GUILLERMO AYALA GALLEGO

ESTADÍSTICA BÁSICA

Guillermo Ayala
Guillermo.Ayala@uv.es

Índice general

1 *Datos y R* 5

2 *Estadística descriptiva* 21

3 *Probabilidad* 39

4 *Distribución muestral* 71

5 *Estimación* 79

6 *Contraste de hipótesis* 105

7 *Comparación de dos poblaciones normales* 127

8 *Correlación y regresión* 155

9 *Bibliografía* 187

1
Datos y R

Introducción

Este texto no es más que unas notas de clase con una introducción a la Estadística básica. Los datos que utilizamos son, en lo posible, de aplicaciones medioambientales. Sin embargo, las técnicas estadíticas son esencialmente las mismas y con una presentación similar a la que podemos encontrar en, por ejemplo, un texto de Bioestadística. Por esta razón en la bibliografía se incluyen buenas referencias bibliográficas de Estadística aplicada a datos medioambientales y a otros tipos de datos. No asumimos ningún tipo de conocimiento previo de la Probabilidad. Intentar estudiar Estadística con una orientación (muy fuertemente) aplicada es inútil si no se dispone de una herramienta informática de calidad. En nuestro caso la herramienta es R Core Team [2015a, R]. [1]

La sección 1.1 comenta las referencias básicas de donde sacamos el material de la asignatura. En la sección 1.9 describimos algunos de los ficheros de datos que utilizamos. [2]

[1] A lo largo de estos apuntes se introducen notas bien a pié de página bien al margen. No es necesario leerlas para seguir el curso. Es material que no se considera evaluable en este curso. De hecho, llevan material adicional y no necesario para seguir las notas. En ocasiones son comentarios para hacer más llevadera la lectura de este tostón. Una de las muchas pruebas que Dios nos manda.

[2] Se pueden encontrar en el directorio DATOS de Aula Virtual.

1.1 *Bibliografía comentada*

Estas notas están tomadas (como todo material docente) de otros textos que son la referencia fundamental. No hay nada de mayor importancia en el estudio que la consulta de los libros, de varios libros, de modo que busquemos problemas para intentar resolverlos, que veamos distintas notaciones (similares pero no iguales), que veamos cómo se pueden decir las cosas (otra vez, similares pero no iguales). Que nos acostumbremos a la búsqueda del material. En esta (desgraciada) época lo que no falta es el acceso a la información (y, con más frecuencia, a la desinformación que surge del exceso de material a consultar). Sin duda, lo primero en una asignatura es hablar de los libros de los cuales se nutre y, a los cuales, no se pretende sustituir.

Vamos a comentar algunas referencias bibliográficas generales que se utilizan en esta asignatura. En cada tema indicaremos la referencia a consultar indicando siempre que se pueda las páginas concretas. Es importante utilizar la bibliografía. Estas notas son una especie de guión para estas clases presenciales pero no hay un tratamiento extenso de los conceptos. Para ello está la bibliografía.

Nuestra referencia básica es Wilcox [2009]. Es un texto no específicamente dedicado a Estadística medio ambiental pero es correcto y breve. Un buen libro de texto (muy largo para nuestros objetivos) que incluye lo que vemos aquí y más cosas es [3]. Es un texto maravilloso y, con tiempo, el que utilizaríamos como referencia única. Su mayor inconveniente para esta asignatura es su orientación médica.

[3] Bernard Rosner. *Fundamentals of Biostatistics*. Brooks/Cole Cengage Learning, seven edition, 2010

Todos los tratamientos estadíticos los realizaremos con [R Core Team, 2015a]. En particular los autores del texto [Reimann et al., 2008] han desarrollado y lo vamos a usar sus datos el paquete StatDA.

Otros libros que utilizamos (fundamentalmente para conseguir problemas) en lo que sigue son Reimann et al. [2008], Berthouex and Brown [2002], Manly [2009], Millard and Neerchal [2001], Piegorsch and Bailer [2005]. De todos ellos es especialmente útil [4].

[4] P.M. Berthouex and L.C. Brown. *Environmental Engineers*. Lewis Publishers, second edition, 2002. URL `/home/gag/BIBLIOGRAFIA/MISLIBROS/Berthouex_Brown_Statistics_for_Environmental_Engineers_2nd_Ed_CRC_Press_2002.pdf`

Para un uso básico de R sin un interés especial sobre datos medioambientales se puede consultar Verzani [2005], Dalgaard [2002]. En particular, el texto [Verzani, 2005] es muy adecuado para el manejo del programa aunque sufre la parte de explicación estadítica. Recomendable para aprender R pero no Estadística. En versiones libres se puede encontrar en

- http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf
- http://www.math.csi.cuny.edu/Statistics/R/simpleR/.

Un texto muy recomendable aunque hace un uso más sofisticado de R que los anteriores es Cohen and Cohen [2008].

Una referencia en castellano muy útil puede ser `http://knuth.uca.es/moodle/course/view.php?id=37`.

1.2 *Lo primero*

Vamos a empezar con el manejo básico del lenguaje R. Una visión general del software y sus posibilidades la podemos encontrar en http://www.r-project.org/.

Instalación Los pasos a seguir son los siguientes:

1. Bajamos el programa de la siguiente dirección http://cran.r-project.org/.

2. Podemos bajar y utilizar la versión que necesitemos. En este curso utilizaremos la versión para Windows [5] aunque se pueden utilizar las versiones para Linux/GNU o bien para MacOS X.
3. Una vez hemos bajado el paquete se instala ejecutándolo con las opciones por defecto.

[5] Los comentarios se referirán a la versión de Windows. Un consejo: usad cualquiera de las otras y abandonad Windows. Si alguna persona usa Linux que me consulte.

Inicio de una sesión En el escritorio tenemos el icono de R. Simplemente clicando el icono iniciamos la sesión de trabajo.

Instalación de un paquete R tiene muchos paquetes que extienden el R base que acabamos de instalar. De hecho, es casi imposible realizar un análisis estadístico por sencillo que sea sin utilizar paquetes de R. Vamos a instalar el paquete **?**, UsingR. Es un paquete con herramientas para la enseñanza de Estadística básica. Lo podemos hacer de distintas formas.

1. Utilizando **Paquetes-Instalar paquetes**. Elegimos previamente el espejo desde donde vamos a bajar este paquete.
2. Utilizando **install.packages**, es el procedimiento más simple.

```
install.packages(``UsingR'')
```

Cargando un paquete Una vez instalado el paquete, para poder usar las funciones o datos que contenga, debemos *cargarlo* mediante

```
library(UsingR)
```

Ahora podemos utilizar las extensiones que proporciona a R este paquete.

1.3 *Lectura y escritura de datos*

Supongamos que hemos recogido unos datos y queremos analizarlos con R. *Hemos de leer estos datos desde R.* Hay muchas opciones para hacerlo.

Lo primero que necesitamos es algún programa para introducirlos en un fichero que luego leeremos desde R. Utilizamos unos datos reales. En concreto son las temperaturas máxima y mínima en la ciudad de Alicante en los años 1939 y 1940 durante cada uno de los 12 meses del año. Tenemos pues 12 observaciones (meses) y, para cada observación, las temperaturas máxima y mínima en el año 1939 y 1940. Estas temperaturas son las variables observadas. Los datos (en décimas de grado) son los siguientes: tenemos 12 observaciones (corresponden con las distintas filas) y hemos observado 4 variables (que corresponden con las cuatro columnas).

1	182	154	82	62
2		180		79
3	183	209	64	93
4	205	224	88	100
5	239	255	112	123
6	267	275	149	150
7	302	303	183	180
8	310	313	189	195
9	291	287	181	164
10	241	237	139	119
11	211	201	90	90
12	176	166	70	54

Hemos de darle un nombre a las columnas que nos identifique la variable. Por ejemplo, los nombres que vamos a usar pueden ser las siguientes

Mes Nos indica el mes del año.

MaxAlicante39 Temperatura máxima en Alicante en 1939.

MaxAlicante40 Temperatura máxima en Alicante en 1940.

MinAlicante39 Temperatura mínima en Alicante en 1939.

MinAlicante39 Temperatura mínima en Alicante en 1940.

Es importante notar que hay huecos. ¿A qué corresponden estos huecos? Son datos faltantes. En concreto podemos ver en febrero (segunda fila) del año 1939 no se observó la temperatera máxima y mínima (segunda y cuarta columna).

¿Cómo podemos introducir estos datos en un fichero de algún tipo que luego leamos desde R?

1.3.1 Con Calc de LibreOffice

Podemos utilizar una aplicación como Calc (de LibreOffice) [6] que nos produzca un fichero texto y luego utilizaremos la función *read.table* para leer los datos.

Es frecuente que se introduzcan datos utilizando una hoja de cálculo. Una opción cómoda y segura es *Calc*, versión libre de Excel, incluida en LibreOffice. Se encuentra disponible en la versión Windows y todos los sistemas operativos. Los pasos a seguir son los siguientes

1. Empezamos por abrir un documento LibreOffice y elegimos la opción Calc.

[6] La suite LibreOffice la podemos conseguir en http://www.libreoffice.org/. En este curso todo el software que se utiliza es software libre. Soy de la opinión de que en educación se debe utilizar exclusivamente software libre.

2. En la primera línea es conveniente (pero no imprescindible) poner los nombres de las variables. Podemos utilizar los nombres que sugerimos para nuestros datos. En cualquier caso los nombres de las variables no han de tener blancos y, para no complicarse la vida, lo mejor es que empiecen con alguna letra y contenga letras y números exclusivamente. Notemos que R diferencia mayúsculas y minúsculas.

3. En la columna correspondiente introducimos los datos anteriores, cada variable en una columna distinta. Si no disponemos del dato (como así es en un par de casos) dejamos en blanco la casilla. Dependiendo de la configuración que tengamos es probable que Calc necesite la coma para indicar la coma decimal.

4. En *Archivo-Guardar como* elegimos *Texto CSV*. Indicamos un nombre para el fichero, por ejemplo, temperaturas_Alicante_39-40.csv.

5. Marcamos *Editar configuración de filtros*.

6. Como separador de campo elegimos ";".[7]

7. Y ya tenemos un fichero texto.

[7] Tiene la ventaja de que no confundimos con la coma decimal.

Nos falta *leerlo desde R*.

1.3.2 *Lectura de un fichero texto*

Si tenemos los datos en un fichero texto de modo que en cada línea tenemos una observación y en cada columna una variable. El fichero a leer será *temperaturas_Alicante_39-40.txt*. ¿Cómo leemos estos datos desde R? Una opción sencilla es con *read.table*.

```
x = read.table(file="../data/temperaturas_Alicante_39-40.csv",
    dec=".",sep = ";",header = TRUE)
```

Si el punto decimal lo hemos con una coma entonces cambiamos *dec="."* por *dec=","*.[8] en la primera fila de los datos hemos puesto los nombres de las variables y esto lo indicamos con `header = TRUE`.

[8] Un consejo, olvidad el sistema español e indicad siempre el decimal con un punto. Evitaréis problemas. Bueno casi olvidaros de España y viviréis mejor. El Polo Sur es un buen lugar para establecerse. Al menos, mejor que nuestro país.

Podemos ver los datos que acabamos de introducir con

```
x
```

```
##    Mes MaxAlicante39 MaxAlicante40 MinAlicante39
## 1    1           182           154            82
## 2    2            NA           180            NA
## 3    3           183           209            64
## 4    4           205           224            88
## 5    5           239           255           112
```

```
## 6    6           267           275           149
## 7    7           302           303           183
## 8    8           310           313           189
## 9    9           291           287           181
## 10  10           241           237           139
## 11  11           211           201            90
## 12  12           176           166            70
##    MinAlicante40
## 1             62
## 2             79
## 3             93
## 4            100
## 5            123
## 6            150
## 7            180
## 8            195
## 9            164
## 10           119
## 11            90
## 12            54
```

1.4 *Sobre lo imprescindible en R*

En esta sección vamos a ver las funciones que, en mi opinión, son básicas cuando trabajamos con R.

1.4.1 *La función c*

Otro modo (menos elegante) de declararle a R datos es la siguiente.

```
x = c(35.84122,28.95458,36.02971,33.13809,39.55091,39.48182,27.52009,32.58105,
31.54865,36.73312,33.87558,30.05730,29.45515,38.70321,34.80034,35.86523,
32.76480,35.94576,30.44356,38.75483,31.21475,33.15148,36.17373,28.34059,
40.52086,39.34035,34.26828,41.92718,34.83630,43.46855)
```

La función *c* nos sirva para *concatenar* uno detrás de otro los datos numéricos. Veamos si lo hemos hecho bien.

```
x

##  [1] 35.84122 28.95458 36.02971 33.13809 39.55091 39.48182
##  [7] 27.52009 32.58105 31.54865 36.73312 33.87558 30.05730
## [13] 29.45515 38.70321 34.80034 35.86523 32.76480 35.94576
```

```
## [19] 30.44356 38.75483 31.21475 33.15148 36.17373 28.34059
## [25] 40.52086 39.34035 34.26828 41.92718 34.83630 43.46855
```

1.4.2 *Selección de casos*

¿Cuál es el primer valor de este vector de datos?

```
x[1]

## [1] 35.84122
```

¿Y el que ocupa la posición 13?

```
x[13]

## [1] 29.45515
```

Podemos ver los datos que están entre el 13 y el 23. Para ello fijémonos en el siguiente código.

```
13:23

##  [1] 13 14 15 16 17 18 19 20 21 22 23
```

Cuando ponemos dos enteros separados por : nos devuelve todos los enteros entre el primero y el segundo. Ahora podemos ver los datos que ocupan estas posiciones en el vector x.

```
x[13:23]

##  [1] 29.45515 38.70321 34.80034 35.86523 32.76480 35.94576
##  [7] 30.44356 38.75483 31.21475 33.15148 36.17373
```

Podemos tener interés en saber los valores de los datos que ocupan las posiciones 7, 9 y de la 20 a la 25. Estas posiciones las podemos obtener con

```
c(7,9,20:25)

## [1]  7  9 20 21 22 23 24 25
```

y los valores de x serían

```
x[c(7,9,20:25)]

## [1] 27.52009 31.54865 38.75483 31.21475 33.15148 36.17373
## [7] 28.34059 40.52086
```

Puede que nuestro interés en ver los datos no venga dado por la posición que ocupan sino por su valor. Por ejemplo, queremos saber cuántos de estos datos superan o son iguales a 35. ¿Cómo lo hacemos? Lo lógico es comparar los valores de x con 35. Lo hacemos con

```
x >= 35
```

```
##  [1]  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE
## [10]  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE
## [19] FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE
## [28]  TRUE FALSE  TRUE
```

Vemos que nos devuelve un vector diciéndonos si es cierta o no la condición que hemos preguntado, si es mayor o igual a 35. Pero: ¿qué valores son? Si hacemos

```
x[x >= 35]
```

```
##  [1] 35.84122 36.02971 39.55091 39.48182 36.73312 38.70321
##  [7] 35.86523 35.94576 38.75483 36.17373 40.52086 39.34035
## [13] 41.92718 43.46855
```

Nos devuelve los datos que ocupan las posiciones donde se daba la condición, donde la condición era cierta. Podemos saber qué valores toman los datos que son mayores que 37 con

```
x[x > 37]
```

```
## [1] 39.55091 39.48182 38.70321 38.75483 40.52086 39.34035
## [7] 41.92718 43.46855
```

o bien los datos que son mayores que 35 y menores o iguales que 37.

```
x[x > 35 & x <= 37]
```

```
## [1] 35.84122 36.02971 36.73312 35.86523 35.94576 36.17373
```

Podemos querer que los casos que estamos seleccionando estén en un nuevo vector.

```
y = x[x > 35 & x <= 37]
```

y podemos ver los valores de y.

```
y
```

```
## [1] 35.84122 36.02971 36.73312 35.86523 35.94576 36.17373
```

1.5 Algunas cosas útiles R

1.5.1 De cómo guardar un dibujo

Primero hemos de hacerlo. Supongamos que queremos un histograma de los datos que tenemos en el vector x (figura 1.1).

El dibujo anterior podemos querer guardarlo en un fichero externo posiblemente para incorporarlo después a un documento. [9] En el siguiente código lo guardamos en un fichero pdf.

```
pdf("histograma_ejemplo.pdf") #Fijamos el nombre del fichero
hist(x)
dev.off()
```

```
hist(x)
```

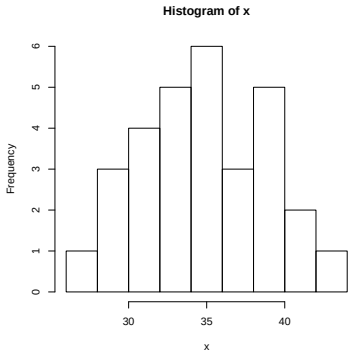


Figura 1.1: Histograma

[9] Espero que no sea un documento Word. Pero asumo que, por desgracia, lo será.

1.5.2 De cómo conseguir ayuda con R

Supongamos que buscamos ayuda sobre las opciones de la función que nos dibuja un histograma, *hist*.

Lo más simple es utilizar la ayuda en html. Utilizamos la siguiente función.

```
help.start()
```

Vemos que nos abre el navegador y nos ofrece distintas opciones. Quizás la opción más simple sea utilizar la herramienta de búsqueda.

Otra opción es, en línea de comandos escribir

```
?hist
```

O simplemente,

```
help(hist)
```

1.5.3 De cómo trabajar con R

Hay dos formas de trabajar con R. La primera opción es utilizar un editor en el que escribimos código de R. Lo copiamos y luego lo pegamos en la línea de comandos de R.[10] Dentro de esta manera de trabajar podemos utilizar distintas editores (que llevan herramientas para facilitarnos el trabajo).

[10] Es la opción más educativa en donde aprendemos realmente a trabajar con el programa.

Con el editor de R El propio programa lleva un editor incorporado. Es básico pero suficiente. Es la opción que utilizaremos en las clases prácticas.

TinnR En Windows se puede usar el programa TinnR que lo podemos descargar desde http://sourceforge.net/projects/tinn-r/.

Bloc de notas Por último una opción simple y efectiva es abrir un editor de textos (como el bloc de notas pero no el Word) e ir escribiendo el código allí. Luego aplicamos el famoso copiar y pegar.

RStudio Quizás ahora mismo sea la mejor opción. Es un programa que incorpora el editor, nos muestra las salidas, los gráficos y la historia previa. De manejo muy simple. http://rstudio.org/

Otra opción es utilizar un interfaz gráfico. La opción a la que más acostumbrados estamos y, creemos, es más sencilla.[11]

[11] No es mi opinión. Encuentro mucho más difícil esta opción pero para gustos ...

RStudio 1.1. *Una vez hemos instalado RStudio lo ejecutamos pinchando en el icono y obtenemos una pantalla como aparece en la figura 1.2.*

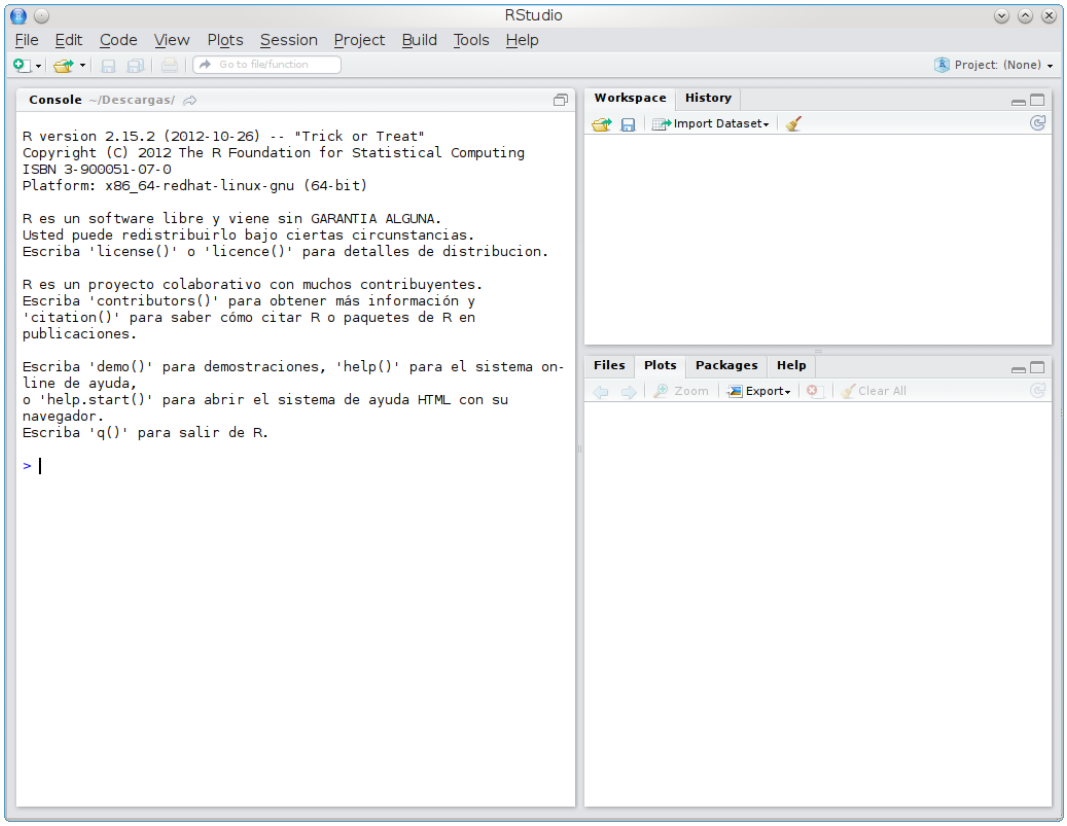


Figura 1.2: Pantalla de inicio de RStudio.

1.6 Seguimos con lo básico de R

En esta sección pretendemos ver cómo trabajar de un modo simple con las variables que componen el banco de datos. Vamos a utilizar unos datos sobre calidad del aire. Estos datos los tenemos en el paquete [12]. por lo que empezamos cargando el paquete.

[12]

```
library(datasets)
```

Los datos se llaman *airquality*. Con el siguiente código podemos obtener información sobre los datos que vamos a manejar.

```
help(airquality)
```

Adjuntamos para poder usar los nombres de las variables.

```
attach(airquality)
```

Nos fijamos en lo que sigue en las variables *Ozone* y *Month* (cuyo significado podemos consultar con la ayuda anterior).

Podemos ver las cinco primeras mediciones de ozono con

```
Ozone[1:5]

## [1] 41 36 12 18 NA
```

o bien la muestra que ocupa la posición 100 en el vector de datos.

```
Ozone[100]

## [1] 89
```

También podemos queremos ver conjuntamente los datos de la muestra 23, esto es, tanto el ozono como el mes con

```
c(Ozone[23],Month[23])

## [1] 4 5
```

Podemos querer conocer qué cuáles son los datos correspondientes a mayo con

```
which(Month == 5)

##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
## [19] 19 20 21 22 23 24 25 26 27 28 29 30 31
```

y las concentraciones serían

```
Ozone[which(Month == 5)]
```

```
##  [1]  41  36  12  18  NA  28  23  19   8  NA   7  16  11  14
## [15]  18  14  34   6  30  11   1  11   4  32  NA  NA  NA  23
## [29]  45 115  37
```

Podemos contar el número de datos que tenemos de cada mes con

```
length(which(Month == 5))
```

```
## [1] 31
```

```
length(which(Month == 6))
```

```
## [1] 30
```

```
length(which(Month == 7))
```

```
## [1] 31
```

Una forma bastante más sencilla de hacerlo es hacer una tabla que nos lo cuente. Por ejemplo con

```
table(Month)
```

```
## Month
##  5  6  7  8  9
## 31 30 31 31 30
```

¿ Qué observación corresponde con la máxima concentración de ozono?

```
which.max(Ozone)
```

```
## [1] 117
```

y toda la información de dicha observación vendría dada por

```
airquality[which.max(Ozone),]
```

```
##     Ozone Solar.R Wind Temp Month Day
## 117   168     238  3.4   81     8  25
```

Cuando nos planteamos transformar los datos originales aplicando alguna función como la raiz cuadrada simplemente hemos de hacer lo siguiente (mostramos los cinco primeros datos)

```
sqrt(Ozone)[1:5]

## [1] 6.403124 6.000000 3.464102 4.242641       NA
```

Podemos plantearnos ver cuántos valores superan una concentración de 40 unidades. Esto lo podemos hacer con

```
which(Ozone >= 40)

##  [1]   1  29  30  40  62  63  66  67  68  69  70  71  77  79
## [15]  80  81  85  86  88  89  90  91  92  96  98  99 100 101
## [29] 104 106 109 112 116 117 118 120 121 122 123 124 125 126
## [43] 127 128 134 139
```

1.7 *Dato faltante*

Con mucha frecuencia no tenemos todas las variables observadas sobre todas las observaciones, esto es, no tenemos observaciones *completas*. Por ejemplo, en la serie de temperaturas en Alicante nos faltan datos del año 1939. ¿Cómo manejamos los datos faltantes con R? Tenemos distintas opciones.

Si los datos los hemos introducido en un fichero utilizando Calc (o Excel) entonces cuando no tenemos el dato correspondiente dejamos en blanco la casilla. Cuando luego leemos los datos en R ese dato lo tendrá como dato faltante.

Si utilizamos un editor de texto (como Bloc de Notas en Windows) entonces con NA indicamos que tenemos un dato faltante. No podemos dejar el hueco sin más.

Por ejemplo, consideremos los siguientes datos.

```
 1      182      154       82       62
 2       NA      180       NA       79
 3      183      209       64       93
 4      205      224       88      100
 5      239      255      112      123
 6      267      275      149      150
 7      302      303      183      180
 8      310      313      189      195
 9      291      287      181      164
10      241      237      139      119
11      211      201       90       90
12      176      166       70       54
```

Podemos copiar y pegar las filas anteriores en un fichero de texto. Cuando leemos los datos con read.table las variables donde hemos indicado NA las entiende como dato faltante.

1.8 Estadística medioambiental en R

Este curso es básico. No pretende más que introducir lo fundamental de las técnicas estadísticas. Sin embargo, la elección de R como herramienta viene motivada por la gran cantidad de paquetes en R para el análisis de datos medioambientales. La dirección http://cran.r-project.org/web/views/Environmetrics.html tiene una relación actualizada y comentada de estos paquetes.

1.9 Datos

En este curso usaremos distintos bancos de datos. Algunos de ellos son datos propios que podéis encontrar en el directorio DATOS del Aula Virtual.

1.9.1 Datos de precipitaciones y temperaturas en la Comunidad Valenciana

En las secciones 1.9.1,1.9.1,1.9.1 y 1.9.1 comentamos distintos datos de precipitaciones y temperaturas correspondientes a las poblaciones de Orihuela, Alicante, Utiel, Valencia, Castellón y Morella. [13] Todos los datos están en décimas de grados (para las temperaturas) o en décimas de mm (1 mm de lluvia = 1 litro/m2).

[13] Estos datos han sido proporcionados amablemente por la Agencia Estatal de Metereología. Agradezco especialmente a Braulio Aguilar su colaboración.

Pdiaria Nuestro banco de datos se refiere a la precipitación diaria de 7 de la mañana de un día a 7 de la mañana del día siguiente en 2010. Las variables son que aparecen en este banco de datos son:

INDICATIVO Un código indicativo de la población.

ANYO Año.

MES Mes del año indicado con el número.

NOMBRE Nombre de la población.

ALTITUD En metros.

CX Coordenada geográfica x.

CY Coordenada geográfica y.

P1 Precipitación el día 1. De un modo análogo P2 indica precipitación el segundo día y así sucesivamente.

Las precipitaciones son indicadas en décimas de milímetro de modo que un milímetro de lluvia corresponde con un litro por metro cuadrado. El valor -3 indica que la cantidad registrada es inapreciable. El

valor -4 indica que el valor que observamos es un valor acumulado de días consecutivos (no se realizó la medida en sus días correspondientes).

Tdiaria Nuestro banco de datos se refiere a las temperaturas máxima y mínima en cada día. Las variables son:

INDICATIVO Un código indicativo de la población.

ANYO Año.

MES Mes del año indicado con el número.

NOMBRE Nombre de la población.

ALTITUD

CX Coordenada geográfica.

CY Coordenada geográfica.

TMAX1 Temperatura máxima el día 1. Análogamente tenemos las variables que nos dan la temperatura para cada día del mes.

TMIN1 Temperatura mínima el día 1. Análogamente tenemos las variables que nos dan la temperatura mínima para cada día del mes.

Las temperaturas aparecen en décimas de grado.

Precipitacion1964-2011 Recoge la precipitación mensual, desde que se tienen registros, para distintas localidades de la Comunidad Valenciana.

INDICATIVO Un código indicativo de la población.

ANYO Año.

MES Mes del año indicado con el número.

NOMBRE Nombre de la población.

ALTITUD En metros.

CX Coordenada geográfica.

CY Coordenada geográfica.

PMES77 Es la precipitación mensual obtenida sumando las precipitaciones diarias de 7 de la mañana de un día a 7 de la mañana del día siguiente.

PMAX Es la precipitación diaria máxima en ese mes. No se indica el día que se produjo.

Temperatura1964-2011 Es la temperatura máxima y mínima mensual de distintas estaciones desde que se tienen registros para distintas localidades de la Comunidad Valenciana.

alicante_temperaturas_mes_1939_2010 Son datos relativos a temperaturas en la ciudad de Alicante desde el año 1939 hasta el año 2010. Las observaciones corresponden con meses del año. Como variables tenemos el mes así como las temperaturas máxima y mínima en cada uno de los años considerados. Las variables *tmax1939* y *tmin1939* corresponden con la temperatura máxima y mínima en el año 1939.

alicante_temperaturas_anyo_1939_2010 Son datos relativos a temperaturas en la ciudad de Alicante desde el año 1939 hasta el año 2010. Las observaciones corresponden con los años. Como variables tenemos el año en que observamos así como las temperaturas máxima y mínima en cada uno de los meses. En concreto las variables *tmax1* y *tmin1* corresponden con la temperatura máxima y mínima en el primer mes, es decir, enero.

1.9.2 Concentraciones de ozono

El fichero *ozono2011.txt* tiene los datos sobre las concentraciones de ozono en distintas ciudades durante el año 2011 (son datos modificados a partir de datos reales). La unidad son partes por cada mil millones. Una concentración alta en el aire es un problema de salud de importancia. Las normativas de los distintos países fijan los niveles máximos (y el tiempo que se pueden sobrepasar). Las variables que aparecen el banco de datos son:

dia Día del año (de 1 a 365).

estacion La estación.

Santiago La concentración media durante el día en Santiago. De un modo análogo las variables **Madrid, Murcia, Godella, Zaragoza** contiene la concentración en el día correspondiente en estas ciudades.

2
Estadística descriptiva

2.1 Introducción

Tenemos unos datos numéricos. Ejemplos de datos medio ambientales son temperaturas o precipitaciones observados en una localización geográfica y un día determinado. Podemos tener muestras de agua en las que determinamos la demanda química o biológica de oxígeno.

Ejemplo 2.1. *Se han tomado mediciones de la concentración de nitrito en agua utilizando el método de electrodo directo selectivo de iones (ISE). Los datos son los siguientes:*

```
0.32 0.36 0.24 0.11 0.11 0.44 2.79 2.99 3.47 0.23 0.55
3.21 4.02 0.23
```

En el ejemplo 2.1 tenemos unos datos que hemos observados en distintas muestras de agua. De un modo genérico denotaremos por x_1 el primer valor observado, por x_2 el segundo valor observado y así sucesivamente. En el lenguaje estadístico a esto se le llama una **muestra**. Por ello, diremos que tenemos una muestra $x_1, \ldots, x_n$ de n datos. Se entiende que estos datos se han tomado en unas condiciones similares. ¿Cómo son estos datos? Pretendemos describirlos de un modo sencillo. Esta descripción será de dos tipos: una descripción numérica, describimos muchos números con unos pocos números que tengan un sentido claro; y una descripción gráfica. Describimos los números con gráficos que destaquen sus propiedades básicas. Al conjunto de técnicas que nos dan descripciones numéricas y gráficas de un conjunto de datos reciben el nombre de **Estadística descriptiva** y, con frecuencia, simplemente **descriptiva**. Se habla de la *descriptiva de los datos*. Veremos que son ideas sencillas pero de un uso constante.

Cuando se describen los datos las preguntas básicas que hemos de tener en la cabeza pueden ser:

1. ¿De qué orden son?

2. ¿Cómo de dispersos están?

3. ¿Hay datos anormales que estén muy alejados de los demás?

En la primera pregunta intentamos localizar los valores: ¿estamos alrededor de 2?, ¿o bien alrededor de 20?, ¿o alrededor de 200000? Pretendemos localizar la muestra, dar un valor representativo de todos ellos. En la segunda pregunta nos preguntamos si los datos se agrupan si están próximos entre sí. Por último, nos planteamos si tenemos datos que son anormales. Obviamente lo que es anormal depende de cómo son los otros. Por ejemplo, si medimos nivel de radioactividad cerca de un reactor nuclear podemos observar valores que serían anormales si los agrupamos con mediciones tomadas en una zona muy alejada de cualquier central nuclear. Sin embargo, no lo son en el entorno de dicha central.

2.2 *Descriptivas numéricas*

Empezamos con las descripciones numéricas. Leemos los datos del ejemplo 2.1.

```
x = c(0.32,0.36,0.24,0.11,0.11,0.44,2.79,2.99,3.47,0.23,0.55,
    3.21,4.02,0.23)
```

De un modo genérico se pretende describir un conjunto de datos numéricos mediante unos pocos números (elegidos, eso sí, con gracia). En particular vamos a considerar *medidas de localización* y *medidas de dispersión*. Las medidas de localización intentan responder la pregunta que veíamos antes de: ¿de qué orden son los datos? Las medidas de dispersión intentan responder a la segunda pregunta, ¿cómo de dispersos son los datos? ¿Cómo de variables son los datos?

Como medidas de localización veremos la media y medianas muestrales fundamentalmente y como medidas de dispersión básicas veremos la varianza y desviación típica muestrales.

2.2.1 *Media muestral*

La medida de localización más utilizada es la media aritmética o *media muestral* (que es el nombre habitualmente usado en Estadística) que se define como

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}. \tag{2.1}$$

La podemos calcular con

```
mean(x)

## [1] 1.362143
```

Notemos que si conocemos solamente la variable x y queremos saber cuántos datos tenemos lo mejor es hacer

```
length(x)

## [1] 14
```

Y, aunque no muy recomendable, otro modo de calcular la media muestral es calcular la suma y dividir por el número de términos que estamos sumando. Esto lo podemos hacer utilizando las funciones *sum* y *length*.

```
sum(x)/length(x)

## [1] 1.362143
```

2.2.2 *Media ajustada*

La media muestral es, sin duda, la mejor forma de localizar una muestra. No obstante es muy sensible a datos anómalos o extremos (y que con frecuencia son errores introducidos en el banco de datos). Por ejemplo, a nuestros datos originales les vamos a añadir un dato anómalo. Le añadimos el valor 34.

```
(xx = c(x,34))

##  [1]  0.32  0.36  0.24  0.11  0.11  0.44  2.79  2.99  3.47
## [10]  0.23  0.55  3.21  4.02  0.23 34.00
```

Podemos comparar la media original y la nueva media muestral.

```
mean(x)

## [1] 1.362143

mean(xx)

## [1] 3.538
```

Se ha modificado muchísimo. Puede que el dato sea real pero puede que sea un error. En cualquier caso es un único valor entre otros muchos. Puede interesarnos localizar nuestra muestra sin atender a estos datos anómalos. ¿Cómo? Una opción simple es fijarnos en una

cierta proporción de los datos, por ejemplo, una proporción α y eliminar el α por uno de los datos más pequeños y el α por uno de los más grandes. La media de los que quedan es la *media ajustada*. La media ajustada se obtiene con *mean* indicándole un parámetro adicional.

```
mean(x,trim=.1)

## [1] 1.245
```

Estamos eliminando el 10 % de los datos mayores y el 10 % de los datos menores. La media de los restantes es nuestra media ajustada. Ahora podemos comparar la media ajustada de los datos originales y de los datos con la observación anómala.

```
mean(x,trim=.1)

## [1] 1.245

mean(xx,trim=.1)

## [1] 1.458462
```

Vemos cómo no se ha modificado demasiado. Estamos describiendo la localización de la parte central de los datos despreciando los datos extremos a ambos lados.

2.2.3 *Percentiles*

Otro manera de localizar los datos es utilizar los *percentiles muestrales*. Supongamos que tomamos un valor p entre 0 y 1. El *percentil de orden p* es un valor que tiene por debajo el $100 \times p$ por ciento de los datos y por encima el $100 \times (1 - p)$ por ciento. Denotaremos el percentil de orden p como q_p. Ordenamos nuestros datos de menor a mayor con la función *sort*.

```
sort(x)

##  [1] 0.11 0.11 0.23 0.23 0.24 0.32 0.36 0.44 0.55 2.79 2.99
## [12] 3.21 3.47 4.02
```

¿Cuántos de los datos originales son menores o iguales a 1? Podemos contar a mano. [1] Otra posibilidad es utilizar la función *ecdf*.

[1] Cosa antigua en franco retroceso y que no está prohibido hacer. El inconveniente es cuando tenemos centenares o miles de observaciones.

```
Fn = ecdf(x)
Fn(1)

## [1] 0.6428571
```

Vemos pues que la proporción de datos que son inferiores a 1 es de 0.6428571. O dicho de otro modo: el valor 1 es el percentil de orden 0.6428571 de los datos.

La función básica para calcular los cuantiles es *quantile*. El detalle exacto del procedimiento utilizado para estimar estos valores se puede consultar con *help(quantile)*. [2]

[2] Se pueden ver hasta nueve procedimientos distintos.

La *mediana muestral* es el percentil de orden 0,5 esto es por debajo tiene al menos la mitad de los datos y por encima la otra mitad. La podemos obtener con

```
median(x)

## [1] 0.4
```

De hecho, podemos plantearnos cómo conseguir un percentil de orden p (con $0 < p < 1$) arbitrario. Tomemos $p = 0{,}27$.

```
quantile(x,probs = 0.27)

##    27%
## 0.2351
```

O bien $p = 0{,}76$

```
quantile(x,probs = 0.76)

##   76%
## 2.966
```

Cuando $p = 0{,}25$ al percentil le llamamos *cuartil inferior*. Si $p = 0{,}75$ tenemos el *cuartil superior*.

```
quantile(x,probs = c(0.25,0.75))

##    25%    75%
## 0.2325 2.9400
```

2.2.4 *Varianza y desviación estándar muestrales*

Ahora pretendemos cuantificar lo dispersos que están nuestros datos. Las dos medidas más utilizadas son la varianza y la desviación estándar. La *varianza muestral* se define como

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}, \tag{2.2}$$

y la desviación estándar (o típica) muestral se define como

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}. \tag{2.3}$$

La varianza se puede calcular con

```
var(x)

## [1] 2.32071
```

y la desviación estándar la obtenemos con

```
sd(x)

## [1] 1.523388
```

2.2.5 *Rango*

El mínimo y el máximo lo podemos obtener con

```
range(x)

## [1] 0.11 4.02
```

o bien con

```
min(x)

## [1] 0.11

max(x)

## [1] 4.02
```

El rango, esto es, el máximo valor menos el mínimo valor lo podemos obtener con

```
max(x)-min(x)

## [1] 3.91
```

o bien con

```
diff(range(x))

## [1] 3.91
```

2.2.6 *Rango intercuartílico*

Una medida más robusta que el rango es el rango intercuartílico. Se define como la diferencia entre los percentiles de orden 0,75 y 0,25, es decir, el cuartil superior menos el cuartil inferior. Se puede obtener con

```
IQR(x)

## [1] 2.7075
```

2.2.7 *La función summary*

Es una función que nos proporciona una descripción básica de los datos. En concreto, nos da el mínimo, el primer cuartil, la media, la mediana, el tercer cuartil y el máximo.

```
summary(x)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1100  0.2325  0.4000  1.3620  2.9400  4.0200
```

Sin duda, es la opción más simple para obtener una descripción rápida de los datos.

2.3 *Ejercicios*

Ejercicio 2.1. *Consideremos los siguientes datos.*

```
43.86 33.14 37.04 29.29 21.49 34.98 18.09 18.84 36.20 27.82 22.86 32.11
22.45 38.22 44.55 39.03 33.25 18.27 34.44 24.88 24.58 42.12 30.04 19.58
34.00 32.98 28.35 25.75 22.78 15.88 38.97 13.47 21.42 34.19 16.49 15.17
31.42 17.00 37.06 30.35 19.65 34.62 16.48 19.42 42.89 23.89 29.26 45.64
32.29 22.96 29.60 39.98 21.86 18.25 35.96 30.57 40.79 17.21 27.07 28.56
15.59 23.51 18.78 37.72 14.40 28.40 43.17 22.65 27.85 41.56 42.44 16.57
23.55 29.66 20.72 28.28 42.10 13.76 27.27 19.69 20.18 23.80 14.37 22.02
29.06 34.52 21.91 19.98 16.24 44.56 18.54 35.96 30.12 32.82 45.76 28.75
32.01 19.39 23.76 41.72 32.90 31.47 15.04 12.74 44.11 38.65 27.18 35.52
15.70 38.95 30.59 15.43 45.60 14.98 23.11 22.11 23.03 19.91 34.95 16.05
```

Se pide:

1. *Leer los datos utilizando el método que se prefiera.*
2. *Calcular la media, mediana, media recortada con una proporción del* 0,05*, los percentiles de orden* 0,1 *y* 0,9*.*
3. *Supongamos que se han seguido recogiendo datos. En concreto una segunda muestra con los siguientes valores.*

   ```
   123.34 78.23 89.6 1.2
   ```

   Incorporar estas nuevas observaciones a los datos originales y calcular las descriptivas numéricas anteriores sobre los nuevos datos. Indicar cuáles de ellas varían y cuáles no justificando la respuesta.

2.4 *Descripciones gráficas de los datos*

En esta sección pretendemos ver algunas de las descripciones gráficas para variables numéricas y categóricas.

2.4.1 *Añadimos variables y seleccionamos casos o variables*

Vamos a añadir más información en nuestro banco de datos. Seguimos con las concentraciones de nitritos del ejemplo 2.1. En concreto se sabe que las muestras de agua se tomaron en dos localizaciones de la Albufera distintas. Las primeras 8 muestras se tomaron en el puerto de Catarroja y las 6 últimas muestras se tomaron en El Palmar. Esta información la guardamos en un vector

```
y = c(1,1,1,1,1,1,1,1,2,2,2,2,2,2)
```

2.4.2 *Frecuencias*

La segunda variable que hemos introducido en el banco de datos es la zona en que tomamos la medida. Es pues una variable categórica que nos indica la pertenencia del dato a una categoría, en este caso, la zona en que se observa el dato. La descripción básica más simple son los *conteos o frecuencias absolutas*. Contamos el número de veces que se repiten cada una de las categorías. Tendremos el número de datos que se ha observado en cada zona. Los obtenemos de un modo simple con la función *table*.

```
table(y)

## y
## 1 2
## 8 6
```

Si dividimos las frecuencias absolutas por el número total de datos tenemos las *frecuencias relativas*. Por ejemplo, con

```
prop.table(table(y))

## y
##         1         2
## 0.5714286 0.4285714
```

O, de otro modo, si sumamos la tabla nos da el total de casos

```
sum(table(y))

## [1] 14
```

y podemos obtener las frecuencias relativas dividiendo los conteos o frecuencias absolutas por esta suma.

```
table(y)/sum(table(y))

## y
##         1         2
## 0.5714286 0.4285714
```

Las frecuencias absolutas y relativas las podemos representar gráficamente con un diagrama de barras. Bien las frecuencias absolutas (figura 2.1) o relativas (figura 2.2). Como dibujo es el mismo y solamente nos cambia la escala que observamos en ordenadas.

Una representación que nos permite representar frecuencias (absolutas o relativas pues el dibujo no cambia) es el *diagrama de sectores.*

2.4.3 *Histograma*

Para una variable cuantitativa una buena opción para observar la distribución de los datos es un histograma. La idea de un histograma es (demasiado) simple. Si $x_1, \ldots, x_n$ son los datos de los cuales queremos construir el histograma consideramos el intervalo que va del mínimo al máximo, es decir, el intervalo

$$[a, b] = [\text{mín}\{x_1, \ldots, x_n\}, \text{máx}\{x_1, \ldots, x_n\}]$$

y lo subdivimos en un número de subintervalos con la misma longitud. A estos subintervalos se les suele llamar *clases*. Supongamos que elegimos k clases. Entonces los subintervalos que consideramos son

$$[a, a+\delta), [a+\delta, a+2\delta), \ldots, [a+(k-1)\delta, b]$$

donde

$$\delta = \frac{b-a}{k}$$

Dependiendo del software que utilicemos los valores de a y b suelen elegirse como un poco menos que el mínimo y un poco más que el mínimo. El número de clases se elige de un modo automático pero siempre modificable por el usuario. Contamos el número de datos que hay en cada clase. Representamos una barras (que se representan pegadas una con otra lo que también nos permite diferenciarlo de un diagrama de barras) cuya base coincide con el subintervalo y cuya altura es proporcional al número de datos que hemos observado en

```
barplot(table(y))
```

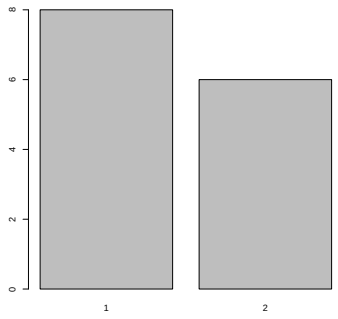


Figura 2.1: Diagrama de barras con las frecuencias absolutas.

```
barplot(prop.table(table(y)))
```

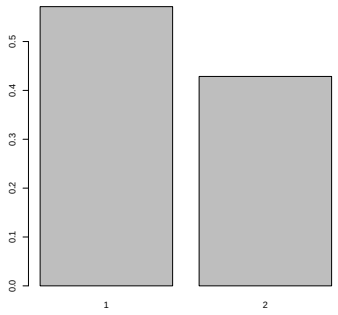


Figura 2.2: Diagrama de barras con las frecuencias relativas.

```
pie(table(y))
```

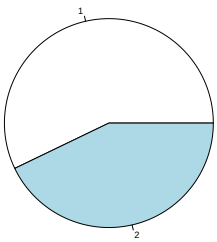


Figura 2.3: Diagrama de sectores.

dicho subintervalo. Este es el dibujo. Veamos cómo hacerlo con R. Si no le indicamos nada el programa decide el número de clases o subintervalos (figura 2.4).

Podemos cambiar el número de clases y vemos que el dibujo cambia mucho (figura 2.5).

Un dibujo que es demasiado sensible a la elección del usuario (que no tiene porqué tener mucha idea del número de clases a elegir) no es demasiado recomendable. Tiene un uso muy extendido.[3]

2.4.4 *Diagramas de cajas*

Es un dibujo basado en los cuartiles fundamentalmente. [4] La idea es representar una caja como la parte que más destaca en el dibujo. Dentro de la caja se representa con una línea (algo más gruesa habitualmente) la mediana. Los extremos de la caja coinciden con los percentiles del 25 % y del 75 %. La longitud de la caja es la diferencia de estos percentiles, es decir, el rango intercuartílico. La caja muestra la distribución de la mitad de los datos, la mitad de los datos que están centrados.

Se le añaden unos bigotes que salen de la caja. Son unas líneas que describen la variabilidad de los datos que están en los extremos. Hay varias opciones para elegir el punto al que llegan los bigotes. Entre otras opciones las dos más habituales son:

1. Que lleguen al mínimo y al máximo de los datos.
2. Que lleguen al mínimo y al máximo de los datos que están en el intervalo que tiene por extremo inferior el percentil del 25 % menos 1.5 veces el rango intercuartílico y por extremo superior el percentil del 75 % mas 1.5 veces el rango intercuartílico. Este valor de 1.5 obviamente podemos modificarlo.

Supongamos que nos planteamos representar un diagrama de cajas de toda la muestra. En la figura 2.6 tenemos el resultado.

Vamos a añadir a los datos que tenemos unos cuantos valores extremos. En concreto que sean mucho mayores que los que tenemos

```
x1 = c(x,c(7,9))
```

y representamos el diagrama de cajas con los nuevos datos (figura 2.7).

Puede que no nos interese mostrar los puntos extremos. En la figura 2.8 vemos cómo hacerlo.

En figura 2.9 vemos cómo conseguir que los bigotes lleguen a los valores más extremos.

```
hist(x)
```

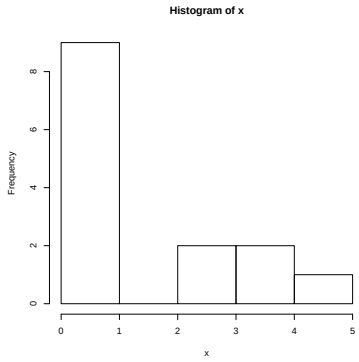


Figura 2.4: Histograma sin elegir número de clases.

```
hist(x,breaks=10)
```

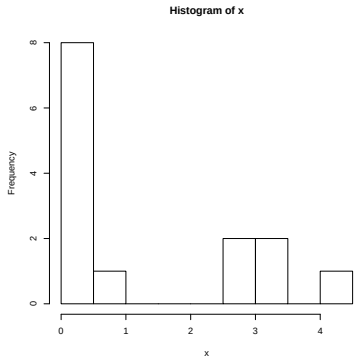


Figura 2.5: Histograma eligiendo el número de clases.

[3] Pero no todo lo que se hace frecuentemente ha de ser bueno.

[4] Una buena explicación se puede consultar en http://en.wikipedia.org/wiki/Box_plot.

```
boxplot(x)
```

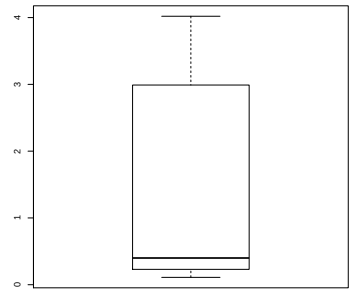


Figura 2.6: Diagrama de cajas.

```
boxplot(x1)
```

No parece que sea la mejor opción (es una opinión claro). Quizás la mayor utilidad de un diagrama de cajas es comparar submuestras, esto es, partes distintas de la muestra. Por ejemplo, para los grupos definidos por la variable categórica y (figura 2.10).

Es un dibujo simple que nos proporciona una comparación rápida de las dos muestras.

```
boxplot(x ~ y)
```

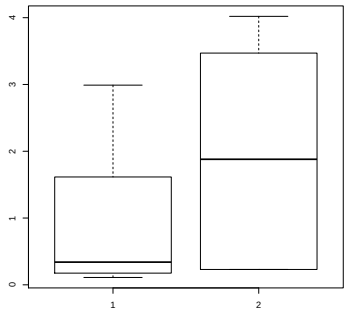


Figura 2.10: Diagrama de cajas comparando muestras.

2.4.5 *Estimadores kernel de la densidad*

Un gráfico alternativo (y de mucha más calidad) es el estimador kernel de la densidad cuya expresión viene dada por

$$\hat{f}(x) = \sum_{i}^{n} \frac{1}{n} K\left(\frac{x_i - x}{h}\right),$$

donde la función K es no negativa y verifica

$$\int_{-\infty}^{+\infty} K(u)du = 1,$$

y

$$K(-u) = K(u),$$

es decir, es simétrica respecto del origen. Una función K que se suele utilizar es la gaussiana dada por

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

En la figura 2.11 aparece un estimador kernel de la densidad utilizando una función kernel gaussiana.

```
plot(density(x))
```

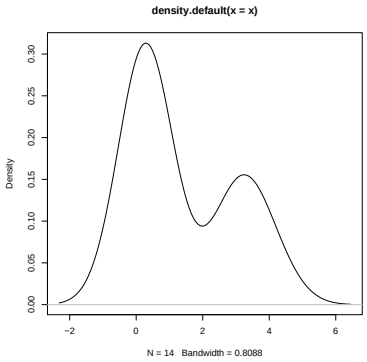


Figura 2.11: Estimador kernel de la función de densidad.

2.4.6 *Función de distribución muestral*

Dados los datos $x_1, \ldots, x_n$, la función de distribución muestral se define como

$$F_n(x) = \frac{|\{x_i : x_i \leq x\}|}{n}, \tag{2.4}$$

donde $|\cdot|$ denota el cardinal, esto es, el número de elementos del conjunto. Para cada punto x consideramos el conjunto formado por todos los datos x_i que son menores o iguales que x. Contamos el número de puntos en este conjunto (su cardinal) y lo dividimos por el total de datos. En resumen, $F_n(x)$ nos da la proporción de datos que son menores o iguales que x. Y esto lo consideramos para todos los valores de x posibles.

Vamos a considerar dos funciones para obtener la función de distribución muestral. La primera es la función *ecdf* y podemos verla en la figura 2.12

Si queremos conocer el valor de F_n en un valor determinado, por ejemplo para $x = 37$ podemos hacer

```
plot(ecdf(x))
```

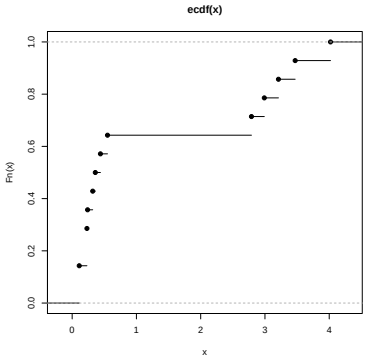


Figura 2.12: Función de distribución muestral con la función *ecdf*.

```
ecdf(x)(37)

## [1] 1
```

o bien en 40,

```
ecdf(x)(40)

## [1] 1
```

La segunda opción es la función *Ecdf* del paquete *Hmisc*. Aparece en la figura 2.13.

```
library(Hmisc)
Ecdf(x)
```

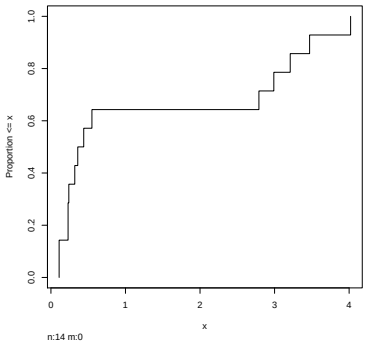


Figura 2.13: Función de distribución muestral con la función *Ecdf*.

2.4.7 *Buscando datos anómalos*

Tenemos unos datos numéricos $x_1, \ldots, x_n$ y queremos saber si hay alguno que se sale de madre. Si hay alguno que está muy alejado de los demás. Que es anómalo. Lo primero es precisar qué entendemos por dato anómalo.[5] Vamos a ver las dos definiciones más habitualmente utilizadas. En la primera utilizamos media y desviación estándar. En la segunda utilizamos cuartiles.

La primera es la que más tradición tiene. Dados los datos calculamos su media y desviación típica muestrales: $\bar{x}$ y s. Se entiende por dato anómalo aquél que está fuera del intervalo

$$[\bar{x} - 3s, \bar{x} - 3s],$$

es decir, que o bien es extremo porque es menor que $\bar{x} - 3s$ o bien es extremo porque es mayor que $\bar{x} + 3s$.

Vamos a añadir a nuestros datos dos datos extremos por arriba y uno por abajo.

```
x2 = c(x,c(0,9,14))
```

Y veamos si son anómalos en el sentido que acabamos de definir.

```
x2[x2 < mean(x2) - 3 * sd(x2)]

## numeric(0)
```

Vemos que el valor que añadimos por debajo no es detectado.

¿Y los dos valores grandes?

```
x2[x2 > mean(x2) + 3 * sd(x2)]

## [1] 14
```

[5] La expresión inglesa es outlier.

Vemos que solamente detectamos uno de ellos. Si no somos tan exigentes (como se suele ser por cierto) y simplemente consideramos que el dato se aleje de la media en dos desviaciones típicas por arriba o por abajo entonces se detectan los siguientes

```
x2[x2 < mean(x2) - 2 * sd(x2)]

## numeric(0)

x2[x2 > mean(x2) + 2 * sd(x2)]

## [1] 14
```

es decir, el mismo.

El segundo procedimiento utiliza los cuartiles. Denotemos por q_{25} y q_{75} los percentiles de orden 25 % y 75 %, esto es, los cuartiles inferior y superior. El rango intercuartílico sería

$$IQR = q_{75} - q_{25}.$$

La segunda forma de considerar un dato como extremo es considerar que lo es si está fuera del intervalo

$$[q_{25} - 1{,}5 \times IQR, q_{75} + 1{,}5 \times IQR].$$

Puede ser extremo por abajo si es menor que $q_{25} - 1{,}5 \times IQR$ o por arriba si es mayor que $q_{75} + 1{,}5 \times IQR$. Determinemos los extremos del intervalo.

```
(lw = quantile(x2,probs=0.25) - 1.5 * IQR(x2))

##   25%
## -4.24

(up = quantile(x2,probs=0.75) + 1.5 * IQR(x2))

##  75%
## 7.68
```

Y veamos si hay puntos extremos por abajo

```
x2[x2 < lw]

## numeric(0)
```

Detecta el punto añadido por nosotros. Y por arriba.

```
x2[x2 > up]

## [1]  9 14
```

Detecta los dos puntos extremos por arriba. En fin, no nos ha ido mal con este método de detección. No obstante, el más habitual es el primero de los procedimientos propuestos.

2.5 *Ejercicios*

Ejercicio 2.2. *Vamos a realizar distintas representaciones gráficas con los datos del ejercicio 2.1. Se pide lo siguiente:*

1. *Realizar distintos histogramas de los datos que aparecen en el ejercicio 2.1 modificando el número de clases. ¿Hay un comportamiento consistente en la representación gráfica?*

2. *Representar gráficamente un estimador kernel de la densidad. Observar el valor que se ha utilizado para el ancho de banda.*

3. *Modificar el valor del ancho de banda observado en el apartado 2 doblando su valor y volver a representar el estimador kernel de la densidad.*

4. *Modificar el valor del ancho de banda observado en el apartado 2 considerando la mitad de su valor y volver a representar el estimador kernel de la densidad.*

5. *Comparar los tres estimadores kernel que hemos obtenido. ¿Qué ocurre cuando incrementamos el ancho de banda? ¿Y cuando lo disminuimos?*

Ejercicio 2.3. *Consideramos los datos del ejercicio 2.1. La muestra inicial la denotamos por x mientras que los datos ampliados los denotamos por xx. Supongamos que los datos x han sido obtenidos en distintas localizaciones. En concreto las localizaciones las tenemos codificadas de un modo numérico. Son las siguientes.*

```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2
```

Se pide:

1. *Introducir estos datos en un vector en R que denotaremos por y utilizando la función* **c()** *de concatenación.*

2. *Realizar un diagrama de cajas de la variable x.*

3. *Realizar diagramas de barras de los valores que tenemos en la variable x para las distintas localizaciones que nos vienen indicadas por la variable y.*

4. *Los datos adicionales que aparecen en el vector xx han sido obtenidos en una cuarta localizacion. Completar el vector yy para incluir esta localización, es decir, en xx tenemos los datos ampliados mientras que en yy tendremos las localizaciones ampliadas.*

5. *Muestra un diagrama de barras comparativo de los valores xx para las distintas localizaciones que aparecen en yy.*

Ejercicio 2.4. *Consideremos las dos muestras siguientes.*

```
x
     22.24 21.04 23.89 22.49 25.22 22.15 22.49 27.51 23.57 25.22 23.47 18.91
     21.64 24.29 21.68 24.51 22.32 24.77 18.30 23.03 22.03 21.09 23.32 21.15
     21.21 25.53 19.34 25.89 23.06 25.90 20.09 25.65 27.76 29.14 22.88 31.40
     22.79 23.68 22.15 21.50 22.40 24.39 20.34 17.53 25.59 20.25 20.76 23.08
     20.66 20.47
y
     27.60 24.02 33.97 27.84 33.12 37.32 37.53 38.95 29.80 32.72 30.04 26.45
     26.34 32.82 28.91 29.37 32.39 29.43 37.83 24.46 37.82 32.19 34.51 32.64
     30.44 38.70 29.84 29.35 32.78 34.01 36.24 41.86 35.96 35.57 33.84 27.69
     29.32 41.71 34.08 27.64 33.06 39.98 36.62 29.72 33.51 31.49 33.51 33.24
     25.02 39.78 31.96 37.69 44.01 29.07 32.94 30.47 33.33 24.34 35.99 32.25
     36.51 33.47 35.37 31.82 38.49 25.67 29.36 36.64 24.14 39.54
```

Se pide:

1. *Representar en dos gráficos distintos los estimadores kernel de ambas densidades.*

2. *Repetir el apartado anterior pero en la misma gráfica.*

3. *Representar las funciones de distribución de la primera muestra. Haced lo mismo con la función de distribución de la segunda muestra.*

4. *Representad las dos funciones de distribución en un mismo gráfico.*

2.6 Un dibujo

En esta sección vamos a realizar un dibujo a partir de unos datos. El objetivo es mostrar cómo estudiar la relación entre dos variables, cómo hacerlo con R mostrando las dificultades con las que nos encontramos. Los datos que vamos a utilizar son las temperaturas mínimas y máximas por mes en Alicante desde el año 1939 hasta el año 2010. Leemos los datos y adjuntamos para poder usar los nombres de las variables.

```
x = read.table("../data/alicante_temperaturas_anyo_1939_2010.txt")
attach(x)
```

Nos fijamos en el mes de enero. Las variables que nos dan la temperatura mínima y máxima (en décimas de grado) en enero son *tmin1* y *tmax1* respectivamente. La variable *anyo* nos indica el año.

Parece natural estudiar la variación de la temperatura mínima a lo largo de los años. En la figura 2.14(a) representamos en abscisas el año y en ordenadas la temperatura mínima en enero. Parece lo natural para ver la evolución temporal.

```
plot(anyo,tmin1)
```

¿Es correcto este dibujo? Por defecto la función *plot* utiliza puntos para representar. Sin embargo, en este caso para seguir la evolución es parece preferible unir los puntos consecutivos mediante segmentos. En la figura 2.14(b) tenemos el código y el resultado.

```
plot(anyo,tmin1,type = "l")
```

En nuestros datos también nos aparece la temperatura máxima en enero. ¿Por qué no representarla junto con la mínima? De este modo podemos ver para cada mes la variación. Para superponer las máximas utilizamos la función *lines*. Hace lo mismo que el plot anterior pero conserva el dibujo anterior. En la figura 2.14(c) vemos el resultado.

```
plot(anyo,tmin1,type = "l")
lines(anyo,tmax1)
```

No hay error, la figura no ha cambiado. Pero los nuevos datos los hemos representado. El problema lo tenemos en que el eje de ordenadas se ha escalado para las mínimas y no llega a los valores que corresponden a las máximas. Hemos de indicarle una nueva escala al eje de ordenadas, al eje y, de modo que aparezcan todos los datos. Por ejemplo, supongamos que damos una escala que tiene un mínimo de 35 y un máximo de 200 (en décimas de grado). Esto lo indicamos con `ylim = c(35,200)`. En figura 2.14(d) tenemos el resultado. Vamos bien. Podemos observar que para el año 1941 no tenemos el dato y por ello no aparece el punto correspondiente.

```
plot(anyo,tmin1,type = "l",ylim=c(35,200))
lines(anyo,tmax1)
```

En la figura 2.14(d) vemos que las etiquetas que utiliza para las abscisas y ordenadas no son muy adecuadas. Usa el nombre de las

variables que utilizamos en el *plot*. Vamos a cambiarlas indicando para abscisas la etiqueta "año" y para ordenadas "Temperatura". Esto lo indicamos con los argumentos xlab e ylab. En la figura 2.14(e) tenemos cómo hacerlo.

```
plot(anyo,tmin1,type = "l",ylim=c(35,200),xlab="Año",ylab="Temperatura")
lines(anyo,tmax1)
```

Nos va quedando bien el dibujo. Vamos a mejorarlo. Podemos indicar con dos líneas horizontales las medias de la mínima y la media de la máxima. Para ello utilizamos la función abline. Primero hemos de calcular la media de la mínima con

```
mean(tmin1)

## [1] NA
```

Tenemos un problema. Nos da que no lo puede calcular porque le falta el dato para 1941. Le indicamos que nos calcule la media de los demás.

```
mean(tmin1,na.rm=T)

## [1] 62.39437
```

También calculamos la media de las temperaturas máximas.

```
mean(tmax1,na.rm=T)

## [1] 167.0563
```

En la figura 2.14(f) tenemos la figura añadiendo las dos líneas horizontales.

```
plot(anyo,tmin1,type = "l",ylim=c(35,200),xlab="Año",
     ylab="Temperatura")
lines(anyo,tmax1)
abline(h = mean(tmin1,na.rm=T))
abline(h = mean(tmax1,na.rm=T))
```

Nos ha quedado un dibujo apañado. Contentos con ello decidimos guardarlo en un fichero para luego insertarlo en algún documento. Supongamos que nos queremos guardarlo como una imagen jpeg y queremos que el fichero se llame *tmin-max-Alicante-1939-2010.jpg*. Esto lo hacemos con

```
jpeg("tmin-max-Alicante-1939-2010.jpg")
plot(anyo,tmin1,type = "l",ylim=c(35,200),xlab="Año",ylab="Temperatura")
lines(anyo,tmax1)
abline(h = mean(tmin1,na.rm=T))
abline(h = mean(tmax1,na.rm=T))
dev.off()

## pdf
##   2
```

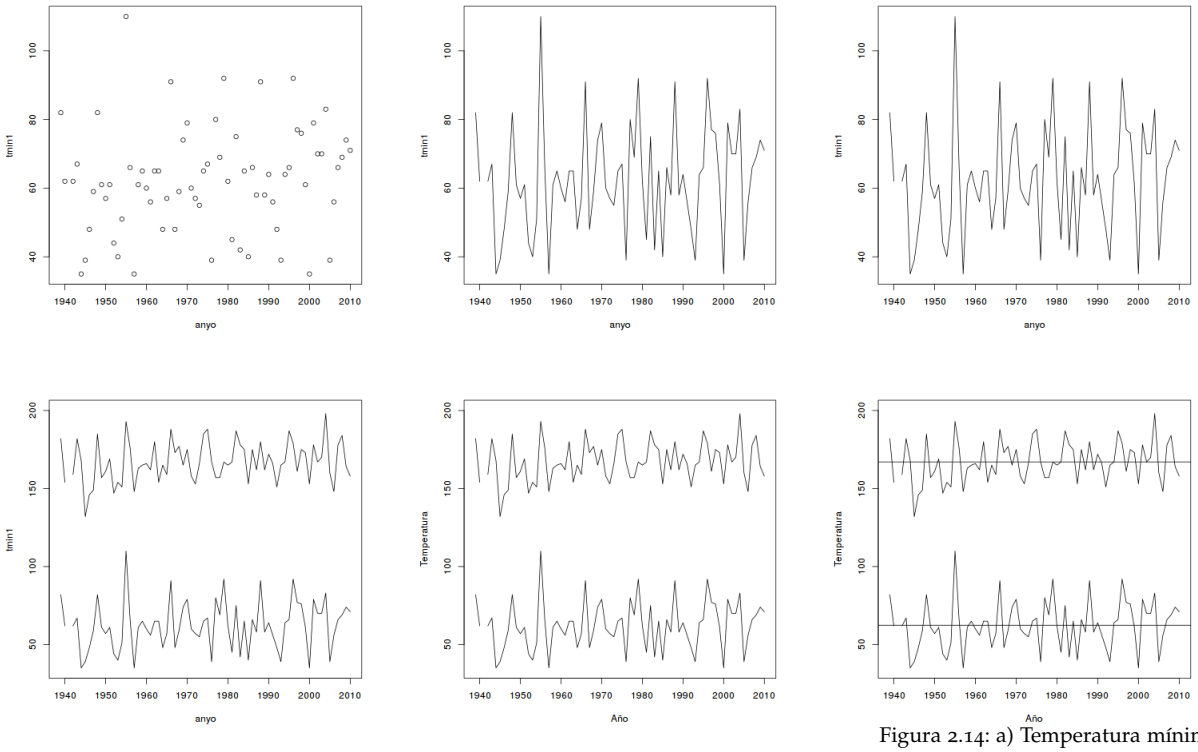


Figura 2.14: a) Temperatura mínima en enero en Alicante desde el año 1939 hasta el año 2010. b) Unimos los puntos. c) La máxima no se ve, ¿por qué? d) Temperatura mínima y máxima en enero en Alicante desde el año 1939 hasta el año 2010. Hemos modificado la escala en ordenadas y podemos ver las temperaturas máximas. e) Respecto de la figura d hemos modificado las etiquetas de los ejes. Respecto de la figura d hemos modificado las etiquetas de los ejes.

3
Probabilidad

3.1 Experimento, suceso y probabilidad

Nos movemos constantemente en situaciones en que no podemos predecir qué va a ocurrir.[1] Realizamos un viaje en coche entre dos ciudades dadas. El tiempo que tardamos en realizar el viaje depende de muchos factores que hacen incierto el resultado. Nos sometemos a una operación quirúrgica. El resultado también es incierto. Quizás los ejemplos más simples y, por ello, los que se suelen utilizar están asociados con juegos de azar. Lanzamos un dado, el resultado que obtenemos no lo podemos predecir. Y el número de ejemplos es infinito. Nos encontramos en situaciones en donde no podemos predecir lo que va a ocurrir.

[1] Lo cual no es de nuestro agrado. Pero la vida es así.

Experimento Estas situaciones son lo que llamamos *experimentos*. Dado un conjunto de condiciones no podemos predecir el resultado. Los experimentos más simples de analizar suelen estar ligados a juegos de azar.

1. Un experimento puede ser lanzar una moneda en donde no podemos predecir si el resultado va a ser que sale una cara o bien que sale una cruz.
2. El experimento puede ser lanzar un dado. Aquí tenemos seis posibles resultados (cada una de las caras del dado). También es un experimento con resultado incierto si jugamos a la lotería. Aquí el resultado será el número que sale premiado en un sorteo.

Espacio muestral En estos ejemplos sabemos qué cosas pueden pasar, el conjunto de posibles resultados al que se le llama *espacio muestral* y se suele denotar con la letra Ω. Sin embargo, cuando realizamos el experimento no sabemos exactamente *qué* va a pasar. No sabemos qué resultado entre los posibles se va a producir exactamente.

En el experimento consistente en lanzar una moneda tenemos dos posibles resultados: el resultado cara y el resultado cruz. Por ello podemos denotar el espacio muestral como

$$\Omega = \{cara, cruz\}.$$

Si lanzamos el dado podríamos representar el espacio muestral como

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

donde el número 2 significa que cuando lanzamos el dado *sale el número 2* y así sucesivamente.

En una lotería en la que se juegan 15000 números los posibles resultados son cada uno de estos números. Por tanto, el espacio muestral sería

$$\Omega = \{1, 2, \ldots, 15000\}.$$

Suceso aleatorio Se llama suceso aleatorio a cualquier conjunto de resultados. Por ejemplo, si lanzamos un dado, un suceso puede ser saber si obtenemos un número par. Este suceso es el conjunto formado por tres resultados. Lo representamos como

$$A = \{2, 4, 6\}.$$

Nos puede interesar saber si el número que obtenemos es un valor mayor o igual a 4. Este es el suceso

$$A = \{4, 5, 6\}.$$

Dado un suceso que nos interesa nuestro problema es saber cómo de probable es que se produzca. Y conocer esto antes de realizar el experimento.

Probabilidad Consideremos el experimento en que lanzamos una moneda. Conocemos el conjunto de resultados que se pueden producir, el espacio muestral: $\Omega = \{1, 2, \ldots, 6\}$. Nos planteamos la siguiente pregunta: ¿Qué probabilidad tengo de obtener un número par? Si el dado está bien construido no esperamos que un resultado ocurra con más frecuencia que otro. No esperamos que un resultado *sea más probable que otro*. Estamos usando la palabra *probabilidad* en su uso del lenguaje común. El término se usa de un modo impreciso pero comprensible por todos. Diríamos que los resultados son equiprobables o que tienen la misma probabilidad. En un experimento no queremos saber qué va a pasar. No tiene mucho sentido plantearnos esto porque es aleatorio sino que nos conformamos con saber la *probabilidad de lo que puede pasar*. En el

ejemplo del dado parece razonable decir que la probabilidad de un número par sería

$$P(\text{Obtenemos un número par}) = P(\{2,4,6\}) = \frac{3}{6}.$$

¿Por qué? Si todos los resultados son equiprobables entonces la probabilidad parece razonable considerar que sería el número de resultados que corresponden con *salir un número par* dividido por el número de resultados que se pueden producir. En resumen, el número de resultados que son *favorables* a lo que me interesa (que salga un número par) dividido por el número total de resultados. En el mismo experimento: ¿qué probabilidad tenemos de un número mayor o igual a 4?

$$P(\text{Mayor o igual a } 4) = P(\{5,5,6\}) = \frac{3}{6}.$$

¿Qué probabilidad tenemos de obtener una cara al lanzar una moneda? Por el mismo razonamiento:

$$P(\text{Cara}) = P(\{Cara\}) = \frac{1}{2}.$$

Un resultado es favorable y dos resultados son posibles cuando lanzamos la moneda.

Tenemos en R una función que nos sirve para elegir al azar de un grupo de elementos previamente definidos y de un modo equiprobable entre los posibles resultados del experimento. Es la función *sample*.

Nota de R 3.1 (Lanzamos una moneda muchas veces)**.** *Veamos cómo lanzar una moneda con R. Le decimos cuál es el espacio muestral*

```
Omega = c("cara","cruz")
```

y luego elegimos uno al azar en el espacio muestral.

```
sample(Omega,1)

## [1] "cruz"
```

Volvamos a lanzar la moneda.

```
sample(Omega,1)

## [1] "cruz"
```

Y una tercera vez.

```
sample(Omega,1)

## [1] "cruz"
```

De continuar iríamos obteniendo una serie de resultados cara o cruz. Lancemos 30 veces la moneda y veamos qué pasa.

```
sample(Omega,30,replace=TRUE)

##  [1] "cruz" "cara" "cara" "cara" "cruz" "cruz" "cruz" "cruz"
##  [9] "cara" "cruz" "cruz" "cruz" "cara" "cara" "cara" "cara"
## [17] "cara" "cara" "cruz" "cruz" "cara" "cruz" "cara" "cara"
## [25] "cruz" "cara" "cruz" "cara" "cara" "cruz"
```

Y otras 30 veces.

```
sample(Omega,30,replace=TRUE)

##  [1] "cara" "cara" "cara" "cara" "cara" "cruz" "cara" "cruz"
##  [9] "cruz" "cruz" "cara" "cruz" "cara" "cruz" "cara" "cruz"
## [17] "cruz" "cara" "cruz" "cruz" "cara" "cara" "cara" "cruz"
## [25] "cruz" "cruz" "cara" "cruz" "cara" "cara"
```

Podemos contar cuántas veces nos ha salido cara y cruz (el que quiera puede hacerlo manualmente).

```
x = sample(Omega,30,replace=TRUE)
table(x)

## x
## cara cruz
##   16   14
```

Si dividimos por el total de lanzamientos tenemos la frecuencia relativa de veces que nos ha salido cada uno de los dos posibles resultados. En nuestro caso tenemos las siguientes frecuencias relativas observadas:

```
table(x) / 30

## x
##      cara      cruz
## 0.5333333 0.4666667
```

Vamos a lanzar 100 veces la moneda y calculamos las frecuencias relativas.

```
x = sample(Omega,100,replace=TRUE)
table(x) / 100

## x
## cara cruz
## 0.52 0.48
```

¿Y por qué no lanzar la moneda 1000 veces?

```
x = sample(Omega,1000,replace=TRUE)
table(x) / 1000

## x
##  cara  cruz
## 0.522 0.478
```

Y para acabar con la experiencia vamos a lanzarla 100000 veces.

```
x = sample(Omega,100000,replace=TRUE)
table(x) / 100000

## x
##    cara    cruz
## 0.49986 0.50014
```

Es claro que conforme repetimos el experimento y observamos la frecuencia relativa de veces que ocurre cada resultado nos acercamos cada vez más al valor 0,5 *para cada uno de los posibles resultados.*

En la figura 3.1 representamos en el eje de abscisas el número de veces que lanzamos la moneda y en ordenadas la frecuencia relativa de veces que se ha observado cara. Podemos ver cómo las frecuencias relativas de aparición de cara van apróximandose al valor 0,5.

Supongamos que nos fijamos en la frecuencia relativa de aparición de las cruces. En la figura 3.2 representamos en abscisas el número de lanzamientos y en ordenadas la frecuencia relativa de aparición de cruces. Vemos cómo se estabiliza la frecuencia alrededor del valor 0,5.

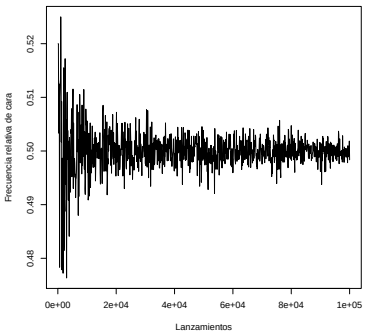


Figura 3.1: Frecuencias relativas de aparición de cara en sucesivos lanzamientos de una moneda correcta.

Nota de R 3.2 (Lanzamiento de un dado). *¿Cómo lanzamos un dado con R? Pues la función* sample *es adecuada. Empezamos definiendo el espacio muestral.*

```
(Omega = 1:6)

## [1] 1 2 3 4 5 6
```

Y ahora lanzamos el dado.

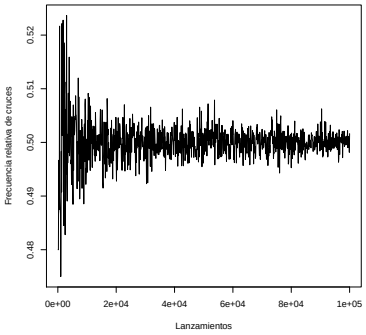


Figura 3.2: Frecuencias relativas de aparición de cruz en sucesivos lanzamientos de una moneda correcta.

```
sample(Omega,1)

## [1] 3
```

O bien lo lanzamos 20 veces.

```
sample(Omega,20,replace=TRUE)

##  [1] 6 1 4 1 5 4 4 1 3 4 3 6 5 2 1 1 5 6 5 3
```

Esperamos que cuando lo lanzamos un gran número de veces la frecuencia de veces que ocurre cada resultado se aproxime a $\frac{1}{6}$.

```
x = sample(Omega,1000,replace=TRUE)
table(x) / 1000

## x
##     1     2     3     4     5     6
## 0.176 0.168 0.167 0.176 0.147 0.166
```

Como así podemos comprobar.

En este tipo de experimentos con resultados equiprobables la probabilidad de cualquier suceso A viene dada por el cociente entre el número de resultados que tiene A (a los que se suele llamar *casos favorables* a la ocurrencia de A) y el número total de resultados o *casos posibles*.

Si denotamos el cardinal o número de elementos de A con $|A|$ entonces, cuando los resultados son equiprobables, podemos definir la probabilidad del suceso A como

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Casos favorables a que ocurra A}}{\text{Casos posibles}}. \tag{3.1}$$

Se suele conocer como la definión de probabilidad de Laplace. Describe con precisión situaciones como las descritas asociadas a algunos juegos de azar y algún otro caso de interés pero, desde luego, no todos los posibles experimentos en que estamos interesados.

Observemos que, en la definición de Laplace, si tomamos un suceso formado por un solo resultado entonces

$$P(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{\text{Casos posibles}}. \tag{3.2}$$

La probabilidad de cara es $1/2$. La probabilidad de cruz también es $1/2$. La probabilidad de un seis cuando lanzamos un dado es $1/6$. También lo es la de obtener un 5. O un 4. Finalmente vemos que

$$P(A) = \sum_{\omega \in A} \frac{1}{|\Omega|}.$$

3.1.1 Contando: variaciones, permutaciones y combinaciones

En experimentos con resultados equiprobables hay que contar el número total de resultados y el número de casos que contiene el suceso aleatorio de interés, esto es, hemos de contar casos favorables y casos posibles.

Para ellos es fundamental un breve repaso de combinatoria: variaciones, permutaciones y combinaciones.

Variaciones sin repetición Supongamos que tenemos un conjunto de n elementos (por ejemplo, el conjunto $\{1, \ldots, n\}$ y pretendemos saber cuántas secuencias ordenadas de k elementos (con $k < n$) podemos formar. Para entender el problema supongamos que $n = 3$ y $k = 2$. Entonces las posibles secuencias serían

```
12 21 13 31  23 32
```

Tenemos seis secuencias porque consideramos distintas 12 y 21. ¿Cómo contarlas sin enumerarlas todas ellas? El razonamiento es sencillo. Para la primera posición tenemos 3 posibilidades. Una vez elegido el número que ocupa la primera posición nos quedan 2 posibilidades para el segundo. En total 3×2.

En general, si consideramos n y k tenemos n posibilidades para el primero. Dada la elección nos quedan $(n-1)$ elecciones para el segundo. Dadas las dos primeras elecciones nos quedan $(n-2)$ elecciones para el tercero. Para la última posición nos quedarán $(n-k+1)$ posibles elecciones. En total tendremos

$$n \times (n-1) \times \ldots \times (n-k+1).$$

Esto recibe el nombre de **variaciones sin repetición**.

Permutaciones ¿De cuántas maneras podemos ordenar n elementos distintos? Podemos seguir con el razonamiento del párrafo anterior. Cuando ordenamos n elementos tenemos que elegir el primero de ellos, con n posibles elementos. Una vez tenemos el primero, para el segundo tenemos $n-1$ y así sucesivamente. Podemos ver que tenemos variaciones sin repetición donde $n = k$. Por tanto el número total de **permutaciones** es

$$n! = n \times (n-1) \times \ldots \times 1.$$

Combinaciones ¿Cuántos conjuntos distintos de k elementos podemos formar con un total de n? Estamos en una situación similar a las variaciones sin repetición salvo que no queremos que intervenga el orden. Por ejemplo, las secuencias

```
12 21
```

son distintas como secuencias. Sin embargo, los conjuntos

```
{1,2} {2,1}
```

son el mismo. Una vez hemos elegido los elementos de un conjunto (por ejemplo, los elementos 1 y 2) hemos de plantearnos cuántas secuencias ordenadas podemos formar, 2! en el ejemplo. En resumen, de un conjunto de k elementos podemos formar $k!$ secuencias (ordenadas) distintas. Tenemos n elementos. Con ellos podemos formar $n \times (n-1) \times \ldots \times (n-k+1)$ secuencias ordenadas. Pero cada $k!$ de estas secuencias tenemos los mismos elementos. En resumen, el número de conjuntos distintos de k elementos que podemos formar con n elementos distintos es

$$\binom{n}{k} = \frac{n \times (n-1) \times \ldots \times (n-k+1)}{k!}.$$

Fácilmente podemos comprobar que

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

El número $\binom{n}{k}$ recibe el nombre de **combinaciones de n elementos tomados de k en k**.

Nota de R 3.3 (Factorial y combinaciones)**.** *Dado un valor de n, por ejemplo, $n = 10$ podemos obtener el valor de* 10! *con*

```
factorial(10)

## [1] 3628800
```

Las combinaciones de n elementos tomados de k en k para $n = 10$ y $k = 5$ las obtenemos con

```
choose(10,5)

## [1] 252
```

Obviamente estamos eligiendo *5 elementos de un total de 10. De ahí el nombre de la función.*

Ejemplo 3.1 (Póquer)**.** *Supongamos el póquer cerrado sin comodines. Nos planteamos la probabilidad de que nos sirvan en una mano exactamente una pareja. Estamos en una situación de resultados equiprobables. Los resultados posibles son todos los posibles subconjuntos de 5 elementos del total de 52 cartas. Por tanto, el número de manos distintas será*

$$\binom{52}{5}.$$

Contemos ahora el número de manos que contienen exactamente una pareja (y no dos parejas o un trío). Vamos contando. Primero elegimos el número del cual formamos la pareja. Tenemos 13 posibilidades. Una vez elegido el palo tenemos cuatro cartas con el mismo número, por tanto, $\binom{4}{2}$ posibles parejas con ese número. Ahora hemos de elegir los otros tres números que nos aparecen en la mano. Tenemos 12 número disponibles (quitamos el que hemos utilizado para formar la pareja) y elegimos tres números con un total de $\binom{12}{3}$ posibilidades. Pero una vez elegidos los números tenemos cuatro cartas de cada número. Por lo tanto, por cada elección de números tenemos 4^3. En total como casos favorables nos encontramos con $13\binom{4}{2}\binom{12}{3}4^3$ casos favorables. La probabilidad buscada es

$$\frac{13\binom{4}{2}\binom{12}{3}4^3}{\binom{52}{5}}.$$

La podemos calcular con R.

```
(casosfavorables = 13 * choose(4,2) * choose(12,3) * 4^3)

## [1] 1098240

(casosposibles = choose(52,5))

## [1] 2598960

casosfavorables / casosposibles

## [1] 0.422569
```

Tenemos una probabilidad de 2.366.

Esto es como se debe de hacer. Y ahora como no se debe de hacer. Vamos a jugar con R. Empezamos definido las cartas que tenemos

```
(cartas = rep(1:13,4))

##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13  1  2  3  4  5
## [19]  6  7  8  9 10 11 12 13  1  2  3  4  5  6  7  8  9 10
## [37] 11 12 13  1  2  3  4  5  6  7  8  9 10 11 12 13
```

Y ahora extraemos al azar una mano.

```
(mano = sample(cartas,5))

## [1] 11 10 12  8 11
```

¿Cómo sé que tenemos una pareja? Una forma sencilla es contar la frecuencia de cada número.

```
(conteosmano = table(mano))

## mano
##  8 10 11 12
##  1  1  2  1
```

Y ver cuántos conteos me devuelve.

```
length(conteosmano)

## [1] 4
```

Si me devuelve 5 quiere decir que no se repite ningún número. Si me devuelve 4 quiere decir que hay una pareja exactamente. Contemos pues el número de veces que se produce esta situación y repitamos la experiencia muchas veces. La frecuencia relativa de éxitos nos ha de dar la probabilidad. Un poco de código más complicado de R.

```
nsimulaciones = 1000
exitos = 0
for(i in 1:nsimulaciones){
  mano = sample(cartas,5)
  conteosmano = table(mano)
  if(length(conteosmano) == 4) exitos = exitos  + 1
}
exitos / nsimulaciones

## [1] 0.416
```

3.1.2 *Un poco de teoría*

[2]

En lo anterior hemos visto ejemplos de experimentos aleatorios donde todos los resultados son *equiprobables* y por ello la probabilidad de cada suceso no era más que el número de elementos que contiene (casos favorables) dividido por el total de resultados posibles (casos posibles). ¿Este es el único caso que nos encontramos de experimento aleatorio? Desde luego que no. De hecho, una probabilidad es cualquier función que a los sucesos les da valores entre 0 y 1 verificando algunos principios (o axiomas) razonables.

[2] Esta sección incluye algo de teoría. Su lectura no es muy necesaria para seguir el curso aunque sí que es conveniente. No hay nada tan práctico como la teoría.

Definición 3.1 (Probabilidad)**.** *Una función de conjunto, P, definida sobre los sucesos es una probabilidad si verifica*

1. $P(A) \geq 0$ *para todo suceso* A.
2. $P(\Omega) = 1$.

3. *P es aditiva, es decir, si* $\{A_i\}_{i=1,\ldots,n}$ *son sucesos disjuntos entonces*

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i).$$

A partir de la definición anterior se deducen algunas propiedades muy útiles.

1. La probabilidad del vacío es cero: $P(\emptyset) = 0$.
2. Si tenemos sucesos tales que $A \subset B$ entonces $P(A) \leq P(B)$.
3. Si los sucesos $A_1, \ldots, A_n$ no son disjuntos entonces

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n) \quad (3.3)$$

En particular si tenemos dos sucesos

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2). \quad (3.4)$$

4. A partir del punto anterior es inmediato que

$$P(A^c) = 1 - P(A). \quad (3.5)$$

5. Dados los sucesos $A_1, \ldots, A_n$, la relación existente entre la probabilidad de la unión de los A_i y la probabilidad de cada uno de ellos es la siguiente:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

3.2 *Ejercicios*

Ejercicio 3.1. *Seis personas se sientan a comer en un restaurante. Hay seis sillas alrededor de la mesa.*

1. *¿De cuántas formas distintas pueden sentarse?*
2. *Supongamos que las sillas están enfrentadas, tres en un lado de la mesa y otras tres al otro lado. Además las seis personas son tres parejas que han quedado a cenar. Siguiendo una antigua costumbre se sientan a un lado los hombres y al otro lado las mujeres. ¿De cuántas formas distintas se pueden sentar?*

Ejercicio 3.2. *Supongamos el póquer cerrado sin comodines. Calcular la probabilidad de obtener un póquer cuando nos dan una mano.*

Ejercicio 3.3. *Consideremos el experimento aleatorio consistente en lanzar dos veces un dado. Se pide:*

1. *¿Qué probabilidad tenemos de obtener dos veces el número 6?*

2. *¿Y de obtener el par de valores 1 y 5?*

3. *¿Qué probabilidad tenemos de que coincidan el primer y el segundo resultado, esto es, de que el primer lanzamiento sea un uno y el segundo también o de que el primer lanzamiento sea un dos y el segundo también, etc?*

4. *¿Qué probabilidad tenemos de que la suma de los dos valores sea 7? ¿Y de que la suma de los dos valores sea mayor o igual que 7? ¿Y de que sea mayor que 7?*

3.3 *Variable aleatoria*

Supongamos el experimento consistente en elegir al azar o aleatoriamente a una individuo de la Comunidad Valenciana. Obviamente el espacio muestral está formado por los distintos individuos. Si los numeramos tenemos $\Omega = \{\omega_i\}_{i=i}^{N}$ donde N es el número total de personas de la Comunidad. Elección al azar supone que cada individuo tiene la misma probabilidad de ser elegido. Por tanto tenemos

$$P(\{\omega_i\}) = \frac{1}{N}.$$

El resultado mismo no suele tener un interés especial para nosotros. Seleccionamos aleatoriamente estas personas porque tenemos interés en sus ingresos, en el número de personas que conviven con el o ella, si está casada o es soltera, su glucosa o tensión arterial. Son ejemplos de valores que nos pueden interesar del individuo seleccionado dependiendo de que estemos realizando un estudio de tipo socio económico o bien un estudio de salud. Por ello nuestro interés no está en el individuo ω que obtenemos cuando seleccionamos a una persona al azar sino que nuestro interés es un valor asociado a ω que denotamos por $X(\omega)$ y que puede ser su tensión arterial. Lo que es aleatorio es ω porque lo seleccionamos aleatoriamente. Una vez tenemos ω el valor $X(\omega) = x$ viene dado. Elegimos al azar a una persona, una vez elegida su edad no es aleatoria, es la que tiene. Elegimos al azar una muestra de agua en una planta de tratamiento de aguas residuales. Luego determinamos la demanda química de oxígeno. A la función que asocia a un resultado un valor numérico se le llama **variable**

aleatoria. Lo podemos denotar como

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = x. \end{aligned}$$

3.3.1 *Variable aleatoria discreta*

Es una variable aleatoria que toma un número finito de valores o bien toma un número infinito de posibles valores que podemos numerar. [3]

Un ejemplo muy simple y habitual es la edad de un individuo seleccionado al azar. La edad suele cuantificarse con valores enteros. Decimos que una persona tiene 24 años o que tiene 25 años. Por lo tanto es un valor entero. Si cuantificamos la edad en años entonces los valores posibles de la variable aleatoria son $0, 1, 2, \ldots$. En cualquier caso un número finito de posibles valores.

Otro ejemplo puede ser el número de organismos que observamos en una placa al microscopio, el número de huevos en una buitrera, el número de píxeles defectuosos en un monitor de ordenador, el número de árboles dentro de un quadrat en un muestreo espacial. Todos son ejemplos de variable aleatoria discreta. Como vemos habitualmente son valores que resultan de contar. Es lo más frecuente pero no siempre es así.

Supongamos que denotamos por $\mathbb{D} = \{x_1, x_2, \ldots, \}$ el conjunto de valores posibles para la variable. Las probabilidades que hemos de conocer son

$$P(X = x_i),$$

es decir, la probabilidad de que la variable tome cada uno de sus valores posibles. Estas probabilidades $P(X = x_i)$ con $i = 1, \ldots, n$ reciben el nombre de **función de probabilidad** de la variable aleatoria X. Cualquier otra probabilidad que nos interese la podemos obtener a partir de la función de probabilidad.

Por ejemplo, supongamos $\mathbb{D} = \{0, 1, \ldots, 10\}$ y la función de probabilidad aparece en la tabla 3.1.

x	0.000	1.000	2.000	3.000	4.000	5.000	6.000	7.000	8.000	9.000	10.000
$P(X = x)$	0.107	0.268	0.302	0.201	0.088	0.026	0.006	0.001	0.000	0.000	0.000

Cuadro 3.1: Función de probabilidad de una variable discreta. En la primera fila el valor y en la segunda la probabilidad de tomar este valor.

A partir de la función de probabilidad $P(X = x_i)$ podemos calcular cualquier otra probabilidad. Por ejemplo:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2),$$

[3] Para ser rigurosos una variable discreta puede tomar también un número infinito numerable de valores. En fin, detalles técnicos a los que no hay que prestar demasiado interés en un curso como este.

o bien,

$$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10).$$

También

$$P(4 \leq X \leq 7) = P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7).$$

$$P(4 < X \leq 7) = P(X = 5) + P(X = 6) + P(X = 7).$$

$$P(4 < X < 7) = P(X = 5) + P(X = 6).$$

De un modo genérico podemos escribir que

$$P(X \in A) = \sum_{x \in A} P(X = x),$$

siendo A cualquier conjunto (por ejemplo, un intervalo).

3.3.2 *Ejercicios*

Ejercicio 3.4. *Consideremos el experimento aleatorio consistente en lanzar dos veces un dado. Un resultado del experimento puede ser $\omega = (1,3)$ indicando que en primer lugar hemos obtenido un 1 y en el segundo lanzamiento hemos obtenido un 3. Consideramos la variable aleatoria que asocia al resultado obtenido la suma de los valores que obtenemos en el primer y en el segundo lanzamiento. Si $\omega = (i,j)$ entonces $X(\omega) = i + j$.*

1. *Indicar qué valores puede tomar la variable X.*
2. *Obtener la función de probabilidad de la variable X.*
3. *Obtener las probabilidades siguientes: $P(X \leq 1), P(X \leq 2), P(X > 2), P(X \leq 4), P(4 \leq X \leq 6), P(4 < X \leq 6), P(4 \leq X < 6)$.*

3.3.3 *Variable aleatoria continua*

Consideremos el siguiente experimento. Cogemos una muestra de panga vietnamita y medimos la concentración de mercurio en dicha muestra. El resultado ω es la muestra que hemos tomado. De esta muestra nos interesa solamente la concentración de mercurio. Nos interesa el valor asociado a la muestra y no la muestra misma. Por ello podemos definir $X(\omega) = x$ donde x es la concentración medida de mercurio. El valor aleatorio que observamos los denotamos por X, la variable aleatoria. Una vez hemos observado el valor, esto es, una vez se ha realizado la determinación de mercurio el valor ya no es aleatorio. Es un valor dado, por ejemplo, una concentración de 0,5 miligramos por kilogramo. Este valor ya no lo denotamos con la letra mayúscula sino con la letra en minúscula, x. ¿Qué nos interesa conocer sobre estos valores aleatorios? Posiblemente muchas cosas

pero lo más básico (y de lo cual se deduce cualquier otra) es conocer la probabilidad que tenemos de que el valor que observemos esté entre dos números. ¿Qué probabilidad tenemos de que la muestra que analicemos de un valor entre 0,3 y 0,6? ¿Qué probabilidad hay de observar un valor entre 0,4 y 0,8? O bien, si una cierta normativa afirma que un valor por encima de 0,5 no se permite entonces parece natural plantearse: ¿cuál es la probabilidad de observar una muestra por encima de este valor? O por el contrario: ¿con qué frecuencia observamos muestras que no alcance el valor 0,5? Las probabilidades que acabamos de indicar se denotan como: $P(0{,}3 \leq X \leq 0{,}6)$, $P(0{,}4 \leq X \leq 0{,}8)$, $P(X \geq 0{,}5)$ y $P(X \leq 0{,}5)$. Sin tener que referirnos a valores concretos podemos denotar de un modo genérico todos los casos anteriores como $P(a \leq X \leq b)$ donde a y b toman los valores que queramos. Cuando consideramos $P(0{,}3 \leq X \leq 0{,}6)$ estamos tomando $a = 0{,}3$ y $b = 0{,}6$. También $P(X \geq 0{,}5)$ tiene esta forma ya que estamos tomando $a0 = {,}5$ y $b = +\infty$. Obviamente, $P(X \leq 0{,}5)$ corresponde con $a = -\infty$ y $b = 0{,}5$.

En resumen, cuando trabajamos con una variable aleatoria lo fundamental es conocer las probabilidades $P(a \leq X \leq b)$ donde a y b son números reales o $a = -\infty$ o $b = +\infty$.

Cuando una variable es continua entonces la probabilidad anterior se puede calcular como

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

La función f recibe el nombre de **función de densidad** (de probabilidad) de la variable X.

De hecho se tiene que

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

Ejemplo 3.2 (Uniforme en el intervalo unitario)**.** *Una variable aleatoria uniforme en el intervalo* $[0, 1]$ *es un experimento que ya hemos visto. En las calculadoras suele haber una función conocida como* rand *que cuando la usamos nos devuelve un valor entre 0 y 1. La idea es que es imprevisible el valor y no esperamos que aparezca alrededor de nada. Simplemente dentro del intervalo. La función de densidad viene dada por*

$$f(x) = \begin{cases} 1 & si \quad 0 \leq x \leq 1 \\ 0 & si \quad x < 0 \text{ ó } x > 1. \end{cases}$$

La podemos ver representada en la figura 3.3.

Supongamos que tomamos $0 \leq a \leq b \leq 1$ *entonces*

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

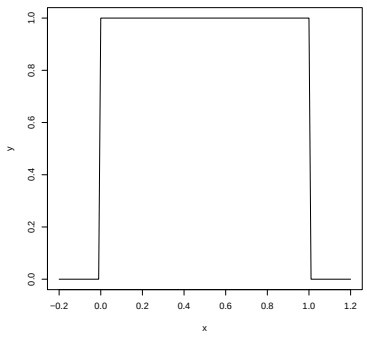


Figura 3.3: Función de densidad de una variable uniforme en el intervalo $[0, 1]$.

es decir, la probabilidad de que la variable está en el intervalo $[a,b]$ depende solamente de la longitud del intervalo y no de dónde está colocado dentro del intervalo $[0,1]$.

¿Podemos generar un valor aleatorio que se comporte como una uniforme en $[0,1]$? Simplemente con

```
runif(1,min=0,max=1)

## [1] 0.6736497
```

De hecho, podemos simular el número de puntos que queramos. Por ejemplo, generemos 20 valores.[4]

[4] Más no que ocupa demasiado espacio. En cualquier caso podemos probar a cambiar el valor 20 por el número de valores que nos apetezca simular.

```
runif(20,min=0,max=1)

##  [1] 0.4254709 0.1286097 0.8850760 0.6972555 0.3407570
##  [6] 0.2811757 0.7881495 0.3535969 0.3210780 0.8187089
## [11] 0.6592973 0.2631164 0.4046266 0.4530579 0.7039735
## [16] 0.2226044 0.6317559 0.4801175 0.4283014 0.0634963
```

Ejemplo 3.3. *También podemos considerar la uniforme en un intervalo arbitrario $[a,b]$ donde a y b son números arbitrarios siendo a menor que b. La función de densidad de la variable es*

$$f(x) = \begin{cases} \frac{1}{b-a} & si \quad a \leq x \leq b \\ 0 & en\ otro\ caso. \end{cases}$$

Supongamos $a = 2$ y $b = 5$. La función de densidad la podemos ver en la figura 3.4.

Además, si consideramos un par de puntos c y d tales que $a \leq c \leq d \leq b$ entonces:

$$P(c \leq X \leq d) = \int_c^d f(x)dx = \int_c^d \frac{1}{b-a}dx = \frac{d-c}{b-a}.$$

Otra vez la probabilidad de que el valor aleatorio de X esté en el intervalo $[c,d]$ solamente depende de lo largo que es el intervalo y no de dónde está dentro de $[a,b]$.

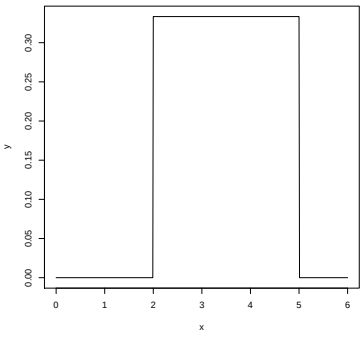


Figura 3.4: Función de densidad de una variable uniforme en el intervalo $[2,5]$.

Ejemplo 3.4 (Muestreo espacial)**.** *¿Cómo podemos obtener un punto al azar en una cierta región? En estudios medioambientales es frecuente que tengamos que elegir al azar un punto en una cierta zona. Por simplificar supongamos que queremos generar un punto en el cuadrado $[-1,1] \times [-1,1]$. ¿Cómo hacerlo? Lo más simple es generar la abscisa aleatoriamente según una uniforme en el intervalo $[-1,1]$, la ordenada según una uniforme en el intervalo $[-1,1]$. Si X e Y denotan las coordenadas aleatorias, entonces el punto aleatorio uniforme en dicho cuadrado será el punto (X,Y). En la figura 3.5 podemos ver el cuadrado y 50 puntos generados al azar.*

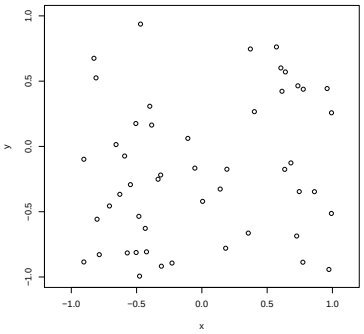


Figura 3.5: Generamos 50 puntos aleatoriamente sobre el cuadrado $[-1,1] \times [-1,1]$.

¿Cómo podemos generar puntos al azar sobre el círculo centrado en el origen de coordenadas y con radio 0,5*? Lo maś simple sería generar puntos en el cuadrado* $[-1,1] \times [-1,1]$ *y los que caen dentro del círculo los conservamos. Los demás los eliminamos. Los que conservamos son puntos* **uniformes** *sobre el círculo indicado. En la figura 3.6 aparecen estos puntos.*

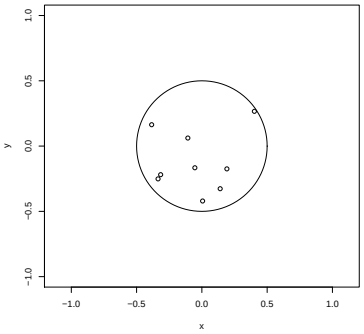


Figura 3.6: Generamos 50 puntos aleatoriamente sobre el cuadrado $[-1,1] \times [-1,1]$.

3.3.4 Ejercicios

Ejercicio 3.5. *Consideremos una variable aleatoria uniforme en el intervalo* $[0,1]$*. Se pide:*

1. *¿Qué probabilidad tenemos de que la variable sea menor o igual que* 0,5*? En otras palabras: ¿cuánto vale* $P(X \leq 0{,}5)$*?*
2. *¿Y* $P(X < 0{,}5)$*?*
3. *Calcular* $P(X \geq 0{,}5)$ *y* $P(X > 0{,}5)$*.*
4. *Determinar las siguientes probabilidades:* $P(0{,}6 < X \leq 0{,}9)$*,* $P(0{,}6 \leq X < 0{,}9)$ *y* $P(0{,}6 \leq X \leq 0{,}9)$*.*

Ejercicio 3.6. *Consideremos una variable aleatoria uniforme en el intervalo* $[0,8]$*. Se pide:*

1. *¿Qué probabilidad tenemos de que la variable sea menor o igual que* 0,5*? En otras palabras: ¿cuánto vale* $P(X \leq 0{,}5)$*?*
2. *¿Y* $P(X < 0{,}5)$*?*
3. *Calcular* $P(X \geq 0{,}5)$ *y* $P(X > 0{,}5)$*.*
4. *Determinar las siguientes probabilidades:* $P(0{,}6 < X \leq 0{,}9)$*,* $P(0{,}6 \leq X < 0{,}9)$ *y* $P(0{,}6 \leq X \leq 0{,}9)$*.*

3.3.5 Función de distribución

Dada una variable aleatoria X se define la función de distribución (o función de distribución acumulada) como la función que para valor real x nos da la probabilidad de que la variable sea menor o igual que este valor, es decir,

$$F_X(x) = F(x) = P(X \leq x). \tag{3.6}$$

Obviamente si la variable que tenemos es discreta y toma valores en $\{x_1, x_2, \ldots\}$ entonces

$$F_X(x) = \sum_{x_i : x_i \leq x} P(X = x_i),$$

es decir, sumamos la probabilidad de que la variable tome cada uno de los valores que puede tomar y que son menores o iguales al valor x.

Ejemplo 3.5. *Consideramos la variable aleatoria discreta tal que su función de probabilidad aparece en la tabla 3.1. Vamos a determinar la función de distribución. Las probabilidades que aparecen en la tabla 3.1 aparecen representadas en la figura* **??**. *La función de distribución la hemos representada en la figura 3.8.*

En el caso de una variable que sea continua se tiene la igualdad

$$F_X(x) = \int_{-\infty}^{x} f(t)dt.$$

En general se tiene la siguiente importante igualdad.

$$P(a < X \leq b) = F(b) - F(a) \text{ con } a \leq b, a, b \in \mathbb{R}. \tag{3.7}$$

La igualdad afirma que la probabilidad de que la variable sea estrictamente mayor que a y menor o igual que b lo podemos obtener como la diferencia de la función de distribución en los extremos.

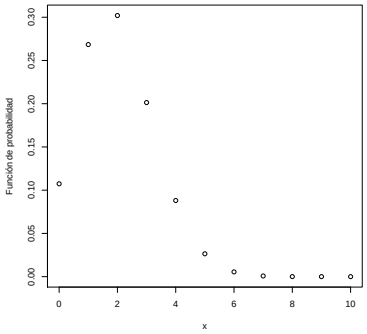


Figura 3.7: Probabilidades de la tabla 3.1.

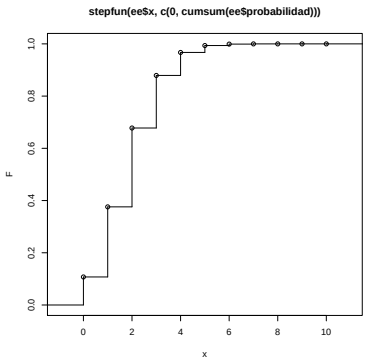


Figura 3.8: Función de distribución de la variable con función de probabilidad en la tabla 3.1.

3.3.6 Ejercicios

Ejercicio 3.7. *Consideremos el experimento aleatorio consistente en lanzar dos veces un dado. Un resultado del experimento puede ser $\omega = (1,3)$ indicando que en primer lugar hemos obtenido un 1 y en el segundo lanzamiento hemos obtenido un 3. Consideramos la variable aleatoria que asocia al resultado obtenido la suma de los valores que obtenemos en el primer y en el segundo lanzamiento. Si $\omega = (i,j)$ entonces $X(\omega) = i + j$. Se pide:*

1. *Determinar la función de distribución de la variable aleatoria X.*
2. *Representar de un modo manual la función de distribución que hemos determinado en el punto 1.*
3. *Representar la función de distribución utilizando la función* stepfun.

Ejercicio 3.8. *Supongamos una variable uniforme en el intervalo $[2,6]$ que denotamos como $X \sim U(2,6)$. Se pide:*

1. *Determinar la función de distribución de la variable aleatoria X.*
2. *Representar gráficamente la función de distribución de la variable aleatoria X.*

Ejercicio 3.9. *Supongamos una variable uniforme en el intervalo $[2,6]$ que denotamos como $X \sim U(2,6)$. Se pide:*

1. *Representar gráficamente la función de distribución utilizando las funciones* plot *y* punif.

3.4 Media y varianza

Una variable suele describirse de un modo simple mediante su media y su varianza. La media nos da una idea de alrededor de qué valor se producen los valores aleatorios de la variable mientras que la varianza cuantifica la dispersión de estos valores alrededor de la media.

3.4.1 Media de una variable aleatoria discreta

Ejemplo 3.6 (La media como límite de medias muestrales)**.** *Supongamos que tenemos una variable que puede tomar los valores* $\{0, 9\}$ *con probabilidades dadas en la tabla 3.2. El experimento supongamos que consiste en elegir al azar una vivienda en una gran población (por ejemplo, Valencia) y observar el número de personas que habitan la vivienda. En la fila etiquetada con x tenemos el número de personas y en la fila etiquetada* $P(X = x)$ *la frecuencia de cada valor posible que asumimos conocidas.*

x	0.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00
$P(X = x)$	0.20	0.11	0.13	0.24	0.27	0.02	0.01	0.01	0.00	0.01

Cuadro 3.2: Función de probabilidad de la variable aleatoria que nos da el número de personas que residen en una vivienda.

Podemos simular la selección aleatoria de una vivienda en esa población utilizando la función sample*.*

```
x = 0:9
probabilidades = c(0.20,0.11,0.13,0.24,0.27,0.02,0.015,0.009,
  0.0009,0.0051)
sample(x,size=1,replace=TRUE,prob=probabilidades)

## [1] 0
```

Supongamos que repetimos el proceso de extracción 100 veces.

```
n = 100
(y = sample(x,size=n,replace=TRUE,prob=probabilidades))

##   [1] 9 1 2 4 6 2 1 4 2 2 4 0 3 1 3 3 3 6 0 1 0 2 1 0 4 3 3
##  [28] 4 1 0 4 0 4 4 3 0 8 3 2 1 3 3 4 4 4 1 3 3 6 3 4 1 3 4
##  [55] 3 2 2 3 4 0 3 3 4 2 3 2 0 2 4 3 3 3 0 4 0 1 0 9 4 0 1
##  [82] 4 0 3 4 5 4 3 4 3 3 0 3 4 3 4 4 4 3 3
```

Las frecuencias que observamos de cada tipo son las siguientes

```
prop.table(table(y))

## y
```

```
##    0    1    2    3    4    5    6    8    9
## 0.15 0.11 0.11 0.30 0.26 0.01 0.03 0.01 0.02
```

¿Se parecen? En las categorías más probables bastante. Repitamos el proceso con 1000 muestras.

```
y = sample(x,size=1000,replace=TRUE,prob=probabilidades)
prop.table(table(y))

## y
##     0     1     2     3     4     5     6     7     8     9
## 0.192 0.098 0.140 0.253 0.265 0.023 0.012 0.010 0.001 0.006
```

Vemos cómo se parecen más las frecuencias observadas a las probabilidades de cada uno de los valores. Denotemos los valores simulados por $\{y_1, \ldots, y_n\}$. *Su media muestral será* $\bar{y}_n = \frac{\sum_{i=1}^n y_i}{n}$ *pero*

$$\bar{y}_n = \sum_{i=1}^{n} \frac{y_i}{n} = \sum_{x=0}^{n} x \frac{n_x}{n}$$

donde n_x *denota el número de veces que aparece el resultado* x. *Obviamente cuando el número de datos que generamos va creciendo la frecuencia relativa de veces que aparece el resultado* x *que viene dada por el cociente* n_x/n, *se va aproximando a la probabilidad* $P(X = x)$. *Por ello tendremos que*

$$\bar{y}_n = \sum_{x=0}^{n} x \frac{n_x}{n} \longrightarrow \sum_{x=0}^{n} xP(X = x).$$

En nuestro caso el número medio de personas que viven en la vivienda vendría dada por

```
(mu = sum(x * probabilidades))

## [1] 2.4761
```

En la figura 3.9 hemos representado con una línea horizontal cuya ordenada es la media que acabamos que calcular. Luego vamos simulando valores según la variable discreta que acabamos de proponer. En el eje de abscisas consideramos el número de valores que vamos promediando. Vemos cómo la media muestral se va aproximando al valor dado de la media.

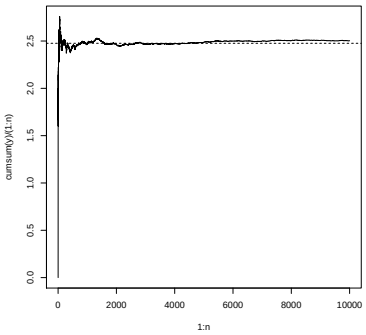


Figura 3.9: Medias muestrales del número de personas que habitan una vivienda en función del tamaño de la muestra. La línea horizontal punteada indica la media poblacional.

Definición 3.2. *Si* X *es una variable aleatoria discreta que toma los valores* $\{x_1, x_2, \ldots\}$ *entonces la media de* X *se define como*

$$EX = \mu_X = \mu = \sum_{i=1}^{+\infty} x_i P(X = x_i).$$

Ejemplo 3.7. *Lanzamos la moneda si sale cara la variable* X *vale uno y si sale cruz la variable* X *vale cero. ¿Cuál es su media?*

$$\mu = 1 \times p + 0 \times (1 - p) = p.$$

3.4.2 *Varianza y desviación típica*

Definición 3.3 (Varianza y desviación típica)**.** *Si X es una variable aleatoria discreta que toma los valores $\{x_1, x_2, \ldots\}$ entonces la varianza de X se define como*

$$var(X) = \sigma^2 = E(X-\mu)^2 = \sum_{i=1}^{+\infty}(x_i-\mu)^2 P(X=x_i).$$

Habitualmente además de la varianza se suele utilizar para medir variabilidad la desviación típica *dada por*

$$\sigma = \sqrt{var(X)} = \sqrt{\sum_{i=1}^{+\infty}(x_i-\mu)^2 P(X=x_i)}.$$

En variables continuas las definiciones de media y varianza son las análogas sustituyendo sumatorios por integrales.

Definición 3.4. *Si X es una variable continua con función de densidad f entonces la media de X se define como*

$$EX = \mu = \int_{-\infty}^{+\infty} x f(x) dx$$

Definición 3.5. *Si X es una variable continua con función de densidad f entonces la varianza de X se define como*

$$var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx,$$

mientras que la desviación típica de X se define como

$$\sigma_X = \sigma = \sqrt{\int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx}.$$

Hemos visto antes que si vamos observando valores de una variable entonces las sucesivas medias muestrales se van aproximando a la media poblacional. ¿Ocurre algo similar con la varianza y con la desviación típica? Por supuesto que sí.

3.5 *Variable binomial*

Consideremos un experimento con dos posibles resultados. A uno de ellos le llamamos éxito y al otro le llamamos fracaso. Suponemos que hay una probabilidad p de obtener un éxito (donde $0 \leq p \leq 1$) y, por lo tanto, una probabilidad $1-p$ de obtener un fracaso. Un experimento de este tipo recibe el nombre de *prueba de Bernoulli*.

El ejemplo más simple, lanzamiento de una moneda donde identificamos salir cara con éxito y salir cruz como fracaso. En este caso, la probabilidad de éxito es $p = 0{,}5$.

Otro ejemplo, elegimos al azar a una persona de la población española e identificamos éxito con que la persona tenga un cierto atributo (sea diabético por ejemplo) y fracaso con que no tenga el atributo. La probabilidad de éxito p coincide con la proporción que realmente hay de diabéticos en la población.

Repetimos n veces una prueba de Bernoulli independientemente una de otra (lanzamos n veces una moneda) y observamos la variable aletoria que nos da el número total de éxitos. Una variable de este tipo recibe el nombre de *variable binomial con n pruebas y una probabilidad de éxito p* y se suele denotar como

$$X \sim Bi(n, p).$$

Observemos que los valores que puede tomar esta variable son $0, 1, 2, \ldots, n$, esto es, desde cero éxitos hasta n éxitos. ¿Qué probabilidad tenemos de observar un número determinado de éxitos? Esta probabilidad, $P(X = x)$ es la función de probabilidad de la binomial y se prueba que tiene la siguiente expresión.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{3.8}$$

Si tenemos una variable binomial X con n pruebas y una probabilidad de éxito p, $X \sim Bi(n, p)$, entonces su media es

$$EX = \mu = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = np, \tag{3.9}$$

mientras que su varianza viene dada por

$$var(X) = \sigma^2 = \sum_{x=0}^{n} (x - np)^2 \binom{n}{x} p^x (1-p)^{n-x} = np(1-p). \tag{3.10}$$

Nota de R 3.4 (Función de probabilidad de la binomial)**.** *La función* dbinom *nos permite calcular la función de probabilidad de una variable binomial. Por ejemplo, ¿qué probabilidad tenemos de obtener 70 caras si lanzamos 123 veces una moneda. La respuesta es*

```
dbinom(70,size=123,prob=0.5)

## [1] 0.02230619
```

Nota de R 3.5 (Simulación de una variable binomial)**.** *Supongamos que queremos lanzar una moneda 30 veces con R.*[5] *Esto lo podemos hacer con*

[5] Algo no demasiado raro aunque pueda parecerlo.

```
rbinom(30,size=1,prob=.5)

##  [1] 0 1 1 0 1 0 0 0 0 0 0 1 0 0 1 1 1 1 0 0 0 0 1 0 0 0 1 0
## [29] 1 1
```

El uno corresponde con éxito (cara) y el cero con fracaso. Si repetimos el proceso posiblemente no obtengamos los mismos resultados.

```
rbinom(30,size=1,prob=.5)

##  [1] 0 1 0 1 1 0 0 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1
## [29] 0 0
```

Realmente nos interesa el número de éxitos y no el orden en que se producen. Esto lo podemos hacer con

```
rbinom(1,size=30,prob=.5)

## [1] 14
```

Si lo repetimos posiblemente no obtengamos el mismo número de unos.

```
rbinom(1,size=30,prob=.5)

## [1] 13
```

Supongamos que queremos simular 40 veces el experimento consistente en lanzar 30 veces una moneda y contamos en cada caso el número de unos.

```
rbinom(40,size=30,prob=.5)

##  [1] 13 14 18 20 17 14 17 15 19 14 16 21 18 13 18 15 14 16
## [19] 15 16 16 15 17 15 18 12 11 19 12  7 15 10 14 14 16 16
## [37] 15 14 10 11
```

Si la moneda no está bien construida y pretendemos que la cara tenga una probabilidad de 0,6, entonces repetimos el experimento anterior con

```
rbinom(40,size=30,prob=.6)

##  [1] 17 14 17 18 14 14 21 20 18 13 17 13 20 18 18 18 20 16
## [19] 17 19 17 21 17 13 17 22 20 22 23 15 15 17 20 17 19 17
## [37] 18 19 19 12
```

Nota de R 3.6 (De cómo calcular probabilidades de la distribución binomial)**.** *Supongamos que queremos calcular la probabilidad de obtener 23 éxitos cuando realizamos 30 pruebas de Bernoulli donde la probabilidad de éxito es 0,6, es decir, pretendemos calcular para $X \sim Bi(30, 0{,}6)$ la función de probabilidad en $x = 23$ dada por*

$$P(X = x) = \binom{30}{23}(0{,}6)^{30}(1 - 0{,}6)^{30-23}. \tag{3.11}$$

```
dbinom(23,size=30,prob=.6)

## [1] 0.02634109
```

Podemos conocer las probabilidades de cada uno de los posibles resultados, es decir, la función de probabilidad $P(X = x)$, con

```
dbinom(0:30,size=30,prob=.6)

##  [1] 1.152922e-12 5.188147e-11 1.128422e-09 1.579791e-08
##  [5] 1.599538e-07 1.247640e-06 7.797748e-06 4.010270e-05
##  [9] 1.729429e-04 6.341240e-04 1.997491e-03 5.447702e-03
## [13] 1.293829e-02 2.687184e-02 4.894513e-02 7.831221e-02
## [17] 1.101265e-01 1.360387e-01 1.473752e-01 1.396186e-01
## [21] 1.151854e-01 8.227527e-02 5.048710e-02 2.634109e-02
## [25] 1.152423e-02 4.148722e-03 1.196747e-03 2.659437e-04
## [29] 4.274096e-05 4.421478e-06 2.210739e-07
```

En la figura 3.10 tenemos la representación gráfica de estas probabilidades.

También podemos obtener la función de la distribución binomial en cualquier punto, es decir, la probabilidad $P(X \leq 12)$ es

```
pbinom(12,size=30,prob=.6)

## [1] 0.02123988
```

La figura 3.12 muestra la función de distribución de una variable aleatoria con distribución binomial con $n = 30$ pruebas y una probabilidad de éxito $p = 0{,}6$.

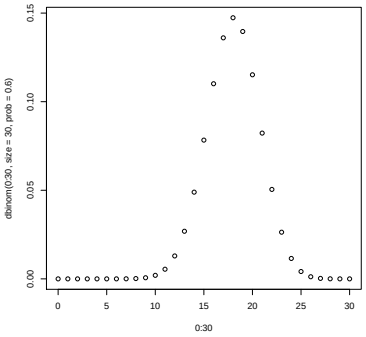


Figura 3.10: Para una variable binomial con $n = 30$ y una probabilidad de éxito de $p = 0{,}6$ mostramos la función de probabilidad que para cada x nos da la probabilidad de que la variable tome ese valor, $P(X = x)$.

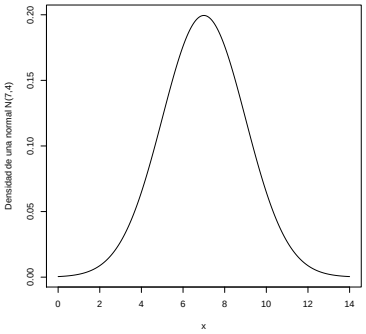


Figura 3.11: Función de densidad de una normal con media 7 y varianza 4.

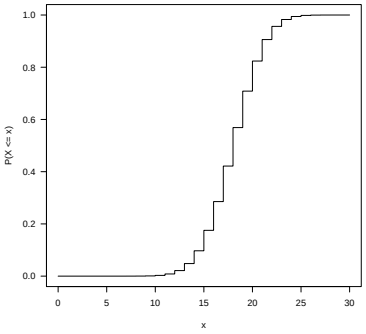


Figura 3.12: Función de distribución de una binomial con $n = 30$ pruebas y una probabilidad de éxito $p = 0{,}6$. Para cada abcisa x tenemos la probabilidad de que la variable sea menor o igual que ese valor x, $P(X \leq x)$.

3.5.1 Ejercicios

Ejercicio 3.10. *Se pide:*

1. *Simular 100 valores con distribución binomial con 20 pruebas y una probabilidad de éxito en cada prueba de 0,3. Guardar estos valores en el vector x.*
2. *Calcular la media y varianza muestrales de los valores generados.*
3. *Comparar la media muestral observada con $20 \times 0{,}3$ y la varianza muestral observada con $20 \times 0{,}3 \times 0{,}7$ que corresponden con la media y la varianza teóricas.*
4. *Repetir los apartados anteriores sustituyendo las 100 simulaciones por 1000, por 10000 y por 100000. Comparar en cada caso los valores teóricos con los valores muestrales.*

Ejercicio 3.11. *Generamos 10000 valores con distribución binomial con 20 pruebas y una probabilidad de éxito por prueba de* 0,3*. Comparar la función de distribución muestral (vista en la práctica anterior) con la función de distribución teórica que acabamos de ver.*

Ejercicio 3.12. *Consideremos una variable aleatoria con distribución binomial con 45 pruebas y una probabilidad de éxito de* 0,67*. Se pide:*

1. $P(X \leq 23)$.
2. $P(X < 23)$.
3. $P(X > 29)$.
4. $P(X \geq 29)$.
5. $P(34 < X \leq 45)$.
6. $P(34 \leq X \leq 45)$.

Ejercicio 3.13. *Dos especialistas en plagas vegetales difieren en su apreciación de la proporción de palmeras afectadas por el picudo en la Comunidad Valenciana. Uno de ellos (especialista A) afirma que un 30 % de las palmeras están afectadas. El otro, especialista B, afirma que es un 45 % de las palmeras. Un tercer especialista decide tomar una muestra aleatoria simple de la población total de palmeras. En total muestrea un total de 325 palmeras y observa 133 palmeras afectadas. Se pide:*

1. *Calcula la probabilidad de las afirmaciones de cada uno de los especialistas.*
2. *¿Qué afirmación es más probable? ¿Con cuál de los dos juicios nos quedaríamos?*
3. *Si es cierta la afirmación del especialista A: ¿qué probabilidad tenemos de observar 133 o menos?*
4. *Si es cierta la afirmación del especialista B: ¿qué probabilidad tenemos de observar 133 o menos?*
5. *Se decide continuar el muestreo y observamos el estado de 145 palmeras más de las cuales están afectadas 56.* Utilizando solamente la nueva muestra *responde a las preguntas 1, 2, 3 y 4.*
6. *Responder las preguntas 1, 2, 3 y 4 utilizando* conjuntamente *toda la muestra.*

3.6 Distribución normal

Definición 3.6 (Variable normal)**.** *Una variable aleatoria X se dice que sigue una distribución normal con media μ y varianza σ^2 (o, simplemente, que es una variable aleatoria normal) y se denota con $X \sim N(\mu, \sigma^2)$[6] si su función de densidad viene dada por*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{3.12}$$

De otro modo, si tomamos dos valores arbitrarios a y b con $a \leq b$ entonces

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

En esta sección asumimos siempre que la variable aleatoria sigue una distribución normal con media μ y con varianza σ^2, $X \sim N(\mu, \sigma^2)$.

En la figura 3.11 aparece un ejemplo de la función definida en 3.12. En concreto es una normal con media 7 y varianza 4.

Ya lo indicamos en la propia definición pero se puede demostrar que la media y varianza de una normal son μ y σ^2. En definitiva estamos afirmando que se verifican las siguientes ecuaciones.

$$\begin{aligned} \mu &= \int_{-\infty}^{+\infty} x f(x)dx, \\ \sigma^2 &= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx. \end{aligned}$$

Nota 3.1 (La normal estándar)**.** *Una variable aleatoria Z se dice que tiene una distribución normal estándar cuando su media es cero y su varianza es uno: $Z \sim N(0,1)$. Su función de densidad es*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{3.13}$$

La representación gráfica de esta densidad la tenemos en la figura 3.13.

Si Z es una normal estándar entonces la función de distribución, esto es, la función que para cada valor z nos da la probabilidad de que la variable sea menor o igual que este valor z es la siguiente

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx. \tag{3.14}$$

Dado un punto z el valor de la función $\Phi(z)$ nos da el área bajo la curva de la densidad normal entre $-\infty$ y el punto z. En la figura 3.14 hemos rayado en negro esta zona para $z = 1{,}3$

Hay tablas que nos proporcionan el valor de esta función para diferentes valores de z.[7] Esto era necesario cuando no teníamos herramientas informáticas. Ahora lo lógico es utilizar software. En concreto el valor de $\Phi(1{,}3)$

[6] También se denota con frecuencia $X \sim N(\mu, \sigma)$, es decir, se indica la media μ y la desviación típica o estándar σ.

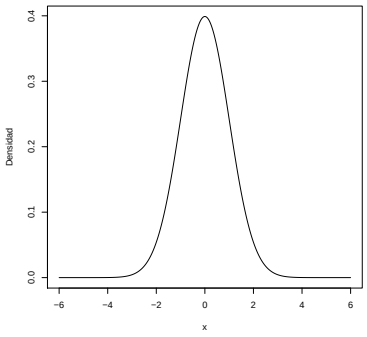


Figura 3.13: Función de densidad de una normal estándar o típica.

[7] Simplemente poniendo en Google "tablas de la normal" nos aparecen un montón de tablas. Cualquier libro de texto de hace unos años lleva al final del texto unas tablas de la normal.

(área de la zona rayada en negro en la figura 3.14 lo obtendríamos con R del siguiente modo.

```
pnorm(1.3)

## [1] 0.9031995
```

En la figura 3.15 tenemos representada la función Φ.

Nota 3.2 (Estandarización o tipificación)**.** *Si tenemos una variable aleatoria con distribución normal con media* μ *y varianza* σ^2 *entonces la variable aleatoria* $Z = \frac{X-\mu}{\sigma}$ *sigue una distribución normal con media 0 y con varianza 1, esto es, se verifica*

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1). \tag{3.15}$$

[8] *Esta transformación recibe el nombre de tipificación o estandarización de la variable aleatoria X.*

En la figura 3.16 mostramos la densidad de una variable X normal con media 7 y varianza 4 y la densidad de la variable tipificada Z.

Nota 3.3 (De cómo calculaban los antiguos las probabilidades con la normal)**.** *¿Qué problema nos planteamos? Suponemos que una cierta cantidad sigue una distribución normal con unos parámetros (media y varianza) dados. Nos planteamos cómo calcular la probabilidad de que la variable esté en un cierto intervalo. Por ejemplo, sabemos que el valor aleatorio que observamos sigue una distribución normal con media 56 y desviación típica 9. ¿Qué probabilidad tenemos de que la variable aleatoria tome un valor entre 60 y 63? Nos planteamos el valor de la siguiente probabilidad:* $P(60 \leq X \leq 63)$*. Esta probabilidad corresponde con la zona rayada de negro en la figura 3.17. Para calcular este área se aplicaban las siguientes igualdades donde* $Z = (X-56)/3$*,*

$$P(60 \leq X \leq 63) = P\left(\frac{60-56}{3} \leq \frac{X-56}{3} \leq \frac{63-56}{3}\right) = P\left(\frac{60-56}{3} \leq Z \leq \frac{63-56}{3}\right) = P\left(Z \leq \frac{63-56}{3}\right) - P\left(Z \leq \frac{60-56}{3}\right). \tag{3.16}$$

Pero la variable Z es una normal estándar por lo que

$$P\left(Z \leq \frac{63-56}{3}\right) - P\left(Z \leq \frac{60-56}{3}\right) = \Phi\left(\frac{63-56}{3}\right) - \Phi\left(\frac{60-56}{3}\right).$$

De un modo genérico lo que acabamos de indicar es que si $X \sim N(56,9)$ *entonces*

$$P(60 \leq X \leq 63) = \Phi\left(\frac{63-56}{3}\right) - \Phi\left(\frac{60-56}{3}\right),$$

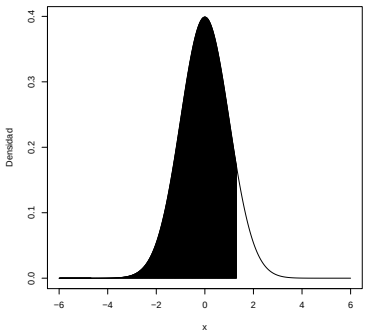


Figura 3.14: La función de distribución de la normal estándar en el punto 1,3 corresponde con la zona rayada.

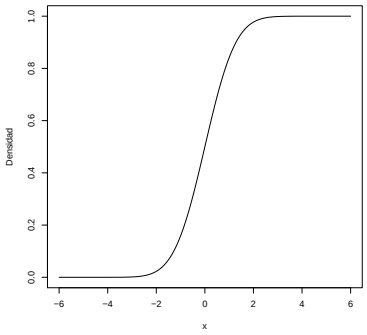


Figura 3.15: La función de distribución de la normal estándar.

[8] Esta afirmación es simplemente consecuencia de la siguiente igualdad (y perdón por la ecuación)

$$\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx.$$

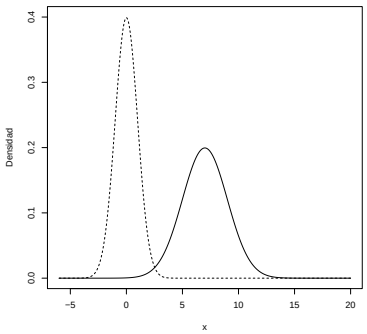


Figura 3.16: Función de densidad de una normal con media 7 y varianza 4 (trazo continuo) y de una normal típica con media 0 y varianza 1 (trazo discontinuo).

siendo Φ la función de ditribución de una normal estándar.

Si suponemos que $X \sim N(\mu, \sigma^2)$ y tomamos dos números a y b tales que $a \leq b$ entonces

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \tag{3.17}$$

Nota de R 3.7 (Calculando la función de densidad de una normal)**.** *Supongamos que $\mu = 16$ y $\sigma^2 = 4$. La función de densidad en un punto la podemos calcular con*

```
dnorm(14,mean= 16, sd= 2)

## [1] 0.1209854
```

o, en un conjunto de puntos, con

```
x0 = seq(10,22,1)
dnorm(x0,mean= 16, sd= 2)

##  [1] 0.002215924 0.008764150 0.026995483 0.064758798
##  [5] 0.120985362 0.176032663 0.199471140 0.176032663
##  [9] 0.120985362 0.064758798 0.026995483 0.008764150
## [13] 0.002215924
```

En la figura 3.18 aparecen tres densidades normales con parámetros distintos de modo que veamos el efecto de modificar la media y la varianza. En concreto se representan las densidades de las distribuciones normales $N(16,4)$, $N(24,4)$ y $N(16,9)$.

Nota de R 3.8. *Podemos generar valores aleatórios con distribución normal.*

```
rnorm(20,mean= 16, sd= 2)

##  [1] 16.88077 13.79176 18.19256 17.50622 15.09785 11.11469
##  [7] 15.22471 16.74930 19.32057 13.41720 15.99190 14.50691
## [13] 19.89402 14.96238 15.46071 15.50408 18.33124 14.07603
## [19] 12.90852 18.57567
```

La función de distribución de la variable X, es decir, $F(x) = P(X \leq x)$ la obtenemos con

```
pnorm(14,mean= 16, sd= 2)

## [1] 0.1586553
```

Podemos representar esta función (figura 3.19).

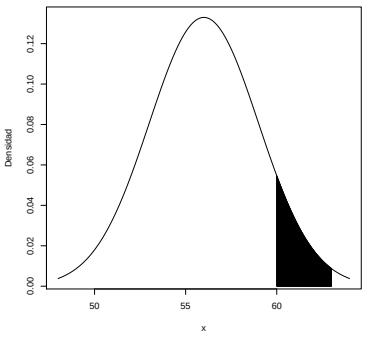


Figura 3.17: Densidad de una $N(56,9)$. El área de la zona rayada en negro corresponde a la probabilidad de que la variable esté entre 60 y 63.

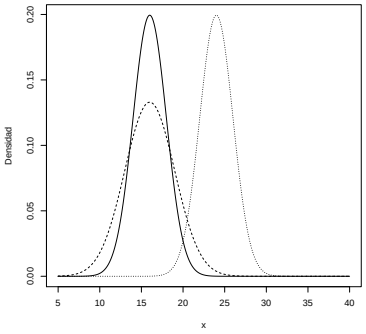


Figura 3.18: Funciones de densidad de una normal con media 16 y desviación típica 2 (trazo continuo), de una normal con media 16 y desviación típica 3 (trazo discontinuo) y una normal con media 24 y desviación típica 2 (trazo punteado).

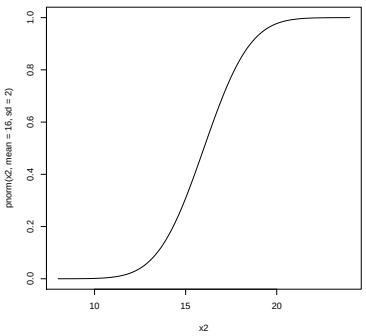


Figura 3.19: Función de distribución (acumulada) de la distribución normal con meida 16 y desviación estándar 2.

También podemos plantear el problema inverso. Consideramos una probabilidad, por ejemplo 0,34*, y buscamos el valor de* x *donde* $P(X \leq x) = 0{,}34$ *o dicho de otro modo el percentil de orden* 0,34*.*

```
qnorm(0.34,mean= 16, sd= 2)

## [1] 15.17507
```

Nota 3.4 (¿Cómo interpretar la desviación típica?)**.** *Hemos visto cómo una desviación típica mayor supone una mayor variabilidad. La variable aleatorio tiende a producir valores más dispersos. Hemos representado la función de densidad de distintas distribuciones normales y vemos cómo cuando la desviación típica es mayor entonces la gráfica es más plana. Los valores normales se observan alrededor de la media* μ *y están más o menos dispersos según el valor de* σ *sea mayor o menor. Hay una interpretación sencilla de la desviación estándar. Consideremos el intervalo* $[\mu - \sigma, \mu + \sigma]$*, ¿qué probabilidad tenemos de que una variable aleatoria con distribución normal esté en este intervalo?*

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = P(-1 \leq \frac{X - \mu}{\sigma} \leq 1) = P(-1 \leq Z \leq 1) \tag{3.18}$$

siendo Z *una variable con distribución normal estándar,* $Z \sim N(0, 1)$*. Pero,*

$$P(-1 \leq Z \leq 1) = \int_{-1}^{+1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^{+1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx - \int_{-\infty}^{-11} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \Phi(1) - \Phi(-1). \tag{3.19}$$

Vamos a calcular la diferencia anterior utilizando R.

```
pnorm(1,mean=0,sd=1) - pnorm(-1,mean=0,sd=1)

## [1] 0.6826895
```

Por tanto, si X *es una variable aleatoria con distribución normal con media* μ *y desviación típica* σ *entonces la probabilidad de que la variable esté entre* $\mu - \sigma$ *y* $\mu + \sigma$ *es*

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0{,}6826895. \tag{3.20}$$

De un modo análogo si consideramos el intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ *entonces*

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P(-2 \leq Z \leq 2) \tag{3.21}$$

que viene dado por

```
pnorm(2,mean=0,sd=1) - pnorm(-2,mean=0,sd=1)

## [1] 0.9544997
```

Y finalmente si consideramos el intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ *se tiene*

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = P(-3 \leq Z \leq 3) \tag{3.22}$$

que es igual a

```
pnorm(3,mean=0,sd=1) - pnorm(-3,mean=0,sd=1)

## [1] 0.9973002
```

En la tabla 3.3 tenemos las probabilidades que hemos calculado. De un modo sencillo podemos decir: **la variable dista de la media en una desviación estándar con una probabilidad de 0.68, en dos desviaciones con una probabilidad de 0.95 y en tres desviaciones estándar con una probabilidad de 0.99**.

$P(\mu - \sigma \leq X \leq \mu + \sigma)$	0.6826895
$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$	0.9544997
$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma)$	0.9973002

Cuadro 3.3: Probabilidad de que la variable diste de la media en un número dado de desviaciones típicas

3.7 Ejercicios

Ejercicio 3.14. *Se pide:*

1. *Simular 100 valores con distribución normal con media 20 y desviación típica 3. Guardar estos valores en el vector x.*
2. *Calcular la media y varianza muestrales de los valores generados.*
3. *Comparar la media muestral observada con 20 y la varianza muestral observada con 9 que corresponden con la media y la varianza teóricas.*
4. *Repetir los apartados anteriores sustituyendo las 100 simulaciones por 1000, por 10000 y por 100000. Comparar en cada caso los valores teóricos con los valores muestrales.*

Ejercicio 3.15. *Consideremos una variable aleatoria con distribución normal con media 20 y desviación típica 3. Se pide:*

1. $P(X \leq 23)$.
2. $P(X < 23)$.

3. $P(X > 29)$.
4. $P(X \geq 29)$.
5. $P(34 < X \leq 45)$.
6. $P(34 \leq X \leq 45)$.

4
Distribución muestral

4.1 Población y muestra aleatoria

Podemos tener interés en estudiar alguna característica en una población grande. Por ejemplo, la población puede ser una población animal o vegetal: los patos de la Albufera; las palmeras de la provincia de Valencia; una especie de pájaros en la provincia de Alicante; toda la población española, etc.

Fijémonos en las palmeras de la provincia de Valencia. Supongamos que la característica que nos interesa es si la palmera está afectada por el picudo rojo. Queremos conocer la proporción p de palmeras afectadas por la plaga. Para ello lo que podemos hacer es recorrer toda la provincia e ir palmera por palmera observando si el picudo ha afectado a la palmera. Cuando hayamos observado todas las palmeras la proporción de palmeras afectadas será el cociente entre el número de palmeras afectadas y el número total de palmeras que hay. Realmente no es difícil. Sin embargo, parece laborioso, caro y, posiblemente, innecesario. Es más barato (y esperemos que suficiente) elegir al azar un conjunto de n palmeras (con un número n no muy grande) y observar su estado. Es decir, tomar una *muestra aleatoria* de palmeras y, con lo que observamos en la muestra, intentar estimar el valor de la proporción en toda la población o *proporción poblacional*.

4.2 Distribución muestral de una variable binomial

Ejemplo 4.1. *Tenemos una población de individuos. Intentamos estudiar la prevalencia de una enfermedad no muy frecuente, la hidatidosis. Vamos a suponer que la proporción real de personas con la enfermedad es* $p = 0{,}034$.

Hemos numerada a la población y guardado los datos en el vector X. Si la i-ésima persona tiene la enfermedad entonces guardamos en la posición i del vector X un valor uno, si no la tiene guardamos un valor cero. Por ejemplo, los 10 primeros individuos de la población son

```
X[1:10]

##  [1] 0 0 0 0 1 0 0 0 0 0
```

Y el individuo que ocupa la posición 100000 es

```
X[100000]

## [1] 0
```

Ahora vamos a simular el muestreo aleatorio de la población. Tomamos una muestra de tamaño $n = 100$ *y observamos la proporción de personas enfermas en la muestra.*

```
n = 100
x = sample(X,n)
x

##   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [28] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [55] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [82] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

De hecho el total de individuos enfermos en la muestra lo podemos ver con

```
sum(x)

## [1] 2
```

Repitiendo el proceso obtenemos

```
x = sample(X,100)
sum(x)

## [1] 6
```

Y si lo hacemos como unas 20 veces obtenemos los siguientes valores observados

```
##  [1]  3  5  4  3  2  8  2  4 10  3  3  6  4  4  5  4  6  3
## [19]  4  4
```

Si, en lugar de contar el número de enfermos observados, nos fijamos en la proporción observada tenemos

```
sumas/n

##  [1] 0.03 0.05 0.04 0.03 0.02 0.08 0.02 0.04 0.10 0.03 0.03
## [12] 0.06 0.04 0.04 0.05 0.04 0.06 0.03 0.04 0.04
```

¿Con qué frecuencia observamos 3 enfermos?
La obtenemos con la expresión

$$\binom{n}{3},034^3(1-0{,}034)^{97}$$

En general si en una selección aleatoria de n individuos contamos el número de individuos con la enfermedad (que entendemos como el número de éxitos) entonces

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

siendo p la proporción en la población de enfermos (que en nuestro ejemplo estamos suponiendo $p=0{,}034$).

4.2.1 Ejercicios

Ejercicio 4.1. *Wilcox [2009, pág. 79, problemas 2-3] Muchos equipos de investigación pretenden realizar un estudio sobre el porcentaje de personas que tienen cáncer de colon. Si una muestra aleatoria de diez personas se pudo obtener, y si la probabilidad de probabilidad de tener cáncer de colon es 0,05, ¿cuál es la probabilidad de que un equipo de investigación obtenga $\hat{p}=0{,}1$? ¿Y la de $\hat{p}=0{,}05$?*

Ejercicio 4.2. *Wilcox [2009, pág. 80, problema 4] Alguien afirma que la probabilidad de perder dinero cuando se utiliza una estrategia de inversión para la compra y venta de los productos básicos es de 0,1. Si esta afirmación es correcta: ¿cuál es la probabilidad de obtener $\hat{p}\leq 0{,}05$ sobre la base de una muestra aleatoria de 25 de los inversores?*

Ejercicio 4.3. *Wilcox [2009, pág. 80, problemas 6-7] Imaginemos que un millar de equipos de investigación extraen una muestra al azar de una distribución binomial con $p=0{,}4$, cada estudio está basado en una muestra de tamaño 30. Tendremos 1000 valores de $\hat{p}$. Si promediamos estos 1000 valores: ¿Cuál sería aproximadamente el resultado? Si calculamos la varianza muestral de los valores de $\hat{p}$: ¿Cuál sería aproximadamente el resultado?*

4.3 Distribución muestral de la media bajo normalidad

Supondremos que es una población normal con media $\mu=160$ y una desviación estándar de 10.23. Podría corresponder con la altura de los individuos. Suponemos que hay $N=237456$ personas en la población en estudio. Como no vamos a medir a tantas personas ahorramos tiempo generando aleatoriamente estos valores.

```
N = 237456
X = rnorm(N,mean=160,sd=10.23)
```

Es una población de $2{,}37456 \times 10^5$ individuos. Por X estamos denotando *toda la población*. Es decir, suponemos (de un modo irreal) que tenemos a toda la población. Podemos ver los 10 primeros elementos de la población con

```
X[1:10]

##  [1] 155.6790 159.2694 166.8456 166.4420 168.7246 153.1952
##  [7] 159.5500 152.5387 155.1360 145.7707
```

La figura 4.1 tenemos un histograma de toda la población. La figura 4.2 tiene un estimador kernel de la densidad de probabilidad.

Como conocemos toda la población podemos conocer su media o *media poblacional* y vale:

```
(mu=mean(X))

## [1] 160.0131
```

Tomamos una muestra de $n = 100$ individuos de la población con la función *sample*.

```
n = 100
x = sample(X,n)
```

Parece natural *aproximar* o, dicho con propiedad, *estimar* el valor desconocido de μ con la media muestral. Veamos el valor:

```
mean(x)

## [1] 159.4892
```

Si repetimos la selección de los individuos y el cálculo de la media muestral tenemos

```
x = sample(X,n)
mean(x)

## [1] 158.4999
```

Podemos repetirlo muchas veces e iremos obteniendo valores distintos. En la figura 4.3 estimador kernel de la densidad de los valores generados.

Supongamos que repetimos lo anterior pero incrementando el tamaño de la muestra. En lugar de tomar muestras de tamaño 100

```
hist(X)
```

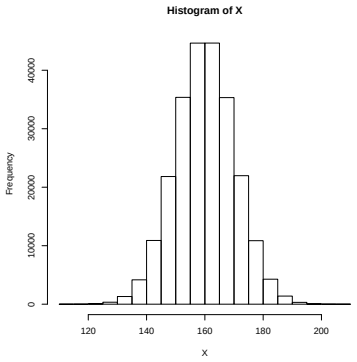


Figura 4.1: Histograma de la población de alturas.

```
plot(density(X))
```

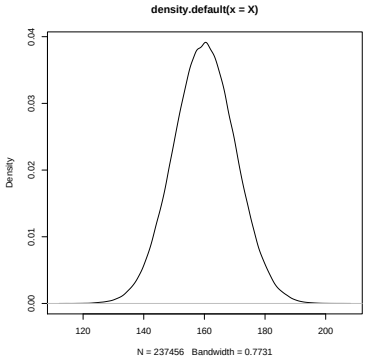


Figura 4.2: Estimador kernel de la densidad de las alturas.

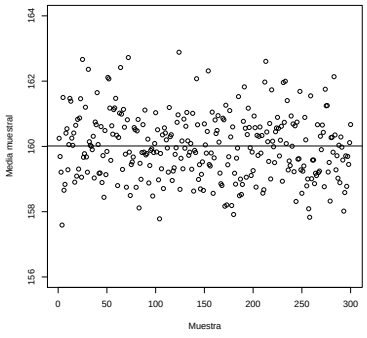


Figura 4.3: Generamos muestras de tamaño 100 de la población. Calculamos la media muestral de cada muestra. En el eje de abscisas mostramos el número de la muestra que hemos generado. En el eje de ordenadas en valor observado. La línea horizontal muestra la media poblacional.

pasamos a tomar muestras de tamaño 400. En la figura 4.4 mostramos las medias muestrales obtenidas para diferentes muestras. Las medias muestrales están mucho más próximas a la media poblacional.

Finalmente supongamos que tomamos muestras de tamaño creciente y mostramos en abscisas el tamaño de la muestra y en ordenadas la media muestral observada. En la figura 4.5 tenemos el resultado. Lo repetimos. En la figura 4.6 tenemos las medias observadas. No obtenemos las mismas medias muestrales pero si un comportamiento aleatorio similar.

Si denotamos la muestra aleatoria que estamos extrayendo de la población con $X_1, \ldots, X_n$ entonces la media muestral (que utilizaremos para estimar la media poblacional) tiene una distribución (o se distribuye como) una normal con media μ (la media poblacional) y con varianza $\sigma^2_{\bar{X}_n} = \frac{\sigma^2}{n}$, la varianza poblacional dividida por el tamaño de la muestra. De un modo resumido esto se expresa con

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \tag{4.1}$$

si $X_1, \ldots, X_n$ son variables aleatorias independientes y con distribución

$$X_i \sim N(\mu, \sigma^2), \text{ con } i = 1, \ldots, n.$$

El resultado dado en 4.1 también lo podemos expresar como

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim N(0, 1) \tag{4.2}$$

Nos interesa conocer probabilidades como

$$P(\bar{X} \leq b)$$

o en general probabilidades como

$$P(a \leq \bar{X} \leq b)$$

donde a y b son valores que nos pueden ir interesando dependiendo del problema.

Nota de R 4.1. *Por ejemplo, supongamos que nos interesa saber qué probabilidad tiene la media muestral de ser menor que* 162*. Como estamos suponiendo que conocemos toda la población podemos tomar como varianza la de toda la población.*

```
sigma = sd(X)
```

Y la probabilidad la podemos obtener con

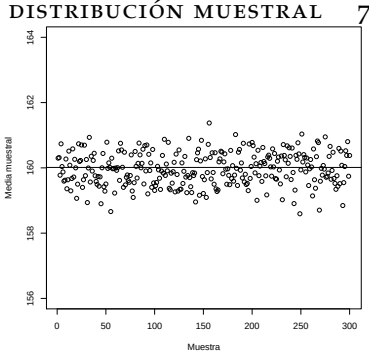


Figura 4.4: Generamos muestras de tamaño 400 de la población. Calculamos la media muestral de cada muestra. En el eje de abscisas mostramos el número de la muestra que hemos generado. En el eje de ordenadas en valor observado. La línea horizontal muestra la media poblacional.

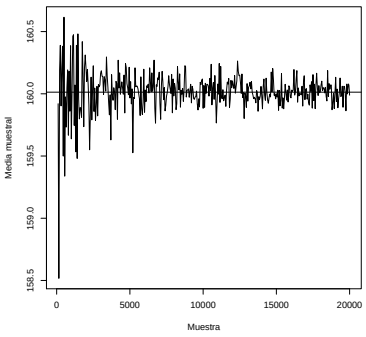


Figura 4.5: Generamos muestras de tamaño creciente. Calculamos la media muestral de cada muestra. En el eje de abscisas mostramos el tamaño de la muestra que hemos generado. En el eje de ordenadas en valor observado de la media muestral. La línea horizontal muestra la media poblacional. Vemos cómo las medias muestrales se aproximan a la media de la población.

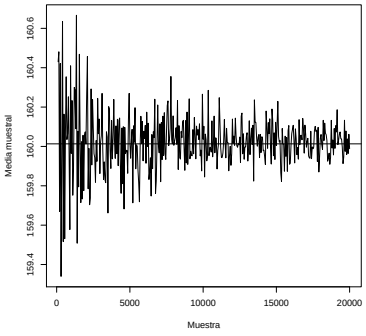


Figura 4.6: Generamos muestras de tamaño creciente. Calculamos la media muestral de cada muestra. En el eje de abscisas mostramos el tamaño de la muestra que hemos generado. En el eje de ordenadas en valor observado de la media muestral. La línea horizontal muestra la media poblacional. Vemos cómo las medias muestrales se aproximan a la media de la población.

```
pnorm(162,mean=mu,sd=sigma/sqrt(n))

## [1] 0.9999501
```

¿Qué probabilidad tenemos de que la media esté entre 159 y 162?

```
pnorm(162,mean=mu,sd=sigma/sqrt(n)) -
    pnorm(159,mean=mu,sd=sigma/sqrt(n))

## [1] 0.9763304
```

¿O de que sea mayor que 160?

```
1 - pnorm(160,mean=mu,sd=sigma/sqrt(n))

## [1] 0.510255
```

Si

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(t-\mu)^2}{2\sigma^2} dt.$$

lo que estamos haciendo con R es simplemente aplicar que

$$P(\bar{X} \leq b) = \Phi(b)$$

o

$$P(a \leq \bar{X} \leq b) = \Phi(b) - \Phi(a)$$

o

$$P(a \leq \bar{X}) = 1 - \Phi(a)$$

4.3.1 Ejercicios

Ejercicio 4.4. *Wilcox [2009, pág. 84, problema 8] Supongamos $n = 16$, $\sigma = 2$ y $\mu = 30$. Supongamos normalidad. Determinar:*

1. $P(\bar{X} \leq 29)$,
2. $P(\bar{X} > 30{,}5)$,
3. $P(29 \leq \bar{X} \leq 31)$.

Ejercicio 4.5. *Wilcox [2009, pág. 84, problemas 10-11] Alguien dice que dentro de un determinado barrio, el coste medio de una casa es de $\mu = 100000$ euros con una desviación estándar de $\sigma = 10{,}000$ euros. Supongamos que, basándonos en $n = 16$ viviendas, observamos una media muestral $\bar{X} = 95{,}000$. Suponiendo normalidad, ¿cuál es la probabilidad de obtener una media muestral como la observada o menor si las afirmaciones sobre la media y desviación estándar son verdaderas? ¿Y la probabilidad de tener una media muestral entre 97500 y 102500 euros?*

Ejercicio 4.6. *Wilcox [2009, pág. 85, problema 13] Supongamos que eres un profesional de la salud interesado en los efectos de la medicació en la presión arterial diastólica de las mujeres adultas. Para un medicamento en particular que está siendo estudiado, se encuentra que para $n = 9$ mujeres, la media muestral es $\bar{X} = 85$ y la varianza muestral es $s^2 = 160{,}78$. Estimar el error estándar de la media muestral asumiendo que tenemos una muestra aleatoria.*

Ejercicio 4.7. *Wilcox [2009, pág. 85, problema 12] Una compañía afirma que las primas pagadas por sus clientes para el seguro de automóviles tiene una distribución normal con media $\mu = 750$ euros y desviación estándar $\sigma = 100$ euros. Suponiendo normalidad, ¿cuál es la probabilidad de que para $n = 9$ clientes elegidos al azar, ¿la media muestral tome un valor entre 700 y 800 euros?*

4.4 Distribución muestral de la media en poblaciones no normales. Teorema central del límite

El resultado que damos en 4.1 es aproximadamente cierto incluso aunque los datos no sigan aproximadamente una distribución normal.

4.4.1 Aproximación de la distribución binomial

Con datos binomiales, el estimador de la proporción p, $\hat{p}$ podemos verlo como una media muestral

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n},$$

donde $X_i = 1$ si hay un éxito en la i-ésima prueba de Bernoulli. Utilizando el teorema central del límite tenemos que

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}. \tag{4.3}$$

Si n es suficientemente grande tenemos que para cualesquiera valores reales a y b tenemos

$$\begin{aligned} P(a \leq \hat{p} \leq b) = \\ P(\hat{p} \leq b) - P(\hat{p} \leq a) = \\ P(Z \leq \frac{b-p}{\sqrt{p(1-p)/n}}) - P(Z \leq \frac{a-p}{\sqrt{p(1-p)/n}}). \end{aligned} \tag{4.4}$$

La calidad de la aproximación depende los valores de n y de p. Una regla simple es que la aproximación es buena si $np \geq 15$ y $n(1-p) \geq 15$.

4.4.2 *Ilustración del teorema central del límite*

Hemos utilizado en la sección anterior este resultado probabilístico. El teorema central del límite dice que si tenemos variables aleatorias $X_1, X_2, \ldots$ independientes entre sí y con una misma distribución[1] entonces la media muestral se comporta asintóticamente según una distribución normal. En concreto, si la media y varianza común a todas las variables son μ and σ^2 entonces

[1] En definitiva repetimos independientemente un mismo experimento y observamos una misma cantidad cada vez.

$$\lim_{n\to+\infty} P\left(\frac{\bar{X}_n - \mu}{\sqrt{n}\sigma} \leq z\right) = P(Z \leq z) \tag{4.5}$$

donde Z es una variable con distribución normal con media 0 y varianza 1, es decir, una normal estándar o típica.

4.4.3 *Ejercicios*

Ejercicio 4.8. *1. Supongamos una distribución binomial con $p = 0{,}5$ y $n = 10$ y queremos calcular la probabilidad de que $\hat{p}$ sea menor o igual a $7/10$. Obtener el valor exacto y el valor aproximado utilizando la aproximación dada por el teorema central del límite.*

2. Repetir el punto anterior obteniendo el valor exacto y el valor aproximado de $P(0{,}3 \leq \hat{p} \leq 0{,}7)$.

Ejercicio 4.9. *1. Supongamos una distribución binomial con $p = 0{,}5$ y $n = 100$ y queremos calcular la probabilidad de que $\hat{p}$ sea menor o igual a $0{,}55$. Obtener el valor exacto y el valor aproximado utilizando la aproximación dada por el teorema central del límite.*

2. Repetir el punto anterior obteniendo el valor exacto y el valor aproximado de $P(0{,}45 \leq \hat{p} \leq 0{,}55)$.

5
Estimación

5.1 Introducción

Tratamos el problema de la estimación. En concreto, en poblaciones normales, nos planteamos la estimación de la media y varianza. También consideramos la estimación de una proporción. Se aborda el problema del cálculo del tamaño de la muestra para estimar los parámetros con un error máximo dado.

5.2 La población

¿Qué es una población? La Estadística se ocupa del estudio de grandes poblaciones. Pero, otra vez, ¿y qué es una población? La respuesta no es simple ni tampoco es única.

El primer sentido que podemos dar al término *población* es una gran colección de elementos de los cuales queremos conocer algo. Algunos ejemplos son:

1. la población española a día 2 de noviembre de 2011;
2. la población de samaruc el 2 de diciembre de 2010;
3. la población de fartet en la Comunidad Valenciana en febrero de 2012;
4. Los pinos de los Montes Universales;

y muchísimos ejemplos similares. Todos los ejemplos que acabamos de proponer se caracterizan porque *toda la población está ahí* y, en principio, podríamos observarla. Podemos tener la santa paciencia de observar todos y cada uno de los pinos de los Montes Universales. Estos ejemplos son poblaciones *finitas*, es decir, son grandes conjuntos de individuos pero un número finito. Tenemos interés en alguna característica de la población. Por ejemplo, ¿cuál es la longitud media

de un samaruc adulto de la Albufera? ¿Cuál es la proporción de pinos en los Montes Universales afectados por la procesionaria?

¿Cómo podemos obtener este valor? Es una población finita y los peces están ahí. Pues los cogemos y medimos cada uno de ellos. La cantidad buscada no es más que la media de las longitudes de cada uno de los peces. Simple de decir sí pero, obviamente, impracticable. Aunque, sobre el papel, es una población accesible; realmente, es una población inaccesible. *No podemos acceder a toda la población*. Por ello, hemos de considerar el experimento consistente en elegir al azar un individuo de la población. Si el individuo elegido lo denotamos por ω, una vez lo tenemos, podemos medirlo, podemos observar la característica que nos interesa, podemos observar la variable X en el individuo ω, siendo x el valor que observamos. Este proceso es lo que se denota abreviadamente como $X(\omega) = x$. Sin embargo, no observamos la variable aleatoria X una sola vez. Elegimos independientemente y de la misma población de peces n individuos. Antes de elegirlos tenemos una colección de valores aleatorios que son independientes entre sí y que repiten el experimento. Esto es lo que se conoce como *muestra aleatoria* y denotamos $\{X_1, \ldots, X_n\}$. Los valores observamos una vez hemos elegido a los n peces los denotamos con $x_1, \ldots, x_n$.

En otras ocasiones la población no es un concepto tan concreto. Por ejemplo, estamos la demanda biológica de oxígeno en muestras de agua tomadas en la playa de Canet d'en Berenguer. En principio, el número de muestras es infinito. Además, aunque asumamos que las muestras tienen un comportamiento aleatorio similar en la playa de Canet, es claro que necesitamos indicar, al menos, el día en que las tomamos. En este caso, la población no existe. Son las infinitas repeticiones que podemos hacer del experimento consistente en tomar una muestra de agua y determinar la demanda biológica de oxígeno.

5.3 *Estimación puntual*

Nota 5.1 (Estimando una media)**.** *Vamos a ilustrar los conceptos con una población finita. Y muy grande. Suponemos que tenemos una población de 237456 personas que conocemos la altura de cada una de las personas.*[1] *Los datos los tenemos en el vector X.*

Por ejemplo, las estaturas de las diez primeras personas (en centímetros) son

```
##  [1] 157.7 161.8 156.5 158.7 159.2 153.2 154.4 155.9 162.7
## [10] 164.0
```

Por X estamos denotando toda la población*. Por lo tanto la media de la población o media poblacional la conocemos. Simplemente hemos de realizar*

[1] Hemos tenido la santa paciencia de medir la estatura de cada uno de ellos. Y ellos se han dejado.

la media muestral de todas las estaturas.

```
(mu=mean(X))

## [1] 159.9984
```

Pretendemos estimar la media poblacional (que en nuestro ejemplo artificial conocemos) utilizando una muestra aleatoria de 10 individuos (una muestra no muy grande ciertamente). Podemos elegir la muestra aleatoria con la función sample*. Por ejemplo, una primera muestra estaría compuesta por los individuos*

```
n = 10
(x = sample(X,n))

##  [1] 155.9479 154.8805 159.8398 158.3481 161.9378 160.6372
##  [7] 152.8162 159.2216 163.1594 166.1877
```

Nuestra estimación sería

```
(mediamuestral = mean(x))

## [1] 159.2976
```

de modo que el error que cometemos es

```
mediamuestral - mu

## [1] -0.7008141
```

Supongamos que repetimos la estimación varias veces y veamos los diez primeros valores que obtenemos

```
##  [1] 159.6961 161.0358 159.9950 156.5019 159.8882 159.3672
##  [7] 161.3581 159.7155 161.3998 159.8084
```

Veamos un resumen de los errores observados.

```
errores = estimaciones-mu
summary(errores)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.78700 -1.04900 -0.05683 -0.12220  0.98350  3.96100
```

De hecho, en la figura 5.1 hemos representado un estimador kernel de las estimaciones y una línea vertical mostrando la media real.

Así es como funciona la estimación puntual. Hemos visto una situación irreal en que tenemos todos los valores que componen la población y,

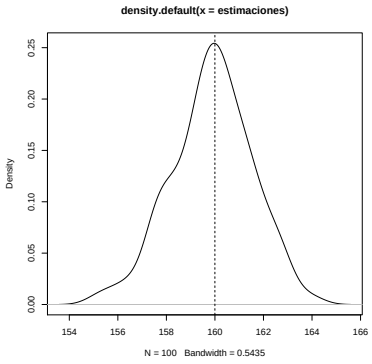


Figura 5.1: Estimador kernel de la densidad de las estimaciones obtenidas eligiendo muestras de tamaño 10 de la población. La línea vertical tiene como abscisa común la media poblacional.

por lo tanto, podemos conocer la media de la población. Esto no es nunca así. Conocemos una muestra aleatoria de la población que denotaremos por $X_1, \ldots, X_n$ *y, con estos valores, pretendemos estimar la media de la población* μ*. De hecho, hemos utilizado estos valores y obtenido un método para estimar* μ*, a esto le podemos llamar un estimador de* μ *y denotarlo como*

$$\hat{\mu} = \bar{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Antes de tomar la muestra, $\bar{X}_n$ *es un valor aleatorio pero* después de tomar la muestra *tenemos que* $X_1 = x_1, \ldots, X_n = x_n$*, es decir, que no son valores aleatorios sino unos valores dados. En el ejemplo anterior, antes de utilizar la función* sample *tenemos un valor aleatorio pero después tenemos el valor observado correspondiente. Por lo tanto una vez seleccionada la muestra decimos que tenemos los valores observados* $x_1, \ldots, x_n$ *y la variable aleatoria* $\bar{X}_n$ *toma el valor fijo* $\bar{x}_n = \sum_{i=1}^{n} x_i/n$*.*

5.4 *Algunas definiciones*

Definición 5.1 (Muestra aleatoria)**.** *Una muestra aleatoria (de tamaño n) son n valores aleatorios* $X_1, \ldots, X_n$ *que tienen la misma distribución y son independientes entre sí.*

Definición 5.2 (Estimador)**.** *Un* estimador *es cualquier función de la muestra aleatoria* $X_1, \ldots, X_n$ *que toma valores admisibles para el parámetro que estimamos.*

5.5 *Estimación puntual de la media*

Si pretendemos estimar la media μ de una población su estimador usual es la media muestral, $\bar{X}_n$. Si tenemos una muestra aleatoria donde cada X_i sigue una distribución normal entonces

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \tag{5.1}$$

Si las distintas variables X_i que componen la muestra no siguen una distribución muestral entonces asumiendo que tenemos una muestra grande el resultado que damos en 5.1 es aproximadamente cierto. En cualquier caso, la varianza de la media muestral $\bar{X}_n$

5.6 *Intervalo de confianza para la media*

Hemos considerado en el apartado anterior la estimación puntual de la media poblacional, μ, mediante la media muestral, $\bar{X}$. Es una opción natural. Otra opción puede ser estimar la media de la

población μ dando un intervalo que la contenga. En lugar de decir "estimamos la media poblacional en 158.55", esto es utilizar un valor numérico exclusivamente podemos utilizar una expresión como "la media poblacional es mayor que 158,37 y menor que 162,79". Este segundo tipo de estimación es la que se hace con un intervalo de confianza. En resumen estimamos la media poblacional μ o bien mediante un punto (estimador puntual) o bien mediante un intervalo (estimación por intervalos). Al primer método se le llama *estimador puntual* y al segundo *intervalo de confianza*.

5.6.1 *Asumimos que conocemos la varianza*

Vamos a asumir en un primer momento algo que no tiene ninguna realidad (raramente nos lo vamos a encontrar en una aplicación real) pero nos facilita la presentación. En concreto asumimos que no conocemos la media μ (es lo que queremos estimar) pero, sin embargo, **sí** conocemos la desviación estándar, σ.

Una segunda hipótesis (que sí se verifica con frecuencia) que vamos a asumir es que los datos proceden de una población normal. Ya discutiremos qué hacemos después con cada una de estas hipótesis.

Siendo μ y σ la media y la desviación típica reales, por el teorema central del límite se verifica que aproximadamente (o con más precisión si el tamaño de la muestra aleatoria, n, es grande) la media muestral tiene una distribución normal con media μ y desviación estándar $\sigma/\sqrt{n}$,

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}),$$

o, lo que es equivalente, que

$$\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim N(0,1),$$

donde $N(0,1)$ es una normal con media cero y varianza uno, lo que se conoce como una *normal estándar* o *normal típica*. [2]

Notemos que, *asumiendo la desviación estándar conocida*, en la expresión $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ conocemos todos los términos que aparecen ($\bar{X}, \sigma$ y n) una vez hemos tomado la muestra salvo el valor de la media poblacional μ. Precisamente es lo que queremos estimar.

Fijemos una probabilidad alta, por ejemplo, una probabilidad de 0,95. Podemos determinar un valor positivo c tal que

$$P\left(-c \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq c\right) = 0{,}95$$

Por las propiedades de simetría de la normal c ha de verificar que

$$P\left(\frac{\bar{X}-\mu}{\sigma} \leq c\right) = 0{,}975$$

[2] Es equivalente porque si una variable aleatoria X suponemos que tiene distribución normal con media μ y varianza σ^2 entonces la variable aleatoria $(X-\mu)/\sigma$ sigue una distribución normal con media o y varianza 1 que se conoce como una normal típica o normal estándar. Esto se suele indicar como: $X \sim N(\mu, \sigma^2)$ entonces $(X-\mu)/\sigma \sim N(0,1)$

Podemos determinarlo con la función *qnorm* del siguiente modo.

```
qnorm(0.975,mean=0,sd=1)

## [1] 1.959964
```

En resumen que

$$P\left(-1{,}96 \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq 1{,}96 \right) = 0{,}95$$

o, lo que es equivalente,

$$P\left(\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}} \right) = 0{,}95.$$

Vemos que el intervalo $[\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}]$ tiene una probabilidad de 0,95 de contener a μ o también de *cubrir* a μ. Si ahora sustituimos los valores aleatorios X_i con $i = 1, \ldots, n$ con los valores observados en la muestra entonces el intervalo aleatorio $[\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}]$ pasa a ser un intervalo fijo $[\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}]$ que conocemos como *intervalo de confianza con nivel de confianza* 0,95.

Nota de R 5.1 (Intervalo de confianza con R)**.** *Vamos a evaluarlo con R. Empezamos tomando los datos.*

```
(x = sample(X,10))

##  [1] 160.8437 150.2768 154.0478 169.9773 161.1820 148.4176
##  [7] 160.6448 159.4391 161.9439 161.9326
```

Determinamos la media muestral y la desviación estándar muestral.

```
media = mean(x)
s = sd(x)
```

Por lo tanto el intervalo tendrá por extremo inferior

```
(extremo.inferior = mean(x) - qnorm(0.975,mean=0,sd=1)*sd(x)/sqrt(n))

## [1] 154.9438
```

y por extremo superior.

```
(extremo.superior = mean(x) + qnorm(0.975,mean=0,sd=1)*sd(x)/sqrt(n))

## [1] 162.7974
```

Si en lugar de 0,95 *elegimos como nivel de confianza (utilizando la notación habitual)* $1-\alpha$ *con* α *un valor pequeño (en el ejemplo anterior* $\alpha = 0{,}05$*) podemos determinar* c *que verifique*

$$P(-c \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq c) = 1-\alpha$$

o que

$$P(\frac{\bar{X}-\mu}{\sigma} \leq c)) = 1-\frac{\alpha}{2}.$$

Denotamos el valor de c *que verifica lo anterior como* $Z_{1-\frac{\alpha}{2}}$*. Finalmente el intervalo*

$$[\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

cubre *a* μ *con una probabilidad de* $1-\alpha$ *y el intervalo de confianza es el que obtenemos cuando sustituimos los valores aleatorios por los valores observados. Es decir, el intervalo de confianza viene dado por*

$$[\bar{x} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

De un modo genérico: ¿Cómo podemos determinar el intervalo de confianza que acabamos de ver? Fijamos el nivel de confianza $1-\alpha$ *(en este caso a* $1-\alpha = 0{,}99$ *o equvalentemente* $\alpha = 0{,}01$*).*

```
alpha = 0.01
```

Determinamos el extremo inferior:

```
(extremo.inferior = mean(x) - qnorm(1-alpha/2,mean=0,sd=1)*sd(x)/sqrt(n))

## [1] 153.7099
```

y el superior

```
(extremo.superior = mean(x) + qnorm(1-alpha/2,mean=0,sd=1)*sd(x)/sqrt(n))

## [1] 164.0313
```

El intervalo es el siguiente:

```
c(extremo.inferior,extremo.superior)

## [1] 153.7099 164.0313
```

El último paso es rogar a Dios que las cosas hayan ido bien. Tenemos una confianza de $1-\alpha$ *(*0,99 *en el ejemplo) de que el valor real de la media esté en este intervalo. Pero esto no quiere decir que realmente lo está. Si repetimos un gran número de veces el valor real de la media está en el intervalo un* $(1-\alpha)\times 100$ *% de la veces (un* 99 *% en el ejemplo) pero puede ocurrir (desgracias de la vida) que estemos en el* $\alpha \times 100$ *(en el ejemplo un* 1 *%) restante. En general la cosa va bien porque elegimos un nivel de confianza grande (próximo a uno) pero no siempre va bien.*

5.6.2 *No asumimos la varianza conocida*

No es realista asumir que conocemos la varianza σ^2 (o la desviación típica σ). En absoluto lo es. ¿Por qué vamos a desconocer la media y conocer la varianza? [3] Salvo situaciones realmente esotéricas esto no es así. En consecuencia lo lógico es sustituir la desviación típica poblacional σ por su estimador natural $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$. Si lo hacemos tendremos la cantidad

$$\sqrt{n}\frac{\bar{X} - \mu}{S}. \tag{5.2}$$

En la expresión anterior lo conocemos todo (una vez tenemos los datos) excepto el valor de μ. Sin embargo, la distribución de probabilidad de esta cantidad *ya no es una normal estándar*. Nos aparece una distribución de probabilidad nueva que se conoce como la distribución t de Student. [4]

De hecho, William Sealy Gossett demostró que la cantidad $\sqrt{n}\frac{\bar{X}-\mu}{S}$ se comporta como una t de Student con $n-1$ grados de libertad. Esto lo denotaremos como

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S} \sim t_{n-1}. \tag{5.3}$$

En la figura 5.2 hemos representado la función de densidad de una t de Student con 9 grados de libertad.

[3] ¿Van por ahí los datos diciendo *somos normales no te decimos la media pero te decimos la varianza*? Los pobre datos no dicen nada.

[4] La expresión exacta de esta densidad es $f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}}\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$ donde $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$ es la función Gamma de Euler. La miráis y ya está. No hace falta saberla pero por lo menos es bueno verla una vez en la vida.

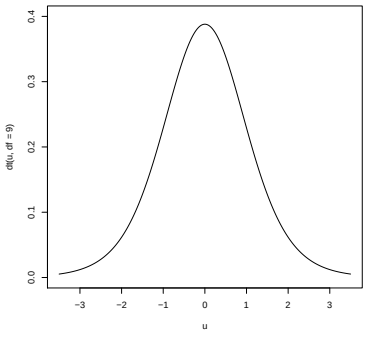


Figura 5.2: Función de densidad de una t de Student con 9 grados de libertad

Nota 5.2 (¿Qué relación tiene la t de Student con la normal?)**.** *Hay dos puntos interesantes a tener en cuenta cuando utilizamos una distribución t de Student. La primera es qué relación tiene con una distribución normal estándar y la segunda es qué ocurre cuando modificamos los grados de libertad.*

Para ver la relación con la normal estándar hemos representado en la figura 5.3 la densidad de la normal en trazo continuo y la densidad de una t de Student con 2 grados de libertad en trazo discontinuo. Vemos que tienen una forma similar, ambas están centradas en cero. Sin embargo, la densidad de la normal está más concentrada alrededor de cero. La densidad de la t de Student está más repartida. Y esto ocurre para cualquier número de grados de libertad.

¿Qué ocurre cuando incrementamos el número de grados de libertad? Cuando se va incrementando el número de grados la densidad de la t de Student se aproxima a la densidad de la normal. En la figura 5.4 se ilustra y comenta este hecho.

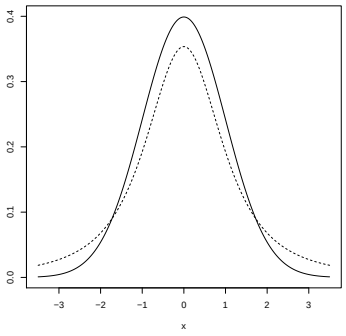


Figura 5.3: Funciones de densidad de la normal estándar (trazo continuo) y de densidades t de Student con 2 grados de libertad.

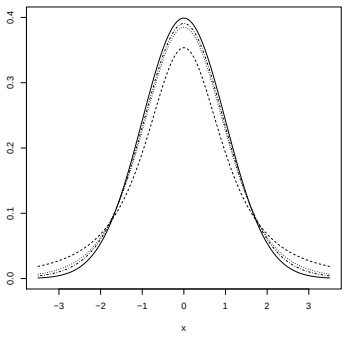


Figura 5.4: Funciones de densidad de la normal estándar (trazo continuo) y de densidades t de Student con 2, 7 y 12 grados de libertad. Según el número de grados de libertad de la t es mayor más se aproxima la densidad de la t a la normal. Por ello, la más alejada es la t(2) y la más próxima es la t(12).

Y ahora vamos a repetir lo visto en la sección anterior sustituyendo a la normal estándar con la densidad de la t de Student con n-1 grados de libertad. Dada una probabilidad, por ejemplo 0,95, podemos determinar el valor c tal que

$$P(-c \leq T \leq c) = 0{,}95.$$

En concreto verificará que

$$P(T \leq c) = 0{,}975$$

y el valor de c (por ejemplo, para 9 grados de libertad) lo obtendremos con

```
qt(0.975,df=9)

## [1] 2.262157
```

Denotamos el valor de c tal que

$$P(T \leq c) = 1 - \frac{\alpha}{2},$$

como $t_{n-1,1-\alpha/2}$. Entonces

$$P(\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}) = 1 - \alpha. \qquad (5.4)$$

El intervalo de confianza lo obtenemos sustituyendo los valores aleatorios por los valores observados.

Resultado 5.1. *Si suponemos que tenemos una muestra aleatoria de datos normales $X_1, \ldots, X_n$ y observamos los datos $X_1 = x_1, \ldots, X_n = x_n$ entonces el intervalo*

$$\left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

es un intervalo de confianza con nivel de confianza $1-\alpha$ para la media μ. De un un modo abreviado el intervalo anterior se puede escribir como

$$\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$$

Nota de R 5.2 (Cálculo del intervalo de confianza con R)**.** *¿Y cómo hacerlo con R? Fijamos el nivel de confianza (en este caso a 0,99).*

```
alpha = 0.01
```

Determinamos el extremo inferior:

```
(extremo.inferior = mean(x) - qt(1-alpha/2,df=n-1)*sd(x)/sqrt(n))

## [1] 152.3595
```

y el superior

```
(extremo.superior = mean(x) + qt(1-alpha/2,df=n-1)*sd(x)/sqrt(n))

## [1] 165.3817
```

El intervalo es el siguiente

```
c(extremo.inferior,extremo.superior)

## [1] 152.3595 165.3817
```

[5]

Nota 5.3 (Obtención del intervalo de confianza utilizando *t.test*)**.** *No hace falta escribir todo lo anterior para calcular el intervalo de confianza. El intervalo de confianza para la media lo podemos obtener con la función* t.test*. De hecho, nos da el intervalo de confianza y alguna cosa más que, de momento, no necesitamos.*

```
alpha = 0.05
t.test(x,conf.level=1-alpha)

##
##  One Sample t-test
```

[5] Podemos comprobar que el intervalo que obtenemos asumiendo varianza conocida es más pequeño que el que obtenemos asumiendo varianza desconocida.

```
##
## data:  x
## t = 79.296, df = 9, p-value = 4.084e-14
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  154.3383 163.4028
## sample estimates:
## mean of x
##  158.8706
```

Si queremos una salida en la que solamente nos aparezca el intervalo de confianza podemos hacer

```
alpha = 0.05
t.test(x,conf.level=1-alpha)$conf.int

## [1] 154.3383 163.4028
## attr(,"conf.level")
## [1] 0.95
```

Nota 5.4 (Evaluando el intervalo de confianza)**.** *Podemos repetir el proceso de estimación muchas veces y ver en cuántas ocasiones el intervalo contiene al verdadero valor de la media. Generamos 100 intervalos con muestras de tamaño* $n = 100$*. En la figura 5.5 hemos representado los 100 intervalos que hemos obtenido. La línea vertical indica la media poblacional. Cada segmento horizontal se ha dibujado de modo que las abcisas de sus extremos corresponden con el extremo inferior y superior del correspondiente intervalo de confianza.*

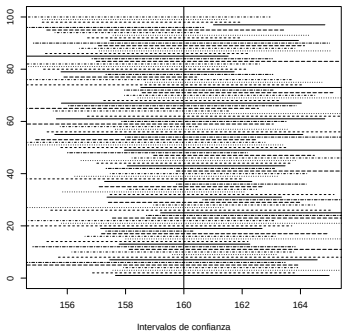


Figura 5.5: Intervalos de confianza de la media en una población normal. Hemos simulado 100 intervalos. La línea vertical tiene como abscisa el valor real de la media. ¿Cuántos intervalos no contienen a la media real? Cuéntalos.

Ejemplo 5.1 (Datos Kola)**.** *Empezamos cargando los datos del proyecto Kola.* [6] *En concreto vamos a utilizar los datos* chorizon*. Con la función* attach *podemos usar los nombres de las variables.* [7]

```
load("../data/chorizon.rda")
attach(chorizon)
```

Vamos a obtener el intervalo de confianza para la concentración media de escandio en Noruega. En primer lugar guardamos en Sc.Nor *los datos correspondientes a Noruega.*

```
Sc.Nor = Sc[COUN == "NOR"]
```

La función básica es t.test*.*

[6] http://www.ngu.no/Kola/

[7] Se recomienda hacer *help(chorizon)* para conocer más sobre los datos.

```
t.test(Sc.Nor)

##
##  One Sample t-test
##
## data:  Sc.Nor
## t = 18.342, df = 127, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.479107 3.078706
## sample estimates:
## mean of x
##  2.778906
```

Si queremos obtener el intervalo de confianza solamente podemos hacer lo siguiente.

```
Sc.t = t.test(Sc.Nor)
Sc.t$conf.int

## [1] 2.479107 3.078706
## attr(,"conf.level")
## [1] 0.95
```

o simplemente

```
t.test(Sc.Nor)$conf.int

## [1] 2.479107 3.078706
## attr(,"conf.level")
## [1] 0.95
```

Por defecto el nivel de confianza que se elige es 0,95*. Podemos modificarlo. Bien bajándolo*

```
t.test(Sc.Nor,conf.level=.90)$conf.int

## [1] 2.527873 3.029940
## attr(,"conf.level")
## [1] 0.9
```

o bien tomando un nivel de confianza mayor, 0,99*.*

```
t.test(Sc.Nor,conf.level=.99)$conf.int

## [1] 2.382708 3.175105
## attr(,"conf.level")
## [1] 0.99
```

Ejemplo 5.2. *Lo vamos a ilustrar con datos de los resultados de la selectividad en la Comunidad Valenciana. Estos datos se pueden encontrar en http://www.edu.gva.es/univ/val/prueba_acceso.htm Los datos nos dan la media y desviación estándar muestrales de las notas obtenidas en Matemáticas II en el examen de selectividad en la Comunidad Valenciana el año 2010. Tenemos estos valores distinguiendo por universidad y si los resultados se han obtenido en la fase general o en la fase específica. Tenemos unos datos agregados. Conocemos unos resúmenes de los datos pero no los datos mismos.*[8]

[8] No es tan extraña esta situación. Con frecuencia cuando lees un informe o bien un artículo científico no sueles disponer de los datos originales sino de los resúmenes que de los mismos proporcionan los autores en la publicación. Tiene sentido e interés ver cómo calcular estos intervalos a partir de los datos resumidos.

Universidad	Matric.	Present.	Aptos	Media	DE
UA	783	771	401	4,989	2,124
UJI	530	518	249	4,857	2,088
UMH	693	673	262	4,369	1,922
UPV	1073	1049	589	5,274	2,070
UV	1525	1488	851	5,212	2,053
SUV	4604	4499	2352	5,021	2,077

Universidad	Present. FG	Present. FE	Aprob. FE
UA	276	495	254
UJI	213	305	148
UMH	267	406	154
UPV	422	627	357
UV	546	942	537
SUV	1724	2775	1450

Universidad	Media FG	DE FG	Media FE	DE FE
UA	5,053	1,839	4,954	2,267
UJI	4,852	1,874	4,860	2,225
UMH	4,304	1,812	4,412	1,989
UPV	5,135	1,959	5,367	2,137
UV	5,170	1,89	5,237	2,141
SUV	4,969	1,909	5,054	2,174

Cuadro 5.1: Resumen de los resultados de Matemáticas II. Las etiquetas indican: *Matric.*, matriculados; *Present.*,presentados; *Aptos*, aptos; *Media*, nota media; *DE*, desviación estándar; *Present. FG*, presentados fase general; *Present. FE*, presentados fase específica; *Aprob. FE*, aprobados fase específica; *Media FG*, media fase general; *DE FG*, desviación típica fase general; *Media FE*, media fase específica; *DE FG*, desviación típica fase específica. En filas tenemos las universidades de Alicante (UA), la Jaume I de Castellón (UJI), la Miguel Hernández de Elche (UMH), la Universidad de Valencia (UV) y todos los estudiantes en el Sistema de Universidades Valencianas (SUV).

Utilizando los datos de la tabla 5.1 vamos a calcular intervalos de confianza para la nota media en cada una de las universidades y en el total de todas las universidades. Vamos a centrarnos en los resultados globales, esto es, la última línea de la tabla. Empezamos construyendo un intervalo de confianza para la media global. En el siguiente código

```
media = 5.021
desviacion.estandar = 2.077
n = 4499
alpha = .01
(extremoinferior =  media - qt(1-alpha/2,df=n-1) *
 desviacion.estandar/ sqrt(n))

## [1] 4.941204

(extremosuperior =  media + qt(1-alpha/2,df=n-1) *
 desviacion.estandar / sqrt(n))

## [1] 5.100796
```

De modo que el intervalo de confianza viene dado por [4.941,5.101].
Y ahora para la nota en la fase general.

```
media = 4.969
desviacion.estandar = 1.909
n = 1724
alpha = .01
(extremoinferior =
 media - qt(1-alpha/2,df=n-1) *  desviacion.estandar/ sqrt(n))

## [1] 4.850441

(extremosuperior =
 media + qt(1-alpha/2,df=n-1) * desviacion.estandar / sqrt(n))

## [1] 5.087559
```

El intervalo es [4.85, 5.088]. Y terminamos con la media en la fase específica.

```
media = 5.054
desviacion.estandar = 2.174
n = 2775
alpha = .01
(extremoinferior = media - qt(1-alpha/2,df=n-1) *
 desviacion.estandar/ sqrt(n))

## [1] 4.947624

(extremosuperior = media + qt(1-alpha/2,df=n-1) *
 desviacion.estandar / sqrt(n))

## [1] 5.160376
```

5.6.3 Ejercicios

Ejercicio 5.1. *Determinar el intervalo de confianza para la concentración media de escandio en Finlandia y Rusia. Se pide utilizar como niveles de confianza* 0,90, 0,95 *y* 0,99.

Ejercicio 5.2. *Determinar el intervalo de confianza para la concentración media de escandio en toda la zona de estudio, esto es, considerando los tres países conjuntamente. Se pide utilizar como niveles de confianza* 0,90, 0,95 *y* 0,99.

Ejercicio 5.3. *En este ejercicio utilizamos los datos de la tabla 5.1. Se pide:*

1. *Determinar los intervalos de confianza con niveles de confianza* 0,95 *y* 0,99 *para la nota media* en la fase específica *en cada una de las cinco universidades de la Comunidad Valenciana.*

2. *Determinar los intervalos de confianza con niveles de confianza* 0,95 *y* 0,99 *para la nota media* en la fase general *en cada una de las cinco universidades de la Comunidad Valenciana.*

5.7 Error absoluto y tamaño de la muestra

En esta sección nos planteamos (por primera vez) un problema de cálculo de tamaño de la muestra. ¿Cuántos datos hemos de tener para que nuestro estudio tenga validez? Es una pregunta muy genérica sin respuesta. La respuesta necesita que concretemos más lo que queremos de nuestros datos. En esta sección nos planteamos la siguiente pregunta: ¿Cuántos datos necesitamos para que cuanto estimamos una media poblacional el error máximo que cometemos sea menor que una cantidad que previamente especificamos? Por ejemplo, queremos conocer la concentración media de un elemento en una zona. Queremos conocer esta cantidad, denotada por μ, con un error máximo de δ unidades siendo δ una cantidad positiva que fijamos nosotros. Lo primero que hemos de tener en cuenta es que en Estadística nunca podemos afirmar *con seguridad* nada. Podemos pedir que sea muy probable o, mejor, que tengamos *mucha confianza* en que ocurra pero que *seguro* que ocurra es mucho pedir. Siempre hacemos afirmaciones basadas en la probabilidad de sucesos que pueden o no ocurrir y, por lo tanto, afirmar que nuestro error va a ser menor que un cierto nivel δ *siempre* o *seguro* no es posible.

Vamos a responder a la pregunta anterior utilizando los datos *chorizon* del proyecto Kola. Pretendemos estimar la concentración media μ de escandio en Noruega (dentro del proyecto Kola). Ya hemos determinado un intervalo de confianza para μ con un nivel $1-\alpha$. Por ejemplo, con $\alpha = 0{,}05$ el intervalo es

```
t.test(Sc.Nor,conf.level=.95)$conf.int
```

```
## [1] 2.479107 3.078706
## attr(,"conf.level")
## [1] 0.95
```

Tenemos una confianza de 0,95 de que el valor de μ esté dentro (esté cubierto por) el intervalo anterior. Asumamos que la cosa ha ido bien, esto es, asumamos que μ está dentro. No lo sabemos pero *confiamos* con un nivel de confianza 0,95 que esto sea así. Pues bien, lo asumimos. A la pregunta: ¿en cuánto estimas μ?, respondemos (sin dudar) que en

```
mean(Sc.Nor)
```

```
## [1] 2.778906
```

esto es, respondemos dando el valor de la media muestral $\bar{x}$ de las concentraciones utilizadas para calcular el intervalo. Esta media muestral es el punto medio del intervalo. Por lo tanto si μ está en intervalo entonces la diferencia entre μ y el centro es menor o igual que la mitad de la longitud de dicho intervalo que vale

```
## [1] 0.5995994
```

¿Qué error absoluto tenemos con la muestra que se tomó? Guardamos el intervalo de confianza en *Sc.ci*.

```
Sc.ci  = t.test(Sc.Nor,conf.level=.95)$conf.int
```

Sabemos que el error absoluto es la mitad de la longitud del intervalo de confianza. Lo calculamos.

```
(Sc.ci[2] - Sc.ci[1]) / 2
```

```
## [1] 0.2997997
```

Supongamos que nos parece excesivo y pretendemos un error absoluto máximo de 0,2.

¿Cuál era el tamaño muestral en Noruega? ¿Cuántas muestras se han tomado en Noruega?

```
length(Sc.Nor)
```

```
## [1] 128
```

De hecho, es más fácil saber el número de muestras que hemos tomado en cada país.

```r
table(COUN)

## COUN
## FIN NOR RUS
## 187 128 290
```

El intervalo de confianza para la media de una población normal viene dado por la siguiente expresión.

$$\left[\bar{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

En consecuencia, el error absoluto vendrá dado por

$$t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

Esta expresión en R se calcula con el siguiente código.

```r
alpha = .05
n = length(Sc.Nor)
qt(1-alpha/2,df=n-1)*sd(Sc.Nor)/sqrt(n)

## [1] 0.2997997
```

Suponemos que la desviación estándar no se va a modificar mucho si tomamos más muestras.

```r
sd0 = sd(Sc.Nor)
```

Fijamos un valor para el error que queremos cometer.

```r
delta = 0.2
```

Se tiene que verificar

$$t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \delta$$

Si asumimos que s es aproximadamente constante.

```r
m = n + 10
qt(1-alpha/2,df=m-1)*sd(Sc.Nor)/sqrt(m)

## [1] 0.2885306
```

Vemos que no es suficiente. ¿Y 100 observaciones más?

```
m = n + 100
qt(1-alpha/2,df=m-1)*sd(Sc.Nor)/sqrt(m)

## [1] 0.2236826
```

¿Y 200 observaciones más?

```
m = n + 200
qt(1-alpha/2,df=m-1)*sd(Sc.Nor)/sqrt(m)

## [1] 0.1861879
```

Quizás nos hemos pasado. Veamos con 150.

```
m = n + 150
qt(1-alpha/2,df=m-1)*sd(Sc.Nor)/sqrt(m)

## [1] 0.2023752
```

Una manera de hacerlo sería

```
for(m in 280:290)
  print(c(m,qt(1-alpha/2,df=m-1)*sd(Sc.Nor)/sqrt(m)))

## [1] 280.0000000   0.2016448
## [1] 281.0000000   0.2012826
## [1] 282.0000000   0.2009223
## [1] 283.0000000   0.2005639
## [1] 284.0000000   0.2002074
## [1] 285.0000000   0.1998529
## [1] 286.0000000   0.1995002
## [1] 287.0000000   0.1991493
## [1] 288.0000000   0.1988003
## [1] 289.0000000   0.1984532
## [1] 290.0000000   0.1981078
```

5.7.1 *Ejercicios*

Ejercicio 5.4. *Utilizamos los datos* chorizon *del paquete* StatDA. *Se quiere estimar la la concentración media de escandio en Finlandia y en Rusia. Queremos estimar estas medias con un error máximo de* 0,20.

1. *Determinar el número de datos (tamaño de la muestra) que necesitamos en Finlandia para tener un error máximo de* 0,20. *¿Y si queremos un error máximo de* 0,1*?*
2. *Repetir el apartado 1 para Rusia.*

Ejercicio 5.5. *Supongamos que asumimos que el nivel medio de escandio no es distinto para los distintos países y por lo tanto usamos todos los valores. ¿Cuál es el error máximo observado?*

5.8 Estimación de la varianza en poblaciones normales

Si $X_1, \ldots, X_n$ es una muestra aleatoria de una normal con media μ y varianza σ^2 entonces

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X}_n)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Estamos diciendo que la variable aleatoria $\frac{(n-1)S^2}{\sigma^2}$ tiene una distribución ji-cuadrado con n-1 grados de libertad. [9] En la figura 5.6 hemos representado la función de densidad de una ji-cuadrado con 10 grados de libertad.

Con objeto de ver cómo cambia la forma de la función de densidad cuando cambia el número de grados de libertad en la figura 5.7 mostramos las densidades de la ji-cuadrado con 10 grados de libertad (trazo continuo) y con 20 grados de libertad (trazo discontinuo).

Denotamos el percentil de orden p de una ji-cuadrado con k grados de libertad como $\chi_{p,k}$. Es decir, si la variable X se distribuye como una ji-cuadrado con k grados de libertad, $X \sim\sim \chi^2_k$, entonces

$$P(X \leq \chi^2_{p,k}) = p.$$

El valor anterior lo podemos obtener con la función *qchisq*.

```
p = 0.75
k = 13
qchisq(p,df=k)

## [1] 15.98391
```

Entonces tenemos que

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

Y por lo tanto,

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

Es decir el intervalo

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

[9] La densidad de una distribución ji-cuadrado con k grados de libertad es

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \text{ para } x \geq 0 \text{ y cero en otro caso}$$

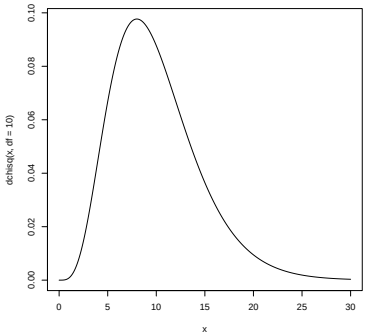


Figura 5.6: Función de densidad de una ji-cuadrado con 10 grados de libertad.

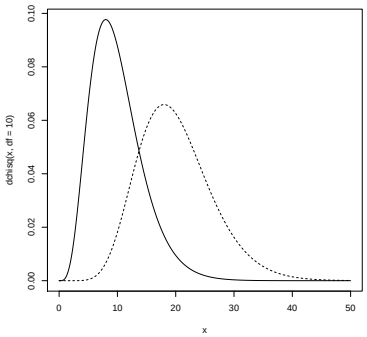


Figura 5.7: Función de densidad de una ji-cuadrado con 10 grados de libertad (trazo continuo) y con 20 grados de libertad (trazo discontinuo).

es un intervalo de confianza para σ^2 con nivel de confianza $1-\alpha$.

Es claro que el correspondiente intervalo de confianza (con nivel de confianza $1-\alpha$) para la desviación estándar poblacional vendrá dado por

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}\right]$$

Veamos cómo hacerlo con R. La función que nos da los percentiles de la ji-cuadrado es *pchisq*. Tomamos los datos.

```
n = 100
x = sample(X,100)
```

Y construimos el intervalo de confianza.

```
alpha = .05
s2 = var(x)
(extremoinferior = (n-1)*s2 / qchisq(1-alpha/2,df=n-1))

## [1] 20.83801

(extremosuperior = (n-1)*s2 / qchisq(alpha/2,df=n-1))

## [1] 36.47791
```

El intervalo de confianza para σ^2 con un nivel de confianza de 0,95 es $[20{,}838, 36{,}478]$.

5.8.1 Ejercicios

Ejercicio 5.6. *Consideremos los datos StatDA::chorizon. Se pide:*

1. *Utilizando el código*

   ```
   help(chorizon)
   ```

   consulta qué tipo de datos son.

2. *Calcular un intervalo de confianza para la varianza de la concentración de níquel en Rusia con un nivel de confianza de* 0,99.

3. *Repite el apartado anterior utilizando un nivel de confianza de* 0,9. *¿Qué intervalo es más grande? ¿Por qué?*

4. *Repite los apartados 2 y 3 para el nivel medio de níquel en Finlandia y para el nivel medio de níquel en Noruega.*

Ejercicio 5.7. *Con los datos de la tabla 5.1 se pide:*

1. *Construir un intervalo de confianza para la varianza y otro para la desviación estándar de la nota en Matemáticas II.*

2. *Repetir el apartado 1 para la varianza y la desviación estándar de la nota de Matemáticas II en la fase general.*

3. *Repetir el apartado 1 para la varianza y la desviación estándar de la nota de Matemáticas II en la fase específica.*

5.9 Estimación de una proporción

Supongamos que denotamos por p la proporción a estimar (proporción de aprobados en un examen, proporción de votantes de un cierto partido político). Realmente disponemos (antes de tomar los datos) de una muestra aleatoria $X_1, \ldots, X_n$ donde cada valor aleatorio X_i puede tomar los valores 1 y cero con probabilidades p y $1 - p$. El estimador de p viene dado por

$$\hat{p} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Por el teorema central del límite sabemos que aproximadamente (cuando el número de datos n es grande) $\hat{p}$ tiene una distribución normal con media p y varianza $p(1-p)/n$, es decir,

$$\hat{p} \sim N(p, p(1-p)/n).$$

Esto lo podemos escribir como

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1).$$

Sin embargo, esto no es utilizable para determinar un intervalo de confianza. Una opción es estimar la varianza mediante $\hat{p}(1-\hat{p})/n$ y considerar que aproximadamente (lo cual quiere decir que hemos de tener una muestra grande) se verifica

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1).$$

Si denotamos (como hemos indicado antes) el percentil de orden γ ($0 < \gamma < 1$) de una normal estándar como Z_γ entonces se verifica para un α dado

$$P\left(-Z_{1-\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq Z_{1-\alpha/2} \right) = 1 - \alpha.$$

o, escrito de otro modo, que

$$P\left(\hat{p} - Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) = 1 - \alpha.$$

Resultado 5.2 (Intervalo de confianza para una proporción)**.** *Si $X_1, \ldots, X_n$ son una muestra aleatoria de variables binomiales con una prueba y probabilidad de éxito p entonces el intervalo*

$$\hat{p} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

es un intervalo de confianza para la proporción p siendo

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Nota de R 5.3 (R y lo buena que es la aspirina)**.** *Vamos a obtener este intervalo de confianza utilizando R. Los datos con los que vamos a trabajar consideran si una persona toma aspirina o placebo y si ha sufrido un ataque cardíaco o no. Aparecen en la tabla 5.2.*

	Ataque fatal y no fatal	No ataque
Placebo	189	10845
Aspirina	104	10933

Cuadro 5.2: Efecto preventivo de la aspirina

Vamos a estimar la proporción de personas que tomando placebo tienen un ataque cardíaco.

```
library(Hmisc,T)
binconf(x=189, n=11034, method="asymptotic")

##    PointEst      Lower      Upper
##  0.01712887 0.01470788 0.01954987
```

Del mismo modo podemos estimar la proporción de los que, tomando aspirina, tienen ataque cardíaco. La estimación puntual y el intervalo de confianza vienen dados en la siguiente salida.

```
binconf(x=104, n=11037, method="asymptotic")

##    PointEst       Lower      Upper
##  0.00942285 0.007620423 0.01122528
```

En principio parece que si tomas aspirina tienes una probabilidad menor de tener el ataque. En cualquier caso en lo que sigue veremos el problema de comparar las proporciones.

5.9.1 Ejercicios

Ejercicio 5.8. *Para los datos de la tabla 5.1 se pide:*

1. *Estimar la proporción de aptos para cada una de las universidades y para todo el sistema universitario valenciano.*

Ejercicio 5.9. *Wilcox [2009, pág. 120, problema 21] Se observan los siguientes éxitos y fracasos: 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0. Calcule un intervalo de confianza con nivel 0,95 para la probabilidad de éxito p.*

Ejercicio 5.10. *Wilcox [2009, pág. 120, problema 22] Teniendo en cuenta los siguientes resultados para una muestra de una distribución binomial, calcule el error estándar de $\hat{p}$ cuando:*

1. $n = 25$, $X = 5$.
2. $n = 48$, $X = 12$.
3. $n = 100$, $X = 80$.
4. $n = 300$, $X = 160$.

Ejercicio 5.11. *Wilcox [2009, pág. 120, problema 23] Entre los 100 adultos seleccionados al azar, 10 se encontraron desempleados. Dar una intervalo de confianza con un nivel de 0,99 para el porcentaje de adultos desempleados.*

Ejercicio 5.12. *Wilcox [2009, pág. 121, problema 31] Una compañía de cosméticos encontró que 180 de cada 1000 mujeres seleccionadas al azar en Nueva York han visto el anuncio de televisión de la empresa. Calcule un intervalo de confianza al 0,95 para el porcentaje de mujeres en la ciudad de Nueva York que han visto el anuncio.*

5.10 Tamaño de la muestra en la estimación de una proporción

En la sección anterior hemos visto un intervalo de confianza para estimar una proporción. Es el intervalo $\left[\hat{p} - Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$. Este intervalo tiene nivel de confianza $1-\alpha$. Dados unos datos tenemos el intervalo. Supongamos que la cosa ha ido bien: El intervalo que calculamos cubre al valor verdadero de p. Tenemos una confianza $1-\alpha$ de que esto sea cierto. Bien. El intervalo cubre a p pero nuestra estimación puntual de p es $\hat{p}$. Si nos piden que demos un valor para p responderemos dando la estimación puntual, dando el valor $\hat{p}$. La diferencia entre la estimación puntual que damos y el valor real de p (que desconocemos y siempre desconoceremos) es como mucho la mitad de la longitud del intervalo de confianza (siempre asumiendo que p está dentro de este intervalo). En otras palabras:

$$|\hat{p} - p| \leq Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Un intervalo que acierte mucho es bueno, esto es, un nivel de confianza alto es bueno. De hecho, menos de un 90 % no se suele utilizar.

Pero no sólo importa el nivel de confianza. Por ejemplo, el intervalo $[0, 1]$ tiene un nivel de confianza de 1. ¿Cuál es la proporción de personas que sufren ataque de corazón tomando aspirina? Si respondemos que esta proporción está entre 0 y 1 estamos haciendo una afirmación con una confianza de uno. No nos equivocamos en absoluto. *Pero también la afirmación es absolutamente inútil.* No nos dice nada que no supieramos previamente. En resumen el intervalo que damos para estimar p tiene que ser tal que *confiemos* que contiene el valor que queremos conocer pero también tiene que ser preciso, tan estrecho como podamos para que sea informativo. Es fácil suponer que precisión supone más muestra, más datos.

El problema se plantea normalmente del siguiente modo. Quiero estimar la proporción p y quiero que hacerlo con un error máximo dado δ (por ejemplo, $\delta = 0{,}02$). Necesitamos una primera muestra. ¿Para qué? Para tener una estimación inicial de p. Denotemos el tamaño de esta primera muestra como n_0 y la estimación puntual de p obtenida como $\hat{p}_0$. Entonces podemos plantearnos qué valor de n verifica que

$$Z_{1-\alpha/2}\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}} \leq \delta$$

Esto es simple. Despejamos en la desigualdad anterior el valor de n y tenemos que

$$n \geq Z^2_{1-\alpha/2}\frac{\hat{p}_0(1-\hat{p}_0)}{\delta^2}$$

La estimación inicial de la proporción no necesariamente la tenemos que obtener de una muestra inicial. Puede que tengamos alguna muestra previa de esa población o alguna estimación que obtengamos de alguna publicación.

5.10.1 Ejercicios

Ejercicio 5.13. *Pretendemos estimar la proporción de palmeras afectadas por el picudo. Se ha tomado una primera muestra de 100 palmeras al azar. Se han observado 23 palmeras afectadas. Se pide:*

1. *¿Cuál es el error máximo observado con un nivel de confianza de* 0,95*?*
2. *Tomamos como estimación inicial de p el valor observado con la primera muestra de 100 palmeras. Supongamos que nos planteamos estimar la proporción de palmeras afectadas con un error máximo de* 0,04 *y un nivel de confianza de* 0,95*. ¿Cuál ha de ser el número total de palmeras a observar?*
3. *Responde a la pregunta del apartado 2 suponiendo que deseamos un nivel de confianza en el error máximo de* 0,99*. Mantenemos el error máximo en* 0,04*.*

4. *Responde a la pregunta del apartado 2 suponiendo que deseamos un nivel de confianza en el error máximo de* 0,95 *pero queremos un error máximo de* 0,02.

5. *¿Más nivel de confianza supone más muestra? Responde la pregunta considerando los apartados anteriores.*

6. *¿Menor error supone una mayor muestra? Responde la pregunta utilizando los apartados anteriores?.*

6
Contraste de hipótesis

6.1 Introducción

Se introduce el problema del contraste de hipótesis en una población.

En concreto, vamos a asumir que tenemos una muestra aleatoria de una distribución normal, $X_1, \ldots, X_n$, variables aleatorias independientes y con una misma distribución. La distribución común que asumimos es normal con media μ y varianza σ^2. Es decir, estamos asumiendo que

$$X_i \; N(\mu, \sigma^2)$$

y que los distintos valores son independientes entre si.

En este tema nos planteamos el problema del contraste de hipótesis y lo estudiamos estudiando, fundamentalmente, los contrastes sobre la media μ de la variable que observamos (concentración de contaminante, nivel de radiación o de ruido).

Asumimos normalidad en los datos con los que trabajamos. Pero: ¿es razonable está hipótesis? Nos ocupamos al final de este tema de lo que se conoce como contrastes de normalidad.

6.2 Constrastes para una muestra

Vamos a plantear el problema del contraste de hipótesis utilizando el problema de una muestra en poblaciones normales. Sobre aplicaciones medioambientales es interesante leer la introducción del capítulo 3 de Ginevan and Splitstone [2004] en donde se comentan distintas aplicaciones al contexto medioambiental de lo que tratamos en esta sección. También es interesante leer el capítulo 16 de Berthouex and Brown [2002].

Algunas preguntas que podremos responder son:

1. La contaminación observada en una localización: ¿es peligrosa para la salud?

2. ¿Está la descarga efectuada de acuerdo con las limitaciones establecidas?
3. El nivel de contaminantes: ¿es significativamente mayor que los niveles habituales?
4. ¿Se ha logrado limpiar suficientemente la zona contaminada?
5. En un control de calidad de un laboratorio, medimos la concentración de muestras que han sido preparadas o calibradas de un modo preciso. En estas muestras la concentración a medir se ha fijado de un modo preciso. Medimos la concentración en estas muestras utilizando un método analítico determinado y tenemos que comparar los valores medidos con la concentración previamente fijada. Estamos evaluando el funcionamiento del procedimiento de determinación, es pues, un control de calidad del laboratorio en cuestión.

Estas preguntas se pueden (y deben para poder responderlas) concretar en otras como:

1. ¿La concentración efluente media durante 30 días en la descarga de aguas residuales en un cierto emisario supera los 137 mg/l?[1]
2. ¿La concentración de torio en la superficie del suelo promediada sobre un cuadrado de 100 metros de lado es mayor de 10 picocuries por gramo?

[1] Siento obviamente este el máximo valor permitido.

6.2.1 *Un contraste unilateral*

Empezamos comentando un ejemplo que nos ayude a entender el problema del contraste de hipótesis.

Ejemplo 6.1 (Un problema inventado)**.** *Un fabricante de bombillas afirma que sus bombillas tienen una duración media de* al menos 1500 horas*. Muy bien, esta persona afirma esto. Nuestro problema es tomar una entre dos decisiones. Admitir que lo que afirma es correcto o bien que no lo es y la duración media real es claramente menor que 1500. Lo primero que necesitamos para tomar la decisión son datos. Hemos tomado una muestra de bombillas y hemos repetido el experimento consistente en tenerlas en funcionamiento ininterrumpido hasta que la bombilla deja de funcionar.*[2] *En definitiva tenemos una muestra aleatoria de duraciones de bombillas de este fabricante. Supongamos que las duraciones observadas son*

[2] Alguna bombilla dura algo más. http://www.centennialbulb.org/index.htm.

```
##   [1] 1518.3 1150.3 1586.5 1064.2 1457.0 1473.1  837.0
##   [8] 1789.1 1792.9 1931.5 1419.0 1256.1 1294.1 1809.1
##  [15] 1667.9 1754.3 1962.9 1593.9 1309.7 1599.1 1369.6
##  [22] 1606.6 1796.0 1913.1  991.1  986.1 1714.3 1263.1
```

```
##  [29] 1290.6 1381.3 1942.5 1585.7 2180.0 1603.4 1660.6
##  [36] 1732.5 1262.0 1729.9 1490.7 1366.7 2048.5 1838.2
##  [43] 2107.4 1755.3 1506.4 1895.4 1339.3 1736.7 1754.3
##  [50] 1430.9 2160.6 1738.7 1270.8 2093.7 1629.4 1456.1
##  [57] 1693.6 2153.2 1799.3 1345.3 2175.6 1939.3 1843.8
##  [64] 1083.0 2185.3 1743.1 1708.1 1385.4  706.9 1439.8
##  [71] 1417.3 1918.8 1555.6 1709.6 1743.1  829.8 1834.8
##  [78] 1364.1 1539.3 1716.0 1534.9 2136.2 1620.5 1233.1
##  [85] 1272.2 1486.9 1499.6 1489.4 1733.6 1928.9 1922.9
##  [92] 2243.8 1255.5 1394.3 1192.8 1327.6 1418.5 1909.1
##  [99] 1428.2 1796.4
```

La afirmación del fabricante la consideramos como una hipótesis que hemos de evaluar. *En concreto nos planteamos dicha hipótesis y su negación. De un modo más preciso la hipótesis de una duración media menor o igual a 1500 y su negación serían*

$$\begin{aligned} H_0: &\quad \mu \leq 1500, \\ H_1: &\quad \mu > 1500, \end{aligned}$$

donde H_0 y H_1 son las hipótesis nula y alternativa *respectivamente.*[3] *Hemos de elegir entre una de ellas. Elegir supone decidir. Hemos de decidir, con los datos que tenemos sobre las duraciones de las bombillas, cuál de las hipótesis es más compatible con los datos: la nula o la alternativa. Esto supone tomar una decisión: bien* rechazamos la hipótesis nula *o bien* no rechazamos la hipótesis nula.

[3] En algunos textos se denotan las hipótesis nula y alternativa como H_0 y H_a respectivamente.

¿Y cómo decidimos? El procedimiento que se utiliza es tomar una función de la muestra aleatoria, $T(X_1, \ldots, X_n)$, y en función de los valores que toma decidir. En este contraste, el estadístico habitualmente utilizado es el siguiente

$$T = \frac{\bar{X}_n - 1500}{S/\sqrt{n}}. \tag{6.1}$$

Si consideramos los valores observados tenemos

$$t_0 = \frac{\bar{x}_n - 1500}{s/\sqrt{n}}, \tag{6.2}$$

que toma el siguiente valor

```
(t0 = (mean(x) - 1500) / (sd(x)/sqrt(n)))

## [1] 2.955805
```

Si observamos la definición de T parece razonable definir como regla de decisión: rechazamos la hipótesis nula si

$$T \geq c$$

donde c es un valor que elegiremos adecuadamente. Si asumimos que la media poblacional es $\mu = 1500$*, es decir, el valor límite entre ambas hipótesis entonces se verifica que T (definido en 6.1) sigue una distribución t de Student con n-1 grados de libertad, es decir,*

$$T = \frac{\bar{X}_n - 1500}{S/\sqrt{n}} \sim t_{n-1}. \tag{6.3}$$

Supongamos que no queremos equivocarnos rechazando la hipótesis nula cuando es cierta más que un 5 % de las ocasiones. Al error que cometemos cuando hacemos esto se le llama error tipo I. *Esto supone que, colocándonos en la situación límite entre las hipótesis nula y alternativa, hemos de elegir el valor de la constante c de modo que*

$$P(T \geq c) = 0{,}05, \tag{6.4}$$

o, equivalentemente, que

$$P(T \leq c) = 1 - 0{,}05 = 0{,}95. \tag{6.5}$$

La constante c es el percentil o cuantil de orden 0,95 *de una distribución t de Student con n-1 grados de libertad que denotamos* $t_{n-1,0{,}95}$*. El valor de c lo podemos calcular con R con el siguiente código*

```
qt(.95,df=99)

## [1] 1.660391
```

ya que $n = 100$*. Ya lo tenemos todo. El valor observado de T es* t_0*.*

```
t0

## [1] 2.955805
```

Como t_0 *es mayor que* $t_{n-1,0{,}95}$ *rechazamos la hipótesis nula lo que indica que el fabricante no mentía: sus bombillas tienen una duración media superior a 1500.*

Gráficamente podemos representar lo que acabamos de hacer. En la figura 6.1 representamos la densidad cuando la media vale $\mu_0 = 1500$*. En línea discontinua más a la izquierda (trazo discontinuo) indicamos la posición del estadístico* t_0*.*

Y ahora planteamos el problema de un modo genérico. Consideramos el contraste de hipótesis.

$$\begin{aligned} H_0 &: \quad \mu \leq \mu_0, \\ H_1 &: \quad \mu > \mu_0. \end{aligned}$$

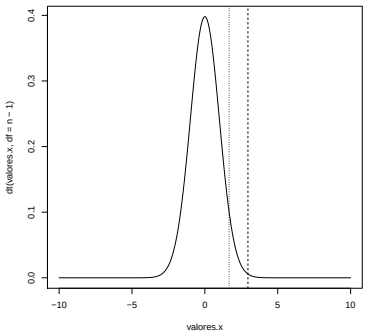


Figura 6.1: Contraste unilateral sobre la media.

Hemos de tomar una entre dos posibles decisiones: *rechazamos la hipótesis nula* o bien *no rechazamos la hipótesis nula*. Un problema de contraste de hipótesis consiste en tomar una decisión. Cuando decidimos nos equivocamos.[4] De hecho, nos podemos equivocar de dos modos. El primer tipo de error consiste en rechazar la hipótesis nula cuando realmente es cierta, el error tipo I. Es el error al que tradicionalmente se le ha prestado más atención. Se considera el *error a controlar*. Por ello, se le pone una cota, un valor máximo. Esta cota es el *nivel de significación* α. El valor más habitual para α es 0,05. Otros valores que habitualmente se utilizando son 0,01 o 0,1. El otro posible error consiste en no rechazar la hipótesis nula cuando realmente no es cierta. Es el error tipo II. Como todo error ha de ser pequeño. Este tipo de error lo podemos controlar utilizando muestras grandes. Fijamos un α y con más muestra tendremos un menor error tipo II.

[4] Algunos sostienen que lo mejor en la vida es no tomar decisiones. Que las cosas pasen y asumirlas.

El procedimiento para realizar un contraste de hipótesis es el siguiente:

1. Planteamos el contraste de hipótesis.
2. Fijamos un nivel de significación α.
3. Tomamos los datos.
4. Evaluamos el valor del estadístico.
5. Si el estadístico está en la región crítica rechazamos la hipótesis nula. En otro caso, no rechazamos dicha hipótesis nula.

	Realidad	
Decisión	H_0	H_1
Rechazamos H_0	Error tipo I	
No rechazamos H_0		Error tipo II

Cuadro 6.1: Errores que podemos cometer en un problema de contraste de hipótesis.

Suponemos fijado α. Calculamos el estadístico.

$$T = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \tag{6.6}$$

y el valor observado

$$t_0 = \frac{\bar{x}_n - \mu_0}{s/\sqrt{n}}. \tag{6.7}$$

Bajo la hipótesis de que la media poblacional μ vale μ_0, $\mu = \mu_0$, se tiene que

$$T = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \sim t_{n-1}, \tag{6.8}$$

y t_0 sería un valor observado de una variable aleatoria con distribución t con n-1 grados de libertad.

Supongamos que queremos (es una elección del decisor que somos nosotros) un error tipo I que sea menor o igual a α (habitualmente 0,05, 0,01 o 0,1) entonces la regla de decisión es:

- *Rechazamos* H_0 *si* $T > t_{n-1,1-\alpha}$, o dicho de otro modo, rechazamos si $T \in [t_{n-1,1-\alpha}, +\infty)$.
- En otro caso, no rechazamos H_0.

Si denotamos $C = [t_{n-1,1-\alpha}, +\infty)$ entonces estamos rechazando cuando el valor del estadístico está en C. Rechazamos H_0 si $T \in C$ y no la rechazamos en otro caso. A este intervalo le llamamos *región crítica*.

Lo que acabamos de hacer se puede hacer de otro modo.

1. Supongamos que calculamos el área a la derecha de t_0 en una t de Student con $n-1$ grados de libertad. Este valor lo vamos a denotar con la letra p y recibe el nombre de *p-valor*. Es decir, el p-valor viene dado por

$$p = P(T \geq t_0) \text{d}onde T \sim t_{n-1}.$$

   El p-valor lo podemos calcular con el siguiente código.

```
(pvalor=1-pt(t0,df=n-1))

## [1] 0.001949062
```

2. Las dos afirmaciones siguientes son equivalentes:

   a) Se verifica que $t_0 > t_{n-1,1-\alpha}$ y por lo tanto rechazamos la hipótesis nula.

   b) El área a la derecha de t_0 sea menor que α.

Es equivalente que a que el área a la derecha de t_0 sea menor que el área a la derecha de $t_{n-1,1-\alpha}$. El área a la derecha de t_0 es el valor de p o el p-valor mientras que el área a la derecha de $t_{n-1,1-\alpha}$ es α. Por tanto, la misma regla de decisión que estamos empleando se puede formular como: *rechazamos la hipótesis nula* H_0 *si* $p < \alpha$ *y no rechazamos si* $p \geq \alpha$. En la figura 6.2 el p-valor es el área de la zona en color negro.

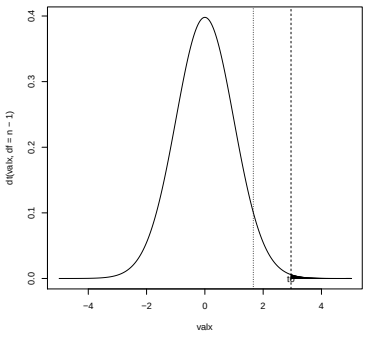


Figura 6.2: El p-valor corresponde con el área de la zona negra.

Nota de R 6.1 (Función t.test para contrastar)**.** *Vamos a realizar el contraste utilizando la función* t.test.

```
t.test(x,mu=1500,alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  x
## t = 2.9558, df = 99, p-value = 0.001949
## alternative hypothesis: true mean is greater than 1500
## 95 percent confidence interval:
##  1541.843      Inf
## sample estimates:
## mean of x
##  1595.474
```

Vemos cómo la función no nos da el valor a partir del cual rechazamos. Simplemente nos indica el p-valor y la hipótesis alternativa. ¿Qué hipótesis nula tenemos? La hipótesis nula la obtenemos por negación de la alternativa que nos da la salida. Y la decisión la hemos de tomar nosotros que, al fin y al cabo, somos el decisor. *El nivel de significación α lo hemos elegido a priori. Si trabajamos con un nivel de significación $\alpha = 0{,}05$ entonces podemos ver en la salida anterior que dicho p-valor es menor que 0,05 y rechazamos la hipótesis nula. Ahí se acaba la historia. Un contraste de hipótesis se acaba cuando se ha tomado una decisión.*

Ejemplo 6.2. *Vamos a fijarnos en la concentración de níquel en Rusia en los datos* chorizon *del paquete* StatDA **?**. *Vamos a suponer que una concentración media por encima de 20 es peligrosa. Con los datos observados, ¿podemos considerar que la concentración media es 20 o mayor? Formulamos el siguiente contraste. Hay que demostrar que efectivamente estamos por debajo del nivel de peligrosidad.*

$$H_0: \quad \mu \leq 20,$$
$$H_1: \quad \mu > 20.$$

Vamos a realizar el contraste. Cargamos los datos y seleccionamos las concentraciones relativas a Rusia.

```
load("../data/chorizon.rda")
attach(chorizon)
Ni.Rus = Ni[which(COUN == "RUS")]
```

Ahora podemos aplicar el test.

```
t.test(Ni.Rus,alternative="greater",mu=20)

##
##  One Sample t-test
```

```
##
## data:  Ni.Rus
## t = 2.8777, df = 289, p-value = 0.002152
## alternative hypothesis: true mean is greater than 20
## 95 percent confidence interval:
##  21.69218      Inf
## sample estimates:
## mean of x
##   23.9669
```

Vemos que el p-valor vale o.9978 que es menor que 0,05. *Rechazamos la hipótesis nula.*

6.2.2 *Otro problema de contraste unilateral*

Supongamos que una empresa vierte sus residuos a un rio. La concentración media de una cierta sustancia contaminante no puede superar un cierto valor μ_0. ¿Cómo formulamos el contraste y cómo contrastamos las hipótesis?

El contraste lo podemos formularíamos como:

$$\begin{aligned} H_0: &\quad \mu \geq \mu_0, \\ H_1: &\quad \mu < \mu_0. \end{aligned}$$

La empresa debe de probar que las concentraciones medidas en las muestras de agua no superan en media el valor que se le indica. El contraste se puede hacer con dos procedimientos equivalentes.

Primer procedimiento

1. En primer lugar elegimos el nivel de significación α.
2. Calculamos el estadístico $T = t_0$ definido en ecuación 6.6.
3. Rechazamos la hipótesis nula si $t_0 < t_{n-1,\alpha}$ y no rechazamos en otro caso.

Segundo procedimiento

1. Elegimos el nivel de significación α.
2. Calculamos el estadístico $T = t_0$ definido en ecuación 6.6.
3. Calculamos el área a la izquierda de t_0 en una t de Student con n-1 grados de libertad. Este es el valor p o p-valor.
4. Rechazamos la hipótesis nula si $p < \alpha$ y no rechazamos en otro caso.

Hagámoslo con R. Supongamos que el valor por debajo del cual ha de mantenerse en media es $\mu_0 = 34$ unidades (en la unidad que queráis).

Supongamos que los datos con los que trabajamos son los siguientes:

```
##   [1] 20.973 39.539 32.493 26.072 41.707 36.018 29.690
##   [8] 33.443 29.465 36.475 26.587 26.519 25.945 29.246
##  [15] 32.949 35.980 30.122 19.944 27.732 32.752 34.471
##  [22] 31.101 30.948 34.889 38.423 15.096 30.686 25.827
##  [29] 35.518 20.583 30.270 40.163 24.304 29.901 27.796
##  [36] 29.757 31.565 33.469 43.764 38.501 29.701 25.604
##  [43] 29.370 27.745 30.444 36.138 29.218 42.374 33.488
##  [50] 26.963 43.596 37.485 31.784 29.643 31.317 36.104
##  [57] 29.045 23.118 36.749 21.059 28.418 31.972 28.440
##  [64] 31.853 32.684 25.085 17.989 34.004 30.220 33.363
##  [71] 32.711 33.406 30.941 33.308 37.603 35.756 34.475
##  [78] 32.479 31.590 38.261 23.675 27.966 25.178 28.546
##  [85] 32.591 35.659 44.179 33.872 26.764 34.152 35.569
##  [92] 35.228 34.226 36.482 26.873 21.433 34.175 31.271
##  [99] 32.274 44.706 43.041 33.643 33.492 32.990 23.928
## [106] 37.604 20.463 31.747 34.955 23.572 34.865 20.958
## [113] 38.566 26.956 32.186 25.578 22.428 27.077 16.561
## [120] 27.135 31.037 28.506 25.053 33.964 22.703 29.521
## [127] 28.459 31.293 33.983 37.472 31.521 35.512 29.767
## [134] 32.345
```

Tenemos $n = 134$. Tomamos $\alpha = 0{,}05$. El estadístico vale

```
n = 134
mu0 = 34
(t0 = (mean(x) - mu0) / (sd(x)/sqrt(n)))

## [1] -5.795389
```

Determinamos el punto a partir del cual rechazamos: $t_{n-1,\alpha}$

```
alpha = 0.05
qt(alpha,df=n-1)

## [1] -1.656391
```

Como $t_0 < t_{n-1,\alpha}$ entonces rechazamos la hipótesis nula.

El segundo procedimiento es más simple de aplicar y es el que usaremos.

```
t.test(x,mu=mu0,alternative="less")

##
```

```
##  One Sample t-test
##
## data:  x
## t = -5.7954, df = 133, p-value = 2.35e-08
## alternative hypothesis: true mean is less than 34
## 95 percent confidence interval:
##      -Inf 31.94208
## sample estimates:
## mean of x
##  31.11851
```

Como el p-valor es notablemente menor que 0,05 rechazamos la hipótesis nula.

6.2.3 Y, finalmente, el contraste bilateral

Suponemos que queremos valorar si la media de la variable de interés podemos considerar que toma un valor próximo a μ_0. ¿Cómo formulamos el contraste? El contraste lo formularíamos como:

$$\begin{aligned} H_0: &\quad \mu = \mu_0, \\ H_1: &\quad \mu \neq \mu_0. \end{aligned}$$

El contraste se puede hacer con dos procedimientos equivalentes.

Primer procedimiento

1. En primer lugar elegimos el nivel de significación α.
2. Calculamos el estadístico $T = t_0$ (definido en ecuación 6.6).
3. Rechazamos la hipótesis nula si $t_0 < 1 - t_{n-1,1-\alpha/2}$ o bien si $t_0 > t_{n-1,1-\alpha/2}$ y no rechazamos en otro caso.

Segundo procedimiento

1. Elegimos el nivel de significación α.
2. Calculamos el estadístico $T = t_0$ (definido en ecuación 6.6).
3. Calculamos el área a la izquierda de $-|t_0|$ más a la derecha de $|t_0|$ en una t de Student con n-1 grados de libertad. Este es el valor p o **p-valor**.
4. Rechazamos la hipótesis nula si $p < \alpha$ y no rechazamos en otro caso.

Ejemplo 6.3. *Vamos a trabajar con las concentraciones de níquel en Rusia en los datos* chorizon **?**. *Supongamos que nos planteamos contrastar si estamos alrededor de una concentración de 23. El contraste es pues el siguiente:*

$$\begin{aligned} H_0: &\quad \mu = 23, \\ H_1: &\quad \mu \neq 23. \end{aligned}$$

```
t.test(Ni.Rus,alternative="two.sided",mu=23)

##
##  One Sample t-test
##
## data:  Ni.Rus
## t = 0.70141, df = 289, p-value = 0.4836
## alternative hypothesis: true mean is not equal to 23
## 95 percent confidence interval:
##  21.25373 26.68006
## sample estimates:
## mean of x
##   23.9669
```

Con el p-valor 0.4836 no podemos rechazar la hipótesis nula a ninguno de los niveles de significación habituales de 0,01 *o* 0,05 *o* 0,1.

¿Y de 24? El contraste es ahora:

$$\begin{aligned} H_0: &\quad \mu = 24, \\ H_1: &\quad \mu \neq 24. \end{aligned}$$

Realizamos el contrate.

```
t.test(Ni.Rus,alternative="two.sided",mu=24)

##
##  One Sample t-test
##
## data:  Ni.Rus
## t = -0.024014, df = 289, p-value = 0.9809
## alternative hypothesis: true mean is not equal to 24
## 95 percent confidence interval:
##  21.25373 26.68006
## sample estimates:
## mean of x
##   23.9669
```

El p-valor es todavía mayor, 0.9809. No rechazamos la hipótesis nula con un nivel de significación de 0,05. *De hecho, tampoco con un nivel de significación de* 0,1 *Notemos que el argumento* alternative *toma el valor* two.sided *por defecto. @*

6.3 *Intervalo de confianza y contraste de hipótesis*

Hemos estudiado cómo estimar, mediante un intervalo de confianza, la media de una variable normal. El intervalo de confianza *con un nivel de confianza* $1-\alpha$ viene dado por $[\bar{x}_n - t_{n-1,1-\alpha/2}s/\sqrt{n}, \bar{x}_n - t_{n-1,1-\alpha/2}s/\sqrt{n}]$. Supongamos que tomamos μ_0 en este intervalo. Nos planteamos el siguiente contraste:

$$\begin{aligned} H_0: &\quad \mu = \mu_0, \\ H_1: &\quad \mu \neq \mu_0 \end{aligned}$$

y pretendemos contrastar estas hipótesis *con un nivel de significación* α. Si calculamos el valor del estadístico $t_0 = (\bar{x}_n - \mu_0)/(s/\sqrt{n})$ tenemos que se verifica que $|t_0| \leq t_{n-1,1-\alpha/2}$ y por lo tanto no rechazamos la hipótesis nula consistente en que la verdadera media de la población μ toma el valor μ_0.

Este resultado lo hemos formulado referido a la media de una población normal. Sin embargo, es un resultado válido en general. Si suponemos que tenemos un parámetro de toda la población (media, varianza, desviación estandar en poblaciones normales o bien la probabilidad de éxito en poblaciones binomiales) que vamos a denotar de un modo genérico como θ. Tenemos un intervalo de confianza (con un nivel de confianza $1-\alpha$) para θ que depende de la muestra $\{x_1, \ldots, x_n\}$ y que denotamos por $[A(x_1, \ldots, x_n), B(x_1, \ldots, x_n)]$ o simplemente como $[A, B]$. Entonces, si $A \leq \theta_0 \leq B$ y consideramos el contraste bilateral

$$\begin{aligned} H_0: &\quad \theta = \theta_0, \\ H_1: &\quad \theta \neq \theta_0 \end{aligned}$$

no rechazaremos la hipótesis nula con un nivel de significación α. Y viceversa: si consideremos todos los valores θ_0 para los cuales no rechazamos la hipótesis nula (con un nivel de significación α) en el contraste formulado anteriormente tenemos un intervalo de confianza para θ con nivel de confianza $1-\alpha$.

Ejemplo 6.4. *Vamos a ilustrar con datos lo dicho. Recuperamos el contraste de si la media de níquel en Rusia es de* 23.

```
t.test(Ni.Rus,alternative="two.sided",mu=23,conf.level=0.95)

##
##  One Sample t-test
##
## data:  Ni.Rus
```

```
## t = 0.70141, df = 289, p-value = 0.4836
## alternative hypothesis: true mean is not equal to 23
## 95 percent confidence interval:
##  21.25373 26.68006
## sample estimates:
## mean of x
##   23.9669
```

Podemos comprobar que si tomamos como valores para la media μ cualquier valor μ_0 que esté en el intervalo de confianza el p valor que observaremos para el contraste $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ será mayor que $\alpha = 0{,}05$ y por lo tanto no la rechazamos.

Y al revés, si vamos contrastando $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ para todos los posibles valores de μ_0 y nos quedamos con los valores de μ_0 para los cuales no rechazamos con un nivel de significación α entonces tenemos un intervalo de confianza con nivel de confianza $1 - \alpha$.

6.4 Ejercicios

Ejercicio 6.1. *Berthouex and Brown [2002, ejercicio 16.1] Una empresa advierte que un producto químico tiene un 90 % de efectividad en la limpieza y cita como prueba que en una muestra de diez aplicaciones se observó un promedio de limpieza del 81 %. El gobierno dice que esto es publicidad engañosa porque el 81 % no igual al 90 %. La compañía dice que el valor observado es de 81 %, pero fácilmente podría ser del 90 %. Los datos observados fueron 92, 60, 77, 92, 100, 90, 91, 82, 75, 50. ¿Quién está en lo cierto y por qué?*

Ejercicio 6.2. *Berthouex and Brown [2002, ejercicio 16.2] Fermentación. El gas producido a partir de una fermentación biológica es puesto a la venta con la garantía de que el contenido medio en metano es de 72 %. Una muestra aleatoria de n = 7 muestras de gas dió un contenido de metano (en %) de 64, 65, 75, 67, 65, 74 y 75.*

1. *Llevar a cabo el contraste de hipótesis con niveles de significación de 0.10, 0.05 y 0.01 para determinar si es justo afirmar que el contenido medio es de 72 %.*

2. *Calcular los intervalos de confianza al 90 %, 95 % y 99 % para evaluar la afirmación que el promedio es de un 72 %.*

3. *¿Cuál es el error máximo observado en la estimación de la media con un nivel de confianza del 95 %?*

4. *Supongamos que el productor se compromete a que contiene al menos un 72 por cierto? Formula el contraste y realízalo.*

Ejercicio 6.3. *Berthouex and Brown [2002, ejercicio 16.3] Un protocolo de control de calidad en un laboratorio introduce soluciones estándar que contienen 50 mg / L de carbono orgánico total de un modo aleatorio en el trabajo del laboratorio. Los analistas del laboratorio desconocen estos estándares introducidos en su trabajo. Estima el sesgo y la precisión de las 16 observaciones más recientes de tales estándares. ¿Está bajo control el procedimiento de medida? Los valores observados fueron: 50.3 51.2 50.5 50.2 49.9 50.2 50.3 50.5 49.3 50.0 50.4 50.1 51.0 49.8 50.7 50.6*

Ejercicio 6.4. *Berthouex and Brown [2002, ejercicio 16.4] Permiso de Descarga. El permiso de descarga para una industria que requiere la DQO (demanda química de oxígeno) media mensual sea menor de 50 mg / L. La industria quiere que esto se interprete como que el valor 50 mg / L cae dentro del intervalo de confianza de la media que a su vez se calcula a partir de 20 observaciones por mes. Si los siguientes 20 valores observados son los siguientes: ¿Está cumpliendo la industria la norma?*

57 60 49 50 51 60 49 53 49 56 64 60 49 52 69 40 44 38 53 66

Ejercicio 6.5. *Un artículo publicado en Transactions of the American Fisheries Society recogía los resultados de un estudio para investigar las concentraciones de mercurio en róbalos de boca grande (con perdón). Se tomaron muestras de 53 lagos en Florida y se midió la concentración de mercurio en el tejido muscular (ppm). Los valores observados fueron:*

```
1.230, 1.330, 0.040, 0.044, 1.200, 0.270, 0.490,  0.190, 0.830, 0.810,
0.710, 0.500, 0.490, 1.160, 0.050, 0.150,  0.190, 0.770, 1.080, 0.980,
0.630, 0.560, 0.410, 0.730, 0.590, 0.340, 0.340,  0.840, 0.500, 0.340,
0.280, 0.340, 0.750, 0.870, 0.560,   0.170, 0.180, 0.190, 0.040, 0.490,
1.100, 0.160, 0.210, 0.860,  0.520, 0.650, 0.270, 0.940, 0.400, 0.430,
0.250, 0.270.
```

Se pide:

1. *Construir un intervalo de confianza para la media con nivel de confianza* 0,90.
2. *Contrastar la hipótesis nula de una concentración media menor o igual a* 0,7.
3. *Contrastar la hipótesis nula de una concentración media igual a* 0,7.

Ejercicio 6.6. *Se les envió a 14 laboratorios soluciones estandarizadas que fueron preparadas de modo que contenían cada una de ellas 1.2 mg/L de oxígeno disuelto (DO). Se les pidió que midieran la concentración de oxígeno disuelto utilizando el método* Winkler. *Las concentraciones obtenidas por cada uno de los laboratorios fueron las siguientes:*

```
1.2 1.4 1.4 1.3 1.2 1.35 1.4 2.0 1.95 1.1 1.75 1.05 1.05 1.4
```

Se pide:

1. *Construir un intervalo de confianza con nivel de confianza* 0,90 *para la concentración media de oxígeno disuelto.*
2. *Teniendo en cuenta el intervalo que hemos construido en el apartado anterior se pide responder la siguiente pregunta: ¿miden los laboratorios en promedio una concentración de 1.2 mg/L o, por el contrario hay un sesgo?*
3. *Contrastar la hipótesis de que la concentración media es igual a 1.2 mg/L con un nivel de significación* $\alpha = 0{,}1$.

6.5 Contraste de normalidad

Hemos vistos que, con mucha frecuencia, los procedimientos estadísticos que hemos utilizado (y mucho de lo que sigue) asumen que los datos que estamos manejando pueden considerarse una muestra de una distribución normal. ¿Y esto es así sin más? ¿Hemos de asumirlo y confiar en nuestra suerte? No. Podemos contrastar esta hipótesis. La hipótesis de que los datos que estamos usando siguen una distribución normal es una hipótesis a contrastar. Y estamos en condiciones de poder hacerlo. El contraste lo podemos formular de un modo informal como:

H_0: Tenemos una muestra de una distribución normal.

H_1: No tenemos una muestra de una distribución normal.

Quizás una formulación más formalista del contraste puede ser la siguiente donde X es el valor aleatorio que estamos observando n veces.

H_0: $X \sim N(\mu, \sigma^2)$ con $-\infty < \mu < +\infty$ y $\sigma^2 > 0$.

H_1: X no sigue una distribución normal.

Consideremos unos datos y veamos cómo evaluar si han sido generados según una distribución normal. Realmente vamos a considerar dos conjuntos de datos. Uno de ellos (que denotamos por x) sí que sigue una distribución normal mientras que la segunda muestra (que denotamos por y) no sigue una distribución normal. Vamos a estudiar la normalidad de estas dos muestras y recordemos que nos ha de salir siempre una valoración afirmativa para x y negativa para y.

6.5.1 Gráficos para evaluar la normalidad

Parece que lo primero es ver gráficamente cómo son estos datos. En esta sección vemos el uso del dibujo q-q o dibujo cuantil-cuantil.

Nota 6.1 (Estimando la densidad de puntos)**.** *Una primera opción (bastante natural pero nada aconsejable) para evaluar la normalidad es utilizar un histograma o un estimador kernel de la densidad. Nos dan una idea de cómo se distribuyen los puntos. En las figuras 6.3(a) y 6.3(b) mostramos el histograma y el estimador kernel de la densidad respectivamente para la muestra x. Las figuras 6.3(c) y 6.3(d) son las análogas para la muestra y. ¿Qué se espera ver en estas dos figuras si los datos son normales? La mejor respuesta a esta pregunta es otra pregunta: ¿se parecen las figuras a una densidad normal? Yo diría que las figuras 6.3(a) y 6.3(b) correspondientes a la muestra sí que recuerdan la forma de la densidad normal mientras que esto no es cierto para las figuras 6.3(c) y 6.3(d).*

No es muy fácil evaluar hasta qué punto es normal la muestra con aproximaciones de la densidad normal que es lo que son el histograma y el estimador kernel de la densidad. De hecho, este tipo de representaciones no son nada aconsejables para este propósito.

Hay una representación gráfica muy popular para este problema concreto: **el dibujo *q-q* o dibujo cuantil-cuantil**. ¿Qué se pretende evaluar? Si los datos pueden haber sido generados según un modelo normal con la media y varianza que sean. Si X es la variable que estamos observando (n veces), pretendemos evaluar hasta qué punto es cierto que

$$X \sim N(\mu, \sigma^2), \tag{6.9}$$

para algún μ y algún σ^2. Supongamos que es cierta la afirmación, suponemos cierta que la variable sigue una distribución normal. Elegimos una serie de probabilidades p_i.[5]. En concreto estos valores tienen la forma

$$p_i = \frac{i - \alpha}{n - 2\alpha + 1}, \tag{6.10}$$

donde $i = 1, \ldots, n$. Dos son los valores de α que suelen utilizarse [6]

$$\alpha = 0{,}375, \tag{6.11}$$

o bien

$$\alpha = 0{,}5. \tag{6.12}$$

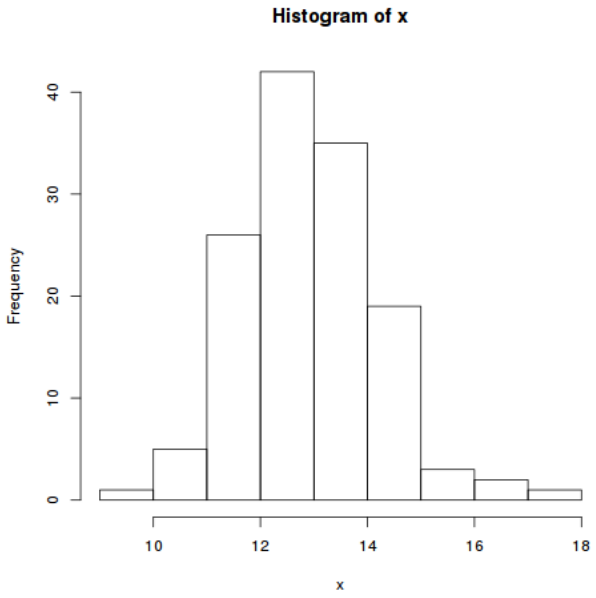
Una vez hemos elegido estos valores p_i hemos de determinar los valores de la abscisa y la ordenada del i-ésimo punto. Si x_i con $i = 1, \ldots, n$ son los datos entonces los ordenamos obteniendo los estadísticos ordenados $x_{(i)}$ que verifican

$$x_{(1)} \leq \ldots \leq x_{(n)}.$$

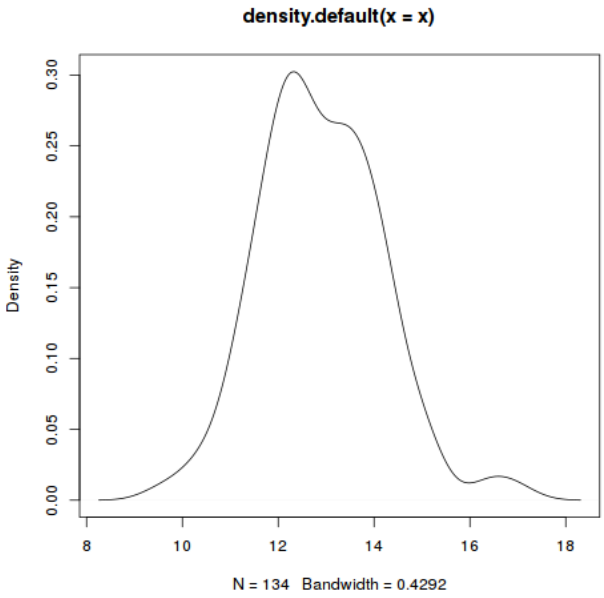
Notemos que el i-ésimo estadístico ordenado $x_{(i)}$ es aproximadamente el percentil de orden p_i. Ahora consideramos una distribución

[5] Una explicación detallada de cómo elegir estos valores lo podemos encontrar en o en **?**, páginas 18 y siguientes

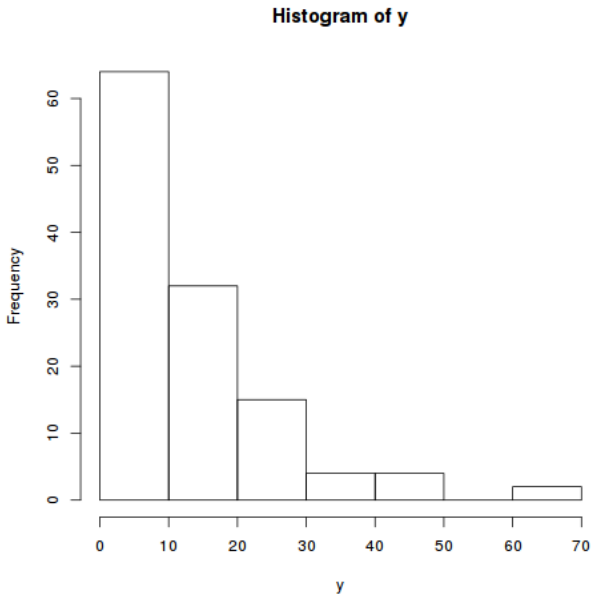
[6] La función *pppoints* nos indica los valores que realmente se utilizan.



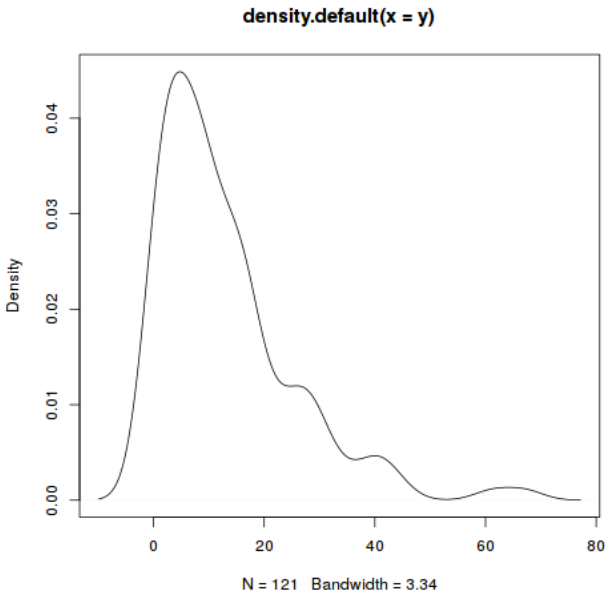
(a)



(b)



(c)



(d)

Figura 6.3: Datos x: histograma (a) y estimador kernel de la densidad de x (b). Datos y: histograma (c) y estimador kernel de la densidad de x (d).

normal estándar y consideramos el valor q_i tal que si Z es una variable aleatoria con distribución normal estándar entonces

$$p_i = P(Z \leq q_i) = \int_{-\infty}^{q_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

De hecho, es habitual denotar

$$\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Entonces

$$p_i = \Phi(q_i),$$

es decir,

$$q_i = \Phi^{-1}(p_i).$$

En el dibujo q-q representamos los puntos $(q_i, x_{(i)})$ con $i = 1, \ldots, n$.

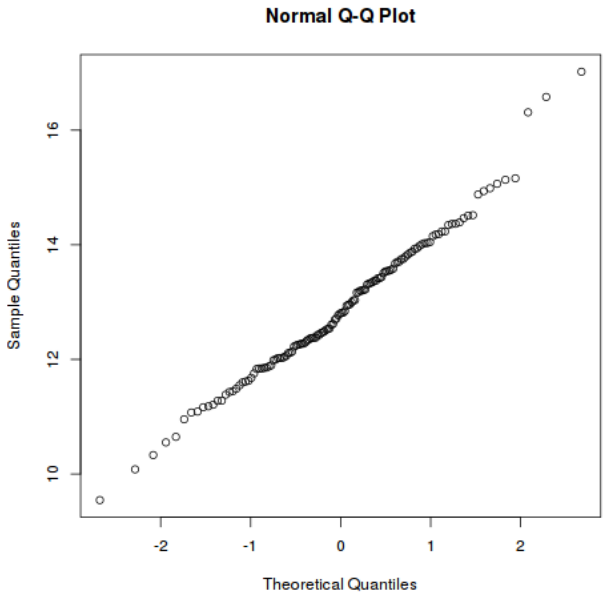
La nota 6.2 nos muestra cómo realizar el dibujo de un modo simple utilizando las funciones *qqnorm* y *qqline*.

Nota de R 6.2 (Dibujo q-q con las funciones *qnorm* y *qqline*)**.** *La función* qqnorm *nos proporciona el dibujo q-q fácilmente. En la figura 6.4(a) tenemos la representación gráfica para la muestra x.*

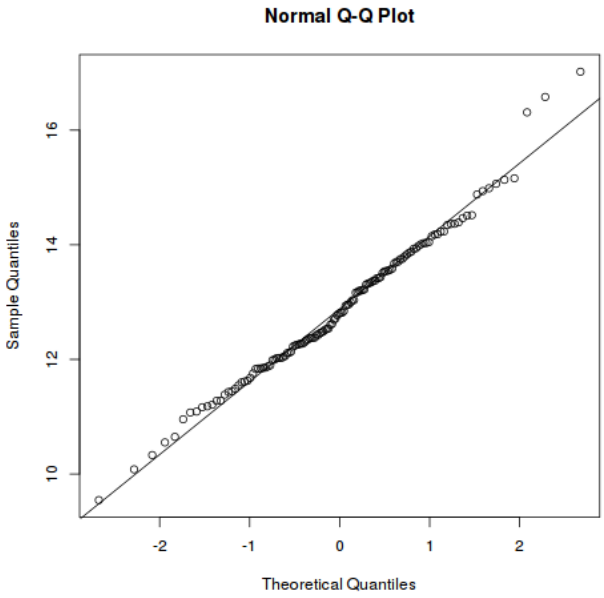
```
qqnorm(x)
```

La figura 6.4(c) muestra el mismo dibujo para la segunda muestra. ¿Podemos considerar que los puntos de la figura 6.4(a) están sobre una línea recta, están alineados? Yo diría que sí. ¿Y los puntos de la figura 6.4(c)? Creo que coincidiríamos que no.

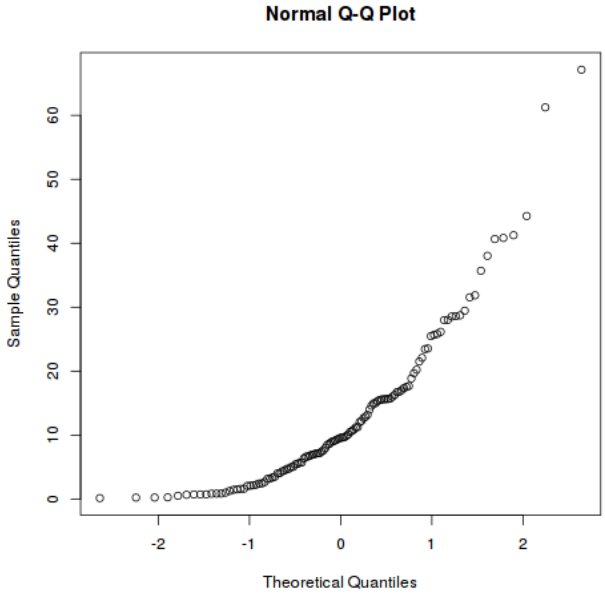
En las dos figuras estamos intentando ver si podemos superponer una línea recta a los puntos que estamos representando. No viene mal una ayuda visual que nos coloque una buena *línea y sobre ella veamos si los puntos están cercanos a la línea. Se pueden superponer muchas líneas. Existe una práctica a la hora de elegir esta línea. Para trazar una línea necesitamos dos puntos. Se eligen dos* buenos *puntos y consideramos la línea que pasa por ellos. El primero es el punto correspondiente a los percentiles de orden* 0,25*, esto es, el cuartil inferior. Determinamos este cuartil inferior en los datos observados (y es el valor de la ordenada) y en una distribución normal estándar (y es el valor de la abscisa). El segundo punto es el correspondiente al cuartil superior o percentil de orden* 0,75*. Su abscisa y ordenada son los percentiles de orden* 0,75 *en la distribución normal estándar y el observado en los datos. En las figuras 6.4(b) y 6.4(d) tenemos los dibujos q-q añadiendo la línea que pasa por los cuartiles inferior y superior. Para los datos x se obtiene con el siguiente código.*



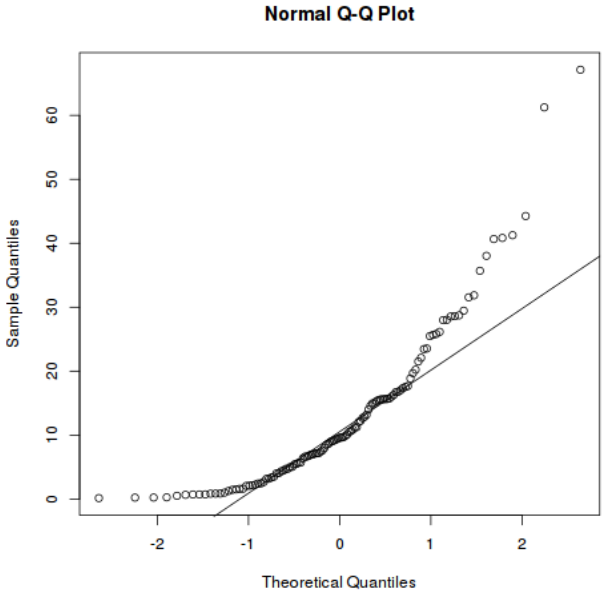
(a)



(b)



(c)



(d)

Figura 6.4: (a) Dibujo q-q o cuantil-cuantil para datos x. (b) Dibujo q-q o cuantil-cuantil para la muestra x añadiendo la línea que pasa por el primer y tercer cuartil. Vemos cómo los puntos están muy próximos a la línea. No podemos rechazar la normalidad de los datos utilizando este dibujo. (c) Dibujo q-q o cuantil-cuantil para datos y. (d) Dibujo q-q o cuantil-cuantil para la muestra y añadiendo la línea que pasa por el primer y tercer cuartil. Los puntos están alejados de la línea. Parece razonable rechazar la normalidad de los datos utilizando este gráfico.

```
qqnorm(x)
qqline(x)
```

¿Están sobre una línea recta los puntos en cada una de las gráficas? Podemos ver que para la figura 6.4(b) correspondiente a la muestra x los datos parecen bien alineados. Esto no parece tan cierto para los datos de la muestra y que aparecen en la figura 6.4(d). Rechazaríamos gráficamente la normalidad de la muestra y mientras que no la rechazaríamos para la muestra x.

En http://en.wikipedia.org/wiki/Q-Q_plot se tiene una explicación muy completa de este gráfico.

6.6 Constrastes de normalidad

Un procedimiento gráfico siempre ayuda a descartar situaciones claras. En la sección anterior hemos visto cómo podíamos rechazar la normalidad de los datos y mediante un dibujo q-q. Para los datos x no hemos podido rechazarla. ¿Nos quedamos con esto? ¿Admitimos que son datos normales y en paz? No. El siguiente paso a dar consiste en utilizar un contraste de hipótesis. Vamos a comentar rápidamente los contrastes más utilizados. En este sección utilizamos el paquete de R *nortest* **?** para los test ji-cuadrado y Kolmogorov-Smirnov. El test de Shapiro-Wilk lo aplicamos con la función *shapiro.test* de R Core Team [2015a, stats].

6.6.1 Test de Shapiro–Wilk

Es otro test para determinar si podemos considerar unos datos normales Apliquémoslo a nuestros datos. Para la muestra x

```
shapiro.test(x)

##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.98705, p-value = 0.2398
```

No rechazamos con un nivel de significación de 0,05. Para la segunda muestra,

```
shapiro.test(y)

##
##  Shapiro-Wilk normality test
```

```
##
## data:  y
## W = 0.84304, p-value = 5.271e-10
```

Vemos como rechazamos claramente.

Una buena explicación de este test se puede encontrar en http://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

6.6.2 *Test ji-cuadrado*

Lo podemos hacer con la función *pearson.test* del paquete **?**, nortest. La aplicamos a ambos conjuntos de datos. Primero a los datos x.

```
library(nortest)
pearson.test(x)

##
##  Pearson chi-square normality test
##
## data:  x
## P = 10.179, p-value = 0.6003
```

Y después a la muestra y.

```
pearson.test(y)

##
##  Pearson chi-square normality test
##
## data:  y
## P = 78.587, p-value = 2.767e-12
```

Vemos cómo rechazamos la normalidad de la muestra y mientras que no la rechazamos para la muestra x a un nivel de significación $\alpha = 0{,}05$.

6.6.3 *Test de Kolmogorov-Smirnov*

Para aplicar este test utilizamos la función *lillie.test* del paquete **?**, nortest. Empezamos con los datos x.

```
lillie.test(x)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  x
## D = 0.061707, p-value = 0.2415
```

Y ahora para los datos y.

```
lillie.test(y)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  y
## D = 0.14959, p-value = 5.243e-07
```

Vemos cómo rechazamos la normalidad de la muestra y mientras que no la rechazamos para la muestra x a un nivel de significación $\alpha = 0{,}05$.

6.7 Ejercicios

Ejercicio 6.7. 1. *Bajad de Aula Virtual (Recursos) el fichero* ejer93datos-input.rar. *En este fichero encontraremos a su vez los ficheros* dat1.txt, . . ., dat6.txt.

2. *Leer los datos de cada uno de los ficheros utilizando la función* read.table.

3. *Para cada uno de los ficheros se pide realizar un dibujo q-q y valorar la hipótesis de normalidad de un modo gráfico.*

4. *Para cada fichero, se pide contrastar, utilizando los tests de Shapiro-Wilk, ji-cuadrado y Kolmogorov-Smirnov, la hipótesis de normalidad.*

7
Comparación de dos poblaciones normales

7.1 Introducción

Distintos problemas relativos a comparar dos poblaciones se tratan en este tema. Empezamos abordando el problema de la comparación mediante herramientas puramente descriptivas de las dos muestras de que disponemos, una por población en estudio. Seguimos con la comparación de dos poblaciones normales, en particular, la comparación de sus medias y varianzas. Continuamos con la comparación mediante el test de Kolmogorov-Smirnov para dos muestras. Terminamos con un test de Montecarlo para comparar las medias de dos poblaciones.

7.2 Comparación descriptiva de las muestras

Pretendemos comparar dos poblaciones. De cada una de ellas disponemos de una muestra. Hablamos de dos **muestras independientes** porque proceden de poblaciones distintas.

Por ejemplo, podemos tener una muestra de concentraciones de un cierto elemento químico en una zona y otra muestra en otra zona y pretendemos compararlas.

Para ilustrar tomamos muestras concretas. Tenemos dos muestras x e y, una por población. Lo primero es la comparación descriptiva de estas muestras.

La primera muestra es

```
##  [1] 26.7 22.7 25.2 22.5 25.7 25.1 21.1 21.3 21.0 22.2 22.7
## [12] 25.3 21.9 24.0 26.5 23.8 27.3 23.1 25.8 21.6 24.9 23.0
## [23] 18.3 20.6 23.4 22.7 19.9 21.6 20.9 19.8 24.4 20.7 22.2
## [34] 22.7 25.3 27.7 17.5 24.2 21.8 22.3 24.1 23.4 24.0 20.7
## [45] 20.9
```

mientras que la segunda muestra es

```
round(y,1)
```

```
##  [1] 29.4 34.6 27.4 30.2 30.3 33.7 32.8 25.6 35.0 34.6 27.6
## [12] 29.1 26.7 29.9 29.7 23.8 36.1 29.8 27.8 32.2 30.2 30.4
## [23] 26.1 29.6 27.0 26.1 31.9 37.2 28.1 26.8 26.3 29.1 34.0
## [34] 35.3 32.1 19.0 27.0 27.5 27.3 26.7 40.4 31.1 28.4 24.6
## [45] 28.2 27.2 28.3 31.2 27.1 27.7 31.5 29.0 33.5 27.3
```

¿Qué significa comparar las dos muestras? Supongamos que son concentraciones de un cierto elemento en dos zonas distintas. Desde un punto de vista estadístico, son realizaciones de variables aleatorias. Si repetimos el muestreo obtendremos valores distintos en cada una de las dos zonas observadas. Por tanto, la comparación no puede ser comparar los valores numéricos uno a uno. Tampoco puede ser comparar sin más algún resumen de los datos como la media y la varianza muestrales. En nuestro caso, los valores observados son para la primera muestra

```
mean(x)
```

```
## [1] 22.9355
```

```
sd(x)
```

```
## [1] 2.283711
```

y para la segunda

```
mean(y)
```

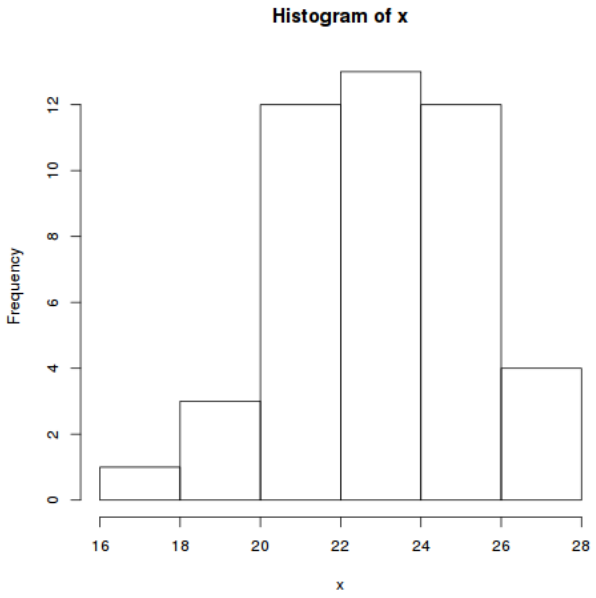
```
## [1] 29.61508
```

```
sd(y)
```

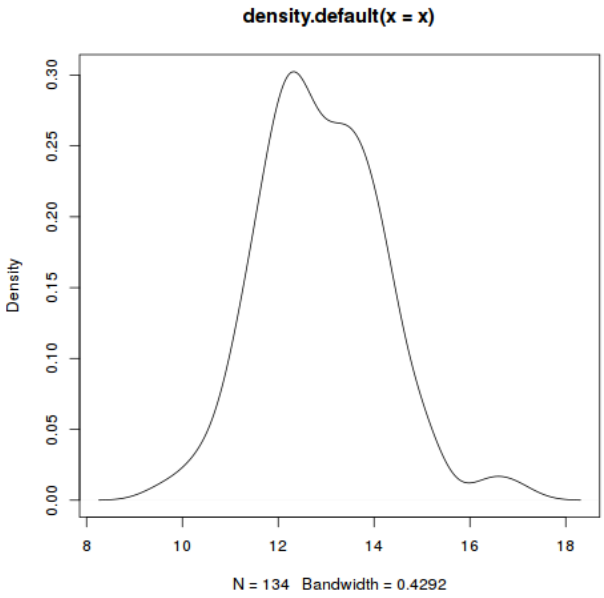
```
## [1] 3.708588
```

Casi tiene un mayor interés la comparación gráfica de ambas muestras. En la figura 7.1 tenemos el histograma y el estimador kernel de la densidad de las muestras x e y. ¿Podemos considerar que ambas muestras han sido obtenidas de una misma población? O, por el contrario: ¿proceden de distintas poblaciones?

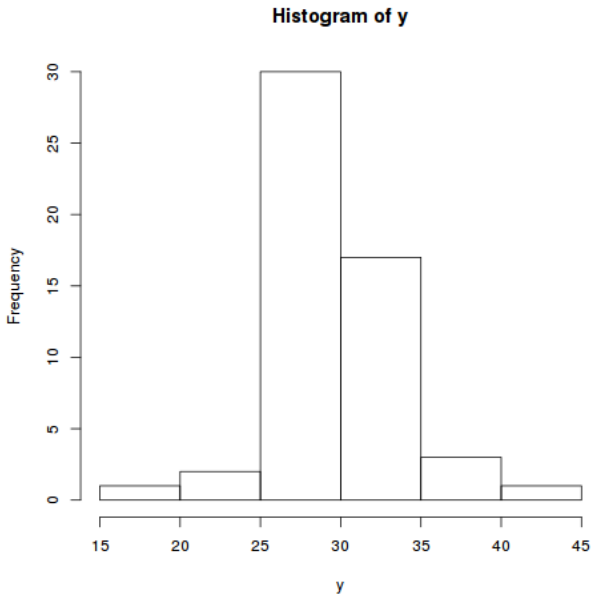
Una primera cuestión es que *pretendemos comparar muestras obtenidas de posiblemente dos poblaciones distintas*. No parece una buena opción utilizar figuras distintas (bien histogramas bien estimadores kernel). Ambos dibujos se configuran para su propia muestra. Por ejemplo, el número de clases de cada histograma depende del número de datos y por lo tanto es distinta. La anchura de banda de los



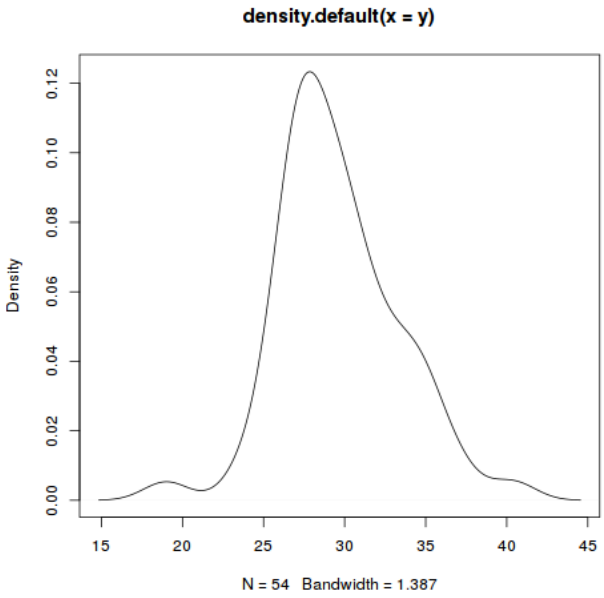
(a)



(b)



(c)



(d)

Figura 7.1: Datos x: histograma (a) y estimador kernel de la densidad de x (b). Datos y: histograma (c) y estimador kernel de la densidad de x (d).

estimadores kernel también son distintos. En resumen, si queremos comparar lo lógico es representar conjuntamente ambas muestras. ¿Cómo? Utilizar histogramas conjuntos no es especialmente útil o de interpretación fácil. Es preferible utilizar colocar en una misma gráfica los dos estimadores kernel. Un detalle, el nivel de suavizado *debe* de ser el mismo para los dos estimadores, o dicho de otro modo, la anchura de banda ha de ser la misma. En la figura 7.2 se muestra una representación conjunta de ambos estimadores kernel de la densidad. Los datos muestran una forma simétrica y vemos que la muestra y toma valores mayores que la muestra x. Sin embargo, es una pura interpretación gráfica. Hemos de estimar las diferencias entre las muestras y contrastar si ambas muestras proceden de distintas poblaciones con contrastes de hipótesis formales. Es lo que vamos a hacer.

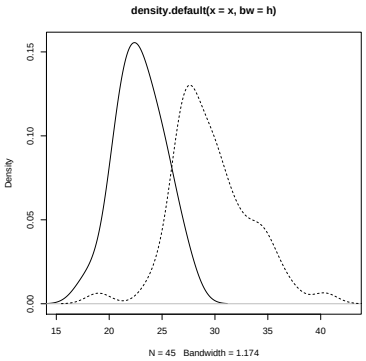


Figura 7.2: Estimadores kernel de la densidad de x (trazo continuo) e y (trazo discontinuo). Los datos de la muestra y tienden a tomar valores mayores que los de x.

Cuando comparamos dos poblaciones hay que hacerse varias preguntas y en función de las respuestas que obtengamos planteamos la comparación. ¿Son los datos normales? Es la primera pregunta que debemos responder: ¿Podemos considerar que los datos proceden de dos poblaciones normales? De otro modo: ¿la primera muestra procede de una población normal con media y varianza μ_X y σ_X^2 desconocidos y la segunda muestra procede de una población normal con media y varianza μ_Y y σ_Y^2 desconocidos? Obviamente esto supone pasar test de normalidad a nuestros datos. Cada una de las dos muestras ha de pasar el test de normalidad. Supongamos que la respuesta es afirmativa. El problema de comparar las poblaciones se simplifica. ¿Cómo es la densidad de una normal? Si la variable aleatoria $X \sim N(\mu_X, \sigma_Y^2)$ entonces su función de densidad es

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

La de la variable Y con distribución $Y \sim N(\mu_Y, \sigma_Y^2)$ tiene la misma expresión en donde cambiamos μ_x por μ_Y y σ_X por σ_Y. En resumen si las medias son la misma, $\mu_X = \mu_Y$ y las varianzas son iguales $\sigma_X^2 = \sigma_Y^2$ entonces las densidades son la misma. Tenemos la misma población normal. En resumen, si asumimos que las dos poblaciones son normales entonces comparar las poblaciones se reduce a comparar las medias y las varianzas. Además cuando no sean iguales podremos saber a qué se debe. Las situaciones en que nos podemos encontrar son las siguientes:

1. La misma media y varianza.

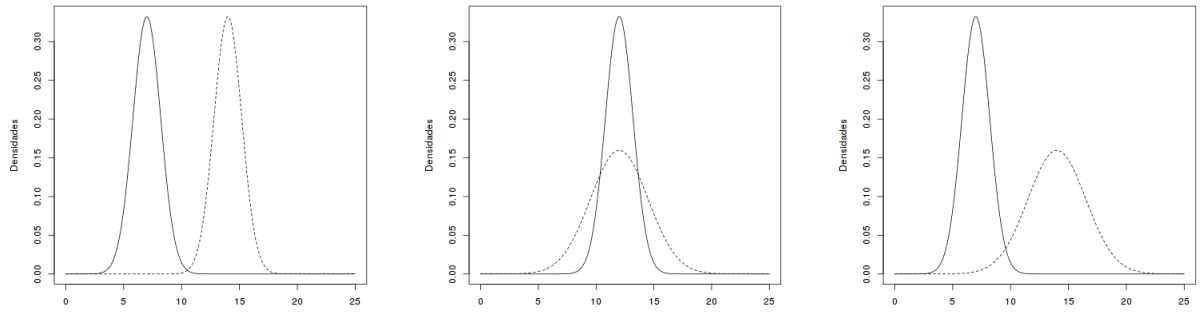
2. Distinta media y la misma varianza. En la figura **??** mostramos dos densidades normales verificándolo.

3. La misma media y distinta varianza. En la figura **??** vemos las funciones de densidad.

4. Distinta media y varianza. En la figura **??** tenemos un ejemplo de densidades normales verificando esto.

En la situación 1 de la enumeración anterior *no* tenemos dos poblaciones. Tenemos una sola población. Sin embargo, en los casos 2, 3 y 4 tenemos dos poblaciones distintas bien porque la media, la varianza o ambas son distintas. Podemos evaluar (contrastar) si la diferencia entre las poblaciones se da en la variabilidad (en las varianzas) en las medias o en ambas cosas. Tenemos una visión clara de las diferencias entre las poblaciones.

En la sección 7.4.2 planteamos el correspondiente contraste de hipótesis en donde comparamos las varianzas de dos poblaciones normales. Obviamente si rechazamos la hipótesis de igualdad de las varianzas tenemos dos poblaciones distintas ya que sus varianzas lo son. Lo que no sabemos es sus medias son o no iguales. Esto va después. Si no hemos rechazado que las varianzas sean iguales entonces comparamos las medias asumiendo una misma varianza en las dos poblaciones, es decir, asumimos que la primera muestra aleatoria es de una población normal con media μ_X y varianza σ^2 mientras que la segunda muestra es de una normal con media μ_Y y varianza σ^2. Finalmente, también podemos realizar el contraste de hipótesis para comparar las medias cuando las varianzas son distintas.



(a) (b) (c)

Figura 7.3: Densidades normales: (a) distinta media y la misma varianza; (b) misma media y distinta varianza; (c) distintas medias y varianzas.

7.3 *Comparando las medias de dos poblaciones normales*

Asumimos, en esta sección, que las muestras son de poblaciones normales. Desde el punto de vista del usuario el interés suele estar en comparar medias independientemente de si las varianzas son

iguales o no. Esto es una visión algo miope porque la diferencia de la varianza también indica que tenemos poblaciones distintas.

Tenemos una muestra $X_1, \ldots, X_n$ de una variable $X \sim N(\mu_X, \sigma_X^2)$ y otra muestra $Y_1, \ldots, Y_m$ de $Y \sim N(\mu_Y, \sigma_Y^2)$.

7.3.1 *Estimación de la diferencia de medias*

Nos interesa estudiar si μ_X y μ_Y son iguales o no. Esto se puede formular de varias formas. En primer lugar podemos plantearnos estimar la diferencia de los dos valores. El estimador de esta cantidad es

$$\widehat{\mu_X - \mu_Y} = \hat{\mu}_X - \hat{\mu}_Y = \bar{X}_n - \bar{Y}_m.$$

La expresión anterior significa que el estimador de la diferencia $\mu_X - \mu_Y$, que denotamos como $\widehat{\mu_X - \mu_Y}$, tomamos la diferencia de los estimadores de μ_X y μ_Y que son respectivamente $\bar{X}_n$ y $\bar{Y}_m$. El valor a estimar $\mu_X - \mu_Y$ es un número desconocido, lo que llamamos parámetro, que estimamos mediante el estimador puntual $\bar{X}_n - \bar{Y}_m$.

Hemos visto antes como construir intervalos de confianza para cada una de las medias μ_X y μ_Y. Estos intervalos (asumiendo normalidad) tienen la expresión $\bar{x}_n \pm t_{n-1,1-\alpha/2} s_X / \sqrt{n}$ y $\bar{y}_m \pm t_{m-1,1-\alpha/2} s_X / \sqrt{m}$.

Nota de R 7.1. *Supongamos que tomamos un nivel de confianza* $1 - \alpha = 0{,}95$. *Con los datos que tenemos los intervalos son, para la primera muestra,*

```
t.test(x)$conf.int

## [1] 22.2494 23.6216
## attr(,"conf.level")
## [1] 0.95
```

y para la segunda

```
t.test(y)$conf.int

## [1] 28.60283 30.62733
## attr(,"conf.level")
## [1] 0.95
```

El primer intervalo contiene a μ_X con un nivel de confianza $1 - \alpha$ y el segundo a μ_Y con el mismo nivel de confianza. *Pero esto ya lo sabía y no me resuelve nada*. Queremos un intervalo para la diferencia de medias $\mu_X - \mu_Y$ con un nivel de confianza previamente especificado

$1-\alpha$. Para ello se utiliza la siguiente cantidad.

$$T = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{SE(\bar{X}_n - \bar{Y}_m)} \tag{7.1}$$

donde $SE(\bar{X}_n - \bar{Y}_m)$ denota la desviación estándar de la variable $\bar{X}_n - \bar{Y}_m$. Esta cantidad no tiene una expresión única. De hecho, depende de si las varianzas son iguales o distintas. Además el comportamiento aleatorio de T definido en la ecuación 7.1 también depende de si son la misma varianza o son distintas. En cualquier caso T va a tener una distribución de probabilidad que es una t de Student donde serán distintos los grados de libertad. Si de momento denotamos como ν estos grados (tomará dos valores posibles) entonces el intervalo de confianza para $\mu_X - \mu_Y$ con un nivel de confianza $1-\alpha$ tendrá la expresión general

$$\bar{X}_n - \bar{Y}_m \pm t_{\nu,1-\alpha/2} \times SE(\bar{X}_n - \bar{Y}_m).$$

Como vemos no es muy distinto al caso de una sola muestra en que estimamos una sola media.

Nota de R 7.2. *En nuestro caso, y antes de ver las expresiones, podemos obtener el intervalo de confianza asumiendo varianzas iguales con*

```
t.test(x,y,var.equal=TRUE)$conf.int

## [1] -7.938818 -5.420354
## attr(,"conf.level")
## [1] 0.95
```

y asumiendo que son distintas (que es la opción por defecto) con

```
t.test(x,y,var.equal=FALSE)$conf.int

## [1] -7.889036 -5.470136
## attr(,"conf.level")
## [1] 0.95
```

Como vemos son bastante parecidos los dos intervalos. Posiblemente no deben de ser muy distintas las varianzas de las dos poblaciones.

Si asumimos que las dos varianzas son iguales, esto es, asumimos la hipótesis de que $\sigma_X^2 = \sigma_Y^2$, denotaremos por σ^2 el valor común: $\sigma^2 = \sigma_X^2 = \sigma_Y^2$. El valor común σ^2 de la varianza se puede estimar con

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

De hecho, lo que tenemos es que $SE(\bar{X}_n - \bar{Y}_m) = S_p\sqrt{\frac{1}{n}+\frac{1}{m}}$ y

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}, \tag{7.2}$$

tiene una distribución t de Student con $n+m-2$ grados de libertad.

El intervalo de confianza lo podemos construir utilizando el resultado dado en 7.2 y vendría dado por

$$\bar{X}_n - \bar{Y}_m \pm t_{n+m-2,1-\alpha/2}\frac{S_p}{n+m-2}.$$

Nota de R 7.3. *Este intervalo viene dado, para un nivel de* $1-\alpha = 0{,}95$ *por*

```
t.test(x,y,var.equal=TRUE,conf.level=0.95)$conf.int

## [1] -7.938818 -5.420354
## attr(,"conf.level")
## [1] 0.95
```

y, para un nivel de $1-\alpha = 0{,}99$ *por*

```
t.test(x,y,var.equal=TRUE,conf.level=0.99)$conf.int

## [1] -8.346615 -5.012557
## attr(,"conf.level")
## [1] 0.99
```

Como vemos un mayor nivel de confianza supone un intervalo más grande. Tenemos más confianza en que el valor verdadero de $\mu_X - \mu_Y$ *esté en el intervalo pero el intervalo al ser mayor nos estima la diferencia de medias con una precisión menor.*

Y ahora consideremos el caso con varianzas posiblemente distintas, esto es, suponemos que $\sigma_X^2 \neq \sigma_Y^2$. No tiene ahora sentido estimar una varianza común que no existe. Estimamos cada una de ellas con su estimador natural: σ_X^2 con S_X^2 y σ_Y^2 con S_Y^2. El error estándar de $\bar{X}_n - \bar{Y}_m$, $SE(\bar{X}_n - \bar{Y}_m)$, viene dado por $SE(\bar{X}_n - \bar{Y}_m) = \sqrt{\frac{S_X^2}{n}+\frac{S_y^2}{m}}$ y la expresión que adopta 7.1 y su distribución **aproximada** es

$$T = \frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n}+\frac{S_y^2}{m}}} \sim t_{\nu_0} \tag{7.3}$$

con

$$\nu_0 = \frac{\left(\frac{S_X^2}{n}+\frac{S_y^2}{m}\right)}{\frac{(S_X^2/n)^2}{n-1}+\frac{(S_Y^2/m)^2}{m-1}}.$$

Es decir, que T se distribuye **aproximadamente** como una distribución t de Student ν_0 grados de libertad donde ν_0 no tiene porqué ser un número entero. El intervalo de confianza con nivel de confianza $1-\alpha$ viene dado por

$$\bar{X}_n - \bar{Y}_m \pm t_{\nu_0,1-\alpha/2}\sqrt{\frac{S_X^2}{n}+\frac{S_y^2}{m}}.$$

Nota de R 7.4. *Con los datos que estamos analizando los intervalos para la diferencia de medias con niveles de confianza* 0,95 *y* 0,99 *son*

```
t.test(x,y,var.equal=FALSE,conf.level=0.95)$conf.int

## [1] -7.889036 -5.470136
## attr(,"conf.level")
## [1] 0.95
```

y

```
t.test(x,y,var.equal=FALSE,conf.level=0.99)$conf.int

## [1] -8.281661 -5.077511
## attr(,"conf.level")
## [1] 0.99
```

7.3.2 *Contraste de hipótesis*

Otra manera de comparar las medias de dos poblaciones es contrastar si sus medias son iguales o bien si alguna de ellas es mayor que la otra.

Vamos a considerar el contraste de igualdad de medias frente a la desigualdad lo que también se llama contraste bilateral, bidireccional o de dos colas.

$$\begin{aligned} H_0: &\quad \mu_X = \mu_Y, \\ H_1: &\quad \mu_X \neq \mu_Y. \end{aligned}$$

La región crítica, es decir, los valores para los cuales rechazamos la hipótesis nula es:

$$|T_0| > t_{\nu,1-\alpha/2}$$

siendo

$$T_0 = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}, \tag{7.4}$$

y $\nu = n+m-2$ si las varianzas se asumen iguales. Para varianzas desiguales:

$$T_0 = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{S_X^2}{n}+\frac{S_y^2}{m}}} \tag{7.5}$$

y $\nu = \nu_0$.

El p-valor viene dado por

$$p = P(T_0 \geq |t_0|)$$

siendo t_0 el valor observado de T_0 en cada caso.

Nota de R 7.5 (Test de la t para comparar medias)**.** *Supongamos que el nivel de significación elegido es $\alpha = 0{,}01$. Vamos a contrastar la igualdad de medias en los dos casos. Con varianzas iguales tendremos:*

```
t.test(x,y,var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  x and y
## t = -10.528, df = 97, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.938818 -5.420354
## sample estimates:
## mean of x mean of y
##  22.93550  29.61508
```

Vemos que el p-valor o es muy pequeño (menor que α) y por ello rechazamos la igualdad de las medias. ¿Qué pasa si no asumimos la igualdad de las varianzas? ¿Cambia el resultado del test? Lo hacemos.

```
t.test(x,y,var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -10.972, df = 89.809, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.889036 -5.470136
## sample estimates:
## mean of x mean of y
##  22.93550  29.61508
```

El p-valor o es muy pequeño por lo que rechazamos la hipótesis nula.

7.3.3 *Los contrastes unilaterales o direccionales*

Los otros dos contrastes serían los correspondientes unilaterales, o direccionales o de una cola.

$$\begin{aligned} H_0 &: \quad \mu_X \leq \mu_Y, \\ H_1 &: \quad \mu_X > \mu_Y. \end{aligned}$$

$$\begin{aligned} H_0 &: \quad \mu_X \geq \mu_Y, \\ H_1 &: \quad \mu_X < \mu_Y. \end{aligned}$$

Nota de R 7.6. *Vamos a comparar las concentraciones de bario en Rusia y Noruega para los datos del proyecto Kola. En principio vamos a suponer que las varianzas de dichas concentraciones las podemos considerar iguales. Primero leemos los datos.*

```
load("../data/chorizon.rda")
attach(chorizon)
```

Los intervalos de confianza para la diferencia de medias y el contraste de hipótesis para la igualdad de las medias frente a la alternativa de medias distintas asumiendo que las varianzas son la misma lo podemos obtener con

```
t.test(Ba[COUN=="RUS"],Ba[COUN=="NOR"],var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  Ba[COUN == "RUS"] and Ba[COUN == "NOR"]
## t = 2.618, df = 416, p-value = 0.009166
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.865386 41.212405
## sample estimates:
## mean of x mean of y
##  69.18655  45.64766
```

Estimamos la diferencia de medias como 23.539 y el intervalo de confianza con nivel 0,95 *es [5.865,41.212]. Además el p-valor observado es 0.0091663 menor que* 0,01 *por lo que rechazamos la hipótesis nula de igualdad de medias.*

Vamos a repetir el estudio sin asumir que la varianza es la misma en las dos poblaciones.

```r
t.test(Ba[COUN=="RUS"],Ba[COUN=="NOR"],var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  Ba[COUN == "RUS"] and Ba[COUN == "NOR"]
## t = 2.618, df = 416, p-value = 0.009166
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.865386 41.212405
## sample estimates:
## mean of x mean of y
##  69.18655  45.64766
```

El estimador puntual de la diferencia de medias no cambia, 23.539. Sin embargo, el intervalo de confianza con nivel 0,95 *sí que cambia. Ahora es [10.459,36.619]. Además el p-valor observado es* $4{,}5 \times 10^{-4}$ *sigue siendo menor que* 0,01 *por lo que rechazamos la hipótesis nula de igualdad de medias.*

La pregunta obvia es: ¿qué opción elegir? Desde el punto de vista aplicado lo lógico es asumir varianzas desiguales (que por otra parte suele ser el caso). Por ello la opción por defecto de t.test *es precisamente asumir varianzas distintas. Además cuando las varianzas son realmente iguales o casi iguales los resultados que se obtienen asumiendo que las varianzas son la misma o sin asumirlo son prácticamente las mismas.*

7.3.4 Ejercicios

Ejercicio 7.1. *Los biosólidos de una planta de tratamiento de aguas residuales industriales se aplicaron a 10 parcelas que fueron seleccionados aleatoriamente de un total de 20 parcelas de ensayo de las tierras agrícolas. El maíz se cultiva en el grupo tratado (T) y no tratados (UT), las parcelas, con los siguientes rendimientos (fanegas / acre).*

```
Grupo T
126 122 90 135 95 180 68 99 122 113

Grupo no tratado NT
144 122 135 122 77 149 122 117 131 149
```

Se pide:

1. *Calcular el intervalo de confianza con un nivel de confianza del 95 % para la diferencia de las medias.*
2. *Existen diferencias significativas entre las medias.*

Ejercicio 7.2. *Las mediciones de plomo. A continuación damos las concentraciones medidas de plomo en soluciones que son idénticas, salvo por la cantidad de plomo que se ha añadido. Catorce muestras contenían* 1,25*mg / L y 14 contenían* 2,5 *mg / L. ¿Es consistente la diferencia de los medias muestrales observadas con la diferencia (real) de* 1,25 *mg / l?*

```
Con 1.25 mg/L
 1.1 2.0 1.3 1.0 1.1 0.8 0.8 0.9 0.8 1.6 1.1 1.2 1.3 1.2
Con 2.5 mg/L
 2.8 3.5 2.3 2.7 2.3 3.1 2.5 2.5 2.5 2.7 2.5 2.5 2.6 2.7
```

7.4 Inferencia sobre las varianzas de dos poblaciones normales

Hasta ahora no nos hemos preocupado de las varianzas de las poblaciones normales con las que estamos trabajando. Ya es hora de hacerlo. Además es enormemente importante estimar y contrastar cosas sobre las varianzas. Tenemos dos muestras correspondientes a dos poblaciones. Tenemos, como siempre, un doble interés: estimar y contrastar.

7.4.1 Estimación del cociente de varianzas

Nota de R 7.7. *Vamos a generar unos datos con distribución normal de los cuales sabemos sus medias y varianzas.*

```
n = 45
m = 54
x = rnorm(n,mean=23,sd=2.45)
y = rnorm(m,mean=30,sd=3.45)
```

Si llamamos a la función var.test *tenemos lo siguiente:*

```
var.test(x,y)

##
##  F test to compare two variances
##
## data:  x and y
## F = 0.59444, num df = 44, denom df = 53, p-value =
## 0.07819
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3380477 1.0614600
## sample estimates:
## ratio of variances
##          0.5944433
```

Si miramos la última línea vemos que no nos está estimando cada varianza separadamente. En lugar de ello, como nos interesa compararlas, *lo que estimamos es el cociente* σ_X^2/σ_Y^2.

Las varianzas de las dos poblaciones se comparan utilizando el cociente σ_X^2/σ_Y^2. Podemos estimarlo o bien contrastar hipótesis sobre el cociente. El estimador puntual de esta cantidad es

$$\frac{S_X^2}{S_Y^2} \tag{7.6}$$

que con nuestros datos vale 0.594 indicando que la primera varianza es menor que la segunda. Bien, nos hemos centrado en la estimación de σ_X^2/σ_Y^2. Siempre que estimamos damos una estimación puntual (que acabamos de ver en 7.6) y un intervalo de confianza. Para obtenerlo utilizamos la cantidad pivotal siguiente:

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n-1, m-1) \tag{7.7}$$

que tiene una distribución F (de Fisher) con $n-1$ y $m-1$ grados de libertad.

Los tamaños muestrales de nuestros datos son n=45 y m=54. La función de densidad de $\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ aparece en la figura 7.4.

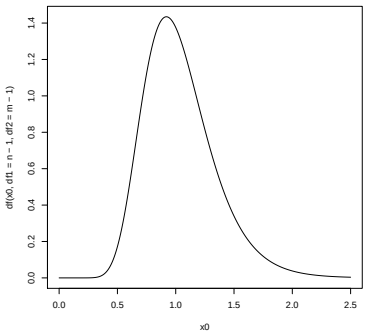


Figura 7.4:

Denotemos por $F_p(n-1, m-1)$ el percentil de orden p de la distribución $F(n-1, m-1)$, es decir, el punto que a su izquierda tiene un área p. Por ejemplo, si tomamos los valores de n y m anteriores y $p = 0{,}975$ entonces el percentil viene dado como

```
qf(0.975,df1=n-1,df2=m-1)

## [1] 1.75846
```

La figura 7.5 muestra el área p y el punto en el eje de abscisas más a la derecha de la zona rayada es el correspondiente percentil.

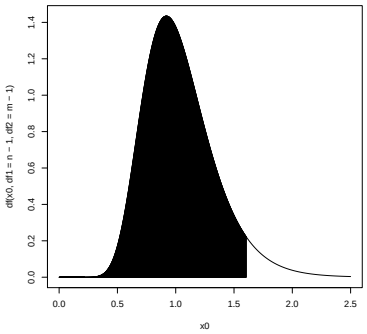


Figura 7.5: Percentil de una F de Fisher.

Teniendo en cuenta el resultado 7.7 el intervalo de confianza para $\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ es el siguiente:

$$\left[\frac{S_Y^2}{S_X^2}\frac{1}{F_{1-\alpha/2}(n-1, m-1)}, \frac{S_Y^2}{S_X^2}\frac{1}{F_{\alpha/2}(n-1, m-1)}\right]$$

7.4.2 *Contraste de hipótesis para el cociente de varianzas*

Nuestro interés fundamental es valorar si podemos considerar que las varianzas son iguales o que son distintas. Tenemos un problema de contraste de hipótesis. De hecho, estamos interesados en el

siguiente contraste:

$$\begin{aligned} H_0: &\quad \sigma_X^2 = \sigma_Y^2, \\ H_1: &\quad \sigma_X^2 \neq \sigma_Y^2. \end{aligned}$$

que lo reformulamos como

$$\begin{aligned} H_0: &\quad \frac{\sigma_X^2}{\sigma_Y^2} = 1, \\ H_1: &\quad \frac{\sigma_X^2}{\sigma_Y^2} \neq 1. \end{aligned}$$

Bajo la hipótesis de que $H_0 : \sigma_X^2 = \sigma_Y^2$ (o $H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$) tenemos que

$$F = \frac{S_X^2}{S_Y^2} \sim F(n-1, m-1) \tag{7.8}$$

y podemos contrastar que las varianzas sean la misma rechazando la hipótesis nula de igualdad con un nivel de significación α si

$$F < F_{\alpha/2}(n-1, m-1) \text{ o } F > F_{1-\alpha/2}(n-1, m-1).$$

Nota de R 7.8. *Vamos a plantearnos si podemos considerar que las concentraciones de bario en Rusia y Noruega tienen varianzas similares o no. En primer lugar podemos observar las varianzas muestrales de ambas muestras*

```
var(Ba[COUN=="RUS"])

## [1] 9730.406

var(Ba[COUN=="NOR"])

## [1] 1372.308
```

Vemos que la variabilidad en Rusia es mayor que la que tenemos en Noruega.

```
var.test(Ba[COUN=="RUS"],Ba[COUN=="NOR"])

##
##  F test to compare two variances
##
## data:  Ba[COUN == "RUS"] and Ba[COUN == "NOR"]
## F = 7.0905, num df = 289, denom df = 127, p-value <
## 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
```

```
##  5.227669 9.452443
## sample estimates:
## ratio of variances
##           7.090539
```

El cociente lo estimamos como 7.091. El intervalo de confianza para el cociente de las varianzas es [5.228,9.452]. Como vemos el valor uno no está en el intervalo de confianza y no podemos considerar que la variabilidad de las medidas sean semejantes.

Por consiguiente lo correcto es elegir la opción por defecto de varianzas desiguales que nos da el siguiente código.

```
var.test(Ba[COUN=="NOR"],Ba[COUN=="FIN"])

##
##  F test to compare two variances
##
## data:  Ba[COUN == "NOR"] and Ba[COUN == "FIN"]
## F = 0.84466, num df = 127, denom df = 186, p-value =
## 0.3082
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6161697 1.1692619
## sample estimates:
## ratio of variances
##          0.8446567
```

7.4.3 Ejercicios

Ejercicio 7.3. *Estamos analizando dos catalizadores con objeto de determinar como afectan a la producción media de un proceso químico. Teniendo en cuenta que el segundo catalizador es más barato, éste sería el elegido suponiendo que la producción media no se modifica manifiestamente. Se tomaron dos muestras, una por catalizador, y se obtuvieron los resultados siguientes: en la muestra 1;*

`91.50, 94.18, 92.18, 95.39, 91.79, 89.07, 94.72, 89.21`

y en la segunda muestra,

`89.19, 90.95, 90.46, 93.21, 97.19, 97.04, 91.07, 92.75.`

Se pide:

1. *¿Podemos considerar que la varianza de las muestras es la misma?*
2. *Comparar las medias teniendo en cuenta la respuesta que hemos dado en el apartado anterior.*

Ejercicio 7.4 (Los biosólidos)**.** *Los biosólidos de una planta de tratamiento de aguas residuales industriales se aplicaron a 10 parcelas seleccionadas aleatoriamente de un total de 20 parcelas de ensayo de las tierra agrícola. El maíz se cultiva en las parcelas tratadas (T) y no tratadas (UT) con los siguientes rendimientos (fanegas / acre).*

```
UT 126 122 90 135 95 180 68 99 122 113
T  144 122 135 122 77 149 122 117 131 149
```

Se pide calcular el intervalo de confianza al 90 % para la diferencia de las producciones medias.

Ejercicio 7.5. *La presencia de arsénico en el agua potable de la red pública es un riesgo para la salud. Se han tomado medidas de la concentración de arsénico en 20 poblaciones de la región de Murcia y en 20 poblaciones de la Comunidad Valenciana. Los valores observados son los siguientes:*

```
 9.304729  7.521734 12.709850 16.142454  7.475735 11.741802 16.330484
15.435159 14.835557  9.759411 12.250387  9.426963 15.992894 17.116555
15.507461 11.764259 11.381000 19.548472 11.412476 13.776614
```

y

```
 27.02096 20.59237 18.38069 23.80335 25.07857 22.43169 21.49625 20.86017
 24.24796 22.88794 20.23317 23.99709 18.30906 25.35573 19.92300 23.65616
 23.58910 17.72509 17.93796 26.33419 25.73931 24.29759 20.00022
```

Se pide:

1. *Introducir los datos en un fichero texto (por ejemplo, con Calc de OpenOffice.org). Hay que introducir dos variables. En una de ellas hemos de poner todas las concentraciones indicadas correspondientes a las dos comunidades autónomas. En la otra indicamos con un código numérico si estamos en una u otra comunidad. Llamad a este fichero* arsenico.txt.
2. *¿Podemos considerar que la varianza de las muestras es la misma?*
3. *Comparar las medias teniendo en cuenta la respuesta que hemos dado en el apartado anterior.*

7.5 Comparación de medias con muestras apareadas

Volvemos al problema de comparar dos muestras. Ahora ya no hablamos de muestras independientes. No son dos zonas distintas en las que medimos una concentración y tomamos mediciones en cada una de ellas. O dos grupos de enfermos distintos de modo que en un grupo administramos una medicación y en el otro grupo administramos la otra. Ahora vamos a suponer que tenemos muestras *apareadas*

o *emparejadas* (paired es la expresión inglesa). Por ejemplo, si medimos la humedad en una localización un día y repetimos la medición otro día pero en el mismo lugar entonces tenemos dos observaciones apareadas. Son el mismo punto y lo que cambia es el momento en que observamos. Esto lo hacemos en n localizaciones distintas y tenemos: (x_i, y_i) con $i = 1, \ldots, n$. El factor que empareja es lo localización en que medimos.

Un ejemplo médico habitual es medir algo en un enfermo antes y después de tomar una medicación. Los datos son apareados pues corresponden a la misma persona. El factor que empareja es la persona.

Hablando con (un poco) de precisión lo que tenemos son n observaciones *independientes* de dos variables aleatorias que denotamos (X_i, Y_i) con $i = 1, \ldots, n$. Notemos que las dos variables aleatorias X_i e Y_i *no son independientes*, están relacionadas porque bien corresponden a la misma localización o a una misma persona. Siendo μ_X y μ_Y las medias de las variables X e Y nos interesa (lo que no es mucha novedad) conocer la diferencia de medias $\mu_X - \mu_Y$. Como es sabido la media de la variable diferencia es la diferencia de las medias de cada variable, esto es, $\mu_{X-Y} = \mu_X - \mu_Y$. Teniendo en cuenta este resultado lo que vamos a hacer es olvidarnos de las variables X e Y y trabajar con la variable $D = X - Y$ y contrastar si la media de la variable D es nula.

Nota de R 7.9. *Supongamos las siguientes mediciones de humedad en n localizaciones con una diferencia de un día en la observación.*

Los valores de las muestras x e y (mostramos los 10 primeros solamente por razones de espacio) son:

```
cbind(x[1:10],y[1:10])
```

```
##           [,1]     [,2]
##  [1,] 40.09233 33.78533
##  [2,] 34.91605 47.96437
##  [3,] 31.60288 34.10186
##  [4,] 44.87360 39.25625
##  [5,] 31.46792 38.72550
##  [6,] 31.08674 30.87797
##  [7,] 32.08353 31.95785
##  [8,] 35.81122 44.87951
##  [9,] 32.60339 24.67302
## [10,] 27.63668 44.56006
```

Si hacemos lo indicado. Empezamos calculando las diferencias y vemos los 10 primeros valores.

```
d = x -y
d[1:10]

##  [1]   6.3070036 -13.0483230  -2.4989813   5.6173465
##  [5]  -7.2575778   0.2087732   0.1256797  -9.0682863
##  [9]   7.9303700 -16.9233845
```

Ahora podemos obtener el estimador puntual de la diferencia de medias, el intervalo de confianza y el contraste de si la diferencia de medias vale cero utilizando t.test *sobre las diferencias.*

```
t.test(d)

##
##  One Sample t-test
##
## data:  d
## t = -3.6672, df = 144, p-value = 0.0003443
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.630276 -1.387014
## sample estimates:
## mean of x
## -3.008645
```

Otra opción en la que no necesitamos calcular previamente las diferencias es la siguiente:

```
t.test(x,y,paired=TRUE)

##
##  Paired t-test
##
## data:  x and y
## t = -3.6672, df = 144, p-value = 0.0003443
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.630276 -1.387014
## sample estimates:
## mean of the differences
##               -3.008645
```

Como vemos lo que sale es lo mismo.

Ejemplo 7.1. *Se trata de evaluar el efecto de la dieta y el ejercicio en el nivel de colesterol. Se midió la concentración antes y después de un programa de*

ejercicio aeróbico y cambio a una dieta baja en grasas. Los datos son los siguientes

```
(x = c(265,240,258,295,251,245,287,314,260,279,283,240,238,225,247))
```

```
##  [1] 265 240 258 295 251 245 287 314 260 279 283 240 238 225
## [15] 247
```

```
(y = c(229,231,227,240,238,241,234,256,247,239,246,218,219,226,233))
```

```
##  [1] 229 231 227 240 238 241 234 256 247 239 246 218 219 226
## [15] 233
```

donde x corresponde a los niveles antes e y a los niveles después. Podemos contrastar la igualdad de medias frente a la desigualdad con

```
t.test(x,y,paired=T)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 5.4659, df = 14, p-value = 8.316e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  16.32430 37.40904
## sample estimates:
## mean of the differences
##                26.86667
```

Podemos observar que el valor 0 *no está en el intervalo de confianza. Por el tipo de dato que tenemos parece más natural contrastar las hipótesis* $H_0 : \mu_X \leq \mu_Y$ *frente a* $H_1 : \mu_X > \mu_Y$.

```
t.test(x,y,paired=T,alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 5.4659, df = 14, p-value = 4.158e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  18.20922      Inf
## sample estimates:
## mean of the differences
##                26.86667
```

Vemos que el p-valor o es menor que $\alpha = 0{,}05$ y por lo tanto rechazamos la hipótesis nula. De hecho, también rechazamos con un nivel de significación $\alpha = 0{,}01$.

7.5.1 Ejercicios

Ejercicio 7.6. *Berthouex and Brown [2002, problema 17.1] Se ha evaluado la concentración de antimonio en tres muestras de pescado. Cada muestra fue evaluada con el método oficial y con un nuevo método. ¿Difieren significativamente los procedimientos?*

Muestra	*1*	*2*	*3*
Método nuevo	*2.964*	*3.030*	*2.994*
Método estándar	*2.913*	*3.000*	*3.024*

Ejercicio 7.7. *Seguimos con los datos de arsénico que hemos analizado en el ejercicio 7.5. Tanto en Murcia como en la Comunidad Valenciana se modificó el procedimiento de depuración de las aguas potables. Se repitieron las medidas en las mismas poblaciones después de la modificación indicada. Los nuevos valores observados fueron los siguientes:*

```
10.070679  9.715279  8.267526 10.214574 12.710386 11.647418  7.818922
9.805545  5.662589 14.808491  5.809585  8.732512 12.487733  8.146030
6.216196  5.901716  6.489493  7.509617  3.819655  7.923631
```

y

```
 12.901566 14.898103 19.722906 16.463436 19.689758 22.498737 20.082501
12.142602 15.509206 13.028103 13.517140 10.484259  4.246902 11.312456
18.731691 19.317701 14.559483 20.541582 10.797239 15.124459 20.216890
 9.673123 14.935897
```

Se pide:

1. *Vamos a introducir en el fichero que hemos construido en el primer apartado del ejercicio 7.5 (que sugería llamar* arsenico.txt *aunque casi seguro que nadie me ha hecho caso) una nueva variable. En ella vamos a colocar las nuevas concentraciones de arsénico.*
2. *Estimar el cambio medio que se ha producido en la concentración de arsénico en Murcia y en la Comunidad Valenciana.*
3. *¿Ha sido significativo el cambio en cada una de las comunidades autónomas con un nivel de significación de* 0,01.
4. *¿Es significativamente distinto el cambio observado en una y otra comunidad autónoma?*

Ejercicio 7.8. *Berthouex and Brown [2002, ejercicio 17.2] Medición de nitrito. Los siguientes datos se obtuvieron a partir de mediciones apareadas*

del nitrito en agua y en aguas residuales por el método del electrodo directo selectivo de iones (ISE) y un método colorimétrico. ¿Son los dos métodos consistentes?

```
ISE   0.32 0.36 0.24 0.11 0.11 0.44 2.79 2.99 3.47
Color 0.36 0.37 0.21 0.09 0.11 0.42 2.77 2.91 3.52
```

Ejercicio 7.9. *Berthouex and Brown [2002, ejercicio 17.3] Pruebas de demanda de oxígeno bioquímico. Los datos que figuran abajo son comparaciones por pares de las pruebas de demanda de oxígeno bioquímico hechos utilizando la botella estándar de 300 ml y con botellas experimentales de 60 ml. Se pide estimar la diferencia entre los resultados obtenidos con los dos tamaños de botella. Por estimar la diferencia entendemos tanto el estimador puntual de la diferencia media como el intervalo de confianza (que lo pedimos a un 90 %).*

```
300 mL 7.2 4.5 4.1 4.1 5.6 7.1 7.3 7.7 32 29 22 23 27
60 mL  4.8 4.0 4.7 3.7 6.3 8.0 8.5 4.4 30 28 19 26 28
```

Ejercicio 7.10. *Berthouex and Brown [2002, ejercicio 17.5] Seguimiento de una corriente. Una industria voluntariamente monitoriza un arroyo para determinar si su objetivo de elevar el nivel de contaminación en 4 mg / L o menos se verifica. Las mediciones que siguen para septiembre y abril se realizaron cada cuatro días de trabajo. ¿Se está cumpliendo el objetivo de la industria?*

Septiembre		Abril	
Rio arriba	Rio abajo	Rio arriba	Rio abajo
7.5	12.5	4.6	15.9
8.2	12.5	8.5	25.9
8.3	12.5	9.8	15.9
8.2	12.2	9.0	13.1
7.6	11.8	5.2	10.2
8.9	11.9	7.3	11.0
7.8	11.8	5.8	9.9
8.3	12.6	10.4	18.1
8.5	12.7	12.1	18.3
8.1	12.3	8.6	14.1

Ejercicio 7.11. *Un procedimiento importante para certificar la calidad del trabajo que se hace en un laboratorio es el análisis de muestras estándar que contienen cantidades conocidas de una cierta substancia. Estas muestras son introducidas en la rutina del laboratorio de modo que el analista no conoce la identidad de la muestra. A menudo el analista no conoce que estas muestras introducidas para evaluar la calidad del trabajo que se realiza han sido introducidas. En este ejemplo, se propuso a los analistas que midieran*

la concentración de oxígeno disuelto en una misma muestra con dos métodos distintos. Se enviaron muestras a 14 laboratorios preparadas con una baja concentración de oxígeno disuelto (1.2 mg/L). Cada laboratorio realizó sus determinaciones utilizando el método de Winkler y el método del electrodo. La cuestión que nos planteamos es si los dos métodos predicen distintas concentraciones de oxígeno disuelto.

Laboratorio	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
Winkler	*1.2*	*1.4*	*1.4*	*1.3*	*1.2*	*1.3*	*1.4*	*2.0*	*1.9*	*1.1*	*1.8*	*1.0*	*1.1*	*1.4*
Electrodo	*1.6*	*1.4*	*1.9*	*2.3*	*1.7*	*1.3*	*2.2*	*1.4*	*1.3*	*1.7*	*1.9*	*1.8*	*1.8*	*1.8*

Se pide:

1. *¿Podemos considerar que las mediciones realizadas con el método de Winkler difieren significativamente del valor* 1,2*? Responded a esta pregunta utilizando el intervalo de confianza y el contraste de hipótesis correspondiente.Trabajar con un nivel de confianza de* 0,9 *y un nivel de significación de* 0,1.

2. *¿Podemos considerar que las mediciones realizadas con el método del electrodo difieren significativamente del valor* 1,2*? Responded a esta pregunta utilizando el intervalo de confianza y el contraste de hipótesis correspondiente. Trabajar con un nivel de confianza de* 0,9 *y un nivel de significación de* 0,1.

3. *¿Difieren entre si los dos métodos de medición de la concentración de oxígeno disuelto? La pregunta hemos de responderla utilizando intervalos de confianza para la diferencia de medias y el contraste de hipótesis correspondiente. El nivel de confianza a utilizar es* 0,9 *y el nivel de significación* 0,1.

7.6 Test de Kolmogorov-Smirnov para dos muestras

En esta sección nos ocupamos del test no paramétrico conocido como test de Kolmogorov-Smirnov para dos muestras. [1]

Hasta ahora hemos comparado poblaciones asumiendo que ambas poblaciones tienen una distribución normal. Las variables que observábamos seguían una distribución normal. ¿Qué ocurre cuando no lo podemos asumir? ¿Qué ocurre cuando nuestros datos son *marcadamente* no normales?

En la figura 7.6 tenemos los estimadores kernel de las densidades de dos muestras que pretendemos comparar. Vemos claramente que las formas de las densidades estimadas no se parecen en nada a los de una normal.

Un dibujo q-q para la muestra x aparece en la figura 7.7 y para la muestra y en la figura 7.8. Vemos que en ambas figuras los puntos se alejan de la línea. Ninguna de las dos muestras puede considerarse

[1] Es aconsejable consultar `http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test`.

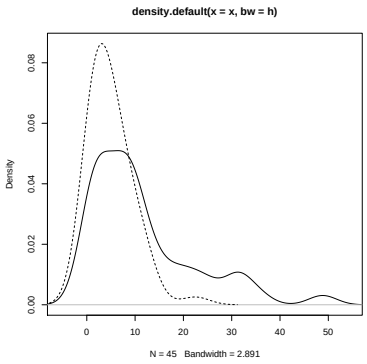


Figura 7.6: Estimadores kernel de la densidad de x (trazo continuo) e y (trazo discontinuo).

normal. Finalmente, si aplicamos un test de normalidad a las muestras (en concreto, vamos a utilizar el test de Shapiro-Wilk) obtenemos los siguientes resultados.

```
shapiro.test(x)

##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.84909, p-value = 3.469e-05

shapiro.test(y)

##
##  Shapiro-Wilk normality test
##
## data:  y
## W = 0.86647, p-value = 2.386e-05
```

Vemos cómo el test de normalidad rechaza claramente que podamos considerarlas muestras normales. No parece muy razonable utilizar un test de comparación de medias basada en la distribución t. ¿Hay otras opciones? Sí que las hay. Además no son malas soluciones. En esta sección nos ocupamos del test de Kolmogorov-Smirnov para dos muestras. Es un procedimiento *no paramétrico*. Un procedimiento se dice no paramétrico cuando no asume ningún modelo particular para los datos. No suponemos que son normales o que son binomiales o que son exponenciales, etc. A veces se les llama de distribución libre indicando que no estamos sujetos a un modelo específico. Esto suena bien. De hecho es bueno pero también tiene su pago. Asumimos menos hipótesis pero también son procedimientos con menos potencia. Les cuesta más rechazar la hipótesis nula.

La hipótesis nula que pretendemos contrastar la podemos formular como:

H_0: Las muestras han sido extraídas de una misma población.

H_1: Las muestras han sido extraídas de poblaciones distintas.

Para contrastar, como siempre, necesitamos un estadístico del contraste, un estadístico que compare ambas muestras. El estadístico no se basa en la comparación de las medias y varianzas muestrales como en los test anteriores. Denotamos las muestras como $x_1, \ldots, x_n$ e $y_1, \ldots, y_m$ de tamaños respectivos n y m. Consideramos las funciones de distribución empíricas o muestrales que denotamos F_n para

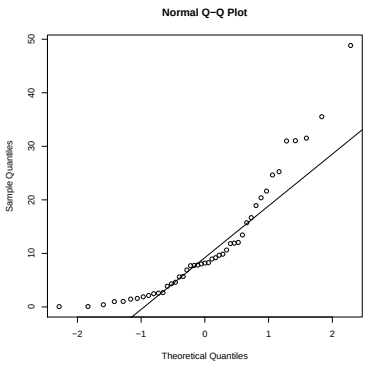


Figura 7.7: Dibujo q-q de la muestra x. Vemos que los punto se alejan de la línea indicando que no hay normalidad en los datos.

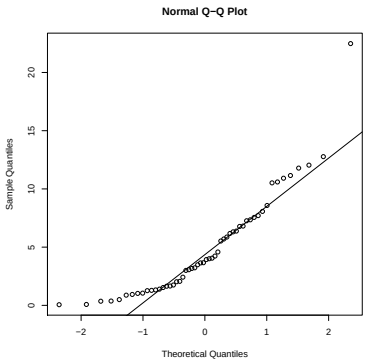


Figura 7.8: Dibujo q-q de la muestra y. Vemos que los punto se alejan de la línea indicando que no hay normalidad en los datos.

muestra de las x's y por G_m para la muestra de las y's. Es decir:

$$F_n(z) = \frac{|\{x_i : x_i \leq z\}|}{n}.$$

Donde $|\cdot|$ denota el cardinal del conjunto. En resumen, $F_n(z)$ está contando el número de valores en la muestra x que son menores o iguales que z. La función G_m se define de un modo análogo con la segunda muestra. En la figura 7.9 mostramos las dos funciones F_n y G_m.

El estadístico del test es

$$D = \max_z |F_n(z) - G_m(z)|, \tag{7.9}$$

es decir, D nos da la máxima diferencia que observamos entre las funciones de distribución muestrales F_n y G_m. En la figura 7.9 representamos con un segmento vertical la máxima diferencia entre ambas funciones, esto es, el valor del estadístico D. Por la definición del estadístico D es claro que rechazamos para valores grandes de D. Si d es el valor observado entonces el p-valor vendría dado por

$$p = P(D \geq d),$$

donde en la probabilidad anterior asumimos la hipótesis nula.

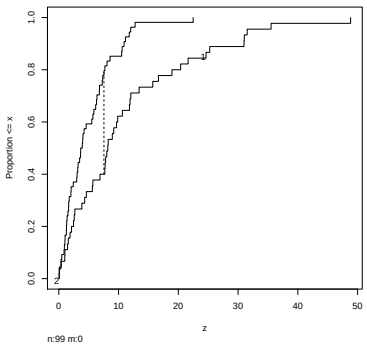


Figura 7.9: Funciones de distribución empíricas de ambas muestras. Vemos que la función de distribución de la segunda muestra (la muestra y indicada en la gráfica con el número 2) es mayor que la función de distribución empírica de la primera muestra (muestra x indicada con 1 en la gráfica). La longitud de la línea punteada nos muestra la máxima diferencia entre ambas funciones de distribución. Esta longitud es el valor del estadístico de

Nota de R 7.10 (Función *ks.test*)**.** *El test de Kolmogorov-Smirnov para dos muestras lo podemos aplicar con la función* ks.test *del siguiente modo:*

```
ks.test(x,y)

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.3963, p-value = 0.0005895
## alternative hypothesis: two-sided
```

La salida se autoexplica. Observamos un p-valor de 6×10^{-4}*. Si trabajamos con un nivel de significación de* $\alpha = 0{,}05$ *entonces como el p-valor es menor que este nivel rechazamos la hipótesis nula. Podemos decir que hay diferencias significativas en la distribución de los datos a un nivel de significación* 0,05*.*

Ejercicio 7.12. *En el fichero* tmmaxVALENCIA.txt *tenemos las temperaturas mínimas medias en la ciudad de Valencia desde el año 1937 hasta el 2011.*

1. *Leer los datos utilizando la función* read.table*.*

2. *Comparar las temperaturas máximas medias desde 1937 hasta 1950 con las temperaturas máximas medias desde el año 2000 hasta el 2011 utilizando un test de Kolmogorov-Smirnov para dos muestras.*

Ejercicio 7.13. *En el fichero* arsenico_por_compasion *(que podemos leer con la función* read.table*) tenemos las concentraciones de arsenico en distintas localidades de Murcia y de la Comunidad Valenciana. Tenemos el valor antes y después de una modificación del sistema de depuración de las aguas. La variable region indica la comunidad autónoma (1 para Murcia y 2 para la Comunidad Valenciana). Utilizando el test de Kolmogorov-Smirnov para dos muestras comparar las concentraciones de arsénico entre comunidades. Haced esto antes de la modificación. Repetirlo para los datos observados después de la modificación.*

7.7 Ejercicios globales

Ejercicio 7.14. *En el fichero* Temperatura1964-2011.csv *tenemos datos de temperatura en distintas poblaciones. Se pide:*

1. *Leer los datos utilizando el siguiente código (asumimos que el directorio de trabajo de R es el que tiene el fichero de datos).*

```
library(foreign)
x = read.csv("Temperatura1964-2011.csv",header=TRUE,sep=";")
attach(x)
```

2. *Nos fijamos en las temperaturas mínimas observadas en Morella y en Utiel en el año 1965.*

   a) *Obtener el estimador puntual de la diferencia entre las temperaturas medias mínimas entre las dos poblaciones.*

   b) *Obtener un intervalo de confianza al 90 % para la diferencia entre las temperaturas medias mínimas en las dos poblaciones.*

   c) *¿Podemos considerar, con un nivel de significación de* 0,05 *que no hay diferencia entre las temperaturas medias mínimas en las dos poblaciones?*

3. *Repetir al apartado 2 sustituyendo las poblaciones de Morella y Utiel por Valencia y Castellón. Ahora vamos a considerar el año 1976.*

4. *Cuando comparamos las temperatura media mínima entre Valencia y Castellón en 1976: ¿podemos admitir un valor de 16.5 con un nivel de significación de 0.1?*

5. *Repetir los apartado 2 utilizando las temperaturas máximas para las poblaciones de Valencia y Castellón. Estamos comparando ahora la temperatura media máxima en ambas poblaciones.*

Ejercicio 7.15. *En el fichero* tmminVALENCIA.txt *tenemos las temperaturas mínimas medias en la ciudad de Valencia desde el año 1937 hasta el 2011.*

1. *Leer los datos utilizando la función* read.table*.*
2. *Comparar las temperaturas medias mínimas observadas en el periodo de 1937 hasta 1962 con las temperaturas medias mínimas observadas desde 1990 en adelante utilizando un test de la t. Para ello previamente hemos de responder las siguientes preguntas.*
   a) *¿Podemos considerar que nuestros datos son normales?*
   b) *¿Podemos considerar que la varianza es la misma en ambos conjuntos de medidas?*
   c) *Comparar las medias teniendo en cuenta el apartado anterior?*
3. *Repetir el apartado 2 utilizando un test de Kolmogorov-Smirnov para dos muestras.*

8
Correlación y regresión

8.1 Curva de descenso de residuo

Un problema de interés en la investigación de calidad medioambiental es el de las curvas de descenso de residuo. Podemos tener un material tóxico disperso en una zona (parece razonable recordar el ejemplo del Prestige en la costa gallega), dioxinas en sedimentos acuáticos, pesticidas en campos de cultivo. Dado que tenemos este material tóxico en la zona en cuestión: ¿qué tiempo tardará en desaparecer? ¿O simplemente en reducirse? Parece razonable para responder estas preguntas tomar muestras en distintos instantes temporales y estudiar cómo se va modificando la concentración de dicho material según pasa el tiempo. Nuestros datos serían (x_i, z_i) con $i = 1, \ldots, n$ donde x_i es el i-ésimo tiempo de observación mientras que z_i sería la concentración medida en el instante x_i.

Perfectamente los datos podría corresponder a los mostrados en la figura 8.1. En abscisas tenemos tiempo en días y en ordenadas tenemos la concentración.

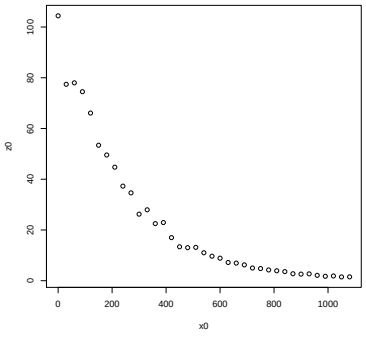


Figura 8.1: Datos para la curva de reducción de residuo. En abscisas tenemos el tiempo y en ordenadas la concentración de contaminante.

Un modelo que razonablemente podría aproximar el decrecimiento en la concentración ([1]) es

$$z = C_0 e^{-c_1 x} \tag{8.1}$$

Notemos que si en la ecuación 8.1 tomamos logaritmos (naturales) tendremos

$$ln(z) = ln(C_0) - c_1 x \tag{8.2}$$

Si denotamos $y = ln(z)$, $\beta_0 = ln(C_0)$ y $\beta_1 = -c_1$ realmente podemos escribir la ecuación anterior como

$$y = \beta_0 + \beta_1 x \tag{8.3}$$

En la figura 8.2 representamos los valores (x_i, y_i) donde $y_i = ln(z_i)$, es decir, mantenemos el tiempo original (en días) pero en lugar de

[1] Michael E. Ginevan and Douglas E. Splitstone. *Statistical Tools for Environmental Quality Measurement*. Chapman & Hall / CRC, 2004. URL /home/gag/BIBLIOGRAFIA/MISLIBROS/Ginevan_Splitstone_Statistical_Tools_for_Environmental_Quality_Measurement_Chapman_Hall_CRC_2004.pdf

considerar las concentraciones originales z tomamos el logaritmo natural de estos valores.

Vemos que hay una relación aproximadamente lineal entre el tiempo en que observamos y la concentración medida. Asumir que vamos a observar una relación como la anterior de un modo perfecto es absurdo. Errores de medida, el modelo (cualquier modelo) es una aproximación y muchas otras variables que no observamos y que no consideramos (lluvias en la zona por ejemplo) en este modelo hacen que lo que tenemos sea una *aproximación*. De hecho, cualquier modelo siempre será una aproximación (esperemos que buena) a la realidad.

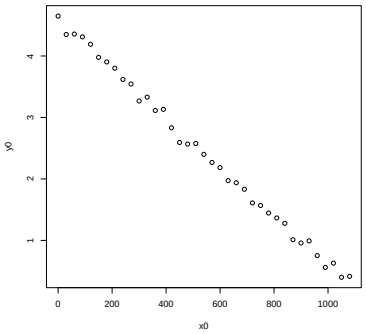


Figura 8.2: Datos para la curva de reducción de residuo. En abscisas tenemos el tiempo y en ordenadas el logaritmo natural de la concentración de contaminante.

Sin entrar en más detalles de modelo matemático. Partimos de unos valores observados (x_i, y_i) con $i = 1, \ldots, n$. ¿Cómo podemos determinar unos buenos valores para β_0 y β_1. La idea de Legendre fue encontrar los valores de β_0 y β_1 minimizando la siguiente función

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{8.4}$$

que como vemos es una de los cuadrados de la diferencia del valor exacto y_i y lo que debiera de valer si la relación fuera perfectamente lineal. Denotemos por $\hat{\beta_0}$ y $\hat{\beta_1}$ los valores de β_0 y β_1 que minimizan la función dada en 8.4. Se prueba (sin dificultad pero no es nuestro problema aquí) que tienen la siguiente expresión:

$$\hat{\beta_1} = \frac{S_{xy}}{s_x^2}, \quad \hat{\beta_0} = \bar{y}_n - \hat{\beta_1}\bar{x}_n, \tag{8.5}$$

donde

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{n-1}, \tag{8.6}$$

es la *covarianza muestral* de los valores (x_i, y_i) y

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2. \tag{8.7}$$

es la varianza muestral de los valores x_i.

8.2 *Ejemplos*

Veamos algunos otros ejemplos que vamos a analizar más tarde.

Ejemplo 8.1 (Datos orange)**.** *Los datos Orange están en el paquete* **?**, *datasets. Por ello lo podemos cargar con*

```
data(Orange)
```

En la figura 8.3 mostramos la circunferencia del tronco frente a la edad. En este caso pretendemos predecir la circunferencia del tronco a partir de la edad del arbol.

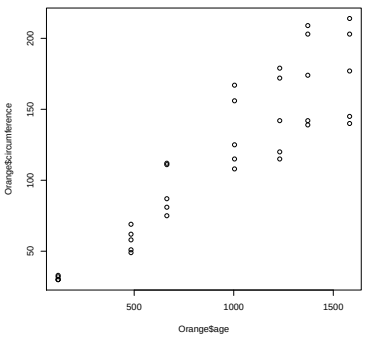


Figura 8.3: Datos Orange. En abscisas tenemos la edad del arbol y en ordenadas tenemos la circunferencia a la altura del pecho. Se aprecia una cierta dependencia lineal.

Ejemplo 8.2 (Temperatura en Valencia y Alicante)**.** *En el fichero valencia_alicante_temperaturas_anyo_1939_2010.txt Tenemos como variables las temperaturas mínimas y máximas para cada mes de los años que van de 1939 a 2010.*

En la figura 8.4 tenemos los datos. ¿Podemos predecir el valor de la temperatura en Alicante si tenemos la temperatura en Valencia?

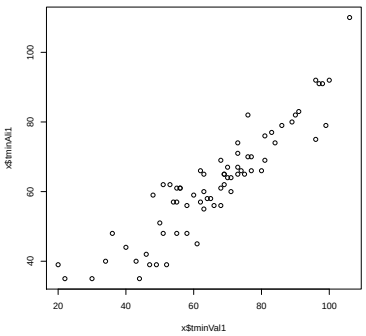


Figura 8.4: En abscisas las temperaturas mínimas (en el mes de enero desde el año 1939 hasta el 2010) en Valencia y en ordenadas las temperaturas mínimas (en el mismo mes y años) en Alicante.

8.3 Regresión lineal simple

En todos los ejemplos antes comentados el problema común es determinar el valor de Y a partir del valor de X. Obviamente la respuesta más simple sería buscar una función que podemos denotar por f de modo que para un valor dado x simplemente calculamos $y = f(x)$. Un poco de imaginación y conocimiento de la posible relación entre x e y podrían darnos una idea de qué función f buscar. Este planteamiento es de base muy restrictivo. ¿Por qué? Pues en primer lugar porque estamos asumiendo que, para un valor de x, existe un *único* valor de y asociado. Y esto nunca (o casi) es así. Un detalle, a veces X es una variable aleatoria que observamos simultáneamente con Y, en otras ocasiones es un valor que nosotros prefijamos (dosis de medicación, tratamiento en un problema de diseño de experimentos). Sin embargo, desde el punto de vista de la regresión X siempre lo consideramos fijo y estudiamos cómo se comporta Y dado el valor de $X = x$. Es decir, de la distribución condicionada de Y al valor de $X = x$.

Un ejemplo muy famoso de Francis Galton. Se tomaba como variable predictora la estatura del padre y como variable respuesta o a predecir, la estatura de un hijo. Es claro que para un mismo padre la estatura de sus hijos es variable. No todos los hijos de un mismo padre miden lo mismo. No tiene ningún sentido asumir una relación funcional entre la estatura de un padre y la de un hijo.

Tan tontos no son los estadísticos. De hecho, lo que se modeliza es la relación entre el valor x y *el valor medio* de la variable Y dado ese valor x. Siguiendo con el ejemplo de Galton. Si consideramos un padre de estatura $X = 178$ centímetros. Supondremos que la media de la variable Y que nos da la estatura aleatoria de un hijo es la que se relaciona con x. Denotemos por $E[Y \mid x]$ esta media (estatura media de todos los hijos de un padre con estatura 178 centímetros). Hemos de admitir que además de lo que mide el padre, algo tendrá que decir la madre, y también otros muchos factores que todos podemos imaginar. De modo que Y, conocida la estatura del padre, sigue siendo una cantidad aleatoria. De hecho, se asume que la distribución de Y es normal cuya media depende de Y, $E[Y \mid x]$, pero cuya varianza no depende de x, es decir, es una cantidad constante que

denotaremos por σ^2. En resumen, estamos asumiendo que

$$Y \sim N(E[Y \mid x], \sigma^2). \tag{8.8}$$

En el modelo de regresión más simple con el que se trabaja se asume que la media condicionada $E[Y \mid x]$ es una función lineal de x, en otras palabras, se asume que

$$E[Y \mid x] = \beta_0 + \beta_1 x. \tag{8.9}$$

Las hipótesis asumidas en 8.8 y 8.9, podemos expresarlas conjuntamente diciendo que la variable respuesta Y se puede expresar como

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{8.10}$$

donde

$$\epsilon \sim N(0, \sigma^2). \tag{8.11}$$

En la formulación de 8.10 expresamos el valor aleatorio de Y como suma de una parte que *sistemáticamente* depende de x (la componente sistemática del modelo) y un término aleatorio con distribución normal, un término de error o desajuste del modelo. En esta variable normal con media cero y varianza constante σ^2 estamos incluyendo todas las posibles causas que influyen el valor de Y y que no vienen dadas por la variable predictora.

No consideramos un solo valor aleatorio de Y dado un valor fijo de x. Realmente, tenemos n valores observados cuyos valores son *independientes entre sí pero no tienen la misma distribución*. Hemos de pensar que cada Y_i tiene una variable predictora distinta que influye en la distribución de Y_i. Tenemos pares (x_i, Y_i) donde la x_i viene dada y consideramos la distribución de Y_i condicionada a x_i, es decir, $Y_i \mid x_i$.

Resumiendo, estamos asumiendo que $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ y que los distintos Y_i son independientes entre si. Este modelo probabilístico es conocido como el **modelo de regresión lineal simple**.

Al vector $\beta = (\beta_0, \beta_1)$ le llamaremos el *vector de coeficientes*. Los estimadores de β los obtenemos minimizando la suma de cuadrados siguiente

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

y por ello reciben el nombre de *estimadores mínimo-cuadráticos*. Los denotaremos mediante $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$. Una vez tenemos las estimaciones $\hat{\beta}$ podemos obtener las *predicciones* de las observaciones con

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{8.12}$$

y los residuos (diferencia de la observación con la predicción) mediante

$$\hat{\epsilon}_i = y_i - \hat{y}_i. \tag{8.13}$$

Finalmente la varianza del error aleatorio la estimamos mediante

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\hat{\epsilon}_i^2}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2}. \tag{8.14}$$

La **suma de cuadrados residual** o **suma de cuadrados del error** que viene dada por

$$SS(Error) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{8.15}$$

De hecho,

$$\hat{\sigma}^2 = \frac{SS(Error)}{n-2}. \tag{8.16}$$

Nota de R 8.1 (La función *lm*)**.** *Vamos a realizar el análisis de regresión para la curva de descenso de residuo. La función básica es* lm.

```
lm(y0 ~ x0)

##
## Call:
## lm(formula = y0 ~ x0)
##
## Coefficients:
## (Intercept)           x0
##    4.575663    -0.003984
```

En la cual podemos ver los estimadores de los coeficientes. ¿Cómo interpretar estos coeficientes? La constante $\hat{\beta}_0$ sería la concentración en instante inicial (medida en escala logarítmica mientras que $\hat{\beta}_1$ es el cambio que se produce en la concentración por cada cambio unitario en la variable x que, en este caso, denota el tiempo en días. El resto de la información (y más cosas que veremos más tarde) la podemos obtener con la función genérica summary *aplicada al ajuste. Por ejemplo, con el siguiente código.*

```
y0.fit = lm(y0 ~ x0)
summary(y0.fit)

##
## Call:
## lm(formula = y0 ~ x0)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.189735 -0.036399 -0.000206  0.049226  0.142053
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  4.576e+00  2.512e-02  182.12   <2e-16 ***
## x0          -3.984e-03  4.002e-05  -99.56   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07797 on 35 degrees of freedom
## Multiple R-squared:  0.9965,Adjusted R-squared:  0.9964
## F-statistic:  9911 on 1 and 35 DF,  p-value: < 2.2e-16
```

De momento podemos ver que los coeficientes estimados aparecen con la etiqueta Estimate *(estimación). También podemos ver un resumen de los residuos (mínimo, máximo, primer y tercer cuartil y el segundo cuartil o mediana). En esta salida vemos que la estimación de la desviación estándar σ es 0.078.*

Las predicciones de las observaciones (mostramos las diez primeras solamente) las obtenemos con

```
predict(y0.fit)[1:10]
```

```
##        1        2        3        4        5        6
## 4.575663 4.456146 4.336629 4.217112 4.097595 3.978077
##        7        8        9       10
## 3.858560 3.739043 3.619526 3.500009
```

Y los residuos (también los correspondientes a las diez primeras observaciones) vendrían dados por

```
y0[1:10] - predict(y0.fit)[1:10]
```

```
##             1             2             3             4
##  0.0727865378 -0.1067740252  0.0200518051  0.0938117992
##             5             6             7             8
##  0.0933779005 -0.0002063006  0.0446152623  0.0623259440
##             9            10
## -0.0009408336  0.0444120672
```

o bien simplemente con

```
residuals(y0.fit)[1:10]
```

```
##             1             2             3             4
##  0.0727865378 -0.1067740252  0.0200518051  0.0938117992
##             5             6             7             8
##  0.0933779005 -0.0002063006  0.0446152623  0.0623259440
##             9            10
## -0.0009408336  0.0444120672
```

8.3.1 *Intervalos de confianza y contrastes para los coeficientes*

¿Depende realmente el logaritmo de la concentración del tiempo o más o menos es constante y no observamos ninguna modificación? Si observamos el modelo de regresión lineal simple

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

donde $\epsilon_i \sim N(0, \sigma^2)$, entonces si el coeficiente β_1 vale cero significa que la variable predictora x no aparece en el modelo y podemos considerar que (salvo variaciones aleatorias) la variable Y no varía con el tiempo. En resumen que responder la pregunta anterior lo podemos plantear como contrastar la hipótesis nula de que el coeficiente β_1 es nulo frente a la alternativa de que no lo es. El contraste lo formulamos como

$$H_0 \quad : \quad \beta_1 = 0, \tag{8.17}$$
$$H_1 \quad : \quad \beta_1 \neq 0. \tag{8.18}$$

Para contrastar estas hipótesis hemos de tener en cuenta que el estimador $\hat{\beta}_1$ tiene una distribución normal

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right) \tag{8.19}$$

En particular, la varianza de $\hat{\beta}_1$ es $\sigma^2 / \sum_{i=1}^n (x_i - \bar{x}_n)^2$. Como sabemos la raiz cuadrada de la varianza del estimador es lo que llamamos su error estándar. En resumen, el error estándar de $\hat{\beta}_1$ es $SE(\hat{\beta}_1) = \sqrt{\sigma^2 / \sum_{i=1}^n (x_i - \bar{x}_n)^2}$. No conocemos (obviamente) la varianza σ^2 del error aleatorio. Hemos visto cómo estimarla en la ecuación 8.16. El error estándar estimado de $\hat{\beta}_1$ será

$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}.$$

Se verifica que

$$\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \sim t_{n-2}. \tag{8.20}$$

Utilizando este resultado el intervalo de confianza para β_1 se sigue inmediatamente y para un nivel de confianza de $1 - \alpha$ sería

$$[\hat{\beta}_1 - t_{n-2,1-\alpha/2} \widehat{SE}(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,1-\alpha/2} \widehat{SE}(\hat{\beta}_1)].$$

Obviamente si suponemos que se verifica la hipótesis nula $H_0 : \beta_1 = 0$ entonces lo enunciado en 8.20 se puede reformular como

$$T_1 = \sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \frac{\hat{\beta}_1}{\hat{\sigma}} \sim t_{n-2}. \tag{8.21}$$

Utilizando este resultado podemos contrastar hipótesis sobre β_1. En concreto, un contraste con un nivel de significación de α supone rechazar la hipótesis nula cuando

$$|T_1| > t_{n-2,1-\alpha/2}.$$

Nota de R 8.2. *En particular si vemos el resumen del ajuste.*

```
summary(y0.fit)

##
## Call:
## lm(formula = y0 ~ x0)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.189735 -0.036399 -0.000206  0.049226  0.142053
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.576e+00  2.512e-02  182.12   <2e-16 ***
## x0          -3.984e-03  4.002e-05  -99.56   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07797 on 35 degrees of freedom
## Multiple R-squared:  0.9965,Adjusted R-squared:  0.9964
## F-statistic:  9911 on 1 and 35 DF,  p-value: < 2.2e-16
```

Podemos ver que el valor del error estándar de $\hat{\beta}_1$ es 0, el valor del estadístico T_1 es $-99{,}5548$. El área de las dos colas (izquierda de $-|T_1|$ y derecha de $|T_1|$ es 0. Los intervalos de confianza para los coeficientes los podemos obtener con

```
confint(y0.fit)

##                   2.5 %       97.5 %
## (Intercept)  4.524656977  4.626669674
## x0          -0.004065146 -0.003902667
```

Vemos que el nivel de confianza de estos intervalos es del 95 %. Podemos modificar el nivel de confianza, por ejemplo, vamos a considerar un nivel de confianza del 99 %.

```
confint(y0.fit,level=0.99)

##                    0.5 %       99.5 %
## (Intercept)  4.507227862  4.644098789
## x0          -0.004092905 -0.003874907
```

Ejercicio 8.1. *En el fichero valencia_alicante_temperaturas_mes_1939_2010.txt tenemos como variables los años de 1939 a 2010 y como observaciones los distintos meses del año. Vamos a considerar como variable predictora la temperatura mínima en 1962 (buen año) y como respuesta la temperatura mínima en 2002 (mal año). Se pide:*

1. *Ajusta un modelo de regresión lineal simple. Obtener el valor de los coeficientes.*
2. *¿Es un buen ajuste atendiendo al coeficiente de determinación.*
3. *Realizar un dibujo que en abscisas tenga las predicciones y en ordenadas los residuos. ¿Qué indica este dibujo? Interpretarlo.*
4. *¿Cuál es el máximo residuo observado? ¿A qué observación corresponde?*

Ejercicio 8.2. *Consideremos los datos Orange. Vamos a considerar como variable predictora la edad del arbol y como variable respuesta la circunferencia observada. Se pide:*

1. *Ajusta un modelo de regresión lineal simple. Obtener el valor de los coeficientes.*
2. *¿Es un buen ajuste atendiendo al coeficiente de determinación.*
3. *Realizar un dibujo que en abscisas tenga las predicciones y en ordenadas los residuos. ¿Qué indica este dibujo? Interpretarlo.*
4. *¿Cuál es el máximo residuo observado? ¿A qué observación corresponde?*

Ejercicio 8.3. *Consideremos los datos Orange. Vamos a considerar como variable predictora la circunferencia y como variable respuesta la edad del arbol. Se pide:*

1. *Ajusta un modelo de regresión lineal simple. Obtener el valor de los coeficientes.*
2. *¿Es un buen ajuste atendiendo al coeficiente de determinación.*
3. *Realizar un dibujo que en abscisas tenga las predicciones y en ordenadas los residuos. ¿Qué indica este dibujo? Interpretarlo.*
4. *¿Cuál es el máximo residuo observado? ¿A qué observación corresponde?*

8.4 *Coeficiente de correlación de Pearson*

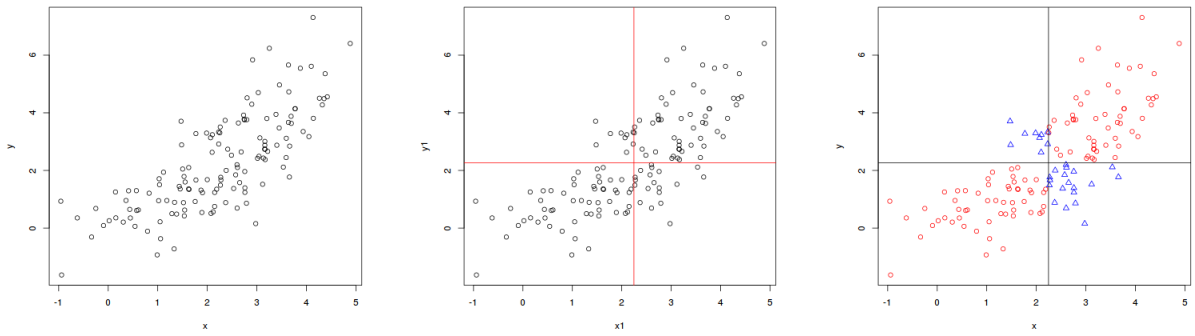
En la sección anterior nos planteábamos cómo, a partir de una variable a la que llamamos predictora, aproximar el valor de una variable a la que llamamos variable respuesta. Supongamos que nos planteamos una pregunta más simple: ¿están relacionadas linealmente las dos variables que estamos considerando: las variables x e y? Y la respuesta la hemos de dar utilizando los datos (x_i, y_i) con $i = 1, \ldots, n$. Un valor que se utiliza frecuentemente para responder esta pregunta es el *coeficiente de correlación de Pearson*. Se define del siguiente modo.

Fué propuesto por Karl Pearson.

Definición 8.1 (Coeficiente de correlación de Pearson)**.**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}.$$

Veamos una ilustración gráfica del concepto. En la figura 8.5(a) mostramos los datos con los que vamos a trabajar.



(a) (b) (c)

Figura 8.5: a) Datos. b) Los datos con dos líneas: la línea horizontal que corta al eje de ordenadas en $\bar{y}$ y la línea vertical que corta al eje de abscisas en $\bar{x}$. c) Los datos, la línea horizontal que corta al eje de ordenadas en $\bar{y}$ y la línea vertical que corta al eje de abscisas en $\bar{x}$. Representamos en rojo aquellos puntos donde el producto cruzado $(x_i - \bar{x})(y_i - \bar{y})$ es positivo y en azul aquellos puntos donde el producto toma un valor negativo.

En la figura 8.5(b) añadimos al diagrama de puntos mostrado en la figura 8.5(a) un par de líneas que se cortan en el punto $(\bar{x}, \bar{y})$: la línea horizontal que corta al eje de ordenadas en $\bar{y}$ y la línea vertical que corta al eje de abscisas en $\bar{x}$.

Consideremos los productos cruzados $(x_i - \bar{x})(y_i - \bar{y})$. En la figura 8.5(c) reproducimos el dibujo de la figura **??**(b). Simplemente, representamos en rojo aquellos puntos donde el producto cruzado anterior es positivo y en azul aquellos puntos donde el producto toma un valor negativo.

Tenemos una relación razonablemente lineal entre las dos variables y por ello vemos muchos más puntos rojos que azules. El coeficiente de correlación lineal de Pearson vale

```
cor(x1,y1)

## [1] 0.7629915
```

Vamos a repetir el mismo análisis con otros dos conjuntos de datos. En uno de ellos la asociación lineal es mayor mientras que en el otro buscaremos que no haya prácticamente asociación lineal entre las variables.

En la figura 8.6(a) mostramos los datos con fuerte asociación lineal. Distinguimos por el signo del producto cruzado en la figura 8.6(b) Vemos cómo hay muchos puntos azules. El nuevo coeficiente de correlación lineal de Pearson es

```
cor(x2,y2)

## [1] 0.990152
```

Mayor que en el caso anterior. Y un tercer ejemplo donde la asociación lineal es casi inexistente. En las figuras 8.6(c) y (d) mostramos los dibujos análogos al caso anterior. Vemos cómo hay muchos más puntos azules que en los dos casos anteriores. El nuevo coeficiente de correlación lineal de Pearson es

```
cor(x3,y3)

## [1] 0.02126659
```

Esta es la ilustración gráfica de la idea que subyace a este concepto. Cuanto más se aproximan los puntos a una línea recta mayor es el valor absoluto del coeficiente de correlación, más próximo a uno. Si cuando x crece y crece entonces la recta el valor de coeficiente de correlación es positivo. Si cuando x crece y decrece entonces el coeficiente de correlación es negativo.

Si vemos la definición 8.1 que acabamos de dar se tiene que

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

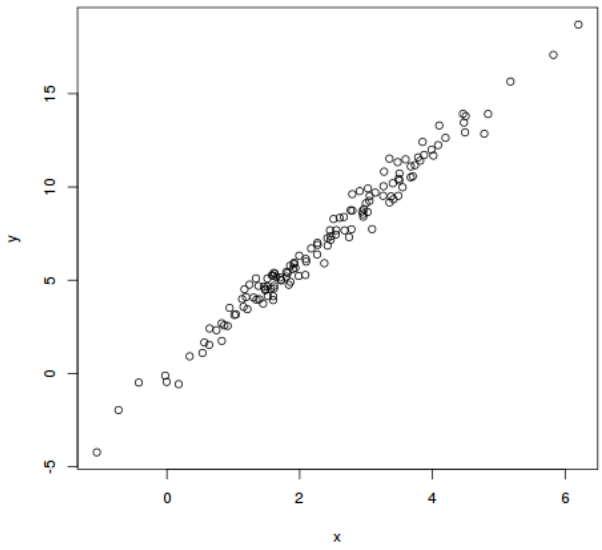
La *covarianza muestral*

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{n-1},$$

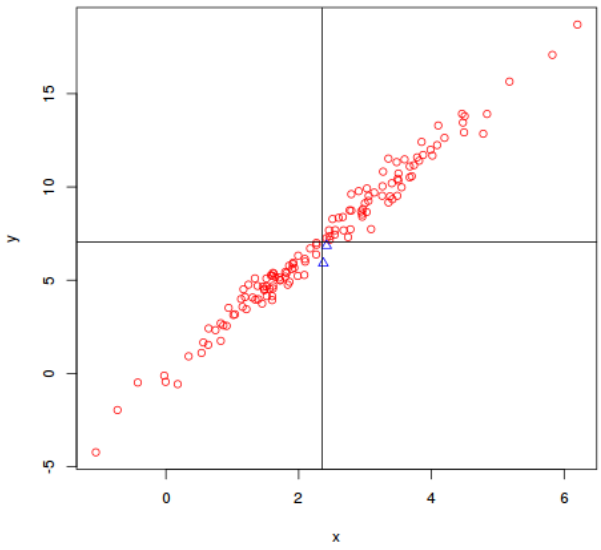
está *estimando* la cantidad

$$E(X - \mu_X)(Y - \mu_Y),$$

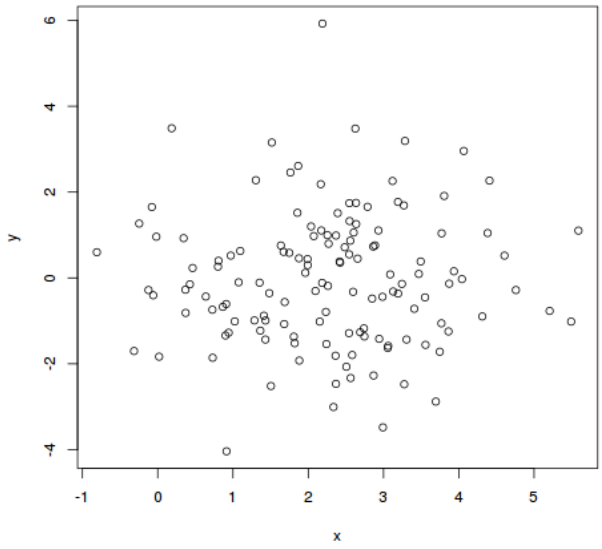
mientras que las desviaciones típicas muestrales, s_x y s_y, no son más que estimaciones de las desviaciones típicas poblacionales σ_X



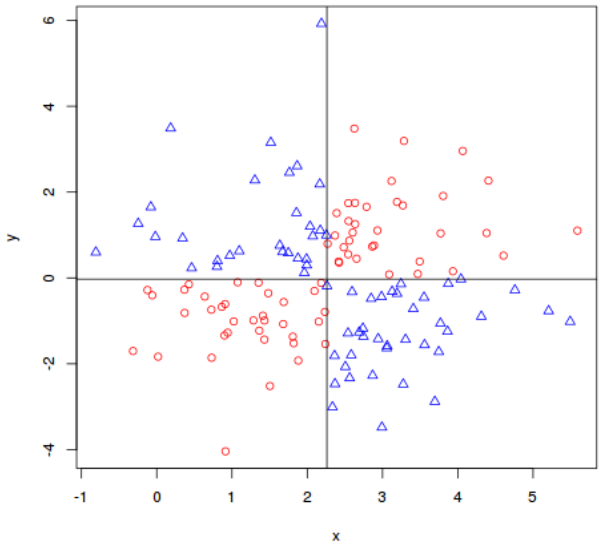
(a)



(b)



(c)



(d)

Figura 8.6: Ejemplo con fuerte asociación lineal: datos (a) y los datos diferenciando el signo del producto cruzado. Las figuras (c) y (d) son los dibujos análogos con datos en los que apenas hay asociación lineal entre las abscisas y las ordenadas.

y σ_Y. En resumen que $r = s_{xy}/(s_x s_y)$ no es más que un estimador del coeficiente de correlación de Pearson que damos en la siguiente definición.

Definición 8.2. *Dadas dos variables aleatorias X e Y definimos su coeficiente de correlación de Pearson como*

$$\rho = \frac{E(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y}.$$

Esta cantidad es una cuantificación de la asociación lineal entre X e Y. Es importante notar que aquí no hay predictor y respuesta. Ambas variables son tratadas de un modo simétrico. De hecho si intercambiamos las variables el valor del coeficiente de correlación no se modifica, es el mismo. Además se tiene que

$$-1 \leq \rho \leq 1.$$

Se puede probar que si $\rho = 1$ entonces existe dos constantes a y b (con b positiva) tales que $Y = a + bX$ (con probabilidad uno). Si $\rho = -1$ entonces existen a y b (con b negativa) tales que $Y = a + bX$ (con probabilidad uno). En resumen que si el coeficiente de correlación es 1 o -1 entonces una variable es función lineal de la otra. La recta es creciente si la correlación es positiva y decreciente si es negativa.

Ejemplo 8.3 (Curva de descenso de residuo)**.** *Veamos el grado de asociación lineal del tiempo con la concentración de contaminante.*

```
cor(x0,z0)

## [1] -0.8746311
```

Y ahora vamos a ver lo mismo pero con el logaritmo natural de la concentración del contaminante.

```
cor(x0,y0)

## [1] -0.998239
```

Es claramente mayor la segunda.

8.5 Regresión lineal múltiple

Pretendemos determinar la relación que liga a una variable respuesta Y como función de $p-1$ variables predictoras, $x_1, \ldots, x_{p-1}$. Siguiendo el razonamiento anterior podemos plantearnos un modelo muy general como el que sigue.

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} + \epsilon, \tag{8.22}$$

donde ϵ es el término del error. Realmente observamos n vectores $(y_i, x_{i1}, \ldots, x_{i,p-1})$ en consecuencia nuestro modelo estocástico ha de considerar el modelo para los n valores aleatorios Y_i, donde cada Y_i tiene asociado un vector x_i. Vamos a suponer que para una combinación de valores $(x_{i1}, \ldots, x_{i,p-1})$ vamos a observar un valor aleatorio Y_i con distribución normal cuya media es $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1}$ y cuya varianza va a ser constante e igual a σ^2. Además los distintos Y_i son independientes entre si.

En realidad nuestro modelo probabilístico es

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \epsilon_i$$

donde los ϵ_i son independientes y con la misma distribución normal con media nula y varianza σ^2.

8.6 *Estimación de β*

¿Cómo estimamos los parámetros $\beta = (\beta_0, \ldots, \beta_{p-1})$? Nuestros datos son $(y_i, x_{i1}, \ldots, x_{i,p-1})$ con $i = 1, \ldots, n$. Nuestro objetivo es estimar los coeficientes β de modo que $\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$ esté próximo a y. En concreto vamos a minimizar

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}) \right)^2 \tag{8.23}$$

Si consideramos la siguiente matriz

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n1} & \ldots & x_{n,p-1} \end{bmatrix}$$

y asumimos que la matriz $X'X$ (X' es la matrix traspuesta de X) es una matriz no singular (tiene inversa) entonces tendremos que los estimadores de los coeficientes $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ vienen dados por

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = (X'X)^{-1} X'y, \tag{8.24}$$

con $y = (y_1, \ldots, y_n)'$.

Si $(X'X)^{-1} = [a_{ij}]_{i,j=1,\ldots,p}$ entonces el estimador de la varianza de $\hat{\beta}_i$, $var(\hat{\beta}_i)$, sería $a_{ii}\hat{\sigma}^2$. Finalmente el error estándar de $\hat{\beta}_i$, es decir, su desviación típica (raíz cuadrada de su varianza) sería

$$\widehat{SE}(\hat{\beta}_i) = \sqrt{a_{ii}}\hat{\sigma}. \tag{8.25}$$

Se tiene que

$$\hat{\beta}_i \sim N(\beta_i, a_{ii}\hat{\sigma}^2). \tag{8.26}$$

Para la observación i-ésima tendremos la predicción

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{ij}. \tag{8.27}$$

Los residuos esto es las diferencias entre los valores observados originalmente y las predicciones que de ellos hacemos, vienen dados por

$$\hat{\epsilon}_i = y_i - \hat{y}_i. \tag{8.28}$$

Finalmente, hemos determinado los coeficientes que nos minimizaban la suma de cuadrados.

Ejemplo 8.4 (Ahorro)**.** *Los datos que vamos a utilizar son los datos* savings *contenido en el paquete* faraway*. Se pretende estudiar la relación que liga la fracción de ahorro con la proporción de población menor de 15 años, mayor de 75 y las variables dpi y ddpi.*

```
library(faraway)
data(savings)
attach(savings)
```

Ajustamos el modelo.

```
savings.lm = lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
```

Y vemos un resumen del ajuste.

```
summary(savings.lm)

##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
```

```
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

En este resumen los estimadores de los coeficientes vienen dados por

```
coefficients(savings.lm)
```

```
##   (Intercept)         pop15         pop75           dpi
## 28.5660865407 -0.4611931471 -1.6914976767 -0.0003369019
##          ddpi
##  0.4096949279
```

8.7 Bondad de ajuste

Hemos supuesto una relación lineal entre la media de la variable respuesta y las variables predictoras. La primera pregunta que hay que responder es: ¿tenemos un ajuste razonable? La respuesta se da utilizando medidas que comparan los valores observados con las predicciones asumiendo el modelo, es decir, comparando y_i con $\hat{y}_i$ para los distintos datos. En concreto, con diferencia la más utilizada es el **coeficiente de determinación** que se denota por R^2 y se define como

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}. \tag{8.29}$$

El ajuste que estamos realizando se supone que será tanto mejor cuanto más pequeña sea $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. El coeficiente de determinación toma valores entre 0 y 1 y cuanto más cerca de 1 mejor es el ajuste.

Tiene un pequeño inconveniente y es que no tiene en cuenta el número de variables predictoras que estamos utilizando. Una pequeña modificación de R^2 para incorporar esta información es el **coeficiente de determinación ajustado** que podemos denotar R^2-ajustado y se define como

$$R^2_{ajustado} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}, \tag{8.30}$$

donde suponemos que tenemos $p-1$ variables predictoras.

Ejemplo 8.5 (Ahorro). *Para el ahorro los coeficientes de determinación sin ajustar y ajustado los tenemos en el resumen del ajuste.*

```
summary(savings.lm)

##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

siendo el coeficiente de determinación 0.3385 y el coeficiente de determinación ajustado 0.2797.

8.8 Inferencia sobre el modelo

Hemos formulado un modelo probabilístico en donde relacionamos la variable respuesta con una serie de variables predictoras. Es claro que el experimentador introduce en el modelo como variables predictoras variables que a priori sospecha que pueden ser relevantes a la hora de predecir. Esto no quiere decir que luego podamos prescindir de alguna o algunas de ellas. Bien porque se demuestra que dicha variable no es relevante o bien porque la información que contiene esa variable predictora está contenida en las otras.

Supongamos que nos planteamos el siguiente contraste de hipóte-

sis:

$$H_0 \quad : \quad \beta_{i_1} = \ldots = \beta_{i_r} = 0 \tag{8.31}$$

$$H_1 \quad : \quad \text{existe algún j tal que } \beta_{i_j} \neq 0. \tag{8.32}$$

Si un coeficiente determinado β_i es nulo entonces la variable respuesta Y no dependería de la variable asociada a dicho coeficiente. En definitiva, la hipótesis nula considerada se podría formular diciendo que la variable Y no depende de las variables $x_{i_1}, \ldots, x_{i_r}$. ¿Cómo contrastar las hipótesis indicadas?

Vamos a considerar en primer lugar el modelo en que tenemos todas las p-1 posibles variables predictoras, esto es, el modelo más completo que podemos considerar. En este modelo tendremos una suma de cuadrados del error que denotamos por $SS(Error)$. Ahora vamos a considerar el modelo que se verifica bajo la hipótesis nula, el modelo que no tiene las variables predictoras $x_{i_1}, \ldots, x_{i_r}$. En este segundo modelo (simplificado) tendremos una suma de cuadrados (mayor) que denotamos por $SS(Error)_0$. Se verifica si la hipótesis nula es cierta (asumiendo que es cierta) que

$$F = \frac{(SS(Error)_0 - SS(Error))/r}{SS(Error)/(n-p)} \sim F_{r,n-p}. \tag{8.33}$$

De modo que rechazaremos la hipótesis nula de que $H_0 : \beta_{i_1} = \ldots = \beta_{i_r} = 0$ si

$$F > F_{r,n-p,1-\alpha}$$

donde $F_{r,n-p,1-\alpha}$ es el percentil $1-\alpha$ de una F con r y $n-p$ grados de libertad.

8.8.1 ¿Podemos prescindir de todas las variables predictoras?

¿Realmente depende la variable respuesta de alguna de las variables predictoras? Realmente nos estamos planteando la hipótesis de que todos los coeficientes, salvo el término constante β_0, valen cero, es decir, la hipótesis nula $H_0 : \beta_1 = \ldots = \beta_{p-1} = 0$. En este caso tendremos que

$$F = \frac{(SS(Total) - SS(Error))/(p-1)}{SS(Error)/(n-p)} \sim F_{p-1,n-p}.$$

Ejemplo 8.6 (Ahorro). *En el resumen*

```
summary(savings.lm)

##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

tenemos que el p-valor cuando contrastamos si todas las variables las podemos considerar nulas viene en la última línea.

8.8.2 ¿Podemos prescindir de una variable predictora?

Como segundo caso tendríamos la situación en que contrastamos que un solo coeficiente vale cero, es decir, la hipótesis nula $H_0 : \beta_i = 0$ frente a la alternativa $H_1 : \beta_i \neq 0$. Tenemos que bajo la hipótesis nula indicada

$$t_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \sim t_{n-p}$$

donde $SE(\hat{\beta}_i)$ es el error estándar de $\hat{\beta}_i$ y viene dado en ecuación 8.25. Se tiene, de hecho, que

$$F = t_i^2.$$

Rechazaremos la hipótesis nula si

$$|t_i| > t_{n-p,1-\frac{\alpha}{2}}$$

o bien si

$$F = t_i^2 > F_{1,n-p,1-\frac{\alpha}{2}}.$$

Ambos procedimientos son equivalentes.

Ejemplo 8.7 (Ahorro)**.** *En el resumen*

```
summary(savings.lm)

##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

los p-valores correspondientes a si el coeficiente asociado a cada variable lo podemos considerar nulo lo podemos observar en la línea correspondiente a cada variable.

8.8.3 ¿Podemos prescindir de un conjunto dado de variables?

El contraste planteado en 8.34 nos permite valorar, dado un conjunto de variables predictoras si un subconjunto dado puede ser eliminadas del modelo de regresión sin que el ajuste global empeore de un modo apreciable.

Ejemplo 8.8 (Ahorro)**.** *Con los datos savings vamos a plantearnos si podemos prescindir* simultánemante *de las variables* pop75 *y de la variable* dpi*. Podemos formular el contraste como:*

$$H_0 \quad : \quad \beta_{pop75} = \beta_{dpi} = 0 \tag{8.34}$$

$$H_1 \quad : \quad \beta_{pop75} \neq 0 \text{ ó } \beta_{dpi} \neq 0. \tag{8.35}$$

Empezamos ajustando los modelos por separado. Primero el modelo más completo.

```
savings.lm.1 = lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
```

Y ahora ajustamos el modelo al que quitamos las variables pop75 y dpi.

```
savings.lm.2 = lm(sr ~ pop15 + ddpi, savings)
```

Notemos que el modelo 2 es un submodelo, quitamos variables que estaban en el modelo 1 pero no añadimos ninguna variable que no estuviera ya en el modelo 1. El contraste planteado en 8.34 se puede realizar con la función anova.

```
anova(savings.lm.1,savings.lm.2)

## Analysis of Variance Table
##
## Model 1: sr ~ pop15 + pop75 + dpi + ddpi
## Model 2: sr ~ pop15 + ddpi
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     45 650.71
## 2     47 700.55 -2   -49.839 1.7233   0.19
```

Podemos ver que el p-valor asociado es muy grande por lo que no rechazamos la hipótesis de que ambos coeficientes sean nulos. Podemos eliminarlos del modelo original y quedarnos con el modelo simplificado.

8.9 Ejemplos de regresión lineal múltiple

Ejemplo 8.9 (Curva de reducción de residuo)**.** *Vamos a continuar con el ejemplo de la curva de reducción de residuos. Trabajamos con otros datos. Se tomaron cada 10 días. En la toma de la información se eligió una zona homogénea (un trozo de playa o bien un trozo de acantilado). Una vez elegida la zona se tomaron al azar dentro de esa zona 30 puntos de muestreo. En cada uno de estos puntos se midió la concentración de petróleo.*

Los datos originales aparecen en la figura 8.7.

Transformamos como antes la variable respuesta, concentración del contaminante a la escala logarítmica. Y representamos de nuevo los datos en la figura 8.8.

Podemos ver que los datos no parecen ajustarse a una recta. Más bien sugiere que hay alguna componente cuadrática. Es decir, que el logaritmo (natural) de la concentración podemos aproximarlo con una función del tipo

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

siendo x el tiempo. Vamos a probar los dos modelos. El modelo que solamente utiliza como predictora x (modelo 1) y el modelo que utiliza como variables predictoras el tiempo x y el cuadrado del tiempo x^2. Empezamos ajustando el primer modelo.

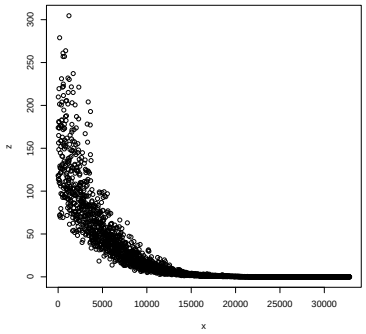


Figura 8.7: Curva de reducción de residuo.

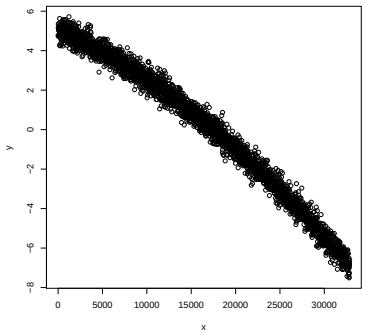


Figura 8.8: Curva de reducción de residuo. La concentración es dada en escala logarítmica.

```
prestige.ajuste1 = lm(y ~ x)
```

El segundo modelo lo podemos ajustar con

```
prestige.ajuste2 = lm(y ~ x + x*x)
```

o bien con

```
prestige.ajuste2 = lm(y ~ poly(x,2))
```

Como no cuesta mucho esfuerzo. Podemos probar con un tercer modelo en el que tengamos un polinomio de orden 4 (o de grado 3 como se quiera).

```
prestige.ajuste3 = lm(y ~ poly(x,3))
```

Veamos un resumen de cada uno de los dos ajustes. El resumen del primer modelo es

```
summary(prestige.ajuste1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60744 -0.34855  0.03014  0.37552  1.66456
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.938e+00  1.823e-02   325.7   <2e-16 ***
## x           -3.632e-04  9.611e-07  -377.9   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5226 on 3284 degrees of freedom
## Multiple R-squared:  0.9775,Adjusted R-squared:  0.9775
## F-statistic: 1.428e+05 on 1 and 3284 DF,  p-value: < 2.2e-16
```

y el resumen del segundo ajuste es

```
summary(prestige.ajuste2)

##
## Call:
```

```
## lm(formula = y ~ poly(x, 2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14817 -0.23281 -0.00191  0.23006  1.24704
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.719e-02  6.013e-03   -4.523 6.32e-06 ***
## poly(x, 2)1 -1.975e+02  3.447e-01 -572.981  < 2e-16 ***
## poly(x, 2)2 -2.252e+01  3.447e-01  -65.327  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3447 on 3283 degrees of freedom
## Multiple R-squared:  0.9902,Adjusted R-squared:  0.9902
## F-statistic: 1.663e+05 on 2 and 3283 DF,  p-value: < 2.2e-16
```

Finalmente, el resumen del modelo con un polinomio hasta grado 3 es

```
summary(prestige.ajuste3)

##
## Call:
## lm(formula = y ~ poly(x, 3))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13862 -0.23425 -0.00212  0.22674  1.25148
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.719e-02  6.011e-03   -4.524 6.29e-06 ***
## poly(x, 3)1 -1.975e+02  3.446e-01 -573.089  < 2e-16 ***
## poly(x, 3)2 -2.252e+01  3.446e-01  -65.339  < 2e-16 ***
## poly(x, 3)3  5.149e-01  3.446e-01    1.494    0.135
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3446 on 3282 degrees of freedom
## Multiple R-squared:  0.9902,Adjusted R-squared:  0.9902
## F-statistic: 1.109e+05 on 3 and 3282 DF,  p-value: < 2.2e-16
```

Ejemplo 8.10 (Precio de una vivienda)**.** *Vamos a trabajar con un banco de datos relativo a precios de la vivienda. Es un fichero que viene con el paquete SPSS. Tenemos las siguientes variables:*

VALTERR Valor de tasación del terreno.

VALMEJOR Valor de tasación de las mejoras.

VALTOT Valor de tasación total.

PRECIO Precio de venta.

TASA Razón del precio de venta sobre el valor de tasación total.

BARRIO Barrio en el que se encuentra la vivienda.

Nos planteamos predecir el precio de venta de la vivienda utilizando como variables predictoras el valor de tasación del terreno y de las mejoras. Notemos que el valor total no es más que la suma de la tasación del terreno más el valor de las mejoras. @

Comenzamos leyendo los datos. Notemos que por estar en formato de SPSS utilizamos el paquete R Core Team [2015b, foreign].

```
library(foreign)
x = read.spss(file='../data/venta_casas.sav',to.data.frame=T)
attach(x)
```

Nos planteamos predecir el precio de la vivienda utilizando como variables predictoras el precio de terreno y el valor de las mejoras.

```
(casas.lm  = lm(precio ~ valterr + valmejor))

##
## Call:
## lm(formula = precio ~ valterr + valmejor)
##
## Coefficients:
## (Intercept)      valterr     valmejor
##    767.4080       3.1916       0.4779
```

Veamos un resumen del ajuste.

```
summary(casas.lm)

##
## Call:
## lm(formula = precio ~ valterr + valmejor)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -153634  -10451    -576    8690  356418
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.674e+02  1.290e+03   0.595    0.552
## valterr     3.192e+00  5.339e-02  59.777   <2e-16 ***
## valmejor    4.779e-01  2.552e-02  18.728   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28070 on 2437 degrees of freedom
## Multiple R-squared:  0.6756,Adjusted R-squared:  0.6754
## F-statistic:  2538 on 2 and 2437 DF,  p-value: < 2.2e-16
```

Ejemplo 8.11 (Esperanza de vida por estados). *Son unos datos sobre esperanza de vida en los estados de Estados Unidos.*

```
data(state)
statedata = data.frame(state.x77, row.names = state.abb, check.names = T)
g = lm(Life.Exp ~ ., data = statedata)
summary(g)
```

```
##
## Call:
## lm(formula = Life.Exp ~ ., data = statedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
## Population   5.180e-05  2.919e-05   1.775   0.0832 .
## Income      -2.180e-05  2.444e-04  -0.089   0.9293
## Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
## Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
## Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
## Area        -7.383e-08  1.668e-06  -0.044   0.9649
## ---
## Signif. codes:
```

```
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362,Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

Quitamos la variable Area.

```
g = update(g, . ~ . - Area)
summary(g)

##
## Call:
## lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +
##     HS.Grad + Frost, data = statedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49047 -0.52533 -0.02546  0.57160  1.50374
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
## Population   5.188e-05  2.879e-05   1.802   0.0785 .
## Income      -2.444e-05  2.343e-04  -0.104   0.9174
## Illiteracy   2.846e-02  3.416e-01   0.083   0.9340
## Murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
## HS.Grad      4.847e-02  2.067e-02   2.345   0.0237 *
## Frost       -5.776e-03  2.970e-03  -1.945   0.0584 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7361 on 43 degrees of freedom
## Multiple R-squared:  0.7361,Adjusted R-squared:  0.6993
## F-statistic: 19.99 on 6 and 43 DF,  p-value: 5.362e-11

g = update(g, . ~ . - Illiteracy)
summary(g)

##
## Call:
## lm(formula = Life.Exp ~ Population + Income + Murder + HS.Grad +
##     Frost, data = statedata)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4892 -0.5122 -0.0329  0.5645  1.5166
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
## Population   5.115e-05  2.709e-05   1.888   0.0657 .
## Income      -2.477e-05  2.316e-04  -0.107   0.9153
## Murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
## HS.Grad      4.776e-02  1.859e-02   2.569   0.0137 *
## Frost       -5.910e-03  2.468e-03  -2.395   0.0210 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7277 on 44 degrees of freedom
## Multiple R-squared:  0.7361,Adjusted R-squared:  0.7061
## F-statistic: 24.55 on 5 and 44 DF,  p-value: 1.019e-11
```

```
g = update(g, . ~ . - Income)
summary(g)
```

```
##
## Call:
## lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
##     data = statedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
## Population   5.014e-05  2.512e-05   1.996  0.05201 .
## Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
## Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,Adjusted R-squared:  0.7126
```

```
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12

g = update(g, . ~ . - Population)
summary(g)

##
## Call:
## lm(formula = Life.Exp ~ Murder + HS.Grad + Frost, data = statedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5015 -0.5391  0.1014  0.5921  1.2268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
## Murder      -0.283065   0.036731  -7.706 8.04e-10 ***
## HS.Grad      0.049949   0.015201   3.286  0.00195 **
## Frost       -0.006912   0.002447  -2.824  0.00699 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7427 on 46 degrees of freedom
## Multiple R-squared:  0.7127,Adjusted R-squared:  0.6939
## F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

Ejemplo 8.12 (Consumo de agua). *Se trata de unos datos utilizados para reducir costes de producción. En concreto se pretende valorar el consumo de agua en una fábrica. Se tiene el consumo de agua en distintos meses (en galones) como variable respuesta. Las variables predictoras serán la temperatura media en el mes, la producción (en libras), número de días que ha funcionado la fábrica durante ese mes, número de personas trabajando.*

```
x = read.table(file = "../data/agua.txt", header = T)
attach(x)
```

Ajustamos el modelo.

```
a.lm = lm(agua ~ temperatura + produccion + dias + personas)
summary(a.lm)

##
## Call:
## lm(formula = agua ~ temperatura + produccion + dias + personas)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.99 -131.52    2.58  108.97  368.52
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6360.33733 1314.39161   4.839 0.000406 ***
## temperatura   13.86886    5.15982   2.688 0.019748 *
## produccion     0.21170    0.04554   4.648 0.000562 ***
## dias        -126.69036   48.02234  -2.638 0.021647 *
## personas     -21.81796    7.28452  -2.995 0.011168 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 249 on 12 degrees of freedom
## Multiple R-squared:  0.767,Adjusted R-squared:  0.6894
## F-statistic: 9.877 on 4 and 12 DF,  p-value: 0.0008958
```

8.10 Ejercicios

Ejercicio 8.4. *Vamos a utilizar los datos* homedata *(contenidos en el paquete* UsingR*). Son datos sobre valores asegurados de viviendas en el año 1970 y en el año 2000. Queremos estudiar la relación entre el primer valor asegurado (en el año 1970 que corresponde con la variable y1970) y el último valor asegurado (en el año 2000 que corresponde con la variable y2000). Utilizamos como variable predictora la que nos da el primer valor.*

1. *Uno de los precios asegurados en el año 1970 es 0. Esto es un error. Declarar ese valor como dato faltante con el siguiente código.*
2. *Representar gráficamente el precio asegurado en el 2000 frente al precio asegurado en 1970. ¿Sugiere el dibujo una relación lineal entre las variables?*
3. *Ajustar un modelo de regresión lineal simple donde la variable predictora es y1970 y la variable respuesta es y2000.*
4. *¿Cuáles son los coeficientes del ajuste?*
5. *¿Cuál es el coeficiente de determinación? ¿Podemos considerar que el ajuste es bueno atendiendo al valor del coeficiente de determinación?*
6. *¿Cuál es la predicción del valor asegurado en el año 2000 para una casa que estaba asegurada en 1970 en 75000 euros?*

7. *Considerar como variable predictora el logaritmo natural del valor asegurado en y1970 y como variable respuesta el logaritmo de la cantidad asegurada en el año 2000, y2000. Representa gráficamente los nuevos datos. Realiza el ajuste y compara los coeficientes de determinación. ¿Se ha incrementado o decrementado el coeficiente de determinación? ¿Cuál de los dos ajustes es preferible?*

Ejercicio 8.5. *Leer los datos* babies *el paquete* UsingR.

```
library("UsingR")
data("babies")
attach(babies)
```

El banco de datos contiene información sobre recién nacidos y sus madres en un total de 1236 observaciones.

1. *Leer la ayuda de los datos.*

   ```
   help(babies)
   ```

   En esta ayuda veremos que cada variable tiene un código de dato faltante. Declarar los datos faltantes del siguiente modo.

   ```
   sex[sex == 9] = NA
   wt[wt == 999]  = NA
   parity[parity == 99]  = NA
   race[race == 99]  = NA
   age[age == 99]  = NA
   ed[ed == 9]  = NA
   ht[ht == 99]  = NA
   wt1[wt1 == 999]  = NA
   smoke[smoke == 9]  = NA
   time[time == 99]  = NA
   time[time == 98]  = NA
   number[number == 98 | number == 99] = NA
   ```

2. *Determinar el coeficiente de correlación de Pearson entre las variables* age *(edad) y* wt *(peso). Calculadlo también para las variables* ht *(altura) y* wt *(peso). Haz un diagrama de puntos para cada par de variables y analiza gráficamente el tipo de relación que las liga.*

3. *Supongamos que pretendemos predecir el peso del niño utilizando como variables predictoras las variables* gestation, ht, age, wt1 *que corresponden con el tiempo de gestación, la altura de la madre, la edad de la madre, el peso de la madre antes del nacimiento. Se pide:*

a) *Realizar el correspondiente ajuste.*

b) *Evaluar el coeficiente de determinación.*

c) *Contrastar la hipótesis de que todos los coeficientes excepto la constante son nulos.*

d) *Determinar para cada uno de los predictores si podemos considerar que el correspondiente coeficiente es nulo.*

Ejercicio 8.6. *El banco de datos* teengamb *(paquete* faraway*) contiene datos relativos a hábitos de juego en Gran Bretaña. Ajusta un modelo de regresión lineal donde la cantidad gastada en juego sea la variable respuesta y el sexo, estatus, ingresos y la puntuación en la prueba verbal sean los predictores.*

1. *¿Qué porcentaje de variación de la respuesta es explicada por los predictores?*
2. *¿Qué observación tiene el residuo positivo mayor? Dar el número de caso.*
3. *Determinar la media y la mediana de los residuos.*
4. *Determinar la correlación de los residuos con los valores ajustados.*
5. *Determinar la correlación de los residuos con los ingresos.*
6. *Supongamos que mantenemos constantes todos los demás predictores, ¿qué diferencia tenemos en los valores predichos cuando comparamos hombres con mujeres?*
7. *Contrastar la hipótesis de que todos los coeficientes excepto la constante son nulos.*
8. *Determinar para cada uno de los predictores si podemos considerar que el correspondiente coeficiente es nulo.*

Ejercicio 8.7. *El banco de datos* prostate *procede de un estudio de 97 hombres con cancer de próstata que habían sufrido una prostatectomía radical. Ajusta un modelo con* lpsa *como variable respuesta y* lcavol *como predictores. Registra el error estándar residual y el valor del coeficiente de determinación* R^2*. Añade ahora las variables* lweight, svi, lpph, age, lcp, pgg45, gleason *al modelo y valora el modelo.*

9
Bibliografía

P.M. Berthouex and L.C. Brown. *Environmental Engineers*. Lewis Publishers, second edition, 2002. URL /home/gag/BIBLIOGRAFIA/MISLIBROS/Berthouex_Brown_Statistics_for_Environmental_Engineers_2nd_Ed_CRC_Press_2002.pdf.

Y. Cohen and J.Y. Cohen. *Statistics and Data with R: An applied approach through examples*. John Wiley & Sons, 2008.

P. Dalgaard. *Introductory Statistics with R*. Springer, 2002.

Michael E. Ginevan and Douglas E. Splitstone. *Statistical Tools for Environmental Quality Measurement*. Chapman & Hall / CRC, 2004. URL /home/gag/BIBLIOGRAFIA/MISLIBROS/Ginevan_Splitstone_Statistical_Tools_for_Environmental_Quality_Measurement_Chapman_Hall_CRC_2004.pdf.

B.F.J. Manly. *Statistics for Environmental Science and Management*. Chapman & Hall/CRC Press, second edition, 2009. URL /home/gag/BIBLIOGRAFIA/MISLIBROS/Manly_Statistics_for_Environmental_Science_and_Management_2nd_edition_Chapman_Hall_2009.pdf.

Steven P. Millard and Nagaraj K. Neerchal. *Environmental Statistics with S-PLUS*. Applied Environmental Statistics. CRC Press LLC, 2001. URL /home/gag/BIBLIOGRAFIA/MISLIBROS/Millard_Neerchal_Environmental_Statistics_with_S-PLUS_CRC_2001.pdf.

Walter W. Piegorsch and A. John Bailer. *Analyzing Environmental Data*. Wiley, 2005. URL /home/gag/BIBLIOGRAFIA/MISLIBROS/Piegorsch_Bailer_Analysing_Environmental_Data_Wiley_2005.pdf.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015a. URL http://www.R-project.org/.

R Core Team. *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...*, 2015b. URL http://CRAN.R-project.org/package=foreign. R package version 0.8-63.

Clemens Reimann, Peter Filzmoser, Robert Garrett, and Rudolf Dutter. *Statistical Data Analysis Explained. Applied Environmental Statistics with R*. Wiley, Chichester, UK, 2008. URL /home/gag/BIBLIOGRAFIA/MISLIBROS/Hastie_Tibshirani_Friedman_The_Elements_of_Statistical_Learning_2nd_edition_2008.pdf.

Bernard Rosner. *Fundamentals of Biostatistics*. Brooks/Cole Cengage Learning, seven edition, 2010.

J. Verzani. *Using R for Introductory Statistics*. Chapman & Hall / CRC, 2005.

Rand R. Wilcox. *Basic Statistics*. Oxford University Press, 2009. URL /home/gag/BIBLIOGRAFIA/MISLIBROS/Wilcox_Basic_Statistics_Oxford_University_Press_2009.pdf.