  -0,735   0,879
Carb    0,608  -0,284   0,387   0,524   0,729       1  -0,594  -0,217   0,224   -0,64   0,822
Prof   -0,196   0,154  -0,227  -0,312  -0,368  -0,594       1   0,174  -0,275   0,417  -0,452
pH     -0,159   0,205   -0,37  -0,405  -0,125  -0,217   0,174       1   0,115   0,227  -0,167
Cond    0,577   0,351   0,022   0,267   0,599   0,224  -0,275   0,115       1  -0,488   0,594
OD     -0,674   -0,15  -0,457  -0,609  -0,735   -0,64   0,417   0,227  -0,488       1  -0,675
Temp    0,721  -0,061   0,411   0,642   0,879   0,822  -0,452  -0,167   0,594  -0,675       1
```

De la matriz de correlaciones obtenida, se observa que la mayoría de las correlaciones entre las variables que se estudian son muy bajas e indican asociaciones lineales pobres entre las mismas, que sugieren una falta de adecuación de los datos para la ejecución de un AF. Sin embargo, esto es una inspección descriptiva que es preciso formalizar a través de los test de adecuación que se han discutido durante el desarrollo de esta sección, específicamente el test de esfericidad de Bartlett y la medida de adecuación KMO como veremos en la siguiente salida de resultados de R.

```
> library(psych)
> cortest.bartlett(r,n=29)
$chisq
[1] 225.2213

$p.value
[1] 1.804675e-22

$df
[1] 55
```

Nótese que el p-valor = 1.805×10^{-22}, al ser menor que un nivel de significancia elegido de 0.05, brinda evidencia suficiente para rechazar la hipótesis nula y afirmar que el determinante de la matriz de correlación de nuestros datos es diferente de uno, y por lo tanto, se concluye que los valores de correlación son significativamente diferentes de cero y existen relaciones entre las variables estudiadas. No obstante, una calificación de la adecuación de los datos a partir de las correlaciones entre las variables la podemos establecer a través de la medida KMO como se muestra a continuación

```
> library(psych)
> KMO(r)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = r)
Overall MSA =  0.66
MSA for each item =
 NO2  NO3  NH4  PO4 SiO2 Carb Prof   pH Cond   OD Temp
0.74 0.27 0.76 0.81 0.71 0.61 0.57 0.55 0.55 0.72 0.65
```

Observamos que el valor resultante KMO = 0.66, se sitúa dentro del rango de regular, por lo que la adecuación de nuestros datos para la ejecución de una AF no es muy buena. Sin embargo, seguiremos con la ejecución del análisis como un intento de alcanzar el objetivo que nos hemos planteado en el desarrollo de este ejercicio.

El paso a seguir sería ejecutar el análisis factorial en un espacio rotado a través del procedimiento *Varimax* sin extracción de las puntuaciones factoriales. A continuación se muestra la salida de resultados de R respectiva.

```
> FA<-factanal(Lagos[,3:13], factors = 2, method = "mle",
rotation = "varimax", scores = "none")
> FA

Call:

factanal(x = Lagos[, 3:13], factors = 2, scores = "none",
rotation = "varimax",     method = "mle")

Uniquenesses:

  NO2   NO3   NH4   PO4  SiO2  Carb  Prof    pH  Cond    OD
Temp
0.357 0.724 0.744 0.594 0.155 0.151 0.743 0.876 0.005 0.419
0.084

Loadings:

     Factor1 Factor2
NO2   0.802
NO3           0.526
NH4   0.364  -0.351
PO4   0.592  -0.235
SiO2  0.915
Carb  0.778  -0.494
Prof -0.492   0.123
pH   -0.145   0.321
Cond  0.724   0.686
OD   -0.757
Temp  0.947  -0.135

                Factor1 Factor2
SS loadings       4.826   1.323
Proportion Var    0.439   0.120
Cumulative Var    0.439   0.559

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 63.17 on 34 degrees of freedom.
The p-value is 0.00173
```

En la salida de resultados del análisis se muestra en primera medida las especificidades del modelo factorial (*Uniquenesses*). Se observa que cuatro variables tiene especificidades bastante bajas: SiO4 (0.155), Carb. (0.151), Cond. (0.005) y Temp. (0.084), por lo tanto, la comunalidad de estas variables es alta. Los anterior, quiere decir que la mayor parte de la variabilidad de estas variables esta explicada por los factores comunes y estas variables se corresponden con el modelo factorial planteado. Las otras variables por el contrario, al tener especifidades muy elevadas, y por lo tanto, comunalidades altas, no son explicadas adecuadamente por nuestro modelo.

Otra elemento importante de la salida de resultados obtenida para nuestro AF es el test de bondad de ajuste de máxima verosimilitud, de este se observa que arroja un p-valor = 0.00173 (cantidad sombreada en amarillo), valor bastante inferior al nivel de significancia elegido (0.05), lo que nos lleva a rechazar la hipótesis nula que establece que los dos factores elegidos son suficientes para describir nuestros datos, es decir, se requiere especificar un modelo más completo. Esta conclusión se corresponde con las altas especificidades que presentan ciertas variables que indican que no se encuentran explicadas por el modelo.

Analizando las cargas factoriales (valores sombreados en turquesa) dadas en la salida de resultados de nuestro AF, podemos observar que el primer factor está asociado a la cantidad de carbono orgánico disuelto, el NO_2, SiO_4, es decir, nutrientes y, por lo tanto, indicadores de la capacidad de producción del lago. Sin embargo, también está asociado a indicadores fisicoquímicos, como la concentración de oxígeno disuelto y la temperatura. En función de lo anterior, este factor opone a los lagos con bajas temperaturas, baja capacidad de producción y altas concentraciones de oxígeno disuelto de los lagos más productivos y con menos concentración de oxígeno. Lo anterior se observa en el gráfico de saturación de la Figura 10.4. El segundo factor se encuentra mayormente asociado con la conductividad del lago. Sin embargo, esta variable también se encuentra asociada al primer factor y, por lo tanto, no aporta información clara a nuestros propósitos. Además, si se observa el porcentaje de varianza explicado (valor sombreado en verde) por este factor, nos damos cuenta que este solo absorbe un porcentaje muy pequeño de la variabilidad de los datos (12.0%).

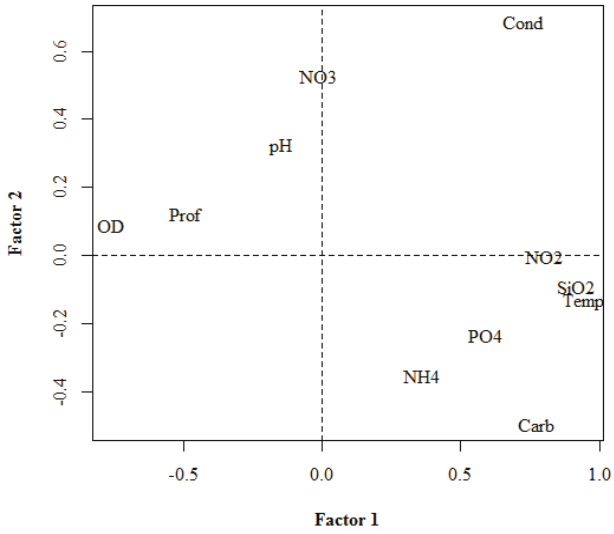


Figura 10.4. Gráfico de saturación del Ejemplo 10.2.

En general, los resultados de nuestro análisis muestran que nuestro modelo solo es adecuado para algunas de las variables estudiadas: conductividad, temperatura, oxigeno, carbono disuelto, NO_2 y SiO_4. El resto de variables no responden un modelo factorial, por lo se sugiere que este sea reformulado.

10.4. Análisis de correspondencias

10.4.1. Generalidades

El análisis de correspondencias (AC) se concibe como un método multivariado de reducción de la dimensionalidad de una tabla de casos-variables con datos cualitativos con el propósito de conseguir un número reducido de dimensiones, cuya posterior interpretación permitirá un estudio más simple del problema investigado (Pérez, 2004). Dicho de otro modo, el AC es una forma de representar datos cualitativos es un espacio de dimensión reducida, de forma análoga al ACP (Peña, 2002), pero con un tratamiento estadístico distinto dada la naturaleza cualitativa de los datos.

Por realizarse el tratamiento de variables cualitativas (o cuantitativas categorizadas), imprimen una característica diferencial a esta técnica estadística, pues los datos de análisis no son mediciones individuales, sino

que el análisis se realiza sobre una tabla de frecuencias absolutas, es decir, el número de veces que se encuentra un individuo en una casilla, de forma análoga al tratamiento que se discutió para variables cualitativas en el desarrollo de capítulo 6 cuando se introdujo la **prueba de independencia o asociación**, que realizaba un tratamiento conjunto para dos caracteres o variables cualitativas proporcionando información sobre la relación significativa o no entre ambas, sin explicar que categorías o modalidades estaban implicadas. Por su parte, el AC permite extraer estas relaciones entre categorías definiendo relaciones de similaridad o disimilaridad entre ellas, lo que permite su agrupamiento si se detectan que se corresponden, y todo esto es plasmado gráficamente en un espacio de dimensión reducida de variables sintéticas o factores que posteriormente deben ser interpretados o nombrados, y además, tengan la característica añadida de condensar el máximo posible de información (en tanto por ciento).

El AC, tiene dos clasificaciones diferenciadas: análisis de correspondencias simple, o comúnmente llamado simplemente análisis de correspondencia (CA), aplicado cuando el análisis se realiza solo sobre dos caracteres o variables cualitativas, cada una de las cuales representa varias modalidades o categorías, es decir, es un análisis aplicado sobre una tabla de contingencia, discutidas en el capítulo 6. De otro modo, cuando el método se extiende a más de dos caracteres o variables cualitativas, hablamos de un análisis de correspondencias multiple (ACM). Ambos métodos, serán objeto de discusión de este texto.

10.4.2. Análisis de correspondencias simple (AC)

Ya se dejó por sentado que el AC, tiene aplicación directa sobre tablas de contingencia donde se cruzan los efectivos existentes de las diferentes modalidades o categorías de dos caracteres o variables cualitativas (Husson *et al.*, 2013, Pérez, 2004). De este modo, si cruzamos en una tabla de contingencia el carácter I con modalidades desde $i = 1$ hasta $i = n$ (en filas), con el carácter J con modalidades desde $j = 1$ hasta $j = p$ (en columnas), podemos representar el número de unidades estadísticas que pertenecen simultáneamente a la modalidad i del carácter I y a la modalidad j del carácter J mediante k_{ij}. En este caso la distinción entre observaciones y variables es arbitraria. Sin embargo, por similitud con el análisis de componentes principales, suele hablarse de individuos u observaciones cuando nos referimos al conjunto de modalidades del carácter I (filas) y de variables cuando nos referimos al conjunto de

modalidades del carácter *J* (columnas), como se puede observar en la siguiente representación de una tabla de contingencia:

		J					
		1	2	⋯	*J*	⋯	*p*
I	1						
	2						
	⋮		⋯	k_{ij}	⋯		
	i						
	⋮						
	n						

De forma genérica puede considerarse que el AC y el ACP, tiene propósitos similares, específicamente de estudiar las relaciones existentes en el interior del conjunto de modalidades del carácter *I*, así como en el interior del conjunto de modalidades del carácter *J*, y estudiar las relaciones entre las modalidades del carácter *I* y las modalidades del carácter *J*.

El desarrollo matemático del AC, a pesar de ser relativamente simple y no involucrar técnicas matemáticas complejas, no será objeto de discusión de este texto, dado que al igual que las demás técnicas multivariadas, su desarrollo se ejecuta a través de paquetes estadísticos, que en nuestro caso particular corresponde al lenguaje de programación R. Por ello, presentaremos un ejemplo de aplicación práctica de esta técnica, y la modelaremos en el entorno de R a través de la función ***ca*** del paquete "*ca*" (Greenacre, 2015), interpretando las salidas de resultados más relevantes.

Ejemplo 10.3. A continuación se presenta datos relativos a un estudio fenológico del *Stenocereus griseus* realizado en el municipio de Manaure, La Guajira, durante un tiempo de ejecución del estudio de un año con realización de muestreos mensuales. En esencia se muestra una tabla de contingencia entre los estadios fenológicos y los meses de muestreo, donde se muestran las frecuencias absolutas de los individuos que presentan cada estadio fenológico en cada mes de muestreo. A partir de estos se desea conocer la relación entre los diferentes estadios fenológicos y los meses de muestreo para determinar cuál es el patrón hipotético de la secuencia de cada uno de los estadios fenológicos de esta especie.

Meses	BtnF	BtnA	FlrT	FlrM	FrtA	FrtI	FrtM	FrtD
Ene	7	1	1	0	2	4	0	0
Feb	11	3	2	0	2	10	0	0
Mar	11	2	4	1	6	13	1	3
Abr	13	1	4	1	2	12	0	3
May	8	0	1	1	1	12	0	1
Jun	12	3	2	4	4	9	0	1
Jul	13	2	7	4	4	10	0	3
Ago	7	5	5	3	4	4	14	2
Sep	12	1	3	2	2	11	1	0
Oct	5	1	0	2	0	10	1	0
Nov	4	0	1	0	1	8	0	3
Dic	4	0	0	0	0	6	0	3

Solución

Inicialmente, tabulamos esta tabla de datos, tal cual como se presenta, en una hoja de cálculo de Excel y la guardamos en nuestro directorio de trabajo bajo la extensión *.csv*. Luego ejecutamos R, seleccionamos nuestro directorio de trabajo e importamos este data frame como se muestra en la siguiente salida de resultados.

```
> Datos<-read.csv2("Fenofases.csv",header=TRUE,row.names=1,encoding
="latin1")
```

Obsérvese que en la importación de la base de datos, se anexo el argumento ***row.manes*** a la función ***read.csv2***, al que se le indica que las categorías de la primera columna representará las etiquetas o nombres de las filas.

Luego, cargamos el paquete "*ca*" y a través de la función ***ca*** del mismo paquete, ejecutamos el análisis de correspondencias simple (AC), tal como se muestra en las siguientes líneas de programación.

```
> library(ca)
> AC<-ca(Datos)
```

Los resultados más relevantes del AC, lo podemos obtener a través de un resumen (***summary***) del objeto que contiene los cálculos efectuados por la función ***ca***.

```
> summary(AC)

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.269211  61.8  61.8  ***************
 2      0.070663  16.2  78.0  ****
 3      0.045099  10.4  88.4  ***
 4      0.026059   6.0  94.4  *
 5      0.012198   2.8  97.2  *
 6      0.007946   1.8  99.0
 7      0.004136   1.0 100.0
        -------- -----
 Total: 0.435312 100.0

Rows:
     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1  |  Ene |   43  483   28 | -211 157   7 | -305 326  57 |
2  |  Feb |   81  416   44 | -212 188  13 | -233 228  62 |
3  |  Mar |  118  103   25 |  -93  92   4 |   32  11   2 |
4  |  Abr |  104  714   27 | -267 629  28 |   98  85  14 |
5  |  May |   69  606   42 | -381 553  37 |  118  53  14 |
6  |  Jun |  101  566   45 | -117  71   5 | -308 495 136 |
7  |  Jul |  124  173   51 | -119  80   7 | -129  94  29 |
8  |  Ago |  127  999  521 | 1333 993 837 |  104   6  19 |
9  |  Sep |   92  479   20 | -111 129   4 | -183 351  44 |
10 |  Oct |   55   21   54 |  -95  21   2 |   -7   0   0 |
11 |  Nov |   49  971   60 | -385 276  27 |  610 695 258 |
12 |  Dic |   37  933   82 | -453 215  29 |  830 719 365 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | BtnF |  308  744   44 | -182 535  38 | -114 209  57 |
2 | BtnA |   55  717   57 |  464 471  44 | -336 246  87 |
3 | FlrT |   86  168   50 |  156  96   8 | -134  71  22 |
4 | FlrM |   52  263   68 |  204  73   8 | -328 190  79 |
5 | FrtA |   81  242   46 |  101  41   3 | -223 201  57 |
6 | FrtI |  314  697  108 | -298 594 103 |  124 103  68 |
7 | FrtM |   49  987  504 | 2083 968 789 |  287  18  57 |
8 | FrtD |   55  786  123 | -181  33   7 |  860 753 573 |
```

Nótese que de esta salida de resultados se obtiene una inercia de $\phi^2 = 0.435$, este valor expresa la variabilidad total de nuestra tabla de contingencia. Cuando el valor del mismo es bajo, como en nuestro caso, el perfil de cada una de las variables se encuentra cerca de su perfil medio y en tal caso es un indicio de una pobre asociación o correlación entre las variables.

Ahora, el AC como todo método de reducción de la dimensionalidad busca la representación del perfil de las variables medidas en un espacio de dimensión reducida, para ello se realiza una descomposición de la inercia total (variabilidad) en la construcción de ejes ortogonales, y en nuestro caso observamos que las dos primeros ejes o dimensiones logran explicar un 78.0% de la inercia total, un valor que aunque se desea sea lo más grande posible, para nuestros fines prácticos resulta ser aceptable. Así mismo, se muestra un gráfico de sedimentación que corrobora lo anteriormente expuesto.

Luego se observan los resultados del AC efectuado para los meses de muestreo (filas) y para los estadios fenológicos (columnas), donde se exponen los valores de sus masas (*mass*), la calidad de representación (*qlt*) en el subespacio de las dos primeras dimensiones, las inercias de cada uno de los perfiles (*inr*), ya sean fila o columna, coordenadas ($k = 1$, $k = 2$), para la representación de los perfiles en las primeras dos dimensiones, las contribuciones relativas (*cor*) y absolutas (*ctr*) de cada uno de los perfiles (filas y columnas) a la inercia total. Todos los valores de esta tabla se encuentran expresados en "*tantos por mil*" (‰), con el objetivo de incluir tres cifras significativas sin la necesidad de utilizar decimales (Greenacre, 2008).

Así, los estadios fenológicos BtnF (botones florales) y FrtM (frutos maduros) son los que mayor peso (masa) poseen en el perfil medio de esta variable. La distinción entre masas grandes y pequeñas, se establece al compararse con un valor umbral definido como la media de las masas todos los perfiles (filas o columnas). Para los estadios fenológicos este valor umbral es 125. Los meses de muestreo, Mar, Abr, Jun, Jul y Ago, son los meses con mayores pesos.

Se observa que todas las modalidades de los estadios fenológicos se encuentran bien representadas en el plano de las dos primeras dimensiones (valores por encima del 60% o 600‰), solo pocos estadios fenológicos, específicamente, FlrT (floración total), FlrM (flores marchitas) y FrtA (frutos apareciendo), se encuentran regularmente representadas en este plano. Un análisis similar se realiza para los meses de monitoreo. El gráfico

de los perfiles de los estadios fenológicos y los meses de muestreo se ilustran en la Figura 10.5.

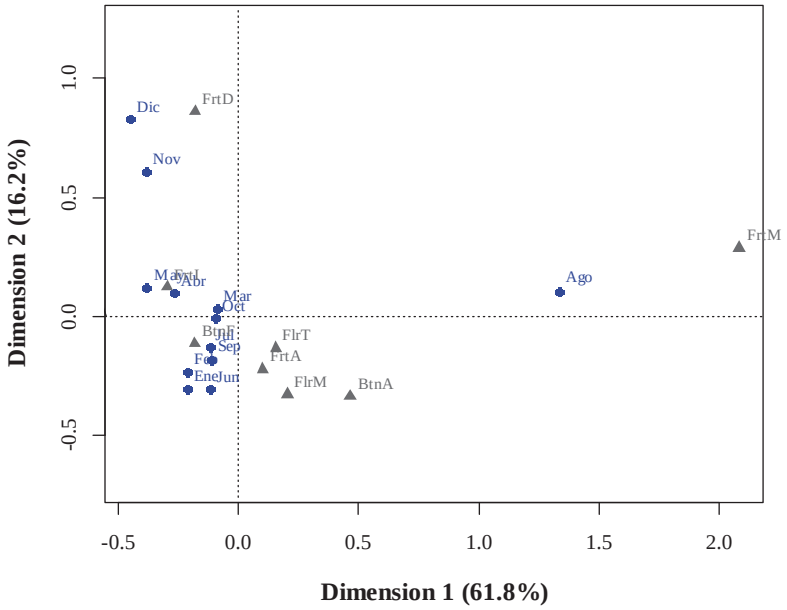


Figura 10.5. Resultado del AC del Ejemplo 10.3.

Las contribuciones absolutas, definidas como aquella parte de la inercia de cada eje o dimensión que es explicada por cada perfil (Guisande *et al.*, 2011), nos ayudan a determinar que perfiles son los que tiene mayor peso en la construcción de cada uno de los ejes dimensionales. De esta forma, el estadio FrtM es quien tiene mayor contribución en la construcción (78.9%) de la primera dimensión de nuestro plano de representación, análogo al mes Ago (83.7%). La segunda dimensión es mejor explicada por el estadios FrtD (57.3%), con asociación a los meses de Nov (25.8%) y Dic (36.5%). Los restantes estadios fenológicos y meses de muestreo al poseer contribuciones absolutas con valores bajos, exhiben una influencia pobre en la inercia de las dimensiones, lo que explica su concentración alrededor del origen del gráficos 10.5.

Lo anterior es confirmado con el análisis de las contribuciones relativas, expresadas como correlaciones al cuadrado o cuadrados de los cosenos del ángulo formado entre cada perfil y el eje dimensional (análogo a las cargas factoriales en AF), este indicador explica la parte de la inercia de cada perfil que es explicado por las dimensiones e indica hasta qué punto cada perfil esta adecuadamente explicado por las dimensiones (Guisande *et al.*, 2011).

Para nuestro ejemplo se observa que la primera dimensión explica mejor al estadio fenológico FrtM (96.8%) y en menor medida a FrtI (59.4%), estadios asociados a los meses de Ago, Abr y May con contribuciones relativas del 99.3, 62.9 y 55.3%, respectivamente. La segunda dimensión, explica mejor la inercia del perfil del estadio FrtD, con una contribución relativa del 75.3%, estadio asociado a los meses de Nov (69.5%) y Dic (71.9%).

En conclusión, el AC efectuado en el estudio fenológico del *S. griseus*, revela que el estadio frutos inmaduros, se presentan mayoritariamente en los meses de abril y mayo, mientras que la fructificación total, tiene mayor ocurrencia en el mes de agosto, y la dehiscencia de los frutos, ocurre en los meses de noviembre y diciembre. Los demás estadios fenológicos, exhiben un patrón regular de ocurrencia en todos los meses de muestreo.

10.4.3. Análisis de correspondencias múltiple (ACM)

En la sección anterior se discutieron los aspectos teórico-prácticos del análisis de correspondencia simple (AC), donde se analizaban simultáneamente las diferentes categorías o modalidades de dos variables cualitativas, a través del estudio de los efectivos existentes para las dos modalidades representadas en una tabla de contingencia. Este procedimiento es generalizable para tratar más de dos variables cualitativas de forma simultánea, tomando la denominación de análisis de correspondencias múltiple (ACM) (Pérez, 2004; Lagrand & Pinzón, 2009), es decir, es una aplicación particular del AC de tablas que cruzan individuos y sus respuestas para diversas variables cualitativas (Husson *et al.*, 2013).

Por realizarse el análisis sobre más de dos caracteres cualitativos ya no es posible la construcción de una tabla de contingencia y la representación de los datos se vuelve compleja. Sin embargo, no se pierde el horizonte del AC de estudiar las relaciones entre las diferentes modalidades de los caracteres cualitativos considerados. Como solución a esta complejidad de representación de los datos, en ACM, los datos se tabulan en una tabla denominada **tabla disyuntiva completa** (que representaremos como *Z*) que consta de un conjunto de individuos $I = 1, \ldots, i, \ldots n$ (en filas), un conjunto de variables o caracteres cualitativos $J_1, \ldots, J_2, \ldots J_Q$ (en columnas) y un conjunto de modalidades excluyentes $1, \ldots, m_k$. La totalidad de modalidades excluyentes será entonces $J = \sum_{k=1}^{Q} m_k$. La tabla disyuntiva completa es de dimensiones $I \times J$, como la mostrada a continuación.

	J_1		J_k		J_Q
	$1.....m_1$	…	$1.....m_k$	…	$1.....m_Q$
1					
⋮			⋮		
i			$\cdots\ z_{ij}\ \cdots$		
⋮			⋮		
n					

El elemento z_{ij} toma el valor 0 o 1 según que el individuo *i* haya elegido (este afectado) por la modalidad *j*. De allí, el nombre de tabla disyuntiva completa, porque cada individuo debe pertenecer a una y solo una de las modalidades, entonces aparece siempre 'uno' en un solo lugar bajo las modalidades pertenecientes a una sola variable, razón por la cual, algunos autores denominan a esta tabla como **tabla binaria**.

El procedimiento analítico del ACM, es efectuar un AC simple sobre la tabla disyuntiva completa y realizar la interpretación de los resultados. Este desarrollo matemático no está dentro de los alcances de este texto, de modo que se realizará la ejemplificación del ACM a través de una situación problema que corresponda a la realidad, modelada en R con la interpretación de las salidas de resultados más relevantes.

Otra alternativa para la realización del ACM, es a través de la denominada **tabla de Burt** (denotada por *B*), que consiste en una tabla que yuxtapone todas las tablas de contingencia de las variables cruzadas de dos en dos, por lo que también es común encontrar referencias de la misma en la literatura bajo la denominación de **tabla de contingencia múltiple**. Esta tabla, posee una apariencia similar a la mostrada a continuación

	J_1	J_2	…	J_Q
J_1	$0^{\cdot}\cdot.0$	C_{12}	…	C_{1Q}
J_2	C_{21}	$0^{\cdot}\cdot.0$	…	C_{2Q}
⋮	⋮	⋮	⋱	⋮
J_Q	C_{Q1}	C_{Q2}	…	$0^{\cdot}\cdot.0$

Los bloques de la diagonal de la tabla de Burt, son tablas diagonales que cruzan una variable con ella misma, siendo los elementos de la diagonal los efectivos de cada modalidad. Los bloques fuera de la diagonal corresponden a tablas de contingencia que se obtiene del cruce de las variables de dos en dos, cuyos elementos los las frecuencias de asociación de las dos modalidades correspondientes (Pérez, 2004). La tabla de Burt es simétrica y por lo tanto es suficiente mostrar la parte triangular inferior o superior. Las subtablas o bloques de la parte inferior son las transpuestas de las tablas de contingencias mostradas en la parte superior de la diagonal de *B*.

La realización de un ACM a partir de la tabla disyuntiva completa o a través de la tabla de Burt es idénticos para el cálculo de las coordenadas estándar de cada una de las modalidades. Sin embargo, el cálculo de las inercias principales a partir de la tabla de Burt, son los cuadrados de las inercias principales calculados a partir del análisis de la tabla disyuntiva completa. Como consecuencia de ello los porcentajes de inercia explicados serán mayores cuando se emplea la tabla de Burt y habrá una reducción de la escala del gráfico (mapa) bidimensional de correspondencias (Greenacre, 2008; Husson *et al.*, 2013).

La modelación del ACM, ya sea a través de la tabla disyuntiva completa o la tabla de Burt, se lleva a cabo en el entorno de programación de R a través de la función ***mjca*** del paquete "*ca*" (Greenacre *et al.*, 2015), la distinción de los dos procedimientos se realiza a través de un argumento propio de esta función como observaremos en el ejemplo práctico que veremos en breve.

Ejemplo 10.4. A continuación se presentan datos extraídos de Fine (1996), donde se presentan la caracterización de 27 razas de perros, de acuerdo a algunas variables físicas (tamaño, peso y velocidad), psíquicas (inteligencia, efectividad y agresividad) y una variable adicional que clasifica a los perros según su función. A partir de estos datos se desea realizar una clasificación de las diferentes razas estudiadas en función de las modalidades de los caracteres cualitativos considerados.

RAZ	TAM	PES	VEL	INT	AFE	AGR	FUN
bass	peq	liv	len	baj	baj	alt	caz
beau	gra	med	alt	med	alt	alt	uti
boxe	med	med	med	med	alt	alt	com

buld	peq	liv	len	med	alt	baj	com
bulm	gra	pes	len	alt	baj	alt	uti
cani	peq	liv	med	alt	alt	baj	com
chih	peq	liv	len	baj	alt	baj	com
cock	med	liv	len	med	alt	alt	com
coll	gra	med	alt	med	alt	baj	com
dalm	med	med	med	med	alt	baj	com
dobe	gra	med	alt	alt	baj	alt	uti
dogo	gra	pes	alt	baj	baj	alt	uti
foxh	gra	med	alt	baj	baj	alt	caz
foxt	peq	liv	med	med	alt	alt	com
galg	gra	med	alt	baj	baj	baj	caz
gasc	gra	med	med	baj	baj	alt	caz
labr	med	med	med	med	alt	baj	caz
masa	gra	med	alt	alt	alt	alt	uti
mast	gra	pes	len	baj	baj	alt	uti
peki	peq	liv	len	baj	alt	baj	com
podb	med	med	med	alt	alt	baj	caz
podf	gra	med	med	med	baj	baj	caz
poin	gra	med	alt	alt	baj	baj	caz
sett	gra	med	alt	med	baj	baj	caz
stbe	gra	pes	len	med	baj	alt	uti
teck	peq	liv	len	med	alt	baj	com
tern	gra	pes	len	med	baj	baj	uti

Solución

Antes de emprender nuestro ACM, tabulamos los datos en la hoja de cálculo de Excel, que guardaremos bajo la extensión *.csv* con el nombre *Razas perros.csv* para ser importada desde R como se muestra en la siguiente línea de código

```
>     Datos<-read.csv2("Razas     perros.csv",     header=TRUE,
row.names=1, encoding="latin1")
```

Nótese que se insertó el argumento ***row.names***, con el cual se indicó que la primera columna de nuestra base de datos sea tomada como el nombre de las filas de la misma.

Luego, cargamos la librería (paquete) *ca* y ejecutamos el ACM a partir de la tabla disyuntiva completa (o indicadora), siguiendo las estructuras de programación que se muestran a continuación en la que se establece en su argumento ***lambda***, la cadena de caracteres "*indicator*". Si se desea realizar el ACM a partir de la tabla de Burt, se modifica el argumento ***lambda***, especificando en él la opción "*Burt*".

```
> library(ca)
> ACM<-mjca(Datos,lambda="indicator")
```

Acto seguido a lo anterior, efectuamos un resumen de nuestro ACM para observar las salidas de resultados más relevantes del análisis. Esto lo conseguimos a través de la función ***summary***, como se muestra en la siguiente salida de resultados de R.

```
> summary(ACM)

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.521890  30.4  30.4  ********
 2      0.358471  20.9  51.4  *****
 3      0.222279  13.0  64.3  ***
 4      0.161566   9.4  73.7  **
 5      0.135078   7.9  81.6  **
 6      0.121884   7.1  88.7  **
 7      0.070051   4.1  92.8  *
 8      0.058946   3.4  96.3  *
 9      0.028306   1.7  97.9
 10     0.020155   1.2  99.1
 11     0.009942   0.6  99.7
 12     0.005716   0.3 100.0
        -------- -----
 Total: 1.714286 100.0

Columns:
       name   mass  qlt  inr     k=1 cor ctr     k=2 cor ctr
1  | TAM:gra |   79  839   55 |   819 838 102 |   -31   1   0 |
2  | TAM:med |   26  408   58 |  -809 149  33 | -1069 260  84 |
3  | TAM:peq |   37  726   78 | -1177 485  98 |   831 241  71 |
4  | PES:liv |   42  828   80 | -1168 575 111 |   776 253  71 |
5  | PES:med |   74  886   45 |   267  77  10 |  -867 810 155 |
6  | PES:pes |   26  606   72 |  1122 286  64 |  1187 320 104 |
7  | VEL:alt |   48  407   52 |   804 323  59 |  -409  84  22 |
8  | VEL:len |   53  721   50 |  -258  39   7 |  1077 682 171 |
9  | VEL:med |   42  473   49 |  -582 143  27 |  -886 330  93 |
```

```
10 | INT:alt |   32   79   36 |   390  44   9 |  -350  35  11 |
11 | INT:baj |   42  158   39 |   306  39   8 |   531 119  33 |
12 | INT:med |   69  151   29 |  -369 126  18 |  -165  25   5 |
13 | AFE:alt |   74  672   52 |  -777 650  86 |  -140  21   4 |
14 | AFE:baj |   69  672   56 |   837 650  92 |   151  21   4 |
15 | AGR:alt |   69  268   32 |   449 188  27 |   294  80  17 |
16 | AGR:baj |   74  268   30 |  -417 188  25 |  -273  80  15 |
17 | FUN:caz |   48  370   46 |   325  53  10 |  -796 317  84 |
18 | FUN:com |   53  767   72 | -1122 741 128 |   208  25   6 |
19 | FUN:uti |   42  623   70 |  1037 453  87 |   636 170  48 |
```

De esta salida de resultados, merece hacerse hincapié en que las dos primeras dimensiones explican un 51.4 % de la inercia total de nuestra tabla de datos, porcentaje suficientemente aceptables para fines prácticos de enseñanza del método. Inmediatamente se exponen los valores de los diferentes elementos de interpretación (especialmente las contribuciones absolutas y relativas), que son analizados de la misma forma que se hizo en el análisis de correspondencias simple.

Para ilustrar el procedimiento, construimos el mapa de las dos primeras dimensiones de nuestro ACM (Figura 10.6), de forma simple utilizando la función ***plot***.

```
> plot(ACM)
```

Este gráfico puede ser mejorado para aumentar su poder de ilustración, adicionando las razas caninas a las cuales están asociadas las modalidades de los caracteres cualitativos considerados en el estudio. Esta mejora, la realizamos a través de la construcción de un *biplot* (Figura 10.7), gráfico de amplio uso en análisis de datos multivariados que permite la representación de dos matrices de datos simultáneamente (Guisande & Vaamonde, 2012). Para ello es preciso realizar indexaciones (extracción de subconjuntos de datos) de los conjuntos de datos que serán representados en el biplot, uno de ellos corresponderá a las coordenadas de las razas caninas estudiadas y el segundo será las coordenadas de las variables consideradas en el estudio con sus respectivas modalidades. Así mismo, se crearán objetos que contendrán las etiquetas que serán representadas en el gráfico. Las instrucciones para las indexaciones y la construcción del biplot.

```
> x<-ACM$rowcoord
> y<-ACM$colcoord
> EtiX<-rownames(Datos)
> EtiY<-ACM$levelnames
> biplot(x, y, xlabs=EtiX, ylabs=EtiY,xlab="Dimensión 1
(30.4%)", ylab="Dimensión 2 (20.9%)")
> abline(v=0,lty=3)
> abline(h=0,lty=3)
```

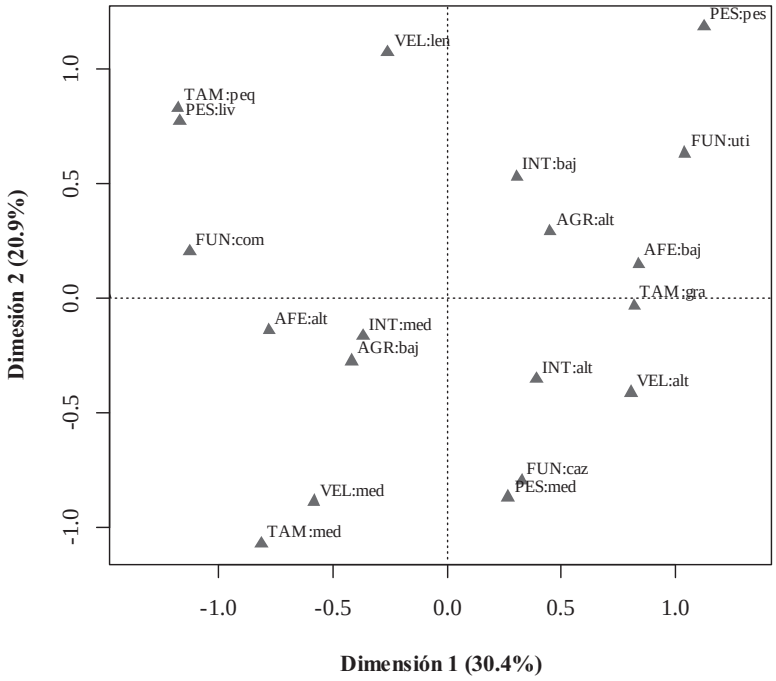


Figura 10.6. Mapa bidimensional del ACM del ejemplo 10.4.

Observamos que en la salida gráfica (mapa) del ACM, la primera dimensión opone a las razas caninas de tamaño grande, velocidad y agresividad alta, inteligencia media y alta y afectividad alta con funciones útiles (policivas, rescates, emergencias, etc.), a razas con pequeño tamaño, peso liviano, afectividad alta y baja agresividad con funciones de compañía (esto se establece con base a las contribuciones).

Por otra parte, la segunda dimensión distingue a razas poco veloces, pesados y de poca inteligencia a aquellos de tamaño, peso y velocidad media destinados a funciones de caza.

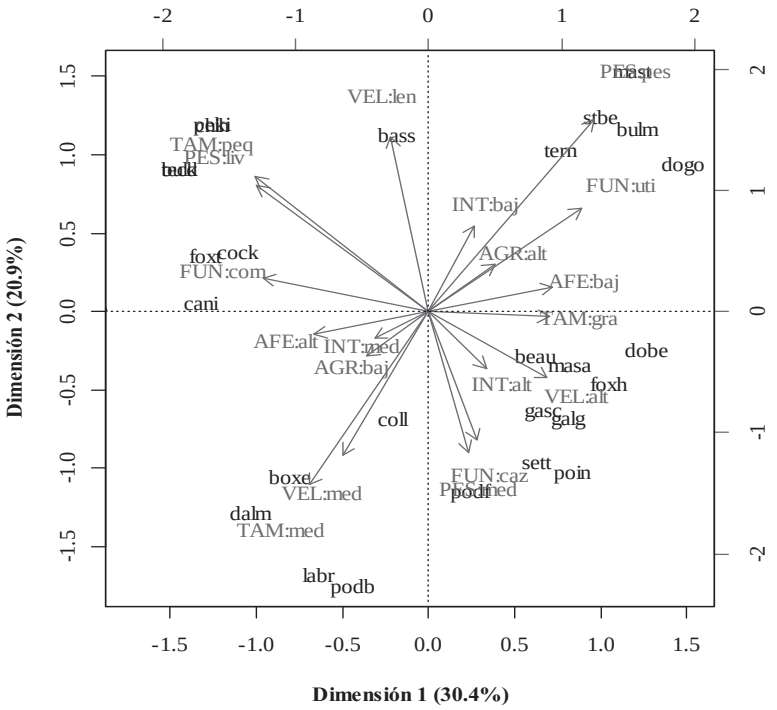


Figura 10.8. Biplot ACM ejemplo 10.4.

10.5. Métodos de clasificación: Análisis clúster

10.5.1. Generalidades

En el análisis multivariado de datos existe un conjunto de técnicas denominadas **métodos de clasificación**, utilizados con el propósito de realizar agrupaciones o clasificación de los elementos de una muestra de objetos descritos a través de la medición de varias variables (Guisande *et al.*, 2011), sin importar la naturaleza de las mismas (cuantitativas o cualitativas). En general, se busca la obtención de grupos homogéneos, denominados *clúster*, cuya principal característica es que los elementos que se encuentren contenidos en cada grupo sean similares entre sí, y a su vez, sean diferentes de los elementos contenidos en otros grupos. De esta forma, se puede comprender mejor la estructura de los datos que describen la realidad estudiada.

En oposición a las demás técnicas multivariadas que se han estudiado a lo largo del desarrollo de este capítulo, el planteamiento del análisis clúster

no tiene un enfoque algebráico, sino logarítmico, y depende del uso de la informática y equipos computacionales para llevarlos a cabo.

Una característica importante del análisis clúster, es que la creación de los clúster (agrupamientos), puede realizarse desde los individuos o las variables de nuestra matriz de datos, dependiendo de los objetivos que se persigan con la aplicación del método. Todos los algoritmos existentes en el análisis clúster utilizan como criterio de agrupamiento para la conformación de los clúster, la *proximidad* que existe entre los individuos (o variables). Desde el punto de vista de los individuos, sus proximidades son medidas a través de las distancias entre los mismos, y en análisis multivariado son muchas las formas para medir estas distancias, entre las que se citan las distancias *euclidianas* (usado por defecto en la mayoría de paquetes estadísticos), distancia de *Mahalanobis*, *Manhattan* o *city – block*, *Minkowski*, etc. (se invita al lector consultar acerca de ellas), los valores de todas estas medidas de distancias serán tanto mayor cuanto más alejados se encuentren los individuos. Por otra parte, las proximidades entre variables se miden a través de sus similitudes o grado de asociación que existe entre ellas, destacando el ya conocido coeficiente de correlación de Pearson, igual al coeficiente de correlación de Spearman y de Kendall, que contrario a las medidas de distancias tendrán valores tan altos como más cercanos (próximos) se encuentran los elementos considerados.

Otro aspecto a tener en cuenta en la aplicación de análisis clúster, es la forma en que queremos realizar el agrupamiento de nuestros datos y el tipo de algoritmo que se usará para ello. En el primero de los casos el método propone dos enfoques: la clasificación a través de **clúster jerárquicos**, en los que la agrupación se configura con una estructura arborescente, de forma que clúster de niveles más bajos van siendo englobados en otros niveles superiores; o a través de **clúster no jerárquicos**, donde los casos son asignados a grupos diferenciados que el mismo procedimiento de análisis configura, sin que unos dependan de otros (Pérez, 2004). En segundo lugar, el algoritmo o criterio de agrupamiento o construcción de los clúster, va a depender especialmente del objetivo que se persiga con nuestro a análisis (agrupación de individuos o variables) y de la medida de proximidad seleccionada. Son varios los métodos de aglomeración o análisis clúster disponibles, y los más destacados se discutirán brevemente en la siguiente sección.

10.5.2. Clúster jerárquicos

En el análisis de datos, son numerosas las situaciones experimentales en las que existe la necesidad de realizar clasificaciones de nuestros datos en grupos homogéneos que exhiban un patrón de dependencia a través de diferentes niveles de jerarquía, por ejemplo, la clasificación de especies biológicas (animales y vegetales), donde se realizan agrupaciones desde lo particular, es decir, las diferentes especies en grupos jerárquicamente superiores como los géneros, familias, ordenes, etc., hasta alcanzar un grupo que generalice a todos los individuos, que para este ejemplo en particular, lo constituye en reino, phylum o grupo al que pertenecen los individuos. Este tipo de agrupación es la denominada **clasificación jerárquica**, ampliamente utilizada en el tratamiento estadístico de datos y cuyo principio de aplicación consiste en realizar clasificaciones de los individuos (o variables) de nuestra tabla de datos, partiendo de tantos grupos iniciales como individuos se tengan y conseguir agrupaciones sucesivas entre ellos de forma que progresivamente se vayan integrando en clúster, los cuales a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde juntarán hasta llegar a un clúster final que contiene todos los casos analizados (Pérez, 2004).

La salida de resultados principal es un gráfico en forma de árbol invertido denominado dendograma o árbol jerárquico (Figura 10.9) donde se representan las diferentes etapas de formación de los grupos.

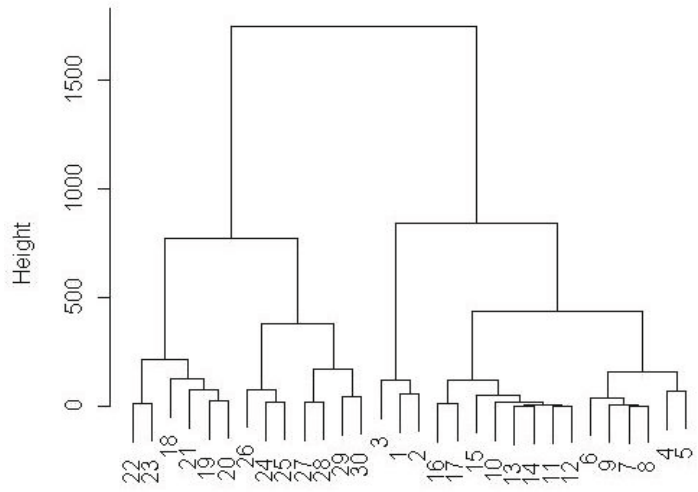


Figura 10.9. Dendograma o árbol jerárquico.

Evidentemente, la agrupación de los datos en cada uno de los conglomerados (clúster) se realiza a través de la similaridad (o distancia) entre el grupo de variables o de individuos que se están estudiando, ya que en cada nivel de jerarquía se unen los dos clúster más cercanos. De esta forma, es importante como paso previo a un análisis clúster jerárquicos, la elección de una adecuada medida de similaridad o disimilaridad.

Otra decisión importante que el investigador debe tomar es la elección del procedimiento de agrupación o algoritmo de formación de clúster, siendo muy variada y continuamente ampliada la oferta de dichos algoritmos. Pérez (2004), hace una exposición de los principales algoritmos de clasificación jerárquica, con una descripción de cada una de ellos y comentarios sus ventajas y desventajas, que sirven de ayuda al investigador a la hora de elegir el método que mejor se ajuste a sus necesidades. No obstante, el algoritmo con mejores resultados prácticos y de uso más extendido (Pérez, 2004, Langrand & Pinzón, 2009; Guisande *et al.*, 2011, Husson *et al.*, 2013) debido a que en cada paso de agrupación, se unen los elementos que dan lugar a una menor perdida de información, considerada como la suma de cuadrados de las distancias de cada objeto al centro de su clase (mínima varianza o inercia) (Guisande *et al.*, 2011).

Como se comentado anteriormente, el desarrollo de técnicas de análisis multivariado de datos han surgido y se han masificado gracias al creciente desarrollo informático de este siglo, por lo que se hace casi impensable la ejecución de un análisis clúster de forma mecánica, es imperativamente necesaria la utilización de computadoras y software`s para ello. R permite la realización de análisis clúster a través de diferentes rutas de programación. En primera medida en necesario la formación de una matriz de distancias (o similitud) como insumo a la formación de los clústeres, que el paquete de instalación básico de R se calcula a través de la función ***dist***, siguiendo la siguiente línea de código

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE)
```

Donde ***x*** es una matriz numérica o dataframe que contiene los datos a partir de los cuales se calculara la matriz de distancias, ***method*** establece la medida de distancia a ser utilizada y por defecto hace uso de la distancia euclidea, sin embargo permite el cálculo de otras medidas de distancia que pueden ser consultadas en la ayuda de esta función, el argumento ***diag***, asume un valor lógico que indica cuando la diagonal de la matriz de distancias será mostrada y ***upper***, asume un valor lógico que indica si la triangular superior de la matriz de distancias se mostrará en pantalla.

Otra función de R que permite realizar cálculo de la matriz de una variedad de medidas de distancias y de disimilaridad de uso más generalizado en estudios ecológicos es la función ***distance*** del paquete "*ecodist*" (Goslee & Urban, 2015), cuya línea de programación para su utilización se muestra a continuación

```
distance(x, method = "euclidean")
```

Donde similar a la función ***dist*** del paquete de instalación básico de R, ***x*** es una matriz numérica o dataframe que contiene los datos a partir de los cuales se calculara la matriz de distancias o de disimilitud, ***method*** establece la medida de distancia a ser utilizada y por defecto hace uso de la distancia euclidea, sin embargo permite el cálculo de otras medidas de distancia que pueden ser consultadas en la ayuda de esta función.

Una vez construida la matriz de distancia o disimilitud se procede con la ejecución del análisis clúster jerárquico, que bajo el paquete de instalación básico de R se ejecuta a través de la función ***hclust***, siguiendo la siguiente orden de programación

```
hclust(d, method = "complete")
```

Donde ***d*** especifica la matriz de disimilitud o distancias usada la para la ejecución del análisis clúster, ***method*** especifica el método de aglomeración (agrupación) usado para la formación de los clúster, por defecto, utiliza el método de enlace completo, pero se pueden elegir entre una variedad de algoritmos que se encuentran programados en esta función y que se pueden consultar en la ayuda de la misma, incluyendo el comentado método de Ward, que se especifica asignando al argumento ***method*** la expresión "*ward.D*" o "*ward.D2*", dada la existencia de dos procedimientos.

Otras procedimientos para la ejecución de análisis clúster jerárquico en R se pueden realizar a través de las funciones ***agnes*** del paquete "*cluster*" (Maechler *et al.*, 2015) y ***HCPC*** del paquete "*FactoMineR*" (Husson *et al.*, 2015), esta última realiza el análisis sobre la salida de resultados de un análisis de componentes principales (función PCA del mismo paquete).

A continuación ilustraremos la aplicación del análisis cluster a través de un ejemplo práctico, que será modelado con la ayuda del ambiente de programación de R

Ejemplo 10.5. A continuación se muestra un conjunto de datos tomados de Guisande & Vaamonde (2012), sobre concentración relativa de 19 pigmentos, en relación a la concentración de clorofila a, de especies de fitoplancton pertenecientes a diferentes clases: Diatomeas, Clorofíceas, Dinofíceas, Cianofíceas y Criptofíceas. Se busca a partir de esta matriz de datos determinar si es posible diferenciar las clases de algas en función de su composición relativa de pigmentos.

Especies	Chc2	Chc1	Mca1	Prdn	Fcxn	X9cn	Vlxn	Ddnx	Dnxn	Allx	Zxnt	Lutn	Chb	Chbe	Chaa	Chae	ßucr	ßecr	ßßcr
Diatomeas-P. t.	0.422	0.106	0.064	0	1.992	0	0	0	0	0	0	0	0	0	0	0.044	0	0	0.093
Diatomeas-P. t.	0.415	0.117	0.111	0	1.815	0	0	0	0.031	0	0	0	0	0	0	0.013	0	0	0.099
Diatomeas-Ch. g.	0.242	0.377	0	0	1.208	0	0	0	0.008	0	0	0	0	0	0.146	0.026	0	0	0.065
Diatomeas-Ch. g.	0.271	0.852	0	0	1.153	0	0	0	0.006	0	0	0	0	0	0.061	0.026	0	0	0
Diatomeas-S. c.	0.595	0.322	0.062	0	1.031	0	0	0	0	0	0	0	0	0	0.089	0	0	0	0.006
Diatomeas-S. c.	0.845	0.493	0.021	0	1.415	0	0	0	0	0	0	0	0	0	0.058	0.017	0	0	0
Clorofíceas-Ch. sp.	0	0	0.028	0	0	0.173	0.031	0	0.007	0	0.235	0.981	0.199	0	0.012	0.011	0	0	0.051
Clorofíceas-Ch. sp.	0	0	0	0	0	0.179	0.02	0	0.007	0	0.264	1.047	0.222	0	0	0.012	0	0	0.034
Clorofíceas-Ch. a.	0	0	0	0	0	0.13	0.077	0	0.007	0	0.003	0.534	0.199	0.007	0.014	0.032	0	0.004	0.115
Clorofíceas-Ch. a.	0	0	0	0	0	0.145	0.083	0	0.003	0	0.002	0.565	0.202	0.004	0	0.011	0	0.001	0.117
Clorofíceas-D. sp.	0	0	0	0	0	0.112	0.096	0	0.02	0	0.006	0.584	0.347	0.009	0.139	0.017	0	0.005	0.122
Clorofíceas-D. sp.	0	0	0	0	0	0.126	0.116	0	0.024	0	0	0.603	0.186	0	0.124	0.01	0	0.001	0.008
Clorofíceas-N. sp.	0	0	0	0	0	0.133	0.034	0	0.001	0	0	0.758	0.231	0.005	0.005	0.013	0	0.277	0.063
Clorofíceas-N. sp.	0	0	0	0	0	0.147	0.039	0	0	0	0	0.787	0.225	0.003	0.017	0.015	0	0.054	0.023
Dinofíceas-A. m.	0.706	0	0	1.155	0	0	0	0.199	0	0	0	0	0	0	0	0	0	0	0.049
Dinofíceas-P. m.	0.793	0	0	1.17	0	0	0	0.066	0	0	0	0	0	0	0.024	0.053	0	0	0.058
Dinofíceas-P. m.	0.871	0	0	1.389	0	0	0	0.066	0	0	0	0	0	0	0.031	0.041	0	0	0.069
Dinofíceas-H. sp.	0.938	0	0	1.004	0	0	0	0.162	0	0	0	0	0	0	0.032	0.03	0	0	0.077
Dinofíceas-H. sp.	1.125	0	0	1.257	0	0	0	0.19	0	0	0	0	0	0	0.012	0.016	0	0	0.081
Cianofíceas-C. sp.	0	0	0	0	0	0	0	0	0	0	0.354	0	0	0	0.02	0.018	0	0	0.357
Cianofíceas-C. sp.	0	0	0	0	0	0	0	0	0	0	0.35	0	0	0	0.007	0.002	0	0	0.335
Cianofíceas-S. sp.	0	0	0	0	0	0	0	0	0	0	0.339	0	0	0	0.021	0.014	0	0	0.272
Cianofíceas-S. sp.	0	0	0	0	0	0	0	0	0	0	0.361	0	0	0	0.015	0.023	0	0	0.275
Criptofíceas-R. b.	0.383	0	0	0	0	0	0	0	0	0.808	0	0	0	0	0.056	0.013	0	0.097	0.023
Criptofíceas-R. b.	0.392	0	0	0	0	0	0	0	0	0.867	0	0	0	0	0.081	0.014	0	0.12	0.024
Criptofíceas-C. sp.	0.579	0	0	0	0	0	0	0	0	1.023	0	0	0	0	0.035	0.033	0.003	0.115	0.014
Criptofíceas-C. sp.	0.644	0	0	0	0	0	0	0	0	1.074	0	0	0	0	0.014	0.036	0.007	0.107	0

Solución

El procedimiento de análisis para este ejemplo, inicia con la tabulación de los datos, procedimiento que hemos acostumbrado realizar en la hoja de cálculo de Excel y exportamos bajo la extensión *.csv* con el nombre *Pigmentos_Algas*, para su consecuente lectura en el ambiente de programación de R, siguiendo las siguientes líneas de código

```
> Algas<-read.csv2("Pigmentos_Algas.csv", header=TRUE,
encoding = "latin1")
```

Siguiendo con el procedimiento, realizamos el cálculo de la matriz de distancias, lógicamente este procedimiento se hace con datos de naturaleza numérica, por ello debemos excluir la primera columna de la matriz de datos que corresponden a las etiquetas de las clases de algas.

```
> Distancia<-dist(Algas[,-1],method="euclidea")
```

A continuación, ejecutamos el análisis cluster sobre la matriz de distancias, utilizando como método de agrupamiento el algoritmo de Ward.

```
> Cluster<-hclust(Distancia,method="ward.D")
```

Por último, generamos el dendograma de nuestro análisis (Figura 10.10), para extraer las conclusiones más relevantes que salten a la vista. Para resaltar las agrupaciones que se formen en el árbol jerárquico construido, utilizamos la función ***polygon***, y con la indicación de coordenadas x e y, trazamos rectángulos que nos ayuden con la identificación de los grupos, como se muestra en las siguientes ordenes de programación

```
> polygon(x = c(0.6,0.6,6.4,6.4), y = c(-3.9,1.8,1.8,-3.9),
border = "brown")
> polygon(x = c(6.6,6.6,11.4,11.4), y = c(-3.9,0.6,0.6,-3.9),
border = "yellow")
> polygon(x = c(11.6,11.6,15.4,15.4), y = c(-3.9,0.6,0.6,-
3.9), border = "red")
> polygon(x = c(15.6,15.6,19.4,19.4), y = c(-3.9,0.3,0.3,-
3.9), border="blue")
> polygon(x = c(19.6,19.6,27.4,27.4), y = c(-3.9,1.3,1.3,-
3.9), border = "green")
```

Del gráfico generado, se puede observar que las diferentes clases de algas se agrupan perfectamente en función del tipo de pigmentos que tienen (Guisande *et al.*, 2011)

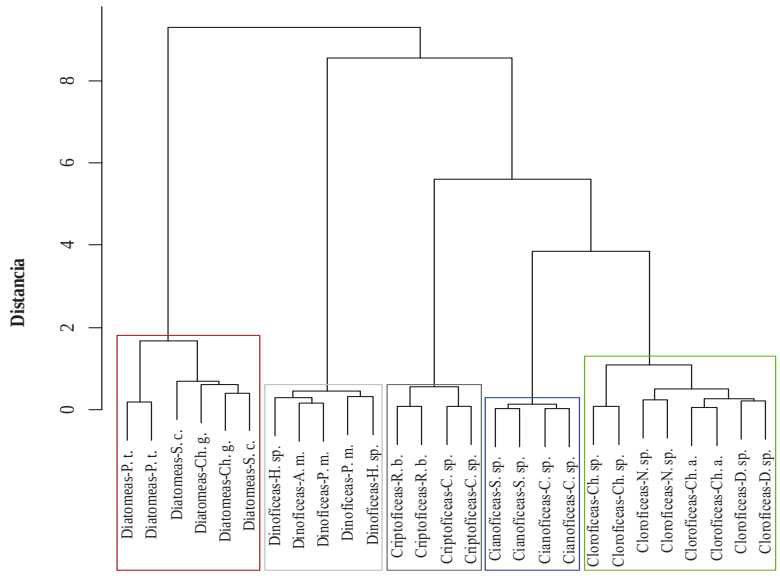


Figura 10.10. Clasificación de algas en función de sus pigmentos.

10.5.3. Cluster no jerárquicos: Clasificación de k medias

Anteriormente se hizo la distinción entre cluster jerárquicos y no jerárquicos, consistiendo estos últimos en procedimientos que permiten el agrupamientos de individuos de una matriz de datos sin ceñirse a una estructura jerárquica de dependencia entre los diferentes agrupamientos que se configuren en el desarrollo del análisis. En general, se busca la partición de los individuos de una matriz de datos en *k* grupos fijados de antemano por el analista, característica que lo diferencia respecto a los métodos jerárquicos. Esta cualidad, hace en muchas ocasiones necesario que el análisis sea realizado en varias repeticiones con la fijación de diferentes agrupamientos, a fin de encontrar la clasificación que mejor se ajuste a los objetivos del problema que se estudia y que brinde la mejor interpretabilidad.

Otra característica distintiva de los métodos no jerárquicos es que estos solo son aplicables para el agrupamiento de individuos (no de variables) y la medida exclusiva de proximidad utilizada para la conformación de los conglomerados en la distancia euclidiana. Así mismo, la aplicación del algoritmo de clasificación se realiza a través de la matriz de datos originales, es decir, contrario a los métodos jerárquicos no es necesario la conversión previa a una matriz de proximidades (Pérez, 2004).

Dada la no dependencia de las agrupaciones establecidas con el análisis, como un mecanismo para seleccionar la mejor clasificación, nos hemos de basar en aquella cuya variabilidad dentro de cada grupo (varianza intra-grupos) sea mínima, y máxima entre grupos diferentes (varianza inter-grupos). Este criterio, es denominado *criterio de la varianza* para la elección de la mejor clasificación de nuestros datos desde el punto de vista matemático, es preciso tener en cuenta la interpretabilidad de las agrupaciones realizadas.

La revisión de la literatura muestra la existencia de diferentes procedimientos de clasificación no jerárquica, basados en minimizar progresivamente la varianza intra-grupos, y el método de *k* medias es el más importante y de más amplio uso por sus eficiencia desde un punto de vista práctico y conceptual (Pérez, 2004), sobresaliendo los algoritmos de MacQueen (1967), Lloyd (1957), Forgy (1965) y Hartigan & Wong (1979), siendo este último el que mejores resultados provee en la asignación de los individuos en los diferentes grupos conformados.

Hasta este momento deben asaltar dudas de cuándo es preciso la utilización de métodos de agrupamiento jerárquico y no jerárquico y la respuesta va a depender en primera medida de los objetivos que se persigan en el desarrollo del análisis, en cuando al deseo de obtener una clasificación que siga una estructura de dependencia o no, a las dimensiones de nuestra matriz de datos, pues los métodos jerárquicos tiene limitaciones de cálculo cuando se tienen tablas de grandes dimensiones, problema que no afecta en lo absoluto a la aplicación de los métodos no jerárquicos, y por último, la existencia de datos atípicos que pueden ser identificados a través de los métodos no jerárquicos para después correr el análisis con la supresión de ellos y la consecuente obtención de mejores resultados y mejores interpretaciones de acuerdo al fenómeno que se intenta explicar.

El modelamiento del método de *k* medias en el entorno de programación de R ser realiza a través de la función ***kmean*** del paquete de instalación básica del software, siguiendo las siguiente ruta de programación

```
kmeans(x, centers, iter.max = 10, nstart = 1, algorithm =
c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE)
```

Donde ***x*** es la matriz de datos numérica con los individuos a clasificar y las variables medidas para tal caso, ***centers*** especifica el número de grupos *k*

fijados para el desarrollo del análisis, ***inter.max*** establece el número máximo de iteraciones que debe desarrollar el algoritmo para la clasificación de los datos, generalmente este número de iteraciones no se alcanza, ya que antes de que esto suceda se suele repetir la clasificación y se da por finalizado el algoritmo (Guisande & Vaamonde, 2012), ***nstart*** el número de centros iniciales de forma aleatoria, a partir de las cuales se realizará la reasignación de los individuos a nuevos grupos, ***algorithm*** especifica el algoritmo de clasificación a utilizarse, por defecto R establece el algoritmo de Hartigan – Wong, ***trace*** toma valores lógicos (*TRUE* o *FALSE*), actualmente solo usado si se selecciona el algoritmo de Hartigan-Wong, si se le asigna el valor *TRUE* a este argumento, se genera información del proceso de seguimiento del algoritmo.

A continuación, ilustraremos un ejemplo de aplicación de análisis cluster no jerárquicos a través del método de *k* medias sobre datos reales modelados en el ambiente de programación de R con la interpretación y discusión de los resultados más relevantes.

Ejemplo 10.6. A partir de los datos del ejemplo 10.5 sobre concentración relativa de 19 pigmentos, en relación a la concentración de clorofila a, de especies de fitoplancton pertenecientes a diferentes clases. Se persigue el mismo objetivo de determinar si es posible diferenciar las clases de algas en función de su composición relativa de pigmentos a través de un proceso de clasificación no jerárquica.

Solución

Iniciamos el análisis cargando en R nuestra matriz de datos, guardada en el archivo *Pigmentos_Algas.csv*.

```
>  Algas<-read.csv2("Pigmentos_Algas.csv",  header  =  TRUE,
encoding = "latin1")
```

Una vez cargada la matriz de datos ejecutamos el análisis utilizando el algoritmo propuesto por Hartigan & Wong (1979), con cinco agrupamientos (o grupos), pues contamos con cinco especies de algas, el número máximo de interacciones lo dejamos por defecto (10) y correremos el algoritmo con 10 centros aleatorios iniciales.

```
> kmedias<-kmeans(Algas[,-1], centers = 5, iter.max = 10,
nstart=10, algorithm = "Hartigan-Wong")
```

Dentro de los resultados más importantes del análisis podríamos observar el tamaño de los cluster formados o número de observaciones en cada agrupamiento, a través de la siguiente instrucción

```
> kmedias$size
[1] 8 4 4 5 6
```

Se observa que el primer cluster contiene 8 elementos, el segundo y tercer cluster contiene 4 individuos y el cuarto y quinto cluster poseen, respectivamente, 5 y 6 casos. Ahora, para observar la asignación que el algoritmo realiza para la conformación de los cluster, ingresamos la siguiente orden de programación

```
> kmedias$cluster
 [1] 5 5 5 5 5 5 1 1 1 1 1 1 1 1 4 4 4 4 4 3 3 3 3 2 2 2 2
```

El algoritmo de agrupamiento, realiza la asignación de forma arbitraria, de modo que es posible observar que el cluster nombrado automáticamente por R como 1 es quien contiene los 8 elementos observados en la salida de resultados anterior, 2 y 3, contienen cada uno cuatro elementos y así sucesivamente para los cluster restantes.

También es posible visualizar los centros de gravedad o valores medios de cada una de las variables involucradas en el análisis en cada uno de los cluster, para observar que grupos presentan fuertes asociaciones con las variables medidas. Los códigos de programación para tal caso se muestran en la siguiente salida de resultados

```
> kmedias$center
   Chc2  Chc1  Mca1  Prdn  Fcxn  X9cn  Vlxn  Ddnx  Dnxn  Allx  Zxnt  Lutn   Chb  Chbe  Chaa  Chae  ßucr  ßecr  ßßcr
1 0.000 0.000 0.004 0.000 0.000 0.143 0.062 0.000 0.009 0.000 0.064 0.732 0.226 0.003 0.039 0.015 0.000 0.043 0.067
2 0.500 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.943 0.000 0.000 0.000 0.000 0.047 0.024 0.002 0.110 0.015
3 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.351 0.000 0.000 0.000 0.016 0.014 0.000 0.000 0.310
4 0.887 0.000 0.000 1.195 0.000 0.000 0.000 0.137 0.000 0.000 0.000 0.000 0.000 0.000 0.020 0.028 0.000 0.000 0.067
5 0.465 0.378 0.043 0.000 1.436 0.000 0.000 0.000 0.008 0.000 0.000 0.000 0.000 0.000 0.059 0.021 0.000 0.000 0.044
```

Los resultados de mayor importancia del análisis son las sumas total de interna dentro de cada grupo (medida de la varianza intra-grupos) y la suma de cuadrados externa o entre grupos (medida de la varianza inter-grupos), pues estas permiten evaluar la clasificación resultante. Las líneas de programación para su cálculo se muestran en la siguiente salida de resultados.

```
> kmedias$tot.withinss
[1] 2.233424
> kmedias$betweenss
[1] 26.41152
```

Estos resultados muestran que la varianza intra-grupos (2.23) es bastante pequeña comparada con la varianza inter-grupos (26.41), evidencia de que el análisis generó una buena clasificación de los individuos, pues la dispersión entre los elementos del mismo grupo es pequeña (grupos homogéneos) y la dispersión entre grupos es grande (grupos separados unos de otros).

Para finalizar con el análisis, realizaremos la representación gráfica de los datos. Dado que se evaluaron diferentes variables (concentración de pigmentos) para clasificar las diferentes especies de algas, la representación gráfica de los datos la realizaremos a través de la construcción de un biplot sobre el análisis de componentes principales de un modelo matricial sin intercepto (centrado en el origen de coordenadas), con las etiquetas de las especies de algas dadas en la matriz de datos, en vez de las asignadas arbitrariamente por el algoritmo de formación de cluster. Las órdenes de programación de R para conseguir lo anterior se muestran a continuación y el gráfico resultante se muestra en la figura 10.11.

```
> biplot(princomp(model.matrix(~-
1+Chc2+Chc1+Mca1+Prdn+Fcxn+X9cn+
Vlxn+Ddnx+Dnxn+Allx+Zxnt+Lutn+Chb+Chbe+Chaa+Chae+ßucr+ßecr+ßßcr,
data=Algas)),xlabs=Algas$Especies)
```

En el gráfico se logran observar que las diferentes clases de algas se lograr agrupar perfectamente, y tiene la característica añadida respecto a la construcción de un dendograma, que permite mostrar cuales son las variables que tiene mayor importancia o contribución en la formación de los cluster, en este caso Piridina (Prdn), Clorofila c_2 (Chc2), Luteína (Lutn) y Fucoxantína (Fcxn) (Guisande & Vaamonde, 2012).

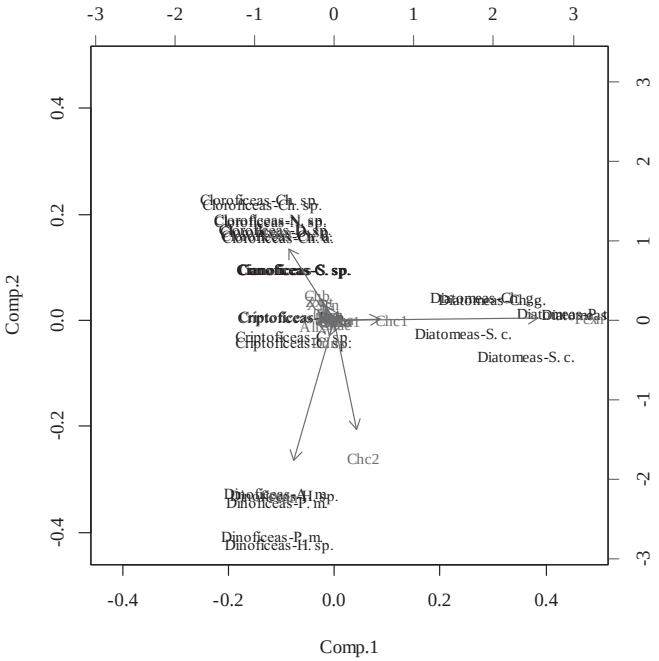


Figura 10.11. Cluster no jerárquico Ejemplo 10.6.

10.5.4. Análisis discriminante

En ocasiones sin importar la disciplina de estudio, surge la necesidad de identificar las características que permitan diferenciar a dos o más grupos, generalmente con el objetivo que al obtener nuevos individuos o casos, poderlos clasificar como perteneciente a cualquiera de los grupos existentes. La técnica de análisis multivariado de datos que cumple con este propósito es el **análisis discriminante (AD)**, cuyo planteamiento estadístico supone la disponibilidad de un conjunto amplio de elementos que pueden venir de dos o más poblaciones distintas (grupos), y en cada elemento se ha observado o medido *p* variables continuas cuya distribución se conoce en las poblaciones consideradas, buscando establecer una *regla discriminante* que permita asignar un nuevo individuo u objeto a una de las poblaciones, clases o grupos previamente identificados (Peña, 2002; Pozo & Carrasco, 2005). Por ejemplo, la primera aplicación del análisis discriminante consistió en clasificar los restos de un cráneo descubierto en una excavación como humano, utilizando la distribución de medidas físicas para los cráneos humanos y los de antropoides (Peña, 2002).

Otro punto de vista, considera al AD como un análisis de regresión en el que a partir de *n* individuos se miden *p* variables cuantitativas independientes o "*explicativas*", como perfil de cada uno de ellos y una variable cualitativa adicional, dependiente o "*clasificativa*", con dos o más categorías que define el grupo de pertenecia de cada uno de los individuos (Pérez, 2004). A partir de las variables medidas, se obtiene una modelo matemáticos, denominado *funciones discriminantes*, que permitirá establecer en que grupo se debe asignar cada individuo, con la máxima probabilidad (Guisande *et al.*, 2011; Pérez, 2004).

Las funciones discriminantes se construyen como combinaciones lineales de las variables originales, y para una mejor asignación de los individuos en cada grupo, minimiza la variabilidad dentro cada uno de los grupos con la consecuente maximización de variabilidad entre grupos. Así, ante la existencia de nuevos elementos, estos se asignan a los grupos existentes tras contrastar el modelo matemático discriminante con el perfil de los nuevos individuos, o medidas de las mismas *p* variables utilizadas para la construcción del modelo. Esta asignación se realiza a través del criterio de probabilidad de Bayes: cada elemento se asigna a la clase para la que es mayor la probabilidad de pertenencia condicionada por los valores que toman las funciones discriminantes (Guisande *et al.*, 2011; Pérez, 2004).

En R, es bastante fácil e intuitiva la aplicación de AD a través de la función ***lda*** (análisis discriminante lineal por su siglas en inglés) del paquete "*MASS*" (Ripley *et al.*, 2015), cuya estructura de programación se muestra a continuación

```
lda(x, grouping, prior = proportions, CV = FALSE)
```

Donde ***x*** es una matriz o dataframe que contiene los valores de las variables explicativas (perfil de los individuos), ***grouping*** es un factor que especifica las clases o grupos de cada observación, ***prior*** son las probabilidades a priori de pertenencia de cada individuo a una clase, generalmente se fijan iguales para todos los grupos, y de no especificarse el programa entenderá que deseamos que las probabilidades a priori se determinen por el número de elementos que tiene clase en la muestra. Por último el argumento ***CV*** asume valores lógicos para indicarle al software la validación de las funciones discriminantes (como se verá más adelante).

Es importante tener en cuenta en la ejecución del AD, que al ser un modelo estadístico formal, como los modelos de regresión y análisis de varianza, se requiere que el mismo dé cumplimiento a ciertos supuestos o hipótesis.

Específicamente el cumplimiento de la hipótesis de homogeneidad de varianzas, es decir, que las matrices de covarianzas de cada uno de los grupos sean iguales (no difieran significativamente), o que la dispersión en cada clase sea similar para todas las variables medidas; el test estadístico utilizado para evaluar el cumplimiento de este supuesto es el test M de Box (Pérez, 2004; Guisande *et al.*, 2011). Este, se simula en el entorno de programación de R a través de la función ***boxM*** del paquete "*biotools*" (da Silva, 2015), siguiendo la siguiente línea de programación

```
boxM(data, grouping)
```

Donde el argumento ***data*** corresponde a una matriz de datos o dataframe que contiene las mediciones de *p* variables en las *n* observaciones y ***grouping***, define un vector de longitud *n* que contiene las clases a la que pertenece cada observación.

Otro supuesto de importancia en la ejecución del AD, es que las variables medidas deben seguir una distribución normal multivariante en los diferentes grupos existentes. No obstante, el método es robusto a un ligero incumplimiento de este supuesto y funciona razonablemente bien (Calderón, 2008; Pérez, 2004; Guisande *et al.*, 2011). En la práctica se suele verificar este supuesto variable a variable, dada la complejidad de cálculo para realizarlo tomando todas las variables simultáneamente (Pérez, 2004) y la no incorporación de este procedimiento en muchos paquetes estadísticos. Sin embargo, R posee un procedimiento para la ejecución del test de Shapiro-Wilk para normalidad multivariada, a través de la función ***mshapiro.test*** del paquete "*mvnormtest*" (Jarek, 2015), cuya línea de código se muestra a continuación

```
mshapiro.test(U)
```

Con ***U*** definiendo una matriz de datos de valores numéricos. Para la aplicación del test en los diferentes grupos o clases de be usarse en asocio a la función ***by*** del paquete de instalación básica de R, que ha sido utilizado en capítulos anteriores.

Hasta el momento, se han descrito los procedimientos analíticos para la construcción del modelo discriminante y la verificación de los supuestos del mismo. No obstante, un aspecto de suma importancia en la utilización del AD para el tratamiento de nuestros datos es la validación del modelo, dado que las funciones discriminantes se construyen con datos proporcionados por una muestra aleatoria, estas clasifican con un elevado

grado de acierto a los elementos contenidos en dicha muestra. Sin embargo, el objetivo es muchas ocasiones es establecer una regla discriminante para la clasificación de individuos nuevos, por lo que la utilización de las funciones discriminantes para la clasificación de nuevas observaciones disminuye notablemente el grado de acierto de las mismas y ello hace necesario la utilización de algún procedimiento de validación que permita conocer la eficacia en la clasificación de nuevos individuos no solo de la muestra, sino a nivel poblacional.

El método generalmente utilizado para la validación de las funciones discriminantes es la llamada clasificación cruzada dejando uno fuera (*leave one out cross validation*), cuyo procedimiento de análisis, según Guisande *et al.*, (2011) consiste en realizar un análisis discriminante excluyendo el primer elemento de la muestra, y a continuación ese elemento es asignado mediante las funciones discriminantes obtenidas. Posteriormente se realiza el mismo procedimiento con cada una de las observaciones de la muestra. De este modo se asignan todos los elementos de la muestra sin que el elemento correspondiente hay sido utilizado en los callos de las funciones discriminantes y, por lo tanto, hemos utilizado todos los datos como una muestra de validación. Si el porcentaje de acierto con esta validación cruzada es alto y similar, o solo ligeramente inferior, al porcentaje de acierto del análisis con toda la muestra, la validación es positiva, y el análisis discriminante es adecuado para esos datos. En R, la aplicación de clasificación cruzada se realiza simplemente dado el valor lógico *TRUE* al argumento ***CV*** de la función ***lda***.

Otra forma de validar las funciones discriminantes que suele utilizarse es reservar una parte de la muestra como muestra de validación pero de la cual se conoce previamente a que grupo pertenece, es decir, no se considera en la construcción de las funciones discriminantes. Posteriormente, se utiliza como nuevos individuos que se desean clasificar en los grupos existentes a través de las funciones discriminantes obtenidas. Si resulta que la proporción de observaciones que son clasificados correctamente con las funciones discriminantes es elevada, se establece como adecuado el modelo planteado. En el ejemplo de aplicación del método se ilustrará las dos metodologías de validación descritas.

Ejemplo 10.7. A continuación se muestran las mediciones de diferentes variables fisicoquímicas como oxígeno disuelto (mg/L), conductividad (μS/cm), pH, amonio, nitritos, nitratos y fosfatos (μg/L) y solidos suspendidos totales (mg/L) en cuatro diferentes cuerpos de agua: laguna

costera, estuario, mar y pozo. A través de estas, de desea determinar si es posible discriminar o clasificar a los diferentes cuerpos de agua.

Ecosistema	OD	Cond	pH	NH4	NO_2	NO_3	PO_4	SST
Laguna	5.2	14800	8.0	0.200	0.400	5.000	0.800	785
Laguna	5.6	36300	7.9	0.500	0.500	10.000	0.100	635
Laguna	5.4	52300	7.9	0.500	0.400	25.000	0.050	784
Laguna	5.2	36300	8.5	1.000	0.010	25.000	4.100	559
Laguna	5.1	32900	8.1	1.000	0.010	25.000	4.700	476
Laguna	4.4	53600	8.2	1.000	0.010	23.000	5.600	386
Laguna	5.3	35500	8.0	0.500	0.300	23.000	0.900	580
Laguna	4.8	25600	8.3	0.300	0.400	15.000	0.600	645
Laguna	5.2	21000	8.1	1.000	0.010	25.000	3.500	395
Laguna	5.5	25800	7.9	0.400	0.010	25.000	4.700	495
Estuario	8.8	243	7.3	0.390	0.430	0.700	2.430	220
Estuario	5.0	451	8.2	0.280	0.300	0.900	0.870	320
Estuario	7.0	605	8.0	0.440	0.260	0.280	2.310	402
Estuario	5.8	609	8.3	0.150	0.200	0.660	1.120	398
Estuario	5.8	594	8.4	0.060	0.050	0.530	0.420	396
Estuario	5.8	580	6.7	0.150	0.130	0.200	1.040	390
Estuario	4.7	370	8.5	0.060	0.250	0.530	1.490	370
Estuario	4.4	330	8.2	0.030	0.380	0.340	1.900	268
Estuario	4.2	287	8.0	0.010	0.520	0.160	2.480	242
Estuario	5.2	374	7.9	0.400	0.580	0.380	9.100	280
Mar	6.0	52500	8.1	0.027	0.009	0.047	0.040	316
Mar	6.6	54400	8.1	0.057	0.003	0.006	0.020	345
Mar	6.8	50200	8.0	0.002	0.000	0.005	0.020	303
Mar	7.3	54600	8.1	0.034	0.006	0.008	0.000	314
Mar	7.5	53400	8.1	0.104	0.002	0.014	0.010	322
Mar	6.7	42500	8.1	0.085	0.135	0.155	0.070	322
Mar	6.9	54500	8.0	0.054	0.011	0.001	0.050	323
Mar	8.2	55500	8.1	0.103	0.002	0.002	0.010	332
Mar	7.9	55500	8.2	0.029	0.001	0.002	0.080	540
Mar	7.6	52800	8.2	0.021	0.001	0.008	0.010	506

Pozo	3.5	143	7.3	0.002	0.000	0.087	0.014	54
Pozo	3.5	451	6.9	0.002	0.004	0.051	0.000	50
Pozo	3.6	305	7.0	0.001	0.004	0.034	0.013	65
Pozo	3.6	409	6.9	0.002	0.003	0.049	0.122	72
Pozo	5.4	494	7.0	0.002	0.004	0.011	0.047	62
Pozo	3.1	380	6.9	0.002	0.007	0.016	0.037	50
Pozo	5.9	370	8.5	0.001	0.002	0.010	0.017	40
Pozo	5.8	330	8.7	0.002	0.001	0.011	0.000	73
Pozo	5.7	287	8.0	0.002	0.001	0.013	0.015	72
Pozo	5.4	374	8.1	0.002	0.004	0.002	0.050	73

Solución

Iniciamos el análisis de los datos con la acostumbrada tabulación de los mismos en la hoja de cálculo de Excel, guardado bajo la extensión *.csv* e importación desde el ambiente de programación de R

```
>Datos<-read.csv2("Ejemplo D.csv",header=TRUE,encoding="latin1")
```

Continuamos con la verificación de los supuestos del modelo discriminante lineal, iniciando con el supuesto de normalidad. Para ello, realizaremos indexaciones en nuestra matriz de datos para extraer subgrupos que contengan los valores de las variables fisicoquímicas para cada uno de los sistemas estudiados y luego se aplica el test de Shapiro-Wilk multivariante a cada subconjunto

```
> library(mvnormtest)
> Laguna<-Datos[Datos$Ecosistema=="Laguna",]
> Estuario<-Datos[Datos$Ecosistema=="Estuario",]
> Mar<-Datos[Datos$Ecosistema=="Mar",]
> Pozo<-Datos[Datos$Ecosistema=="Pozo",]
> mshapiro.test(t(Laguna[,-1]))

        Shapiro-Wilk normality test

data:  Z
W = 0.40107, p-value = 2.566e-07

> mshapiro.test(t(Estuario[,-1]))

        Shapiro-Wilk normality test
```

```
data:  Z
W = 0.36662, p-value = 1.028e-07

> mshapiro.test(t(Mar[,-1]))

        Shapiro-Wilk normality test

data:  Z
W = 0.37654, p-value = 1.337e-07

> mshapiro.test(t(Pozo[,-1]))

        Shapiro-Wilk normality test

data:  Z
W = 0.39113, p-value = 1.97e-07
```

Antes de comentar las salidas de resultado para este test, evaluaremos el supuesto de homogeneidad de las matrices de covarianza entre los grupos a través del test M de box, para luego realizar los comentarios a que haya lugar.

```
> library(biotools)
> boxM(Datos[,2:9],Datos[,1])

        Box's M-test for Homogeneity of Covariance Matrices

data:  Datos[, 2:9]
Chi-Sq (approx.) = 619.16, df = 108, p-value < 2.2e-16
```

En general, la inspección de la salida de resultados (p-valores) para los test de normalidad multivariada de los datos y homogeneidad de varianzas, muestra claramente que estos supuestos son transgredidos fuertemente, por lo tanto, no es aconsejable continuar con el análisis a menos que se realice un tratamiento especial de los datos (transformaciones) que permitan cambiar la escala de los mismos y corregir esta violación de los supuestos, es decir, volver normales los datos y conseguir que las matrices de covarianza entre los grupos sean homogéneas. Sin embargo, para fines práctico de este libro y no extendernos en procedimientos que ya han sido discutidos en capítulos anteriores, continuaremos el análisis con los datos originales. Además, Calderón (2008), afirma haber obtenido buenos porcentajes de discriminación acertada a través de la aplicación del análisis discriminante lineal, bajo transgresiones de los supuestos del modelo.

A continuación se muestra la salida de resultados para la aplicación del AD a los dados crudos (originales)

```
> AD <- lda(Datos[,2:9], grouping= Datos[,1], prior =
c(1,1,1,1)/4)
> AD
Call:
lda(Datos[, 2:9], grouping = Datos[, 1], prior = c(1, 1, 1,
1)/4)
Prior probabilities of groups:
Estuario   Laguna      Mar     Pozo
    0.25     0.25     0.25     0.25
Group means:
           OD    Cond   pH    NH4   NO2     NO3    PO4   SST
Estuario 5.67   444.3 7.95 0.1970 0.310  0.4680 2.3160 328.6
Laguna   5.17 33410.0 8.09 0.6400 0.205 20.1000 2.5050 574.0
Mar      7.15 52590.0 8.10 0.0516 0.017  0.0248 0.0310 362.3
Pozo     4.55   354.3 7.53 0.0018 0.003  0.0284 0.0315  61.1

Coefficients of linear discriminants:
               LD1           LD2           LD3
OD    0.3441942739 -0.1570674918 -3.251132e-01
Cond  0.0001036604 -0.0001409973  1.574934e-05
pH   -0.1679255216 -0.1115849951 -3.243425e-01
NH4  -4.2566664128  0.6095390661 -1.025092e+00
NO2  -5.3890860706  1.0283909050 -1.718443e+00
NO3  -0.2922174777  0.0171427764  1.755565e-01
PO4   0.0624825903 -0.1633013578 -4.004528e-01
SST  -0.0103992922 -0.0051745918 -4.454674e-03

Proportion of trace:
   LD1    LD2    LD3
0.6402 0.3259 0.0339
```

Inicialmente se observan las probabilidades a priori de cada uno de los grupos, fijadas iguales para cada uno de ellos (25%), luego se observar las medias o centroides de las variables observadas en cada uno de los grupos. Así mismo, se obtienen los coeficientes de las funciones discriminantes y la proporción o porcentaje de la varianza total que es explicado por las funciones discriminantes. Nótese que se obtuvieron tres funciones discriminantes, es decir, el número de grupos menos uno (g - 1). En general esto siempre se cumple, a menos que el número de variables explicativas sea menor que el número de grupos existentes, dicho de otro modo, el número de funciones discriminantes siempre será el valor mínimo entre el número de grupos menos uno y el número de variables explicativas

(min(g – 1, p)) (Pérez, 2004; Celedón, 2008; Guisande *et al.*, 2011). En cuanto a la proporción de variabilidad explicada, se observa que la primera y segunda función discriminantes, explican un 64.02 y 32.59 %, respectivamente, de la variabilidad total de los datos, es decir, en conjunto recogen un 96.61 % de la variabilidad total y un gráfico bidimensional de las mismas es suficiente para representar correctamente los datos.
Ahora, continuaremos con el análisis haciendo la validación de nuestro modelo discriminante a través del comentado método de validación cruzada dejando uno fuera, como se puede observar en la siguiente salida de resultados

```
> AD <- lda(Datos[,2:9], grouping= Datos[,1], prior =
c(1,1,1,1)/4, CV= TRUE)
```

Enseguida, construimos una tabla cruzada de las clasificaciones realizadas por el modelo discriminante y las clases dadas en la muestra, para evaluar la eficacia en la clasificación de nuestro modelo comparando las clasificaciones realizadas correctamente (situadas en la diagonal principal) con las clasificaciones incorrectas (situadas fuera de ella).

```
> T<- table(AD$class,Datos[,1])
> T

           Estuario Laguna Mar Pozo
  Estuario       10      1   0    0
  Laguna          0      9   0    0
  Mar             0      0  10    0
  Pozo            0      0   0   10
```

Podemos expresar esta tabla de forma porcentual para sacarle mayor provecho e interpretabilidad a la misma, solo basta con ceñirse a la línea de programación que se muestra en la siguiente salida de resultados

```
> prop.table(T,2)*100

           Estuario Laguna Mar Pozo
  Estuario      100     10   0    0
  Laguna          0     90   0    0
  Mar             0      0 100    0
  Pozo            0      0   0  100
```

Nótese que todas las clases fueron clasificadas correctamente, a excepción de las muestras de agua provenientes de lagunas costeras, donde el 10% de las mismas (1 caso) fue asignado incorrectamente a estuarios. No obstante, el porcentaje de acierto para esta clase es significativamente elevado.

Un resultado interesante del AD es el porcentaje global de acierto del modelo discriminante, calculado a través de la línea de programación que se muestra a continuación (Guisande *et al.*, 2011)

```
> Acierto=100-(sum(AD$class!=Datos[,1])/sum(T))*100
> Acierto
[1] 97.5
```

Se observa que nuestras funciones discriminantes tiene un porcentaje global de aciertos o clasificación correcta del 97.5%, porcentaje bastante elevado en la práctica, aun cuando se ha incurrido en transgresiones de los supuestos del modelo.

Para dar por terminado el análisis, construiremos el gráfico de las dos primeras dimensiones (Figura 10.12) haciendo uso de la función ***candisc*** del paquete "*candisc*" (Friendly & Fox, 2015). Las órdenes de programación respectivas para la ejecución de esta función se muestran en la siguiente salida de resultados

```
>    mod<-lm(cbind(OD,Cond,pH,NH4,NO3,NO2,PO4,SST)~Ecosistema,
data=Datos)
> can<-candisc(mod,term="Ecosistema")
> plot(can, which= c(1,2), conf= 0.95, pch= c(15,16,17,18),
col= 1:4, var.col="purple", var.lwd=1, prefix= "Función ",
suffix=TRUE)
```

Nótese que es necesario la creación de un modelo lineal multivariante, a través de la función ***lm*** donde se consideren las variables explicativas medidas y la variable clasificativa que contiene los diferentes grupos o clases. Luego, se inserta el modelo lineal en la función ***candisc*** y se agrega el argumento ***term*** que especifica la variables de agrupación de nuestros datos. Finalmente, se construye el gráfico ordenando al programa que tome solo las dos primeras dimensiones haciendo uso del argumento ***which***, el trazado de circunferencias de confiabilidad del 95% con ***conf***, que definen la probabilidad de un elemento de esa clase este dentro del circulo; las circunferencias sin solapamiento indican clases bien definidas o

identificables (Guisande & Vaamonde, 2012). Así mismo, se especifican los marcadores y color de los mismos (***pch*** y ***col***), el color de las etiquetas de las variables explicativas y el grosor de los vectores asociados a las mismas con los argumentos **var.col** y **var.lwd**. Los argumentos ***prefix*** y ***suffix***, especifican, respectivamente, el prefijo de las etiquetas de los ejes del gráfico y la representación del porcentaje de variabilidad (inercia) explicado por cada dimensión. Por último se inserta una legenda que representa la correspondencia de cada uno de los puntos sobre el gráfico para cada uno de los grupos.

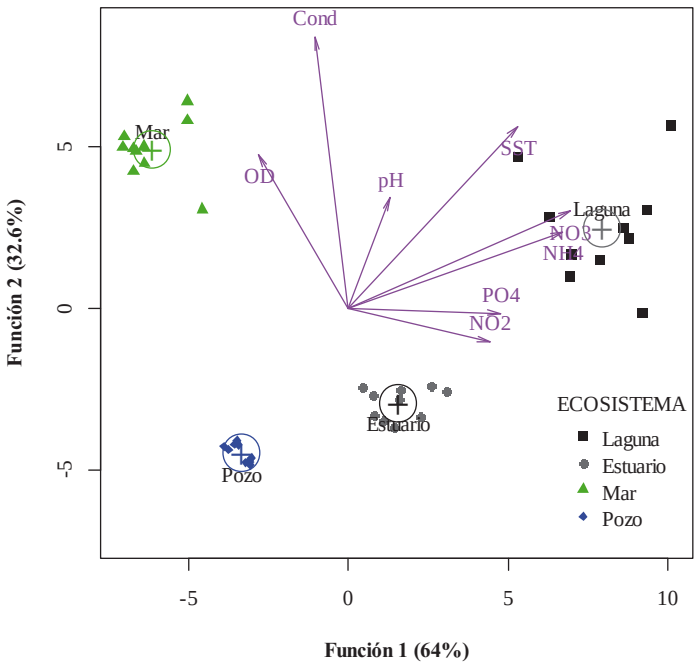


Figura 10.12. Representación de funciones discriminantes en el plano.

Obsérvese que el gráfico resultante muestra que todos los ecosistemas estudiados están bien definidos y diferenciados al no encontrarse solapamiento, ni alguna cercanía de los puntos sobre el plano y de sus circunferencias de confiabilidad.

Referencias bibliográficas

Aitchison, J. & Silvey, S. 1960. Maximum-Likelihood Estiamtion of Parameters Subject to Restraints. *Annals of Mathematical Statistics*. 29: 813-828.

Arrianza, A., Fernández, F., López, M., Muñoz, M., Pérez, S. & Sánchez, A. 2008. Estadística Básica con R Y R-Commander. Universidad de Cadíz. Servicio de Publicaciones. p. 160.

Bartlett, M. 1937. Propieties of Sufficiency and Statistical Test. *Proceedings of the Royal Society of London. Series A, Mathemattical and Physical Sciences*. 160(901): 268-282.

Box, G. & Cox, D. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*. 26(2): 211-252.

Box, G., Hunter, S. & Hunter, W. 2008. Estadística para Investigadores: Diseño, Innovación y Descubrimiento. Segunda Edición. Editorial Reverté. Barcelona España. p. 639.ISBN 978-958-648-766-5.

Breusch, T. & Pagam, A. 1979. A Simple Test for Heterocedasticity and Random Coerfficient Variation. *Econometrica*. 47(5).

Brown, M. & Forsythe, A. 1974. Robust Test for Equality of Variances. *Jorunal of the American Satatistical Association.* 69(346): 364-367.

Calderón, I. 2008. Detección del Perfil del Infractor Tributario en el SENIAT. Tesis de Pregrado. Universidad de los Andes. Mérida, Zulia.

Cardoso, G. & Veitía, N. 2008. Aplicación de Métodos de Comparaciones Múltiples en Biotecnología Vegetal. *Biotectnología Vegetal*. 8(2): 67-71. ISBN 1609-1841.

Cochran, W. 1941. The Distributions of the largest of a set of estimated variances as a fraction of their total. *Annals of Eugenics*. 11: 47-52.

Conavos, G. 1988. Probabilidad y Esstadística. Aplicaciones y Métodos. Editorial McGraw-Hill. México. p. 667. ISBN 968-451-856-0.

Conover, W., Johnson, E. & Johnson, M. 1981. A comparative study of tests for homogeneity of variances with applications to the outer continental shelf bidding data. *Technometrics.* 23(4): 351-361.

Córdova, V. & Cortés, A. 2010. Probabilidad y Estadística I. Colegio de bachilleres del estado de Sonora. México. p. 134.

Da Silva, A. 2015. biootools: Tools for Biometry and Applied in Agricultural Science. R Package version 2.2.

De la Huerta, V. 2012. Pruab de Homogeneidad de varianzas para muestras normales cesuradas. Tesis de maestría. Instituto de Enseñanzas e Investigación en Ciencias Agricolas. Campus Montecillo. Postgrado de Socioeconomía, Estadística e Informática Estadística. Texcoco, México. p. 78.

Devore, J. 2008. Probabilidad y Estadística para Ingeniería y Ciencias. Séptima Edición. Editorial Cengage Learning Editores S.A. México. p.723. ISBN 978-970-686-831-2.

Di Rienzo, J., Casanoves, F., González, L., Tablada, E., DÍAZ, M., Robledo, C. & Balzarini, M. 2005. Estadística para las Ciencias Agripecuarias. Sexta Edición. p. 329.

Duncan, C. 1955. Multiple Range and Multiple F Test. *Biometrics*. 11: 1-42.

Duncan, D. 1955. Multiple Range and Multiple *F* Test. *Biometrics*. 11(4): 1-42.

Erickson, F. & Nosanchuck, T. 1977. Understandig Data. Editorial McGraw-Hill. Toronto.

Febrero, M., San, P., González, J. & Pateiro, B. 2008. Prácticas de Estadística en R. Ingeniería Técnica en Informática de Sistemas. Dpartamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela. Coruña, España. p. 114. ISBN 978-84-691-0975-1.

Ferrer, O. 2007. Estadística Aplicada. Universidad del Zulia. Departamento de Bilogía. p. 250.

Fine, J. 1996. Iniciación a los analisis de datos multidimensionales a partir de ejemplos. Notas de clase. Montevideo.

Fox, J., Bouchet, M., Andronic, L., Ash, M., Boye, T., calca, S., Chang, A., Grosjean, P., Heiberger, P., Pour, K., Kerns, G., Lancelot, R., Lesnoff, M., Ligges, U., Messad, S., Maechler, M., Muenchen, R., Murdoch, D., Neuwith, E., Putler, D., Ripley, B., Ristic, M., Wolf, P. & Wrigth, K. 2015. Rcmdr: R Commander. R Package version 2.2.1.

Fox, J., Muecvhen, R. & Putler, D. 2015. RcmdrMisc: R Commander Miscellaneous Funtions. R Package version 1.0.3.

Fox, J., Weisberg, S., Adler, D., Baetes, D., Baud-Bovy, G., Ellison, E., Fierth, D., Friendly, M., Gorjanc, G., Graves, S., Heisberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. Zeileis, A. & R-Core Team. 2015. car: Companion to Applied regression. R Package version 2.1.0.

Friendly, M. & Fox, B. 2015. candisc: Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R Package version 0.6.7.

Gaete, A. 1978. Teoria y Aplicación de la Transformación de variables en Geografía. Universidad Católica de Chile. Instituto de Geografía. Santiago, Chile.

Gastwirth, J., Gel, Y., Wallace, W., Lyubchich, V., Miao, W. & Noguchi, K. 2015. lawstat: Tools for Biostatistics, Public Policy, and Law. R Package version 3.0.

Goslee, S. & Urban, D. 2015. ecodist: Dissimilarity-Based Functions for Ecological Analysis. R Package version 1.2.9.

Greeenacre, M., Nenadic, O. & Friendly, M. 2015. ca: Simple, Multiple and Joint Correspondence Analysis. R Package version 0.58.

Greenacre, M. 2008. La Prática del Análisis de Correspondencias. Fundación BBVA. Bilbao, España. ISBN 978-84-96515-71-0.

Gross, J. 2015. nortest: Test for Normality. R Package version 1.0.4.

Guajarati, D. & Porter, D. 2010. Econometría. Quinta Edición. Editorial McGraw-Hill. México. ISBN 978-607-15-0294-0.

Guisande, C. & Vaamonde, A. 2012. Gráficos Estadísticos y Mapas con R. Editorial Díaz de Santos. España. p. 367. ISBN 978-84-9969-211-1.

Guisande, C., Vaamonde, A. & Barreiro, A. 2011. Tratamiento de Datos von R, Statistica y SPSS. Ediciones Díaz de Santos. España. p.978. ISBN 978-84-7978-998-5.

Gutiérrez, H. & de la Vara, R. 2008. Análisis y diseño de experimentos. Segunda Edición. Editorial McGraw-Hill. México. p. 546. ISBN 978-970-10-6526-6.

Hartigan, J. & Wong, M. 1979. "Algorithm AS 136: A K-means Clustering Algorithm". *Applied Statistics.* 28(1): 100–108.

Hasler, M. 2015. SimComp: Simultaneus Comparasons for Multiple Endpoints. R package version 2.2.

Horthon, T., Zeileis, A., Farebrother, R., Cummins, C., Millo, G. & Mitchell, D. 2015. lmtest: Testing Linear Tegression Models. R Package version 0.9.34.

Husson, F., Josse, J., Le, Sebastien. & Mazet, Jeremy. 2015. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R Package version 1.31.3.

Husson, F., Lê. & Pagès. 2013. Análisis de Datos con R. Editorial Escuela Colombiana de Ingeniería. Colombia. p. 314. ISBN 978-058-8726-05-2.

Hyndman, R. & Fan, Y. 1996. Sample Quantiles in Statistical Packages. *Journal of the American Statisticia Associationl.* 50(4): 361-365.

Instituto Nacional de Estadística e Informática, INEI. 2009. Guía para la Presentación de Gráficos Estadísticos. Lima, Perú. p. 59.

Jarek, S. 2015. mvnormtest: Normality Test for Multivariate Variables. R Package version 0.1.9.

Johnson, D. 2000. Métodos Multivariados Aplicados al Análisis de Datos. Editorial International Thomson Editores. p. 566. ISBN 968-7529-90-3.

Keuls, M. 1952. The Use of the Studentized in Connection with an Analisys of Variance. *Euphytica.* 1: 122-122.

Kmenta, J. 1986. Elementos de Econometría. Universidad de Vicens.

Kuehl, R. 2001. Diseño de Experimentos. Segunda Edición. Editorial Thomson Learning. México. p. 678. ISBN 970.686-048-7.

Lagrande, C. & Pimzón, L. 2009. Análisis de Datos. Métodos y Ejemplos. Editorial Escuela Colombiana de Ingeniría. Colombia. p. 388. ISBN 978-958-8060-90-3.

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson B., Wickham, H., Tyagi, A., Eterradossi, O., Grothendieck, G., Toews, M., Kane, J., Turner, R., Witthoft, C., Stander, J., Petzoldt, T., Duurma, R., Biancotto, E., Levy, O., Dutang, C., Solymos, P., Engelmann, R., Hecker, M., Steinbeck, F., Borchers, H., Singmann, H., Toal, T. & Ogle, D. 2015. plotrix: various Plotting Functions. R Packages version 3.5.12.

Levene, H. 1960. Robust Test for Equality of Variances. in I. Olkin, Ed. *Cobtributions to Probability and Statistics*. p. 278-292.

Lilliefors, H. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknow. *Journal of de American Statistical Association*. 62(318): 339-402.

Lloyd, S. 1982. Least Squares Quantization in PCM. *Ieee transactions on information theory*. 28: 129–137. Originally as an Unpublished Bell Laboratorios Technical Note (1957).

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M. & Roudier, P. 2015. cluster: "Finding Groups in Data": Cluster Analysis Extended Rousseauw et al. R Package version 2.0.3.

Martínez, C. 2011. Estadística Básica Aplicada. Cuarta Edición. Editorial Ecor Ediciones. p. 388. ISBN 978-958-648-766-5.

McQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Fifth Berkeley Symposium on Mathematics. Statistics and Probability*. University of California Press. p. 281–297.

Mendiburu, F. 2015. agricolae: Satatistical Procedures for Agricultural Research. R Package version 1.2-2.

Milton, S. 2004. Estadística para Biología y Ciencias de la Salud. Tercera Edición. Editorial McGraw-Hill. Madrid, España. p. 592. ISBN 84-486-0321-4.

Monteiro, L. & Gomes, J. 2006. Introdução ã Biometria Utilizando R. Laboratório de Ciéncias Ambientais, CBB. Universidade Estadual do Norte Fluminense.

Montgomery, D. 2001. Diseño y Análisis de Experimentos. Segunda Edición. Editorial Limusa Wiley. México. p. 668. ISBN 968-18-6156-6.

Morales, P. 2008. Estadística Aplicada a las Ciencias Sociales. Universidad Pontifica Comillas. Madrid, España.

Morales, P. 2011. Análisis de Varianza para Varias Muestras Independientes. Universidad Pontifica Comillas. Facultad de Ciencias Hymanas y Sociales. Madrid, España.

Myer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. & Lin, C. 2015. e1071: Misc Functions of the Departament of Statitical, Probability Theory Group (Formerly: E1071), TU Wien. R Package version 1.6.7.

Nakazawa, M. 2015. fmsb: Functions for Medical Statistics Book with Some Demographic Data. R Package version 0.5.2.

Newman, D. 1939. The Distribution of the Range in Samples from a Normal Population, Expressed in Terms of an Independient Estimate of Standard Deviation. *Biometrika*. 30: 20-30.

Orlandoni, G. 2010. Escalas de medición en Estadística. *Red de Revistas Científicas de América Latina, el Caribe, España y Portugal*. Redalyc (URBE). 12(2): 243-347. ISBN 1317-0570.

Peña, D. 2002. Análisis de Datos Multivariantes. Editorial McGraw-Hill. México. ISBN 8448136101.

Pérez, C. 2004. Técnicas de Análisis Multivariante de Datos. Aplicaciones con SPSS. Editorial Pearson Educaciín S.A. Madrid, España. p.672. ISBN 978-84-205-4104-4.

Pozo, M. & carrasco, G. 2005. Aplicación del Análisis Discriminante a un Conjunto de Datos Vinícolas Mediante el Paquete Estadístico SPSS v10. *Tecnociencia*. 7(1): 7-21.

Quevedo, H. 2006. Métodos Estadísticos para la Ingenieria Ambiental y las Ciencias. Universidad Autónoma de Ciudad Juaréz. Instituto de Ingeniería y Tecnología. Departamento de Ingeniería Cívil y Ambiental. p.848.

Revelle, W. 2015. psych: Procedures for Psychological, Psychometric, and Personality Research. R Package version 1.5.8.

Ripley, B. & Venalbles, W. 2015. class: Function for Classification. R Package version 7.3.14.

Rodríguez, L. 2007. Probabilidad y Estadística Básica para Ingenieros: con el soporte de MATLAB para cálculos y gráficos cálculos estadísticos. Instituto de Ciencias Matemáticas. Escuela Superior Politécnica del Litoral, ESPOL. Guayaquil, Ecuador. p. 311.

Rosado, J. 2009. Laguna Salá: su biología y ambiente. Editorial Gente Nueva. Universidad de La Guajira. p. 190. ISBN 978-958-8530-08-6.

Saénz, A. 2012. Apuntes de Estadística para Ingenieros. Departamento de Estadística e Investigación Operativa. Universidad de Jaén. Andalucía, España. p.235.

Sandrini, L. & Camargo, M. 2015. GAD: Analysis of Variance from General Principles. R Package version 1.1.1.

Sarkar, D. 2015. lattice: Trellis Graphics for R. R Package version 0.20.33.

Scheffé, H. 1953. A Method for Judging All Contrasts in the Analisys of Variance. *Biometrika*. 40: 87-104.

Shapiro, S. & Wilk, M. 1965. An Analysis of Variance Test for normality (Complete Samples). *Biometrika*. 52(3-4): 591-611.

Tukey, J. 1953. The Problem of Multiple Comparisons. Unpublished Manuscript. Princeton University.

Tukey. J. 1977. Exploratory Data Analysis. Reading, Mass: Adisson-Wesley.

Vargas, V. 2007. Estadistica Descriptiva para Ingeniería Ambiental con SPSS. Universidad Nacional de Colombia-Sede Palmira. p. 312. ISBN 978-958-33-9319-3.

Vorapongsathorn, T., Taejarenkul, S. & Viwatwongkasem, C. 2004. A Comparison of Type I Error and Power of Bartlett's test, Levene's Test and Cochran's Test Under Violation of Assumptions. *Songklanakarin J. Sci. Technol.* 26(4): 537-547.

Walpole, R., Myers, R., Myers, S. & Ye, K. 2007. Probabilidad y Estadística para Ingeniería y Ciencias. Octava edición. Editorial Pearson Educación. México. p. 840. ISBN 978-970-26-0936-0.

Wilcoxon, F. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1(6): 80-83.

Yakir, B. 2011. Introduction to Sattistical Thinking (With R, Without Calculus). The Hebrew University. p. 324.

Zimmermann, F. 2004. Estadística para Investigadores. Editorial Escuela Colombiana de Ingeniería. Colombia. p. 423. ISBN 958-8060-X.

Apéndices

Tablas y pruebas estadísticas

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$.

		p														
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
1	0	0.9900	0.9500	0.9000	0.8000	0.7500	0.7000	0.6000	0.5000	0.4000	0.3000	0.2500	0.2000	0.1000	0.0500	0.0100
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	10000	1.0000
2	0	0.9801	0.9025	0.8100	0.6400	0.5625	0.4900	0.3600	0.2500	0.1600	0.0900	0.0625	0.0400	0.0100	0.0025	00001
	1	0.9999	0.9975	0.9900	0.9600	0.9375	0.9100	0.8400	0.7500	0.6400	0.5100	0.4375	0.3600	0.1900	0.0975	0.0199
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	0.9703	0.8574	0.7290	0.5120	0.4219	0.3430	0.2160	0.1250	0.0640	0.0270	0.0156	0.0080	0.0010	0.0001	0.0000
	1	0.9997	0.9928	0.9720	0.8960	0.8438	0.7840	0.6480	0.5000	0.3520	0.2160	0.1562	0.1040	0.0280	0.0073	0.0003
	2	1.0000	0.9999	0.9990	0.9920	0.9844	0.9730	0.9360	0.8750	0.7840	0.6570	0.5781	0.4880	0.2710	0.1426	0.0297
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	0.9606	0.8145	0.6561	0.4096	0.3164	0.2401	0.1296	0.0625	0.0256	0.0081	0.0039	0.0016	0.0001	0.0000	0.0000
	1	0.9994	0.9860	0.9477	0.8192	0.7383	0.6517	0.4752	0.3125	0.1792	0.0837	0.0508	0.0272	0.0037	0.0005	0.0000
	2	1.0000	0.9995	0.9963	0.9728	0.9492	0.9163	0.8208	0.6875	0.5248	0.3483	0.2617	0.1808	0.0523	0.0140	0.0006
	3	1.0000	1.0000	0.9999	0.9984	0.9961	0.9919	0.9744	0.9375	0.8704	0.7599	0.6836	0.5904	0.3439	0.1855	0.0394
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	0.9510	0.7738	0.5905	0.3277	0.2373	0.1681	0.0778	0.0312	0.0102	0.0024	0.0010	0.0003	0.0000	0.0000	0.0000
	1	0.9990	0.9774	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0156	0.0067	0.0005	0.0000	0.0000
	2	1.0000	0.9988	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.1035	0.0579	0.0086	0.0012	0.0000
	3	1.0000	1.0000	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.3672	0.2627	0.0815	0.0226	0.0010
	4	1.0000	1.0000	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.7627	0.6723	0.4095	0.2262	0.0490
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	0.9415	0.7351	0.5314	0.2621	0.1780	0.1176	0.0467	0.0156	0.0041	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000
	1	0.9985	0.9672	0.8857	0.6554	0.5339	0.4202	0.2333	0.1094	0.0410	0.0109	0.0046	0.0016	0.0001	0.0000	0.0000
	2	1.0000	0.9978	0.9842	0.9011	0.8306	0.7443	0.5443	0.3437	0.1792	0.0705	0.0376	0.0170	0.0013	0.0001	0.0000
	3	1.0000	0.9999	0.9987	0.9830	0.9624	0.9295	0.8208	0.6562	0.4557	0.2557	0.1694	0.0989	0.0158	0.0022	0.0000
	4	1.0000	1.0000	0.9999	0.9984	0.9954	0.9891	0.9590	0.8906	0.7667	0.5798	0.4661	0.3446	0.1143	0.0328	0.0015
	5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9959	0.9844	0.9533	0.8824	0.8220	0.7379	0.4686	0.2649	0.0585
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	0.9321	0.6983	0.4783	0.2097	0.1335	0.0824	0.0280	0.0078	0.0016	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
	1	0.9980	0.9556	0.8503	0.5767	0.4449	0.3294	0.1586	0.0625	0.0188	0.0038	0.0013	0.0004	0.0000	0.0000	0.0000
	2	1.0000	0.9962	0.9743	0.8520	0.7564	0.6471	0.4199	0.2266	0.0963	0.0288	0.0129	0.0047	0.0002	0.0000	0.0000
	3	1.0000	0.9998	0.9973	0.9667	0.9294	0.8740	0.7102	0.5000	0.2898	0.1260	0.0706	0.0333	0.0027	00002	0.0000
	4	1.0000	1.0000	0.9998	0.9953	0.9871	0.9712	0.9037	0.7734	0.5801	0.3529	0.2436	0.1480	0.0257	0.0038	0.0000
	5	1.0000	1.0000	1.0000	0.9996	0.9987	0.9962	0.9812	0.9375	0.8414	0.6706	0.5551	0.4233	0.1497	0.0444	0.0020
	6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9984	0.9922	0.9720	0.9176	0.8665	0.7903	0.5217	0.3017	0.0679
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$ (continuación).

		p														
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
8	0	0.9227	0.6634	0.4305	0.1678	0.1001	0.0576	0.0168	0.0039	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9973	0.9428	0.8131	0.5033	0.3671	0.2553	0.1064	0.0352	0.0085	0.0013	0.0004	0.0001	0.0000	0.0000	0.0000
	2	0.9999	0.9942	0.9619	0.7969	0.6785	0.5518	0.3154	0.1445	0.0498	0.0113	0.0042	0.0012	0.0000	0.0000	0.0000
	3	1.0000	0.9996	0.9950	0.9437	0.8862	0.8059	0.5941	0.3633	0.1737	0.0580	0.0273	0.0104	0.0004	0.0000	0.0000
	4	1.0000	1.0000	0.9996	0.9896	0.9727	0.9420	0.8263	0.6367	0.4059	0.1941	0.1138	0.0563	0.0050	0.0004	0.0000
	5	1.0000	1.0000	1.0000	0.9988	0.9958	0.9887	0.9502	0.8555	0.6846	0.4482	0.3215	0.2031	0.0381	0.0058	0.0001
	6	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9915	0.9648	0.8936	0.7447	0.6329	0.4967	0.1869	0.0572	0.0027
	7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9961	0.9832	0.9424	0.8999	0.8322	0.5695	0.3366	0.0773
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	0.9135	0.6302	0.3874	0.1342	0.0751	0.0404	0.0101	0.0020	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9966	0.9288	0.7748	0.4362	0.3003	0.1960	0.0705	0.0195	0.0038	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.9999	0.9916	0.9470	0.7382	0.6007	0.4628	0.2318	0.0898	0.0250	0.0043	0.0013	0.0003	0.0000	0.0000	0.0000
	3	1.0000	0.9994	0.9917	0.9144	0.8343	0.7297	0.4826	0.2539	0.0994	0.0253	0.0100	0.0031	0.0001	0.0000	0.0000
	4	1.0000	1.0000	0.9991	0.9804	0.9511	0.9012	0.7334	0.5000	0.2666	0.0988	0.0489	0.0196	0.0009	0.0000	0.0000
	5	1.0000	1.0000	0.9999	0.9969	0.9900	0.9747	0.9006	0.7461	0.5174	0.2703	0.1657	0.0856	0.0083	0.0006	0.0000
	6	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9750	0.9102	0.7682	0.5372	0.3993	0.2618	0.0530	0.0084	0.0001
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9962	0.9805	0.9295	0.8040	0.6997	0.5638	0.2252	0.0712	0.0034
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9980	0.9899	0.9596	0.9249	0.8658	0.6126	0.3698	0.0865
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	0.9044	0.5987	0.3487	0.1074	0.0563	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9957	0.9139	0.7361	0.3758	0.2440	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9999	0.9885	0.9298	0.6778	0.5256	0.3828	0.1673	0.0547	0.0123	0.0016	0.0004	0.0001	0.0000	0.0000	0.0000
	3	1.0000	0.9990	0.9872	0.8791	0.7759	0.6496	0.3823	0.1719	0.0548	0.0106	0.0035	0.0009	0.0000	0.0000	0.0000
	4	1.0000	0.9999	0.9984	0.9672	0.9219	0.8497	0.6331	0.3770	0.1662	0.0473	0.0197	0.0064	0.0001	0.0000	0.0000
	5	1.0000	1.0000	0.9999	0.9936	0.9803	0.9527	0.8338	0.6230	0.3669	0.1503	0.0781	0.0328	0.0016	0.0001	0.0000
	6	1.0000	1.0000	1.0000	0.9991	0.9965	0.9894	0.9452	0.8281	0.6177	0.3504	0.2241	0.1209	0.0128	0.0010	0.0000
	7	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9877	0.9453	0.8327	0.6172	0.4744	0.3222	0.0702	0.0115	0.0001
	8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.7560	0.6242	0.2639	0.0861	0.0043
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9990	0.9940	0.9718	0.9437	0.8926	0.6513	0.4013	0.0956
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$ (continuación).

		p														
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
11	0	0.8953	0.5688	0.3138	0.0859	0.0422	0.0198	0.0036	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9948	0.8981	0.6974	0.3221	0.1971	0.1130	0.0302	0.0059	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9998	0.9848	0.9104	0.6174	0.4552	0.3127	0.1189	0.0327	0.0059	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9984	0.9815	0.8389	0.7133	0.5696	0.2963	0.1133	0.0293	0.0043	0.0012	0.0002	0.0000	0.0000	0.0000
	4	1.0000	0.9999	0.9972	0.9496	0.8854	0.7897	0.5328	0.2744	0.0994	0.0216	0.0076	0.0020	0.0000	0.0000	0.0000
	5	1.0000	1.0000	0.9997	0.9883	0.9657	0.9218	0.7535	0.5000	0.2465	0.0782	0.0343	0.0117	0.0003	0.0000	0.0000
	6	1.0000	1.0000	1.0000	0.9980	0.9924	0.9784	0.9006	0.7256	0.4672	0.2103	0.1146	0.0504	0.0028	0.0001	0.0000
	7	1.0000	1.0000	1.0000	0.9998	0.9988	0.9957	0.9707	0.8867	0.7037	0.4304	0.2867	0.1611	0.0185	0.0016	0.0000
	8	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9941	0.9673	0.8811	0.6873	0.5448	0.3826	0.0896	0.0152	0.0002
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9993	0.9941	0.9698	0.8870	0.8029	0.6779	0.3026	0.1019	0.0052
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9964	0.9802	0.9578	0.9141	0.6862	0.4312	0.1047
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
12	0	0.8864	0.5404	0.2824	0.0687	0.0317	0.0138	0.0022	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9938	0.8816	0.6590	0.2749	0.1584	0.0850	0.0196	0.0032	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9998	0.9804	0.8891	0.5583	0.3907	0.2528	0.0834	0.0193	0.0028	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9978	0.9744	0.7946	0.6488	0.4925	0.2253	0.0730	0.0153	0.0017	0.0004	0.0001	0.0000	0.0000	0.0000
	4	1.0000	0.9998	0.9957	0.9274	0.8424	0.7237	0.4382	0.1938	0.0573	0.0095	0.0028	0.0006	0.0000	0.0000	0.0000
	5	1.0000	1.0000	0.9995	0.9806	0.9456	0.8822	0.6652	0.3872	0.1582	0.0386	0.0143	0.0039	0.0001	0.0000	0.0000
	6	1.0000	1.0000	0.9999	0.9961	0.9857	0.9614	0.8418	0.6128	0.3348	0.1178	0.0544	0.0194	0.0005	0.0000	0.0000
	7	1.0000	1.0000	1.0000	0.9994	0.9972	0.9905	0.9427	0.8062	0.5618	0.2763	0.1576	0.0726	0.0043	0.0002	0.0000
	8	1.0000	1.0000	1.0000	0.9999	0.9996	0.9983	0.9847	0.9270	0.7747	0.5075	0.3512	0.2054	0.0256	0.0022	0.0000
	9	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9972	0.9807	0.9166	0.7472	0.6093	0.4417	0.1109	0.0196	0.0002
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9968	0.9804	0.9150	0.8416	0.7251	0.3410	0.1184	0.0062
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9978	0.9862	0.9683	0.9313	0.7176	0.4596	0.1136
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	0	0.8775	0.5133	0.2542	0.0550	0.0238	0.0097	0.0013	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9928	0.8646	0.6213	0.2336	0.1267	0.0637	0.0126	0.0017	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9997	0.9755	0.8661	0.5017	0.3326	0.2025	0.0579	0.0112	0.0013	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9969	0.9658	0.7473	0.5843	0.4206	0.1686	0.0461	0.0078	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9997	0.9935	0.9009	0.7940	0.6543	0.3530	0.1334	0.0321	0.0040	0.0010	0.0002	0.0000	0.0000	0.0000
	5	1.0000	1.0000	0.9991	0.9700	0.9198	0.8346	0.5744	0.2905	0.0977	0.0182	0.0056	0.0012	0.0000	0.0000	0.0000
	6	1.0000	1.0000	0.9999	0.9930	0.9757	0.9376	0.7712	0.5000	0.2288	0.0624	0.0243	0.0070	0.0001	0.0000	0.0000
	7	1.0000	1.0000	1.0000	0.9988	0.9944	0.9818	0.9023	0.7095	0.4256	0.1654	0.0802	0.0300	0.0009	0.0000	0.0000
	8	1.0000	1.0000	1.0000	0.9998	0.9990	0.9960	0.9679	0.8666	0.6470	0.3457	0.2060	0.0991	0.0065	0.0003	0.0000
	9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9922	0.9539	0.8314	0.5794	0.4157	0.2527	0.0342	0.0031	0.0000
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9987	0.9888	0.9421	0.7975	0.6674	0.4983	0.1339	0.0245	0.0003
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9983	0.9874	0.9363	0.8733	0.7664	0.3787	0.1354	0.0072
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9987	0.9903	0.9762	0.9450	0.7458	0.4867	0.1225
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$ (continuación).

		p														
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
14	0	0.8687	0.4877	0.2288	0.0440	0.0178	0.0068	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9916	0.8470	0.5846	0.1979	0.1010	0.0475	0.0081	0.0009	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9997	0.9699	0.8416	0.4481	0.2811	0.1608	0.0398	0.0065	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9958	0.9559	0.6982	0.5213	0.3552	0.1243	0.0287	0.0039	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9996	0.9908	0.8702	0.7415	0.5842	0.2793	0.0898	0.0175	0.0017	0.0003	0.0000	0.0000	0.0000	0.0000
	5	1.0000	1.0000	0.9985	0.9561	0.8883	0.7805	0.4859	0.2120	0.0583	0.0083	0.0022	0.0004	0.0000	0.0000	0.0000
	6	1.0000	1.0000	0.9998	0.9884	0.9617	0.9067	0.6925	0.3953	0.1501	0.0315	0.0103	0.0024	0.0000	0.0000	0.0000
	7	1.0000	1.0000	1.0000	0.9976	0.9897	0.9685	0.8499	0.6047	0.3075	0.0933	0.0383	0.0116	0.0002	0.0000	0.0000
	8	1.0000	1.0000	1.0000	0.9996	0.9978	0.9917	0.9417	0.7880	0.5141	0.2195	0.1117	0.0439	0.0015	0.0000	0.0000
	9	1.0000	1.0000	1.0000	1.0000	0.9997	0.9983	0.9825	0.9102	0.7207	0.4158	0.2585	0.1298	0.0092	0.0004	0.0000
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9961	0.9713	0.8757	0.6448	0.4787	0.3018	0.0441	0.0042	0.0000
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9994	0.9935	0.9602	0.8392	0.7189	0.5519	0.1584	0.0301	0.0003
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9991	0.9919	0.9525	0.8990	0.8021	0.4154	0.1530	0.0084
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9932	0.9822	0.9560	0.7712	0.5123	0.1313
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0	0.8601	0.4633	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9904	0.8290	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9996	0.9638	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9945	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9994	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9999	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0008	0.0001	0.0000	0.0000	0.0000
	6	1.0000	1.0000	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0042	0.0008	0.0000	0.0000	0.0000
	7	1.0000	1.0000	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0173	0.0042	0.0000	0.0000	0.0000
	8	1.0000	1.0000	1.0000	0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0566	0.0181	0.0003	0.0000	0.0000
	9	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.1484	0.0611	0.0022	0.0001	0.0000
	10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.3135	0.1642	0.0127	0.0006	0.0000
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.5387	0.3518	0.0556	0.0055	0.0000
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9963	0.9729	0.8732	0.7639	0.6020	0.1841	0.0362	0.0004
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9948	0.9647	0.9198	0.8329	0.4510	0.1710	0.0096
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9953	0.9866	0.9648	0.7941	0.5367	0.1399
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$ (continuación).

		p														
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
16	0	0.8515	0.4401	0.1853	0.0281	0.0100	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9891	0.8108	0.5147	0.1407	0.0635	0.0261	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9995	0.9571	0.7892	0.3518	0.1971	0.0994	0.0183	0.0021	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9930	0.9316	0.5981	0.4050	0.2459	0.0651	0.0106	0.0009	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9991	0.9830	0.7982	0.6302	0.4499	0.1666	0.0384	0.0049	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9999	0.9967	0.9183	0.8103	0.6598	0.3288	0.1051	0.0191	0.0016	0.0003	0.0000	0.0000	0.0000	0.0000
	6	1.0000	1.0000	0.9995	0.9733	0.9204	0.8247	0.5272	0.2272	0.0583	0.0071	0.0016	0.0002	0.0000	0.0000	0.0000
	7	1.0000	1.0000	0.9999	0.9930	0.9729	0.9256	0.7161	0.4018	0.1423	0.0257	0.0075	0.0015	0.0000	0.0000	0.0000
	8	1.0000	1.0000	1.0000	0.9985	0.9925	0.9743	0.8577	0.5982	0.2839	0.0744	0.0271	0.0070	0.0001	0.0000	0.0000
	9	1.0000	1.0000	1.0000	0.9998	0.9984	0.9929	0.9417	0.7728	0.4728	0.1753	0.0796	0.0267	0.0005	0.0000	0.0000
	10	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9809	0.8949	0.6712	0.3402	0.1897	0.0817	0.0033	0.0001	0.0000
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9951	0.9616	0.8334	0.5501	0.3698	0.2018	0.0170	0.0009	0.0000
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9991	0.9894	0.9349	0.7541	0.5950	0.4019	0.0684	0.0070	0.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9979	0.9817	0.9006	0.8029	0.6482	0.2108	0.0429	0.0005
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9967	0.9739	0.9365	0.8593	0.4853	0.1892	0.0109
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9967	0.9900	0.9719	0.8147	0.5599	0.1485
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
17	0	0.8429	0.4181	0.1668	0.0225	0.0075	0.0023	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9877	0.7922	0.4818	0.1182	0.0501	0.0193	0.0021	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9994	0.9497	0.7618	0.3096	0.1637	0.0774	0.0123	0.0012	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9912	0.9174	0.5489	0.3530	0.2019	0.0464	0.0064	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9988	0.9779	0.7582	0.5739	0.3887	0.1260	0.0245	0.0025	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9999	0.9953	0.8943	0.7653	0.5968	0.2639	0.0717	0.0106	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
	6	1.0000	1.0000	0.9992	0.9623	0.8929	0.7752	0.4478	0.1662	0.0348	0.0032	0.0006	0.0001	0.0000	0.0000	0.0000
	7	1.0000	1.0000	0.9999	0.9891	0.9598	0.8954	0.6405	0.3145	0.0919	0.0127	0.0031	0.0005	0.0000	0.0000	0.0000
	8	1.0000	1.0000	1.0000	0.9974	0.9876	0.9597	0.8011	0.5000	0.1989	0.0403	0.0124	0.0026	0.0000	0.0000	0.0000
	9	1.0000	1.0000	1.0000	0.9995	0.9969	0.9873	0.9081	0.6855	0.3595	0.1046	0.0402	0.0109	0.0001	0.0000	0.0000
	10	1.0000	1.0000	1.0000	0.9999	0.9994	0.9968	0.9652	0.8338	0.5522	0.2248	0.1071	0.0377	0.0008	0.0000	0.0000
	11	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9894	0.9283	0.7361	0.4032	0.2347	0.1057	0.0047	0.0001	0.0000
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9975	0.9755	0.8740	0.6113	0.4261	0.2418	0.0221	0.0012	0.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9936	0.9536	0.7981	0.6470	0.4511	0.0826	0.0088	0.0000
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9988	0.9877	0.9226	0.8363	0.6904	0.2382	0.0503	0.0006
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9979	0.9807	0.9499	0.8818	0.5182	0.2078	0.0123
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9977	0.9925	0.9775	0.8332	0.5819	0.1571
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$ (continuación).

		p														
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
18	0	0.8345	0.3972	0.1501	0.0180	0.0056	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9862	0.7735	0.4503	0.0991	0.0395	0.0142	0.0013	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9993	0.9419	0.7338	0.2713	0.1353	0.0600	0.0082	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9891	0.9018	0.5010	0.3057	0.1646	0.0328	0.0038	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9985	0.9718	0.7164	0.5187	0.3327	0.0942	0.0154	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9998	0.9936	0.8671	0.7175	0.5344	0.2088	0.0481	0.0058	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	6	1.0000	1.0000	0.9988	0.9487	0.8610	0.7217	0.3743	0.1189	0.0203	0.0014	0.0002	0.0000	0.0000	0.0000	0.0000
	7	1.0000	1.0000	0.9998	0.9837	0.9431	0.8593	0.5634	0.2403	0.0576	0.0061	0.0012	0.0002	0.0000	0.0000	0.0000
	8	1.0000	1.0000	1.0000	0.9957	0.9807	0.9404	0.7368	0.4073	0.1347	0.0210	0.0054	0.0009	0.0000	0.0000	0.0000
	9	1.0000	1.0000	1.0000	0.9991	0.9946	0.9790	0.8653	0.5927	0.2632	0.0596	0.0193	0.0043	0.0000	0.0000	0.0000
	10	1.0000	1.0000	1.0000	0.9998	0.9988	0.9939	0.9424	0.7597	0.4366	0.1407	0.0569	0.0163	0.0002	0.0000	0.0000
	11	1.0000	1.0000	1.0000	1.0000	0.9998	0.9986	0.9797	0.8811	0.6257	0.2783	0.1390	0.0513	0.0012	0.0000	0.0000
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9942	0.9519	0.7912	0.4656	0.2825	0.1329	0.0064	0.0002	0.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9987	0.9846	0.9058	0.6673	0.4813	0.2836	0.0282	0.0015	0.0000
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9962	0.9672	0.8354	0.6943	0.4990	0.0982	0.0109	0.0000
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9993	0.9918	0.9400	0.8647	0.7287	0.2662	0.0581	0.0007
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9987	0.9858	0.9605	0.9009	0.5497	0.2265	0.0138
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9984	0.9944	0.9820	0.8499	0.6028	0.1655
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
19	0	0.8262	0.3774	0.1351	0.0144	0.0042	0.0011	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9847	0.7547	0.4203	0.0829	0.0310	0.0104	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9991	0.9335	0.7054	0.2369	0.1113	0.0462	0.0055	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9868	0.8850	0.4551	0.2631	0.1332	0.0230	0.0022	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9980	0.9648	0.6733	0.4654	0.2822	0.0696	0.0096	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9998	0.9914	0.8369	0.6678	0.4739	0.1629	0.0318	0.0031	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	6	1.0000	1.0000	0.9983	0.9324	0.8251	0.6655	0.3081	0.0835	0.0116	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
	7	1.0000	1.0000	0.9997	0.9767	0.9225	0.8180	0.4878	0.1796	0.0352	0.0028	0.0005	0.0000	0.0000	0.0000	0.0000
	8	1.0000	1.0000	1.0000	0.9933	0.9713	0.9161	0.6675	0.3238	0.0885	0.0105	0.0023	0.0003	0.0000	0.0000	0.0000
	9	1.0000	1.0000	1.0000	0.9984	0.9911	0.9674	0.8139	0.5000	0.1861	0.0326	0.0089	0.0016	0.0000	0.0000	0.0000
	10	1.0000	1.0000	1.0000	0.9997	0.9977	0.9895	0.9115	0.6762	0.3325	0.0839	0.0287	0.0067	0.0000	0.0000	0.0000
	11	1.0000	1.0000	1.0000	1.0000	0.9995	0.9972	0.9648	0.8204	0.5122	0.1820	0.0775	0.0233	0.0003	0.0000	0.0000
	12	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9884	0.9165	0.6919	0.3345	0.1749	0.0676	0.0017	0.0000	0.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9969	0.9682	0.8371	0.5261	0.3322	0.1631	0.0086	0.0002	0.0000
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9994	0.9904	0.9304	0.7178	0.5346	0.3267	0.0352	0.0020	0.0000
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9978	0.9770	0.8668	0.7369	0.5449	0.1150	0.0132	0.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9996	0.9945	0.9538	0.8887	0.7631	0.2946	0.0665	0.0009
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9896	0.9690	0.9171	0.5797	0.2453	0.0153
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9989	0.9958	0.9856	0.8649	0.6226	0.1738
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$ (continuación).

		p														
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
20	0	0.8179	0.3585	0.1216	0.0115	0.0032	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9831	0.7358	0.3917	0.0692	0.0243	0.0076	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9990	0.9245	0.6769	0.2061	0.0913	0.0355	0.0036	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	1.0000	0.9841	0.8670	0.4114	0.2252	0.1071	0.0160	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9974	0.9568	0.6296	0.4148	0.2375	0.0510	0.0059	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9997	0.9887	0.8042	0.6172	0.4164	0.1256	0.0207	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	1.0000	1.0000	0.9976	0.9133	0.7858	0.6080	0.2500	0.0577	0.0065	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	7	1.0000	1.0000	0.9996	0.9679	0.8982	0.7723	0.4159	0.1316	0.0210	0.0013	0.0002	0.0000	0.0000	0.0000	0.0000
	8	1.0000	1.0000	0.9999	0.9900	0.9591	0.8867	0.5956	0.2517	0.0565	0.0051	0.0009	0.0001	0.0000	0.0000	0.0000
	9	1.0000	1.0000	1.0000	0.9974	0.9861	0.9520	0.7553	0.4119	0.1275	0.0171	0.0039	0.0006	0.0000	0.0000	0.0000
	10	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.8725	0.5881	0.2447	0.0480	0.0139	0.0026	0.0000	0.0000	0.0000
	11	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9435	0.7483	0.4044	0.1133	0.0409	0.0100	0.0001	0.0000	0.0000
	12	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9790	0.8684	0.5841	0.2277	0.1018	0.0321	0.0004	0.0000	0.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.2142	0.0867	0.0024	0.0000	0.0000
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9984	0.9793	0.8744	0.5836	0.3828	0.1958	0.0113	0.0003	0.0000
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9941	0.9490	0.7625	0.5852	0.3704	0.0432	0.0026	0.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9987	0.9840	0.8929	0.7748	0.5886	0.1330	0.0159	0.0000
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9964	0.9645	0.9087	0.7939	0.3231	0.0755	0.0010
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9924	0.9757	0.9308	0.6083	0.2642	0.0169
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9968	0.9885	0.8784	0.6415	0.1821
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$ (continuación).

		p														
n	r	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.500	0.60	0.70	0.75	0.80	0.90	0.95	0.99
25	0	0.7778	0.2774	0.0718	0.0038	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9742	0.6424	0.2712	0.0274	0.0070	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9980	0.8729	0.5371	0.0982	0.0321	0.0090	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9999	0.9659	0.7636	0.2340	0.0962	0.0332	0.0024	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9928	0.9020	0.4207	0.2137	0.0905	0.0095	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9988	0.9666	0.6167	0.3783	0.1935	0.0294	0.0020	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	1.0000	0.9998	0.9905	0.7800	0.5611	0.3407	0.0736	0.0073	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	1.0000	1.0000	0.9977	0.8909	0.7265	0.5118	0.1536	0.0216	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	1.0000	1.0000	0.9995	0.9532	0.8506	0.6769	0.2735	0.0539	0.0043	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	9	1.0000	1.0000	0.9999	0.9827	0.9287	0.8106	0.4246	0.1148	0.0132	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
	10	1.0000	1.0000	1.0000	0.9944	0.9703	0.9022	0.5858	0.2122	0.0344	0.0018	0.0002	0.0000	0.0000	0.0000	0.0000
	11	1.0000	1.0000	1.0000	0.9985	0.9893	0.9558	0.7323	0.3450	0.0778	0.0060	0.0009	0.0001	0.0000	0.0000	0.0000
	12	1.0000	1.0000	1.0000	0.9996	0.9966	0.9825	0.8462	0.5000	0.1538	0.0175	0.0034	0.0004	0.0000	0.0000	0.0000
	13	1.0000	1.0000	1.0000	0.9999	0.9991	0.9940	0.9222	0.6550	0.2677	0.0442	0.0107	0.0015	0.0000	0.0000	0.0000
	14	1.0000	1.0000	1.0000	1.0000	0.9998	0.9982	0.9656	0.7878	0.4142	0.0978	0.0297	0.0056	0.0000	0.0000	0.0000
	15	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9868	0.8852	0.5754	0.1894	0.0713	0.0173	0.0001	0.0000	0.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9957	0.9461	0.7265	0.3231	0.1494	0.0468	0.0005	0.0000	0.0000
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9988	0.9784	0.8464	0.4882	0.2735	0.1091	0.0023	0.0000	0.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9927	0.9264	0.6593	0.4389	0.2200	0.0095	0.0002	0.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9980	0.9706	0.8065	0.6217	0.3833	0.0334	0.0012	0.0000
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9905	0.9095	0.7863	0.5793	0.0980	0.0072	0.0000
	21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9976	0.9668	0.9038	0.7660	0.2364	0.0341	0.0001
	22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9996	0.9910	0.9679	0.9018	0.4629	0.1271	0.0020
	23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9984	0.9930	0.9726	0.7288	0.3576	0.0258
	24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9962	0.9282	0.7226	0.2222
	25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.1. Sumas de probabilidad binomial $\sum_{x=0}^{r} b(x; n, p)$ (continuación).

		p														
n	r	0.01	0.05	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95	0.99
30	0	0.7397	0.2146	0.0424	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.9639	0.5535	0.1837	0.0105	0.0020	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.9967	0.8122	0.4114	0.0442	0.0106	0.0021	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9998	0.9392	0.6474	0.1227	0.0374	0.0093	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	1.0000	0.9844	0.8245	0.2552	0.0979	0.0302	0.0015	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	1.0000	0.9967	0.9268	0.4275	0.2026	0.0766	0.0057	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	1.0000	0.9994	0.9742	0.6070	0.3481	0.1595	0.0172	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	1.0000	0.9999	0.9922	0.7608	0.5143	0.2814	0.0435	0.0026	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	1.0000	1.0000	0.9980	0.8713	0.6736	0.4315	0.0940	0.0081	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	1.0000	1.0000	0.9995	0.9389	0.8034	0.5888	0.1763	0.0214	0.0009	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	1.0000	1.0000	0.9999	0.9744	0.8943	0.7304	0.2915	0.0494	0.0029	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	1.0000	1.0000	1.0000	0.9905	0.9493	0.8407	0.4311	0.1002	0.0083	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	12	1.0000	1.0000	1.0000	0.9969	0.9784	0.9155	0.5785	0.1808	0.0212	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
	13	1.0000	1.0000	1.0000	0.9991	0.9918	0.9599	0.7145	0.2923	0.0481	0.0021	0.0002	0.0000	0.0000	0.0000	0.0000
	14	1.0000	1.0000	1.0000	0.9998	0.9973	0.9831	0.8246	0.4278	0.0971	0.0064	0.0008	0.0001	0.0000	0.0000	0.0000
	15	1.0000	1.0000	1.0000	0.9999	0.9992	0.9936	0.9029	0.5722	0.1754	0.0169	0.0027	0.0002	0.0000	0.0000	0.0000
	16	1.0000	1.0000	1.0000	1.0000	0.9998	0.9979	0.9519	0.7077	0.2855	0.0401	0.0082	0.0009	0.0000	0.0000	0.0000
	17	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9788	0.8192	0.4215	0.0845	0.0216	0.0031	0.0000	0.0000	0.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9917	0.8998	0.5689	0.1593	0.0507	0.0095	0.0000	0.0000	0.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9971	0.9506	0.7085	0.2696	0.1057	0.0256	0.0001	0.0000	0.0000
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9991	0.9786	0.8237	0.4112	0.1966	0.0611	0.0005	0.0000	0.0000
	21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9919	0.9060	0.5685	0.3264	0.1287	0.0020	0.0000	0.0000
	22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9974	0.9565	0.7186	0.4857	0.2392	0.0078	0.0001	0.0000
	23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9993	0.9828	0.8405	0.6519	0.3930	0.0258	0.0006	0.0000
	24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9943	0.9234	0.7974	0.5725	0.0732	0.0033	0.0000
	25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	0.9698	0.9021	0.7448	0.1755	0.0156	0.0000
	26	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9907	0.9626	0.8773	0.3526	0.0608	0.0002
	27	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9979	0.9894	0.9558	0.5886	0.1878	0.0033
	28	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9980	0.9895	0.8163	0.4465	0.0361
	29	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9988	0.9576	0.7854	0.2603
	30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.2. Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x;\lambda)$.

	λ								
r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865
4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

	λ								
r	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
0	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067
1	0.7358	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404
2	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247
3	0.9810	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650
4	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405
5	0.9994	0.9955	0.9834	0.9580	0.9161	0.8576	0.7851	0.7029	0.6160
6	0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622
7	1.0000	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666
8	1.0000	1.0000	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319
9	1.0000	1.0000	1.0000	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682
10	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9972	0.9933	0.9863
11	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991	0.9976	0.9945
12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.2. Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x;\lambda)$ (Continuación).

	λ								
r	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5
0	0.0041	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001
1	0.0266	0.0174	0.0113	0.0073	0.0047	0.0030	0.0019	0.0012	0.0008
2	0.0884	0.0620	0.0430	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042
3	0.2017	0.1512	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149
4	0.3575	0.2851	0.2237	0.1730	0.1321	0.0996	0.0744	0.0550	0.0403
5	0.5289	0.4457	0.3690	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885
6	0.6860	0.6063	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649
7	0.8095	0.7440	0.6728	0.5987	0.5246	0.4530	0.3856	0.3239	0.2687
8	0.8944	0.8472	0.7916	0.7291	0.6620	0.5925	0.5231	0.4557	0.3918
9	0.9462	0.9161	0.8774	0.8305	0.7764	0.7166	0.6530	0.5874	0.5218
10	0.9747	0.9574	0.9332	0.9015	0.8622	0.8159	0.7634	0.7060	0.6453
11	0.9890	0.9799	0.9661	0.9467	0.9208	0.8881	0.8487	0.8030	0.7520
12	0.9955	0.9912	0.9840	0.9730	0.9573	0.9362	0.9091	0.8758	0.8364
13	0.9983	0.9964	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981
14	0.9994	0.9986	0.9970	0.9943	0.9897	0.9827	0.9726	0.9585	0.9400
15	0.9998	0.9995	0.9988	0.9976	0.9954	0.9918	0.9862	0.9780	0.9665
16	0.9999	0.9998	0.9996	0.9990	0.9980	0.9963	0.9934	0.9889	0.9823
17	1.0000	0.9999	0.9998	0.9996	0.9992	0.9984	0.9970	0.9947	0.9911
18	1.0000	1.0000	0.9999	0.9999	0.9997	0.9993	0.9987	0.9976	0.9957
19	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9995	0.9989	0.9980
20	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9991
21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996
22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla A.2. Sumas de probabilidad de Poisson $\sum_{x=0}^{r} p(x;\lambda)$ (Continuación).

	λ								
r	10	11	12	13	14	15	16	17	18.8
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0028	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
3	0.0103	0.0049	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000
4	0.0293	0.0151	0.0076	0.0037	0.0018	0.0009	0.0004	0.0002	0.0000
5	0.0671	0.0375	0.0203	0.0107	0.0055	0.0028	0.0014	0.0007	0.0002
6	0.1301	0.0786	0.0458	0.0259	0.0142	0.0076	0.0040	0.0021	0.0006
7	0.2202	0.1432	0.0895	0.0540	0.0316	0.0180	0.0100	0.0054	0.0017
8	0.3328	0.2320	0.1550	0.0998	0.0621	0.0374	0.0220	0.0126	0.0044
9	0.4579	0.3405	0.2424	0.1658	0.1094	0.0699	0.0433	0.0261	0.0099
10	0.5830	0.4599	0.3472	0.2517	0.1757	0.1185	0.0774	0.0491	0.0203
11	0.6968	0.5793	0.4616	0.3532	0.2600	0.1848	0.1270	0.0847	0.0381
12	0.7916	0.6887	0.5760	0.4631	0.3585	0.2676	0.1931	0.1350	0.0659
13	0.8645	0.7813	0.6815	0.5730	0.4644	0.3632	0.2745	0.2009	0.1062
14	0.9165	0.8540	0.7720	0.6751	0.5704	0.4657	0.3675	0.2808	0.1603
15	0.9513	0.9074	0.8444	0.7636	0.6694	0.5681	0.4667	0.3715	0.2281
16	0.9730	0.9441	0.8987	0.8355	0.7559	0.6641	0.5660	0.4677	0.3077
17	0.9857	0.9678	0.9370	0.8905	0.8272	0.7489	0.6593	0.5640	0.3958
18	0.9928	0.9823	0.9626	0.9302	0.8826	0.8195	0.7423	0.6550	0.4878
19	0.9965	0.9907	0.9787	0.9573	0.9235	0.8752	0.8122	0.7363	0.5788
20	0.9984	0.9953	0.9884	0.9750	0.9521	0.9170	0.8682	0.8055	0.6644
21	0.9993	0.9977	0.9939	0.9859	0.9712	0.9469	0.9108	0.8615	0.7410
22	0.9997	0.9990	0.9970	0.9924	0.9833	0.9673	0.9418	0.9047	0.8065
23	0.9999	0.9995	0.9985	0.9960	0.9907	0.9805	0.9633	0.9367	0.8600
24	1.0000	0.9998	0.9993	0.9980	0.9950	0.9888	0.9777	0.9594	0.9019
25	1.0000	0.9999	0.9997	0.9990	0.9974	0.9938	0.9869	0.9748	0.9334
26	1.0000	1.0000	0.9999	0.9995	0.9987	0.9967	0.9925	0.9848	0.9562
27	1.0000	1.0000	0.9999	0.9998	0.9994	0.9983	0.9959	0.9912	0.9720
28	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991	0.9978	0.9950	0.9827
29	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9989	0.9973	0.9896
30	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9986	0.9939
31	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9966
32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9981
33	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9990
34	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995
35	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997
36	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
37	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999

Tabla A.3. Área bajo la curva normal.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Tabla A.3. Área bajo la curva normal (Continuación).

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Tabla A.4. Valores críticos de la distribución *t*.

	α						
v	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Tabla A.4. Valores críticos de la distribución t (Continuación).

	α						
v	**0.02**	**0.015**	**0.01**	**0.0075**	**0.005**	**0.0025**	**0.0005**
1	15.895	21.205	31.821	42.433	63.657	127.321	636.619
2	4.849	5.643	6.965	8.073	9.925	14.089	31.599
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.290

Tabla A.5. Valores críticos de la distribución chi cuadrado.

	α									
v	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.75	0.70	0.50
1	0.000	0.000	0.001	0.001	0.004	0.016	0.064	0.102	0.148	0.455
2	0.010	0.020	0.040	0.051	0.103	0.211	0.446	0.575	0.713	1.386
3	0.072	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	9.299	9.926	12.340
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.037	11.721	14.339
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	9.260	10.196	11.293	11.689	13.091	14.848	17.187	18.137	19.021	22.337
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867	24.337
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	11.808	12.879	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.336
30	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336
40	20.707	22.164	23.838	24.433	26.509	29.051	32.345	33.660	34.872	39.335
50	27.991	29.707	31.664	32.357	34.764	37.689	41.449	42.942	44.313	49.335
60	35.534	37.485	39.699	40.482	43.188	46.459	50.641	52.294	53.809	59.335

Tabla A.5. Valores críticos de la distribución chi cuadrado (Continuación).

	α									
v	0.30	0.25	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.828
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.816
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.467
5	6.064	6.626	7.289	9.236	11.070	12.833	13.388	15.086	16.750	20.515
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.458
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.559	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	29.246	30.435	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	30.319	31.528	32.912	36.741	40.113	43.195	44.140	46.963	49.645	55.476
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.993	56.892
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.336	58.301
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.703
40	44.165	45.616	47.269	51.805	55.758	59.342	60.436	63.691	66.766	73.402
50	54.723	56.334	58.164	63.167	67.505	71.420	72.613	76.154	79.490	86.661
60	65.227	66.981	68.972	74.397	79.082	83.298	84.580	88.379	91.952	99.607

Tabla A.6. Valores críticos de la distribución *F*.

	$f_{0.05}(v_1, v_2)$								
$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Tabla A.6. Valores críticos de la distribución F (Continuación).

	$f_{0.05}(v_1, v_2)$									
$v_2 \backslash v_1$	10	12	15	20	24	30	40	60	120	∞
1	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
50	2.03	1.95	1.87	1.78	1.74	1.69	1.63	1.58	1.51	1.44
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Tabla A.6. Valores críticos de la distribución F (Continuación).

	$f_{0.01}(v_1, v_2)$								
$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Tabla A.6. Valores críticos de la distribución F (Continuación).

$v_2 \backslash v_1$	$f_{0.01}(v_1, v_2)$									
	10	12	15	20	24	30	40	60	120	∞
1	6055.85	6106.32	6157.28	6208.73	6234.63	6260.65	6286.78	6313.03	6339.39	6365.86
2	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
50	2.70	2.56	2.42	2.27	2.18	2.10	2.01	1.91	1.80	1.68
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Tabla A.7. Valores críticos de la distribución del estadístico de Kolmogorov-Smirnov.

	α				
n	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**
1	0.900	0.950	0.975	0.990	0.995
2	0.684	0.776	0.842	0.900	0.929
3	0.565	0.636	0.708	0.785	0.829
4	0.493	0.565	0.624	0.689	0.734
5	0.447	0.509	0.563	0.627	0.669
6	0.410	0.468	0.519	0.577	0.617
7	0.381	0.436	0.483	0.538	0.576
8	0.358	0.410	0.454	0.507	0.542
9	0.339	0.387	0.430	0.480	0.513
10	0.323	0.369	0.409	0.457	0.489
11	0.308	0.352	0.391	0.437	0.468
12	0.296	0.338	0.375	0.419	0.449
13	0.285	0.325	0.361	0.404	0.432
14	0.275	0.314	0.349	0.390	0.418
15	0.266	0.304	0.338	0.377	0.404
16	0.258	0.295	0.327	0.366	0.392
17	0.250	0.286	0.318	0.355	0.381
18	0.244	0.279	0.309	0.346	0.371
19	0.237	0.271	0.301	0.337	0.361
20	0.232	0.265	0.294	0.329	0.352
21	0.226	0.259	0.287	0.321	0.344
22	0.221	0.253	0.281	0.314	0.337
23	0.216	0.247	0.275	0.307	0.330
24	0.212	0.242	0.269	0.301	0.323
25	0.208	0.238	0.264	0.295	0.317
26	0.204	0.233	0.259	0.290	0.311
27	0.200	0.229	0.254	0.284	0.305
28	0.197	0.225	0.250	0.279	0.300
29	0.193	0.221	0.246	0.275	0.295
30	0.190	0.218	0.242	0.270	0.290
31	0.187	0.214	0.238	0.266	0.285
32	0.184	0.211	0.234	0.262	0.281
33	0.182	0.208	0.231	0.258	0.277
34	0.179	0.205	0.227	0.254	0.273
35	0.177	0.202	0.224	0.251	0.269
36	0.174	0.199	0.221	0.247	0.265
37	0.172	0.196	0.218	0.244	0.262
38	0.170	0.194	0.215	0.241	0.258
39	0.168	0.191	0.213	0.238	0.255
40	0.165	0.189	0.210	0.235	0.252
> 40	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Tabla A.8. Valores críticos de la distribución del estadístico de Lilliefors.

	α				
n	**0.2**	**0.15**	**0.1**	**0.05**	**0.01**
4	0.300	0.319	0.352	0.381	0.417
5	0.285	0.299	0.315	0.337	0.405
6	0.265	0.277	0.294	0.319	0.364
7	0.247	0.258	0.276	0.300	0.348
8	0.233	0.244	0.261	0.285	0.331
9	0.223	0.233	0.249	0.271	0.311
10	0.215	0.224	0.239	0.258	0.294
11	0.206	0.217	0.230	0.249	0.284
12	0.199	0.212	0.223	0.242	0.275
13	0.190	0.202	0.214	0.234	0.268
14	0.183	0.194	0.207	0.227	0.261
15	0.177	0.187	0.201	0.220	0.257
16	0.173	0.182	0.195	0.213	0.250
17	0.169	0.177	0.189	0.206	0.245
18	0.166	0.173	0.184	0.200	0.239
19	0.163	0.169	0.179	0.195	0.235
20	0.160	0.166	0.174	0.190	0.231
25	0.149	0.153	0.165	0.180	0.203
30	0.131	0.136	0.144	0.161	0.187
> 30	$\frac{0.736}{\sqrt{n}}$	$\frac{0.768}{\sqrt{n}}$	$\frac{0.805}{\sqrt{n}}$	$\frac{0.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

Tabla A.9. Coeficientes a_{n-i+1} para el test de normalidad de Shapiro-Wilk.

	n									
i	2	3	4	5	6	7	8	9	10	
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	
2		0.0000	0.1877	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	
3				0.0000	0.0875	0.1401	0.1743	0.1976	0.2141	
4						0.0000	0.0561	0.0947	0.1224	
5								0.0000	0.0399	

	n									
i	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1420	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8					0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9							0.0000	0.0163	0.0303	0.0422
10									0.0000	0.0140

	n									
i	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	0.3285	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944
3	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.2510	0.2499	0.2487
4	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148
5	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836	0.1848	0.1857	0.1864	0.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316	0.1346	0.1372	0.1395	0.1415
8	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219
9	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036
10	0.0263	0.0368	0.0549	0.0539	0.0610	0.0672	0.0728	0.0778	0.1822	0.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12			0.0000	00107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537
13					0.0000	0.0094	0.0178	0.0253	0.0320	0.0381
14							0.000	0.0084	0159	0.0227
15									0.000	0076

Tabla A.9. Coeficientes a_{n-i+1} para el test de normalidad de Shapiro-Wilk (Continuación).

	n									
i	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964
2	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774	0.2755	0.2737
3	0.2475	0.2463	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391	0.2380	0.2368
4	0.2145	0.2141	0.2137	0.2132	0.2127	0.2121	0.2116	0.2110	0.2104	0.2098
5	0.2874	0.1878	0.1880	0.1882	0.1883	0.1883	0.1883	0.1881	0.1880	0.1878
6	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691
7	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513	0.1520	0.1526
8	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356	0.1366	0.1376
9	0.1066	0.1093	0.1118	0.1140	0.1160	0.1179	0.1196	0.1211	0.1225	0.1237
10	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075	0.1092	0.1108
11	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986
12	0.0585	0.0629	0.0669	0.0706	0.0739	0.0770	0.0798	0.0824	0.0848	0.0870
13	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706	0.0733	0.0759
14	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592	0.0622	0.0651
15	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481	0.0515	0.0546
16	0.0000	0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444
17			0.0000	0.0062	0.0119	0.0172	0.0220	0.0264	0.0305	0.0343
18					0.0000	0.0057	0.0110	0.0158	0.0203	0.0244
19							0.0000	0.0053	0.0101	0.0146
20									0.0000	0.0049

	n									
i	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
3	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
4	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
8	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
9	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
10	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
11	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
13	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	0.0288	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
21	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22			0.0000	0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23					0.0000	0.0039	0.0076	0.0111	0.0143	0.0174
24							0.0000	0.0037	0.007	0.0104
25									0.0000	0.0035

Tabla A.10. Valores críticos para el test de Shapiro-Wilk.

	α								
n	0.01	0.02	0.05	0.10	0.50	0.90	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.964	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	0.878	0.892	0.911	0.926	0.961	0.980	0.984	0.987	0.989
23	0.881	0.895	0.914	0.928	0.962	0.981	0.984	0.987	0.989
24	0.884	0.898	0.916	0.930	0.963	0.981	0.984	0.987	0.989
25	0.888	0.901	0.918	0.931	0.964	0.981	0.985	0.988	0.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	0.894	0.906	0.923	0.935	0.965	0.982	0.985	0.988	0.990
28	0.896	0.908	0.924	0.936	0.966	0.982	0.985	0.988	0.990
29	0.898	0.910	0.926	0.937	0.966	0.982	0.985	0.988	0.990
30	0.900	0.912	0.927	0.939	0.967	0.983	0.985	0.988	0.990
31	0.902	0.914	0.929	0.940	0.967	0.983	0.986	0.988	0.990
32	0.904	0.915	0.930	0.941	0.968	0.983	0.986	0.988	0.990
33	0.906	0.917	0.931	0.942	0.968	0.983	0.986	0.989	0.990
34	0.908	0.919	0.933	0.943	0.969	0.983	0.986	0.989	0.990
35	0.910	0.920	0.934	0.944	0.969	0.984	0.986	0.989	0.990
36	0.912	0.922	0.935	0.945	0.970	0.984	0.986	0.989	0.990
37	0.914	0.924	0.936	0.946	0.970	0.984	0.987	0.989	0.990
38	0.916	0.925	0.938	0.947	0.971	0.984	0.987	0.989	0.990
39	0.917	0.927	0.939	0.948	0.971	0.984	0.987	0.989	0.991
40	0.918	0.928	0.940	0.949	0.972	0.985	0.987	0.989	0.991
41	0.920	0.929	0.941	0.950	0.972	0.985	0.987	0.989	0.991
42	0.922	0.930	0.942	0.951	0.972	0.985	0.987	0.989	0.991
43	0.923	0.932	0.943	0.951	0.973	0.985	0.987	0.990	0.991
44	0.924	0.933	0.944	0.952	0.973	0.985	0.987	0.990	0.991
45	0.926	0.934	0.945	0.953	0.973	0.985	0.988	0.990	0.991
46	0.927	0.935	0.945	0.953	0.974	0.985	0.988	0.990	0.991
47	0.928	0.936	0.946	0.954	0.974	0.985	0.988	0.990	0.991
48	0.929	0.937	0.947	0.954	0.974	0.985	0.988	0.990	0.991
49	0.929	0.937	0.947	0.955	0.974	0.985	0.988	0.990	0.991
50	0.930	0.938	0.947	0.955	0.974	0.985	0.988	0.990	0.991

Tabla A.11. Valores críticos para el test de Cochran.

	$\alpha = 0.01$													
	n													
k	2	3	4	5	6	7	8	9	10	11	17	37	145	∞
2	0.9999	0.9950	0.9794	0.9586	0.9373	0.9172	0.8988	0.8823	0.8674	0.8539	0.7949	0.7067	0.6062	0.5000
3	0.9933	0.9423	0.8831	0.8355	0.7933	0.7606	0.7335	0.7107	0.6912	0.6743	0.6059	0.5153	0.4230	0.3333
4	0.9676	0.8643	0.7814	0.7212	0.6761	0.6410	0.6129	0.5897	0.5702	0.5536	0.4884	0.4057	0.3251	0.2500
5	0.9279	0.7885	0.6957	0.6329	0.5875	0.5531	0.5259	0.5037	0.4854	0.4697	0.4094	0.3351	0.2644	0.2000
6	0.8828	0.7218	0.6258	0.5635	0.5195	0.4866	0.4608	0.4401	0.4229	0.4084	0.3529	0.2858	0.2229	0.1667
7	0.8376	0.6644	0.5685	0.5080	0.4659	0.4347	0.4105	0.3911	0.3751	0.3616	0.3105	0.2494	0.1929	0.1429
8	0.7945	0.6152	0.5209	0.4627	0.4226	0.3932	0.3704	0.3522	0.3373	0.3248	0.2779	0.2214	0.1700	0.1250
9	0.7544	0.5727	0.4810	0.4251	0.3870	0.3592	0.3378	0.3207	0.3067	0.2950	0.2514	0.1992	0.1521	0.1111
10	0.7175	0.5358	0.4469	0.3934	0.3572	0.3308	0.3106	0.2945	0.2813	0.2704	0.2297	0.1811	0.1376	0.1000
12	0.6528	0.4751	0.3919	0.3428	0.3099	0.2861	0.2680	0.2535	0.2419	0.2320	0.1961	0.1535	0.1157	0.0833
15	0.5747	0.4069	0.3317	0.2882	0.2593	0.2386	0.2228	0.2104	0.2002	0.1918	0.1612	0.1251	0.0934	0.0667
20	0.4799	0.3297	0.2654	0.2288	0.2048	0.1877	0.1748	0.1646	0.1567	0.1501	0.1248	0.0960	0.0709	0.0500
24	0.4247	0.2871	0.2295	0.1970	0.1759	0.1608	0.1495	0.1406	0.1338	0.1283	0.1060	0.0810	0.0595	0.0417
30	0.3632	0.2412	0.1913	0.1635	0.1454	0.1327	0.1232	0.1157	0.1100	0.1054	0.0867	0.0658	0.0480	0.0333
40	0.2940	0.1915	0.1508	0.1281	0.1135	0.1033	0.0957	0.0898	0.0853	0.0816	0.0668	0.0503	0.0363	0.0250
60	0.2151	0.1371	0.1069	0.0902	0.0796	0.0722	0.0668	0.0625	0.0594	0.0567	0.0461	0.0344	0.0245	0.0167
120	0.1225	0.0759	0.0585	0.0489	0.0429	0.0387	0.0357	0.0334	0.0316	0.0302	0.0242	0.0178	0.0125	0.0083
∞	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Tabla A.11. Valores críticos para el test de Cochran (Continuación).

	$\alpha = 0.01$													
	n													
k	2	3	4	5	6	7	8	9	10	11	17	37	145	∞
2	0.9985	0.9750	0.9392	0.9057	0.8772	0.8534	0.8332	0.8159	0.8010	0.7880	0.7341	0.6602	0.5813	0.5000
3	0.9669	0.8709	0.7977	0.7457	0.7071	0.6771	0.6530	0.6333	0.1667	0.6025	0.5466	0.4748	0.4031	0.3333
4	0.9065	0.7679	0.6841	0.6287	0.5895	0.5598	0.5365	0.5175	0.5017	0.4884	0.4366	0.3720	0.3093	0.2500
5	0.8412	0.6838	0.5981	0.5441	0.5065	0.4783	0.4564	0.4387	0.4241	0.4118	0.3645	0.3066	0.2513	0.2000
6	0.7808	0.6161	0.5321	0.4803	0.4447	0.4184	0.3980	0.3817	0.3682	0.3568	0.3135	0.2612	0.2119	0.1667
7	0.7271	0.5612	0.4800	0.4307	0.3974	0.3726	0.3535	0.3384	0.3259	0.3154	0.2756	0.2278	0.1833	0.1429
8	0.6798	0.5157	0.4377	0.3910	0.3595	0.3362	0.3185	0.3043	0.2926	0.2829	0.2462	0.2022	0.1616	0.1250
9	0.6385	0.4775	0.4027	0.3584	0.3286	0.3067	0.2901	0.2768	0.2659	0.2568	0.2226	0.1820	0.1446	0.1111
10	0.6020	0.4450	0.3733	0.3311	0.3029	0.2823	0.2666	0.2541	0.2439	0.2353	0.2032	0.1655	0.1308	0.1000
12	0.5410	0.3924	0.3264	0.2880	0.2624	0.2439	0.2299	0.2187	0.2098	0.2020	0.1737	0.1403	0.1100	0.0833
15	0.4709	0.3346	0.2758	0.2419	0.2195	0.2034	0.1911	0.1815	0.1736	0.1671	0.1429	0.1144	0.0889	0.0667
20	0.3894	0.2705	0.2205	0.1921	0.1735	0.1602	0.1501	0.1422	0.1357	0.1303	0.1108	0.0879	0.0675	0.0500
24	0.3434	0.2354	0.1907	0.1656	0.1493	0.1374	0.1286	0.1216	0.1160	0.1113	0.0942	0.0743	0.0567	0.0417
30	0.2929	0.1980	0.1593	0.1377	0.1237	0.1137	0.1061	0.1002	0.0958	0.0921	0.0771	0.0604	0.0457	0.0333
40	0.2370	0.1576	0.1259	0.1082	0.0968	0.0887	0.0827	0.0780	0.0745	0.0713	0.0595	0.0462	0.0347	0.0250
60	0.1737	0.1131	0.9895	0.0765	0.0682	0.0623	0.0583	0.0552	0.0520	0.0497	0.0411	0.0316	0.0234	0.0167
120	0.0998	0.0632	0.0495	0.0419	0.0371	0.0337	0.0312	0.0292	0.0279	0.0266	0.0218	0.0165	0.0120	0.0083
∞	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Tabla A.12. Puntos porcentuales de de la distribución de rango estudentizado.

	$\alpha = 0.05$																		
	p																		
υ	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	18.1	26.7	32.8	37.2	40.5	43.1	45.4	47.3	49.1	50.6	51.9	53.2	54.3	55.4	56.3	57.2	58.0	58.8	59.6
2	6.09	8.28	9.80	10.89	11.73	12.43	13.03	13.54	13.99	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.36	16.57	16.77
3	4.50	5.88	6.83	7.51	8.04	8.47	8.85	9.18	9.46	9.72	9.95	10.16	10.35	10.52	10.69	10.84	10.98	11.12	11.24
4	3.93	5.00	5.76	6.31	6.73	7.06	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.67	8.80	8.92	9.03	9.14	9.24
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21
6	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65	6.79	6.92	7.04	7.14	7.24	7.34	7.43	7.51	7.59
7	3.34	4.16	4.68	5.06	5.35	5.59	5.80	5.99	6.15	6.29	6.42	6.54	6.65	6.75	6.84	6.93	7.01	7.08	7.16
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87
9	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.65
10	3.15	3.88	4.33	4.66	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.12	6.20	6.27	6.34	6.41	6.47
11	3.10	3.82	4.26	4.58	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.14	6.20	6.27	6.33
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21
13	3.06	3.73	4.15	4.46	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	6.93	6.00	6.06	6.11
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.56	5.64	5.72	5.79	5.86	5.92	5.98	6.03
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.79	5.85	5.91	5.96
16	3.00	3.65	4.05	4.34	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90
17	2.98	3.62	4.02	4.31	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.55	5.61	5.68	5.74	5.79	5.84
18	2.97	3.61	4.00	4.28	4.49	4.67	4.83	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79
19	2.96	3.59	3.98	4.26	4.47	4.64	4.79	4.92	5.04	5.14	5.23	5.32	5.39	5.46	5.53	5.59	5.65	5.70	5.75
20	2.95	3.58	3.96	4.24	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.50	5.56	5.61	5.66	5.71
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.50	5.55	5.59
30	2.89	3.48	3.84	4.11	4.30	4.46	4.60	4.72	4.83	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.48
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82	4.90	4.98	5.05	5.11	5.17	5.22	5.27	5.32	5.36
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24
120	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13
∞	2.77	3.32	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	6.62	4.68	4.74	4.8–0	4.84	4.89	4.89	4.97	5.01

Tabla A.12. Puntos porcentuales de de la distribución de rango estudentizado (Continuación).

	$\alpha = 0.01$																		
	k																		
v	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	90.0	135	164	186	202	216	227	237	246	253	260	266	272	272	282	286	290	294	298
2	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7	32.6	33.4	34.1	34.8	35.4	36.0	36.5	37.0	37.5	37.9
3	8.26	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7	17.1	17.5	17.9	18.2	18.5	18.8	19.1	19.3	19.5	19.8
4	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3	12.6	12.8	13.1	13.3	13.5	13.7	13.9	14.1	14.2	14.4
5	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.49	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65
8	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65	7.78	7.91	8.03	8.13	8.23	8.32	8.41	8.49	8.57
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.48	7.60	7.71	7.81	7.91	7.99	8.07	8.15	8.22
11	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95
12	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19	7.27	7.34	7.42	7.48	7.55
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.12	7.20	7.27	7.33	7.39
15	4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26
16	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.45	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.80	6.87	6.94	7.00	7.05
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65	6.72	6.79	6.85	6.91	6.96
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.29	6.37	6.45	6.52	6.59	6.65	6.71	6.76	6.82
24	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41
40	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69	5.77	5.84	5.90	5.96	6.02	6.07	6.12	6.17	6.21
60	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.79	5.84	5.89	5.93	5.98	6.02
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38	5.40	5.51	5.56	5.61	5.66	5.71	5.75	5.79	5.83
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.62	5.65

Tabla A.13. Valores críticos para el test de Duncan

	$\alpha = 0.05$								
	k								
v	2	3	4	5	6	7	8	9	10
1	19.97	19.97	19.97	19.97	19.97	19.97	19.97	19.97	19.97
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.927	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.797	3.814	3.814	3.814	3.814	3.814	3.814
6	3.461	3.587	3.649	3.680	3.694	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.626	3.626	3.626
8	3.261	3.399	3.475	3.521	3.549	3.566	3.575	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522
11	3.113	3.256	3.342	3.397	3.435	3.462	3.480	3.493	3.501
12	3.082	3.225	3.313	3.370	3.410	3.439	3.459	3.474	3.484
13	3.055	3.200	3.289	3.348	3.389	3.419	3.442	3.458	3.470
14	3.033	3.178	3.268	3.329	3.372	3.403	3.426	3.444	3.457
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.432	3.446
16	2.998	3.144	3.235	3.298	3.343	3.376	3.402	3.422	3.437
17	2.984	3.130	3.222	3.285	3.331	3.366	3.392	3.412	3.429
18	2.971	3.118	3.210	3.274	3.321	3.356	3.383	3.405	3.421
19	2.960	3.107	3.199	3.264	3.311	3.347	3.375	3.397	3.415
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.391	3.409
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390
30	2.888	3.035	3.131	3.199	3.350	3.290	3.322	3.349	3.371
40	2.858	3.006	3.102	3.171	3.224	3.266	3.300	3.328	3.352
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.287	3.314
∞	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294

Tabla A.13. Valores críticos para el test de Duncan (Continuación).

	$\alpha = 0.01$								
	k								
v	2	3	4	5	6	7	8	9	10
1	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03
2	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04
3	8.261	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321
4	6.512	6.667	6.740	6.756	6.756	6.756	6.756	6.756	6.756
5	5.702	5.893	5.989	6.040	6.065	6.074	6.074	6.074	6.074
6	5.243	5.439	5.549	5.614	5.655	5.680	5.694	5.701	5.703
7	4.949	5.145	5.260	5.334	5.383	5.416	5.439	5.454	5.464
8	4.746	4.939	5.057	5.135	5189	5.227	5.256	5.276	5.291
9	4.596	4.787	4.906	4.986	5.043	5.086	5.118	5.142	5.160
10	4.482	4.671	4.790	4.871	4.931	4.975	5.010	5.037	5.058
11	4.392	4.579	4.697	4.780	4.841	4.887	4.924	4.952	4.975
12	4.320	4.504	4.622	4.706	4.767	4.815	4.852	4.883	4.907
13	4.260	4.442	4.560	4.644	4.706	4.755	4.793	4.824	4.850
14	4.210	4.391	4.508	4.591	4.654	4.704	4.743	4.775	4.802
15	4.168	4.347	4.463	4.547	4.610	4.660	4.700	4.733	4.760
16	4.131	4.309	4.425	4.509	4.572	4.622	4.663	4.696	4.724
17	4.099	4.275	4.391	4.475	4.539	4.589	4.630	4.664	4.693
18	4.071	4.246	4.362	4.445	4.509	4.560	4.601	4.635	4.664
19	4.046	4.220	4.335	4.419	4.483	4.534	4.575	4.610	4.639
20	4.024	4.197	4.312	4.395	4.459	4.510	4.552	4.587	4.617
24	3.956	4.126	4.239	4.322	4.386	4.437	4.480	4.516	4.546
30	3.889	4.056	4.168	4.250	4.314	4.366	4.409	4.445	4.477
40	3.825	3.988	4.098	4.180	4.244	4.296	4.339	4.376	4.408
60	3.762	3.922	4.031	4.111	4.174	4.226	4.270	4.307	4.340
120	3.702	3.858	3.965	4.044	4.107	4.158	4.202	4.239	4.272
∞	3.643	3.796	3.900	3.978	4.040	4.091	4.135	4.172	4.205

Tabla A.14. Valores de $d_{\alpha,k-1,\upsilon}$ para el test de Dunnett bilateral.

	$\alpha = 0.05$												
	k												
υ	2	3	4	5	6	7	8	9	10	11	12	15	20
5	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97	4.03	4.09	4.14	4.26	4.42
6	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71	3.76	3.81	3.86	3.97	4.11
7	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53	3.58	3.63	3.67	3.78	3.91
8	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41	3.46	3.50	3.54	3.64	3.76
9	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32	3.36	3.40	3.44	3.53	3.65
10	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24	3.29	3.33	3.36	3.45	3.57
11	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19	3.23	3.27	3.30	3.39	3.50
12	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14	3.18	3.22	3.25	3.34	3.45
13	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10	3.14	3.18	3.21	3.29	3.40
14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07	3.11	3.14	3.18	3.26	3.36
15	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04	3.08	312	3.15	3.23	3.33
16	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02	3.06	3.09	3.12	3.20	3.30
17	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00	3.01	3.07	3.10	3.18	3.27
18	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98	3.00	3.05	3.08	3.16	3.25
19	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96	2.98	3.03	3.06	3.14	3.23
20	2.39	2.54	2.65	2.73	2.80	2.86	2.90	2.95	2.94	3.01	3.05	3.12	3.22
24	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90	2.89	2.97	3.00	3.07	3.16
30	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86	2.85	2.92	2.95	3.02	3.11
40	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81	2.80	2.87	2.90	2.97	3.06
60	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77	2.83	2.83	2.86	2.92	3.00
120	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73	2.79	2.79	2.81	2.87	2.95
∞	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69	2.74	2.74	2.77	2.83	2.91

Tabla A.15. Valores de $d_{\alpha,k-1,\upsilon}$ para el test de Dunnett unilateral.

	$\alpha = 0.05$												
	k												
υ	2	3	4	5	6	7	8	9	10	11	12	15	20
5	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30	3.36	3.41	3.45	3.57	3.72
6	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12	3.17	3.22	3.26	3.37	3.50
7	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01	3.05	3.10	3.13	3.23	3.36
8	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92	3.96	3.01	3.04	3.14	3.25
9	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86	2.90	2.94	2.97	3.06	3.18
10	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81	2.85	2.89	2.92	3.01	3.12
11	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77	2.81	2.85	2.88	2.96	3.07
12	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74	2.78	2.81	2.84	2.93	3.03
13	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71	2.75	2.78	2.82	2.90	3.00
14	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69	2.72	2.76	2.79	2.87	2.97
15	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67	2.70	2.74	2.77	2.85	2.95
16	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	2.69	2.72	2.75	2.83	2.93
17	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64	2.67	2.71	2.74	2.81	2.91
18	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62	2.66	2.69	2.72	2.80	2.89
19	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61	2.65	2.68	2.71	2.79	2.88
20	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	2.64	2.67	2.70	2.77	2.87
24	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	2.60	2.64	2.66	2.74	2.83
30	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54	2.57	2.60	2.63	2.70	2.79
40	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	2.54	2.57	2.60	2.67	2.75
60	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48	2.51	2.54	2.56	2.63	2.72
120	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45	2.48	2.51	2.63	2.60	2.68
∞	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	2.45	2.48	2.50	2.56	2.64

Tabla A.16. Valores de $d_{\alpha,k-1,\upsilon}$ para el test de Dunnett bilateral.

	$\alpha = 0.01$												
	k												
υ	2	3	4	5	6	7	8	9	10	11	12	15	20
5	4.63	4.98	5.22	5.41	5.56	5.69	5.80	5.89	5.98	6.05	6.12	6.30	6.52
6	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28	5.35	5.41	5.47	5.62	5.81
7	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89	4.95	5.01	5.06	5.19	5.36
8	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62	4.68	4.73	4.78	4.90	5.05
9	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43	4.48	4.53	4.57	4.68	4.82
10	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28	4.33	4.37	4.42	4.52	4.65
11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16	4.21	4.25	4.29	4.39	4.52
12	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07	4.12	4.16	4.19	4.29	4.41
13	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99	4.04	4.08	4.11	4.20	4.32
14	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93	3.97	4.01	4.05	4.13	4.24
15	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88	3.92	3.95	3.99	4.07	4.18
16	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83	3.87	3.91	3.94	4.02	4.13
17	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79	3.83	3.86	3.90	3.98	4.08
18	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75	3.79	3.83	3.86	3.94	4.04
19	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72	3.76	3.79	3.83	3.90	4.00
20	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69	3.73	3.77	3.80	3.87	3.97
24	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61	3.64	3.68	3.70	3.78	3.87
30	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52	3.56	3.59	3.62	3.69	3.78
40	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44	3.48	3.51	3.53	3.60	3.68
60	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37	3.40	3.42	3.45	3.51	3.59
120	2.85	2.97	3.06	3.12	3.18	3.22	3.16	329	3.32	3.35	3.37	3.43	3.51
∞	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22	3.25	3.27	3.29	3.35	3.42

Tabla A.17. Valores de $d_{\alpha,k-1,\upsilon}$ para el test de Dunnett unilateral.

	$\alpha = 0.01$												
	k												
υ	2	3	4	5	6	7	8	9	10	11	12	15	20
5	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30	3.36	3.41	3.45	3.57	3.72
6	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12	3.17	3.22	3.26	3.37	3.50
7	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01	3.05	3.10	3.13	3.23	3.36
8	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92	2.96	3.01	3.04	3.14	3.25
9	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86	2.90	2.94	2.97	3.06	3.18
10	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81	2.85	2.89	2.92	3.01	3.12
11	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77	2.81	2.85	2.88	2.96	3.07
12	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74	2.78	2.81	2.84	2.93	3.03
13	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71	2.75	2.78	2.82	2.90	3.00
14	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69	2.72	2.76	2.79	2.87	2.97
15	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67	2.70	2.74	2.77	2.85	2.95
16	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	2.69	2.72	2.75	2.83	2.93
17	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64	2.67	2.71	2.74	2.81	2.91
18	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62	2.66	2.69	2.72	2.80	2.89
19	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61	2.65	2.68	2.71	2.79	2.88
20	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	2.64	2.67	2.70	2.77	2.87
24	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	2.60	2.64	2.66	2.74	2.83
30	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54	2.57	2.60	2.63	2.70	2.79
40	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	2.54	2.57	2.60	2.67	2.75
60	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48	2.51	2.54	2.56	2.63	2.72
120	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45	2.48	2.51	2.53	2.60	2.68
∞	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	2.45	2.48	2.50	2.56	2.64

Tabla A.18. Estadístico *d* de Durbin-Watson: limites d_L y d_U a una significancia de 0.05.

	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$		$k' = 6$		$k' = 7$		$k' = 8$		$k' = 9$		$k' = 10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.610	1.400																		
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.368	2.287	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	—	—	—	—	—	—	—	—	—	—	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	—	—	—	—	—	—	—	—	—	—
11	0.927	1.324	0.658	1.604	0.595	1.928	0.444	2.283	0.316	2.645	0.203	3.005	—	—	—	—	—	—	—	—
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506	0.268	2.832	0.171	3.149	—	—	—	—	—	—
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	—	—
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.472	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.257	0.502	2.461	0.407	2.667	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.692	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.732	2.124	0.637	2.290	0.547	2.460	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.734
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.751	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.012	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.958	0.874	2.071	0.798	2.188	0.723	2.309	0.650	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.682	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.795	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.080	1.891	1.015	1.979	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.877	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.198
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

Tabla A.18. Estadístico d de Durbin-Watson: limites d_L y d_U a una significancia de 0.05
(Continuación)

	$k' = 11$		$k' = 12$		$k' = 13$		$k' = 14$		$k' = 15$		$k' = 16$		$k' = 17$		$k' = 18$		$k' = 19$		$k' = 20$	
n	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L
16	0.098	3.503	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.138	3.378	0.087	3.557	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.177	3.265	0.123	3.441	0.078	3.603	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.220	3.159	0.160	3.335	0.111	3.496	0.070	3.642	—	—	—	—	—	—	—	—	—	—	—	—
20	0.263	3.063	0.200	3.234	0.145	3.395	0.100	3.542	0.063	3.676	—	—	—	—	—	—	—	—	—	—
21	0.307	2.976	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583	0.058	3.705	—	—	—	—	—	—	—	—
22	0.349	2.897	0.281	3.057	0.220	3.211	0.166	3.358	0.120	3.495	0.083	3.619	0.052	3.731	—	—	—	—	—	—
23	0.391	2.826	0.322	2.979	0.259	3.128	0.202	3.272	0.153	3.409	0.110	3.535	0.076	3.650	0.048	3.753	—	—	—	—
24	0.431	2.761	0.362	2.908	0.297	3.053	0.239	3.193	0.186	3.327	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773	—	—
25	0.470	2.702	0.400	2.844	0.335	2.983	0.275	3.119	0.221	3.251	0.172	3.376	0.130	3.494	0.094	3.604	0.065	3.702	0.041	3.790
26	0.508	2.649	0.438	2.784	0.373	2.919	0.312	3.051	0.256	3.179	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.544	2.600	0.475	2.730	0.409	2.859	0.348	2.987	0.291	3.112	0.238	3.233	0.191	3.349	0.149	3.460	0.112	3.563	0.081	3.658
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.928	0.325	3.050	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.104	3.592
29	0.612	2.515	0.544	2.634	0.479	2.755	0.418	2.874	0.359	2.992	0.305	3.107	0.254	3.219	0.208	3.327	0.166	3.431	0.129	3.528
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.392	2.937	0.337	3.050	0.286	3.160	0.238	3.266	0.195	3.368	0.156	3.465
31	0.674	2.443	0.608	2.553	0.545	2.665	0.484	2.776	0.425	2.887	0.370	2.996	0.317	3.103	0.269	3.208	0.224	3.309	0.183	3.406
32	0.703	2.411	0.638	2.517	0.576	2.625	0.515	2.733	0.457	2.840	0.401	2.946	0.349	3.050	0.299	3.153	0.253	3.252	0.211	3.348
33	0.731	2.382	0.668	2.484	0.606	2.588	0.546	2.692	0.488	2.796	0.432	2.899	0.379	3.000	0.329	3.100	0.283	3.198	0.239	3.293
34	0.758	2.355	0.695	2.454	0.634	2.554	0.575	2.654	0.518	2.754	0.462	2.854	0.409	2.954	0.359	3.051	0.312	3.147	0.267	3.240
35	0.783	2.330	0.722	2.425	0.662	2.521	0.604	2.619	0.547	2.716	0.492	2.813	0.439	2.910	0.388	3.005	0.340	3.099	0.295	3.190
36	0.808	2.306	0.748	2.398	0.689	2.492	0.631	2.586	0.575	2.680	0.520	2.774	0.467	2.868	0.417	2.961	0.369	3.053	0.323	3.142
37	0.831	2.285	0.772	2.374	0.714	2.464	0.657	2.555	0.602	2.646	0.548	2.738	0.495	2.829	0.445	2.920	0.397	3.009	0.351	3.097
38	0.854	2.265	0.796	2.351	0.739	2.438	0.683	2.526	0.628	2.614	0.575	2.703	0.522	2.792	0.472	2.880	0.424	2.968	0.378	3.054
39	0.875	2.246	0.819	2.329	0.763	2.413	0.707	2.499	0.653	2.585	0.600	2.671	0.549	2.757	0.499	2.843	0.451	2.929	0.404	3.013
40	0.896	2.228	0.840	2.309	0.785	2.391	0.731	2.473	0.678	2.557	0.626	2.641	0.575	2.724	0.525	2.808	0.477	2.892	0.430	2.974
45	0.988	2.156	0.938	2.225	0.887	2.296	0.838	2.367	0.788	2.439	0.740	2.512	0.692	2.586	0.644	2.659	0.598	2.733	0.553	2.807
50	1.064	2.103	1.019	2.163	0.973	2.225	0.927	2.287	0.882	2.350	0.836	2.414	0.792	2.479	0.747	2.544	0.703	2.610	0.660	2.675
55	1.129	2.062	1.087	2.116	1.045	2.170	1.003	2.225	0.961	2.281	0.919	2.338	0.877	2.396	0.836	2.454	0.795	2.512	0.754	2.571
60	1.184	2.031	1.145	2.079	1.106	2.127	1.068	2.177	1.029	2.227	0.990	2.278	0.951	2.330	0.913	2.382	0.874	2.434	0.836	2.487
65	1.231	2.006	1.195	2.049	1.160	2.093	1.124	2.138	1.088	2.183	1.052	2.229	1.016	2.276	0.980	2.323	0.944	2.371	0.908	2.419
70	1.272	1.986	1.239	2.026	1.206	2.066	1.172	2.106	1.139	2.148	1.105	2.189	1.072	2.232	1.038	2.275	1.005	2.318	0.971	2.362
75	1.308	1.970	1.277	2.006	1.247	2.043	1.215	2.080	1.184	2.118	1.153	2.156	1.121	2.195	1.090	2.235	1.058	2.275	1.027	2.315
80	1.340	1.957	1.311	1.991	1.283	2.024	1.253	2.059	1.224	2.093	1.195	2.129	1.165	2.165	1.136	2.201	1.106	2.238	1.076	2.275
85	1.369	1.946	1.342	1.977	1.315	2.009	1.287	2.040	1.260	2.073	1.232	2.105	1.205	2.139	1.177	2.172	1.149	2.206	1.121	2.241
90	1.395	1.937	1.369	1.966	1.344	1.995	1.318	2.025	1.292	2.055	1.266	2.085	1.240	2.116	1.213	2.148	1.187	2.179	1.160	2.211
95	1.418	1.929	1.394	1.956	1.370	1.984	1.345	2.012	1.321	2.040	1.296	2.068	1.271	2.097	1.247	2.126	1.222	2.156	1.197	2.186
100	1.439	1.923	1.416	1.948	1.393	1.974	1.371	2.000	1.347	2.026	1.324	2.053	1.301	2.080	1.277	2.108	1.253	2.135	1.229	2.164
150	1.579	1.892	1.564	1.908	1.550	1.924	1.535	1.940	1.519	1.956	1.504	1.972	1.489	1.989	1.474	2.006	1.458	2.023	1.443	2.040
200	1.654	1.885	1.643	1.896	1.632	1.908	1.621	1.919	1.610	1.931	1.599	1.943	1.588	1.955	1.576	1.967	1.565	1.979	1.554	1.991

Tabla A.18. Estadístico *d* de Durbin-Watson: limites d_L y d_U a una significancia de 0.01 (Continuación)

	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$		$k' = 6$		$k' = 7$		$k' = 8$		$k' = 9$		$k' = 10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.390	1.142	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.435	1.036	0.294	1.676	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.497	1.003	0.345	1.489	0.229	2.102	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	—	—	—	—	—	—	—	—	—	—	—	—
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	—	—	—	—	—	—	—	—	—	—
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	—	—	—	—	—	—	—	—
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053	—	—	—	—	—	—
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	—	—	—	—
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287	—	—
15	0.811	1.070	0.700	1.252	0.591	1.464	0.488	1.704	0.391	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.737	1.252	0.633	1.446	0.532	1.663	0.437	1.900	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.874	1.102	0.772	1.255	0.672	1.432	0.574	1.630	0.480	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.613	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.132	0.835	1.265	0.742	1.415	0.650	1.584	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.863	1.271	0.773	1.411	0.685	1.567	0.598	1.737	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.714
21	0.975	1.161	0.890	1.277	0.803	1.408	0.718	1.554	0.633	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.914	1.284	0.831	1.407	0.748	1.543	0.667	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	1.018	1.187	0.938	1.291	0.858	1.407	0.777	1.534	0.698	1.673	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	0.960	1.298	0.882	1.407	0.805	1.528	0.728	1.658	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.211	0.981	1.305	0.906	1.409	0.831	1.523	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362
26	1.072	1.222	1.001	1.312	0.928	1.411	0.855	1.518	0.783	1.635	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168	0.441	2.313
27	1.089	1.233	1.019	1.319	0.949	1.413	0.878	1.515	0.808	1.626	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131	0.473	2.269
28	1.104	1.244	1.037	1.325	0.969	1.415	0.900	1.513	0.832	1.618	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098	0.504	2.229
29	1.119	1.254	1.054	1.332	0.988	1.418	0.921	1.512	0.855	1.611	0.788	1.718	0.723	1.830	0.658	1.947	0.595	2.068	0.533	2.193
30	1.133	1.263	1.070	1.339	1.006	1.421	0.941	1.511	0.877	1.606	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041	0.562	2.160
31	1.147	1.273	1.085	1.345	1.023	1.425	0.960	1.510	0.897	1.601	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017	0.589	2.131
32	1.160	1.282	1.100	1.352	1.040	1.428	0.979	1.510	0.917	1.597	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995	0.615	2.104
33	1.172	1.291	1.114	1.358	1.055	1.432	0.996	1.510	0.936	1.594	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975	0.641	2.080
34	1.184	1.299	1.128	1.364	1.070	1.435	1.012	1.511	0.954	1.591	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957	0.665	2.057
35	1.195	1.307	1.140	1.370	1.085	1.439	1.028	1.512	0.971	1.589	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940	0.689	2.037
36	1.206	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.988	1.588	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925	0.711	2.018
37	1.217	1.323	1.165	1.382	1.112	1.446	1.058	1.514	1.004	1.586	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911	0.733	2.001
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.585	0.966	1.658	0.913	1.735	0.860	1.816	0.807	1.899	0.754	1.985
39	1.237	1.337	1.187	1.393	1.137	1.453	1.085	1.517	1.034	1.584	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887	0.774	1.970
40	1.246	1.344	1.198	1.398	1.148	1.457	1.098	1.518	1.048	1.584	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876	0.749	1.956
45	1.288	1.376	1.245	1.423	1.201	1.474	1.156	1.528	1.111	1.584	1.065	1.643	1.019	1.704	0.974	1.768	0.927	1.834	0.881	1.902
50	1.324	1.403	1.285	1.446	1.245	1.491	1.205	1.538	1.164	1.587	1.123	1.639	1.081	1.692	1.039	1.748	0.997	1.805	0.955	1.864
55	1.356	1.427	1.320	1.466	1.284	1.506	1.247	1.548	1.209	1.592	1.172	1.638	1.134	1.685	1.095	1.734	1.057	1.785	1.018	1.837
60	1.383	1.449	1.350	1.484	1.317	1.520	1.283	1.558	1.249	1.598	1.214	1.639	1.179	1.682	1.144	1.726	1.108	1.771	1.072	1.817
65	1.407	1.468	1.377	1.500	1.346	1.534	1.315	1.568	1.283	1.604	1.251	1.642	1.218	1.680	1.186	1.720	1.153	1.761	1.120	1.802
70	1.429	1.485	1.400	1.515	1.372	1.546	1.343	1.578	1.313	1.611	1.283	1.645	1.253	1.680	1.223	1.716	1.192	1.754	1.162	1.792
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.587	1.340	1.617	1.313	1.649	1.284	1.682	1.256	1.714	1.227	1.748	1.199	1.783
80	1.466	1.515	1.441	1.541	1.416	1.568	1.390	1.595	1.364	1.624	1.338	1.653	1.312	1.683	1.285	1.714	1.259	1.745	1.232	1.777
85	1.482	1.528	1.458	1.553	1.435	1.578	1.411	1.603	1.386	1.630	1.362	1.657	1.337	1.685	1.312	1.714	1.287	1.743	1.262	1.773
90	1.496	1.540	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636	1.383	1.661	1.360	1.687	1.336	1.714	1.312	1.741	1.288	1.769
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.642	1.403	1.666	1.381	1.690	1.358	1.715	1.336	1.741	1.313	1.767
100	1.522	1.562	1.503	1.583	1.482	1.604	1.462	1.625	1.441	1.647	1.421	1.670	1.400	1.693	1.378	1.717	1.357	1.741	1.335	1.765
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693	1.543	1.708	1.530	1.722	1.515	1.737	1.501	1.752	1.486	1.767
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725	1.613	1.735	1.603	1.746	1.592	1.757	1.582	1.768	1.571	1.779

Tabla A.18. Estadístico *d* de Durbin-Watson: limites d_L y d_U a una significancia de 0.01 (Continuación)

	$k' = 11$		$k' = 12$		$k' = 13$		$k' = 14$		$k' = 15$		$k' = 16$		$k' = 17$		$k' = 18$		$k' = 19$		$k' = 20$	
n	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L
16	0.060	3.446	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.084	3.286	0.053	3.506	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.113	3.146	0.075	3.358	0.047	3.357	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.145	3.023	0.102	3.227	0.067	3.420	0.043	3.601	—	—	—	—	—	—	—	—	—	—	—	—
20	0.178	2.914	0.131	3.109	0.092	3.297	0.061	3.474	0.038	3.639	—	—	—	—	—	—	—	—	—	—
21	0.212	2.817	0.162	3.004	0.119	3.185	0.084	3.358	0.055	3.521	0.035	3.671	—	—	—	—	—	—	—	—
22	0.246	2.729	0.194	2.909	0.148	3.084	0.109	3.252	0.077	3.412	0.050	3.562	0.032	3.700	—	—	—	—	—	—
23	0.281	2.651	0.227	2.822	0.178	2.991	0.136	3.155	0.100	3.311	0.070	3.459	0.046	3.597	0.029	3.725	—	—	—	—
24	0.315	2.580	0.260	2.744	0.209	2.906	0.165	3.065	0.125	3.218	0.092	3.363	0.065	3.501	0.043	3.629	0.027	3.747	—	—
25	0.348	2.517	0.292	2.674	0.240	2.829	0.194	2.982	0.152	3.131	0.116	3.274	0.085	3.410	0.060	3.538	0.039	3.657	0.025	3.766
26	0.381	2.460	0.324	2.610	0.272	2.758	0.224	2.906	0.180	3.050	0.141	3.191	0.107	3.325	0.079	3.452	0.055	3.572	0.036	3.682
27	0.413	2.409	0.356	2.552	0.303	2.694	0.253	2.836	0.208	2.976	0.167	3.113	0.131	3.245	0.100	3.371	0.073	3.490	0.051	3.602
28	0.444	2.363	0.387	2.499	0.333	2.635	0.283	2.772	0.237	2.907	0.194	3.040	0.156	3.169	0.122	3.294	0.093	3.412	0.068	3.524
29	0.474	2.321	0.417	2.451	0.363	2.582	0.313	2.713	0.266	2.843	0.222	2.972	0.182	3.098	0.146	3.220	0.114	3.338	0.087	3.450
30	0.503	2.283	0.447	2.407	0.393	2.533	0.342	2.659	0.294	2.785	0.249	2.909	0.208	3.032	0.171	3.152	0.137	3.267	0.107	3.379
31	0.531	2.248	0.475	2.367	0.422	2.487	0.371	2.609	0.322	2.730	0.277	2.851	0.234	2.970	0.196	3.087	0.160	3.201	0.128	3.311
32	0.558	2.216	0.503	2.330	0.450	2.446	0.399	2.563	0.350	2.680	0.304	2.797	0.261	2.912	0.221	3.026	0.184	3.137	0.151	3.246
33	0.585	2.187	0.530	2.296	0.477	2.408	0.426	2.520	0.377	2.633	0.331	2.746	0.287	2.858	0.246	2.969	0.209	3.078	0.174	3.184
34	0.610	2.160	0.556	2.266	0.503	2.373	0.452	2.481	0.404	2.590	0.357	2.699	0.313	2.808	0.272	2.915	0.233	3.022	0.197	3.126
35	0.634	2.136	0.581	2.237	0.529	2.340	0.478	2.444	0.430	2.550	0.383	2.655	0.339	2.761	0.297	2.865	0.257	2.969	0.221	3.071
36	0.658	2.113	0.605	2.210	0.554	2.310	0.504	2.410	0.455	2.512	0.409	2.614	0.364	2.717	0.322	2.818	0.282	2.919	0.244	3.019
37	0.680	2.092	0.628	2.186	0.578	2.282	0.528	2.379	0.480	2.477	0.434	2.576	0.389	2.675	0.347	2.774	0.306	2.872	0.268	2.969
38	0.702	2.073	0.651	2.164	0.601	2.256	0.552	2.350	0.504	2.445	0.458	2.540	0.414	2.637	0.371	2.733	0.330	2.828	0.291	2.923
39	0.723	2.055	0.673	2.143	0.623	2.232	0.575	2.323	0.528	2.414	0.482	2.507	0.438	2.600	0.395	2.694	0.354	2.787	0.315	2.879
40	0.744	2.039	0.694	2.123	0.645	2.210	0.597	2.297	0.551	2.386	0.505	2.476	0.461	2.566	0.418	2.657	0.377	2.748	0.338	2.838
45	0.835	1.972	0.790	2.044	0.744	2.118	0.700	2.193	0.655	2.269	0.612	2.346	0.570	2.424	0.528	2.503	0.488	2.582	0.448	2.661
50	0.913	1.925	0.871	1.987	0.829	2.051	0.787	2.116	0.746	2.182	0.705	2.250	0.665	2.318	0.625	2.387	0.586	2.456	0.548	2.526
55	0.979	1.891	0.940	1.945	0.902	2.002	0.863	2.059	0.825	2.117	0.786	2.176	0.748	2.237	0.711	2.298	0.674	2.359	0.637	2.421
60	1.037	1.865	1.001	1.914	0.965	1.964	0.929	2.015	0.893	2.067	0.857	2.120	0.822	2.173	0.786	2.227	0.751	2.283	0.716	2.338
65	1.087	1.845	1.053	1.889	1.020	1.934	0.986	1.980	0.953	2.027	0.919	2.075	0.886	2.123	0.852	2.172	0.819	2.221	0.786	2.272
70	1.131	1.831	1.099	1.870	1.068	1.911	1.037	1.953	1.005	1.995	0.974	2.038	0.943	2.082	0.911	2.127	0.880	2.172	0.849	2.217
75	1.170	1.819	1.141	1.856	1.111	1.893	1.082	1.931	1.052	1.970	1.023	2.009	0.993	2.049	0.964	2.090	0.934	2.131	0.905	2.172
80	1.205	1.810	1.177	1.844	1.150	1.878	1.122	1.913	1.094	1.949	1.066	1.984	1.039	2.022	1.011	2.059	0.983	2.097	0.955	2.135
85	1.236	1.803	1.210	1.834	1.184	1.866	1.158	1.898	1.132	1.931	1.106	1.965	1.080	1.999	1.053	2.033	1.027	2.068	1.000	2.104
90	1.264	1.798	1.240	1.827	1.215	1.856	1.191	1.886	1.166	1.917	1.141	1.948	1.116	1.979	1.091	2.012	1.066	2.044	1.041	2.077
95	1.290	1.793	1.267	1.821	1.244	1.848	1.221	1.876	1.197	1.905	1.174	1.934	1.150	1.963	1.126	1.993	1.102	2.023	1.079	2.054
100	1.314	1.790	1.292	1.816	1.270	1.841	1.248	1.868	1.225	1.895	1.203	1.922	1.181	1.949	1.158	1.977	1.136	2.006	1.113	2.034
150	1.473	1.783	1.458	1.799	1.444	1.814	1.429	1.830	1.414	1.847	1.400	1.863	1.385	1.880	1.370	1.897	1.355	1.913	1.340	1.931
200	1.561	1.791	1.550	1.801	1.539	1.813	1.528	1.824	1.518	1.836	1.507	1.847	1.495	1.860	1.484	1.871	1.474	1.883	1.462	1.896

Tabla A.19. Valores críticos para el test de rangos con signos de Wilcoxon.

n	Unilateral $\alpha = 0.01$ Bilateral $\alpha = 0.02$	Unilateral $\alpha = 0.025$ Bilateral $\alpha = 0.05$	Unilateral $\alpha = 0.05$ Bilateral $\alpha = 0.1$
5			1
6		1	2
7	0	2	4
8	2	4	6
9	3	6	8
10	5	8	11
11	7	11	14
12	10	14	17
13	13	17	21
14	16	21	26
15	20	25	30
16	24	30	36
17	28	35	41
18	33	40	47
19	38	46	54
20	43	52	60
21	49	59	68
22	56	66	75
23	62	73	83
24	69	81	92
25	77	90	101
26	85	98	110
27	93	107	120
28	102	117	130
29	111	127	141
30	120	137	152

Tabla A.20. Valores críticos para el test U de Mann Whitney.

Prueba unilarteral $\alpha = 0.001$ con o prueba bilateral con $\alpha = 0.002$

	n_2														
n_1	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1															
2															
3												0	0	0	0
4					0	0	0	1	1	1	2	2	3	3	3
5		0	0	1	1	2	2	3	3	4	5	5	6	7	7
6	0	1	2	2	3	4	4	5	6	7	8	9	10	11	12
7		2	3	3	5	6	7	8	9	10	11	13	14	15	16
8			5	5	6	8	9	11	12	14	15	17	18	20	21
9				7	8	10	12	14	15	17	19	21	23	25	26
10					10	12	14	17	19	21	23	25	27	29	32
11						15	17	20	22	24	27	29	32	34	37
12							20	23	25	28	31	34	37	40	42
13								26	29	32	35	38	42	45	48
14									32	36	39	43	46	50	54
15										40	43	47	51	55	59
16											48	52	56	60	65
17												57	61	66	70
18													66	71	76
19														77	82
20															88

Prueba unilarteral $\alpha = 0.01$ con o prueba bilateral con $\alpha = 0.02$

	n_2															
n_1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																
2									0	0	0	0	0	0	1	1
3			0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6		3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7			6	8	9	11	12	14	16	17	19	21	23	24	26	28
8				10	11	13	15	17	20	22	24	26	28	30	32	34
9					14	16	18	21	23	26	28	31	33	36	38	40
10						19	22	24	27	30	33	36	38	41	44	47
11							25	28	31	34	37	41	44	47	50	53
12								31	35	38	42	46	49	53	56	60
13									39	43	47	51	55	59	63	67
14										47	51	56	60	65	69	73
15											56	61	66	70	75	80
16												66	71	76	82	87
17													77	82	88	93
18														88	94	100
19															101	107
20																114

Tabla A.20. Valores críticos para el test U de Mann Whitney (continuación).

Prueba unilarteral $\alpha = 0.025$ con o prueba bilateral con $\alpha = 0.05$

	n_2																
n_1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																	
2					0	0	0	0	1	1	1	1	1	2	2	2	2
3		0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5		2	3	5	6	7	8	9	11	12	13	14	16	17	18	19	20
6			5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7				8	10	12	14	16	18	20	22	24	26	28	30	32	34
8					13	15	17	19	22	24	26	29	31	34	36	38	41
9						17	20	23	26	28	31	34	37	39	42	45	48
10							23	26	29	33	36	39	42	45	48	52	55
11								30	33	37	40	44	47	51	55	58	62
12									37	41	45	49	53	57	61	65	69
13										45	50	54	59	63	67	72	76
14											55	59	64	67	74	78	83
15												64	70	75	80	85	90
16													75	81	86	92	98
17														87	93	99	105
18															99	106	112
19																113	119
20																	127

Prueba unilarteral $\alpha = 0.05$ con o prueba bilateral con $\alpha = 0.1$

	n_2																
n_1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																		
2			0	0	0	1	1	1	1	2	2	3	3	3	3	4	4	4
3	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
4		1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5			4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6				7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7					11	13	15	17	19	21	24	26	28	30	33	35	37	39
8						15	18	20	23	26	28	31	33	36	39	41	44	47
9							21	24	27	30	33	36	39	42	45	48	51	54
10								27	31	34	37	41	44	48	51	55	58	62
11									34	38	42	46	50	54	57	61	65	69
12										42	47	51	55	60	64	68	72	77
13											51	56	61	65	70	75	80	84
14												61	66	71	77	82	87	92
15													72	77	83	88	94	100
16														83	89	95	101	107
17															96	102	109	115
18																109	116	123
19																	123	130

ISBN 978-958-8942-58-2