PRÁCTICAS DE ESTADÍSTICA CON R

Autores:
Santiago Angulo Díaz-Parreño (sangulo@ceu.es)
José Miguel Cárdenas Rebollo (cardenas@ceu.es)
José Rojo Montijano (jrojo.eps@ceu.es)
Anselmo Romero Limón (arlimon@ceu.es)
Alfredo Sánchez Alberca (asalber@ceu.es)

Curso 2011-2012

Version control information:
Head URL: https://practicas-r.googlecode.com/svn/trunk/practicas_r.tex
Last changed date: 2011-12-05 13:41:27 +0100 (lun 05 de dic de 2011)
Last changes revision: 17
Version: Revision 17
Last changed by: Alfredo Sánchez Alberca (asalber)

Prácticas de Estadística con R

Santiago Angulo Díaz-Parreño, José Miguel Cárdenas Rebollo, Anselmo Romero Limón y Alfredo Sánchez Alberca (asalber@gmail.com).

Índice general

Práctica de Estadística con R 1

Introducción a R y R-commander

1 Introducción

La gran potencia de cálculo alcanzada por los ordenadores ha convertido a los mismos en poderosas herramientas al servicio de todas aquellas disciplinas que, como la estadística, requieren manejar un gran volumen de datos. Actualmente, prácticamente nadie se plantea hacer un estudio estadístico serio sin la ayuda de un buen programa de análisis estadístico.

R es un potente lenguaje de programación que incluye multitud de funciones para la representación el análisis de datos. Fue desarrollado por Robert Gentleman y Ross Ihaka en la Universidad de Auckland en Nueva Zelanda, aunque actualmente es mantenido por una enorme comunidad científica en todo el mundo.

La ventajas de R frente a otros programas habituales de análisis de datos, como pueden ser SPSS, SAS, SPlus, Matlab o Minitab, son múltiples:

- Es software libre y por tanto gratuito. Puede descargarse desde la web http://www.r-project.org/.
- Es multiplataforma. Existen versiones para Windows, Macintosh, Linux y otras plataformas.
- Está avalado y en constante desarrollo por una amplia comunidad científica que lo utiliza como estándar para el análisis de datos.
- Cuenta con multitud de paquetes para todo tipo de análisis estadísticos y representaciones gráficas, desde los más habituales, hasta los más novedosos y sofisticados que no incluyen otros programas. Los paquetes están organizados y documentados en un repositorio CRAN (Comprehensive R Archive Network) desde donde pueden descargarse libremente. En España hay una copia de este repositorio en la web http://cran.es.r-project.org/.
- Es programable, lo que permite que el usuario pueda crear fácilmente sus propias funciones o paquetes para análisis de datos específicos.
- Existen multitud de libros, manuales y tutoriales libres que permiten su aprendizaje e ilustran el análisis estadístico de datos en distintas disciplinas científicas como las matemáticas, la física, la biología, la psicología, la medicina, etc.

Por defecto el entorno de trabajo de R es en línea de comandos, lo que significa que los cálculos y los análisis se relizan mediante comandos o instrucciones que el usuario teclea en una ventana de texto. No obstante, existen distintas interfaces gráficas de usuario que facilitan su uso, sobre todo para usuarios

noveles. Una de las interfaces gráficas de usuario más extendidas es *R-commander*, desarrollada por John Fox, y será la que se utilizará a lo largo de las prácticas.

El objetivo de esta práctica es introducir al alumno en la utilización de este programa, enseñándole a realizar las operaciones básicas más habituales de carga y manipulación de datos.

2 Instalación

2.1 Instalación de R

Linux En la distribución Debian y cualquiera de sus derivadas (Ubuntu, Kubuntu, etc.) basta con teclear en la línea de comandos

```
> sudo apt-get install r-base-html r-cran-rcmdr r-cran-rodbc r-doc
  -html r-recommended
```

Windows Descargar de http://cran.es.r-project.org/bin/windows/base/release.htm el programa de instalación de R, ejecutarlo y seguir las instrucciones de instalación.

2.2 Instalación de la interfaz gráfica R-commander y otros paquetes

La interfaz gráfica de usuario R-commander viene en el paquete `Rcmdr`, que puede instalarse como cualquier otro paquete para R. Para instalarlo, una vez arrancado R, hay que teclear el comando

```
> install.packages("Rcmdr", dependencies=TRUE)
```

La instalación de cualquier otro paquete se realiza con el mismo comando, cambiando el nombre del paquete por el deseado.

En Windows, también puede instalarse desde la ventana de R mediante el menú `Paquetes`→`Instalar Paquete(s)`, eligiendo el respositorio desde el cual se quiere instalar el paquete, por ejemplo Spain (Madrid), y seleccionando el paquete deseado.

3 Arranque

Como cualquier otra aplicación de Windows, para arrancar el programa hay que hacer click sobre la opción correspondiente del menú `Inicio`→`Programas`→`R`, o bien sobre el icono de escritorio

Una vez arrancado, para arrancar la interfaz gráfica de usuario R-commander hay que cargar el paquete `Rcmdr` y para ello hay que teclear el comando

```
> library("Rcmdr")
```

Para cargar cualquier otro paquete se utiliza el mismo comando, cambiando el nombre del paquete por el deseado.

Si se cierra R-commander, para volver a cargarlo se utiliza el comando `Commander()`.

La interfaz gráfica de usuario R-commander se muestra en la figura 1.1 por los siguientes elementos:

- **Barra de menús**. Contiene distintos menús con operaciones que pueden realizarse con R.
- **Barra de botones**. Contiene botones para seleccionar, editar o visualizar el conjunto de datos activos y también el modelo activo.

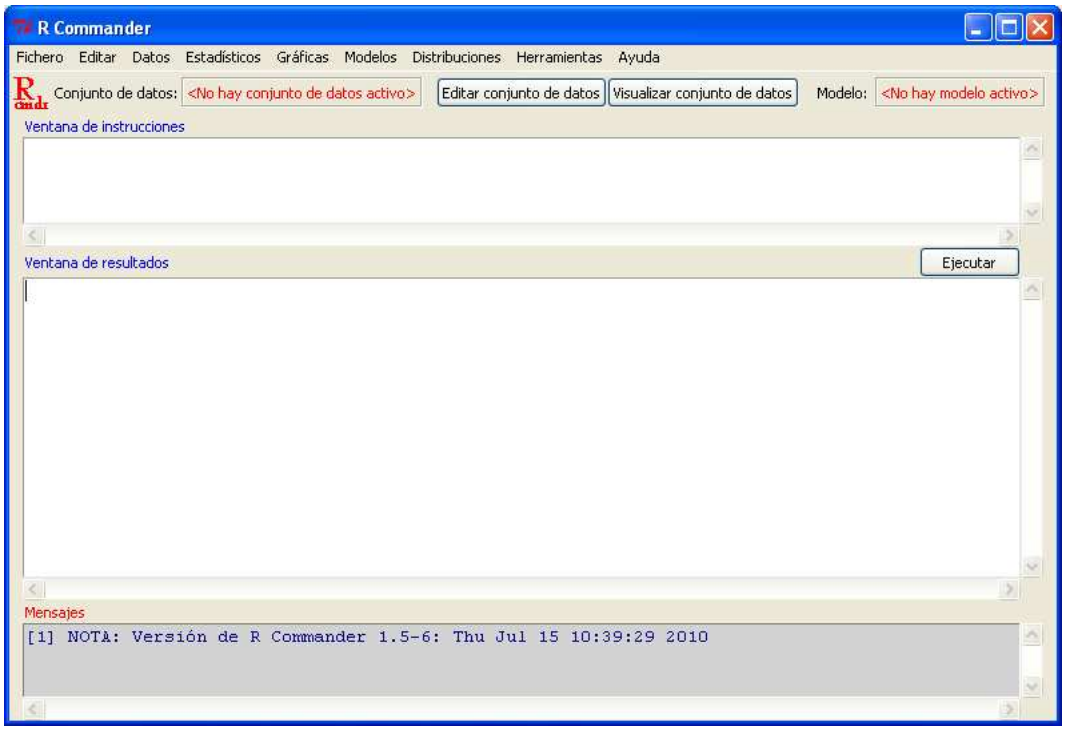


Figura 1.1 – Interfaz gráfica de usuario de R-commander.

- **Ventana de instrucciones**. Es una ventana de texto similar a la línea de comandos de R donde se pueden introducir directamente comandos para R. Se puede ejecutar un comando o un bloque de comandos situándose en la línea que contiene el comando y pulsando sobre el botón Ejecutar o bien tecleando `Ctrl+r`.

  Cada vez que se seleccione un menú que lleve asociado la ejecución de algún comando, dicho comando aparecerá en esta ventana. Esto permite modificar fácilmente los parámetros del comando y volver a ejecutarlo rápidamente sin necesidad de volver al menú.

- **Ventana de resultados**. Es una ventana de texto en la que aparecerán las salidas que generen los comandos que se ejecuten en R. Las instrucciones aparecen en color rojo y los resultados en azul.

- **Mensajes**. Es una ventana de texto donde se muestra información adicional sobre errores, advertencias u otra información auxiliar al ejecutar un comando.

4 Tipos de datos y operadores aritméticos y lógicos

En R existen distintos tipos de datos. Los más básicos son:

Numeric : Es cualquier número decimal. Se utiliza el punto como separador de decimales. Por defecto, cualquier número que se teclee tomará este tipo.

Integer : Es cualquier número entero. Para convertir un número de tipo Numeric en un entero se utiliza el comando `as.integer()`

Logical : Puede tomar cualquiera de los dos valores lógicos `TRUE` (verdadero) o `FALSE` (falso).

Character : Es cualquier cadena de caracteres alfanuméricos. Deben introducirse entre comillas. Para convertir cualquier número en una cadena de caracteres se utiliza el comando `as.character()`.

Los valores de estos tipos de datos pueden operarse utilizando distintos operadores o funciones predefinidas para cada tipo de datos. Los más habituales son:

Operadores aritméticos : `+` (suma), `-` (resta), `*` (producto), `/` (cociente), `^` (potencia).

Operadores de comparación : `>` (mayor), `<` (menor), `>=` (mayor o igual), `<=` (menor o igual), `==` (igual), `!=` (distinto).

Operadores lógicos : `&` (conjunción y), `|` (disyunción o), `!` (negación no).

Funciones predefinidas : `sqrt()` (raíz cuadrada), `abs()` (valor absoluto), `log()` (logarítmo neperiano), `exp()` (exponencial), `sin()` (seno), `cos()` (coseno), `tan()` (tangente).

Al evaluar las expresiones aritméticas existe un orden de prioridad entre los operadores de manera que primero se evaluan las funciones predefinidas, luego las potencias, luego los productos y cocientes, luego las sumas y restas, luego los operadores de comparación, luego las negaciones, luego las conjunciones y finalmente las disyunciones. Para forzar un orden de evaluación distinto del predefinido se pueden usar paréntesis. Por ejemplo

```
> 2^2+4/2
[1] 6
> (2^2+4)/2
[1] 4
> 2^(2+4/2)
[1] 16
> 2^(2+4)/2
[1] 32
> 2^((2+4)/2)
[1] 8
```

También es posible asignar valores a variables mediante el operador de asignación `=`. Una vez definidas, las variables pueden usarse en cualquier expresión aritmética o lógica. Por ejemplo,

```
> x=2
> y=x+2
> y
[1] 4
> y>x
[1] TRUE
> x>=y
[1] FALSE
> x==y-2
[1] TRUE
> x!=0 & !y<x
[1] TRUE
```

5 Introducción y manipulación de datos

Antes de realizar cualquier análisis de datos hay que introducir los datos que se quieren analizar.

5.1 Introducción de datos en línea de comandos

Existen muchas formas de introducir datos en R pero aquí sólo veremos las más habituales. La forma más sencilla de introducir datos es crear un vector de datos, mediante el comando `c()`. Por ejemplo, para introducir las notas de 5 alumnos se debe teclear

```
> nota = c(5.6,7.2,3.5,8.1,6.4)
```

Esto crea el vector nota con el que posteriormente se pueden realizar cálculos como por ejemplo la media

```
> mean(nota)
[1] 6.16
```

Otra forma habitual de introducir los datos de una muestra es crear un conjunto de datos mediante el comando `data.frame()`. Por ejemplo, para crear un conjunto de datos a partir de las notas anteriores, hay que teclear

```
> curso = data.frame(nota)
```

Esto crea una matriz de datos en la que cada columna se corresponde con una variable y cada fila con un individuo de la muestra. En el ejemplo la matriz curso sólo tendría una columna que se correspondería con las notas y 5 filas, cada una de ellas correspondiente a un alumno de la muestra. Es posible acceder a las variables de un conjunto de datos con el operador dolar `$`. Por ejemplo, para acceder a las notas hay que teclear

```
> curso$nota
[1] 5.6 7.2 3.5 8.1 6.4
```

Es fácil añadir nuevas variables a un conjunto de datos, pero siempre deben tener el mismo tamaño muestral. Por ejemplo, para añadir una nueva variable con el grupo (mañana o tarde) de los alumnos, hay que teclear

```
> curso$grupo = c("m","t","t","m","m")
```

Ahora el conjunto de datos curso tendría dos columnas, una para la nota y otra para el grupo de los alumnos. Tecleando el nombre de cualquier objeto, se muestra su información:

```
> curso
   nota   grupo
1   5.6      m
2   7.2      t
3   3.5      t
4   8.1      m
5   6.4      m
```

Cuando se introducen datos se puede utilizar el código `NA` (not available), para indicar la ausencia del dato.

Las variables definidas en cada sesión de trabajo quedan almacenas en la memoria interna de R. Es posible obtener un listado de todos los objetos almacenados en la memoria mediante los comandos `ls()`. Si se desea más información, el comando `ls.str()` además de mostrar los objetos de la memoria indica sus tipos y sus valores.

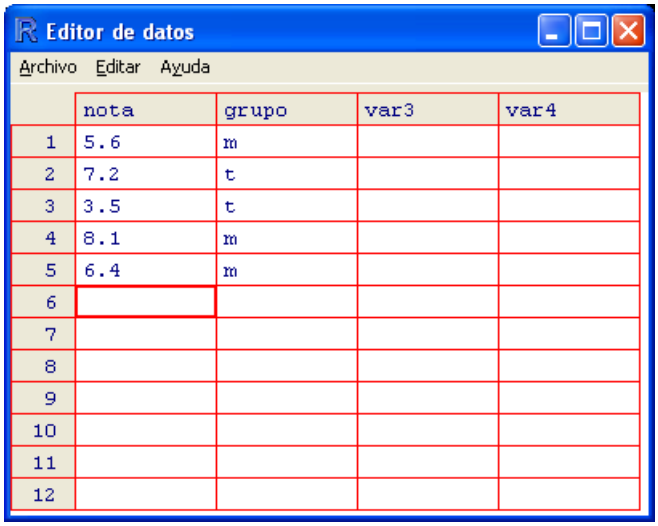
```
> ls()
[1] "curso" "nota"  "x"     "y"
> ls.str()
curso : 'data.frame':   5 obs. of  2 variables:
 $ nota : num  5.6 7.2 3.5 8.1 6.4
 $ grupo: chr  " m " " t " " t " " m " ...
nota :  num [1:5] 5.6 7.2 3.5 8.1 6.4
x :  num 2
y :  num 4
```

Para eliminar un objeto de la memoria se utiliza el comando `rm()`.

```
> ls()
[1] "curso" "nota"  "x"     "y"
> rm(x,y)
> ls()
[1] "curso" "nota"
```

5.2 Introducción de datos en R-comander

Para introducir los datos desde R-commander hay que ir al menu `Datos→Nuevo conjunto de datos`. Con esto aparecerá una ventana donde hay que darle un nombre al conjunto de datos y tras esto aparece la ventana de la figura 1.2 con una tabla en la que se pueden introducir los datos de la muestra. Al igual que antes, cada variable debe introducirse en una columna y cada individuo en una fila.



	nota	grupo	var3	var4
1	5.6	m		
2	7.2	t		
3	3.5	t		
4	8.1	m		
5	6.4	m		
6				
7				
8				
9				
10				
11				
12				

Figura 1.2 – Ventana de introducción de datos

Haciendo click en la cabecera de cada fila es posible cambiar el nombre de la variable y su tipo. Los nombres de variables deben comenzar con una letra o un punto y pueden contener cualquier letra, punto, subrayado (_) o número. En particular, no se pueden utilizar espacios en blanco. Además, R es distingue entre mayúsculas y minúsculas.

R permite definir más de un conjunto de datos en una misma sesión de trabajo, pero sólo puede haber uno activo en cada momento. Para cambiar el conjunto de datos activo se utiliza el menú `Datos→Conjunto de datos activo→Seleccionar conjunto de datos activo`.

Otra forma habitual de introducir datos, que utilizaremos a menudo en las prácticas, es por medio de una hoja de cálculo de Excel. Para ello basta con introducir los datos en la hoja de cálculo siguiendo los mismos criterios de antes. Es conveniente poner los nombres de las variables en la primera fila de la hoja. Una vez guardada la hoja de cálculo es posible cargarla en R como se indica en la sección de abrir datos.

5.3 Guardar datos

Una vez introducidos los datos, conviene guardarlos en un fichero para no tener que volver a introducirlos en futuras sesiones. Para guardar el conjunto de datos activo en un fichero, se utiliza el menú `Datos→Conjunto de datos activo→Guardar el conjunto de datos activo`. Con esto aparece una ventana donde hay que darle un nombre al fichero y seleccionar la carpeta donde se guardará. Los conjuntos de datos se guardan siempre en ficheros de R con extensión `rda`.

También es posible guardar los datos en un fichero de texto plano mediante el menú `Datos→Conjunto de datos activo→Exportar el conjunto de datos activo`. Tras esto aparece una venta donde se puede indicar entre otras cosas el separador de los datos, que puede ser un espacio, tabuladores, comas u otro caracter, y tras aceptar aparece otra venta donde hay que darle un nombre al fichero de texto y seleccionar la carpeta donde se guardará.

5.4 Abrir datos

Si los datos con los que se pretende trabajar ya están guardados en un fichero de R, entonces tendremos que abrir dicho fichero. Para ello se utiliza el `Datos→Cargar conjunto de datos` y en la ventana que

aparece se selecciona el fichero que se desea abrir. Automáticamente se cargará el conjunto de datos del fichero y pasará a ser el conjunto de datos activo.

También es posible cargar datos de ficheros con otros formatos, como por ejemplo una hoja de cálculo Excel. Para ello se utiliza el menú `Datos→Importar datos→desde Excel` y en la ventana que aparece se selecciona el fichero con la hoja de cálculo Excel que se desea abrir. Si el fichero sólo contiene una hoja se cargará automáticamente, mientras que si tiene varias, aparecerá una ventana en la que habrá que seleccionar la hoja que se quiere abrir.

5.5 Modificación del conjunto de datos

A menudo en los análisis hay que realizar transformaciones en los datos originales. A continuación se presentan las transformaciones más habituales.

5.6 Eliminación de datos

Para eliminar una variable del conjunto de datos se utiliza el menú `Datos→Modificar variables del conjunto de datos activo→Eliminar variables del conjunto de datos` y en el cuadro de diálogo que aparece basta con seleccionar la variable del conjunto de datos activo que se quiere eliminar.

Para eliminar individuos del conjunto de datos se utiliza el menú `Datos→Conjunto de datos activo →Borrar fila(s) del conjunto de datos activo` y en el cuadro de diálogo que aparece hay que indicar los números de las filas que se desean eliminar.

5.7 Filtrado de datos

Cuando se desea realizar un análisis con un subconjunto de individuos del conjunto de datos activo que cumplen una determinada condición, o bien con un subconjunto de variables, es posible filtrar el conjunto de datos activo para quedarse con esos individuos y esas variables. Para ello se utiliza el menú `Datos→Conjunto de datos activo→Filtrar el conjunto de datos activo`. Con esto aparece un cuadro de diálogo en el que hay que indicar las variables que se desean y en el cuadro `Expresión de selección` indicar la condición lógica que tienen que cumplir los individuos seleccionados. También se puede indicar el nombre del nuevo conjunto de datos ya que si se omite se sobreescribirá el conjunto de datos activo. Por ejemplo, para seleccionar los alumnos del grupo de la mañana habría que indicar la condición `grupo==''m''` tal y como se muestra en la figura 1.3.

Figura 1.3 – Ventana de filtrado de datos.

5.8 Cálculo de variables

Para calcular una nueva variable a partir de las ya existentes en el conjunto de datos activo se utiliza el menú `Datos→Modificar variables del conjunto de datos activo→Calcular una nueva`

`variable`. Con esto aparece un cuadro de diálogo en la que hay que indicar el nombre de la nueva variable y la expresión a partir de la que se calculará la nueva variable. Aquí puede introducirse cualquier expresión aritmética o lógica de R, y dentro de las expresiones puede utilizarse cualquiera de las variables del conjunto de datos activo. Por ejemplo, para eliminar los decimales de la variable nota podría crearse una nueva variable puntuacion multiplicando por 10 las notas, tal y como se muestra en la figura 1.4.

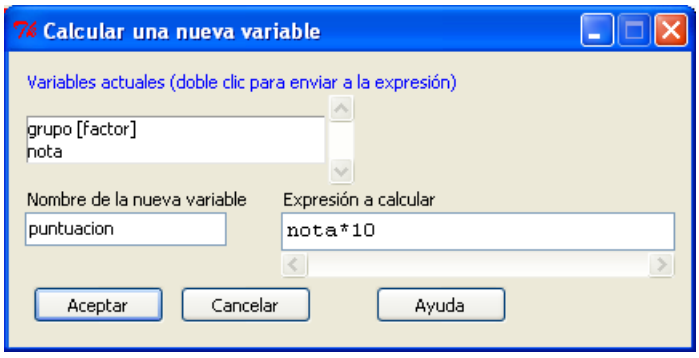


Figura 1.4 – Ventana de cálculo de nuevas variables.

5.9 Recodificación de variables

Otra transformación habitual es la recodificación de variables que permite transformar los valores de una variable de acuerdo a un conjunto de reglas de reescritura. Normalmente se utiliza para convertir una variable numérica en una variable categórica que pueda usarse como un factor.

Para recodificar una variable se utiliza el menú `Datos→Modificar variables del conjunto de datos activo→Recodificar variable`. Con esto aparece una ventana en la que hay que seleccionar la variable que se desea recodificar, introducir el nombre de la nueva variable recodificada e introducir las reglas de recodificación en el cuadro de texto `Introducir directrices de recodificación`. Las reglas de recodificación siempre siguen la sintaxis `valor o rango de valores = nuevo valor` y pueden introducirse tantas reglas como se desee, cada una en una línea. Al lado izquierdo de la igualdad puede introducirse un único valor, varios valores separados por comas, o un rango de valores indicando el límite inferior y el límite superior del intervalo separados por el operador `:`. A la hora de definir el límite inferior puede utilizarse la palabra clave `lo` para referirse al menor de los valores de la muestra y `hi` para referirse al mayor de los valores. Por ejemplo, para recodificar la variable nota en categorías correspondientes a las calificaciones ([0-5) Suspenso, [5,7) Aprobado, [7,9) Notable y [9,10] Sobresaliente), habría que introducir las reglas que se muestran en la figura 1.5.

6 Manipulación de ficheros de resultados e instrucciones

6.1 Guardar los resultados

Para guardar el contenido de la ventana de resultados en un fichero de texto plano se utiliza el menú `Fichero→Guardar los resultados`. Con esto aparece un cuadro de diálogo en el que hay que indicar el nombre del fichero y la carpeta donde se desea guardar.

También es posible copiar los resultados para después pegarlos en un procesador de texto como Word. Existen también paquetes como por ejemplo `R2HTML` o `xtable` que permite formatear la salida en LaTeXo HTML.

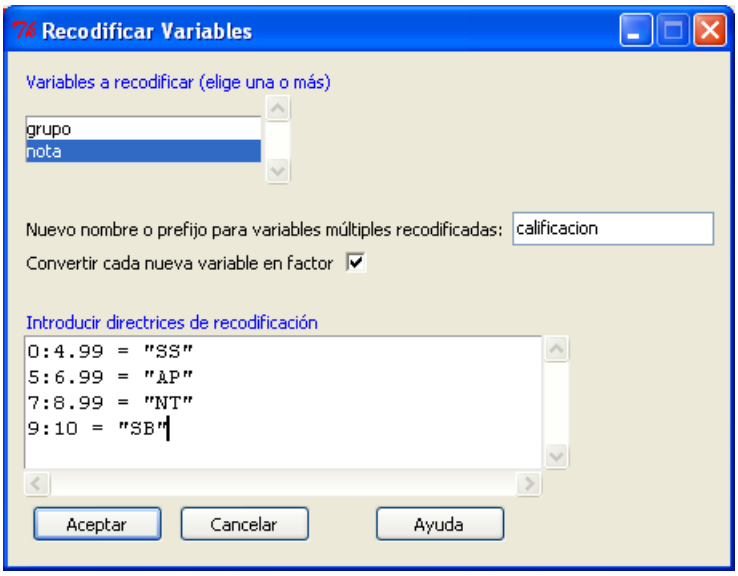


Figura 1.5 – Ventana de recodificación de variables

6.2 Guardar las instrucciones

Las instrucciones de la ventana de instrucciones también pueden guardarse en un fichero de texto plano mediante el menú `Fichero`→`Guardar las instrucciones` e indicando el nombre del fichero y la carpeta donde se guardará en el cuadro de diálo que aparece. Esto opción es muy interesante de cara a repetir análisis o automatizar tareas repetitivas.

6.3 Abrir un fichero de instrucciones

Para abrir un fichero de texto con instrucciones de R se utiliza el menú `Fichero`→`Abrir fichero de instrucciones` y después seleccionar el fichero que se desea abrir en el cuadro de diálogo que aparece. Con esto las instrucciones del fichero se aparecen en la ventana de instrucciones y pueden seleccionarse y ejecutarse.

6.4 Guardar el entorno de trabajo

Durante cualquier sesión de trabajo, R guarda en memoria las variables o los modelos que se definan. Cuando cierra el programa esta información se pierde, de manera que si se desea continuar con un análisis conviene guardar el contenido del entorno de trabajo antes de acabar la sesión de trabajo. Para ello se utiliza el menú `Fichero`→`Guardar el entorno de trabajo de R` y como siempre hay que indicar el nombre del fichero y la carpeta donde se guardará.

6.5 Cargar un entorno de trabajo

Cuando se inice una nueva sesión de trabajo R cargará automáticamente en el entorno de trabajo cualquier fichero con extensión `.RData` que se encuentre en el directorio de trabajo. Para cambiar de directorio, y por tanto de entorno de trabajo, se utiliza el menú `Fichero`→`Cambiar directorio de trabajo` y se selecciona el nuevo directorio en el cuadro de diálogo que aparece.

7 Extensión de Rcommander mediante plugins

Una de las ventajas de Rcommander frente a otras interfaces gráficas de R, es que permite incorporar menús con nuevas funcionalidades mediante un sistema de plugins.

Para realizar algunos de los ejercicios que se plantean a lo largo de estas prácticas es necesario instalar los paquetes TeachingExtras y RcmdrPlugin.TeachingExtras en R. Para ello hay que descargarse estos paquetes y luego, en la ventana de R, seleccionar el menú `Paquetes→Instalar paquetes(s) a partir de archivos zip locales`, y en el cuadro de diálogo que aparece seleccionar el fichero TeachingExtras.zip y hacer click en el botón Aceptar. Esto mismo debe repetirse para el paquete `RcmdrPlugin.TeachinExtras.zip`. Finalmente, en la ventana de Rcommander debe seleccionarse el menú `Herramientas→Cargar plugin(s) de Rcmdr`, seleccionar el plugin RcmdrPlugin.TeachinExtras y hacer click sobre el botón Aceptar.

8 Ayuda

Otra de las ventajas de R es que tiene un sistema de ayuda muy documentado. Es posible conseguir ayuda sobre cualquier función, prodecimiento o paquete simplemente tecleando el comando `help()`. Por ejemplo, para obtener ayuda sobre el comando `mean` se tecletaría

```
> help("mean")
```

y con esto aparecerá una ventana de ayuda donde se describe la función y también aparecen ejemplos que ilustran su uso. Si no se conoce exactamente el nombre de la función o comando, se puede hacer una búsqueda aproximada con el comando `help.search()`. Por emplo, si no se recuerda el nombre de la función logarítmica, se podría teclear

```
> help("logarithm")
```

y con esto aparecerá una ventana con todos los ficheros de ayuda que contienen la palabra logarithm.

Finalmente, también es posible invocar la ayuda general de R con el menú `Ayuda→Star R help system` con lo que aparecerá una página web desde donde podremos navegar a la información deseada.

Para más información sobre R se recomienda visitar la página http://www.r-project.org/, y para más información sobre R-commander se recomienda visitar la página http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/. También puede encontrarse un buen manual de introducción a R-commander en el menú de ayuda de la propia interfaz gráfica de usuario.

9 Ejercicios resueltos

1. Crear un conjunto de datos con los datos de la siguiente muestra y guardarlo con el nombre `colesterol.rda`

Nombre	Sexo	Peso	Altura	Colesterol
José Luis Martínez Izquierdo	H	85	179	182
Rosa Díaz Díaz	M	65	173	232
Javier García Sánchez	H	71	181	191
Carmen López Pinzón	M	65	170	200
Marisa López Collado	M	51	158	148
Antonio Ruiz Cruz	H	66	174	249

Indicación
Para crear el conjunto de datos:

a) Seleccionar el menú `Datos→Nuevo conjunto de datos`.

b) En el cuadro de diálogo que aparece introducir el nombre del conjunto de datos y hacer click en el botón Aceptar.

c) En la ventana del editor de datos hay que definir una variable en cada columna haciendo click sobre la cabecera de la columna. Con ello aparece un cuadro de diálogo en el que hay que introducir el nombre de la variable, seleccionar el tipo de variable y pulsar la tecla `Intro`.

d) Una vez definidas las variables hay que introducir los datos de cada variable en la columna correspondiente.

e) Una vez introducidos los datos hay que cerrar la ventana del editor de datos.

Para guardar los datos:

a) Selecionar el menú `Datos→Conjunto de datos activo→Guardar el conjunto de datos activo`.

b) En el cuadro de diálogo que aparece hay que darle un nombre al fichero, seleccionar la carpeta donde guardarlo y hacer click en el botón Aceptar.

2. Abrir el fichero creado en el ejercicio anterior y realizar las siguientes operaciones:

a) Insertar una nueva variable Edad con las edades de todos los individuos de la muestra.

Nombre	Edad
José Luis Martínez Izquierdo	18
Rosa Díaz Díaz	32
Javier García Sánchez	24
Carmen López Pinzón	35
Marisa López Collado	46
Antonio Ruiz Cruz	68

Indicación
Para abrir el conjunto de datos del ejercicio anterior:

1) Seleccionar el menú `Datos→Cargar conjunto de datos`.
2) En el cuadro de diálogo que aparece seleccionar la carpeta donde se encuentra el fichero con los datos del ejercicio anterior, seleccionar el fichero y hacer click en el botón Aceptar.

Para insertar la variable Edad:

1) Hacer click en el botón Editar conjunto de datos.
2) En la ventana del editor de datos hacer doble click sobre la cabecera de la primera columna vacía.
3) En el cuadro de diálogo que aparece introducir el nombre de la variable Edad, seleccionar como tipo `numeric` y pulsar la tecla `Intro`.
4) En la ventana del editor de datos introducir los datos de las edades en la columna correspondiente y cerrar la ventana del editor de datos.

b) Insertar un nuevo individuo con siguientes datos

Nombre: Cristóbal Campos Ruiz.
Edad: 44 años.
Sexo: Hombre.

Peso: 70 Kg.
Altura: 178 cm.
Colesterol: 220 mg/dl.

Indicación

1) Hacer click en el botón Editar conjunto de datos.
2) En la ventana del editor de datos introducir los datos de del nuevo individuo en la primera fila vacía y cerrar la ventana del editor de datos.

c) Crear una nueva variable donde se calcule el índice de masa corporal de cada paciente mediante la formula:

$$\text{imc} = \frac{\text{Peso (en Kg)}}{\text{Altura (en mt)}^2}$$

Indicación

1) Seleccionar el menú `Datos→Modificar variables del conjunto de datos activo→Calcular una nueva variable`.
2) En el cuadro de diálogo que aparece darle un nombre a la nueva variable, introducir la fórmula para calcular el índice de masa corporal en el campo `Expresión a calcular` y hacer click sobre el botón Aceptar.

d) Recodificar el índice de masa corporal de acuerdo a las siguientes categorías:

Menor de 18,5	Bajo peso
De 18,5 a 24,5	Saludable
De 24,5 a 30	Sobrepeso
Mayor de 30	Obeso

Indicación

1) Selecionar el menú `Datos→Modificar variables del conjunto de datos activo→Recodificar variables`.
2) En el cuadro de diálogo que aparece darle un nombre a la nueva variable, introducir las reglas de recodificación en el campo `Introducir directrices de recodificación`:

```
lo:18.5 = ‘‘Bajo peso’’
18.5:24.5 = ‘‘Saludable’’
24.5:30 = ‘‘Sobrepeso’’
30:hi = ‘‘Obeso’’
```

y hacer click sobre el botón Aceptar.

Distribuciones de Frecuencias y Representaciones Gráficas

1 Fundamentos teóricos

Uno de los primeros pasos en cualquier estudio estadístico es el resumen y la descripción de la información contenida en una muestra. Para ello se van a aplicar algunos métodos de análisis descriptivo, que nos permitirán clasificar y estructurar la información al igual que representarla gráficamente.

Las características que estudiamos pueden ser o no susceptibles de medida; en este sentido definiremos una *variable* como un carácter susceptible de ser medido, es decir, cuantitativo y cuantificable mediante la observación, (por ejemplo el peso de las personas, la edad, etc...), y definiremos un *atributo* como un carácter no susceptible de ser medido, y en consecuencia observable tan sólo cualitativamente (por ejemplo el color de ojos, estado de un paciente, etc...). Se llaman modalidades a las posibles observaciones de un atributo.

Dentro de los atributos, podemos hablar de *atributos ordinales*, los que presentan algún tipo de orden entre las distintas modalidades, y de *atributos nominales*, en los que no existe ningún orden entre ellas.

Dentro de las variables podemos diferenciar entre *discretas*, si sus valores posibles son valores aislados, y *continuas*, si pueden tomar cualquier valor dentro de un intervalo.

En algunos textos no se emplea el término *atributo* y se denominan a todos los caracteres *variables*. En ese caso se distinguen *variables cuantitativas* para designar las que aquí hemos definido como *variables*, y *variables cualitativas* para las que aquí se han llamado *atributos*. En lo sucesivo se aplicará este criterio para simplificar la exposición.

1.1 Cálculo de Frecuencias

Para estudiar cualquier característica, lo primero que deberemos hacer es un recuento de las observaciones, y el número de repeticiones de éstas. Para cada valor x_i de la muestra se define:

Frecuencia absoluta Es el número de veces que aparece cada uno de los valores x_i y se denota por n_i.

Frecuencia relativa Es el número de veces que aparece cada valor x_i dividido entre el tamaño muestral y se denota por f_i

$$f_i = \frac{n_i}{n}$$

Generalmente las frecuencias relativas se multiplican por 100 para que representen el tanto por ciento.

En el caso de que exista un orden entre los valores de la variable, a veces nos interesa no sólo conocer el número de veces que se repite un determinado valor, sino también el número de veces que aparece dicho valor y todos los menores. A este tipo de frecuencias se le denomina *frecuencias acumuladas*.

Frecuencia absoluta acumulada Es la suma de las frecuencias absolutas de los valores menores que x_i más la frecuencia absoluta de x_i, y se denota por N_i

$$N_i = n_1 + n_2 + \ldots + n_i$$

Frecuencia relativa acumulada Es la suma de las frecuencias relativas de los valores menores que x_i más la frecuencia relativa de x_i, y se denota por F_i

$$F_i = f_1 + f_2 + \ldots + f_i$$

Los resultados de las observaciones de los valores de una variable estadística en una muestra suelen representarse en forma de tabla. En la primera columna se representan los valores x_i de la variable colocados en orden creciente, y en la siguiente columna los valores de las frecuencias absolutas correspondientes n_i.

Podemos completar la tabla con otras columnas, correspondientes a las frecuencias relativas, f_i, y a las frecuencias acumuladas, N_i y F_i. Al conjunto de los valores de la variable observados en la muestra junto con sus frecuencias se le conoce como *distribución de frecuencias muestral*.

Ejemplo En una encuesta a 25 matrimonios, sobre el número de hijos que tienen, se obtienen los siguientes datos:

1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.

Los valores distintos de la variable son: 0, 1, 2, 3 y 4. Así la tabla será:

x_i	Recuento	n_i
0	II	2
1	IIII I	6
2	IIII IIII IIII	14
3	II	2
4	I	1

La distribución de las frecuencias quedaría:

x_i	n_i	f_i	N_i	F_i
0	2	0,08	2	0,08
1	6	0,24	8	0,32
2	14	0,56	22	0,88
3	2	0,08	24	0,96
4	1	0,04	25	1
Suma	25	1		

Cuando el tamaño de la muestra es grande en el caso de variables discretas con muchos valores distintos de la variable, y en cualquier caso si se trata de variables continuas, se agrupan las observaciones en *clases*, que son intervalos contiguos, preferiblemente de la misma amplitud.

Para decidir el número de clases a considerar, una regla frecuentemente utilizada es tomar el entero más próximo a $\sqrt{n}$ donde n es el número de observaciones en la muestra. Pero conviene probar con distintos números de clases y escoger el que proporcione una descripción más clara. Así se prefijan los intervalos $(a_{i-1}, a_i], i = 1, 2, \ldots, l$ siendo $a = a_0 < a_1 < .... < a_l = b$ de tal modo que todos los valores observados estén dentro del intervalo $(a, b]$, y sin que exista ambigüedad a la hora de decidir a qué intervalo pertenece cada dato.

Llamaremos *marca de clase* al punto medio de cada intervalo. Así la *marca de la clase* $(a_{i-1}, a_i]$ es el punto medio x_i de dicha clase, es decir

$$x_i = \frac{a_{i-1} + a_i}{2}$$

En el tratamiento estadístico de los datos agrupados, todos los valores que están en una misma clase se consideran iguales a la marca de la clase. De esta manera si en la clase $(a_{i-1}, a_i]$ hay n_i valores observados, se puede asociar la marca de la clase x_i con esta frecuencia n_i.

1.2 Representaciones Gráficas

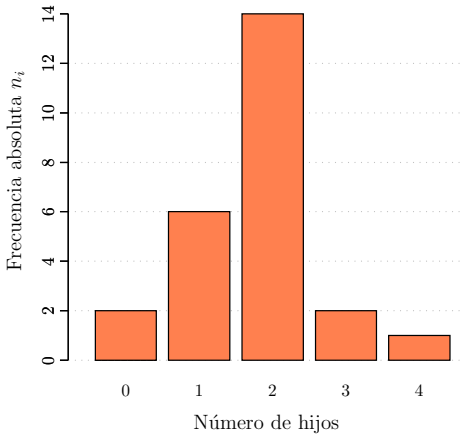
Hemos visto que la tabla estadística resume los datos de una muestra, de forma que ésta se puede analizar de una manera más sistemática y resumida. Para conseguir una percepción visual de las características de la población resulta muy útil el uso de gráficas y diagramas. Dependiendo del tipo de variable y de si trabajamos con datos agrupados o no, se utilizarán distintos tipos.

Diagrama de barras y polígono de frecuencias

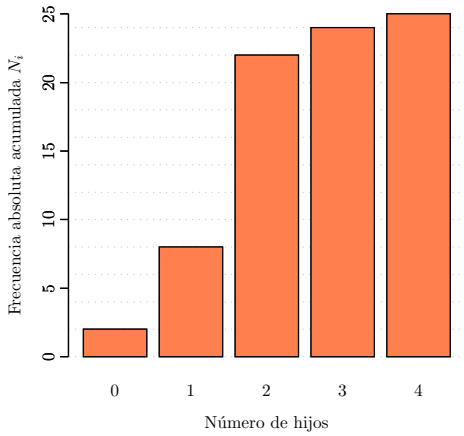
Consiste en representar sobre el eje de abscisas de un sistema de ejes coordenados los distintos valores de la variable X, y levantar sobre cada uno de esos puntos una barra cuya altura sea igual a la frecuencia absoluta o relativa correspondiente a ese valor, tal y como se muestra en la figura 2.1(a). Esta representación se utiliza para distribuciones de frecuencias con pocos valores distintos de la variable, tanto cuantitativas como cualitativas, y en este último caso se suele representar con rectángulos de altura igual a la frecuencia de cada modalidad.

En el caso de variables cuantitativas se puede representar también el diagrama de barras de las frecuencias acumuladas, tal y como se muestra en la figura 2.1(b).

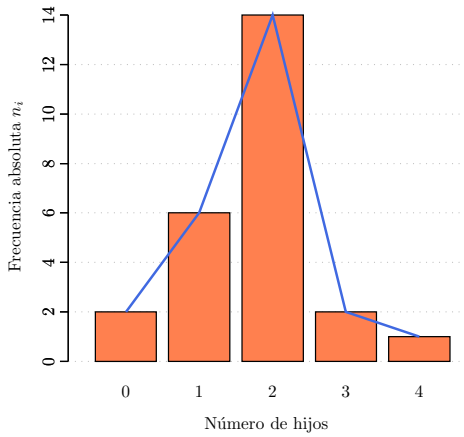
Otra representación habitual es el *polígono de frecuencias* que consiste en la línea poligonal cuyos vertices son los puntos (x_i, n_i), tal y como se ve en la figura 2.1(c), y si en vez de considerar las frecuencias absolutas o relativas se consideran las absolutas o relativas acumuladas, se obtiene el *polígono de frecuencias acumuladas*, como se ve en la figura 2.1(d).



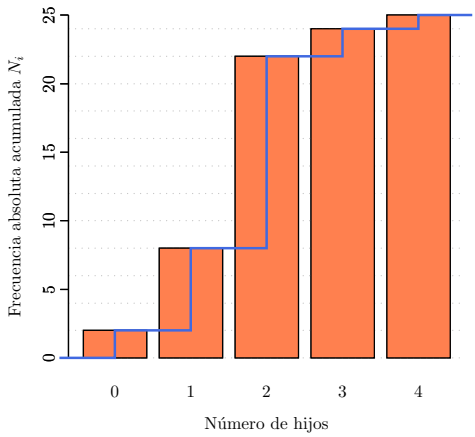
(a) Diagrama de barras de frecuencias absolutas.



(b) Diagrama de barras de frecuencias absolutas acumuladas.



(c) Polígono de frecuencias absolutas.



(d) Polígono de frecuencias absolutas acumuladas

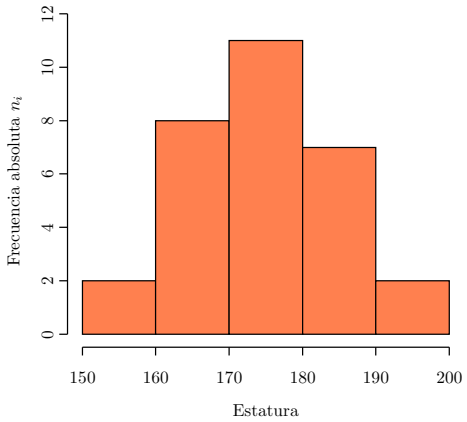
Figura 2.1 – Diagramas de barras y polígonos asociados para datos no agrupados.

Histogramas

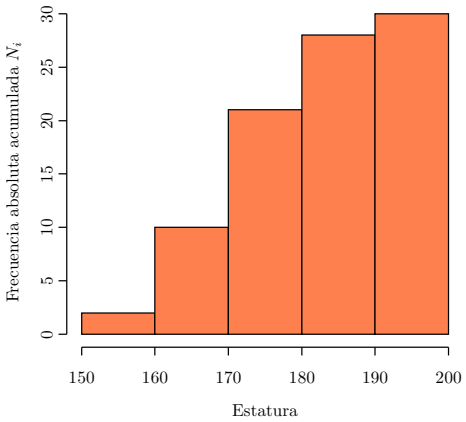
Este tipo de representaciones se utiliza en variables continuas y en variables discretas en que se ha realizado una agrupación de las observaciones en clases. Un *histograma* es un conjunto de rectángulos, cuyas bases son los intervalos de clase $(a_{i-1}, a_i]$ sobre el eje OX y su altura la correspondiente frecuencia absoluta , relativa, absoluta acumulada, o relativa acumulada, tal y como se muestra en la figuras 2.2(a) y 2.2(b).

Si unimos los puntos medios de las bases superiores de los rectángulos del histograma, se obtiene el *polígono de frecuencias* correspondiente a datos agrupados (figura 2.2(c)).

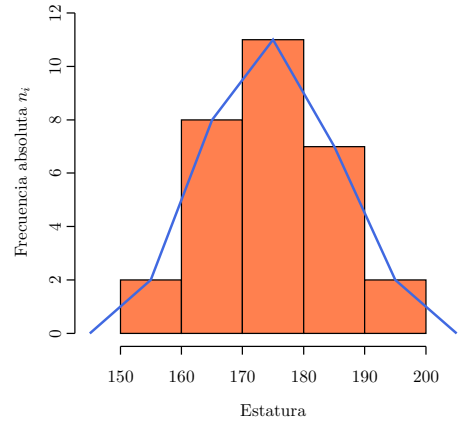
El polígono de frecuencias también se puede utilizar para representar las frecuencias acumuladas, tanto absolutas como relativas. En este caso la línea poligonal se traza uniendo los extremos derechos de las bases superiores de los rectángulos del histograma de frecuencias acumuladas, en lugar de los puntos centrales (figura 2.2(d)).



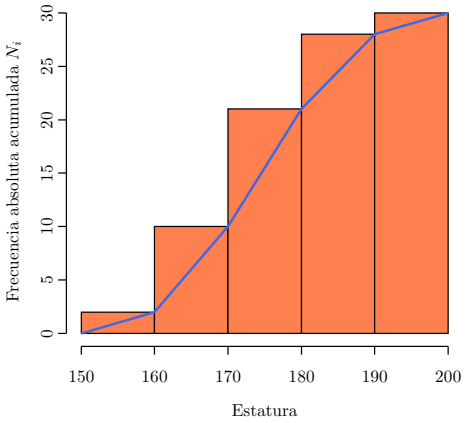
(a) Histograma de frecuencias absolutas.



(b) Histograma de frecuencias absolutas acumuladas.



(c) Polígono de frecuencias absolutas.



(d) Polígono de frecuencias absolutas acumuladas

Figura 2.2 – Histograma y polígonos asociados para datos agrupados.

Para variables cualitativas y cuantitativas discretas también se pueden usar las superficies representativas; de éstas, las más empleadas son los *sectores circulares*.

Sectores circulares o diagrama de sectores

Es una representación en la que un círculo se divide en sectores, de forma que los ángulos, y por tanto las áreas respectivas, sean proporcionales a la frecuencia.

Ejemplo Se está haciendo un estudio en una población del grupo sanguíneo de sus ciudadanos. Para ello disponemos de una muestra de 30 personas, con los siguientes resultados: 5 personas con grupo 0, 14 con grupo A, 8 con grupo B y 3 con grupo AB. El el diagrama de sectores de frecuencias relativas correspondiente aparece en la figura 2.3.

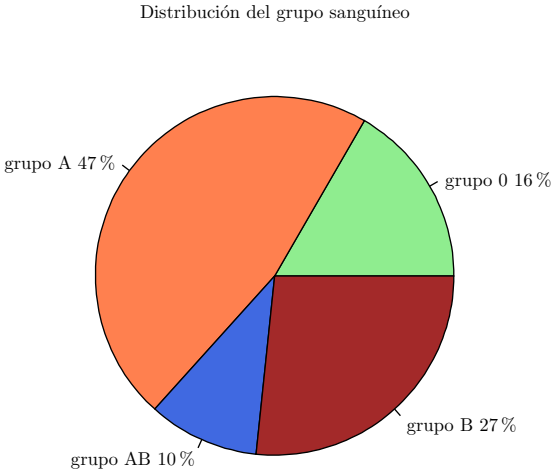


Figura 2.3 – Diagrama de sectores de frecuencias relativas del grupo sanguíneo.

Diagrama de cajas y datos atípicos

Los datos extremadamente altos o bajos, en comparación con los del resto de la muestra, reciben el nombre de datos influyentes o *datos atípicos*. Tales datos que, como su propio nombre indica, pueden modificar las conclusiones de un estudio, deben ser considerados atentamente antes de aceptarlos, pues no pocas veces podrán ser, simplemente, datos erróneos. La representación gráfica más apropiada para detectar estos datos es el *diagrama de cajas*. Este diagrama está formado por una caja que contiene el 50 % de los datos centrales de la distribución, y unos segmentos que salen de la caja, que indican los límites a partir de los cuales los datos se consideran atípicos. En la figura 2.4 se puede observar un ejemplo en el que aparecen dos datos atípicos.

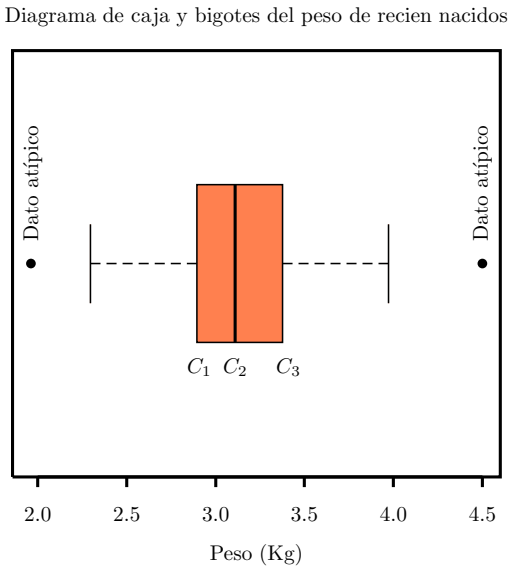


Figura 2.4 – Diagrama de cajas para una muestra de recién nacidos. Existen dos niños con pesos atípicos, uno con peso extremadamente bajo 1,9 kg, y otro con peso extremadamente alto 4,3 kg.

2 Ejercicios resueltos

1. En una encuesta a 25 matrimonios sobre el número de hijos que tenían se obtuvieron los siguientes datos:

   1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

   Se pide:

   a) Crear un conjunto de datos con la variable hijos e introducir los datos.

   b) Construir la tabla de frecuencias.

   Indicación
   1) Seleccionar el menú `Estadísticos→Distribuciones de frecuencias→Tabla de frecuencias (datos numéricos no agrupados)` .
   2) En el cuadro de diálogo que aparece, seleccionar la variable hijos y hacer click en el botón Aceptar.

   c) Dibujar el diagrama de barras de las frecuencias absolutas.

   Indicación
   1) Seleccionar el menú `Gráficas→Gráfica de barras`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable hijos y hacer click en el botón Aceptar.

   d) Para la misma tabla de frecuencias anterior, dibujar también el diagrama de barras de las frecuencias relativas, el de absolutas acumuladas y el de relativas acumuladas, además de sus correspondientes polígonos.

   Indicación
   Repetir los pasos del apartado anterior activando la opción `Frecuencias relativas` si se desea el diagrama de barras de frecuencias relativas, activando la opción `Frecuencias acumuladas` si se desea el diagrama de barras de frecuencias acumuladas y activando la opción `Polígono` para obtener el polígono asociado.

2. En un hospital se realizó un estudio sobre el número de personas que ingresaron en urgencias cada día del mes de noviembre. Los datos observados fueron:

   15, 23, 12, 10, 28, 50, 12, 17, 20, 21, 18, 13, 11, 12, 26
   30, 6, 16, 19, 22, 14, 17, 21, 28, 9, 16, 13, 11, 16, 20

   Se pide:

   a) Crear un conjunto de datos con la variable urgencias e introducir los datos.

   b) Dibujar el diagrama de cajas. ¿Existe algún dato atípico? En el caso de que exista, eliminarlo y proceder con los siguientes apartados.

   Indicación
   1) Seleccionar el menú `Gráficas→Diagrama de cajas`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable urgencias, marcar la opción `Identificar atípicos con el ratón` y hacer click sobre el botón Aceptar.
   3) En la ventana que aparece con el diagrama de barras hacer click sobre el dato atípico para identificarlo.
   4) Seleccionar el menú `Datos→Conjunto de datos activo→Borrar fila(s) del conjunto de datos`.
   5) En el cuadro de diálogo que aparece introducir el índice del individuo atípico en el campo `Índices o nombres de la(s) fila(s) para borrar` y hacer click en el botón Aceptar.

   c) Construir la tabla de frecuencias agrupando en 5 clases.

   Indicación
   1) Seleccionar el menú `Estadísticos->Distribuciones de frecuencias->Tabla de frecuencias (datos numéricos agrupados)`.
   2) En el cuadro de diálogo que aparece seleccionar la variable urgencias, marcar la opción `Nº de intervalos`, introducir el número deseado de intervalos en el campo `Intervalos` y hacer click sobre el botón Aceptar

d) Dibujar el histograma de frecuencias absolutas correspondiente a la tabla anterior.

Indicación
1) Seleccionar el menú `Gráficas→Histograma`.
2) En el cuadro de diálogo que aparece seleccionar la variable urgencias, marcar la opción `Nº de intervalos`, introducir el número deseado de intervalos en el campo `Intervalos`, poner el título deseado en el campo `Título` y hacer click sobre el botón Aceptar.

e) Para la misma tabla de frecuencias anterior, dibujar también el histograma de las frecuencias relativas, el de absolutas acumuladas y el de relativas acumuladas, además de sus correspondientes polígonos.

Indicación
Repetir los pasos del apartado anterior activando la opción `Frecuencias relativas` si se desea el histograma de frecuencias relativas, activando la opción `Frecuencias acumuladas` si se desea el histograma de frecuencias acumuladas y activando la opción `Polígono` para obtener el polígono asociado.

3. Los grupos sanguíneos de una muestra de 30 personas son:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, A, AB,
A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

Se pide:

a) Crear un conjunto de datos con la variable grupo_sanguineo e introducir los datos.

b) Construir la tabla de frecuencias.

Indicación
1) Seleccionar el menú `Estadísticos->Distribuciones de frecuencias->Tabla de frecuencias (datos categóricos)`.
2) En el cuadro de diálogo que aparece seleccionar la variable grupo_sanguineo y hacer click sobre el botón Aceptar

c) Dibujar el diagrama de sectores.

Indicación
1) Seleccionar el menú `Gráficas→Gráfica de sectores`.
2) En el cuadro de diálogo que aparece, seleccionar la variable grupo_sanguineo y hacer click sobre el botón Aceptar.

4. En un estudio de población se tomó una muestra de 27 personas, y se les preguntó por su edad y estado civil, obteniendo los siguientes resultados:

Estado civil	Edad								
Soltero	31	45	35	65	21	38	62	22	31
Casado	62	39	62	59	21	62			
Viudo	80	68	65	40	78	69	75		
Divorciado	31	65	59	49	65				

Se pide:

a) Crear un conjunto de datos con la variables estado_civil y edad e introducir los datos.

b) Dibujar los diagramas de cajas de la edad según el estado civil. ¿Existen datos atípicos? ¿En qué grupo hay mayor dispersión?

Indicación
1) Seleccionar el menú `Gráficas→Diagrama de cajas`.
2) En el cuadro de diálogo que aparece, seleccionar la variable edad, marcar la opción `Identificar atípicos con el ratón` y hacer click sobre el botón Gráfica por grupos.
3) En el cuadro de diálogo que aparece, seleccionar la variable estado_civil y hacer click sobre el botón Aceptar.
4) En la ventana que aparece con los diagramas de barras hacer click sobre el dato atípico para identificarlo.

3 Ejercicios propuestos

1. El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

   0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1

   Se pide:

   a) Construir la tabla de frecuencias.

   b) Dibujar el diagrama de barras de las frecuencias relativas y de frecuencias relativas acumuladas.

   c) Dibujar el diagrama de sectores.

2. Para realizar un estudio sobre la estatura de los estudiantes universitarios, seleccionamos, mediante un proceso de muestreo aleatorio, una muestra de 30 estudiantes, obteniendo los siguientes resultados (medidos en centímetros):

   179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
   162, 187, 198, 177, 178, 165, 154, 188, 166, 171,
   175, 182, 167, 169, 172, 186, 172, 176, 168, 187.

   Se pide:

   a) Dibujar el histograma de las frecuencias absolutas agrupando desde 150 a 200 en clases de amplitud 10.

   b) Dibujar el diagrama de cajas. ¿Existe algún dato atípico?.

Práctica de Estadística con R 3

Estadísticos Muestrales

1 Fundamentos teóricos

Hemos visto cómo podemos presentar la información que obtenemos de la muestra, a través de tablas o bien a través de gráficas. La tabla de frecuencias contiene toda la información de la muestra pero resulta difícil sacar conclusiones sobre determinados aspectos de la distribución con sólo mirarla. Ahora veremos cómo a partir de esos mismos valores observados de la variable estadística, se calculan ciertos números que resumen la información muestral. Estos números, llamados *Estadísticos*, se utilizan para poner de manifiesto ciertos aspectos de la distribución, tales como la dispersión o concentración de los datos, la forma de su distribución, etc. Según sea la característica que pretenden reflejar se pueden clasificar en Medidas de posiPara esta práctica es necesario instalar los paquetes `TeachingExtras` y `RcmdrPlugin.TeachingExtras` en R.

Indicación

1. Descargar los paquetes `TeachingExtras.zip` y `RcmdrPlugin.TeachingExtras.zip` desde Moodle.
2. En la ventana de R seleccionar el menú `Paquetes→Instalar paquetes(s) a partir de archivos zip locales`.
3. En el cuadro de diálogo que aparece seleccionar los ficheros de los paquetes `TeachingExtras.zip` y `RcmdrPlugin.TeachingExtras.zip` y hacer click en el botón Aceptar.
4. En la ventana de Rcommander seleccionar el menúCargar plugin(s) de Rcmdr.
5. En el cuadro de diálogo que aparece seleccionar el plugin `RcmdrPlugin.TeachinExtras` y hacer click sobre el botón Aceptar.

ción, Medidas de dispersión y Medidas de forma.

1.1 Medidas de posición

Son valores que indican cómo se sitúan los datos. Los más importantes son la Media aritmética, la Mediana y la Moda.

Media aritmética $\overline{x}$

Se llama *media aritmética* de una variable estadística X, y se representa por $\overline{x}$, a la suma de todos los resultados observados, dividida por el tamaño muestral. Es decir, la media de la variable estadística X, cuya distribución de frecuencias es (x_i, n_i), viene dada por

$$\overline{x} = \frac{x_1 + \ldots + x_1 + \ldots + x_k + \ldots + x_k}{n_1 + \ldots + n_k} = \frac{x_1 n_1 + \ldots + x_k n_k}{n} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

La media aritmética sólo tiene sentido en variables cuantitativas.

Mediana *Me*

Se llama *mediana* y lo denotamos por Me, a aquel valor de la muestra que, una vez ordenados todos los valores de la misma en orden creciente, tiene tantos términos inferiores a él como superiores. En consecuencia, divide la distribución en dos partes iguales.

La mediana sólo tiene sentido en atributos ordinales y en variables cuantitativas.

Moda *Mo*

La *moda* es el valor de la variable que presenta una mayor frecuencia en la muestra. Cuando haya más de un valor con frecuencia máxima diremos que hay más de una moda. En variables continuas o discretas agrupadas llamaremos clase modal a la que tenga la máxima frecuencia. Se puede calcular la moda tanto en variables cuantitativas como cualitativas.

Cuantiles

Si el conjunto total de valores observados se divide en r partes que contengan cada una $\frac{n}{r}$ observaciones, los puntos de separación de las mismas reciben el nombre genérico de *cuantiles*.

Según esto la mediana también es un cuantil con $r = 2$. Algunos cuantiles reciben determinados nombres como:

Cuartiles. Son los puntos que dividen la distribución en 4 partes iguales y se designan por C_1, C_2, C_3. Es claro que $C_2 = Me$.

Deciles. Son los puntos que dividen la distribución en 10 partes iguales y se designan por $D_1, D_2, \ldots, D_9$.

Percentiles. Son los puntos que dividen la distribución en 100 partes iguales y se designan por $P_1, P_2, \ldots, P_{99}$.

1.2 Medidas de dispersión

Miden la separación existente entre los valores de la muestra. Las más importantes son el Rango o Recorrido, el Rango Intercuartílico, la Varianza, la Desviación Típica y el Coeficiente de Variación.

Rango o Recorrido *Re*

La medida de dispersión más inmediata es el rango. Llamamos *recorrido* o *rango* y lo designaremos por Re a la diferencia entre los valores máximo y mínimo que toma la variable en la muestra. Es decir

$$Re = max\{x_i, i = 1, 2, \ldots, n\} - min\{x_i, i = 1, 2, \ldots, n\}$$

Este estadístico sirve para medir el campo de variación de la variable, aunque es la medida de dispersión que menos información proporciona sobre la mayor o menor agrupación de los valores de la variable alrededor de las medidas de tendencia central. Además tiene el inconveniente de que se ve muy afectado por los datos atípicos.

Rango Intercuartílico *RI*

El *rango intercuartílico* RI es la diferencia entre el tercer y el primer cuartil, y mide, por tanto, el campo de variación del 50 % de los datos centrales de la distribución. Por consiguiente

$$RI = C_3 - C_1$$

La ventaja del rango intercuartílico frente al recorrido es que no se ve tan afectado por los datos atípicos.

Varianza s_x^2

Llamamos *varianza* de una variable estadística X, y la designaremos por s_x^2, a la media de los cuadrados de las desviaciones de los valores observados respecto de la media de la muestra. Así

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i$$

Desviación Típica s_x

La raíz cuadrada positiva de la varianza se conoce como *desviación típica* de la variable X, y se representa por s

$$s = +\sqrt{s_x^2}$$

Coeficiente de Variación de Pearson Cv_x

Al cociente entre la desviación típica y el valor absoluto de la media se le conoce como *coeficiente de variación de Pearson* o simplemente *coeficiente de variación*:

$$Cv_x = \frac{s_x}{|\overline{x}|}$$

El coeficiente de variación es adimensional, y por tanto permite hacer comparaciones entre variables expresadas en distintas unidades. Cuanto más próximo esté a 0, menor será la dispersión de la muestra en relación con la media, y más representativa será ésta última del conjunto de observaciones.

1.3 Medidas de forma

Indican la forma que tiene la distribución de valores en la muestra. Se pueden clasificar en dos grupos: Medidas de *asimetría* y medidas de *apuntamiento o curtosis*.

Coeficiente de asimetría de Fisher g_1

El *coeficiente de asimetría de Fisher*, que se representa por g_1, se define como

$$g_1 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^3 f_i}{s_x^3}$$

Dependiendo del valor que tome tendremos:

- $g_1 = 0$. Distribución simétrica.
- $g_1 < 0$. Distribución asimétrica hacia la izquierda.
- $g_1 > 0$. Distribución asimétrica hacia la derecha.

Coeficiente de apuntamiento o curtosis g_2

El grado de apuntamiento de las observaciones de la muestra, se caracteriza por el *coeficiente de apuntamiento o curtosis* y se representa por g_2

$$g_2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^4 f_i}{s_x^4} - 3$$

Dependiendo del valor que tome tendremos:

- $g_2 = 0$. La distribución tiene un apuntamiento igual que el de la distribución normal de la misma media y desviación típica. Se dice que es una distribución *mesocúrtica*.
- $g_2 < 0$. La distribución es menos apuntada que la distribución normal de la misma media y desviación típica. Se dice que es una distribución *platicúrtica*.
- $g_2 > 0$. La distribución es más apuntada que la distribución normal de la misma media y desviación típica. Se dice que es una distribución *leptocúrtica*.

Tanto g_1 como g_2 suelen utilizarse para comprobar si los datos muestrales provienen de una población no normal. Cuando g_1 está fuera del intervalo [-2,2] se dice que la distribución es demasiado asimétrica como para que los datos provengan de una población normal. Del mismo modo, cuando g_2 está fuera del intervalo [-2,2] se dice que la distribución es, o demasiado apuntada, o demasiado plana, como para que los datos provengan de una población normal.

1.4 Estadísticos de variables en las que se definen grupos

Ya sabemos cómo resumir la información contenida en una muestra utilizando una serie de estadísticos. Pero hasta ahora sólo hemos estudiado ejemplos con un único carácter objeto de estudio.

En la mayoría de las investigaciones no estudiaremos un único carácter, sino un conjunto de caracteres, y muchas veces será conveniente obtener información de un determinado carácter, en función de los grupos creados por otro de los caracteres estudiados en la investigación. A estas variables que se utilizan para formar grupos se les conoce como *variables clasificadoras* o *factores*.

Por ejemplo, si se realiza un estudio sobre un conjunto de niños recién nacidos, podemos estudiar su peso. Pero si además sabemos si la madre de cada niño es fumadora o no, podremos hacer un estudio del peso de los niños de las madres fumadoras por un lado y los de las no fumadoras por otro, para ver si existen diferencias entre ambos grupos.

2 Ejercicios resueltos

1. En una encuesta a 25 matrimonios sobre el número de hijos que tenían se obtuvieron los siguientes datos:

   1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

   Se pide:

   a) Crear un conjunto de datos con la variable hijos e introducir los datos. Si ya se tienen los datos, simplemente recuperarlos.

   b) Calcular la media aritmética, varianza y desviación típica de dicha variable. Interpretar los estadísticos.

   Indicación
   1) Seleccionar el menú `Estadísticos→Resúmenes→Resumen descriptivo`.
   2) En el cuadro de diálogo que aparece seleccionar la variable hijos y marcar las opciones `Media`, `Desviación típica`, y hacer click sobre el botón Aceptar.

   c) Calcular los cuartiles, el recorrido, el rango intercuartílico, el tercer decil y el percentil 68.

   Indicación
   Repetir los pasos del apartado anterior, pero activar sólo la opción `Cuantiles` y escribiendo en el campo `cuantiles` las frecuencias relativas de los cuantiles deseados.

2. En un hospital se realizó un estudio sobre el número de personas que ingresaron en urgencias cada día del mes de noviembre. Los datos observados fueron:

   15, 23, 12, 10, 28, 50, 12, 17, 20, 21, 18, 13, 11, 12, 26
   30, 6, 16, 19, 22, 14, 17, 21, 28, 9, 16, 13, 11, 16, 20

   Se pide:

   a) Crear un conjunto de datos con la variable urgencias e introducir los datos.

   b) Calcular la media aritmética, varianza, desviación típica y coeficiente de variación de dicha variable. Interpretar los estadísticos.

   Indicación
   1) Seleccionar el menú `Estadísticos→Resúmenes→Resumen descriptivo`.
   2) En el cuadro de diálogo que aparece seleccionar la variable urgencias y marcar las opciones `Media`, `Varianza`, `Desviación típica`, `Coeficiente de variación` y hacer click sobre el botón Aceptar.

   c) Calcular el coeficiente de asimetría y el de curtosis e interpretar los resultados

   Indicación
   Seguir los mismos pasos del apartado anterior, seleccionando ahora los estadísticos que se piden.

3. En un grupo de 20 alumnos, las calificaciones obtenidas en Matemáticas fueron:

   SS, AP, SS, AP, AP, NT, NT, AP, SB, SS
   SB, SS, AP, AP, NT, AP, SS, NT, SS, NT

   Se pide:

   a) Crear un conjunto de datos con la variable calificaciones e introducir los datos.

b) Recodificar esta variable, asignando 2,5 al SS, 5,5 al AP, 7,5 al NT y 9,5 al SB.

Indicación

1) Seleccionar el menú `Datos→Modificar variables del conjunto activo→Recodificar variable`.
2) En el cuadro de diálogo que aparece seleccionar la variable calificaciones, introducir como nombre de la variable recodificada nota y desmarcar la opción `Convertir cada nueva variable en factor`.
3) En el campo `Introducir directrices de recodificación` introducir las reglas de recodificación y hacer click en el botón Aceptar.

c) La mediana y el rango intercuartílico.

Indicación

1) Seleccionar el menú `Estadísticos→Resúmenes→Resumen descriptivo`.
2) En el cuadro de diálogo que aparece seleccionar la variable nota, marcar la opción `Cuantiles` y hacer click sobre el botón Aceptar.

4. Para realizar un estudio sobre la estatura de los estudiantes universitarios se ha seleccionado mediante un proceso de muestreo aleatorio, una muestra de 30 estudiantes, obteniendo los siguientes resultados (medidos en centímetros):

Mujeres:	173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168.
Hombres:	179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187.

Se pide:

a) Crear un conjunto de datos con las variables estatura y sexo e introducir los datos.

b) Obtener un resumen de estadísticos en el que se muestren la media aritmética, mediana, varianza, desviación típica y cuartiles según el sexo. Interpretar los estadísticos.

Indicación

1) Seleccionar el menú `Estadísticos→Resúmenes→Resumen descriptivo`.
2) En el cuadro de diálogo que aparece seleccionar la variable estatura y marcar las opciones `Media`, `Varianza`, `Desviación típica`, `Cuantiles` y hacer click sobre el botón Resumir por grupos.
3) En el cuadro de diálogo que aparece, seleccionar la variable sexo y hacer click sobre el botón Aceptar.

3 Ejercicios propuestos

1. El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1

Se pide:

a) Calcular la media aritmética, mediana, varianza y desviación típica de las lesiones e interpretarlas.

b) Calcular los coeficientes de asimetría y curtosis e interpretarlos.

c) Calcular el cuarto y el octavo decil e interpretarlos.

2. En un estudio de población se tomó una muestra de 27 personas, y se les preguntó por su edad y estado civil, obteniendo los siguientes resultados:

Estado civil	Edad								
Soltero	31	45	35	65	21	38	62	22	31
Casado	62	39	62	59	21	62			
Viudo	80	68	65	40	78	69	75		
Divorciado	31	65	59	49	65				

Se pide:

a) Calcular la media y la desviación típica de la edad según el estado civil e interpretarlas.

b) ¿En qué grupo es más representativa la media?

3. En un estudio se ha medido la tensión arterial de 25 individuos. Además se les ha preguntado si fuman y beben:

Fumador	si	no	si	si	si	no	no	si	no	si	no	si	no
Bebedor	no	no	si	si	no	no	si	si	no	si	no	si	si
Tensión arterial	80	92	75	56	89	93	101	67	89	63	98	58	91

Fumador	si	no	no	si	no	no	no	si	no	si	no	si
Bebedor	si	no	si	si	no	no	si	si	si	no	si	no
Tensión arterial	71	52	98	104	57	89	70	93	69	82	70	49

Calcular la media aritmética, desviación típica, coeficiente de asimetría y curtosis de la tensión arterial por grupos dependiendo de si beben o fuman e interpretarlos.

Práctica de Estadística con R 4

Regresión Lineal Simple y Correlación

1 Fundamentos teóricos

1.1 Regresión

La *regresión* es la parte de la estadística que trata de determinar la posible relación entre una variable numérica Y, que suele llamarse *variable dependiente*, y otro conjunto de variables numéricas, $X_1, X_2, \ldots, X_n$, conocidas como *variables independientes*, de una misma población. Dicha relación se refleja mediante un modelo funcional $y = f(x_1, \ldots, x_n)$.

El caso más sencillo se da cuando sólo hay una variable independiente X, y entonces se habla de *regresión simple*. En este caso el modelo que explica la relación entre X e Y es una función de una variable $y = f(x)$.

Dependiendo de la forma de esta función, existen muchos tipos de regresión simple. Los más habituales son los que aparecen en la siguiente tabla:

Modelo	Ecuación genérica
Lineal	$y = a + bx$
Parabólico	$y = a + bx + cx^2$
Polinómico de grado n	$y = a_0 + a_1x + \cdots + a_nx^n$
Potencial	$y = ax^b$
Exponencial	$y = e^{a+bx}$
Logarítmico	$y = a + b\log x$
Inverso	$y = a + b/x$
Curva S	$y = e^{a+b/x}$

Para elegir un tipo de modelo u otro, se suele representar el *diagrama de dispersión*, que consiste en dibujar sobre unos ejes cartesianos correspondientes a las variables X e Y, los pares de valores (x_i, y_j) observados en cada individuo de la muestra.

Ejemplo En la figura la figura 4.1 aparece el diagrama de dispersión correspondiente a una muestra de 30 individuos en los que se ha medido la estatura en cm (X) y el peso en kg (Y). En este caso la forma de la nube de puntos refleja una relación lineal entre la estatura y el peso.

Según la forma de la nube de puntos del diagrama, se elige el modelo más apropiado (figura 4.2), y se determinan los parámetros de dicho modelo para que la función resultante se ajuste lo mejor posible a la nube de puntos.

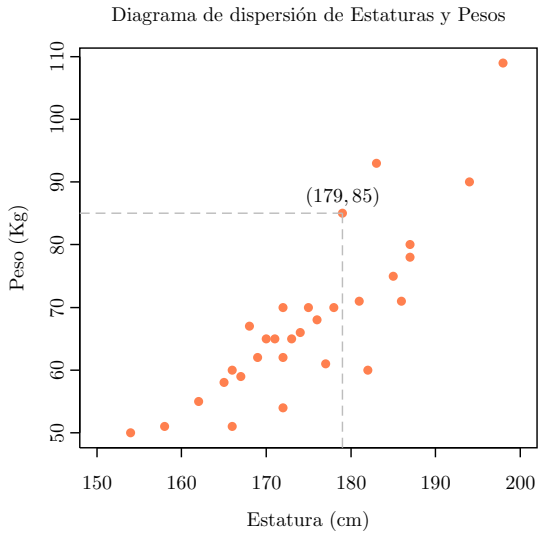
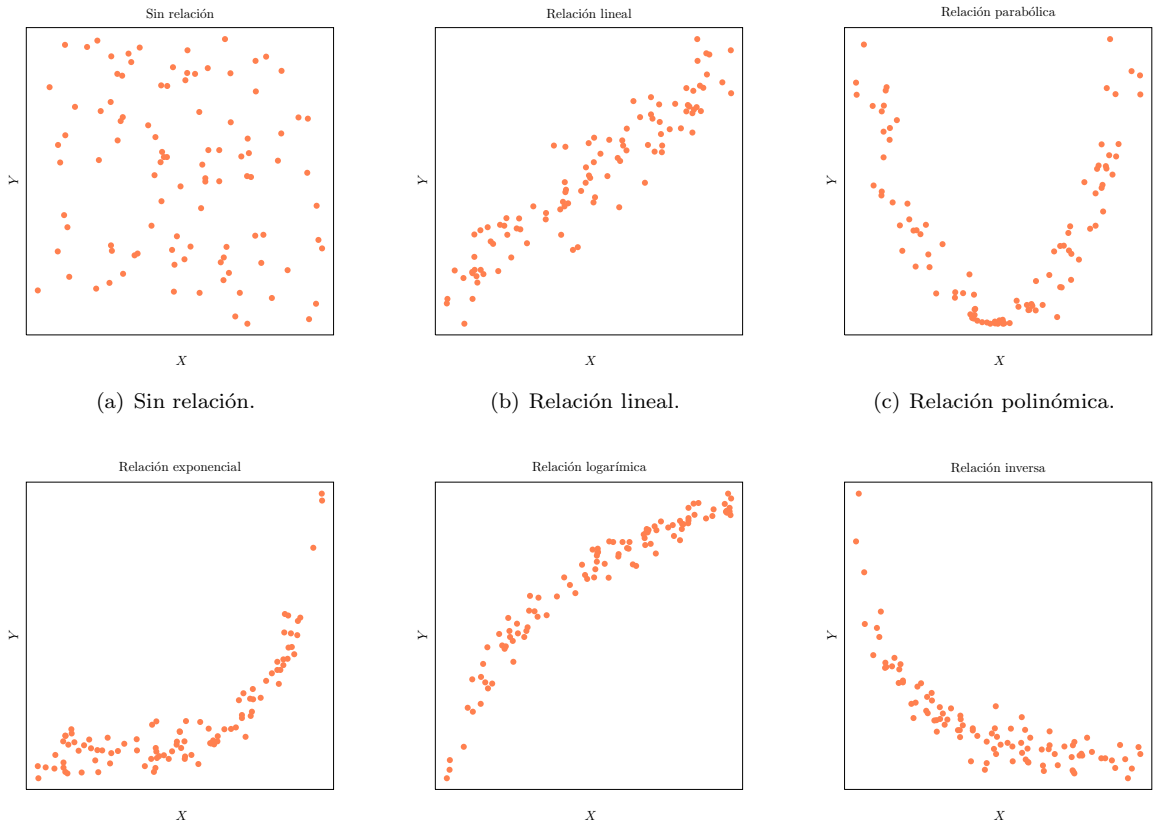


Figura 4.1 – Diagrama de dispersión. El punto (179,85) indicado corresponde a un individuo de la muestra que mide 179 cm y pesa 85 Kg.



(a) Sin relación. (b) Relación lineal. (c) Relación polinómica.

(d) Relación exponencial. (e) Relación logarítmica. (f) Relación inversa.

Figura 4.2 – Diagramas de dispersión correspondientes a distintos tipos de relaciones entre variables.

El criterio que suele utilizarse para obtener la función óptima, es que la distancia de cada punto a la curva, medida en el eje Y, sea lo menor posible. A estas distancias se les llama *residuos* o *errores* en Y (figura 4.3). La función que mejor se ajusta a la nube de puntos será, por tanto, aquella que hace mínima la suma de los cuadrados de los residuos.[1]

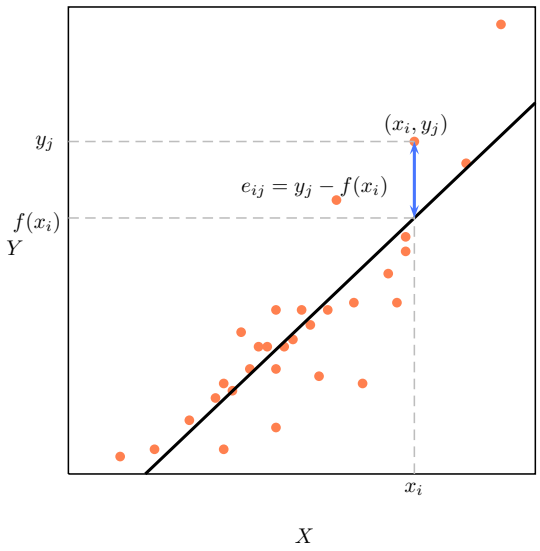


Figura 4.3 – Residuos o errores en Y. El residuo correspondiente a un punto (x_i, y_j) es la diferencia entre el valor y_j observado en la muestra, y el valor teórico del modelo $f(x_i)$, es decir, $e_{ij} = y_j - f(x_i)$.

Rectas de regresión

En el caso de que la nube de puntos tenga forma lineal y optemos por explicar la relación entre X e Y mediante una recta $y = a + bx$, los parámetros a determinar son a (punto de corte con el eje de ordenadas) y b (pendiente de la recta). Los valores de estos parámetros que hacen mínima la suma de residuos al cuadrado, determinan la recta óptima. Esta recta se conoce como *recta de regresión de Y sobre X* y explica la variable Y en función de la variable X. Su ecuación es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}),$$

donde s_{xy} es un estadístico llamado *covarianza* que mide el grado de relación lineal, y cuya fórmula es

$$s_{xy} = \frac{1}{n}\sum_{i,j}(x_i - \bar{x})(y_j - \bar{y})n_{ij}.$$

Ejemplo En la figura 4.4 aparecen las rectas de regresión de Estatura sobre Peso y de Peso sobre Estatura del ejemplo anterior.

La pendiente de la recta de regresión de Y sobre X se conoce como *coeficiente de regresión de Y sobre X*, y mide el incremento que sufrirá la variable Y por cada unidad que se incremente la variable X, según la recta.

Cuanto más pequeños sean los residuos, en valor absoluto, mejor se ajustará el modelo a la nube de puntos, y por tanto, mejor explicará la relación entre X e Y. Cuando todos los residuos son nulos, la recta pasa por todos los puntos de la nube, y la relación es perfecta. En este caso ambas rectas, la de Y sobre X y la de X sobre Y coinciden (figura 4.5(a)).

Por contra, cuando no existe relación lineal entre las variables, la recta de regresión de Y sobre X tiene pendiente nula, y por tanto la ecuación es $y = \bar{y}$, en la que, efectivamente no aparece x, o $x = \bar{x}$ en el caso de la recta de regresión X sobre Y, de manera que ambas rectas se cortan perpendicularmente (figura 4.5(b)).

[1] Se elevan al cuadrado para evitar que en la suma se compensen los residuos positivos con los negativos.

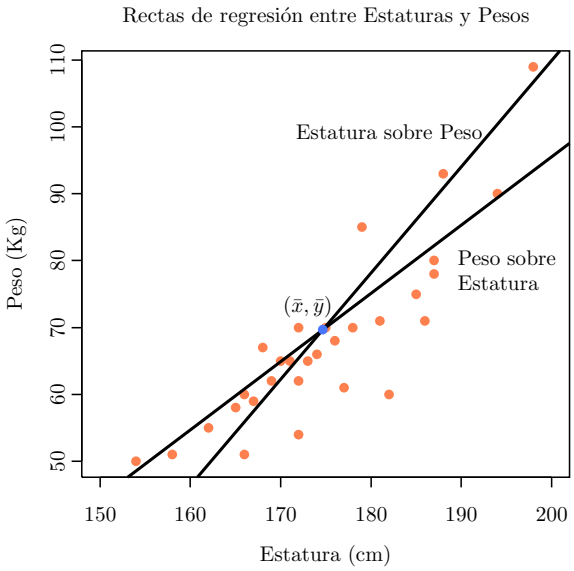


Figura 4.4 – Rectas de regresión de Estatura sobre Peso y de Peso sobre Estatura. Las rectas de regresión siempre se cortan en el punto de medias $(\bar{x}, \bar{y})$

1.2 Correlación

El principal objetivo de la regresión simple es construir un modelo funcional $y = f(x)$ que explique lo mejor posible la relación entre dos variables X (variable independiente) e Y (variable dependiente) medidas en una misma muestra. Generalmente, el modelo construido se utiliza para realizar inferencias predictivas de Y en función de X en el resto de la población. Pero aunque la regresión garantiza que el modelo construido es el mejor posible, dentro del tipo de modelo elegido (lineal, polinómico, exponencial, logarítmico, etc.), puede que aún así, no sea un buen modelo para hacer predicciones, precisamente porque no haya relación de ese tipo entre X e Y. Así pues, con el fin de validar un modelo para realizar predicciones fiables, se necesitan medidas que nos hablen del grado de dependencia entre X e Y, con respecto a un modelo de regresión construido. Estas medidas se conocen como medidas de *correlación*.

Dependiendo del tipo de modelo ajustado, habrá distintos tipos de medidas de correlación. Así, si el modelo de regresión construido es una recta, hablaremos de correlación lineal; si es un polinomio, hablaremos de correlación polinómica; si es una función exponencial, hablaremos de correlación exponencial, etc. En cualquier caso, estas medidas nos hablarán de lo bueno que es el modelo construido, y como consecuencia, de si podemos fiarnos de las predicciones realizadas con dicho modelo.

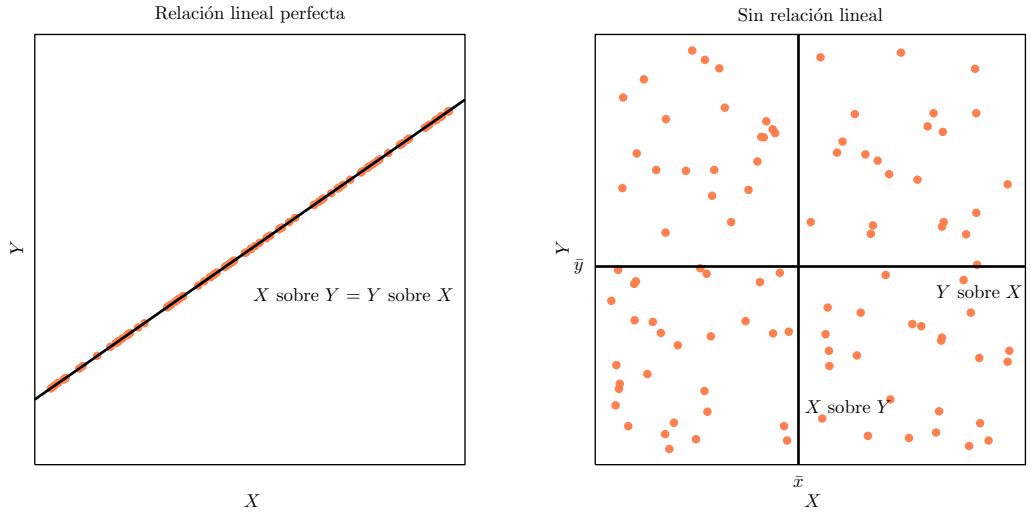
La mayoría de las medidas de correlación surgen del estudio de los residuos o errores en Y, que son las distancias de los puntos del diagrama de dispersión a la curva de regresión construida, medidas en el eje Y, tal y como se muestra en la figura (4.3). Estas distancias, son en realidad, los errores predictivos del modelo sobre los propios valores de la muestra.

Cuanto más pequeños sean los residuos, mejor se ajustará el modelo a la nube de puntos, y por tanto, mejor explicará la relación entre X e Y. Cuando todos los residuos son nulos, la curva de regresión pasa por todos los puntos de la nube, y entonces se dice que la relación es perfecta, o bien que existe una dependencia funcional entre X e Y (figura 4.5(a)). Por contra, cuando los residuos sean grandes, el modelo no explicará bien la relación entre X e Y, y por tanto, sus predicciones no serán fiables (figura 4.5(b)).

Varianza residual

Una primera medida de correlación, construida a partir de los residuos es la *varianza residual*, que se define como el promedio de los residuos al cuadrado:

$$s^2_{ry} = \frac{\sum_{i,j} e^2_{ij} n_{ij}}{n} = \frac{\sum_{i,j} (y_j - f(x_i))^2 n_{ij}}{n}.$$



(a) Dependencia funcional lineal.

(b) Independencia lineal.

Figura 4.5 – Distintos grados de dependencia. En el primer caso, la relación es perfecta y los residuos son nulos. En el segundo caso no existe relación lineal y la pendiente de la recta es nula.

Cuando los residuos son nulos, entonces $s^2_{ry} = 0$ y eso indica que hay dependencia funcional. Por otro lado, cuando las variables son independientes, con respecto al modelo de regresión ajustado, entonces los residuos se convierten en las desviaciones de los valores de Y con respecto a su media, y se cumple que $s^2_{ry} = s^2_y$. Así pues, se cumple que

$$0 \leq s^2_{ry} \leq s^2_y.$$

Según esto, cuanto menor sea la varianza residual, mayor será la dependencia entre X e Y, de acuerdo al modelo ajustado. No obstante, la varianza tiene como unidades las unidades de Y al cuadrado, y eso dificulta su interpretación.

Coeficiente de determinación

Puesto que el valor máximo que puede tomar la varianza residual es la varianza de Y, se puede definir fácilmente un coeficiente a partir de la comparación de ambas medidas. Surge así el *coeficiente de determinación* que se define como

$$R^2 = 1 - \frac{s^2_{ry}}{s^2_y}.$$

Se cumple que

$$0 \leq R^2 \leq 1,$$

y además no tiene unidades, por lo que es más fácil de interpretar que la varianza residual:

- $R^2 = 0$ indica que existe independencia según el tipo de relación planteada por el modelo de regresión.
- $R^2 = 1$ indica dependencia funcional.

Por tanto, cuanto mayor sea R^2, mejor será el modelo de regresión.

Si multiplicamos el coeficiente de determinación por 100, se obtiene el porcentaje de variabilidad de Y que explica el modelo de regresión. El porcentaje restante corresponde a la variabilidad que queda por explicar y se corresponde con el error predictivo del modelo. Así, por ejemplo, si tenemos un coeficiente de determinación $R^2 = 0{,}5$, el modelo de regresión explicaría la mitad de la variabilidad de Y, y en consecuencia, si se utiliza dicho modelo para hacer predicciones, estas tendrían la mitad de error que si no se utilizase, y se tomase como valor de la predicción el valor de la media de Y.

Coeficiente de determinación lineal

En el caso de que el modelo de regresión sea lineal, la fórmula del coeficiente de determinación se simplifica y se convierte en

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2},$$

que se conoce como *coeficiente de determinación lineal*.

Coeficiente de correlación

Otra medida de dependencia bastante habitual es el *coeficiente de correlación*, que se define como la raíz cuadrada del coeficiente de determinación:

$$R = \pm\sqrt{1 - \frac{s_{ry}^2}{s_y^2}},$$

tomando la raíz del mismo signo que la covarianza.

La única ventaja del coeficiente de correlación con respecto al coeficiente de determinación, es que tiene signo, y por tanto, además del grado de dependencia entre X e Y, también nos habla de si la relación es directa (signo +) o inversa (signo -). Su interpretación es:

- $R = 0$ indica independencia con respecto al tipo de relación planteada por el modelo de regresión.
- $R = -1$ indica dependencia funcional inversa.
- $R = 1$ indica dependencia funcional directa.

Por consiguiente, cuanto más próximo esté a -1 o a 1, mejor será el modelo de regresión.

Coeficiente de correlación lineal Al igual que ocurría con el coeficiente de determinación, cuando el modelo de regresión es lineal, la fórmula del coeficiente de correlación se convierte en

$$r = \frac{s_{xy}}{s_x s_y},$$

y se llama *coeficiente de correlación lineal*.

Por último, conviene remarcar que un coeficiente de determinación o de correlación nulo, indica que hay independencia según el modelo de regresión construido, pero puede haber dependencia de otro tipo. Esto se ve claramente en el ejemplo de la figura 4.6.

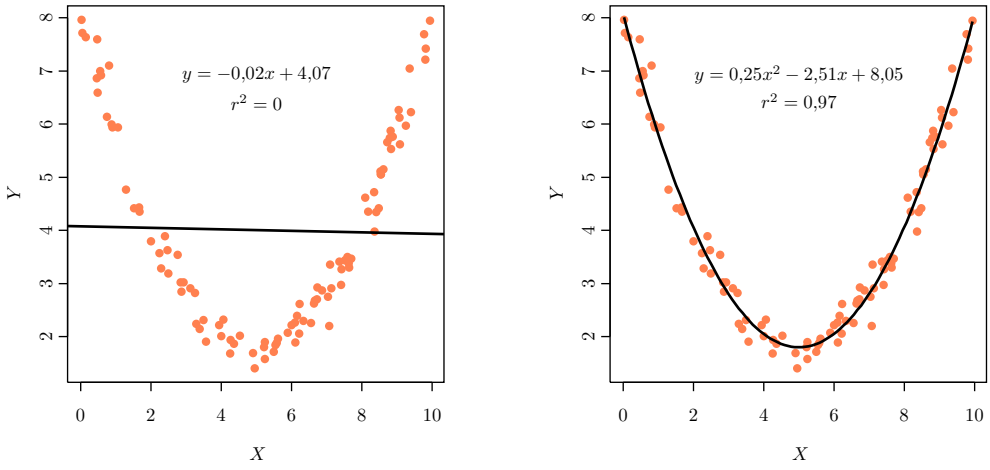
Fiabilidad de las predicciones

Aunque el coeficiente de determinación o de correlación nos hablan de la bondad de un modelo de regresión, no es el único dato que hay que tener en cuenta a la hora de hacer predicciones.

La fiabilidad de las predicciones que hagamos con un modelo de regresión depende de varias cosas:

- El coeficiente de determinación: Cuando mayor sea, menores serán los errores predictivos y mayor la fiabilidad de las predicciones.
- La variablidad de la población: Cuanto más variable es una población, más difícil es predecir y por tanto menos fiables serán las predicciones del modelo.
- El tamaño muestral: Cuanto mayor sea, más información tendremos y, en consecuencia, más fiables serán las predicciones.

Además, hay que tener en cuenta que un modelo de regresión es válido para el rango de valores observados en la muestra, pero fuera de ese rango no tenemos información del tipo de relación entre las variables, por lo que no deberíamos hacer predicciones para valores que estén lejos de los observados en la muestra.



(a) Dependencia lineal débil.

(b) Dependencia parabólica fuerte.

Figura 4.6 – En la figura de la izquierda se ha ajustado un modelo lineal y se ha obtenido un $R^2 = 0$, lo que indica que el modelo no explica nada de la relación entre X e Y, pero no podemos afirmar que X e Y son independientes. De hecho, en la figura de la derecha se observa que al ajustar un modelo parabólico, $R^2 = 0{,}97$, lo que indica que casi hay una dependencia funcional parabólica entre X e Y.

2 Ejercicios resueltos

1. Se han medido dos variables X e Y en 10 individuos obteniendo los siguientes resultados:

X	0	1	2	3	4	5	6	7	8	9
Y	2	5	8	11	14	17	20	23	26	29

Se pide:

a) Crear un conjunto de datos con las variables X y Y e introducir estos datos.

b) Dibujar el diagrama de dispersión correspondiente.

Indicación
1) Seleccionar el menú `Gráficos→Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar como `Variable x` la variable X y como `Variable y` la variable Y, desmarcar todas las opciones y hacer click en el botón Aceptar.

En vista del diagrama, ¿qué tipo de modelo crees que explicará mejor la relación entre X y Y?

c) Calcular la recta de regresión de Y sobre X.

Indicación
1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable Y como `Variable explicada` y la variable X como `Variable explicativa`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.
3) La recta de regresión es de la forma Y=a+bX donde a es el término independiente y b es la pendiente. Las estimaciones de ambos valores aparecen en la ventana de resultados en la columna Estimated, el término independiente corresponde a la fila Intercept y la pendiente a la fila con el nombre de la variable independiente, en este caso X.

d) Dibujar dicha recta sobre el diagrama de dispersión.

Indicación
Repetir los pasos del apartado anterior para dibujar el diagrama de dispersión pero activando la opción `Línea de mínimos cuadrados`.

e) Calcular la recta de regresión de X sobre Y y dibujarla sobre el correspondiente diagrama de dispersión.

Indicación
Repetir los pasos de los apartados anteriores pero escogiendo como `Variable explicada` la variable X, y como `Variable explicativa` la variable Y

f) ¿Son grandes los residuos? Comentar los resultados.

2. En una licenciatura se quiere estudiar la relación entre el número medio de horas de estudio diarias y el número de asignaturas suspensas. Para ello se obtuvo la siguiente muestra:

Horas	Suspensos	Horas	Suspensos	Horas	Suspensos
3,5	1	2,2	2	1,3	4
0,6	5	3,3	0	3,1	0
2,8	1	1,7	3	2,3	2
2,5	3	1,1	3	3,2	2
2,6	1	2,0	3	0,9	4
3,9	0	3,5	0	1,7	2
1,5	3	2,1	2	0,2	5
0,7	3	1,8	2	2,9	1
3,6	1	1,1	4	1,0	3
3,7	1	0,7	4	2,3	2

Se pide:

a) Crear un conjunto de datos con las variables horas estudio y suspensos e introducir estos datos.

b) Calcular la recta de regresión de suspensos sobre horas estudio y dibujarla.

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable suspensos como `Variable explicada` y la variable horas estudio como `variable explicativa`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.
3) La recta de regresión es de la forma suspensos=a+$b$horas estudio donde a es el término independiente y b es la pendiente. Las estimaciones de ambos valores aparecen en la ventana de resultados en la columna Estimated, el término independiente corresponde a la fila Intercept y la pendiente a la fila con el nombre de la variable independiente, en este caso horas estudio.
4) Seleccionar el menú `Gráficos→Diagrama de Dispersión`.
5) En el cuadro de diálogo que aparece, seleccionar como `variable x` la variable horas estudio y como `variable y` la variable suspensos, marcar la opción `Línea de mínimos cuadrados` y hacer click en el botón Aceptar.

c) Indicar el coeficiente de regresión de suspensos sobre horas estudio. ¿Cómo lo interpretarías?

Indicación
El coeficiente de regresión es la pendiente de la recta de regresión.

d) La relación lineal entre estas dos variables, ¿es mejor o peor que la del ejercicio anterior? Comentar los resultados a partir las gráficas de las rectas de regresión y sus residuos.

e) Calcular los coeficientes de correlación y de determinación lineal. ¿Es un buen modelo la recta de regresión? ¿Qué porcentaje de la variabilidad del número de suspensos está explicada por el modelo?

Indicación
El coeficiente de determinación aparece en la ventana de resultados como Multiple R-squared, y el coeficiente de correlación es su raíz cuadrada.

f) Utilizar la recta de regresión para predecir el número de suspensos correspondiente a 3 horas de estudio diarias. ¿Es fiable esta predicción?

Indicación

1) Seleccionar en el botón de modelos el modelo con el que hacer predicciones.
2) Seleccionar el menú `Modelos→Predicciones de regresión simple`.
3) En el cuadro de diálogo que aparece introducir los valores para los que se desea la predicción en el campo `Predicciones para` y hacer click sobre el botón Aceptar.

g) Según el modelo lineal, ¿cuántas horas diarias tendrá que estudiar como mínimo un alumno si quiere aprobarlo todo?

Indicación
Seguir los mismos pasos de los apartados anteriores, pero escogiendo como variable dependiente horas estudio, y como independiente suspensos.

3. Después de tomar un litro de vino se ha medido la concentración de alcohol en la sangre en distintos instantes, obteniendo:

Tiempo después (minutos)	30	60	90	120	150	180	210
Concentración (gramos/litro)	1,6	1,7	1,5	1,1	0,7	0,2	2,1

Se pide:

a) Crear las variables tiempo y alcohol e introducir estos datos.

b) Calcular el coeficiente de correlación lineal entre el alcohol y el tiempo e interpretarlo. ¿Es bueno el modelo lineal?

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable alcohol como `Variable explicada` y la variable tiempo como `variable explicativa`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.
3) El coeficiente de determinación aparece en la ventana de resultados como Multiple R-squared, y el coeficiente de correlación es su raíz cuadrada. Aceptar.

c) Dibujar la recta de regresión del alcohol sobre el tiempo. ¿Existe algún individuo con un residuo demasiado grande? Si es así, eliminar dicho individuo de la muestra y volver a calcular el coeficiente de correlación. ¿Ha mejorado el modelo?

Indicación

1) Seleccionar el menú `Gráficos→Diagrama de dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar como `variable x` la variable tiempo y como `variable y` la variable alcohol, marcar la opción `Línea de mínimos cuadrados` y la opción `Identificar observaciones` y hacer click en el botón Aceptar.
3) En la ventana con el gráfico de dispersión, si existe algún individuo con un residuo demasiado grande, hacer click sobre él para identificarlo.
4) Seleccionar el menú `Datos→Conjunto de datos activo→Borrar fila(s) del conjunto de datos activo`.
5) En el cuadro de diálogo que aparece introducir los índices de los datos con residuos grandes en el campo `Índices o nombres de la(s) fila(s) para borrar` y hacer click sobre el botón Aceptar.
6) Repetir los pasos del apartado anterior.
7) Repetir los pasos para dibujar el diagrama de dispersión.

d) Si la concentración máxima de alcohol en la sangre que permite la ley para poder conducir es 0,5 g/l, ¿cuánto tiempo habrá que esperar después de tomarse un litro de vino para poder conducir sin infringir la ley? ¿Es fiable esta predicción?

Indicación

1) Repetir los pasos del primer apartado pero tomando alcohol como `Variable explicativa` y tiempo como `Variable explicada`.
2) Seleccionar en el botón de modelos el modelo con el que hacer predicciones.
3) Seleccionar el menú `Modelos→Predicciones de regresión simple`.
4) En el cuadro de diálogo que aparece introducir los valores para los que se desea la predicción en el campo `Predicciones para` y hacer click sobre el botón Aceptar.

4. En un estudio se ha medido la altura y la edad de 30 personas y se han guardado en el fichero `edad_estatura.txt`. Se pide:

a) Importar los datos del fichero `edad_estatura.txt` en un conjunto de datos.

b) Calcular la recta de regresión de la altura sobre la edad. ¿Es un buen modelo la recta de regresión?

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable altura como `Variable explicada` y la variable edad como `variable explicativa`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar. Aceptar.

c) Dibujar la recta de regresión de la altura sobre la edad. ¿Alrededor de qué edad se observa un cambio en la tendencia?

Indicación

1) Seleccionar el menú `Gráficos→Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar como `variable x` la variable edad y como `variable y` la variable altura, marcar la opción `Línea de mínimos cuadrados` y hacer click en el botón Aceptar.

d) Recodificar la variable edad en dos grupos para mayores y menores de 20 años.

Indicación

1) Seleccionar el menú `Datos→Modificar variables del conjunto activo→Recodificar variable`.
2) En el cuadro de diálogo que aparece seleccionar la variable edad e introducir como nombre de la variable recodificada grupo_edad.
3) En el campo `Introducir directrices de recodificación` introducir las reglas de recodificación y hacer click en el botón Aceptar.

e) Calcular la recta de regresión de la altura sobre la edad para cada grupo de edad. ¿En qué grupo explica mejor la recta de regresión la relación entre la altura y la edad? Justificar la respuesta.

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable altura como `Variable explicada` y la variable edad como `variable explicativa`, en el campo `Expresión de selección` introducir la condición `grupo_edad=="menores"`, introducir el nombre Recta.menores para el modelo y hacer click sobre el botón Aceptar.
3) Repetir los mismo pero con la condición `grupo_edad="mayores"` y llamando al modelo Recta.mayores.

f) Dibujar las rectas de regresión anteriores.

Indicación

1) Seleccionar el menú `Gráficos→Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar como `variable x` la variable edad y como `variable y` la variable altura, marcar la opción `Línea de mínimos cuadrados` y hacer click en el botón Gráfica por grupos.
3) En el cuadro de diálogo que aparece, seleccionar la variable grupo_edad y hacer click en el botón Aceptar.

g) ¿Qué altura se espera que tenga una persona de 14 años? ¿Y una de 38?

Indicación

1) Hacer click en el botón de modelos y seleccionar el modelo Recta.menores
2) Seleccionar el menú `Modelos→Predicciones de regresión simple`.
3) En el cuadro de diálogo que aparece introducir los valores para los que se desea la predicción en el campo `Predicciones para` y hacer click sobre el botón Aceptar.
4) Para la segunda predicción repetir los mismos pasos pero seleccionando el modelo Recta.mayores.

5. El fichero `nations.txt` contiene información sobre el desarrollo de distintos países (tasa de uso de anticonceptivos (contraception), producto interior bruto per cápita (GDP), tasa de mortalidad infantil (infant.mortality) y tasa de fertilidad (TFR)). Se pide:

a) Importar el fichero `nations.txt` en un conjunto de datos.

b) ¿Entre qué variables existe relación lineal?

Indicación

1) Seleccionar el menú `Estadísticos→Resúmenes→Matriz de correlaciones`.
2) En el cuadro de diálogo que aparece seleccionar todas las variables y hacer click sobre el botón Aceptar.

c) ¿Existe relación lineal entre la tasa de mortalidad infantil y tasa de fertilidad en Europa?

Indicación

1) Seleccionar el menú `Datos→Conjunto de datos activo→Filtrar conjunto de datos activo`.
2) En el cuadro de diálogo que aparece introducir la condición `region=="Europe"` en el campo `Expresión de selección`, introducir el nombre Europa en el campo `Nuevo nombre del conjunto de datos` y hacer click en el botón Aceptar.
3) Repetir los pasos del apartado anterior.

6. La siguiente tabla recoge la información de las calificaciones obtenidas por un grupo de alumnos en dos asignaturas X e Y.

Alumno	1	2	3	4	5	6	7	8	9	10	11	12
X	NT	AP	SS	SS	AP	AP	SS	NT	SB	SS	AP	AP
Y	SB	SS	AP	SS	AP	NT	SS	NT	NT	AP	AP	NT

Se pide:

a) Crear un conjunto de datos con las variables X e Y e introducir los datos.

b) ¿Existe relación entre las calificaciones de X e Y? Justificar la respuesta.

Indicación

Primero hay que crear dos nuevas variables con los rangos (números de orden) de las variables X e Y.

1) Seleccionar el menú `Datos→Modificar variables del conjunto activo→Recodificar variable`.
2) En el cuadro de diálogo que aparece seleccionar la variable X e introducir como nombre de la variable recodificada rangoX.
3) En el campo `Introducir directrices de recodificación` introducir las reglas de recodificación `''SS''=1`, `''AP''=2`, `''NT''=3` y `''SB''=4`, desmarcar la opción `Convertir cada nueva variable en factor` y hacer click en el botón Aceptar.
4) Repetir lo mismo para la variable Y.

Ahora ya se puede calcular el coeficente de correlación de Spearman:

1) Seleccionar el menú `Estadísticos→Resúmenes→Matriz de correlaciones`.
2) En el cuadro de diálogo que aparece seleccionar las variables rangoX y rangoY, seleccionar la opción `Coeficiente de Spearman` y hacer click en el botón Aceptar.

3 Ejercicios propuestos

1. Se determina la pérdida de actividad que experimenta un medicamento desde el momento de su fabricación a lo largo del tiempo, obteniéndose el siguiente resultado:

Tiempo (en años)	1	2	3	4	5
Actividad restante (%)	96	84	70	58	52

Se desea calcular:

a) La relación fundamental (recta de regresión) entre actividad restante y tiempo transcurrido.

b) ¿En qué porcentaje disminuye la actividad cada año que pasa?

c) ¿Cuándo tiempo debe pasar para que el fármaco tenga una actividad del 80 %? ¿Cuándo será nula la actividad? ¿Son igualmente fiables estas predicciones?

2. Al realizar un estudio sobre la dosificación de un cierto medicamento, se trataron 6 pacientes con dosis diarias de 2 mg, 7 pacientes con 3 mg y otros 7 pacientes con 4 mg. De los pacientes tratados con 2 mg, 2 curaron al cabo de 5 días, y 4 al cabo de 6 días. De los pacientes tratados con 3 mg diarios, 2 curaron al cabo de 3 días, 4 al cabo de 5 días y 1 al cabo de 6 días. Y de los pacientes tratados con 4 mg diarios, 5 curaron al cabo de 3 días y 2 al cabo de 4 días. Se pide:

a) Calcular la recta de regresión del tiempo de curación con respecto a la dosis suministrada.

b) Calcular el coeficiente de regresión del tiempo de curación con respecto a la dosis e interpretarlo.

c) Calcular el coeficiente de correlación lineal e interpretarlo.

d) Determinar el tiempo esperado de curación para una dosis de 5 mg diarios. ¿Es fiable esta predicción?

e) ¿Qué dosis debe aplicarse si queremos que el paciente tarde 4 días en curarse? ¿Es fiable la predicción?

3. En una clase de alumnos universitarios se ha medido la estatura, el peso y el sexo de cada uno y se han guardado en el fichero `estaturas_pesos_alumnos.txt`. Se pide:

a) Importar los datos del fichero `estaturas_pesos_alumnos.txt` en un conjunto de datos.

b) Calcular la recta de regresión del peso sobre la estatura y dibujarla.

c) Calcular las rectas de regresión del peso sobre la estatura para cada sexo y dibujarlas.

d) Calcular los coeficientes de determinación de ambas rectas. ¿Qué recta es mejor modelo? Justificar la respuesta.

e) ¿Qué peso tendrá un hombre que mida 170 cm? ¿Y una mujer de la misma estatura?

Práctica de Estadística con R 5

Regresión no lineal

1 Fundamentos teóricos

La regresión simple tiene por objeto la construcción de un modelo funcional $y = f(x)$ que explique lo mejor posible la relación entre dos variables Y (variable dependiente) y X (variable independiente) medidas en una misma muestra.

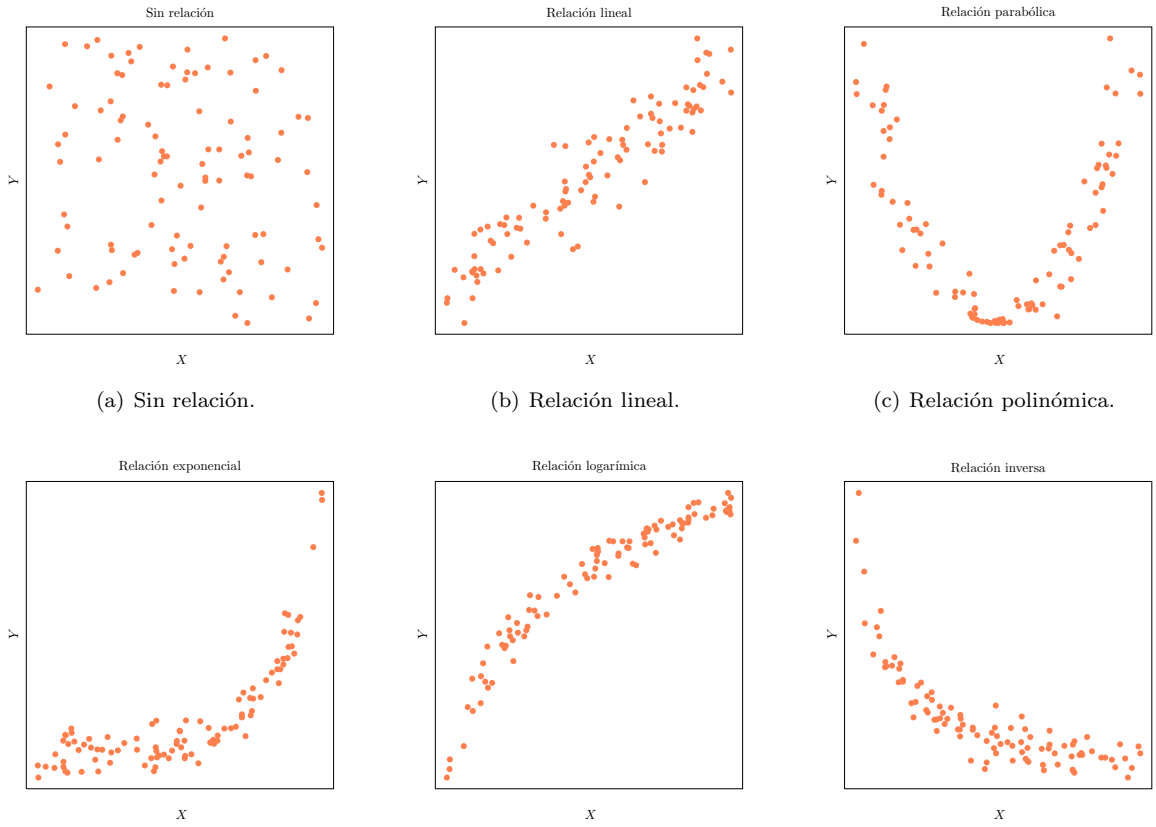
Ya vimos que, dependiendo de la forma de esta función, existen muchos tipos de regresión simple. Entre los más habituales están:

Modelo	Ecuación genérica
Lineal	$y = a + bx$
Parabólico	$y = a + bx + cx^2$
Polinómico de grado n	$y = a_0 + a_1x + \cdots + a_nx^n$
Potencial	$y = ax^b$
Exponencial	$y = e^{a+bx}$
Logarítmico	$y = a + b\log x$
Inverso	$y = a + b/x$
Curva S	$y = e^{a+b/x}$

La elección de un tipo de modelo u otro suele hacerse según la forma de la nube de puntos del diagrama de dispersión. A veces estará claro qué tipo de modelo se debe construir, tal y como ocurre en los diagramas de dispersión de la figura 5.1. Pero otras veces no estará tan claro, y en estas ocasiones, lo normal es ajustar los dos o tres modelos que nos parezcan más convincentes, para luego quedarnos con el que mejor explique la relación entre Y y X, mirando el coeficiente de determinación[1] de cada modelo.

Ya vimos en la práctica sobre regresión lineal simple, cómo construir rectas de regresión. En el caso de que optemos por ajustar un modelo no lineal, la construcción del mismo puede realizarse siguiendo los mismos pasos que en el caso lineal. Básicamente se trata de determinar los parámetros del modelo que minimizan la suma de los cuadrados de los residuos en Y. En los modelos multiplicativo y exponencial, el sistema aplica transformaciones logarítmicas a las variables y después ajusta un modelo lineal a los datos transformados. En el modelo recíproco, el sistema sustituye la variable dependiente por su recíproco antes de estimar la ecuación de regresión.

[1] Ver la práctica de correlación.



(a) Sin relación. (b) Relación lineal. (c) Relación polinómica.

(d) Relación exponencial. (e) Relación logarítmica. (f) Relación inversa.

Figura 5.1 – Diagramas de dispersión correspondientes a distintos tipos de relaciones entre variables.

2 Ejercicios resueltos

El procedimiento más sencillo para construir un modelo no lineal, siempre que sea posible, es transformar las variables para convertirlo en un modelo lineal. En el caso de los modelos de regresión simple más comunes las transformaciones que convierten cada modelo en un modelo lineal aparecen en la tabla siguiente:

Modelo	Modelo no lineal	Modelo lineal	Transformación
Potencial	$y = ax^b$	$\log(y) = \log(a) + b\log(x)$	Se toma el logaritmo de ambas variables
Exponencial	$y = e^{a+bx}$	$\log(y) = a + bx$	Se toma el logaritmo de la variable dependiente
Logarítmico	$y = a + b\log x$	$y = a + b\log x$	Se toma el logaritmo de la variable independiente
Inverso	$y = a + b/x$	$y = a + b\frac{1}{x}$	Se toma el inverso de la variable independiente
Curva S	$y = e^{a+b/x}$	$\log(y) = a + b\frac{1}{x}$	Se toma el logaritmo de la variable dependiente y el inverso de la independiente

1. En un experimento se ha medido el número de bacterias por unidad de volumen en un cultivo, cada hora transcurrida, obteniendo los siguientes resultados:

Horas	0	1	2	3	4	5	6	7	8
Nº Bacterias	25	28	47	65	86	121	190	290	362

Se pide:

a) Crear un conjunto de datos con las variables horas y bacterias e introducir estos datos.

b) Dibujar el diagrama de dispersión correspondiente. En vista del diagrama, ¿qué tipo de modelo crees que explicará mejor la relación entre el número de bacterias y el tiempo transcurrido?

Indicación

1) Seleccionar el menú `Gráficos→Diagrama de dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar como `variable x` la variable horas y como `variable y` la variable bacterias, desmarcar todas las opciones y hacer click en el botón Aceptar.

c) Calcular los modelos exponencial y cuadrático de las bacterias sobre las horas. ¿Qué tipo de modelo es el mejor?

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable bacterias como `Variable explicada` y la variable horas como `Variable explicativa`, marcar la opción `Exponencial`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.
3) El modelo de regresión exponencial es bacterias $= e^{a+b\text{horas}}$. Las estimaciones de los parámetros a y b, aparecen en la ventana de resultados en la columna Estimated, el valor de a corresponde a la fila Intercept y el del b al nombre de la variable independiente, en este caso horas.
4) El modelo mejor será aquel que tenga un coeficiente de determinación mayor.

d) Dibujar la curva del mejor de los modelos anteriores.

Indicación

1) Seleccionar el menú `Gráficas→Gráfica de regresión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable bacterias como `Variable explicada` y la variable horas como `Variable explicativa`, marcar la opción `Exponencial`, introducir un título para el gráfico y hacer click sobre el botón Aceptar.

e) Según el modelo anterior, ¿cuántas bacterias habrá al cabo de 3 horas y media del inicio del cultivo? ¿Y al cabo de 10 horas? ¿Son fiables estas predicciones?

Indicación

1) Seleccionar en el botón de modelos el modelo con el que hacer predicciones.
2) Seleccionar el menú `Modelos→Predicciones de regresión simple`.
3) En el cuadro de diálogo que aparece introducir los valores para los que se desea la predicción en el campo `Predicciones para` y hacer click sobre el botón Aceptar.
4) Como se trata de un modelo exponencial, las predicciones obtenidas corresponden al logaritmo de bacterias. Para obtener la predicción de bacterias basta con aplicar la función exponencial a los valores obtenidos.

f) Dar una predicción lo más fiable posible del tiempo que tendría que transcurrir para que en el cultivo hubiese 100 bacterias.

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable horas como `Variable explicada` y la variable bacterias como `Variable explicativa`, marcar la opción `Logarítmica`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.
3) Seleccionar el menú `Modelos→Predicciones de regresión simple`.
4) En el cuadro de diálogo que aparece introducir los valores para los que se desea la predicción en el campo `Predicciones para` y hacer click sobre el botón Aceptar.

2. El fichero `nations.txt` contiene información sobre el desarrollo de distintos países (tasa de uso de anticonceptivos (contraception), producto interior bruto per cápita (GDP), tasa de mortalidad infantil (infant.mortality) y tasa de fertilidad (TFR)). Se pide:

a) Importar el fichero `nations.txt` en un conjunto de datos.

b) ¿Entre qué variables existe relación no lineal?

Indicación

1) Seleccionar el menú `Gráficas→Matriz de diagramas de dispersión`.
2) En el cuadro de diálogo que aparece seleccionar todas las variables y hacer click sobre el botón Aceptar.

c) Construir el mejor modelo de regresión de la tasa de mortalidad infantil sobre el producto interior bruto. ¿Cómo explicarías esta relación?

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable infant.mortality como `Variable explicada` y la variable GDP como `Variable explicativa`, marcar el modelo, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.
3) Repetir lo mismo para cada tipo de modelo. El modelo mejor será aquel que tenga un coeficiente de determinación mayor.

d) Dibujar el modelo del apartado anterior.

Indicación

1) Seleccionar el menú `Gráficas→Gráfica de regresión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable infant.mortality como `Variable explicada` y la variable GDP como `Variable explicativa`, marcar la opción correspondiente al mejor modelo, introducir un título para el gráfico y hacer click sobre el botón Aceptar.

3. El fichero `dieta.txt` contiene los datos de un estudio llevado a cabo por un centro dietético para probar una nueva dieta de adelgazamiento. Para cada individuo se ha medido el número de días que lleva con la dieta, el número de kilos perdidos desde entonces y si realizó o no un programa de ejercicios. Se pide:

a) Importar el fichero `dieta.txt` en un conjunto de datos.

b) Dibujar el diagrama de dispersión. Según la nube de puntos, ¿qué tipo de modelo explicaría mejor la relación entre los kilos perdidos y los días de dieta?

Indicación

1) Seleccionar el menú `Gráficos→Diagrama de dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar como `variable x` la variable dias y como `variable y` la variable kilos, desmarcar todas las opciones y hacer click en el botón Aceptar.

c) Construir el modelo de regresión que mejor explique la relación entre los kilos perdidos y los días de dieta.

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos como `Variable explicada` y la variable dias como `Variable explicativa`, marcar el modelo, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.
3) Repetir lo mismo para cada tipo de modelo. El modelo mejor será aquel que tenga un coeficiente de determinación mayor.

d) Dibujar el modelo del apartado anterior.

Indicación

1) Seleccionar el menú `Gráficas→Gráfica de regresión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos como `Variable explicada` y la variable dias como `Variable explicativa`, marcar la opción `scurve`, introducir un título para el gráfico y hacer click sobre el botón Aceptar.

e) Dibujar el diagrama de dispersión distinguiendo los que hacen ejercicio de los que no.

Indicación

1) Seleccionar el menú `Gráficos→Diagrama de dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar como `variable x` la variable dias y como `variable y` la variable kilos, desmarcar todas las opciones y hacer click en el botón Gráfica por grupos.
3) En el cuadro de diálogo que aparece, seleccionar la variable ejercicio y hacer click en el botón Aceptar.

f) Construir el modelo de regresión curva S que mejor explique la relación entre los kilos perdidos y los días de dieta para los que no hacen ejercicio.

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos como `Variable explicada` y la variable dias como `Variable explicativa`, marcar el modelo `Curva S`, introducir un nombre para el modelo, introducir la condición `ejercicio=="no"` en el campo `Expresión de selección` y hacer click sobre el botón Aceptar.

g) Dibujar el modelo del apartado anterior.

Indicación

1) Seleccionar el menú `Gráficas→Gráfica de regresión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos como `Variable explicada` y la variable dias como `Variable explicativa`, marcar el modelo `Curva S`, introducir un título para el gráfico, introducir la condición `ejercicio=="no"` en el campo `Expresión de selección` y hacer click sobre el botón Aceptar.

h) Construir el modelo de regresión inverso que mejor explique la relación entre los kilos perdidos y los días de dieta para los que si hacen ejercicio.

Indicación

1) Seleccionar el menú `Estadísticos→Ajustes de modelos→Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos como `Variable explicada` y la variable dias como `Variable explicativa`, marcar el modelo `Inverso`, introducir un nombre para el modelo, introducir la condición `ejercicio=="si"` en el campo `Expresión de selección` y hacer click sobre el botón Aceptar.

i) Dibujar el modelo del apartado anterior.

Indicación

1) Seleccionar el menú `Gráficas→Gráfica de regresión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos como `Variable explicada` y la variable dias como `Variable explicativa`, marcar el modelo `Inverso`, introducir un título para el gráfico, introducir la condición `ejercicio=="si"` en el campo `Expresión de selección` y hacer click sobre el botón Aceptar.

j) Utilizar el modelo construido para predecir el número de kilos perdidos tras 40 y 500 días de dieta, tanto para los que hacen ejercicio como para los que no. ¿Son fiables estas predicciones?

Indicación

1) Hacer click en el botón de modelos y seleccionar el mejor de los modelos para los que no hacen ejercicio.
2) Seleccionar el menú `Modelos→Predicciones de regresión simple`.
3) En el cuadro de diálogo que aparece introducir los valores para los que se desea la predicción en el campo `Predicciones para` y hacer click sobre el botón Aceptar.
4) Como se trata de un modelo de curva S, las predicciones obtenidas corresponden al logaritmo de los kilos. Para obtener la predicción de los kilos basta con aplicar la función exponencial a los valores obtenidos.
5) Repetir los pasos anteriores seleccionando el mejor de los modelos para los que si hacen ejercicio.

3 Ejercicios propuestos

1. La concentración de un fármaco en sangre, C en mg/dl, es función del tiempo, t en horas, y viene dada por la siguiente tabla:

t	2	3	4	5	6	7	8
C	25	36	48	64	86	114	168

Se pide:

a) Según el modelo exponencial, ¿qué concentración de fármaco habría a las 4,8 horas? ¿Es fiable la predicción? Justificar adecuadamente la respuesta.

b) Según el modelo logarítmico, ¿qué tiempo debe pasar para que la concentración sea de 100 mg/dl?

2. El fichero `nations.txt` contiene información sobre el desarrollo de distintos países (tasa de uso de anticonceptivos (contraception), producto interior bruto per cápita (GDP), tasa de mortalidad infantil (infant.mortality) y tasa de fertilidad (TFR)). Se pide:

a) Importar el fichero `nations.txt` en un conjunto de datos.

b) Construir el mejor modelo de regresión de la tasa de fertilidad sobre el producto interior bruto. ¿Cómo explicarías esta relación?

c) Dibujar el modelo del apartado anterior.

d) ¿Qué tasa de fertilidad le corresponde a una mujer que viva en un país con un producto interior bruto per cápita de 10000 $? ¿Y si la mujer vive en Europa?

Práctica de Estadística con R 6

Variables Aleatorias Discretas

1 Fundamentos teóricos

1.1 Variables Aleatorias

Se define una *variable aleatoria* asignando a cada resultado del experimento aleatorio un número. Esta asignación puede realizarse de distintas maneras, obteniéndose de esta forma diferentes variables aleatorias. Así, en el lanzamiento de dos monedas podemos considerar el número de caras o el número de cruces. En general, si los resultados del experimento son numéricos, se tomarán dichos números como los valores de la variable, y si los resultados son cualitativos, se hará corresponder a cada modalidad un número arbitrariamente.

Formalmente, una *variable aleatoria* X es una función real definida sobre los puntos del espacio muestral E de un experimento aleatorio.

$$X : E \to \mathbb{R}$$

De esta manera, la distribución de probabilidad del espacio muestral E, se transforma en una distribución de probabilidad para los valores de X.

El conjunto formado por todos los valores distintos que puede tomar la variable aleatoria se llama *Rango* o *Recorrido* de la misma.

Las variables aleatorias pueden ser de dos tipos: discretas o continuas. Una variable es *discreta* cuando sólo puede tomar valores aislados, mientras que es *continua* si puede tomar todos los valores posibles de un intervalo.

1.2 Variables Aleatorias Discretas (v.a.d.)

Se considera una v.a.d. X que puede tomar los valores x_i, $i = 1, 2, ..., n$.

Función de probabilidad

La *distribución de probabilidad* de X se suele caracterizar mediante una función $f(x)$, conocida como *función de probabilidad*, que asigna a cada valor de la variable su probabilidad. Esto es

$$f(x_i) = P(X = x_i),\ i = 1, .., n$$

verificándose que

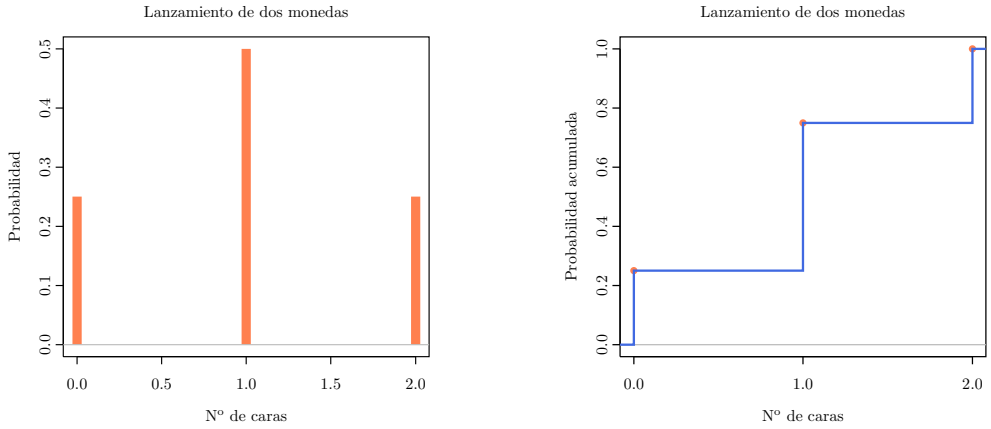
$$\sum_{i=1}^{n} f(x_i) = 1$$

Función de distribución

Otra forma equivalente de caracterizar la distribución de probabilidad de X es mediante otra función $F(x)$, llamada *función de distribución*, que asigna a cada $x \in \mathbb{R}$ la probabilidad de que X tome un valor menor o igual que dicho número x. Así,

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

Tanto la función de probabilidad como la función de distribución pueden representarse de forma gráfica, tal y como se muestra en la figura 6.1.



(a) Función de probabilidad. (b) Función de distribución.

Figura 6.1 – Función de probabilidad y función de distribución de la variable aleatoria X que mide el número de caras obtenido en el lanzamiento de dos monedas.

Estadísticos poblacionales

Los parámetros descriptivos más importantes de una v.a.d. X son:

Media o Esperanza

$$E[X] = \mu = \sum_{i=1}^{n} x_i f(x_i)$$

Varianza

$$V[X] = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i) = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2$$

Desviación típica

$$D[X] = \sigma = +\sqrt{\sigma^2}$$

La media es una medida de tendencia central, mientras que la varianza y la desviación típica son medidas de dispersión.

Entre las v.a.d. cabe destacar las denominadas *Binomial* y de *Poisson*.

Variable Binomial

Se considera un experimento aleatorio en el que puede ocurrir el suceso A o su contrario $\overline{A}$, con probabilidades p y $1-p$ respectivamente.

Si se realiza el experimento anterior n veces, la v.a.d. X que recoge el número de veces que ha ocurrido el suceso A, se denomina *Variable Binomial* y se designa por $X \sim B(n,\ p)$.

El recorrido de la variable X es $\{0, 1, ..., n\}$ y su función de probabilidad viene dada por

$$f(x) = \binom{n}{x} p^x \left(1-p\right)^{n-x}$$

cuya gráfica se puede apreciar en la figura 6.2.

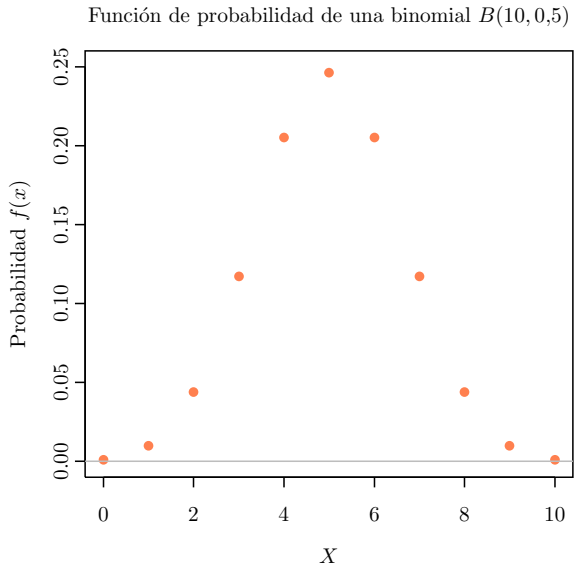


Figura 6.2 – Función de probabilidad de una variable aleatoria binomial de 10 repeticiones y probabilidad de éxito 0.5

A partir de la expresión anterior se puede demostrar que

$$\mu = np$$
$$\sigma^2 = np(1-p)$$
$$\sigma = +\sqrt{np(1-p)}$$

En el caso particular de que el experimento se realice una sola vez, la variable aleatoria recibe el nombre de Variable de Bernouilli. Una variable Binomial $X \sim B(n,\ p)$ se puede considerar como suma de n variables de Bernouilli idénticas con distribución $B(1,\ p)$.

Variable de Poisson

Las variables de Poisson surgen de la observación de un conjunto discreto de fenómenos puntuales en un soporte continuo de tiempo, longitud o espacio. Por ejemplo: nº de llamadas que llegan a una centralita telefónica en un tiempo establecido, nº de hematíes en un volumen de sangre, etc. Se supone además que en un soporte continuo suficientemente grande, el número medio de fenómenos ocurridos por unidad de soporte considerado, es una constante que designaremos por λ.

A la v.a.d. X, que recoge el número de fenómenos puntuales que ocurren en un intervalo de amplitud establecida, se le denomina *Variable de Poisson* y se designa por $X \sim P(\lambda)$.

El recorrido de la variable X es $\{0, 1, 2, ...\}$, no existiendo un valor máximo que pueda alcanzar. Su función de probabilidad viene dada por

$$f(x) = \frac{\lambda^x}{x!}\ e^{-\lambda}$$

y su gráfica aparece en la figura 6.3

Se puede demostrar que

$$\mu = \lambda$$
$$\sigma^2 = \lambda$$
$$\sigma = +\sqrt{\lambda}$$

La distribución de Poisson aparece como límite de la distribución Binomial cuando el número n de repeticiones del experimento es muy grande y la probabilidad p de que ocurra el suceso A considerado

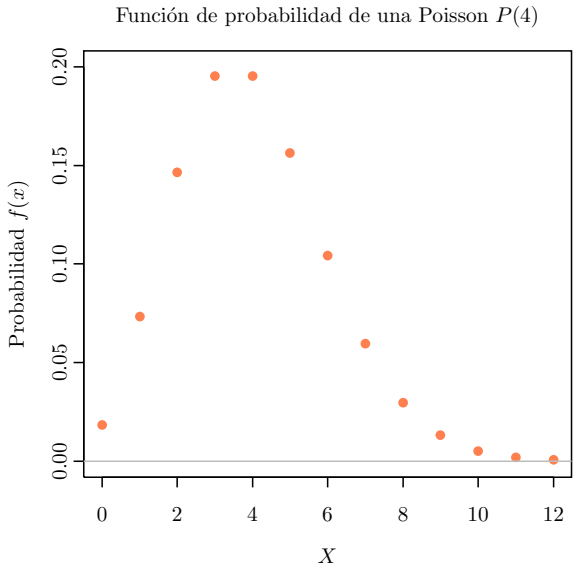


Figura 6.3 – Función de probabilidad de una variable aleatoria Poisson de media $\lambda = 4$

es muy pequeña. Por ello, la distribución de Poisson se llama también *Ley de los Casos Raros*. En la práctica se considera aceptable realizar los cálculos de probabilidades correspondientes a una variable $B(n,\ p)$ mediante las fórmulas correspondientes a una variable $P(\lambda)$ con $\lambda = np$, siempre que $n \geq 50$ y $p < 0{,}1$.

2 Ejercicios resueltos

1. La ley de los grandes números establece que cuando un experimento aleatorio se repite de manera indefinida, la frecuencia relativa de cada suceso tiende a estabilizarse en torno a un valor que es la probabilidad del suceso. Para comprobar la ley se realiza un experimento que consiste en lanzar un dado varias veces y anotar la frecuencia relativa de cada cara. Se pide:

   a) Lanzar el dado 10 veces y calcular las frecuencias relativas de las caras obtenidas y el diagrama de barras asociado.

   Indicación
   1) Seleccionar el menú `Distribuciones→Simulaciones→Lanzador de dados`.
   2) En el cuadro de diálogo que aparece, introducir 10 en el campo `Número de lanzamientos`, introducir un nombre para el conjunto de datos y hacer click en el botón aceptarAceptar.
   3) Hacer click en el botón del Conjunto de datos y en el cuadro de diálogo que aparece seleccionar el conjunto de datos creado y hacer click en el botón Aceptar.
   4) Seleccionar el menú `Estadísticos→Distribución de frecuencias→Tabla de frecuencias (datos numéricos no agrupados)`.
   5) En el cuadro de diálogo que aparece, seleccionar la variable V1 y hacer click en el botón Aceptar.
   6) Seleccionar el menú `Gráficas→Gráfica de barras`.
   7) En el cuadro de diálogo que aparece seleccionar la variable V1, marcar la opción `Frecuencias relativas` y hacer click en el botón Aceptar.
   8) Observar las diferencias entre las frecuencias relativas y en la altura de las barras.

   b) Repetir el apartado anterior para 100, 1000 y 1000000 lanzamientos. ¿Se cumple la ley de los grandes números? ¿En torno a qué valor se estabilizan las frecuencias relativas?

2. Sea X la variable que mide el número de caras obtenidas al lanzar 10 monedas. Para ver de manera experimental la distribución de probabilidad de X se realiza un experimento aleatorio que consiste en lanzar varias veces las 10 monedas y anotar el número de caras obtenido en cada lanzamiento. Se pide:

   a) Lanzar las 10 monedas 1000 veces y calcular las frecuencias relativas de las caras obtenidas y el diagrama de barras asociado.

   Indicación
   1) Seleccionar el menú `Distribuciones→Simulaciones→Lanzador de monedas`.
   2) En el cuadro de diálogo que aparece, introducir 10 en el campo `Número de monedas`, 1000 en el campo `Número de lanzamientos`, introducir un nombre para el conjunto de datos y hacer click en el botón aceptarAceptar.
   3) Hacer click en el botón del Conjunto de datos y en el cuadro de diálogo que aparece seleccionar el conjunto de datos creado y hacer click en el botón Aceptar.
   4) Seleccionar el menú `Estadísticos→Distribución de frecuencias→Tabla de frecuencias (datos numéricos no agrupados)`.
   5) En el cuadro de diálogo que aparece, seleccionar la variable sum y hacer click en el botón Aceptar.
   6) Seleccionar el menú `Gráficas→Gráfica de barras`.
   7) En el cuadro de diálogo que aparece seleccionar la variable sum, marcar la opción `Frecuencias relativas` y hacer click en el botón Aceptar.

   b) Generar la distribución de probabilidad de una variable Binomial $B(10\,,\,0{,}5)$ y compararla con la distribución de frecuencias relativas del apartado anterior.

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución binomial →Probabilidades binomiales`.
   2) En el cuadro de diálogo que aparece, introducir 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, y hacer click en el botón Aceptar.

   c) Dibujar la gráfica de la función de probabilidad de la Binomial $X \sim B(10\,,\,0{,}5)$ y compararla con el diagrama de barras de frecuencias relativas del primer apartado.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución binomial →Gráfica de la distribución binomial`.
2) En el cuadro de diálogo que aparece, introducir 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, marcar la opción `Gráfica de la función de probabilidad` y hacer click en el botón Aceptar.

d) Dibujar la gráfica de la función de distribución.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución binomial →Gráfica de la distribución binomial`.
2) En el cuadro de diálogo que aparece, introducir 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, marcar la opción `Gráfica de la función de distribución` y hacer click en el botón Aceptar.

e) Calcular $P(X = 7)$.

Indicación
Las probabilidades de valores aislados aparecen en la distribución de probabilidad obtenida en el primer apartado.

f) Calcular $P(X \leq 4)$.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución binomial →Probabilidades binomiales acumuladas`.
2) En el cuadro de diálogo que aparece, introducir 4 en el campo `Valor(es) de la variable`, 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

g) Calcular $P(X > 5)$.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución binomial →Probabilidades binomiales acumuladas`.
2) En el cuadro de diálogo que aparece, introducir 5 en el campo `Valor(es) de la variable`, 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

h) Calcular $P(2 \leq X < 9)$.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución binomial →Probabilidades binomiales acumuladas`.
2) En el cuadro de diálogo que aparece, introducir los valores 1, 8 en el campo `Valor(es) de la variable`, 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.
3) La probabilidad del intervalo $P(2 \leq X < 9)$ es la resta de las probabilidades obtenidas $P(X < 9) = P(X \leq 8)$ y $P(X < 2) = P(X \leq 1)$.

3. El número de nacimientos diarios en una determinada población sigue una distribución de Poisson de media 6 nacimientos al día. Se pide:

a) Generar la distribución de probabilidad de la variable.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución de Poisson →Probabilidades de Poisson`.
2) En el cuadro de diálogo que aparece, introducir el valor 6 en el campo `Media`, y hacer click en el botón Aceptar.

b) Dibujar la gráfica de la función de probabilidad.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución de Poisson →Gráfica de la distribución de Poisson`.
2) En el cuadro de diálogo que aparece, introducir el valor 6 en el campo `Media`, marcar la opción `Gráfica de la función de probabilidad` y hacer click en el botón Aceptar.

c) Dibujar la gráfica de la función de distribución.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución de Poisson →Gráfica de la distribución de Poisson`.
2) En el cuadro de diálogo que aparece, introducir el valor 6 en el campo `Media`, marcar la opción `Gráfica de la función de distribución` y hacer click en el botón Aceptar.

d) Calcular la probabilidad de un día no haya nacimientos.

Indicación
Las probabilidades de valores aislados aparecen en la distribución de probabilidad obtenida en el primer apartado.

e) Calcular la probabilidad de que un día haya menos de 6 nacimientos.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución de Poisson →Probabilidades de Poisson acumuladas`.
2) En el cuadro de diálogo que aparece, introducir 5 en el campo `Valor(es) de la variable`, 6 en el campo `Media`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

f) Calcular la probabilidad de que un día haya 4 o más nacimientos.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución de Poisson →Probabilidades de Poisson acumuladas`.
2) En el cuadro de diálogo que aparece, introducir 3 en el campo `Valor(es) de la variable`, 6 en el campo `Media`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

g) Calcular la probabilidad de que un día haya entre 4 y 8 nacimientos, inclusives.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución de Poisson →Probabilidades de Poisson acumuladas`.
2) En el cuadro de diálogo que aparece, introducir los valores 3,8 en el campo `Valor(es) de la variable`, 6 en el campo `Media`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.
3) La probabilidad del intervalo $P(4 \leq\leq 8)$ es la resta de las probabilidades obtenidas $P(X \leq 8)$ y $P(X < 4) = P(X \leq 3)$.

4. La ley de los casos raros dice que en una distribución Binomial $B(n\,,\,p)$, cuando $n \geq 30$ y $p \leq 0{,}1$ la distribución se parece mucho a una distribución Poisson $P(np)$. Para comprobar hasta qué punto se parecen esta distribuciones, se pide:

a) Generar la distribución de probabilidad de una variable Binomial $B(30\,,0{,}1)$.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones discretas→Distribución binomial →Probabilidades binomiales`.
2) En el cuadro de diálogo que aparece, introducir el valor 30 en el campo `Ensayos binomiales`, 0,1 en el campo `Probabilidad de éxito` y hacer click en el botón Aceptar.

b) Generar la distribución de probabilidad de una variable Poisson $P(3)$ y compararla con la de la binomial $B(30\,,0{,}1)$.

Indicación

1) Seleccionar el menú `Distribuciones`→`Distribuciones discretas`→`Distribución de Poisson`→`Probabilidades de Poisson`.
2) En el cuadro de diálogo que aparece, introducir el valor 3 en el campo `Media`, y hacer click en el botón Aceptar.

c) Generar la distribución de probabilidad de una variable Binomial $B(100\,,0{,}03)$ y compararla con la de la Poisson $P(3)$. ¿Se parecen más estas distribuciones que las anteriores?

Indicación

1) Seleccionar el menú `Distribuciones`→`Distribuciones discretas`→`Distribución binomial`→`Probabilidades binomiales`.
2) En el cuadro de diálogo que aparece, introducir el valor 100 en el campo `Ensayos binomiales`, 0,03 en el campo `Probabilidad de éxito` y hacer click en el botón Aceptar.

d) Dibujar las gráficas de las distribuciones anteriores y ver cuáles se parecen más. ¿Se cumple la ley de los casos raros?

Indicación

1) Seleccionar el menú `Distribuciones`→`Simulaciones`→`Ley de los casos raros`.
2) En el cuadro de diálogo que aparece, desplazar el deslizador de `n` hasta 30 y el de `p` hasta 0.1.
3) Después desplazar el deslizador de `n` hasta 100 y el de `p` hasta 0.03.

3 Ejercicios propuestos

1. Al lanzar 100 veces una moneda, ¿cuál es la probabilidad de obtener entre 40 y 60 caras inclusive?
2. La probabilidad de curación de un paciente al ser sometido a un determinado tratamiento es 0,85. Calcular la probabilidad de que en un grupo de 6 enfermos sometidos a tratamiento:
   a) Se curen la mitad.
   b) Se curen al menos 4.
3. La probabilidad de que al administrar una vacuna dé una determinada reacción es 0,001. Si se vacunan 2000 personas ¿cuál es la probabilidad de que aparezca alguna reacción adversa?
4. El número medio de llamadas por minuto que llegan a una centralita telefónica es igual a 120. Se pide:
   a) Dar la distribución de probabilidad del número de llamadas en 2 segundos y dibujar su gráfica.
   b) Calcular al probabilidad de que durante 2 segundos lleguen a la centralita menos de 4 llamadas.
   c) Calcular la probabilidad de que durante 3 segundos lleguen a la centralita 3 llamadas como mínimo.
5. Se sabe que la probabilidad de que aparezca una bacteria en un mm^3 de cierta disolución es de 0,002. Si en cada mm^3 a los sumo puede aparecer una bacteria, determinar la probabilidad de que en un cm^3 haya como máximo 5 bacterias.

Práctica de Estadística con R 7

Variables Aleatorias Continuas

1 Fundamentos teóricos

1.1 Variables Aleatorias

Se define una *variable aleatoria* asignando a cada resultado del experimento aleatorio un número. Esta asignación puede realizarse de distintas maneras, obteniéndose de esta forma diferentes variables aleatorias. Así, en el lanzamiento de dos monedas podemos considerar el número de caras o el número de cruces. En general, si los resultados del experimento son numéricos, se tomarán dichos números como los valores de la variable, y si los resultados son cualitativos, se hará corresponder a cada modalidad un número arbitrariamente.

Formalmente, una *variable aleatoria* X es una función real definida sobre los puntos del espacio muestral E de un experimento aleatorio.

$$X : E \to \mathbb{R}$$

De esta manera, la distribución de probabilidad del espacio muestral E, se transforma en una distribución de probabilidad para los valores de X.

El conjunto formado por todos los valores distintos que puede tomar la variable aleatoria se llama *Rango* o *Recorrido* de la misma.

Las variables aleatorias pueden ser de dos tipos: discretas o continuas. Una variable es *discreta* cuando sólo puede tomar valores aislados, mientras que es *continua* si puede tomar todos los valores posibles de un intervalo.

1.2 Variables Aleatorias Continuas (v.a.c.)

Se considera una v.a.c. X. En este tipo de variables, a diferencia de las discretas, la probabilidad de que la variable tome un valor aislado cualquiera es nula, y sólo hablaremos de probabilidades asociadas a intervalos.

Función de densidad

La *distribución de probabilidad* de X se suele caracterizar mediante una función $f(x)$, conocida como *función de densidad*. Formalmente, una función de densidad es una función no negativa, integrable en $\mathbb{R}$, que cumple

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

A partir de esta función, se puede calcular la probabilidad de que el valor de la variable pertenezca a un intervalo $[a, b]$, midiendo el área encerrada por dicha función y el eje de abscisas entre los límites del intervalo, como se observa en la figura 7.1, es decir

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

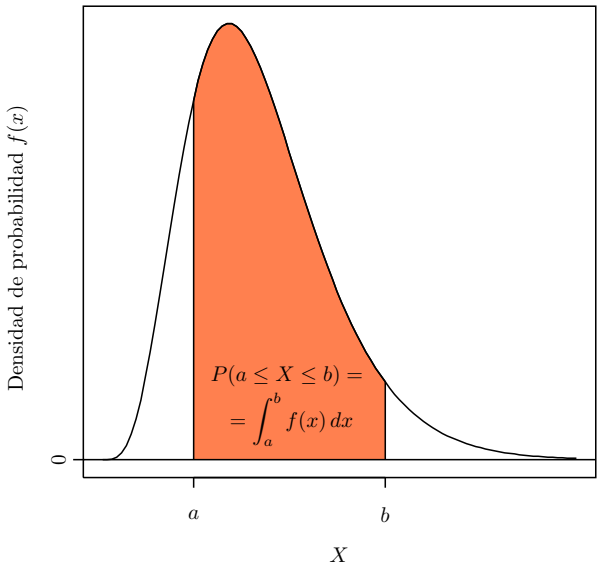


Figura 7.1 – En una v.a.c. la probabilidad asociada a un intervalo $[a, b]$, es el área que queda encerrada por la función de densidad y el eje de abscisas entre los límites del intervalo.

Función de distribución

Otra forma equivalente de caracterizar la distribución de probabilidad de X es mediante otra función $F(x)$, llamada *función de distribución*, que asigna a cada $x \in \mathbb{R}$ la probabilidad de que X tome un valor menor o igual que dicho número x. Así,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

A partir de la definición anterior es claro que la probabilidad de que la variable tome un valor en el intervalo [a,b] puede calcularse a partir de la función de distribución de la siguiente forma:

$$P(a \leq X \leq b) = \int_a^b f(x)dx = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = F(b) - F(a)$$

Estadísticos poblacionales

Los parámetros descriptivos más importantes de una v.a.c. X son:

Media o Esperanza

$$E[X] = \mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

Varianza

$$V[X] = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

Desviación típica

$$D[X] = \sigma = +\sqrt{\sigma^2}$$

La media es una medida de tendencia central, mientras que la varianza y la desviación típica son medidas de dispersión.

Distribución Uniforme Continua

Una v.a.c. X se dice que sigue una *Distribución Uniforme Continua* de parámetros a y b, y se designa por $X \sim U(a,\ b)$, si su recorrido es el intervalo $[a, b]$ y su función de densidad es

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b, \\ 0 & \text{en el resto} \end{cases}$$

Esta función es constante en el intervalo $[a, b]$ y nula fuera de él. Se cumple que

$$\mu = \frac{a+b}{2} \qquad \sigma = +\frac{b-a}{\sqrt{12}}.$$

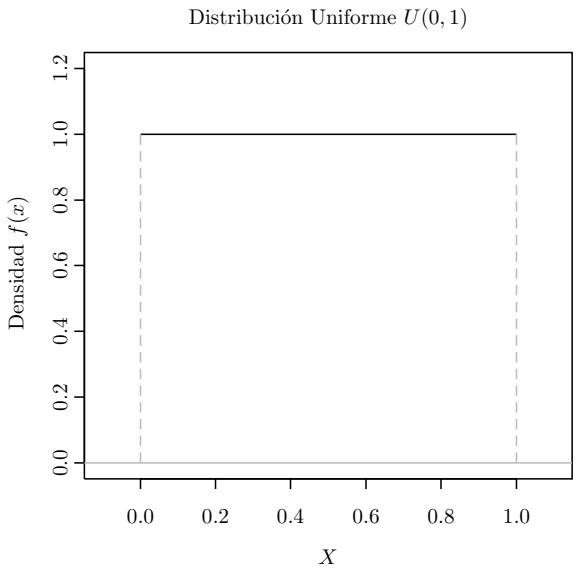


Figura 7.2 – Función de densidad de una variable aleatoria uniforme continua $U(0,\ 1)$.

Distribución Normal

Una v.a.c. X se dice que sigue una *Distribución Normal* o *Gaussiana* de media μ y desviación típica σ, y se designa por $X \sim N(\mu,\ \sigma)$, si su recorrido es todo $\mathbb{R}$ y su función de densidad es

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Esta función tiene forma acampanada y es simétrica con respeto a la media μ.

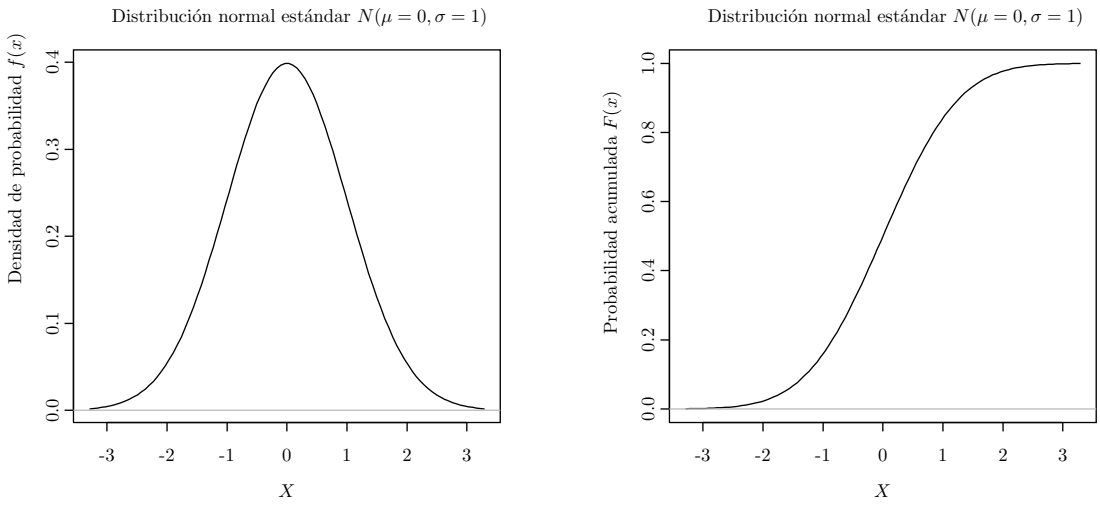
La distribución Normal es la distribución continua más importante, ya que muchos de los fenómenos que aparecen en la naturaleza presentan esta distribución. Ello es debido a que, como establece el *Teorema Central del Límite*, cuando los resultados de un experimento están influidos por muchas causas independientes que actúan sumando sus efectos, se puede esperar que dichos resultados sigan una distribución normal.

La v.a.c normal de media 0 y desviación típica 1, $Z \sim N(0, 1)$, se conoce como *variable normal estándar* o *tipificada* y se utiliza muy a menudo. Su función de densidad aparece en la figura 7.3(a) y su función de distribución en la figura 7.3(b).

Distribución Chi-cuadrado

Si $Z_1, \ldots, Z_n$ son n v.a.c. normales estándar independientes, entonces la variable

$$X = Z_1^2 + \cdots + Z_n^2$$



(a) Función de densidad.

(b) Función de distribución.

Figura 7.3 – Función de densidad y función de distribución de la variable aleatoria continua Z Normal de media 0 y desviación típica 1 $\quad Z \sim N(0, 1)$

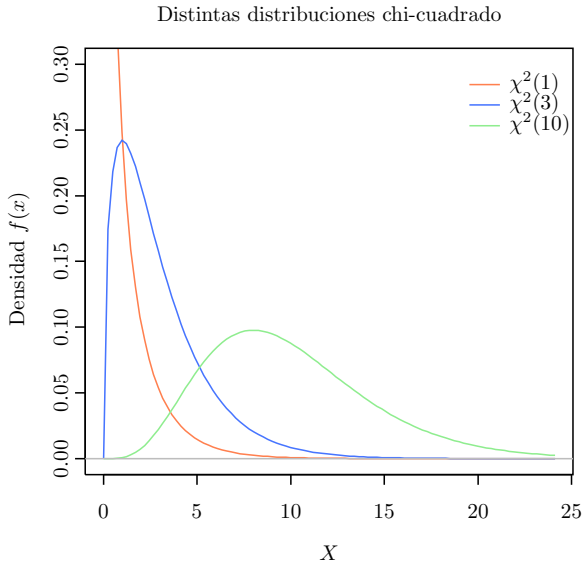


Figura 7.4 – Función de densidad de una variable aleatoria Chi cuadrado de 6 grados de libertad

se dice que sigue una distribución *Chi-cuadrado* con n grados de libertad, y se nota $X \sim \chi^2(n)$.

Se cumple que

$$\mu = n$$
$$\sigma = +\sqrt{2n}$$

La distribución Chi-cuadrado se utiliza en inferencia estadística para cálculos de intervalos de confianza y contrastes de hipótesis sobre la varianza de la población.

Distribución T de Student

Si Z es una v.a.c. normal estándar y X es una v.a.c chi-cuadrado con n grados de libertad, ambas variables independientes, entonces la variable

$$T = \frac{Z}{\sqrt{X/n}}$$

se dice que sigue una distribución T *de Student* con n grados de libertad, y se nota $T \sim T(n)$.

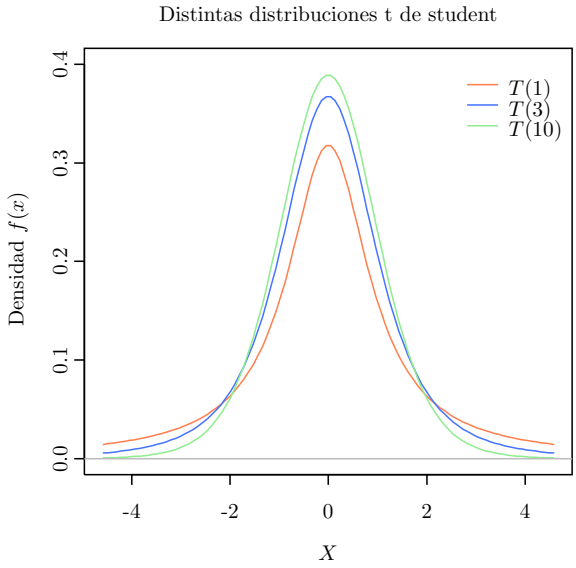


Figura 7.5 – Función de densidad de una variable aleatoria t de student de 10 grados de libertad

Esta variable es muy parecida a la normal estándar pero un poco menos apuntada, y se parece más a ésta a medida que aumentan los grados de libertad, de manera que para $n \geq 30$ ambas distribuciones se consideran prácticamente iguales. Se cumple que

$$\mu = 0$$
$$\sigma = +\sqrt{n/(n-2)} \quad \text{con } n > 2$$

La distribución T de Student se utiliza en inferencia estadística para cálculos de intervalos de confianza y contrastes de hipótesis sobre la media de la población.

Distribución F de Fisher-Snedecor

Si X e Y son dos v.a.c chi-cuadrado con m y n grados de libertad respectivamente, ambas variables independientes, entonces la variable

$$F = \frac{X/m}{Y/n}$$

se dice que sigue una distribución F de Fisher-Snedecor con m y n grados de libertad, y se denota $F \sim F(m,\ n)$.

$$\mu = \frac{n}{n-2}$$
$$\sigma = +\sqrt{\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}} \quad \text{con } n > 4$$

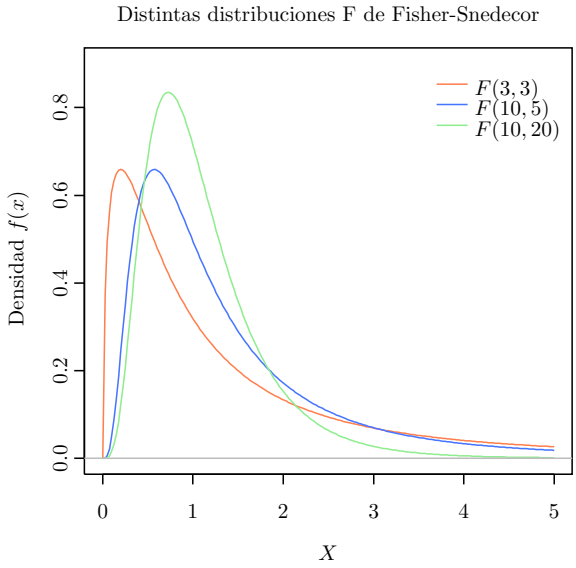


Figura 7.6 – Función de densidad de una variable aleatoria F de Fisher-Snedecor de 6 y 8 grados de libertad

De la definición se deduce fácilmente que $F(m,\ n) = \dfrac{1}{F(n,\ m)}$, y si llamamos $F(m,\ n)_p$ al valor que cumple que $P(F(m,\ n) \leq F(m,\ n)_p) = p$, entonces se verifica

$$F(m,\ n)_p = \frac{1}{F(n,\ m)_{1-p}}$$

La distribución F de Fisher-Snedecor se utiliza en inferencia estadística para cálculos de intervalos de confianza y contrastes de hipótesis sobre el cociente de varianzas de dos poblaciones, y en análisis de la varianza.

2 Ejercicios resueltos

1. Supongase que un autobús pasa por una parada cada 15 minutos y que una persona puede llegar a la parada en cualquier instante, entonces la variable que mide el tiempo que la persona espera al autobús es una variable Uniforme contnua $U(0, 15)$, ya que cualquier valor entre 0 y 15 minutos es equiprobable. Se pide:

   a) Dibujar la gráfica de la función de densidad de la Uniforme $X \sim U(0, 15)$.

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución uniforme →Gráfica de la distribución uniforme`.
   2) En el cuadro de diálogo que aparece, introducir el valor 0 en el campo `Mínimo`, 15 en el campo `Máximo`, marcar la opción `Gráfica de la función de densidad` y hacer click en el botón Aceptar.

   b) Dibujar la gráfica de la función de distribución.

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución uniforme →Gráfica de la distribución uniforme`.
   2) En el cuadro de diálogo que aparece, introducir valor 0 en el campo `Mínimo`, 15 en el campo `Máximo`, marcar la opción `Gráfica de la función de distribución` y hacer click en el botón Aceptar.

   c) Calcular la probabilidad de esperar al autobús menos de 5 minutos.

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución uniforme →Probabilidades uniformes`.
   2) En el cuadro de diálogo que aparece, introducir el valor 5 en el campo `Valor(es) de la variable`, 0 en el campo `Mínimo`, 15 en el campo `Máximo`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

   d) Calcular la probabilidad de esperar al autobús más de 12 minutos.

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución uniforme →Probabilidades uniformes`.
   2) En el cuadro de diálogo que aparece, introducir el valor 12 en el campo `Valor(es) de la variable`, 0 en el campo `Mínimo`, 15 en el campo `Máximo`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

   e) Calcular la probabilidad de esperar al autobús entre 5 y 10 minutos.

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución uniforme →Probabilidades uniformes`.
   2) En el cuadro de diálogo que aparece, introducir los valores 10,5 en el campo `Valor(es) de la variable`, 0 en el campo `Mínimo`, 15 en el campo `Máximo`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.
   3) La probabilidad del intervalo $P(5 \leq X \leq 10)$ es la resta de las probabilidades obtenidas $P(X \leq 10) - P(X \leq 5)$.

   f) ¿Por debajo de qué tiempo esperará al autobús la mitad de las veces?

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución uniforme →Cuantiles uniformes`.
   2) En el cuadro de diálogo que aparece, introducir la probabilidad 0.5 en el campo `Probabilidades`, 0 en el campo `Mínimo`, 15 en el campo `Máximo`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

g) ¿Por encima de qué tiempo esperará al autobús el 10 % de las veces?

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución uniforme →Cuantiles uniformes`.
2) En el cuadro de diálogo que aparece, introducir la probabilidad 0.1 en el campo `Probabilidades`, 0 en el campo `Mínimo`, 15 en el campo `Máximo`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

2. La variable aleatoria normal de media 0 y desviación típica 1, $Z \sim N(0,1)$, se conoce como normal estándar y es la variable normal más importante. Se pide:

a) Dibujar la gráfica de la función de densidad.

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución normal →Gráfica de la distribución normal`.
2) En el cuadro de diálogo que aparece, introducir el valor 0 el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Gráfica de la función de densidad` y hacer click en el botón Aceptar.

b) ¿Cómo afectan los dos parámetros de la normal, su media y su desviación típica, a la forma de la campana de Gauss?

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución normal →Gráfica interactiva de la distribución normal`.
2) En el cuadro de diálogo que aparece, desplazar el deslizador de la media por distintos valores y ver cómo cambia la forma de la campana.
3) Después desplazar el deslizador de la desviación típica por distintos valores y ver cómo cambia la forma de la campana.

c) Dibujar la gráfica de la función de distribución.

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución normal →Gráfica de la distribución normal`.
2) En el cuadro de diálogo que aparece, introducir el valor 0 en el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Gráfica de la función de distribución` y hacer click en el botón Aceptar.

d) Calcular la probabilidad de que la normal estándar tome un valor menor que -1.

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución normal →Probabilidades normales`.
2) En el cuadro de diálogo que aparece, introducir el valor -1 en el campo `Valor(es) de la variable`, 0 en el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

e) Calcular la probabilidad de que la normal estándar tome un valor mayor que 1.

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución normal →Probabilidades normales`.
2) En el cuadro de diálogo que aparece, introducir el valor 1 en el campo `Valor(es) de la variable`, 0 en el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

f) Calcular la probabilidad de que la normal estándar tome un valor entre -1 (la media menos la desviación típica) y 1 (la media más la desviación típica).

Indicación

1) Seleccionar el menú `Distribuciones`→`Distribuciones continuas`→`Distribución normal`→`Probabilidades normales`.
2) En el cuadro de diálogo que aparece, introducir los valores 1,−1 en el campo `Valor(es) de la variable`, 0 en el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.
3) La probabilidad del intervalo $P(-1 \leq Z \leq 1)$ es la resta de las probabilidades obtenidas $P(Z \leq 1) - P(Z \leq -1)$.

g) Calcular la probabilidad de que la normal estándar tome un valor entre −2 (la media menos dos veces la desviación típica) y 2 (la media más dos veces la desviación típica).

Indicación

1) Seleccionar el menú `Distribuciones`→`Distribuciones continuas`→`Distribución normal`→`Probabilidades normales`.
2) En el cuadro de diálogo que aparece, introducir los valores 2,−2 en el campo `Valor(es) de la variable`, 0 en el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.
3) La probabilidad del intervalo $P(-2 \leq Z \leq 2)$ es la resta de las probabilidades obtenidas $P(Z \leq 2) - P(Z \leq -2)$.

h) Calcular la probabilidad de que la normal estándar tome un valor entre −3 (la media menos tres veces la desviación típica) y 3 (la media más tres veces la desviación típica).

Indicación

1) Seleccionar el menú `Distribuciones`→`Distribuciones continuas`→`Distribución normal`→`Probabilidades normales`.
2) En el cuadro de diálogo que aparece, introducir los valores 3,−3 en el campo `Valor(es) de la variable`, 0 el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.
3) La probabilidad del intervalo $P(-3 \leq Z \leq 3)$ es la resta de las probabilidades obtenidas $P(Z \leq 3) - P(Z \leq -3)$.

i) Calcular los cuartiles.

Indicación

1) Seleccionar el menú `Distribuciones`→`Distribuciones continuas`→`Distribución normal` →`Cuantiles normales`.
2) En el cuadro de diálogo que aparece, introducir las probabilidades 0.25, 0.5, 0.75 en el campo `Probabilidades`, 0 en el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

j) Calcular el valor que deja acumulada por debajo una probabilidad 0,95.

Indicación

1) Seleccionar el menú `Distribuciones`→`Distribuciones continuas`→`Distribución normal` →`Cuantiles normales`.
2) En el cuadro de diálogo que aparece, introducir la probabilidad 0.95 en el campo `Probabilidades`, 0 en el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

k) Calcular el valor que deja acumulada por encima una probabilidad 0,025.

Indicación

1) Seleccionar el menú `Distribuciones`→`Distribuciones continuas`→`Distribución normal` →`Cuantiles normales`.
2) En el cuadro de diálogo que aparece, introducir la probabilidad 0.025 en el campo `Probabilidades`, 0 en el campo `media`, 1 en el campo `desviación típica`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

3. El teorema central del límite establece que la variable resultante de sumar 30 o más variables independientes sigue una distribución normal de media la suma de las medias de cada una de las variables y de varianza la suma de sus varianzas. Esta es la explicación de que una gran parte de las variables

continuas que aparecen en la naturaleza sean variables normales. Para observar de manera experimental el teorema central del límite se realiza un experimento que consiste en lanzar varios dados muchas veces y sumar los valores obtenidos. Se pide:

a) Simular el lanzamiento de un dado 100000 veces y dibujar el diagrama de barras asociado. ¿Tiene forma de campana de Gauss?

Indicación

1) Seleccionar el menú `Distribuciones→Simulaciones→Lanzador de dados`.
2) En el cuadro de diálogo que aparece, introducir 100000 en el campo `Número de lanzamientos`, introducir un nombre para el conjunto de datos y hacer click en el botón aceptarAceptar.
3) Hacer click en el botón del Conjunto de datos y en el cuadro de diálogo que aparece seleccionar el conjunto de datos creado y hacer click en el botón Aceptar.
4) Seleccionar el menú `Gráficas→Gráfica de barras`.
5) En el cuadro de diálogo que aparece seleccionar la variable sum, marcar la opción `Frecuencias relativas` y hacer click en el botón Aceptar.

b) Repetir el apartado anterior con 2 y 30 dados. ¿Se cumple el teorema central del límite?

4. La suma de n variables normales estándar independientes elevadas al cuadrado es una variable con distribución Chi-cuadrado con n grados de libertad $\chi^2(n)$. Sea X una variable Chi-cuadrado con 6 grados de libertad $\chi^2(6)$. Se pide:

a) Dibujar la gráfica de la función de densidad.

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución Chi-cuadrado →Gráfica de la distribución Chi-cuadrado`.
2) En el cuadro de diálogo que aparece, introducir el valor 6 en el `Grados de libertad`, marcar la opción `Gráfica de la función de densidad` y hacer click en el botón Aceptar.

b) Calcular la probabilidad de que la variable tome un valor menor que 6.

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución Chi-cuadrado →Probabilidades Chi-cuadrado`.
2) En el cuadro de diálogo que aparece, introducir el valor 6 en el campo `Valor(es) de la variable`, 6 en el campo `Grados de libertad`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

c) Calcular el valor que deja acumulada por debajo una probabilidad 0,05.

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución Chi-cuadrado →Cuantiles Chi-cuadrado`.
2) En el cuadro de diálogo que aparece, introducir la probabilidad 0.05 en el campo `Probabilidades`, 6 en el campo `Grados de libertad`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

d) Calcular el valor que deja acumulada por arriba una probabilidad 0,1.

Indicación

1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución Chi-cuadrado →Cuantiles Chi-cuadrado`.
2) En el cuadro de diálogo que aparece, introducir la probabilidad 0.1 en el campo `Probabilidades`, 6 en el campo `Grados de libertad`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

5. La variable que se obtiene al dividir una normal estándar entre la raíz de una variable Chi-cuadrado de n grados de libertad dividida por n, sigue una distribución t de student de n grados de libertad $T(n)$. Sea X una variable t de student de 8 grados de libertad $T(8)$. Se pide:

a) Dibujar la gráfica de la función de probabilidad y compararla con la de la normal estándar.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución t →Gráfica de la distribución t`.
2) En el cuadro de diálogo que aparece, introducir el valor 8 en el campo `Grados de libertad`, marcar la opción `Gráfica de la función de densidad` y hacer click en el botón Aceptar.

b) Calcular el percentil octavo.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución t →Cuantiles t`.
2) En el cuadro de diálogo que aparece, introducir la probabilidad 0.08 en el campo `Probabilidades`, 8 en el campo `Grados de libertad`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

c) Calcular el valor por encima del cual está el 5 % de la población.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución t →Cuantiles t`.
2) En el cuadro de diálogo que aparece, introducir la probabilidad 0.05 en el campo `Probabilidades`, 8 en el campo `Grados de libertad`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

6. La variable resultante de dividir una variable Chi-cuadrado de n grados de libertad dividida por n, entre una variable Chi-cuadrado de m grados de libertad dividida por m, sigue un modelo de distribución F de Fisher de n y m grados de libertad $F(n, m)$. Sea X una variable F de Fisher de 10 y 20 grados de libertad $F(10, 20)$. Se pide:

a) Dibujar la gráfica de la función de densidad

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución F →Gráfica de la distribución F`.
2) En el cuadro de diálogo que aparece, introducir 10 el campo `Grados de libertad del numerador`, introducir 20 en el campo `Grados de libertad del denominador`, marcar la opción `Gráfica de la función de densidad` y hacer click en el botón Aceptar.

b) Calcular la probabilidad acumulada por encima de 1.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución F →Probabilidades F`.
2) En el cuadro de diálogo que aparece, introducir el valor 1 en el campo `Valor(es) de la variable`, 10 en el campo `Grados de libertad del numerador`, 20 en el campo `Grados de libertad del denominador`, marcar la opción `Cola derecha` y hacer click en el botón Aceptar.

c) Calcular el rango intercuartílico.

Indicación
1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución F →Cuantiles F`.
2) En el cuadro de diálogo que aparece, introducir las probabilidades 0.75, 0.25 en el campo `Probabilidades`, 10 en el campo `Grados de libertad del numerador`, 20 en el campo `Grados de libertad del denominador`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.
3) El rango intercuartílico es la resta de los valores obtenidos correspondientes al tercer y primer cuartiles.

3 Ejercicios propuestos

1. Entre los diabéticos, el nivel de glucosa en la sangre en ayunas X, puede suponerse de distribución aproximadamente normal, con media 106mg/100ml y desviación típica 8mg/100ml.

a) Hallar $P(X \leq 120\text{mg}/100\text{ml})$

b) ¿Qué porcentaje de diabéticos tendrá niveles entre 90 y 120mg/100ml?

c) Encontrar un valor que tenga la propiedad de que el 25 % de los diabéticos tenga un nivel de glucosa por debajo de dicho valor.

2. Se sabe que el nivel de colesterol en varones de más de 30 años de una determinada población sigue una distribución normal, de media 220mg/dl y desviación típica 30mg/dl. Si la población tiene 20000 varones mayores de 30 años,

   a) ¿Cuántos se espera que tengan su nivel de colesterol entre 210mg/dl y 240mg/dl?

   b) ¿Cuántos se espera que tengan su nivel de colesterol por encima de 250mg/dl?

   c) ¿Cuál será el nivel de colesterol por encima del cual se espera que esté el 20 % de la población?

3. Calcular la probabilidad de obtener entre 40 y 60 caras, inclusive, al lanzar 100 veces una moneda. Utilizar la aproximación de la distribución binomial mediante una normal.

Intervalos de Confianza para Medias y Proporciones

1 Fundamentos teóricos

1.1 Inferencia Estadística y Estimación de Parámetros

El objetivo de un estudio estadístico es doble: describir la muestra elegida de una población en la que se quiere estudiar alguna característica, y realizar inferencias, es decir, sacar conclusiones y hacer predicciones sobre la población de la que se ha extraído dicha muestra.

La metodología que conduce a obtener conclusiones sobre la población, basadas en la información contenida en la muestra, constituye la *Inferencia Estadística*.

Puesto que la muestra contiene menos información que la población, las predicciones serán aproximadas. Por eso, uno de los objetivos de la inferencia estadística es determinar la probabilidad de que una conclusión obtenida a partir del análisis de una muestra sea cierta, y para ello se apoya en la teoría de la probabilidad.

Cuando se desea conocer el valor de alguno de los parámetros de la población, el procedimiento a utilizar es la *Estimación de Parámetros,* que a su vez se divide en *Estimación Puntual*, cuando se da un único valor como estimación del parámetro poblacional considerado, y *Estimación por Intervalos*, cuando interesa conocer no sólo un valor aproximado del parámetro sino también la precisión de la estimación. En este último caso el resultado es un intervalo, dentro del cual estará, con una cierta confianza, el verdadero valor del parámetro poblacional. A este intervalo se le denomina *intervalo de confianza*. A diferencia de la estimación puntual, en la que se utiliza un único estimador, en la estimación por intervalo se emplean dos estimadores, uno para cada extremo del intervalo.

1.2 Intervalos de Confianza

Dados dos estadísticos muestrales L_1 y L_2, se dice que el intervalo $I = (L_1,\ L_2)$ es un *Intervalo de Confianza* para un parámetro poblacional θ, con *nivel de confianza* $1-\alpha$ (o *nivel de significación* α), si la probabilidad de que los estadísticos que determinan los límites del intervalo tomen valores tales que θ esté comprendido entre ellos, es igual a $1-\alpha$, es decir,

$$P\left(L_1 < \theta < L_2\right) = 1-\alpha$$

Los extremos del intervalo son variables aleatorias cuyos valores dependen de la muestra considerada. Es decir, los extremos inferior y superior del intervalo serían $L_1\left(X_1,...,X_n\right)$ y $L_2\left(X_1,...,X_n\right)$ respectivamente, aunque habitualmente escribiremos L_1 y L_2 para simplificar la notación. Designaremos mediante l_1 y l_2 los valores que toman dichas variables para una muestra determinada $\left(x_1,...,x_n\right)$.

Cuando en la definición se dice que la probabilidad de que el parámetro θ esté en el intervalo $(L_1,\ L_2)$ es $1-\alpha$, quiere decir que en el $100\left(1-\alpha\right)$ % de las posibles muestras, el valor de θ estaría en los correspondientes intervalos $(l_1,\ l_2)$.

Una vez que se tiene una muestra, y a partir de ella se determina el intervalo correspondiente $(l_1,\ l_2)$, no tendría sentido hablar de la probabilidad de que el parámetro θ esté en el intervalo $(l_1,\ l_2)$, pues al ser l_1 y l_2 números, el parámetro θ, que también es un número, aunque desconocido, estará o no estará en dicho intervalo, y por ello hablamos de confianza en lugar de probabilidad.

Así, cuando hablemos de un intervalo de confianza para el parámetro θ con nivel de confianza $1-\alpha$, entenderemos que antes de tomar una muestra, hay una probabilidad $1-\alpha$ de que el intervalo que se

construya a partir de ella, contenga el valor del parámetro θ. O, dicho de otro modo, si tomasemos 100 muestras del mismo tamaño y calculásemos sus respectivos intervalos, el $1-\alpha\,\%$ de estos contendrían el verdadero valor del parámetro a estimar (ver figura 8.1).

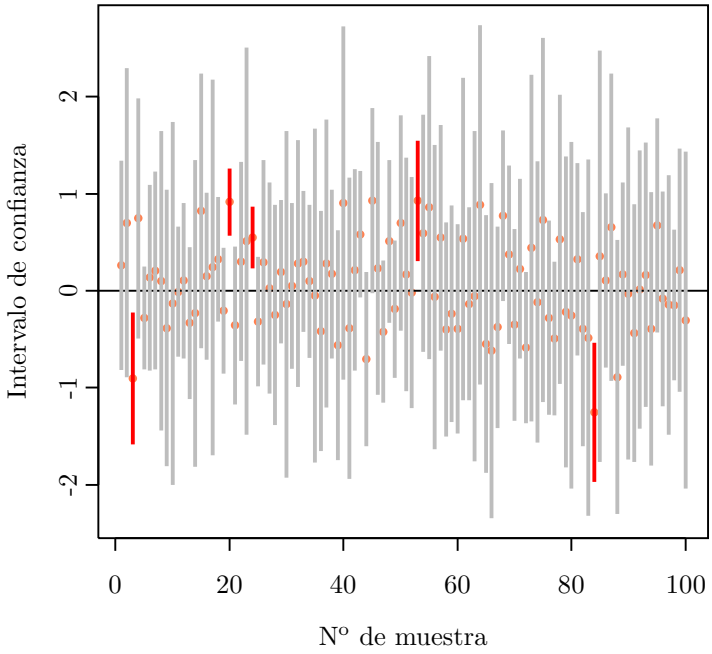


Figura 8.1 – Intervalos de confianza del 95 % para la media de 100 muestras tomadas de una población normal $N(0,1)$. Como se puede apreciar, de los 100 intervalos, sólo 5 no contienen el valor de la media real $\mu = 0$.

Cuando se realiza la estimación de un parámetro mediante un intervalo de confianza, el nivel de confianza se suele fijar a niveles altos (los más habituales son 0,90, 0,95 ó 0,99), para tener una alta confianza de que el parámetro está dentro del intervalo. Por otro lado, también interesa que la amplitud del intervalo sea pequeña para delimitar con precisión el valor del parámetro poblacional (esta amplitud del intervalo se conoce como *imprecisión* de la estimación). Pero a partir de una muestra, cuanto mayor sea el nivel de confianza deseado, mayor amplitud tendrá el intervalo y mayor imprecisión la estimación, y si se impone que la estimación sea más precisa (menor imprecisión), el nivel de confianza correspondiente será más pequeño. Por consiguiente, hay que llegar a una solución de compromiso entre el nivel de confianza y la precisión de la estimación. No obstante, si con la muestra disponible no es posible obtener un intervalo de amplitud suficientemente pequeña (imprecisión pequeña) con un nivel de confianza aceptable, hay que emplear una muestra de mayor tamaño. Al aumentar el tamaño muestral se consiguen intervalos de menor amplitud sin disminuir el nivel de confianza, o niveles de confianza más altos manteniendo la amplitud.

Intervalos de confianza para la media

Apoyándose en conclusiones extraídas del Teorema Central del Límite se obtiene que, siempre que las muestras sean grandes (como criterio habitual se toma que el tamaño muestral, n, sea mayor o igual que 30), e independientemente de la distribución original de la variable de partida X, de media μ y desviación típica σ, la variable

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

sigue una distribución Normal tipificada, $N(0,1)$.

Si la desviación típica σ de la variable de partida es desconocida, se utiliza como estimación la

cuasidesviación típica muestral:

$$\hat{S} = \sqrt{\frac{\sum \left(X_i - \overline{X}\right)^2}{n-1}}$$

y con ello, la nueva variable

$$\frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}}$$

sigue una distribución t de Student con $n-1$ grados de libertad, $T(n-1)$.

Para muestras pequeñas ($n < 30$) también pueden aplicarse los resultados anteriores, siempre y cuando la variable aleatoria de partida X, siga una distribución Normal.

A partir de lo anterior y teniendo en cuenta los tres factores de clasificación expuestos: si la población de partida en la que obtenemos la muestra sigue o no una distribución Normal, si la varianza de dicha población es conocida o desconocida, y si la muestra es grande ($n \geq 30$) o no, pueden deducirse las siguientes expresiones correspondientes a los diferentes intervalos de confianza.

Intervalo de confianza para la media de una población normal con varianza conocida en muestras de cualquier tamaño

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

En la figura 8.2 aparece un esquema explicativo de la construcción de este intervalo.

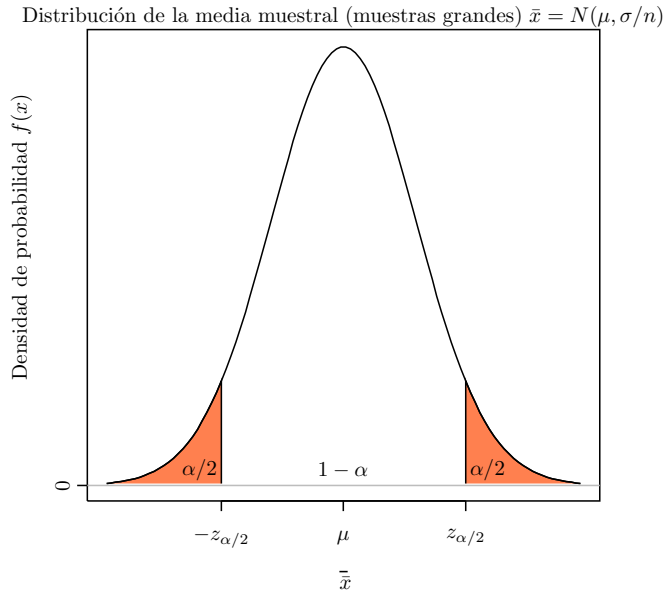


Figura 8.2 – Cálculo del intervalo de confianza para la media de una población normal con varianza conocida, a partir de las distribución de la media muestral $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$ para muestras grandes ($n \geq 30$).

Intervalo de confianza para la media de una población normal con varianza desconocida en muestras de cualquier tamaño

$$\left(\overline{x} - t^{n-1}_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + t^{n-1}_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Si las muestras son grandes ($n \geq 30$) el anterior intervalo puede aproximarse mediante:

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Intervalo de confianza para la media de una población no normal, varianza conocida y muestras grandes ($n \geq 30$)

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Intervalo de confianza para la media de una población no normal, varianza desconocida y muestras grandes ($n \geq 30$)

$$\left(\overline{x} - t^{n-1}_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + t^{n-1}_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Al tratarse de muestras grandes, el anterior intervalo puede aproximarse por:

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Si la población de partida no es normal, y las muestras son pequeñas, no puede aplicarse el Teorema Central del Límite y no se obtienen intervalos de confianza para la media.

Para cualquiera de los anteriores intervalos:

- n es el tamaño de la muestra.
- $\overline{x}$ es la media muestral.
- σ es la desviación típica de la población.
- $\hat{s}$ es la cuasidesviación típica muestral: $\hat{s}^2 = \dfrac{\sum (x_i - \overline{x})^2}{n-1}$.
- $z_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución Normal tipificada.
- $t^{n-1}_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución t de Student con $n-1$ grados de libertad.

Intervalos de confianza para la proporción poblacional p

Para muestras grandes ($n \geq 30$) y valores de p (probabilidad de "éxito") cercanos a 0,5, la distribución Binomial puede aproximarse mediante una Normal de media np y desviación típica $\sqrt{np(1-p)}$. En la práctica, para que sea válida dicha aproximación, se toma el criterio de que tanto np como $n(1-p)$ deben ser mayores que 5. Esto hace que también podamos construir intervalos de confianza para proporciones tomando éstas como medias de variables dicotómicas en las que la presencia o ausencia de la característica objeto de estudio ("éxito" ó "fracaso") se expresan mediante un 1 ó un 0 respectivamente.

De este modo, en muestras grandes y con distribuciones binomiales no excesivamente asimétricas (tanto np como $n(1-p)$ deben ser mayores que 5), si denominamos $\widehat{p}$ a la proporción de individuos que presentan el atributo estudiado en la muestra concreta, entonces el intervalo de confianza para la proporción con un nivel de significación α viene dado por:

$$\left(\widehat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p} \cdot (1-\widehat{p})}{n}}\ ,\ \widehat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p} \cdot (1-\widehat{p})}{n}}\right)$$

donde:

- n es el tamaño muestral.

– $\widehat{p}$ a la proporción de individuos que presentan el atributo estudiado en la muestra concreta.

– $z_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución Normal tipificada.

En muestras pequeñas o procedentes de una Binomial fuertemente asimétrica ($np \leq 5$ ó $n(1-p) \leq 5$) no puede aplicarse el Teorema Central del Límite y la construcción de intervalos de confianza debe realizarse a partir de la distribución Binomial.

2 Ejercicios resueltos

1. Se analiza la concentración de principio activo en una muestra de 10 envases tomados de un lote de un fármaco, obteniendo los siguientes resultados en mg/mm^3:

$$17{,}6 - 19{,}2 - 21{,}3 - 15{,}1 - 17{,}6 - 18{,}9 - 16{,}2 - 18{,}3 - 19{,}0 - 16{,}4$$

   Se pide:

   a) Crear un conjunto de datos con la variable concentracion.

   b) Calcular el intervalo de confianza para la media de la concentración del lote con nivel de confianza del 95 % (nivel de significación $\alpha = 0{,}05$).

   Indicación
   1) Seleccionar el menú `Estadísticos→Medias→Test t para una muestra`.
   2) En el cuadro de diálogo que aparece seleccionar la variable concentracion, introducir 0,95 en el campo `Nivel de confianza` y hacer click sobre el botón Aceptar.
   3) El intervalo de confianza aparece en la ventana de resultados justo después de la frase 95 percent confidence interval:.

   c) Calcular los intervalos de confianza para la media con niveles del 90 % y del 99 % (niveles de significación $\alpha = 0{,}1$ y $\alpha = 0{,}01$).

   Indicación
   Repetir los mismos pasos del apartado anterior, cambiando el nivel de confianza para cada intervalo.

   d) Si definimos la precisión del intervalo como la inversa de su amplitud, ¿cómo afecta a la precisión del intervalo de confianza el tomar niveles de significación cada vez más altos? ¿Cuál puede ser la explicación?

   e) ¿Qué tamaño muestral sería necesario para obtener una estimación del contenido medio de principio activo con un margen de error de $\pm 0{,}5$ mg/mm^3 y una confianza del 95 %?

   Indicación
   1) Seleccionar el menú `Estadísticos→Resúmenes→Resumen descriptivo`.
   2) En el cuadro de diálogo que aparece seleccionar la variable concenctracion, marcar los estadísticos `Media` y `Cuasidesviación típica` y hacer click en el botón Aceptar.
   3) Seleccionar el menú `Estadísticos→Medias→Cálculo del tamaño muestral→Cálculo del tamaño muestral para una media`.
   4) En el cuadro de diálogo que aparece introducir la media muestral en el campo `Media`, la cuasidesviación típica muestral en el campo `Desviación típica`, el nivel de significación deseado, en este caso 0,05, en el campo `Nivel de significación`, el margen de error deseado, en este caso 0,5, en el campo `Error`, y hacer click en el botón Aceptar.
   5) El tamaño muestral requerido aparece en la ventana de resultados como n.

   f) Si, para que sea efectivo, el fármaco debe tener una concentración mínima de 16 mg/mm^3 de principio activo, ¿se puede aceptar el lote como bueno? Justificar la respuesta.

2. Una central de productos lácteos recibe diariamente la leche de dos granjas X e Y. Para analizar la calidad de la leche, durante una temporada, se controla el contenido de materia grasa de la leche que proviene de ambas granjas, con los siguientes resultados:

X		Y	
0,34	0,34	0,28	0,29
0,32	0,35	0,30	0,32
0,33	0,33	0,32	0,31
0,32	0,32	0,29	0,29
0,33	0,30	0,31	0,32
0,31	0,32	0,29	0,31
		0,33	0,32
		0,32	0,33

a) Crear un conjunto de datos con las variables grasa y granja.

b) Calcular el intervalo de confianza con un 95 % de confianza para el contenido medio de materia grasa de la leche sin tener en cuenta si la misma procede de una u otra granja.

Indicación

1) Seleccionar el menú `Estadísticos→Medias→Test t para una muestra`.
2) En el cuadro de diálogo que aparece seleccionar la variable grasa, introducir 0,95 en el campo `Nivel de confianza` y hacer click sobre el botón Aceptar.
3) El intervalo de confianza aparece en la ventana de resultados justo después de la frase 95 percent confidence interval:.

c) Calcular los intervalos de confianza con un 95 % de confianza para el contenido medio de materia grasa de la leche dividiendo los datos según la granja de procedencia de la leche.

Indicación

1) Seleccionar el menú `Datos→Conjunto de datos activo→Filtrar el conjunto de datos activo`.
2) En el cuadro de diálogo que aparece introducir la condición `granja=="X"` en el campo `Expresión de selección`, introducir un nombre para el nuevo conjunto de datos en el campo `Nombre del nuevo conjunto de datos` y hacer click sobre el botón Aceptar.
3) Seleccionar el menú `Estadísticos→Medias→Test t para una muestra`.
4) En el cuadro de diálogo que aparece seleccionar la variable grasa, introducir 0,95 en el campo `Nivel de confianza` y hacer click sobre el botón Aceptar.
5) El intervalo de confianza aparece en la ventana de resultados justo después de la frase 95 percent confidence interval:.
6) Repetir los mismos pasos para el intervalo de confianza de la granja Y, introduciendo la condición `granja=="Y"` en el campo `Expresión de selección`.

d) A la vista de los intervalos obtenidos en el punto anterior, ¿se puede concluir que existen diferencias significativas en el contenido medio de grasa según la procedencia de la leche? Justificar la respuesta.

3. En una encuesta realizada en una facultad, sobre si el alumnado utiliza habitualmente (al menos una vez a la semana) la biblioteca de la misma, se han obtenido los siguientes resultados:

Alumno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Respuesta	no	si	no	no	no	si	no	si	si	si	si	no	si	no	si	no	no

Alumno	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Respuesta	no	si	si	si	no	no	si	no	no	si	si	no	no	si	no	si	no

a) Crear un conjunto de datos con la variable respuesta.

b) Calcular el intervalo de confianza con $\alpha = 0{,}01$ para la proporción del alumnado que utiliza habitualmente la biblioteca.

Indicación

1) Seleccionar el menú `Datos→Modificar variables del conjunto de datos activo→Reordenar niveles de factor`.
2) En el cuadro de diálogo que aparece seleccionar el factor respuesta y hacer click sobre el botón Aceptar.
3) En el cuadro de diálgo que aparece asignar el valor 1 al nivel `si`, el valor 2 al nivel `no` y hacer click en el botón Aceptar.
4) Seleccionar el menú `Estadísticos→Proporciones→Test de proporciones para una muestra`.
5) En el cuadro de diálogo que aparece seleccionar la variable respuesta, introducir 0,99 en el campo `Nivel de confianza` y hacer click en el botón Aceptar.

El comando para el test de las proporciones siempre toma la proporción del primer nivel del factor, de ahí que haya que reordenar los niveles antes.

c) ¿Qué interpretación tiene dicho intervalo? ¿Cómo es su precisión?

d) ¿Qué tamaño muestral sería necesario para obtener una estimación del porcentaje de alumnos que utilizan regularmente la biblioteca con un margen de error de un 1 % y una confianza del 95 %? ¿Y si se sabe que el número de alumnos en la universidad es de 5000?

Indicación

1) Seleccionar el menú `Estadísticos→Resúmenes→Distribución de frecuencias`.
2) Copiar de la ventana de resultados la proporción muestral de personas que si utilizan regularmente la biblioteca.
3) Seleccionar el menú `Estadísticos→Proporciones→Cálculo del tamaño muestral→Cálculo del tamaño muestral para una proporción`.
4) En el cuadro de diálogo que aparece introducir la proporción muestral en el campo `p`, el nivel de significación deseado, en este caso 0,05, en el campo `Nivel de significación`, el margen de error deseado, en este caso 0,01, en el campo `Error`, y hacer click en el botón Aceptar.
5) El tamaño muestral requerido aparece en la ventana de resultados como n.

Para obener el tamaño muestral para una población finita, hay que repetir los pasos de antes pero introduciendo el tamaño de la población, 5000 en este caso, en el campo `Tamaño poblacional`.

4. El Ministerio de Sanidad está interesado en la elaboración de un intervalo de confianza para la proporción de personas mayores de 65 años con problemas respiratorios que han sido vacunadas en una determinada ciudad. Para ello, después de preguntar a 200 pacientes mayores de 65 años con problemas respiratorios en los hospitales de dicha ciudad, 154 responden afirmativamente.

   a) Calcular el intervalo de confianza al 95 % para la proporción de pacientes vacunados.

   Indicación

   1) Seleccionar el menú `Estadísticos→Proporciones→Test para una proporción`.
   2) En el cuadro de diálogo que aparece introducir 154 en el campo `Frecuencia muestral`, introducir 200 en el campo `Tamaño muestral`, introducir 0,95 en el campo `Nivel de confianza` y hacer click en el botón Aceptar.

   b) Si entre los objetivos del Ministerio se encontraba alcanzar una proporción del al menos un 70 % de vacunados en dicho colectivo, ¿se puede concluir que se han cumplido los objetivos? Justificar la respuesta.

3 Ejercicios propuestos

1. Para determinar el nivel medio de colesterol (en mg/dl) en la sangre de una población, se realizaron análisis sobre una muestra de 8 personas, obteniéndose los siguientes resultados:

   196 212 188 206 203 210 201 198

   Hallar los intervalos de confianza para la media del nivel de colesterol con niveles de significación 0,1, 0,05 y 0,01. ¿Se puede afirmar que el nivel de colesterol medio de la población está por debajo de 210 mg/dl?

2. Para tratar un determinado síndrome neurológico se utilizan dos técnicas A y B. En un estudio se tomó una muestra de 60 pacientes con dicho síndrome y se le aplicó la técnica A a 25 de ellos y la técnica B a los 35 restantes. De los pacientes tratados con la técnica A, 18 se curaron, mientras que de los tratados con la técnica B, se curaron 21. Calcular un intervalo de confianza del 95 % para la proporción de curaciones con cada técnica. ¿Qué intervalo es más preciso?

3. A las siguientes elecciones locales en una ciudad se presentan tres partidos: A, B y C. Con el objetivo de hacer una estimación sobre la proporción de voto que cada uno de ellos obtendrá, se realiza una encuesta en la que responden 300 personas, de las cuales 60 piensan votar a A, 80 a B, 90 a C, 15 en blanco y 55 abstenciones. Calcular un intervalo de confianza para la proporción de votos, sobre el total del censo, de cada uno de los partidos que se presentan.

4. El fichero `nations.txt` contiene información sobre el desarrollo de distintos países (tasa de uso de anticonceptivos (contraception), producto interior bruto per cápita (GDP), tasa de mortalidad infantil (infant.mortality) y tasa de fertilidad (TFR)). Se pide:

a) Importar el fichero `nations.txt` en un conjunto de datos.

b) Calcular el intervalo de confianza de la tasa de uso de anticonceptivos y de la tasa de fertilidad para los países con un producto interior bruto per cápita superior a 10000 US$ e inferiores a dicha cantidad. Interpretar los intervalos.

Práctica de Estadística con R 9

Intervalos de Confianza para la Comparación de 2 Poblaciones

1 Fundamentos teóricos

1.1 Inferencia Estadística y Estimación de Parámetros

El objetivo de un estudio estadístico es doble: describir la muestra elegida de una población en la que se quiere estudiar alguna característica, y realizar inferencias, es decir, sacar conclusiones y hacer predicciones sobre la población de la que se ha extraído dicha muestra.

La metodología que conduce a obtener conclusiones sobre la población, basadas en la información contenida en la muestra, constituye la *Inferencia Estadística*.

Puesto que la muestra contiene menos información que la población, las predicciones serán aproximadas. Por eso, uno de los objetivos de la inferencia estadística es determinar la probabilidad de que una conclusión obtenida a partir del análisis de una muestra sea cierta, y para ello se apoya en la teoría de la probabilidad.

Cuando se desea conocer el valor de alguno de los parámetros de la población, el procedimiento a utilizar es la *Estimación de Parámetros,* que a su vez se divide en *Estimación Puntual*, cuando se da un único valor como estimación del parámetro poblacional considerado, y *Estimación por Intervalos*, cuando interesa conocer no sólo un valor aproximado del parámetro sino también la precisión de la estimación. En este último caso el resultado es un intervalo, dentro del cual estará, con una cierta confianza, el verdadero valor del parámetro poblacional. A este intervalo se le denomina *intervalo de confianza*. A diferencia de la estimación puntual, en la que se utiliza un único estimador, en la estimación por intervalo emplearemos dos estimadores, uno para cada extremo del intervalo.

1.2 Intervalos de Confianza

Dados dos estadísticos muestrales L_1 y L_2, se dice que el intervalo $I = (L_1,\ L_2)$ es un *Intervalo de Confianza* para un parámetro poblacional θ, con *nivel de confianza* $1-\alpha$ (o *nivel de significación* α), si la probabilidad de que los estadísticos que determinan los límites del intervalo tomen valores tales que θ esté comprendido entre ellos, es igual a $1-\alpha$, es decir,

$$P\left(L_1 < \theta < L_2\right) = 1-\alpha$$

Los extremos del intervalo son variables aleatorias cuyos valores dependen de la muestra considerada. Es decir, los extremos inferior y superior del intervalo serían $L_1\left(X_1,...,X_n\right)$ y $L_2\left(X_1,...,X_n\right)$ respectivamente, aunque habitualmente escribiremos L_1 y L_2 para simplificar la notación. Designaremos mediante l_1 y l_2 los valores que toman dichas variables para una muestra determinada $\left(x_1,...,x_n\right)$.

Cuando en la definición se dice que la probabilidad de que el parámetro θ esté en el intervalo $(L_1,\ L_2)$ es $1-\alpha$, quiere decir que en el $100\left(1-\alpha\right)$ % de las posibles muestras, el valor de θ estaría en los correspondientes intervalos $(l_1,\ l_2)$.

Una vez que se tiene una muestra, y a partir de ella se determina el intervalo correspondiente $(l_1,\ l_2)$, no tendría sentido hablar de la probabilidad de que el parámetro θ esté en el intervalo $(l_1,\ l_2)$, pues al ser l_1 y l_2 números, el parámetro θ, que también es un número, aunque desconocido, estará o no estará en dicho intervalo, y por ello hablamos de confianza en lugar de probabilidad.

Así, cuando hablemos de un intervalo de confianza para el parámetro θ con nivel de confianza $1-\alpha$, entenderemos que antes de tomar una muestra, hay una probabilidad $1-\alpha$ de que el intervalo que se construya a partir de ella, contenga el valor del parámetro θ.

Cuando se realiza la estimación de un parámetro mediante un intervalo de confianza, el nivel de confianza se suele fijar a niveles altos (los más habituales son 0,90, 0,95 ó 0,99), para tener una alta confianza de que el parámetro está dentro del intervalo. Por otro lado, también interesa que la amplitud del intervalo sea pequeña para delimitar con precisión el valor del parámetro poblacional (esta amplitud del intervalo se conoce como *imprecisión* de la estimación). Pero a partir de una muestra, cuanto mayor sea el nivel de confianza deseado, mayor amplitud tendrá el intervalo y mayor imprecisión la estimación, y si se impone que la estimación sea más precisa (menor imprecisión), el nivel de confianza correspondiente será más pequeño. Por consiguiente, hay que llegar a una solución de compromiso entre el nivel de confianza y la precisión de la estimación. No obstante, si con la muestra disponible no es posible obtener un intervalo de amplitud suficientemente pequeña (imprecisión pequeña) con un nivel de confianza aceptable, hay que emplear una muestra de mayor tamaño. Al aumentar el tamaño muestral se consiguen intervalos de menor amplitud sin disminuir el nivel de confianza, o niveles de confianza más altos manteniendo la amplitud.

Intervalos de confianza para la diferencia de medias

De igual manera a como ocurría con los intervalos de confianza para la media de una variable, apoyándose en conclusiones extraídas del Teorema Central del Límite se puede demostrar que, en muestras grandes ($n_1 \geq 30$ y $n_2 \geq 30$), procedentes de poblaciones de dos variables X_1 y X_2, con distribuciones no necesariamente Normales, de medias μ_1 y μ_2 y desviaciones típicas σ_1 y σ_2 respectivamente, la variable

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

sigue una distribución Normal tipificada, $N(0,\ 1)$.

De igual manera, si las varianzas de las variables son desconocidas, utilizando como estimadores muestrales sus correspondientes cuasivarianzas $\hat{S}_1^2$ y $\hat{S}_2^2$, donde

$$\hat{S}_1^2 = \frac{\sum (x_{1,i} - \overline{x}_1)^2}{n_1 - 1} \quad \text{y} \quad \hat{S}_2^2 = \frac{\sum (x_{2,i} - \overline{x}_2)^2}{n_2 - 1}$$

entonces la variable

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

sigue una distribución t de Student, en la que el número de grados de libertad dependerá de si las varianzas, aún siendo desconocidas, pueden considerarse iguales o no.

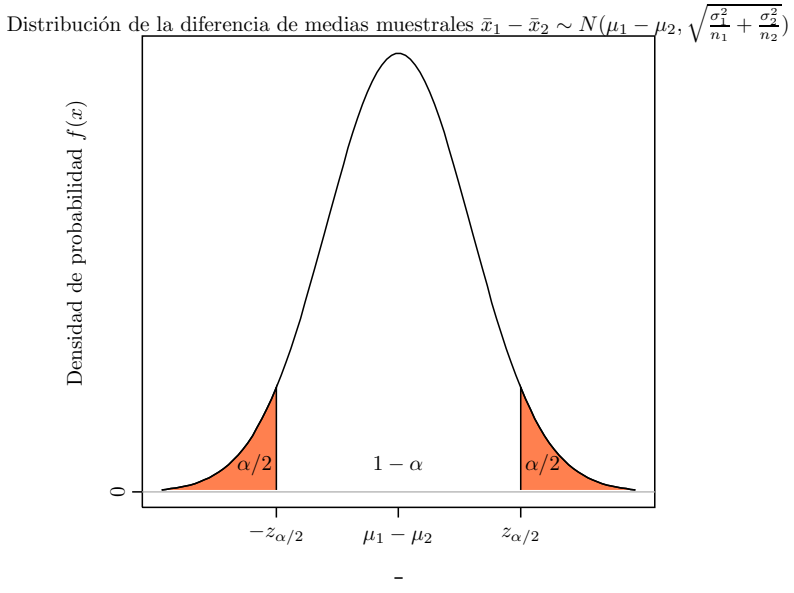
Para muestras pequeñas ($n_1 < 30$ ó $n_2 < 30$), las distribuciones anteriores son también aplicables siempre que las variables de partida sigan distribuciones Normales.

A partir de todo ello y teniendo en cuenta los tres factores de clasificación comentados: si las poblaciones de partida en las que obtenemos las muestras siguen o no distribuciones Normales, si las varianzas de dichas poblaciones son conocidas o desconocidas, y si la muestras son grandes o no, obtenemos las siguientes expresiones correspondientes a los diferentes intervalos de confianza.

Intervalo de confianza para la diferencia de dos medias en poblaciones normales, con varianzas poblacionales conocidas, independientemente del tamaño de la muestra

$$\left(\overline{x}_1 - \overline{x}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ ,\ \overline{x}_1 - \overline{x}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

En la figura 9.1 aparece un esquema explicativo de la construcción de este intervalo.



$$P\left(\mu_1 - \mu_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \bar{x}_1 - \bar{x}_2 \leq \mu_1 - \mu_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Figura 9.1 – Cálculo del intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas conocidas a partir de la distribución de la diferencia de medias muestrales $\bar{x_1} - \bar{x_2} \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$.

Intervalo de confianza para la diferencia de dos medias en poblaciones normales, con varianzas poblacionales desconocidas, independientemente del tamaño de la muestra

Si aún siendo desconocidas, las varianzas pueden considerarse iguales, el intervalo es:

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\alpha/2}^{n_1+n_2-2} \cdot \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\ ,\ \overline{x}_1 - \overline{x}_2 + t_{\alpha/2}^{n_1+n_2-2} \cdot \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

donde s_p^2 es una cuasivarianza ponderada:

$$s_p^2 = \frac{(n_1 - 1) \cdot \hat{s}_1^2 + (n_2 - 1) \cdot \hat{s}_2^2}{n_1 + n_2 - 2}$$

Si las varianzas, desconocidas, no pueden considerarse como iguales, el intervalo es:

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\alpha/2}^{\nu} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}\ ,\ \overline{x}_1 - \overline{x}_2 + t_{\alpha/2}^{\nu} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}\right)$$

donde ν es el número entero más proximo al valor de la expresión:

$$\frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{s}_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{\hat{s}_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

Si los tamaños muestrales son grandes ($n_1 \geq 30$ y $n_2 \geq 30$) las $t_{\alpha/2}^{\nu}$ y $t_{\alpha/2}^{n_1+n_2-2}$ pueden sustituirse por $z_{\alpha/2}$.

Intervalo de confianza para la diferencia de dos medias en poblaciones no normales, y muestras grandes ($n_1 \geq 30$ y $n_2 \geq 30$)

En este caso, como ya sucedía con la media muestral, los intervalos para la diferencia de medias son los mismos que sus correspondientes en poblaciones normales y, de nuevo, habría que distinguir si las varianzas son conocidas o desconocidas (iguales o diferentes), lo cual se traduce en que sus correspondientes fórmulas son las mismas que las dadas en los párrafos anteriores. No obstante, por tratarse de muestras grandes, también es válida la aproximación de $t^{\nu}_{\alpha/2}$ y $t^{n_1+n_2-2}_{\alpha/2}$ por $z_{\alpha/2}$, y habitualmente tan sólo se distingue entre varianzas conocidas y desconocidas.

Para varianzas conocidas:

$$\left(\overline{x}_1 - \overline{x}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \ , \ \overline{x}_1 - \overline{x}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

Y para varianzas desconocidas:

$$\left(\overline{x}_1 - \overline{x}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \ , \ \overline{x}_1 - \overline{x}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right)$$

Si las poblaciones de partida no son normales y las muestras son pequeñas, no puede aplicarse el Teorema Central de Límite y no se obtienen intervalos de confianza para la diferencia de medias.

Para cualquiera de los anteriores intervalos:

- n_1 y n_2 son los tamaños muestrales.
- $\overline{x}_1$ y $\overline{x}_2$ son las medias muestrales.
- σ_1 y σ_2 son las desviaciones típicas poblacionales.
- $\hat{s}_1$ y $\hat{s}_2$ son las cuasidesviaciones típicas muestrales: $\hat{s}_1^2 = \dfrac{\sum (x_{1,i} - \overline{x}_1)^2}{n_1 - 1}$, y análogamente $\hat{s}_2^2$.
- $z_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución Normal tipificada.
- $t^{n_1+n_2-1}_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución t de Student con $n_1 + n_2 - 1$ grados de libertad.
- $t^{\nu}_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución t de Student con ν grados de libertad.

Intervalos de confianza para la media de la diferencia en datos emparejados

En muchas ocasiones hay que estudiar una característica en una población en dos momentos distintos, para estudiar cómo evoluciona con el tiempo, o para analizar la incidencia de algún hecho ocurrido entre dichos momentos.

En estos casos se toma una muestra aleatoria de la población y en cada individuo de la misma se observa la característica objeto de estudio en los dos momentos citados. Así se tienen dos conjuntos de datos que no son independientes, pues los datos están emparejados para cada individuo. Por consiguiente, no se pueden aplicar los procedimientos vistos anteriormente, ya que se basan en la independencia de las muestras.

El problema se resuelve tomando para cada individuo la diferencia entre ambas observaciones. Así, la construcción del intervalo de confianza para la diferencia de medias, se reduce a calcular el intervalo de confianza para la media de la variable diferencia. Además, si cada conjunto de observaciones sigue una distribución Normal, su diferencia también seguirá una distribución Normal.

Intervalos de confianza para la diferencia de dos proporciones poblacionales p_1 y p_2

Para muestras grandes ($n_1 \geq 30$ y $n_2 \geq 30$) y valores de p_1 y p_2 (probabilidad de "éxito") cercanos a 0,5, las correspondientes distribuciones Binomiales pueden aproximarse mediante distribuciones Normales de medias respectivas n_1p_1 y n_2p_2, y desviaciones típicas respectivas $\sqrt{n_1p_1(1-p_1)}$ y $\sqrt{n_2p_2(1-p_2)}$. En la práctica, para que sea válida dicha aproximación, se toma el criterio de que tanto n_1p_1 y n_2p_2 como $n_1(1-p_1)$ y $n_2(1-p_2)$ deben ser mayores que 5. Lo anterior hace que también podamos construir intervalos de confianza para la diferencia de proporciones tomando éstas como medias de variables dicotómicas en las que la presencia o ausencia de la característica objeto de estudio ("éxito" ó "fracaso") se expresan mediante un 1 ó un 0 respectivamente.

De este modo, en muestras grandes y con distribuciones Binomiales no excesivamente asimétricas (tanto n_1p_1 y n_2p_2 como $n_1(1-p_1)$ y $n_2(1-p_2)$ deben ser mayores que 5), si denominamos $\widehat{p}_1$ y $\widehat{p}_2$ a la proporción de individuos que presentan el atributo estudiado en la primera y segunda muestras respectivamente, entonces el intervalo de confianza para la diferencia de proporciones con un nivel de significación α viene dado por:

$$\left(\widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1 \cdot (1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2 \cdot (1-\widehat{p}_2)}{n_2}} \ , \ \widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1 \cdot (1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2 \cdot (1-\widehat{p}_2)}{n_2}} \right)$$

donde:

- n_1 y n_2 son los respectivos tamaños muestrales.
- $\widehat{p}_1$ y $\widehat{p}_2$ son las proporciones de individuos que presentan los atributos estudiados en sus respectivas muestras.
- $z_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución Normal tipificada.

En muestras pequeñas o procedentes de unas distribuciones Binomiales fuertemente asimétricas ($n_1p_1 \leq 5$, $n_2p_2 \leq 5$, $n_1(1-p_1) \leq 5$ ó $n_2(1-p_2) \leq 5$) no puede aplicarse el Teorema Central del Límite y la construcción de intervalos de confianza debe realizarse basándose en la distribución Binomial.

Intervalo de Confianza para la Razón de dos Varianzas σ_1^2 y σ_2^2 de Poblaciones Normales

Como ya hemos visto en la sección de los intervalos de confianza para la diferencia de dos medias en poblaciones normales con varianzas desconocidas, los mismos dependen de si las varianzas, aún siendo desconocidas, pueden considerarse iguales o no. Para dar respuesta a esta cuestión, previa al cálculo del intervalo para la diferencia de medias, se construye un intervalo para la razón (cociente) de varianzas de ambas poblaciones. Para ello tenemos en cuenta que si partimos de dos variables X_1 y X_2 que siguen distribuciones normales con varianzas σ_1^2 y σ_2^2 respectivamente, y tomamos muestras de tamaños n_1 y n_2 de las respectivas poblaciones se tiene que la variable

$$F = \frac{\frac{\hat{S}_1^2}{\sigma_1^2}}{\frac{\hat{S}_2^2}{\sigma_2^2}}$$

sigue una distribución F de Fisher de $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador.

De lo anterior se deduce que el intervalo de confianza con nivel de significación α para $\dfrac{\sigma_2^2}{\sigma_1^2}$ es

$$\left(\frac{\hat{s}_2^2}{\hat{s}_1^2} \cdot F_{1-\alpha/2}^{(n_1-1,n_2-1)}, \ \frac{\hat{s}_2^2}{\hat{s}_1^2} \cdot F_{\alpha/2}^{(n_1-1,n_2-1)} \right)$$

Si dentro del intervalo de confianza obtenido está el número 1 (el cociente de varianzas vale la unidad), no habrá, por tanto, evidencia estadística suficiente, con un nivel de significación α, para rechazar que las varianzas sean iguales.

2 Ejercicios resueltos

1. Para ver si una campaña de publicidad sobre un fármaco ha influido en sus ventas, se tomó una muestra de 8 farmacias y se midió el número de unidades de dicho fármaco vendidas durante un mes, antes y después de la campaña, obteniéndose los siguientes resultados:

Antes	147	163	121	205	132	190	176	147
Después	150	171	132	208	141	184	182	145

   a) Crear un conjunto de datos con las variables antes y despues.

   b) Obtener un resumen estadístico en el que aparezcan la media y la desviación típica de ambas variables. A la vista de los resultados: ¿son las medias diferentes?, ¿ha aumentado la campaña el nivel de ventas?, ¿crees que los resultados son estadísticamente significativos?

   Indicación
   1) Seleccionar el menú `Estadísticos→Resúmenes→Resúmenes descriptivo`.
   2) En el cuadro de diálogo que aparece seleccionar las variables antes y despues, activar la casilla de selección para la `Media` y la `Desviación típica` y hacer click en el botón Aceptar.

   c) Obtener los intervalos de confianza para la media de la diferencia entre ambas variables con niveles de significación 0,05 y 0,01.

   Indicación
   1) Seleccionar el menú `Estadísticos→Medias→Test t para datos relacionados`.
   2) En el cuadro de diálogo que aparece seleccionar la variable antes en el campo `Primera variable`, la variable después en el campo `Segunda variable`, introducir 0.95 en el campo `Nivel de confianza` y hacer click en el botón Aceptar.
   3) El intervalo de confianza para la diferencia aparece en la ventana de resultados justo después de la frase 95 percent confidence interval:.
   4) Repetir los pasos para el intervalo de confianza con nivel de significación 0,01 poniendo 0,99 en el campo `Nivel de confianza`.

   d) ¿Existen pruebas suficientes para afirmar con un 95 % de confianza que la campaña de publicidad ha aumentado las ventas? ¿Y si cambiamos los dos últimos datos de la variable despues y ponemos 190 en lugar de 182 y 165 en lugar de 145? Observar qué le ha sucedido al intervalo para la diferencia de medias y darle una explicación.

   Indicación
   1) Hacer click sobre el botón Editar conjunto de datos.
   2) En la ventana de edición de datos, cambiar los datos de las dos últimas farmacias y cerrar la ventana.
   3) Repetir los pasos del apartado anterior.

   Existen diferencias entre las medias con el nivel de confianza fijado siempre que el intervalo resultante no contenga el valor 0.

2. Una central de productos lácteos recibe diariamente la leche de dos granjas X e Y. Para analizar la calidad de la leche, durante una temporada, se controla el contenido de materia grasa de la leche que proviene de ambas granjas, con los siguientes resultados:

X		Y	
0,34	0,34	0,28	0,29
0,32	0,35	0,30	0,32
0,33	0,33	0,32	0,31
0,32	0,32	0,29	0,29
0,33	0,30	0,31	0,32
0,31	0,32	0,29	0,31
		0,33	0,32
		0,32	0,33

   a) Crear un conjunto de datos con las variables grasa y granja.

b) Calcular el intervalo de confianza para el cociente de varianzas del contenido de materia grasa de la leche procedente de ambas granjas.

Indicación

1) Seleccionar el menú `Estadísticos→Varianzas→Test F para dos varianzas`.
2) En el cuadro de dialogo que aparece seleccionar la variable grasa al campo `Variable explicada`, seleccionar la variable granja al campo `Grupos`, introducir 0.95 en el campo `Nivel de confianza` y hacer click sobre el botón Aceptar.

Se mantiene la hipótesis de igualdad de varianzas con la confianza fijada si el intervalo resultante contiene el valor 1.

c) Calcular el intervalo de confianza con un 95 % de confianza para la diferencia en el contenido medio de materia grasa de la leche procedente de ambas granjas.

Indicación

1) Seleccionar el menú `Estadísticos→Medias→Test t para muestras independientes`.
2) En el cuadro de dialogo que aparece seleccionar la variable grasa al campo `Variable explicada`, seleccionar la variable granja al campo `Grupos`, introducir 0.95 en el campo `Nivel de confianza`, marcar la opción `Si` en el campo `¿Suponer varianzas iguales?` y hacer click sobre el botón Aceptar.

d) A la vista del intervalo obtenido en el punto anterior, ¿se puede concluir que existen diferencias significativas en el contenido medio de grasa según la procedencia de la leche? Justificar la respuesta.

Indicación

Existen diferencias entre las medias con el nivel de confianza fijado siempre que el intervalo resultante no contenga el valor 0.

3. En una encuesta realizada en una facultad, sobre si el alumnado utiliza habitualmente (al menos una vez a la semana) la biblioteca de la misma, se han obtenido los siguientes resultados:

Alumno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Respuesta	no	si	no	no	no	si	no	si	si	si	si	no	si	no	si	no	no
Sexo	H	M	M	H	H	H	M	M	M	M	H	H	M	H	M	H	H

Alumno	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Respuesta	no	si	si	si	no	no	si	no	no	si	si	no	no	si	no	si	no
Sexo	M	H	M	M	M	H	M	H	H	M	M	H	H	M	M	M	H

a) Crear un conjunto de datos con las variables respuesta y sexo.

b) ¿Existen diferencias significativas entre las proporciones de chicos y chicas que usan habitualmente la biblioteca? Justificar la respuesta.

Indicación

1) Seleccionar el menú `Datos→Modificar variables del conjunto de datos activo→Reordenar niveles de factor`.
2) En el cuadro de diálogo que aparece seleccionar el factor respuesta y hacer click sobre el botón Aceptar.
3) En el cuadro de diálgo que aparece asignar el valor 1 al nivel `si`, el valor 2 al nivel `no` y hacer click en el botón Aceptar.
4) Seleccionar el menú `Estadísticos→Proporciones→Test de proporciones para dos muestras`.
5) En el cuadro de dialogo que aparece seleccionar la variable respuesta al campo `Variable explicada`, seleccionar la variable sexo al campo `Grupos`, introducir 0.95 en el campo `Nivel de confianza` y hacer click sobre el botón Aceptar.

Hay diferencias entre las proporciones con el nivel de confianza fijado si el intervalo resultante no contiene el valor 0.

4. Un profesor universitario ha tenido dos grupos de clase a lo largo del año: uno con horario de mañana y otro de tarde. En el de mañana, sobre un total de 80 alumnos, han aprobado 55; y en el de tarde, sobre un total de 90 alumnos, han aprobado 32. ¿Existen diferencias significativas en el porcentaje de aprobados en ambos grupos? ¿Pueden ser debidas al turno horario? Justificar la respuesta.

Indicación

a) Seleccionar el menú `Estadísticos`→`Proporciones`→`Test para dos proporciones`.

b) En el cuadro de diálogo que aparece introducir 55 en el campo `Frecuencia muestral 1`, introducir 80 en el campo `Tamaño muestral 1`, introducir 32 en el campo `Frecuencia muestral 2`, introducir 90 en el campo `Tamaño muestral 2`, introducir 0.95 en el campo `Nivel de confianza` y hacer click en el botón Aceptar.

3 Ejercicios propuestos

1. Se ha realizado un estudio para investigar el efecto del ejercicio físico en el nivel de colesterol en la sangre. En el estudio participaron once personas, a las que se les midió el nivel de colesterol (en mg/dl) antes y después de desarrollar un programa de ejercicios. Los resultados obtenidos fueron los siguientes:

Nivel Previo	182	232	191	200	148	249	276	213	241	280	262
Nivel Posterior	198	210	194	220	138	220	219	161	210	213	226

   a) Hallar el intervalo de confianza del 95 % para la diferencia del nivel medio de colesterol antes y después del ejercicio.

   b) Hallar el intervalo de confianza del 99 % para la diferencia del nivel medio de colesterol antes y después del ejercicio.

   c) A la vista de los intervalos anteriores, ¿se concluye que el ejercicio físico disminuye el nivel de colesterol?

2. En una encuesta realizada en los dos hospitales de una ciudad se pregunta a los pacientes hospitalizados cuando salen del hospital por si consideran que el trato recibido ha sido correcto. En el primero de ellos se pregunta a 200 pacientes y 140 responden que sí, mientras que en el segundo, se pregunta a 300 pacientes y 180 responden que sí.

   a) Calcular el intervalo de confianza para la diferencia de proporciones de pacientes satisfechos con el trato recibido.

   b) ¿Hay pruebas significativas para un nivel de significación $\alpha = 0{,}01$ de que el trato recibido en un hospital es mejor que en el otro?

3. El fichero `nations.txt` contiene información sobre el desarrollo de distintos países (tasa de uso de anticonceptivos (contraception), producto interior bruto per cápita (GDP), tasa de mortalidad infantil (infant.mortality) y tasa de fertilidad (TFR)). Se pide:

   a) Importar el fichero `nations.txt` en un conjunto de datos.

   b) Crear una nueva variable nivel_economico que tome el valor `Ricos` para los países con un producto interior bruto per cápita superior a 10000 US$ y el valor `Pobres` a los países con un producto interior bruto per cápita inferior a dicha cantidad.

   c) ¿Existen diferencias significativas en el uso de anticonceptivos entre los países ricos y pobres? Justificar la respuesta.

   d) ¿Existen diferencias significativas en la tasa de fertiliad entre los países ricos y pobres? Justificar la respuesta.

   e) ¿Existen diferencias significativas en la tasa de mortalidad infantil entre los países ricos y pobres? Justificar la respuesta.

Contraste de Hipótesis

1 Fundamentos teóricos

1.1 Inferencia Estadística y Contrastes de Hipótesis

Cualquier afirmación o conjetura que determina, parcial o totalmente, la distribución de una población se realiza mediante un *Hipótesis Estadística*.

En general, nunca se sabe con absoluta certeza si una hipótesis es cierta o falsa, ya que, para ello tendríamos que medir a todos los individuos de la población. Las decisiones se toman sobre una base de probabilidad y los procedimientos que conducen a la aceptación o rechazo de la hipótesis forman la parte de la Inferencia Estadística que se denomina *Contraste de Hipótesis*.

Una hipótesis se contrasta comparando sus predicciones con la realidad que se obtiene de las muestras: si coinciden, dentro del margen de error probabilísticamente admisible, mantendremos la hipótesis; en caso contrario, la rechazamos y buscaremos nuevas hipótesis capaces de explicar los datos observados.

1.2 Tipos de Contrastes de Hipótesis

Los contrastes de hipótesis se clasifican como:

- *Contrastes Paramétricos*. Que a su vez son de dos tipos según que:
  - Se contraste un valor concreto o intervalo para los parámetros de la distribución de una variable aleatoria. Por ejemplo: podemos contrastar la hipótesis de que la media del nivel de colesterol en sangre en una población es de 180 mg/dl.
  - Se comparen los parámetros de las distribuciones de dos o más variables. Por ejemplo: podemos contrastar la hipótesis de que la media del nivel de colesterol en sangre es más baja en las personas que ingieren por debajo de una cierta cantidad de grasas en su dieta.
- *Contrastes No Paramétricos*. En los que se contrastan las hipótesis que se imponen como punto de partida en los contrastes paramétricos, y que reciben el nombre de *Hipótesis Estructurales*. Entre ellas, el modelo de distribución de los datos y la independencia de los mismos. Por ejemplo: en muchos de los contrastes paramétricos se exige como hipótesis de partida que los datos muestrales provengan de una población normal, pero precisamente éste sería el primer contraste al que habría que dar respuesta, puesto que si los datos no provienen de una población normal, las conclusiones obtenidas gracias a los contrastes paramétricos derivados pueden ser completamente erróneas.

1.3 Elementos de un Contraste

Hipótesis Nula e Hipótesis Alternativa

El primer punto en la realización de un contraste de hipótesis es la formulación de la Hipótesis Nula y su correspondiente Hipótesis Alternativa.

Llamaremos *Hipótesis Nula* a la hipótesis que se contrasta. Se suele representar como H_0 y representa la hipótesis que mantendremos a no ser que los datos observados en la muestra indiquen su falsedad, en términos probabilísticos.

El rechazo de la hipótesis nula lleva consigo la aceptación implícita de la *Hipótesis Alternativa*, que se suele representar como H_1. Para cada H_0 tenemos dos H_1 diferentes según que el contraste sea de tipo *Bilateral*, si desconocemos la dirección en que H_0 puede ser falsa, o *Unilateral*, si sabemos en qué dirección H_0 puede ser falsa. Y el Unilateral se clasifica como *Con Cola a la Derecha*, si en H_1 sólo englobamos valores mayores del parámetro para el que planteamos el contraste que el que aparecen en H_0, y *Con Cola a la Izquierda*, si en H_1 sólo englobamos valores menores.

En la siguiente tabla se formulan tanto H_0 como H_1 en contrastes paramétricos, para un parámetro cualquiera, que denominaremos θ, de una población, y para la comparación de dos parámetros, θ_1 y θ_2, de dos poblaciones.

	H_0	H_1
Bilateral en una población	$\theta = \theta_0$	$\theta \neq \theta_0$
Unilateral en una población	$\theta = \theta_0$	$\theta > \theta_0$ (Cola a la dcha.) ó $\theta < \theta_0$ (Cola a la Izda.)
Bilateral en dos poblaciones	$\theta_1 = \theta_2$	$\theta_1 \neq \theta_2$
Unilateral en dos poblaciones	$\theta_1 = \theta_2$	$\theta_1 > \theta_2$ ó $\theta_1 < \theta_2$

Ejemplo Supongamos que, gracias a datos previos, conocemos que la media del nivel de colesterol en sangre en una determinada población es 180 mg/dl, y suponemos que la aplicación de una cierta terapia ha podido influir (ya sea para aumentar o para disminuir) en dicha media. Para formular H_0 debemos tener en cuenta que la hipótesis nula siempre es conservadora, es decir, no cambiaremos nuestro modelo si no hay evidencias probabilísticamente fuertes de que ha dejado de ser válido. Según esto, la hipótesis nula será que la media no ha cambiado:

$$H_0 : \mu = 180.$$

Una vez fijada la hipótesis nula, para formular la hipótesis alternativa debemos tener en cuenta que se trata de un contraste bilateral, ya que no conocemos, a priori, el sentido de la variación de la media (si será mayor o menor de 180 mg/dl). Por tanto, la hipótesis alternativa es que la media es distinta de 180 mg/dl:

$$H_1 : \mu \neq 180.$$

Por otro lado, si presumimos que la aplicación de la terapia ha servido para disminuir el nivel de colesterol, estamos ante un contraste unilateral en el que la hipótesis nula sigue siendo que la media no ha cambiado, y la alternativa es que ha disminuido:

$$H_0 : \mu = 180,$$
$$H_1 : \mu < 180.$$

Normalmente, el objetivo del investigador es rechazar la hipótesis nula para probar la certeza de la hipótesis alternativa, y esto sólo lo hará cuando haya pruebas suficientemente significativas de la falsedad de H_0. Si los datos observados en la muestra no aportan estas pruebas, entonces se mantiene la hipótesis nula, y en este sentido se dice que es la hipótesis conservadora. Pero conviene aclarar que aceptar la hipótesis nula no significa que sea cierta, sino que no tenemos información suficiente o evidencia estadística para rechazarla.

Errores en un Contraste. Nivel de significación y Potencia

Como ya hemos comentado, la aceptación o rechazo de H_0 siempre se realiza en términos probabilísticos, a partir de la información obtenida en la muestra. Esto supone que nunca tendremos absoluta seguridad de conocer la certeza o falsedad de una hipótesis, de modo que al aceptarla o rechazarla es posible que nos equivoquemos.

Los errores que se pueden cometer en un contraste de hipótesis son de dos tipos:

- **Error de tipo I**: se produce cuando rechazamos H_0 siendo correcta.
- **Error de tipo II**: se produce cuando aceptamos H_0 siendo falsa.

La probabilidad de cometer un error de tipo I se conoce como *Nivel de Significación* del contraste y se designa por

$$\alpha = P(\text{Rechazar } H_0 | H_0 \text{ es cierta}).$$

Y la probabilidad de cometer un error de tipo II se designa por

$$\beta = P(\text{Aceptar } H_0 | H_0 \text{ es falsa}).$$

Así pues, al realizar un contraste de hipótesis, pueden darse las cuatro situaciones que aparecen esquematizadas en el cuadro 10.1.

	Realidad	
Decisión	H_0 cierta	H_0 falsa
Aceptar H_0	Decisión correcta (Probabilidad $1-\alpha$)	Error de Tipo II (Probabilidad β)
Rechazar H_0	Error de Tipo I (Probabilidad α)	Decisión correcta (Probabilidad $1-\beta$)

Cuadro 10.1 – Tipos de errores en un contraste de hipótesis.

Puesto que lo interesante en un contraste es rechazar la hipótesis nula, lo que más interesa controlar es el riesgo de equivocación si se rechaza, es decir el error del tipo I. Por tanto, α se suele fijar a niveles bajos, ya que cuanto más pequeño sea, mayor seguridad tendremos al rechazar la hipótesis nula. Los niveles más habituales a los que se fija α son 0,1, 0,05 y 0,01.

Una vez controlado el error de tipo I, también es interesante controlar el error del tipo II. Ahora bien, el valor de β se calcula partiendo de que la hipótesis nula es falsa, es decir $\theta \neq \theta_0$ (o $\theta_1 \neq \theta_2$ en el caso de dos poblaciones), pero esto engloba infinitas posibilidades, de manera que para poder calcularlo no queda más remedio que fijar H_1 dando un único valor al parámetro. En este caso, se define la *Potencia del Contraste* como la probabilidad de rechazar H_0 cuando la hipótesis alternativa fijada es verdadera, y vale $1-\beta$. Resulta evidente que un contraste será mejor cuanta más potencia tenga.

Como la potencia depende del valor del parámetro fijado en la hipótesis alternativa, se puede definir una función para la potencia como

$$\text{Potencia}(x) = P(\text{Rechazar } H_0 | \theta = x),$$

que indica la probabilidad de rechazar H_0 para cada valor del parámetro θ. Esta función se conoce como *curva de potencia* (figura 10.1).

Por otro lado, β también depende de α ya que al disminuir α, cada vez es más difícil rechazar H_0, y por tanto, la probabilidad de aceptar la hipótesis nula siendo falsa aumenta. En consecuencia, y como veremos más adelante, la única forma de disminuir β y ganar potencia, una vez fijado α, es aumentando el tamaño de la muestra.

Estadístico del Contraste y Regiones de Aceptación y Rechazo

La decisión entre la aceptación o el rechazo de H_0 que se plantea en el contraste, se realiza en base a un estadístico en el muestreo, relacionado con el parámetro o característica que queremos contrastar, y cuya distribución debe ser conocida suponiendo cierta H_0 y una vez fijado el tamaño de la muestra. Este estadístico recibe el nombre de *Estadístico del Contraste*.

Para cada muestra el estadístico del contraste toma un valor concreto recibe el nombre de *estimación del estadístico*. Será a partir de esta estimación que tomaremos la decisión de aceptar o rechazar la hipótesis nula. Si la estimación difiere demasiado del valor que propone H_0 para el parámetro, entonces rechazaremos H_0, mientras que si no es demasiado diferente la aceptaremos.

La magnitud de la diferencia que estamos dispuestos a tolerar entre la estimación y el valor de parámetro para mantener la hipótesis nula, depende de la probabilidad de error de tipo I que estemos dispuestos a asumir. Si α es grande, pequeñas diferencias pueden ser suficientes para rechazar H_0, mientras que si α es muy pequeño, sólo rechazaremos H_0 cuando la diferencia entre el estimador y el valor del parámetro según H_0, sea muy grande. De esta manera, al fijar el nivel de significación α,el conjunto de valores que puede tomar el estadístico del contraste queda divido en dos partes: la de las estimaciones

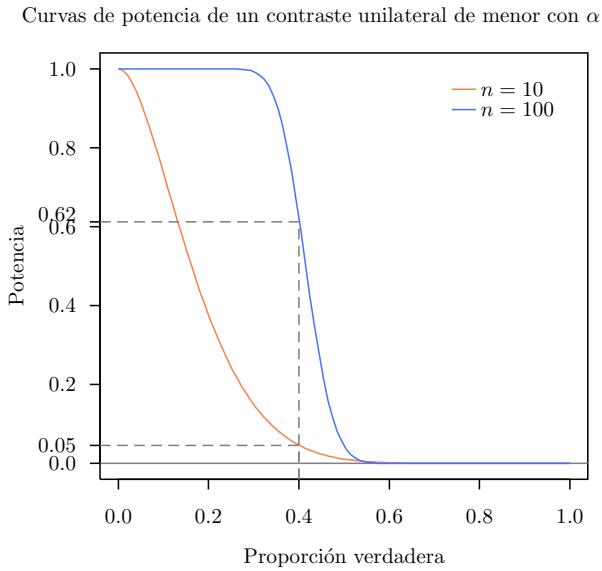


Figura 10.1 – Curvas de potencia.

que conducirían a la aceptación de H_0, que se denomina *Región de Aceptación*, y la de las estimaciones que conducirían al rechazo de H_0, que se denomina *Región de Rechazo*.

Si llamamos al estadístico del contraste $\hat{\theta}$, entonces, dependiendo de si el contraste es unilateral o bilateral, tendremos las siguientes regiones de aceptación y rechazo:

Contraste	Región de Aceptación	Región de Rechazo
$H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$	$\{\hat{\theta}_{1-\alpha/2} \leq \hat{\theta} \leq \hat{\theta}_{\alpha/2}\}$	$\{\hat{\theta} < \hat{\theta}_{1-\alpha/2}\} \cup \{\hat{\theta} > \hat{\theta}_{\alpha/2}\}$
$H_0: \theta = \theta_0$ $H_1: \theta < \theta_0$	$\{\hat{\theta} \geq \hat{\theta}_{1-\alpha}\}$	$\{\hat{\theta} < \hat{\theta}_{1-\alpha}\}$
$H_0: \theta = \theta_0$ $H_1: \theta > \theta_0$	$\{\hat{\theta} \leq \hat{\theta}_{\alpha}\}$	$\{\hat{\theta} > \hat{\theta}_{\alpha}\}$

donde $\hat{\theta}_{1-\alpha}$ y $\hat{\theta}_{\alpha}$ son valores tales que $P(\theta < \hat{\theta}_{1-\alpha} | \theta = \theta_0) = \alpha$ y $P(\theta > \hat{\theta}_{\alpha} | \theta = \theta_0) = \alpha$, tal y como se muestra en las figuras 10.2 y 10.3.

En resumen, una vez que tenemos el estadístico del contraste y hemos fijado el nivel de significación α, las regiones de aceptación y rechazo quedan delimitadas, y ya sólo queda tomar una muestra, aplicar el estadístico del contraste a la muestra para obtener la estimación, y aceptar o rechazar la hipótesis nula dependiendo de si la estimación cae en la región de aceptación o en la de rechazo respectivamente.

El p-Valor de un Contraste

Aunque ya disponemos de los elementos necesarios para realizar un contraste de hipótesis que nos permita tomar una decisión respecto a aceptar o rechazar la hipótesis nula, en la práctica, la decisión que se toma suele acompañarse del grado de confianza que tenemos en la misma. Si por ejemplo, tenemos una región de rechazo $\{\hat{\theta} > \hat{\theta}_{\alpha}\}$, siempre que la estimación del estadístico del contraste caiga dentro de esta región rechazaremos H_0, pero obviamente, si dicha estimación es mucho mayor que $\hat{\theta}_{\alpha}$ tendremos más confianza en el rechazo que si la estimación está cerca del límite entre las regiones de aceptación y rechazo $\hat{\theta}_{\alpha}$. Por este motivo, al realizar un contraste, también se calcula la probabilidad de obtener una discrepancia mayor o igual que la observada entre el valor del parámetro, suponiendo cierta H_0, y la estimación que se obtiene de los datos muestrales. Esta probabilidad se conoce como *p-valor del contraste*, y en cierto modo, expresa la confianza que se tiene al tomar la decisión en el contraste, ya que si . Si H_0 es cierta y el p-valor es pequeño, es porque bajo la hipótesis nula resulta poco probable

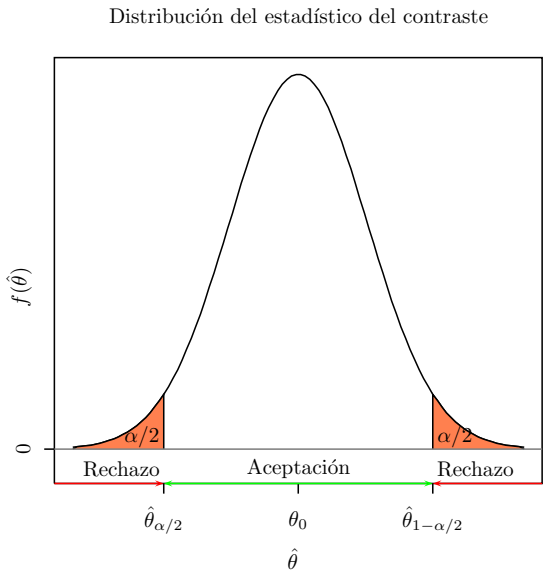


Figura 10.2 – Regiones de aceptación y rechazo en un contraste bilateral.

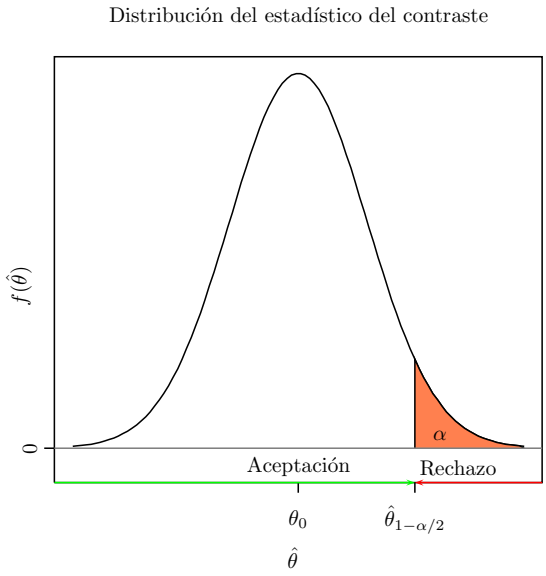


Figura 10.3 – Regiones de aceptación y rechazo en un contraste unilateral.

encontrar una discrepancia como la observada, y por tanto, tendremos bastante seguridad a la hora de rechazar H_0. En general, cuanto más próximo esté p a 1, mayor seguridad existe al aceptar H_0, y cuanto más próximo esté a 0, mayor seguridad hay al rechazarla.

El cálculo del p-valor dependerá de si el contraste es bilateral o unilateral, y en este último caso de si es unilateral con cola a la derecha o con cola a la izquierda. El p-valor que se obtiene para los diferentes tipos de contrastes es el que aparece en la tabla siguiente:

Contraste	p-valor
Bilateral	$2P(\hat{\theta} > \hat{\theta}_0 \mid H_0 \text{ es cierta})$
Unilateral con cola a la derecha	$P(\hat{\theta} > \hat{\theta}_0 \mid H_0 \text{ es cierta})$
Unilateral con cola a la izquierda	$P(\hat{\theta} < \hat{\theta}_0 \mid H_0 \text{ es cierta})$

En la figura 10.4 se observa que el p-valor es el área de la cola de la distribución (o colas si se trata de un contraste bilateral) definida a partir del estadístico del contraste.

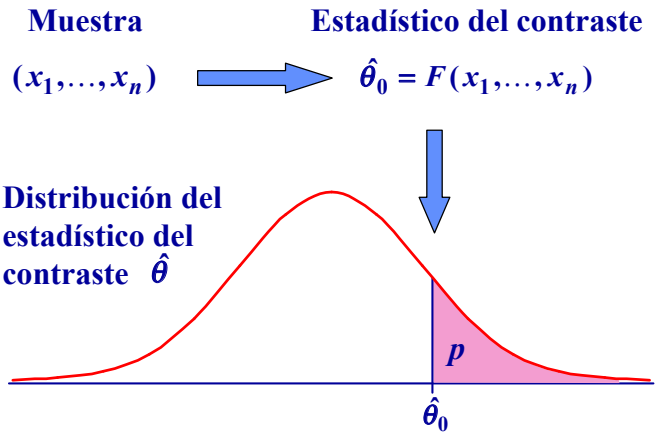


Figura 10.4 – El p-valor de un contraste unilateral con cola a la derecha.

Una vez calculado el p-valor, si hemos fijado el nivel de significación α y han quedado delimitadas las regiones de aceptación y rechazo, el que la estimación caiga dentro de la región de rechazo es equivalente a que $p < \alpha$, mientras que si cae dentro de la región de aceptación, entonces $p \geq \alpha$. Esta forma de abordar los contrastes, nos da una visión más amplia, ya que nos da información de para qué niveles de significación puede rechazarse la hipótesis nula, y para cuales no se puede.

Contrastes y Estadísticos de Contraste

Apoyándose en las distintas distribuciones en el muestreo comentadas en las prácticas sobre intervalos de confianza, a continuación se presentan las fórmulas para los principales estadísticos de contraste.

Contraste para la media de una población normal con varianza conocida

- Hipótesis Nula: $H_0 : \mu = \mu_0$
- Estadístico del contraste:

$$\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

que sigue una distribución normal tipificada, $N(0,1)$.

Este contraste también es válido para la media de una población no normal, siempre y cuando las muestras sean grandes ($n \geq 30$), con varianza conocida; y para la media de la diferencia de datos emparejados, siempre y cuando la variable diferencia siga una distribución normal con varianza conocida, o una distribución cualquiera si la muestra es grande.

Contraste para la media de una población normal con varianza desconocida

- Hipótesis Nula: $H_0 : \mu = \mu_0$
- Estadístico del contraste:

$$\frac{\overline{X} - \mu_0}{\frac{S_{n-1}}{\sqrt{n}}}$$

que sigue una distribución t de Student con $n-1$ grados de libertad, $T(n-1)$.

Este contraste también es válido para la media de una población no normal en muestras grandes ($n \geq 30$), con varianza desconocida; y para la media de la diferencia de datos emparejados, siempre y cuando la variable diferencia siga una distribución normal con varianza desconocida, o una distribución cualquiera si la muestra es grande.

Contraste para la proporción en muestras grandes y distribuciones simétricas (tanto np como $n(1-p)$ deben ser mayores que 5)

- Hipótesis Nula: H_0:$p = p_0$
- Estadístico del contraste:

$$\frac{p - p_0}{\sqrt{\frac{p(1-p)}{n}}}$$

que sigue una distribución normal tipificada, $N(0,1)$.

Contraste para la varianza de una población normal

- Hipótesis Nula: H_0: $\sigma^2 = \sigma_0^2$
- Estadístico del contraste:

$$\frac{(n-1)\, S_{n-1}^2}{\sigma_0^2}$$

que sigue una distribución Chi-cuadrado con $n-1$ grados de libertad.

Contraste para la diferencia de medias de poblaciones normales con varianzas conocidas

- Hipótesis Nula: $H_0 : \mu_1 = \mu_2$
- Estadístico del contraste:

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

que sigue una distribución normal tipificada, $N(0,1)$.

Este contraste también es válido para la diferencia de medias de dos poblaciones no normales, siempre y cuando las muestras sean grandes ($n_1 \geq 30$ y $n_2 \geq 30$), con varianzas conocidas.

Contraste para la diferencia de medias de poblaciones normales con varianzas desconocidas

- Hipótesis Nula: $H_0 : \mu_1 = \mu_2$
- Estadístico del contraste:

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_{1,n_1-1}^2}{n_1} + \frac{S_{2,n_2-1}^2}{n_2}}}$$

que sigue una distribución t de Student con ν grados de libertad, donde ν es el número entero más próximo al valor de la expresión:

$$\frac{\left(\frac{s_{1,n_1-1}^2}{n_1} + \frac{s_{2,n_2-1}^2}{n_2}\right)^2}{\frac{\left(\frac{s_{1,n_1-1}^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{s_{2,n_2-1}^2}{n_2}\right)^2}{n_2+1}} - 2$$

.

Este contraste también es válido para la diferencia de medias de dos poblaciones no normales, siempre y cuando las muestras sean grandes ($n_1 \geq 30$ y $n_2 \geq 30$), con varianzas desconocidas.

Contraste para la diferencia de proporciones en muestras grandes y distribuciones simétricas (n_1p_1, $n_1(1-p_1)$, n_2p_2, $n_2(1-p_2)$ deben ser mayores que 5)

- Hipótesis Nula: H_0:$p_1 = p_2$
- Estadístico del contraste:

$$\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

que sigue una distribución normal tipificada, $N(0,1)$.

Contraste para la igualdad de varianzas de poblaciones normales

- Hipótesis Nula: H_0: $\sigma_1^2 = \sigma_2^2$
- Estadístico del contraste:

$$\frac{S^2_{1,n_1-1}}{S^2_{2,n_2-1}}$$

que sigue una distribución F de Fisher con $n_1 - 1$ y $n_2 - 1$ grados de libertad.

2 Ejercicios resueltos

1. Para averiguar si en una determinada población existen menos hombres que mujeres se plantea un contraste de hipótesis sobre la proporción de hombres que hay en la población: $H_0: \ p = 0{,}5$ frente a $H_1: \ p < 0{,}5$ y para ello se toma una muestra aleatoria de 10 personas. Se pide:

   a) Suponiendo cierta la hipótesis nula, ¿qué distribución sigue la variable que mide el número de hombres en la muestra de tamaño 10?

   b) Suponiendo cierta la hipótesis nula, ¿cuál es la probabilidad de que en la muestra se obtengan 0 hombres? ¿Se aceptaría la hipótesis nula en tal caso? Justificar la respuesta.

   Indicación
   1) Seleccionar el menú `Distribuciones`→`Distribuciones discretas`→`Distribución binomial`→`Probabilidades binomiales acumuladas`.
   2) En el cuadro de diálogo que aparece, introducir 0 en el campo `Valor(es) de la variable`, 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

   c) Suponiendo cierta la hipótesis nula, si se decide rechazarla cuando en la muestra haya 2 o menos hombres, ¿cuál es el riesgo de equivocarse?

   Indicación
   1) Seleccionar el menú `Distribuciones`→`Distribuciones discretas`→`Distribución binomial`→`Probabilidades binomiales acumuladas`.
   2) En el cuadro de diálogo que aparece, introducir 2 en el campo `Valor(es) de la variable`, 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

   d) Si el máximo riesgo de error α que se tolera es 0,05, ¿qué número de hombres en la muestra formarían la región de rechazo de la hipótesis nula?

   Indicación
   1) Seleccionar el menú `Distribuciones`→`Distribuciones discretas`→`Distribución binomial`→`Probabilidades binomiales acumuladas`.
   2) En el cuadro de diálogo que aparece, introducir 1 en el campo `Valor(es) de la variable`, 10 en el campo `Ensayos binomiales`, 0,5 en el campo `Probabilidad de éxito`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

   e) Suponiendo que la proporción real de hombres en la población fuese de 0,4, ¿cuál es la potencia del contraste para la región de rechazo del apartado anterior?

   Indicación
   1) Seleccionar el menú `Distribuciones`→`Distribuciones discretas`→`Distribución binomial`→`Probabilidades binomiales acumuladas`.
   2) En el cuadro de diálogo que aparece, introducir 1 en el campo `Valor(es) de la variable`, 10 en el campo `Ensayos binomiales`, 0,4 en el campo `Probabilidad de éxito`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

   f) Si en lugar de una muestra de tamaño 10 se tomase una muestra de tamaño 100, y haciendo uso de la aproximación de una distribución binomial mediante una normal, ¿qué número de hombres en la muestra formarían la región de rechazo para un riesgo $\alpha = 0{,}05$? ¿Qué potencia tendría ahora el contraste si la proporción real de hombres fuese de 0,4? ¿Es mejor o peor contraste que el anterior? Justificar la respuesta.

Indicación
Una distribución binomial $B(100,\ 0{,}5)$ puede aproximarse mediante una normal $N(50, 5)$.

1) Seleccionar el menú `Distribuciones`→`Distribuciones continuas`→`Distribución normal` →`Cuantiles normales`.
2) En el cuadro de diálogo que aparece, introducir las probabilidad 0,05 en el campo `Probabilidades`, 50 en el campo `media`, 5 en el campo `desviación típica`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

El valor obtenido es la frontera entre la región de aceptación y la región de rechazo. Si en la muestra se obtienen menos hombres de dicho valor se rechazará la hipótesis nula, mientras que si se obtienen más se aceptará. Para calcular la potencia de contraste:

1) Seleccionar el menú `Distribuciones`→`Distribuciones continuas`→`Distribución normal` →`Probabilidades normales`.
2) En el cuadro de diálogo que aparece, introducir el valor de la frontera en el campo `Valor(es) de la variable`, 40 en el campo `media`, 4,899 en el campo `desviación típica`, marcar la opción `Cola izquierda` y hacer click en el botón Aceptar.

g) Si se toma una muestra de tamaño 100 y se observan 41 hombres, ¿cuál es p-valor del contraste? Podría rechazarse la hipótesis nula pra un riesgo $\alpha = 0{,}05$? ¿y para un riesgo $\alpha = 0{,}01$?

Indicación

1) Seleccionar el menú `Estadísticos`→`Proporciones`→`Test para una proporción`.
2) En el cuadro de diálogo que aparece introducir 41 en el campo `Frecuencia muestral`, introducir 100 en el campo `Tamaño muestral`, introducir 0,5 en el campo `Hipótesis nula`, marcar la opción `Proporción de la población <p0` en el campo `Hipótesis alternativa` y hacer click en el botón Aceptar.
3) El p-valor del contraste aparece en la ventana de resultados como p-value.

2. Se analiza la concentración de principio activo en una muestra de 10 envases tomados de un lote de un fármaco, obteniendo los siguientes resultados en mg/mm^3:

$$17{,}6 - 19{,}2 - 21{,}3 - 15{,}1 - 17{,}6 - 18{,}9 - 16{,}2 - 18{,}3 - 19{,}0 - 16{,}4$$

Se pide:

a) Crear un conjunto de datos con la variable concentracion.

b) Realizar el contraste de hipótesis bilateral: H_0: $\mu = 18$ y H_1: $\mu \neq 18$ con un nivel de significación 0,05.

Indicación

1) Seleccionar el menú `Estadísticos`→`Medias`→`Test t para una muestra`.
2) En el cuadro de diálogo que aparece seleccionar la variable concentracion, introducir 18 en el campo `Hipótesis nula`, marcar la opción `Media poblacional != mu0` y hacer click sobre el botón Aceptar.
3) El p-valor del contraste aparece en la ventana de resultados como p-value.

c) De igual manera realizar los contrastes bilaterales: H_0: $\mu = 19{,}5$ y H_1: $\mu \neq 19{,}5$ con un niveles de significación 0,05 y 0,01. ¿Cómo afecta la disminución en el nivel de significación en la facilidad para rechazar H_0?

Indicación
Seguir los mismos pasos del apartado anterior introduciendo 19,5 en el campo `Hipótesis nula`.

d) Realizar los contrastes bilaterales y unilaterales para la hipótesis nula H_0: $\mu = 17$ con un nivel de significación de 0,05. ¿Qué relación hay entre el p-valor de los contrastes bilateral y unilaterales?

Indicación

1) Seleccionar el menú `Estadísticos`→`Medias`→`Test t para una muestra`.
2) En el cuadro de diálogo que aparece seleccionar la variable concentracion, introducir 17 en el campo `Hipótesis nula`, marcar la opción `Media poblacional != mu0` y hacer click sobre el botón Aceptar.
3) Para el contraste de meyor repetir lo mismo marcando la opción `Media poblacional <mu0`.
4) Para el contraste de mayor repetir lo mismo marcando la opción `Media poblacional >mu0`.

e) Si el fabricante del lote asegura haber aumentado la concentración de principio activo con respecto a anteriores lotes, en los que la media era de 17 mg/mm^3, ¿se acepta o se rechaza la afirmación del fabricante?

f) ¿Cuál sería el tamaño muestral requerido para poder detectar una diferencia de 0.5 mg/mm^3 más con un nivel de significación $\alpha = 0{,}05$ y una potencia $1 - \beta = 0{,}8$?

Indicación
Para calcular el tamaño muestral se necesita saber la desviación típica de la población o una estimación suya. Para ello se calcula previamente la cuasivarianza muestral

1) Seleccionar el menú `Estadísticos→Resúmenes→Resumen descriptivo`.
2) En el cuadro de diálogo que aparece seleccionar la variable concentracion, marcar la opción `Cuasidesviación típica` y hacer click en el botón Aceptar.

Para calcular el tamaño muestral:

1) Seleccionar el menú `Estadísticos→Medias→Cálculo del tamaño muestral para el test T`.
2) En el cuadro de diálogo que aparece introducir en el campo `Diferencia en las medias` el valor 0.5, introducir en el campo `Desviación típica` el valor de la cuasivesviación típica obtenido, introducir el valor 0,05 en el campo `Nivel de significación`, introducir el valor 0,8 en el campo `Potencia`, marcar la opción `Una muestra` en el campo `Tipo de test`, marcar la opción `Unilateral` en el campo `Hipótesis alternativa` y hacer click sobre el botón Aceptar.

3. En una encuesta realizada en una facultad, sobre si el alumnado utiliza habitualmente (al menos una vez a la semana) la biblioteca de la misma, se han obtenido los siguientes resultados:

Alumno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Respuesta	no	si	no	no	no	si	no	si	si	si	si	no	si	no	si	no	no

Alumno	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Respuesta	no	si	si	si	no	no	si	no	no	si	si	no	no	si	no	si	no

a) Crear un conjunto de datos con la variable respuesta.

b) Contrastar si el porcentaje de alumnos que utiliza regularmente la biblioteca es superior al 40 %.

Indicación

1) Seleccionar el menú `Datos→Modificar variables del conjunto de datos activo→Reordenar niveles de factor`.
2) En el cuadro de diálogo que aparece seleccionar el factor respuesta y hacer click sobre el botón Aceptar.
3) En el cuadro de diálgo que aparece asignar el valor 1 al nivel `si`, el valor 2 al nivel `no` y hacer click en el botón Aceptar.
4) Seleccionar el menú `Estadísticos→Proporciones→Test de proporciones para una muestra`.
5) En el cuadro de diálogo que aparece seleccionar la variable respuesta, introducir 0,4 en el campo `Hipótesis nula`, marcar la opción `Proporción de la población >p0` en el campo `Hipótesis alternativa` y hacer click en el botón Aceptar.

El comando para el test de las proporciones siempre toma la proporción del primer nivel del factor, de ahí que haya que reordenar los niveles antes.

4. Varios investigadores desean saber si es posible concluir que dos poblaciones de niños difieren respecto a la edad promedio en la cual pueden caminar por sí solos. Los investigadores obtuvieron los siguientes datos para la edad al comenzar a andar (expresada en meses):

Muestra en la población A: $9{,}5 - 10{,}5 - 9{,}0 - 9{,}8 - 10{,}0 - 13{,}0 - 10{,}0 - 13{,}5 - 10{,}0 - 9{,}8$

Muestra en la población B: $12{,}5 - 9{,}5 - 13{,}5 - 13{,}8 - 12{,}0 - 13{,}8 - 12{,}5 - 9{,}5 - 12{,}0 - 13{,}5 - 12{,}0 - 12{,}0$

a) Crear un conjunto de datos con las variables población y edad.

b) Realizar un contraste de hipótesis con un nivel de significación de 0,05 para dar respuesta a la conclusión que buscan los investigadores.

Indicación
Primero hay que realizar un constraste de comparación de varianzas.

1) Seleccionar el menú `Estadísticos→Varianzas→Test F para dos varianzas`.
2) En el cuadro de dialogo que aparece seleccionar la variable edad al campo `Variable explicada`, seleccionar la variable población al campo `Grupos`, marcar la opción `Bilateral` en el campo `Hipótesis alternativa` y hacer click sobre el botón Aceptar.

Se mantiene la hipótesis de igualdad de varianzas con la confianza fijada si el el p-valor es mayor que 0,5. Después se realiza el contraste de comparación de medias.

1) Seleccionar el menú `Estadísticos→Medias→Test t para muestras independientes`.
2) En el cuadro de dialogo que aparece seleccionar la variable edad al campo `Variable explicada`, seleccionar la variable población al campo `Grupos`, marcar la opción `Bilateral` en el campo `Hipótesis alternativa`, marcar la opción `Si` en el campo `¿Suponer varianzas iguales?` y hacer click sobre el botón Aceptar.

Hay diferencias entre las poblaciones si el p-valor es menor que 0,05.

5. Algunos investigadores han observado una mayor resistencia de las vías respiratorias en fumadores que en no fumadores. Para confirmar dicha hipótesis, se realizó un estudio para comparar el porcentaje de retención traqueobronquial en las mismas personas cuando aún eran fumadoras y transcurrido un año después de dejarlo. Los resultados se indican en la tabla siguiente:

Porcentaje de retención	
Cuando fumaba	Transcurrido un año sin fumar
60,6	47,5
12,0	13,3
56,0	33,0
75,2	55,2
12,5	21,9
29,7	27,9
57,2	54,3
62,7	13,9
28,7	8,90
66,0	46,1
25,2	29,8
40,1	36,2

   a) Crear un conjunto de datos con las variables antes y después e introducir los datos.

   b) Plantear el contraste de hipótesis adecuado para confirmar o denegar la hipótesis de los investigadores.

Indicación

1) Seleccionar el menú `Estadísticos→Medias→Test t para datos relacionados`.
2) En el cuadro de diálogo que aparece seleccionar la variable `antes` en el campo `Primera variable`, la variable después en el campo `Segunda variable`, marcar la opción `Diferencia >0` en el campo `Hipótesis alternativa` y hacer click en el botón Aceptar.

6. Un profesor universitario ha tenido dos grupos de clase a lo largo del año: uno con horario de mañana y otro de tarde. En el de mañana, sobre un total de 80 alumnos, han aprobado 55; y en el de tarde, sobre un total de 90 alumnos, han aprobado 32. ¿Se puede afirmar que hay diferencias significativas entre los porcentajes de aprobados en ambos grupos? Justificar la respuesta.

Indicación

1) Seleccionar el menú `Estadísticos→Proporciones→Test para dos proporciones`.
2) En el cuadro de diálogo que aparece introducir 55 en el campo `Proporción muestral 1`, introducir 80 en el campo `Tamaño muestral 1`, introducir 32 en el campo `Proporción muestral 2`, introducir 90 en el campo `Tamaño muestral 2`, marcar la opción `Bilateral` en el campo `Hipótesis alternativa` y hacer click en el botón Aceptar.

3 Ejercicios propuestos

1. El fichero `pulse.txt` contiene información sobre el pulso de un grupo de pacientes que han realizado distintos ejercicios: pulso en reposo (pulse1), pulso después de hacer ejercicio (pulse2), tipo de ejercicio (ran, 1=correr, 2=andar), sexo (sex, 1=hombre, 2=mujer) y peso (weight). Se pide:

   a) Contrastar si el pulso en reposo está por debajo de 75 pulsaciones.

   b) ¿Qué tamaño muestral sería necesario para detectar una diferencia de 2 pulsaciones más en la media de las pulsaciones en reposo, con un nivel de significación 0,05 y una potencia de 0,9?

   c) Contrastar si el pulso después de correr está por encima de 85 pulsaciones.

   d) Contrastar si el porcentaje de personas con taquicardia leve (número de pulsaciones en reposo por encima de 90) supera el 5 %.

   e) ¿Se puede afirmar que el ejercicio aumenta las pulsaciones con una significación de 0,05? ¿y con una significación 0,01? Justificar la respuesta.

   f) ¿Existen diferencias entre las pulsaciones después de andar y después de correr? Justificar la respuesta.

   g) ¿Existen diferencias entre las pulsaciones en reposo entre hombres y mujeres? ¿Y entre las pulsaciones después de correr? Justificar la respuesta.

Análisis de la Varianza de 1 Factor

1 Fundamentos teóricos

El *Análisis de la Varianza con un Factor* es una técnica estadística de contraste de hipótesis, cuyo propósito es estudiar el efecto de la aplicación de varios *niveles*, también llamados *tratamientos*, de una variable aleatoria cualitativa, llamada *factor*, en una variable cuantitativa, llamada *respuesta*.

Por ejemplo: Supongamos que estamos interesados en conocer si el sueldo medio de los médicos que entran a formar parte de la plantilla de un hospital, depende de la comunidad autónoma en la que trabajan. En este problema, la variable factor es la comunidad autónoma, con sus distintos niveles que son las distintas comunidades, mientras que la variable respuesta es el sueldo cobrado. A diferencia de un análisis de regresión simple, en el que se intenta explicar la variable respuesta mediante otra variable cuantitativa (como por ejemplo, el sueldo en función de las horas de permanencia en el hospital, o de la antigüedad en el puesto de trabajo), en el análisis de la varianza el factor, que es la variable independiente, es una variable cualitativa.

Por otro lado, el análisis de la varianza de 1 factor se parece a un contraste de comparación de medias, sólo que en dicho contraste se comparan las medias de dos poblaciones, mientras que en el análisis de la varianza se comparan las medias de las k poblaciones correspondientes a los k niveles del factor.

Para comparar las medias de la variable respuesta según los diferentes niveles del factor, se realiza un contraste de hipótesis en el que la hipótesis nula, H_0, es que la variable respuesta tiene igual media en todos los niveles, mientras que la hipótesis alternativa, H_1, es que hay diferencias estadísticamente significativas en al menos dos de las medias; y dicho contraste de hipótesis se basa en la comparación de dos estimadores de la varianza total de los datos de la variable respuesta; de ahí procede el nombre de esta técnica: *ANOVA* (Analysis of Variance).

1.1 Notación, Modelo y Contraste

La notación habitual en ANOVA es la siguiente:

k es el número de niveles del factor.

n_i es el tamaño de la muestra aleatoria correspondiente al nivel i-ésimo del factor.

$n = \sum_{i=1}^{k} n_i$ es el número total de observaciones.

X_{ij} $(i = 1, ..., k;\ j = 1, ..., n_i)$ es una variable aleatoria que indica la respuesta de la j-ésima unidad experimental al i-ésimo nivel del factor.

x_{ij} es el valor concreto, en una muestra dada, de la variable X_{ij}.

Nivel del Factor			
1	2	$\cdots$	k
X_{11}	X_{21}	$\cdots$	X_{k1}
X_{12}	X_{22}	$\cdots$	X_{k2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}
	X_{2n_2}		

μ_i es la media de la población del nivel i.

$\overline{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$ es la variable media muestral del nivel i, y estimador de μ_i.

$\overline{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ es la estimación concreta para una muestra dada de la variable media muestral del nivel i.

μ es la media de la población incluidos todos los niveles.

$\overline{X} = \sum_{i=1}^{k} \overline{X}_i/k = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}/n$ es la variable media muestral de todas las respuestas, y estimador de μ.

$\overline{x} = \sum_{i=1}^{k} \overline{x}_i/k = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}/n$ es la estimación concreta para una muestra dada de la variable media muestral.

Con esta notación podemos expresar la variable respuesta mediante un modelo matemático que la descompone en componentes atribuibles a distintas causas:

$$X_{ij} = \mu + (\mu_i - \mu) + (X_{ij} - \mu_i)\,,$$

es decir, la respuesta j-ésima en el nivel i-ésimo puede descomponerse como resultado de una media global, más la desviación con respecto a la media global debida al hecho de que recibe el tratamiento i-ésimo, más una nueva desviación con respecto a la media del nivel debida a influencias aleatorias.

Sobre este modelo se plantea la hipótesis nula: las medias correspondientes a todos los niveles son iguales; y su correspondiente alternativa: al menos hay dos medias de nivel que son diferentes.

$$H_0:\ \mu_1 = \mu_2 = ... = \mu_k$$

$$H_1:\ \mu_i \neq \mu_j \text{ para algún } i \text{ y } j$$

Para poder realizar el contraste con este modelo es necesario plantear ciertas hipótesis estructurales (supuestos del modelo):

- Las k muestras, correspondientes a los k niveles del factor, representan muestras aleatorias independientes de k poblaciones con medias $\mu_1 = \mu_2 = ... = \mu_k$ desconocidas.
- Cada una de las k poblaciones es normal.
- Cada una de las k poblaciones tiene la misma varianza, σ^2.

Teniendo en cuenta la hipótesis nula y los supuestos del modelo, podemos construir un estadístico del contraste con distribución conocida, tal que permite aceptar o rechazar H_0; pero hasta poder dar el valor de dicho estadístico, aún debemos seguir ampliando la notación habitual en los test de ANOVA.

Si sustituimos en el modelo las medias poblacionales por sus correspondientes estimadores muestrales tenemos

$$X_{ij} = \overline{X} + \left(\overline{X}_i - \overline{X}\right) + \left(X_{ij} - \overline{X}_i\right),$$

o lo que es lo mismo,

$$X_{ij} - \overline{X} = \left(\overline{X}_i - \overline{X}\right) + \left(X_{ij} - \overline{X}_i\right).$$

Elevando al cuadrado y teniendo en cuenta las propiedades de los sumatorios, se llega a la ecuación que recibe el nombre de *identidad de la suma de cuadrados*:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}\right)^2 = \sum_{i=1}^{k} n_i\left(\overline{X}_i - \overline{X}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2,$$

donde:

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2$ recibe el nombre de *suma total de cuadrados*, (STC), y es la suma de cuadrados de las desviaciones con respecto a la media global; por lo tanto, una medida de la variabilidad total de los datos.

$\sum_{j=1}^{k} n_i(\overline{X}_i - \overline{X})^2$ recibe el nombre de *suma de cuadrados de los tratamientos o suma de cuadrados intergrupos*, $(SCInter)$, y es la suma ponderada de cuadrados de las desviaciones de la media de cada nivel con respecto a la media global; por lo tanto, una medida de la variabilidad atribuida al hecho de que se utilizan diferentes niveles o tratamientos.

$\sum_{i=1}^{k} \sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$ recibe el nombre de *suma de cuadrados residual o suma de cuadrados intragrupos*, $(SCIntra)$, y es la suma de cuadrados de las desviaciones de las observaciones con respecto a las medias de los sus respectivos niveles o tratamientos; por lo tanto, una medida de la variabilidad en los datos atribuida a las fluctuaciones aleatorias dentro del mismo nivel.

Con esta notación la identidad de suma de cuadrados se expresa:

$$SCT = SCInter + SCIntra$$

Y un último paso para llegar al estadístico que permitirá contrastar H_0, es la definición de los *Cuadrados Medios*, que se obtienen al dividir cada una de las sumas de cuadrados por sus correspondientes grados de libertad. Para SCT el número de grados de libertad es $n-1$; para $SCInter$ es $k-1$; y para $SCIntra$ es $n-k$. Por lo tanto,

$$CMT = \frac{SCT}{n-1}$$
$$CMInter = \frac{SCInter}{k-1}$$
$$CMIntra = \frac{SCIntra}{n-k}$$

Y se podría demostrar que, en el supuesto de ser cierta la hipótesis nula y los supuestos del modelo, el cociente:

$$\frac{CMInter}{CMIntra}$$

sigue una distribución F de Fisher con $k-1$ y $n-k$ grados de libertad.

De forma que, si H_0 es cierta, el valor del cociente para un conjunto de muestras dado, estará próximo a 1 (aún siendo siempre mayor que 1); pero si no se cumple H_0 crece la variabilidad intergrupos y la estimación del estadístico crece. En definitiva realizaremos un contraste de hipótesis unilateral con cola a la derecha de igualdad de varianzas, y para ello calcularemos el p-valor de la estimación de F obtenida y aceptaremos o rechazaremos en función del nivel de significación fijado.

Tabla de ANOVA

Todos los estadísticos planteados en el punto anterior se recogen en una tabla denominada Tabla de ANOVA, en la que se ponen los resultados de las estimaciones de dichos estadísticos en las muestras concretas objeto de estudio. Esas tablas también son las que aportan como resultado de cualquier ANOVA los programas estadísticos, que suelen añadir al final de la tabla el p-valor del F calculado, y que permite aceptar o rechazar la hipótesis nula de que las medias correspondientes a todos los niveles del factor son iguales.

	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico F	p-valor
Intergrupos	$SCInter$	$k-1$	$CMInter = \frac{SCInter}{k-1}$	$f_0 = \frac{CMInter}{CMIntra}$	$P(F > f_0)$
Intragrupos	$SCIntra$	$n-k$	$CMIntra = \frac{SCIntra}{n-k}$		
Total	SCT	$n-1$			

Test de Comparaciones Múltiples y por Parejas

Una vez realizado el ANOVA de un factor para comparar las k medias correspondientes a los k niveles o tratamientos del factor, nos encontramos en una de las dos siguientes situaciones:

- No hemos podido rechazar H_0. En este caso se da por concluido el análisis de los datos en cuanto a detección de diferencias entre los niveles.
- Tenemos razones estadísticas para rechazar H_0. En este caso es natural continuar con el análisis para tratar de localizar con precisión dónde está la diferencia, cuáles son el nivel o niveles cuyas respuestas son estadísticamente diferentes.

En el segundo supuesto, hay varios métodos que permiten detectar las diferencias entre las medias de los diferentes niveles, y que reciben el nombre de *Test de Comparaciones Múltiples*. A su vez este tipo de test se suelen clasificar en:

- *Test de comparaciones por parejas*, cuyo objetivo es la comparación una a una de todas las posibles parejas de medias que se pueden tomar al considerar los diferentes niveles. Su resultado es una tabla en la que se reflejan las diferencias entre todas las posibles parejas y los intervalos de confianza para dichas diferencias, con la indicación de si hay o no diferencias significativas entre las mismas. Hay que aclarar que los intervalos obtenidos no son los mismos que resultarían si considerásemos cada pareja de medias por separado, ya que el rechazo de H_0 en el contraste general de ANOVA implica la aceptación de una hipótesis alternativa en la que están involucrados varios contrastes individuales a su vez; y si queremos mantener un nivel de significación α en el general, en los individuales debemos utilizar un α' considerablemente más pequeño.
- *Test de rango múltiple*, cuyo objetivo es la identificación de subconjuntos homogéneos de medias que no se diferencian entre sí.

Entre otros, para los primeros, el test de Bonferroni; para los segundos, el test de Duncan; y para ambas categorías a la vez los test HSD de Tukey y Scheffé.

2 Ejercicios resueltos

1. Se realiza un estudio para comparar la eficacia de tres programas terapéuticos para el tratamiento del acné. Se emplean tres métodos:

   a) Lavado, dos veces al día, con cepillo de polietileno y un jabón abrasivo, junto con el uso diario de 250 mg de tetraciclina.

   b) Aplicación de crema de tretinoína, evitar el sol, lavado dos veces al día con un jabón emulsionante y agua, y utilización dos veces al día de 250 mg de tetraciclina.

   c) Evitar el agua, lavado dos veces al día con un limpiador sin lípidos y uso de crema de tretinoína y de peróxido benzoílico.

   En el estudio participan 35 pacientes. Se separó aleatoriamente a estos pacientes en tres subgrupos de tamaños 10, 12 y 13, a los que se asignó respectivamente los tratamientos I, II, y III. Después de 16 semanas se anotó para cada paciente el porcentaje de mejoría en el número de lesiones.

Tratamiento					
I		II		III	
48,6	50,8	68,0	71,9	67,5	61,4
49,4	47,1	67,0	71,5	62,5	67,4
50,1	52,5	70,1	69,9	64,2	65,4
49,8	49,0	64,5	68,9	62,5	63,2
50,6	46,7	68,0	67,8	63,9	61,2
		68,3	68,9	64,8	60,5
				62,3	

   a) Crear un conjunto de datos con las variables tratamiento y mejora e introducir los datos de la muestra.

   b) Dibujar el diagrama de puntos. ¿Se observan diferencias entre los tratamientos en el diagrama?

   Indicación
   1) Seleccionar el menú `Gráficas→Diagrama de puntos`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable tratamiento en el campo `Factores`, seleccionar la variable mejora en el campo `Variable explicada` y hacer click sobre el botón Aceptar.

   c) Realizar el contraste de ANOVA. ¿Se puede concluir que los tres tratamientos tienen el mismo efecto medio con un nivel de significación de 0,05?

   Indicación
   1) Seleccionar el menú `Estadísticos→Medias→ANOVA de un factor`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable tratamiento en el campo `Grupos`, la variable mejora en el campo `Variable explicada` y hacer click sobre el botón Aceptar.

   d) Obtener la tabla de ANOVA correspondiente al problema pero que además muestre los intervalos de confianza de comparación de los tres tratamientos con una significación de 0,05. ¿Entre qué parejas de tratamientos hay diferencias estadísticamente significativas?

   Indicación
   Repetir los mismos pasos del apartado anterior activando la opción `Comparaciones dos a dos de las medias`.

   e) Dibujar el gráfico de los intervalos de confianza para la media de cada tratamiento.

   Indicación
   1) Seleccionar el menú `Gráficas→Diagrama de las medias`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable tratamiento en el campo `Factores`, la variable mejora en el campo `Variable explicada`, seleccionar la opción `Intervalos de confianza` y hacer click sobre el botón Aceptar.

2. Se sospecha que hay diferencias en la preparación del examen de selectividad entre los diferentes centros de bachillerato de una ciudad. Con el fin de comprobarlo, de cada uno de los 5 centros, se eligieron 8 alumnos al azar, con la condición de que hubieran cursado las mismas asignaturas, y se anotaron las notas que obtuvieron en el examen de selectividad. Los resultados fueron:

Centros				
1	2	3	4	5
5,5	6,1	4,9	3,2	6,7
5,2	7,2	5,5	3,3	5,8
5,9	5,5	6,1	5,5	5,4
7,1	6,7	6,1	5,7	5,5
6,2	7,6	6,2	6,0	4,9
5,9	5,9	6,4	6,1	6,2
5,3	8,1	6,9	4,7	6,1
6,2	8,3	4,5	5,1	7,0

a) Crear un conjunto de datos con las variables nota y centro e introducir los datos de la muestra.

b) Dibujar el diagrama de puntos. ¿Se observan diferencias entre los centros en el diagrama?

Indicación
Antes de nada hay que convertir la variable centro en un factor ya que puede confundirse con una variable numérica.

1) Seleccionar el menú `Datos→Modificar variables del conjunto de datos activo→Convertir variable numérica en factor`.
2) En el cuadro de diálogo que aparece seleccionar la variable centro, marcar la opción `Utilizar números` en el campo `Niveles del factor` y hacer click sobre el botón Aceptar.

Ahor ya se puede dibujar el diagrama de puntos.

1) Seleccionar el menú `Gráficas→Diagrama de puntos`.
2) En el cuadro de diálogo que aparece, seleccionar la variable centro en el campo `Factores`, seleccionar la variable nota en el campo `Variable explicada` y hacer click sobre el botón Aceptar.

c) Realizar el contraste de ANOVA. ¿Se puede confimar la sospecha de que hay diferencias entre las notas medias de los centros?

Indicación

1) Seleccionar el menú `Estadísticos→Medias→ANOVA de un factor`.
2) En el cuadro de diálogo que aparece, seleccionar la variable centro en el campo `Grupos`, la variable nota en el campo `Variable explicada` y hacer click sobre el botón Aceptar.

d) ¿Qué centros son los mejores en la preparación de la selectividad?

Indicación
Repetir los mismos pasos del apartado anterior activando la opción `Comparaciones dos a dos de las medias`.

3 Ejercicios propuestos

1. Se midió la frecuencia cardíaca (latidos por minuto) en cuatro grupos de adultos; controles normales (A), pacientes con angina (B), individuos con arritmias cardíacas (C) y pacientes recuperados del infarto de miocardio (D). Los resultados son los siguientes:

A	B	C	D
83	81	75	61
61	65	68	75
80	77	80	78
63	87	80	80
67	95	74	68
89	89	78	65
71	103	69	68
73	89	72	69
70	78	76	70
66	83	75	79
57	91	69	61

¿Proporcionan estos datos la suficiente evidencia para indicar una diferencia en la frecuencia cardiaca media entre esos cuatro tipos de pacientes?. Considerar $\alpha = 0{,}05$.

2. Se midió la frecuencia respiratoria (inspiraciones por minuto) en ocho animales de laboratorio y con tres niveles diferentes de exposición al monóxido de carbono. Los resultados son los siguientes:

Nivel de exposición		
Bajo	Moderado	Alto
36	43	45
33	38	39
35	41	33
39	34	39
41	28	33
41	44	26
44	30	39
45	31	29

Con base en estos datos, ¿es posible concluir que los tres niveles de exposición, en promedio, tienen un efecto diferente sobre la frecuencia respiratoria? Tomar $\alpha = 0,05$.

Práctica de Estadística con R 12

ANOVA de Múltiples Factores y ANOVA de Medidas Repetidas

1 Fundamentos teóricos

Como ya se vio en una práctica anterior, el *Análisis de la Varianza de un Factor*, *ANOVA* o también *ANOVA de una Vía*, es una técnica estadística de contraste de hipótesis cuyo propósito es estudiar el efecto de la aplicación de varios *niveles* (también llamados *tratamientos*) de una variable aleatoria cualitativa, llamada *factor* o *vía*, en una variable cuantitativa, llamada *respuesta*. Si se supone que la variable cualitativa independiente, es decir el factor, presenta k niveles diferentes, entonces para comparar las k medias de la variable respuesta según los diferentes niveles del factor se realiza un contraste de hipótesis, cuya hipótesis nula, H_0, es que la variable respuesta tiene igual media en todos los niveles, mientras que la hipótesis alternativa, H_1, es que hay diferencias estadísticamente significativas en al menos dos de las medias. Dicho contraste de hipótesis se basa en la comparación de dos estimadores de la varianza total de los datos de la variable respuesta; de ahí procede el nombre de esta técnica: *ANOVA* (Analysis of Variance).

No obstante, en muchos problemas aparece no ya un único factor que permite clasificar los individuos de la muestra en k diferentes niveles, sino que pueden presentarse dos o más factores que permiten clasificar a los individuos de la muestra en múltiples grupos según diferentes criterios, que se pueden analizar para ver si hay o no diferencias significativas entre las medias de la variable respuesta. Para tratar con este tipo de problemas surge el *ANOVA Múltiples Factores* (o también *ANOVA de Varias Vías*) como una generalización del proceso de un factor, que además de permitir el análisis de la influencia de cada uno de los factores por separado también hace posible el estudio de la *interacción* entre ellos.

Por otra parte, también son frecuentes los problemas en los que se toma más de una medida de una variable cuantitativa (respuesta) en cada sujeto de la muestra, y se procede al análisis de las diferencias entre las diferentes medidas. Si sólo se toman dos, el procedimiento adecuado es la T de Student de datos pareados, o su correspondiente no paramétrico, el test de Wilcoxon; pero si se han tomado tres o más medidas, el test paramétrico correspondiente a la T de Student de datos pareados es el *ANOVA de Medidas Repetidas*.

Incluso también se puede dar el caso de un problema en el que se analice una misma variable cuantitativa medida en varias ocasiones en cada sujeto de la muestra pero teniendo en cuenta a la vez la influencia de uno, dos o más factores que permiten clasificar a los individuos en varios subgrupos diferentes. En definitiva, pueden aparecer problemas donde a la par que un ANOVA de medidas repetidas se requiera realizar un ANOVA de dos o más vías.

Por último, la situación más compleja que se puede plantear en el análisis de una respuesta cuantitativa se presenta cuando, añadida a medidas repetidas y dos o más vías o factores de clasificación, se tienen una o más variables cuantitativas, llamadas *Covariables*, que se piensa que pueden influir en la variable respuesta. Se procede entonces a realizar un *ANCOVA* o *Análisis de Covarianza*, con el que se pretende analizar la influencia de los factores y también ver si hay diferencias entre las medidas repetidas pero habiendo eliminado previamente la influencia (variabilidad) debida a la presencia de las covariables que se pretenden controlar.

1.1 ANOVA de múltiples factores

ANOVA de dos factores con dos niveles cada factor

Para entender qué es un ANOVA de múltiples factores, conviene partir de un caso sencillo con dos factores y dos niveles en cada factor. Por ejemplo, se puede plantear un experimento con individuos que siguen o no una dieta (primer factor: dieta, con dos niveles: sí y no), y que a su vez toman o no un determinado fármaco (segundo factor: fármaco, con dos niveles: sí y no) para reducir su peso corporal (variable respuesta numérica: reducción del peso corporal expresada en Kg). En esta situación, se generan cuatro grupos diferentes: los que no hacen dieta ni toman fármaco (No-No), los que no hacen dieta pero sí toman fármaco (No-Sí), los que hacen dieta y no toman fármaco (Sí-No), y los que hacen dieta y toman fármaco (Sí-Sí). Y se pueden plantear tres efectos diferentes:

- El de la dieta: viendo si hay o no diferencias significativas en los Kg perdidos entre los individuos que la han seguido y los que no.
- El del fármaco: viendo si hay o no diferencias significativas en los Kg perdidos entre los individuos que lo han tomado y los que no.
- El de la interacción: viendo si el efecto combinado de dieta y fármaco es diferente del que tendrían sumando sus efectos por separado, y entonces se diría que sí que hay interacción; o si por el contrario el efecto de la combinación de dieta y fármaco es el mismo que la suma de los efectos por separado, y entonces se diría que no hay interacción. A su vez, si hay interacción se puede dar en dos sentidos: si la combinación de dieta y fármaco ha hecho perder más kilos a los pacientes de los que cabría esperar con la suma de dieta y fármaco por separado, entonces la interacción de ambos factores ha actuado en sinergia con los mismos, mientras que si la combinación ha hecho perder menos kilos de los que cabría esperar con dieta y fármaco por separado, entonces la interacción ha actuado en antagonismo con ambos.

Siguiendo con el ejemplo, supongamos que la tabla que aparece a continuación refleja la media de Kg perdidos dentro de cada uno de los grupos comentados. Por simplificar el ejemplo, no se reflejan los Kg en cada individuo con la consiguiente variabilidad de los mismos, pero el ANOVA de dos vías sí que tendría en cuenta esa variabilidad para poder hacer inferencia estadística, plantear contrastes de hipótesis y calcular sus correspondientes p-valores.

	Fármaco No	Fármaco Sí
Dieta No	0	5
Dieta Sí	3	8

Si los resultados obtenidos fuesen los de la tabla anterior, se diría que no hay interacción entre fármaco y dieta, ya que el efecto del fármaco en el grupo de los que no hacen dieta ha hecho perder 5 Kg en media a los individuos, el efecto de la dieta en el grupo de los que no toman fármaco les ha hecho perder 3 Kg en media, y el efecto combinado de dieta y fármaco ha hecho perder 8 Kg con respecto a los que no hacen dieta y tampoco toman fármaco. Estos 8 Kg son iguales a la suma de 3 y 5, es decir iguales a la suma de los efectos de los factores por separado, sin ningún tipo de interacción (de término añadido) que cambie el resultado de la suma.

Con las medias de los cuatro grupos que se generan en el cruce de los dos factores, cada uno con dos niveles (2×2), se representan los gráficos de medias que aparecen más adelante. En estos gráficos, cuando no hay interacción las rectas que unen las medias correspondientes a un mismo nivel de uno de los factores son paralelas dentro de cierto margen de variabilidad.

Por el contrario, también podría obtenerse una tabla en la que la suma de los efectos por separado fuese menor que el efecto combinado de dieta y fármaco:

	Fármaco No	Fármaco Sí
Dieta No	0	5
Dieta Sí	3	12

En este caso, dejando al margen las variabilidad dentro de cada uno de los grupos y suponiendo que la misma es lo suficientemente pequeña como para que las diferencias sean significativas, los 8 Kg en

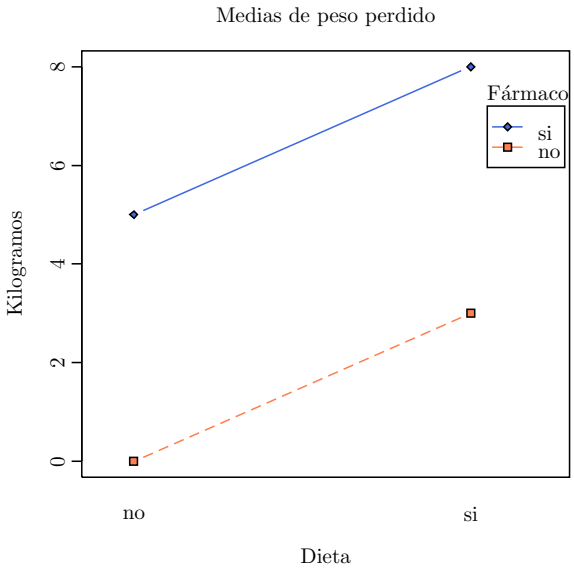


Figura 12.1 – Gráfico de medias de dos factores sin interacción

media que se perderían al sumar los efectos por separado de dieta y fármaco son menores que los 12 que, en media, han perdido los individuos que han tomado el fármaco y han seguido la dieta a la vez. Por lo tanto, se ha producido una interacción de los dos factores que, al unirlos, ha servido para potenciar sus efectos por separado. Dicho de otra forma, para explicar el resultado final de los individuos que han tomado el fármaco y también han seguido la dieta habría que introducir un nuevo término en la suma, el término de interacción, que contribuiría con 4 Kg de pérdida añadidos a los 8 Kg que se perderían considerando simplemente la suma de dieta y fármaco. Como este nuevo término contribuye a aumentar la pérdida que se obtendría al sumar los efectos por separado de ambos factores, se trataría de un caso de interacción en sinergia con los dos factores de partida.

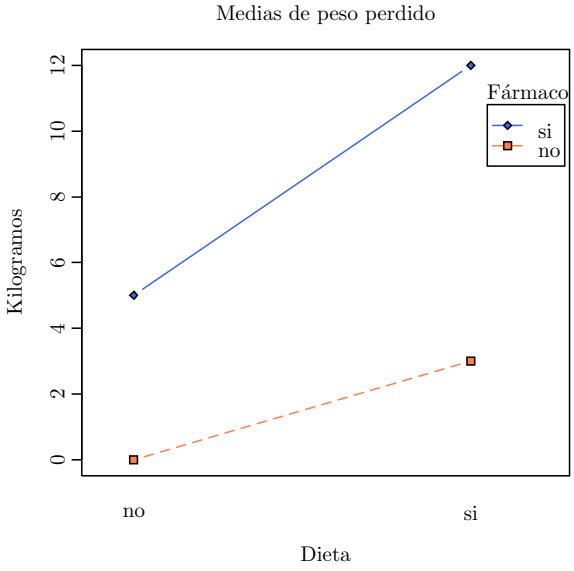


Figura 12.2 – Gráfico de medias de dos factores con interacción sinérgica.

Por último, también se podría obtener una tabla en la que la suma de los efectos por separado fuese mayor que el efecto combinado de los dos factores:

	Fármaco No	Fármaco Sí
Dieta No	0	5
Dieta Sí	3	4

Igualmente, en este nuevo ejemplo los 8 Kg en media que se perderían al sumar los efectos por separado de los dos factores son mayores que los 4 que en realidad pierden, en media, los individuos que han seguido la dieta y utilizado el fármaco. Por lo tanto, para explicar el resultado obtenido en el grupo de los que toman el fármaco y siguen la dieta habría que introducir un término añadido a la suma de efectos sin más, que se restaría a los 8 Kg hasta dejarlos en 4 Kg. Se trataría de un caso de interacción en antagonismo con los dos factores de partida.

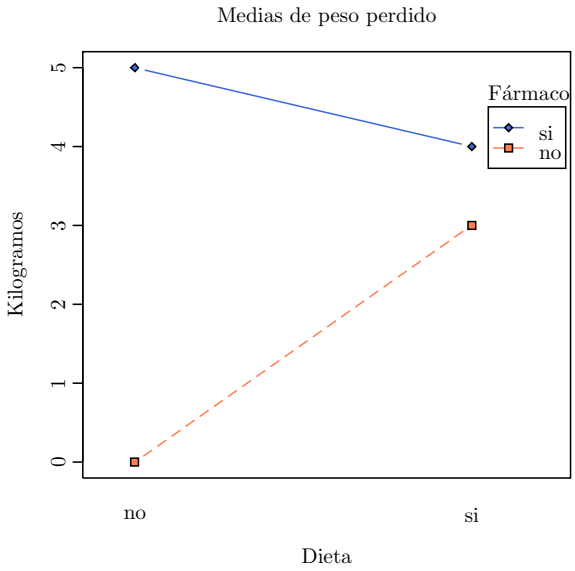


Figura 12.3 – Gráfico de medias de dos factores con interacción antagónica.

En realidad, la interacción también puede producirse en sinergia con uno de los factores y en antagonismo con el otro, ya que a veces los dos factores pueden producir un efecto con signo contrario. Por ejemplo, al hablar del factor dieta, se tiende a pensar que se trata de una dieta que sirve para bajar el peso, pero también cabe plantearse un experimento con personas que siguen una dieta de alto contenido calórico que en principio debería hacerles subir peso y ver qué evolución siguen cuando a la vez toman un fármaco para bajarlo.

Como puede deducirse fácilmente de las tablas y gráficas anteriores, la presencia de interacción implica que la diferencia entre las medias de los dos grupos dentro de un mismo nivel de uno de los factores no es la misma que para el otro nivel. Por ejemplo, en la segunda tabla, la diferencia entre las medias de Kg perdidos entre los que sí que toman el fármaco y los que no lo toman vale: 5-0=5 Kg en los que no hacen dieta, y 12-3=9 Kg en los que sí que hacen dieta. Lo cual gráficamente se traduce en que la pendiente de la recta que une las medias dentro del grupo de los que sí que toman el fármaco es diferente de la pendiente que une las medias dentro del grupo de los que no lo toman. En las ideas anteriores se basará el planteamiento del contraste de hipótesis para ver si la interacción ha resultado o no significativa.

Como ya se ha comentado, en cualquiera de las tablas anteriores se podrían analizar tres efectos diferentes: el de la dieta, el del fármaco y el de la interacción de dieta con fármaco; lo cual, en términos matemáticos, se traduce en tres contrastes de hipótesis diferentes:

1. Efecto de la dieta sobre la cantidad de peso perdido:

$$H_0 : \mu_{\text{con dieta}} = \mu_{\text{sin dieta}}$$
$$H_1 : \mu_{\text{con dieta}} \neq \mu_{\text{sin dieta}}$$

2. Efecto del fármaco sobre la cantidad de peso perdido:

$$H_0 : \mu_{\text{con fármaco}} = \mu_{\text{sin fármaco}}$$
$$H_1 : \mu_{\text{con fármaco}} \neq \mu_{\text{sin fármaco}}$$

3. Efecto de la interacción entre dieta y fármaco, que a su vez se puede plantear de dos formas equivalentes:

   a) Viendo si dentro dentro de los grupos definidos en función de la dieta la diferencia de Kg perdidos entre los que toman fármaco y los que no lo toman es la misma:

$$H_0 : (\mu_{\text{con fármaco}} - \mu_{\text{sin fármaco}})_{\text{sin dieta}} = (\mu_{\text{con fármaco}} - \mu_{\text{sin fármaco}})_{\text{con dieta}}$$
$$H_1 : (\mu_{\text{con fármaco}} - \mu_{\text{sin fármaco}})_{\text{sin dieta}} \neq (\mu_{\text{con fármaco}} - \mu_{\text{sin fármaco}})_{\text{con dieta}}$$

   b) Viendo si dentro de los grupos definidos en función del fármaco la diferencia de Kg perdidos entre los que hacen dieta y los que no la hacen es la misma:

$$H_0 : (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{sin fármaco}} = (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{con fármaco}}$$
$$H_1 : (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{sin fármaco}} \neq (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{con fármaco}}$$

Aunque los detalles matemáticos más precisos sobre cómo el ANOVA de dos o más vías da respuesta a los contrastes expuestos quedan fuera del nivel de esta práctica, la idea general es sencilla y muy parecida a la explicada con más detalle en la práctica de ANOVA de una vía. En el ANOVA de una vía, la variabilidad total de los datos, expresada como suma de distancias al cuadrado con respecto a la media global (llamada Suma de Cuadrados Total), se descompone en dos diferentes fuentes de variabilidad: las distancias al cuadrado de los datos de cada grupo con respecto a la media del grupo, *Suma de Cuadrados Intra*, más las distancias al cuadrado entre las diferentes medias de los grupos y la media general, *Suma de Cuadrados Inter*. La suma de cuadrados intra-grupos es también llamada *Variabilidad Residual* o *Suma de Cuadrados Residual*, ya que su cuantía es una medida de la dispersión residual, remanente incluso después de haber dividido los datos en grupos. Estas sumas de cuadrados, una vez divididas por sus correspondientes grados de libertad, generan varianzas llamadas *Cuadrados Medios*, y el cociente de cuadrados medios (cuadrado medio inter dividido entre cuadrado medio intra) bajo la hipótesis nula de igualdad de medias en todos los grupos sigue una distribución F de Fisher que se puede utilizar para calcular un p-valor del contraste de igualdad de medias. En el ANOVA de dos factores, en lugar de dos fuentes de variabilidad tenemos cuatro: una por el primer factor, otra por el segundo, otra por la interacción y otra más que contempla la variabilidad residual o variabilidad intragrupos. En el ejemplo anterior, las cuatro fuentes de variabilidad son:

1. La debida al primer factor: la dieta.
2. La debida al segundo factor: el fármaco.
3. La debida a la interacción entre ambos.
4. La residual.

Las tres primeras fuentes de variabilidad llevan asociadas sus correspondientes sumas de cuadrados, similares a la suma de cuadrados inter del ANOVA de una vía, mientras que la variabilidad residual lleva asociada su suma de cuadrados residual, similar a la suma de cuadrados intra del ANOVA de una vía. Dividiendo las sumas de cuadrados entre sus respectivos grados de libertad se obtienen varianzas, que divididas entre la varianza residual generan, bajo la hipótesis nula de igualdad de medias, valores f de la distribución F de Fisher que pueden utilizarse para calcular el p-valor del correspondiente contraste.

Lo anterior se resume en forma de tabla de un ANOVA de dos vías, considerando un primer factor con k_1 niveles, un segundo factor con k_2 niveles y un total de datos n. Si se denomina F_1 al primer factor, F_2 al segundo, I a la interacción y R al residual, la tabla de un ANOVA de dos vías tiene la siguiente forma:

Fuente	Suma Cuadrados	Grados Libertad	Cuadrados Medios	Estadístico f	p-valor
F_1	SF_1	$k_1 - 1$	$CF_1 = \frac{SF_1}{k_1 - 1}$	$f_1 = \frac{CF_1}{CR}$	$P(F > f_1)$
F_2	SF_2	$k_2 - 1$	$CF_2 = \frac{SF_2}{k_2 - 1}$	$f_2 = \frac{CF_2}{CR}$	$P(F > f_2)$
Interacción	SI	$(k_1 - 1)(k_2 - 1)$	$CI = \frac{SI}{(k_1 - 1)(k_2 - 1)}$	$f_I = \frac{CI}{CR}$	$P(F > f_I)$
Residual	SR	$n - k_1 k_2$	$CR = \frac{SR}{n - k_1 k_2}$		
Total	ST	$n - 1$			

Una vez obtenida la tabla, habitualmente mediante un programa de estadística para evitar realizar la gran cantidad de cálculos que conlleva (los distintos programas pueden proporcionar tablas ligeramente diferentes a la expuesta en esta práctica, en las que pueden aparecer filas añadidas cuya interpretación dependerá del programa utilizado), el siguiente paso es la interpretación de los p-valores obtenidos en cada uno de los factores y en la interacción. Para ello, resulta clave el p-valor de la interacción porque condicionará completamente el análisis:

- Si la interacción no ha resultado significativa (p-valor de la interacción mayor que el nivel de significación, habitualmente 0,05), se puede considerar por separado la actuación de los dos factores y ver si hay o no diferencias significativas en sus niveles atendiendo al p-valor que aparece en la tabla para cada uno de ellos. Por ejemplo, en la primera de las tablas del análisis de Kg perdidos en función de la dieta y el fármaco, se obtendría que la interacción no es significativa, lo cual implicaría que habría que analizar el efecto de los factores por separado. Para ello, se acudiría al p-valor del factor dieta y si es menor que el nivel de significación fijado, entonces el factor dieta habría resultado significativo, lo cual quiere decir que habría diferencias significativas (más allá de las asumibles por azar) entre los Kg perdidos por los individuos que hacen dieta y los que no; y todo ello, independientemente de si los individuos están tomando o no el fármaco, ya que no hay una interacción significativa que ligue los resultados de la dieta con el fármaco. Igualmente, con el factor fármaco, se acudiría a su p-valor y se vería si hay o no diferencias significativas entre los Kg perdidos por los que toman el fármaco y los que no lo hacen, independientemente de si siguen o no la dieta.

- Si la interacción ha resultado significativa (p-valor de la interacción menor que el nivel de significación, habitualmente 0,05), no se puede considerar por separado la actuación de los dos factores, la presencia de uno de los factores condiciona lo que sucede en el otro y el análisis de diferencias debidas al segundo factor debe realizarse por separado dentro de cada uno de los niveles del primero; y a la inversa, el análisis de diferencias debidas al primero debe realizarse por separado dentro de cada uno de los niveles del segundo. Por ejemplo, en la segunda de las tablas del análisis de Kg perdidos en función de la dieta y el fármaco, muy probablemente se obtendría que la interacción sí que es significativa, con lo cual no habría un único efecto del fármaco: en el grupo de los que no toman el fármaco, la diferencia de Kg perdidos entre los que sí que hacen dieta y los que no la hacen no sería la misma que en el grupo de los que sí que toman el fármaco. E igualmente, tampoco habría un único efecto de la dieta: en el grupo de los que no hacen dieta, la diferencia de Kg perdidos entre los que sí que toman el fármaco y los que no lo hacen no sería la misma que en el grupo de los que sí que hacen dieta.

Una aclaración final importante es que en ningún caso un ANOVA de dos factores con dos niveles en cada vía equivale a hacer por separado una T de Student de datos independientes en cada uno de los factores. Ni siquiera en el caso de que no haya interacción el p-valor que se obtiene en cada uno de los

dos factores coincide con el que se obtendría en la comparación de los niveles mediante la T de Student. El ANOVA de dos factores es una técnica multivariante que cuantifica la influencia de cada una de las variables independientes en la variable dependiente después de haber eliminado la parte de la variabilidad que se debe a las otras variables independientes que forman parte del modelo. En el ejemplo de los Kg perdidos, no sería lo mismo analizar la influencia de la variable dieta después de eliminar la variabilidad explicada mediante la variable fármaco e incluso la interacción entre dieta y fármaco, que es lo que haría el ANOVA de dos factores, que analizar simplemente la influencia de la variable dieta sin más, o fármaco sin más, que es lo que podríamos hacer mediante una T de Student de datos independientes. Tampoco el análisis de la interacción en el ANOVA de dos factores equivale a realizar un ANOVA de una vía considerando una nueva variable independiente con cuatro categorías diferentes (1:Sí-Sí, 2:Sí-No, 3:No-Sí, 4:No-No), por el mismo motivo: las conclusiones del ANOVA de dos vías hay que entenderlas en el contexto de una técnica multivariante en que la importancia de cada variable independiente se obtiene después de eliminar de los datos la variabilidad debida a las demás.

ANOVA de dos factores con tres o más niveles en algún factor

El planteamiento y resolución de un ANOVA de dos factores con tres o más niveles en algún factor es muy parecido al ya expuesto de dos niveles en cada factor. Únicamente cambian ligeramente las hipótesis nulas planteadas en los factores en las que habría que incluir la igualdad de tantas medias como niveles tenga el factor analizado, y las alternativas en las que se supone que alguna de las medias es diferente. En cuanto a las interacciones, también se contemplarían diferencias de medias pero teniendo en cuenta que hay más diferencias posibles al tener más niveles dentro de cada factor.

En cuanto a la interpretación final de los resultados de la tabla del ANOVA, si no hay interacción y sin embargo hay diferencias significativas en cualquiera de los factores con 3 o más niveles, el siguiente paso sería ver entre qué medias se dan esas diferencias. Por ejemplo, si no hay interacción y se ha rechazado la hipótesis nula de igualdad de medias entre los tres niveles del factor 1, habría que ver si esas diferencias aparecen entre los niveles 1 y 2, o entre el 1 y 3, e incluso entre el 2 y el 3, independientemente del factor 2; e igualmente con el factor 2. Para poder ver entre qué niveles hay diferencias, habría que realizar *Test de Comparaciones Múltiples y por Parejas*; por ejemplo un test de Bonferroni o cualquier otro de los vistos en la práctica de ANOVA de una vía. Si la interacción saliese significativa, habría que hacer lo mismo pero considerando las posibles diferencias entre los 3 niveles del factor 1 dentro de cada nivel del factor 2 y viceversa.

Como ya se ha comentado para el ANOVA de dos factores con dos niveles en cada factor y la T de Student de datos independientes, igualmente el ANOVA de dos factores con tres o más niveles en algún factor no equivale a dos ANOVAS de una vía. El p-valor que se obtiene en el de dos factores no es el mismo que que se obtendría en los ANOVAS de una vía realizados teniendo en cuenta cada uno de los factores por separado, incluso si la interacción no es significativa.

ANOVA de tres o más factores

Aunque los fundamentos del ANOVA de tres o más factores son muy parecidos a los de dos y la tabla obtenida es muy similar, la complejidad en la interpretación sube un escalón. Por ejemplo, en un ANOVA de tres factores la tabla presentaría los tres efectos de cada uno de los factores por separado, las tres interacciones dobles (1 con 2, 1 con 3 y 2 con 3), e incluso también podría mostrar la interacción triple (los programas de estadística permiten considerar o no las interacciones de cualquier orden). Si la interacción triple fuese significativa, entonces no se podría hablar del efecto general del factor 1, sino que habría que analizar el efecto del factor 1 dentro de cada nivel del 2 y a su vez dentro de cada nivel del 3, y así sucesivamente. Si la interacción triple no fuese significativa pero sí que lo fuese la del factor 1 con el 2, entonces habría que analizar el efecto del factor 1 dentro de cada uno de los niveles del 2 pero independientemente del factor 3. Y así hasta completar un conjunto muy grande de análisis posibles y de Test de Comparaciones Múltiples aplicados. No obstante, es el propio experimentador el que debe limitar el conjunto de análisis a realizar con un planteamiento muy claro del experimento, reduciendo en la medida de lo posible el número de factores considerados y teniendo claro que no merece la pena considerar interacciones triples, o de órdenes superiores, si no hay forma clara de interpretar su resultado.

En ningún caso un ANOVA de tres o más factores equivale a tres ANOVAS de una vía realizados teniendo en cuenta los factores considerados por separado.

Factores fijos y Factores aleatorios

A la hora de realizar un ANOVA de varios factores, el tratamiento de la variabilidad debida a cada uno de ellos y también las conclusiones que se pueden obtener después de realizarlo, son diferentes dependiendo de que los factores sean fijos o aleatorios.

Se entiende como *Factor Fijo o Factor de Efectos Fijos* aquel cuyos niveles los establece, los fija de antemano, el investigador (por ejemplo, cantidades concretas de fármaco o de tiempo transcurrido), o vienen dados por la propia naturaleza del factor (por ejemplo, el sexo o la dieta). Su variabilidad es más fácil de controlar y también resulta más sencillo su tratamiento en los cálculos que hay que hacer para llegar a la tabla final del ANOVA, pero tienen el problema de que los niveles concretos que toma el factor constituyen la población de niveles sobre los que se hace inferencia. Es decir, no se pueden sacar conclusiones poblacionales que no se refieran a esos niveles fijos con los que se ha trabajado.

Por contra, un *Factor Aleatorio o Factor de Efectos Aleatorios* es aquel cuyos niveles son seleccionados de forma aleatoria entre todos los posibles niveles del factor (por ejemplo, cantidad de fármaco, con niveles 23 mg, 132 mg y 245 mg, obtenidos al escoger 3 niveles de forma aleatoria entre 0 y 250 mg). Su tratamiento es más complicado, pero al constituir una muestra aleatoria de niveles, se pretende sacar conclusiones extrapolables a todos los niveles posibles.

Supuestos del modelo de ANOVA de dos o más vías

Como ya sucedía con el ANOVA de una vía, el de dos o más vías es un test paramétrico que supone que:

- Los qdatos deben seguir distribuciones normales dentro de cada categoría, entendiendo por categorías todas las que se forman del cruce de todos los niveles de todos los factores. Por ejemplo, en un ANOVA de 2 factores con 3 niveles en cada factor, se tienen 3^2 categorías diferentes.
- Todas las distribuciones normales deben tener igualdad de varianzas (homocedasticidad).

Cuando no se cumplen las condiciones anteriores y además las muestras son pequeñas, no se debería aplicar el ANOVA de dos o más vías, con el problema añadido de que no hay un test no paramétrico que lo sustituya. Mediante test no paramétricos (sobre todo mediante el test de Kruskall-Wallis) se podría controlar la influencia de cada uno de los factores por separado en los datos, pero nunca el importantísimo papel de la interacción.

1.2 ANOVA de medidas repetidas

Concepto de ANOVA de medidas repetidas

En muchos problemas se cuantifica el valor de una variable dependiente en varias ocasiones en el mismo sujeto (por ejemplo: en un grupo de individuos que están siguiendo una misma dieta, se puede anotar el peso perdido al cabo de un mes, al cabo de dos y al cabo de tres), y se intenta comparar la media de esa variable en las diferentes ocasiones en que se ha medido, es decir, ver si ha habido una evolución de la variable a lo largo de las diferentes medidas (en el ejemplo anterior, una evolución del peso perdido). Conceptualmente es una situación análoga a la estudiada al comparar dos medias con datos emparejados mediante una T de Student de datos emparejados, o su correspondiente no paramétrico, el test de Wilcoxon, pero ahora hay más de dos medidas emparejadas, realizadas en el mismo individuo. En estas situaciones se utiliza el ANOVA de medidas repetidas.

El ANOVA de medidas repetidas, como también sucede con cualquier otro test que utilice datos emparejados, tiene la ventaja de que las comparaciones que se realizan están basadas en lo que sucede dentro de cada sujeto (intra-sujetos), lo cual reduce el ruido o variabilidad que se produce en comparaciones entre diferentes grupos de sujetos. Por ejemplo, en el estudio sobre la evolución del peso perdido con personas que siguen la misma dieta, se podría haber cuantificado la variable al cabo de uno, dos y tres meses, pero en tres grupos diferentes que hubiesen seguido la misma dieta, pero con este diseño del estudio no se controlan otras variables que pueden influir en el resultado final, por ejemplo el sexo, la edad, o la cantidad de ejercicio que se hace al día. Dicho de otra forma, en el diseño con grupos independientes es posible que alguno de los grupos tenga una media de edad superior, o no haya igual número de hombres que de mujeres, y todo ello tener su reflejo en el número de Kg perdidos. Mientras

que, con el diseño de datos emparejados, la segunda medida se compara con la primera que también se ha realizado en la misma persona, y por lo tanto es igual su sexo, su edad y la cantidad de deporte que realiza; y así con todas las demás medidas que se comparan entre sí pero dentro del mismo individuo. Eso permite controlar la variabilidad y detectar pequeñas diferencias que de otra forma serían indetectables.

ANOVA de medidas repetidas como ANOVA de dos vías sin interacción

El ANOVA de medidas repetidas puede realizarse como un ANOVA de dos vías sin interacción sin más que realizar los cálculos oportunos introduciendo adecuadamente los datos en un programa estadístico.

En la situación de partida, si suponemos que tenemos k medidas emparejadas de una variable dependiente numérica y n individuos en los que hemos tomado las medidas, los datos se pueden organizar como aparecen en la tabla siguientes:

	Medida 1	Medida 2	...	Medida k
Individuo 1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,k}$
Individuo 2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,k}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Individuo n	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,k}$

Pero esos mismos datos también se pueden ordenar en un formato de tabla mucho más conveniente para poderles aplicar un ANOVA de dos vías:

	Variable Dependiente	Individuo	Medida
Fila 1	$x_{1,1}$	1	1
Fila 2	$x_{2,1}$	2	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Fila n	$x_{n,1}$	n	1
Fila $n+1$	$x_{1,2}$	1	2
Fila $n+2$	$x_{2,2}$	2	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Fila $2n$	$x_{n,2}$	n	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Fila $(k-1)n+1$	$x_{1,k}$	1	k
Fila $(k-1)n+2$	$x_{2,k}$	2	k
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Fila kn	$x_{n,k}$	n	k

Con ello, tanto Individuo como Medida son variables categóricas que dividen la muestra total ($n \cdot k$ datos de la variable dependiente) en grupos: n grupos en la variable Individuo y k grupos en la variable Medida. Además, considerando el cruce de ambas variables (Medida x Individuo) se forman $n \cdot k$ grupos con un único dato de la variable dependiente en cada grupo.

Para explicar la variabilidad de los datos de la variable dependiente cuantitativa se pueden considerar tres fuentes: la debida a la variable Medida, la debida a la variable Individuo, y la residual. Ahora no cabe hablar de la variabilidad debida a la interacción entre Medida e Individuo ya que los grupos que surgen del cruce de los dos factores sólo tienen un dato y no es viable calcular medias y dispersiones dentro de un grupo con un único dato. Y el análisis de la influencia de cada uno de los factores se realiza mediante un ANOVA de dos factores sin interacción, que genera la siguiente tabla:

Fuente	Suma Cuadrados	Grados Libertad	Cuadrados Medios	Estadístico f	p-valor
$F_1 =$ Medida	SF_1	$k-1$	$CF_1 = \frac{SF_1}{k-1}$	$f_1 = \frac{CF_1}{CR}$	$P(F > f_1)$
$F_2 =$ Individuo	SF_2	$n-1$	$CF_2 = \frac{SF_2}{n-1}$	$f_2 = \frac{CF_2}{CR}$	$P(F > f_2)$
Residual	SR	$nk-n-k+1$	$CR = \frac{SR}{nk-n-k+1}$		
Total	ST	$n-1$			

Y permite dar respuesta a los siguientes contrastes:

1. En la variable Medida:

$$H_0 : \mu_{\text{Medida 1}} = \mu_{\text{Medida 2}} = ... = \mu_{\text{Medida k}}$$
$$H_1 : \text{Alguna de las medias es diferente.}$$

   Si el p-valor obtenido es menor que el nivel de significación fijado querrá decir que alguna de las medias es significativamente diferente del resto. Este es el contraste más importante del ANOVA de medidas repetidas y supone que la variabilidad dentro de cada individuo (intra-sujeto) es lo suficientemente grande como para que se descarte el azar como su causa. Por lo tanto la variable Medida ha tenido un efecto significativo.

2. En la variable Individuo:

$$H_0 : \mu_{\text{Individuo 1}} = \mu_{\text{Individuo 2}} = ... = \mu_{\text{Individuo n}}$$
$$H_1 : \text{Alguna de las medias es diferente.}$$

   Si el p-valor obtenido es menor que el nivel de significación fijado querrá decir que alguna de las medias es significativamente diferente del resto, y por lo tanto alguno de los individuos analizados ha tenido un comportamiento en la variable dependiente diferente del resto. En realidad no es un contraste importante en el ANOVA de medidas repetidas ya que supone un análisis de la variabilidad entre individuos (inter-sujetos), pero es muy difícil que en un experimento dado esta variabilidad no esté presente.

Si la conclusión del ANOVA es que hay que rechazar alguna de las dos hipótesis nulas, ya sea la de igualdad de medias en los grupos formados por la variable Medida o la de igualdad de medias en los grupos formados por la variable Individuo, entonces en el siguiente paso se podría aplicar un Test de Comparaciones Múltiples y por Parejas, por ejemplo un test de Bonferroni, para ver qué medias son diferentes, especialmente para ver entre qué niveles del la variable Medida se dan las diferencias.

Supuestos del ANOVA de medidas repetidas

Como en cualquier otro ANOVA, en el de medidas repetidas se exige que:

- Los datos de la variable dependiente deben seguir distribuciones normales dentro de cada grupo, ya sea formado por la variable Medida o por la variable Individuo. Como el contraste más importante se realiza en la variable Medida, resultará especialmente importante que sean normales las distribuciones de todas las Medidas .

- Todas las distribuciones normales deben tener igualdad de varianzas (homocedasticidad), especialmente las de las diferentes Medidas.

Cuando en un ANOVA de medidas repetidas se cumple la normalidad y la homocedasticidad de todas las distribuciones se dice que se cumple la *Esfericidad* de los datos, y hay tests estadísticos especialmente diseñados para contrastar la esfericidad como la *prueba de Mauchly*.

Cuando no se cumplen las condiciones anteriores y además las muestras son pequeñas, no se debería aplicar el ANOVA de medidas repetidas, pero al menos sí que hay una prueba no paramétrica que permite realizar el contraste de si hay o no diferencias significativas entre los distintos niveles de la variable Medida, que es el *test de Friedman*.

1.3 ANOVA de medidas repetidas + ANOVA de una o más vías

No son pocos los problemas en los que, además de analizar el efecto intra-sujetos en una variable dependiente cuantitativa medida varias veces en los mismos individuos para el que cabría plantear un ANOVA de medidas repetidas, también aparecen variables cualitativas que se piensa que pueden estar relacionadas con la variable dependiente. Estas últimas variables introducen un efecto que aunque habitualmente es catalogado como inter-sujetos más bien se trataría de un efecto inter-grupos, ya que permiten definir grupos entre los que se podría plantear un ANOVA de una o más vías. Por ejemplo, se podría analizar la pérdida de peso en una muestra de individuos al cabo de uno, dos y tres meses de tratamiento (ANOVA de medidas repetidas), pero teniendo en cuenta que los individuos de la muestra han sido divididos en seis grupos que se forman por el cruce de dos factores, Dieta y Ejercicio, con tres dietas diferentes: a, b y c, y dos niveles de ejercicio físico diferentes: bajo y alto. Para analizar la influencia de estos dos factores inter-sujetos, habría que plantear un ANOVA de dos vías con interacción. Para un ejemplo como el comentado, aunque los datos podrían disponerse de una forma similar a la que permite realizar el ANOVA de medidas repetidas como un ANOVA de dos factores (variables Medida e Individuo), y añadirle dos factores más (Dieta y Ejercicio), no resulta cómodo tener que introducir en la matriz de datos varias filas para un mismo individuo (tantas como medidas repetidas diferentes se hayan realizado). Por ello, determinados programas de estadística, como PASW, permiten realizar ANOVAS de medidas repetidas introduciendo los datos en el formato clásico, una fila para cada individuo y una variable para cada una de las medidas repetidas, definiendo factores intra-sujeto que en realidad estarían compuestos por todas las variables que forman parte de las medidas repetidas. Además, a los factores intra-sujeto permiten añadirle nuevos factores inter-sujeto (categorías) que pueden influir en las variables respuesta (las diferentes medidas), e incluso comprobar si hay o no interacción entre los factores inter-sujeto entre sí y con los factores intra-sujeto. Por lo tanto, son procedimientos que realizan a la vez un ANOVA de medidas repetidas y un ANOVA de una o más vías, con la ventaja de que se pueden introducir los datos en la forma clásica: una fila para cada individuo.

El resultado de la aplicación de estos procedimientos es muy parecido a los comentados en apartados previos: se generan tablas de ANOVA en las que se calcula un p-valor para cada uno de los factores, ya sean intra-sujeto (medidas repetidas) o inter-sujeto (categorías), y también para la interacción, ya sea de los factores inter-sujeto entre sí o de factores inter-sujeto con los intra-sujeto.

2 Ejercicios resueltos

1. En un estudio diseñado para analizar la influencia de un tipo de dieta y de un fármaco en el peso corporal perdido, expresado en Kg, se ha anotado el número de Kg perdidos en un grupo de personas al cabo de 3 meses de dieta y de tomar el fármaco, obteniendo los siguientes resultados (si algún individuo presenta un dato negativo significa que en lugar de perder Kg de peso los ha ganado):

	Fármaco NO	Fármaco SÍ
Dieta NO	1,5; 0,5; 0,0; −1,0; −1,0	6,5; 5,0; 7,0; 3,0; 4,5; 4,0
Dieta SÍ	3,5; 3,0; 4,0; 2,5; 2,0	9,5; 8,0; 7,5; 7,0; 8,5; 7,5

   a) Crear un conjunto de datos con las variables kilos_perdidos, dieta y farmaco.

   b) Mostrar el gráfico de medias de los kilos perdidos para los distintos grupos según la dieta y el fármaco. ¿Qué conclusiones cualitativas pueden sacarse del gráfico obtenido?

   Indicación
   1) Seleccionar el menú `Gráficas→Gráfica de las medias`.
   2) En el cuadro de diálogo que aparece seleccionar la variable kilos_perdidos como `Variable explicada`, las variables dieta y farmaco como `Factores`, seleccionar la opción `Sin barra de errores` y hacer click sobre el botón Aceptar.

   Se observa claramente que no hay interacción (líneas paralelas), que los dos puntos del grupo de los que no hacen dieta están por debajo de los que sí que la hacen, lo cual hace sospechar que el factor dieta será significativo, e igualmente los dos puntos de los que no toman fármaco están por debajo de los que sí que lo toman, lo cual hace sospechar que el factor fármaco también será significativo.

   c) Realizar un contraste de ANOVA de dos vías con los datos e interpretar la tabla de ANOVA obtenida.

   Indicación
   1) Seleccionar el menú `Estadísticos->Medias->ANOVA de múltiples factores`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable kilos_perdidos como `Variable explicada` y las variables dieta y farmaco como `Factores`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.

   Para la interpretación de la tabla de ANOVA, prestar especial atención a las siguientes líneas de la tabla:
   1) dieta: muestra si la dieta resulta o no significativa para explicar la variabilidad del peso perdido.
   2) farmaco: muestra si el fármaco resulta o no significativo.
   3) dieta:farmaco: muestra si la interacción de dieta y fármaco resulta o no significativa.

   Una conclusión muy importante a la luz de los resultados es que no hay una interacción significativa entre dieta y fármaco, es decir que el efecto del fármaco no dependerá de si una persona toma o no dieta, y a la inversa, que el efecto de la dieta no dependerá de si se toma o no fármaco.

   d) Calcular las medias y desviaciones típicas de los Kg perdidos en todos los grupos.

   Indicación
   El procedimiento anterior para obtener el contraste de ANOVA también muestra las medias y desviaciones típicas para cada grupo.

   e) Teniendo en cuenta que no hay interacción significativa, calcular el intervalo de confianza para la diferencia de medias en los kg perdidos según la variable dieta e igualmente con la variable fármaco.

   Indicación
   1) Seleccionar el menú `Estadísticos->Medias->ANOVA de múltiples factores`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable kilos_perdidos como `Variable explicada` y las variables dieta y farmaco como `Factores`, seleccionar la opción `Intervalos de comparación de medias de Tukey`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.

2. En un estudio diseñado para analizar la influencia de un tipo de dieta y de un fármaco en el peso corporal perdido, expresado en Kg, se ha anotado el número de Kg perdidos en un grupo de personas al cabo de 3 meses de dieta y de tomar el fármaco, obteniendo los siguientes resultados (si algún individuo presenta un dato negativo significa que en lugar de perder Kg de peso los ha ganado):

	Fármaco NO	Fármaco SÍ
Dieta NO	1,5; 0,5; 0,0; −1,0; −1,0	6,5; 5,0; 7,0; 3,0; 4,5; 4,0
Dieta SÍ	3,5; 3,0; 4,0; 2,5; 2,0	12,5; 12,0; 11,5; 13,5; 12,5; 10,0

a) Crear un conjunto de datos con las variables kilos_perdidos, dieta y farmaco.

b) Mostrar el gráfico de medias de los kilos perdidos para los distintos grupos según la dieta y el fármaco. ¿Qué conclusiones cualitativas pueden sacarse del gráfico obtenido?

Indicación

1) Seleccionar el menú `Gráficas→Gráfica de las medias`.
2) En el cuadro de diálogo que aparece seleccionar la variable kilos_perdidos como `Variable explicada`, las variables dieta y farmaco como `Factores`, seleccionar la opción `Sin barra de errores` y hacer click sobre el botón Aceptar.

Ahora se observa claramente que hay interacción (líneas no paralelas), que los dos puntos del grupo de los que no hacen dieta están por debajo de los que sí que la hacen, lo cual hace sospechar que el factor dieta será significativo, e igualmente los dos puntos de los que no toman fármaco están por debajo de los que sí que lo toman, lo cual hace sospechar que el factor fármaco también será significativo.

c) Realizar un contraste de ANOVA de dos vías con los datos e interpretar la tabla de ANOVA obtenida. ¿Hay interacción significativa? ¿Cómo se interpretaría?

Indicación

1) Seleccionar el menú `Estadísticos->Medias->ANOVA de múltiples factores`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos_perdidos como `Variable explicada` y las variables dieta y farmaco como `Factores`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.

Ahora puede concluirse que sí hay interacción significativa, y eso implica que no hay la misma diferencia en Kg perdidos entre los que hacen dieta y los que no si consideramos el grupo de los que no toman fármaco, que si consideramos el grupo de los que sí lo toman.

d) Teniendo en cuenta que hay interacción significativa, calcular el intervalo de confianza para la diferencia de medias en los kg perdidos según la variable dieta y fármaco, así como entre los grupos que surgen de su interacción.

Indicación

1) Seleccionar el menú `Estadísticos->Medias->ANOVA de múltiples factores`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos_perdidos como `Variable explicada` y las variables dieta y farmaco como `Factores`, seleccionar la opción `Intervalos de comparación de medias de Tukey`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.

3. Se ha realizado un experimento que consiste en que se ha anotado el tiempo, en días, que han tardado en contestar correctamente a un cuestionario 30 personas, 15 hombres y 15 mujeres, distribuidos en grupos que han seguido tres métodos diferentes de aprendizaje de la materia del cuestionario. Los resultados aparecen en la siguiente tabla:

	Método a	Método b	Método c
Hombre	15, 16, 18, 19, 14	25, 27, 28, 23, 29	21, 22, 18, 17, 20
Mujer	24, 27, 29, 25, 23	17, 15, 13, 16, 18	20, 19, 22, 17, 23

a) Crear un conjunto de datos con las variables sexo, método y días.

b) Mostrar el gráfico de medias de los del tiempo de aprendizaje para los distintos grupos según el sexo y el método de aprendizaje. ¿Qué se puede decir de la interacción de las variables ?

Indicación

1) Seleccionar el menú `Gráficas→Gráfica de las medias`.
2) En el cuadro de diálogo que aparece seleccionar la variable días como `Variable explicada`, las variables método y sexo como `Factores`, seleccionar la opción `Sin barra de errores` y hacer click sobre el botón Aceptar.

Es evidente que las líneas se cruzan, lo cual indica que hay interacción.

c) Realizar un contraste de ANOVA de dos vías con interacción e interpretar los resultados.

Indicación

1) Seleccionar el menú `Estadísticos->Medias->ANOVA de múltiples factores`.
2) En el cuadro de diálogo que aparece, seleccionar la variable días como `Variable explicada` y las variables sexo y método como `Factores`, introducir un nombre para el modelo y hacer click sobre el botón Aceptar.

Se puede observar que no hay diferencias significativas asociadas al sexo ni al método. Sin embargo sí que hay interacción, es decir la diferencia en el tiempo de respuesta entre hombres y mujeres depende del método seguido, e igualmente las diferencias entre los tiempos de respuesta según los diferentes métodos dependen del sexo.

d) Calcular los intervalos de confianza para la diferencia de medias en el tiempo de aprendizaje entre los grupos que surgen de la interacción del sexo con el método de aprendizaje.

Indicación

Repetir los pasos del apartado anterior pero seleccionando la opción `Intervalos de comparación de medias de Tukey`.

4. Se desea comparar la rapidez con la que aparece el efecto de tres nuevos agentes repigmentadores: A, B y C. Con esta intención, se aplican de manera tópica dosis equivalentes de los tres repigmentadores en zonas de la piel con pérdida total de pigmentación en los mismos ocho pacientes con vitíligo. A continuación, se recoge el tiempo, en días, que tardan en aparecer los primeros signos de repigmentación:

A	B	C
19	3	31
11	2	9
7	4	16
4	1	6
3	2	8
5	7	18
7	1	5
4	3	9

a) Crear un conjunto de datos con las variables individuo, tiempo y repigmentador.

Indicación

Aunque todos los repigmentadores se aplican a cada individuo, para hacer un ANOVA de medidas repetidas hay que introducir el individuo como el factor inter-sujetos, mientras que el repigmentador sería el factor-intrasujetos

.

Si los datos de la variable individuo se introducen como números, es preciso convertirla en un factor:

1) Seleccionar el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Convertir variable numérica en factor`.
2) En el cuadro de diálogo que aparece seleccionar la variable individuo, activar la opción `Utilizar números` y hacer click en el botón Acetar.
3) En la ventana que aparece preguntando si se desea sobreescribir la variable hacer click en el botón Si.

b) Realizar un ANOVA de medidas repetidas e interpretar el resultado obtenido.

Indicación

1) Seleccionar el menú `Medias`→`ANOVA de medidas repetidas`.
2) En el cuadro de diálogo que aparece, seleccionar como variable explicada el tiempo, como factor la variable individuo y como grupos la variable repigmentador, darle un nombre al modelo y hacer click en el botón Aceptar.

c) ¿Entre qué medidas del tratamiento repigmentador se dan diferencias estadísticamente significativas?

Indicación

Repetir los pasos del apartado anterior pero seleccionando la opción `Intervalos de comparación de medias de Tukey`.

3 Ejercicios propuestos

1. En un estudio se quiere analizar la influencia sobre la ansiedad social, cuantificada mediante una escala numérica que va de 0 a 10, de la edad, dividida en tres categorías, y si se fuma o no. Los datos obtenidos fueron:

	Fumar No	Fumar Sí
Edad 1	3,91; 5,01; 4,47; 3,33; 4,71	4,83; 3,95; 4,04; 3,66; 9,44
Edad 2	5,65; 6,49; 5,50; 5,72; 5,44	9,66; 7,68; 9,57; 7,98; 7,39
Edad 3	4,94; 7,13; 5,54; 5,94; 6,16	5,92; 5,48; 5,19; 6,12; 4,45

   a) Considerando la posibilidad de interacción entre las variables independientes, ¿se puede considerar que la edad, expresada en forma de categorías, influye en la ansiedad? ¿Y el fumar? ¿Se puede considerar que el fumar o no influye de forma diferente en la ansiedad dependiendo de la categoría de edad analizada?

   b) Dependiendo de los resultados del apartado anterior, ¿entre qué medias habría diferencias estadísticamente significativas? Calcular los intervalos de confianza para las diferencias.

2. En un estudio se quiere analizar la eficacia de dos tipos de entrenamiento (A1: entrenamiento sólo físico, A2: entrenamiento físico + entrenamiento psicológico) para mejorar el rendimiento físico. Para ello, se dispone de una muestra de 8 individuos con los que se generan dos grupos de 4 asignados aleatoriamente, y se mide su rendimiento físico mediante un test de rendimiento numérico que va de 0 a 15 puntos. Los 8 individuos son sometidos al test en 4 momentos diferentes (B1: al cabo de una semana de entrenamiento, B2: al cabo de dos, B3: al cabo de tres y B4: al cabo de 4). Los datos obtenidos fueron:

	B1	B2	B3	B4
A1	3	4	7	7
	6	5	8	8
	3	4	7	9
	3	3	6	8
A2	1	2	5	10
	2	3	6	10
	2	4	5	9
	2	3	6	11

   a) ¿Influye significativamente la semana en la que se realiza el test en el resultado? ¿Y el tipo de entrenamiento? ¿Es significativa la interacción entre tipo de entrenamiento y la semana en la que se realiza el test?

   b) Entre qué medias hay diferencias estadísticamente significativas? Calcular los intervalos de confianza para las medias y para las diferencias.

3. Se ha aplicado un dispositivo electrónico que mide la frecuencia cardíaca a 10 estudiantes. Se realizó una primera medición un minuto antes de que comenzasen a hacer un examen, la segunda medición se hizo cuando llevaban 15 minutos realizando el examen, la tercera un minuto después de entregarlo y la cuarta 15 minutos después de terminar. Los resultados fueron:

Estudiante	Medida1	Medida2	Medida3	Medida4
1	57	61	77	70
2	73	87	88	83
3	75	89	89	65
4	75	60	67	68
5	77	87	67	67
6	88	96	84	55
7	89	65	89	60
8	101	80	77	60
9	103	85	76	66
10	107	73	69	60

¿Son las mediciones significativamente distintas entre sí? Si hay diferencia, ¿entre qué mediciones se dan?

Contrastes de Hipótesis No Paramétricos

1 Fundamentos teóricos

Gran parte de los procedimientos estadísticos diseñados para hacer inferencia (prueba T para contrastar hipótesis sobre medias, prueba F para contrastar hipótesis sobre varianzas...), presentan tres características comunes:

- Permiten contrastar hipótesis referidas a algún parámetro poblacional (μ, σ, ...), o relaciones entre parámetros poblacionales ($\mu_1 - \mu_2$, σ_1/σ_2, ...).
- Exigen el cumplimiento de determinados supuestos sobre las poblaciones originales de las que se extraen las muestras, como la normalidad o la igualdad de varianzas (homogeneidad de varianzas u homocedasticidad), e incluso sobre la manera de obtener los datos, como la aleatoriedad en las observaciones.
- Están diseñados para trabajar con variables cuantitativas.

A los procedimientos estadísticos que presentan las tres características anteriores se les denomina **Contrastes Paramétricos** o **Pruebas Paramétricas**, y son las técnicas estadísticas más utilizadas. No obstante, presentan dos inconvenientes que hacen que su utilidad se vea reducida: por un lado, exigen el cumplimiento de supuestos que en determinadas ocasiones pueden resultar demasiado restrictivos (a menudo las distribuciones no son normales, o no se tiene homogeneidad de varianzas...); por otro, no siempre se puede trabajar con variables cuantitativas (en ciencias sociales y de la salud, en muchos casos serán cualitativas y en el mejor de los casos cualitativas ordinales).

Afortunadamente, existen contrastes que permiten poner a prueba hipótesis que no se refieren específicamente a parámetros poblacionales (como μ o σ), o que no exigen supuestos demasiado restrictivos en las poblaciones originales, e incluso que tampoco necesitan trabajar con variables cuantitativas. Este tipo de contrastes que no cumplen alguna de las tres características de los contrastes paramétricos reciben el nombre de **Contrastes No Paramétricos** o **Pruebas No Paramétricas**.

En realidad, en términos estrictos, para tener un contraste no paramétrico sería suficiente con que en el mismo no se plantease una hipótesis sobre algún parámetro poblacional. Por eso, algunos autores distinguen entre contrastes no paramétricos y contrastes de distribución libre (distribution-free), o exentos de distribución, que serían los que no impondrían supuestos restrictivos en la distribución de la población original. No obstante, como casi todos los contrastes no paramétricos son a su vez de distribución libre, hoy en día se denomina contraste no paramétrico a cualquiera que no cumpla alguna de las tres características señaladas.

Los mayores atractivos de los contrastes no paramétricos son:

- Al exigir condiciones menos restrictivas a la muestra sobre las características de la distribución de la población de la que se ha extraído, son más generales que los paramétricos (se pueden aplicar tanto a situaciones en las que no se deberían aplicar paramétricos como a situaciones en las que sí que se podrían aplicar).
- Son especialmente útiles cuando hay que analizar muestras pequeñas, ya que, con muestras grandes ($n \geq 30$), aunque la población de partida no siga una distribución normal, el Teorema Central del Límite garantiza que la variable media muestral sí que será normal, y no habrá ningún problema con

contrastes sobre la media poblacional basados en la media muestral cuyo comportamiento es normal (por ejemplo, contrastes con la T de Student). Con muestras pequeñas ($n < 30$), si la distribución de los datos de la muestra no es normal tampoco podemos garantizar que lo sea la distribución de la media muestral, y no quedará más remedio que acudir a contrastes no paramétricos cuando se pretenda ver si hay o no diferencias significativas entre distribuciones.

- Permiten trabajar tanto con variables cualitativas como cuantitativas. En éstas últimas, en lugar de trabajar con los valores originales, generalmente se trabaja con su rangos, es decir, simplemente con el número de orden que ocupa cada valor, por lo que incluso se simplifican los cálculos necesarios para aplicar los contrastes.

Por otro lado, también hay inconvenientes:

- Si en un problema concreto se puede aplicar tanto una prueba paramétrica como una no paramétrica, la paramétrica presentará mayor potencia, es decir, mayor capacidad de detección de diferencias significativas; o lo que es lo mismo, la no paramétrica necesitará mayores diferencias muestrales para concluir que hay diferencias poblacionales.
- Con las pruebas paramétricas no sólo se pueden realizar los contrastes de hipótesis planteados sobre los adecuados parámetros poblacionales para llegar al p-valor del contraste, sino que también se pueden generar intervalos de confianza con los que delimitar adecuadamente el posible valor de dichos parámetros. Con el p-valor se genera un número que cuantifica si el efecto analizado ha sido o no estadísticamente significativo, pero puede que el efecto sea tan pequeño que no resulte clínicamente relevante, mientras que con el intervalo de confianza sí que se aprecia la magnitud del efecto, y por lo tanto se puede concluir si ha sido o no clínicamente relevante. Igualmente, la aplicación de una prueba no paramétrica permitirá obtener un p-valor del contraste, pero en muy pocas de ellas hay procedimientos desarrollados que permitan obtener intervalos de confianza.

En las técnicas no paramétricas juega un papel fundamental la ordenación de los datos, hasta el punto de que en gran cantidad de casos ni siquiera es necesario hacer intervenir en los datos observados (variables cuantitativas observadas) más que para establecer una relación de menor a mayor entre los mismos. El número de orden que cada dato ocupa recibe el nombre de rango del dato. Por ello, gran parte de los contrastes no paramétricos, además de sobre variables cuantitativas, también pueden aplicarse sobre variables cualitativas ordinales, a cuyos valores también se les podrá asignar un rango (un número de orden).

Por otra parte, el proceso para llevar a cabo un contraste no paramétrico es muy parecido al de uno paramétrico:

1. Planteamos una hipótesis nula H_0 y su correspondiente alternativa H_1.
2. Suponiendo cierta la H_0, se calcula el valor de un estadístico muestral de distribución conocida.
3. A partir del estadístico, se calcula el p-valor del contraste y se acepta o se rechaza H_0 dependiendo de que el p-valor obtenido haya sido, respectivamente, mayor o menor que el nivel de significación marcado en la prueba (habitualmente $0,05$).

1.1 Contrastes no paramétricos más habituales

Los contrastes no paramétricos más habituales son:

- Para analizar la **relación entre dos variables cualitativas nominales** (diferencia de proporciones en las tablas de contingencia), el contraste no paramétrico más importante es el que se basa en el **Test de la Chi-Cuadrado**, al que se ha dedicado toda una práctica previa (Contrastes basados en el estadístico χ^2. Comparación de proporciones).
- Para analizar la **aleatoriedad de una muestra**, sobre todo se utiliza el **Test de Rachas**.
- Para analizar la **normalidad de los datos de una muestra**, sobre todo se utiliza el **Contraste de Kolmogorov-Smirnov**, aunque también se utilizan criterios basados en los estadísticos de Asimetría y Curtosis, el Contraste de Shapiro y Wilk, e incluso diversos métodos que se apoyan en gráficos, como los gráficos P-P o Q-Q.

- Para analizar las **diferencias entre dos muestras, con datos independientes, en variables cuantitativas u ordinales**, sobre todo se utiliza el **Test de la U de Mann-Whitney**, cuyo correspondiente paramétrico para variables cuantitativas es la T de Student de datos independientes.

- Para analizar las **diferencias entre dos muestras, con datos pareados, en variables cuantitativas u ordinales**, sobre todo se utiliza el **Test de Wilcoxon**, cuyo correspondiente paramétrico para variables cuantitativas es la T de Student de datos independientes.

- Para analizar las **diferencias entre varias muestras (más de dos), con datos independientes, en variables cuantitativas u ordinales**, sobre todo se utiliza el **Test de Kruskal-Wallis**, cuyo correspondiente paramétrico para variables cuantitativas es el ANOVA.

- Para analizar las **diferencias entre varias muestras, con datos pareados, en variables cuantitativas u ordinales**, sobre todo se utiliza el **Test de Friedman**, cuyo correspondiente paramétrico para variables cuantitativas es el ANOVA de Medidas Repetidas.

- Para analizar la **homogeneidad de varianzas de varias muestras** (dos o más), sin tener en cuenta la normalidad de los datos, sobre todo se utiliza el **Test de Levene**, cuyo correspondiente paramétrico cuando se comparan dos muestras es el test de igualdad de varianzas mediante la distribución F de Fisher.

- Por último, para analizar la **correlación entre dos variables cuantitativas u ordinales**, sobre todo se utiliza el **coeficiente de correlación Rho de Spearman**, junto con su test asociado para ver si la correlación es o no significativa

A continuación se analizan todos los contrastes enumerados.

1.2 Aleatoriedad de una muestra: Test de Rachas

Frecuentemente suponemos que la muestra que utilizamos es aleatoria simple, es decir, que las n observaciones se han tomado de manera aleatoria e independientemente unas de otras, y de la misma población. Pero, en general, esta hipótesis no siempre se puede admitir, siendo por tanto necesario, en muchos casos, contrastar si realmente estamos frente a una muestra aleatoria simple o no.

El Test de Rachas es una prueba que mide hasta qué punto el valor de una observación en una variable puede influir en las observaciones siguientes. Por ejemplo, si la variable analizada es el sexo y la muestra es aleatoria extraída de una población con aproximadamente el mismo número de mujeres que de hombres, entonces resultará prácticamente imposible que después de preguntar a un individuo que resulta ser hombre, los 10 individuos siguientes también sean hombres; en todo caso, podrían aparecer "rachas"de 3 o 4 hombres seguidos, pero no de 10.

Consideremos una muestra de tamaño n que ha sido dividida en dos categorías X e Y con n_1 y n_2 observaciones cada una. Se denomina racha a una sucesión de valores de la misma categoría. Por ejemplo, si estudiamos una población de personas y anotamos el sexo: X = Hombre e Y= Mujer, y obtenemos la secuencia: $XXXYYXYYY$, tendríamos $n_1 = 4$, $n_2 = 5$, $n = n_1 + n_2 = 9$, y el número de rachas igual a 4. En función de las cantidades n_1 y n_2 se espera que el número de rachas no sea ni muy pequeño ni muy grande, lo cual equivale a decir que no pueden aparecer rachas demasiado largas, ni demasiadas rachas cortas.

Si las observaciones son cantidades numéricas, pueden dividirse en dos categorías que posean aproximadamente el mismo tamaño sin más que considerar la mediana de las observaciones como valor que sirve para dividir la muestra entre los valores que están por encima de la mediana y los que están por debajo.

Con todo ello, el Test de Rachas permite determinar si la variable aleatoria número observado de rachas, R, en una muestra de tamaño n (dividida en dos categorías de tamaños n_1 y n_2), es lo suficientemente grande o pequeño como para rechazar la hipótesis de independencia (aleatoriedad) entre las observaciones. Para el contraste generalmente se utiliza el estadístico:

$$Z = \frac{R - E(R)}{\sigma_R}$$

donde:

$$E(R) = \frac{2n_1n_2}{n+1}$$
$$\sigma_R = \sqrt{\frac{2n_1n_2\left(2n_1n_2 - n\right)}{n^2\left(n+1\right)}}$$

En el supuesto de que se cumpla la hipótesis nula de aleatoreidad de la muestra, se puede demostrar que el estadístico Z sigue una distribución normal tipificada, que es la que se utiliza para dar el p valor de la prueba.

Por último, conviene no confundir la hipótesis de aleatoriedad de una muestra con la hipótesis de bondad de ajuste de los datos de una muestra mediante una distribución dada. Por ejemplo, si en 10 lanzamientos de una moneda obtenemos 5 caras y 5 cruces, sin duda los datos se ajustan adecuadamente a una binomial con probabilidad de éxito igual a $0,5$, y para comprobarlo, por ejemplo, se podría utilizar un test de Chi-cuadrado; pero si las 5 caras salen justo en los 5 primeros lanzamientos y las 5 cruces en los últimos, difícilmente se podría cumplir la hipótesis de independencia o aleatoriedad.

1.3 Pruebas de Normalidad

Ya hemos comentado que muchos procedimientos estadísticos (intervalos de confianza, tests de hipótesis, análisis de la varianza...) únicamente son aplicables si los datos provienen de distribuciones normales. Por ello, antes de su aplicación conviene comprobar si se cumple la hipótesis de normalidad.

No obstante, las inferencias respecto a medias son en general robustas (no les afecta demasiado la hipótesis de normalidad), sea cual sea la población base, si las muestras son grandes ($n \geq 30$) ya que la distribución de la media muestral es asintóticamente normal. Por ello, gran parte de los métodos paramétricos son válidos con muestras grandes incluso si las distribuciones de partida dejan de ser normales, pero, aunque válidos, las inferencias respecto a la media dejan de ser óptimas; es decir los métodos paramétricos pierden precisión, y esto se traduce en intervalos innecesariamente grandes o contrastes poco potentes.

Por otro lado, las inferencias respecto a varianzas son muy sensibles a la hipótesis de normalidad, por lo que no conviene construir intervalos o contrastes para varianzas si no tenemos cierta seguridad de que la población es aproximadamente normal.

Entre los múltiples métodos que se utilizan para contrastar la hipótesis de normalidad destacan:

- El cálculo de los estadísticos de asimetría y curtosis de la distribución.
- La prueba de Kolmogorov-Smirnov, especialmente en su versión corregida por Lilliefors.
- La prueba de Saphiro-Wills.
- Gráficos Q-Q y P-P de comparación con la distribución normal.

Conviene hacer la aclaración de que la prueba de Kolmogorov-Smirnov no tiene aplicación específica al contraste de normalidad, sino que, como también sucede con la Chi-cuadrado, la prueba de Kolmogorov-Smirnov, ha sido desarrollada para establecer el ajuste de los datos observados mediante un modelo teórico (una distribución de probabilidad) que no tiene que ser específicamente normal. La corrección de Lilliefors adapta la prueba de Kolmogorov-Smirnov para un contraste específico de normalidad.

En cuanto a la conveniencia de una u otra prueba, hay que destacar que no existe un contraste óptimo para probar la hipótesis de normalidad. La razón es que la potencia relativa de un contraste de normalidad depende del tamaño muestral y de la verdadera distribución que genera los datos. Desde un punto de vista poco riguroso, el contraste de Shapiro-Wilks es, en términos generales, el más conveniente en pequeñas muestras ($n < 30$), mientras que el de Kolmogorov-Smirnov, en la versión modificada de Lilliefors, es más adecuado para muestras grandes.

Pasamos a dar una pequeña explicación más detallada de cada uno de los contrastes citados.

Estadísticos de asimetría y curtosis

Una primera idea de si los datos provienen de una distribución normal, nos la pueden dar los estadísticos de asimetría, g_1, y de curtosis, g_2, junto con sus correspondientes errores estándar, EE_{g1} y EE_{g2}. Sabemos que las distribuciones normales son simétricas, y por tanto $g_1 = 0$, y tienen un apuntamiento normal, $g_2 = 0$. Según esto, si una muestra proviene de una población normal sus coeficientes de asimetría y de apuntamiento no deberían estar lejos de 0, y se acepta que no están demasiado lejos de 0 cuando:

$$g_1 < 2EE_{g1} \quad \text{o} \quad g_2 < 2EE_{g2}.$$

Contraste de Kolmogorov-Smirnov y corrección de Lilliefors

Este contraste general, válido para comprobar si los datos de una variable aleatoria continua se ajustan adecuadamente mediante un modelo de distribución continuo, compara la función de distribución teórica con la empírica de una muestra. Tiene la ventaja de que no requiere agrupar los datos. Utiliza el estadístico:

$$D_n = \sup |F_n(x) - F(x)|$$

donde $F_n(x)$ es la función de distribución empírica muestral (la probabilidad acumulada hasta un cierto valor de la variable suponiendo que en cada uno de los n valores de la muestra acumulamos una probabilidad igual a $1/n$) y $F(x)$ es la función de distribución teórica (probabilidad acumulada hasta un cierto valor de la variable, o lo que es lo mismo $\int_{-\infty}^{x} f(t)dt$, donde $f(t)$ es la función de densidad teórica).

Con el estadístico D_n, se calcula:

$$Z_{K-S} = \frac{D_n}{\sqrt{n}}$$

que sigue una distribución normal tipificada.

Para el caso específico en que se contrasta si la muestra proviene de una distribución normal, más que el estadístico Z_{K-S} se utiliza directamente la D_n pero con una distribución tabulada corregida introducida por Lilliefors, por lo que a veces se habla del test de Kolomogorov-Smirnov-Lilliefors, específico para el contraste de normalidad.

En general, el test de Kolmogorov-Smirnov sin corrección resulta menos potente, y por lo tanto será más difícil rechazar la normalidad de los datos, que el corregido por Lilliefors, especialmente si la presenta falta de normalidad se produce porque la distribución de partida presenta valores atípicos.

Ya sea mediante el método de Kolmogorov-Smirnov en general o mediante la corrección introducida por Lilliefors, obtenemos un p valor del contraste con el que aceptar o rechazar la hipótesis nula de normalidad de los datos de la muestra.

Contraste de Shapiro-Wilk

Es una prueba sólo válida para contrastar el ajuste de unos datos muestrales mediante una distribución normal. Además, por su potencia, se considera que es la más adecuada para analizar muestras pequeñas ($n < 30$).

En el proceso, se calcula un estadístico llamado W de Shapiro-Wilk (la forma detallada de obtenerlo va más allá del nivel de esta práctica), cuya distribución está tabulada. Con ello, de nuevo se genera un p valor que se utiliza para aceptar o no la hipótesis nula de normalidad de los datos de la muestra.

Gráficos Q-Q y P-P de comparación con la distribución normal

Son procedimientos gráficos que permiten llegar a conclusiones cualitativas sobre si los datos se ajustan adecuadamente mediante una distribución normal.

- El gráfico Q-Q: presenta en el eje de abcisas los valores de la variable, y en el eje de ordenadas el valor que le correspondería en una distribución normal tipificada según las probabilidades acumuladas obtenidas gracias a la función de distribución empírica. Junto con los puntos obtenidos, también se representa la recta que se obtendría sin más que tipificar los datos teniendo en cuenta la media y la desviación típica de los mismos, de manera que si los datos se ajustasen de forma perfecta mediante una distribución normal, los puntos quedarían justo en la recta, mientras que si las distancias entre los puntos y la recta son grandes no cabría aceptar la normalidad del conjunto de datos.

- El gráfico P-P: es muy parecido al Q-Q pero directamente representa en un eje la probabilidad acumulada observada hasta cada valor de la variable teniendo en cuenta la función de distribución empírica y en el otro la probabilidad acumulada teórica que le correspondería si los datos se ajustasen perfectamente mediante una distribución normal. De nuevo se representa la recta que se produciría si el ajuste fuese perfecto.

1.4 Test de la U de Mann-Whitney

Este test sustituye a la T de Student para comparar las medias de dos grupos independientes cuando no se cumplen los supuestos en los que se basa la prueba T. Como requiere ordenar los valores antes de hacer el test, no compara realmente las medias, sino los denominados rangos o números de orden de los datos.

Se debe usar cuando:

- Alguna de las dos muestras contiene menos de 30 observaciones y no se puede asumir normalidad.
- La comparación se realiza en una variable ordinal en vez de ser realmente cuantitativa.
- La muestra es muy pequeña (menos de 10 observaciones en alguno de las dos grupos).

Intuitivamente, si se ordenan de menor a mayor conjuntamente los datos de dos muestras y se les asigna a cada uno de ellos su rango (es decir su número de orden dentro de la lista obtenida al unir los datos de las dos muestras, teniendo en cuenta que si hay dos o más datos iguales se les asigna a todos ellos el rango promedio de los que les corresponderían si fuesen distintos) es natural que, si se cumple que las dos muestras originales están igualmente distribuidas, los rangos se mezclen al azar y no que los rangos de una de las muestras aparezcan al principio y los de la otra al final, pues esto sería indicio de que una de las muestras tendría valores sistemáticamente mayores o menores que la otra. Con esta idea, se plantean las hipótesis:

- H_0: Las poblaciones de las que provienen las muestras están igualmente distribuidas.
- H_1: Las poblaciones difieren en su distribución.

Para llevar a cabo el contraste de hipótesis, se obtienen las sumas de rangos de la muestra 1, R_1, y la de la muestra 2, R_2. Con R_1 y R_2 se calculan los estadísticos:

$$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$
$$U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2 = n_1 n_2 - U_1$$

Y teniendo en cuenta U_1 y U_2, se toma U, que es el mínimo de U_1 y U_2, del que se conoce su distribución de probabilidad y con el que se puede calcular el p valor del contraste. No obstante, para muestras grandes, ($n \geq 30$), los cálculos con la distribución exacta pueden ser complicados y por eso se suele trabajar con un estadístico Z que sigue una distribución normal tipificada:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{n (n-1)} \left(\frac{n^3 - n}{12} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12} \right)}}$$

donde k es el número de rangos distintos en los que existen empates y t_i el número de puntuaciones distintas empatadas en el rango i.

1.5 Test de Wilcoxon para datos emparejados

Este test, también llamado prueba de rangos con signo de Wilcoxon, se debe utilizar en lugar de la T para datos emparejados cuando:

- Los datos a comparar son ordinales.

- Son datos cuantitativos pero la muestra es pequeña (< 30) y además no sigue una distribución normal en la variable diferencia entre las dos mediciones emparejadas.

Intuitivamente, si calculamos las diferencias individuo a individuo en las dos variables emparejadas, suponiendo que no hay diferencia global entre las dos variables se obtendrán aproximadamente tantas diferencias positivas como negativas. Además, sería conveniente trabajar con un test que no sólo detecte si la diferencia ha sido positiva o negativa, sino que también debería tener en cuenta la cuantía de la diferencia, o al menos el orden en la cuantía de la diferencia; con ello se podrían compensar situaciones en las que hay pocas diferencias negativas de mucha cuantía frente a muchas positivas pero de poca cuantía, o a la inversa. Con esta idea, se plantean las hipótesis:

- H_0: No hay diferencia entre las observaciones emparejadas.
- H_1: Sí que las hay.

Para realizar el contraste de hipótesis, en primer lugar se calculan las diferencias entre los datos emparejados de las variables X e Y para cada uno de los n individuos:

$$D_i = X_i - Y_i \quad i = 1, ..., n$$

Posteriormente se desechan la diferencias nulas y se toman los valores absolutos de todas las diferencias, $|D_i|$, asignándoles rangos (órdenes), R_i (si hay empates, se asignan rangos promedio). Después se suman por separado los rangos de las diferencias que han resultado positivas, $S_+ = \sum R_i^+$, y por otra parte los rangos de las diferencias que han resultado negativas, $S_- = \sum R_i^-$. Y con ello, si H_0 fuese correcta, la suma de rangos positivos debería ser muy parecida a la suma de rangos negativos: $S_+ = S_-$.

Por último, teniendo en cuenta que se conoce la distribución de probabilidad de los estadísticos S_+ y S_-, se puede calcular el p valor del contraste con la hipótesis nula de su igualdad. No obstante, para muestras grandes, los programas de estadística suelen utilizar un nuevo estadístico Z que sigue una distribución normal tipificada:

$$Z = \frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{48}}}$$

S es el mínimo de S_+ y S_-, k el número de rangos distintos en los que existen empates, y t_i el número de puntuaciones empatadas en el rango i.

1.6 Test de Kruskal-Wallis: comparación no paramétrica de k medias independientes

Es el test no paramétrico equivalente en su uso al ANOVA, de tal forma que permite contrastar si k muestras independientes (con $k \geq 3$) han sido obtenidas de una misma población y por lo tanto presentan igual distribución.

Se debe usar el Test de Kruskal-Wallis en lugar del ANOVA cuando:

- Los datos son ordinales.
- No hay normalidad en alguna de las muestras.
- No hay homogeneidad de varianzas (la homogeneidad de varianzas recibe el nombre de homocedasticidad).

Su uso está también especialmente indicado en el caso de muestras pequeñas y/o tamaños muestrales desiguales, ya que entonces es muy arriesgado suponer normalidad y homocedasticidad de los datos.

En esencia, el Test de Kruskal-Wallis es una extensión del test de la U de Mann-Whitney, pero trabajando con k muestras en lugar de 2.

Las hipótesis que se contrastan son:

- H_0: Las poblaciones de las que provienen las k muestras están igualmente distribuidas.

- H_1: Alguna de las poblaciones difiere en su distribución con respecto a las demás.

Para realizar el contraste, primero se ordenan de menor a mayor todos los valores observados en las k muestras. Luego se asigna el rango 1 al valor inferior, el rango 2 al 2º valor y así sucesivamente. En caso de empate entre dos valores se asigna la media de los números de orden de los individuos empatados. Después se suman por separado los rangos asignados a las observaciones de cada grupo y se obtiene la suma de rangos de cada grupo: $R_i, i = 1, ..., k$. Con la suma de rangos dentro de cada grupo se puede obtener el rango medio de cada grupo sin más que dividir la suma entre el número total de observaciones en el grupo: $R_i/n_i, i = 1, ..., k$. Y por último, si la hipótesis nula fuera cierta, los rangos medios de cada grupo serían muy parecidos entre sí y muy parecidos al rango medio total.

Basándose en lo anterior, el estadístico H de Kruskall-Wallis se calcula mediante la fórmula:

$$H = \frac{12}{n\,(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\,(n+1)$$

Y se puede demostrar que, si la hipótesis nula es cierta, H sigue una distribución Chi-cuadrado con $k-1$ grados de libertad, que es lo que se utiliza para calcular el p-valor del contraste. Si el número de muestras es 3 y el número de observaciones en alguna no pasa de 5, el estadístico H no sigue una distribución Chi-cuadrado, pero su distribución está convenientemente tabulada y los programas estadísticos aplican las adecuadas correcciones.

Si el resultado del test fuera significativo, para buscar entre qué grupos existen diferencias, se harán comparaciones por parejas con la U de Mann-Whitney, pero penalizando los p valores obtenidos; esto quiere decir que los p valores de cada una las parejas comparadas deben multiplicarse por el número de comparaciones realizadas. Con ello se logra controlar el error de tipo I en el contraste global, es decir la probabilidad de que haya aparecido, simplemente por azar y no porque realmente haya diferencias, algún p valor menor que el nivel de significación fijado. Por ejemplo, en un problema con 4 grupos, habría que hacer 6 comparaciones de parejas distintas, y trabajando con un p-valor frontera de 0,05, la probabilidad de que apareciese alguna diferencia significativa entre las parejas simplemente por azar viene dada por: $1 - 0,05^0 \cdot 0,95^6 = 0,265$ (probabilidad de algún éxito en una binomial de 6 intentos y probabilidad de éxito 0,05), que queda muy alejado del 0,05 global con el que se suele trabajar; mientras que con un $p - valor$ frontera de $0,05/4 = 0,0125$, $1 - 0,0125^0 \cdot 0,9875^5 = 0,073$, que es muy parecido $0,05$. Igualmente, con n grupos, habría que exigir un p-valor en cada comparación de $0,05/n$; o lo que es lo mismo, no considerar como significativas ninguna de las comparaciones entre parejas en las que el p-valor obtenido multiplicado por n sea mayor que $0,05$.

1.7 Test de Friedman: equivalente no paramétrico del ANOVA con medidas repetidas

qEs el test no paramétrico equivalente al ANOVA de medidas repetidas, y por lo tanto se aplica en situaciones en las que para cada individuo tenemos varias medidas (3 o más) en diferentes tiempos, en una situación muy similar a la del Test de Wilcoxon para datos emparejados, pero en este último caso con sólo 2 medidas en cada individuo. Generalmente cada medida representa el resultado de un tratamiento, por lo que habitualmente se habla de que tenemos n individuos con k tratamientos, siendo $k \geq 3$.

Se usa el Test de Friedman en lugar del ANOVA de medidas repetidas cuando:

- Se comparan medidas repetidas ordinales en lugar de cuantitativas.
- Alguna de las variables diferencia, generadas mediante la diferencia de todos los tratamientos tomados dos a dos, no siga una distribución normal.
- Alguna de las variables diferencia, generadas mediante la diferencia de todos los tratamientos tomados dos a dos, no tenga la misma varianza.

Cuando todas las variables diferencia sigan distribuciones normales con la misma varianza, se dice que los datos cumplen con el supuesto de Esfericidad, necesario para aplicar el ANOVA de medidas repetidas. Para contrastar el supuesto de esfericidad de los datos se suele utilizar el Test de Esfericidad de Mauchly.

De nuevo, con el test de Friedman, las hipótesis del contraste a realizar son:

- H_0: No hay diferencia entre los diferentes tratamientos.
- H_1: Al menos alguno de los tratamientos presenta un comportamiento diferente del resto.

Para realizar el contraste, se reemplazan los datos de cada sujeto por su rango dentro de cada fila, es decir por su posición, una vez ordenados de menor a mayor los datos correspondientes a las diferentes observaciones de cada uno de los sujetos. En el caso de empates se asignará el rango promedio de los valores empatados. Con ello, tenemos una matriz de rangos R_{ij}, donde $i = 1, ..., n$ siendo n el número de individuos, y $j = 1, ..., k$, siendo k el número de tratamientos. Después se suman los rangos correspondientes a cada una de las mediciones realizadas:

$$R_j = \sum_{i=1}^{n} R_{ij} \quad j = 1, ..., k$$

y se calculan sus correspondientes promedios sin más que dividir entre n:

$$\bar{R}_j = \frac{R_j}{n} \quad j = 1, ..., k$$

En estas condiciones, si la hipótesis nula fuera cierta, los rangos promedio dentro de cada tratamiento deberían ser similares, por lo que es posible plantearse de nuevo un contraste basado en la Chi-cuadrado. El estadístico, debido a Friedman, en el que se va a basar el contraste es:

$$\chi^2 = \frac{12}{n\,k\,(k+1)} \sum_{j=1}^{k} R_j^2 - 3n\,(k+1)$$

Este estadístico sigue una distribución Chi-cuadrado con $k-1$ grados de libertad y es el que se utiliza para calcular el p valor del contraste. Si una vez realizado el contraste se rechaza la hipótesis nula, habrá que determinar qué tratamientos son los que presentan un comportamiento diferente del resto. Para ello, habrá que aplicar un test de Wilcoxon para datos emparejados a cada una de las parejas que se puedan formar teniendo en cuenta los diferentes tratamientos (los diferentes tiempos en los que se ha medido la respuesta). Por ejemplo, si tenemos 3 tratamientos para cada individuo, se pueden generar tres comparaciones pareadas: 1-2, 1-3 y 2-3. Después de aplicar el test de Wilcoxon, debemos corregir el p valor obtenido en cada cruce multiplicándolo por el número de comparaciones realizado. Por ejemplo, si en la comparación 1-2 hemos obtenido un p valor igual a 0,03, teniendo en cuenta que hay 3 comparaciones posibles, el p valor corregido sería 0,09.

1.8 Test de Levene para el contraste de homogenidad de varianzas

Como ya se ha comentado, en algunos test paramétricos que realizan la comparación de medias de diferentes muestras se exige la condición de que las muestras tengan igualdad de varianzas (homogeneidad de varianzas u homocedasticidad), especialmente en el ANOVA y también en algunos tipos concretos de T de Student. En realidad, se puede utilizar la distribución F de Fisher para comprobar la homogeneidad de varianzas de dos poblaciones de las que se han extraído muestras aleatorias, pero tiene el problema de que exige la normalidad de los datos y además está limitado a la comparación de dos varianzas. Por lo tanto, sería conveniente disponer de un test no paramétrico que permita realizar el contraste de homogeneidad de varianzas sin suponer la normalidad de los datos, y además también permitir el contraste de homogeneidad de varianzas entre más de dos muestras. Eso mismo es lo que realiza el Test de Levene, cuyas hipótesis vinculadas son:

- H_0: Las varianzas poblacionales de las que se han extraído las k muestras son iguales ($\sigma_1 = \sigma_2 = ... = \sigma_k$).
- H_1: Al menos alguna de las varianzas es diferente al resto.

Para hacer el contraste se utiliza el estadístico de Levene:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i\,(Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2}$$

donde:

- k es el número de grupos diferentes.
- N es el número total de individuos.
- N_i es el número de individuos en el grupo i.
- Y_{ij} es el valor del individuo j en el grupo i.
- Para Z_{ij} se utiliza habitualmente uno de los dos siguientes criterios:

$$Z_{ij} = \begin{cases} \left|Yij - \bar{Y}_i.\right| \\ \left|Yij - \tilde{Y}_i.\right| \end{cases}$$

  donde $\bar{Y}_i.$ es la media del grupo i y $\tilde{Y}_i.$ es la mediana del grupo i. Cuando se utilizan las medianas en lugar de las medias, el Test de Levene suele recibir el nombre de Test de Brown-Forsythe.

- $Z..$ es la media de todas las Z_{ij}:

$$Z.. = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} Z_{ij}$$

- $Z_i.$ es la media de los Z_{ij} en el grupo i:

$$Z_i. = \frac{1}{N_i}\sum_{j=1}^{N_i} Z_{ij}$$

Al final, si se cumple la hipótesis nula de igualdad de varianzas se podría demostrar que el estadístico W debería seguir una distribución $F(k-1, N-k)$ de Fisher con $k-1$ y $N-k$ grados de libertad, que es la que se utiliza para calcular el p valor del contraste.

1.9 El coeficiente de correlación de Spearman

El coeficiente de correlación de Spearman, ρ, es el equivalente no paramétrico al coeficiente de correlación lineal, r. Se utiliza en lugar de r cuando:

- Las variables entre las que se analiza la correlación no siguen distribuciones normales.
- Cuando se quiere poner de manifiesto la relación entre variables ordinales.

Además, a diferencia del coeficiente de correlación lineal de Pearson, r, el coeficiente de Spearman no estima específicamente una asociación lineal entre las variables, sino que es capaz de detectar asociación en general. De hecho, las hipótesis del contraste de asociación de dos variables mediante ρ serían:

- H_0: No hay asociación entre las dos variables.
- H_1: Sí que hay asociación.

Para calcular el coeficiente ρ se utilizan los rangos de los valores (el orden de cada valor), tomando rangos promedio cuando tengamos dos o más valores iguales en alguna de las variables. Si tenemos dos variables X e Y, entre las que queremos ver si hay relación, con n valores, el primer paso es obtenemos los rangos: R_{x_i} y R_{y_i}, $i = 1, ..., n$. Después se calcula la diferencia entre rangos: $d_i = R_{x_i} - R_{y_i}$, $i = 1, ..., n$, y posteriormente se calcula ρ mediante la fórmula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}$$

Otra forma de obtener el coeficiente ρ, que conduce exactamente a los mismos resultados que la fórmula anterior, es calcular el coeficiente de correlación lineal r pero de los rangos en lugar del r de los valores de partida. El resultado obtenido para ρ, igual que con r, siempre está comprendido entre -1 y 1. Si ρ está cercano o es igual a 1, quiere decir que los rangos crecen a la ven en las dos variables, mientras que si su valor está cercano o es igual a -1, quiere decir que cuando en una variable crecen los rangos en la otra decrecen. Si está cercano a 0 quiere decir que no hay correlación entre los rangos.

Por último, una vez obtenido ρ el contraste de si existe o no asociación entre las variables se puede reformular en términos de si ρ puede o no ser igual a 0:

- H_0: $\rho = 0$.
- H_1: $\rho \neq 0$.

Para ello se utiliza el estadístico:

$$t_\rho = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

Que, bajo la hipótesis nula de no asociación entre las variables ($\rho = 0$) y siempre que el tamaño muestral no sea demasiado pequeño (por ejemplo, $n \geq 10$), sigue una distribución T de Student con $n-2$ grados de libertad, lo cual se puede utilizar para calcular el p valor del contraste.

2 Ejercicios resueltos

1. Sea una variable con distribución normal estándar. Se pide:

   a) Generar una muestra aleatoria de tamaño 100.

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución normal→Muestra de una distribución normal`.
   2) En el cuadro de diálogo que aparece introducir 100 en el campo `Numero de muestras`, 1 en el campo `Número de observaciones`, darle un nombre al conjunto de datos y hacer click en el botón Aceptar.
   3) Hacer click en el botón del Conjunto de datos y seleccionar el conjunto de datos con la muestra generada.

   b) Comprobar la normalidad de los datos mediante un diagrama Q-Q.

   Indicación
   1) Seleccionar el menú `Gráficas→Gráfica de comparación de cuantiles`.
   2) En el cuadro de diálogo que aparece seleccionar la variable obs, activar la opción correspondiente a la distribución normal y hacer click en el botón Aceptar.

   Como los datos provienen de una población normal, los puntos del diagrama deberían quedar alineados alrededor de la línea recta de la diagonal.

   c) Comprobar la normalidad de los datos mediante un contraste de Shapiro-Wilk.

   Indicación
   1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de normalidad de Shapiro-Wilk`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable obs en el campo `Variable` y hacer click en el botón Aceptar.

2. Sea una variable con distribución uniforme continua $U(0,1)$. Se pide:

   a) Generar una muestra aleatoria de tamaño 100.

   Indicación
   1) Seleccionar el menú `Distribuciones→Distribuciones continuas→Distribución uniforme→Muestra de una distribución uniforme`.
   2) En el cuadro de diálogo que aparece introducir 100 en el campo `Numero de muestras`, 1 en el campo `Número de observaciones`, darle un nombre al conjunto de datos y hacer click en el botón Aceptar.
   3) Hacer click en el botón del Conjunto de datos y seleccionar el conjunto de datos con la muestra generada.

   b) Comprobar la normalidad de los datos mediante un diagrama Q-Q.

   Indicación
   1) Seleccionar el menú `Gráficas→Gráfica de comparación de cuantiles`.
   2) En el cuadro de diálogo que aparece seleccionar la variable obs, activar la opción correspondiente a la distribución normal y hacer click en el botón Aceptar.

   Obsérvese que ahora los puntos del diagrama no están alineados sobre la linea recta de la diagonal ya que los datos no provienen de una población normal.

   c) Comprobar la normalidad de los datos mediante un contraste de Shapiro-Wilk.

   Indicación
   1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de normalidad de Shapiro-Wilk`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable obs en el campo `Variable` y hacer click en el botón Aceptar.

3. Las notas obtenidas en un examen en dos grupos de alumnos que han seguido metodologías de estudio distintas han sido:

Metodología A:	5,8	3,2	8,0	7,3	7,1	2,1	5,0	4,4	4,2	6,7
Metodología B:	8,1	5,4	7,2	7,5	6,3	8,2	6,0	7,8		

a) Crear un conjunto de datos con las variables Nota y Metodología.

b) Comprobar la hipótesis de normalidad de los datos en cada grupo.

Indicación

1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de normalidad de Shapiro-Wilk`.
2) En el cuadro de diálogo que aparece, seleccionar la variable Nota en el campo `Variable` y hacer click en el botón Grupos según.
3) En el cuadro de diálogo que aparece, seleccionar la variable Metodología y hacer click en el botón Aceptar.
4) En el cuadro de diálogo que quedaba abierto, hacer click en el botón Aceptar

c) Contrastar la hipótesis de homocedasticidad (igualdad de varianzas) entre las leches de las dos granjas.

Indicación

1) Seleccionar el menú `Estadísticos→Varianzas→Test de Levene`
2) En el cuadro de diálogo que aparece, seleccionar la variable Nota en el campo `Variable explicada`, la variable Metodología en el campo `Grupos` y hacer click en el botón Aceptar.

d) Utilizando el contraste más adecuado, ¿se puede concluir que existen diferencias en la nota media según cada metodología?

Indicación

Aunque según el análisis anterior, no podemos rechazar la hipótesis de normalidad en los grupos que queremos comparar y tampoco la igualdad de varianzas en los datos de los dos grupos, como la muestra es muy pequeña e incluso uno de los grupos tiene menos de 10 observaciones, la más correcto sería aplicar el contraste de la U de Mann Whitney.

1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de Wilcoxon para dos muestras`.
2) En el cuadro de diálogo que aparece seleccionar la variable Nota en el campo `Variable explicada`, la variable Metodología en el campo `Grupos` y hacer click en el botón Aceptar.

4. Para ver si una campaña de publicidad sobre un fármaco ha influido en sus ventas, se tomó una muestra de 8 farmacias y se midió el número de unidades de dicho fármaco vendidas durante un mes, antes y después de la campaña, obteniéndose los siguientes resultados:

Antes	147	163	121	205	132	190	176	147
Después	150	171	132	208	141	184	182	149

a) Crear un conjunto de datos con las variables Antes y Despues.

b) Comprobar la hipótesis de normalidad de la variable diferencia.

Indicación

1) Seleccionar el menú `Datos→Modificar variable del conjunto de datos activo→Calcular una nueva variable`.
2) En el cuadro de diálogo que aparece, introducir Diferencia en el campo `Nombre de la nueva variable`, introducir la expresión `Antes-Despues` en el campo `Expresión a calcular` y hacer click en el botón Aceptar.
3) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de normalidad de Shapiro-Wilk`.
4) En el cuadro de diálogo que aparece, seleccionar la variable Diferencia en el campo `Variable` y hacer click en el botón Aceptar.

c) Utilizando el contraste más adecuado, ¿se puede concluir que la campaña de publicidad ha aumentado las ventas?

Indicación

Aunque, según el análisis anterior, no podemos rechazar la hipótesis de normalidad en la variable diferencia, como la muestra es muy pequeña, lo más correcto sería aplicar el contraste de Wilcoxon.

1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de Wilcoxon para muestras pareadas`.
2) En el cuadro de diálogo que aparece seleccionar la variable Antes en el campo `Primera variable`, la variable Despues en el campo `Segunda variable`, seleccionar la opción `Diferencia <0` en el campo `Hipótesis alternativa` y hacer click en el botón Aceptar.

5. Se quiere contrastar la eficacia de tres fármacos para reducir la tensión arterial. Para ello se ha medido la variación en la presión arterial sistólica (antes del tratamiento con el fármaco menos después del tratamiento, en mm Hg) en quince pacientes, que se dividieron en tres grupos, aplicando a cada grupo un fármaco diferente. Los resultados obtenidos tras un año de tratamiento fueron:

Fármaco A	Fármaco B	Fármaco C
12	-3	1
15	5	5
16	-8	19
6	-2	45
8	4	3

   a) Crear un conjunto de datos con las variables Farmaco y Variacion_presion.

   b) Dibujar el diagrama de puntos de la variación de la presión sistólica según el fármaco recibido. A la vista del diagrama, ¿crees que los datos presentas homogeneidad de varianzas? ¿crees que hay algún grupo con un cambio en la presión arterial diferente del resto?

   Indicación
   1) Seleccionar el menú `Gráficas→Diagrama de puntos`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable Farmaco en el campo `Factores`, seleccionar la variable Variacion_presion en el campo `Variable explicada` y hacer click sobre el botón Aceptar.

   c) Analizar la normalidad de los datos de los tres grupos.

   Indicación
   1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de normalidad de Shapiro-Wilk`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable Variacion_presion en el campo `Variable`, y hacer click en el botón Grupos según.
   3) En el cuadro de diálogo que aparece, seleccionar la variable Farmaco y hacer click en el botón Aceptar.
   4) En el cuadro de diálogo que quedaba abierto, hacer click en el botón Aceptar

   d) Analizar la homogeneidad de varianzas de los tres grupos.

   Indicación
   1) Seleccionar el menú `Estadísticos→Varianzas→Test de Levene`
   2) En el cuadro de diálogo que aparece, seleccionar la variable Variacion_presion en el campo `Variable explicada`, la variable Farmaco en el campo `Grupos`, activar la opción `media` y hacer click en el botón Aceptar.

   e) Utilizando el contraste más adecuado, ¿se puede concluir que existen diferencias en los cambios de la presión sistólica en función del fármaco recibido?

   Indicación
   En este caso no se cumple la homogeneidad de varianzas, por lo que no se puede aplicar una ANOVA y tendremos que recurrir al test de Kruskal-Wallis.
   1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de Kruskal-Wallis`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable Variacion_presion en el campo `Variable explicada`, la variable Farmaco en el campo `Grupos` y hacer click en el botón Aceptar.

   f) ¿Entre qué grupos se dan las diferencias significativas?

   Indicación
   Repetir los pasos del apartado anterior pero activando la opción `Comparación por pares`.

6. Se quiere contrastar la dificultad de cuatro modelos de examen que se van a poner en la convocatoria ordinaria de la asignatura de bioestadística. Para ello se pide a cinco profesores diferentes que valoren cada uno de los modelos de 0 a 10, y los resultados fueron:

	Modelo1	Modelo2	Modelo3	Modelo4
Profesor 1	6	8	5	8
Profesor 2	5	4	7	9
Profesor 3	5	4	5	6
Profesor 4	7	4	6	7
Profesor 5	6	3	7	8

a) Crear las variables Modelo1, Modelo2, Modelo3 y Modelo4 e introducir los datos de la muestra.

b) ¿Podemos afirmar que el grado de dificultad de los modelos es diferente?

Indicación
Debe utilizarse el test de Friedman ya que las puntuaciones en los modelos son medidas repetidas en los profesores.

1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de suma de rangos de Friedman`.
2) En el cuadro de diálogo que aparece seleccionar las variables Modelo1, Modelo2, Modelo3 y Modelo4 al campo `Variables de medidas repetidas` y hacer click sobre el botón `Aceptar`.

7. El test de Apgar es un examen clínico de neonatología en donde el médico realiza una prueba medida en 3 estándares sobre el recién nacido para obtener una primera valoración simple (macroscópica) y clínica sobre el estado general del neonato después del parto. El recién nacido es evaluado de acuerdo a cinco parámetros fisioanatómicos simples, que son: color de la piel, frecuencia cardíaca, reflejos, tono muscular y respiración. A cada parámetro se le asigna una puntuación entre 0 y 2, y sumando las cinco puntuaciones se obtiene el resultado del test. El test se realiza al minuto, a los cinco minutos y, en ocasiones, a los diez minutos de nacer. La puntuación al primer minuto evalúa el nivel de tolerancia del recién nacido al proceso del nacimiento y su posible sufrimiento, mientras que la puntuación obtenida a los 5 minutos evalúa el nivel de adaptabilidad del recién nacido al medio ambiente y su capacidad de recuperación.

   En la siguiente tabla se refleja la puntuación obtenida por 22 recién nacidos en el test de Apgar al minuto y a los cinco minutos de haber nacido:

Apgar 1	10	3	8	9	8	9	8	8	8	8	7	8	6	8	9	9	9	9	8	9	3	9
Apgar 5	10	6	9	10	9	10	9	9	9	9	9	9	6	8	9	9	9	9	8	9	3	9

   Con los datos anteriores se pretende realizar un contraste de hipótesis para analizar si existe o no relación entre las dos puntuaciones. Para ello, se pide:

   a) Crear un conjunto de datos con las variables Apgar1 y Apgar5.

   b) Comprobar si las variables siguen distribuciones normales.

   Indicación
   1) Seleccionar el menú `Estadísticos→Test no paramétricos→Test de normalidad de Shapiro-Wilk`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable Apgar1 en el campo `Variable` y hacer click en el botón Aceptar.
   3) Repetir lo mismo para la variable Apgar5.

   c) Realizar el contraste de correlación bilateral.

   Indicación
   Como las variables analizadas no siguen distribuciones normales, para realizar un contraste de relación entre ambas hay que obtener el coeficiente de correlación de Spearman y ver si es o no significativamente diferente de 0.
   1) Seleccionar el menú `Estadísticos→Resúmenes→Test de correlación`.
   2) En el cuadro de diálogo que aparece, seleccionar las variables Apgar1 y Apgar5 en el campo `Variables`, seleccionar la opción `Coeficiente de Spearman` en el campo `Tipo de correlación` y hacer click en el botón Aceptar.

3 Ejercicios propuestos

1. Se ha realizado un estudio para investigar el efecto del ejercicio físico en el nivel de colesterol en la sangre. En el estudio participaron once personas, a las que se les midió el nivel de colesterol antes y después de desarrollar un programa de ejercicios. Los resultados obtenidos fueron los siguientes:

Nivel Previo, 223, 212, 221, 210, 225, 202, 198, 200, 185, 220
Nivel Posterior, 226, 211, 222, 212, 225, 201, 196, 217, 130, 220

Utilizando el contraste más adecuado, ¿se puede concluir que el ejercicio físico disminuye el nivel de colesterol?

2. Dos químicos A y B realizan respectivamente 14 y 16 determinaciones de la actividad radiactiva de una muestra de material. Sus resultados en Curios:

A		B	
263.10	262.60	286.53	254.54
262.10	259.60	284.55	286.30
257.60	262.20	272.52	282.90
261.70	261.20	283.85	253.75
260.70	259.20	252.01	245.26
269.13	268.63	275.08	266.08
268.13	217.00	267.53	252.05
		253.82	269.81

Utilizando el contraste más adecuado, ¿se puede concluir que existen diferencias significativas en la actividad detectada por cada químico?

3. En un hospital se están evaluando dos tratamientos diferentes para ver si existen diferencias entre ellos, para lo cual se seleccionaron dos grupos de 32 pacientes cada uno y se aplicó un tratamiento a cada grupo. Los resultados fueron:

	Empeoraron	Igual	Mejoraron	Curaron
A	9	12	5	6
B	5	6	11	10

Utilizando el contraste más adecuado, ¿se puede concluir que existen diferencias significativas entre ambos tratamientos?

4. Queremos comparar las notas iniciales de un grupo de 20 alumnos, con las obtenidas al final del curso, pera ver si existen diferencias, las notas fueron (SS suspenso, A aprobado, N notable y SB sobresaliente):

Alumno	1	2	3	4	5	6	7	8	9	10
Nota Inicial	SS	A	A	SS	N	SS	SS	SB	A	A
Nota Final	A	A	SS	A	SB	N	A	N	SS	A

Alumno	11	12	13	14	15	16	17	18	19	10
Nota Inicial	SS	N	A	SS	SB	A	N	SS	A	SB
Nota Final	SB	A	N	A	SB	A	N	A	N	SB

Utilizando el contraste más adecuado, ¿se puede concluir que existen diferencias significativas entre las notas al comienzo y al final del curso?

5. Disponemos de la evaluación que han obtenido tres grupos de prácticas de las asignatura de bioestadística (MM muy mal, M mal, R regular, B bien y MB muy bien):

Grupo 01	R	B	R	M	MM	B	MB	R	M	B	M	R	R	MM	M
Grupo 02	B	R	M	B	R	M	B	MB	M	R	M	R			
Grupo 03	MB	B	M	R	B	MB	B	R	B	MB	B	R	MB		

Utilizando el contraste más adecuado, ¿se puede concluir que existen diferencias significativas en la evaluación de los diferentes grupos?

6. Para comparar las dificultades presentados por un grupo de problemas de lógica, se han seleccionado aleatoriamente a ocho individuos a los que se les ha planteado tres pruebas iguales, a cada uno y se han anotado los tiempos, en minutos, que han tardado en resolverlos. Los resultados obtenidos son:

Prueba 1	Prueba 2	Prueba 3
38	6	35
22	4	9
14	8	8
8	2	4
6	4	8
10	14	10
14	2	5
8	6	3

   Utilizando el contraste más adecuado, ¿se puede concluir que existen diferencias significativas en los tiempos de resolución de las tres pruebas?

7. La siguiente tabla muestra los datos de 9 pacientes con anemia aplástica:

Reticulocitos (%)	3,6	2,0	0,3	0,3	0,2	3,0	0,0	1,0	2,2
Linfocitos (por mm^2)	1700	3078	1820	2706	2086	2299	676	2088	2013

   Mediante el adecuado contraste de hipótesis basado en el coeficiente de correlación de Spearman, ¿existe relación entre ambas variables?

Contrastes Basados en el Estadístico χ^2. Comparación de Proporciones

1 Fundamentos teóricos

Existen multitud de situaciones en el ámbito de la salud, o en cualquier otro ámbito, en las que el investigador está interesado en determinar posibles relaciones entre variables cualitativas. Un ejemplo podría ser el estudio de si existe relación entre las complicaciones tras una intervención quirúrgica y el sexo del paciente, o bien el hospital en el que se lleva a cabo la intervención. En este caso, todas las técnicas de inferencia vistas hasta ahora para variables cuantitativas no son aplicables, y para ello utilizaremos un contraste de hipótesis basado en el estadístico χ^2 (Chi-cuadrado).

Sin embargo, aunque éste sea su aspecto más conocido, el uso del test no se limita al estudio de la posible relación entre variables cualitativas, y también se aplica para comprobar el ajuste de la distribución muestral de una variable, ya sea cualitativa o cuantitativa, a su hipotético modelo teórico de distribución.

En general, este tipo de tests consiste en tomar una muestra y observar si hay diferencia significativa entre las *frecuencias observadas* y las especificadas por la ley teórica del modelo que se contrasta, también denominadas *frecuencias esperadas*.

Podríamos decir que existen dos grandes bloques de aplicaciones básicas en el uso del test de la χ^2:

1. **Test de ajuste de distribuciones**. Es un contraste de significación para saber si los datos de la población, de la cual hemos extraído una muestra, son conforme a una ley de distribución teórica que sospechamos que es la correcta.

   Por ejemplo: disponemos de 400 datos que, a priori, siguen una distribución de probabilidad uniforme, pero ¿es estadísticamente cierto que se ajusten a dicho tipo de distribución?

2. **Test para tablas de contingencia.** En las que se parte de la tabla de frecuencias bidimensional para las distintas modalidades de las variables cualitativas. Aunque muy a menudo el test de la χ^2 aplicado en tablas de contingencia se denomina prueba de independencia, en realidad se aplica en dos diseños experimentales diferentes, que hacen que se clasifique en dos bloques diferentes:

   a) **Prueba de independencia**. Mediante la que el investigador pretende estudiar la relación entre dos variables cualitativas en una población.

   Por ejemplo: tenemos una muestra de 200 enfermos (el investigador tan sólo controla el total en una muestra) operados de apendicitis en 4 hospitales diferentes y queremos ver si hay relación entre la posible infección postoperatoria y el hospital en el que el paciente ha sido operado.

   b) **Prueba de homogeneidad**. Mediante la que el investigador pretende ver si la proporción de una determinada característica es la misma en poblaciones, tal vez, diferentes.

   Por ejemplo: tenemos dos muestras diferentes, una de ellas de 100 individuos VIH positivos, y otra de 600 VIH negativos (el investigador controla el total en ambas muestras), y queremos analizar si la proporción de individuos con problemas gastrointestinales es la misma en ambas.

Por último, aunque el test de la Chi-cuadrado es muy importante en el análisis de las relaciones entre variables cualitativas, su aplicación puede conducir a errores en determinadas situaciones; sobre todo cuando los tamaños muestrales son pequeños, lo cual conduce a que en algunas categorías apenas

tengamos individuos y ello invalida los supuestos de aplicación del test; y también cuando tenemos variables cualitativas con valores sí o no analizadas en los mismos individuos pero en diferentes tiempos, es decir, mediante datos pareados. Para el primer caso, cuando el número de individuos en alguna categoría es muy pequeño, se utiliza el test Exacto de Fisher, mientras que en el segundo, con datos pareados, se utiliza el test de McNemar.

1.1 Contraste χ^2 de Pearson para ajuste de distribuciones

Es el contrate de ajuste más antiguo y es válido para todo tipo de distribuciones. Para analizar una muestra de una variable agrupada en categorías (aunque sea cuantitativa), evaluando una hipótesis previa sobre probabilidad de cada modalidad o categoría, se realiza un contraste de hipótesis Chi-cuadrado de bondad de ajuste.

El contraste se basa en hacer un recuento de los datos y comparar las frecuencias observadas de cada una de las modalidades con las frecuencias esperadas por el modelo teórico que se contrasta. De este modo, se calcula es estadístico:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

donde O_i son las frecuencias observadas en la muestra en la modalidad i, y E_i son las frecuencias esperadas para la misma modalidad según el modelo teórico. Las frecuencias esperadas se calculan multiplicando el tamaño de la muestra por la probabilidad de la correspondiente modalidad según el modelo teórico, es decir $E_i = np_i$, siendo p_i la probabilidad de la modalidad i.

Si la población de la que se ha obtenido la muestra sigue el modelo de distribución teórica, el estadístico anterior se distribuye como χ^2 con $k-1$ grados de libertad, donde k es el número de modalidades de la variable. Un valor del estadístico χ^2 grande indica que las distribuciones de las frecuencias observadas y esperadas son bastantes diferentes, mientras que un valor pequeño del estadístico indica que hay poca diferencia entre ellas.

La prueba χ^2 de bondad del ajuste es válida si todas las frecuencias esperadas son mayores o iguales que 1 y no más de un 20 % de ellas tienen frecuencias esperadas menores que 5. Si no se cumple lo anterior, entonces las categorías implicadas deben combinarse con categorías adyacentes para garantizar que todas cumplen la condición. Si las categorías corresponden a variables cuantitativas categorizadas, no tienen necesariamente que corresponder a la misma amplitud de variable.

1.2 Contraste χ^2 en tablas de contingencia

Como ya hemos visto, el contraste de la χ^2 en tablas de contingencia sirve para establecer relaciones entre variables cualitativas (o cuantitativas categorizadas), entre las que no puede realizarse un análisis de regresión y correlación, y tanto para determinar independencia entre variables, como homogeneidad entre poblaciones (igual proporción de una determinada característica). Para ello, describimos el proceso metodológico en el caso de independencia entre variables, que en la práctica, y aunque conceptualmente son casos diferentes, es el mismo también para la homogeneidad entre poblaciones.

Por tablas de contingencia se entiende aquellas tablas de doble entrada donde se realiza una clasificación de la muestra de acuerdo a un doble criterio de clasificación. Por ejemplo, la clasificación de unos individuos de acuerdo a su sexo y su grupo sanguíneo crearía una tabla donde cada celda de la tabla representaría la frecuencia bivariante de las características correspondientes a su fila y columna (por ejemplo mujeres de grupo sanguíneo A). Si se toma una muestra aleatoria de tamaño n en la que se miden ambas variables y se representan las frecuencias de los pares observados en una tabla bidimensional, tenemos:

X/Y	y_j	
x_i	n_{ij}	n_i
	n_j	n

Donde n_{ij} es la frecuencia absoluta del par (x_i, y_j), n_i es la frecuencia marginal de la modalidad x_i y n_j es la frecuencia marginal de la modalidad y_j. Dichas frecuencias aparecen en los márgenes de la tabla de contingencia sumando las frecuencias por filas y columnas, y por ello se conocen como frecuencias marginales.

Siguiendo un procedimiento parecido al del apartado anterior, se comparan las frecuencias observadas en la muestra (frecuencias reales) con las frecuencias esperadas (frecuencias teóricas). Para ello, calculamos la probabilidad de cada casilla de la tabla teniendo en cuenta que si ambas variables son independientes la probabilidad de cada celda surge como un producto de probabilidades (probabilidad de la intersección de dos sucesos independientes) $p_{ij} = p_i p_j = \frac{n_i}{n}\frac{n_j}{n}$. De este modo, obtenemos la frecuencia esperada como:

$$E_i = np_{ij} = n\frac{n_i}{n}\frac{n_j}{n} = \frac{n_i n_j}{n},$$

Y con ello se calcula el estadístico de la Chi-cuadrado de Pearson:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

En el caso de que X e Y fuesen independientes, este estadístico presenta una distribución Chi-cuadrado con $(f-1)(c-1)$ grados de libertad, donde f es el número de filas de la tabla de contingencia y c el número de columnas. Un valor del estadístico Chi-cuadrado grande indica que las distribuciones de las frecuencias observadas y esperadas son bastantes diferentes, y por lo tanto falta de independencia; mientras que un valor pequeño del estadístico indica que hay poca diferencia entre ellas, lo cual nos indica que son independientes.

Este test es adecuado si las frecuencias esperadas para cada celda valen como mínimo 1 y no más de un 20 % de ellas tienen frecuencias esperadas menores que 5. En el caso de una tabla 2x2, estas cifras se alcanzan sólo cuando ninguna frecuencia esperada es menor que 5. Si esto no se cumple, puede, entre otras, utilizarse una prueba para pequeñas muestras llamada prueba exacta de Fisher.

1.3 Test Exacto de Fisher

Este test se puede utilizar cuando no se cumplan las condiciones necesarias para aplicar el test de la Chi cuadrado (más de un 20 % de las frecuencias esperadas para cada celda son menores que 5). Aunque, dada la gran cantidad de cálculos necesarios para llegar al resultado final del test, los programas de Estadística sólo lo calculaban para tablas de contingencia 2x2.

El test Exacto de Fisher está basado en la distribución exacta de los datos y no en aproximaciones asintóticas, y presupone que los marginales de la tabla de contingencia están fijos. El procedimiento para su cálculo consiste en evaluar la probabilidad asociada, bajo el supuesto de independencia, a todas las tablas que se pueden formar con los mismos totales marginales que los datos observados y variando las frecuencias de cada casilla para contemplar todas las situaciones en las que hay un desequilibrio de proporciones tan grande o más que en la tabla analizada. Para el cálculo de la probabilidad asociada a cada tabla se utiliza la función de probabilidad de una variable discreta hipergeométrica.

Aunque generalmente el test Exacto de Fisher es más conservador que la Chi cuadrado (resulta más complicado que detecte diferencias estadísticamente significativas entre las proporciones), no obstante tiene la ventaja de que se puede aplicar sin ninguna restricción en las frecuencias de las casillas de la tabla de contingencia.

1.4 Test de McNemar para datos emparejados

Hasta ahora hemos supuesto que las muestras a comparar eran independientes, es decir dos grupos diferentes en los que se había mirado una determinada característica. Por lo tanto, hemos realizado comparaciones de proporciones de individuos que presentan una determinada característica en dos grupos distintos, pero también nos podemos plantear comparar la proporción de individuos que presentan esa característica en un mismo grupo de individuos pero analizados en dos momentos diferentes. En este último caso se habla comparación de proporciones en datos emparejados, pareados o apareados.

Por ejemplo, si queremos ver si existen o no diferencias en la mejora de los síntomas de una determinada enfermedad, y para ello aplicamos dos fármacos distintos a un grupo de individuos en dos momentos diferentes en los que hayan contraído la misma enfermedad. En este caso, podría pensarse que resultaría

adecuado aplicar tanto la chi cuadrado como el test exacto de Fisher para determinar si existe diferencias entre ambos fármacos en la proporción de pacientes curados, pero aquí hay una diferencia fundamental con los casos anteriores y es que sólo tenemos un grupo de pacientes y no dos. En este tipo de estudios se reduce considerablemente la variabilidad aleatoria, ya que es un mismo individuo el que se somete a los dos tratamientos, y el que manifieste mejoría en los síntomas no dependerá de otros factores tan importantes como, por ejemplo, la edad, el sexo o el tipo de alimentación, que pueden influir pero que tal vez no se controlen adecuadamente en un diseño de grupos independientes. Al reducir la variabilidad aleatoria mediante datos emparejados, pequeñas diferencias entre las proporciones pueden llegar a ser significativas, incluso con tamaños muestrales pequeños, lo cual se traduce en que este tipo de diseños del experimento resultan más eficientes a la hora de obtener resultados estadísticamente significativos.

No obstante, nuevos diseños implican nuevas formas de tratar los datos, y el procedimiento más adecuado es el que se utiliza en el test de McNemar para datos emparejados. Para su aplicación en nuestro ejemplo, se debería construir una tabla con 4 casillas en las que se contabilicen: las personas que han obtenido una mejoría de los síntomas con los dos fármacos, los que han obtenido con el primero y no con el segundo, los que han obtenido con el segundo y no con el primero y los que no han obtenido mejoría con ninguno.

Mejoría con 1º \ Mejoría con 2º	Sí	No	Totales
Sí	a	b	$a+b$
No	c	d	$c+d$
Totales	$a+c$	$b+d$	$n=a+b+c+d$

Con ello, la proporción muestral de pacientes que han experimentado mejoría con el medicamento 1 vale: $\widehat{p}_1 = (a+b)/n$, e igualmente con el 2: $\widehat{p}_2 = (a+c)/n$, y podemos plantear el contraste cuya hipótesis nula es que no hay diferencia de proporciones poblacionales entre ambos medicamentos: $H_0 : p_1 = p_2$, que puede realizarse sin más que tener en cuenta el oportuno intervalo de confianza para la diferencia de proporciones, o también que, en el supuesto de igualdad de proporciones:

- $z = \dfrac{b-c}{\sqrt{b+c}}$, es un estadístico que sigue una distribución normal tipificada.
- $\chi^2 = \dfrac{(b-c)^2}{b+c}$, es un estadístico que sigue una distribución Chi-cuadrado con un grado de libertad.

Con cualquiera de ellos, se podría calcular el p-valor del contraste.

2 Ejercicios resueltos

1. Dadas dos parejas de genes Aa y Bb, la descendencia del cruce efectuado según las leyes de Mendel, debe estar compuesto del siguiente modo:

Fenotipo	Frecuencias Relativas
AB	$9/16 = 0{,}5625$
Ab	$3/16 = 0{,}1875$
aB	$3/16 = 0{,}1875$
ab	$1/16 = 0{,}0625$

Elegidos 300 individuos al azar de cierta población, se observa la siguiente distribución de frecuencias:

Fenotipo	Frecuencias Observadas
AB	165
Ab	47
aB	67
ab	21

Se pide

a) Crear un conjunto de datos con las variables probabilidad_teorica y frecuencia_observada.

b) Comprobar si esta muestra cumple las leyes de Mendel.

Indicación

1) Seleccionar el menú `Analizar`→`Pruebas no paramétricas`→`Test de bondad de ajuste Chi-cuadrado`.
2) En el cuadro de diálogo que aparece seleccionar la variable probabilidad_teorica en el campo `Probabilidad teórica`, seleccionar la variable frecuencia_observada en el campo `Frecuencia observada` y hacer click en el botón Aceptar.

c) A la vista de los resultados del contraste, ¿se puede aceptar que se cumplen las leyes de Mendel en los individuos de dicha población?

2. En un estudio sobre úlceras pépticas se determinó el grupo sanguíneo de 1655 pacientes ulcerosos y 10000 controles, los datos fueron:

	O	A	B	AB
Paciente	911	579	124	41
Controles	4578	4219	890	313

a) Construir la tabla de contingencia y realizar el contraste Chi-cuadrado.

Indicación

1) Seleccionar el menú `Estadísticos`→`Tablas de contingencia`→`Introducir y analizar una tabla de doble entrada`.
2) En el cuadro de diálogo que aparece, seleccionar 2 filas, 4 columnas, introducir las frecuencias de la muestra en tabla de frecuencias, activar las opciones `Porcentajes totales`, `Test de independencia Chi-cuadrado` y `Componentes del estadístico Chi-cuadrado`, `Imprimir las frecuencias esperadas` y hacer click en el botón Aceptar.

b) A la vista de los resultados del contraste, ¿existe alguna relación entre el grupo sanguíneo y la úlcera péptica?, es decir, ¿se puede concluir que la proporción de pacientes y de controles es diferente dependiendo del grupo sanguíneo?

3. Mitchell et al. (1976, Annals of Human Biology), partiendo de una muestra de 478 individuos, estudiaron la distribución de los grupos sanguíneos en varias regiones del sur-oeste de Escocia, obteniendo

los resultados que se muestran:

	Eskdale	Annandale	Nithsdale	
A	33	54	98	185
B	6	14	35	55
O	56	52	115	223
AB	5	5	5	15
	100	125	253	478

a) Construir la tabla de contingencia y realizar el contraste Chi-cuadrado.

Indicación

1) Seleccionar el menú `Estadísticos→Tablas de contingencia→Introducir y analizar una tabla de doble entrada`.
2) En el cuadro de diálogo que aparece, seleccionar 4 filas, 3 columnas, introducir las frecuencias de la muestra en tabla de frecuencias, activar las opciones `Porcentajes totales`, `Test de independencia Chi-cuadrado` y `Componentes del estadístico Chi-cuadrado`, `Imprimir las frecuencias esperadas` y hacer click en el botón Aceptar.

b) En vista de los resultados del contraste, ¿se distribuyen los grupos sanguíneos de igual manera en las diferentes regiones?

4. En un estudio para saber si el habito de fumar está relacionado con el sexo, se ha preguntado a 26 personas. De los 9 hombres consultados 2 respondieron que fumaban, mientras que de las 17 mujeres consultadas, 6 fumaban. ¿Podemos afirmar que existe relación entre ambas variables?

a) Crear un conjunto de datos con las variables sexo y fuma.

b) Construir la tabla de contingencia y realizar el contraste Chi-cuadrado.

Indicación

1) Seleccionar el menú `Estadísticos→Tablas de contingencia→Tabla de doble entrada`.
2) En el cuadro de diálogo que aparece, seleccionar la variable `sexo` en el campo `Variable de fila`, la variable `fuma` en el campo `Variable de columna`, activar las opciones `Porcentajes totales`, `Test de independencia Chi-cuadrado` y `Componentes del estadístico Chi-cuadrado`, `Imprimir las frecuencias esperadas`, `Test exacto de Fisher` y hacer click en el botón Aceptar.

c) En vista de los resultados del contraste, ¿se distribuyen los fumadores de igual manera en ambos sexos?

Indicación

En este caso el procedimiento a seguir es igual que para la Chi cuadrado, pero vemos que ahora no se cumplen las condiciones para poder aplicar esta prueba, ya que el número de hombres fumadores es menor que 5, y por eso nos tendremos que fijar en el p-valor del test exacto de Fisher, que si podemos aplicar, teniendo en cuenta si estamos realizando un contraste bilateral o unilateral.

5. Para probar la eficacia de dos fármacos diferentes contra las migrañas, se seleccionaron a 20 personas que padecían migrañas habitualmente, y se les dió a tomar a cada uno los fármacos en momentos diferentes. Luego se les preguntó si habían obtenido mejoría o no con el fármaco tomado. Los resultados fueron los siguientes:

	1	2	3	4	5	6	7	8	9	10
Fármaco 1	Sí	Sí	Sí	Sí	Sí	No	Sí	No	Sí	Sí
Fármaco 2	No	No	Sí	No	Sí	Sí	No	No	No	No

	11	12	13	14	15	16	17	18	19	20
Fármaco 1	Sí	No	Sí	No	Sí	Sí	Sí	No	Sí	Sí
Fármaco 2	Sí	No	Sí	No	No	Sí	No	Sí	No	No

a) Crear un conjunto de datos con las variables Mejora_Farmaco1, y Mejora_Farmaco2.

b) Construir la tabla de contingencia y realizar el contraste Chi-cuadrado.

Indicación

1) Seleccionar el menú `Estadísticos`→`Tablas de contingencia`→`Tabla de doble entrada`.
2) En el cuadro de diálogo que aparece, seleccionar la variable Mejora_Farmaco1 en el campo `Variable de fila`, la variable Mejora_Farmaco2 en el campo `Variable de columna`, activar las opciones `Porcentajes totales`, `Test de McNemar` y hacer click en el botón Aceptar.

c) En vista de los resultados del contraste, ¿podemos afirmar que existen diferencias significativas entre los dos fármacos?

3 Ejercicios propuestos

1. Supongamos que queremos comprobar si un dado está bien equilibrado o no. Lo lanzamos 1200 veces, y obtenemos los siguientes resultados:

Número	Frecuencias de aparición
1	120
2	275
3	95
4	310
5	85
6	315

   a) A la vista de los resultados, ¿se puede aceptar que el dado está bien equilibrado?
   b) Nos dicen que, en este dado, los números pares aparecen con una frecuencias 3 veces superior a la de los impares. Contrastar dicha hipótesis.

2. Se realiza un estudio en una población de pacientes críticos hipotéticos y se observan, entre otras, dos variables, la evolución (si sobreviven SV o no NV) y la presencia o ausencia de coma, al ingreso. Se obtienen los siguientes resultados:

	No coma	Coma	
SV	484	37	521
NV	118	89	207
	602	126	728

   Nos preguntamos: ¿es el coma al ingreso un factor de riesgo para la mortalidad?

3. La recuperación producida por dos tratamientos distintos A y B, se clasifican en tres categorías: muy buena, buena y mala. Se administra el tratamiento A a 32 pacientes y el B a otros 28. De las 22 recuperaciones muy buenas, 10 corresponden al tratamiento A; de las 24 recuperaciones buenas, 14 corresponden al tratamiento A y de las 14 que tienen una mala recuperación, 8 corresponden al tratamiento A. ¿Son igualmente efectivos ambos tratamientos para la recuperación de los pacientes?

4. Para contrastar la hipótesis de que las mujeres tienen más éxito en sus estudios que los hombres, se ha tomado una muestra de 10 chicos y otra de 10 chicas que han sido examinados por un profesor que aprueba siempre al 40 % de los alumnos presentados a examen. Teniendo en cuenta que sólo aprobaron 2 chicos, utiliza el test de hipótesis más adecuado para decidir si la citada hipótesis es cierta.

5. Se ha preguntado a los 150 alumnos de un curso, si estaban de acuerdo o no, con la metodología de enseñanza de dos profesores distintos que les han dado clase en la asignatura de bioestadística. Los resultados se recogen en la siguiente tabla:

Profesor 1 \ Profesor 2	Opinión favorable	Opinión desfavorable
Opinión favorable	37	48
Opinión desfavorable	44	21

   ¿Podemos afirmar que existe diferente opinión por parte de los alumnos, sobre los dos profesores?

Análisis de Concordancia

1 Fundamentos teóricos

1.1 Introducción

La medición es un proceso que aparece siempre tanto en la práctica como en la investigación clínica. Hay variables que son sencillas de medir, como puede ser el peso, pero hay otras que conllevan cierto grado de subjetividad, como la intensidad del dolor, que hacen especialmente difícil su medición. En cualquier caso, el proceso de medición lleva asociado algún grado de error. Existen factores asociados a los individuos, al observador o al instrumento de medida que pueden influir en los resultados de las mediciones. Se denomina validez de una medición a la capacidad de poder medir lo que realmente se quiere medir mientras que se denomina fiabilidad o reproducibilidad de una medición a la capacidad de poder obtener un mismo valor cuando la medición se realiza sobre la misma muestra en más de una ocasión en condiciones similares. Los términos concordancia y acuerdo son sinónimos de reproducibilidad. En los estudios que tratan de evaluar la validez de una medida se comparan sus resultados con los obtenidos mediante una prueba de referencia (gold standard) que se sabe válida y fiable. Cuando se trata de estudiar la fiabilidad de una medición, se repite el proceso de medida para evaluar la concordancia obtenida entre las diferentes mediciones. En un estudio de la fiabilidad pueden valorarse los siguientes aspectos:

Concordancia intraobservador: tiene por objetivo evaluar el grado de coincidencia de las mediciones efectuadas por un mismo observador en las mismas condiciones.

Concordancia interobservador: tiene por objetivo evaluar el grado de coincidencia de las mediciones efectuadas por dos observadores en un mismo individuo.

Concordancia entre métodos de medición: tiene por objetivo evaluar el grado de coincidencia de las mediciones efectuadas con diferentes métodos de medida.

La concordancia entre mediciones es de gran interés en la práctica clínica. Las técnicas de análisis de la concordancia dependen del tipo de variable a estudiar. En el caso de variables cuantitativas se utiliza habitualmente el coeficiente de correlación intraclase, mientras que en el caso de variables cualitativas el estadístico más utilizado es el índice kappa.

1.2 Análisis de la Concordancia entre dos Variables Cuantitativas

En la investigación clínica resulta muy frecuente la evaluación de la fiabilidad, o concordancia, de las medidas realizadas, pudiéndose distinguir entre dos tipos de situaciones diferentes:

- Aquellas en las que se determina la concordancia en los resultados cuando se repite la medición con el mismo instrumento en condiciones idénticas.
- Aquellas en las que se determina hasta qué punto los resultados obtenidos con diferentes instrumentos de medida, o con diferentes observadores, concuerdan.

Por ejemplo, podríamos plantearnos el acuerdo entre las medidas de la presión arterial sistólica final, tomadas en los mismos pacientes y en idénticas condiciones (también por el mismo observador), excepto

que una de las medidas se ha hecho con el tensiómetro habitual y la otra con un tensiómetro electrónico de muñeca.

Para analizar la concordancia entre este tipo de variables numéricas, muy a menudo se ha utilizado r, el coeficiente de correlación de Pearson, pero no resulta un índice adecuado ya que dos instrumentos pueden medir sistemáticamente cantidades diferentes uno del otro, y sin embargo la correlación podría incluso ser perfecta. Por ejemplo, supongamos que la medida del tensiómetro de muñeca es sistemáticamente $10mmHg$ inferior a la obtenida con el tensiómetro habitual. De esta forma, si llamamos y a la medida del tensiómetro de muñeca y x a la del habitual, ambas expresadas en $mmHg$, concluiremos que $y = x - 10$ con una correlación lineal perfecta ($r = 1$); sin embargo, evidentemente, la concordancia entre ambos métodos de medida deja bastante que desear. En definitiva, el coeficiente de correlación lineal mide la intensidad de la asociación lineal pero no proporciona información sobre el nivel de acuerdo entre las medidas.

El índice estadístico que permite cuantificar el acuerdo entre variables numéricas es el *Coeficiente de Correlación Intraclase (CCI)*, cuyo cálculo se basa en un Análisis de la Varianza (ANOVA) en el que se tiene en cuenta que la variabilidad total de los datos puede dividirse en tres componentes:

- La variabilidad debida a las diferencias entre los distintos pacientes: P.
- La debida a las diferencias entre observadores o métodos de observación: O.
- La residual aleatoria inherente a toda medición: R.

A partir de ello, se define el CCI como el cociente entre la variabilidad debida a los pacientes y la variabilidad total:

$$CCI = \frac{P}{P + O + R},$$

El valor del CCI siempre se encuentra entre 0 y 1, de tal forma que si la variabilidad debida al observador (o método de observación), junto con la residual, son muy pequeñas, el CCI será muy próximo a 1, y la concordancia entre las medidas muy buena. Por el contrario, si la variabilidad entre los pacientes es muy pequeña comparada con la que introduce el observador, el CCI será muy próximo a 0 y la concordancia mala o muy mala.

Generalmente, para delimitar entre qué valores del CCI podemos considerar que la concordancia es muy buena, buena, moderada, mediocre o mala, se utiliza la siguiente tabla:

Valor de CCI	Fuerza de la Concordancia
$CCI \geq 0{,}9$	Muy buena
$0{,}7 \leq CCI < 0{,}9$	Buena
$0{,}5 \leq CCI < 0{,}7$	Moderada
$0{,}3 \leq CCI < 0{,}5$	Mediocre
$CCI < 0{,}3$	Mala o nula

La tabla anterior nos indica el grado de bondad estadístico de la concordancia, pero para establecer si una concordancia es clínicamente buena, la respuesta no la puede dar la tabla, sino la experiencia previa del experimentador, que es quien marca el valor del CCI que considera necesario.

1.3 Análisis de la Concordancia entre dos Variables Cualitativas

La existencia o no de relación entre variables cualitativas se apoya, habitualmente, en el contraste de Chi Cuadrado (o similares); no obstante, como ya ocurría con las variables cuantitativas, la existencia de relación no implica concordancia.

Para entender cómo se cuantifica la concordancia entre variables cualitativas, supongamos que tenemos, por ejemplo, dos médicos diferentes que observan a los mismos 100 pacientes, y los diagnostican como enfermos o sanos, con los resultados que aparecen en la siguiente tabla de contingencia:

Médico 1 \ Médico 2	Enfermos	Sanos	Totales
Enfermos	8	12	20
Sanos	17	63	80
Totales	25	75	100

Evidentemente, la concordancia en el diagnóstico sería tanto mejor cuantos más pacientes se situasen en las casillas en que coinciden los diagnósticos de los dos médicos y menos en las casillas en que los diagnósticos son diferentes. Si como criterio de concordancia simplemente nos fijáramos en la proporción de diagnósticos coincidentes, tendríamos 71 pacientes de un total de 100, es decir un 71 % de acuerdo. No obstante lo anterior, simplemente por azar cabría esperar que, en la casilla correspondiente a un diagnóstico de enfermo por parte de los dos médicos, hubiese una frecuencia esperada de 5 pacientes $(25 * 20/100)$, y en la casilla correspondiente a un diagnóstico de sano por parte de ambos médicos 60 $(75 * 80/100)$. Por lo tanto, simplemente por azar se hubiesen producido 65 coincidencias (un 65 % de acuerdo), con lo cual sólo se han logrado 6 coincidencias más de las que cabría esperar por azar. Dividiendo esas 6 coincidencias entre las 35 que se hubiesen podido dar como máximo sin tener en cuenta el azar, obtenemos un porcentaje del 17,1 %, que, ahora sí que cuantifica la concordancia sin tener en cuenta el azar, y que en nada se parece al 71 % de acuerdo que parecía deducirse de los datos originales.

El ejemplo anterior pone de manifiesto la discrepancia tan grande que puede haber entre la proporción o porcentaje de observaciones concordantes P_o, también llamada concordancia simple, y la proporción o porcentaje de concordancia más allá del azar, que es precisamente lo que cuantifica el índice Kappa de Cohen, κ . Para el cálculo de κ se utiliza la expresión:

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

Donde P_o es la proporción de observaciones concordantes, y P_e es la proporción de observaciones concordantes simplemente por azar, suponiendo independientes las dos variables que cruzamos en la tabla de contingencia.

Sin más que sustituir en la fórmula las proporciones del ejemplo anterior, comprobamos cómo el índice κ tiene un valor de 0,171:

Para interpretar los valores del índice Kappa κ se suele utilizar la siguiente tabla:

Valor de κ	Fuerza de la Concordancia
$\kappa \geq 0{,}8$	Muy buena
$0{,}6 \leq \kappa < 0{,}8$	Buena
$0{,}4 \leq \kappa < 0{,}6$	Moderada
$0{,}2 \leq \kappa < 0{,}4$	Mediocre
$\kappa < 0{,}2$	Mala o nula

Por lo tanto el índice Kappa κ obtenido indica que la concordancia del diagnóstico entre ambos médicos es mala.

2 Ejercicios resueltos

1. Se ha medido la presión arterial sistólica en un grupo de 20 pacientes con el tensiómetro habitual y con un tensiómetro de muñeca, obteniéndose los siguientes resultados:

Tensiómetro habitual	Tensiómetro de muñeca	Tensiómetro habitual	Tensiómetro de muñeca
112	124	133	126
124	116	115	121
96	88	104	98
106	110	86	94
138	144	93	102
155	150	144	132
86	82	125	120
126	118	112	104
114	120	108	108
92	97	98	98

Se pide

a) Crear un conjunto de datos con las varaibles tens_habitual y tens_muñeca.

b) Calcular el coeficiente de correlación intraclase e interpretar el resultado.

Indicación

1) Seleccionar el menú `Estadísticos→Análisis de concordancia→Coeficiente de correlación intraclase`.
2) Seleccionar las variables tens_habitual y tens_muñeca en el campo `Medidas` y hacer click en el botón Aceptar.

El coeficiente de correlación intraclase aparece en la fila `Single raters absolute` y vale 0,93, lo que indica que el nivel de concordancia entre las medidas de la presión arterial obtenidas con el tensiómetro habitual y con el de muñeca es muy buena.

c) Dibujar un gráfico de dispersión en el que aparezca la recta de regresión de la presión arterial medida con el tensiómetro de muñeca en función de la medida con el tensiómetro habitual.

Indicación

1) Seleccionar el menú `Gráficas->Diagrama de dispersión`.
2) En el cuadro de diálogo que aparece seleccionar la variable tens_muñeca como `variable x` y la variable tens_habitual como `variable y`, activar únicamente la opción `Línea de mínimos cuadrados` y hacer click en el botón Aceptar.

d) Añadir al gráfico obtenido en el apartado anterior, la recta que correspondería si ambas tensiómetros diesen la misma medida.

Indicación

Sin cerrar la ventana gráfica del apartado anterior, ejecutar el comando `abline(0,1)` para dibujar la recta $y = x$, que tiene término independiente 0 y pendiente 1.

2. Se entregaron a dos radiólogos A y B un conjunto de radiografías de tórax de pacientes oncológicos, para que informaban si presentaban metástasis en los pulmones o no. Ambos radiólogos analizaron todas las radiografías y cada uno de ellos emitió su informe, indicando en cuáles de ellas se apreciaban metástasis y en cuáles no. Como resultado de dichos informes hubo 32 radiografías en que ambos radiólogos apreciaron metástasis, 68 radiografías en que ninguno las apreció, 6 radiografías en que el radiólogo A las apreció y el B no, y 10 en que el B las apreció y el A no.

a) Construir la tabla de contingencia y calcular el índice Kappa de Cohen e interpretar el resultado.

Indicación

1) Seleccionar el menú `Estadísticos→Tablas de contingencia→Introducir y analizar una tabla de doble entrada`.
2) En el cuadro de diálogo que aparece, seleccionar 2 filas, 2 columnas, introducir las frecuencias de la muestra en tabla de frecuencias, seleccionar la opción `Índice kappa de Cohen` y hacer click en el botón Aceptar.

El valor del índice de Kappa obtenido es 0,7 por lo que la concordancia es buena.

3 Ejercicios propuestos

1. Se ha medido la concentración de ácido úrico en sangre en un grupo de diez pacientes con el equipo tradicional y con un equipo nuevo, obteniéndose los siguientes resultados, expresados en mg./dl.:

Tradicional	Nuevo
5,4	5,8
6,2	6,9
3,7	3,4
7,6	6,4
4,5	4,5
3,8	4,4
5,2	5,8
4,7	5,6
4,9	4,2
5,5	6,8

   Se desea:

   a) Calcular el coeficiente de correlación intraclase e interpretarlo en términos de concordancia de las medidas de la concentración de ácido úrico en sangre obtenidas con ambos equipos.

   b) Dibujar el diagrama de dispersión, en el que aparezca la recta de regresión de las concentraciones de ácido úrico obtenidas con el equipo nuevo sobre las obtenidas con el equipo tradicional, y la recta de regresión que se obtendría si el equipo nuevo siempre diera un resultado $0,8 mg/dl$ superior al que proporciona el equipo tradicional.

2. Se plantean a un grupo de personas dos tests A y B para determinar si su régimen alimenticio es adecuado o no. Como resultado de los test hubo 72 personas cuyo régimen alimenticio evaluado con ambos tests resultó adecuado, 34 personas en que con ninguno de los tests resultó adecuado, 12 personas cuyo régimen era adecuado según el test A pero no según el test B y 10 personas cuyo régimen era adecuado según el test B pero no según el test A. Se pide:

   a) Obtener la tabla de contingencia correspondiente a los resultados obtenidos con ambos tests.

   b) ¿Hay mucha coincidencia entre los resultados obtenidos con ambos tests? Calcular el índice de Kappa y contestar a partir del resultado obtenido.