UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Departamento de Matemáticas

ESTADÍSTICA PARA CIENCIAS DEL MAR

Estadística Descriptiva con 

Estadística Descriptiva con R

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Departamento de Matemáticas

Fecha: 15 de septiembre de 2011

Índice general

Capítulo 0

Estadística Descriptiva con R

1. Introducción.

La *estadística descriptiva* es el conjunto de métodos diseñados para organizar, resumir y representar los datos recogidos en el curso de algún estudio. Su finalidad es convertir los datos brutos en información que pueda ser fácilmente entendida y asimilada. En este sentido, la estadística descriptiva es una herramienta indispensable para la *exploración* de los datos: descubrir tendencias, asociaciones, características relevantes, ...

Para poder aplicar los métodos de la estadística descriptiva de manera eficiente se hace necesario disponer de programas informáticos adecuados para ello, con capacidad para capturar datos desde distintas fuentes, procesarlos, transformarlos si es necesario, y generar tablas, gráficos y medidas de síntesis.

 http://www.r-project-org

En este curso proponemos la utilización del paquete estadístico R, que cuenta con numerosas ventajas: es gratuito, se actualiza constantemente, dispone de librerías adicionales para múltiples aplicaciones (genética, climatología, pesquerías, economía, ...), permite la realización de gráficos de alta calidad, incluye un lenguaje de programación que permite al usuario desarrollar funciones a medida y funciona en todas las plataformas (Windows, Linux y Mac).

Pretendemos además que este capítulo sea interactivo y que el alumno vaya aplicando las técnicas y métodos que en él se explican a medida que avanza en su lectura. Con este fin se han dispuesto en la web de la asignatura diversas bases de datos que pueden ser utilizadas libremente para el aprendizaje.

2. Objetivos.

Al finalizar el estudio de este tema, se espera que el alumno sea capaz de:

- Comprender la importancia de la exploración de los datos mediante tablas y gráficos.
- Distinguir los distintos tipos de variables y sus características.
- Calcular e interpretar correctamente la información aportada por las diferentes medidas de síntesis.
- Conocer los métodos de estadística descriptiva para el estudio conjunto de dos variables.
- Utilizar el programa R para la exploración y descripción de datos.

3. Población y Muestra.

Cuando se realiza un estudio de cualquier tipo (de investigación, de mercado, de evaluación de calidad, etc.), generalmente se observan características o magnitudes correspondientes a los elementos de una *población* de interés. Normalmente dicha población no suele ser accesible en su totalidad, y el estudio ha de reducirse a unos cuantos elementos escogidos de la misma. El subconjunto de objetos (o sujetos) de la población que son incluidos en el estudio, recibe el nombre de *muestra*. Así, por ejemplo, en el ámbito de las Ciencias Marinas:

- El estudio de las poblaciones biológicas –cefalópodos, crustáceos, peces, mamíferos marinos, ...– se realiza a partir de los datos aportados por los ejemplares que se capturan o se observan durante una campaña de muestreo.
- El estudio de parámetros físicos o químicos –temperatura, salinidad, velocidad de corriente, concentración de CO_2 disuelto, ...– se realiza a partir de los datos obtenidos por sensores que se colocan en los lugares de interés durante periodos concretos.

El proceso mediante el cual los resultados *particulares* obtenidos en un muestreo se emplean para responder cuestiones *generales* sobre la población recibe el nombre de *inferencia.* Cuando el muestreo es *aleatorio* (todos los elementos de la población tienen, a priori, la misma probabilidad de formar parte de la muestra[1]) el proceso de inferencia se lleva a cabo mediante métodos estadísticos basados en la probabilidad, y recibe el nombre de *Inferencia Estadística.*

[1]Ello garantiza al mismo tiempo que la muestra es *representativa* de la población, es decir, tiene sus mismas características generales. Un muestreo no aleatorio, en el que se seleccionan los objetos con unas características determinadas, puede resultar *tendencioso* y no representar para nada a la población de interés.

4. Tipos de datos.

Las magnitudes o atributos medidos sobre cada objeto de la muestra reciben el nombre de *variables estadísticas* (longitud, peso, duración, temperatura, ...). Los *datos* son los valores que toma la variable en cada objeto. Formalmente, una variable estadística X definida sobre una población Ω y con valores en un conjunto V es una función $X : \Omega \longrightarrow V$, que a cada objeto ω de Ω , le asigna un único valor en V. Cuando este conjunto es numérico ($V \subseteq \mathbb{R}$), la variable se dice *cuantitativa* o *numérica*, y en caso contrario *cualitativa* o *categórica*.

Las variables cuantitativas son *continuas* si pueden tomar cualquier valor dentro de un rango numérico (temperatura, peso, longitud, etc.); son *discretas* si no admiten todos los valores intermedios de un rango. Las variables discretas suelen tomar sólo valores *enteros* (número de hijos de una familia, número de fallos en un equipo técnico durante un año, etc.).

Las variables categóricas son *binarias* si solo toman dos valores (sano/enfermo, observado/no observado, etc.). Pueden ser además *nominales*, si los datos corresponden a categorías sin relación de orden entre sí (color, sexo, profesión,...), u *ordinales* cuando sí que hay relación de orden (curso escolar, posición en una cola,...).

Una vez que se han observado los valores que toman las variables de nuestro estudio es preciso guardar los datos en un archivo que pueda ser leido fácilmente por un programa estadístico, en nuestro caso R. Si la muestra está formada por n objetos $\omega_1, \omega_2, \ldots, \omega_n$, sobre los que se han medido p variables $X_1, X_2, \ldots, X_p$, los datos resultantes deberán organizarse, en general, en forma de una matriz con n filas (cada fila corresponde a un objeto) y p columnas (cada columna corresponde a una variable), tal como se muestra en la tabla 1. Denotamos por x_{ij} al valor observado de la variable X_j sobre el objeto ω_i.

	Variables					
Objetos	X_1	X_2	$\ldots$	X_j	$\ldots$	X_p
ω_1	x_{11}	x_{12}	$\ldots$	x_{1j}	$\ldots$	x_{ip}
ω_2	x_{21}	x_{22}	$\ldots$	x_{2j}	$\ldots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
ω_i	x_{i1}	x_{i2}	$\ldots$	x_{ij}	$\ldots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
ω_n	x_{n1}	x_{n2}	$\ldots$	x_{nj}	$\ldots$	x_{np}

Tabla 1: Organización de los datos para su tratamiento estadístico.

En la mayor parte de los casos la matriz de datos en bruto, aunque contiene toda la información recogida en el muestreo, no permite interpretar la información de forma clara. La percepción y resumen de las características de los datos se consigue fundamentalmente a través de:

1. Tablas de Frecuencias.
2. Representaciones Gráficas.
3. Medidas de Síntesis de datos numéricos.

4.1. Datos de ejemplo: acceso y lectura.

Para ilustrar los distintos métodos de la Estadística Descriptiva utilizaremos los datos que se encuentran en el archivo sargos.csv, que puede descargarse de la web de la asignatura[2]. Este archivo corresponde a un muestreo de sargos realizado sobre capturas de esta especie en las Islas Canarias durante el año 2005. La tabla 2 muestra datos relativos a 10 ejemplares, si bien la base de datos completa contiene 200. Sobre cada ejemplar se han medido las variables: *isla* (donde fue capturado), *sexo, long* (longitud total), *ldors* (longitud medida desde el morro hasta la aleta dorsal), *lpect* (longitud hasta la aleta pectoral), *loper* (longitud hasta el opérculo), *altop* (altura del pez en la región del opérculo), *peso* (peso total), *pgon* (peso de las gónadas), *phig* (peso del hígado), *ptdo* (variable que vale 1 si el pez está parasitado por larvas de anisákidos y 0 si no no está) y *larvas* (número de larvas de anisákidos encontradas en la cavidad abdominal del pez). Como puede apreciarse, el peso de las gónadas no está disponible para todos los peces. A estos valores no disponibles nos referiremos como *valores perdidos*.

isla	sexo	long	ldors	lpect	loper	altop	peso	pgon	phig	ptdo	larvas
GC	Macho	22,59	5,14	5,32	4,08	8	163,81		17,3	0	0
HI	Macho	26,35	6,44	6,02	5,36	8,89	277,04	6,86	22,3	0	0
FV	Macho	21,23	5,11	4,63	4,39	6,39	135,69	1,98	5,4	0	0
TF	Macho	22,7	5,35	4,61	4,95	7,33	167,54	1,65	27	1	5
LZ	Hembra	20,2	4,84	4,58	4,38	6,63	131,68		7,1	0	0
TF	Macho	21,6	5,5	5,56	3,83	6,08	176,21	4,54	22,9	0	0
GC	Hembra	25,18	5,73	5,52	5,72	8,14	257,38	37,01	12,4	0	0
GC	Macho	21,68	5,02	5,19	4,74	6,62	145,14		18,2	0	0
LP	Macho	23,29	6,03	5,4	5,34	6,95	201,82	3,55	12,7	0	0
TF	Hembra	16,39	4,31	3,54	3,57	5,21	78,54		6,4	0	0

Tabla 2: Datos recogidos en un muestreo de ejemplares de Sargo (*Diplodus Sargus*) en las Islas Canarias. Se muestran solo 10 ejemplares.

El archivo está en formato csv (*Comma Separated Values*), que es un archivo ASCII plano (es decir, sin información de formato de ningún tipo), en el que los distintos valores están separados por el símbolo *punto y coma* (;). Puede abrirse con cualquier editor de texto, si

[2]Este archivo puede descargarse también desde http://dl.dropbox.com/u/7610774/sargos.csv.

bien las hojas de cálculo estándar (OpenOffice o Microsoft Excel) nos lo muestran en forma de tabla visualmente más atractiva. En la primera fila del archivo se encuentran los nombres de las variables.

Supondremos que una vez descargado el archivo lo hemos guardado en el directorio[3]:

c:\documents and settings\fcmar\data\

Para leer este archivo con R utilizaremos los siguientes comandos:

```
> setwd("c:/documents and settings/fcmar/data/")
> sargos = read.table(file = "sargos.csv", sep = ";", dec = ",",
     header = TRUE)
```

El primer comando *setwd()* (acrónimo de *set working directory*) se encarga de indicar a R el directorio de trabajo, en el que se encuentran los datos (y en el que previsiblemente guardaremos los resultados).

Importante: Las barras empleadas para especificar el directorio deben ser de la forma "/" y no la habitual "\" en Windows.

La segunda linea es la que lee el fichero sargos.csv y asigna su contenido al objeto *sargos*. Indicamos además que los datos están separados por punto y coma (*sep=";"*), que el símbolo decimal que se usa en los valores numéricos es la coma (*dec=","*), y que el archivo tiene una cabecera con los nombres de las variables (*header=TRUE*).

Nota: si disponemos de un ordenador con conexión directa a internet, el fichero sargos.csv puede ser importado directamente desde la red con R mediante:

```
> sargos = read.table(file = "http://dl.dropbox.com/u/7610774/sargos.csv",
     sep = ";", dec = ",", header = TRUE)
```

El objeto en que R almacena la matriz de datos con la que vamos a trabajar –en el ejemplo, la tabla leída del archivo sargos.csv se ha almacenado en el objeto *sargos*– recibe el nombre de *data.frame*. En esencia, un *data.frame* es una matriz de datos cuyas columnas representan variables identificadas por su nombre.

[3]Suponemos que se utiliza un ordenador con sistema operativo Windows, que es la situación más habitual. En caso de utilizar Linux o Mac las rutas de directorio pueden ser ligeramente distintas. En lo que se refiere al funcionamiento de R, es idéntico en todos los sistemas operativos.

4.2. Acceso directo a las variables dentro de una matriz de datos.

En general, cuando deseamos acceder a una variable que está dentro de un *data.frame* deberemos anteponer al nombre de la variable el nombre del objeto que la contiene, separados por el símbolo $. Por ejemplo, para ver el contenido de la variable long deberíamos escribir sargos$long. Si hemos de trabajar con muchas variables, tener que escribir siempre el nombre de la matriz de datos puede llegar a hacerse muy tedioso. Podemos habilitar un *"acceso directo"* a las variables por su nombre utilizando la función:

```
> attach(sargos)
```

A partir de ahora todas las variables estarán disponibles directamente por su nombre. Para cancelar este acceso directo, deberemos ejecutar *detach(sargos)*.

4.3. Tipos de datos en R.

Hemos visto al comienzo de esta sección que las variables estadísticas pueden clasificarse en categóricas y numéricas, y estas últimas en discretas o continuas. R distingue las variables según su *clase*:

- *numeric*: variables numéricas continuas.
- *integer*: variables numéricas discretas.
- *character*: variables alfanuméricas; sus valores son combinaciones de cifras y letras.
- *factor*: variables categóricas; R almacena internamente los valores de un factor como números enteros, pero los muestra como valores alfanuméricos.

La función *str()* (acrónimo de *estructura*) muestra la estructura del objeto especificado. Así, si aplicamos esta función a nuestros datos de ejemplo obtenemos:

```
> str(sargos)
```

```
'data.frame':        200 obs. of  12 variables:
 $ isla  : Factor w/ 7 levels "FV","GC","HI",..: 2 3 1 7 6 7 2 2 5 7 ...
 $ sexo  : Factor w/ 2 levels "Hembra","Macho": 2 2 2 2 1 2 1 2 2 1 ...
 $ long  : num  22.6 26.4 21.2 22.7 20.2 ...
 $ ldors : num  5.49 5.49 5.36 4.5 5.36 5 5.66 4.78 4.83 3.79 ...
```

```
 $ lpect : num  5.32 6.02 4.63 4.61 4.58 5.56 5.52 5.19 5.4 3.54 ...
 $ loper : num  4.08 5.36 4.39 4.95 4.38 3.83 5.72 4.74 5.34 3.57 ...
 $ altop : num  8 8.89 6.39 7.33 6.63 6.08 8.14 6.62 6.95 5.21 ...
 $ peso  : num  164 277 136 168 132 ...
 $ pgon  : num  NA 6.86 1.98 1.65 NA ...
 $ phig  : num  17.3 22.3 5.4 27 7.1 22.9 12.4 18.2 12.7 6.4 ...
 $ ptdo  : int  0 0 0 1 0 0 0 0 0 0 ...
 $ larvas: int  0 0 0 5 0 0 0 0 0 0 ...
```

Podemos ver que las variables isla y sexo han sido identificadas como factores (*factor*); las variables long, ldors, lpect, loper, altop, peso, pgon y phig han sido identificadas como *numeric* (valores reales, variables numéricas continuas); y las variables ptdo y larvas han sido identificadas como *integer* (valores enteros, variables numéricas discretas).

La variable isla es un factor; ello significa que si pedimos a R que nos muestre sus valores, nos los mostrará como alfanuméricos:

```
> isla
```

```
  [1] GC HI FV TF LZ TF GC GC LP TF GC GC LP LP GC HI GC FV FV FV GC
 [22] GC TF GC HI LZ GC GC LZ HI LG TF GC HI LZ HI LP LZ TF GC TF LP
 [43] LZ TF LP TF LG LZ FV TF TF GC GC LP TF FV LZ LZ TF TF LG FV GC
 [64] GC HI LZ LZ FV GC GC LG TF GC LZ LZ LP TF LP LZ LZ GC FV TF GC
 [85] LG FV FV GC TF FV TF GC LG LZ LZ TF HI TF LZ FV HI FV FV TF TF
[106] GC GC FV LP LZ FV LP GC HI LP LZ HI FV LZ TF TF FV LZ HI GC FV
[127] GC FV LG GC LZ GC FV LG FV GC FV LP FV FV LG TF HI TF TF GC LP
[148] LZ GC LP GC GC LZ LZ FV TF GC GC FV TF GC LP FV LP TF LP LZ TF
[169] LP LP TF TF GC GC LP GC LP GC TF TF LP TF LP LZ GC HI LZ FV HI
[190] TF FV FV GC GC GC LZ LZ LZ TF TF
Levels: FV GC HI LG LP LZ TF
```

Pero si ejecutamos la función *unclass()* vemos que internamente los valores de esta variable están almacenados como números enteros:

```
> unclass(isla)
```

```
  [1] 2 3 1 7 6 7 2 2 5 7 2 2 5 5 2 3 2 1 1 1 2 2 7 2 3 6 2 2 6 3 4 7
 [33] 2 3 6 3 5 6 7 2 7 5 6 7 5 7 4 6 1 7 7 2 2 5 7 1 6 6 7 7 4 1 2 2
 [65] 3 6 6 1 2 2 4 7 2 6 6 5 7 5 6 6 2 1 7 2 4 1 1 2 7 1 7 2 4 6 6 7
```

```
 [97] 3 7 6 1 3 1 1 7 7 2 2 1 5 6 1 5 2 3 5 6 3 1 6 7 7 1 6 3 2 1 2 1
[129] 4 2 6 2 1 4 1 2 1 5 1 1 4 7 3 7 7 2 5 6 2 5 2 2 6 6 1 7 2 2 1 7
[161] 2 5 1 5 7 5 6 7 5 5 7 7 2 2 5 2 5 2 7 7 5 7 5 6 2 3 6 1 3 7 1 1
[193] 2 2 2 6 6 6 7 7
attr(,"levels")
[1] "FV" "GC" "HI" "LG" "LP" "LZ" "TF"
```

4.4. Recodificación y etiquetado de niveles de los factores.

En muchas ocasiones, los niveles de un factor son poco ilustrativos de su significado. En los datos de nuestro ejemplo, la variable que indica si un pez está parasitado o no, ptdo, toma los valores 0 y 1, y éstos son los valores que aparecerán en las tablas y gráficos que podamos hacer con esta variable. Sería deseable que en su lugar apareciesen los términos "*No Parasitado*" y *"Parasitado",* ya que de esta forma la salida de resultados sería más clara e interpretable. Podemos conseguir este efecto creando un nuevo factor a partir de esta variable, y asignando etiquetas a sus valores mediante la siguiente sintaxis:

```
> fptdo = factor(ptdo, levels = c(0, 1), labels = c("No Parasitado",
     "Parasitado"))
```

Con ello hemos creado una nueva variable fptdo de clase *factor*; esta variable se construye a partir de ptdo, asignando a sus niveles originales, *levels=c(0,1)*, unas nuevas *etiquetas,* *labels=c("No Parasitado","Parasitado")* (las etiquetas deben asignarse en el mismo orden que en *levels()*). De esta manera, a partir de ahora, en todos los resultados que involucren a la variable fptdo (gráficos, tablas, etc.) sus valores aparecerán identificados como "*No Parasitado*" y *"Parasitado"*.

Nota: al crear una variable de clase *factor*, R almacena internamente sus valores como enteros consecutivos (1, 2, ...), si bien en todas las salidas se mostrarán exclusivamente las etiquetas que hayamos puesto. Puede observarse la codificación interna que se ha hecho de la variable fptdo mediante *unclass(fptdo)*.

Importante: si la variable que convertimos en factor tiene otros valores distintos que no han sido especificados en *levels*, tales valores se pierden: se convierten en *No Asignados* (*NA*), y no serán utilizados en los análisis que posteriormente podamos hacer de los datos.

¿Crear variables o recodificar variables existentes?

Acabamos de ver como se crea un factor (fptdo) a partir de una variable existente (ptdo). Si hubiésemos utilizado la sintaxis:

```
> ptdo = factor(ptdo, levels = c(0, 1), labels = c("No Parasitado",
     "Parasitado"))
```

en lugar de *crear* un nuevo factor, habríamos *recodificado* la variable ptdo ya existente, que de esta forma quedaría convertida directamente en factor (y habría perdido sus valores originales, en este caso 0 y 1)[4]. Podemos comprobarlo, por ejemplo, utilizando el comando ***unique()***, que muestra los valores *distintos* que toma la variable:

```
> unique(ptdo)
```

```
[1] No Parasitado Parasitado
Levels: No Parasitado Parasitado
```

¿Es mejor crear nuevas variables o recodificar las que ya existen? Si somos principiantes en R lo mejor es crear nuevas variables; de esta forma las variables originales estarán siempre disponibles y en caso de error podemos volver a utilizarlas. Si las recodificamos y nos hemos equivocado en la recodificación, tendríamos que recuperar la variable original, lo que a veces puede resultar complicado.

En este caso particular la recuperación resulta sencilla, ya que los valores originales de ptdo siguen almacenados en el data.frame sargos (vinculado al *entorno de trabajo* actual mediante el comando ***attach***). Si borramos la variable ptdo mediante:

```
> rm(ptdo)
```

en realidad sólo borramos la variable recodificada; la variable ptdo del data.frame original, que permanecía en el entorno de trabajo vuelve a ser accesible:

```
> unique(ptdo)
```

```
[1] 0 1
```

[4] En sentido estricto, la variable ptdo que pertenece al data.frame sargos, no se elimina de éste, sino que queda *oculta* por la nueva definición que se ha dado de dicha variable.

5. Tablas de frecuencias y representaciones gráficas.

5.1. Variables categóricas o numéricas discretas.

Cuando se observan variables categóricas tales como la isla en que fue capturado un pez, su sexo, y si está o no parasitado, muchos de sus valores aparecen repetidos. La *frecuencia absoluta* de la i-ésima categoría es el número de veces n_i que se repite dicha categoría en el total de observaciones. La *frecuencia relativa* es la proporción:

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

siendo $n = \sum_{i=1}^{k} n_i$ el número total de observaciones (k es el número de categorías). La frecuencia relativa suele también expresarse en porcentaje:

$$f_i = 100 \cdot \frac{n_i}{n} \,\%$$

Estas definiciones se extienden también a la construcción de tablas de frecuencias para variables numéricas discretas. En este último caso se suele considerar también la *frecuencia acumulada* hasta el valor x_i como el número $N_i = \sum_{j=1}^{i} n_j$ de observaciones menores o iguales que x_i. La *frecuencia acumulada relativa* es la proporción:

$$F_i = \frac{N_i}{n}$$

Estas frecuencias suelen presentarse como se muestra en la tabla 3. En la columna de la variable X se anotan sólo las k categorías o valores *distintos* que toma la variable, en orden creciente si X es numérica. Asimismo las frecuencias acumuladas sólo se incluyen cuando X es numérica.

X	Frecuencia Absoluta	Frecuencia Relativa	Frec. Acum. Absoluta	Frec. Acum. Relativa
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k

Tabla 3: Tabla de frecuencias para variables categóricas o numéricas discretas.

Tablas de frecuencias para variables categóricas o discretas en R.

Los siguientes comandos nos muestran las tablas de frecuencias absolutas y relativas para la isla en que se han capturado los peces de nuestro ejemplo:

```
> table(isla)

isla
FV GC HI LG LP LZ TF
32 48 15  9 24 32 40

> prop.table(table(isla))

isla
   FV    GC    HI    LG    LP    LZ    TF
0.160 0.240 0.075 0.045 0.120 0.160 0.200
```

De igual modo, para el número de larvas:

```
> table(larvas)

larvas
  0   3   4   5   6   7   8   9
170   4   2   4   2   3   9   6

> prop.table(table(larvas))

larvas
    0     3     4     5     6     7     8     9
0.850 0.020 0.010 0.020 0.010 0.015 0.045 0.030
```

Para las frecuencias acumuladas utilizamos la función *cumsum()*:

```
> cumsum(table(larvas))

  0   3   4   5   6   7   8   9
170 174 176 180 182 185 194 200

> cumsum(prop.table(table(larvas)))
```

```
    0     3     4     5     6     7     8     9
0.850 0.870 0.880 0.900 0.910 0.925 0.970 1.000
```

Podemos construir una tabla más compacta para estas frecuencias del siguiente modo:

```
> tbl = table(larvas)
> nlarvas = names(tbl)
> fi = as.vector(tbl)
> fri = as.vector(prop.table(tbl))
> Fi = cumsum(fi)
> Fri = cumsum(fri)
> data.frame(nlarvas, fi, fri, Fi, Fri)
```

```
  nlarvas  fi   fri  Fi   Fri
1       0 170 0.850 170 0.850
2       3   4 0.020 174 0.870
3       4   2 0.010 176 0.880
4       5   4 0.020 180 0.900
5       6   2 0.010 182 0.910
6       7   3 0.015 185 0.925
7       8   9 0.045 194 0.970
8       9   6 0.030 200 1.000
```

Aquí hemos utilizado los siguientes comandos de R:

- *tbl=table(larvas)*: asignamos el contenido de la tabla de frecuencias al objeto *tbl*.
- *nlarvas=names(tbl)*: asigna a *nlarvas* los nombres (categorías) de la tabla anterior; en este ejemplo, las categorías son los distintos números de larvas encontrados. Utilizaremos estos nombres como primera columna de nuestra tabla compacta.
- *f=as.vector(tbl)*: la función *table(larvas)* como hemos visto antes, crea una tabla de frecuencias absolutas. En R una tabla es un objeto con una estructura muy particular, que contiene los nombres de las distintas categorías de la variable que se tabula y sus frecuencias. Al aplicar la función *as.vector()* a esta estructura, la convierte en un simple vector, sin nombres, sólo con los valores de las frecuencias, que se van a usar como segunda columna en la tabla.
- *data.frame()*: crea la matriz de datos que contiene la tabla de frecuencias que se presenta por pantalla.

Sugerencia: Si necesitáramos hacer frecuentemente tablas como ésta, resulta conveniente definir una función en R para ello, que nos ahorre tener que escribir todas estas líneas cada vez. Esta función podría ser, por ejemplo:

```
> tablaFrec = function(x) {
      tbl = table(x)
      categ = names(tbl)
      fi = as.vector(tbl)
      fri = as.vector(prop.table(tbl))
      Fi = cumsum(fi)
      Fri = cumsum(fri)
      tabla = data.frame(categ, fi, fri, Fi, Fri)
      names(tabla)[1] = deparse(substitute(x))
      return(tabla)
 }
```

Observemos que la función usa prácticamente los mismos comandos que acabamos de ver. Se ha añadido una linea al final para mejorar la presentación:

- *names(tabla)[1]=deparse(substitute(x))*: Nuestra función recibirá en general como argumento una variable arbitraria *x*. La función *deparse(substitute(x))* extrae su nombre, y *names(tabla)[1]=* lo asigna como cabecera de la primera columna de nuestra tabla.

Para aplicar la función que acabamos de definir a la variable *larvas* bastaría con introducir:

```
> tablaFrec(larvas)
```

A medida que vamos trabajando con R podemos ir construyendo nuestra colección de funciones útiles y guardarlas, por ejemplo, en el archivo *MisFunciones.R*. Para tenerlas disponibles cada vez que usemos R bastará con ejecutar al principio de nuestra sesión:

```
> source("MisFunciones.R")
```

Gráficos: diagramas de barras y diagramas de sectores.

Las tablas de frecuencias que hemos visto en esta sección se representan gráficamente mediante:

- *Diagramas de barras,* que en R se obtienen con el comando `barplot()`.
- *Diagramas de sectores,* que en R se obtienen con el comando `pie()`.

En la figura 1 se muestran ambos diagramas para el número de capturas de sargos por isla en la muestra que estamos utilizando como ejemplo. Para generar estos gráficos se ha utilizado la sintaxis:

```
> barplot(table(isla))
> pie(table(isla))
```

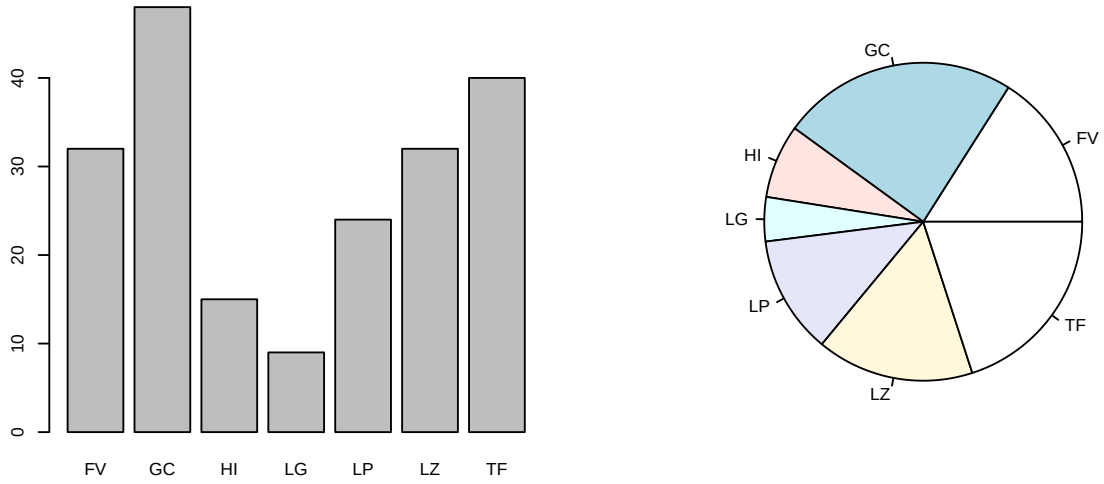

Figura 1: Izquierda: gráfico de barras del número de ejemplares capturados por isla. Derecha: gráfico de sectores con los mismos datos.

Como puede apreciarse en esta figura, en el diagrama de barras la altura de cada barra es igual a la frecuencia absoluta representada. Asimismo, en el diagrama de sectores, el ángulo del sector correspondiente a cada categoría es proporcional a su frecuencia. En el caso del diagrama de barras, si queremos que la altura de las barras represente frecuencias relativas, bastará emplear `prop.table()` del modo siguiente:

```
> barplot(prop.table(table(isla)))
```

Mejorando la presentación de los gráficos.

Los dos gráficos anteriores, si bien representan correctamente las frecuencias observadas, resultan poco informativos: carecen de título; las etiquetas de las barras o sectores (FV, GC, HI, etc) resultan poco claras (el lector del informe estadístico puede no saber qué significan estas siglas); estas etiquetas figuran en orden alfabético y quizás tuviese más sentido colocarlas en orden geográfico, con las islas de este a oeste; incluso el gráfico en tonos grises puede resultar visualmente poco atractivo.

Con R es sencillo mejorar el aspecto de los gráficos. La siguiente sintaxis produce el diagrama de barras mostrado en la figura 2, que mejora bastante al de la figura 1:

```
> isla = factor(isla, levels = c("HI", "LP", "LG", "TF",
     "GC", "FV", "LZ"), ordered = TRUE)
> par(cex.axis = 0.9, las = 1)
> barplot(prop.table(table(isla)), main = "Ejemplares capturados por isla",
     names.arg = c("Hierro", "La\nPalma", "La \nGomera",
         "Tenerife", "Gran \nCanaria", "Fuerte-\nventura",
         "Lanza-\nrote"), col = terrain.colors(12))
```

En la primera linea hemos redefinido el factor isla, simplemente colocando la lista de niveles de este factor en el orden Oeste-Este, e indicando a R que debe mantener esta ordenación (*ordered=TRUE*) en todas las representaciones que afecten a esta variable.

En la segunda linea hemos modificado algunos de los parámetros gráficos que usa R por defecto. En particular, *cex.axis=0.9* disminuye el tamaño de la letra que se usa para etiquetar las barras a un 90 % de su tamaño original (con objeto de que se puedan poner los nombres completos de las islas). A su vez *las=1* produce que las etiquetas en ambos ejes se escriban horizontalmente.

Por último, en la tercera linea se genera el diagrama de barras. Con la opción *main* se indica el título del gráfico. En *names.arg* se especifican los nombres que se van a utilizar como etiquetas de las barras. Si no se incluye esta opción, se usan las etiquetas del factor que se va a tabular. En este caso, hemos incluído los nombres de las islas para poder separar en dos lineas los nombres largos: para ello, hay que indicar con "\n" el lugar de la separación. La última opción, *col*, permite indicar los colores a utilizar. En este caso hemos utilizado la paleta *terrain.colors(n)* que genera n colores dentro de una misma gama[5]. Los colores para un gráfico pueden designarse también por su nombre (en inglés). Así, en este caso

[5]Si el número m de colores a representar es menor que n se utilizan los m primeros de esa gama. Y si el número es mayor, los colores se repiten hasta completar el gráfico.

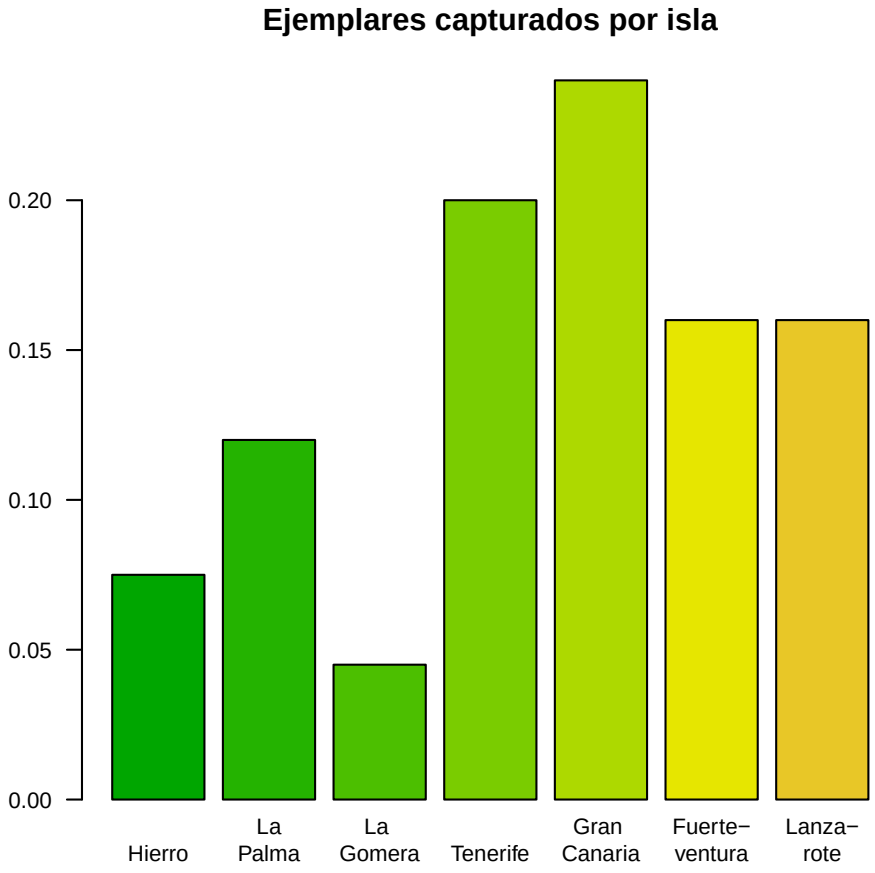

Figura 2: Diagrama de barras mejorado.

podíamos haber especificados los colores de cada barra, por ejemplo, mediante `col=c("red","blue","green","yellow","cyan","orange","magenta")`. Podemos obtener la lista de colores que maneja R mediante el comando `colours()`.

El gráfico de sectores de la figura 1 también puede mejorarse si se incluye el nombre completo de las islas y se indica además el porcentaje de capturas que corresponde a cada una. Requiere algo más de trabajo, pero el código es también muy simple:

```
> noms = c("Hierro", "La Palma", "La Gomera", "Tenerife",
    "Gran Canaria", "Fuerteventura", "Lanzarote")
> pct = prop.table(table(isla)) * 100
> etiquetas = paste(noms, " (", pct, "%)", sep = "")
> pie(table(isla), col = terrain.colors(7), labels = etiquetas,
    main = "Captura por isla")
```

En la primera línea hemos creado el vector `noms` que contiene los nombres de las islas.

En la segunda linea obtenemos la tabla de frecuencias relativas y la multiplicamos por 100; de esta forma sus valores, en lugar de estar expresados en tanto por uno, quedan expresados en tanto por ciento. La tabla se almacena en el objeto `pct`.

En la tercera linea se construyen las etiquetas que se van a añadir al diagrama de sectores; cada etiqueta será el nombre de la isla seguido del porcentaje de capturas obtenido en la misma entre paréntesis. Ello se consigue *"pegando"* mediante la función `paste()` los vectores `noms` y `pct`. La misma función `paste()` nos permite, como vemos, insertar los símbolos de paréntesis y de porcentaje.

Por último, en la cuarta linea, generamos el diagrama de sectores; utilizamos de nuevo la paleta de colores `terrain.colors()`, fijamos como etiquetas (`labels`) las que acabamos de generar, y añadimos un título al gráfico usando `main`. El resultado se muestra en la figura 3.

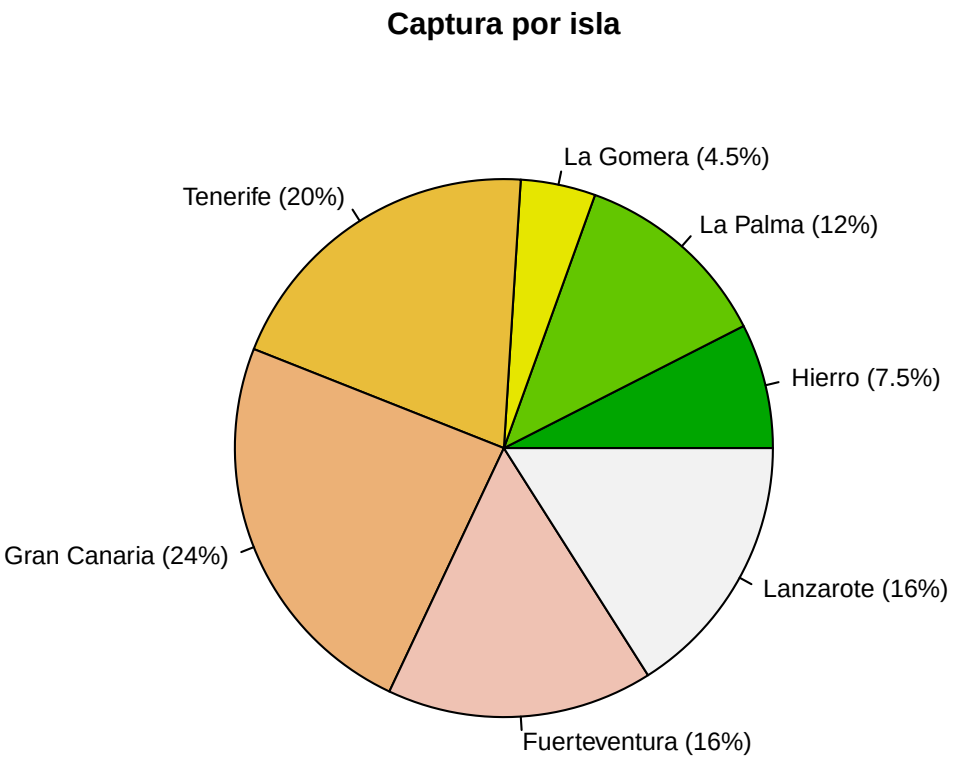

Figura 3: Diagrama de sectores mejorado.

Tablas cruzadas para variables categóricas o numéricas discretas.

Cuando se estudian conjuntamente dos variables categóricas o numéricas discretas, resulta de interés determinar qué valores aparecen juntos con más o menos frecuencia. Con este fin se construyen las *tablas de frecuencias cruzadas*. Si la variable X toma los valores $x_1, x_2, \ldots, x_k$ y la variable Y toma los valores $y_1, y_2, \ldots, y_m$, se denomina *frecuencia absoluta del par* (x_i, y_j) al número de veces n_{ij} que dicha pareja de valores aparecen juntos en la muestra. Las frecuencias absolutas se suelen presentar en una *tabla cruzada* como se muestra en la tabla 4.

	y_1	y_2	$\ldots$	y_m	*Totales*
x_1	n_{11}	n_{12}	$\ldots$	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2m}	$n_{2\bullet}$
$\vdots$					
x_k	n_{k1}	n_{k2}	$\ldots$	n_{km}	$n_{k\bullet}$
Totales	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet m}$	$n_{\bullet\bullet}$

Tabla 4: Tabla de frecuencias cruzadas.

El valor $n_{i\bullet}$ representa el total de la fila i, $\left(n_{i\bullet} = \sum_{j=1}^{m} n_{ij}\right)$, y por tanto es la frecuencia absoluta con que se observa el valor x_i. Asimismo, el valor $n_{\bullet j}$ representa el total de la fila j, $\left(n_{\bullet j} = \sum_{i=1}^{k} n_{ij}\right)$, y por tanto es la frecuencia absoluta con que se observa el valor y_j. Por último $n_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}$ representa el total de valores observados y coincide con el tamaño de la muestra. Las frecuencias $n_{i\bullet}$ y $n_{\bullet j}$ reciben el nombre de *frecuencias marginales* de X e Y, respectivamente.

A partir de una tabla de frecuencias cruzadas absolutas es posible construir tres clases de tablas de frecuencias relativas:

- *Frecuencias relativas globales:* se calculan dividiendo cada frecuencia cruzada por el total de la tabla:
$$f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}$$
- *Frecuencias relativas por filas:* se calculan dividiendo cada frecuencia cruzada por el total de su fila:
$$ff_{ij} = \frac{n_{ij}}{n_{i\bullet}}$$
Representan la frecuencia relativa con que se produce cada valor de Y cuando se fija el valor $X = x_i$. Por esta razón, suelen denominarse *frecuencias relativas de Y condicionadas por $X = x_i$*.
- *Frecuencias relativas por columnas:* se calculan dividiendo cada frecuencia cruzada por

el total de su columna:

$$fc_{ij} = \frac{n_{ij}}{n_{\bullet j}}$$

Representan la frecuencia relativa con que se produce cada valor de X cuando se fija el valor $Y = y_j$. Por esta razón, suelen denominarse *frecuencias relativas de X condicionadas por $Y = y_j$*.

Tablas cruzadas en R.

Las tablas cruzadas en R se generan también mediante la función *table()*, especificando ahora como argumento qué variables se desean cruzar. Así, en nuestros datos de ejemplo, si queremos evaluar el número de peces parasitados por *anisakis* capturados en cada una de las islas durante nuestra campaña de muestreo ejecutaríamos simplemente:

```
> table(fptdo, isla)

              isla
fptdo          HI LP LG TF GC FV LZ
  No Parasitado 14 19  8 31 44 28 26
  Parasitado     1  5  1  9  4  4  6
```

Podemos añadir los totales por filas y columnas mediante *addmargins*:

```
> addmargins(table(fptdo, isla))

              isla
fptdo           HI  LP  LG  TF  GC  FV  LZ Sum
  No Parasitado 14  19   8  31  44  28  26 170
  Parasitado     1   5   1   9   4   4   6  30
  Sum           15  24   9  40  48  32  32 200
```

Las distintas tablas cruzadas de frecuencias relativas se obtienen utilizando *prop.table()*:

- Frecuencias relativas globales:

```
> prop.table(table(fptdo, isla))
```

```
                isla
fptdo              HI    LP    LG    TF    GC    FV    LZ
  No Parasitado 0.070 0.095 0.040 0.155 0.220 0.140 0.130
  Parasitado    0.005 0.025 0.005 0.045 0.020 0.020 0.030
```

- Frecuencias relativas por filas: basta añadir a la función *prop.table()* el argumento *margin=1*. Aquí además redondeamos a tres decimales:

```
> round(prop.table(table(fptdo, isla), margin = 1), 3)

                isla
fptdo              HI    LP    LG    TF    GC    FV    LZ
  No Parasitado 0.082 0.112 0.047 0.182 0.259 0.165 0.153
  Parasitado    0.033 0.167 0.033 0.300 0.133 0.133 0.200
```

- Frecuencias relativas por columnas: Igual que en el caso anterior, pero utilizando el argumento *margin=2*:

```
> round(prop.table(table(fptdo, isla), margin = 2), 3)

                isla
fptdo              HI    LP    LG    TF    GC    FV    LZ
  No Parasitado 0.933 0.792 0.889 0.775 0.917 0.875 0.812
  Parasitado    0.067 0.208 0.111 0.225 0.083 0.125 0.188
```

Nota: Se puede omitir la palabra *margin* en los comandos anteriores. El resultado habría sido idéntico utilizando *prop.table(table(fptdo,isla),1)*.

Presentación gráfica de tablas cruzadas.

Las tablas de frecuencias cruzadas pueden representarse gráficamente también mediante *barplot()*. En la figura 4 se muestran dos diagramas de barras en los que se representa la distribución de sexos por isla. El gráfico (a) ha sido generado con la siguiente sintaxis:

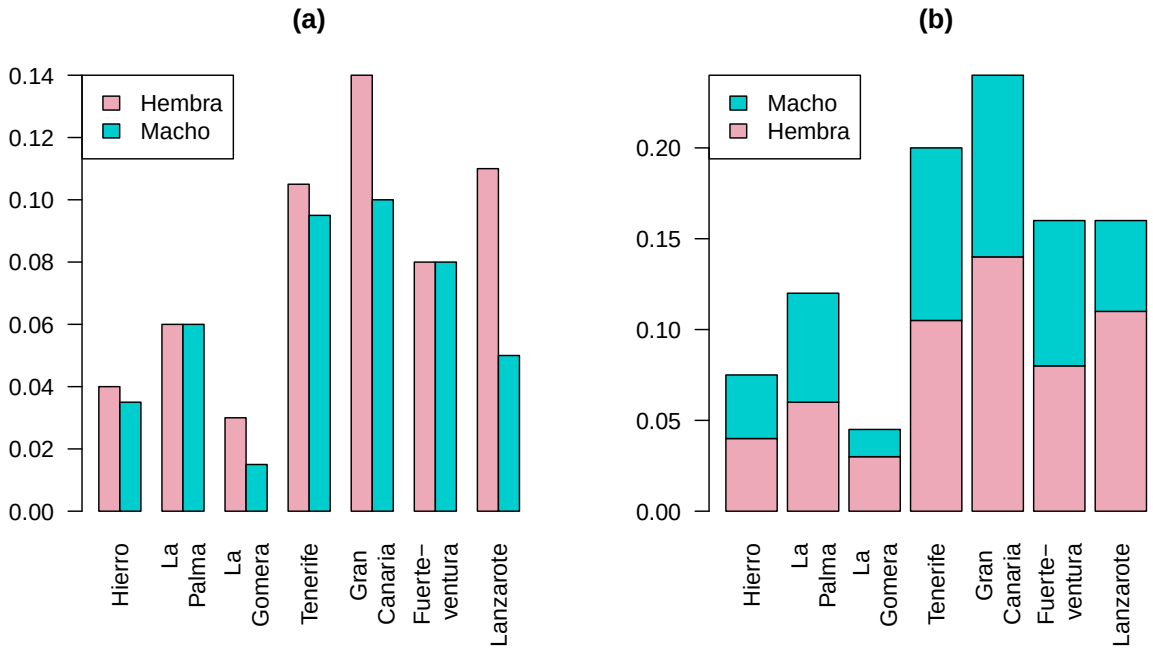

Figura 4: Representación gráfica de la distribución de sexos por isla. (a) Barras emparejadas (*beside=T*). (b) Barras apiladas. (*beside=F*)

```
> barplot(prop.table(table(sexo, isla)), col = c("pink2",
    "cyan3"), beside = TRUE, legend.text = TRUE, names.arg = c("Hierro",
    "La\nPalma", "La \nGomera", "Tenerife", "Gran \nCanaria",
    "Fuerteven-\ntura", "Lanza-\nrote"), las = 2)
```

El gráfico (b) ha sido generado con una sintaxis idéntica salvo que se ha especificado *beside=FALSE* para que las barras se presenten apiladas en lugar de una junto a otra. En este código se han especificado dos colores, uno para hembras y otro para machos. El orden en que se especifican los colores se corresponde con el orden alfabético de las etiquetas de la variable sexo. Por último, la opción *legend.text=TRUE* hace que se dibuje un recuadro en que se especifica qué color corresponde a cada categoría de la variable sexo.

5.2. Variables numéricas continuas.

Si la variable numérica es continua, no cabe esperar repeticiones de un mismo valor de la variable. En este caso, conviene sintetizar el conjunto de valores mediante agrupaciones de la variable en intervalos de clase $(x_{i-1}, x_i]$. En general, los intervalos deben ser de la misma longitud. Denominaremos "marca de clase" al punto medio del intervalo de clase, $m_i = \frac{x_{i-1}+x_i}{2}$. Para determinar el número de intervalos a construir suele emplearse la regla empírica de

Sturges que consiste tomar como número de intervalos un valor próximo a $k \approx 1+3{,}22\log(n)$, siendo n el número total de valores observados. Esta regla es la que emplea R por defecto en la construcción de tablas y gráficos de frecuencias para variables continuas.

Tablas de Frecuencias para variables continuas.

Una vez agrupados los datos en intervalos de clase, el cálculo de las frecuencias es análogo al caso anterior, con la única diferencia de que ahora n_i es el número de observaciones dentro del intervalo $(x_{i-1}, x_i]$, tal como se muestra en la tabla 5.

X (Intervalo)	Marca de clase	Frecuencia Absoluta	Frecuencia Relativa	Frec. Acum. Absoluta	Frec. Acum. Relativa
$[x_0, x_1]$	m_1	n_1	f_1	N_1	F_1
$(x_1, x_2]$	m_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(x_{k-1}, x_k]$	m_k	n_k	f_k	N_k	F_k

Tabla 5: Tabla de frecuencias para variables continuas.

Tablas de frecuencias para variables continuas en R

La configuración básica de R no dispone de ninguna función específica para la construcción de tablas de frecuencias para variables continuas. Sin embargo, si instalamos[6] el paquete *agricolae* tendremos a nuestra disposición la función *table.freq()*, muy similar a la que hemos construido más arriba para variables discretas. Veamos como utilizar esta función para construir una tabla de frecuencias de las longitudes de los sargos de nuestro ejemplo:

```
> library(agricolae)
> table.freq(hist(long, plot = F))

  Inf Sup MC fi   fri  Fi   Fri
    8  10  9  1 0.005   1 0.005
   10  12 11  1 0.005   2 0.010
   12  14 13  4 0.020   6 0.030
```

[6] Antes de usar una nueva librería –paquete de programas– en R por primera vez, será preciso descargarla e instalarla desde internet. Para ello, arrancamos R, y en el menú superior elegimos la opción *Paquetes →Instalar Paquete(s);* se abre una ventana en la que indicamos el país desde el que deseamos descargar el paquete. Elegimos un pais y a continuación se despliega la lista de paquetes disponibles, en la que seleccionamos el que nos interesa instalar.

```
14  16 15 10 0.050  16 0.080
16  18 17 28 0.140  44 0.220
18  20 19 33 0.165  77 0.385
20  22 21 39 0.195 116 0.580
22  24 23 34 0.170 150 0.750
24  26 25 24 0.120 174 0.870
26  28 27 16 0.080 190 0.950
28  30 29  8 0.040 198 0.990
30  32 31  2 0.010 200 1.000
```

Representación gráfica de las tablas de frecuencias para variables continuas.

Histogramas.

La distribución de frecuencias de variables continuas se representa habitualmente en un *histograma.* Este gráfico se construye levantando sobre cada intervalo un rectángulo de área proporcional a la frecuencia que se pretende representar. En R podemos obtener el histograma de las longitudes de los sargos de nuestra muestra mediante:

```
> hist(long, xlab = "longitud", ylab = "Frecuencia", freq = FALSE,
     main = "Longitudes observadas en la muestra", col = topo.colors(40))
```

En esta sintaxis hemos utilizado los comandos *xlab* e *ylab* para especificar etiquetas en los ejes X e Y respectivamente. Asimismo la opción *freq=FALSE* indica a R que en el eje Y represente frecuencias relativas. Las frecuencias absolutas se obtienen con *freq=TRUE*. El gráfico resultante se muestra en la figura 5.

Polígonos de frecuencias.

Los *polígonos de frecuencias* son representaciones similares al histograma, sustituyendo las barras por líneas que unen los distintos valores de frecuencia correspondientes a cada marca de clase. Suelen utilizarse también para representar las frecuencias acumuladas.

En R no existe ninguna función específica para dibujar polígonos de frecuencias. Sin embargo es muy sencillo construirlos a partir de la tabla de frecuencias:

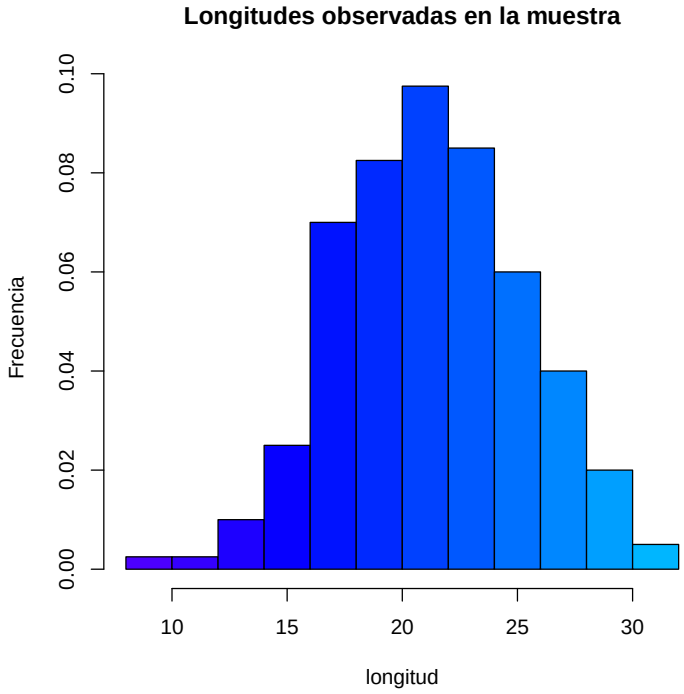

Figura 5: Histograma de longitudes de los sargos de la muestra.

```
> par(mfrow = c(1, 2))
> tbl = data.frame(table.freq(hist(long, plot = FALSE)))
> plot(tbl$MC, tbl$fi, type = "b", col = "red", lwd = 3,
     xlab = "Marca de Clase", ylab = "Frecuencia", sub = "(Longitud del sargo)",
     main = "Poligono de frecuencias absolutas")
> plot(tbl$MC, tbl$Fi, type = "b", col = "darkgreen", lwd = 3,
     xlab = "Marca de Clase", ylab = "Frecuencia", sub = "(Longitud del sargo)",
     main = "Poligono de frecuencias absolutas \nacumuladas")
```

6. Medidas de síntesis o resumen de variables numéricas.

Las variables numéricas pueden resumirse a través de diversas medidas que describen sus características de:

- **Posición:** percentiles y cuartiles
- **Tendencia central:** media, mediana y moda

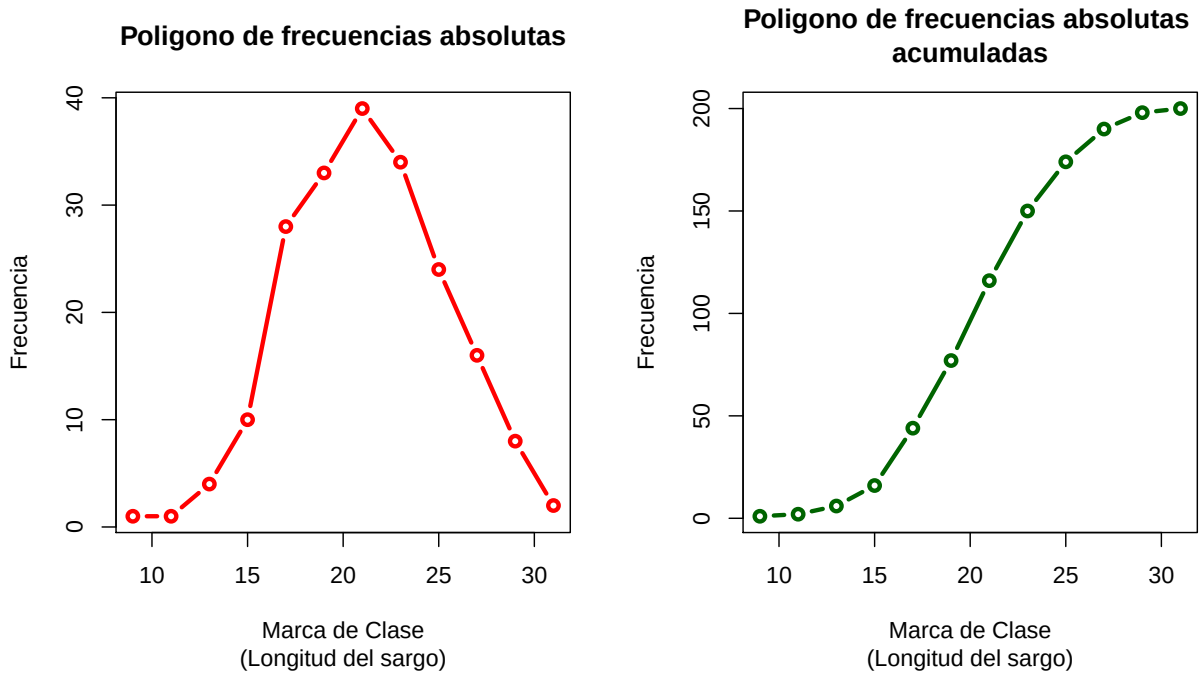

Figura 6: Polígonos de frecuencias para las longitudes de los sargos de la muestra.

- **Dispersión:** Varianza, desviación típica (o estándar), coeficiente de variación y rango.
- **Forma:** Asimetría, Apuntamiento (curtosis).

Pasamos a describir cada una de estas medidas.

6.1. Medidas de posición.

El *k-ésimo percentil* es un valor P_k tal que el $k\,\%$ de las observaciones de la variable tienen un valor menor o igual que P_k. Los percentiles 25, 50 y 75 reciben el nombre de *primer, segundo y tercer cuartiles,* respectivamente.

Los percentiles en R se calculan mediante la función *quantile()*. Así, para calcular los percentiles 0,05, 0,25, 0,50, 0,75, 0,9 y 0,95 de la longitud de los peces obtenidos durante la campaña de muestreo utilizaremos:

```
> quantile(long, probs = c(0.05, 0.25, 0.5, 0.75, 0.9,
    0.95))

    5%     25%     50%     75%     90%     95%
15.470 18.840 21.245 23.980 26.422 27.773
```

6.2. Medidas de tendencia central.

Mediana. Es el valor que ocupa la posición intermedia del conjunto de datos una vez que éstos se han ordenado de menor a mayor. La mediana es, por tanto, aquel valor que es mayor que la primera mitad de los datos, y menor que la segunda mitad. Obviamente, por su definición, coincide con el percentil 50, P_{50} y con el segundo cuartil. Si el número de datos es impar, se toma como mediana el valor que deja a derecha e izquierda el mismo número de datos. Si el número de datos es par , entonces la mediana es igual al promedio de los dos valores centrales.

En R la mediana se calcula mediante el comando *median()*. La longitud mediana de los sargos de la muestra es:

```
> median(long)

[1] 21.245
```

Media aritmética. Si en una muestra de una variable X se han observado los valores $x_1, x_2, \ldots, x_k$, siendo $n_1, n_2, \ldots, n_k$ sus frecuencias absolutas (número de veces que se ha observado cada valor), se define la *media aritmética* como:

$$\overline{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots, + x_k n_k}{n} = \sum_{i=1}^{k} x_i \frac{n_i}{n} = \sum_{i=1}^{k} x_i f_i$$

siendo $n = \sum_{i=i}^{k} n_i$ el número total de observaciones y f_i la frecuencia relativa del valor x_i.

La media aritmética representa el centro de gravedad de los datos, por lo que efectivamente puede entenderse como medida de tendencia central.

En R la media se calcula mediante el comando *mean()*:

```
> mean(long)

[1] 21.3458
```

Moda. Es el valor que más veces se repite (esto es, el valor con mayor frecuencia absoluta). En el caso de datos agrupados suele sustituirse la moda por el *intervalo modal,* que se corresponde con el intervalo de mayor frecuencia absoluta observada. Tanto la moda como el intervalo modal pueden no ser únicos.

R no dispone de ninguna función para calcular la moda. En realidad tal función resulta innecesaria: si la variable que consideramos es categórica o discreta, una simple inspección de la tabla de frecuencias o del diagrama de barras nos indica cuál es el valor más frecuente (o los valores más frecuentes en caso de haber varios). En el caso de variables continuas, la inspección del histograma nos indica el intervalo modal.

En cualquier caso, con variables categóricas podemos construir la siguiente función para obtener la moda:

```
> moda = function(x) {
     tbl = table(x)
     m = which(tbl == max(tbl))
     return(names(m))
 }
```

La aplicamos para determinar de qué isla procede la mayor parte de las capturas de sargos de la muestra:

```
> moda(isla)
```

```
[1] "GC"
```

En el caso de variables continuas, podemos usar la siguiente función para obtener el intervalo modal (o intervalos modales en caso de haber varios) a partir del histograma:

```
> intModal = function(x) {
     tbl = hist(x, plot = FALSE)
     m = which(tbl$counts == max(tbl$counts))
     im = data.frame(tbl$breaks[m], tbl$breaks[m + 1])
     names(im) = c("Inf", "Sup")
     return(im)
 }
```

Aplicamos esta función para hallar el intervalo modal de la longitud de los sargos de la muestra:

```
> intModal(long)
```

```
  Inf Sup
1  20  22
```

Media geométrica. Se define como:

$$\gamma = \{x_1 \cdot x_2 \cdot \cdots \cdot x_n\}^{1/n}$$

Suele utilizarse para promediar incrementos relativos, tales como los que se observan frecuentemente en Economía o Demografía. Por ejemplo, si el tamaño de una población se ha incrementado en un 50 % en un primer año, y ha disminuido un 50 % al año siguiente, la aplicación ingenua de la media aritmética nos llevaría a concluir que, por término medio, el tamaño de la población no cambia. Sin embargo un análisis más atento nos revela que si la población parte inicialmente de, digamos, 1000 individuos, el incremento inicial del 50 % significa una cifra de 1500 individuos al acabar el primer año, y la disminución posterior del 50 % deja la población en 750 individuos; por tanto, en los dos años ha habido un decremento global del 25 %. En realidad, la tasa media de variación interanual en este caso debe calcularse mediante la media geométrica: $\gamma = (1{,}50 \cdot 0{,}50)^{1/2} = 0{,}866$. Su interpretación es que, *por término medio*, cada año el tamaño de la población es un 86.6 % del tamaño del año anterior; dos años sucesivos con esta tasa media producen una tasa acumulada de $0{,}866 \cdot 0{,}866 = 0{,}75$, o lo que es lo mismo, un 75 % del tamaño inicial, lo que sí coincide con la cifra observada.

Si en la definición de media geométrica tomamos logaritmos resulta:

$$\log \gamma = \frac{1}{n} \sum_{i=i}^{n} \log (x_i)$$

Por tanto el logaritmo de la media geométrica coincide con la media aritmética de los logaritmos de los datos originales.

R tampoco dispone de ninguna función para el cálculo de la media geométrica. No obstante, es muy fácil de calcular utilizando la propiedad anterior:

```
> tasas = c(1.5, 0.5)
> exp(mean(log(tasas)))
```

```
[1] 0.8660254
```

O incluso aplicando directamente la definición:

```
> prod(tasas)^(1/length(tasas))
```

```
[1] 0.8660254
```

Hemos utilizado aquí la función *length(tasas)* que nos devuelve la longitud (número de elementos) del vector tasas. En este caso es innecesario (podíamos haber puesto directamente 2), pero de esta forma tenemos una expresión general que nos evita en otros casos tener que contar el número de términos cuya media geométrica se va a calcular.

6.3. Medidas de Dispersión.

Varianza. Si en una muestra de una variable X se han observado los valores $x_1, x_2, \ldots, x_k$, siendo $n_1, n_2, \ldots, n_k$ sus frecuencias absolutas (número de veces que se ha observado cada valor), se define la *varianza muestral* (o *cuasi-varianza*) como:

$$s^2 = var\,(X) = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i = \frac{n}{n-1}\sum_{i=1}^{k}(x_i - \overline{x})^2 \frac{n_i}{n} = \frac{n}{n-1}\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i$$

siendo $n = \sum_{i=i}^{k} n_i$ el número total de observaciones y f_i la frecuencia relativa del valor x_i. Obviamente la varianza es una medida de dispersión ya que cuanto más alejados entre sí se encuentren los valores x_i más lejos estarán de su media aritmética y mayor será el valor de la varianza; y a la inversa, cuánto más próximos entre sí, más cerca estarán de la media y menor será la varianza.

En R la varianza se calcula mediante la función *var()*:

```
> var(long)

[1] 15.12042
```

Desviación típica (o *Desviación estándar*). Es la raiz cuadrada de la varianza. Se obtiene así una medida de dispersión en las mismas unidades que la variable original:

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

En R se obtiene con la función *sd()*:

```
> sd(long)

[1] 3.888498
```

Coeficiente de variación. La varianza y la desviación estándar son medidas de dispersión dependientes de las unidades en las que se mida la variable. El coeficiente de variación es una medida de dispersión adimensional que se define como:

$$cv(X) = \frac{s}{\overline{x}}$$

(siempre que $\overline{x} \neq 0$).

El coeficiente de variación resulta especialmente útil para comparar el grado de dispersión de variables que se miden en unidades diferentes. Por ejemplo si, en la muestra que estamos utilizando, queremos saber si los sargos presentan más dispersión en longitud o en peso, no tiene sentido comparar sus desviaciones típicas, medidas en centímetros y en gramos respectivamente. Sin embargo sus coeficientes de variación:

```
> sd(long)/mean(long)

[1] 0.1821669

> sd(peso)/mean(peso)

[1] 0.4552767
```

nos indican una mayor variabilidad en peso.

Rango y rango intercuartílico. El rango de una variable se define como la distancia entre los valores mínimo y máximo:

$$\text{rango}\,(X) = \text{máx}\,(X) - \text{mín}\,(X)$$

Asimismo, el rango intercuartílico es la distancia entre los cuartiles primero y tercero ($P_{75} - P_{25}$).

La función *range()* de R nos proporciona los valores mínimo y máximo de una variable. A su vez, como ya hemos visto, la función *quantile()* nos proporciona los cuartiles. La función *diff()* nos da la distancia entre valores:

```
> range(long)

[1]  9.74 30.65

> diff(range(long))
```

```
[1] 20.91

> quantile(long, probs = c(0.25, 0.75), names = FALSE)

[1] 18.84 23.98

> diff(quantile(long, probs = c(0.25, 0.75), names = FALSE))

[1] 5.14
```

6.4. Medidas de forma.

Coeficiente de asimetría. En los casos en que los datos estén distribuidos de forma simétrica, la media y mediana son medidas aproximadamente similares. Sin embargo, cuando los datos muestran largas colas a la derecha (valores altos muy alejados del resto de los datos), el valor de la media tenderá a ser mayor que el de la mediana. Así por ejemplo, para el conjunto de datos $\{1, 2, 2, 3, 3, 3, 4, 4, 5\}$ media y mediana coinciden en el valor 3. Por el contrario, si el conjunto de datos es $\{1, 2, 2, 3, 3, 3, 4, 4, 50\}$, la mediana sigue siendo el valor 3, mientras que la media aritmética se desplaza al valor 8. En estos casos, la mediana representa (localiza) mejor el centro de la distribución que la media aritmética.

Dada una muestra de una variable X formada por n observaciones, siendo $\overline{x}$ su media aritmética y s su desviación típica, la asimetría de la variable puede cuantificarse a través del *coeficiente de asimetría de Fisher*, definido como:

$$a_F = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

si bien en la práctica es preferible utilizar la siguiente versión corregida:

$$a_F = \frac{n\sqrt{(n-1)}}{n-2}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}} = \frac{n}{(n-1)(n-2)}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$$

ya que esta última expresión tiende a producir valores más próximos a la asimetría de la variable en la población de la que se ha extraído la muestra. Cuando los datos son perfectamente simétricos este coeficiente es nulo. Cuando los valores se concentran a

la derecha, con largas colas a la izquierda este coeficiente es negativo (*asimetría a la izquierda o negativa*); y cuando los valores tienden a concentrarse a la izquierda, con largas colas a la derecha, el coeficiente es positivo (*asimetría a la derecha o positiva*).

El paquete base de R no contiene ninguna función para el cálculo del coeficiente de asimetría. Podríamos construir una función para su cálculo, pero en este caso ya existen varios paquetes que lo hacen, entre ellos el paquete *agricolae* que ya hemos usado con anterioridad. Para calcular la asimetría utilizamos la función *skewness()*:

```
> require(agricolae)
> skewness(ldors)

[1] -0.3480565

> skewness(phig)

[1] 1.400168
```

Como vemos, la distancia desde el morro del pez a la aleta dorsal (ldors) presenta asimetría negativa y el peso del hígado (phig) asimetría positiva. En la figura 7 podemos observar los histogramas de ambas variables y comprobar que son efectivamente asimétricos.

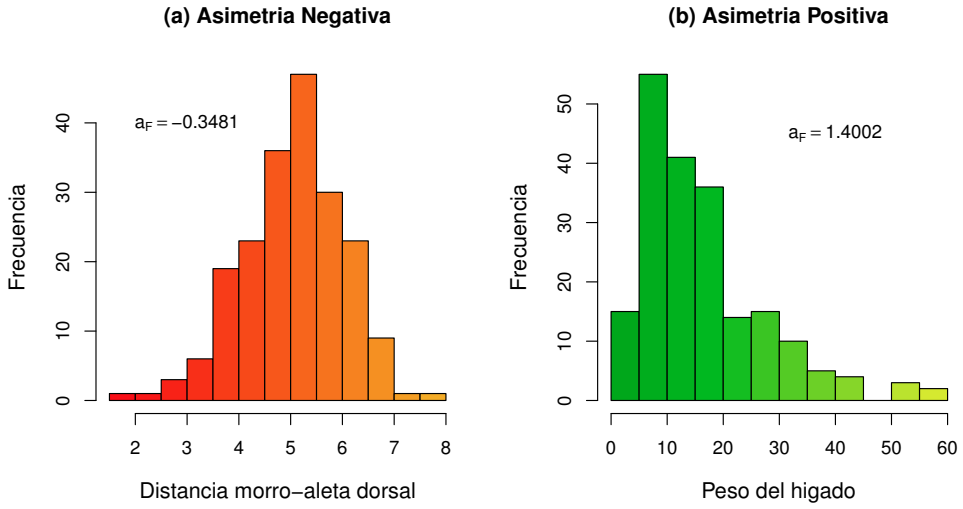

Figura 7: Variables que presentan asimetría (a) Histograma de la distancia del morro a la aleta dorsal (asimetría negativa) (b) Histograma del peso del hígado (asimetría positiva).

Nota: en el fragmento de código anterior hemos usado la función *require()*. Esta función comprueba si una librería –en este caso *agricolae*– ha sido ya cargada mediante *library()*. Si la librería ya ha sido cargada, *require()* no hace nada, y en caso contrario carga la librería.

Coeficiente_de_apuntamiento_(curtosis): mide el grado de concentración que presentan los valores alrededor de la zona central del conjunto de datos. La definición habitual de curtosis es:

$$\kappa = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

si bien, como ocurre con la asimetría, en la práctica se emplea una versión corregida (cuando n es grande produce prácticamente el mismo valor que la anterior, pero para valores de n pequeños tiende a producir valores de curtosis más próximos al verdadero valor en la población de la que se ha extraído la muestra):

$$\kappa = \frac{n\,(n+1)}{(n-1)\,(n-2)\,(n-3)}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{S^4} - 3\frac{(n-1)^2}{(n-2)\,(n-3)}$$

Si $\kappa > 0$ la forma del conjunto de datos es "puntiaguda" (*leptocúrtica*); por el contrario, si $\kappa < 0$, la forma es "aplastada" (*platicúrtica*). El caso $\kappa = 0$ corresponde a una forma "normal" (*mesocúrtica*), ni muy apuntada ni muy aplastada.

Al igual que ocurría con la asimetría, R no dispone en su paquete base de ninguna función para el cálculo de la curtosis, si bien podemos encontrarla en el paquete *agricolae*:

```
> kurtosis(ldors)

[1] 0.2372677

> kurtosis(phig)

[1] 2.168432
```

Como vemos, ambas variables presentan apuntamiento positivo (corresponden a distribuciones leptocúrticas), tal como podemos apreciar visualmente en los histogramas mostrados en la figura 7).

6.5. Valores perdidos.

En muchas ocasiones no se dispone de los valores de todas las variables, bien sea porque no se han podido medir sobre los objetos de la muestra, bien sea porque dichos valores no quedaron registrados en el archivo de datos. En cualquier caso, cuando R encuentra un espacio en blanco en una posición del archivo en la que esperaba encontrar un dato, considera que ese valor está perdido y lo codifica internamente como *NA* (*No Asignado*). A veces cuando un valor de la muestra se ha perdido, en lugar de dejar un espacio en blanco en el archivo de datos, se consigna con un valor identificativo (-1, 9999, "*",...). En tal caso, al leer el archivo hay que indicar a R que ese valor representa un valor perdido mediante la opción *na.strings*. Si, por ejemplo, los valores perdidos se identificaran con 9999, en el comando de lectura deberíamos especificar, junto a las opciones ya vistas en la sección 4.1:

```
> MisDatos = read.table(..., na.strings = "9999", ...)
```

La presencia de valores perdidos afecta a las funciones que calculan las medidas de síntesis (*mean, sd, quantile*, etc). Recordemos que en nuestro archivo de ejemplo, el peso de las gónadas no se había medido para todos los peces. Si quisiéramos calcular el peso medio de las gónadas obtendríamos:

```
> mean(pgon)

[1] NA
```

lo que indica que R no ha podido calcularlo debido a la presencia de valores perdidos. En realidad R sí que puede calcular el peso medio, y el hecho de que no lo calcule directamente significa más bien un aviso para que tengamos en cuenta la presencia de tales valores. Para calcular la media (o cualquier otra medida de síntesis) en estas condiciones, hay que añadir la opción *na.rm=TRUE* (acrónimo de *NA remove*):

```
> mean(pgon, na.rm = T)

[1] 11.48706
```

Nota: Bajo determinadas condiciones la existencia de valores perdidos (sobre todo si éstos constituyen una parte importante de la muestra) podría dar lugar a que la muestra no fuese realmente representativa de la población de la que se ha extraido y por tanto el análisis estadístico que hagamos de la misma tendría escaso valor.

6.6. Diagrama de cajas y barras (*boxplot*)

Estos diagramas representan los percentiles de una variable y son especialmente útiles para una comparación gráfica de varias poblaciones, así como para la detección de posibles valores anómalos (*outliers*). Su construcción se realiza de la siguiente forma: sea $\{x_1, \cdots, x_n\}$ el conjunto de datos correspondientes a una variable numérica X, y representemos por P_{25}, P_{50} y P_{75} los percentiles 25, 50 y 75 respectivamente; se dibuja un rectángulo vertical cuyos lados inferior y superior corresponden a P_{25} (primer cuartil) y P_{75} (tercer cuartil) respectivamente; a la altura P_{50} (mediana) se traza un segmento horizontal. Por último el rectángulo se une mediante lineas a dos barras correspondientes los extremos de la distribución, trazadas a alturas respectivas b y B:

1. *Barra superior*: $B = \text{mín}\left\{\text{máx}\left(X\right),\ P_{75} + 1{,}5\left(P_{75} - P_{25}\right)\right\}$
2. *Barra inferior*: $b = \text{máx}\left\{\text{mín}\left(X\right),\ P_{25} - 1{,}5\left(P_{75} - P_{25}\right)\right\}$

Los valores de los datos que quedan fuera de las barras superior e inferior se marcan con puntos y se entenderá que pueden ser anómalos, y deben ser revisados por si constituyeran errores de medida, datos correspondientes a otra población, etc.

Para obtener en R el boxplot de la variable longitud, por ejemplo, ejecutaríamos simplemente la función:

```
> boxplot(long, col = "orange", main = "longitud")
```

6.7. Medidas de síntesis en subgrupos de la muestra.

En muchas ocasiones los objetos de la muestra pueden clasificarse según los valores de alguna variable categórica. Así, en los datos de nuestro ejemplo, podríamos clasificar los sargos en función de la isla de procedencia, o en función de su sexo. En la sección 5.1 ya hemos visto como construir tablas cruzadas para esta clase de variables. Cuando lo que nos interesa es calcular las distintas medidas de síntesis sobre cada uno de los grupos que forman la muestra, en R podemos utilizar los comandos *by()* y *aggregate()*.

Así, por ejemplo, para calcular la longitud media de los sargos según sexo usaríamos la función:

```
> by(long, sexo, mean)
```

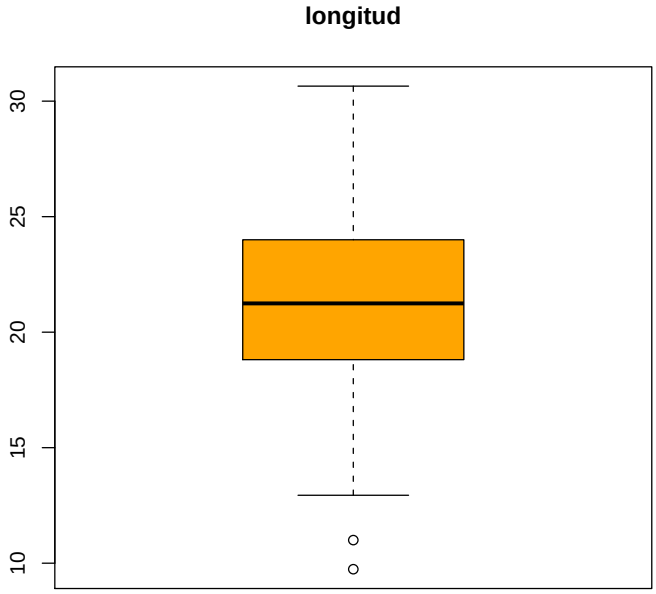

Figura 8: Diagrama de cajas y barras para la variable longitud.

```
sexo: Hembra
[1] 20.84080
------------------------------------------------------
sexo: Macho
[1] 22.00172
```

o de manera equivalente:

```
> aggregate(long, by = list(sexo), mean)

  Group.1        x
1  Hembra 20.84080
2   Macho 22.00172
```

La presentación de la tabla construida con el comando *aggregate()* mejora si:

- La variable (o variables, ya que pueden incluirse varias) a resumir se especifica como subconjunto (*subset()*) del conjunto de datos original.
- La variable (o variables, también podrían incluirse varias) que define los grupos se *renombra* dentro del comando *list()*.

Veamos el efecto de estos cambios, calculando la longitud y el peso medios por sexo y por isla en nuestra muestra:

```
> aggregate(subset(sargos, select = c(long, peso)), by = list(Sexo = sexo,
     Isla = isla), mean)

     Sexo Isla     long     peso
1  Hembra   HI 20.98250 156.9800
2   Macho   HI 22.78571 188.4914
3  Hembra   LP 20.46750 146.8017
4   Macho   LP 23.72500 216.5800
5  Hembra   LG 21.11167 158.4017
6   Macho   LG 22.08667 169.3333
7  Hembra   TF 21.77286 176.5952
8   Macho   TF 21.82632 174.2589
9  Hembra   GC 20.66786 152.8236
10  Macho   GC 22.39400 185.4225
11 Hembra   FV 20.07000 144.1612
12  Macho   FV 21.02563 161.7181
13 Hembra   LZ 20.81000 155.5855
14  Macho   LZ 20.47000 149.0160
```

Si quisiéramos calcular varias medidas de síntesis sobre los subgrupos de la muestra debemos definir una función con las medidas a calcular; así, por ejemplo, si de cada variable quisiéramos obtener la media, desviación típica, mínimo y máximo, construiríamos la función de resumen siguiente:

```
> resumen = function(x, ...) {
     m = mean(x, ...)
     s = sd(x, ...)
     mn = min(x, ...)
     mx = max(x, ...)
     output = round(c(m, s, mn, mx), 2)
     names(output) = c("media", "sd", "min", "max")
     return(output)
 }
```

Nota: los puntos sucesivos permiten que la función reciba otras opciones; por ejemplo, si al llamarla añadiésemos *na.rm=T* podríamos calcular todas las medidas de síntesis especificadas en presencia de valores perdidos.

Utilizamos esta función para resumir la variable peso según sexo:

```
> by(peso, sexo, resumen)

sexo: Hembra
 media     sd    min    max
156.50  73.00  27.09 371.89
------------------------------------------------
sexo: Macho
 media     sd    min    max
178.43  77.51  18.04 382.18
```

O, utilizando *aggregate()* para el peso del hígado, teniendo en cuenta la presencia de valores perdidos:

```
> aggregate(subset(sargos, select = phig), by = list(Sexo = sexo),
     resumen, na.rm = T)

    Sexo phig.media phig.sd phig.min phig.max
1 Hembra      15.36   11.66     1.70    59.00
2  Macho      18.06   10.43     0.70    55.00
```

Para concluir esta sección citemos que es posible utilizar la función *boxplot()* para hacer diagramas de cajas y barras según subgrupos de la muestra. El siguiente código genera los gráficos mostrados en la figura 9

```
> boxplot(peso ~ sexo, main = "Peso", col = c("pink2",
     "cyan3"))
> boxplot(peso ~ isla, main = "Peso", col = heat.colors(14))
```

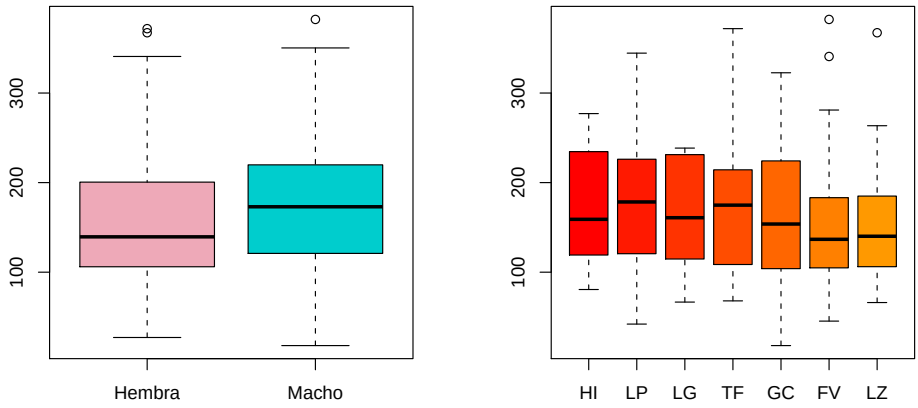

Figura 9: Boxplots para subgrupos de la muestra. Izquierda: peso según sexo. Derecha: peso según isla.

7. Asociación entre variables continuas.

En la sección 5.2 hemos llevado a cabo la descripción de datos correspondientes a variables continuas: tablas de frecuencias, histogramas y polígonos de frecuencias. Asimismo, en la sección 6 hemos presentado las medidas de síntesis que nos permiten resumir las características de estas variables en unos pocos valores. En ambos casos, el análisis de los datos ha sido univariante: cada variable se estudia aisladamente, sin conexión con las restantes variables continuas medidas en la muestra. Todo lo más, en 6.7 hemos visto como varía una variable continua en varios grupos definidos por una variable categórica.

Ahora bien, cuando se realiza el estudio conjunto de dos variables, normalmente el objetivo es determinar si existe algún tipo de asociación entre ellas o si, por el contrario, son independientes. En términos prácticos, la asociación significa que el conocimiento de los valores de una de las variables proporciona alguna información sobre los valores de la otra. Por ejemplo, conocer la estatura de una persona nos informa sobre su peso, ya que las personas más altas tienen, en general, un peso mayor que las personas más bajas. Esta asociación estadística, obviamente no es exacta: dos personas de la misma altura no tienen que tener exactamente el mismo peso, y una persona más alta puede pesar menos que una más baja. La figura 10 ilustra este tipo de asociación: valores altos de X tienden a ir acompañados de valores altos de Y, a la vez que valores bajos de X tienden a ir acompañados de valores bajos de Y, si bien no de manera exacta.

Al estudiar la asociación entre variables continuas podemos encontrarnos ante dos problemas distintos, según cuál sea el objetivo de nuestro estudio:

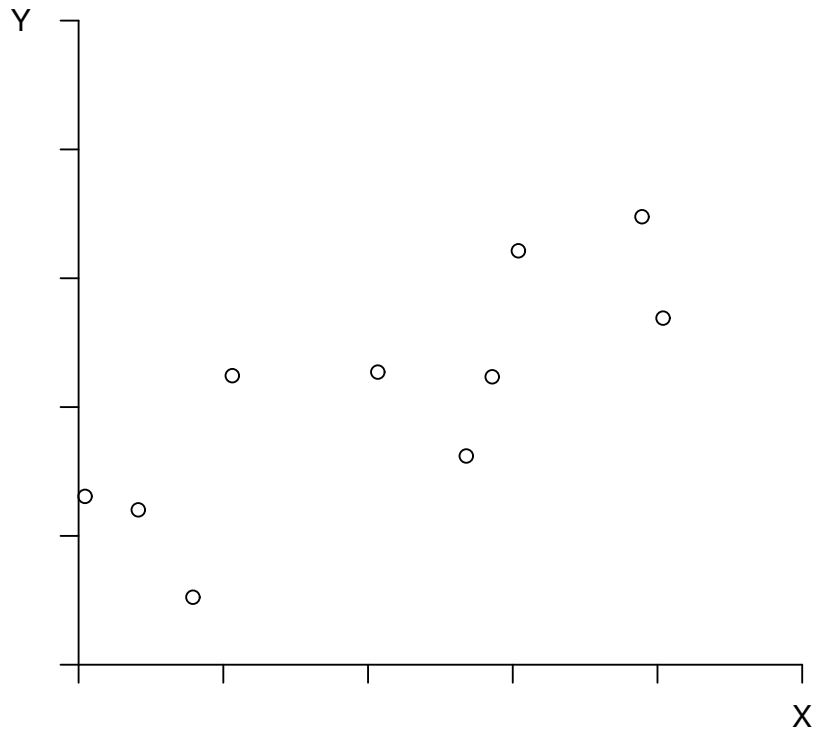

Figura 10: Nube de puntos correspondiente a la observación de dos variables X e Y sobre los sujetos de una muestra.

Análisis de regresión: nuestro objetivo es construir un modelo para *predecir* el valor de una variable Y cuando se conoce el valor de otra variable X. Esto es, si para el sujeto i-ésimo de la muestra sabemos que $X = x_i$, queremos hallar una función f tal que el valor de Y predicho para ese sujeto sea $y_i = f(x_i) + \varepsilon_i$. Los términos ε_i representan los *errores de predicción.* Cuando la función $f(X)$ es lineal nos hallamos ante un problema de *regresión lineal.* En caso contrario estaríamos ante un problema de *regresión no lineal.*

Análisis de correlación: nuestro objetivo es medir la intensidad de la asociación lineal entre dos variables X e Y. Una correlación alta indicaría una fuerte asociación y una correlación baja, una asociación débil. Las variables son tratadas de forma simétrica, no hay una variable predictora y una variable a predecir.

En un análisis de correlación ambas variables X e Y son *aleatorias*, lo que significa que sus valores no se conocen hasta haberlas observado. El observador usa la correlación para medir la asociación entre estas variables tal como se produce en la naturaleza. En la muestra que venimos utilizando como ejemplo, para cada sargo se mide su longitud y su peso; antes de

tomar la muestra estos valores son desconocidos, por lo que ambas variables son aleatorias. Sin embargo, en un análisis de regresión, si bien ambas variables pueder ser también aleatorias, es frecuente que el observador (o experimentador) fije de antemano los valores de la variable X y mida a continuación como responde la variable Y, que sería en tal caso la única aleatoria. Es importante señalar que en estas condiciones la asociación que se produzca entre X e Y puede ser muy distinta de la que se observa en condiciones naturales.

Nota: tanto en el caso de la regresión como en el de la correlación *no debe confundirse asociación con causalidad*. Podemos usar una regresión para predecir la edad de un niño a partir de su estatura, ya que niños más altos probablemente tienen mayor edad; pero evidentemente, la altura *no es la causa* de la edad. Podemos detectar una correlación –asociación– fuerte entre altos niveles de glucosa en sangre e hipertensión; sin embargo ello no quiere decir que la diabetes cause la hipertensión o que la hipertensión cause la diabetes; no puede descartarse la posibilidad de que exista una causa común –en este caso, el *síndrome metabólico*– que sea en realidad la que da lugar a la asociación entre ambas enfermedades.

Sólo los estudios experimentales pueden probar de manera concluyente una posible relación causal entre dos variables: en estos estudios el experimentador controla todos los posibles factores de confusión (terceras variables que puedan influir en la asociación) y las posibles fuentes de "ruido" en los datos; si en tales condiciones la modificación de X produce un cambio en Y, y se cuenta además con un mecanismo para explicar como se produce tal efecto, entonces y sólo entonces se puede hablar de causalidad, o al menos de influencia de X sobre Y.

7.1. Regresión lineal.

Una de las formas más comunes de asociación entre variables es la asociación lineal. Los valores representados en la figura 10 muestran precisamente este tipo de asociación. En la práctica resulta de interés determinar la ecuación de la recta que define esta relación y que permite aproximar el valor de Y cuando se conoce el valor de X. Esta recta se denomina *recta de regresión de Y sobre X*, y su ecuación es de la forma $Y = b_0 + b_1 X$.

La variable X recibe el nombre de *variable explicativa* (o *independiente*) y la Y el de *variable respuesta* (o *dependiente*). El valor de b_1 es la *pendiente* y b_0 es la *ordenada en el origen*. La pendiente representa el incremento (si b_1 es positivo) o decremento (si b_1 es negativo) que experimenta el valor promedio de Y por cada unidad de incremento en el valor de X.

Asimismo, la ordenada en el origen b_0 es el valor de Y cuando $X = 0$. Hay que señalar que, desde el punto de vista del análisis de los datos, esta interpretación solo debe realizarse cuando el valor $X = 0$ ha sido efectivamente observado. Si, por ejemplo, Y fuese el peso de una persona de altura X y se dispusiera de una recta de regresión $Y = b_0 + b_1X$ que relacionase ambas variables, dado que no existen personas de estatura $X = 0$ no tiene sentido decir que b_0 es el peso aproximado de tales personas.

Para calcular la recta de regresión de Y sobre X se utiliza habitualmente el *método de los mínimos cuadrados*. Supongamos que sobre una muestra de n objetos hemos medido el par de variables (X, Y), y que los valores observados han sido $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Supongamos además que estos puntos se encuentran alineados a lo largo de una recta de ecuación $Y = b_0 + b_1X$, y llamemos $\hat{y}_i = b_0 + b_1x_i$ al valor que corresponde sobre la recta al punto x_i (*valor predicho por la recta*). El error de predicción sería entonces $e_i = y_i - \hat{y}_i$. El criterio de los mínimos cuadrados consiste en determinar los valores de b_0 y b_1 de forma que la suma de distancias al cuadrado entre observaciones y predicciones sea mínima, esto es:

$$\text{mín} \sum_{i=1}^{n} e_i^2 = \text{mín} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{b_0,b_1} \sum_{i=1}^{n} (y_i - (b_0 + b_1x_i))^2$$

De esta forma se consigue que la recta pase simultáneamente lo más cerca posible de todos los puntos observados. La figura 11 ilustra gráficamente esta idea.

Llamemos:

$$L(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1x_i)^2$$

Para obtener los valores de b_0 y b_1 que minimizan esta expresión derivamos con respecto a b_0 y a b_1 e igualamos a 0, obteniendo las llamadas ecuaciones normales de mínimos cuadrados:

$$\begin{array}{rcccl} \frac{\partial L(b_0,b_1)}{\partial b_0} & = & -2\sum_{i=1}^{n} (y_i - b_0 - b_1x_i) & = & 0 \\ \frac{\partial L(b_0,b_1)}{\partial b_1} & = & -2\sum_{i=1}^{n} (y_i - b_0 - b_1x_i)x_i & = & 0 \end{array}$$

De la primera ecuación se tiene:

$$-2\sum_{i=1}^{n} (y_i - b_0 - b_1x_i) = 0 \Rightarrow \sum_{i=1}^{n} (y_i - b_0 - b_1x_i) = 0 \Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1x_i = 0$$
$$\Rightarrow \sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0 \Rightarrow b_0 = \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow b_0 = \bar{y} - b_1\bar{x}$$

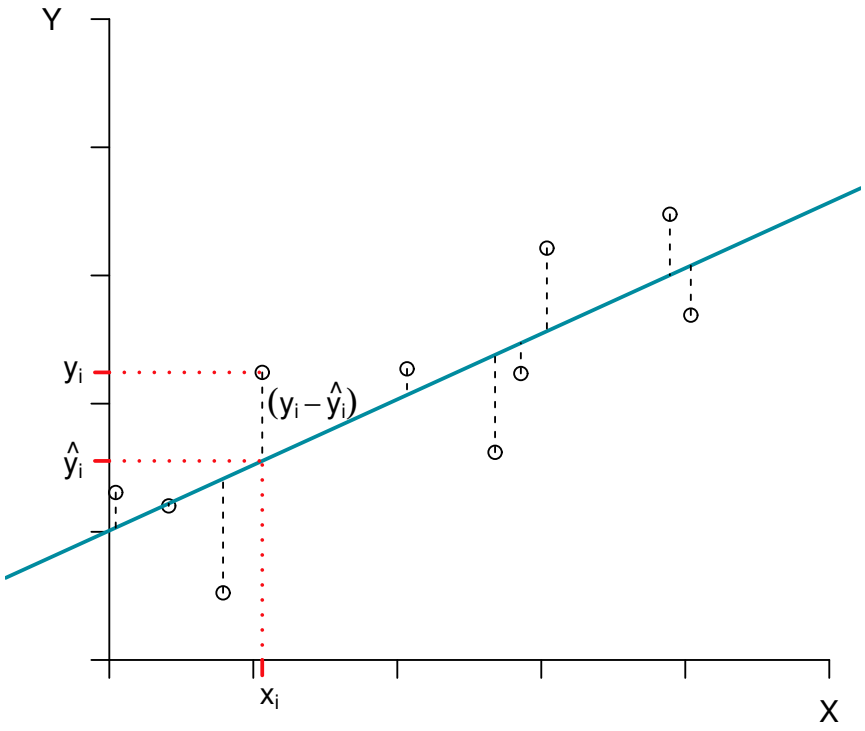

Figura 11: Recta de regresión ajustada a la nube de puntos de la figura 10. Las lineas a trazos verticales representan las distancias de los puntos a la recta. El método de los mínimos cuadrados busca la recta que minimiza la suma de los cuadrados de estas distancias.

Sustituyendo en la segunda ecuación:

$$-2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i \;=\; 0 \Rightarrow \sum_{i=1}^{n}(y_i - (\bar{y} - b_1\bar{x}) - b_1 x_i)x_i = 0 \Rightarrow$$

$$\sum_{i=1}^{n}(y_i - \overline{y})\,x_i - b_1\sum_{i=1}^{n}(x_i - \overline{x})\,x_i = 0 \Rightarrow b_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})\,x_i}{\sum_{i=1}^{n}(x_i - \overline{x})\,x_i} = \frac{\sum_{i=1}^{n} y_i x_i - \overline{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i}$$

Si tenemos en cuenta que:

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

podemos sustituir en la expresión anterior y nos queda:

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Una vez obtenido el valor de b_1, el valor de b_0 se despeja de:

$$b_0 = \bar{y} - b_1\bar{x}$$

En R es muy sencillo obtener la recta de regresión. La siguiente sintaxis muestra como calcular la recta de regresión entre la longitud total del pez y la distancia desde el morro a la aleta dorsal:

```
> lm(peso ~ long)

Call:
lm(formula = peso ~ long)

Coefficients:
(Intercept)         long
    -236.20        18.84
```

El valor indicado como intercept es la ordenada en el origen b_0, mientras que el valor bajo el nombre de la variable es la pendiente b_1. Para representar esta recta gráficamente podemos utilizar la siguiente sintaxis, cuyo resultado se muestra en la figura 12.

```
> plot(long, ldors, xlab = "Longitud total", ylab = "Distancia morro-aleta dorsal",
     main = "Regresión Longitud-Distancia a la aleta dorsal")
> recta = lm(ldors ~ long)
> abline(recta, col = "darkgreen", lwd = 2)
```

Con R es posible dibujar en un mismo gráfico nubes de puntos correspondientes a distintos grupos de datos, mostrando el ajuste de regresión para cada uno. Por ejemplo, la siguiente sintaxis repite el gráfico anterior pero dibujando de color distinto machos y hembras, y ajustando una recta de regresión a cada grupo:

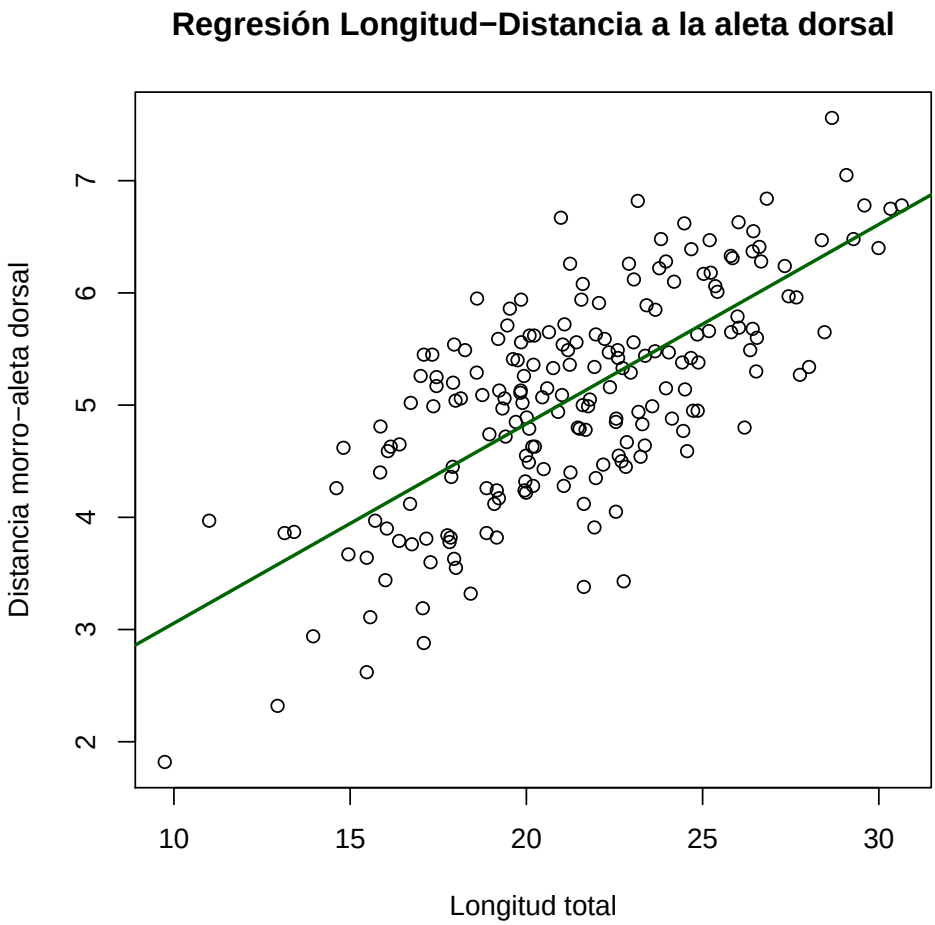

Figura 12: Recta de regresión para la distancia desde el morro a la aleta dorsal frente a la longitud total del pez.

```
> plot(long, ldors, xlab = "Longitud total", ylab = "Distancia morro-aleta dorsal",
     main = "Regresión Longitud-Distancia a la aleta dorsal",
     type = "n")
> with(subset(sargos, sexo == "Hembra"), {
     points(long, ldors, col = "pink3", pch = 19)
     abline(lm(ldors ~ long), col = "pink3", lwd = 2)
 })
> with(subset(sargos, sexo == "Macho"), {
     points(long, ldors, col = "cyan4", pch = 19)
     abline(lm(ldors ~ long), col = "cyan4", lwd = 2)
 })
> legend("topleft", c("Hembra", "Macho"), col = c("pink3",
     "cyan4"), pch = 19, lty = 2, bty = "n")
```

El resultado de esta sintaxis se muestra en la figura 13 .

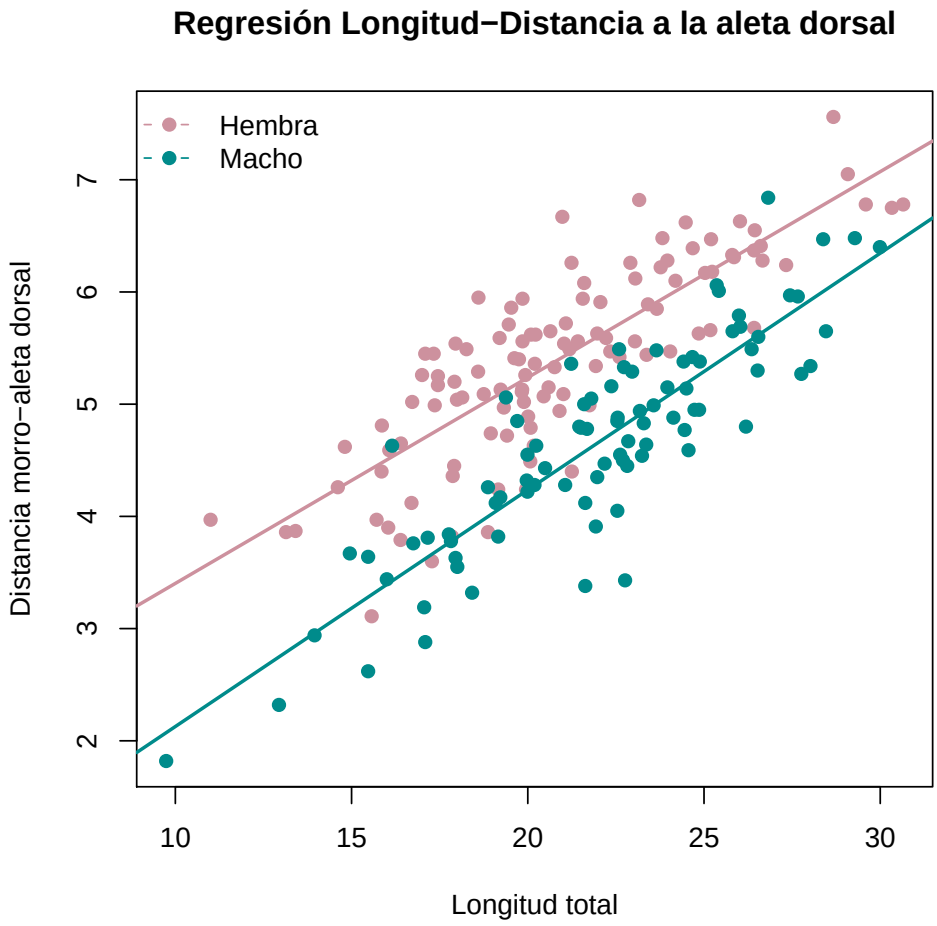

Figura 13: Rectas de regresión para la distancia desde el morro a la aleta dorsal frente a la longitud total del pez, ajustadas para cada sexo.

Nota: el paquete *lattice* contiene funciones gráficas de alto nivel que permiten construir este gráfico (y otros más complejos) de manera mucho más simple.

Si queremos obtener los valores numéricos de las ecuaciones de ambas rectas bastará con ejecutar:

```
> lm(ldors ~ long, data = subset(sargos, sexo == "Hembra"))

Call:
lm(formula = ldors ~ long, data = subset(sargos, sexo == "Hembra"))
```

```
Coefficients:
(Intercept)         long
     1.5677       0.1835

> lm(ldors ~ long, data = subset(sargos, sexo == "Macho"))

Call:
lm(formula = ldors ~ long, data = subset(sargos, sexo == "Macho"))

Coefficients:
(Intercept)         long
    0.01804      0.21091
```

7.2. Covarianza y correlación

La figura 14 nos muestra dos nubes de puntos. Se aprecia claramente que los datos de la nube (a) muestran una asociación lineal muy fuerte, mientras que en la nube (b) esta asociación es más débil.

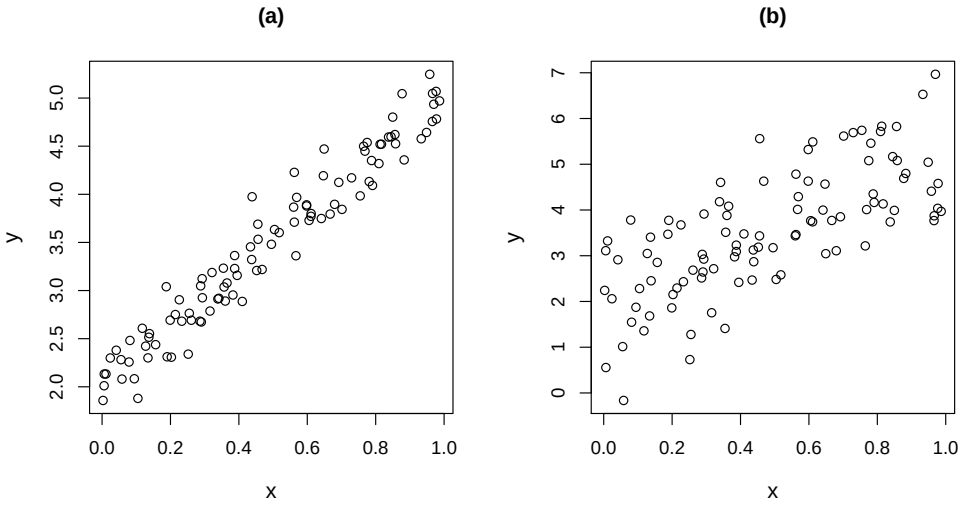

Figura 14: Nubes de puntos con distintos grado de asociación (a) Asociación lineal fuerte. (b) Asociación lineal débil.

Para medir numéricamente la intensidad de la asociación lineal entre dos variables se utiliza

la *covarianza*, definida como:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x}\,\bar{y}\right)$$

Esta medida es positiva si los datos presentan tendencia lineal creciente; es negativa si presentan tendencia lineal decreciente; y es nula si los datos no presentan tendencia lineal.

Nota: La ausencia de tendencia lineal no significa que no exista algún otro tipo de asociación (no lineal) entre X e Y.

La figura 15 muestra cuatro nubes de puntos con distinta covarianza. La figuras (a) y (b) presentan asociación lineal, el caso (a) con pendiente positiva, y por tanto con covarianza positiva, y el caso (b) con pendiente (y por tanto covarianza) negativa. A su vez las figuras (c) y (d) presentan covarianza nula; en el caso (a) porque no existe asociación entre X e Y, y en el caso (d) porque, aún existiendo asociación, esta es claramente no lineal.

La covarianza, como medida de la asociación lineal entre variables presenta un problema práctico: depende de las unidades de X e Y, y por tanto su magnitud, en términos absolutos, sea grande o pequeña puede depender más de las escalas de medida que de la fuerza de la asociación lineal entre ambas variables (por ejemplo, si X e Y son longitudes, el valor de la covarianza entre ambas será un número mucho mayor si X e Y se miden en centímetros que si se miden en metros). Por tanto es preciso introducir una nueva medida de asociación lineal que no dependa de las unidades de X e Y. Esta medida es el *coeficiente de correlación de Pearson,* definido como:

$$r = \frac{S_{XY}}{S_X S_Y}$$

siendo S_X y S_Y las desviaciones típicas respectivas de las variables X e Y. Como éstas son siempre positivas, es obvio que el signo de r coincide con el signo de S_{XY}. Además, se cumple que:

$$-1 \leq r \leq 1$$

siendo el valor absoluto de r igual a 1 cuando los puntos están *exactamente* sobre una recta. La figura 16 muestra cuatro nubes de puntos con distintos valores de correlación lineal.

Así pues:

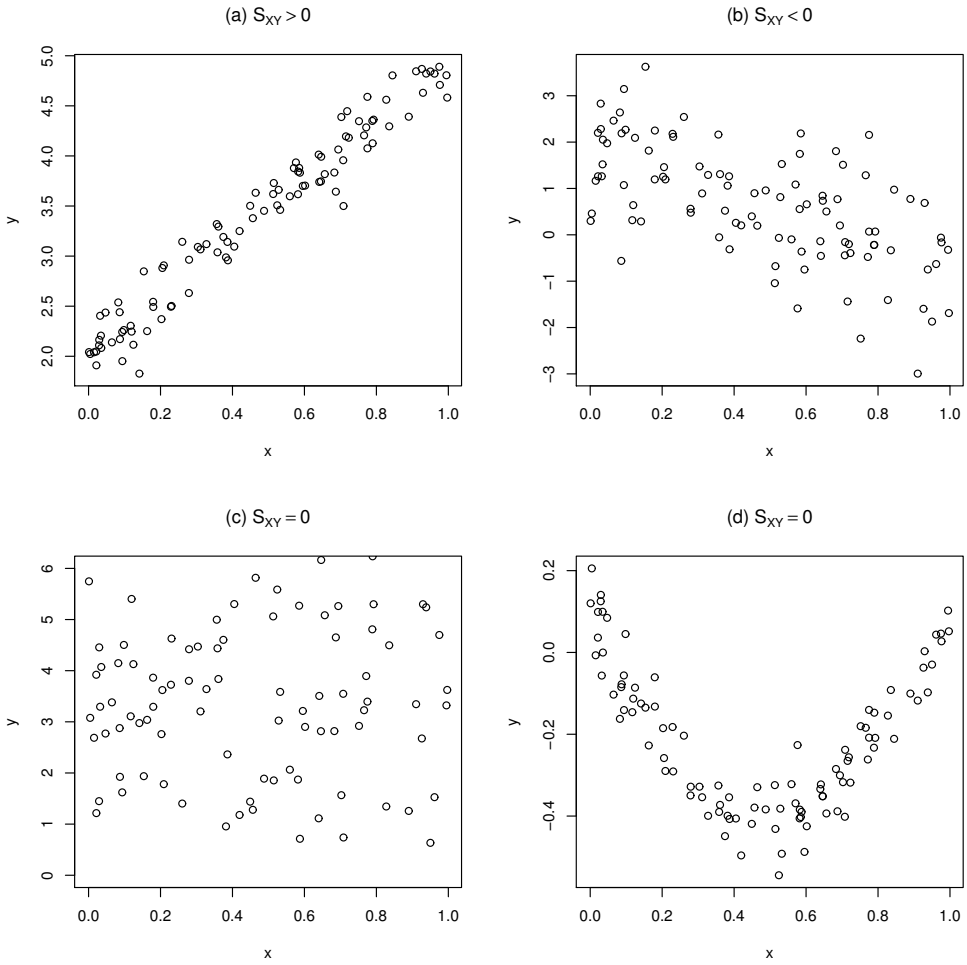

Figura 15: Nubes de puntos con distinta covarianza.

- $r > 0$: indica la presencia de una asociación lineal positiva (recta creciente). Esta asociación es tanto más fuerte (más se ajustan los puntos a la recta) cuanto más se aproxime el valor de r a 1.

- $r < 0$: indica la presencia de una asociación lineal negativa (recta decreciente); cuando aumenta el valor de X, el valor de Y disminuye proporcionalmente). Cuanto más se aproxime r a -1 tanto mejor es el ajuste a una recta.

- $r = 0$: indica la ausencia de asociación lineal entre X e Y: podría haber una ausencia absoluta de asociación como en la figura 15(c), o bien podría existir algún tipo de relación no lineal como en la figura 15(d).

Para determinar si el coeficiente de correlación es una medida adecuada de la asociación entre variables, el primer paso debe ser siempre dibujar un gráfico de la nube de puntos correspondiente a las observaciones. En los siguientes casos no es apropiado utilizar el coeficiente de

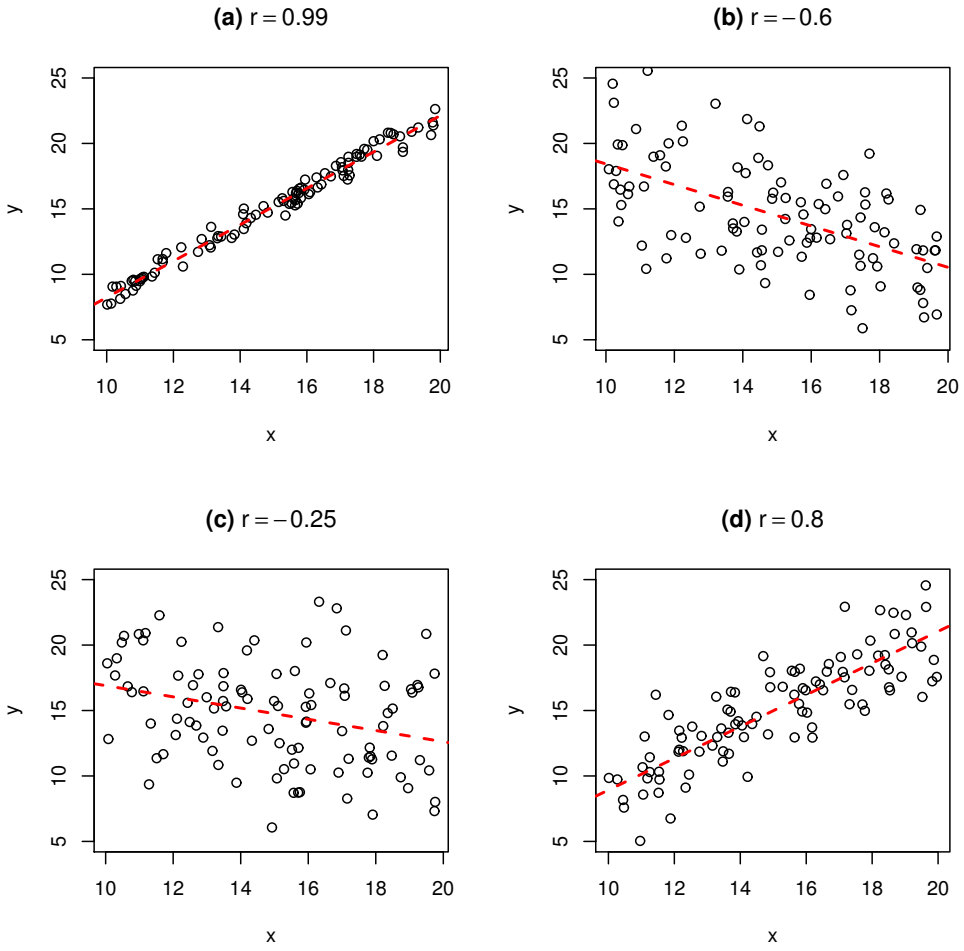

Figura 16: Nubes de puntos con distintos valores de correlación lineal.

correlación:

1. **La relación entre las variables es no lineal:** se observa que los puntos se distribuyen a lo largo de alguna figura geométrica regular distinta de una recta. En este caso lo mejor es tratar de encontrar el modelo matemático que mejor se ajusta a las observaciones. Ello puede significar utilizar, por ejemplo, regresión lineal múltiple (regresión lineal con varias variables independientes) o regresión no lineal. En la figura 17(a) vemos un ejemplo de esta situación. El coeficiente de correlación es alto (0.888), pero la nube de puntos tiene una forma claramente no lineal.

2. **Presencia de valores anómalos (outliers)**: El coeficiente de correlación debe usarse con precaución en presencia de estos valores. Gráficamente, un outlier es un punto que se aparta notoriamente del cuerpo principal de las observaciones y puede incrementar o disminuir artificialmente el valor de r. Así en la figura 17(b) vemos un caso en que hay

una nube de puntos con un ajuste lineal muy bueno. Un único valor alejado de esa nube da lugar a que la correlación sea prácticamente nula (incluso ligeramente negativa, aún cuando la tendencia de la nube de puntos es creciente). En la figura 17(c) vemos la situación contraria: una nube de puntos que no presenta asociación, y un punto aislado; la correlación global de este conjunto de puntos es, sin embargo, muy alta, 0.9.

3. **Presencia de grupos distintos de datos.** El coeficiente de correlación también debe usarse con precaución cuando las variables se miden sobre varios grupos distintos, ya que la correlación global puede llegar a diferir mucho de la correlación en cada grupo. En la imagen mostrada en la figura 17(d) se aprecia que hay dos grupos de datos, cada uno de ellos con una fuerte correlación negativa. Sin embargo, cuando la correlación se calcula globalmente para todos los puntos, sin distinguir grupos, se obtiene un valor positivo relativamente alto (0.743).

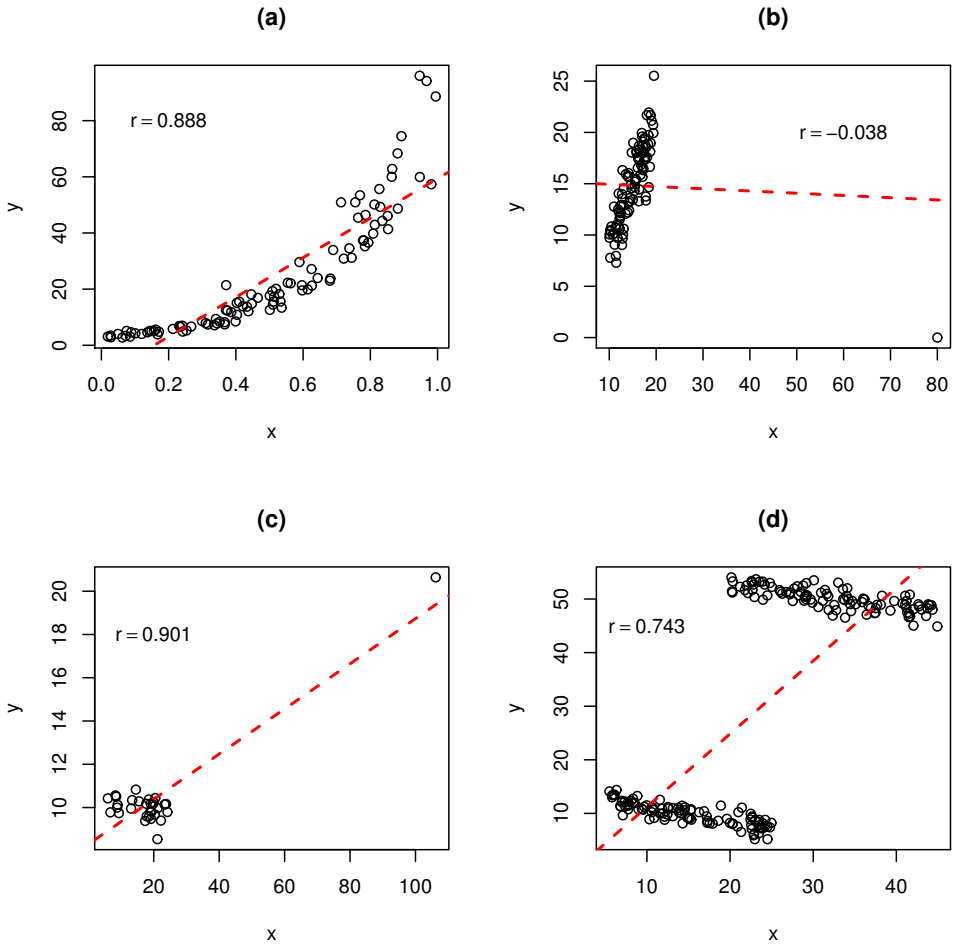

Figura 17: Diversos casos en que el coeficiente de correlación no resulta apropiado como medida de ajuste lineal.

En R la covarianza se calcula mediante la función *cov()* y la correlación mediante *cor()*. Veamos un ejemplo:

```
> cov(long, ldors)

[1] 2.686969

> cor(long, ldors)

[1] 0.7150845
```

Estas funciones pueden aplicarse a más de dos variables, en cuyo caso se obtienen las correspondientes matrices de covarianzas o correlaciones:

```
> cov(data.frame(long, ldors, lpect, peso))

            long      ldors      lpect       peso
long   15.120419  2.6869694  3.6571723  284.92959
ldors   2.686969  0.9337849  0.6619590   50.11847
lpect   3.657172  0.6619590  0.9677932   69.29353
peso  284.929587 50.1184671 69.2935278 5714.58082

> cor(data.frame(long, ldors, lpect, peso))

           long     ldors     lpect      peso
long  1.0000000 0.7150845 0.9560315 0.9693117
ldors 0.7150845 1.0000000 0.6963321 0.6860917
lpect 0.9560315 0.6963321 1.0000000 0.9317710
peso  0.9693117 0.6860917 0.9317710 1.0000000
```

Podemos calcular correlaciones y covarianzas en grupos separados de datos utilizando la función *by* de modo similar a como hemos visto ya en 6.7. La siguiente sintaxis nos proporciona la correlación entre longitud y peso para cada sexo:

```
> by(data.frame(long, peso), sexo, cor)

sexo: Hembra
          long     peso
long 1.000000 0.976949
peso 0.976949 1.000000
```

```
------------------------------------------------------
sexo: Macho
          long     peso
long 1.000000 0.958976
peso 0.958976 1.000000
```

1 Probabilidad

1. Introducción

Muchos fenómenos que habitualmente observamos en los ámbitos de la ciencia o la ingeniería se ven afectados por la presencia de una componente aleatoria[1] más o menos intensa. La presencia de esta componente da lugar a que no podamos responder con exactitud a preguntas como:

- ¿Qué cantidad de agua se va a recoger en un embalse durante el próximo invierno?
- ¿Cuánto tiempo va a durar el termo eléctrico que acabamos de instalar en casa?
- ¿Cuál va a ser el grupo sanguíneo del futuro hijo de una pareja si ambos progenitores son A-?
- ¿Cuántas tortugas nacerán de todos los huevos depositados en los nidos de una playa?
- ¿Cuántas de las personas que empiezan a fumar este año desarrollarán un cáncer de pulmón a lo largo de su vida?

Ahora bien, que no se pueda dar una respuesta exacta no significa que no pueda darse una respuesta aproximada, o incluso que no se puedan dar varias respuestas alternativas, si bien no todas con el mismo grado de certidumbre. Así, en los ejemplos anteriores:

- Si el régimen de lluvias de una región es muy estable a lo largo del tiempo, el agua recogida en inviernos anteriores nos puede dar una idea de la cantidad de agua que podemos esperar este invierno.
- Si disponemos de información de lo que han durado otros termos eléctricos de la misma marca o modelo que el que hemos adquirido, o construídos con los mismos materiales, sujetos a las mismas condiciones de uso, etc., podemos también realizar una estimación de lo que va a durar el nuestro.
- El hijo de la pareja podría ser A- ó 0-, pero seguro que no va a ser A+ ni B-.
- En playas donde anidan tortugas, y supuesto que se dan determinadas condiciones en cuanto a insolación, mareas, depredadores, etc, los estudios disponibles indican que eclosionan aproximadamente el 40 % de los huevos.

[1] *Aleatorio* significa incierto, que depende de la suerte o el azar.

- Si bien a priori no podemos saber si un individuo particular que fuma desarrollará o no cáncer de pulmón, sí sabemos que el riesgo de que lo desarrolle es del orden de 15 veces mayor que en sujetos que no fumen.

Por tanto, el hecho de que el resultado de un fenómeno aleatorio sea incierto, no quiere decir que no se pueda hacer una predicción. Ahora bien, tal predicción habrá de hacerse teniendo en cuenta nuestro grado de incertidumbre relacionado con ese fenómeno. La *probabilidad* es nuestra manera de medir la incertidumbre. Sin ser demasiado precisos por ahora con la definición de probabilidad, las respuestas a las preguntas anteriores podrían redactarse en los siguientes términos:

- En el embalse se recogerán casi seguramente (con una probabilidad del 95 %) entre 60.000 y 80.000 m^3 de agua.
- El termo durará del orden de 5 años, aunque con una probabilidad del 2 % podría durar menos de 4 y con una probabilidad del 1 % podría durar más de 7.
- Con probabilidad 90 % el hijo será A- y con probabilidad 10 % será 0-.
- En condiciones normales, con una probabilidad del 95 % se produce la eclosión de entre el 32 % y el 48 % de los huevos de tortuga de una playa. En condiciones excepcionales son muy probables tasas de eclosión de entre el 55 % y el 62 %.
- Con una probabilidad del 95 % desarrollará cáncer entre el 8 % y el 15 % de los que empiezan a fumar este año.

En este capítulo formalizaremos el concepto de probabilidad, así como sus reglas de cálculo, de tal forma que podamos disponer de herramientas que nos permitan resolver cuestiones como las aquí planteadas.

2. Objetivos

- Entender los conceptos de experimento aleatorio y suceso.
- Comprender el concepto de probabilidad y distinguir los distintos métodos de asignación de probabilidades.
- Ser capaz de calcular probabilidades de resultados de experimentos aleatorios simples, aplicando adecuadamente las propiedades de las operaciones con sucesos.
- Manejar los conceptos de sucesos dependientes e independientes, y ser capaz de identificarlos en casos prácticos.
- Entender y ser capaz de aplicar los teoremas de la probabilidad total y de Bayes.

3. Conceptos básicos

La incertidumbre es una constante en la actividad científico-técnica. La observación reiterada de un mismo fenómeno natural, aún en las mismas condiciones, produce con frecuencia valores distintos y no predecibles con exactitud. En el mucho más controlado ambiente de un laboratorio, experimentos realizados en las mismas condiciones también muestran variabilidad en sus resultados. Esta variabilidad habitualmente se atribuye al efecto del *azar.* En este contexto, el azar puede ser intrínseco al fenómeno que se estudia, tal como ocurre en el mundo cuántico, puede ser una manifestación de nuestro desconocimiento o incapacidad de medir todas las posibles causas involucradas, o puede ser la propia complejidad del fenómeno la que lo vuelve impredecible. En cualquier caso cuando a priori es imposible *predecir el resultado exacto* del fenómeno o experimento, es importante disponer al menos de *una medida del grado de certidumbre con que puede ocurrir cada uno de los resultados posibles*. Para definir una medida de esta clase será preciso introducir algunos conceptos previos:

Experimento (o fenómeno) aleatorio: Es aquel cuyo resultado es incierto y no puede predecirse de antemano con exactitud, aún cuando la experiencia o la observación se realicen en las mismas condiciones. Su opuesto sería un *experimento o fenómeno determinista*, cuyo resultado es perfectamente predecible antes de llevarlo a cabo.

Espacio muestral: se llama espacio muestral asociado a un experimento aleatorio al *conjunto de posibles resultados elementales* del experimento. Representaremos habitualmente el espacio muestral por E.

Consideraremos dos clases de espacios muestrales:

- *Discretos:* son aquellos espacios muestrales asociados a experimentos aleatorios con un conjunto finito o numerable de posibles resultados elementales. Así, los posibles resultados del lanzamiento de un dado constituyen un espacio muestral finito formado por 6 elementos, $E = \{1, 2, 3, 4, 5, 6\}$. Si nuestro experimento aleatorio consiste en contar el número de colisiones entre los átomos en el núcleo de un reactor nuclear, el espacio muestral es infinito numerable, $E = \mathbb{Z}^+ \cup \{0\}$.
- *Continuos:* son los asociados a experimentos aleatorios cuyos posibles resultados elementales constituyen un conjunto infinito no numerable, esto es, formado por intervalos continuos. Así, por ejemplo, si el experimento consiste en medir la distancia alcanzada por un lanzador de jabalina en un campo que mide 120 metros, los posibles resultados van en un rango continuo de 0 (si la jabalina cae a los pies del lanzador) a 120 metros (si la jabalina cae fuera del campo). En este caso $E = [0, 120]$

Suceso elemental: se llama así a cualquier elemento del espacio muestral (resultados más simples del experimento aleatorio).

Suceso: Un suceso es cualquier colección de sucesos elementales (esto es, cualquier subconjunto de E).

Ejemplo 1.1. Sea $E = \{1, 2, 3, 4, 5, 6\}$ el espacio muestral del experimento *"lanzar un dado"*. Entonces:

- Los sucesos *"obtener un número primo mayor que 3"* $= \{5\}$, *"obtener un 2"*$= \{2\}$ son elementales.
- Posibles sucesos no elementales son: *"obtener número par"* = $\{2, 4, 6\}$, *"obtener un número mayor que 3"*$= \{4, 5\}$, *"obtener un número menor que 10"* $= \{1, 2, 3, 4, 5, 6\}$.
- Si S es el conjunto de todos los sucesos de dicho espacio muestral, tenemos:

$$S = \{\emptyset, E, \{1\}, ..., \{6\}, ..., \{1, 3\}, \{4, 6\}, ..., \{2, 4, 6\}, \{1, 3, 5\}, \{1, 2, 3\}, \{4, 5, 6\}, ..., \{2, 3, 4, 5\}, \ldots, \{1, 2, 3, 5, 6\}, \ldots, \}$$

3.1. Sucesos especiales

Suceso seguro: Es aquel que podremos predecir que con seguridad ocurrirá al realizar el experimento aleatorio. Contendrá pues todos los sucesos elementales, por lo que coincide con el propio espacio muestral E.

Ejemplo: Al lanzar un dado al azar, el suceso seguro es *"Obtener un número del 1 a 6"* $= E$.

Suceso imposible: Es aquel que podremos predecir con seguridad que no ocurrirá. Así pues, no contendrá a ningún suceso elemental, por lo podemos representarlo como el conjunto vacío, $\emptyset$.

Ejemplo: Al lanzar un dado al azar, el suceso *"Obtener un número mayor que 6"* es un suceso imposible.

Suceso contrario: Dado un suceso A el suceso contrario, que representaremos por $\bar{A}$ ó A^c, está formado por todos los sucesos elementales que no están en A. La ocurrencia de A supone, por tanto, la no ocurrencia de $\bar{A}$, y viceversa.

Ejemplo: Al lanzar un dado al azar, si $A =$ *"Obtener un número par"*, entonces $\bar{A} =$*"Obtener número impar".*

3.2. Operaciones con sucesos

Dado que los sucesos pueden representarse como subconjuntos del espacio muestral E, las operaciones habituales con conjuntos pueden extenderse a los sucesos.

Inclusión de sucesos: Se dice que un suceso A *está incluido* en otro suceso B (es decir,$A \subset B$), si siempre que ocurre A, ocurre también B. Es decir todos los elementos de A son también elementos de B.

Ejemplo: Al lanzar un dado, sean A = *"Obtener un cinco"*, y B =*"Obtener número impar"*. Se tiene, pues, que $A = \{5\} \subset B = \{1, 3, 5\}$.

Unión de sucesos: Dados dos sucesos A y B, se llama *unión de sucesos*, al nuevo suceso $A \cup B$, que consiste en que ocurra alguno de los dos. Por tanto, $A \cup B$ es la reunión de todos los sucesos elementales de A con los sucesos elementales de B.

Ejemplo: Al lanzar un dado, sean A el suceso *"Obtener un número par"*, y B el suceso *"Obtener número mayor que tres"* $= \{4, 5, 6\}$. Entonces,$A \cup B$ es el suceso *"Obtener número par o mayor que tres"* $= \{2, 4, 5, 6\}$.

Intersección de sucesos: Dados dos sucesos A y B, se llama *intersección de sucesos* al nuevo suceso $A \cap B$, que consiste en que ocurran ambos a la vez. Por tanto, $A \cap B$ es el conjunto los sucesos elementales que están contenidos en A y en B.

Ejemplo: Al lanzar un dado, sean A = *"Obtener un número par"*, y B = *"Obtener número mayor que tres"* $= \{4, 5, 6\}$. Entonces $A \cap B$ = *"Obtener número par mayor que tres"* $= \{4, 6\}$.

Diferencia de sucesos: Dados dos sucesos A y B, se llama diferencia del suceso A menos el B, al suceso $A - B$, formado por todos los sucesos elementales de A que no estén en B.

Ejemplo: Al lanzar un dado, sean A = *"Obtener un número par"*, y B=*"Obtener número mayor que tres"* $= \{4, 5, 6\}$. Entonces, $A - B$ = *"Obtener número par no mayor que tres"* $= \{2\}$.

3.3. Incompatibilidad de sucesos

Sucesos incompatibles: Dados dos sucesos A y B, se dicen *incompatibles* si no pueden ocurrir simultáneamente. Por tanto, si A y B son incompatibles se tiene que $A \cap B = \emptyset$.

Ejemplo: Al lanzar un dado, consideremos los sucesos A = *´´Obtener un número par"* $= \{2, 4, 6\}$, y B = *"Obtener número impar"* $= \{1, 3, 5\}$. Ambos sucesos no pueden ocurrir a la vez; por tanto son incompatibles y $A \cap B = \emptyset$.

3.4. Propiedades de las operaciones con sucesos

Las siguientes propiedades de las operaciones con sucesos son análogas a las del álgebra de conjuntos:

1. $A \cup B = B \cup A$
2. $A \cap B = B \cap A$
3. $A \cup A = A$
4. $A \cap A = A$
5. $A \cup E = E$
6. $A \cap E = A$
7. $A \cup \bar{A} = E$
8. $A \cap \bar{A} = \emptyset$
9. $A \cup \emptyset = A$
10. $A \cap \emptyset = \emptyset$
11. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
12. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
13. $A - B = A \cap \bar{B}$
14. $A - B = A - (A \cap B)$
15. $\overline{(A \cup B)} = \bar{A} \cap \bar{B}$
16. $\overline{(A \cap B)} = \bar{A} \cup \bar{B}$

3.5. Cardinal de un suceso. Propiedades

El cardinal de un suceso arbitrario S, que denotaremos por $N(S)$, se define como el número de sucesos elementales contenidos en S. Dado un espacio muestral E y dos sucesos cualesquiera A y B definidos en E se cumple que:

$$N(A \cup B) = N(A) + N(B) - N(A \cap B)$$

Esta propiedad se generaliza fácilmente a tres sucesos:

$$N(A \cup B \cup C) = N(A) + N(B) + N(C) - N(A \cap B) - N(A \cap C) - N(B \cap C) + N(A \cap B \cap C)$$

La figura 1 ilustra intuitivamente el significado de esta última propiedad. Los valores mostrados en la figura indican el número de sucesos elementales en cada subconjunto. Por tanto, se tiene: $N(A) = 22$, $N(B) = 24$, $N(C) = 16$, $N(A \cap B) = 7$, $N(A \cap B) = 7$, $N(A \cap C) = 5$, $N(B \cap C) = 3$ y $N(A \cap B \cap C) = 2$; es inmediato comprobar la validez del resultado anterior.

3.6. Sistema completo de sucesos.

En un espacio muestral E, una colección de sucesos $A_1, \ldots, A_n$ forman un sistema completo si:

1. $E = A_1 \cup \ldots \cup A_n$ (ocurre con seguridad alguno de ellos)
2. $A_i \cap A_j = \emptyset$, para $i \neq j$ (incompatibilidad por pares).

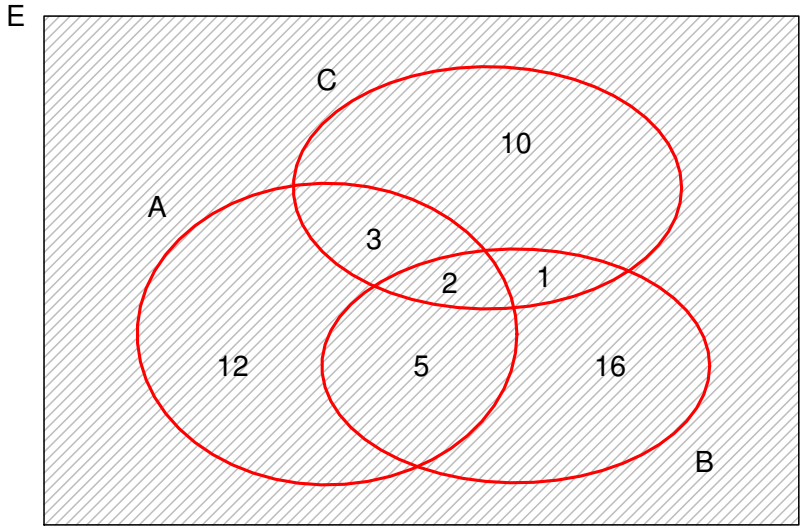

Figura 1: Representación gráfica de tres sucesos A, B y C.

3.7. Álgebra de sucesos

Para construir adecuadamente una medida de incertidumbre (*probabilidad*) sobre los posibles resultados de un experimento aleatorio, el conjunto de sucesos sobre los que se define dicha probabilidad debe tener cierta estructura mínima, que llamamos *álgebra*. Una colección de sucesos $\mathcal{F}$ pertenecientes a un espacio muestral E tiene estructura de *álgebra* si cumple las siguientes propiedades:

1. $E \in \mathcal{F}$ (Esta condición garantiza que no hay resultados que queden fuera de $\mathcal{F}$, ya que cualquiera que sea el resultado del experimento aleatorio, siempre formará parte de E).
2. Si $A \in \mathcal{F}$ entonces $\bar{A} \in \mathcal{F}$ (Esto es, si un suceso está en $\mathcal{F}$ también lo está su contrario).
3. Si $A, B \in \mathcal{F}$ entonces $A \cup B \in \mathcal{F}$ (La unión numerable de sucesos de $\mathcal{F}$ es también un suceso de $\mathcal{F}$).

Es inmediato comprobar que el conjunto S formado por *todos* los sucesos asociados a un espacio muestral E, es un álgebra.

4. Probabilidad

4.1. Definición axiomática de probabilidad

Sea E el espacio muestral asociado a un experimento aleatorio, y sea $\mathcal{F}$ un álgebra de sucesos en E. Una función P es una *medida de probabilidad* sobre $\mathcal{F}$ si cumple los *axiomas de Kolmogórov:*

1. Es una función definida para todos los elementos $A \in \mathcal{F}$, y toma valores en el intervalo $[0, 1]$:

$$\begin{aligned} P: \quad & \mathcal{F} \longrightarrow [0,1] \\ & A \hookrightarrow P(A) \end{aligned}$$

2. El suceso seguro tiene probabilidad 1: $P(E) = 1$
3. Si A y B son dos sucesos incompatibles ($A \cap B = \emptyset$), entonces:

$$P(A \cup B) = P(A) + P(B)$$

La terna $(E, \mathcal{F}, P)$ recibe el nombre de *espacio de probabilidad.* Está formada por el espacio muestral E, un álgebra de sucesos $\mathcal{F}$ definido sobre E, y una medida de probabilidad P definida sobre $\mathcal{F}$.

De la definición de probabilidad pueden deducirse las siguientes propiedades:

1. El suceso imposible tiene probabilidad 0: $P(\emptyset) = 0$
2. Para cualesquiera dos sucesos A y B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

3. Si $A_1, A_2, \ldots, A_n$ son n sucesos incompatibles dos a dos (es decir, $A_i \cap A_j = \emptyset$, con $i \neq j$), entonces:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

4. $P(\bar{A}) = 1 - P(A)$
5. Si $A \subset B \Rightarrow P(B - A) = P(B) - P(A)$

4.2. Asignación de probabilidades a sucesos de un espacio muestral.

La definición formal de probabilidad nos indica qué propiedades debe tener una función para que intuitivamente podamos interpretarla como una medida de incertidumbre. Así el suceso seguro tiene probabilidad 1; el suceso imposible tiene probabilidad 0; el valor de probabilidad se incrementa a medida que el suceso considerado contiene más sucesos elementales. Sin embargo, esta definición formal no nos dice nada respecto a cómo deben asignarse probabilidades a sucesos elementales. Esta asignación puede fundamentarse en alguno de los siguientes criterios.

4.3. Asignación exacta (Regla de Laplace)

Esta asignación es posible en aquellos casos en que el espacio muestral es finito y consideraciones teóricas sobre el mismo nos permiten concluir que los sucesos elementales que lo forman son equiprobables. En tal caso, si el espacio muestral E está formado por n elementos $\omega_1, \omega_2, \ldots, \omega_n$, por ser equiprobables se tiene que $P(\omega_1) = P(\omega_2) = \ldots = P(\omega_n) = p$. Además, de acuerdo con la propiedad 3 vista en la sección anterior:

$$P(E) = P(\{\omega_1, \omega_2, \ldots, \omega_n\}) = P(\omega_1) + P(\omega_2) + \ldots + P(\omega_n) =$$
$$= p + p + \ldots + p = np$$

y como $p(E) = 1$, resulta que la probabilidad de que ocurra un suceso elemental arbitrario es $p = \frac{1}{n}$.

Asimismo, si un suceso A está compuesto por k sucesos elementales del espacio muestral, $A = \{\omega_{i_1}, \omega_{i_2}, \ldots \omega_{i_k}\}$ su probabilidad es:

$$P(A) = P(\{\omega_{i_1}, \omega_{i_2}, \ldots \omega_{i_k}\}) = P(\omega_{i_1}) + P(\omega_{i_2}) + \ldots + P(\omega_{i_k}) =$$
$$= \frac{1}{n} + \frac{1}{n} + \ldots + \frac{1}{n} = \frac{k}{n} = \frac{N(A)}{N(E)}$$

Esta última expresión es la que se conoce como *regla de Laplace* y suele expresarse también como:

$$P(A) = \frac{N(A)}{N(E)} = \frac{\text{Casos favorables a } A}{\text{Casos posibles}}$$

Ejemplo 1.2. Sea $E = \{1, 2, 3, 4, 5, 6\}$ el espacio muestral que se obtiene al realizar el experimento aleatorio *"Lanzar un dado"*. Se tiene que:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

$$P(\text{Obtener múltiplo de 3}) = P(\{3, 6\}) = \frac{2}{6} = \frac{1}{3}$$

$$P(\text{Obtener número par}) = P(\{2, 4, 6\}) = \frac{3}{6} = \frac{1}{2}$$

4.4. Asignación mediante Frecuencias Relativas

En muchas ocasiones el fenómeno, o experimento, de interés no es tan simple como para poder identificar de forma sencilla unos sucesos elementales equiprobables. Sin embargo si es posible observar el fenómeno -o realizar el experimento- repetidas veces en igualdad de

condiciones, podemos asignar como probabilidad de cada suceso A la frecuencia relativa (proporción de veces) con que ocurre el mismo.

Esta definición sólo tiene sentido si la frecuencia relativa con que ocurre un suceso tiende a estabilizarse a medida que el experimento aleatorio se realiza más y más veces, (se probará más adelante que ésto es de hecho lo que ocurre si el experimento realmente se realiza siempre en igualdad de condiciones). En tal caso podemos correctamente definir la probabilidad de un suceso A como:

$$\mathrm{P(A)} = \lim_{n\to\infty} \frac{n_A}{n}$$

donde n es el número de veces que se realiza el experimento y n_A el número de veces que el resultado del experimento ha sido el suceso A.

Ejemplo 1.3. En una pista de bolos se colocan 11 casillas alineadas. Se lanza una pelota pequeña; ¿cuál es la probabilidad de acertar en la casilla central?. Si lanza un tirador inexperto, la pelota podría caer por igual en cualquier casilla. Tras muchos lanzamientos podríamos esperar que acierte un número similar de veces en todas las casillas. Sus resultados podría representarse mediante un diagrama de barras como el de la figura 2(a), donde cada barra representa el número de veces que la pelota ha caído en esa casilla. Sin embargo, si lanza un tirador experto, lo más probable es que la mayoría de las tiradas se acerquen más al centro, y sería de esperar un diagrama de barras como el de la figura 2(b). En ambos casos, la probabilidad de acertar en una casilla determinada se puede calcular como el límite de la frecuencia relativa con que se acierta en esa casilla a medida que el número de tiradas va aumentando. En el primer caso la distribución de frecuencias (y por tanto de probabilidad) tiende a ser uniforme (igualmente repartida entre todas las casillas), mientras que en el segundo caso tiene una forma acampanada (más probabilidad en el centro que en los extremos).

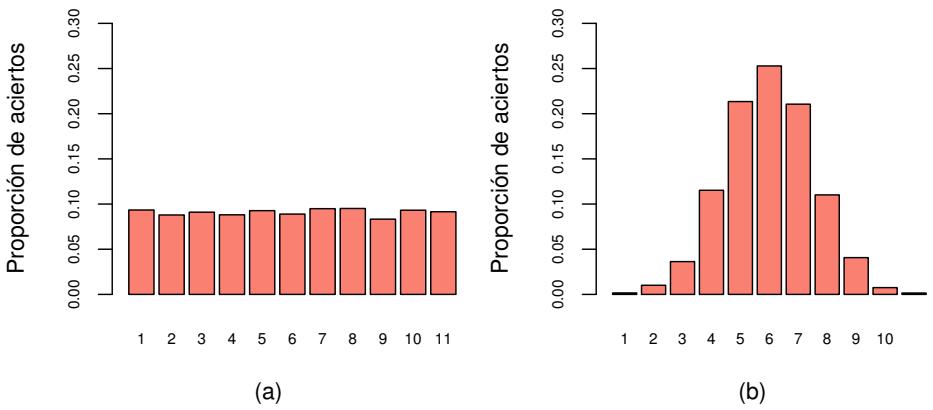

Figura 2: Frecuencias relativas de los resultados del experimento descrito en el ejemplo 1.3

4.5. Asignación subjetiva

En ocasiones no puede aplicarse ninguno de los métodos anteriores para la asignación de probabilidades. ¿Cómo podríamos calcular la probabilidad de que se construyan reactores nucleares de fusión comerciales durante la próxima década? ¿O la probabilidad de que encontremos vida en Marte? ¿O la probabilidad de que un nuevo negocio resulte rentable antes de un año? Es evidente que en estos casos no cabe hablar de modelos teóricos como en la asignación de Laplace, ni es posible tampoco realizar experimentos en igualdad de condiciones para determinar las probabilidades como frecuencias relativas.

La *asignación subjetiva* consiste en asignar probabilidades a sucesos basándonos en nuestro grado (subjetivo) de creencia en la ocurrencia de tales sucesos. Este criterio subjetivo se basa frecuentemente en nuestro conocimiento del fenómeno o en la información que tenemos sobre él. Un médico, por ejemplo, puede juzgar que la probabilidad de que un paciente se cure con cierto tratamiento es 0.85; un economista puede juzgar que la probabilidad de que un negocio quiebre es 0.15. En cualquier caso, hay que señalar que la asignación subjetiva no puede realizarse de manera arbitraria, sino de forma racional y consistente con los axiomas de Kolmogórov.

5. Probabilidad condicionada

Dos sucesos A y B están *asociados* cuando la ocurrencia o no de A afecta a la probabilidad de ocurrencia de B.

Ejemplo 1.4. Al lanzar un dado, consideremos los sucesos $A =$ *"Obtener número Par"* y $B =$*"Obtener un número mayor que 3"*. Tenemos, por tanto, $A = \{2, 4, 6\}$ y $B = \{4, 5, 6\}$

- En ausencia de otra información, la probabilidad de que ocurra el suceso B es

$$P(B) = \frac{N(B)}{N} = \frac{3}{6} = \frac{1}{2}$$

- Si se sabe que ha ocurrido A, y por tanto que ha salido par, la probabilidad de que ocurra B es:

$$\begin{aligned} P(B|A) &= \frac{N\,(\text{casos favorables a B sabiendo que ha ocurrido } A)}{N\,(\text{casos posibles sabiendo que ha ocurrido } A)} \\ &= \frac{N(\{4,6\})}{N(\{2,4,6\})} = \frac{2}{3} \end{aligned}$$

  Nótese como la probabilidad de B ha cambiado cuando se sabe que ha ocurrido A. Por tanto A y B están asociados.

Observemos con algo más de detalle cómo hemos calculado esta probabilidad condicionad

$$P(B|A) = \frac{N(\text{casos favorables a B sabiendo que ha ocurrido } A)}{N(\text{casos posibles sabiendo que ha ocurrido } A)}$$
$$= \frac{N(\{4,6\})}{N(\{2,4,6\})} = \frac{N(B \cap A)}{N(A)} = \frac{N(B \cap A)/N(E)}{N(A)/N(E)} = \frac{P(B \cap A)}{P(A)}$$

Apoyándonos en esta idea, se define la *probabilidad condicionada* de que ocurra un suceso B, dado que ha ocurrido otro suceso A, como:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Esta expresión viene a ser equivalente a calcular la probabilidad de B cuando el espacio muestral queda reducido sólo al suceso A, que es de hecho la condición que se ha producido. Nótese que de esta forma la probabilidad condicionada actúa como una medida de probabilidad, verificando:

1. $P(E|A) = 1$
2. Si $B \cap C = \emptyset$ entonces $P(B \cup C|A) = P(B|A) + P(C|A)$

5.1. Dependencia e independencia de sucesos

Un suceso B se dice independiente de otro suceso A si la probabilidad de B no cambia cuando se sabe que ha ocurrido A, esto es, si:

$$P(B) = P(B/A)$$

Como consecuencia de esta definición se sigue inmediatamente que si B es independiente de A, entonces:

1. $P(A \cap B) = P(A) \cdot P(B)$
2. $P(A) = P(A/B)$, es decir, A es independiente de B

En general, si $A_1, A_2, \ldots, A_n$ son sucesos mutuamente independientes, de la primera propiedad anterior se sigue que:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\,P(A_2)\ldots P(A_n)$$

Ejercicio 1.1. Demostrar las propiedades 1 y 2 anteriores.

Ejemplo 1.5. Consideremos el experimento consistente en extraer dos cartas de una baraja española y sean los sucesos:

A= *Obtener un oro en la primera extracción.*

B= *Obtener un oro en la segunda extracción.*

Calcular la probabilidad de que ocurra B sabiendo que ha ocurrido A.

1. Si después de observar la primera carta, ésta no se repone al mazo de cartas, ambos sucesos son dependientes y:

$$P(A) = \frac{10}{40}; \qquad P\left(B\,|A\right) = \frac{9}{39}$$

2. Si tras observar la primera carta ésta se repone al mazo, y a continuación se extrae la segunda carta, ambos sucesos son independientes y:

$$P(A) = \frac{10}{40}; \qquad P\left(B\,|A\right) = \frac{10}{40} = P(B)$$

Por último señalemos que si dos sucesos A_1 y A_2 no son independientes, la probabilidad de su intersección puede calcularse, a partir de la definición de probabilidad condicionada, como:

$$P\left(A_1 \cap A_2\right) = P\left(A_2\,|A_1\right) P\left(A_1\right)$$

Para más de dos sucesos no independientes, la aplicación reiterada de la propiedad anterior conduce a:

$$P\left(A_1 \cap A_2 \cap \ldots \cap A_n\right) = P\left(A_n\,|A_1 \cap A_2 \cap \ldots \cap A_{n-1}\right) P\left(A_1 \cap A_2 \cap \ldots \cap A_{n-1}\right) =$$

$$= P\left(A_n\,|A_1 \cap A_2 \cap \ldots \cap A_{n-1}\right) \left(A_{n-1}\,|A_1 \cap A_2 \cap \ldots \cap A_{n-2}\right) P\left(A_1 \cap A_2 \cap \ldots \cap A_{n-2}\right) =$$

$$\ldots \quad \ldots \quad \ldots$$

$$= P\left(A_n\,|A_1 \cap A_2 \cap \ldots \cap A_{n-1}\right) \left(A_{n-1}\,|A_1 \cap A_2 \cap \ldots \cap A_{n-2}\right) \ldots\ P\left(A_2\,|A_1\right) P\left(A_1\right)$$

5.2. Teoremas de la probabilidad total y de Bayes

Los siguientes resultados son de especial interés para resolver problemas relacionados con las probabilidades condicionadas.

Teorema de la Probabilidad Total: *Sea $A_1, A_2, \ldots, A_n$ un sistema completo de sucesos y sea B un suceso arbitrario. Se tiene entonces que:*

$$P(B) = \sum_{i=1}^{n} P\left(B\,|A_i\right) P\left(A_i\right)$$

Demostración.

$$\begin{aligned} P(B) = P(B \cap E) = P(B \cap (A_1 \cup A_2 \cup ... \cup A_n)) = \\ = P((B \cap A_1) \cup (B \cap A_2) \cup ... \cup (B \cap A_n)) = \\ = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)\, P(A_i) \end{aligned}$$

□

Ejemplo 1.6. Una marca de automóviles fabrica tres tipos de coches A_1, A_2 y A_3, con una proporción de cada tipo de $4/10$, $5/10$ y $1/10$ respectivamente. Además la probabilidad de que un coche de tipo A_1 se averíe durante el primer año es $0{,}07$, la de que se averíe uno del tipo A_2 es $0{,}04$ y del tipo A_3 es $0{,}09$. ¿Cuál es la probabilidad de que ocurra el suceso $B =$ *"Un coche producido en esa fábrica tenga una avería antes de un año"*?

El espacio muestral E es la producción total de la marca y por tanto $E = A_1 \cup A_2 \cup A_3$. Entonces:

$$\begin{aligned} P(B) &= P(B \cap E) = P(B \cap (A_1 \cup A_2 \cup A_3)) = \\ &= P((B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3)) = \\ &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) = \\ &= P(B/A_1)\, P(A_1) + P(B/A_2)\, P(A_2) + P(B/A_3)\, P(A_3) = \\ &= 0{,}07 \cdot \frac{4}{10} + 0{,}04 \cdot \frac{5}{10} + 0{,}09 \cdot \frac{1}{10} = 0{,}057 \end{aligned}$$

En muchas ocasiones se dispone de una descomposición del espacio muestral en un sistema completo de sucesos $A_1, A_2, \ldots, A_n$, cuyas probabilidades $P(A_i)$ se conocen, en principio, para todos los A_i (*Probabilidades a priori*). En determinadas situaciones los A_i no son directamente observables y nos interesa calcular la probabilidad de que haya ocurrido concretamente el suceso A_j. Si es posible realizar un experimento que produzca un resultado B, cuyas probabilidades condicionadas $P(B/A_i)$ (*verosimilitudes*) se conocen para todos los A_i, entonces el siguiente teorema permite usar la información aportada por B para calcular la probabilidad de que haya ocurrido A_j, esto es, la probabilidad $P(A_j/B)$ (*probabilidad a posteriori*).

Teorema de Bayes: *Sea $A_1, A_2, \ldots, A_n$ un sistema completo de sucesos y sea B un suceso tal que $B \cap A_j \neq \emptyset$. Entonces:*

$$P(A_j|B) = \frac{P(B|A_j)\, P(A_j)}{\sum_{i=1}^{n} P(B|A_i)\, P(A_i)}$$

Demostración:

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} == \frac{P(B \cap A_j)}{P(B)} = \frac{P(B\,|A_j\,) \cdot P(A_j)}{\sum_{i=1}^{n} P(B\,|A_i\,) \cdot P(A_i)}$$

Ejemplo 1.7. Supongamos que en cierta máquina está sujeta a tres tipos de fallo: lógico, eléctrico y mecánico. Por la experiencia se sabe que el 20 % de los fallos son lógicos, el 50 % eléctricos y el 30 % mecánicos. Asimismo, se sabe también que la probabilidad de recuperación inmediata de la máquina después de un fallo lógico es del 95 %; después de uno eléctrico es del 50 %; y después de uno mecánico es del 25 %. Si encontramos que la máquina acaba de recuperarse de un fallo, ¿cuál es la probabilidad de que ese fallo haya sido eléctrico?

De acuerdo con los datos proporcionados:

$$\begin{aligned} P(\text{fallo lógico}) = 0{,}20 \quad & P(\text{Recuperación}/\text{Fallo lógico}) = 0{,}95 \\ P(\text{fallo eléctrico}) = 0{,}50 \quad & P(\text{Recuperación}/\text{Fallo eléctrico}) = 0{,}50 \\ P(\text{fallo mecánico}) = 0{,}30 \quad & P(\text{Recuperación}/\text{Fallo mecánico}) = 0{,}25 \end{aligned}$$

Queremos calcular $P(\text{Fallo Eléctrico}/\text{Recuperación})$. Aplicando el teorema de Bayes:

$$\begin{aligned} P(FE/R) &= \frac{P(R/FE)P(FE)}{P(R/FL)P(FL) + P(R/FE)P(FE) + P(R/FM)P(FM)} = \\ &= \frac{0{,}5 \cdot 0{,}5}{0{,}95 \cdot 0{,}20 + 0{,}5 \cdot 0{,}5 + 0{,}25 \cdot 0{,}30} = 0{,}485 \end{aligned}$$

6. Combinatoria

La combinatoria estudia y cuenta las diferentes formas en que se puede realizar la ordenación o agrupamiento de un determinado número de objetos siguiendo ciertas condiciones. Estos recuentos están íntimamente relacionados con el cálculo de probabilidades, pues son los que permiten determinar en muchos casos el número de casos favorables y de casos posibles asociados a la ocurrencia de sucesos de interés.

6.1. Factorial

Sea n un número natural. Se define el *factorial* de n como el resultado de multiplicar sucesivamente ese número por todos los que le preceden hasta llegar a uno, esto es:

$$n! = n \cdot (n-1) \cdot (n-2) \ldots 3 \cdot 2 \cdot 1$$

Una propiedad inmediata del factorial es que $n! = n \cdot (n-1)!$

Nota: Muchas veces, como veremos, al calcular números combinatorios nos aparece $0!$ ¿Cuánto vale $0!$? Con la definición que hemos dado no tiene sentido calcular el factorial de cero, ya que al ser menor que 1 no puede multiplicarse por los que le preceden hasta llegar a 1. Ahora bien, esta definición sí nos indica que $1! = 1$. Como de la propiedad anterior podemos deducir que $(n-1)! = \frac{n!}{n}$, si sustituimos n por 1 obtenemos $0! = \frac{1!}{1} = \frac{1}{1} = 1$. Por tanto, aunque por definición el valor $0!$ carezca de sentido, resulta razonable asumir que $0! = 1$.

6.2. Variaciones sin repetición (de n objetos tomados de r en r)

Son todas las formas de ordenar n objetos en grupos de r objetos, con $r \leq n$, sin que los objetos se repitan.

El número de estas formas es:

$$V_n^r = \frac{n!}{(n-r)!}$$

Ejemplo. ¿Cuántas claves de 4 letras distintas pueden escribirse utilizando 6 letras distintas?

$$V_6^4 = \frac{6!}{(6-4)!} = \frac{6!}{2!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = 360$$

6.3. Variaciones con repetición (de n objetos tomados de r en r)

Son todas las formas de ordenar n objetos en grupos de r objetos, pudiendo repetir objetos.

El valor de r puede ser mayor, menor o igual que n. El número de variaciones con repetición se calcula mediante:

$$VR_n^r = n^r$$

Ejemplo. ¿Cuántas claves de 12 letras pueden formarse con las letras de la palabra COMPUTER?

$$VR_8^{12} = 8^{12} = 68{,}719{,}476{,}736$$

6.4. Permutaciones (de n objetos)

Son todas las formas de ordenar n objetos sin repetirlos.

El número de permutaciones de n objetos viene dado por:

$$P_n = n!$$

De la definición es obvio que:

$$P_n = V_n^n$$

Ejemplo. ¿De cuántas formas se pueden ordenar 5 libros distintos en una estantería?

$$P_5 = 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

6.5. Permutaciones con repetición

Son todas las formas de ordenar n objetos, entre los cuales hay sólo k distintos, el primero de ellos repetido n_1 veces, el segundo n_2 veces, . . ., y el k-ésimo n_k veces, con $n_1+n_2+...+n_k = n$.

El número total de esas permutaciones viene dado por:

$$P^n_{n_1,n_2,...,n_k} = \frac{n!}{n_1!n_2!...n_k!}$$

Ejemplo. ¿Cuántas palabras distintas pueden escribirse con las letras de la palabra RELEER?

$$P^6_{2,3,1} = \frac{n!}{n_1!n_2!n_3!} = \frac{6!}{2!3!1!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1 \cdot 1} = 60$$

6.6. Combinaciones sin repetición: (de n objetos tomados de r en r)

Son todas las formas de agrupar n objetos en grupos de r objetos, $0 \leq r \leq n$, sin que importe el orden, y sin repetir objetos.

El número de combinaciones se calcula como:

$$C^r_n = \begin{pmatrix} n \\ r \end{pmatrix} = \frac{n!}{r!(n-r)!}$$

Ejemplo. ¿De cuántas formas se pueden elegir 3 representantes para el claustro de un grupo formado por 40 alumnos?

$$C^3_{40} = \frac{40!}{3!(40-3)!} = \frac{40!}{3!37!} = \frac{40 \cdot 39 \cdot 38}{3 \cdot 2 \cdot 1} = 40 \cdot 13 \cdot 19 = 9980$$

6.7. Combinaciones con repetición: (de n objetos tomados de r en r)

Son todas las formas de agrupar n objetos en grupos de r objetos, sin que importe el orden, y pudiendo repetir objetos.

El número de combinaciones con repetición se calcula como:

$$CR_n^r = C_{n+r-1}^r = \begin{pmatrix} n+r-1 \\ r \end{pmatrix}$$

Ejemplo. ¿De cuántas formas se pueden escoger 12 cartulinas de colores (pudiendo escogerse colores repetidos) en un almacén donde hay cartulinas de 20 colores distintos?

$$CR_{20}^{12} = C_{31}^{12} = \begin{pmatrix} 31 \\ 12 \end{pmatrix} = \frac{31!}{12!\,(31-12)!} = \frac{31!}{12!19!} = 141\,,120\,,525$$

Capítulo 2

Variables Aleatorias

2.1. Introducción

En el capítulo anterior hemos visto ejemplos de fenómenos aleatorios en los que resulta sencillo identificar el espacio muestral y llevar a cabo una asignación de probabilidades. Sin embargo, en muchas aplicaciones ésta no es ni mucho menos una tarea inmediata. Pensemos, por ejemplo, que nuestro objetivo es caracterizar el peso que alcanzan las doradas adultas cultivadas en una piscifactoría. Para conseguir este objetivo necesitaremos un instrumento de medida –en este caso una simple balanza–, que nos dé el peso de cada pez. Es obvio que aún cuando todas las doradas hayan sido cultivadas en las mismas condiciones (misma temperatura, salinidad, alimentación, etc.), habrá diferencias en el peso final alcanzado por cada una. Pesar cada dorada es, pues, un experimento aleatorio en el sentido apuntado en el capítulo anterior: su resultado no se conoce hasta haberlo realizado.

Tras pesar muchas doradas adultas observamos que su peso oscila entre los 300 y los 600 gramos. Podemos entonces asignar como espacio muestral el intervalo [300, 600] (o quizás uno un poco mayor, por ejemplo el [200, 700], si queremos darnos un margen para incluir pesos que quizás puedan darse pero que no se han registrado durante nuestro periodo de observación). ¿Cómo realizamos ahora la asignación de probabilidades? O dicho de otra forma, ¿cómo repartimos (distribuimos) la probabilidad total (que debe ser 1) entre todos los valores de ese intervalo?. Como este intervalo (en realidad, cualquier intervalo) contiene infinitos valores, la regla de Laplace no resulta útil. La asignación mediante frecuencias relativas, todo lo más, nos permitiría asignar probabilidades a subintervalos del espacio muestral; quizás ocurre que un 5 % de las doradas observadas pesan entre 300 y 350 gramos, un 15 % pesan entre 350 y 400, un 30 % entre 400 y 450, etc. Podríamos entonces utilizar estas proporciones como aproximaciones de la probabilidad de que el peso de una dorada se encuentre en cada uno

de estos intervalos. Pero, ¿qué hacemos si queremos saber cuál es la probabilidad de que una dorada pese entre 352 y 353 gramos? Sí, podemos construir subintervalos más finos y volver a evaluar las proporciones; pero para ello necesitaremos muchos más datos experimentales que pueden ser difíciles de conseguir.

Por tanto se hace precisa una herramienta matemática que permita modelar y manejar probabilidades en situaciones como ésta. En este capítulo veremos que los conceptos de variable aleatoria y su distribución de probabilidad son la clave para alcanzar este objetivo. Estos conceptos nos proporcionarán, como veremos, una colección de modelos con la suficiente flexibilidad para adaptarse a un gran número de situaciones. Para conseguir este objetivo deberemos aprender a identificar la estructura probabilista subyacente al problema que nos ocupa; si en lugar de caracterizar el peso de las doradas de piscifactoría, nuestro objetivo fuese caracterizar el peso de las doradas salvajes, o la longitud de las lubinas, o el diámetro del opérculo de las percas, es muy posible que podamos utilizar el mismo modelo, adaptando en cada caso los parámetros de ajuste necesarios.

2.2. Objetivos

Al finalizar este capítulo el alumno deberá:

- Comprender el concepto de variable aleatoria y su función de distribución.
- Saber distinguir variables aleatorias discretas y continuas.
- Entender y saber manejar los conceptos de función de probabilidad (caso discreto) y densidad de probabilidad (caso continuo). Ser capaz de pasar de función de distribución a densidad y viceversa.
- Conocer y saber calcular las principales medidas resumen de una variable aleatoria: momentos, esperanza, varianza y cuantiles. Conocer otras medidas de forma: asimetría y apuntamiento.
- Comprender el concepto de distribución conjunta de variables aleatorias, en particular en el caso de variables independientes.
- Conocer y saber calcular medidas de asociación lineal entre variables continuas: covarianza y correlación.
- Conocer y saber aplicar la desigualdad de Chebyshev.

2.3. Concepto de variable aleatoria

Frecuentemente el resultado de un experimento aleatorio puede medirse de formas distintas, dependiendo de la finalidad con que se haya realizado el experimento. Si se lanza una moneda al aire, el resultado será cara o cruz; pero si hemos apostado 10 euros a que sale cara, desde nuestra perspectiva el resultado del lanzamiento será ganar 10 euros o perder 10 euros. En el curso de una campaña oceanográfica se escogen numerosos puntos de observación; dependiendo del tipo de sensor que se utilice, en un mismo punto se podrán medir velocidad de corriente, temperatura, salinidad, concentración de clorofila,... En un estudio sobre pesca se pueden escogen al azar varias nasas situadas en una misma zona; de cada nasa se puede medir el peso de las capturas, el número de ejemplares capturados, la proporción relativa de sujetos de distintas especies, ...

Así pues, el valor numérico obtenido en un experimento aleatorio resulta de aplicar algún instrumento de medida[1] al objeto observado. La formalización del concepto de instrumento de medida conduce a la definición de *variable aleatoria*.

Formalmente, una *variable aleatoria* es una función que a cada suceso elemental de un espacio muestral le asigna un valor numérico. Más concretamente, dado un experimento aleatorio cuyo *espacio de probabilidad*[2] asociado es $(E, \mathcal{F}, P)$, una *variable aleatoria* es una función X definida de E en $\mathbb{R}$ tal que para todo valor $x \in \mathbb{R}$ el conjunto $\{w \in E : X(w) \leq x\}$ pertenece a $\mathcal{F}$.

Ejemplo 2.1. Consideremos el experimento aleatorio consistente en lanzar dos dados equilibrados. El espacio muestral es el conjunto de parejas de valores:

$$E = \{(i, j), i, j \in \{1, 2, 3, 4, 5, 6\}\}$$

(i es el resultado del primer dado y j el del segundo). Sobre este espacio muestral definimos la variable aleatoria $X =$*"Suma de las caras superiores de los dados"*:

$$X(i, j) = i + j$$

Si consideramos el álgebra $\mathcal{F}$ de las partes de E (esto es, el conjunto de todos los conjuntos que pueden formarse con elementos de E), es obvio que para todo $x \in \mathbb{R}$ el conjunto

[1] El término *instrumento de medida* se entiende aquí en sentido amplio; puede ser un termómetro que sirve para medir temperatura, o puede ser simplemente nuestro cerebro que traduce la cara de una moneda en una ganancia de 10 euros.

[2] Recordemos del capítulo anterior que un espacio de probabilidad es una terna $(E, \mathcal{F}, P)$ donde E es el espacio muestral, $\mathcal{F}$ es un álgebra de sucesos asociados a dicho espacio y P es una probabilidad definida sobre $\mathcal{F}$.

$\{w \in E : X(w) \leq x\}$ pertenece a $\mathcal{F}$. Así, por ejemplo:

- si $x = 5$, se tiene que:
  $\{w \in E : X(w) \leq 5\} = \{(1,1),(1,2),(2,1),(2,2),(1,3),(3,1),(2,3),(3,2)\} \in \mathcal{F}$;
- si $x = 0$, $\{w \in E : X(w) \leq 0\} = \emptyset \in \mathcal{F}$;
- si $x = 17$, $\{w \in E : X(w) \leq 17\} = E \in \mathcal{F}$;
- si $x = 2{,}83$, $\{w \in E : X(w) \leq 2{,}83\} = \{(1,1)\} \in \mathcal{F}$

2.4. Función de distribución de una variable aleatoria.

La condición de que el conjunto $B_x = \{w \in E : X(w) \leq x\}$ sea un suceso perteneciente a $\mathcal{F}$ para todo $x \in \mathbb{R}$, nos asegura que tiene asignada una probabilidad, pues ésta está definida para todos los elementos de $\mathcal{F}$. La función F_X que a cada valor x le asigna la probabilidad del suceso B_x, esto es,

$$F_X(x) = P(X \leq x) = P(\{w \in E : X(w) \leq x\})$$

recibe el nombre de *función de distribución acumulativa* de la variable X. Esta función toma valores en toda la recta real y tiene por recorrido el intervalo $[0,1]$.

Ejemplo 2.2. Consideremos de nuevo el experimento aleatorio consistente en lanzar dos dados equilibrados. El resultado de la suma de sus caras superiores es un número entero entre 2 y 12. Si llamamos A_k al suceso consistente en que la suma sea k, tenemos:

$$\begin{aligned}
A_2 &= \{(1,1)\} \\
A_3 &= \{(1,2),(2,1)\} \\
A_4 &= \{(1,3),(3,1),(2,2)\} \\
A_5 &= \{(1,4),(4,1),(2,3)(3,2))\} \\
A_6 &= \{(1,5),(5,1),(2,4),(4,2),(3,3)\} \\
A_7 &= \{(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)\} \\
A_8 &= \{(2,6),(6,2),(3,5),(5,3),(4,4)\} \\
A_9 &= \{(3,6),(6,3),(4,5),(5,4)\} \\
A_{10} &= \{(4,6),(6,4),(5,5)\} \\
A_{11} &= \{(5,6),(6,5)\} \\
A_{12} &= \{(6,6)\}
\end{aligned}$$

La probabilidad de cada uno de estos sucesos puede calcularse como $P(A_k) = \frac{N(A_k)}{N(E)} = \frac{N(A_k)}{36}$. Por tanto las probabilidades de los distintos resultados son:

k	2	3	4	5	6	7	8	9	10	11	12
$P(A_k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Los sucesos B_k, consistentes en que la suma de puntos sea menor o igual que k, pueden obtenerse como:

$$B_k = \{(i,j) : i + j \leq k\} = A_2 \cup A_3 \cup \ldots \cup A_k, \quad k = 2, \ldots, 12.$$

por lo que la probabilidad de cualquiera de los B_k para $k = 2, 3, \ldots, 12$, será:

$$P(B_k) = P(A_2 \cup A_3 \cup \ldots \cup A_k) = \sum_{j=1}^{k} P(A_j) = \sum_{j=1}^{k} \frac{N(A_j)}{N(E)}$$

Si tenemos en cuenta que, obviamente, $B_x = \varnothing$ si $x < 2$ (no es posible sacar una suma menor que dos al tirar dos dados), $B_x = E$ si $x \geq 12$, y además para cualquier x real tal que $k \leq x < k + 1$ (con $k = 2, 3, \ldots 11$) se tiene que $B_x = B_k$ es inmediato construir la función de distribución de X:

$$F_X(x) = P(X \leq x) = P(B_x) = \begin{cases} 0 & x < 2 \\ 1/36 & 2 \leq x < 3 \\ 3/36 & 3 \leq x < 4 \\ 6/36 & 4 \leq x < 5 \\ 10/36 & 5 \leq x < 6 \\ 15/36 & 6 \leq x < 7 \\ 21/36 & 7 \leq x < 8 \\ 26/36 & 8 \leq x < 9 \\ 30/36 & 9 \leq x < 10 \\ 33/36 & 10 \leq x \leq 11 \\ 35/36 & 11 \leq x < 12 \\ 1 & x \geq 12 \end{cases}$$

La figura 2.1 muestra gráficamente esta función de distribución.

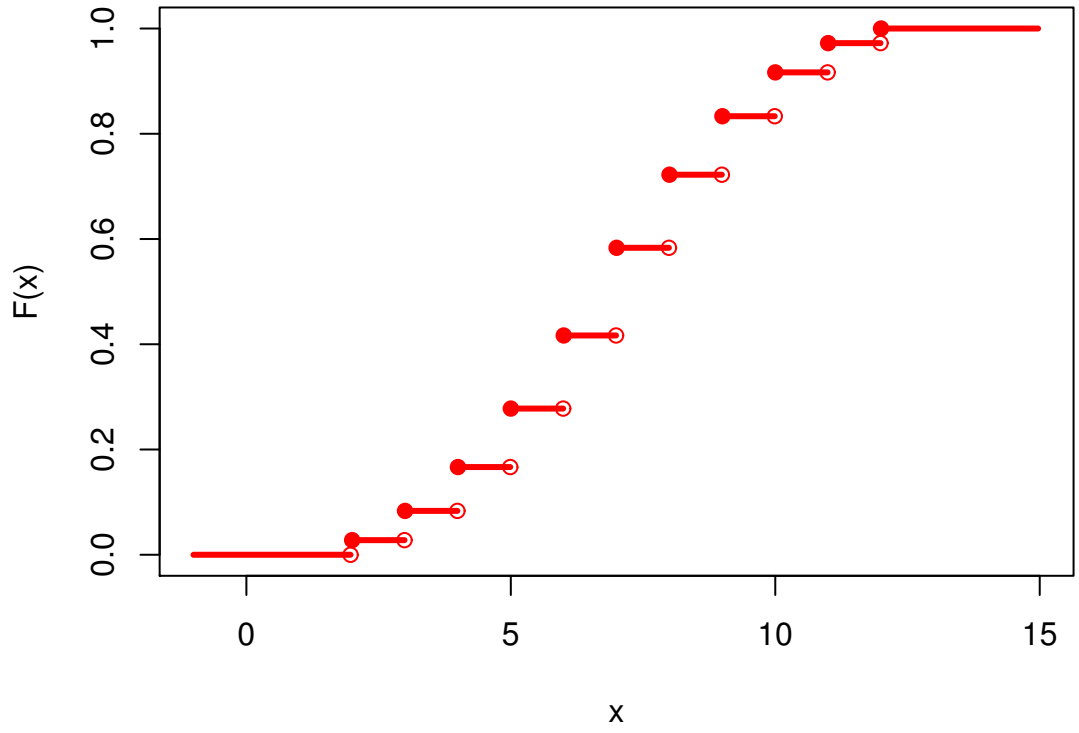

Figura 2.1: Función de distribución de la suma de caras al lanzar dos dados (ejemplo 2.2)

Propiedades de la función de distribución de una variable aleatoria.

1. $0 \leqslant F(x) \leqslant 1 \quad \forall x \in \mathbb{R}$
2. $\lim_{x \to -\infty} F_X(x) = 0$, y $\lim_{x \to \infty} F_X(x) = 1$
3. $F_X(x)$ es una función monótona no decreciente, esto es, si $a < b$ entonces $F_X(a) \leq F_X(b)$
4. Si $a < b$ entonces $P(a < X \leq b) = F_X(b) - F_X(a)$

2.5. Clasificación de variables aleatorias

Las variables aleatorias pueden clasificarse como *discretas* o *continuas*. Las primeras son aquellas que distribuyen la probabilidad sobre un conjunto finito o numerable de valores. Las variables continuas, por su parte, distribuyen la probabilidad sobre un rango continuo

de valores.

2.5.1. Variables aleatorias discretas,

Una variable aleatoria X es *discreta* cuando el conjunto de valores que puede tomar es finito o numerable. En tal caso, su distribución de probabilidad queda plenamente especificada por la *función de probabilidad* $P(X = k)$, donde k es cualquier valor que pueda tomar la variable. Obviamente se tiene que $\sum_k P(X = k) = 1$.

Ejemplo 2.3. (variable discreta con un número finito de valores). Consideremos el experimento aleatorio consistente en tirar una moneda equilibrada tres veces. Definimos la variable aleatoria $X =$*"Número de caras"*. Para este experimento el espacio muestral es

$$E = \{ccc, ccx, cxc, xcc, cxx, xcx, xxc, xxx\}$$

Los únicos valores posibles de X en este experimento son $k = 0, 1, 2, 3$. Para cada k la probabilidad $P(X = k) = P(\{w \in E : X(w) = k\})$ puede obtenerse de manera sencilla utilizando la regla de Laplace y se resume en la tabla siguiente:

k	0	1	2	3
$\mathrm{P}(X = k)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

La función de distribución acumulativa de esta variable aleatoria es:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/8 & 0 \leq x < 1 \\ 4/8 & 1 \leq x < 2 \\ 7/8 & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

La figura (2.2) muestra gráficamente las funciones de probabilidad y de distribución acumulativa de esta variable aleatoria.

Ejemplo 2.4. (variable discreta con un número infinito numerable de valores) Se realiza el experimento aleatorio consistente en lanzar sucesivas veces una moneda hasta que sale cara por primera vez. El espacio muestral asociado a este experimento es entonces $E = \{c, xc, xxc, xxxc, \ldots\}$. Si denotamos por X a la variable aleatoria *"Número de lanzamientos*

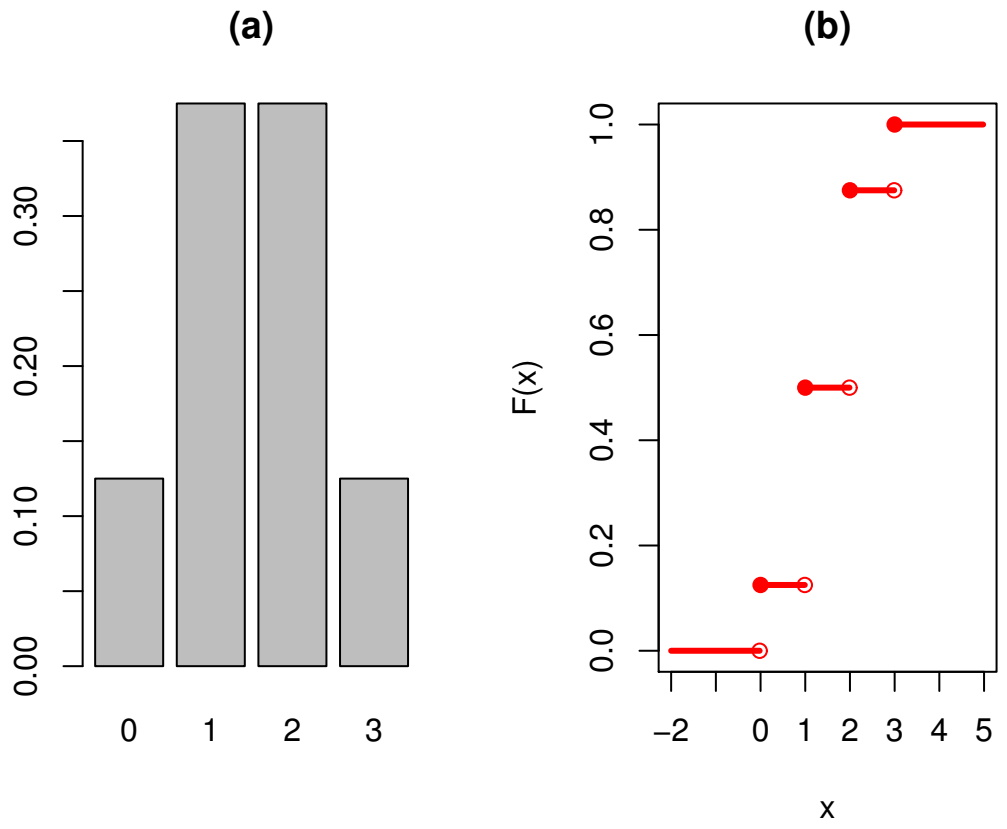

Figura 2.2: (a)Función de probabilidad y (b) Función de distribución acumulativa del número de caras en el lanzamiento de tres monedas (ejemplo 2.3)

hasta que sale cara", teniendo en cuenta que los resultados de los sucesivos lanzamientos constituyen sucesos independientes se tiene:

$$
\begin{aligned}
P(X=1) &= P(\{c\}) = \frac{1}{2} \\
P(X=2) &= P(\{xc\}) = P(\{x\} \cap \{c\}) = P(\{x\})\,P(\{c\}) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2^2} = \frac{1}{4} \\
P(X=3) &= P(\{xxc\}) = P(\{x\} \cap \{x\} \cap \{c\}) = P(\{x\})\,P(\{x\})\,P(\{c\}) = \\
&= \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2^3} = \frac{1}{8} \\
&\vdots \\
P(X=k) &= P(\{x\ldots xc\}) = P(\{x\} \cap \ldots \cap \{x\} \cap \{c\}) = \\
&= P(\{x\})^{k-1}\,P(\{c\}) = \frac{1}{2^{k-1}}\cdot\frac{1}{2} = \frac{1}{2^k} \\
&\vdots
\end{aligned}
$$

(Obsérvese que esta variable aleatoria podría tomar infinitos valores ya que, al menos en teoría, cabe la posibilidad de que en los sucesivos lanzamientos salga siempre cruz, por lo que el experimento no se detiene nunca). Por tanto la función de distribución de esta variable aleatoria, para $n = 1, 2, 3, \ldots$, viene dada por[3]:

$$F(n) = P(X \leq n) = \sum_{k=1}^{n} P(X = k) = \sum_{k=1}^{n} \frac{1}{2^k} = \frac{\frac{1}{2} - \frac{1}{2^{n+1}}}{1 - \frac{1}{2}} = 1 - \frac{1}{2^n}$$

La figura 2.3 muestra las gráficas de la función de probabilidad $P(X = k)$ y la función de distribución acumulativa $F(x)$, sólo para los valores $x \in [0, 10]$.

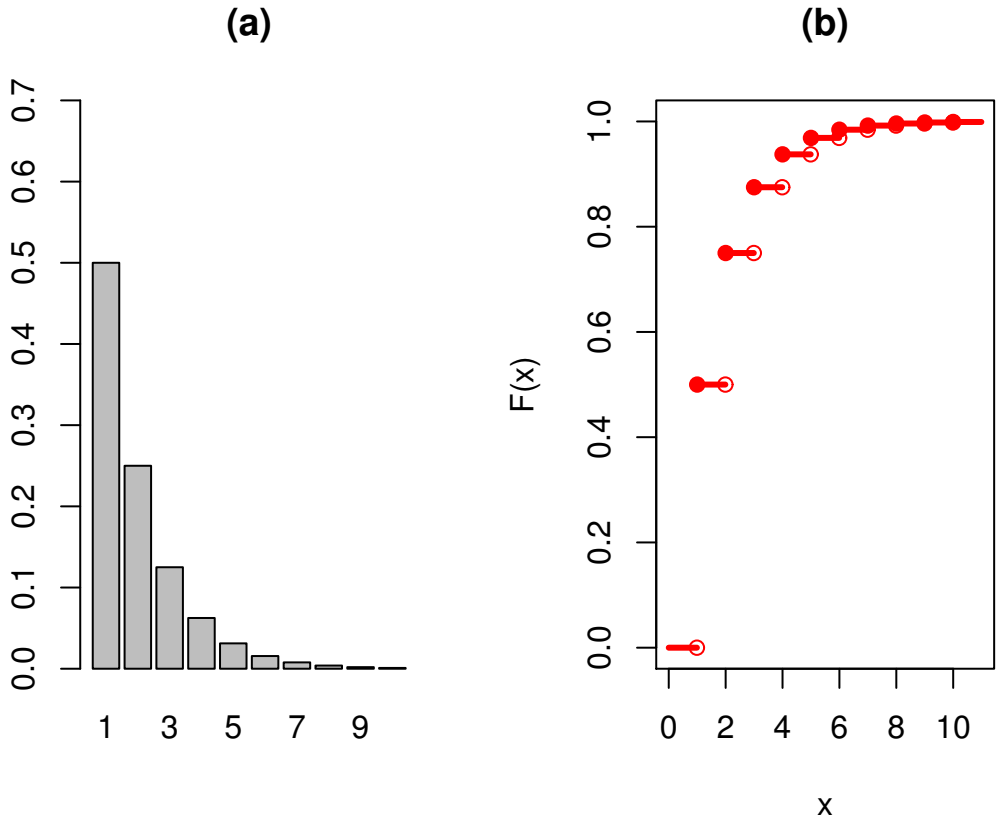

Figura 2.3: (a) Función de probabilidad y (b) Función de distribución acumulativa del número de lanzamientos de una moneda hasta que sale cara por primera vez (ejemplo 2.4).

Como hemos visto en los ejemplos 2.3 y 2.4, las variables aleatorias discretas se caracterizan por tener funciones de distribución acumulativa *escalonadas,* que se van incrementando a

[3]Es preciso utilizar que la suma de los n primeros términos de una progresión geométrica de razón menor que la unidad es $\sum_{k=1}^{n} \rho^k = \frac{1-\rho^{n+1}}{1-\rho}$

saltos. Las posiciones de los saltos corresponden a los valores que puede tomar la variable aleatoria. A su vez, la magnitud de cada salto es igual a la probabilidad de observar el valor correspondiente al punto de salto. Entre salto y salto, la función de distribución es constante.

Propiedades de la función de probabilidad de una variable aleatoria discreta

Sea $M = \{n_1, n_2, n_3, \ldots\}$ el conjunto (finito o infinito numerable) de todos los posibles valores que puede tomar una variable aleatoria discreta X. Suponemos además que $n_1 < n_2 < n_3 < \ldots$, y llamemos $f(n) = P(X = n)$. Las siguientes propiedades se siguen inmediatamente de la definición de $f(n)$:

1. $0 \leqslant f(x) \leqslant 1$ para todo $x \in \mathbb{R}$
2. $\sum_{n_j \in M} f(n_j) = 1$
3. $F(n_k) = \sum_{j \leqslant k} f(n_j)$
4. $f(n_k) = F(n_k) - F(n_{k-1})$

2.5.2. Variables aleatorias continuas.

Las variables aleatorias cuya función de distribución acumulativa es continua reciben el nombre de *variables aleatorias continuas.* Se caracterizan por tomar valores en un rango continuo (intervalo), sin que haya puntos en los que se acumule la probabilidad; dicho de otra forma, si X es una v.a. continua, $P(X = x) = 0$ para cualquier valor $x \in \mathbb{R}$.

Ejemplo 2.5. Realizamos el experimento consistente en tirar de los extremos de una cuerda de 1 metro de longitud hasta que se parte. Suponemos que la cuerda está fabricada con un material completamente homogéneo, de forma que a priori es igualmente probable que se rompa en cualquier punto. Consideremos la variable aleatoria $X =$*"Posición del punto en que se parte la cuerda".*

Dado que existen infinitos puntos entre 0 y 1 en los que la cuerda puede romperse (todos equiprobables, por ser la cuerda homogénea), la regla de Laplace nos indicaría que la probabilidad de que se rompa en un punto x concreto es 0, cualquiera que sea x:

$$P(X = x) = 0 \ \ \forall x \in [0, 1]$$

Ahora bien, si consideramos el punto medio ($x = \frac{1}{2}$), por ser la cuerda homogénea la probabilidad de que se parta a la izquierda de ese punto debe ser igual a la probabilidad de que

se parta a la derecha; por tanto $P\left(X \leq \frac{1}{2}\right) = \frac{1}{2}$. De igual forma, si consideramos el punto $x = \frac{1}{3}$, como el trozo a la izquierda de este punto mide una tercera parte de la longitud total de la cuerda, nuevamente la homogeneidad de ésta implica que $P\left(X \leq \frac{1}{3}\right) = \frac{1}{3}$. En general, el mismo argumento nos permite concluir que para cualquier $x \in [0,1]$, $P(X \leq x) = x$. Asimismo, como la cuerda no puede partirse antes de $x = 0$, se tiene $P(X < 0) = 0$; y como tampoco puede partirse después de $x = 1$, resulta $P(X \leqslant x) = 1$ para los $x > 1$.

Observemos, pues, que aunque para esta variable sea $P(X = x) = 0 \ \forall x \in [0,1]$, el razonamiento anterior nos ha permitido construir su función de distribución acumulativa $F(x) = P(X \leq x)$ para cualquier valor $x \in \mathbb{R}$:

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

La figura 2.4 muestra gráficamente esta función de distribución.

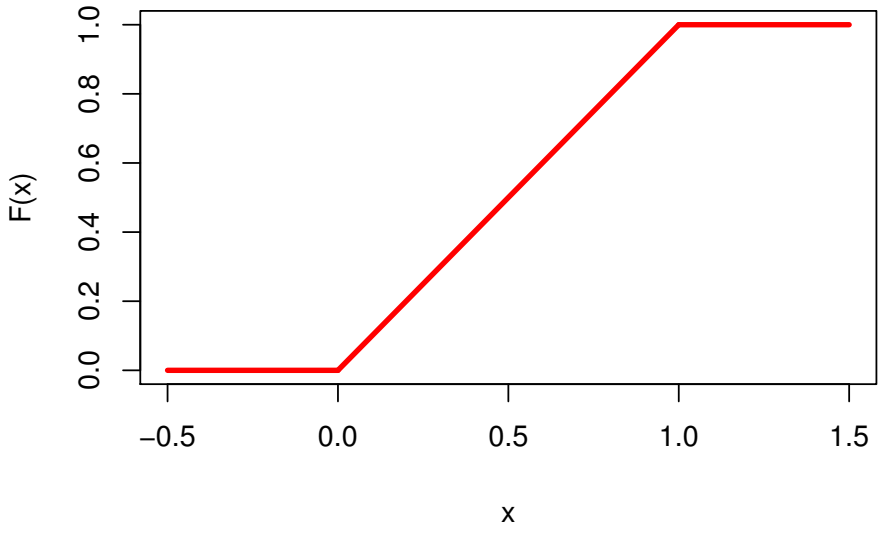

Figura 2.4: Función de distribución acumulativa descrita en el ejemplo 2.5.

Un caso particular de variables aleatorias continuas son las *absolutamente continuas,* que se caracterizan porque su función de distribución es absolutamente continua. Esto significa que existe una función real f, positiva e integrable en el conjunto de números reales, tal que la función de distribución acumulativa F se puede expresar como:

$$F(x) = \int_{-\infty}^{x} f(u)\, du \tag{2.1}$$

La función f recibe el nombre de *función de densidad de probabilidad de la variable aleatoria* X. Este nombre no es arbitrario, ya que $f(x)$ admite una interpretación análoga a la del concepto físico de densidad. En efecto de la ecuación (2.1) se sigue que $f(x)$ es la derivada de $F(x)$ y por tanto:

$$\begin{aligned} f(x) &= F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = \\ &= \lim_{\Delta x \to 0} \frac{P(X \leq x + \Delta x) - P(X \leq x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x} \end{aligned}$$

lo que nos indica que $f(x)$ representa la cantidad de probabilidad en un entorno próximo de x, dividida por la medida Δx de ese entorno. Utilizando un símil físico, $P(x \leq X \leq x + \Delta x)$ puede entenderse como la *masa* total de probabilidad que se concentra en un *volumen* Δx alrededor de x. *Masa* partido por *volumen* es precisamente la definición clásica de densidad, lo que justifica el nombre de la función f.

Asimismo, de la expresión anterior se sigue también que para un valor Δx suficientemente pequeño:

$$P(X \in (x, x + \Delta x]) \cong f(x)\Delta x$$

lo que significa que la probabilidad de que la variable aleatoria X esté dentro de un intervalo muy pequeño que contenga a un valor x es aproximadamente igual a $f(x)$ veces la amplitud de dicho intervalo. Geométricamente, el término $f(x)\Delta x$ representa el área de un rectángulo de base Δx y altura $f(x)$.

Continuación del ejemplo 2.5: Recordemos que en este ejemplo considerábamos la variable aleatoria $X =$*"punto donde se rompe una cuerda homogénea de 1 metro de longitud al tirar de sus extremos"*. La función de distribución de esta variable era de la forma:

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x > 1 \end{cases}$$

Derivando obtenemos la función de densidad :

$$f(x) = \begin{cases} 0 & x < 0 \\ 1 & 0 \leq x \leq 1 \\ 0 & x > 1 \end{cases}$$

Como vemos, esta función es constante en el intervalo $[0, 1]$, lo que se corresponde con la idea intuitiva de que, por ser la cuerda homogénea, es igualmente probable que se rompa en cualquier punto; por tanto la densidad de dicha probabilidad debe ser constante a lo largo de todo el recorrido de la cuerda.

Nota: Si bien es posible definir variables aleatorias continuas que no sean absolutamente continuas, constituyen la excepción antes que la regla. La inmensa mayoría de las variables aleatorias continuas que nos encontramos en las aplicaciones son también absolutamente continuas. Por ello, con el objetivo de simplificar la terminología, cuando en este texto utilicemos la expresión *variable aleatoria continua* nos estaremos refiriendo en realidad a variables aleatorias *absolutamente continuas*, y por tanto con función de densidad bien definida.

Propiedades de la función de densidad de probabilidad de variables aleatorias continuas.

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

La última propiedad nos indica que la probabilidad de que una variable aleatoria continua X tome valores entre dos puntos a y b coincide con el área bajo la función de densidad entre esos dos puntos.

Continuación del ejemplo 2.5: La probabilidad de que la cuerda se parta entre los puntos 0.3 y 0.7 puede calcularse como:

$$P(0{,}3 < X \leqslant 0{,}7) = \int_{0{,}3}^{0{,}7} f(x)\,dx = \int_{0{,}3}^{0{,}7} 1dx = [x]_{0{,}3}^{0{,}7} = 0{,}7 - 0{,}3 = 0{,}4$$

donde hemos tenido en cuenta que $f(x) = 1$ para $x \in [0, 1]$. La figura 2.5 muestra el significado geométrico de esta integral. La probabilidad que se ha calculado es el área bajo la función $f(x) = 1$ entre 0,3 y 0,7, que en este caso corresponde simplemente a un rectángulo.

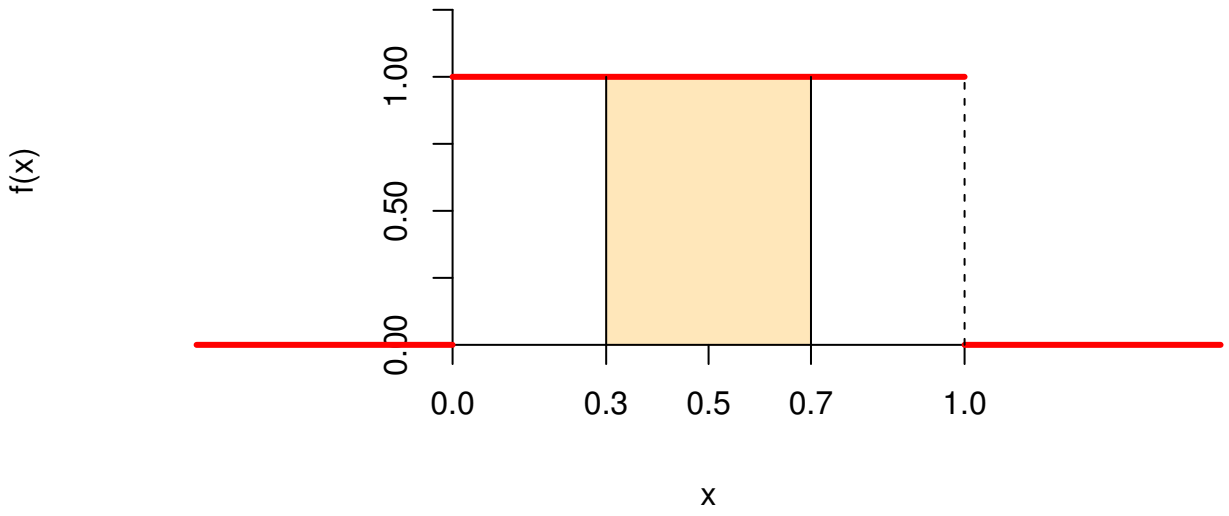

Figura 2.5: La linea de trazo grueso representa la función de densidad $f(x)$ de la variable aleatoria descrita en el ejemplo 2.5 (punto aleatorio en que se rompe una cuerda de un metro). El área coloreada representa la probabilidad de que la cuerda se rompa entre los puntos 0,3 y 0,7.

Ejemplo 2.6. En la desembocadura de muchos ríos es frecuente encontrar radioisótopos (plomo 210, cesio 137 y otros) que pueden ser utilizados como trazadores del arrastre de materiales sedimentarios. Se ha comprobado que la probabilidad de detectar uno de estos radioisótopos disminuye exponencialmente con la profundidad de muestreo en el lecho marino. En particular, en el estuario de cierto río, la variable X=*"Profundidad (en cm.) a la que es detectable la presencia de* ^{210}Pb*"* tiene como función de densidad

$$f(x) = \begin{cases} 0{,}1e^{-0{,}1x}, & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Obviamente $f(x)$ está bien definida como función de densidad, ya que $f(x) \geq 0, \forall x$ y además:

$$\int_0^\infty 0{,}1e^{-0{,}1x}dx = \left[-e^{-0{,}1x}\right]_0^\infty = 1$$

Si se desea obtener la probabilidad de detectar ^{210}Pb entre 5 y 15 cm. de profundidad calcu-

lamos simplemente:

$$P\left(5 \leq X \leq 15\right) = \int_5^{15} 0{,}1e^{-0{,}1x}dx = \left[-e^{-0{,}1x}\right]_5^{15} =$$
$$= e^{-0{,}1\cdot 5} - e^{-0{,}1\cdot 15} = 0{,}38$$

La figura 2.6 muestra la función de densidad de esta variable. La probabilidad que se acaba de calcular corresponde al área bajo esta función entre los valores 5 y 15, que se ha representado también en esta gráfica.

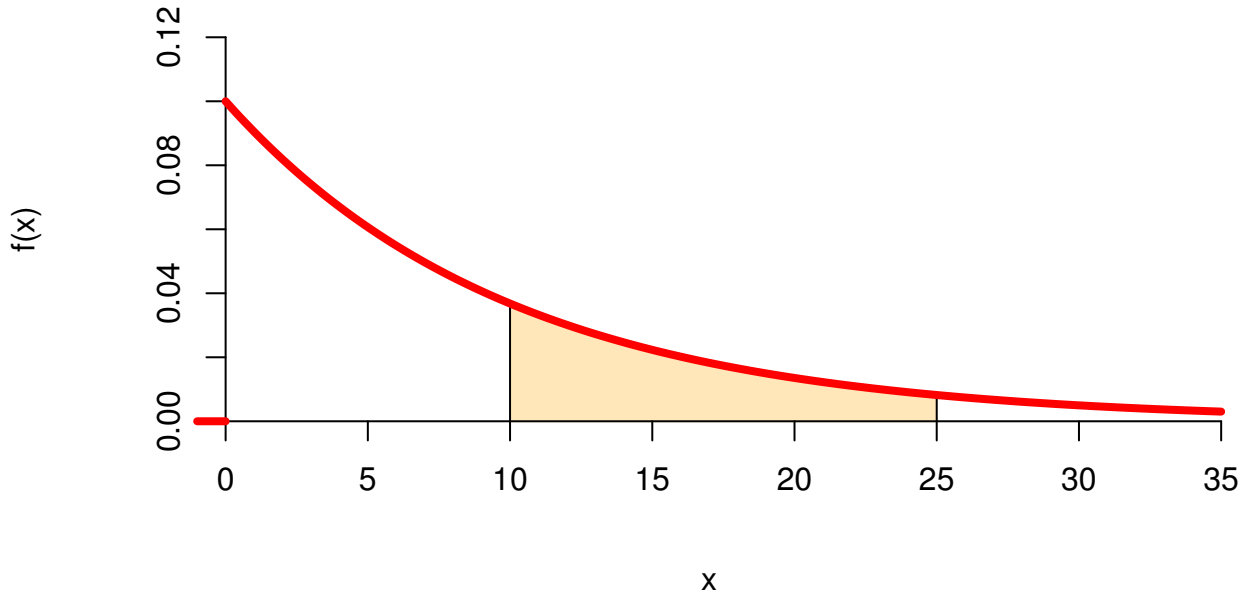

Figura 2.6: Función de densidad de la variable descrita en el ejemplo 2.6

Ejemplo 2.7. En ingeniería de costas resulta de interés modelar la distribución de probabilidad de la altura de ola. En particular es importante conocer la probabilidad de que dicha altura supere ciertos valores, ya que ello determina las características que han de tener las construcciones costeras. La función de densidad:

$$f\left(x\right) = \begin{cases} \vartheta x e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

constituye un modelo simple que puede emplearse en algunos casos. Para que esta función de densidad esté bien definida, el área total bajo la misma debe ser 1, esto es:

$$\int_0^{\infty} \vartheta x e^{-\lambda x}dx = 1$$

Resolvemos esta integral (es sencillo integrar por partes):

$$\int_0^\infty \vartheta x e^{-\lambda x} dx = \vartheta \left[-\frac{x}{\lambda} e^{-\lambda x} - \frac{1}{\lambda^2} e^{-\lambda x} \right]_0^\infty = \frac{\vartheta}{\lambda^2}$$

Por tanto, para que esta integral valga 1 deberá ocurrir que si $\vartheta = \lambda^2$, en cuyo caso $f(x)$ corresponde a una función de densidad correctamente definida cualquiera que sea el valor de λ. Supongamos que $\lambda = 0{,}9$ y que se desea calcular la probabilidad de que la altura de ola supere los 4 metros. Entonces, si X=*"Altura de ola"*:

$$\begin{aligned} P(X \geq 4) &= \int_4^\infty \vartheta x e^{-\lambda x} dx = \lambda^2 \left[-\frac{x}{\lambda} e^{-\lambda x} - \frac{1}{\lambda^2} e^{-\lambda x} \right]_4^\infty = \\ &= 0{,}9^2 \left(\frac{4}{0{,}9} e^{-0{,}9 \cdot 4} + \frac{1}{0{,}9^2} e^{-0{,}9 \cdot 4} \right) = e^{-0{,}9 \cdot 4} (0{,}9 \cdot 4 + 1) = 0{,}126 \end{aligned}$$

La figura 2.7 muestra la gráfica de esta función de densidad. La probabilidad que se acaba de calcular corresponde al área bajo esta curva desde el valor 4 en adelante.

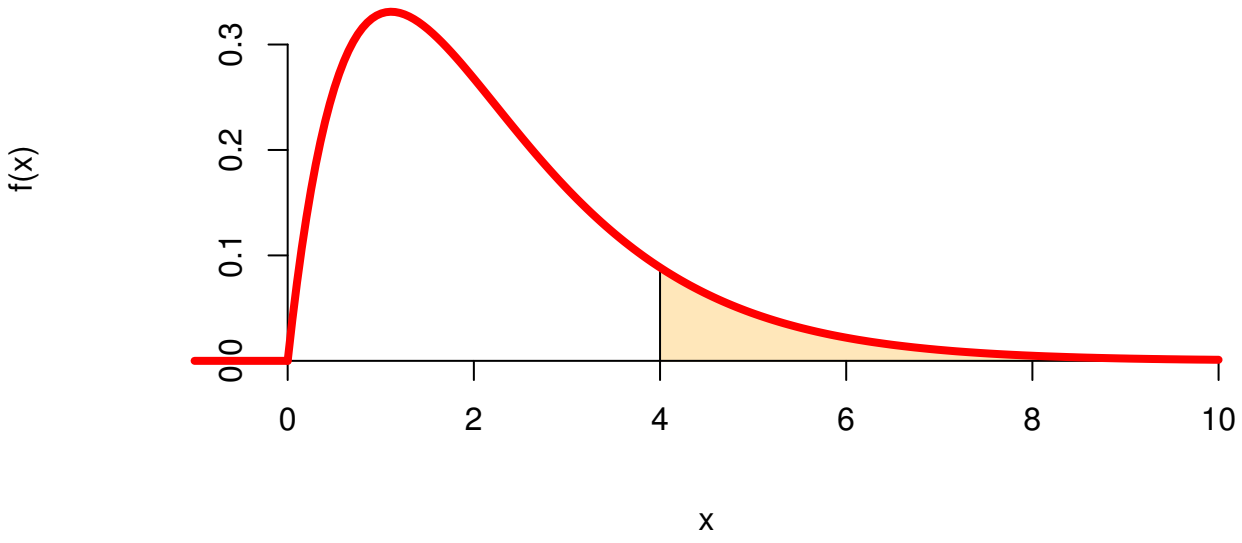

Figura 2.7: Función de densidad de la altura de ola (ejemplo 2.7). Se ha sombreado la probabilidad de que una ola supere los 4 metros.

Así pues, la función de distribución de una variable aleatoria (o sus derivadas, la función de probabilidad en el caso discreto y la función de densidad en el caso continuo) es la herramienta que permite modelar la incertidumbre presente en los procesos de observación o

experimentación. En los ejemplos que acabamos de ver –punto de rotura de una cuerda, profundidad a la que se detecta un isótopo radiactivo, altura de ola– el valor que toma la variable es impredecible *a priori*, pero las funciones de densidad de probabilidad asociadas a estas variables determinan qué rangos de valores tienen más o menos probabilidad de ocurrir. La distribución de probabilidad, pues, modela el efecto del conjunto de causas que dan origen a dichos valores. Permitiéndonos cierto abuso del lenguaje, podemos decir que la distribución de probabilidad es la que *genera* los valores que observamos en las variables aleatorias, produciendo más valores en las regiones con mayor probabilidad y menos en el resto. La figura 2.8 representa esta idea. Se han reproducido de nuevo las funciones de densidad de los últimos ejemplos, pero representando en la base de cada figura puntos correspondientes a 300 observaciones de las respectivas variables (puntos de rotura de 300 cuerdas homogéneas, altura de 300 olas, y profundidad a la que se ha detectado ^{210}Pb en 300 muestras). Como puede apreciarse, en (a) las observaciones se reparten uniformemente en el intervalo $[0, 1]$, en consonancia con una densidad de probabilidad constante; en (b) y en (c) se observa que los valores observados tienden a concentrarse en las regiones con mayor densidad de probabilidad, disminuyendo su número a medida que disminuye la densidad.

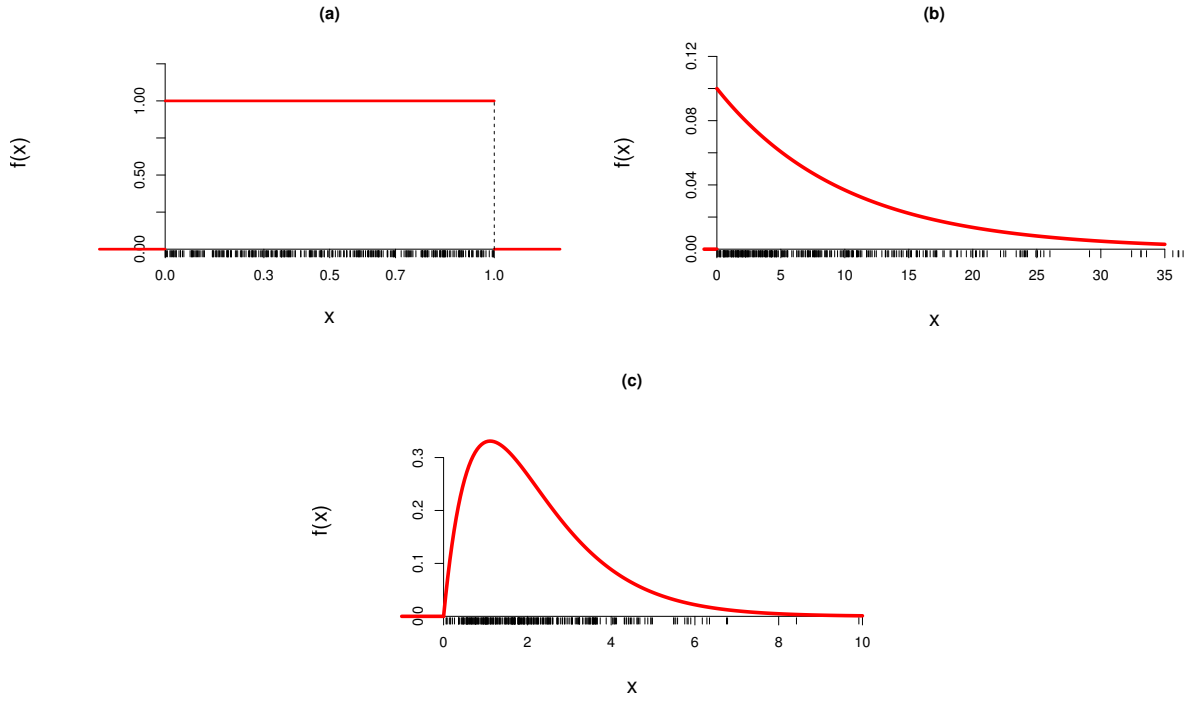

Figura 2.8: Densidades de probabilidad de las variables descritas en los ejemplos 2.5, 2.6 y 2.7. Sobre los ejes de abcisas se han representado las posiciones de 300 valores observados en estas variables.

2.5.3. Variables aleatorias mixtas.

En el caso de que la función de distribución tenga saltos, y además tramos continuos en los que sea estrictamente creciente (no constante), la variable aleatoria es *mixta*. Una variable aleatoria mixta se caracteriza, por tanto, porque toma valores en intervalos continuos, a la vez que existen uno o más valores discretos para los que $P(X = x) > 0$. En este curso no nos ocuparemos de este tipo de variables.

2.6. Variables aleatorias multidimensionales.

En muchas ocasiones se realizan múltiples medidas sobre los objetos de nuestro estudio. Así por ejemplo, en el curso de un trabajo de campo sobre tortugas marinas, en cada ejemplar podemos medir su longitud (X), peso (Y) y perímetro de la concha (Z). De esta forma, cada observación da lugar a un vector (x, y, z). Este vector es una variable aleatoria dado que a priori, antes de capturar cada ejemplar, no podemos predecir su valor. Por ello este vector recibe el nombre de *variable aleatoria multidimensional* (o *vector aleatorio*).

2.6.1. Distribución conjunta de variables aleatorias.

Los conceptos de función de distribución aumulativa, función de probabilidad y función de densidad de probabilidad se generalizan fácilmente al caso multidimensional. Por simplicidad, a continuación enunciamos estos conceptos sólo para el caso bidimensional. Dado un vector aleatorio (X, Y):

- La función $F(x, y) = P(X \leq x \cap Y \leq y)$ recibe el nombre de *función de distribución conjunta* del vector (X, Y).
- Cuando X e Y son discretas, la función $f(x, y) = P(X = x \cap Y = y)$ recibe el nombre de *función de probabilidad conjunta* del vector (X, Y).
- Cuando X e Y son continuas y existe una función $f : \mathbb{R}^2 \to \mathbb{R}$, integrable y no negativa, tal que:
$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\, dsdt$$
si dice entonces que el vector (X, Y) tiene distribución absolutamente continua. En tal caso:
$$f(x, y) = \lim_{\Delta x \to 0, \Delta y \to 0} \frac{P(\{x < X \leq x + \Delta x\} \cap \{y < Y \leq y + \Delta y\})}{\Delta x \Delta y} = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

recibe el nombre de *función de densidad de probabilidad* del vector (X, Y).

Ejemplo 2.8. (Vector de variables discretas) Supongamos que en el lanzamiento de dos dados equilibrados consideramos la variable bidimensional (X, Y), donde X=*"Producto de las caras superiores"* e Y= *"Suma de las caras superiores"*. La tabla 2.1 muestra los posibles valores de la variable (X, Y), así como los sucesos que los generan y su probabilidad. La figura 2.9 representa la función de probabilidad de esta variable aleatoria.

Suceso	(X, Y)	*Probabilidad*
$\{(1,1)\}$	$(1,2)$	$1/36$
$\{(1,2),(2,1)\}$	$(2,3)$	$2/36$
$\{(1,3),(3,1)\}$	$(3,4)$	$2/36$
$\{(2,2)\}$	$(4,4)$	$1/36$
$\{(1,4),(4,1)\}$	$(4,5)$	$2/36$
$\{(2,3),(3,2)\}$	$(6,5)$	$2/36$
$\{(1,5),(5,1)\}$	$(5,6)$	$2/36$
$\{(2,4),(4,2)\}$	$(8,6)$	$2/36$
$\{(3,3)\}$	$(9,6)$	$1/36$
$\{(1,6),(6,1)\}$	$(6,7)$	$2/36$
$\{(2,5),(5,2)\}$	$(10,7)$	$2/36$

Suceso	(X, Y)	*Probabilidad*
$\{(3,4),(4,3)\}$	$(12,7)$	$2/36$
$\{(2,6),(6,2)\}$	$(12,8)$	$2/36$
$\{(3,5),(5,3)\}$	$(15,8)$	$2/36$
$\{(4,4)\}$	$(16,8)$	$1/36$
$\{(3,6),(6,3)\}$	$(18,9)$	$2/36$
$\{(4,5),(5,4)\}$	$(20,9)$	$2/36$
$\{(4,6),(6,4)\}$	$(24,10)$	$2/36$
$\{(5,5)\}$	$(25,10)$	$1/36$
$\{(5,6),(6,5)\}$	$(30,11)$	$2/36$
$\{(6,6)\}$	$(36,12)$	$1/36$

Tabla 2.1: Función de probabilidad de la variable (X, Y) descrita en el ejemplo 2.8 (X=*"Producto de las caras superiores resultantes de lanzar dos dados"* e Y= *"Suma de las caras superiores"*).

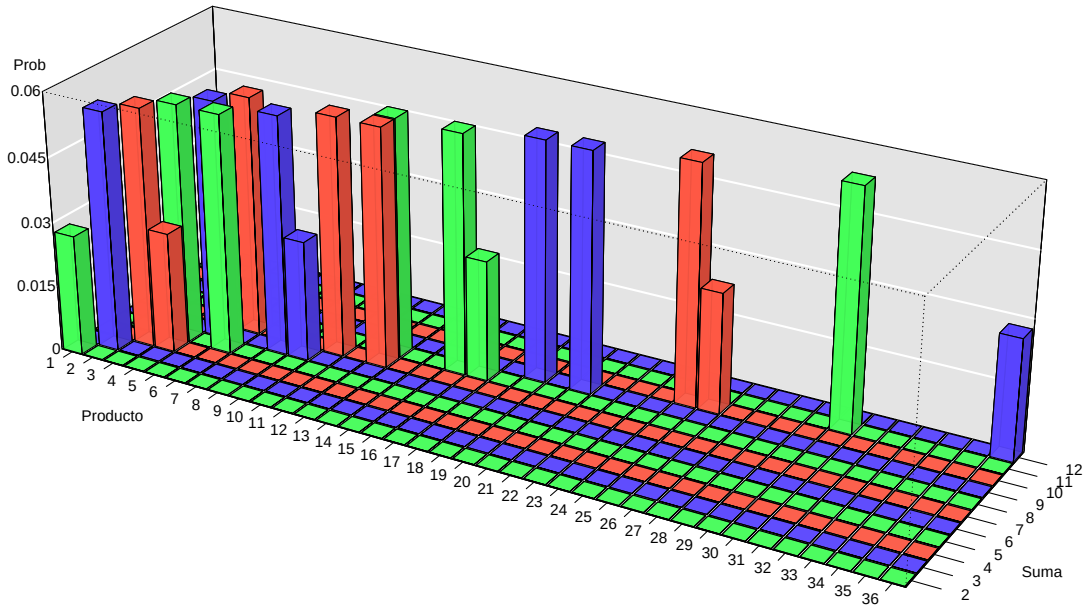

Figura 2.9: Representación gráfica de la función de probabilidad de la variable (X, Y) descrita en el ejemplo 2.8 (Tabla 2.1).

Ejemplo 2.9. (Vector de variables continuas) Un estudio morfométrico de peces de la familia de los *Serránidos,* subfamilia *Epinephelinae* ha permitido obtener una aproximación de la función de densidad conjunta $f(x, y)$ del vector aleatorio (X, Y) siendo X=*"Longitud (cm)"* e Y=*"Peso (kg)"* de los ejemplares de esta familia[4]. Esta aproximación se muestra en la figura 2.10. Del mismo modo que el área entre dos puntos bajo la función de densidad de una variable aleatoria unidimensional da la probabilidad de que la variable tome valores en ese rango, el *volumen* bajo la función de densidad bivariante sobre un entorno determinado da la probabilidad de que el vector aleatorio (X, Y) tome valores en dicho entorno.

La figura 2.11 muestra un conjunto de 1000 observaciones de *(Longitud, Peso)* que obedecen a esta distribución de probabilidad. Como puede apreciarse, donde la densidad de probabilidad encierra un mayor volumen (mayor probabilidad) se produce un mayor número de observaciones, disminuyendo este número a medida que disminuye el volumen; donde la densidad es cero (probabilidad nula), no se producen observaciones.

Obviamente el cálculo de probabilidades con variables aleatorias multidimensionales es más complejo que en el caso unidimensional, y no nos ocuparemos de él en este curso. No obstante

[4]El vector (X, Y) se entiende como aleatorio en el sentido de que, *a priori,* antes de medir cualquier ejemplar de esta familia no se pueden predecir su longitud ni su peso.

existe un caso, que se presenta con frecuencia en las aplicaciones prácticas, en el que las funciones que se acaban de definir adquieren una estructura simple. Es el caso de las *variables aleatorias independientes.*

2.6.2. Independencia de variables aleatorias.

Recordemos que dos sucesos A y B se dicen independientes si $P(A \cap B) = P(A) \cdot P(B)$. Esta definición puede generalizarse al caso de variables aleatorias. Así, dos variables aleatorias X e Y se dicen *estocásticamente independientes* o simplemente, *independientes*, si para cualesquiera $a, b, c, d \in \mathbb{R}$ los sucesos $\{a < X \leq b\}$ y $\{c < Y \leq d\}$ son independientes, esto es:

$$P(\{a < X \leq b\} \cap \{c < Y \leq d\}) = P(a < X \leq b) \cdot P(c < Y \leq d) \tag{2.2}$$

En lo que sigue llamaremos $F_X(x)$ y $F_Y(y)$ a las funciones de distribución respectivas de las variables X e Y. Asimismo, denotaremos por $f_X(x)$ y $f_Y(y)$ las respectivas funciones de probabilidad o densidad de probabilidad (según que X e Y sean discretas o continuas).

En el caso de que dos variables aleatorias X e Y sean independientes se cumplen las siguientes propiedades:

1. $F(x, y) = F_X(y) \cdot F_Y(y)$
2. $f(x, y) = f_X(x) \cdot f_Y(y)$

La demostración de estas propiedades puede encontrarse en el apéndice.

2.7. Parámetros característicos de las distribuciones de probabilidad.

En esta sección presentaremos algunas medidas que tienen como objetivo sintetizar –resumir– la distribución de probabilidad de una variable aleatoria en unos pocos valores característicos:

- *Esperanza:* Valor que describe dónde se encuentra el *"centro"* de la distribución de probabilidad.
- *Varianza*: Valor que describe el grado de dispersión de los valores que toma la variable aleatoria.

- *Momentos:* Valores que describen la *forma* de la distribución de probabilidad (asimetría, apuntamiento).
- *Cuantiles*: Valores por debajo de los cuales se acumula una determinada probabilidad (normalmente el 1 %, 2.5 %,5 %, 25 %, 50 %, 75 %, 95 %, 97.5 %, 99 %).
- *Covarianza y Correlación:* Valores que cuantifican el grado de asociación lineal entre dos variables X e Y.

2.7.1. Esperanza matemática

La *esperanza matemática* de una variable aleatoria X se define como:

- Si X es discreta: $E[X] = \sum_k k \cdot P(X = k)$
- Si X es continua y tiene función de densidad $f(x)$: $E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$

Si en el caso discreto identificamos la probabilidad de un valor con su *masa,* y en el caso continuo la densidad de probabilidad de un valor con la *densidad de masa en un entorno del mismo*, podemos interpretar la esperanza de una variable aleatoria como el *centro de gravedad* de su distribución de probabilidad. Más concretamente, si imaginamos la gráfica de la función de probabilidad (caso discreto) o de la densidad de probabilidad (caso continuo) como un objeto físico, la esperanza coincide con la posición del eje X en que deberíamos apoyar este objeto para que permanezca en equilibrio. La figura 2.12 muestra sendos ejemplos de la posición de la esperanza: en la figura (a) se muestra la función de probabilidad de una variable aleatoria discreta (concretamente la del ejemplo 2.2), y en la figura (b) la función de densidad de probabilidad de la altura de ola vista en el ejemplo 2.7. En ambos casos la posición de la esperanza se ha marcado con un pequeño triángulo. Se puede apreciar a simple vista que la esperanza corresponde al centro de gravedad en ambas figuras.

En ocasiones se requiere calcular la esperanza de alguna función[5] g de la variable aleatoria X. En tal caso la esperanza de la variable aleatoria $g(X)$ se define de modo análogo a la anterior:

- Si X es discreta: $E[g(X)] = \sum_k g(k) \cdot P(X = k)$

[5]Por ejemplo, si tiramos una moneda y el resultado es una variable X que vale 1 si sale cara y 0 si sale cruz. En este caso $E[X]$ representa el número esperado de caras. Si decidimos apostar y ganamos 10 € cada vez que sale cara, y perdemos 10€ cada vez que sale cruz, podemos representar nuestra apuesta mediante la función $g(X)$, que vale 10 cuando $X = 1$ (cara) y -10 cuando $X = 0$ (cruz). En este caso $E[g(X)]$ representa nuestra ganancia (o pérdida) esperada durante el juego.

- Si X es continua y tiene función de densidad $f(x)$: $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$

En el caso de variables aleatorias bidimensionales (X, Y) la esperanza de una función $g(X, Y)$ se define como:

- Si (X, Y) es un vector de variables discretas,

$$E[g(X,Y)] = \sum_x \sum_y g(x,y) \cdot P(X = x, Y = y)$$

- Si (X, Y) tiene distribución absolutamente continua con función de densidad $f(x, y)$:

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dxdy$$

Ejemplo 2.10. Para la variable aleatoria X definida en el ejemplo 2.2, correspondiente al resultado de la suma de las caras superiores resultantes al lanzar dos dados, la esperanza se obtiene fácilmente como:

$$E[X] = \sum_{k=2}^{12} kP(X = k) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + \\ + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} + = 7$$

Ejemplo 2.11. Para la variable aleatoria X definida en el ejemplo 2.5 (punto en que se parte una cuerda homogénea de un metro), la esperanza es:

$$E[X] = \int_{-\infty}^{\infty} xf(x) dx = \int_0^1 x \cdot 1 \cdot dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

Ejercicio 2.1. Calcular la esperanza de las variables aleatorias definidas en los ejemplos 2.6 y 2.7.

Propiedades de la esperanza matemática.

1. *Para cualquier constante arbitraria* c :

$$E[c] = c$$

2. *Dadas una variable aleatoria* X*, y una constante arbitraria* c*:*

$$E[cX] = cE[X]$$

3. *Dadas dos variables aleatorias* X *e* Y *:*

$$E[X + Y] = E[X] + E[Y]$$

4. *Si* X *e* Y *son independientes, entonces:*

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

La demostración de estas propiedades se encuentra en el apéndice.

2.7.2. Medidas de dispersión de una variable aleatoria.

La *varianza* es una medida de dispersión de los valores de una variable aleatoria X. Si la esperanza es $\mu = E[X]$, la varianza se define como:

$$Var(X) = E\left[(X - \mu)^2\right]$$

La varianza es, pues, el valor esperado de la distancia al cuadrado entre los valores que toma la variable aleatoria y su esperanza[6]; si los valores están muy agrupados, estarán muy cerca de su centro (la esperanza) y la varianza será pequeña; por contra, si los valores de X está muy alejados entre sí, lo estarán también de su centro, y la varianza será grande. Por tanto la varianza es, efectivamente, una medida de dispersión.

Dada su definición, es obvio que las unidades en que se mide la varianza corresponden al cuadrado de las unidades en que se mide la variable X. Esto resulta poco práctico en muchas ocasiones, por lo que se suele emplear como medida de dispersión la *desviación típica* definida

[6]Esta distancia se toma al cuadrado para evitar la presencia de valores negativos, que pueden falsear su significado.

como[7]:

$$\text{sd}(X) = \sqrt{\text{var}(X)}$$

Es habitual denotar la desviación típica de una variable aleatoria mediante la letra griega σ. De la misma forma, la varianza suele denotarse como σ^2.

La figura 2.13 muestra tres funciones de densidad correspondientes a variables aleatorias con la misma esperanza $E[X] = 0$, y con distintas desviaciones típicas. Como puede apreciarse, a medida que aumenta la desviación típica, la densidad se distribuye en un rango más amplio (la variable toma valores más dispersos). Nótese también que como el área total bajo la función de densidad debe ser siempre 1, cuando se incrementa el rango que abarca dicha función, su altura disminuye.

Propiedades de la varianza.

1. *Dadas una variable aleatoria X, y una constante arbitraria c:*

$$\text{var}(cX) = c^2\text{var}(X)$$
$$\text{var}(c + X) = \text{var}(X)$$

2. $\text{var}(X) = E[X^2] - (E[X])^2$

3. *Si X e Y son variables aleatorias independientes,* $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$

La demostración de estas propiedades se encuentra en el apéndice.

Desigualdad de Chebyshev.

La desigualdad de Chebyshev permite utilizar la varianza de una variable aleatoria para acotar el valor de ciertas probabilidades que resultan de interés práctico. Concretamente, si X es una variable aleatoria tal que $E[X] = \mu$ y $\text{var}(X) = \sigma^2$ esta desigualdad establece que para todo $k \geq 1$:

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

En otras palabras, la probabilidad de que X tome valores que disten de su esperanza menos de k veces su desviación típica es al menos $1 - \frac{1}{k^2}$. Así, por ejemplo:

- Eligiendo $k = 2$: $P(|X - \mu| \leq 2\sigma) \geq \frac{3}{4} = 0{,}75$

[7]Utilizamos aquí la notación *sd* para la desviación típica, que deriva de su denominación inglesa *standard deviation*.

- Eligiendo $k = 3 : P(|X - \mu| \leq 3\sigma) \geq 1 - \frac{1}{9} = 0{,}89$
- Eligiendo $k = 4 : P(|X - \mu| \leq 4\sigma) \geq 1 - \frac{1}{16} = 0{,}9375$

En cualquier caso, es importante darse cuenta de que la desigualdad de Chebyshev establece una cota inferior para estas probabilidades y puede alejarse mucho de la probabilidad exacta. Así por ejemplo (con $k = 2$) la desigualdad nos indica que la probabilidad de que los valores de X se diferencien de μ en menos de 2 desviaciones típicas es *al menos* 0.75, pero según como sea la distribución de X, esa probabilidad podría en realidad ser 0.8, 0.9, ó 0.95, por ejemplo.

Relación entre esperanza y media aritmética.

Supongamos que la variable aleatoria X mide alguna característica de los sujetos de una población (peso, talla, temperatura, ...), y sean $\mu = E[X]$ y $\sigma^2 = \text{var}(X)$. Se eligen *al azar* y *de manera independiente* n sujetos de esa población. Llamaremos *muestra aleatoria simple* a los valores $X_1, X_2, \ldots, X_n$ que toma la variable X cuando se evalúa sobre cada uno de esos sujetos. $X_1, X_2, \ldots, X_n$ son a su vez variables aleatorias, toda vez que sus valores no se conocen antes de haber sido medidos. Asimismo, como todos los sujetos proceden de la misma población, las X_i tendrán la misma distribución de probabilidad de X, por lo que $E[X_i] = \mu$ y $\text{var}(X_i) = \sigma^2$ para $i = 1, \ldots, n$.

La media aritmética de las observaciones, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, es también una variable aleatoria, ya que no es posible conocer su valor antes de haber obtenido la muestra. Cada posible muestra producirá unos valores distintos de $X_1, X_2, \ldots, X_n$, y por tanto un valor distinto de $\overline{X}$. Tiene sentido, por tanto, que nos preguntemos por cuál es el valor esperado de $\overline{X}$ (el centro de masas de todos los posibles valores que puede tomar) y cuál es su varianza. Ambos valores son fáciles de obtener. Aplicando las propiedades de la esperanza, tenemos:

$$E\left[\overline{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}n\mu = \mu$$

Asimismo, aplicando las propiedades de la varianza:

$$\text{var}\left(\overline{X}\right) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\,\text{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Por tanto, a medida que aumenta el valor de n, la varianza de $\overline{X}$ se va reduciendo, de tal forma que cuando n es grande $\text{var}\left(\overline{X}\right) \cong 0$. Ello significa que para valores grandes de n

el valor de $\overline{X}$ apenas se aparta de su valor esperado μ. De esta forma, cuando n es grande $\overline{X} \cong \mu$. Ello nos permite interpretar la esperanza de una variable aleatoria como la media aritmética de los valores observados de la misma en muestras aleatorias muy grandes.

2.7.3. Momentos.

Dada una variable aleatoria X, el *momento de orden k respecto al origen* (o simplemente *momento de orden k*), con $k \in \mathbb{N}$, se define como:

$$\mu_k = E\left[X^k\right]$$

Asimismo, si la esperanza de X es $E[X] = \mu$, se define el *momento de orden k respecto a la esperanza* (o *momento central de orden k*) como:

$$M_k = E\left[(X-\mu)^k\right]$$

Obviamente $\mu_1 = E[X]$ y $M_2 = \text{var}(X) = \mu_2 - \mu_1^2$. Los momentos centrales está relacionados con la forma de la distribución de probabilidad. Ya hemos visto que la varianza (que coincide con el momento central de orden 2) es una medida de dispersión. A partir del momento central de orden 3 se define el *coeficiente de asimetría*:

$$A = \frac{1}{\sigma^3} E\left[(X-\mu)^3\right]$$

y a partir del momento central de orden 4, el *coeficiente de apuntamiento o curtosis:*

$$\kappa = \frac{1}{\sigma^4} E\left[(X-\mu)^4\right] - 3$$

La figura 2.14 muestra funciones de densidad con diversos grados de asimetría:

- *Asimetría negativa:* la *masa* de probabilidad tiende a concentrarse a la derecha; en este caso el coeficiente de asimetría es negativo.
- *Asimetría positiva:* la *masa* de probabilidad tiende a concentrarse a la izquierda; en este caso el coeficiente de asimetría es positivo.
- *Simetría:* La masa de probabilidad se reparte simétricamente respecto a su centro (la esperanza). En este caso el coeficiente de asimetría es nulo.

La figura 2.15 muestra las funciones de densidad de tres variables aleatorias con las mismas esperanza y varianza, pero con distintos grados de apuntamiento:

- *Curtosis negativa* ($\kappa < 0$): corresponde a funciones de densidad más bien aplanadas y con *"colas"* cortas. Las curvas con esta forma reciben el nombre de *platicúrticas.*
- *Curtosis positiva* ($\kappa > 0$): corresponde a funciones de densidad más bien *"puntiagudas"* y con colas largas. Las curvas con esta forma se llaman *leptocúrticas.*
- *Curtosis nula* ($\kappa = 0$): corresponde al caso intermedio, con un pico redondeado y colas de tamaño intermedio, como ocurre con la curva en forma de campana. Las curvas de este tipo reciben el nombre de *mesocúrticas.*

2.7.4. Cuantiles

Dada una variable aleatoria X cuya función de distribución acumulativa es $F(x)$, se define el α-*ésimo cuantil* $(0 < \alpha < 1)$ como el valor q_α, tal que $F(q_\alpha) = P(X \leq q_\alpha) = \alpha$.

Cuando $F(x)$ es estrictamente creciente la ecuación anterior tiene solución única. En el caso de que la variable aleatoria sea discreta, ya hemos visto que $F(x)$ es escalonada; y aún cuando X sea continua, podría ocurrir que su función de distribución acumulativa presente intervalos en los que su valor sea constante. En estos casos se define el α-*ésimo* cuantil como $q_\alpha = \text{mín}\{x : F(x) \geq \alpha\}$.

Hay algunos cuantiles de uso muy frecuente, que reciben su propio nombre:

- La *mediana* (Me) es el cuantil 0,5. Por tanto, la probabilidad de que la variable tome valores menores o iguales que la mediana es el 50 %, y que tome valores mayores que ella es otro 50 %. Por esta razón, la mediana se usa habitualmente como medida de posición central.
- Los *cuartiles* (Q_1, Q_2 y Q_3): corresponden a los cuantiles 0.25, 0.5 (mediana) y 0.75.
- Los *centiles* o *percentiles* (P_k): corresponden a los cuantiles de la forma $\frac{k}{100}, k = 1, \ldots, 100$

Ejemplo 2.12. En el ejemplo 2.6 vimos que la profundidad a que se detecta el isótopo ^{210}Pb es una variable aleatoria cuya densidad de probabilidad puede modelarse por $f(x) = 0{,}1e^{-0{,}1x}$. La función de distribución es entonces:

$$P(X \leq x) = F(x) = \int_0^x f(s)\,ds = \int_0^x f(x) = \int_0^x 0{,}1e^{-0{,}1s}ds = 1 - e^{-0{,}1s}$$

Para calcular cualquier cuantil α bastará con resolver la ecuación $F(q_\alpha) = \alpha$, que en este caso queda de la forma:

$$1 - e^{-0,1q_\alpha} = \alpha \Rightarrow e^{-0,1q_\alpha} = 1 - \alpha \Rightarrow q_\alpha = -\frac{1}{0,1}\log(1-\alpha) = -10\log(1-\alpha)$$

Así, por ejemplo, la mediana sería $Me = -10\log 0,5 = 6,93$, y el percentil 95 sería $P_{95} = -10\log 0,05 = 29,96$.

Ejemplo 2.13. En el ejemplo 2.7 hemos visto que la altura de ola (en metros) en cierta zona puede modelarse mediante una variable aleatoria con función de densidad $f(x) = \lambda^2 x e^{-\lambda x}$, $x \geq 0$, $\lambda = 0,9$. Se desean calcular los cuantiles 0.025 y 0.975.

Para ello obtenemos primero la función de distribución acumulativa:

$$F(x) = \int_0^x f(s)\,ds = \int_0^x \lambda^2 s \cdot e^{-\lambda s} ds = \lambda\left[-se^{-\lambda s} - \frac{1}{\lambda}e^{-\lambda s}\right]_0^x =$$
$$= \lambda\left(\frac{1}{\lambda} - xe^{-\lambda x} - \frac{1}{\lambda}e^{-\lambda x}\right) = 1 - e^{-\lambda x}(1+\lambda x)$$

Para encontrar el cuantil α hemos de resolver $F(q_{0,025}) = 0,025$. Por tanto (teniendo en cuenta que $\lambda = 0,9$):

$$1 - e^{-0,9\cdot q_{0,025}}(1 + 0,9q_{0,025}) = 0,025$$
$$0,975 - e^{-0,9\cdot q_{0,025}}(1 + 0,9q_{0,025}) = 0$$

Esta ecuación obviamente no puede resolverse de manera explícita, así que utilizamos la función `uniroot` de R. La figura 2.7 nos indica que el cuantil buscado debe estar en el intervalo $(0, 1)$:

```
Q = function(qa) {
    0.975 - exp(-0.9 * qa) * (1 + 0.9 * qa)
}
uniroot(Q, interval = c(0, 1))$root

## [1] 0.2691
```

El cuantil 0.975 se obtiene de modo análogo, salvo que buscamos en el intervalo $(5, 10)$:

```
Q = function(qa) {
    0.025 - exp(-0.9 * qa) * (1 + 0.9 * qa)
}
uniroot(Q, interval = c(5, 10))$root

## [1] 6.191
```

De esta forma, con una probabilidad 0.95, la altura de ola en esta zona se encuentra entre los 0.269 y los 6.191 metros, esto es, $P\left(0{,}269 < X \leq 6{,}191\right) = 0{,}95$

2.7.5. Asociación lineal entre variables aleatorias.

Covarianza.

Dadas dos variables aleatorias X e Y, con esperanzas respectivas $E\left[X\right]$ y $E\left[Y\right]$, se define la *covarianza* entre ambas variables como:

$$\operatorname{cov}\left(X,Y\right) = E\left[\left(X - E\left[X\right]\right)\left(Y - E\left[Y\right]\right)\right]$$

La covarianza es, pues, el valor esperado del producto $\left(X - E\left[X\right]\right)\left(Y - E\left[Y\right]\right)$, lo que significa que:

- Si este valor es positivo X e Y varían conjuntamente en el mismo sentido: en efecto, el producto $\left(X - E\left[X\right]\right)\left(Y - E\left[Y\right]\right)$ es positivo solo si valores positivos de $\left(X - E\left[X\right]\right)$ tienden a ir acompañados de valores positivos de $\left(Y - E\left[Y\right]\right)$, y valores negativos de $\left(X - E\left[X\right]\right)$ tienden a ir acompañados de valores negativos de $\left(Y - E\left[Y\right]\right)$. O, dicho de otra forma, si valores de X superiores a $E\left[X\right]$ tienden a ir acompañados de valores de Y mayores que $E\left[Y\right]$, y valores de X menores que $E\left[X\right]$ tienden a ir acompañados de valores de Y menores que $E\left[Y\right]$. Cuanto más fuerte sea esta tendencia, mayor será el valor de la covarianza.

- Si este valor es negativo X e Y varían conjuntamente sentidos opuestos: el producto $\left(X - E\left[X\right]\right)\left(Y - E\left[Y\right]\right)$ es negativo solo si valores positivos de $\left(X - E\left[X\right]\right)$ tienden a ir acompañados de valores negativos de $\left(Y - E\left[Y\right]\right)$, y valores negativos de $\left(X - E\left[X\right]\right)$ tienden a ir acompañados de valores positivos de $\left(Y - E\left[Y\right]\right)$. Dicho de otra forma, valores de X mayores que $E\left[X\right]$ tienden a ir acompañados de valores de Y menores que $E\left[Y\right]$, y valores de X menores que $E\left[X\right]$ tienden a ir acompañados de valores de

Y mayores que $E[Y]$. Cuánto más fuerte sea esta tendencia mayor (en valor absoluto) será la covarianza.

- Si este valor es nulo, entonces valores positivos y negativos de $(X - E[X])$ van acompañados indistintamente por valores positivos o negativos de $(Y - E[Y])$, de tal forma que los productos $(X - E[X])(Y - E[Y])$ positivos se cancelan con los negativos.

La figura 2.16(a) muestra la función de densidad de un vector aleatorio (X, Y) para el que $\text{cov}(X, Y) > 0$. Puede apreciarse que esta función de densidad concentra la mayor parte de la probabilidad a lo largo de una linea en el plano XY. La figura 2.16(b) muestra una nube de puntos generada por la densidad anterior (hay mayor densidad de puntos donde la densidad encierra mayor volumen). Se aprecia aún más claramente el alineamiento de los puntos a lo largo de una recta, que tiene pendiente positiva. En trazos punteados se han marcado las posiciones de las esperanzas de X e Y respectivamente, dividiendo el plano XY en cuatro cuadrantes. Como puede verse, precisamente debido a la presencia de esta relación lineal positiva entre la X y la Y, hay más puntos en los cuadrantes (2) y (4), justamente aquellos en los que $(X - E[X])(Y - E[Y]) > 0$; además, estos puntos se alejan más del centro, esto es de la posición de $(E[X], E[Y])$, por lo que la magnitud absoluta de los valores $(X - E[X])(Y - E[Y])$ asociados será también mayor. Todo ello indica que la existencia de una asociación lineal con pendiente positiva entre la X y la Y implica un valor positivo de la covarianza, tanto más grande cuanto mayor sea el grado de asociación lineal entre las variables (mejor el ajuste de los puntos a una recta).

Un razonamiento análogo sobre la figura 2.17 nos muestra que la existencia de una relación lineal de pendiente negativa entre X e Y se asocia con una covarianza negativa, tanto mayor en valor absoluto cuanto mejor sea el ajuste a una recta. Por último, la figura 2.18 nos muestra que cuando no hay asociación lineal entre las variables X e Y, se tiene que $\text{cov}(X, Y) = 0$, ya que los puntos se reparten por igual en los cuatro cuadrantes, cancelándose los términos $(X - E[X])(Y - E[Y])$ positivos con los negativos.

La figura 2.19 nos muestra otra situación. Nuevamente tenemos la densidad a la izquierda y una nube de puntos generada por esta densidad a la derecha. Claramente las variables X e Y no son independientes (*conocer el valor de X nos informa aproximadamente de cuál puede ser el valor de Y*). Sin embargo, en los cuadrantes (1) y (2) los valores de $(X - E[X])(Y - E[Y])$ son iguales y de signo contrario; lo mismo sucede con los cuadrantes (3) y (4), por lo que $\text{cov}(X, Y) = 0$. Por tanto *una covarianza nula no significa que no haya asociación entre las variables,* ya que de hecho podría existir una asociación no lineal como en este caso.

Propiedades de la covarianza.

1. $\operatorname{cov}(aX, bY) = ab \operatorname{cov}(X, Y)$
2. $\operatorname{cov}(X, X) = \operatorname{var}(X)$
3. $\operatorname{cov}(X, Y) = E[XY] - E[X]E[Y]$
4. *Si X e Y son independientes,* $\operatorname{cov}(X, Y) = 0$

La demostración de estas propiedades se encuentra en el apéndice.

Ejercicio 2.2. Calcular la covarianza de las variables $U = X \cdot Y$ y $V = X + Y$ siendo X e Y los resultados de las caras superiores del lanzamiento de dos dados (ver ejemplo 2.8).

Correlación.

Hemos visto que el signo de la covarianza nos indica si entre las variables existe una relación lineal con pendiente positiva o negativa. Sin embargo no nos informa de la *intensidad* de esa relación, ya que el valor de la covarianza depende de las unidades en que se midan las variables X e Y. Para evitar este problema se define el *coeficiente de correlación lineal de Pearson* como*:*

$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

siendo σ_X^2 y σ_Y^2 las varianzas de X e Y respectivamente. De esta definición se sigue inmediatamente que $\rho_{X,Y}$ es adimensional.

Propiedades del coeficiente de correlación.

1. *Si X e Y son independientes, entonces* $\rho_{X,Y} = 0$
2. $-1 \leq \rho \leq 1$
3. *Si* $|\rho| = 1$ *entonces* $Y = aX + b$ *(los valores* (X, Y)*se disponen exactamente a lo largo de una recta)*

La demostración de estas propiedades se encuentra en el apéndice.

Cuando $\rho_{X,Y} = 0$, las variables X e Y se dicen *incorreladas.*

La primera de las propiedades anteriores nos indica que la independencia entre dos variables implica la incorrelación. Lo contrario en general no es cierto como se ha visto con las variables

representadas en la figura 2.19; estas variables están asociadas, pero como su covarianza es cero, también su correlación es cero.

Ejercicio 2.3. Calcular el coeficiente de correlación entre las variables del ejercicio 2.2.

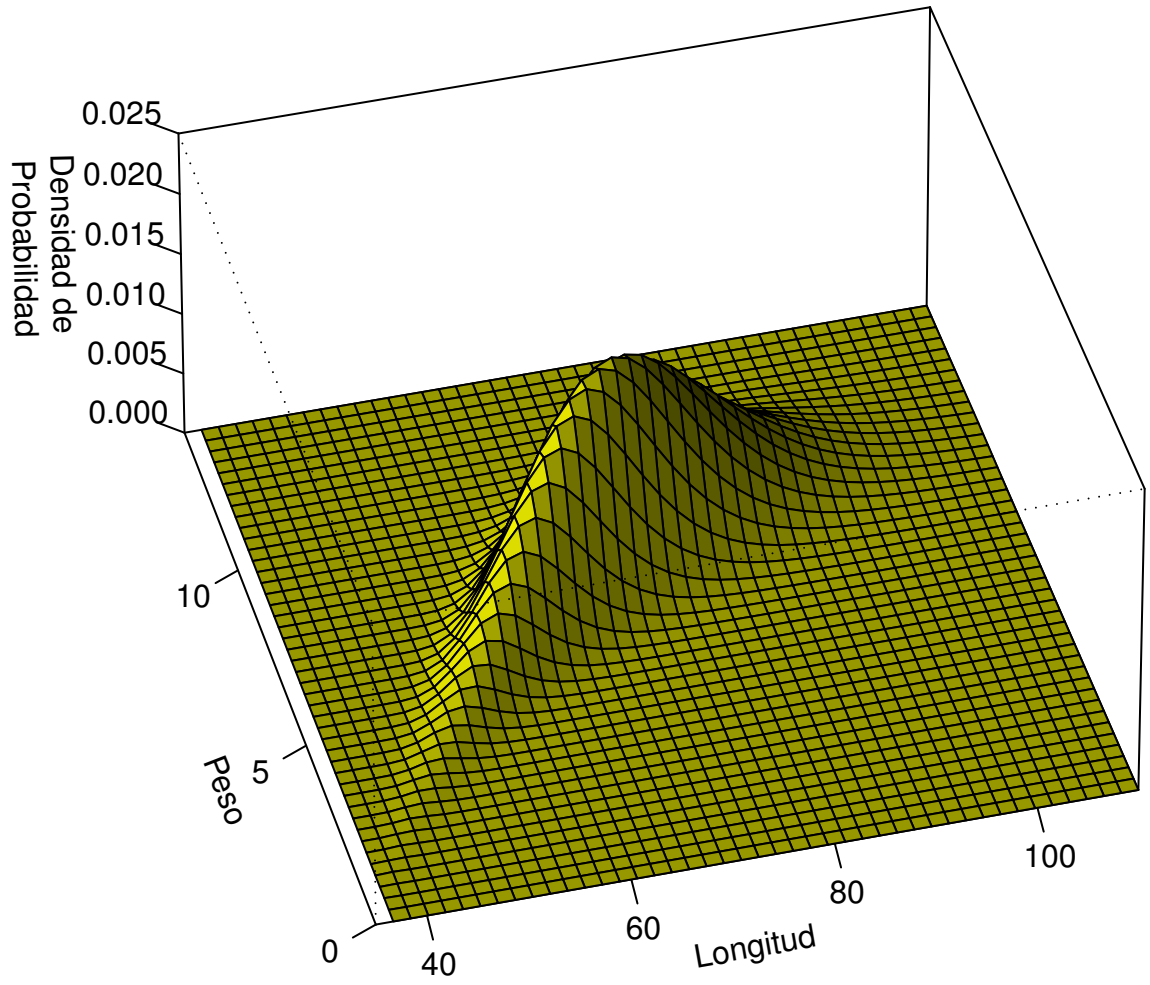

Figura 2.10: Función de densidad del vector aleatorio (X, Y)=(*Longitud, Peso)* para una población de peces de la familia *Serránidos,* subfamilia *Epinephelina* (ejemplo 2.9)

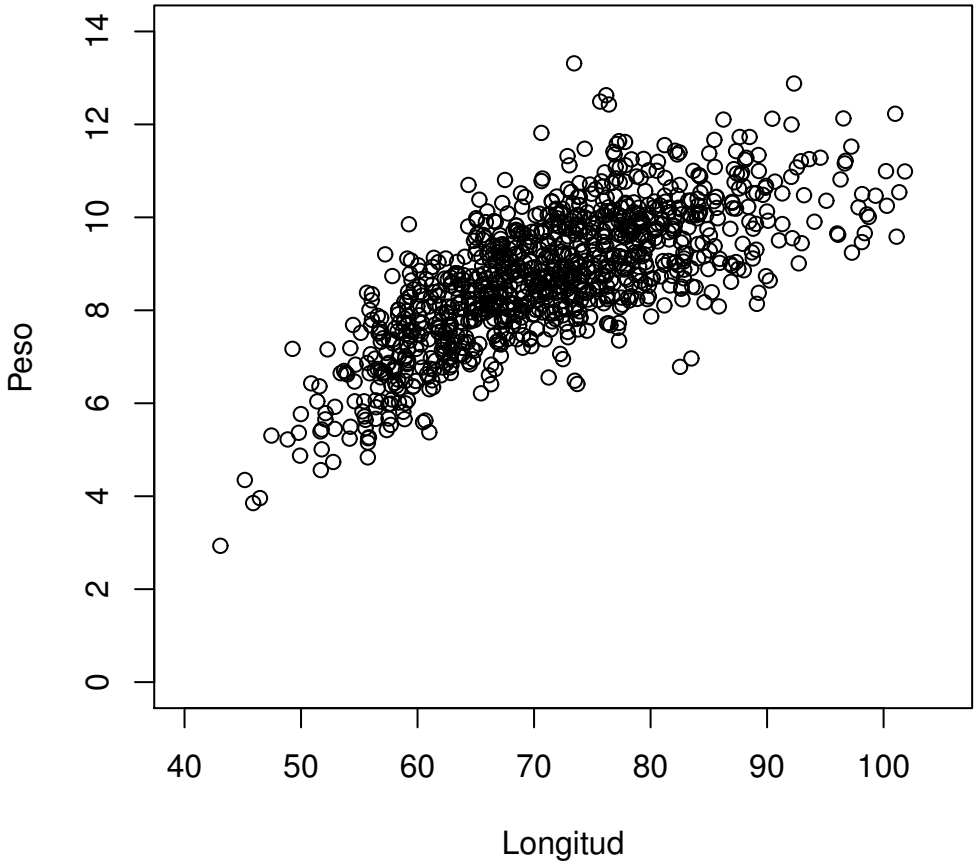

Figura 2.11: Nube de puntos correspondiente a la observación de la longitud y peso de 1000 peces del estudio descrito en el ejemplo 2.9.

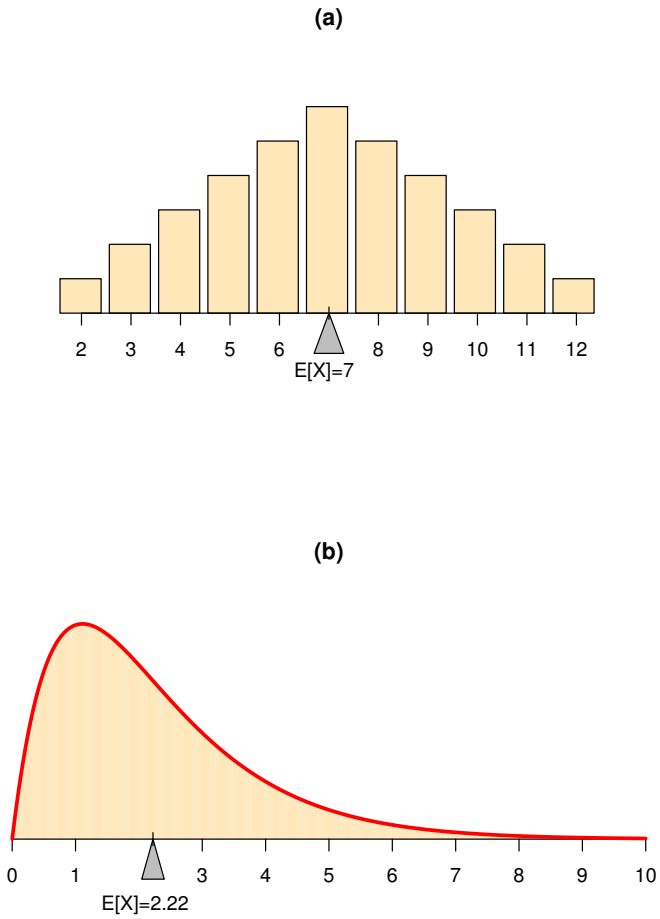

Figura 2.12: (a) Representación de la función de probabilidad de una variable aleatoria discreta (b) Representación de la densidad de probabilidad de una variable aleatoria continua. En ambos casos la posición de su esperanza (centro de gravedad de la figura) se representa mediante un triángulo.

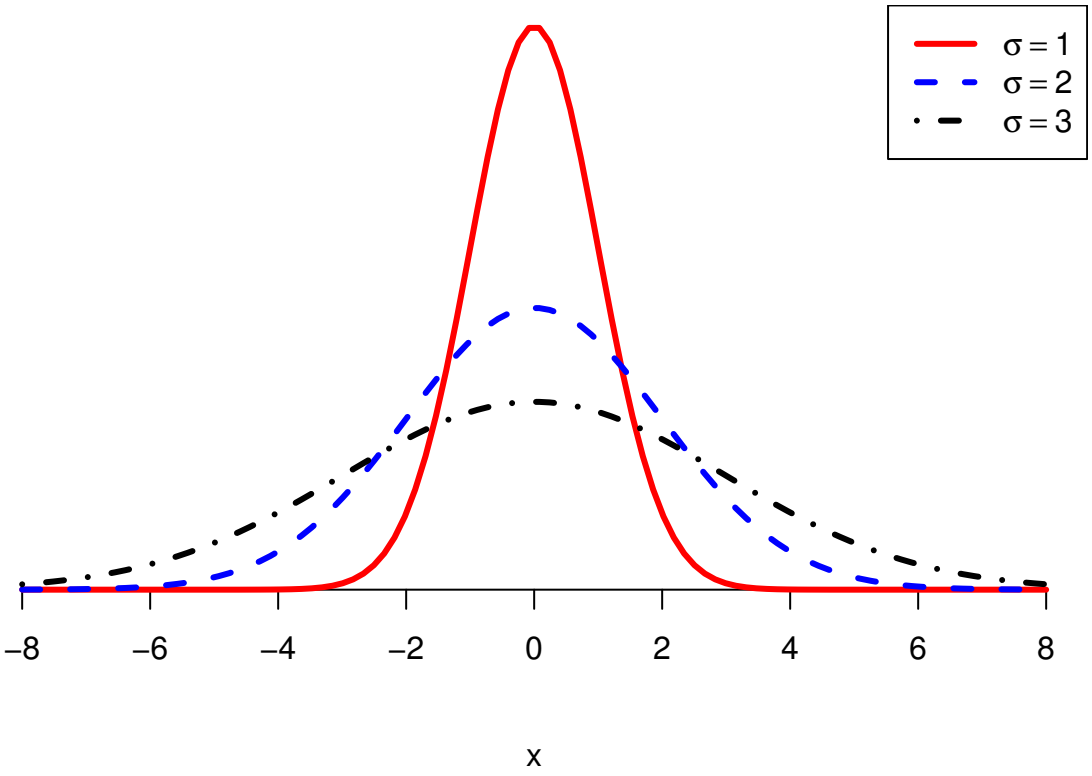

Figura 2.13: Funciones de densidad de tres variables aleatorias con distintas desviaciones típicas.

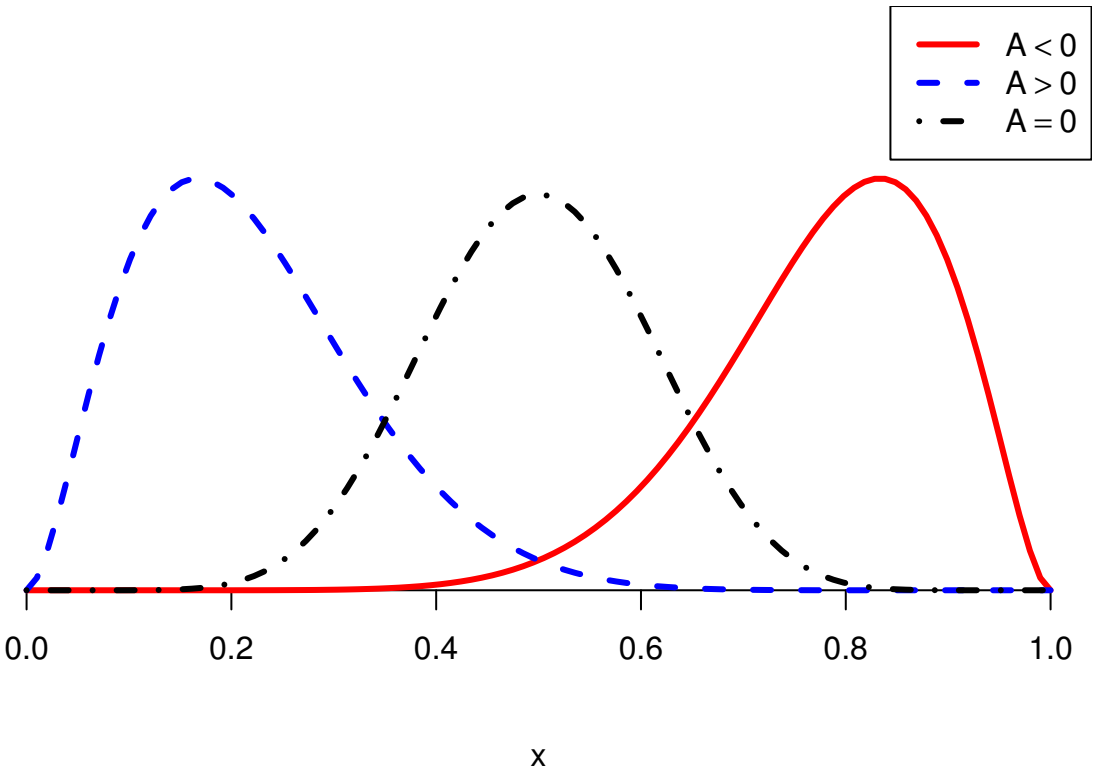

Figura 2.14: Funciones de densidad con diversos grados de asimetría.

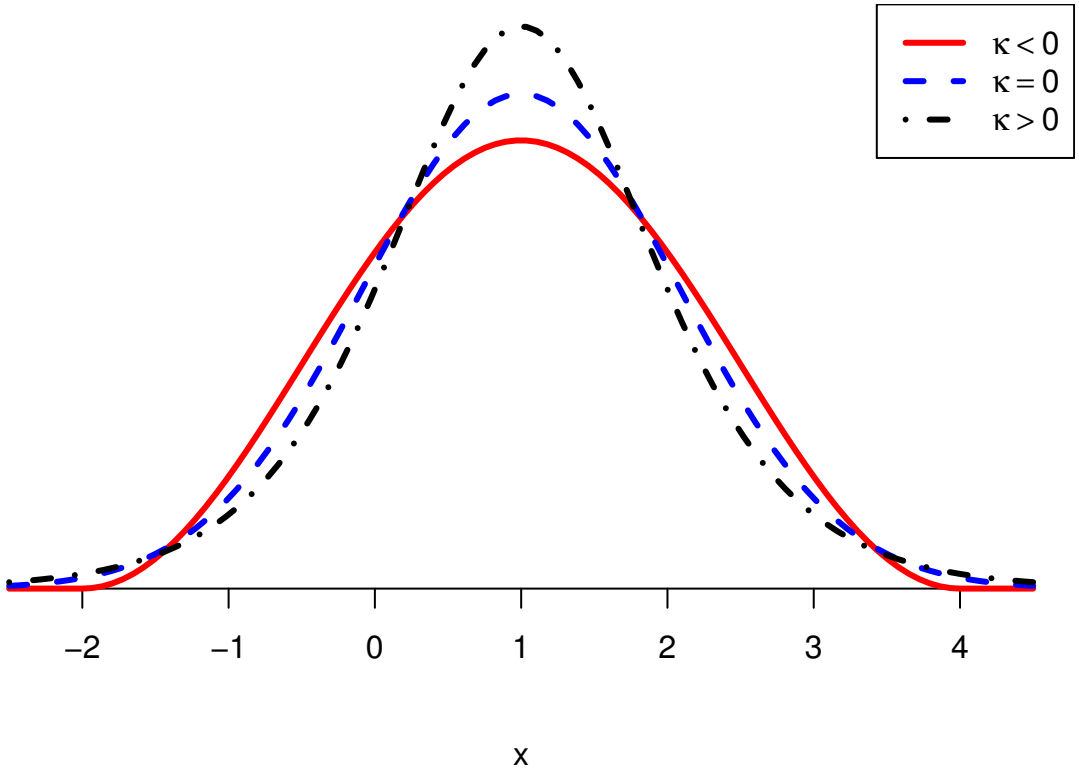

Figura 2.15: Funciones de densidad de tres variables aleatorias con distintos grados de apuntamiento. Las tres variables tienen distribución simétrica y las mismas esperanza y varianza.

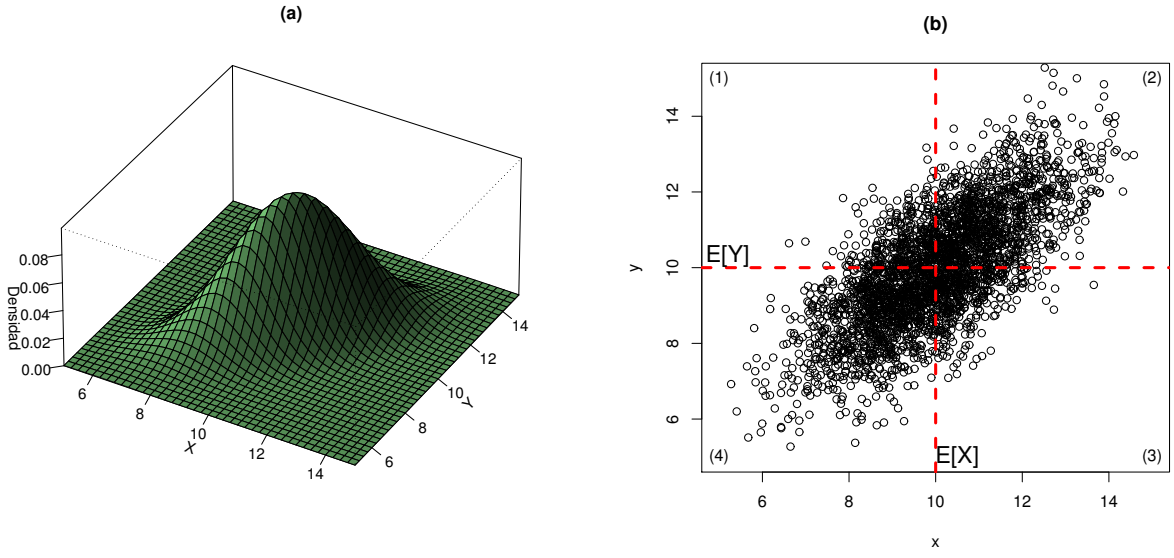

Figura 2.16: (a) Función de densidad de un vector aleatorio (X, Y) para el que cov $(X, Y) > 0$. (b) Nube de puntos generada por la función de densidad anterior.

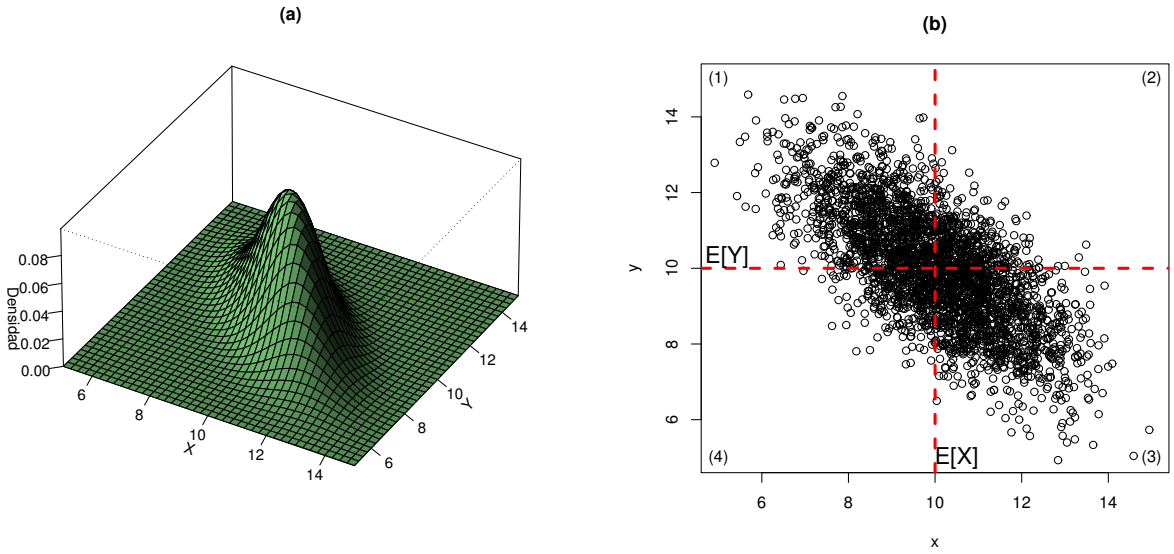

Figura 2.17: (a) Función de densidad de un vector aleatorio (X, Y) para el que $\text{cov}(X, Y) < 0$. (b) Nube de puntos generada por la función de densidad anterior.

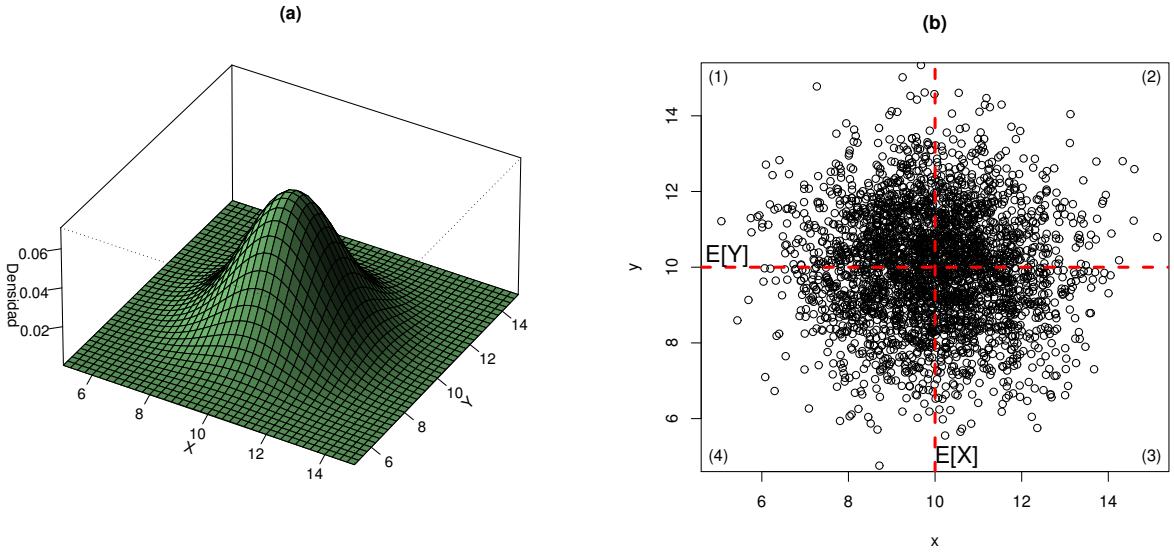

Figura 2.18: (a) Función de densidad de un vector aleatorio (X, Y) para el que $\text{cov}(X, Y) = 0$. (b) Nube de puntos generada por la función de densidad anterior. No se aprecia asociación entre las variables.

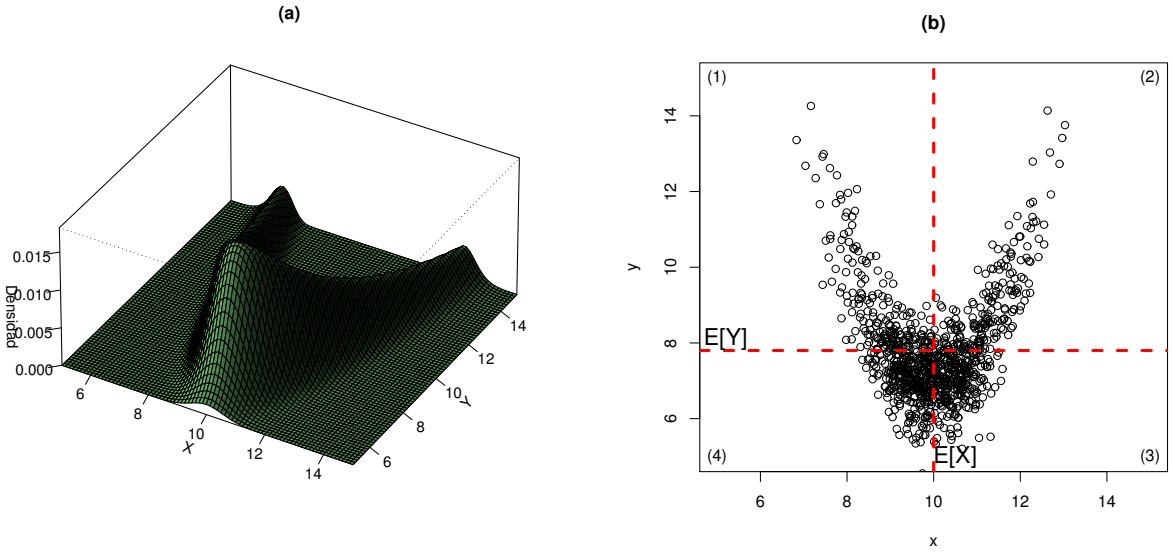

Figura 2.19: (a) Función de densidad de un vector aleatorio (X, Y) para el que cov $(X, Y) = 0$. (b) Nube de puntos generada por la función de densidad anterior. Entre X e Y se aprecia la existencia de una asociación no lineal.

Apéndice A

Demostraciones

Propiedades de la distribución conjunta de variables aleatorias independientes.

En el caso de que dos variables aleatorias X e Y sean independientes se cumplen las siguientes propiedades:

1. $F(x,y) = F_X(y) \cdot F_Y(y)$
2. $f(x,y) = f_X(x) \cdot f_Y(y)$

Demostración.

1. $F(x,y) = P(\{X \leq x\} \cap \{Y \leq y\}) = P(\{-\infty < X \leq x\} \cap \{-\infty < Y \leq y\})) =$
   $= P(-\infty < X \leq x) \cdot P(-\infty < Y \leq y) = P(X \leq x) \cdot P(Y \leq y) == F_X(x) \cdot F_Y(y)$

2. *a)* Si X e Y son discretas:

   $f(x,y) = P(\{X = x\} \cap \{Y = y\}) = P(\{x-1 < X \leq x\} \cap \{y-1 < Y \leq y\})) =$
   $=P(x-1 < X \leq x) \cdot P(y-1 < Y \leq y) = P(X = x) \cdot P(Y = y) = f_X(x) \cdot f_Y(y)$

   b) Si (X,Y) tiene distribución absolutamente continua:

$$f(x,y) = \lim_{\Delta x \to 0, \Delta y \to 0} \frac{P(\{x < X \leq x + \Delta x\}) \cdot P(\{y < Y \leq y + \Delta y\})}{\Delta x \Delta y} =$$
$$= \lim_{\Delta x \to 0} \frac{P(\{x < X \leq x + \Delta x\})}{\Delta x} \lim_{\Delta y \to 0} \frac{P(\{y < Y \leq y + \Delta y\})}{\Delta y} = f_X(x) f_Y(y)$$

□

Propiedades de la esperanza matemática.

1. *Para cualquier constante arbitraria* c :

$$E[c] = c$$

2. *Dadas una variable aleatoria* X*, y una constante arbitraria* c*:*

$$E[cX] = cE[X]$$

3. *Dadas dos variables aleatorias* X *e* Y *:*

$$E[X + Y] = E[X] + E[Y]$$

4. *Si* X *e* Y *son independientes, entonces:*

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

Demostración.

1. Una constante c puede considerarse equivalente a una variable aleatoria I_c que toma el valor c con probabilidad 1. De esta forma, la función de probabilidad de esta variable es:

$$P(I_c = c) = 1$$
$$P(I_c = x) = 0 \ \forall x \neq c$$

   Su esperanza es entonces $E[c] = E[I_c] = \sum_x x \cdot P(I_c = x) = c \cdot P(I_c = c) = c$

2. La demostración de esta propiedad es trivial y se deja como ejercicio.

3. Demostraremos este resultado sólo en el caso de que ambas variables sean discretas. Como $X + Y$ es una función de dos variables aleatorias, para calcular su esperanza

hemos de utilizar su función de probabilidad conjunta:

$$\begin{aligned} E[X+Y] &= \sum_x \sum_y (x+y) \cdot P(X=x, Y=y) = \\ &= \sum_x \sum_y x \cdot P(X=x, Y=y) + \sum_x \sum_y y \cdot P(X=x, Y=y) = \\ &= \sum_x x \cdot \sum_y P(X=x, Y=y) + \sum_y y \cdot \sum_x P(X=x, Y=y) = \\ &= \sum_x x \cdot P(X=x) + \sum_y y \cdot P(Y=y) = \\ &= E[X] + E[Y] \end{aligned}$$

Aquí hemos utilizado que

$$\sum_y P(X=x, Y=y) = P(X=x) \text{ y que } \sum_x P(X=x, Y=y) = P(Y=y)$$

Ambos resultados son triviales: los sucesos de la forma $\{Y=y\}$ forman un sistema completo de sucesos (el espacio muestral es $E = \cup_y \{Y=y\}$ y son incompatibles dos a dos, $\{Y=y_i\} \cap \{Y=y_j\} = \emptyset$ para cualesquiera $y_i \neq y_j$). Por tanto:

$$\begin{aligned} P(X=x) &= P(\{X=x\} \cap E) = P(\{X=x\} \cap (\cup_y \{Y=y\})) = \\ &= P(\cup_y (\{X=x\} \cap \{Y=y\})) = \sum_y P(\{X=x\} \cap \{Y=y\}) \end{aligned}$$

La demostración para el caso continuo es análoga, sustituyendo sumatorias por integrales y la función de probabilidad conjunta por la función de densidad conjunta.

4. En el caso discreto es $E[X \cdot Y] = \sum_x \sum_y x \cdot y \cdot P(X=x, Y=y) = \sum_i \sum_j x \cdot y \cdot f(x,y)$. Como X e Y son independientes $f(x,y) = f_X(x) f_Y(y)$, y por tanto:

$$E[X \cdot Y] = \sum_x \sum_y x \cdot y \cdot f_X(x) f_Y(y) = \left(\sum_x x \cdot f_X(x) \right) \left(\sum_y y \cdot f_Y(y) \right) = E[X] E[Y]$$

La demostración en el caso continuo es análoga cambiando sumatoria por integral.

□

Propiedades de la varianza.

1. *Dadas una variable aleatoria X, y una constante arbitraria c:*

$$\begin{aligned}\text{var}\,(cX) &= c^2\text{var}\,(X)\\ \text{var}\,(c+X) &= \text{var}\,(X)\end{aligned}$$

2. $\text{var}\,(X) = E\,[X^2] - (E\,[X])^2$

3. *Si X e Y son variables aleatorias independientes,* $\text{var}\,(X+Y) = \text{var}\,(X) + \text{var}\,(Y)$

Demostración.

1. La primera propiedad se sigue directamente de la linealidad de la esperanza. En efecto, si $E\,[X] = \mu$ se tiene que $E\,[cX] = c\mu$, y de aquí: $\text{var}\,(cX) = E\left[(cX - c\mu)^2\right] = E\left[c^2\,(X-\mu)^2\right] = c^2 E\left[(X-\mu)^2\right] = c^2\text{var}\,(X)$. Asimismo $\text{var}\,(c+X) = E\left[((c+X) - E\,(c+X))^2\right] = E\left[(c+X-E\,[c]-E\,[X])^2\right] = E\left[(X-E\,[X])^2\right] = \text{var}\,(X)$ ya que $E\,[c] = c$.

2. La segunda propiedad se sigue desarrollando el cuadrado $(X-\mu)^2$ y aplicando la linealidad de la esperanza: $\text{var}\,(X) = E\left[(X-\mu)^2\right] = E\,[X^2 - 2\mu X + \mu^2] = E\,[X^2] - 2\mu E\,[X] + \mu^2 = E\,[X^2] - \mu^2$

3. Para demostrar la tercera propiedad, llamando $\mu_X = E\,[X]$ y $\mu_Y = E\,[Y]$ y teniendo en cuenta que, por la segunda propiedad de la esperanza, $E\,[X+Y] = E\,[X] + E\,[Y] = \mu_X + \mu_Y$:

$$\begin{aligned}\text{var}(X+Y) =& E\left[(X+Y-(\mu_X+\mu_Y))^2\right] = E\left[((X-\mu_X)-(Y-\mu_Y))^2\right] =\\ =& E\left[(X-\mu_X)^2 - 2\,(X-\mu_X)\,(Y-\mu_Y) + (Y-\mu_Y)^2\right] =\\ =& E\left[(X-\mu_X)^2\right] - 2E\left[(X-\mu_X)\,(Y-\mu_Y)\right] + E\left[(Y-\mu_Y)^2\right] =\\ =& \text{var}\,(X) + \text{var}\,(Y) - 2E\left[(X-\mu_X)\,(Y-\mu_Y)\right]\end{aligned}$$

   Veamos ahora que $E\left[(X-\mu_X)\,(Y-\mu_Y)\right] = 0$ cuando X e Y son independientes; en efecto:

$$\begin{aligned}E\left[(X-\mu_X)\,(Y-\mu_Y)\right] &= E\left[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y\right] =\\ &= E\,[XY] - \mu_X E\,[Y] - \mu_Y E\,[X] + \mu_X\mu_Y =\\ &= E\,[XY] - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y = E\,[XY] - \mu_X\mu_Y\end{aligned}$$

   De acuerdo con la tercera propiedad de la esperanza, para variables independientes se tiene $E\,[XY] = E\,[X]\,E\,[Y] = \mu_X\mu_Y$, por lo que $E\,[X]\,E\,[Y] - \mu_X\mu_Y = 0$

□

Desigualdad de Chebyshev.

Si X es una variable aleatoria tal que $E[X] = \mu$ y $\text{var}(X) = \sigma^2$, entonces para todo $k \geq 1$:

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Presentamos la demostración para el caso de variables aleatorias discretas. El caso continuo es análugo sustituyendo sumatorias por integrales.

Demostración. Consideremos el suceso:

$$A = \{x : |x - \mu| \geq k\sigma\}$$

De la definición de varianza se tiene:

$$\sigma^2 = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 P(X = x) = \sum_{x \in A} (x-\mu)^2 P(X = x) + \sum_{x \in \overline{A}} (x-\mu)^2 P(X = x)$$

Como ambos sumandos son positivos:

$$\sigma^2 \geq \sum_{x \in A} (x-\mu)^2 P(X = x)$$

Ahora bien, tal como se ha definido el suceso A, para todos los $x \in A$ se tiene que $|x - \mu| \geq k\sigma$. Por tanto:

$$\sigma^2 \geq \sum_{x \in A} (x-\mu)^2 P(X = x) \geq \sum_{x \in A} (k\sigma)^2 P(X = x) = (k\sigma)^2 \sum_{x \in A} P(X = x) = (k\sigma)^2 P(A)$$

De aquí se sigue que

$$P(A) \leq \frac{1}{k^2}$$

y por tanto □

$$P(|X - \mu| < k\sigma) = 1 - P(A) \geq 1 - \frac{1}{k^2}$$

Propiedades de la covarianza.

1. $\operatorname{cov}(aX, bY) = ab \operatorname{cov}(X, Y)$
2. $\operatorname{cov}(X, X) = \operatorname{var}(X)$
3. $\operatorname{cov}(X, Y) = E[XY] - E[X] E[Y]$
4. *Si X e Y son independientes,* $\operatorname{cov}(X, Y) = 0$

Demostración. La demostración de las tres primeras propiedades es inmediata. La cuarta se sigue de la tercera y de que, como hemos visto en 2.7.1, si X e Y son independientes entonces $E[XY] = E[X] E[Y]$. □

Propiedades del coeficiente de correlación.

1. *Si X e Y son independientes, entonces* $\rho_{X,Y} = 0$
2. $-1 \leq \rho \leq 1$
3. *Si* $|\rho| = 1$ *entonces* $Y = aX + b$ *(los valores* (X, Y)*se disponen exactamente a lo largo de una recta)*

Demostración.

1. La demostración de la primera propiedad es inmediata a partir de la propiedad 4 de la covarianza.
2. Para la segunda propiedad observemos que para cualquier constante a, y para cualesquiera variables aleatorias U y V se tiene que $E\left[(aU + V)^2\right] \geq 0$. Desarrollando el cuadrado y aplicando las propiedades de la esperanza resulta:

$$a^2 E\left[U^2\right] + 2aE[UV] + E\left[V^2\right] \geq 0$$

Esta ecuación representa una parábola que a lo sumo toca al eje de abcisas en un punto; por tanto, la ecuación $a^2 E[U^2] + 2aE[UV] + E[V^2] = 0$ tiene como mucho una solución, lo que significa que su discriminante debe ser menor o igual que cero, esto es, $4(E[UV])^2 - 4E[U^2] E[V^2] \leq 0$, o lo que es lo mismo:

$$(E[UV])^2 \leq E\left[U^2\right] E\left[V^2\right]$$

Si consideramos $U = (X - E[X])$ y $V = (Y - E[Y])$ se obtiene de inmediato la propiedad 2.

3. Por último, si $|\rho| = 1$ entonces siguiendo hacia atrás el argumento que acabamos de emplear, concluimos que existe una constante a tal que $E\left[(aU+V)^2\right] = 0$. Como los términos $(aU+V)^2$ son siempre mayores o iguales que 0 (por ser un cuadrado), la única forma de que su esperanza sea 0, es que $aU + V = 0$. Luego $a\,(X - E\,[X]) + (Y - E\,[Y]) = 0$, de donde $Y = aX - aE\,[X] + E\,[Y]$. Llamando $b = -aE\,[X] + E\,[Y]$ resulta la propiedad 3.

□

Capítulo 3

Distribuciones de Probabilidad Notables. Teorema Central del Límite.

1. Introducción

En este tema estudiaremos las distribuciones de probabilidad más habituales en las aplicaciones prácticas. En primer lugar veremos algunas distribuciones discretas –Bernoulli, binomial, hipergeométrica, geométrica y de Poisson–, y seguidamente algunas distribuciones continuas –uniforme, exponencial, gamma, Weibull y Normal–. De entre las distribuciones continuas destaca la normal ya que bajo determinadas condiciones aparece como límite de muchas variables. Estudiaremos tales condiciones y su interpretación, para finalmente ver las principales distribuciones de probabilidad que aparecen en la inferencia estadística cuando se toman muestras aleatorias de poblaciones que se distribuyen normalmente.

2. OBJETIVOS

Al finalizar este tema alumno deberá:

1. Conocer y saber calcular probabilidades asociadas a las distribuciones discretas notables, en particular, la binomial, la hipergeométrica y la de Poisson
2. Conocer y saber calcular probabilidades asociadas a las distribuciones continuas notables.
3. Entender el significado de los parámetros característicos de cada distribución, y como la elección adecuada de los valores de los parámetros permite modelar variables observadas en la naturaleza.

4. Conocer la distribución normal y su propiedad reproductiva. Utilizar la tabla de la distribución normal estándar. Entender y ser capaz de aplicar en situaciones prácticas el teorema central del límite.

5. Conocer las principales distribuciones que surgen en la inferencia estadística asociadas al muestreo (t de Student, chi-cuadrado y F de Fisher), así como manejar sus tablas.

6. Ser capaz de utilizar R para el cálculo de probabilidades en variables con las distribuciones vistas en este capítulo.

3. Principales distribuciones de probabilidad discretas.

3.1. Distribución Uniforme Discreta.

Definición: Una variable aleatoria X que toma un número finito n de valores $\{x_1, x_2, \ldots, x_n\}$ sigue una *distribución uniforme* si todos sus valores son equiprobables. Por tanto su función de probabilidad es de la forma:

$$f(x) = P(X = x) = \begin{cases} \frac{1}{n} & x \in \{x_1, x_2, \ldots, x_n\} \\ 0 & x \notin \{x_1, x_2, \ldots, x_n\} \end{cases}$$

Esperanza y varianza:

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

$$\sigma^2 = var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p(X = x_i) = \sum_{i=1}^{n} (x_i - \mu)^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Ejemplo: Si X= "*Resultado obtenido al lanzar un dado equilibrado*":

$$\mu = E[X] = \sum_{i=1}^{k} p_i x_i = \frac{1}{6} \sum_{i=1}^{6} i = \frac{1}{6} \cdot 21 = 3,5$$

$$\sigma^2 = var[X] = \sum_{i=1}^{k} p_i (x_i - \mu)^2 = \frac{1}{6} \sum_{i=1}^{6} (i - 3,5)^2 = 2,91$$

3.2. Distribución de Bernoulli $Be\,(p)$

Definición: Una variable aleatoria X sigue una distribución de Bernoulli, $Be\,(p)$, si sólo toma dos posibles valores: 1 ("*éxito*") ó 0 ("*fracaso*"), con probabilidades respectivas p y $1-p$. Su función de probabilidad es, por tanto:

$$f(k) = P(X=k) = \begin{cases} p & k=1 \\ 1-p & k=0 \\ 0 & k \notin \{0,1\} \end{cases}$$

que podemos expresar también como:

$$f\,(k) = p^k\,(1-p)^{1-k}\,,\ k=0,1$$

Esperanza y varianza:

$$\mu = E[X] = \sum_{k\in\{0,1\}} k \cdot p\,(X=k) = 1\cdot p + 0\cdot(1-p) = p$$

$$\sigma^2 = var\,(X) = \sum_{k\in\{0,1\}} (k-\mu)^2\,P\,(X=k) = (1-p)^2 p + (0-p)^2(1-p) = p(1-p)$$

Ejemplo: Se realiza el experimento aleatorio consistente en lanzar una moneda equlibrada y se define la variable aleatoria:

$$X = \begin{cases} 0 & \text{si sale cara} \\ 1 & \text{si sale cruz} \end{cases}$$

Entonces

$$X \approx Be\left(\frac{1}{2}\right)$$

La función de probabilidad en este caso es:

$$P(X=1) = \frac{1}{2};\quad P(X=0) = 1 - \frac{1}{2} = \frac{1}{2}$$

y la media y varianza:

$$\mu = p = \frac{1}{2};\quad \sigma^2 = p(1-p) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$$

3.3. Distribución Binomial $B(n, p)$

Definición: Una variable aleatoria X sigue una distribución *Binomial de parámetros n y p* si representa el número de éxitos obtenidos al realizar n repeticiones independientes de un experimento de Bernoulli, siendo p la probabilidad de éxito en cada experimento.

Obviamente sólo son posibles entre 0 y n éxitos. La función de probabilidad de esta variable es de la forma:

$$f(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \; k \in \{0, 1, 2, \ldots, n\}$$

La figura 1 muestra esta función de probabilidad para diversos valores de n y p

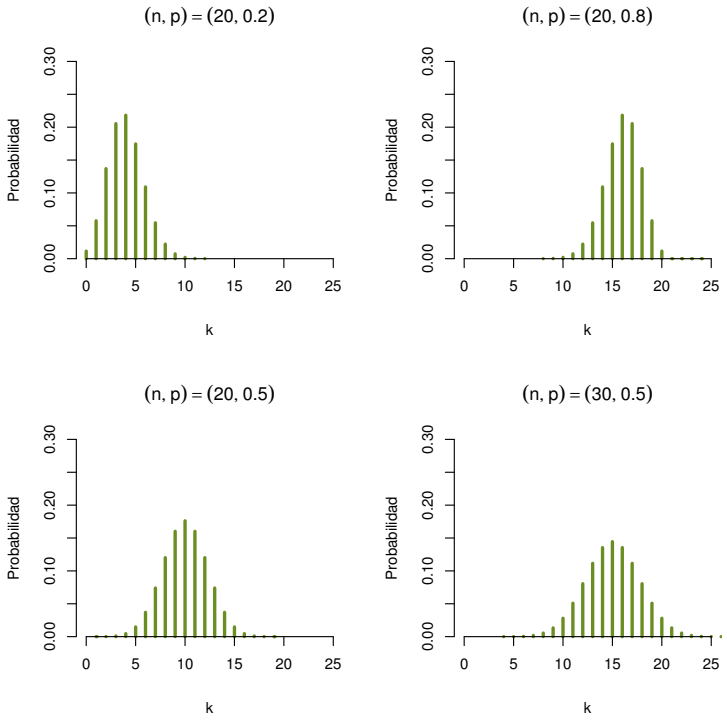

Figura 1: Función de probabilidad de la distribución binomial para diversos valores de n y p. La altura de cada línea representa la $P(X = k)$.

Esperanza y varianza: Por definición, si $X \approx B(n, p)$ entonces $X = X_1 + X_2 + \cdots + X_k$,

siendo las X_i variables de Bernoulli de parámetro p independientes. Por tanto:

$$\begin{aligned}\mu = E[X] = E[X_1 + X_2 + \cdots + X_k] = E[X_1] + E[X_2] + \cdots + E[X_k] = \\ = p + p + \cdots + p = np \\ \sigma^2 = var(X) = var(X_1 + X_2 + \cdots + X_k) = var(X_1) + var(X_2) + \cdots + var(X_k) = \\ = p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p)\end{aligned}$$

Ejemplo: Se sabe que en la puesta de huevos de una tortuga, la probabilidad de que una cría sea macho es 0.30 y de que sea hembra es 0.70. El sexo de cada cría es independiente del resto. Se dispone de una puesta de 10 huevos y se considera la variable X="*Número de machos en la puesta*". ¿Cuál es la probabilidad de que $X = 5$?

De la descripción de esta variable se deduce que $X \approx B(10, 0{,}3)$. Por tanto:

$$P(X = 5) = \binom{10}{5} 0{,}3^5 (1 - 0{,}3)^{10-5} = 0{,}103$$

Cálculo con R : El programa R dispone de varias funciones para el cálculo de probabilidades asociadas a la distribución binomial. Concretamente, si $X \approx B(n, p)$, utilizando R podemos:

- Calcular el valor de la función de probabilidad*:* $P(X = k)$=`dbinom(k,n,p)`
- Calcular el valor de la función de distribución: $P(X \leq k)$=`pbinom(k,n,p)`
- Calcular los cuantiles: $q_\alpha = \text{mín}\{x : F(x) \geq \alpha\}$=`qbinom(`$\alpha$`,n,p)`
- Generar m números aleatorios con distribución $B(n, p)$: `rbinom(m,n,p)`

Ejemplo: La siguiente sintaxis simula una muestra de 1000 valores de una distribución binomial de parámetros $n = 10$ y $p = 0{,}7$, y los representa en un diagrama de barras, junto a la representación gráfica de la función de probabilidad de la $B(10, 0{,}7)$ (figura 2). Asimismo se muestran las proporciones con que aparece cada valor k en la muestra y su correspondiente probabilidad teórica $P(X = k) = \binom{10}{k} 0{,}7^k (1 - 0{,}7)^{10-k}$. Como puede apreciarse, con este valor de n, las probabilidades teóricas son muy similares a las proporciones muestrales observadas.

```
> n=10
> p=0.7
> muestra=rbinom(1000,n,p)
> probabilidades=dbinom(0:n,n,p)
> proporciones=prop.table(table(muestra))
> par(mfrow=c(1,2))
> plot(0:n,probabilidades,type="h",lwd=3,col="olivedrab",ylab="Probabilidad",xlab
> barplot(proporciones,xlab="k",ylab="Proporcion",main="(b)")
> prop=numeric(11);for(k in 0:10) prop[k+1]=length(which(muestra==k))/1000
> data.frame(k=0:10,Prob=round(probabilidades,3),Prop.obs=prop)

    k  Prob Prop.obs
1   0 0.000    0.000
2   1 0.000    0.000
3   2 0.001    0.003
4   3 0.009    0.008
5   4 0.037    0.033
6   5 0.103    0.097
7   6 0.200    0.207
8   7 0.267    0.256
9   8 0.233    0.236
10  9 0.121    0.116
11 10 0.028    0.044

>
```

3.4. Distribución Geométrica $Geo\,(p)$.

Definición: una variable aleatoria X sigue una distribución *Geométrica de parámetro* p si representa el número de experimentos de Bernoulli sucesivos e independientes que acaban en fracaso antes de que ocurra el primer éxito. Su función de probabilidad es por tanto:

$$f\,(k) = P\,(X = k) = (1-p)^k\,p,\ \ k = 0, 1, 2, \ldots$$

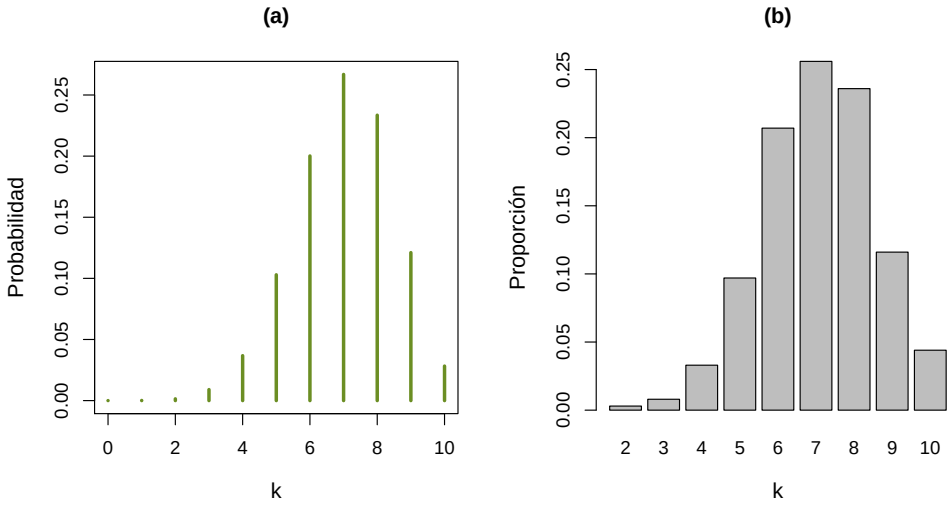

Figura 2: (a) Probabilidades correspondientes a la distribución $B\left(10,0{,}7\right)$ (b) Proporciones observadas en una muestra de tamaño $n = 1000$ de dicha distribución. Puede observarse la coincidencia entre ambas representaciones.

Esperanza y varianza:

$$\mu = E\left[X\right] = \sum_{k=0}^{\infty} k \cdot P\left(X = k\right) = \sum_{k=0}^{\infty} k\left(1-p\right)^{k} p = \frac{1-p}{p}$$

$$\sigma^2 = var\left(X\right) = \sum_{k=0}^{\infty} \left(k-\mu\right)^2 \cdot P\left(X = k\right) = \sum_{k=0}^{\infty} \left(k - \frac{1}{p}\right)^2 \left(1-p\right)^{k} p = \frac{1-p}{p^2}$$

Ejemplo: Sea X="*Número de lanzamientos de un dado equilibrado antes de que salga el primer 6*". Obviamente $X \approx Geo(\frac{1}{6})$. Así, por ejemplo, la probabilidad de que haya que lanzar el dado 9 veces antes del primer 6, sería:

$$P\left(X = 9\right) = \left(1 - \frac{1}{6}\right)^{9} \frac{1}{6} = 0{,}0323$$

El número esperado de veces que habría que lanzar el dado antes de que salga un 6 por primera vez sería $\mu = \frac{1-1/6}{1/6} = 5$

Cálculo con R : Si $X \approx Geo(p)$:

- Valor de la función de probabilidad: $P(X = k)$=`dgeom(k,p)`
- Valor de la función de distribución: $P(X \leq k)$=`pgeom(k,p)`
- Cuantiles: $q_\alpha = \text{mín}\{x : F(x) \geq \alpha\}$=`qgeom(`$\alpha$`,p)`
- Generación de m números aleatorios con distribución $Geo(p)$: `rgeom(m,p)`

Ejemplo: Para calcular con R la probabilidad buscada en el ejemplo anterior ejecutamos:

```
> dgeom(9,1/6)

[1] 0.03230112

>
```

3.5. Distribución Hipergeométrica $H(n, N, N_E)$

Definición: Supongamos que se dispone de una población finita de tamaño N, que está dividida en dos grupos: N_E "*éxitos*" y $N - N_E$ "*fracasos*". Una variable aleatoria X sigue una distribución hipergeométrica si representa el número de éxitos obtenidos al extraer al azar y sin reemplazamiento n objetos de esta población. La función de probabilidad de esta variable aleatoria es:

$$P(X = k) = \frac{\binom{N_E}{k}\binom{N-N_E}{n-k}}{\binom{N}{n}}, \quad x = \text{máx}\{0, n - (N - N_E)\}, \ldots, \text{mín}\{N_E, n\}$$

Esperanza y varianza: Si llamamos $p = \frac{N_E}{N}$ (probabilidad de éxito cuando se extrae un único objeto)

$$\mu = \frac{n \cdot N_E}{N} = np$$
$$\sigma_X^2 = \frac{N_E(N - N_E)\, n\,(N - n)}{N^2(N - 1)} = np(1 - p)\frac{(N - n)}{(N - 1)}$$

Nota: Es evidente que si el experimento donde surge la distribución hipergeométrica se realizara con reemplazamiento, la variable X considerada tendría distribución binomial. Debe señalarse que, aún habiendo reemplazamiento, si N es muy grande en comparación con n, resultaría muy difícil que un mismo objeto de la población fuera elegido aleatoriamente dos ó más veces, lo que es equivalente a que no haya reemplazamiento. Ello significa que la distribución hipergeométrica se va pareciendo cada vez más a la binomial a medida que N crece con respecto a n. Puede observarse incluso en las expresiones de la esperanza y la varianza, que si N se hace grande y n es relativamente pequeño, se obtienen los mismos valores que en la binomial.

Ejemplo: De una urna en la que hay 10 bolas blancas y 5 bolas negras, se extraen 8 bolas sin reemplazamiento. ¿Cual es la probabilidad de que entre estas ocho haya 4 bolas negras? Si llamamos: $X =$ "*número de bolas negras en la muestra*" entonces $X \approx H\,(8, 15, 5)$ y:

$$P(X = 4) = \frac{\binom{5}{4}\binom{15-5}{8-4}}{\binom{15}{8}} = \frac{\binom{5}{4}\binom{10}{4}}{\binom{15}{8}} = 0{,}1632$$

Cálculo con R : la sintaxis a emplear con R para calcular probabilidades asociadas a la distribución geométrica es nuevamente similar a la ya vista en las distribuciones anteriores. Si $X \approx H\,(n, N, N_E)$ y llamamos $N_F = N - N_E$:

- Valor de la función de probabilidad: $P\,(X = k)$=`dhyper(k,NE,NF,n)`
- Valor de la función de distribución: $P\,(X \leq k)$=`phyper(k,NE,NF,n)`
- Cuantiles: $q_\alpha = \text{mín}\,\{x :\ F\,(x) \geq \alpha\}$=`qhyper(`$\alpha$`,NE,NF,n)`
- Generación de m números aleatorios con esta distribución: `rhyper(m,ME,NF,n)`

Para obtener la probabilidad del ejemplo anterior utilizando R emplearíamos la función:

```
> dhyper(4,10,5,8)

[1] 0.1631702
```

Aplicación a la estimación de un tamaño poblacional. (**Método de captura - recaptura**) Una aplicación clásica de la distribución hipergeométrica al campo de las

ciencias biológicas es la siguiente: supongamos que se desea estimar aproximadamente el número de peces que hay en un lago. Para ello realizamos una captura inicial de N_E peces (se capturan al azar, a lo largo de toda la extensión del lago), los marcamos y los devolvemos al agua. De esta forma ahora tenemos en el lago un total de N peces (N es desconocida) de los que N_E *sabemos* que están marcados. Realizamos una segunda captura, ahora de n peces y contamos cuántos hay marcados en esta recaptura. Obviamente el número de peces marcados en la recaptura sigue una distribución hipergeométrica $H(n, N, N_E)$ por lo que el número esperado de peces marcados en dicha recaptura es $n\frac{N_E}{N}$. Si en realidad se observaron k peces marcados, igualamos ambas expresiones (esto es, suponemos que se captura exactamente lo que se esperaba capturar):

$$k = n\frac{N_E}{N}$$

de donde se obtiene el valor de N:

$$\hat{N} = n\frac{N_E}{k}$$

Obviamente este valor de N es una aproximación, ya que la premisa de que lo que se esperaba pescar es lo que se pesca, no tiene que ser válida exactamente. Este es el punto de arranque para los diseños de muestreo más sofisticados que se emplean en la estimación de tamaños poblacionales.

3.6. Distribución de Poisson $P(\lambda)$

Las tortugas marinas suelen cavar sus nidos en la zona supramareal de playas fácilmente accesibles. Supongamos que en determinada playa se ha observado que las posiciones de los nidos se reparten completamente al azar en esa zona, con una densidad media de ϑ nidos por km^2. ¿Cómo podríamos calcular la probabilidad de que en una extensión de S km^2 se encuentren k nidos?

Por simplicidad supongamos que dicha región es rectangular, y que sobre la misma superponemos una malla tal como se muestra en la figura 3. La malla es lo suficientemente fina como para que en cada cuadrícula quepa como mucho un único nido. Las posiciones de los nidos se han marcado mediante puntos en el gráfico resultante. De esta forma el problema de determinar la probabilidad de que en esta zona haya k nidos es equivalente a calcular la probabilidad de que k cuadros de la malla estén ocupados por un nido. Si suponemos que en total la malla tiene n cuadros, que la probabilidad de que un cuadro arbitrario esté ocupado

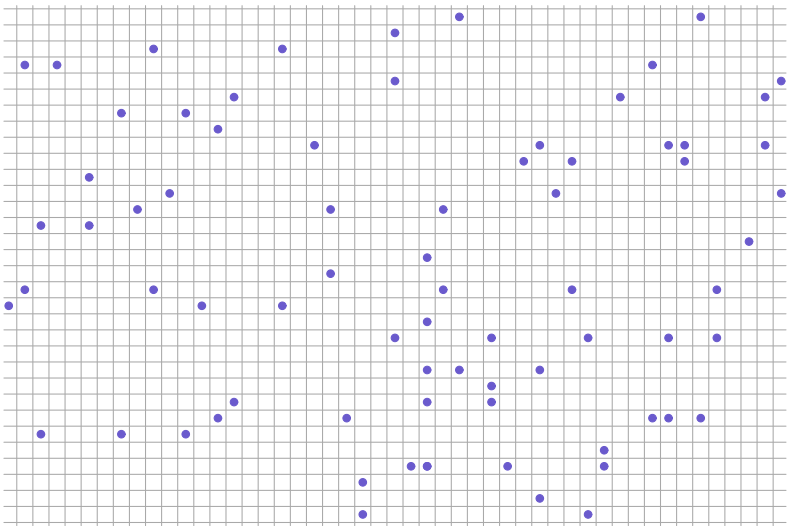

Figura 3: Región rectangular de superficie S situada en la zona supramareal de una playa en la que hay nidos de tortuga. Sobre esta región se ha superpuesto una malla regular y se han marcado las posiciones de los nidos.

es p, y que los cuadros se ocupan independientemente unos de otros (esta última hipótesis es razonable si los nidos están repartidos completamente al azar, es decir, si no tienden a estar concentrados en un único sitio ni a estar sistemáticamente separados unos de otros), entonces la variable $X =$"*número de cuadros ocupados por nidos en la malla*" sigue una distribución binomial $B(n, p)$ donde:

- n es un número muy grande (hay muchos cuadros en la malla).
- p es un número muy pequeño (entre tantos cuadros, la probabilidad de que haya un nido en un cuadro concreto es minúscula).
- Como hay una densidad media de ϑ nidos por km^2 y la región estudiada mide S km^2, el número esperado de nidos en la región es $\lambda = \vartheta S$. Como el valor esperado de la binomial es $n \cdot p$, debe ocurrir entonces que $n \cdot p{=}\lambda$ (de donde $p = \frac{\lambda}{n}$)

Así pues para calcular la probabilidad de k nidos utilizando esta aproximación binomial

tendríamos:

$$
\begin{aligned}
P\left(X=k\right) &= \lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \\
&= \lim_{n\to\infty} \frac{n(n-1)...(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n-k} = \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} 1 \cdot \left(1-\frac{1}{n}\right) \cdot ... \cdot \left(1-\frac{k-1}{n}\right) \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} = \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}
$$

Definición: Una variable aleatoria discreta X sigue una *distribución de Poisson* de parámetro λ, si su función de probabilidad es de la forma:

$$P(X=x) = \frac{\lambda^k}{k!} e^{-\lambda}; \;\; k = 0, 1, 2, 3, ...$$

siendo λ un valor real positivo. La figura 4 muestra la forma de esta función de probabilidad para diversos valores de λ.

En el ejemplo anterior, el número de nidos de tortuga en una región de superficie S sigue una distribución de Poisson de parámetro $\lambda = \vartheta S$, siendo ϑ el número medio de nidos por unidad de superficie.

En general, la distribución de Poisson constituye un modelo de probabilidad adecuado para aquellas variables aleatorias que cuentan el número de puntos que se encuentran en cierto espacio continuo, siempre y cuando estos puntos se encuentren repartidos completamente al azar. A modo de ejemplo podemos citar:

- Número de estrellas en cierta porción del firmamento (los puntos son las estrellas y el espacio continuo es la región estelar observada).
- Número de copépodos en un volumen de agua determinado (los puntos son los copépodos y el espacio continuo donde se encuentran es el volumen de agua).

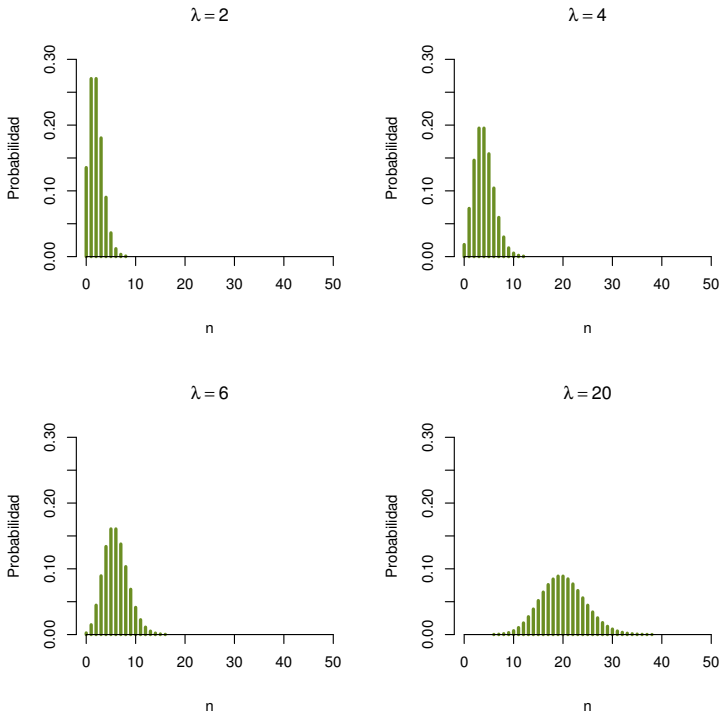

Figura 4: Función de Probabilidad de la distribución de Poisson para varios valores de λ. La altura de cada línea vertical representa la $P(X = k)$

- Número de llamadas telefónicas recibidas en una centralita a lo largo de un día (los puntos son los instantes en que se producen las llamadas, y el espacio continuo en que se sitúan estos puntos es el tiempo transcurrido entre las 0 y las 24 horas).

Esperanza y varianza: Puede probarse que:

$$E[X] = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \lambda$$

$$var(X) = E\left[X^2\right] - E[X]^2 = \sum_{k=0}^{\infty} k^2\frac{\lambda^k}{k!}e^{-\lambda} - \lambda^2 = \lambda$$

Este resultado era de esperar, ya que X es el límite de una binomial cuya esperanza es $np = \lambda$ y cuya varianza es $np(1-p) = \lambda$ (ya que $np = \lambda$ y $p \to 0$, por lo que $(1-p) \cong 1$)

Ejemplo: Si la densidad de nidos de tortuga en una playa es de 0.01 nidos por m^2 (esto es, un nido cada 100 m^2), ¿cuál es la probabilidad de que una zona de 1000 m^2 de extensión haya 8 nidos?

En este ejemplo $\lambda = \vartheta S = 0{,}01 \cdot 1000 = 10$. Aplicando la distribución de Poisson:

$$P(X = 8) = \frac{10^8}{8!} e^{-10} = 0{,}113$$

Cálculo con R :

- Valor de la función de probabilidad: $P(X = k)$=`dpois(k,`λ`)`
- Valor de la función de distribución: $P(X \leq k)$=`ppois(k,`λ`)`
- Cuantiles: $q_\alpha = \text{mín}\{x : F(x) \geq \alpha\}$=`qpois(`$\alpha$`,`$\lambda$`)`
- Generación de m números aleatorios con distribución $P(\lambda)$: `rpois(m,`λ`)`

Continuación del ejemplo: En el ejemplo anterior, si queremos calcular la probabilidad de que en una región de 1 km^2 de extensión haya más de 8 nidos:

$$P(X > 8) = 1 - P(x \leq 8) = 1 - \texttt{ppois(8,10)} = 1\text{-}0.333 = 0.667$$

La probabilidad de que en esa región haya entre 8 y 12 nidos puede hallarse como:

$$\begin{aligned} P(8 \leq X \leq 12) &= P(X \leq 12) - P(X \leq 7) = \\ &= \texttt{ppois(12,10)-ppois(7,10)} = \\ &= 0{,}792 - 0{,}22 = 0{,}572 \end{aligned}$$

Aproximación de la distribución binomial: Hemos obtenido la distribución de Poisson como límite de una binomial cuando $n \to \infty$ y $p \to 0$. La distribución de Poisson constituye en general una buena aproximación de la binomial $B(n, p)$ cuando $n > 20$ y $p < 0{,}05$, en cuyo caso $B(n, p) \cong P(\lambda)$, con $\lambda = n \cdot p$.

Para entender el sentido de esta aproximación consideremos el siguiente ejemplo: se sabe que el 1 % de los huevos de tortuga depositados en una playa son depredados por cangrejos. Si entre cuatro nidos totalizan 280 huevos, ¿cuál es la probabilidad de que ninguno sea depredado por cangrejos?.

Llamando X="Número de huevos depredados en los cuatro nidos", tendríamos que $X \approx B\,(280, 0{,}01)$. La probabilidad de que ningún huevo sea depredado sería:

$$P\,(X = 0) = (1 - 0{,}01)^{280} = 0{,}99^{280} = 0{,}05996$$

Muchas calculadoras no son capaces de realizar este cálculo (aquí lo hemos obtenido con R mediante `dbinom(0,280,0.01)`). La aproximación de Poisson nos indica que $X \approx B\,(280, 0{,}01) \cong P\,(280 \cdot 0{,}01) = P\,(2{,}8)$. Si utilizamos la distribución de Poisson para calcular la probabilidad pedida obtenemos

$$P\,(X = 0) = \frac{2{,}8^0}{0!} e^{-2{,}8} = e^{-2{,}8} = 0{,}06081$$

que se diferencia del verdadero valor en 0,00085, por lo que el error de aproximación es inferior a una milésima. Vemos, pues, que la aproximación mediante la distribución de Poisson funciona razonablemente bien, y es aconsejable su uso cuando no se dispone de medios informáticos avanzados.

Aditividad de la distribución de Poisson. Si dos variables aleatorias independientes X_1 y X_2 siguen sendas distribuciones de Poisson, $X_1 \approx P(\lambda_1)$ y $X_2 \approx P(\lambda_2)$, entonces $X_1 + X_2 \approx P(\lambda_1 + \lambda_2)$. En general, si $X_1, X_2, ..., X_n \approx P(\lambda)$, y además son independientes, entonces $\sum_{i=1}^{n} X_i \approx P(n\lambda)$

4. Principales distribuciones de probabilidad continuas.

4.1. Distribución uniforme $U\,(a, b)$.

Definición: Una variable aleatoria X sigue una *distribución uniforme* en el intervalo real (a, b), si su función de densidad es constante sobre ese intervalo:

$$f\,(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & x \notin (a, b) \end{cases}$$

En la práctica esta distribución corresponde a variables del tipo: X = *"Resultado de elegir al azar un valor del intervalo (a,b)"* cuando la probabilidad de que el valor elegido caiga en un intervalo de amplitud ℓ dentro de (a, b) es siempre la misma independientemente de la posición de dicho intervalo.

Esperanza y varianza:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_a^b x \frac{1}{b-a} dx = \left[\frac{1}{b-a}\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

$$var(X) = E\left[X^2\right] - E[X]^2 = \int_a^b x^2 \frac{1}{b-a} dx - \left(\frac{a+b}{2}\right)^2 = \frac{1}{12}(b-a)^2$$

Ejemplo: la variable aleatoria $X =$ "*Distancia, medida desde el extremo inicial, a la que se rompe una cuerda homogénea de 1 metro cuando se tira con igual fuerza de ambos extremos*" que ya hemos visto en el capítulo anterior sigue una distribución $X \approx U(0,1)$.

Cálculo con R :

- Valor de la función de densidad $f(x)$=`dunif(x,a,b)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`punif(x,a,b)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qunif(`$\alpha$`,a,b)`
- Generación de n números aleatorios con distribución $U(a,b)$: `runif(n,a,b)`

4.2. Distribución exponencial $\exp(\eta)$.

Definición: una variable aleatoria X sigue una *distribución exponencial* de parámetro η si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{1}{\eta} e^{-\frac{1}{\eta}x},\ x \geq 0$$

En la práctica, esta distribución aparece asociada a variables que miden la distancia entre sucesos puntuales que se dispersan completamente al azar en un medio continuo y cuyo número tiene, por tanto, distribución de Poisson. En efecto, supongamos por simplicidad que el medio continuo considerado es el tiempo y que estamos contando el número de eventos que ocurren hasta un instante t. Si el número de tales eventos sigue una distribución de Poisson, siendo λ el número esperado de eventos por unidad de tiempo, ello significa que $\eta = \frac{1}{\lambda}$ es el tiempo esperado entre dos cualesquiera de tales

sucesos. Si llamamos Y_t="Número de sucesos ocurridos en un intervalo de duración t" entonces $Y_t \approx P(\lambda t) = P\left(\frac{1}{\eta}t\right)$. Si acaba de ocurrir uno de estos sucesos, y llamamos X al tiempo que transcurre hasta que ocurre el siguiente, entonces:

$$P(X \geq t) = P(Y_t = 0) = \frac{(\lambda t)^0}{0!}e^{-\lambda t} = \frac{\left(\frac{1}{\eta}t\right)^0}{0!}e^{-\frac{1}{\eta}t} = e^{-\frac{1}{\eta}t}$$

(ya que $X \geq t$ significa que el siguiente suceso ocurre después de t, o lo que es lo mismo, que en un intervalo de duración t no ha ocurrido ningún suceso, esto es $Y_t = 0$). Por tanto:

$$F(t) = P(X \leq t) = 1 - e^{-\frac{1}{\eta}t}$$

de donde:

$$f(t) = F'(t) = \frac{1}{\eta}e^{-\frac{1}{\eta}t},\ t \geq 0$$

La figura 5 muestra la forma de la distribución exponencial para varios valores del parámetro η.

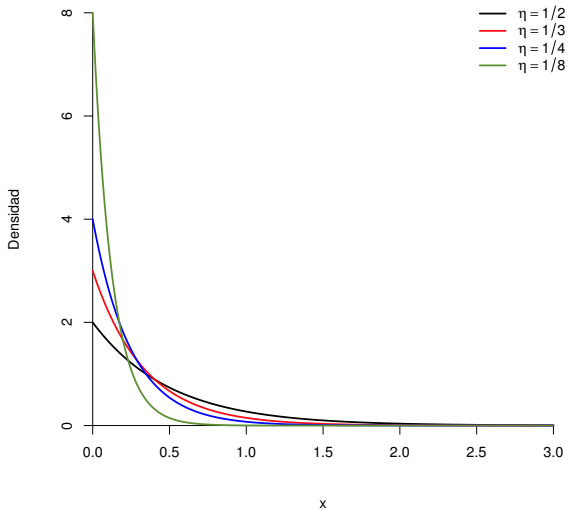

Figura 5: Función de densidad de la distribución exponencial para varios valores de η.

Esperanza y varianza:

$$E[X] = \int_0^\infty \frac{1}{\eta}xe^{-\frac{1}{\eta}x}dx = \eta$$

$$var(X) = E\left[X^2\right] - E[X]^2 = \int_0^\infty \frac{1}{\eta}x^2e^{-\frac{1}{\eta}x}dx - \left(\frac{1}{\eta}\right)^2 = \eta^2$$

Ejemplo: El tiempo que transcurre entre la caída de dos rayos sucesivos durante la fase central de una tormenta tropical sigue una distribución exponencial de parámetro 2.5 segundos. ¿Cuál es la probabilidad de que entre la caída de dos rayos sucesivos transcurran como mucho 3 segundos? ¿Cuál es el tiempo esperado que transcurre entre rayos sucesivos?

Sea X = "*Tiempo transcurrido entre dos rayos sucesivos*"$\approx exp(2{,}5)$. La probabilidad pedida es entonces:

$$P(X \leq 3) = 1 - e^{-\frac{1}{2,5}\cdot 3} = 1 - e^{-1,2} = 0{,}699$$

Dado que en una distribución exponencial el valor esperado coincide con su parámetro, el tiempo esperado entre rayos sucesivos es $E[X] = \eta = 2{,}5$ segundos.

Cálculo con R : Nótese que por defecto R espera recibir como parámetro el valor $1/\eta$ que recibe el nombre de *rate* (tasa).

- Valor de la función de densidad*:* $f(x)$=`dexp(x,`$1/\eta$`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pexp(x,`$1/\eta$`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qexp(`$\alpha$`,`$1/\eta$`)`
- Generación de n números aleatorios con distribución $exp(\lambda)$: `rexp(n,`$1/\eta$`)`

Así, el cálculo de la probabilidad del ejemplo anterior en R sería:

$$P(X \leq 3) = \texttt{pexp(3,1/2.5)} = 0{,}699$$

Falta de memoria de la distribución exponencial. La distribución exponencial tiene una propiedad característica que suele denominarse "*falta de memoria*". Si X es el tiempo entre dos ocurrencias consecutivas de un fenómeno, la *falta de memoria* significa que:

$$P(X \geq t + s \,|\, X \geq s) = P(X \geq t)$$

es decir, si desde la ocurrencia anterior ha transcurrido ya un tiempo s, la probabilidad de que aún falte un tiempo adicional t hasta la próxima ocurrencia es independiente de s. Para entender este enunciado pensemos los siguientes ejemplos:

- Nos encontramos en una estación de metro esperando por el siguiente tren; la linea que esperamos es muy puntual y por término medio pasa un tren cada 10 minutos. Si el último tren pasó hace 9 minutos, podemos estar razonablemente seguros de que el tiempo que aún nos queda por esperar es del orden de 1 minuto. Podemos decir que el tiempo entre llegadas de trenes "*tiene memoria*": el tiempo transcurrido desde la última llegada nos informa sobre el tiempo que aún falta hasta la siguiente.
- En nuestra ciudad cae un premio grande de la lotería por término medio una vez cada 10 años. Si el último de estos premios cayó hace 9 años, eso no nos dice nada sobre cuantos años han de transcurrir aún hasta que vuelva a tocar un premio grande en la ciudad. El tiempo entre premios de la lotería *"no tiene memoria*": el tiempo transcurrido desde el último premio no da ninguna información sobre el tiempo que aún falta hasta el siguiente.

Es fácil comprobar la falta de memoria de la distribución exponencial:

$$P\left(X \geq t+s \,|\, X \geq s\right) = \frac{P\left(\{X \geq t+s\} \cap \{X \geq s\}\right)}{p\left(X \geq s\right)} = \\ = \frac{P\left(X \geq t+s\right)}{p\left(X \geq s\right)} = \frac{e^{-\frac{1}{\eta}(t+s)}}{e^{-\frac{1}{\eta}s}} = e^{-\frac{1}{\eta}t} = P\left(X \geq t\right)$$

Esta propiedad resulta útil para decidir si la distribución exponencial puede ser un buen modelo para el comportamiento de una variable de nuestro interés: podría serlo para el tiempo transcurrido entre premios de la lotería, pero desde luego no lo es para el tiempo entre trenes de una linea de metro.

4.3. Distribución de Weibull $W\left(\kappa, \eta\right)$.

Definición: Una variable aleatoria X sigue una *distribución de Weibull* con parámetro de forma κ y parámetro de escala η si su función de distribución es de la forma:

$$F\left(x\right) = 1 - \exp\left(-\left(x/\eta\right)^{\kappa}\right), \; x \geq 0$$

Su función de densidad es:

$$f\left(x\right) = \frac{\kappa}{\eta}\left(\frac{x}{\eta}\right)^{\kappa-1} \exp\left(-\left(x/\eta\right)^{\kappa}\right), \; x \geq 0$$

En el caso particular de que $\kappa = 1$, la distribución de Weibull coincide con una exponencial de parámetro η.

La distribución de Weibull se utiliza con frecuencia para modelar el tiempo (aleatorio) que transcurre entre dos sucesos de interés, en particular cuando el tiempo transcurrido "*tiene memoria*" en el sentido apuntado más arriba. Así, por ejemplo, suele utilizarse:

- Para modelar la supervivencia: tiempo que sobreviven los enfermos con determinado tratamiento; tiempo que sobreviven las células en un cultivo; tiempo que dura un fenómeno meteorológico.
- Para modelar la fiabilidad: tiempo que dura un componente electrónico, mecánico, etc. en función de su edad y condiciones de uso.
- Para modelar tiempo entre eventos climatológicos: tiempo entre tormentas o ciclones, tiempo entre periodos fríos o cálidos.
- Para modelar tiempo entre determinados fenómenos geofísicos: tiempo entre réplicas de un terremoto, tiempo entre erupciones volcánicas.

Otras aplicaciones de la distribución de Weibull, dado el perfil de su función de densidad, son el modelado de la altura de ola, la velocidad de corriente marina o la velocidad del viento.

La figura 6 muestra la forma de la función de densidad de la distribución de Weibull para varios valores de κ y η.

Esperanza y varianza:

$$\mu = E[X] = \int_0^\infty x f(x)\, dx = \eta \cdot \Gamma\left(1 + \frac{1}{\kappa}\right)$$
$$\sigma^2 = var(X) = \eta^2 \left[\Gamma\left(1 + \frac{2}{\kappa}\right) - \left(\Gamma\left(1 + \frac{1}{\kappa}\right)\right)^2\right]$$

siendo $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$ la función gamma de Euler, que cumple las siguientes propiedades, útiles para el cálculo de sus valores:

1. $\Gamma(a) = (a-1)\Gamma(a-1)$
2. Si $n \in \mathbb{N}: \ \Gamma(n) = (n-1)!$

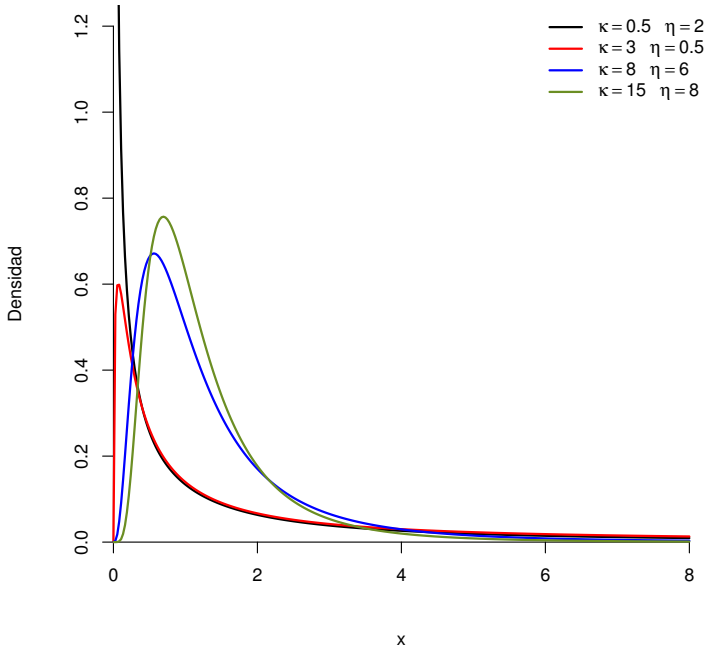

Figura 6: Función de densidad de la distribución de Weibull para varios valores de los parámetros κ y η.

La función gamma de Euler se encuentra implementada en R : $\Gamma(a)$ =`gamma(a)`

Cálculo con R :

- Valor de la función de densidad: $f(x)$=`dweibull(x,`κ`,`η`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pweibull(x,`κ`,`η`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qweibull(`$\alpha$`,`$\kappa$`,`$\eta$`)`
- Generación de n números aleatorios con distribución $exp(\lambda)$: `rweibull(n,`κ`,`η`)`

4.4. Distribución Gamma $\mathcal{G}(\kappa, \eta)$

Definición: Una variable aleatoria X sigue una *distribución gamma* con parámetro de forma κ y parámetro de escala η si su función de densidad es de la forma:

$$f(x) = \frac{1}{\eta^\kappa \Gamma(\kappa)} x^{\kappa-1} \exp(-x/\eta) \ : \ x \geq 0$$

siendo $\Gamma(a)$ la función gamma de Euler. En el caso particular de que $\kappa = 1$, la distribución gamma se reduce a una exponencial de parámetro η.

En la práctica la distribución gamma suele utilizarse para modelar problemas como los ya descritos para la distribución de Weibull. La figura muestra la forma de la función de densidad de la distribución gamma para varios valores de sus parámetros.

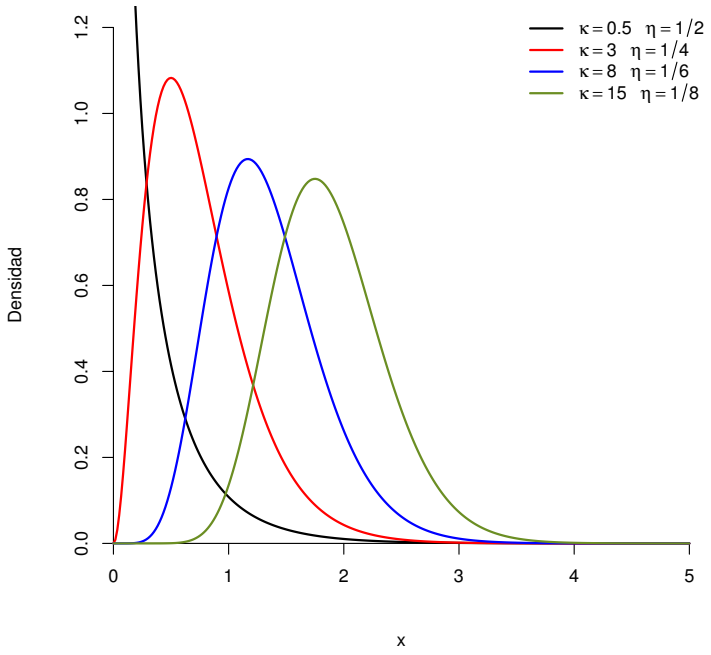

Figura 7: Función de densidad de la distribución Gamma para varios valores de κ y η.

Esperanza y varianza:

$$\mu = E[X] = \kappa \cdot \eta$$
$$\sigma^2 = var(X) = \kappa \cdot \eta^2$$

Cálculo con R : la notación es similar a las distribuciones anteriores. Nótese que por defecto R espera recibir como parámetro el inverso del factor de escala $1/\eta$ que recibe el nombre de *rate* (tasa).

- Valor de la función de densidad*:* $f(x)$=`dgamma(x,`κ`,`$1/\eta$`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pgamma(x,`κ`,`$1/\eta$`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qgamma(`$\alpha$`,`$\kappa$`,`$1/\eta$`)`
- Generación de n números aleatorios con distribución $exp(\lambda)$: `rgamma(n,`κ`,`$1/\eta$`)`

La siguiente proposición resulta de interés en las aplicaciones:

Proposición. *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes y con distribución exponencial de parámetro η. Entonces $\sum_{i=1}^{n} X_i$ sigue una distribución gamma $\mathcal{G}(n, \eta)$.*

4.5. Distribución Normal $N(\mu, \sigma)$

Definición: Una variable aleatoria X sigue una *distribución Normal* de parámetros μ (media) y σ (desviación típica) si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ x \in \mathbb{R}$$

Nótese que $f(x)$ es una función simétrica respecto a x, esto es $f(x) = f(-x)$. La figura 8 muestra la forma de esta función de densidad, que corresponde a la conocida *campana de Gauss.*

En la práctica, la distribución normal aparece asociada a variables aleatorias que se comportan de tal manera que lo más probable es observar valores en torno a la media; y que los valores cada vez más alejados de la media, bien sea hacia arriba o hacia abajo, van siendo progresivamente más difíciles de observar. Muchas variables biológicas se comportan aproximadamente de esta forma: la talla, el peso, la temperatura corporal, etc. También se comportan de esta manera los errores de medida. La distribución normal es una de las más frecuentes en la naturaleza, lo que se justifica de manera teorica por la acción del teorema central del límite, que veremos más adelante. Dicho de una manera intuitiva, este teorema indica que si una variable es el resultado de la suma de efectos de muchas otras variables independientes, la variable resultante tiene necesariamente distribución normal. Si se piensa que las variables que hemos citado –peso,

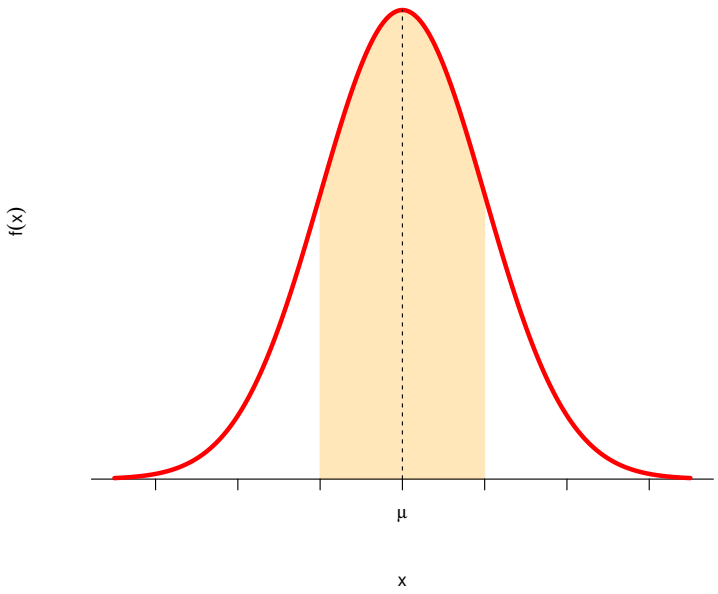

Figura 8: Función de densidad de la distribución normal. Está centrada en la media (μ), valor en torno al cual se concentra la mayor parte de la probabilidad.

talla, errores de medida, ...– son precisamente el efecto de muchas pequeñas causas que operan independientemente se entiende por qué cabe esperar que su distribución de probabilidad sea precisamente normal.

La figura 9 muestra la forma de la función de densidad de la distribución normal con media $\mu = 0$ para varios valores de σ.

Esperanza y varianza: hemos definido la distribución normal precisamente a partir de sus esperanza y varianza. No obstante se puede comprobar resolviendo las integrales correspondientes, que tal como se ha definido la función de densidad $f(x)$ se verifica que:

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

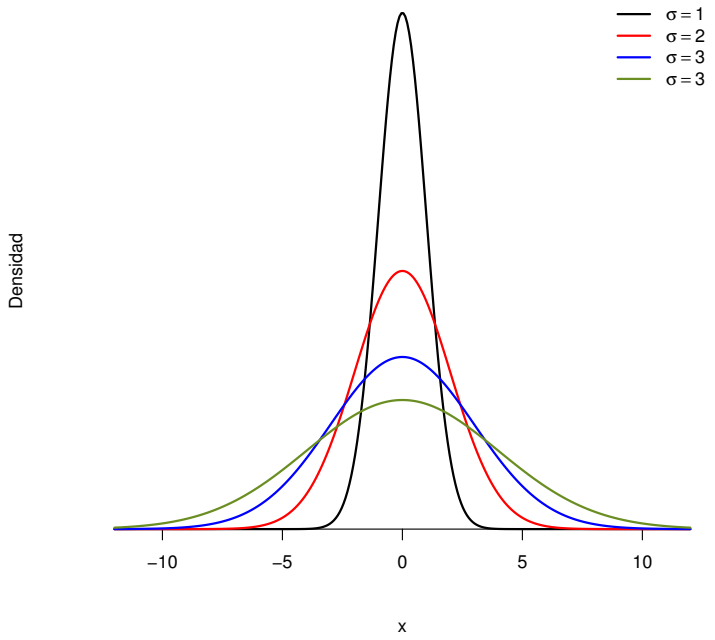

Figura 9: Función de densidad de la distribución normal de media $\mu = 0$ para varios valores de σ.

Distribución normal tipificada: El caso particular de la distribución normal con $\mu = 0$ y $\sigma = 1$ se conoce con el nombre de *distribución normal tipificada o estándar* $N(0,1)$. Si $Z \approx N(0,1)$ *denotaremos como* $\Phi(z) = P(Z \leq z)$.

Una de las dificultades prácticas que presenta la distribución normal es que su función de densidad no tiene una función primitiva, lo que significa que las probabilidades

$$P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du$$

deben calcularse numéricamente. Si bien R calcula esta probabilidad mediante la función `pnorm(x,`μ`,`σ`)` (y existen muchos otros programas que lo hacen, así como la mayoría de las calculadoras científicas), es usual calcularla mediante el uso de tablas. El interés de la distribución normal tipificada es que es la única cuyas probabilidades se encuentran tabuladas.

Uso de la tabla de la distribución normal tipificada. Esta tabla sólo proporciona probabilidades de la forma $P(Z \geq z)$, siendo $Z \approx N(0,1)$, correspondientes al área sombreada en la figura 10. Para aprender a manejar esta tabla, supongamos que queremos

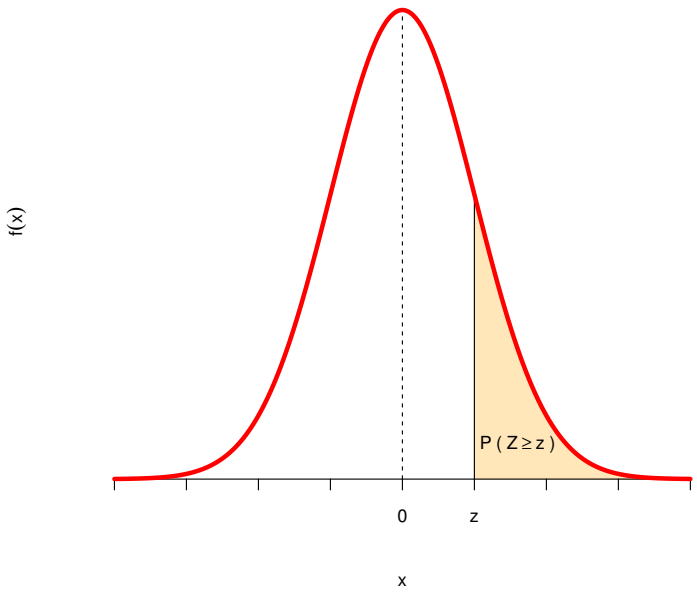

Figura 10: La tabla de la distribución $N(0,1)$ proporciona, para diversos valores de z, el valor de $P(Z \geq z)$, correspondiente al área sombreada.

calcular la probabilidad $P(Z \geq 2{,}16)$. Para ello simplemente separamos el número 2,16 en dos partes: una con la parte entera y las décimas (2,1), y otra con las centésimas (0,06). A continuación vamos a la tabla y buscamos el punto de cruce de la fila etiquetada como 2,1 y la columna etiquetada como 0,06, donde encontramos el valor 0,01539, que corresponde a la probabilidad buscada.

Si queremos calcular probabilidades de la forma $P(Z \leq z)$ simplemente utilizamos que $P(Z \leq z) = 1 - P(Z \geq z)$ y procedemos igual que antes. Si queremos calcular probabilidades para valores negativos de la variable basta tener en cuenta que la distribución normal es simétrica y por tanto que $P(Z \leq -z) = P(Z \geq z)$. Por último la tabla nos indica que si $z \geq 4$ entonces $P(Z \geq z) \cong 0$.

¿Cómo podemos utilizar esta tabla si queremos calcular probabilidades de una $N(\mu, \sigma)$ con $\mu \neq 0$ y $\sigma \neq 1$? En tal caso aplicaríamos el siguiente resultado:

Proposición: *Si* $X \approx N(\mu, \sigma)$ *entonces* $Z = \frac{X-\mu}{\sigma} \approx N(0,1)$

El significado de esta proposición es fácil de entender: los valores de Z se obtienen a partir de los de X por *desplazamiento* (al restar μ) y *cambio de escala* (al dividir por σ). Ninguna de estas transformaciones cambia la *forma* de la función

de densidad; por tanto Z también debe seguir una distribución normal. Asimismo, la simple aplicación de las propiedades de la media y la varianza permite ver de inmediato que $E[Z] = \frac{1}{\sigma}E[X-\mu] = \frac{1}{\sigma}(E[X]-\mu) = \frac{1}{\sigma}(\mu-\mu) = 0$ y $var(Z) = \frac{1}{\sigma^2}var(X-\mu) = \frac{1}{\sigma^2}var(X) = \frac{1}{\sigma^2}\sigma^2 = 1$.

Para calcular entonces probabilidades de la forma $P(X \geq x)$ cuando $X \approx N(\mu, \sigma)$ con $\mu \neq 0$ y $\sigma \neq 1$ bastará con tener en cuenta que

$$P(X \geq x) = P\left(\frac{X-\mu}{\sigma} \geq \frac{x-\mu}{\sigma}\right) = P\left(Z \geq \frac{x-\mu}{\sigma}\right)$$

y localizar el último valor directamente en la tabla. Así, por ejemplo, si $X \approx N(20, 4)$, para calcular $P(X \geq 25)$ procederíamos del siguiente modo:

$$P(X \geq 25) = P\left(\frac{X-20}{4} \geq \frac{25-20}{4}\right) = P\left(Z \geq \frac{5}{4}\right) = P(Z \geq 1{,}25) = 0{,}10565$$

donde hemos encontrado el valor 0,10565 en el cruce de la fila 1,2 con la columna 0,05 de la distribución normal estándar.

Cuantiles de la $N(0,1)$ utilizando la tabla. Un problema frecuente en la práctica es la determinación de cuantiles de la distribución $N(0,1)$. Recordemos que el cuantil α de una variable aleatoria X es el valor q_α tal que $P(X \leq q_\alpha) = \alpha$. En el caso de la distribución normal estándar llamaremos z_α al cuantil $q_{1-\alpha}$; esto es, z_α es el valor tal que $P(Z \leq z_\alpha) = 1-\alpha$, o lo que es lo mismo, $P(Z > z_\alpha) = \alpha$.

Para calcular los cuantiles utilizando la tabla habremos de proceder a la inversa que para el cálculo de probabilidades; por ejemplo, supongamos que deseamos localizar el valor $z_{0,025}$ (es decir, el cuantil 0,975). Buscamos el valor 0,025 (o el que más se le aproxime) en el interior de la tabla; en este caso encontramos el 0,025 en el cruce de la fila 1,9 con la columna 0,06. Por tanto $z_{0,025} = 1{,}96$.

Cálculo con R :

- Valor de la función de densidad: $f(x)$=`dnorm(x,`μ`,`σ`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pnorm(x,`μ`,`σ`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qnorm(`$\alpha$`,`$\mu$`,`$\sigma$`)`
- Generación de n números aleatorios con distribución $N(\mu, \sigma)$: `rnorm(n,`μ`,`σ`)`

Podemos utilizar R para calcular las probabilidades que hemos visto en los ejemplos anteriores. En el caso particular de la normal estándar no es preciso especificar $\mu = 0$ y $\sigma = 1$. Así:

- $P(Z \geq 2{,}16)$=$1 - P(Z \leq 2{,}16)$ =`1-pnorm(2.16)`=0.01539
- si $X \approx N(20, 4)$, entonces $P(X \geq 25)$=`1-pnorm(25,20,4)`=0.10565

Asimismo, el cálculo de los cuantiles es muy simple con R :

- $z_{0{,}025} = q_{1-0{,}025} = q_{0{,}975}$ =`qnorm(0.975)`=1.96

Por último presentamos una importante propiedad de la distribución normal, que nos indica que la suma de variables normales sigue también una distribución normal. Esta propiedad tiene gran aplicación práctica, ya que muchas veces habrán de calcularse probabilidades de sumas de variables normales: peso total de los ejemplares de una muestra, ingresos totales de las sucursales de una empresa durante un día laboral, distancia total recorrida por un animal durante una migración,...

Propiedad reproductiva de la distribución normal: d*adas n variables aleatorias normales e independientes, tales que* $X_i \approx N(\mu_i, \sigma_i)$, $i = 1, \ldots, n$, *su suma* $\sum_{i=1}^{n} X_i$ *sigue también una distribución normal, siendo:*

$$\sum_{i=1}^{n} X_i \approx N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

Como consecuencia de esta propiedad, en el caso particular de que $X_i \approx N(\mu, \sigma)$ para $i = 1, \ldots, n$, aplicando las propiedades de la esperanza y la varianza, se tiene que:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

o, expresado de otra forma,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

4.6. Distribuciones de probabilidad asociadas al muestreo de variables con distribución normal.

En muchas ocasiones nos encontramos con problemas que se refieren a características globales de una variable evaluadas sobre una o varias poblaciones. Por ejemplo ¿la concentración media de cierto contaminante en una zona supera el umbral permitido por la legislación? ¿Es la velocidad media de desplazamiento en los individuos de una especie de delfín superior a la velocidad media en otra especie? ¿Se consigue mayor peso medio en los peces de una piscifactoría cuando se usa una dieta rica en hidratos de carbono o cuando se usa una rica en proteínas? ¿Se observa mayor variabilidad de talla en los machos o en las hembras de una especie? En estos ejemplos la pregunta a responder tiene que ver con los valores medios o las varianzas de estas variables en las poblaciones de interés. Ahora bien, en la práctica estos valores no se conocen, ya que no es posible acceder a todos los sujetos de la población.

Como veremos en el próximo capítulo, la única manera de responder a estas cuestiones consiste en adquirir información sobre las cantidades de interés a partir de una muestra aleatoria. Esto nos conduce a la siguiente cuestión: el valor medio de una variable en una población es único, pero como de una misma población es posible extraer muchas muestras distintas, habrá tantas medias muestrales como muestras sea posible extraer. Lo mismo puede decirse de la varianza. Si el problema es comparar dos poblaciones, pueden extraerse muchas muestras distintas de cada una y por tanto son posibles muchos valores distintos de la diferencia

entre las medias muestrales. Como *a priori,* antes de obtener la muestra (o muestras) es imposible predecir cuáles van a ser los valores resultantes de la media, la varianza o la diferencia de medias, en su caso, resulta que estas cantidades son *variables aleatorias.* Y si son variables aleatorias, debemos preguntarnos cuál es su distribución de probabilidad, ya que es precisamente mediante el uso de dicha distribución que podremos contestar a las preguntas planteadas más arriba.

En el caso particular de que la distribución de probabilidad de la variable de interés sea normal $N(\mu, \sigma)$, se conocen las distribuciones de probabilidad de algunas de las variables aleatorias que se presentan en el muestreo. Describimos a continuación dichas distribuciones y posponemos al próximo capítulo su aplicación concreta en los problemas de inferencia ligados al muestreo.

4.6.1. Distribución Chi-cuadrado χ^2_n

Definición: Una variable aleatoria X sigue una *distribución Chi-Cuadrado de Pearson* con n grados de libertad (χ^2_n) si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}},\ x \geq 0$$

Esta distribución es un caso particular de la gamma, concretamente la $\mathcal{G}\left(\frac{n}{2}, 2\right)$. La importancia práctica de esta distribución deriva de la siguiente propiedad, que constituye el fundamento de la inferencia sobre la varianza en variables con distribución normal.

Proposición: *Si* $Z_1, \ldots, Z_n$ *son* n *variables aleatorias independientes con distribución* $N(0,1)$*, entonces*

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

sigue una distribución χ^2_n*.*

Esperanza y varianza: si $X \approx \chi^2_n$:

$$\mu = E[X] = n$$
$$\sigma^2 = var(X) = 2n$$

La figura 11 muestra la densidad de la χ^2_n para varios valores de n.

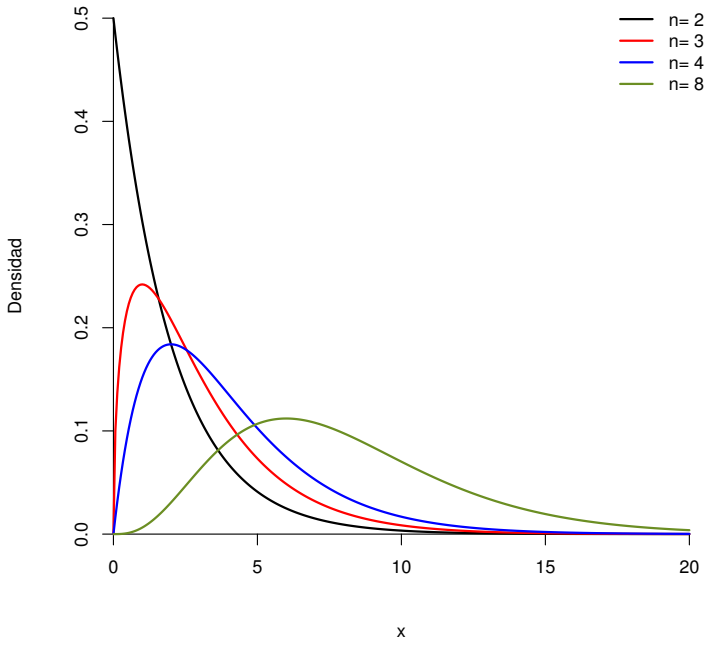

Figura 11: Función de densidad de la distribución χ^2_n para varios valores de n

Cálculo con R :

- Valor de la función de densidad*:* $f(x)$=`dchisq(x,n)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pchisq(x,n)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qchisq(`$\alpha$`,n)`
- Generación de m números aleatorios con distribución χ^2_n: `rchisq(m,n)`

4.6.2. Distribución *t* de Student t_n

Definición: Una variable aleatoria X sigue una *distribución t de Student* con n grados de libertad (t_n) si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \; x \geq 0$$

Por ser una función cuadrática en x, la densidad de la t de Student, al igual que ocurría con la normal, es simétrica respecto al eje de ordenadas, esto es, $f(x) = f(-x)$. En la figura 12 se muestra la forma de esta densidad para varios valores de n. Puede apreciarse la similitud de esta densidad con la normal. De hecho, para valores grandes de n ambas funciones son prácticamente indistinguibles.

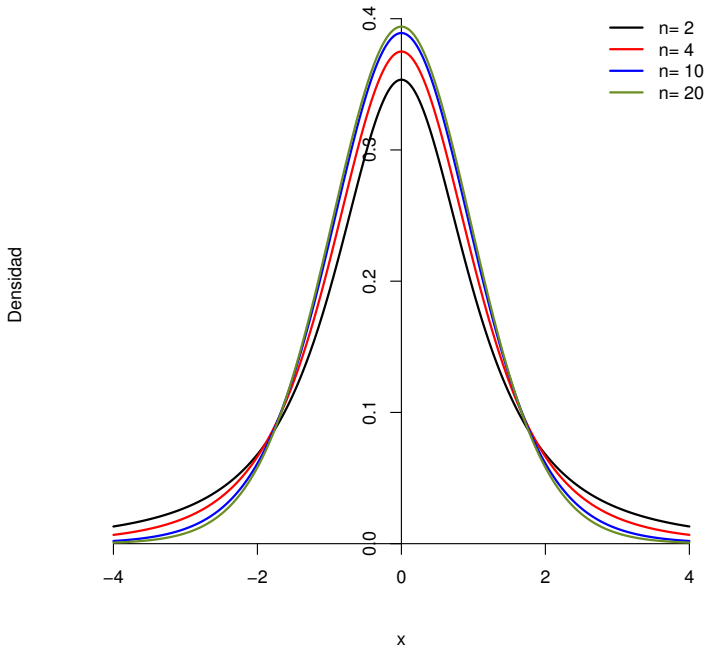

Figura 12: Función de densidad de la distribución t de Student para varios valores de n.

El interés práctico de la distribución t de Student deriva de la siguiente propiedad, que constituye el fundamento de la inferencia sobre la media en variables con distribución normal de varianza desconocida.

Proposición: *Sean* $Z \approx N(0,1)$ *e* $Y \approx \chi^2_n$ *dos variables aleatorias independientes. Entonces*:

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

sigue una distribución t de Student con n grados de libertad.

Esperanza y varianza: Si $X \approx t_n$:

$$\mu = E[X] = 0 \quad (\text{Si } n > 1)$$
$$\sigma^2 = var(X) = \begin{cases} \infty & 1 < n \leq 2 \\ \frac{n}{n-2} & n > 2 \end{cases}$$

Para $n = 1$ no están definidas la media ni la varianza.

Cálculo con R :

- Valor de la función de densidad: $f(x)$=`dt(x,n)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pt(x,n)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qt(`$\alpha$`,n)`
- Generación de m números aleatorios con distribución t_n: `rt(m,n)`

4.6.3. Distribución *F* de Fisher-Snedecor F_{n_1,n_2}.

Definición: Una variable aleatoria X sigue una *distribución F de Fisher-Snedecor* con n_1 y n_2 grados de libertad (F_{n_1,n_2}) si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{n_1/2} \frac{x^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}, \; x \geq 0$$

En realidad, conocer la expresión de la función de densidad de la distribución F de Fisher (al igual que la de la normal, la chi-cuadrado o la t de Student) no nos sirve para calcular probabilidades directamente, ya que no admite primitiva, por lo deberán utilizarse métodos numéricos o tablas. El interés de esta distribución reside en su aplicación en la inferencia relacionada con la comparación de varianzas de variables con distribución normal, cuyo fundamento se encuentra en la siguiente propiedad.

Proposición: *Sean* $Y_1 \approx \chi^2_{n_1}$ *e* $Y_2 \approx \chi^2_{n_2}$ *dos variables aleatorias independientes. Entonces:*

$$X = \frac{Y_1/n_1}{Y_2/n_2}$$

sigue una distribución de probabilidad F *de Fisher-Snedecor con* n_1 *y* n_2 *grados de libertad.*

De aquí se sigue también la siguiente propiedad de la distribución F:

$$X \approx F_{m,n} \Rightarrow \frac{1}{X} \approx F_{n,m}$$

Esperanza y varianza: Si $X \approx F_{n_1,n_2}$:

$$\mu = E[X] = \frac{n_2}{n_2 - 2}, \quad (\text{si } n_2 > 2)$$
$$\sigma^2 = var(X) = 2\left(\frac{n_2}{n_2 - 2}\right)^2 \frac{n_1 + n_2 - 2}{n_1(n_2 - 4)}, \quad (\text{Si } n_2 > 4)$$

La figura 13 muestra la forma de la función de densidad de la distribución F para varios valores de n_1 y n_2.

Cálculo con R :

- Valor de la función de densidad: $f(x)$=`df(x,`n_1`,`n_2`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pf(x,`n_1`,`n_2`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qf(`$\alpha$`,`$n_1$`,`$n_2$`)`
- Generación de m números aleatorios con distribución F_{n_1,n_2}: `rf(m,`n_1`,`n_2`)`

4.7. Utilización de las tablas de la Chi-Cuadrado, t de Student y F de Fisher-Snedecor.

Como ya hemos señalado para el caso de la distribución normal, un problema que se presenta con frecuencia en la práctica es el cálculo de cuantiles de estas distribuciones. Para ello se

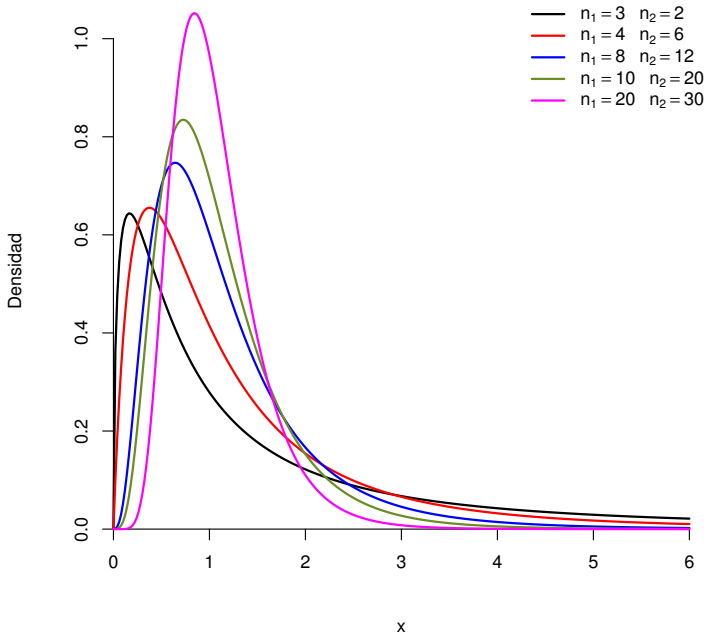

Figura 13: Función de densidad de la distribución F para varios valores de n_1 y n_2.

dispone de tablas de fácil manejo, pero que no incluyen todos los posibles grados de libertad asociados a estas distribuciones (en algunos casos es preciso interpolar). Resulta recomendable en este caso utilizar R u otro software para el cálculo de estos cuantiles (algunas calculadoras lo implementan).

Llamaremos $\chi^2_{n,\alpha}$, $t_{n,\alpha}$ y $F_{n_1,n_2,\alpha}$ a los cuantiles $q_{1-\alpha}$ de las respectivas distribuciones con sus grados de libertad correspondientes. De esta forma:

- Si $X \approx \chi^2_n$, entonces $P\left(X \geq \chi^2_{n,\alpha}\right)=\alpha$
- Si $X \approx t_n$, entonces $P\left(X \geq t_{n,\alpha}\right)=\alpha$
- Si $X \approx F_{n_1,n_2}$ entonces $P\left(X \geq F_{n_1,n_2,\alpha}\right)=\alpha$

La figura 14 muestra la posición de estos cuantiles para cada distribución. El área sombreada es α.

En las tablas de la χ^2_n y la t_n los correspondientes valores de $\chi^2_{n,\alpha}$ y $t_{n,\alpha}$ se encuentran en el cruce de la fila n y la columna α. Los valores de α que figuran en la tabla son los de uso más frecuente. En el caso de la F_{n_1,n_2} se dispone de una tabla para $\alpha = 0{,}025$ y otra para $\alpha = 0{,}05$ (en muchos libros, sobre todo los más antiguos pueden encontrarse tablas para otros

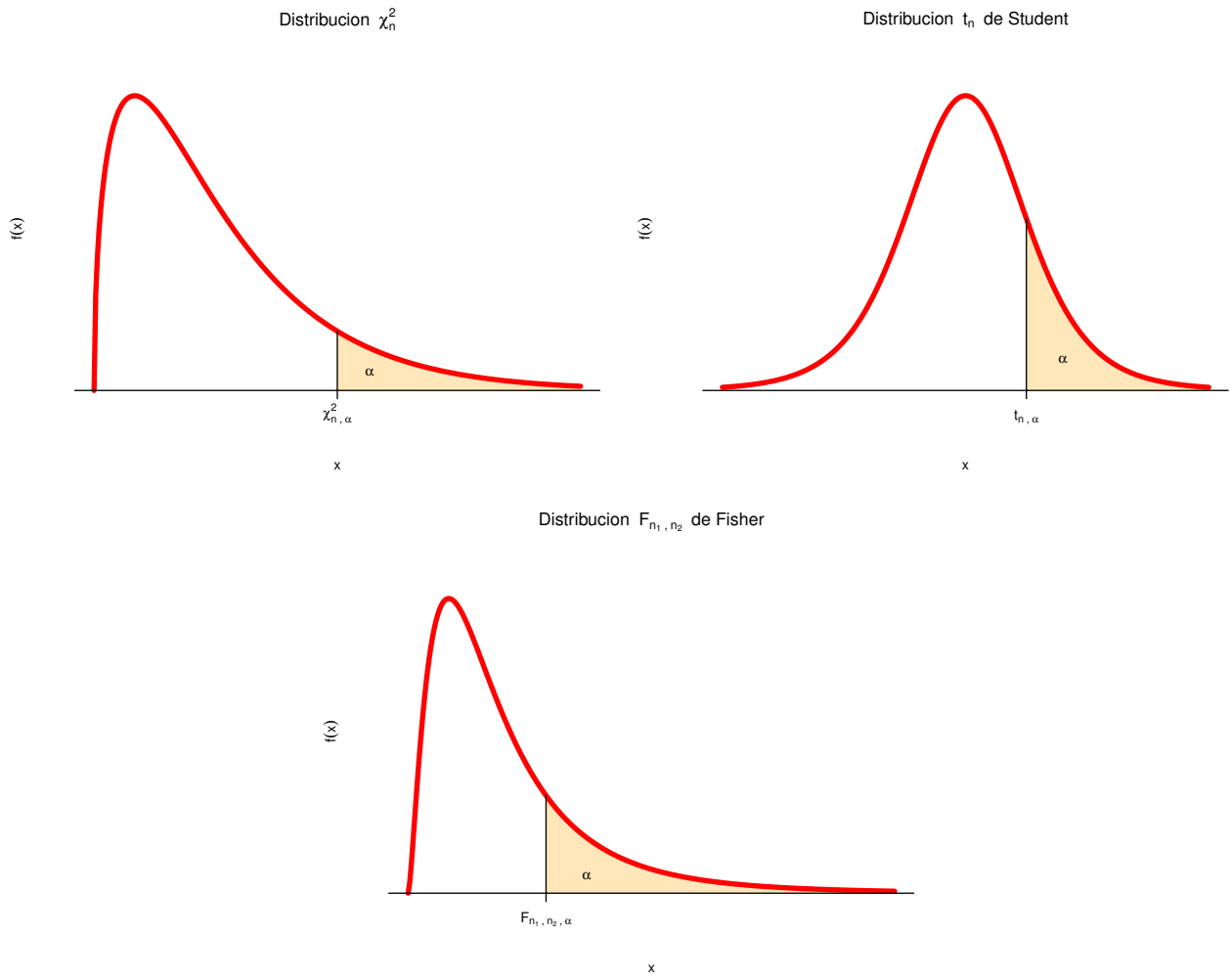

Figura 14: Posición de los cuantiles $q_{1-\alpha}$ de las distribuciones Chi-Cuadrado de Pearson, t de Student y F de Fisher-Snedecor. Estos cuantiles dejan a su derecha un área α (sombreada en las tres figuras).

valores de α; hoy en día, con la ubicuidad de la informática, tales tablas en realidad resultan innecesarias). El valor $F_{n_1,n_2,\alpha}$ se localiza simplemente en el cruce de la fila n_1 con la columna n_2. A veces resulta de interés calcular $F_{n_1,n_2,1-\alpha}$ en cuyo caso se puede utilizar la propiedad siguiente:

$$F_{n_1,n_2,1-\alpha} = \frac{1}{F_{n_2,n_1,\alpha}}$$

Con R estos cuantiles se obtienen directamente como:

- $\chi^2_{n,\alpha}$=`qchisq(1-`α`,n)`
- $t_{n,\alpha}$ =`qt(1-`α`,n)`
- $F_{n_1,n_2,\alpha}$=`qf(1-`α`,`n_1`,`n_2`)`

5. Teorema central del límite.

La propiedad reproductiva de la distribución normal, vista más arriba, nos indica que la suma de variables aleatorias independientes con distribución normal sigue también una distribución normal. El teorema central del límite va un poco más allá, estableciendo condiciones bajo las cuales la suma de variables aleatorias independientes *con distribución no necesariamente normal* sigue una distribución normal. Básicamente tales condiciones son dos: que las variables que se suman tengan todas la misma distribución, y que el número de sumandos sea grande. Estas condiciones se verifican en muchos casos de aplicación práctica; en particular, se cumplen cuando se realiza un muestreo de una variable X con distribución no normal siempre que el número de observaciones sea suficientemente grande, ya que todas las observaciones $X_1, X_2, \ldots, X_n$ proceden de la misma distribución que X.

Teorema Central del Límite *Sea $X_1, \ldots, X_n$ una secuencia de variables aleatorias independientes y con la misma distribución de probabilidad, siendo $E[X_i] = \mu$ y $var(X_i) = \sigma^2$ (finita) para $i = 1, \ldots, n$. Entonces, para $n \to \infty$:*

$$P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq z\right) \to \Phi(z)$$

siendo $\Phi(z)$ la función de distribución de la normal tipificada $N(0,1)$.

Nota: El Teorema Central del Límite, tal como se ha enunciado requiere que todas las variables X_i sean independientes y *tengan la misma distribución.* Existen otras versiones de este teorema, en las que se prueba que, bajo determinadas condiciones[1], si las X_i son independientes *aunque tengan distribuciones de probabilidad diferentes*, su suma también tiene una distribución aproximadamente normal.

[1]Tales condiciones exigen la existencia de determinados momentos de las X_i, y que éstos no crezcan muy deprisa.

Nótese que:

- $E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right] = n\mu$
- $\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}\left(X_i\right) = n\sigma^2 \Rightarrow \text{sd}\left(\sum_{i=1}^{n} X_i\right) = \sigma\sqrt{n}$
- Por tanto, la conclusión del del teorema puede enunciarse diciendo que a medida que n aumenta, la distribución de la suma *tipificada* $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$ se va aproximando a la $N\left(0,1\right)$.

Asimismo, si observamos que:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

el teorema central del límite puede expresarse también como:

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi\left(z\right) \quad \text{para } n \to \infty$$

o, dicho de otra forma, la distribución de probabilidad de la media aritmética *tipificada* $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ de una secuencia de n variables independientes y con la misma distribución, de media μ y desviación típica σ, se va aproximando a la distribución normal $N\left(0,1\right)$ a medida que n aumenta.

En la práctica, el efecto del teorema central del límite puede apreciarse frecuentemente para valores de n que, si bien son grandes, distan mucho de ∞. En muchas ocasiones, con valores de n del orden de entre 30 y 60 ya puede asumirse que, aproximadamente, $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \approx N\left(0,1\right)$ y $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N\left(0,1\right)$, o lo que es lo mismo, que aproximadamente $\sum_{i=1}^{n} X_i \approx N\left(n\mu,\, \sigma\sqrt{n}\right)$ y que $\overline{X} \approx N\left(\mu,\, \sigma/\sqrt{n}\right)$.

En la figura 15 puede apreciarse el significado de este teorema. Cada gráfica corresponde al histograma de 2.000 medias muestrales calculadas sobre muestras de tamaño respectivo 1, 10, 30 y 100 de una distribución exponencial de parámetro $\eta = 100$ (recuérdese que en la distribución exponencial el valor del parámetro coincide con su media). De esta forma cada histograma representa una aproximación a la función de densidad de la media muestral. La linea de trazos corresponde a la estimación de dicha densidad a partir de un suavizado del histograma. La linea roja corresponde a la densidad de una distribución normal cuya media coincide con la de la variable original.

Tal como se puede ver en los gráficos, cuanto mayor es el tamaño de la muestra sobre la que se calcula la media, tanto más se asemeja la distribución de la media a la distribución normal. Asimismo se observa que $E\left[\overline{X}\right]$ se aproxima a $\mu = 100$ y que a medida que n aumenta, $var\left(\overline{X}\right)$ disminuye (de acuerdo con $var\left(\overline{X}\right) = \sigma/\sqrt{n}$).

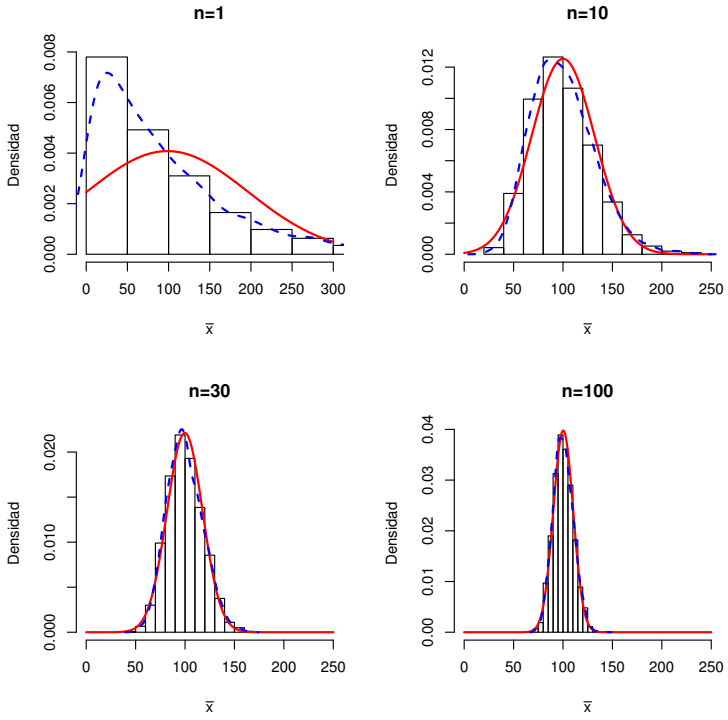

Figura 15: Ilustración del efecto del *Teorema Central del Límite*. A medida que aumenta el tamaño de la muestra (n), la distribución de la media aritmética va asemejándose cada vez más a la normal.

Aproximación de la distribución binomial por la normal

Ya hemos visto en la sección 3.3 que si $X \approx B\left(n,p\right)$ entonces $X = X_1 + X_2 + \cdots + X_n$, siendo las X_i variables de Bernoulli de parámetro p independientes. De acuerdo con el teorema central del límite se tiene que, cuando $n \to \infty$:

$$\frac{X - np}{\sqrt{np\left(1-p\right)}} \approx N\left(0,1\right)$$

En general esta aproximación funciona bien cuando $np \geq 5$, si bien todavía puede mejorarse si se tiene en cuenta el hecho de que la distribución binomial es discreta y la normal es continua. En efecto, la distribución binomial sólo asigna probabilidades a los valores enteros

$0, 1, 2, \ldots, n$ mientras que la normal asignaría probabilidades a todo el rango continuo que contiene a estos valores. Para conseguir una mayor semejanza entre ambas asignaciones se considera que cada valor entero k queda representado por el intervalo $(k - 0{,}5, k + 0{,}5)$. Este procedimiento recibe el nombre de *corrección por continuidad.* De esta forma, la aproximación de las probabilidades binomiales por el teorema central del límite se llevaría a cabo del siguiente modo:

$$\begin{aligned}
P(X = k) \cong & P(k - 0{,}5 \leq X \leq k + 0{,}5) \cong \\
& \cong P\left(\frac{k - 0{,}5 - np}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) = \\
& = P\left(\frac{k - 0{,}5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) \\
P(X \geq k) \cong & P(X \geq k - 0{,}5) \cong P\left(Z \geq \frac{k - 0{,}5 - np}{\sqrt{np(1-p)}}\right) \\
P(X > k) \cong & P(X \geq k + 0{,}5) \cong P\left(Z > \geq \frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) \\
P(X \leq k) \cong & P(X \leq k + 0{,}5) \cong P\left(Z \leq \frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) \\
P(X < k) \cong & P(X \leq k - 0{,}5) \cong P\left(Z \leq \frac{k - 0{,}5 - np}{\sqrt{np(1-p)}}\right)
\end{aligned}$$

siendo $Z \approx N(0, 1)$

Ejemplo: Se dispone de 50 huevos de tortuga; la probabilidad de que un huevo dé lugar a un macho es 0.30. ¿Cuál es la probabilidad de que en total nazcan más de 16 machos?

Si X es el número de machos, se tiene que $X \approx B(50, 0{,}3)$. La probabilidad pedida es

$$P(X > 16) \cong P(X \geq 16{,}5) \cong P\left(Z \geq \frac{16{,}5 - 50 \cdot 0{,}3}{\sqrt{50 \cdot 0{,}3 \cdot 0{,}7}}\right) = P(Z \geq 0{,}46) = 0{,}32276$$

(hemos utilizado la tabla de la $N(0, 1)$ para calcular la última probabilidad). Si utilizamos R para calcular esta probabilidad de manera exacta, obtenemos:

$$\begin{aligned}
P(X > 16) = \sum_{k=17}^{50} P(X = k) = \sum_{k=17}^{50} \binom{50}{k} 0{,}3^k (1 - 0{,}3)^{50-k} = \\
= \texttt{sum(dbinom(17:50,50,0.3))} = 0{,}31612
\end{aligned}$$

Como vemos el error de aproximación es de algo menos de 7 milésimas (0.00664).

Capítulo 4

Inferencia Estadística I: Estimación Puntual.

4.1. Introducción.

La *inferencia estadística* es el proceso mediante el cual se extienden o generalizan a una población las conclusiones o resultados obtenidos a partir de la información proporcionada por una muestra de la misma. Este proceso de inferencia puede perseguir dos objetivos diferentes:

1. *Estimación de parámetros:* utilizar los datos de la muestra para obtener valores aproximados de los parámetros que caracterizan el comportamiento de las variables de interés en la población.

2. *Contraste de hipótesis:* utilizar la información de la muestra para decidir sobre la validez o no de hipótesis relativas a alguna característica de la población.

Dado que la muestra sólo proporciona información parcial sobre la población, los métodos de inferencia estadística se apoyan en el cálculo de probabilidades para cuantificar los márgenes de error probables o para evaluar el riesgo de incurrir en decisiones incorrectas.

Obviamente el desarrollo de los procedimientos de inferencia requiere disponer de una muestra lo suficientemente representativa de la población. En este capítulo presentaremos algunos conceptos elementales sobre muestreo, para a continuación ocuparnos del problema de la estimación de parámetros: qué es un estimador, qué características debe tener y cómo se puede construir un estimador adecuado para un parámetro de interés.

Objetivos.

Al finalizar este capítulo, el alumno deberá:

1. Conocer y comprender los conceptos de población y muestra aleatoria.
2. Entender el significado de la inferencia estadística y distinguir entre inferencia paramétrica e inferencia no paramétrica.
3. Conocer y manejar el concepto de estimador puntual, así como entender el significado de las propiedades de sesgo, varianza y consistencia de un estimador
4. Conocer y ser capaz de aplicar los distintos métodos de obtención de estimadores: momentos, máxima verosimilitud y mínimos cuadrados.
5. Ser capaz de interpretar el significado de los parámetros estimados.
6. Ser capaz de valorar el grado de ajuste conseguido mediante el modelo paramétrico estimado.

4.2. Población y muestra aleatoria.

En la introducción de este capítulo hemos definido la *inferencia estadística* como el proceso mediante el cual se extienden o generalizan a una población las conclusiones o resultados obtenidos a partir de la información proporcionada por una muestra de la misma. Conviene, por tanto, precisar el significado de los términos *población* y *muestra.*

La definición habitual de *población* es la de conjunto formado por *todos* los sujetos u objetos que comparten una o varias características comunes, y sobre los que se desea obtener información. Desde esta perspectiva podemos hablar, por ejemplo, de la población formada por todos los seres humanos que habitan la Tierra, de la población de hormigas de la isla de Gran Canaria, o de la población de delfines mulares hembra del Atántico. Esta definición, sin embargo, presenta dificultades en muchos casos: ¿cuál es la población si el objetivo de nuestro estudio es caracterizar la temperatura del magma volcánico? ¿Y si nuestro objetivo es estudiar la velocidad de una corriente marina? En otro contexto, si deseamos saber si un tratamiento médico es efectivo contra determinada enfermedad, parece lógico considerar como población el conjunto de personas susceptibles de recibir el tratamiento; pero este conjunto incluye tanto aquellos que padecen la enfermedad actualmente, como aquellos que la padecerán en el futuro y a los que podría aplicárseles el tratamiento.

Vemos, pues, que hay poblaciones tangibles (personas, delfines u hormigas), conceptuales (los estados físicos del magma o los comportamientos dinámicos de la corriente marina) e incluso hipotéticas (los sujetos que en el futuro podrían contraer una enfermedad). En cualquier caso, cuando se estudia una población, el objetivo no es, propiamente, el conjunto de sujetos, objetos u entes conceptuales que puedan formar esa población en un instante concreto, sino determinadas *características* que medimos sobre ellos, y que se traducen en *variables aleatorias,* toda vez que sus valores no son conocidos a priori. En este sentido, desde un punto de vista práctico, caracterizar una *población* es equivalente a *conocer la distribución de probabilidad* $\mathbb{P}$ *de la variable aleatoria* X que se mide sobre la misma: la temperatura del magma, la velocidad de la corriente o la variable binaria $1-0$ que indica si un paciente se cura o no.

Normalmente, la población completa no suele ser accesible (por su tamaño, por cuestiones de coste o tiempo, o simplemente porque la población es hipotética), por lo que su estudio habrá de realizarse a partir de sólo una parte de la misma. Se denomina *muestra* a un subconjunto de la población. Para que la información proporcionada por una muestra pueda emplearse aceptablemente para obtener conclusiones sobre la población es necesario:

- Que la muestra sea *representativa*, esto es, que refleje de la mejor manera posible las características de la población. Si una muestra no fuese representativa, es obvio que lo que se pueda deducir de ella no podrá extenderse a la población; en particular la estimación de parámetros en tales condiciones podría estar fuertemente sesgada y los contrastes de hipótesis podrían conducir a decisiones erróneas con mayor frecuencia de lo previsto.

- Que la muestra tenga un tamaño suficiente. En general, cuanto mayor sea el tamaño, más información proporcionará. El tamaño adecuado de la muestra depende de cuál sea el problema que nos planteamos (estimación de parámetros o contraste de hipótesis), de las características de la población (en general, a mayor heterogeneidad de la población con respecto a la variable de interés, mayor habrá de ser el tamaño de la muestra) y de la magnitud de los errores que estamos dispuestos a cometer en nuestro proceso de inferencia.

Como hemos señalado más arriba, habitualmente nuestro interés se centra en el estudio de alguna variable aleatoria X que se mide sobre la población. El comportamiento de dicha variable aleatoria X queda caracterizado por su *distribución de probabilidad* $\mathbb{P}$. En este contexto, definimos una *muestra aleatoria* de tamaño n de una distribución de probabilidad $\mathbb{P}$ como *un conjunto de variables aleatorias* $X_1, \ldots, X_n$ *independientes y con la misma distribución* $\mathbb{P}$. En la práctica, la obtención de una muestra aleatoria se traduce en seleccionar

al azar y de manera independiente n elementos de la población y medir el valor de X en cada uno de ellos. Así, si X es la velocidad de la corriente marina en un punto, $X_1, \ldots, X_n$ serían n observaciones independientes de dicha velocidad en ese punto; si X es la variable binaria $1-0$ que representa la curación (o no) de una enfermedad tras aplicar un tratamiento, $X_1, \ldots, X_n$ sería el efecto del tratamiento en un conjunto de n pacientes elegidos de manera independiente en la misma población.

Podemos preguntarnos de qué manera y hasta qué punto una muestra aleatoria $X_1, \ldots, X_n$ de observaciones de una variable aleatoria X nos informa sobre la distribución de probabilidad de X (evidentemente, si la muestra no contuviese información a este respecto, no tendría sentido el muestreo). Para responder a esta pregunta definimos la función de *distribución empírica* de la muestra como:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

siendo $I(X_i \leq x)$ uno o cero según ocurra o no el suceso $\{X_i \leq x\}$ (por tanto, $\hat{F}_n(x)$ es la proporción de veces que en la muestra se han observado valores menores o iguales que x). El teorema de *Glivenko-Cantelli,* que enunciamos a continuación, prueba que a medida que el tamaño de muestra n se incrementa, la función de distribución empírica $\hat{F}_n(x)$ se va aproximando cada vez más a la función de distribución acumulativa $F(x)$ de la variable X.

Teorema 4.1. ***(de Glivenko-Cantelli)*** *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable aleatoria X con función de distribución acumulativa $F(x)$, y sea $\hat{F}_n(x)$ la función de distribución empírica de la muestra. Entonces para cualquier valor x se verifica, a medida que $n \to \infty$:*

$$E\left[\left(\hat{F}_n(x) - F(x)\right)^2\right] \to 0$$

Demostración. Es inmediato observar que, para cada x, la variable $I(X_i \leq x)$ sigue una distribución de Bernoulli de parámetro $F(x)$, cualquiera que sea i. Por tanto, tal como vimos en el capítulo anterior, $E[I(X_i \leq x)] = F(x)$ y $\text{var}(I(X_i \leq x)) = F(x)(1 - F(x))$. Aplicando ahora las propiedades de la esperanza y la varianza de una suma de variables aleatorias independientes:

$$E\left[\hat{F}_n(x)\right] = \frac{1}{n}\sum_{i=1}^{n} E[I(X_i \leq x)] = F(x)$$

$$\text{var}\left(\hat{F}_n(x)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}\left(I\left(X_i \leq x\right)\right) = \frac{1}{n}F(x)\left(1-F(x)\right)$$

Por tanto:

$$E\left[\left(\hat{F}_n(t) - F(t)\right)^2\right] = \text{var}\left(\hat{F}_n(t)\right) = \frac{1}{n}F(t)\left(1-F(t)\right) \to 0$$

cuando $n \to \infty$. □

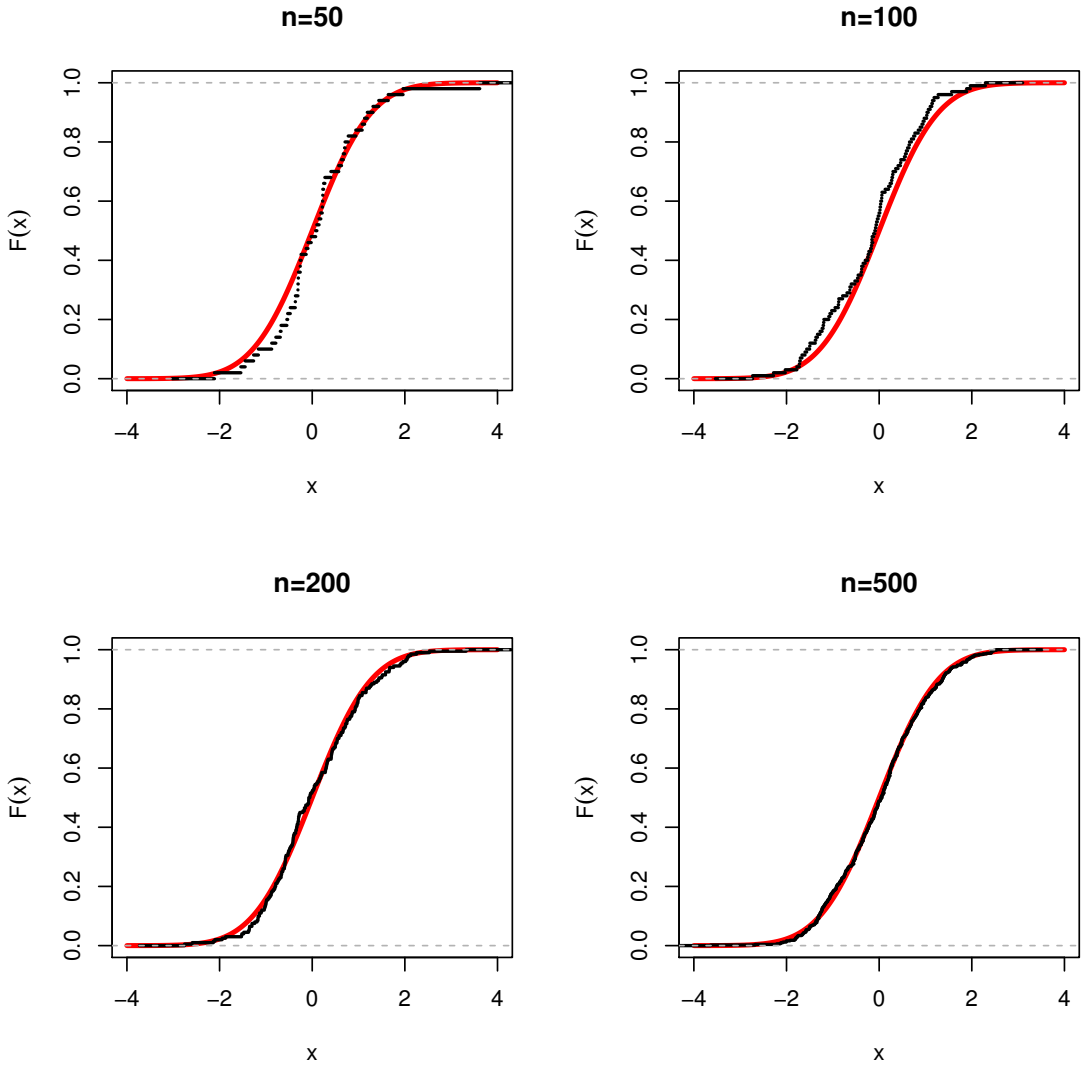

Figura 4.1: Efecto del Teorema de Glivenko-Cantelli: a medida que aumenta el tamaño de la muestra, la función de distribución empírica de la muestra, $\hat{F}_n(x)$, se aproxima cada vez más a la función de distribución acumulativa teórica $F(x)$ de la variable aleatoria.

Así pues, el teorema de Glivenko-Cantelli garantiza que el muestreo aleatorio produce muestras representativas de la variable de interés que, con el tamaño adecuado, permiten aproximar razonablemente la función de distribución acumulativa de dicha variable. Por esta razón este teorema suele conocerse también como **teorema fundamental de la estadística.**

En la figura 4.1 se muestran superpuestas la función de distribución acumulativa de la distribución normal de parámetros $\mu = 0$ y $\sigma = 1$ y la distribución empírica obtenida para muestras aleatorias de tamaños respectivos 50, 100, 200 y 500. Puede apreciarse que a medida que aumenta el tamaño muestral, la función empírica tiende a confundirse con la teórica.

4.3. Inferencia paramétrica vs. inferencia no paramétrica.

Como sabemos, el comportamiento de una variable aleatoria X queda caracterizado mediante su función de distribución acumulativa $F(x)$. Cuando el investigador toma una muestra aleatoria $X_1, X_2, \ldots, X_n$ de esta variable, puede encontrarse en alguno de los siguientes escenarios:

1. Conoce la expresión funcional de $F(x)$, pero no conoce los valores de los parámetros que la caracterizan, y que denotaremos por $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Esto es lo que sucede, por ejemplo, si se sabe (o se sospecha) que los datos proceden de una distribución exponencial (de la que no se conoce el valor del parámetro λ), de una Weibull (de la que no se sabe lo que valen κ y η), de una Normal (de la que no se conocen μ y σ), ...
2. No sabe nada de $F(x)$ salvo, quizás, si es continua o escalonada.

El primer escenario corresponde a la así llamada *inferencia paramétrica*. Cualquier afirmación, en términos de probabilidad, sobre las características de la variable X requiere obtener alguna aproximación del valor del parámetro Θ, proceso que se conoce con el nombre de *estimación*. El segundo escenario corresponde a un problema de *inferencia no paramétrica*. Como veremos, en el primer caso los contrastes de hipótesis se establecen en términos de Θ; en el segundo caso se establecen en términos de características más generales usualmente relacionadas con la forma de $F(x)$.

Señalemos por último que, dado que en la práctica una de las situaciones más habituales es asumir que $F(x)$ corresponde a la distribución normal, es habitual denominar inferencia paramétrica a la inferencia basada en dicha distribución.

4.4. Estimación.

En el capítulo anterior hemos visto una colección de distribuciones de probabilidad que permiten modelar el comportamiento de numerosas variables aleatorias que aparecen en las aplicaciones prácticas: el peso o la longitud de un pez de determinada especie, la altura de ola en una zona costera, el número de nidos de tortuga en una playa, el tiempo entre ocurrencias de un fenómeno meteorológico, etc. Este proceso de modelización requiere ajustar de algún modo los parámetros característicos de la distribución de probabilidad a emplear. Así, por ejemplo, si modelamos la longitud de los peces de una especie mediante una distribución normal, ¿cuáles son los valores de μ y σ adecuados?; si modelamos la altura de ola mediante una distribución de Weibull, ¿cuáles son los valores de los parámetros de localización y escala?; si se modela el número de nidos de tortuga en una playa mediante la distribución de Poisson, ¿cuál es el valor de λ?.

La obtención del valor aproximado de un parámetro se denomina *estimación.* La estimación es *puntual* si proporciona un único valor aproximado para dicho parámetro; es *por intervalo* si proporciona un intervalo que, con cierta confianza, contiene al parámetro.

4.4.1. Definiciones básicas

Estadístico: Dada una muestra aleatoria $X_1, X_2, \ldots, X_n$ se llama *estadístico* a cualquier función de sus valores.

Estimador: Dado un parámetro θ característico de una población, y una muestra aleatoria $X_1, X_2, \ldots, X_n$ de la misma, se llama *estimador* de ϑ a cualquier estadístico $\hat{\theta} = \hat{\theta}\,(X_1, X_2, ..., X_n)$ cuyos valores se aproximen a θ.

Si bien los estimadores muchas veces pueden construirse de forma natural –estimar la esperanza de una variable mediante la media de una muestra aleatoria de la misma, estimar una proporción poblacional mediante la proporción equivalente en la muestra– existen diversos métodos, que veremos en la sección 4.4.3, que permiten construir estimadores en casos más generales, y además con buenas propiedades.

Nótese de la definición anterior que *un estimador es una variable aleatoria*: no puede predecirse su valor mientras no se haya obtenido la muestra. Por tanto, un estimador habrá de caracterizarse en términos de una distribución de probabilidad sobre sus posibles valores.

Como distintas muestras producirán distintos valores del estimador $\hat{\theta}$, es de esperar que algunos de estos valores estén más próximos al valor de θ y otros estén más alejados. Por tanto ¿cuando podemos considerar que $\hat{\theta} = \hat{\theta}\,(X_1, X_2, ..., X_n)$ produce valores próximos a θ?

Como veremos a continuación, la respuesta a esta pregunta está estrechamente relacionada con la distribución de probabilidad de $\hat{\theta}$.

4.4.2. Propiedades deseables de un estimador.

4.4.2.1. Exactitud:

Dado que el estimador puede tomar muchos valores diferentes (según cual sea la muestra que se obtenga), una manera de medir la proximidad entre el estimador y el parámetro es mediante la distancia entre el valor esperado del estimador y el valor del parámetro. Dicha distancia recibe el nombre de *sesgo* del estimador:

$$\text{Sesgo}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Cuando el sesgo del estimador es cero (en cuyo caso $E[\hat{\theta}] = \theta$), el estimador es *exacto* (también se le suele llamar *insesgado* o *centrado*). En caso contrario el estimador es *sesgado*. En general resulta deseable que un estimador sea insesgado. Un sesgo positivo en el estimador significa que sus valores, en media, están por encima del parámetro que pretende estimar y por tanto tiende a sobreestimarlo. De modo similar, los estimadores con sesgo negativo tienden a subestimar el parámetro.

Ejemplo 4.1. La media muestral es un estimador centrado de la media poblacional. En efecto:

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}n\mu = \mu$$

Ejemplo 4.2. La varianza muestral es un estimador sesgado de la varianza poblacional. En efecto, la varianza muestral se define como:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Para calcular la esperanza de S^2 observemos en primer lugar que:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\sum_{i=1}^{n}\left(X_i-\mu+\mu-\bar{X}\right)^2=\sum_{i=1}^{n}\left((X_i-\mu)-\left(\bar{X}-\mu\right)\right)^2=$$
$$=\sum_{i=1}^{n}\left((X_i-\mu)^2-2\left(X_i-\mu\right)\left(\bar{X}-\mu\right)+\left(\bar{X}-\mu\right)^2\right)=$$
$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-2\left(\bar{X}-\mu\right)\sum_{i=1}^{n}\left(X_i-\mu\right)+\sum_{i=1}^{n}\left(\bar{X}-\mu\right)^2=$$

$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-2\left(\bar{X}-\mu\right)n\left(\bar{X}-\mu\right)+n\left(\bar{X}-\mu\right)^2=$$
$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-n\left(\bar{X}-\mu\right)^2$$

Se tiene:

$$E\left[\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right]=\sum_{i=1}^{n}E\left[\left(X_i-\mu\right)^2\right]=n\sigma^2$$

Por ser las X_i independientes:

$$E\left[\left(\bar{X}-\mu\right)^2\right]=var\left(\overline{X}\right)=var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\,var\left(\sum_{i=1}^{n}X_i\right)=$$
$$=\frac{1}{n^2}\sum_{i=1}^{n}var\left(X_i\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}$$

Por tanto:

$$E\left[S^2\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i-\mu\right)^2-n\left(\bar{X}-\mu\right)^2\right]=\frac{1}{n}\left(n\sigma^2-\sigma^2\right)=\frac{n-1}{n}\sigma^2$$

Así pues:

$$\text{Sesgo}(S^2)=E\left[S^2\right]-\sigma^2=\frac{n-1}{n}\sigma^2-\sigma^2=-\frac{1}{n}\sigma^2$$

de donde se sigue que la varianza muestral subestima la varianza poblacional (si bien es cierto que a medida que el tamaño de la muestra n aumenta, el sesgo se hace más pequeño).

Ejemplo 4.3. La cuasivarianza muestral, definida como

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

sí es un estimador centrado de la varianza poblacional. En efecto:

$$E\left[s^2\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] =$$
$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - n\left(\bar{X} - \mu\right)^2\right] = \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \sigma^2$$

Por esta razón, como estimador de la varianza poblacional, en la práctica se prefiere la cuasivarianza muestral.

Ejemplo 4.4. Si X es una variable aleatoria de Bernoulli de parámetro p, la proporción muestral de éxitos $\hat{p}$ es un estimador insesgado de la proporción poblacional p. En efecto, la proporción muestral de éxitos al observar una muestra aleatoria de tamaño n es:

$$\hat{p} = \frac{\text{Número de éxitos}}{\text{Número de Observaciones}} = \frac{N_E}{n}$$

Como X es de Bernoulli, el número N_E de éxitos en n pruebas independientes sigue una distribución $B(n,p)$, y por tanto:

$$E\left[\hat{p}\right] = E\left[\frac{N_E}{n}\right] = \frac{1}{n}E\left[N_E\right] = \frac{1}{n}n\cdot p = p$$

4.4.2.2. Precisión.

Tal como hemos visto, un estimador es una variable aleatoria cuyo valor cambia con la muestra. Si el estimador es centrado, ello indica que el centro de la distribución de valores del estimador coincide con el parámetro que se pretende estimar. Si embargo esto no nos informa de si dicha distribución tiene mucha o poca dispersión en torno al parámetro. Si la dispersión es grande, significa que habrá muestras que darán lugar a estimaciones muy alejadas del valor del parámetro. Si la dispersión es pequeña, aún en la peor de las muestras posibles, la estimación obtenida estará próxima al valor del parámetro. Por tanto, si se dispone de

varios estimadores centrados del mismo parámetro, será preferible (producirá estimaciones más precisas del parámetro) aquél que tenga la menor dispersión. Dado que la dispersión se mide mediante la varianza del estimador[1], el mejor estimador centrado será el de menor varianza (en caso de existir).

La desviación típica del estimador recibe el nombre de *error estándar*. Se suele denotar como

$$\sigma_{\hat{\theta}} = \sqrt{var\left(\hat{\theta}\right)}$$

Puede demostrarse que la media muestral, la cuasivarianza muestral y la proporción muestral son estimadores insesgados y de mínima varianza de sus parámetros respectivos.

4.4.2.3. Menor Error Cuadrático Medio.

Se define el *error cuadrático medio* (ECM) de un estimador $\hat{\theta}$ para un parámetro θ, como:

$$ECM\left[\hat{\theta}\right] = E\left[\left(\hat{\theta} - \theta\right)^2\right] = \left(\text{Sesgo}\left(\hat{\theta}\right)\right)^2 + var\left(\hat{\theta}\right)$$

El ECM constituye una medida conjunta (de hecho es la suma) del sesgo y la varianza de un estimador. Es deseable que el error cuadrático medio de un estimador sea pequeño. El ECM es una medida que resulta útil cuando se debe elegir entre varios estimadores del mismo parámetro con características muy diferentes de sesgo y varianza. Así por ejemplo, puede ser más útil un estimador ligeramente sesgado pero con muy poca varianza (tal que, aunque sesgadas, todas las estimaciones están próximas al parámetro), que uno centrado pero con varianza mucho mayor (que puede dar lugar a muchas estimaciones muy alejadas del parámetro).

4.4.2.4. Consistencia de un estimador.

Un estimador $\hat{\theta}$ de un parámetro θ es consistente si verifica que:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| \leq \varepsilon\right) = 1 \quad \forall \varepsilon > 0$$

[1] O de manera equivalente, mediante la desviación típica. La desviación típica de un estimador recibe el nombre de *error estándar*.

lo que significa que a medida que aumenta el tamaño de la muestra es más probable que el valor del estimador esté cada vez más próximo al valor del parámetro. En general es deseable que los estimadores que utilicemos sean consistentes.

Puede demostrarse que la media muestral, la varianza muestral y la proporción muestral son estimadores consistentes de sus parámetros respectivos. Por ejemplo, para probar que la media muestral es un estimador consistente de la media poblacional basta tener en cuenta que $E\left[\overline{X}\right] = \mu$ y $var\left(\overline{X}\right) = \frac{\sigma^2}{n}$. De acuerdo con el teorema de Chebyshev, para cualquier valor de $k \geq 1$ se tiene:

$$P\left(\left|\overline{X} - \mu\right| > k\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2}$$

Eligiendo entonces $\varepsilon = k\frac{\sigma}{\sqrt{n}}$ (esto es, $k = \frac{\varepsilon\sqrt{n}}{\sigma}$) se tiene que

$$P\left(\left|\overline{X} - \mu\right| > \varepsilon\right) \leq \frac{1}{n}\left(\frac{\sigma}{\varepsilon}\right)^2$$

por lo que cuando $n \to \infty$ resulta $P\left(\left|\overline{X} - \mu\right| > \varepsilon\right) \to 0$, o lo que es lo mismo

$$P\left(\left|\overline{X} - \mu\right| \leq \varepsilon\right) \to 1$$

lo que prueba que la media muestral $\overline{X}$ es un estimador consistente de la media poblacional μ. Ello además vuelve a justificar, como ya hemos visto anteriormente, que el concepto de esperanza de una variable aleatoria puede identificarse con el de media aritmética para grandes valores de n.

4.4.3. Métodos de obtención de estimadores puntuales.

En esta sección abordamos el problema de cómo pueden obtenerse funciones cuyos valores se aproximen al de un parámetro desconocido de cierta distribución de probabilidad. Tres son los métodos que se emplean habitualmente para ello: el método de los momentos, el método de máxima verosimilitud y el método de los mínimos cuadrados.

4.4.3.1. Método de los momentos.

Recordemos que dada una variable aleatoria X, se define el momento de orden k respecto al origen como:

$$\mu_k = E\left[X^k\right] = \begin{cases} \sum\limits_{x_i \in E} x_i^k P\left(X = x_i\right) & \text{si } X\text{es discreta} \\ \int_{-\infty}^{\infty} x^k f\left(x\right) dx & \text{si } X\text{es continua} \end{cases}$$

Ya hemos visto en varias ocasiones que $\mu = \mu_1$ y $\sigma^2 = \mu_2 - \mu_1^2$. De la misma forma que la esperanza y la varianza se pueden poner en función de los momentos, en general si una variable aleatoria X depende de unos parámetros desconocidos $\theta_1, \theta_2, \ldots, \theta_k$, muchas veces será posible expresar estos parámetros como funciones de algunos momentos de la variable, esto es, $\theta_j = g_j(\mu_1, \mu_2, ...)$, $j = 1, 2, ..., k$. El método de los momentos consiste en determinar estas funciones, estimar los momentos correspondientes mediante sus análogos muestrales:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \ ..., \ \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

y por último estimar los θ_j, mediante las funciones anteriores evaluadas en los momentos muestrales: $\hat{\theta}_j = g_j(\hat{\mu}_1, \hat{\mu}_2, ...)$, $j = 1, 2, ..., k$

Este método tiene su fundamento en el hecho de que los momentos muestrales son estimadores insesgados de los momentos poblacionales. Asimismo ya hemos visto que si se toma una muestra aleatoria, a medida que aumenta su tamaño su distribución empírica se va pareciendo cada vez más a la distribución de probabilidad de la variable observada. Intuitivamente ello nos indica que los momentos muestrales se van a ir pareciendo cada vez más a los poblacionales a medida que aumenta el tamaño de la muestra.

Ejemplo 4.5. Supongamos que se desea estimar el parámetro p de una variable Bernoulli $b(p)$. Sabemos que

$$E[X] = p$$

Por lo que p puede expresarse en términos de los momentos simplemente como

$$p = E[X] = \mu_1$$

Para estimar p, simplemente sustituimos μ_1 en esta ecuación por su estimador $\hat{\mu}_1 = \overline{X}$ con lo que como estimador de p se obtiene:

$$\hat{p} = \hat{\mu}_1 = \overline{X}$$

Nótese que al ser $X \approx b(p)$, la variable X sólo toma los valores 1 (éxito) o 0 (fracaso), por lo que la media aritmética de n observaciones de X es:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\text{Nº de éxitos en } n\text{pruebas}}{n}$$

esto es, la proporción de éxitos en la muestra.

Ejemplo 4.6. Se desea estimar el parámetro p de una variable $Geo(p)$. En este caso, sabemos que:

$$\mu_1 = E[X] = \frac{1-p}{p}$$

De aquí despejamos p:

$$p\mu_1 = 1 - p \Rightarrow p(1+\mu_1) = 1 \Rightarrow p = \frac{1}{1+\mu_1}$$

El estimador por el método de los momentos se obtiene sustituyendo el momento poblacional por el correspondiente momento muestral. Por tanto:

$$\hat{p} = \frac{1}{1+\hat{\mu}_1} = \frac{1}{1+\overline{X}}$$

Ejemplo 4.7. Se desea estimar el número de ardillas N que hay en un bosque. Para ello se capturan inicialmente N_M ardillas, que son marcadas y devueltas al bosque. A continuación y durante n días se procede del modo siguiente: se recorre el bosque durante un periodo de tiempo fijo y se van contando las ardillas que se avistan hasta encontrar una ardilla marcada. Sea X_i el número de ardillas no marcadas que se han avistado el día i. Para estimar N por el método de los momentos basta observar que $X_i \approx Geo(p)$ siendo $p = \frac{N_M}{N}$. Por tanto

$$N = \frac{N_M}{p}$$

En el ejemplo anterior ya hemos visto que el estimador de p es $\hat{p} = \frac{1}{1+\overline{X}}$. Por tanto el estimador del número de ardillas en el bosque será:

$$\hat{N} = \frac{N_M}{\hat{p}} = N_M\left(1+\overline{X}\right)$$

siendo $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Ejemplo 4.8. Si $X \approx N(\mu, \sigma)$ y se desea estimar μ y σ por el método de los momentos, basta observar que como:

$$\mu = E[X] = \mu_1, \quad \sigma^2 = E\left[X^2\right] - (E[X])^2 = \mu_2 - \mu_1^2$$

los estimadores serán:

$$\hat{\mu} = \hat{\mu}_1 = \overline{X}$$
$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}\right)^2$$

Ejemplo 4.9. Si $X \approx \mathcal{G}(\kappa, \eta)$, para estimar los parámetros κ y η por el método de los momentos, recordemos que

$$\mu = \kappa \cdot \eta, \quad \sigma^2 = \kappa \cdot \eta^2$$

Teniendo en cuenta que $\mu_1 = \mu$ y $\sigma^2 = \mu_2 - \mu_1^2$, resulta:

$$\kappa \cdot \eta = \mu_1$$
$$\kappa \cdot \eta^2 = \mu_2 - \mu_1^2$$

Para expresar κ y η en función de los momentos poblacionales, dividimos el segundo término entre el primero y obtenemos:

$$\eta = \frac{\mu_2}{\mu_1} - \mu_1$$

Sustituimos este valor en el primer término y despejamos κ:

$$\kappa = \frac{\mu_1}{\eta} = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$$

Los estimadores por el método de los momentos se obtienen entonces sustituyendo en estas expresiones los momentos poblacionales por los muestrales:

$$\hat{\eta} = \frac{1}{n\overline{X}}\sum_{i=i}^{n} X_i^2 - \overline{X}$$
$$\hat{\kappa} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=i}^{n} X_i^2 - \overline{X}^2}$$

Ejemplo 4.10. Si $X \approx W(\kappa, \eta)$, para estimar κ y η por el método de los momentos, al igual que en el caso anterior bastará con tener en cuenta que su esperanza y varianza son:

$$\mu = \eta \cdot \Gamma\left(1 + \frac{1}{\kappa}\right), \quad \sigma^2 = \eta^2\left[\Gamma\left(1 + \frac{2}{\kappa}\right) - \left(\Gamma\left(1 + \frac{1}{\kappa}\right)\right)^2\right]$$

y por tanto:

$$\eta \cdot \Gamma\left(1+\frac{1}{\kappa}\right) = \mu_1$$
$$\eta^2\left[\Gamma\left(1+\frac{2}{\kappa}\right) - \left(\Gamma\left(1+\frac{1}{\kappa}\right)\right)^2\right] = \mu_2 - \mu_1^2$$

Si dividimos el segundo término por el cuadrado del primero nos queda una ecuación en κ :

$$\frac{\Gamma\left(1+\frac{2}{\kappa}\right)}{\left[\Gamma\left(1+\frac{1}{\kappa}\right)\right]^2} = \frac{\mu_2}{\mu_1^2}$$

El estimador de κ se obtiene resolviendo esta ecuación sustituyendo μ_1 y μ_2 por los correspondientes momentos muestrales:

$$\frac{\Gamma\left(1+\frac{2}{\hat{\kappa}}\right)}{\left[\Gamma\left(1+\frac{1}{\hat{\kappa}}\right)\right]^2} = \frac{\frac{1}{n}\sum_{i=i}^{n} X_i^2}{\left(\overline{X}\right)^2} \tag{4.1}$$

Obviamente no es posible despejar de aquí el valor de $\hat{\kappa}$ explícitamente, pero es posible construir un algoritmo numérico que resuelva el problema. Una vez obtenido $\hat{\kappa}$, el valor de $\hat{\eta}$ se obtiene de la ecuación $\eta \cdot \Gamma\left(1+\frac{1}{\kappa}\right) = \mu_1$ mediante:

$$\hat{\eta} = \frac{\overline{X}}{\Gamma\left(1+\frac{1}{\hat{\kappa}}\right)} \tag{4.2}$$

Utilización de R para estimar los parámetros de la distribución de Weibull por el método de los momentos. Veamos como podemos utilizar R para resolver numéricamente la ecuación 4.1 y así obtener $\hat{\kappa}$ y $\hat{\eta}$. Para ello supongamos que se desea ajustar una distribución de Weibull a la siguiente muestra de alturas de ola, correspondiente a 30 olas elegidas al azar entre las registradas en una escollera durante un periodo de marea alta:

```
olas = c(2.1, 2.82, 4.2, 6.34, 2.4, 3.1, 2.15, 2.73, 3.12, 2.41, 4.59, 2.81, 2.61,
    3.81, 3.13, 3.06, 5.85, 3.57, 2.64, 4.08, 3.38, 1.88, 1.94, 3.24, 1.98, 3.29,
    0.21, 2.68, 1.74, 4.25)
```

La figura 4.2 muestra el histograma correspondiente a estos datos.

En primer lugar observemos que a partir de la ecuación 4.1, si llamamos:

$$h\left(\hat{\kappa}\right) = \frac{\Gamma\left(1+\frac{2}{\hat{\kappa}}\right)}{\left[\Gamma\left(1+\frac{1}{\hat{\kappa}}\right)\right]^2} - \frac{\frac{1}{n}\sum_{i=i}^{n} X_i^2}{\left(\overline{X}\right)^2}$$

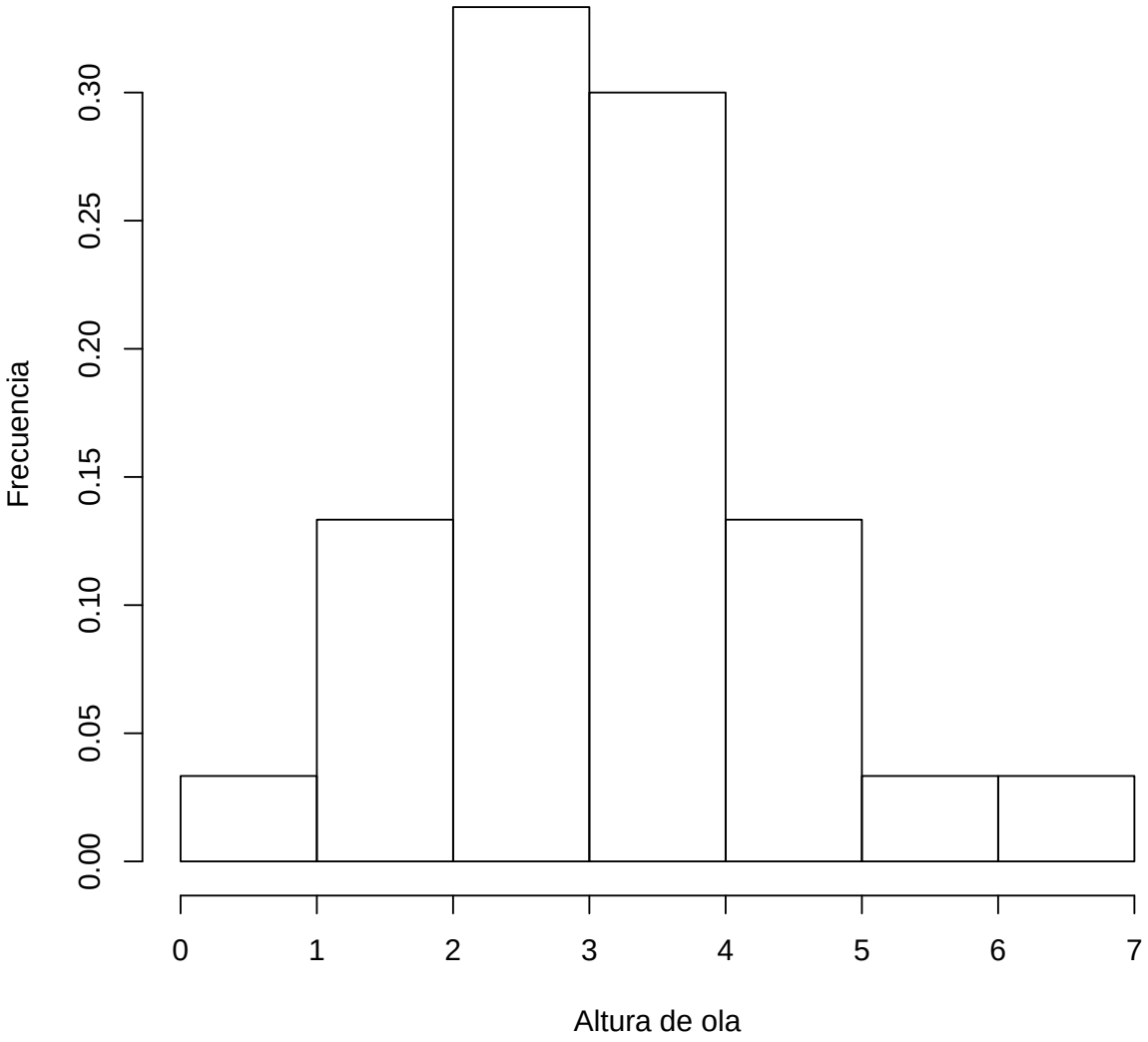

Figura 4.2: Histograma de alturas de ola registradas durante la marea alta en una escollera.

entonces el estimador por el método de los momentos de κ es el valor $\hat{\kappa}$ tal que $h(\hat{\kappa}) = 0$. Por tanto $\hat{\kappa}$ es una raiz de la función h, que puede obtenerse utilizando R mediante la función `uniroot()` que ejecuta un algoritmo de bisección. Ello significa que si proporcionamos un intervalo $[a, b]$ tal que $\text{signo}(h(a)) \neq \text{signo}(h(b))$, `uniroot()` es capaz de encontrar el punto dentro de ese intervalo en el que la función h se anula. Para ello, en primer lugar implementamos la función $h(k)$:

```
h = function(k, x) {
    n = length(x)
    m2 = sum(x^2)/n
    m1 = mean(x)
    return(gamma(1 + 2/k)/gamma(1 + 1/k)^2 - m2/m1^2)
}
```

Nótese que hemos hecho depender la función h no sólo de κ, sino también de la muestra `x` (aquí `x` es un vector que contiene todos los valores de la muestra). Ello permite que dentro de esta función se puedan calcular los momentos de la muestra, necesarios para obtener $h\left(\hat{\kappa}\right)$. Comprobamos que esta función cambia de signo en los extremos del intervalo $[1, 10]$:

```
h(1, olas)

## [1] 0.849

h(10, olas)

## [1] -0.1365
```

lo que indica que esta función tiene una raiz en dicho intervalo. Para obtener esta raiz utilizamos la función *uniroot()*, que nos proporciona el estimador $\hat{\kappa}$ buscado:

```
kappa = uniroot(h, interval = c(1, 10), x = olas)$root
kappa

## [1] 2.785
```

Por último sustituimos este valor en la ecuación 4.2, lo que nos permite obtener $\hat{\eta}$:

```
eta = mean(olas)/gamma(1 + 1/kappa)
eta

## [1] 3.449
```

4.4.3.2. Método de la máxima verosimilitud.

Sea X una variable aleatoria cuya distribución de probabilidad depende uno o varios parámetros desconocidos $\theta_1, \theta_2, \ldots, \theta_k$, y sea $f_\Theta(x)$ su función de probabilidad o de densidad (según que X sea discreta o continua), siendo $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Se desea estimar Θ, y supongamos que para ello disponemos de una muestra aleatoria $(X_1, X_2, \ldots, X_n)$ que ha producido los valores $(x_1, x_2, \ldots, x_n)$. El *método de la máxima verosimilitud* consiste en tomar como estimador de Θ aquel valor que asigna mayor probabilidad al conjunto de valores observado. La idea detrás de este método es que si la muestra aleatoria ha producido los valores $(x_1, x_2, \ldots, x_n)$ es porque debía ser *muy probable* que estos valores se observasen; por tanto los valores que resultan *verosímiles* para Θ son aquellos que hacen que sea muy probable observar $(x_1, x_2, \ldots, x_n)$; y el *más verosímil* es el que maximiza dicha probabilidad.

De un modo más formal, se define la *función de verosimilitud* como:

$$\begin{aligned} L(\Theta) =& L((\theta_1, \theta_2, \ldots, \theta_k) | x_1, x_2, \ldots, x_n) = \\ =& f(x_1, x_2, \ldots, x_n | \Theta = (\theta_1, \theta_2, \ldots, \theta_k)) = f_\Theta(x_1, x_2, \ldots, x_n) \end{aligned}$$

Esta función representa la probabilidad (o densidad) conjunta de las variables $X_1, X_2, \ldots, X_n$ en el punto $(x_1, x_2, \ldots, x_n)$ cuando el valor del parámetro es $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Como $(X_1, X_2, \ldots, X_n)$ una muestra aleatoria, ello significa que las X_i son independientes y con la misma distribución y por tanto su función de probabilidad (o densidad) conjunta es el producto de las funciones de probabilidad (o densidad) de cada variable. Por tanto:

- Si $X_1, X_2, \ldots, X_n$ son variables discretas :

$$L(\Theta) = f_\Theta(x_1, x_2, \ldots, x_n) = P_\Theta(X_1 = x_1) P_\Theta(X_2 = x_2) \cdots P_\Theta(X_n = x_n)$$

  siendo P_Θ la función de probabilidad de las X_i.

- Si $X_1, X_2, \ldots, X_n$ son variables continuas :

$$L(\Theta) = f_\Theta(x_1, x_2, \ldots, x_n) = f_\Theta(x_1) f_\Theta(x_2) \cdots f_\Theta(x_n)$$

  siendo $f_\Theta(x)$ la función de densidad de las X_i.

El *estimador de máxima verosimilitud (estimador MV)* es entonces el valor del parámetro $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ que maximiza esta función:

$$\hat{\Theta} = \arg\max L(\Theta)$$

Este valor puede obtenerse la mayoría de las veces derivando $L(\Theta)$ respecto a cada θ_i, igualando a cero y despejando las θ_i:

$$\frac{\partial}{\partial \theta_i} L(\theta_1, \theta_2, \ldots, \theta_k) = 0,\ i = 1, 2, \ldots, k$$

Notemos que como $L(\Theta)$ es un producto de n términos que dependen de Θ, la obtención de su derivada es en general un proceso complicado (recuérdese como se calcula la derivada de un producto). Por ello, para obtener el máximo de $L(\Theta)$ suele utilizarse en su lugar la log-verosimilitud:

$$\ell(\Theta) = \log(L(\Theta)) = \begin{cases} \sum_{i=1}^{n} \log(P_\Theta(X_i = x_i)) & \text{si las } X_i\text{son discretas.} \\ \sum_{i=1}^{n} \log(f_\Theta(x_i)) & \text{si las } X_i\text{son continuas.} \end{cases}$$

Por ser el logaritmo una función monótona, el máximo de $L(\Theta)$ coincide con el máximo de su logaritmo $\ell(\Theta)$, esto es,

$$\hat{\Theta} = \arg\max\ L(\Theta) = \arg\max \ell(\Theta)$$

siendo la derivada de $\ell(\Theta)$ mucho más sencilla de calcular (ya que la derivada de una suma de términos es simplemente la suma de las derivadas). Por tanto, en la práctica los estimadores de máxima verosimilitud se obtendrán en la mayoría de las ocasiones resolviendo:

$$\frac{\partial}{\partial \theta_i} \ell(\theta_1, \theta_2, \ldots, \theta_k) = 0,\ i = 1, 2, \ldots, k$$

Propiedades de los estimadores de máxima verosimilitud.

Los estimadores de máxima verosimilitud son preferibles a los estimadores obtenidos por el método de los momentos (en algunos casos los estimadores obtenidos por ambos métodos coinciden, aunque no ocurre así en general), ya que gozan de mejores propiedades:

- *Consistencia:* los estimadores MV son consistentes, esto es, a medida que aumenta el tamaño de la muestra es más probable que el valor del estimador esté cada vez más próximo al valor del parámetro.
- *Eficiencia:* a medida que aumenta el tamaño de muestra, los estimadores MV tienen el menor error cuadrático medio de entre los estimadores posibles.
- *Normalidad asintótica:* a medida que aumenta el tamaño de la muestra, los estimadores MV tienden a tener distribución normal.

Ejemplo 4.11. Supongamos que $X \approx exp\left(\frac{1}{\theta}\right)$. En este caso

$$f_\theta(x) = \frac{1}{\theta} e^{-\frac{1}{\theta}x},\ x \geq 0$$

Dada una muestra $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ de esta variable, la función de verosimilitud es:

$$L(\theta) = f_\theta(x_1) \cdot f_\theta(x_2) \cdot \ldots \cdot f_\theta(x_n) = \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta} e^{-\frac{x_2}{\theta}} \cdot \ldots \cdot \frac{1}{\theta} e^{-\frac{x_n}{\theta}} = \left(\frac{1}{\theta}\right)^n e^{-\frac{1}{\theta}(\sum x_i)}$$

Calculando su logaritmo obtenemos la log-verosimilitud:

$$\ell(\theta) = \log(L(\theta)) = n \log\left(\frac{1}{\theta}\right) - \frac{1}{\theta}\sum_{i=1}^{n} x_i = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

Derivamos e igualamos a 0:

$$\ell'(\theta) = 0 \Rightarrow -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

(en el último paso le hemos añadido el símbolo ˆ a θ para indicar que es un estimador). Podemos confirmar que es un máximo hallando la derivada segunda $\ell''(\theta)$ y comprobando que $\ell''(\overline{x}) < 0$.

Ejemplo 4.12. 5. Supongamos que se desea estimar el parámetro p de una variable de Bernoulli, $X \approx Be(p)$ por el método de máxima verosimilitud. Si se ha observado la muestra $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, (donde los x_i son 1 ó 0 según que se obtenga éxito o fracaso), la función de verosimilitud asociada es:

$$\begin{aligned} L(p) &= P(X_1 = x_1) P(X_2 = x_2) \ldots P(X_n = x_n) = \\ &= p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \ldots p^{x_n}(1-p)^{1-x_n} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

La log-verosimilitud será entonces:

$$\ell(p) = \log(L(p)) = \left(\sum_{i=1}^{n} x_i\right)\log(p) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p)$$

Derivamos respecto a p e igualamos a 0:

$$\frac{\partial}{\partial p}\ell\left(p\right)=\left(\sum_{i=1}^{n}x_i\right)\frac{1}{p}-\left(n-\sum_{i=1}^{n}x_i\right)\frac{1}{1-p}=0$$

Despejamos p:

$$\begin{aligned}\left(\sum_{i=1}^{n}x_i\right)\frac{1}{p}&=\left(n-\sum_{i=1}^{n}x_i\right)\frac{1}{1-p}\\ \left(\sum_{i=1}^{n}x_i\right)\left(1-p\right)&=\left(n-\sum_{i=1}^{n}x_i\right)p\\ \sum_{i=1}^{n}x_i&=np\\ \hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i&=\frac{\text{Número de éxitos}}{n}\end{aligned}$$

Como vemos, en este caso hemos obtenido el mismo estimador que por el método de los momentos, si bien en general no tiene por qué ocurrir así.

Ejemplo 4.13. (*modelo de regresión lineal)* Se dispone de n observaciones de dos variables $\{(X_i,Y_i)\,,\,i=1,\ldots,n\}$, siendo las Y_i independientes y tales que, para cada i, $Y_i\approx N\left(\beta_0+\beta_1X_i,\sigma\right)$, con β_0, β_1 y σ parámetros desconocidos. Así, en este modelo se asume que para cada valor fijo $X=x$, la Y sigue una distribución normal con esperanza $E\left[Y\,|X=x\right]=\beta_0+\beta_1x$ y varianza σ^2. Dicho de otra forma, los valores medios de Y siguen la recta $y=\beta_0+\beta_1x$; y los valores individuales de Y se distribuyen alrededor de esta recta, centrados en ella, y con varianza constante σ^2. La figura 4.3 ilustra esta situación.

Este modelo resulta en la práctica adecuado para representar la relación entre muchas variables: talla (X) y peso (Y) de los sujetos adultos de una especie; velocidad del viento (X) y altura de ola (Y); concentración de un compuesto químico (X) y absorbancia medida espectroscópicamente(Y); ...

Para estimar los parámetros β_0, β_1 y σ por máxima verosimilitud debemos determinar primero la función de verosimilitud. Como $Y_i\approx N\left(\beta_0+\beta_1X_i,\sigma\right)$, tenemos que

$$f_{\beta_0,\beta_1,\sigma}\left(y_i\,|X=x_i\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y_i-\left(\beta_0+\beta_1x_i\right)}{\sigma}\right)^2\right)$$

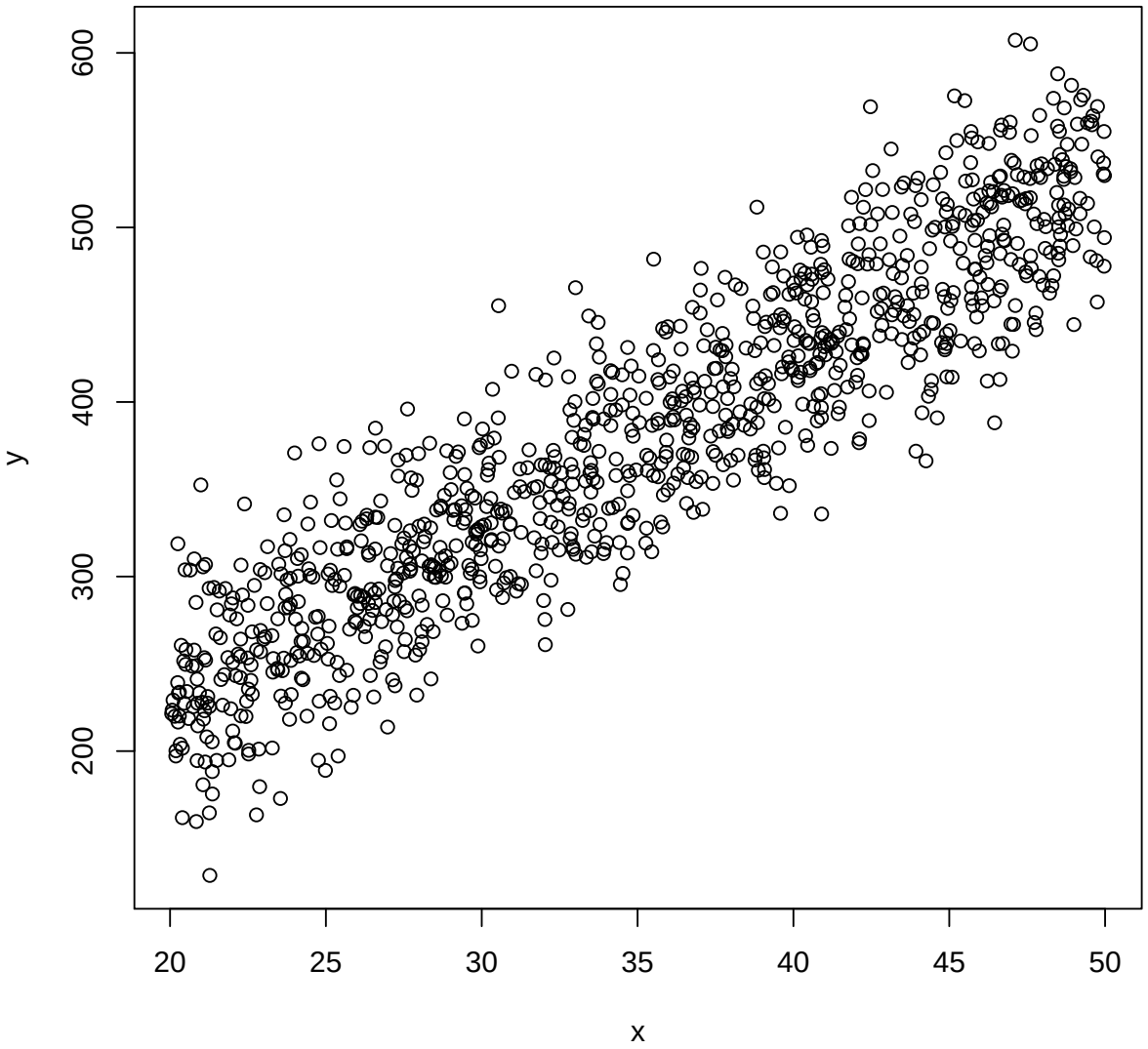

Figura 4.3: Nube de puntos que sigue un modelo de regresión lineal: $Y_i \approx N(\beta_0 + \beta_1 X_i, \sigma)$

Por tanto la función de verosimilitud será:

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f_{\beta_0,\beta_1,\sigma}(y_i) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=i}^{n}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2\right)$$

y la log-verosimilitud:

$$\ell\left(\beta_0,\beta_1,\sigma\right)=-n\log\left(\sigma\right)-n\log\left(\sqrt{2\pi}\right)-\frac{1}{2\sigma^2}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)^2$$

Para obtener los valores de β_0, β_1 y σ que maximizan esta expresión, derivamos e igualamos a 0:

$$\frac{\partial}{\partial\beta_0}\ell\left(\beta_0,\beta_1,\sigma\right)=\frac{1}{\sigma^2}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)=0\Rightarrow\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)=0$$
$$\frac{\partial}{\partial\beta_1}\ell\left(\beta_0,\beta_1,\sigma\right)=\frac{1}{\sigma^2}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)x_i=0\Rightarrow\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)x_i=0$$
$$\frac{\partial}{\partial\sigma}\ell\left(\beta_0,\beta_1,\sigma\right)=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)^2=0\Rightarrow\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)^2=n\sigma^2$$

De la primera ecuación se obtiene:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)=0\Rightarrow\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\beta_0-\sum_{i=1}^{n}\beta_1x_i=0\Rightarrow\sum_{i=1}^{n}y_i-n\beta_0-\beta_1\sum_{i=1}^{n}x_i=0\Rightarrow$$

$$\Rightarrow\beta_0=\frac{\sum_{i=1}^{n}y_i}{n}-\beta_1\frac{\sum_{i=1}^{n}x_i}{n}\Rightarrow\beta_0=\bar{y}-\beta_1\bar{x}\tag{4.3}$$

Sustituyendo en la segunda ecuación:

$$\sum\left(y_i-\beta_0-\beta_1x_i\right)x_i=0\ \ \Rightarrow\ \ \sum_{i=1}^{n}\left(y_i-\left(\bar{y}-\beta_1\bar{x}\right)-\beta_1x_i\right)x_i=0\Rightarrow$$

$$\sum_{i=1}^{n}\left(y_i-\overline{y}\right)x_i-\beta_1\sum_{i=1}^{n}\left(x_i-\overline{x}\right)x_i=0\ \ \Rightarrow\ \ \beta_1=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)x_i}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)x_i}=\frac{\sum_{i=1}^{n}y_ix_i-\overline{y}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2-\overline{x}\sum_{i=1}^{n}x_i}\Rightarrow$$

$$\Rightarrow\hat{\beta}_1=\frac{\sum_{i=1}^{n}y_ix_i-n\overline{x}\,\overline{y}}{\sum_{i=1}^{n}x_i^2-n\left(\overline{x}\right)^2}\tag{4.4}$$

Por último, de la tercera ecuación se obtiene:

$$\sigma^2=\frac{1}{n}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)^2$$

Sustituyendo β_0 por $\overline{y} - \beta_1 \overline{x}$, tras operar y simplificar, queda:

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \overline{y})^2 - \beta_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2\right] \tag{4.5}$$

De esta forma, tras obtener el estimador $\hat{\beta}_1$ utilizando la ecuación 4.4, el estimador $\hat{\beta}_0$ se obtiene sustituyendo $\hat{\beta}_1$ en 4.3 y el estimador $\hat{\sigma}$ sustituyendo $\hat{\beta}_1$ en la ecuación 4.5.

Ejemplo 4.14. Supongamos ahora que tomamos una muestra de n observaciones de una variable con distribución de Weibull de parámetros κ y η. Para estimar estos parámetros por máxima verosimilitud, obtenemos primero la función de verosimilitud:

$$\begin{aligned} L(\kappa, \eta) &= \prod_{i=1}^{n}\left[\frac{\kappa}{\eta}\left(\frac{x_i}{\eta}\right)^{\kappa-1}\exp\left(-(x_i/\eta)^{\kappa}\right)\right] = \\ &= \left(\frac{\kappa}{\eta^{\kappa}}\right)^{n}\left(\prod_{i=1}^{n} x_i\right)^{\kappa-1}\exp\left(-\sum_{i=i}^{n}(x_i/\eta)^{\kappa}\right) \end{aligned} \tag{4.6}$$

La log-verosimilitud es entonces:

$$\ell(\kappa, \eta) = n\log(\kappa) - n\kappa\log(\eta) + (\kappa - 1)\sum_{i=1}^{n}\log(x_i) - \sum_{i=i}^{n}(x_i/\eta)^{\kappa}$$

Para determinar los valores de κ y η que maximizan esta expresión, calculamos las derivadas parciales e igualamos a 0:

$$\begin{aligned} \frac{\partial \ell(\kappa, \eta)}{\partial \kappa} &= \frac{n}{\kappa} - n\log(\eta) + \sum_{i=1}^{n}\log(x_i) - \sum_{i=i}^{n}(x_i/\eta)^{\kappa}\log(x_i/\eta) = 0 \\ \frac{\partial \ell(\kappa, \eta)}{\partial \eta} &= -\frac{n\kappa}{\eta} + \frac{\kappa}{\eta}\sum_{i=i}^{n}(x_i/\eta)^{\kappa} = 0 \end{aligned}$$

De la segunda ecuación se obtiene:

$$\frac{1}{\eta^{\kappa}}\sum_{i=1}^{n} x_i^{\kappa} = n \Rightarrow \eta = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\kappa}\right)^{1/\kappa} \tag{4.7}$$

Reordenamos la primera ecuación:

$$\frac{n}{\kappa} - n\log(\eta) + \sum_{i=1}^{n}\log(x_i) - \frac{1}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa\left(\log(x_i) - \log(\eta)\right) = 0$$

$$\frac{n}{\kappa} - n\log(\eta) + \sum_{i=1}^{n}\log(x_i) - \frac{1}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa\log(x_i) + \frac{\log(\eta)}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa = 0$$

$$\frac{n}{\kappa} + \log(\eta)\left(\frac{1}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa - n\right) + \sum_{i=1}^{n}\log(x_i) - \frac{1}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa\log(x_i) = 0$$

y sustituimos el valor de η:

$$\frac{n}{\kappa} + \frac{1}{\kappa}\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i^\kappa\right)\left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i^\kappa}\sum_{i=i}^{n} x_i^\kappa - n\right) + \sum_{i=1}^{n}\log(x_i) - \frac{n}{\sum_{i=1}^{n} x_i^\kappa}\sum_{i=i}^{n} x_i^\kappa\log(x_i) = 0$$

de donde, tras simplificar, se obtiene:

$$\kappa = \left[\frac{\sum_{i=1}^{n} x_i^\kappa\log(x_i)}{\sum_{i=1}^{n} x_i^\kappa} - \frac{\sum_{i=1}^{n}\log(x_i)}{n}\right]^{-1} \tag{4.8}$$

Esta última ecuación no tiene una solución explícita, debiendo resolverse numéricamente. Una vez que se obtenga de esta manera el valor estimado de $\hat{\kappa}$, se sustituye en la ecuación 4.7 obteniéndose así el estimador máximo verosímil $\hat{\eta}$.

Utilización de R para la estimación de parámetros por el método de máxima verosimilitud.

Como hemos visto en este último ejemplo, la estimación de parámetros por el método de máxima verosimilitud puede ser costosa debido a los cálculos que se deben realizar. Además como también ha ocurrido en este ejemplo, el método no tiene por qué proporcionar soluciones explícitas para los parámetros, por lo que finalmente deben aplicarse métodos numéricos para su obtención. Si bien podríamos proceder con la ecuación 4.8 de modo similar a como ya hicimos para obtener los estimadores por el método de los momentos (definiendo una función que cambie de signo en los extremos y utilizar *uniroot()*), presentamos a continuación un método más general que utiliza la función *optim()* de R para obtener directamente los valores de los parámetros que maximizan la log-verosimilitud.

Para ello es preciso definir primero una función que calcule la log-verosimilitud. En el caso de la distribución de Weibull, la ecuación 4.6 nos da su log-verosimilitud. Su implementación en R es muy sencilla:

```
logver = function(parms, x) {
    k = parms[1]
    eta = parms[2]
    n = length(x)
    lv = n * log(k) - n * k * log(eta) + (k - 1) * sum(log(x)) - sum((x/eta)^k)
    return(lv)
}
```

Como vemos, *logver()* depende de dos vectores: *parms*, que contiene los parámetros de la distribución, y *x* que contiene los valores observados en la muestra. Para obtener ahora los valores de los parámetros que maximizan la log-verosimilitud, utilizaremos *optim()* con los siguientes argumentos:

- *par*: valores iniciales de los parámetros, con los que el algoritmo inicia la búsqueda del óptimo. En este caso usaremos *c(1,1)* (valor que hemos tomado de forma arbitraria). En la siguiente sección presentamos un método que permite obtener estos valores iniciales.
- *logver*: la función a optimizar, en este caso la log-verosimilitud.
- *x=olas*: argumentos adicionales de la función a optimizar, en este caso, los datos muestrales de alturas de ola.
- *control=list(fnscale=-1)*: con esto indicamos que lo que se pretende es *maximizar* la función (por defecto, *optim()* trata de minimizar).

Así pues, la llamada a la función *optim()* se realiza de la forma siguiente:

```
optim(par = c(1, 1), logver, x = olas, control = list(fnscale = -1))$par

## [1] 2.622 3.427
```

La función nos devuelve los valores de los parámetros que maximizan la log-verosimilitud, en el mismo orden en que se definen en la función *logver*, esto es, primero $\hat{\kappa}$ y luego $\hat{\eta}$. Como podemos ver, los valores son ligeramente diferentes a los obtenidos en el ejemplo 4.10 por el

método de los momentos, aún habiendo utilizado los mismos datos. Como hemos señalado más arriba, en general el método de los momentos y el método de máxima verosimilitud no producen exactamente los mismos valores estimados para los parámetros, siendo preferibles los estimadores MV por gozar de mejores propiedades.

Señalemos por último que R implementa las funciones de densidad de muchas distribuciones de probabilidad habituales en la práctica. Ello permite definir la función de log-verosimilitud de una manera alternativa muy simple, teniendo en cuenta que $\ell(\Theta) = \sum_{i=1}^{n} \log(f_\Theta(x_i))$. A modo de ejemplo, en el caso particular de la distribución de Weibull, su función de densidad en R es $f_{\kappa,\eta}(x) =$*dweibull(x,k,eta)*, por lo que la función de log-verosimilitud puede definirse como:

```
logver = function(parms, x) {
    k = parms[1]
    eta = parms[2]
    lv = sum(log(dweibull(x, k, eta)))
    return(lv)
}
```

lo que nos ahorraría tener que escribir explícitamente la función de log-verosimilitud tal como hicimos en la implementación anterior de *logver()*.

Para simplificar aún más las cosas, la librería MASS cuenta con una función específica para el cálculo de estimadores de máxima verosimilitud, la función *fitdistr()*. Para estimar los parámetros de la distribución de Weibull para estos datos simplemente utilizaríamos:

```
library(MASS)
fitdistr(olas, "weibull")
```

```
##    shape    scale
##   2.6214   3.4261
##  (0.3584) (0.2505)
```

Los valores que se muestran entre paréntesis son estimaciones de los errores estándar para el estimador de cada parámetro. Las pequeñas diferencias numéricas que se observan con la solución anterior se deben simplemente a errores de redondeo asociados a los distintos algoritmos de optimización empleados. La función *fitdistr()* reconoce las distribuciones *beta, cauchy, chi-squared, exponential, f, gamma, geometric, log-normal, lognormal,*

logistic, negative binomial, normal, Poisson, t y *weibull*. Si quisiéramos ajustar los parámetros de alguna otra distribución, deberemos implementar una función con la densidad correspondiente (o utilizar el método desarrollado más arriba).

4.4.3.3. Método de los mínimos cuadrados

En el contexto de la estimación de parámetros de una distribución de probabilidad, este método se traduce en localizar los parámetros de la distribución que minimicen los cuadrados de las distancias entre la función de distribución empírica de los datos y la función de distribución teórica correspondiente a dichos parámetros. En la práctica, este método es poco preciso, pero permite obtener estimaciones iniciales de los parámetros que luego se emplean como valores iniciales para la estimación de máxima verosimilitud, tal como hemos visto en la sección anterior.

Para aplicar este método, igual que en los casos anteriores suponemos que se cuenta con una muestra de n observaciones independientes $E = \{x_1, x_2, \ldots, x_n\}$ de una variable aleatoria X con función de distribución acumulativa $F_\Theta(x)$, y que esos valores están ordenados de menor a mayor. Sea $N(x_i)$ el número de observaciones cuyo valor es menor o igual que x_i (obviamente si todas las x_i son distintas, entonces $N(x_i) = i$). Las frecuencias relativas acumuladas $\hat{F}(x_i) = N(x_i)/n$, constituyen una aproximación de la función de distribución $F_\Theta(x)$ de la variable X. Esta aproximación, no obstante, da lugar a que para el valor más alto observado, x_n, se tenga $\hat{F}(x_n) = 1$, lo que de algún modo impone la restricción de que el valor más alto posible es precisamente x_n; ahora bien, que x_n sea el valor más alto observado en esta muestra particular no significa que sea el valor más alto que pueda observarse en general. Para evitar este problema pueden emplearse diversas alternativas, siendo las más frecuentes las siguientes:

$$(a)\ \hat{F}(x_i) = \frac{N(x_i)}{n+1} \qquad (b)\ \hat{F}(x_i) = \frac{N(x_i) - 0{,}5}{n} \qquad (c)\ \hat{F}(x_i) = \frac{N(x_i) - 0{,}3}{N(x_i) + 0{,}4}$$

El método de mínimos cuadrados consiste entonces en encontrar el valor de Θ que minimiza la suma de las diferencias al al cuadrado:

$$SC(\Theta) = \sum_{x_i \in E} \left(\hat{F}(x_i) - F_\Theta(x_i) \right)^2$$

Por tanto el *estimador de mínimos cuadrados* (*estimador MC*) es:

$$\hat{\Theta} = \arg\min SC(\Theta)$$

Ejemplo 4.15. Utilizaremos de nuevo los datos de alturas de ola del ejemplo 4.10, para estimar por mínimos cuadrados los parámetros κ y η de la distribución de Weibull que presumiblemente ha generado esos datos. Para ello consideraremos la estimación *(a)* anterior de la distribución empírica. Asimismo, la función de distribución acumulativa de Weibull que ya hemos visto en el capítulo anterior es de la forma $F_{\kappa,\eta}(x) = 1 - \exp\left(-(t/\eta)^{\kappa}\right)$. Debemos hallar entonces los valores de κ y η que minimizan:

$$SC(\kappa,\eta) = \sum_{i=1}^{n}\left(\hat{F}(x_i) - F_{\kappa,\eta}(x_i)\right)^2 = \sum_{i=1}^{n}\left(\frac{N(x_i)}{n+1} - 1 + \exp\left(-\left(\frac{t}{\eta}\right)^{\kappa}\right)\right)^2$$

Si bien podemos tratar de resolver este problema directamente (derivando con respecto a ambos parámetros, igualando a 0 y resolviendo las ecuaciones resultantes), es más sencillo linealizar el modelo de Weibull. Para ello observemos que:

$$\begin{aligned} 1 - F_{\kappa,\eta}(x) = \exp\left(-\left(\frac{t}{\eta}\right)^{\kappa}\right) \Rightarrow \ln\left(1 - F_{\kappa,\eta}(x)\right) = -\left(\frac{t}{\eta}\right)^{\kappa} \Rightarrow \\ \Rightarrow \ln\left(-\ln\left(1 - F_{\kappa,\eta}(x)\right)\right) = \kappa \ln\left(\frac{x}{\eta}\right) \Rightarrow \\ \Rightarrow \ln\left(-\ln\left(1 - F_{\kappa,\eta}(x)\right)\right) = \kappa \ln(x) - \kappa \ln(\eta) \end{aligned}$$

Esta última ecuación es lineal; llamando:

$$y = \ln\left(-\ln\left(1 - F_{\kappa,\eta}(x)\right)\right); \qquad t = \ln(x); \qquad \theta = -\kappa \ln(\eta)$$

podemos reescribir la ecuación anterior de la forma $y = \kappa t + \theta$. Para estimar entonces κ y η a partir de una muestra ordenada de valores $(x_1, x_2, \ldots, x_n)$ llamaremos:

$$\begin{aligned} \hat{y}_i &= \ln\left(-\ln\left(1 - \hat{F}(x_i)\right)\right) = \ln\left(-\ln\left(1 - \frac{N(x_i)}{n+1}\right)\right) \\ t_i &= \ln(x_i) \end{aligned}$$

y la suma de cuadrados a minimizar será:

$$SC(\kappa,\theta) = \sum_{i=1}^{n}\left(\hat{y}_i - (\kappa t_i + \theta)\right)^2$$

que corresponde a la suma de cuadrados de las distancias entre las observaciones $\hat{y}_i$ y los valores predichos por la recta $y = \kappa t + \theta$. Para obtener los valores de κ y θ que minimizan

$SC(\kappa, \theta)$, derivamos e igualamos a cero:

$$\begin{array}{rcll} \frac{\partial SC(\kappa,\delta)}{\partial \theta} & = & -2\sum_{i=1}^{n}(\hat{y}_i - (\kappa t_i + \theta)) = 0 & \Rightarrow \quad \sum_{i=1}^{n}(\hat{y}_i - (\kappa t_i + \theta)) = 0 \\ \frac{\partial SC(\kappa,\delta)}{\partial \kappa} & = & -2\sum_{i=1}^{n}(\hat{y}_i - (\kappa t_i + \theta))\, t_i = 0 & \Rightarrow \quad \sum_{i=1}^{n}(\hat{y}_i - (\kappa t_i + \theta))\, t_i = 0 \end{array}$$

Estas ecuaciones son análogas a las que ya resolvimos en el ejemplo 4.13 cuando obtuvimos los parámetros de un modelo de regresión lineal por el método de máxima verosimilitud. Por tanto la solución se obtiene del mismo modo, resultando:

$$\hat{\kappa} = \frac{\sum_{i=1}^{n}\hat{y}_i t_i - n\bar{t}\,\bar{\hat{y}}}{\sum_{i=1}^{n} t_i^2 - n\left(\bar{t}\right)^2}$$
$$\hat{\theta} = \bar{\hat{y}} - \hat{\kappa}\bar{t}$$

Por último, como $\theta = -\kappa \ln(\eta)$, se tiene que $\eta = \exp(-\theta/\kappa)$, por lo que $\hat{\eta} = \exp\left(-\hat{\theta}/\hat{\kappa}\right)$.

Podemos utilizar R para realizar esta estimación:

```
x = sort(olas)
Fxi = cumsum(table(x))/(length(x) + 1)
yi = log(-log(1 - Fxi))
ti = log(x)
parms = coef(lm(yi ~ ti))
names(parms) = NULL
kappa = parms[2]
eta = exp(-parms[1]/kappa)
kappa

## [1] 1.689

eta

## [1] 3.78
```

Hemos aprovechado que R cuenta con la función *lm()* que calcula la recta de mínimos cuadrados para predecir *yi* en función de *ti*; asimismo, hemos utilizado la función *coef()* para extraer los coeficientes de esa recta. Tal como puede apreciarse, los valores estimados $\hat{\kappa}$ y $\hat{\eta}$ se

alejan de los que ya obtuvimos por los métodos de los momentos y de máxima verosimilitud pues, como ya se ha dicho, el método de los mínimos cuadrados no es excesivamente preciso. Ahora bien, para utilizar el método de los momentos debimos proporcionar a R un intervalo de búsqueda; y para usar máxima verosimilitud debimos proporcionar unos valores iniciales de los parámetros. Para el método de mínimos cuadrados sólo hemos necesitado los datos. Por tanto, aunque los valores estimados proporcionados por este método no sean muy buenos, pueden utilizarse como valores iniciales para aplicar a continuación el método de máxima verosimilitud.

4.5. Estimación paramétrica con datos censurados.

En ocasiones los datos disponibles para un estudio contienen mediciones incompletas de la variable de interés. Por ejemplo:

1. Se estudia el tiempo que dura la presencia de un contaminante en el entorno costero. Se han realizado 18 ensayos, consistentes en expulsar una cantidad fija del contaminante a través de un emisario submarino y registrar durante cuantos días se detecta en la zona de emisión. Los ensayos duran como mucho una semana y en tres de ellos, al término del ensayo el contaminante aún era detectable. Si X es el número de días que dura la presencia del contaminante, de las 18 observaciones hay tres en las que no se conoce el valor exacto de X, sino sólo que $X \geq 7$.

2. Se dispone de un aparato para medir la altura de ola. Tras sufrir una avería, para las olas de más de 6 metros el aparato registra siempre el valor 6. Si se han observado las alturas de 100 olas y en 12 de ellas el valor registrado es 6, ello quiere decir que en esas 12 observaciones es $X \geq 6$ (siendo X la altura de ola).

3. Se dispone de un aparato para medir la concentración de CO_2 disuelto en el agua de mar. La sensibilidad del aparato es tal que si la concentración está por debajo del valor u, se registra un cero. Por tanto, si el valor 0 se ha registrado k veces durante un periodo de observación, ello significa que en realidad ha habido k valores para los que $X \leq u$ (siendo X la concentración de CO_2).

Cuando se dan estas circunstancias, los datos se dicen *censurados: no se conoce su valor exacto, pero sí que están por debajo (censura por la izquierda) o por encima (censura por la derecha) de cierto valor.* Si se desea estimar los parámetros de las distribuciones de probabilidad de variables como las citadas, sería incorrecto considerar los valores censurados como si fuesen los valores realmente observados en la variable. En el tercero de los ejemplos, si

quisiéramos estimar la concentración media de CO_2 disuelto y considerásemos que los ceros que da el aparato son reales, cuando en realidad son producto de su falta de sensibilidad, es evidente que subestimaríamos la concentración media de CO_2 en la zona de interés.

En presencia de datos censurados, el único método que produce estimaciones fiables es el método de máxima verosimilitud, ya que es posible incluir la presencia de la censura en la función de verosimilitud:

- Si los datos presentan censura por la derecha (como los de los ejemplos 1 y 2 anteriores): sean $x_1, x_2, \ldots, x_r$ las observaciones completas, y $x_{r+1}, x_{r+2}, \ldots, x_n$ las observaciones censuradas (esto es, sólo se sabe que $X_{r+1} \geq x_{r+1}, X_{r+2} \geq x_{r+2}, \ldots, X_n \geq x_n$). La verosimilitud en este caso es:

$$L(\Theta) = f_\Theta(x_1) f_\Theta(x_2) \ldots f_\Theta(x_r) S_\Theta(x_r) S_\Theta(x_{r+2}) \ldots S_\Theta(x_n)$$

  siendo $S_\Theta(x) = 1 - F_\Theta(x) = P(X \geq x)$ la llamada *función de supervivencia de* X.

- Si los datos presentan censura por la izquierda (como los del ejemplo 3 anterior): sean $x_1, x_2, \ldots, x_r$ las observaciones completas, y $x_{r+1}, x_{r+2}, \ldots, x_n$ las observaciones censuradas (esto es, sólo se sabe que $X_{r+1} \leq x_{r+1}, X_{r+2} \leq x_{r+2}, \ldots, X_n \leq x_n$). La verosimilitud en este caso es:

$$L(\Theta) = f_\Theta(x_1) f_\Theta(x_2) \ldots f_\Theta(x_r) F_\Theta(x_r) F_\Theta(x_{r+2}) \ldots F_\Theta(x_n)$$

  siendo $F_\Theta(x) = P(X \leq x)$ la función de distribución acumulativa de X.

Una vez definida la función de verosimilitud con datos censurados, el resto del proceso de estimación es análogo al método de máxima verosimilitud ya visto: derivar la log-verosimilitud con respecto a cada uno de los parámetros, igualar a cero cada derivada y resolver el sistema de ecuaciones resultante.

El lector puede comprobar, a modo de ejemplo, que si $X \approx W(\kappa, \eta)$, los estimadores MV de κ y η en presencia de censura por la derecha se obtienen a partir de:

$$\hat{\kappa} = \left(\frac{\sum_{i=1}^{n} x_i^{\hat{\kappa}} \log(x_i)}{\sum_{i=1}^{n} x_i^{\hat{\kappa}}} - \frac{\sum_{i=1}^{r} \log(x_i)}{r} \right)^{-1}$$

$$\hat{\eta} = \left(\frac{1}{r} \sum_{i=1}^{n} (x_i)^{\kappa} \right)^{1/\kappa}$$

Capítulo 5

Inferencia Estadística II: Estimación por Intervalos de Confianza.

5.1. Introducción.

En el capítulo anterior hemos visto cómo podemos obtener un estimador puntual para un parámetro de una distribución de probabilidad. Si se dan las condiciones adecuadas (error cuadrático medio pequeño, tamaño de muestra suficiente) sabemos que el estimador, al ser evaluado sobre distintas muestras, va a producir valores distintos pero siempre próximos al valor del parámetro que se pretende estimar. Ahora bien, en la práctica, una vez que hemos obtenido la muestra, tenemos un solo valor del estimador, pero ¿cuál es el grado de precisión alcanzado en la estimación? ¿Cuánto se parece este valor estimado al verdadero valor del parámetro? En este capítulo aprenderemos a construir intervalos que podemos confiar en que contienen al parámetro desconocido. La amplitud de estos intervalos, como veremos, nos informa de la precisión alcanzada en la estimación.

Objetivos.

Al finalizar este capítulo el alumno deberá:

1. Conocer y comprender el concepto de intervalo de confianza.
2. Entender la necesidad de acompañar la estimación de parámetros de la estimación de su error estándar y su intervalo de confianza.
3. Ser capaz de calcular los intervalos de confianza más frecuentes en la práctica.

4. Ser capaz de deducir intervalos de confianza a partir de funciones pivote.

5. Ser capaz de deducir intervalos de confianza asintóticos para los estimadores de máxima verosimilitud de una distribución arbitraria.

5.2. Definición de intervalo de confianza.

Dado un parámetro desconocido θ, que caracteriza la distribución de probabilidad de una variable aleatoria determinada, y dada una muestra aleatoria $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$ de dicha variable, diremos que un intervalo de la forma $[\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]$, donde $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$ son variables aleatorias que dependen de la muestra, es un *intervalo de confianza a nivel* $1-\alpha$ *para el parámetro* θ si la probabilidad de que el intervalo contenga a dicho parámetro es $1-\alpha$, esto es:

$$P(\theta \in [\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]) = 1 - \alpha$$

De esta forma, si disponemos de un intervalo de confianza para un parámetro θ desconocido, ya no nos limitaremos a decir que θ tiene un valor parecido a $\hat{\theta}$ (su estimador puntual), sino que además podemos afirmar que con probabilidad $1-\alpha$ (donde α es en general un valor pequeño) el valor de θ se encuentra entre $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$. Ello nos da una idea aproximada de la precisión conseguida en la estimación. Nótese que en la definición de intervalo de confianza, los extremos $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$ son variables aleatorias ya que son funciones de la muestra y ésta es aleatoria. Ello significa que muestras distintas de la misma población producirán intervalos de confianza distintos.

5.3. Intervalo de confianza para la esperanza de una variable $X \approx N(\mu, \sigma)$ con σ conocida.

Supongamos que se desea estimar la esperanza μ de una variable X con distribución normal de varianza σ^2 conocida[1]. Aquí X podría ser el peso que alcanzan los peces de un cultivo marino cuando se les alimenta con cierta dieta experimental, la concentración de un contaminante en la boca de un emisario, el peso mensual de las capturas de una flota, o cualquier otra variable cuya distribución de probabilidad pueda razonablemente considerarse normal.

[1]Debemos confesar que, en la práctica, la varianza σ^2 no se conoce nunca, por lo que el intervalo que vamos a construir carece de interés práctico; no obstante, resulta simple e ilustrativo para entender el concepto y modo de construcción de estos intervalos.

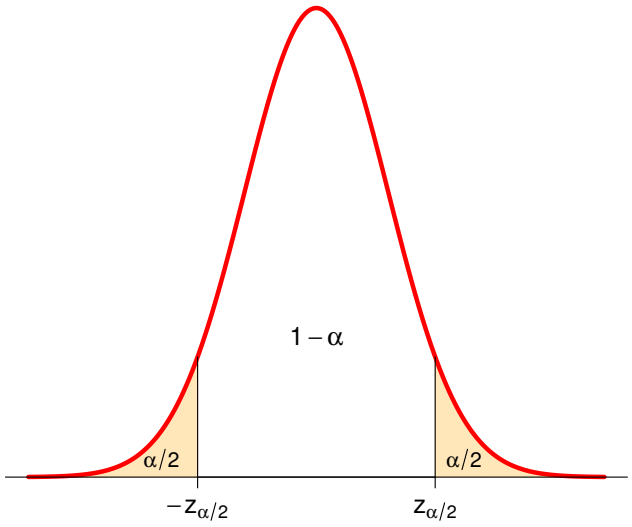

Figura 5.1: Función de densidad de la distribución normal estándar $N(0,1)$. La zona sombreada encierra un área $1-\alpha$. El percentil $z_{\alpha/2}$ es el valor que deja a su derecha un área $\alpha/2$, esto es, $P\left(Z > z_{\alpha/2}\right) = \alpha/2$, por lo que $P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1-\alpha$

En el tema 3 ya hemos visto que, debido a la propiedad reproductiva de la distribución normal, si $\bar{X}$ es la media aritmética de n variables independientes $X_i \approx N(\mu, \sigma)$ entonces:

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

Por tanto, si $z_{\alpha/2}$ es el percentil $1-\alpha/2$ de la distribución normal estándar $N(0,1)$ (véase figura 5.1), se tiene que:

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

de donde:

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

que, tras reordenar términos puede escribirse como:

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

o, de modo análogo:

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

Por tanto, de acuerdo con la definición dada más arriba, el intervalo $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ es un *intervalo de confianza a nivel* $1-\alpha$ para el parámetro μ.

Aplicación a una muestra particular: Se dispone de 20 peces de un cultivo que han estado consumiendo una dieta experimental durante los cuatro últimos meses. Cada pez fue pesado al inicio y al final de este periodo. Los incrementos de peso (en gramos) observados fueron los siguientes:

402	308	261	357	425	378	457	345	372	321
305	370	293	439	363	392	417	452	291	244

Suponiendo que el incremento de peso X experimentado por cada pez en estas condiciones sigue una distribución $N(\mu, \sigma)$, siendo $\sigma = 60$, se desea construir un intervalo de confianza al 95 % para μ.

Para ello basta tener en cuenta que como la confianza buscada es $1-\alpha = 0{,}95$, entonces $\alpha = 0{,}05$ y utilizando la tabla de la $N(0,1)$ encontramos $z_{\alpha/2} = z_{0{,}025} = 1{,}96$. La media aritmética de los 20 valores anteriores es 359.6 gramos, y el intervalo de confianza sería entonces:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[359{,}6 - 1{,}96\frac{60}{\sqrt{20}}, 359{,}6 + 1{,}96\frac{60}{\sqrt{20}}\right] = [333{,}3,\ 385{,}9]$$

Por tanto, con un 95 % de confianza podemos esperar que el incremento medio de peso μ que se obtiene con la citada dieta experimental sea un valor comprendido entre 333.3 y 385.9 gramos.

Cálculo con R : R no incluye ninguna función específica para calcular este intervalo (ya que en la práctica no se presenta nunca una situación en la que se conozca la desviación típica de la población). No obstante, este intervalo de confianza puede calcularse de manera muy sencilla:

```
> incPeso = c(402, 308, 261, 357, 425, 378, 457, 345, 372, 321,
    305, 370, 293, 439, 363, 392, 417, 452, 291, 244)
> sigma = 60
> za2 = qnorm(0.975)
> n = length(incPeso)
> intervalo = mean(incPeso) + c(-1, 1) * za2 * sigma/sqrt(n)
> intervalo

[1] 333.3043 385.8957
```

5.4. Interpretación del intervalo de confianza: ¿por qué el término "confianza"?

Para la determinación del intervalo de confianza que hemos visto en el ejemplo anterior, nos apoyamos en el hecho de que, *antes de obtener la muestra,* la media muestral $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *es una variable aleatoria* con distribución $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. A partir de aquí hemos deducido que:

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

Por tanto, *mientras no se haya obtenido la muestra,* los extremos del intervalo son variables aleatorias y se puede calcular la probabilidad de que dicho intervalo contenga a μ. Ahora bien, una vez que se ha obtenido una muestra, los extremos del intervalo son valores fijos, como 333.3 y 385.9 en el ejemplo anterior. En este momento, el valor de μ estará comprendido entre ellos o no, pero ya no cabe hablar de la probabilidad de que ésto ocurra.

Podemos utilizar el símil del lanzador de cuchillos circense que se dispone a lanzar un cuchillo contra una diana con los ojos vendados. Él sabe, por su experiencia, que la probabilidad de acertar en la diana es del 95 %. Ahora bien, una vez que ha lanzado el cuchillo habrá acertado o no, pero ya no se puede hablar de la probabilidad de que acierte. Si el lanzador continúa con los ojos vendados tras el lanzamiento, *puede confiar* en que ha acertado (incluso, tener mucha confianza en ello, ya que sabe que tiene muy buena puntería), pero no puede estar del todo seguro.

La situación de un investigador que construye un intervalo de confianza a partir de unos datos experimentales es análoga a la del lanzador de cuchillos que nunca se quita la venda de los ojos: *antes de tomar la muestra* sabe que la probabilidad de que el intervalo contenga al parámetro es del 95 %; por tanto, cuando tome los datos y obtenga un intervalo concreto, puede tener mucha confianza (que puede valorar en ese mismo 95 %) en que el intervalo habrá

"*capturado*" al parámetro, pero no puede saber con seguridad si lo ha capturado o no, ya que el valor del parámetro sigue siendo desconocido.

De un modo más general, si para un parámetro θ de una distribución de probabilidad disponemos de dos estadísticos $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$ tales que:

$$P(\theta \in [\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]) = 1 - \alpha$$

siendo $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de dicha distribución, entonces cabe esperar que el $100(1-\alpha)$ % de los intervalos construidos de esta manera contengan a θ y, obviamente, que el restante 100α % no lo contengan. Una vez que obtenemos una muestra particular $(x_1, x_2, \ldots, x_n)$ y calculamos los valores $\hat{\theta}_1 = \theta_1(x_1, x_2, \ldots, x_n)$ y $\hat{\theta}_2 = \theta_2(x_1, x_2, \ldots, x_n)$, tenemos un intervalo concreto $\left[\hat{\theta}_1, \hat{\theta}_2\right]$. En realidad *no sabemos* si este intervalo contiene o no a θ, pero *confiamos* en que sea uno de entre el $100(1-\alpha)$ % de intervalos que contienen al parámetro. De ahí que valoremos nuestra confianza en $1-\alpha$.

El siguiente código en R simula la obtención de 1000 muestras de tamaño 100 de una variable aleatoria $X \approx N(\mu = 10, \sigma = 2)$. Para cada muestra se calculan la media muestral $\bar{X}$ y el intervalo de confianza para μ obtenido en la sección anterior, calculado de acuerdo con la expresión $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$, siendo $\sigma = 2$ y $1 - \alpha = 0{,}95$:

```
> simulaMuestreo = function(n) {
    muestra = rnorm(n, 10, 2)
    intervalo = mean(muestra) + c(-1, 1) * qnorm(0.975) * 2/sqrt(n)
    return(intervalo)
 }
> intervalos = t(replicate(1000, simulaMuestreo(100)))
```

Mostramos los primeros 10 intervalos:

```
> intervalos[1:10, ]
          [,1]      [,2]
 [1,]  9.214422  9.998408
 [2,]  9.868193 10.652178
 [3,]  9.692417 10.476403
 [4,]  9.546502 10.330488
 [5,]  9.560918 10.344904
 [6,]  9.514950 10.298936
 [7,]  9.672468 10.456454
 [8,] 10.120441 10.904426
 [9,]  9.728458 10.512444
[10,]  9.735197 10.519183
```

(obsérvese que en esta simulación particular el octavo intervalo no contiene a la media $\mu = 10$). Ahora contamos cuántos de los 1000 intervalos contienen a μ. Como hemos elegido una confianza del 95 %, esperamos que aproximadamente el 95 % de los intervalos (esto es, unos 950), contengan al parámetro:

```
> numinterv = 0
> for (k in 1:1000) if ((intervalos[k, 1] <= 10) & (10 <= intervalos[k,
      2])) numinterv = numinterv + 1
> numinterv

[1] 944
```

Como vemos, el 94.4 % (muy cerca del 95 %) de los intervalos contiene al parámetro, tal como esperábamos. Se invita al lector a copiar el código anterior y a repetir el experimento varias veces. Podrá comprobar que, efectivamente, en todos los casos el número de intervalos que contienen a la media está siempre en torno al 95 %.

La figura 5.2 representa los 100 primeros intervalos de confianza de la simulación anterior, La línea vertical corresponde al valor de $\mu = 10$. Como vemos, 94 de los intervalos cubren al parámetro y 6 (marcados en rojo) no lo contienen. Remarquemos una vez más, que en la práctica el investigador *toma una única muestra*, no 100 ni 1000. El investigador *confía* (con un nivel de confianza del 95 %) en haber capturado al parámetro. Pero, si ha ocurrido que esa única muestra le lleva a obtener un intervalo de los que se han marcado en rojo entonces, lamentablemente, el parámetro se le habrá escapado, sin que nuestro investigador tenga ningún medio de saberlo.

5.5. Método general de construcción de intervalos de confianza.

El procedimiento de construcción de un intervalo de confianza para un parámetro θ sigue en líneas generales los pasos dados en la sección anterior para obtener el intervalo de confianza para la media μ de una población normal de varianza σ conocida. Partiendo de una muestra aleatoria $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$:

1. Deberemos disponer de una *función pivote* $T(\theta, \mathfrak{X})$ cuya distribución de probabilidad sea conocida y no dependa de θ.

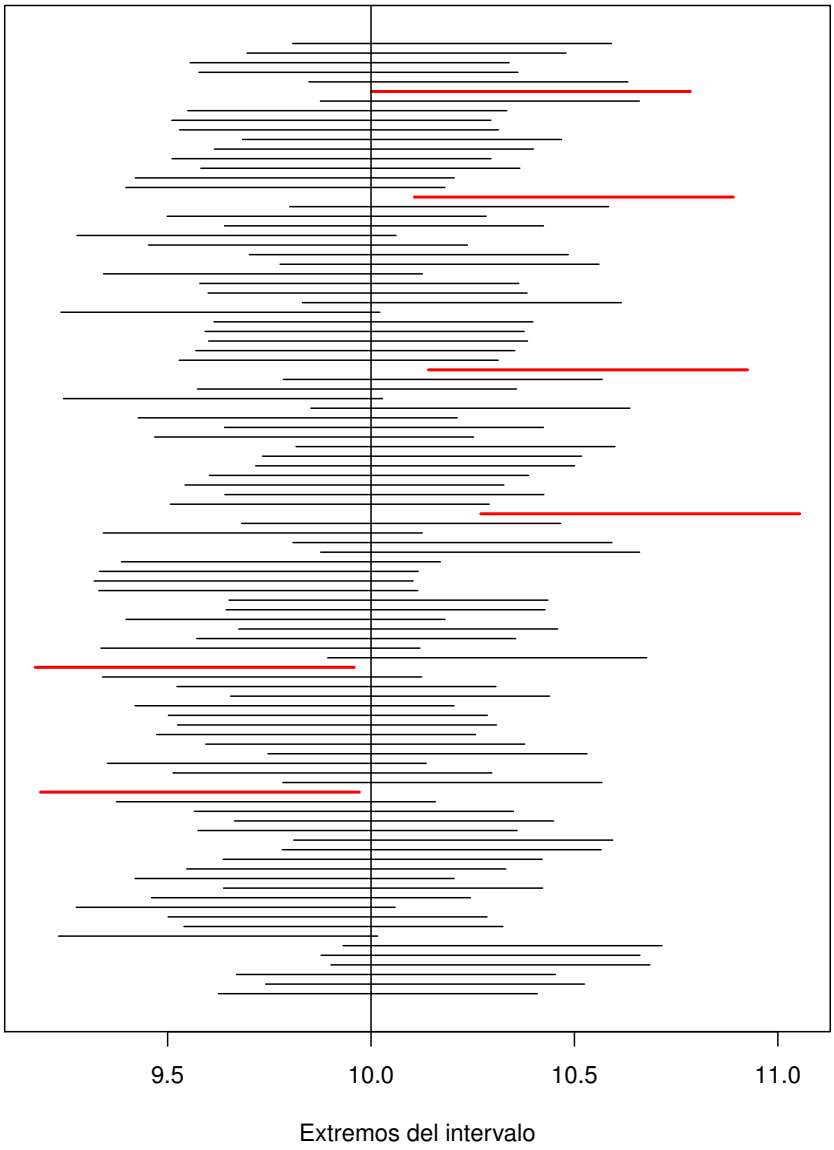

Figura 5.2: 100 intervalos de confianza al 95 % para el parámetro μ de una distribución normal de varianza conocida. En rojo los intervalos que *no* contienen a μ.

2. A partir del pivote y de su distribución de probabilidad deberán encontrarse dos valores $\tau_I(\alpha)$ y $\tau_S(\alpha)$ tales que:

$$P\left(\tau_I(\alpha) \leq T(\theta, X) \leq \tau_S(\alpha)\right) = 1 - \alpha$$

3. Si la función $T(\theta, \mathfrak{X})$ es monótona en θ, las ecuaciones:

$$T(\theta_I, X) = \tau_I(\alpha)$$
$$T(\theta_S, X) = \tau_S(\alpha)$$

tienen solución única. Si $\theta_I(\mathfrak{X}, \alpha)$ y $\theta_S(\mathfrak{X}, \alpha)$ son las respectivas soluciones de estas ecuaciones, se tiene que

$$P(\theta_I(\mathfrak{X}, \alpha) \leq \theta \leq \theta_S(\mathfrak{X}, \alpha)) = 1 - \alpha$$

por lo que el intervalo de confianza a nivel $1 - \alpha$ es $[\theta_I(\mathfrak{X}, \alpha), \theta_S(\mathfrak{X}, \alpha)]$

Ejemplo. Así, para estimar la media μ de una distribución normal de varianza conocida σ^2, la función pivote utilizada fue:

$$T(\mu, \mathfrak{X}) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

cuya distribución de probabilidad es $N(0, 1)$ (y por tanto no depende de μ). En este caso, $\tau_I(\alpha) = -z_{\alpha/2}$ y $\tau_S(\alpha) = z_{\alpha/2}$. Los extremos del intervalo se hallan resolviendo:

$$T(\mu_I, \mathfrak{X}) = \tau_I(\alpha) \Rightarrow \frac{\bar{X} - \mu_I}{\sigma/\sqrt{n}} = -z_{\alpha/2} \Rightarrow \mu_I = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$T(\mu_S, \mathfrak{X}) = \tau_S(\alpha) \Rightarrow \frac{\bar{X} - \mu_S}{\sigma/\sqrt{n}} = z_{\alpha/2} \Rightarrow \mu_s = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

5.6. Intervalo de confianza para la esperanza de una variable $X \approx N(\mu, \sigma)$ con σ desconocida.

Ya hemos visto en la sección 5.3 como contruir un intervalo de confianza para la media de una variable aleatoria con distribución normal de varianza conocida. Este intervalo en la práctica resulta de poca utilidad, toda vez que normalmente la varianza σ^2 es desconocida. Afortunadamente, es posible demostrar que si $X_1, X_2, \ldots$,X es una muestra aleatoria de una distribución $N(\mu, \sigma)$ entonces:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx t_{n-1}$$

siendo $S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$ la desviación típica de la muestra.

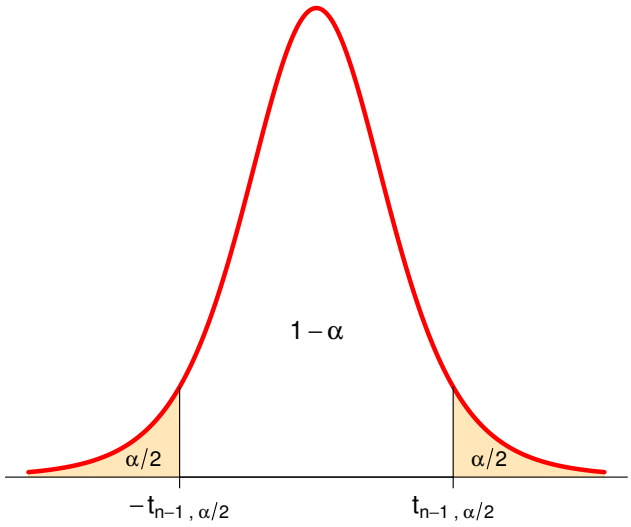

Figura 5.3: Posición de los percentiles $1-\alpha/2$ y $\alpha/2$ de la distribución t de Student denotados, respectivamente, como $t_{n-1,\alpha/2}$ y $-t_{n-1,\alpha/2}$. El área entre estos dos percentiles es $1-\alpha$.

Podemos ahora utilizar las tablas de la t de Student (o R) para encontrar el percentil $t_{n-1,\alpha/2}$ de esta distribución, de tal forma que

$$P\left(-t_{n-1,\alpha/2} \leq t_{n-1} \leq t_{n-1,\alpha/2}\right) = 1-\alpha$$

(ver figura 5.3). Podemos escribir entonces:

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

de donde, operando en el interior del intervalo:

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

o, expresado de otra forma:

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

Así pues, el intervalo de confianza a nivel $1-\alpha$ para la media μ de una distribución $N(\mu,\sigma)$ con σ desconocida es

$$\left[\bar{X}-\frac{S}{\sqrt{n}}t_{n-1,\alpha/2}, \bar{X}+\frac{S}{\sqrt{n}}t_{n-1,\alpha/2}\right]$$

Aplicación a una muestra particular: Consideremos nuevamente los incrementos de peso (en gramos) observados en 20 peces de un cultivo cuando son alimentados con una dieta experimental:

402	308	261	357	425	378	457	345	372	321
305	370	293	439	363	392	417	452	291	244

Si el incremento de peso X experimentado por cada pez en estas condiciones sigue una distribución $N(\mu,\sigma)$, considerando ahora que σ es desconocida, para construir un intervalo de confianza al 95 % para μ, debemos buscar en la tabla de la t de Student el valor $t_{19,0,025} = 2{,}093$. Asimismo, calculamos :

$$s=\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}=\sqrt{\frac{\sum_{i=1}^{20}\left(X_i-359{,}6\right)^2}{19}}=62{,}8$$

El intervalo de confianza es entonces:

$$\left[359{,}6-\frac{62{,}8}{\sqrt{20}}\cdot 2{,}093,\ 359{,}6+\frac{62{,}8}{\sqrt{20}}\cdot 2{,}093, 4{,}8\right]=[359{,}6\pm 29{,}39]=[330{,}21,\ 388{,}99]$$

Por tanto podemos concluir, con una confianza del 95 %, que el incremento medio de peso (en gramos) obtenido en peces alimentados con la dieta experimental se encuentra en el intervalo [330,21, 388,99]; dicho de otro modo, podemos afirmar con una confianza del 95 % que el incremento medio de peso es aproximadamente de 359.6 gramos, con un margen de error de ±29,39 gramos.

Cálculo en R : en R el cálculo del intervalo de confianza es tan simple como escribir el comando:

```
> t.test(incPeso)
```

```
        One Sample t-test

data:  incPeso
t = 25.6066, df = 19, p-value = 3.42e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 330.2072 388.9928
sample estimates:
mean of x
    359.6
```

Como vemos, R proporciona aquí mucha más información que el intervalo de confianza. Además de calcular la media muestral de la variable (mostrada en la última linea) y el intervalo de confianza, R lleva a cabo un *contraste de hipótesis* sobre la media de la población. Explicaremos este concepto en el siguiente capítulo.

Nota: si deseamos que R calcule un intervalo con otro nivel de confianza, por ejemplo 0.9, utilizaríamos la opción *conf.level*:

```
> t.test(incPeso, conf.level = 0.9)
```

5.7. Intervalo de confianza para la varianza σ^2 de una población normal.

Ya hemos visto en el capítulo anterior que la varianza muestral:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

es un estimador centrado de la varianza de la variable aleatoria X cualquiera que sea su distribución de probabilidad. En el caso particular de que $X \approx N(\mu, \sigma)$, dada una muestra aleatoria $\{X_1, X_2, \ldots, X_n\}$ de X, es posible probar que:

$$\frac{(n-1)S^2}{\sigma^2} \approx \chi^2_{n-1}$$

Por tanto, utilizando la tabla de la distribución χ^2_{n-1} (o R) podemos encontrar los percentiles $\chi^2_{n-1,1-\alpha/2}$ y $\chi^2_{n-1,\alpha/2}$ (ver figura 5.4) para los que:

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

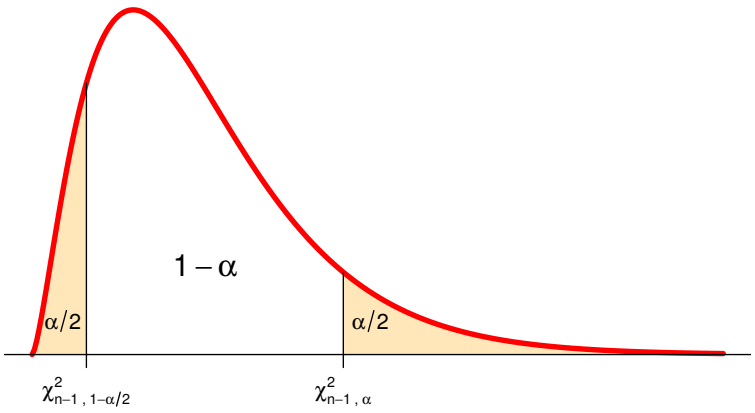

Figura 5.4: Posición de los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución χ^2_{n-1} (denotados, respectivamente, como $\chi^2_{n-1,1-\alpha/2}$ y $\chi^2_{n-1,\alpha/2}$). El área entre estos dos percentiles es $1-\alpha$.

Operando en el interior del intervalo podemos despejar σ^2:

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

Por tanto el intervalo de confianza a nivel $1-\alpha$ para la varianza de una variable aleatoria X con distribución normal $N(\mu,\sigma)$ es:

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

Aplicación a una muestra particular: Siguiendo con nuestro ejemplo del incremento de peso (en gramos) que se obtiene en peces alimentados con una dieta experimental, si deseamos calcular un intervalo de confianza al 95 % para la varianza de esta variable, asumiendo que sigue una distribución normal, y partiendo de la anterior muestra de $n = 20$ peces, en la tabla de la χ^2 encontramos los valores $\chi^2_{19,\,0{,}975} = 8{,}906$ y $\chi^2_{19,\,0{,}025} =$

32,852. La varianza muestral es:

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1} = \frac{\sum_{i=1}^{20} \left(X_i - 359{,}6\right)^2}{19} = 3944{,}25$$

Por tanto, el intervalo de confianza para σ^2 es:

$$\left[\frac{19 \cdot 3944{,}25}{32{,}852}, \frac{19 \cdot 3944{,}25}{8{,}906}\right] = [2281{,}16,\ 8414{,}64]$$

Si queremos calcular el intervalo para la desviación típica $\sigma = \sqrt{\sigma^2}$ basta con aplicar la raiz cuadrada a los extremos del intervalo anterior:

$$\left[\sqrt{\frac{19 \cdot 3944{,}25}{32{,}852}}, \sqrt{\frac{19 \cdot 3944{,}25}{8{,}906}}\right] = [47{,}76,\ 91{,}73]$$

Por tanto podemos concluir, con una confianza del 95 %, que la desviación típica del incremento de peso (en gramos) obtenido por peces alimentados con la nueva dieta experimental se encuentra en el intervalo [47,76, 91,73].

Cálculo en R : en R podemos calcular fácilmente un intervalo de confianza para la varianza del siguiente modo:

```
> n = length(incPeso)
> (n - 1) * var(incPeso)/qchisq(c(0.975, 0.025), n - 1)

[1] 2281.141 8414.154
```

(**Nota:** las diferencias que se observan con el intervalo calculado más arriba obedecen a que en aquel caso hemos utilizado los valores de la tabla de la χ^2, que están redondeados a 3 decimales, mientras que aquí R ha hecho el cálculo con mayor precisión).

En R podemos utilizar también la librería *TeachingDemos*, que implementa la función *sigma.test()* que también calcula el intervalo de confianza para la varianza de una población normal. Para utilizar esta librería debemos cargarla previamente:

```
> library(TeachingDemos)
> sigma.test(incPeso)

        One sample Chi-squared test for variance

data:  incPeso
```

```
X-squared = 74940.8, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
 2281.141 8414.154
sample estimates:
var of incPeso
      3944.253
```

Tal como ocurría también con `t.test()` esta función, además del intervalo de confianza para la varianza, también lleva a cabo un contraste de hipótesis, que se explicará en el siguiente capítulo.

5.8. Intervalo de confianza para el cociente de varianzas de poblaciones normales

En el capítulo 3 hemos visto que si Y_1 e Y_2 son $v$ariables aleatorias independientes con distribuciones de probabilidad respectivas $Y_1 \approx \chi^2_{n_1}$ e $Y_2 \approx \chi^2_{n_2}$, entonces:

$$\frac{Y_1/n_1}{Y_2/n_2} \approx F_{n_1,n_2}$$

Asimismo, en la sección anterior hemos visto también que:

$$\frac{(n-1)S^2}{\sigma^2} \approx \chi^2_{n-1}$$

Así pues, si se dispone de dos muestras aleatorias independientes de tamaños respectivos n_1 y n_2, de dos distribuciones normales con varianzas respectivas σ_1^2 y σ_2^2, llamando $Y_i = (n_i-1)S_i^2/\sigma_i^2$, $i = 1, 2$, de los dos resultados anteriores se sigue que:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \approx F_{n_1-1,n_2-1}$$

Por tanto, utilizando la tabla de la distribución F, podemos encontrar los percentiles $\alpha/2$ y $1-\alpha/2$ de modo que:

$$P\left(F_{n_1-1,n_2-1,1-\alpha/2} \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq F_{n_1-1,n_2-1,\alpha/2}\right) = 1-\alpha$$

Ordenando términos en la desigualdad:

$$P\left(\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right) = 1-\alpha$$

Por tanto el intervalo de confianza a nivel $1-\alpha$ para el cociente de varianzas σ_1^2/σ_2^2 de poblaciones normales es:

$$\left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right]$$

Nota: si sólo se dispone de la tabla F para el nivel $\alpha/2$ utilizaremos la propiedad:

$$F_{n_1-1,n_2-1,1-\alpha/2} = \frac{1}{F_{n_2-1,n_1-1,\alpha/2}}$$

Ejemplo de aplicación: Siguiendo con nuestro ejemplo del incremento de peso (en gramos) que se obtiene en peces alimentados con una dieta experimental, supongamos que se ensaya una segunda dieta en otro tanque con 24 peces, y que los incrementos de peso observados en este caso son:

439	425	345	368	390	424	448	332	452	420	422	311
382	383	419	387	456	500	436	446	385	391	368	405

Obviamente estos incrementos de peso presentan variabilidad (no todos los peces con la misma dieta ganan el mismo peso). Se desea estimar la diferencia entre esta variabilidad y la que se produce cuando se utiliza la primera dieta (ver datos en la página 4).

Las variabilidades de los incrementos de peso con ambas dietas pueden cuantificarse mediante las varianzas muestrales respectivas. Si denotamos por $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$ y $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$ las dos muestras, siendo $n_1 = 20$, $n_2 = 24$, y las medias muestrales respectivas $\bar{X}_1 = 359{,}6$ y $\bar{X}_2 = 405{,}58$, tenemos:

$$s_1^2 = \frac{\sum_{i=1}^{n_1}\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1} = \frac{\sum_{i=1}^{20}\left(X_{1i} - 359{,}6\right)^2}{19} = 3944{,}25$$

$$s_2^2 = \frac{\sum_{i=1}^{n_2}\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1} = \frac{\sum_{i=1}^{24}\left(X_i - 405{,}58\right)^2}{23} = 1895{,}91$$

El cociente entre ambos valores es:

$$\frac{s_1^2}{s_2^2} = \frac{3944{,}25}{1895{,}91} = 2{,}08$$

por lo que la variabilidad *observada* cuando se administra la primera dieta es el doble que cuando se administra la segunda. El intervalo de confianza al 95 % nos ayuda a poner este dato en perspectiva ya que nos proporciona el margen de error probable en esta estimación:

$$\left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right] = \left[\frac{2{,}08}{F_{19,23,0{,}025}}, \frac{2{,}08}{1/F_{23,19,0{,}025}}\right] =$$
$$= \left[\frac{2{,}08}{2{,}374}, \frac{2{,}08}{1/2{,}465}\right] = [0{,}88, 5{,}13]$$

De esta forma vemos que, con la información que tenemos, y con un 95 % de confianza el valor (desconocido) del cociente σ_1^2/σ_2^2 podría llegar a ser tan pequeño como 0.88 o tan grande como 5.13. Nótese que el hecho de que 0.88 sea menor que 1, significa que podría ser que $\sigma_1^2 < \sigma_2^2$; como el valor 1 también está incluído en el intervalo, ello significa que podría ser $\sigma_1^2/\sigma_2^2 = 1$ y por tanto $\sigma_1^2 = \sigma_2^2$; y como el intervalo contiene también valores mayores que 1, ello implicaría que podría ocurrir también que $\sigma_1^2 > \sigma_2^2$. Evidentemente las tres cosas no pueden ocurrir al mismo tiempo, y el resultado que hemos obtenido, en definitiva, nos indica que *no tenemos información suficiente para* distinguir de una manera clara entre las tres situaciones. Por tanto, aunque en las muestras disponibles la varianza observada con la dieta 1 duplique a la varianza observada con la dieta 2, no hay evidencia suficiente para generalizar este resultado, pudiendo achacarse la diferencia observada al puro azar.

Cálculo en R : en R es posible calcular fácilmente un intervalo de confianza para el cociente de varianzas del siguiente modo:

```
> incPeso2 = c(439, 425, 345, 368, 390, 424, 448, 332, 452, 420,
    422, 311, 382, 383, 419, 387, 456, 500, 436, 446, 385, 391,
    368, 405)
> var.test(incPeso, incPeso2)

        F test to compare two variances

data:  incPeso and incPeso2
F = 2.0804, num df = 19, denom df = 23, p-value = 0.0957
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
```

```
 0.8761571 5.1277598
sample estimates:
ratio of variances
          2.080405
```

Al igual que hemos visto en casos anteriores, R no se limita sólo a calcular un intervalo para el cociente de varianzas, sino que presenta además un contraste de hipótesis que se explica en el siguiente capítulo.

5.9. Intervalos de confianza para la diferencia de medias de poblaciones normales.

En muchas ocasiones resulta de interés estimar un intervalo de confianza para la diferencia entre las medias de dos distribuciones normales $X_1 \approx N(\mu_1, \sigma_1)$ y $X_2 \approx N(\mu_2, \sigma_2)$. La diferencia entre las medias muestrales $\bar{X}_1 - \bar{X}_2$ nos permite estimar $\mu_1 - \mu_2$, y el intervalo de confianza nos dará una idea de la precisión conseguida en la estimación. Para ello será preciso disponer de sendas muestras aleatorias de ambas variables. Denotaremos a dichas muestras como $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$ y $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$. El diseño del muestreo puede llevarse a cabo de dos formas:

- **Muestras independientes:** las variables X_1 y X_2 son independientes: el conocimiento de X_1 no aporta información sobre X_2. En general, cuando se utilizan muestras independientes, los sujetos u objetos sobre los que se mide X_1 no tienen relación ni asociación alguna con aquellos sobre los que se mide X_2. Por ejemplo, en un ensayo sobre la ganancia de peso que se consigue con dos dietas distintas, si la primera dieta se experimenta sobre una muestra de n_1 peces en un tanque, y la segunda sobre otros n_2 peces en otro tanque diferente, ambas muestras son independientes. Los valores de n_1 y n_2 pueden ser iguales o distintos.

- **Muestras emparejadas:** las variables X_1 y X_2 están asociadas, y por tanto, el conocimiento de los valores de una aporta información sobre los valores de la otra. En un diseño de muestras emparejadas ambas muestras son del mismo tamaño. Las variables X_1 y X_2 se suelen medir sobre los mismos sujetos u objetos, o bien sobre objetos que han sido cuidadosamente emparejados según características comunes. Por ejemplo, si se desea conocer el incremento medio de peso que se consigue en una semana con una dieta concreta, se pueden utilizar n peces, siendo X_{1i} el peso del pez i-ésimo al inicio del experimento y X_{2i} su peso al final; de esta forma las variables X_1 y X_2 están emparejadas.

5.9.1. Muestras Independientes: Varianzas conocidas.

Si $X_1 \approx N(\mu_1, \sigma_1)$ y $X_2 \approx N(\mu_2, \sigma_2)$, y se toma una muestra de tamaño n_1 de X_1, y una muestra de tamaño n_2 de X_2, siendo ambas muestras independientes, entonces $\bar{X}_1 \approx N\left(\mu_1, \sigma_1/\sqrt{n_1}\right)$ y $\bar{X}_2 \approx N\left(\mu_2, \sigma_2/\sqrt{n_2}\right)$. De acuerdo con la propiedad reproductiva de la distribución normal, se tiene que

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

por lo que:

$$\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

A partir de aquí podemos proceder de modo análogo al caso del intervalo de confianza para la media de una población normal con varianza conocida.

El intervalo de confianza a nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones normales con varianzas conocidas es entonces:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

5.9.2. Muestras Independientes: Varianzas desconocidas e iguales.

Si $X_1 \approx N(\mu_1, \sigma)$ y $X_2 \approx N(\mu_2, \sigma)$, y se dispone de sendas muestras aleatorias independientes de ambas variables, de tamaños respectivos n_1 y n_2 entonces:

$$\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t_{n_1+n_2-2}$$

donde:

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

El intervalo de confianza a nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones normales con la misma varianza (desconocida) es entonces:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

Ejemplo: Siguiendo con nuestro ejemplo del incremento de peso (en gramos) que se obtiene en peces alimentados con dos dietas, sea X_1 el incremento de peso cuando se utiliza la dieta 1 y X_2 el incremento cuando se usa la dieta 2. En este caso podemos asumir que las dos muestras son independientes ya que los datos para cada dieta han sido obtenidos con peces distintos en tanques distintos, sin que haya habido relación ni influencia alguna entre ambos tanques. Si asumimos además que $X_1 \approx N(\mu_1, \sigma_1)$ y $X_2 \approx N(\mu_2, \sigma_2)$, con $\sigma_1 = \sigma_2$, utilizando los datos que hemos visto en las páginas 4 y 16 tenemos:

$$s_p = \sqrt{\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}} = \sqrt{\frac{19 \cdot 3944{,}25 + 23 \cdot 1895{,}91}{42}} = 53{,}13$$

$$\bar{X}_1 = 359{,}6, \qquad \bar{X}_2 = 405{,}58, \qquad \bar{X}_1 - \bar{X}_2 = -45{,}98$$

y por tanto el intervalo de confianza al 95 % es:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = \left[-45{,}98 \pm 2{,}018 \cdot 53{,}13 \cdot \sqrt{\frac{1}{20} + \frac{1}{24}}\right]$$
$$= [-78{,}44, -13{,}52]$$

Así pues, en las muestras disponibles el incremento de peso ha sido, por término medio, casi 46 gramos mayor cuando se usa la dieta 2. Ahora bien, a la hora de generalizar este resultado, con un 95 % de confianza podemos afirmar que con la dieta 2 se ganan, por término medio, entre 13.52 y 78.44 gramos más de peso que con la dieta 1. Por tanto, la dieta 2 produce (con un 95 % de confianza) mayor incremento de peso que la dieta 1.

Cálculo con R : en R es posible calcular fácilmente un intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas iguales utilizando el siguiente comando (nótese el uso del argumento `var.equal=TRUE` con el que se indica que asumimos que las varianzas son iguales):

```
> t.test(incPeso, incPeso2, var.equal = T)

        Two Sample t-test

data:  incPeso and incPeso2
t = -2.8587, df = 42, p-value = 0.006594
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -78.44452 -13.52214
sample estimates:
mean of x mean of y
 359.6000  405.5833
```

Nuevamente, R no se limita sólo a calcular un intervalo para el cociente de varianzas, sino que presenta además un contraste de hipótesis que se explica en el siguiente capítulo.

5.9.3. Muestras Independientes: Varianzas desconocidas y distintas.

En el caso anterior hemos supuesto que las varianzas de las variables X_1 y X_2 son iguales. En la práctica, lo más frecuente es que ambas varianzas sean diferentes. En este caso es posible demostrar que:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_n$$

con

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

El intervalo de confianza a nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones normales con varianzas desconocidas y distintas es entonces:

$$\left[(\bar{X}_1 - \bar{X}_2) \pm t_{n,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Ejemplo: En el caso anterior hemos supuesto la igualdad de las varianzas σ_1^2 y σ_2^2 de los incrementos de peso obtenidos al administrar dos dietas distintas al cultivo de peces de

una misma especie. En la página 17 hemos visto, a partir del cálculo de un intervalo de confianza para el cociente σ_1^2/σ_2^2, que con la evidencia disponible no es posible estar seguros de si ambas varianzas son iguales o distintas. Por ello resulta cuando menos prudente calcular el intervalo de confianza para la diferencia de medias suponiendo que las varianzas son distintas. Bajo este supuesto calculamos en primer lugar:

$$n = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1-1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2-1}} = \frac{\left(\frac{3944{,}25}{20} + \frac{1895{,}91}{24}\right)^2}{\left(\frac{3944{,}25}{20}\right)^2 \frac{1}{19} + \left(\frac{1895{,}91}{24}\right)^2 \frac{1}{23}} = 32{,}91 \cong 33$$

El intervalo de confianza para la diferencia de medias es entonces:

$$\left[(\bar{X}_1 - \bar{X}_2) \pm t_{n,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] = \left[359{,}6 - 405{,}58 \pm t_{33,0{,}025}\sqrt{\frac{3944{,}25}{20} + \frac{1895{,}91}{24}}\right] =$$
$$= [-79{,}79, -12{,}17]$$

Cálculo con R : en R el intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas distintas se calcula mediante el siguiente comando (nótese que ahora NO utilizamos el argumento *var.equal=TRUE*; por defecto R siempre asume que las varianzas de las poblaciones que se comparan son distintas):

```
> t.test(incPeso, incPeso2)

        Welch Two Sample t-test

data:  incPeso and incPeso2
t = -2.7668, df = 32.908, p-value = 0.009215
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -79.79960 -12.16706
sample estimates:
mean of x mean of y
 359.6000  405.5833
```

¿Varianzas iguales o varianzas distintas?: en la práctica, y tal como ha ocurrido en el ejemplo que acabamos de desarrollar, cuando se desea comparar las medias de dos poblaciones normales a partir de la información suministrada por sendas muestras independientes, quien toma los datos no sabe si proceden de poblaciones con varianzas

iguales o de poblaciones con varianzas distintas. ¿Cuál de los dos intervalos anteriores resulta entonces más adecuado?. En general, ambos intervalos resultan muy similares y de hecho, si las muestras son de gran tamaño, ambos intervalos resultan indistinguibles. Cuando las muestras son pequeñas, el intervalo que asume varianzas distintas es siempre algo más amplio que el que asume varianzas iguales. Por tanto el primer intervalo nos garantiza que siempre se alcanza *al menos* la confianza deseada, por lo que resulta preferible. Así, salvo que tengamos razones muy fundadas para pensar que ambas varianzas deban ser iguales, las consideraremos distintas y aplicaremos el intervalo correspondiente a este caso. Como ya hemos mencionado, este es el intervalo que R siempre aplica por defecto.

Variables no normales: Otra cuestión es si las variables cuyas medias se comparan tienen o no distribución normal. Por efecto del teorema central del límite:

En caso de que se disponga de muestras de gran tamaño, aún cuando la distribución de las variables no sea normal, un intervalo de confianza a nivel $1-\alpha$ para la diferencia de medias es:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

En la práctica este intervalo se suele utilizar si n_1 y n_2 son ambos mayores que 30.

En caso de que las variables cuyas medias se comparan no tengan distribución normal, y los tamaños de muestra sean pequeños los intervalos de confianza mostrados en este curso no son de aplicación y debe recurrirse a otras técnicas como el *bootstrap*.

5.10. Muestras emparejadas.

Los intervalos de confianza para las diferencias de medias vistos hasta ahora son de aplicación cuando la comparación se realiza sobre muestras independientes. En el caso de que se utilice un diseño de muestras emparejadas, los valores de X_1 no son independientes de los de X_2. La construcción del intervalo de confianza, no obstante, es sencilla sin más que considerar que si $X_1 \approx N(\mu_1, \sigma_1)$, $X_2 \approx N(\mu_2, \sigma_2)$ y $cov(X_1, X_2) = \sigma_{12}$, entonces la variable $D = X_1 - X_2$ sigue una distribución $N(\mu_D, \sigma_D)$ donde

$$\mu_D = \mu_1 - \mu_2$$
$$\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

Nótese que si $\{X_{11}, X_{12}, \ldots, X_{1n}\}$ y $\{X_{21}, X_{22}, \ldots, X_{2n}\}$, son las muestras de X_1 y X_2, respectivamente, se dispone entonces de una muestra de D, dada por

$$\{D_1, D_2, \ldots, D_n\} = \{X_{11} - X_{21}, X_{12} - X_{22}, \ldots, X_{1n} - X_{1n}\}$$

Por tanto, construir un intervalo para $\mu_1 - \mu_2$ en estas condiciones es equivalente a construir un intervalo de confianza para la media μ_D de una variable normal $N(\mu_D, \sigma_D)$ a partir de la muestra anterior. Si σ_D es desconocida, como suele ser habitual en la práctica, este intervalo según hemos visto en la sección 5.6 es de la forma:

$$\left[\bar{D} - \frac{S_D}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{D} + \frac{S_D}{\sqrt{n}} t_{n-1,\alpha/2}\right]$$

donde

$$\bar{D} = \bar{X}_1 - \bar{X}_2$$

y

$$\begin{aligned} S_D &= \sqrt{\frac{\sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} \left((X_{1i} - X_{2i}) - \left(\bar{X}_1 - \bar{X}_2\right)\right)^2}{n-1}} = \\ &= \sqrt{\frac{\sum_{i=1}^{n} \left(\left(X_{1i} - \bar{X}_1\right) - \left(X_{2i} - \bar{X}_2\right)\right)^2}{n-1}} = \\ &= \sqrt{\frac{\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)^2 + \sum_{i=1}^{n} \left(X_{2i} - \bar{X}_2\right)^2 - 2\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)\left(X_{2i} - \bar{X}_2\right)}{n-1}} = \\ &= \sqrt{S_1^2 + S_2^2 - 2S_{12}} \end{aligned}$$

Por tanto el intervalo de confianza a nivel $1 - \alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de poblaciones normales en muestras emparejadas de tamaño n es:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n-1,\alpha/2} \frac{S_D}{\sqrt{n}}\right]$$

Ejemplo: Se dispone de una muestra de 12 tortugas. De cada ejemplar se han medido las variables X_1=*Longitud* y X_2=*Anchura* del caparazón (en centímetros), con los resultados que se muestran a continuación:

Longitud	82.2	74.5	81.4	81.7	85.8	81.6	82.7	74	78.6	85.9	78	80.3
Anchura	78.4	71.5	74.9	80.1	85.6	80.8	77.5	71.3	76.3	82.7	79.5	79.6

Suponiendo que ambas variables siguen sendas distribuciones normales, se desea calcular un intervalo de confianza al 95 % para la diferencia $\mu_1 - \mu_2$.

Obviamente estos datos corresponden a un diseño de muestras emparejadas, ya que cada pareja de valores Longitud-Anchura se ha medido sobre un mismo ejemplar, por lo que cabe esperar que ambas medidas estén asociadas. Las diferencias entre longitud y anchura observadas para cada tortuga son:

D	3.8	3	6.5	1.6	0.2	0.8	5.2	2.7	2.3	3.2	-1.5	0.7

Se tiene entonces:

$$\bar{X}_1 = 80{,}56 \text{ (Longitud media)}, \qquad \bar{X}_2 = 78{,}18 \text{ (Anchura media)}$$

$$\bar{D} = \bar{X}_1 - \bar{X}_2 = 2{,}38, \qquad S_D = \sqrt{\frac{\sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2}{n-1}} = 2{,}21$$

$$t_{11,0{,}025} = 2{,}201$$

Por tanto, el intervalo de confianza para $\mu_1 - \mu_2$ es

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}}\right] = \left[2{,}38 \pm 2{,}201\frac{2{,}21}{\sqrt{12}}\right] = [2{,}38 \pm 1{,}4] = [0{,}97, 3{,}78]$$

Dicho de otra forma, se estima que la longitud de estas tortugas es, por término medio, 2.38 centímetros mayor que su anchura; con un 95 % de confianza el verdadero valor de esta diferencia se encuentra entre 0.97 y 3.78 centímetros.

Cálculo con R : en R el intervalo de confianza para la diferencia de medias en poblaciones normales con muestras emparejadas se obtiene también con el comando *t.test*, especificando en este caso la opción *paired=TRUE*:

```
> long = c(82.2, 74.5, 81.4, 81.7, 85.8, 81.6, 82.7, 74, 78.6,
    85.9, 78, 80.3)
> anch = c(78.4, 71.5, 74.9, 80.1, 85.6, 80.8, 77.5, 71.3, 76.3,
    82.7, 79.5, 79.6)
> t.test(long, anch, paired = T)
```

```
        Paired t-test

data:  long and anch
t = 3.7187, df = 11, p-value = 0.003390
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9692996 3.7807004
sample estimates:
mean of the differences
                  2.375
```

5.11. Intervalo de confianza para una proporción

La estimación de una proporción es un problema frecuente en la práctica: qué proporción de los huevos de tortuga depositados en una playa eclosionan con éxito, cuál es la proporción de hembras en una especie, qué proporción de los objetos producidos en una fábrica tiene defectos, qué proporción pasa el control de calidad, qué proporción de enfermos experimenta mejoría cuando se les aplica un tratamiento concreto, etc.

Podemos tratar este problema desde un punto de vista general considerando que en estos casos se observa una variable aleatoria X con distribución de Bernoulli de parámetro desconocido π. Recordemos que la variable aleatoria de Bernouilli se caracteriza por tomar uno de dos posibles valores, 1 (éxito*)* ó 0 (*fracaso*), siendo π la probabilidad de éxito. En cada caso particular, el éxito corresponderá a aquel suceso cuya probabilidad queremos estimar: que un huevo de tortuga eclosione, que un ejemplar sea hembra o que un objeto de la producción tenga defectos, por ejemplo.

Sea $\{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de la variable de Bernoulli. Si $N_E = \sum_{i=1}^{n} X_i$ es el número observado de éxitos en la muestra, un estimador de π es:

$$\hat{\pi} = \frac{N_E}{n}$$

esto es, la proporción de éxitos en la muestra. En el capítulo anterior ya hemos visto que este estimador es el que se obtiene tanto por el método de los momentos como por máxima verosimilitud. Sabemos además que el número de éxitos en n pruebas N_E sigue una distribución binomial $B(n, \pi)$, por lo que:

$$E[\hat{\pi}] = E\left[\frac{N_E}{n}\right] = \frac{1}{n}E[N_E] = \frac{1}{n}n\pi = \pi$$

y por tanto $\hat{\pi}$ es un estimador centrado de π.

Ejemplo 5.1. Se han elegido al azar 60 huevos de tortuga en una playa inmediatamente tras la puesta. Transcurrido el periodo de incubación se observa que sólo de 23 de estos huevos nacen tortugas vivas. De esta forma, la proporción de huevos que eclosionan en tortugas vivas puede estimarse como $\hat{\pi} = 23/60 = 0{,}3833 \cong 38{,}33\,\%$.

Para calcular un intervalo de confianza para la proporción π existen varios métodos, que describimos a continuación.

5.11.1. Método de Wilson.

Como $N_E = \sum_{i=1}^{n} X_i \approx B(n, \pi)$, si el valor de n es suficientemente grande (en la práctica si $n\hat{\pi} > 5$ y $n(1-\hat{\pi}) > 5$), entonces, por efecto del teorema central del límite tal como vimos en el capítulo 3:

$$\frac{N_E - n\pi}{\sqrt{n\pi(1-\pi)}} \approx N(0,1)$$

Si observamos que:

$$\frac{N_E - n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{(N_E - n\pi)/n}{\left(\sqrt{n\pi(1-\pi)}\right)/n} = \frac{\frac{N_E}{n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

entonces:

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \approx N(0,1)$$

Por tanto:

$$P\left(-z_{a/2} \leq \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Ahora bien:

$$-z_{a/2} \leq \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq z_{\alpha/2} \Leftrightarrow \left|\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right| \leq z_{\alpha/2} \Leftrightarrow \left(\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right)^2 \leq z_{\alpha/2}^2$$

$$\Leftrightarrow n(\hat{\pi} - \pi)^2 \leq z_{\alpha/2}^2 \pi(1-\pi) \Leftrightarrow \left(n + z_{\alpha/2}^2\right)\pi^2 - \left(2n\hat{\pi} + z_{\alpha/2}^2\right)\pi + n\hat{\pi}^2 \leq 0$$

Si tenemos en cuenta que la función $g(\pi) = \left(n + z_{\alpha/2}^2\right)\pi^2 - \left(2n\hat{\pi} + z_{\alpha/2}^2\right)\pi + n\hat{\pi}^2$ representa una parábola con los brazos abiertos hacia arriba, la desigualdad anterior se verificará para los valores de π comprendidos entre los dos puntos en que esa parábola corta al eje de abcisas.

Estos puntos son las soluciones de la ecuación $\left(n + z^2_{\alpha/2}\right)\pi^2 - \left(2n\hat{\pi} + z^2_{\alpha/2}\right)\pi + n\hat{\pi}^2 = 0$, que se obtienen fácilmente como:

$$\pi = \frac{\left(2n\hat{\pi} + z^2_{\alpha/2}\right) \pm \sqrt{\left(2n\hat{\pi} + z^2_{\alpha/2}\right)^2 - 4\left(n + z^2_{\alpha/2}\right)n\hat{\pi}^2}}{2\left(n + z^2_{\alpha/2}\right)} =$$

$$= \frac{\left(2n\hat{\pi} + z^2_{\alpha/2}\right) \pm \sqrt{4nz^2_{\alpha/2}\hat{\pi}\left(1-\hat{\pi}\right) + z^4_{\alpha/2}}}{2\left(n + z^2_{\alpha/2}\right)} =$$

$$= \frac{\left(n\hat{\pi} + z^2_{\alpha/2}/2\right)}{\left(n + z^2_{\alpha/2}\right)} \pm \frac{z_{\alpha/2}\sqrt{n}}{\left(n + z^2_{\alpha/2}\right)}\sqrt{\hat{\pi}\left(1-\hat{\pi}\right) + z^2_{\alpha/2}/4n}$$

Por tanto, utilizando que $n\hat{\pi} = N_E$:

$$P\left(\pi \in \left[\frac{\left(N_E + z^2_{\alpha/2}/2\right)}{\left(n + z^2_{\alpha/2}\right)} \pm \frac{z_{\alpha/2}\sqrt{n}}{\left(n + z^2_{\alpha/2}\right)}\sqrt{\hat{\pi}\left(1-\hat{\pi}\right) + z^2_{\alpha/2}/4n}\right]\right) = 1 - \alpha$$

Ejemplo de aplicación: Para calcular un intervalo de confianza al 95 % para la proporción de huevos de tortuga que eclosionan con éxito a partir de los datos del ejemplo 5.1, calculamos $\hat{\pi} = 23/60 = 0{,}3833$ y obtenemos $z_{\alpha/2} = z_{0{,}025} = 1{,}96$ en la tabla de la distribución normal. Sustituyendo estos valores en la expresión anterior obtenemos el intervalo:

$$[0{,}39035 \pm 0{,}11947] = [0{,}27088,\ 0{,}50982]$$

por lo que, con un 95 % de confianza dicha proporción se encuentra entre el 27,088 % y el 50,982 % de los huevos que se ponen en esa playa.

Cálculo con R : En el paquete base de R no se encuentra implementado este intervalo. Sí que se encuentra, no obstante, en la librería *binom*, utilizando el comando *binom.confint*. Para los datos de nuestro ejemplo:

```
> library(binom)
> binom.confint(23, 60, method = "wilson")

  method  x  n      mean     lower    upper
1 wilson 23 60 0.3833333 0.2708827 0.509824
```

5.11.2. Método de Agresti-Coull

Este método proporciona un intervalo de confianza para la proporción con una expresión algo más sencilla que la anterior, si bien requiere tamaños muestrales mayores que 40. En estas condiciones se puede utilizar la aproximación:

$$\frac{\pi - \hat{\pi}}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \approx N(0,1)$$

Por tanto:

$$P\left(-z_{a/2} \leq \frac{\hat{\pi} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

y despejando π:

$$P\left(\hat{\pi} - z_{a/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1 - \alpha$$

Así pues, el intervalo de confianza aproximado a nivel $1-\alpha$ para π es:

$$\left[\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

(*Intervalo de Wald*). Este intervalo tiene, no obstante, mal comportamiento para muy diversos valores de n y π, por lo que su uso es desaconsejable. Agresti y Coull han propuesto una modificación de este intervalo que resuelve estos problemas. La modificación consiste en definir:

$$\begin{aligned}\tilde{N}_E &= N_E + z^2_{\alpha/2}/2\\ \tilde{n} &= n + z^2_{\alpha/2}\\ \tilde{\pi} &= \tilde{N}_E/\tilde{n}\end{aligned}$$

y recalcular el intervalo de confianza de Wald sustituyendo $\hat{\pi}$ por $\tilde{\pi}$ y n por $\tilde{n}$. El intervalo de confianza a nivel $1-\alpha$ es entonces de la forma:

$$\left[\tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{\tilde{n}}}\right]$$

(*Intervalo de Agresti y Coull*)

Ejemplo de aplicación: Calculamos de nuevo un intervalo de confianza al 95 % para la proporción de huevos de tortuga que eclosionan con éxito a partir de los datos del ejemplo 5.1, utilizando ahora el método de Agresti-Coull (podemos hacerlo ya que $n > 40$). En este caso se tiene $\tilde{\pi} = 0{,}39035$, $z_{0{,}025} = 1{,}96$ y $\tilde{n} = 63{,}84$. Sustituyendo se obtiene el intervalo:

$$[0{,}39035 \pm 1{,}96 \cdot 0{,}06105] = [0{,}39035 \pm 1{,}96 \cdot 0{,}11964] = [0{,}27069, 0{,}51002]$$

que como puede apreciarse es muy similar al obtenido por el método de Wilson (los extremos se diferencian en menos de una milésima). De hecho, a medida que n aumenta los métodos de Agresti y Coull, y Wilson tienden a producir el mismo intervalo.

Cálculo con R : En el paquete base de R tampoco se encuentra implementado este intervalo, pero al igual que el anterior, podemos encontrarlo en la librería ***binom***, utilizando el comando ***binom.confint*** y especificando el método *"agresti"*. Para los datos de nuestro ejemplo:

```
> library(binom)
> binom.confint(23, 60, method = "agresti")

         method  x  n      mean     lower     upper
1 agresti-coull 23 60 0.3833333 0.2706890 0.5100177
```

Por cierto, que el intervalo de Wald también obtenerse en R con la librería ***binom*** especificando el método "*asymptotic":*

```
> binom.confint(23, 60, method = "asymptotic")

      method  x  n      mean     lower     upper
1 asymptotic 23 60 0.3833333 0.2603104 0.5063562
```

5.11.3. Método de Clopper y Pearson

En el caso de que el tamaño n de la muestra o el valor de la proporción estimada $\hat{\pi}$ sean tan pequeños que no se dan las condiciones para aplicar los métodos de Wilson o Agresti y Coull, puede probarse que el siguiente intervalo garantiza un nivel de confianza de al menos $1 - \alpha$ para la estimación del parámetro π:

$$\left[\frac{N_E}{(n-N_E+1)F_1+N_E}, \frac{(N_E+1)F_2}{(n-N_E)+(N_E+1)F_2}\right]$$

(*Intervalo de Clopper-Pearson*) donde:

$$F_1 = F_{2(n-N_E+1),2N_E,\alpha/2}, \quad F_2 = F_{2(N_E+1),2(n-N_E),\alpha/2}$$

son percentiles de la distribución F de Fisher. Conviene señalar que al ser un intervalo que garantiza que la confianza es al menos $1-\alpha$, en muchas ocasiones el nivel de confianza real será mayor, por lo cual este intervalo resulta en general más amplio y por tanto más impreciso que los anteriores, y sólo debe emplearse si no se dan las condiciones para utilizar alguno de aquéllos.

Ejemplo de aplicación: Si con los datos del ejemplo anterior calculamos el intervalo de Clopper-Pearson, obtenemos:

$$F_1 = F_{2(60-23+1),2\cdot 23,0,025} = F_{76,46,0,025} = 1{,}71636,$$
$$F_2 = F_{2(23+1),2(60-23),0,025} = F_{48,74,0,025} = 1{,}65605$$

y el intervalo es entonces: $\left[\frac{23}{(60-23+1)1{,}71636+23}, \frac{(23+1)\cdot 1{,}65605}{(60-23)+(23+1)\cdot 1{,}65605}\right] = [0{,}26071,\ 0{,}51789]$

Como puede apreciarse este intervalo es similar a los anteriores, aunque algo más amplio. Esta mayor amplitud se debe, como hemos señalado, a que el nivel de confianza de este intervalo es algo mayor que el 95 %.

Cálculo con R : en R el intervalo de Clopper y Pearson se obtiene mediante la función *binom.test*. En la sintaxis debe especificarse primero el número de éxitos N_E, y a continuación el número de pruebas (tamaño de la muestra) n. Así, para los datos del ejemplo anterior utilizaríamos:

```
> binom.test(23, 60)

        Exact binomial test

data:  23 and 60
number of successes = 23, number of trials = 60, p-value = 0.09246
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2607071 0.5178850
sample estimates:
```

```
probability of success
            0.3833333
```

5.12. Intervalos de confianza para la comparación de proporciones en poblaciones independientes.

En ocasiones se desean comparar los parámetros π_1 y π_2 de sendas distribuciones de Bernoulli en poblaciones independientes. Por ejemplo: ¿cuál es la diferencia entre las proporciones de machos en dos especies distintas? ¿Cuál es la diferencia entre las proporciones de enfermos que mejoran con dos tratamientos alternativos? ¿La proporción de microchips defectuosos difiere mucho entre dos técnicas diferentes de fabricación de microchips?. La comparación de dos proporciones puede llevarse a cabo mediante su diferencia $\pi_1 - \pi_2$ o mediante su cociente π_1/π_2. Cada una de las dos proporciones se estima mediante la proporción muestral, por lo que el estimador de la diferencia será $\hat{\pi}_1 - \hat{\pi}_2$ y el del cociente será $\hat{\pi}_1/\hat{\pi}_2$. Como en todos los casos anteriores, en la práctica será conveniente acompañar la estimación por un intervalo de confianza.

Si los tamaños muestrales son grandes, el teorema central del límite nos indica que, aproximadamente:

$$\pi_k \approx N\left(\hat{\pi}_k, \sqrt{\frac{\hat{\pi}_k\left(1-\hat{\pi}_k\right)}{n}}\right), \quad k = 1, 2$$

por lo que

$$\pi_1 - \pi_2 \approx N\left(\hat{\pi}_1 - \hat{\pi}_2, \frac{\hat{\pi}_1\left(1-\hat{\pi}_1\right)}{n_1} + \frac{\hat{\pi}_2\left(1-\hat{\pi}_2\right)}{n_2}\right)$$

de donde se deduce fácilmente que un intervalo de confianza aproximado a nivel $1-\alpha$ para $\pi_1 - \pi_2$ sería de la forma:

$$\left[\left(\hat{\pi}_1 - \hat{\pi}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1\left(1-\hat{\pi}_1\right)}{n_1} + \frac{\hat{\pi}_2\left(1-\hat{\pi}_2\right)}{n_2}}\right]$$

(*intervalo de* Wald*)*. El comportamiento de este intervalo mejora si se introduce una *corrección por continuidad*, tal como se vio en el capítulo 3, en la aproximación de la distribución binomial por la normal. Se obtiene así el *intervalo de Wald corregido*:

$$\left[(\hat{\pi}_1 - \hat{\pi}_2) \pm \left(z_{\alpha/2} \sqrt{\frac{\hat{\pi}_1 (1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2 (1 - \hat{\pi}_2)}{n_2}} + \frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right) \right]$$

Cuando la comparación de las proporciones se realiza a través del cociente, puede probarse que el siguiente intervalo, con muestras grandes, proporciona una confianza aproximada de $1 - \alpha$ para la estimación del logaritmo de π_1/π_2:

$$\ln \left(\frac{\pi_1}{\pi_2} \right) \in \left[\ln \left(\frac{\hat{\pi}_1}{\hat{\pi}_2} \right) \pm z_{\alpha/2} \sqrt{\frac{(1 - \hat{\pi}_1)}{n_1 \hat{\pi}_1} + \frac{(1 - \hat{\pi}_2)}{n_2 \hat{\pi}_2}} \right]$$

Ejemplo: En una playa situada al norte de una isla se han elegido al azar 160 huevos de tortuga, de los cuales 30 habían sido depredados por cangrejos. En otra playa situada al sur, de 125 huevos, 28 presentaban señales de depredación por cangrejos. Se desean calcular intervalos de confianza al 95 % para la diferencia y para el cociente de las proporciones de huevos depredados en ambas playas.

En este caso las proporciones de huevos depredados en cada playa son, respectivamente, $\hat{\pi}_1 = \frac{30}{160} = 0{,}1875 \cong 18{,}75\,\%$ y $\hat{\pi}_2 = \frac{28}{125} = 0{,}224 \cong 22{,}4\,\%$. El intervalo para la diferencia de proporciones es entonces:

$$\left[(0{,}1875 - 0{,}224) \pm \left(1{,}96 \sqrt{\frac{0{,}1875\,(1 - 0{,}1875)}{160} + \frac{0{,}224\,(1 - 0{,}224)}{125}} + \frac{1}{2} \left(\frac{1}{160} + \frac{1}{125} \right) \right) \right]$$

$$= [-0{,}0365 \pm 0{,}1020] = [-0{,}1385, 0{,}0655]$$

Así pues, se estima que en la playa del norte el porcentaje de cangrejos depredados es un 3,65 % inferior al de la playa del sur, si bien el margen de error para esta cifra es tal que con un 95 % de confianza el porcentaje podría oscilar desde un 13.85 % menos a un 6.55 % más, de huevos depredados en el norte que en el sur.

Si deseamos estimar el cociente de proporciones, tenemos que $\hat{\pi}_1/\hat{\pi}_2 = 0{,}1875/0{,}224 = 0{,}8371 \cong 83{,}71\,\%$, esto es, por cada 100 huevos depredados en el sur, sólo se depredan 83.71 en el norte (la tasa de depredación en el norte es un 83.71 % de la del sur). El

intervalo de confianza al 95 % para el logaritmo de este cociente es:

$$\left[\ln(0{,}8371) \pm 1{,}96\sqrt{\frac{(1-0{,}1875)}{30}+\frac{(1-0{,}224)}{28}}\right] = [-0{,}1779 \pm 0{,}4588] =$$

$$= [-0{,}6367, 0{,}2809]$$

y el intervalo al 95 % de confianza para el cociente puede obtenerse sencillamente como:

$$= \left[e^{-0{,}6367}, e^{0{,}2809}\right] = [0{,}5290, 1{,}3244]$$

Por tanto, con un 95 % de confianza podemos decir que, con la incertidumbre que presentan estos datos, la tasa de depredación en el norte podría ser desde poco más de la mitad que la del sur, hasta una vez y un tercio esta última.

Nótese que el intervalo para la diferencia contiene al cero, lo que indica que, con la información que tenemos no es descartable que las tasas de depredación sean iguales en ambas playas. Idéntica conclusión podemos alcanzar observando que el intervalo para el cociente contiene al 1.

Cálculo con R : El intervalo para la diferencia de proporciones puede obtenerse fácilmente en R mediante la función *prop.test(x,n)* donde *x* es un vector con el número de éxitos en cada muestra, y *n* es un vector con los tamaños muestrales. En este caso:

```
> prop.test(c(30, 28), c(160, 125))

        2-sample test for equality of proportions with continuity correction

data:  c(30, 28) out of c(160, 125)
X-squared = 0.3736, df = 1, p-value = 0.5411
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.13849091  0.06549091
sample estimates:
prop 1 prop 2
0.1875 0.2240
```

En cuanto al cociente de proporciones, el paquete base de R no cuenta con ninguna función para la estimación del intervalo de confianza. Aunque es sencillo de calcular, podemos descargar e instalar el paquete *PropCIs*, que incluye la función *riskscoreci(x1,n1,x2,n2,conf)* que sí que implementa este intervalo (con alguna mejora adicional en la aproximación):

```
> library(PropCIs)
> riskscoreci(30, 160, 28, 125, conf = 0.95)

[1] 0.5316599 1.3224231
```

Señalemos, por último, que el cociente de proporciones en la literatura médica se conoce como *Riesgo Relativo*.

5.13. Intervalo de confianza para el parámetro de una distribución exponencial.

Para obtener este intervalo recordemos que si $\{X_1, X_2, \ldots, X_n\}$ es una muestra aleatoria de una distribución $exp(\lambda)$, su suma $T = \sum_{i=1}^{n} X_i$ sigue una distribución gamma $\mathcal{G}\left(n, \frac{1}{\lambda}\right)$ con

$$E[T] = n \cdot \frac{1}{\lambda}$$
$$var(T) = n \cdot \frac{1}{\lambda^2}$$

Si consideramos ahora la variable $V = 2\lambda T = 2\lambda \sum_{i=1}^{n} X_i = 2\lambda n \bar{X}$, como se ha obtenido a partir de T por un simple cambio de escala, entonces V seguirá también una distribución gamma con los parámetros modificados por el mismo factor de misma escala, esto es:

$$E[V] = 2\lambda E[T] = 2\lambda n \frac{1}{\lambda} = 2n$$
$$var(V) = 4\lambda^2 var(T) = 4\lambda^2 n \cdot \frac{1}{\lambda^2} = 4n$$

Por tanto $V = 2\lambda n \bar{X} \approx \mathcal{G}(n, 2) = \chi^2_{2n}$. La tabla de la distribución χ^2 nos permite entonces obtener los percentiles $\chi^2_{2n,1-\alpha/2}$ y $\chi^2_{2n,\alpha/2}$ de forma que:

$$P\left(\chi^2_{2n,1-\alpha/2} \leq V \leq \chi^2_{2n,\alpha/2}\right) = 1 - \alpha$$

Por tanto:

$$P\left(\chi^2_{2n,1-\alpha/2} \leq 2n\lambda\bar{X} \leq \chi^2_{2n,\alpha/2}\right) = 1 - \alpha$$

Dividiendo todos los términos del interior del intervalo por $2n\bar{X}$:

$$P\left(\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}} \leq \lambda \leq \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right) = 1 - \alpha$$

De esta forma el intervalo de confianza a nivel $1-\alpha$ para el parámetro λ de una distribución exponencial calculado a partir de una muestra aleatoria $\{X_1, X_2, \ldots, X_n\}$ con media $\bar{X}$ es:

$$\left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}, \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right]$$

Ejemplo: En una instalación eléctrica, cada vez que se funde un fusible, es reemplazado por otro de iguales características. El tiempo entre reemplazamientos se supone exponencial. A partir de los datos de los últimos 20 fusibles que se han reemplazado, se ha obtenido un tiempo medio entre reemplazamientos de 23 días. Se desea estimar el valor del parámetro λ, así como obtener un intervalo de confianza al 95 % para dicho parámetro.

El estimador de λ es simplemente $\hat{\lambda} = \frac{1}{\bar{X}} = \frac{1}{23} = 0{,}0435$. En la tabla de la distribución χ^2 encontramos los valores $\chi^2_{40,0,975} = 24{,}433$, $\chi^2_{40,0,025} = 59{,}342$. Por tanto el intervalo de confianza al 95 % es:

$$\left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}, \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right] = \left[\frac{24{,}433}{2 \cdot 20 \cdot 23}, \frac{59{,}342}{2 \cdot 20 \cdot 23}\right] = [0{,}0266,\ 0{,}0645]$$

Cálculo con R : R no dispone de ninguna función específica para el cálculo de este intervalo de confianza; no obstante su obtención es elemental. Con los datos del ejemplo anterior:

```
> n = 20
> x = 23
> qchisq(c(0.025, 0.975), 2 * n)/(2 * n * x)
[1] 0.02655765 0.06450186
```

5.14. Intervalo de confianza para el parámetro de una distribución de Poisson

Otra situación frecuente en la práctica es que los datos disponibles procedan de una distribución de Poisson de parámetro λ. Si se dispone de una muestra aleatoria $\{X_1, X_2, \ldots, X_n\}$ de una distribución de Poisson, llamando $T = \sum_{i=1}^{n} X_i$, puede demostrarse que el siguiente intervalo garantiza un nivel de confianza de al menos 1 - α para la estimación del parámetro:

$$\lambda \in \left[\frac{1}{2n}\chi^2_{n_1,1-\alpha/2}, \frac{1}{2n}\chi^2_{n_2,\alpha/2}\right] , \quad n_1 = 2T, \quad n_2 = 2(T+1)$$

Ejemplo. Se realiza un estudio del número de tortugas que acceden diariamente a una playa. Para ello se han seleccionado al azar $n = 40$ días del último año y se ha contado el número de tortugas llegadas a la playa cada día. Durante ese periodo se observó un total de $T = 134$ tortugas. Suponiendo que el número de tortugas diarias sigue una distribución de Poisson, se desea estimar el parámetro de dicha distribución con un intervalo de confianza del 95 %.

El estimador puntual del parámetro, tal como hemos visto en el capítulo anterior es $\hat{\lambda} = \bar{x} = \frac{134}{40} = 3{,}35$. Para obtener el intervalo de confianza calculamos:

$$n_1 = 2T = 2 \cdot 134 = 268, \qquad n_2 = 2\,(134 + 1) = 270$$
$$\chi^2_{268,0,975} = 224{,}5465 \qquad \chi^2_{270,0,025} = 317{,}4092$$

Por tanto, el intervalo de confianza al 95 % es:

$$\left[\frac{1}{80}224{,}5465, \frac{1}{80}317{,}4092\right] = [2{,}807,\ 3{,}968]$$

Cálculo con R : R no dispone de una función específica para el cálculo de este intervalo. No obstante, su cálculo directo es muy simple. Utilizando los datos del ejemplo:

```
> n = 80
> T = 134
> c(qchisq(0.025, 2 * T), qchisq(0.975, 2 * (T + 1)))/(2 * n)

[1] 1.403416 1.983807
```

5.15. Intervalos de confianza aproximados basados en estimadores de máxima verosimilitud.

En todos los casos vistos hasta ahora, la obtención de los intervalos de confianza se ha realizado a través de funciones pivote cuya distribución de probabilidad es conocida y no depende del parámetro a estimar θ, tal como se explicó en la sección 5.5. La obtención de estos pivotes es elemental en algunos casos y más compleja en otros. Pero hay muchos casos en la práctica en que no es posible deducir una función pivote para un parámetro de interés, bien sea por la propia complejidad de la distribución de probabilidad de la variable que se estudia, por la presencia de datos censurados en la muestra[2], o por otras circunstancias. En tales casos, si

[2]Recuérdese del capítulo anterior que un dato censurado es un dato que ofrece sólo información parcial sobre la variable: sabemos de un sujeto que mide más de cierta cantidad, pero no sabemos su longitud exacta;

se dispone de un estimador de máxima verosimilitud para ese parámetro, el siguiente teorema permite utilizarlo para construir intervalos de confianza asintóticos (intervalos de confianza que resultan válidos para tamaños de muestra grandes).

Teorema 5.1. *Sea* $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$ *una muestra aleatoria de una variable* X *con función de densidad* $f_\theta(x)$*, que depende de un parámetro* $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$*. Sea* $L_\mathfrak{X}(\theta)$ *la función de verosimilitud de* θ *dada la muestra* $\mathfrak{X}$*, y sea* $H_\mathfrak{X}(\theta) = \frac{\partial^2 \ln L_X(\theta)}{\partial\theta\,\partial\theta'}$ *la matriz hessiana de segundas derivadas de la log-verosimilitud,* $\ell_\mathfrak{X}(\theta) = \ln(L_\mathfrak{X}(\theta))$*. Bajo las suficientes condiciones de regularidad*[3]*, el estimador de máxima verosimilitud (EMV)* $\hat{\theta}$ *de* θ *es consistente. Además, cuando* $n \to \infty$ *:* $\hat{\theta}_j \approx N\left(\theta_j, \sqrt{\nu_{jj}}\right)$ *siendo* ν_{jj} *el* j*-ésimo elemento de la diagonal de* $-\left(H_\mathfrak{X}(\theta)\right)^{-1}$ *(inversa de la matriz hessiana).*

En la práctica, como el valor de θ no se conoce, la matriz $-\left(H_\mathfrak{X}(\theta)\right)^{-1}$ debe sustituirse por su estimación $-\left(H_\mathfrak{X}\left(\hat{\theta}\right)\right)^{-1}$.

En estas condiciones, el intervalo de confianza aproximado a nivel $1-\alpha$ para el parámetro θ_j, basado en el estimador de máxima verosimilitud $\hat{\theta}$ sería:

$$\left[\hat{\theta}_j \pm z_{\alpha/2}\sqrt{\hat{\nu}_{jj}}\right]$$

siendo $\hat{\nu}_{jj}$ el j-ésimo elemento de la diagonal de $-\left(H_\mathfrak{X}\left(\hat{\theta}\right)\right)^{-1}$

Este resultado, por basarse en la normalidad asintótica de $\hat{\theta}_j$, tal como establece el teorema anterior, es válido sólo cuando $n \to \infty$. En muchas ocasiones se consigue una aproximación razonable a la normalidad para valores de n del orden de 30, si bien ello depende de la distribución de probabilidad de X. Para tamaños de muestra pequeños deben utilizarse otros métodos (*bootstrap, Montecarlo)* que quedan fuera del alcance de este curso.

Nota: la matriz $-\left(H_\mathfrak{X}\left(\hat{\theta}\right)\right)^{-1}$ es un estimador de la *matriz de varianzas-covarianzas* de la *variable aleatoria* $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$. No olvidemos que en muestras distintas se obtienen valores estimados distintos de $\left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$. La variabilidad conjunta de estos valores queda precisamente descrita por su matriz de varianzas-covarianzas. Si $\hat{\nu}_{ij}$ es el término (i,j)-ésimo

sabemos que una célula ha sobrevivido a la acción de un veneno más de 24 horas, pero no sabemos exactamente cuánto ha vivido. Si se utilizan de manera ingenua estos valores censurados para estimar longitudes medias o tiempos medios de supervivencia sin tener en cuenta la presencia de la censura, podemos incurrir en importantes sesgos en la estimación. En el capítulo anterior se señaló como puede construirse una función de verosimilitud que utilice adecuadamente la información de los datos censurados, de forma que el estimador de máxima verosimilitud obtenido a partir de dicha función evita el problema del sesgo.

[3] Condiciones para que exista $H(\Theta)$

de dicha matriz, entonces $\hat{\nu}_{ij}$ es un estimador de $cov\left(\hat{\theta}_i, \hat{\theta}_j\right)$. Asimismo $\hat{\nu}_{jj}$ es un estimador de $var\left(\hat{\theta}_j\right)$.

5.15.1. Ejemplo: cálculo de intervalos de confianza asintóticos para los parámetros de la distribución de Weibull.

Obviamente, calcular los intervalos de confianza asintóticos para los parámetros de una distribución de probabilidad a partir de sus estimadores de máxima verosimilitud puede ser una tarea ardua: calcular la log-verosimilitud, calcular sus derivadas, igualar a cero, despejar los parámetros, calcular las segundas derivadas, ... Afortunadamente R nos permite simplificar enormemente la tarea. Veamos, a modo de ejemplo, como construir intervalos de confianza asintóticos para los parámetros de una distribución de Weibull $W(k, \lambda)$.

Vamos a hacerlo primero de la manera "*difícil*", aplicando paso a paso el teorema anterior. Comenzamos ajustando los parámetros de la distribución $W(\kappa, \lambda)$ por máxima verosimilitud a la variable $X =$"*Altura de ola*". Para ello:

1. Partimos de los datos correspondientes a las alturas medidas en 30 olas:

```
> olas = c(2.1, 2.82, 4.2, 6.34, 2.4, 3.1, 2.15, 2.73, 3.12, 2.41,
    4.59, 2.81, 2.61, 3.81, 3.13, 3.06, 5.85, 3.57, 2.64, 4.08,
    3.38, 1.88, 1.94, 3.24, 1.98, 3.29, 0.21, 2.68, 1.74, 4.25)
```

2. Construimos la función de log-verosimilitud de Weibull, dependiente del vector de parámetros *parms*=(κ, λ), y de la muestra *x*:

```
> logver = function(parms, x) {
    k = parms[1]
    eta = parms[2]
    n = length(x)
    lv = n * log(k) - n * k * log(eta) + (k - 1) * sum(log(x)) -
        sum((x/eta)^k)
    return(lv)
 }
```

3. Hallamos el máximo de esta función de log-verosimilitud mediante la función *optim()*. En este caso, como nos interesa además calcular intervalos de confianza, pediremos a esta función que nos calcule el hessiano mediante la opción *hessian=TRUE*:

```
> EMV = optim(par = c(1, 1), logver, x = olas,
    control = list(fnscale = -1), hessian = TRUE)
> EMV$par

[1] 2.622085 3.426517

> EMV$hessian

          [,1]       [,2]
[1,] -8.571555   3.725557
[2,]  3.725557 -17.562615
```

4. Obtenemos la matriz inversa del hessiano cambiada de signo, $-\left(H_{\mathfrak{x}}\left(\hat{\theta}\right)\right)^{-1}$, y calculamos la raiz de los elementos de su diagonal:

```
> Hinv = solve(EMV$hessian)
> -Hinv

           [,1]       [,2]
[1,] 0.12851401 0.02726167
[2,] 0.02726167 0.06272215

> se = sqrt(diag(-Hinv))
> se

[1] 0.3584885 0.2504439
```

5. Por último construimos los intervalos de confianza para los parámetros:

```
> conf = 0.95
> z = qnorm(1 - (1 - conf)/2)
> EMV$par[1] + c(-1, 1) * z * se[1]

[1] 1.919461 3.324710

> EMV$par[2] + c(-1, 1) * z * se[2]

[1] 2.935656 3.917378
```

Y ahora de la manera *"fácil"* utilizando la función `fitdistr()` de la librería MASS:

```
> library(MASS)
> estimacion = fitdistr(olas, "weibull")
> estimacion
```

```
     shape       scale
 2.6213967   3.4261091
(0.3584319) (0.2504596)

> confint(estimacion)

         2.5 %   97.5 %
shape 1.918883 3.323910
scale 2.935217 3.917001
```

Esta función también proporciona la estimación de la matriz de varianzas-covarianzas $-\left(H_{\mathfrak{x}}\left(\hat{\theta}\right)\right)^{-1}$:

```
> estimacion$vcov

            shape      scale
shape 0.12847341 0.02727454
scale 0.02727454 0.06273002
```

Las ligeras diferencias que se observan entre estos intervalos y los hallados más arriba se deben a errores de redondeo asociados al uso de distintos algoritmos.

5.15.2. Cálculo de intervalos de confianza asintóticos para los parámetros de otras distribuciones.

El procedimiento a seguir es el mismo que acabamos de ver con la distribución de Weibull. El uso de la función *fitdistr()* facilita enormemente esta tarea. Permite estimar los parámetros (e intervalos de confianza) de las siguientes distribuciones de probabilidad: *beta, cauchy, chi-squared, exponential, f, gamma, geometric, log-normal, lognormal, logistic, negative binomial, normal, Poisson, t* y *weibull*.

5.15.3. Intervalos de confianza para funciones de los estimadores de máxima verosimilitud.

En muchas ocasiones el objetivo de la estimación no son los parámetros de la distribución de probabilidad de la variable de interés, sino alguna otra función de los mismos. Si la altura de ola del ejemplo anterior sigue una distribución de Weibull podemos estar interesados no en los parámetros de dicha distribución, sino en estimar cuál es la altura media de ola; o en estimar qué proporción de las olas superará los cuatro metros o quedará por debajo de un metro. Estas cantidades, en general, podrán ponerse como función de los parámetros de la distribución de probabilidad de la altura de ola. Si la estimación de los parámetros de la distribución se

ha llevado a cabo mediante el método de máxima verosimilitud, los siguientes teoremas nos permiten obtener estimaciones de las funciones de interés, e intervalos de confianza, a partir de los estimadores MV (de máxima verosimilitud) de los parámetros.

Teorema 5.2. *Sea $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$ una muestra de n observaciones independientes de una variable aleatoria con función de densidad $f(x)$, que depende de un parámetro $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. Sea $L_{\mathfrak{X}}(\theta)$ la función de verosimilitud de θ dada la muestra $\mathfrak{X}$, y sea $g(\theta)$ una función de $\mathbb{R}^p$ en $\mathbb{R}^k$, $(1 \leq k \leq p)$. Si $\hat{\theta}$ es un estimador MV de θ, entonces $g\left(\hat{\theta}\right)$ es un estimador MV de $g(\theta)$.*

Teorema 5.3. *En las condiciones del teorema anterior, si el valor de parámetro $g(\theta)$ es una función continua y diferenciable, cuando $n \to \infty$:*

$$g\left(\hat{\theta}\right) \approx N\left(g\left(\theta\right), \sigma_g\left(\hat{\theta}\right)\right)$$

siendo $\hat{\theta}$ el estimador MV de θ, y

$$\sigma_g^2\left(\hat{\theta}\right) = \Delta g\left(\hat{\theta}\right)\left\{-H\left(\hat{\theta}\right)\right\}^{-1}\Delta g\left(\hat{\theta}\right)^t$$

$$\Delta g\left(\hat{\theta}\right) = \left(\frac{\partial g}{\partial \theta_1}, ..., \frac{\partial g}{\partial \theta_p}\right)\Big|_{\theta=\hat{\theta}}$$

En estas condiciones, el intervalo de confianza a nivel $1-\alpha$ para $g(\theta)$, basado en el estimador de máxima verosimilitud $\hat{\theta}$ sería:

$$\left[g\left(\hat{\theta}\right) \pm z_{\alpha/2}\sigma_g\left(\hat{\theta}\right)\right]$$

Veamos, a modo de ejemplo, como aplicar estos teoremas para estimar la probabilidad de que la altura de ola supere los 4 metros. Bajo el supuesto de que la altura de ola sigue una distribución $W(\kappa, \lambda)$, la probabilidad de que una ola supere una altura arbitraria h es:

$$g(h) = P(X > h) = \exp\left(-\left(h/\eta\right)^{\kappa}\right) = g_h(\kappa, \eta)$$

1. Implementamos esta función en R , considerando $\theta = (\kappa, \eta)$

```
> g = function(theta, altura) {
      exp(-(altura/theta[2])^theta[1])
  }
```

2. Obtenemos $g\left(\hat{\theta}\right)$ evaluando esta función para $altura = 4$ metros, y utilizando el estimador $\hat{\theta} = (\hat{\kappa}, \hat{\eta}) = (2{,}622, 3{,}427)$ obtenido anteriormente:

```
> gt = g(theta = EMV$par, altura = 4)
> gt

[1] 0.2230288
```

3. Calculamos el gradiente $\Delta g\left(\hat{\theta}\right)$ utilizando la función *grad()* que se encuentra en la librería *numDeriv*:

```
> library(numDeriv)
> Deltag = grad(g, EMV$par, altura = 4)
> Deltag

[1] -0.05178627  0.25608118
```

4. Calculamos $\sigma_g\left(\hat{\theta}\right) = \sqrt{\Delta g\left(\hat{\theta}\right)\left\{-H\left(\hat{\theta}\right)\right\}^{-1}\Delta g\left(\hat{\theta}\right)^t}$:

```
> sg = sqrt(t(Deltag) %*% (-Hinv) %*% Deltag)
> sg

           [,1]
[1,] 0.06111265
```

5. Por último construimos el intervalo de confianza para $g(\theta)$:

```
> conf = 0.95
> z = qnorm(1 - (1 - conf)/2)
> gt + c(-1, 1) * z * sg

[1] 0.1032502 0.3428074
```

De esta forma estimamos que la probabilidad de que una ola supere los 4 metros de altura es 0.223; y además con un 95 % de confianza podemos afirmar que dicha probabilidad se encuentra en el intervalo $[0{,}1033, 0{,}3428]$. Dicho de otra manera, podemos esperar que el 22.3 % de las olas supere los 4 metros, si bien dada la incertidumbre del muestreo, con un 95 % de confianza este porcentaje podría encontrarse en realidad entre el 10.33 % y el 34.28 %.

5.16. Tamaño de la muestra.

Los intervalos de confianza nos permiten determinar el tamaño de muestra necesario para estimar un parámetro con una precisión predeterminada. Para ello, el procedimiento general consiste en fijar el error máximo ε que estamos dispuestos a cometer en la estimación, y el nivel de confianza $1-\alpha$ de la misma. A continuación, utilizando el intervalo de confianza más adecuado para el parámetro que se desea estimar, se iguala el margen de error de dicho intervalo al valor de ε y se despeja el valor de n, que será entonces el tamaño de muestra buscado.

En caso de que el parámetro a estimar dependa de dos muestras de tamaños respectivos n_1 y n_2 (por ejemplo en la estimación de la diferencia de medias, la diferencia de proporciones o el cociente de varianzas), consideraremos que $n_1 = n_2 = n$ y utilizaremos el mismo tamaño muestral para ambas muestras.

Asimismo, en caso de que el intervalo de confianza dependa de alguna cantidad que no se conoce antes de llevar a efecto el muestreo (caso de la varianza muestral o la proporción muestral), podemos recurrir a varias alternativas:

- Tomar una muestra piloto (usualmente una muestra de tamaño reducido que sea posible tomar de forma rápida y con un coste de tiempo y recursos dentro de lo razonable y/o disponible) que nos proporcione un valor aproximado de dicha cantidad.
- Buscar en la literatura referente al problema que nos ocupa valores que puedan resultar razonables en nuestro caso para esa cantidad desconocida.
- Utilizar como valor de n el que resultaría del intervalo más grande posible. Por ejemplo, al estimar una proporción, la longitud del intervalo depende del valor de$\hat{p}$; dicho valor no se conoce antes de tomar la muestra, pero el intervalo más grande (el peor de los posibles) se obtiene cuando $\hat{p} = 1/2$. Este valor es el que se utilizará para despejar n.
- Determinar el tamaño de muestra no para un error absoluto, sino para un error relativo.

5.16.1. Tamaño de muestra para la estimación de la media de una población normal

En este caso, el intervalo de confianza para μ es

$$\left(\bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

Por tanto, si queremos estimar μ con un error máximo ε igualamos:

$$t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = \varepsilon$$

y despejamos n:

$$n = \left(t_{n-1,\alpha/2}\frac{S}{\varepsilon}\right)^2$$

Obviamente, como no se conoce n, no puede calcularse el valor de $t_{n-1,\alpha/2}$. Ahora bien, teniendo en cuenta que para valores grandes de n, la t de Student se aproxima a la normal (y grande en este contexto puede ser del orden de 30), en la ecuación anterior se sustituye el valor $t_{n-1,\alpha/2}$ por $z_{\alpha/2}$ y por tanto el tamaño de la muestra es:

$$n = \left(z_{\alpha/2}\frac{S}{\varepsilon}\right)^2$$

donde el valor de S (desviación típica) habrá de obtenerse por alguno de los métodos señalados anteriormente (muestra piloto o información publicada en la literatura).

Otra alternativa que puede emplearse para resolver este problema es tener en cuenta que:

$$\mu \in \left(\bar{X} \pm t_{n-1,\alpha/2}\tfrac{S}{\sqrt{n}}\right) \Leftrightarrow \mu - \bar{X} \in \left(-t_{n-1,\alpha/2}\tfrac{S}{\sqrt{n}}, t_{n-1,\alpha/2}\tfrac{S}{\sqrt{n}}\right) \Leftrightarrow$$
$$\Leftrightarrow \tfrac{\mu-\bar{X}}{S} \in \left(-\tfrac{t_{n-1,\alpha/2}}{\sqrt{n}}, \tfrac{t_{n-1,\alpha/2}}{\sqrt{n}}\right) \Leftrightarrow \left|\tfrac{\mu-\bar{X}}{S}\right| \leq \tfrac{t_{n-1,\alpha/2}}{\sqrt{n}}$$

y determinar el tamaño n de la muestra de forma que la diferencia relativa (en términos de la desviación típica) entre la media μ desconocida y su estimación muestral $\bar{X}$, sea inferior a un valor δ fijado de antemano, esto es:

$$\left|\frac{\mu - \bar{X}}{S}\right| \leq \delta$$

Para ello basta igualar:

$$\frac{t_{n-1,\alpha/2}}{\sqrt{n}} = \delta$$

y despejar n. Igual que antes, sustituimos $t_{n-1,\alpha/2}$ por $z_{\alpha/2}$, por lo que obtenemos:

$$n = \left(\frac{z_{a/2}}{\delta}\right)^2$$

5.16.2. Tamaño de muestra para la estimación de la varianza de una población normal

El intervalo de confianza a nivel $1-\alpha$ para estimar esta varianza es:

$$\sigma^2 \in \left(\frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}} \right)$$

Si queremos estimar σ^2 con un error máximo ε deberemos determinar n de forma que

$$\frac{1}{2}\left(\frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}} - \frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}} \right) = \varepsilon$$

de donde:

$$(n-1)\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} - \frac{1}{\chi^2_{n-1,\alpha/2}} \right) = \frac{2\varepsilon}{S^2}$$

Esta ecuación no puede resolverse explícitamente, por lo que habrá que probar diversos valores de n. Del mismo modo que en el caso anterior, S^2 no se conoce antes de llevar a cabo el muestreo, por lo que su valor habrá de sustituirse por un valor calculado sobre una muestra piloto, o por un valor máximo razonable que pueda encontrarse en la bibliografía referente al problema en estudio. Otra alternativa es observar que del intervalo de confianza original se sigue que con confianza $1-\alpha$:

$$\frac{\sigma^2}{S^2} \in \left(\frac{(n-1)}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}} \right)$$

y podemos calcular un tamaño de muestra para que, en términos relativos,

$$\left| \frac{\sigma^2}{S^2} - 1 \right| \leq \delta$$

Para conseguir este objetivo bastará con elegir n de tal forma que:

$$(n-1)\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} - \frac{1}{\chi^2_{n-1,\alpha/2}} \right) = 2\delta$$

En R podemos resolver este problema utilizando la función `uniroot()` para encontrar el valor de n tal que:

$$(n-1)\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} - \frac{1}{\chi^2_{n-1,\alpha/2}} \right) - 2\delta = 0$$

Así, por ejemplo, para $\delta = 0{,}4$ y $\alpha = 0{,}05$ el tamaño de muestra necesario puede obtenerse mediante:

```
> dif = function(n, alfa, delta) {
      (n - 1) * (1/qchisq(alfa/2, n - 1) -
          1/qchisq(1 - alfa/2, n - 1)) -
          2 * delta
 }
> n = uniroot(dif, c(2, 1000), alfa = 0.05,
      delta = 0.5)$root
> ceiling(n)

[1] 39
```

La función *ceiling()* se utiliza simplemente para redondear por exceso, ya que habitualmente el valor de n resultante del cálculo anterior no es entero.

5.16.3. Tamaño de muestra para la estimación de la diferencia de medias de poblaciones normales independientes

El intervalo de confianza para la diferencia de medias en poblaciones normales es de la forma:

$$\left((\bar{X}_1 - \bar{X}_2) \mp t_{m,\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

Si hacemos $n = n_1 = n_2$ y aproximamos $t_{m,\alpha/2} \approx z_{\alpha/2}$, el tamaño de muestra n para un error máximo ε se obtiene de:

$$\varepsilon = z_{\alpha/2} \sqrt{\frac{S_1^2 + S_2^2}{n}}$$

esto es:

$$n = \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 \left(S_1^2 + S_2^2 \right)$$

Como siempre, S_1^2 y S_2^2 habrán de obtenerse de una muestra piloto o de alguna otra fuente de información disponible.

5.16.4. Tamaño de muestra para la estimación de una proporción.

Ya hemos visto que si $np > 5$ y $n\,(1-p) > 5$, el intervalo de confianza a nivel $1 - \alpha$ para π es aproximadamente:

$$\pi \in \left(p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right)$$

Entonces, si queremos estimar π con un error inferior a un valor prefijado ε deberemos despejar n de:

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = \varepsilon \Rightarrow n = \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 p\,(1-p)$$

Obviamente, como p es desconocido, esta ecuación no resulta útil. Si se dispone de una estimación previa p (obtenida en una muestra piloto, en una revisión bibliográfica o en un problema similar) puede sustituirse dicha estimación en la fórmula anterior. Otra alternativa consiste en observar que en esta fórmula el valor más grande de n se obtiene cuando $p = 1/2$ (ya que $p(1-p)$ representa una parábola invertida con su máximo en ese valor). Por tanto, en el peor de los casos, si no se tiene información sobre p, sustituiremos el valor $p = 1/2$ en la ecuación anterior, en cuyo caso, el tamaño de muestra es:

$$n = \left(\frac{z_{\alpha/2}}{2\varepsilon}\right)^2$$

que garantiza un error de estimación inferior a ε cualquiera que sea el valor de p.

Capítulo 6

Contrastes de hipótesis

1. Introducción.

En muchas ocasiones el objetivo que se persigue con la realización de un muestreo o de un experimento es poner a prueba alguna hipótesis concebida previamente. Esta es, de hecho, la esencia del método científico: observar, concebir hipótesis y contrastar dichas hipótesis con nuevas observaciones. Ahora bien si, como ocurre frecuentemente, las observaciones están expuestas a fuertes dosis de variabilidad aleatoria, resulta difícil distinguir el efecto que se desea medir de ese "ruido de fondo".

Pongamos un ejemplo sencillo: en un estudio de la morfología de cierta especie, un investigador puede tener *a priori* buenas razones para pensar que los machos deben ser, en promedio, mayores que las hembras. A partir de una muestra aleatoria de 5 machos y 5 hembras, observa en los machos un peso medio de 2,54 kg, frente a 2,77 kg de media en las hembras. ¿Contienen estos datos *evidencia suficiente* para refutar la hipótesis de partida? Es obvio que no todos los animales tienen el mismo peso –*variabilidad natural*– y que, aún siendo cierta la hipótesis de partida, cabe la posibilidad –*por efecto del azar*– de que dicha hipótesis no se verifique.

En este capítulo se desarrollarán los fundamentos básicos para la construcción de contrastes de hipótesis: métodos que, teniendo en cuenta la presencia de la variabilidad y del azar, permitan establecer reglas para decidir si, dentro de ciertos márgenes de error, los datos obtenidos por muestreo o experimentación contienen evidencia suficiente para rechazar la hipótesis de partida o si ésta puede seguir aceptándose como válida.

Una vez establecidos los fundamentos de los contrastes de hipótesis, se estudiarán en particular algunos contrastes de uso frecuente en la práctica, referidos a hipótesis sobre los parámetros de distribuciones de probabilidad conocidas.

Objetivos.

Al finalizar este capítulo el alumno deberá:

1. Conocer y comprender el concepto de contraste de hipótesis.
2. Conocer y comprender los dos tipos de error posibles en un contraste de hipótesis y por tanto los conceptos de nivel de significación y potencia.
3. Conocer, comprender y ser capaz de calcular en algunos casos el p-valor de un contraste.
4. Conocer y ser capaz de aplicar contrastes de hipótesis frecuentes en la práctica, en particular los relativos a medias, varianzas y proporciones.
5. Ser capaz de distinguir las condiciones necesarias para la aplicación de cada contraste de hipótesis.
6. Ser capaz de calcular el tamaño de muestra necesario para la realización de un contraste con significación y potencia predeterminados.
7. Ser capaz de resolver problemas prácticos de contraste de hipótesis utilizando el programa R .

2. Conceptos básicos.

En la actividad científico-técnica práctica, el objetivo que se persigue en muchas ocasiones con la realización de un muestreo o de un experimento es poner a prueba alguna hipótesis concebida previamente.

Por ejemplo:

- Se ha diseñado un nuevo método de depuración de agua, cuyas características físico-químicas inducen a suponer que reducirán la concentración de ciertos contaminantes biológicos con mayor eficiencia que el método que se venía usando hasta ahora. ¿Será verdad esta suposición?
- Se cree que cierto compuesto químico actúa sobre los peces que se crían en tanques de cultivo, reduciendo los niveles de estrés que presentan estos animales al tener que compartir un espacio reducido con un elevado número de congéneres. ¿Es cierta esta conjetura?

- Un método de análisis químico A es mucho más caro que otro método B, pero ¿es realmente mucho más preciso?
- ¿La tasa de mortalidad en cultivos marinos realizados en tanques cerrados es superior a la que se produce en cultivos en mar abierto?

Todos los ejemplos que hemos citado se caracterizan por describir situaciones en las que es imposible realizar un experimento u observación que nos confirme o desmienta *de una manera absolutamente segura* la hipótesis planteada. De ahí que los procedimientos para tomar decisiones sobre la veracidad o falsedad de estas hipótesis hayan de ser necesariamente procedimientos estadísticos, con los que se pretende mantener bajo control el riesgo de tomar decisiones erróneas.

Una hipótesis estadística es una afirmación o conjetura con respecto a alguna característica de interés de la distribución de una variable aleatoria. Llamaremos *hipótesis nula* (H_0) a la hipótesis de partida, que será aceptada como válida si la evidencia en su contra es débil o inexistente. La *hipótesis alternativa* (H_1) será la hipótesis que será aceptada en caso de que se rechace H_0.

Un *contraste de hipótesis* estadístico es una regla de decisión que permita elegir entre la dos hipótesis, H_0 y H_1, en función de la evidencia aportada por los datos disponibles y del riesgo de error que estemos dispuestos a asumir.

Las hipótesis estadísticas pueden plantearse de muy diversas formas:

- En función de los parámetros de la distribución de probabilidad. Por ejemplo, ¿el valor medio de cierta variable en una población es cero?, ¿son iguales las medias de dos poblaciones?, ¿la proporción de sujetos con cierta característica supera el 70 % de la población?
- En términos de la forma de la distribución de la variable de interés: ¿se distribuye una variable de igual forma en dos poblaciones?, ¿es normal la distribución de una variable?.
- En términos de características de asociación: ¿son dos variables independientes?, ¿la relación entre dos variables es lineal?

3. Tipos de Error en los contrastes de hipótesis.

En un contraste de hipótesis es posible cometer dos tipos de error:

Error tipo I: Rechazar la hipótesis nula cuando es verdadera

Error tipo II: Aceptar la hipótesis nula cuando es falsa.

En general, llamaremos:

$$\alpha = P\,(\text{Error Tipo I}) = P\,(\text{Rechazar } H_0\,|H_0 \text{ es cierta})$$
$$\beta = P\,(\text{Error Tipo II}) = P\,(\text{Aceptar } H_0\,|H_0 \text{ es falsa})$$

De esta forma, al realizar un contraste de hipótesis son posibles las siguientes situaciones:

		Realidad	
		H_0 cierta	H_0 falsa
Decisión	Aceptar H_0	Decisión correcta (1-α)	Error II (β)
	Rechazar H_0	Error I (α)	Decisión Correcta (1-β)

La probabilidad α de cometer un error tipo I se conoce como *Nivel de significación* del contraste. Asimismo, la probabilidad de no cometer un error tipo II:

$$1-\beta = P\,(\text{Rechazar } H_0\,|H_0 \text{ es falsa})$$

se conoce como *Potencia del contraste.* Ambas probabilidades, pues, miden la probabilidad de rechazar la hipótesis nula: α cuando es cierta y $1-\beta$ cuando es falsa. La situación ideal es que α sea lo más pequeña posible y $1-\beta$ lo más grande posible. Ello en la práctica se traduce en tener mucha información (muchos datos). Cuando no es posible disponer de toda la información que sería deseable (situación muy frecuente en los estudios reales) en general se procurará que α sea pequeña, aún a costa de que β pueda ser grande (y por ende $1-\beta$ pequeña).

4. Contrastes de Significación.

Supongamos que se desea decidir si el valor (desconocido) de cierto parámetro θ pertenece o no a un conjunto Θ_0. Este parámetro está asociado a la distribución de probabilidad de cierta variable aleatoria X, de la que es posible extraer una muestra aleatoria $(X_1, X_2, \ldots, X_n)$ que contiene información sobre θ. El procedimiento general de los contrastes o pruebas de significación es el siguiente:

1. Fijar las hipótesis nula ($H_0 : \theta \in \Theta_0$) y alternativa ($H_1 : \theta \notin \Theta_0$).
2. Determinar un *estadístico de contraste* dependiente de los datos, $T(X_1, X_2, \ldots, X_n)$, cuya distribución de probabilidad sea conocida cuando H_0 es cierta.
3. Fijar la probabilidad α de error de tipo I (*nivel de significación del* contraste), y determinar una *región crítica* R_C de tal manera que:

$$P(T(X_1, X_2, \ldots, X_n) \in R_C \,|H_o \text{ es cierta}) = \alpha$$

4. Obtener una muestra aleatoria $(X_1, X_2, \ldots, X_n)$ y utilizar la siguiente regla de decisión:

Si $T(X_1, X_2, \ldots, X_n) \in R_C$ *rechazar* H_0*. En caso contrario aceptar* H_0*.*

Observaciones:

1. Con esta regla de decisión se tiene que la probabilidad de error tipo I es:

$$\begin{aligned} P(\text{Error Tipo I}) &= P(\text{Rechazar } H_0 \,|H_0 \text{ es cierta}) = \\ &= P(T(X_1, X_2, \ldots, X_n) \in R_C \,|H_o \text{ es cierta}) = \alpha \end{aligned}$$

2. Al mismo tiempo, la probabilidad de error tipo II queda, en principio, indeterminada:

$$\begin{aligned} P(\text{Error Tipo II}) &= P(\text{Aceptar } H_0 \,|H_0 \text{ es falsa}) = \\ &= P(T(X_1, X_2, \ldots, X_n) \notin R_C \,|H_o \text{ es falsa}) \end{aligned}$$

   si bien, como veremos, puede calcularse para las alternativas de interés, e incluso prefijarse de antemano, fijando un tamaño de muestra adecuado.

3. Para entender el fundamento de los contrastes de significación tengamos en cuenta que, una vez tomados los datos, sólo pueden ocurrir dos cosas: que T caiga en R_C o que no lo haga. Entonces:

a) Si $T \notin R_C$ estaría ocurriendo algo que era muy probable que ocurriese si H_0 fuera cierta ya que, tal como se ha definido R_C, se tiene que:

$$P\left(T\left(X_1, X_2, \ldots, X_n\right) \notin R_C \,|H_o \text{ es cierta}\right) = 1 - \alpha$$

Por tanto, el resultado del test en este caso es el esperado si H_0 es cierta, por lo que nada se opone a aceptar dicha hipótesis. Nótese, no obstante, que aceptar H_0 *no significa* que hayamos demostrado que H_0 sea cierta, sino sólo que los datos no la contradicen. Dicho de otra forma *aceptamos* H_0 *no porque hayamos podido probar que es cierta, sino porque no hemos podido probar que es falsa.*

b) Si $T \in R_C$ estaría ocurriendo algo que*,* de ser H_0 cierta, muy difícilmente podía haber ocurrido. Pero como de hecho ha ocurrido, ello nos indica que los datos contienen una fuerte evidencia de que H_0 es posiblemente falsa o, lo que es lo mismo, una fuerte evidencia de que H_1 es posiblemente cierta.

4. Nótese la no simetría de las dos posibles conclusiones del contraste:

   a) Cuando se acepta H_0 es porque la evidencia en su contra es débil.

   b) Cuando se acepta H_1 es porque la evidencia a su favor es fuerte.

   Por esta razón, cuando planteamos un contraste de hipótesis se debe colocar como hipótesis alternativa aquella de la que queramos tener fuerte evidencia a su favor en caso de que finalmente sea aceptada. La hipótesis nula, en cambio, es la que se aceptará por defecto si no hay fuerte evidencia en su contra (e incluso si no hay fuerte evidencia a su favor).

   Por todo ello, cuando un test concluye con la aceptación de H_0 se dice que ha resultado *no significativo,* y cuando concluye con su rechazo se dice que ha resultado *significativo.*

5. La región crítica R_C suele denominarse también *región de rechazo* (de H_0). La región complementaria se denomina *Región de Aceptación,* R_A. Obviamente

   $$P\left(T\left(X_1, X_2, \ldots, X_n\right) \in R_A \,|H_o \text{ es cierta}\right) = 1 - \alpha$$

   La región de aceptación contiene, pues, los valores del estadístico $T\left(X_1, X_2, \ldots, X_n\right)$ que, con mucha probabilidad, podrían observarse *por puro azar* si H_0 fuese cierta.

Ejemplo 6.1.

Las algas de cierta especie que se cultivan con fines farmacológicos son muy sensibles al pH del agua. Se ha observado que el desarrollo de estas algas es óptimo cuando el pH promedio es 1, y diariamente se realizan controles con el objetivo de aplicar medidas correctoras (añadir aditivos químicos al agua) si el pH se aparta de este valor. Estos controles consisten en tomar 5 muestras de agua y evaluar el pH medio. En un día en que el pH medio de las cinco muestras es de 1.2 con una desviación típica de 0.4. ¿sería preciso aplicar alguna medida correctora? (se supone que la distribución del pH es normal)

1. Si llamamos μ al pH medio real del agua, el problema puede plantearse como el contraste de hipótesis:
$$\begin{cases} H_0 : \mu = 1 \\ H_1 : \mu \neq 1 \end{cases}$$
siendo la información disponible la aportada por una muestra de cinco valores de pH, $\{X_1, X_2, X_3, X_4, X_5\}$.

2. Como no conocemos el valor de μ, podemos estimarlo mediante la media muestral $\bar{X}$. Si H_0 fuera verdad, entonces el valor de $\bar{X}$ debería parecerse a 1. Ello significa que la hipótesis nula H_0 debería rechazarse si $\bar{X}$ se aleja de 1, esto es, si $\left|\bar{X} - 1\right|$ es un valor grande. ¿Como de grande? Para responder a esta pregunta observemos que si H_0 es cierta se tiene que:
$$T\left(X_1, \ldots, X_5\right) = \frac{\bar{X} - 1}{s/\sqrt{5}} \approx t_4$$

3. Podemos usar ahora la tabla de la t de Student para encontrar el valor $t_{4,\alpha/2}$ tal que:
$$P\left(\left|\frac{\bar{X} - 1}{s/\sqrt{5}}\right| > t_{4,\alpha/2} \middle| \, H_0 \text{ cierta}\right) = \alpha$$
De esta forma, la región crítica es $R_C = \left(-\infty, -t_{4,\alpha/2}\right] \cup \left[t_{4,\alpha/2}, \infty\right)$.

4. El contraste consiste entonces en *rechazar* H_0 *si* $\frac{\bar{X}-1}{s/\sqrt{5}} \in R_C$ *y aceptar* H_0 *en caso contrario.* Con los datos de este ejemplo se obtiene $\frac{\bar{X}-1}{s/\sqrt{5}} = \frac{1{,}2-1}{0{,}4/\sqrt{5}} = 1{,}11$. Asimismo, si elegimos $\alpha = 0{,}05$ resulta $t_{4,0{,}025} = 2{,}776$. Como el valor 1.11 no está en la región de rechazo concluimos que puede aceptarse H_0.

Dicho de otra forma, *si* H_0 *fuera cierta*, sería muy improbable que $\left|\frac{\bar{X}-1}{s/\sqrt{5}}\right| > 2{,}776$; o de manera equivalente, lo mas probable sería que $\left|\frac{\bar{X}-1}{s/\sqrt{5}}\right| \leq 2{,}776$. Como el valor observado, 1.11, está dentro de lo que es muy probable observar cuando H_0 es cierta, concluimos que no existe evidencia suficiente para rechazar H_0.

Ejemplo 6.2. Supongamos ahora que las algas de nuestro ejemplo se desarrollan bien si $\mu \leq 1$, pero mueren si $\mu > 1$, siendo μ el pH medio del agua del tanque de cultivo. Si en 7 análisis de agua hemos obtenido un pH medio de 1.1, con desviación típica 0.3, ¿hay evidencia suficiente para rechazar H_0?

En este caso, el contraste que se plantea es de la forma:

$$\begin{cases} H_0: & \mu \leq 1 \\ H_1: & \mu > 1 \end{cases}$$

Obviamente, aún siendo cierta H_0 podría ocurrir por azar que la media muestral $\bar{X}$ fuese *algo mayor* que 1, pero no *mucho mayor.* Por tanto la hipótesis nula H_0 debería rechazarse si el valor de $\bar{X} - 1$ es más grande de lo que cabría esperar por azar cuando $\mu \leq 1$. Para determinar como de grande debe ser $\bar{X} - 1$ para rechazar H_0 podemos utilizar como estadístico de contraste:

$$T(X_1, \ldots, X_7) = \frac{\bar{X} - 1}{S/\sqrt{7}}$$

Cuando H_0 es cierta, el valor de μ para el que cabría esperar valores más altos de $\bar{X}$ por azar es $\mu = 1$, en cuyo caso el estadístico $T(X_1, \ldots, X_7)$ sigue una distribución t de Student con 6 grados de libertad. Por tanto tenemos que:

$$P\left(\frac{\bar{X} - 1}{S/\sqrt{7}} > t_{6,\alpha} \middle| \mu = 1\right) = \alpha$$

Además, si $\mu < 1$ esta probabilidad será más pequeña y por tanto:

$$P\left(\frac{\bar{X} - 1}{S/\sqrt{7}} > t_{6,\alpha} \middle| H_0 \text{ cierta}\right) = P\left(\frac{\bar{X} - 1}{S/\sqrt{7}} > t_{6,\alpha} \middle| \mu \leq 1\right) \leq \alpha$$

De esta forma, si H_0 es cierta, es muy difícil que $T(X_1, \ldots, X_7)$ sea mayor que $t_{6,\alpha}$, por lo que la región crítica o de rechazo para este test es $R_C = [t_{6,\alpha}, \infty)$. Si $T(X_1, \ldots, X_7)$ cayera en este intervalo estaría ocurriendo algo muy difícil de ser H_0 cierta, por lo que H_0 debe rechazarse.

Con los datos aportados en el ejemplo se obtiene $\frac{\bar{X}-1}{S/\sqrt{7}} = \frac{1{,}1-1}{0{,}3/\sqrt{7}} = 0{,}882$. Asimismo, si elegimos $\alpha = 0{,}05$ resulta $t_{6,0{,}05} = 1{,}943$ y la región crítica es $R_C = [1{,}943, \infty)$. Como el valor 0.882 no está en esta región concluimos que puede aceptarse H_0.

Nota: Los contrastes de la forma $\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}$ reciben el nombre de *contrastes bilaterales o de dos colas* (su región crítica es bilateral). Los contrastes de la forma $\begin{cases} H_0: & \theta \geq \theta_0 \\ H_1: & \theta < \theta_0 \end{cases}$ ó $\begin{cases} H_0: & \theta \leq \theta_0 \\ H_1: & \theta > \theta_0 \end{cases}$ se denominan *contrastes unilaterales o de una cola.*

4.1. P-valor de un contraste.

Tal como hemos visto, en la construcción del contraste de hipótesis juega un papel importante el *nivel de significación* α. Este valor representa la probabilidad *que consideramos aceptable* de cometer un error tipo I: rechazar la hipótesis nula cuando es cierta. En este sentido, el valor de α es arbitrario. En el ámbito científico es habitual utilizar los valores 0.05, 0.01 e incluso 0.001. Pero cualquier otro valor podría ser igualmente válido (en la práctica hay consenso en que, en cualquier caso, α nunca debe ser mayor que 0.1).

Obviamente, cuanto más pequeño sea el valor de α, más difícil es rechazar H_0 cuando es cierta. Una vez obtenida una muestra $X_1, X_2, \ldots X_n$, se define el *p-valor* del contraste como el valor mínimo de α para el cual es posible rechazar H_0 con esos datos. Así, por ejemplo:

- Si con los datos disponibles, el valor más pequeño de α que permite el rechazo de H_0 es 0.4, ello querría decir que sólo sería posible rechazar H_0 si estuviéramos dispuestos a aceptar una probabilidad del 40 % de rechazarla siendo cierta (lo que obviamente no resultaría razonable).
- Si con los datos disponibles, el valor mínimo de α que conduce al rechazo de H_0 es 0.02, ello significa que sería posible rechazar esta hipótesis incluso si exigimos un riesgo del 2 % de rechazarla siendo cierta; pero no podríamos rechazarla si el riesgo asumible fuese del 1 %.

De esta forma, una vez obtenida la muestra, podríamos basar nuestra decisión en la siguiente regla basada en el p-valor:

Si $p - valor \geq \alpha$ aceptar H_0. Si $p - valor < \alpha$ rechazar H_0

Ejemplo 6.3. La región crítica para el rechazo de H_0 en el ejemplo 6.1 era de la forma $R_C = \left(-\infty, -t_{4,\alpha/2}\right] \cup \left[t_{4,\alpha/2}, \infty\right)$. Con los datos del ejemplo, el valor del estadístico de contraste fue $\frac{\bar{X}-1}{s/\sqrt{5}} = 1{,}11$. El valor más pequeño de α que permitiría entonces el rechazo de H_0 sería el que produjese $t_{4,\alpha/2} = 1{,}11$ (para que la región de rechazo contenga al valor del estadístico de contraste). Para hallar este valor de α basta tener en cuenta que, por definición:

$$P\left(t_4 \geq t_{4,\alpha/2}\right) = \frac{\alpha}{2}$$

Por tanto

$$P\left(t_4 \geq 1{,}11\right) = \frac{\alpha}{2}$$

La tabla de la t de Student no permite calcular esta probabilidad de forma sencilla, pero podemos calcularla con R :

$$P\left(t_4 \geq 1{,}11\right) = 1 - P\left(t_4 < 1{,}11\right) = \texttt{1-pt(1.11,4)} = 0{,}1646$$

Así pues:

$$\frac{\alpha}{2} = 0{,}1646$$

de donde:

$$\alpha = 0{,}3292$$

De esta forma, para los datos del ejemplo, el *p-valor* (valor mínimo de α que conduce al rechazo de H_0) es 0.3292. Siguiendo la regla del p-valor, sólo rechazaríamos H_0 si estuviésemos dispuestos a asumir una probabilidad 0.3292 de rechazar dicha hipótesis siendo cierta. Como no es el caso (habíamos elegido $\alpha = 0{,}05$), aceptamos H_0.

5. Potencia de un contraste.

Tal como hemos señalado, cuando se realiza un contraste de significación, la regla de decisión se establece de tal forma que el riesgo de cometer un error tipo I –rechazar la hipótesis nula cuando es cierta– es como mucho α, el nivel de significación del test. De esta forma, si se rechaza la hipótesis nula, sabemos *a priori* que existe muy poco riesgo de equivocarnos. Pero ¿qué ocurre si se acepta la hipótesis nula? ¿cuál es el riesgo de aceptar una hipótesis nula falsa? La probabilidad de cometer este error (error tipo II) es la que hemos denotado como β. Su valor complementario $1 - \beta$ recibe el nombre de *potencia del contraste* y representa la probabilidad de

rechazar H_0 cuando es falsa. Tal como hemos definido los contrastes de significación:

$$1-\beta = P\left(T\left(X_1, X_2, \ldots, X_n\right) \in R_C \,|H_o \text{ es falsa}\right)$$

Ejemplo 6.4. Con los datos del ejemplo 6.1 en el contraste:

$$\begin{cases} H_0 : \mu = 1 \\ H_1 : \mu \neq 1 \end{cases}$$

hemos aceptado la hipótesis nula ($\mu = 1$) aún cuando la media muestral era 1.2. ¿Cuál es la probabilidad de que estemos cometiendo un error de tipo II en este contraste? Para responder a esta pregunta observemos que esta probabilidad es:

$$\begin{aligned} P\left(\text{Error Tipo II}\right) &= P\left(\text{Aceptar } H_0 \,|H_0 \text{ es falsa}\right) = \\ &= P\left(T\left(X_1, X_2, \ldots, X_n\right) \notin R_C \,|H_o \text{ es falsa}\right) = \\ &= P\left(\left|\frac{\bar{X}-1}{s/\sqrt{5}}\right| \leq t_{4,\alpha/2} \Big/ \mu \neq 1\right) = P\left(-t_{4,\alpha/2} \leq \frac{\bar{X}-1}{s/\sqrt{5}} \leq t_{4,\alpha/2} \Big/ \mu \neq 1\right) \end{aligned}$$

Para calcular esta probabilidad hemos de tener en cuenta que realizamos el contraste bajo el supuesto de que la variable X que se mide (en este caso el pH) es $N\left(\mu, \sigma\right)$, por lo que el estadístico

$$\frac{\bar{X}-\mu}{s/\sqrt{5}}$$

sigue una distribución t de Student con 4 grados de libertad. Cuando H_0 es falsa se tiene que $\mu \neq 1$ y por tanto:

$$\begin{aligned} \beta\left(\mu\right) =& P\left(-t_{4,\alpha/2} \leq \frac{\bar{X}-1}{s/\sqrt{5}} \leq t_{4,\alpha/2} \Big/ \mu \neq 1\right) = \\ & P\left(-t_{4,\alpha/2} \leq \frac{\bar{X}-\mu+\mu-1}{s/\sqrt{5}} \leq t_{4,\alpha/2} \Big/ \mu \neq 1\right) = \\ & P\left(-t_{4,\alpha/2} \leq \frac{\bar{X}-\mu}{s/\sqrt{5}} + \frac{\mu-1}{s/\sqrt{5}} \leq t_{4,\alpha/2} \Big/ \mu \neq 1\right) = \\ & P\left(-t_{4,\alpha/2} - \frac{\mu-1}{s/\sqrt{5}} \leq \frac{\bar{X}-\mu}{s/\sqrt{5}} \leq t_{4,\alpha/2} - \frac{\mu-1}{s/\sqrt{5}} \Big/ \mu \neq 1\right) = \\ & P\left(-t_{4,\alpha/2} - \frac{\mu-1}{s/\sqrt{5}} \leq t_4 \leq t_{4,\alpha/2} - \frac{\mu-1}{s/\sqrt{5}} \Big/ \mu \neq 1\right) \end{aligned}$$

Así pues, la probabilidad de error tipo II corresponde, geométricamente, al área bajo la función de densidad de una t_4 entre los valores $-t_{4,\alpha/2} - \frac{\mu-1}{s/\sqrt{5}}$ y $t_{4,\alpha/2} - \frac{\mu-1}{s/\sqrt{5}}$. La figura 1 muestra

gráficamente esta área para diversos valores de μ.

Tal como puede apreciarse en esta figura, a medida que el valor de μ se aleja de 1, el término $\frac{\mu-1}{s/\sqrt{5}}$ se hace mayor en valor absoluto, por lo que el intervalo $\left[-t_{4,\alpha/2}-\frac{\mu-1}{s/\sqrt{5}},\, t_{4,\alpha/2}-\frac{\mu-1}{s/\sqrt{5}}\right]$ se va desplazando (hacia la izquierda si $\mu > 1$, o hacia la derecha si $\mu < 1$). Como consecuencia de este desplazamiento, el área que comprende la función de densidad sobre este intervalo –esto es, el valor de la probabilidad de error II, β– se va haciendo cada vez menor. La interpretación de este comportamiento de β es bastante intuitiva: en nuestro contraste estamos tratando de decidir si la verdadera media de la población es 1; será más fácil equivocarse aceptando que es 1 cuando realmente es 0.9 ó 1.1 (el verdadero valor μ está cerca de 1) que cuando la verdadera media es un valor más alejado de 1, como el 0.2 ó el 1.8.

Podemos también calcular numéricamente los valores de β para diversos valores alternativos de μ. Para el contraste del ejemplo 6.1 habíamos elegido $\alpha = 0{,}05$, resultando $t_{4,0{,}025} = 2{,}776$; asimismo, teníamos que $s = 0{,}4$. Por tanto, la probabilidad de error tipo II en este caso es, dependiendo del valor de μ:

$$\begin{aligned}\beta(\mu) =& P\left(-2{,}776-\frac{\mu-1}{0{,}4/\sqrt{5}} \leq t_4 \leq 2{,}776-\frac{\mu-1}{0{,}4/\sqrt{5}}\right) = \\ & P\left(t_4 \leq 2{,}776-\frac{\mu-1}{0{,}4/\sqrt{5}}\right) - P\left(t_4 \leq -2{,}776-\frac{\mu-1}{0{,}4/\sqrt{5}}\right)\end{aligned}$$

La tabla de la t de Student no se presta a calcular estas probabilidades, pero podemos utilizar R :

$\beta(\mu)$=`pt(2.776-(mu-1)/(0.4/sqrt(5)),4)-pt(2.776-(mu-1)/(0.4/sqrt(5)),4)`

La siguiente tabla muestra los valores de la probabilidad de error tipo II, así como la potencia que se alcanza para diversos valores de μ:

μ	$\beta(\mu)$	Potencia = $1-\beta(\mu)$
0	0.0235	0.9765
0.2	0.0816	0.9184
0.4	0.2953	0.7047
0.6	0.6873	0.3127
0.8	0.9049	0.0951
1	0.95	0.05
1.2	0.9049	0.0951
1.4	0.6873	0.3127
1.6	0.2953	0.7047
1.8	0.0816	0.9184
2.0	0.0235	0.9765

Asimismo, la figura 2 representa gráficamente estos valores, mostrando las funciones de error tipo II y potencia para este contraste. En esta figura vemos nuevamente que la probabilidad de error tipo II, $\beta(\mu)$, es tanto mayor cuanto más próximo esté μ a 1, alcanzando su máximo cuando μ coincide con el valor especificado en la hipótesis nula ($\mu = 1$). El comportamiento de la función de potencia –probabilidad de rechazar H_0 cuando es falsa– es, como cabe esperar, justo en inverso: si el verdadero valor de μ está cerca de 1, el contraste apenas tiene potencia para distinguir ambos valores; cuánto más lejos esté μ de 1, mayor es la potencia del contraste.

6. Tamaño de muestra para una significación y potencia preespecificadas.

El contraste de hipótesis que hemos planteado en el 6.1 es un caso particular de contraste de la forma:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

en el que la hipótesis nula que se pone a prueba es si puede aceptarse que el valor de la esperanza μ de una variable X con distribución normal es μ_0. Si se dispone de una muestra aleatoria de n observaciones de esta variable, siendo $\bar{X}$ su media y S su desviación típica, la regla de decisión para este contraste, fijado un nivel de significación α es, generalizando el procedimiento que hemos visto en el ejemplo 6.1:

Rechazar H_0 *si* $\left|\frac{\bar{X}-\mu_0}{s/\sqrt{n}}\right| > t_{n-1,\alpha/2}$ *y aceptar* H_0 *en caso contrario.*

Asimismo, generalizando el resultado obtenido en el ejemplo 6.4, la probabilidad de error tipo II para este contraste viene dada por:

$$\beta(\mu) = P\left(-t_{n-1,\alpha/2} - \frac{\mu-\mu_0}{s/\sqrt{n}} \leq t_{n-1} \leq t_{n-1,\alpha/2} - \frac{\mu-\mu_0}{s/\sqrt{n}} \middle/ \mu \neq \mu_0\right) \tag{6.1}$$

que, como ya hemos visto, representa el área comprendida por la densidad t de Student con $n-1$ grados de libertad sobre el intervalo $\left[-t_{n-1,\alpha/2} - \frac{\mu-\mu_0}{s/\sqrt{n}},\ t_{n-1,\alpha/2} - \frac{\mu-\mu_0}{s/\sqrt{n}}\right]$ (ver figura 1). Obsérvese que este intervalo puede expresarse también de la forma:

$$\left[-t_{n-1,\alpha/2} - \frac{(\mu-\mu_0)\sqrt{n}}{s},\ t_{n-1,\alpha/2} - \frac{(\mu-\mu_0)\sqrt{n}}{s}\right]$$

lo que hace evidente el hecho de que aún cuando $\frac{(\mu-\mu_0)}{s}$ tomase un valor pequeño, eligiendo un valor adecuado de n (tamaño de la muestra) podemos hacer el término $\frac{(\mu-\mu_0)\sqrt{n}}{s}$ todo lo grande que queramos. Ello significa que, tal como vimos en nuestro análisis de la figura 1, podemos desplazar el intervalo anterior (hacia la izquierda o la derecha, según el signo de $\mu-\mu_0$) hasta que el área comprendida sobre el mismo –esto es, la probabilidad de error II– sea tan pequeña como se quiera.

Esto nos permite responder a la cuestión siguiente: *¿cuál debe ser el tamaño n de la muestra si se desea que cuando $\mu = \mu_0 + \Delta$ la probabilidad de error tipo II en el contraste anterior sea un valor prefijado β –o, de modo equivalente, que la potencia sea 1-β–, manteniendo al mismo tiempo un nivel de significación preespecificado α?*

Para ello, utilizando la ecuación 6.1, y teniendo en cuenta que $\mu - \mu_0 = \Delta$, debemos encontrar el valor de n tal que:

$$\begin{aligned}
\beta =& P\left(-t_{n-1,\alpha/2} - \frac{\Delta\sqrt{n}}{s} \leq t_{n-1} \leq t_{n-1,\alpha/2} - \frac{\Delta\sqrt{n}}{s}\right) = \\
=& P\left(t_{n-1} > -t_{n-1,\alpha/2} - \frac{\Delta\sqrt{n}}{s}\right) - P\left(t_{n-1} > t_{n-1,\alpha/2} - \frac{\Delta\sqrt{n}}{s}\right) \cong \\
\cong& P\left(Z > -z_{\alpha/2} - \frac{\Delta\sqrt{n}}{s}\right) - P\left(Z > z_{\alpha/2} - \frac{\Delta\sqrt{n}}{s}\right) \cong \\
\cong& 1 - P\left(Z > z_{\alpha/2} - \frac{\Delta\sqrt{n}}{s}\right) \Rightarrow P\left(Z > z_{\alpha/2} - \frac{\Delta\sqrt{n}}{s}\right) = 1-\beta
\end{aligned}$$

(aquí hemos hecho dos aproximaciones; en primer lugar hemos supuesto que n va a resultar tan grande que la distribución t_n puede aproximarse por la normal estándar Z; y en segundo lugar hemos supuesto que el valor $-z_{\alpha/2} - \frac{\Delta\sqrt{n}}{S}$ es tan grande en valor absoluto que el área a su derecha es prácticamente uno). Utilizando la notación habitual z_β para el percentil de la normal

estándar tal que $P\left(Z > z_{1-\beta}\right) = 1 - \beta$ tenemos que:

$$z_{\alpha/2} - \frac{\Delta\sqrt{n}}{s} = z_{1-\beta} = -z_{\beta}$$

de donde, despejando n, resulta:

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 s^2}{\Delta^2}$$

Obsérvese que el valor de n:

- Es proporcional al cuadrado de la suma $z_{\alpha/2} + z_{\beta}$. Como estos valores son más grandes a medida que α y β son más pequeños, el tamaño de la muestra se incrementa cuando se desea que las probabilidades de los errores I y II disminuyan.
- Es proporcional a la varianza s^2, por lo que cuanto mayor sea la variabilidad en la variable que se mide mayor habrá de ser el tamaño de la muestra. Es intuitivamente claro que debe ser así. Si los valores de X fuesen muy homogéneos (poca variabilidad), una muestra pequeña podría representar bien el comportamiento de la variable; a medida que los valores de X son más heterogéneos será precisa más información –más datos– para representarla.
- Es inversamente proporcional al cuadrado de la diferencia Δ que se pretende detectar entre el verdadero valor medio μ y el valor μ_0 que se pone a prueba. Ello significa que cuanto menor sea la diferencia que se pretende detectar, mayor habrá de ser el tamaño de muestra.

El valor de s^2 no se conoce habitualmente antes de realizar el muestreo, por lo que para planificar el tamaño adecuado de muestra, habrá que utilizar un valor de s^2 obtenido en una muestra piloto o publicado en la literatura en estudios similares.

Señalemos por último que en esta sección hemos desarrollado el cálculo del tamaño de la muestra sólo para contrastar si el valor esperado μ de una variable es igual a un valor preespecificado μ_0. No obstante, el mismo patrón de ideas se aplica para el cálculo del tamaño muestral en otros contrastes de hipótesis, con las lógicas modificaciones derivadas del tipo de datos y de la forma de la regla de decisión. Asimismo, las observaciones que se acaban de realizar sobre la relación del tamaño de muestra con las magnitudes de α, β, Δ y la variabilidad resultan de aplicación general en todos los contrastes de hipótesis.

Ejemplo 6.5. Volviendo al ejemplo 6.1, recordemos que el crecimiento de las algas allí descritas requiere que el pH medio del agua sea 1. Supongamos además que las algas tienen cierta tolerancia a variaciones en el pH y que su desarrollo en cualquier caso es óptimo si el pH medio

se mantiene entre 0.8 y 1.2. Se desea planificar el número de muestras de agua diarias que deben tomarse si se desea realizar el contraste

$$\begin{cases} H_0 : \mu = 1 \\ H_1 : \mu \neq 1 \end{cases}$$

con un nivel de significación 0.05, y garantizando una potencia 0.9 de que se rechazará H_0 si μ cae por debajo de 0.8 o por encima de 1.2.

Usando la información aportada por la muestra del ejemplo 6.1, usaremos como estimador piloto de la varianza el valor $s^2 = 0{,}4^2 = 0{,}16$. La diferencia mínima que interesa detectar en este caso es $\Delta = 0{,}2$, ya que se nos dice que las algas muestran tolerancia con valores de pH que difieran de 1 en 0.2 unidades (entre 0.8 y 1.2). Dado que se desea detectar esta diferencia con potencia $1 - \beta = 0{,}9$, se tiene $\beta = 0{,}1$ y $z_\beta = z_{0{,}1} = 1{,}28$. Para el nivel de significación $\alpha = 0{,}05$ se tiene $z_{\alpha/2} = 1{,}96$, y por tanto:

$$n = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 s^2}{\Delta^2} = \frac{(1{,}96 + 1{,}28)^2 \cdot 0{,}16}{0{,}2^2} \cong 42$$

7. Significación estadística y relevancia práctica.

Ya hemos señalado más arriba que cuando en un contraste se rechaza la hipótesis nula, tal resultado se suele expresar diciendo que *el contraste ha resultado significativo.* Es necesario tener aquí cierta precaución con la terminología, ya que la palabra *"significativo"* en este contexto suele ser mal interpretada. La definición que proporciona el diccionario del adjetivo "*significativo" es* "*que tiene importancia por representar o significar algo".* Por ello, el hablante habitual cuando emplea esta palabra la entiende normalmente como referida a algo *importante*. Sin embargo, en el contexto de un contraste de hipótesis estadístico, el que un resultado haya sido significativo indica simplemente *que dicho resultado no puede explicarse como efecto del azar.* Que sea importante o no, es algo que habrá de ser valorado en función de las implicaciones prácticas que pueda tener dicho resultado.

Así, en el ejemplo 6.5 hemos visto que con una muestra de 42 observaciones del pH del agua hay una probabilidad del 90 % de detectar si el pH medio difiere en más de 0.2 unidades del valor medio deseado $\mu = 1$. El lector puede utilizar la misma fórmula para comprobar que, con la misma potencia, si la muestra fuese de tamaño 672 se podría detectar una diferencia de 0.05 unidades,

y con una muestra de 1867 observaciones se podría detectar una diferencia de 0.03 unidades. Ello significa que, si se hacen los correspondientes contrastes de hipótesis con esos tamaños muestrales, las diferencias citadas, en caso de encontrarse, serían declaradas "*significativas*". Pero desde luego no serían *importantes*: si las algas se desarrollan bien cuando el pH medio se aparta hasta 0.2 unidades de 1, ¿qué importancia tendría haber encontrado que el pH medio es *significativamente* distinto de 1 porque se aparta de ese valor en 0.03 unidades?

Así pues, en general con una muestra lo suficientemente grande cualquier diferencia puede resultar estadísticamente significativa, por muy irrelevante que su valor resulte en la práctica. Obviamente también es cierto lo contrario: si la muestra es demasiado pequeña, diferencias importantes pueden resultar no significativas (recuérdese: aceptar la hipótesis nula no significa que sea cierta). Es responsabilidad del investigador, por tanto, fijar la diferencia mínima Δ que se considera relevante o importante y determinar el tamaño de muestra para que se pueda detectar dicha diferencia con una significación y potencia adecuados. Sólo en estas condiciones podrá ser el resultado de un contraste significativo y relevante a la vez.

8. Relación entre intervalos de confianza y contrastes de hipótesis.

En el capítulo anterior hemos estudiado la construcción de intervalos de confianza para los parámetros de ciertas distribuciones de probabilidad. Recordemos que $[\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]$, donde $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$ son variables aleatorias que dependen de una muestra $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$, es un *intervalo de confianza a nivel* $1-\alpha$ *para el parámetro* θ si la probabilidad de que el intervalo contenga a dicho parámetro es $1-\alpha$, esto es:

$$P(\theta \in [\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]) = 1-\alpha$$

Entonces, si se dispone de un intervalo de confianza para θ, para resolver el contraste de hipótesis:

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}$$

puede utilizarse como regla de decisión:

Si $\theta_0 \in [\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]$ *aceptar* H_0*; en caso contrario, rechazar* H_0.

En efecto, la probabilidad de error tipo I cuando se utiliza esta regla es:

$$P\left(\text{error I}\right) = P\left(\text{rechazar } H_0 \left| H_0 \text{ cierta}\right.\right) = \left(\theta_0 \notin \left[\theta_1\left(\mathfrak{X}\right), \theta_2\left(\mathfrak{X}\right)\right] \left| \theta = \theta_0\right.\right) =$$
$$= P\left(\theta \notin \left[\theta_1\left(\mathfrak{X}\right), \theta_2\left(\mathfrak{X}\right)\right]\right) = \alpha$$

Ejemplo 6.6. En el ejemplo 6.1 debíamos decidir, a partir de 5 muestras de pH de un tanque de agua, si podía aceptarse que el pH medio era 1. Para ello planteábamos el contraste:

$$\begin{cases} H_0: & \mu = 1 \\ H_1: & \mu \neq 1 \end{cases}$$

partiendo del supuesto adicional de que el pH sigue una distribución normal. El intervalo de confianza para la media μ de una distribución normal con varianza σ^2 desconocida es, tal como vimos en el capítulo anterior:

$$\left[\bar{X} - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}\right]$$

Por tanto, podríamos utilizar como regla de decisión para el contraste:

Si $1 \in \left[\bar{X} - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}\right]$*, aceptar* H_0 *y en caso contrario rechazar* H_0*.*

Es fácil comprobar que:

$$1 \in \left[\bar{X} - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}\right] \Leftrightarrow \bar{X} - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2} \leq 1 \leq \bar{X} + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}$$
$$\bar{X} - 1 - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2} \leq 0 \leq \bar{X} - 1 + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2} \Leftrightarrow -\frac{s}{\sqrt{n}} t_{n-1,\alpha/2} \leq \bar{X} - 1 \leq \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}$$
$$-t_{n-1,\alpha/2} \leq \frac{\bar{X} - 1}{s/\sqrt{n}} \leq t_{n-1,\alpha/2} \Leftrightarrow \left|\frac{\bar{X} - 1}{s/\sqrt{n}}\right| \leq t_{n-1,\alpha/2}$$

Por tanto la regla de decisión basada en el intervalo de confianza *es exactamente la misma* que ya habíamos obtenido en el ejemplo 6.1 por otro procedimiento.

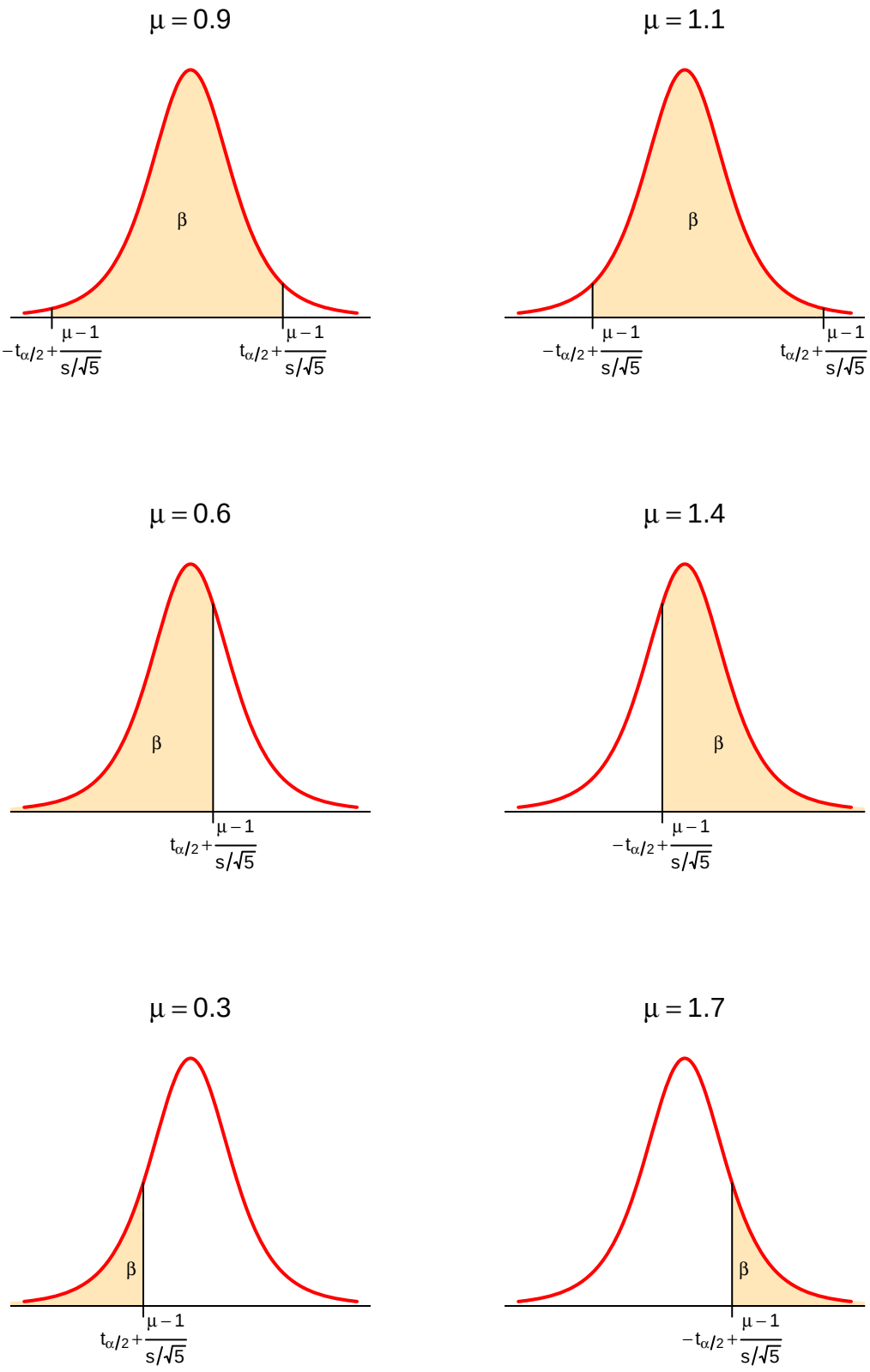

Figura 1: Probabilidad de error tipo II para diversos valores de μ en el contraste de hipótesis del ejemplo 6.1.

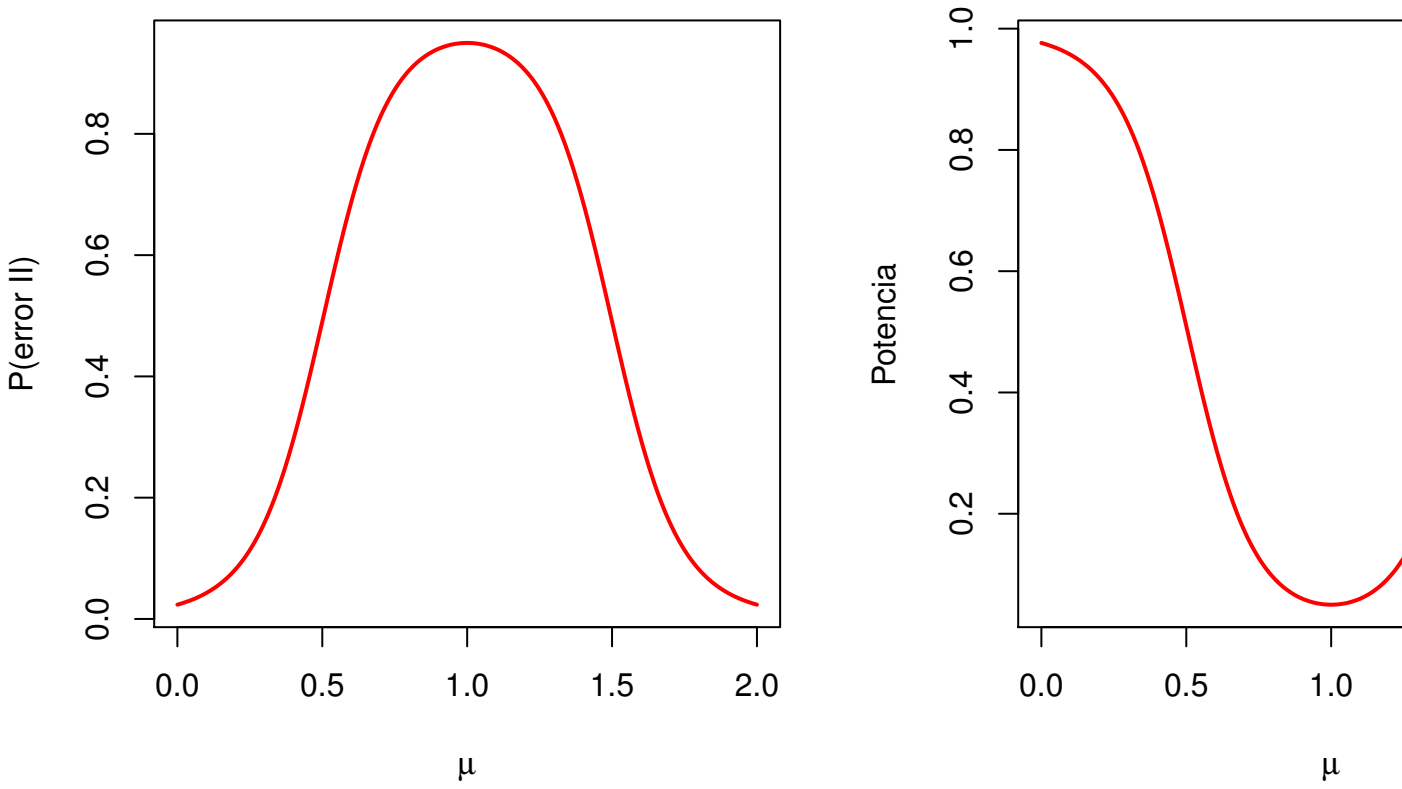

Figura 2: Funciones de error tipo II (izquierda) y potencia (derecha) para el contraste de hipótesis del ejemplo 6.1