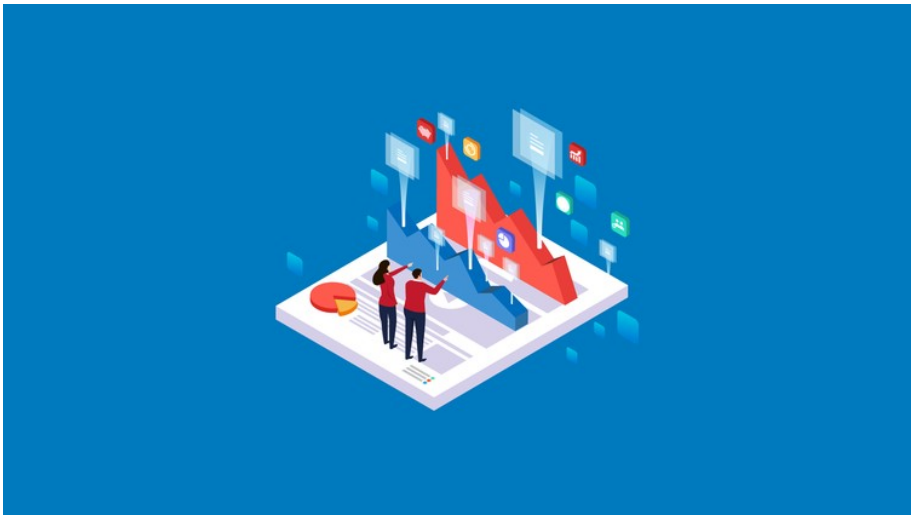
Curso completo de Estadística Inferencial con R y Python

Ricardo Alberich, Juan Gabriel Gomila y Arnau Mir

2021-10-14

Estadística inferencial para Machine Learning con R y Python



R.Alberich, J.G.Gomila y Arnau Mir

Curso online completo
Disponible en Udemy

`https://tinyurl.com/unewvhr`

Índice general

Consulta el curso completo de estadística creado por Ricardo Alberich, Juan Gabriel Gomila y Arnau Mir solamente en Udemy

Asienta las bases para convertirte en el Data Scientist del futuro con todo el contenido de estadística inferencial del curso. En particular verás los mismos contenidos que explicamos en primero de carrera a matemáticos, ingenieros, economistas, biólogos, médicos o informáticos.

1. Muestreo estadístico
2. Estimación puntual
3. Intervalos de confianza
4. Contrastes de hipótesis
5. Bondad de ajuste
6. Bondad de independencia y homogeneidad
7. Análisis de la varianza
8. Regresión lineal
9. Clustering

Y todo con más de 40 horas de vídeo a demanda, cientos de ejercicios, tareas, talleres y trucos de los profesores para que te conviertas en un experto de la materia.

```
## Warning: package 'shiny' was built under R version 3.6.2
```

```
## Warning: package 'reticulate' was built under R version 3.6.2
```

Capítulo 1

Muestreo estadístico

- En todo estudio estadístico distinguiremos entre **población**, (conjunto de sujetos con una o varias características que podemos medir y deseamos estudiar), y **muestra**, (subconjunto de una población.)

- Dos tipos de análisis estadístico:

  - **Exploratorio** o **descriptivo**: **estadística descriptiva**.
  - **Inferencial** o **confirmatorio**: **estadística inferencial**.

Pasos en un estudio inferencial:

- Establecer la característica que se desea estimar o la hipótesis que se desea contrastar.

- Determinar la información (los datos) que se necesita para hacerlo.

- Diseñar un experimento que permita recoger estos datos; este paso incluye:

  - Decidir qué tipo de muestra se va a tomar y su tamaño.

  - Elegir las técnicas adecuadas para realizar las inferencias deseadas a partir de la muestra que se tomará.

- Tomar una muestra y medir los datos deseados sobre los individuos que la forman.

- Aplicar las técnicas de inferencia elegidas con el *software* adecuado.

- Obtener conclusiones.

- Si las conclusiones son fiables y suficientes, redactar un informe; en caso contrario, volver a empezar.

1.1. Tipos de muestreo

1.1.1. Muestreo aleatorio con reposición

Muestreo aleatorio: consiste en seleccionar una muestra de la población de manera que todas las muestras del mismo tamaño sean **equiprobables**.

Consideremos una urna de 100 bolas numeradas del 1 al 100:

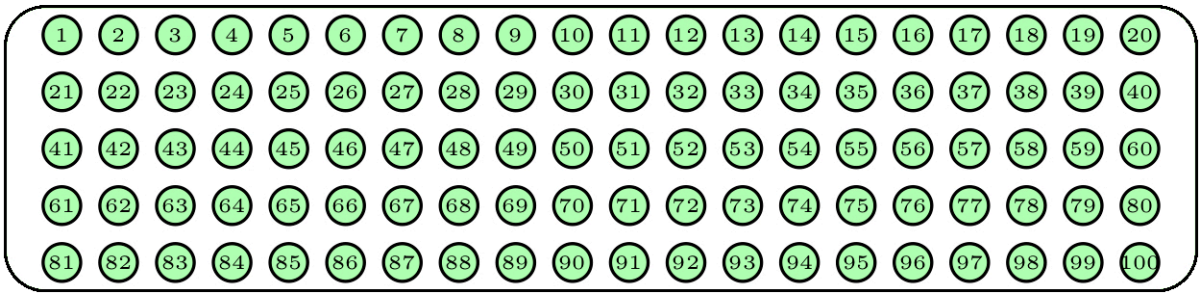


Figura 1.1: Una urna de 100 bolas

Queremos extraer una muestra de 15 bolas. Para ello, podríamos repetir 15 veces el proceso de sacar una bola de la urna, anotar su número y devolverla a la urna. El tipo de muestra obtenida de esta manera recibe el nombre de **muestra aleatoria con reposición**, o simple (una **m.a.s.**, para abreviar).

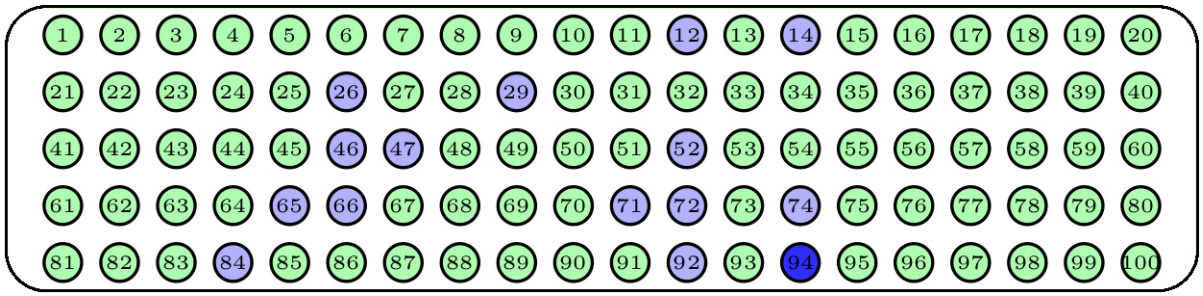


Figura 1.2: Una muestra aleatoria simple

Las bolas violetas son las escogidas para la muestra. La bola azul se ha escogido dos veces al ser el muestreo con reposición.

Para simular un muestreo de 15 bolas con reposición en una urna de 100 en `R`, haríamos los siguiente:

```
sample(1:100, 15, replace=TRUE)
```

```
##  [1] 19 13 31  3 20 93 92 72 85 43 75 30 54 51 58
```

Fijaos que no hemos obtenido la misma muestra. Esto es debido a que no hemos fijado la **semilla de aleatoriedad**.

Ejemplo `iris`

Veamos un ejemplo más elaborado. Consideremos la tabla de datos *iris* que contiene 150 flores de 3 especies diferentes: **setosa**, **versicolor** y **virginica**. La tabla de datos contiene 5 variables: la longitud y amplitud del pétalo, la longitud y la amplitud del sépalo y la especie de la flor.

Las primeras filas de la tabla de datos son:

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Si quisiéramos una muestra de 10 flores con reposición, haríamos lo siguiente:

La función set.seed fija la semilla de aleatoriedad sirve para que siempre dé la misma muestra. A continuación, elegimos las flores de la muestra:

```
set.seed(4)
flores.elegidas.10.con=sample(1:150,10,replace=TRUE)
```

Seguidamente, calculamos la subtabla de las flores de la muestra

```
muestra.iris.10.con = iris[flores.elegidas.10.con,]
```

Por último, mostramos la muestra de las flores:

```
muestra.iris.10.con
```

```
##      Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 75            6.4         2.9          4.3         1.3 versicolor
## 51            7.0         3.2          4.7         1.4 versicolor
## 3             4.7         3.2          1.3         0.2     setosa
## 71            5.9         3.2          4.8         1.8 versicolor
## 115           5.8         2.8          5.1         2.4  virginica
## 51.1          7.0         3.2          4.7         1.4 versicolor
## 56            5.7         2.8          4.5         1.3 versicolor
## 62            5.9         3.0          4.2         1.5 versicolor
## 102           5.8         2.7          5.1         1.9  virginica
## 130           7.2         3.0          5.8         1.6  virginica
```

1.1.2. Muestreo aleatorio sin reposición

Muestra aleatoria sin reposición: Otra manera de extraer nuestra muestra sería repetir 15 veces el proceso de sacar una bola de la urna pero ahora sin devolverla. En este caso se habla de una **muestra**

aleatoria sin reposición.

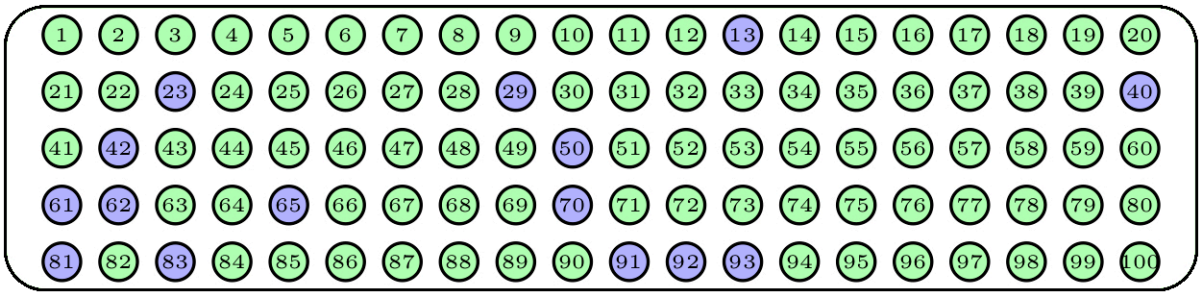


Figura 1.3: Una muestra aleatoria sin reposición

Para simular un muestreo de 15 bolas sin reposición en la urna anterior de 100 en R, haríamos los siguiente:

```
sample(1:100, 15, replace=FALSE)
```

```
##  [1] 24  1 84 35 27 48 95  2 32 47 44 69 15 22 89
```

Ejemplo `iris`

Consideremos de nuevo la tabla de datos `iris`.

Para obtener una muestra de 10 flores sin reposición, haríamos los pasos siguientes:

Primero elegimos las flores de la muestra

```
set.seed(4)
flores.elegidas.10.sin=sample(1:150,10,replace=FALSE)
```

A continuación, calculamos la subtabla de las flores de la muestra

```
muestra.iris.10.sin = iris[flores.elegidas.10.sin,]
```

Por último, mostramos las muestra de las flores:

```
muestra.iris.10.sin
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 75           6.4         2.9          4.3         1.3 versicolor
## 51           7.0         3.2          4.7         1.4 versicolor
## 3            4.7         3.2          1.3         0.2     setosa
## 71           5.9         3.2          4.8         1.8 versicolor
## 115          5.8         2.8          5.1         2.4  virginica
## 149          6.2         3.4          5.4         2.3  virginica
## 56           5.7         2.8          4.5         1.3 versicolor
## 62           5.9         3.0          4.2         1.5 versicolor
## 102          5.8         2.7          5.1         1.9  virginica
## 130          7.2         3.0          5.8         1.6  virginica
```

1.1.3. Muestras aleatorias con reposición vs. sin reposición

Observación: ¿Cuándo se puede considerar equivalente válido realizar una muestra con reposición que sin reposición?

Si el tamaño de la población es muy grande en relación al de la muestra (por dar una regla, digamos que, al menos, unas 1000 veces mayor).

1.1.4. Muestreo sistemático

Muestreo sistemático: Supongamos que los individuos de una población vienen dados en forma de una lista ordenada. El **muestreo sistemático** consiste en tomarlos a intervalos constantes escogiendo al azar el primer individuo que elegimos.

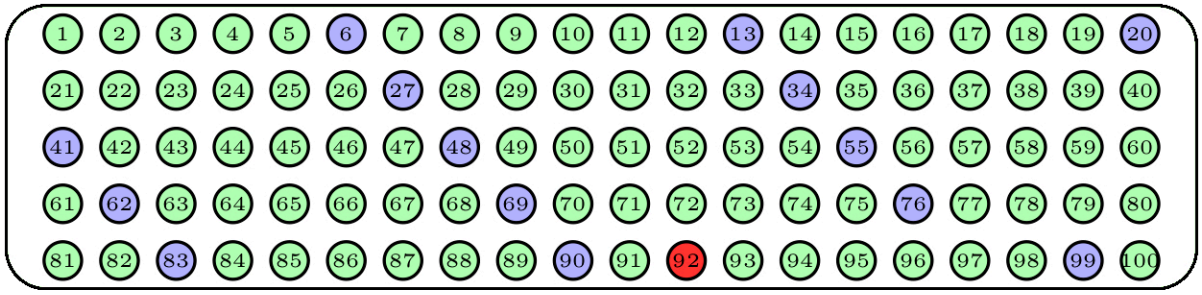


Figura 1.4: Una muestra aleatoria sistemática

La figura anterior describe una muestra aleatoria sistemática de 15 bolas de nuestra urna de 100 bolas: hemos empezado a escoger por la bola roja oscura, que ha sido elegida al azar, y a partir de ella hemos tomado 1 de cada 7 bolas, volviendo al principio cuando hemos llegado al final de la lista de bolas.

Ejemplo `iris`

Vamos a calcular una muestra aleatoria sistemática de la tabla de datos **iris** de tamaño 10.

Primero fijamos la **semilla de aleatoriedad** para la reproducibilidad del experimento:

```
set.seed(15)
```

Seguidamente, hallamos la etiqueta de la primera flor de la muestra (que será una de las 150 de la tabla de datos):

```
(primera.flor=sample(1:150,1))
```

```
## [1] 37
```

A continuación, hallamos el incremento que vamos a ir sumando a la primera etiqueta que hemos elegido:

```
incremento = floor(150/10)
```

el siguiente paso es elegir las flores de la muestra

```
flores.elegidas.10.sis = seq(from=primera.flor,by=incremento,length.out=10)
```

como las etiquetas elegidas no están entre 1 y 150, hemos de transformarlas:

```
flores.elegidas.10.sis = flores.elegidas.10.sis%%150
```

a continuación, calculamos la subtabla de las flores de la muestra

```
muestra.iris.10.sis = iris[flores.elegidas.10.sis,]
```

Y finalmente mostramos la subtabla de la muestra

```
muestra.iris.10.sis
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 37           5.5         3.5          1.3         0.2     setosa
## 52           6.4         3.2          4.5         1.5 versicolor
## 67           5.6         3.0          4.5         1.5 versicolor
## 82           5.5         2.4          3.7         1.0 versicolor
## 97           5.7         2.9          4.2         1.3 versicolor
## 112          6.4         2.7          5.3         1.9  virginica
## 127          6.2         2.8          4.8         1.8  virginica
## 142          6.9         3.1          5.1         2.3  virginica
## 7            4.6         3.4          1.4         0.3     setosa
## 22           5.1         3.7          1.5         0.4     setosa
```

1.1.5. Muestreo aleatorio estratificado

Muestreo aleatorio estratificado: Este tipo de muestreo se utiliza cuando la población está clasificada en **estratos** que son de interés para la propiedad estudiada. Se toma una muestra aleatoria de cada estrato y se unen en una muestra global. A este proceso se le llama **muestreo aleatorio estratificado**.

Supongamos que nuestra urna de 100 bolas contiene 40 bolas de un color y 60 de otro color tal como muestra la figura:

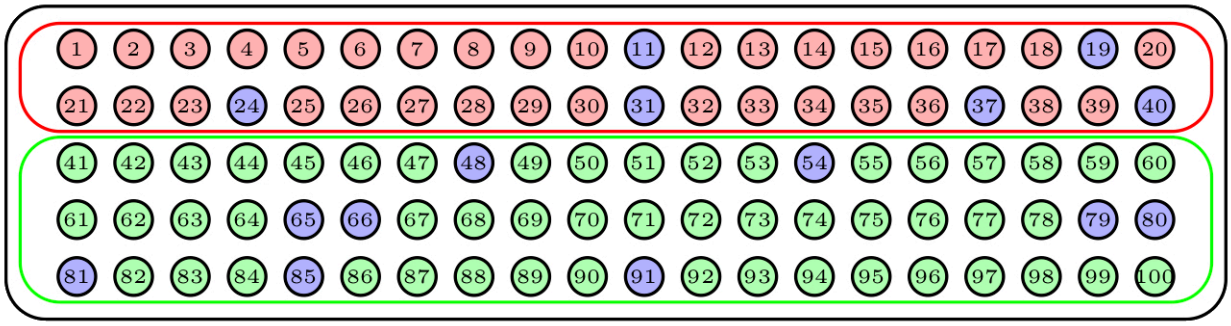


Figura 1.5: Una muestra aleatoria estratificada con dos estratos

Para tomar una muestra aleatoria estratificada de 15 bolas, considerando como estratos los dos colores,

tomaríamos una muestra aleatoria de 6 bolas del primer color y una muestra aleatoria de 9 bolas del segundo color.

Ejemplo `iris`

Vamos a considerar que la tabla de datos iris está estratificada según tres estratos. Cada estrato está compuesto por las 50 flores de la misma especie. Vamos a hallar una muestra de tamaño 12 hallando tres muestras de tamaño 4 de cada especie (estrato) con reposición y después juntaremos la tres submuestras.

En primer lugar, fijamos la semilla de aleatoriedad por reproducibilidad:

```
set.seed(25)
```

a continuación, hallamos las flores de la muestra de cada una de las especies:

```
fls.muestra.setosa=sample(1:50,4,replace=TRUE)
fls.muestra.versicolor=sample(51:100,4,replace=TRUE)
fls.muestra.virginica=sample(101:150,4,replace=TRUE)
```

seguidamente, calculamos y mostramos la muestra estratificada juntando las tres muestras de cada especie

```
(muestra.iris.est=rbind(iris[fls.muestra.setosa,],iris[fls.muestra.versicolor,],
                        iris[fls.muestra.virginica,]))
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 7            4.6         3.4          1.4         0.3     setosa
## 29           5.2         3.4          1.4         0.2     setosa
## 24           5.1         3.3          1.7         0.5     setosa
## 25           4.8         3.4          1.9         0.2     setosa
## 99           5.1         2.5          3.0         1.1 versicolor
## 58           4.9         2.4          3.3         1.0 versicolor
## 91           5.5         2.6          4.4         1.2 versicolor
## 76           6.6         3.0          4.4         1.4 versicolor
## 116          6.4         3.2          5.3         2.3  virginica
## 136          7.7         3.0          6.1         2.3  virginica
## 101          6.3         3.3          6.0         2.5  virginica
## 108          7.3         2.9          6.3         1.8  virginica
```

1.1.6. Muestreo por conglomerados

El proceso de obtener y estudiar una muestra aleatoria en algunos casos es caro o difícil, incluso aunque dispongamos de la lista completa de la población.

Muestreo por conglomerados: una alternativa posible sería, en vez de extraer una muestra aleatoria de todos los individuos de la población, escoger primero al azar unos subconjuntos en los que la población está dividida, a las que llamamos en este contexto **conglomerados** (*clusters*).

Supongamos que las 100 bolas de nuestra urna se agrupan en 20 conglomerados de 5 bolas cada uno según las franjas verticales.

Para obtener una muestra aleatoria por conglomerados de tamaño 15, escogeríamos al azar 3 conglomerados y la muestra estaría formada por sus bolas: los conglomerados escogidos están marcados en azul:

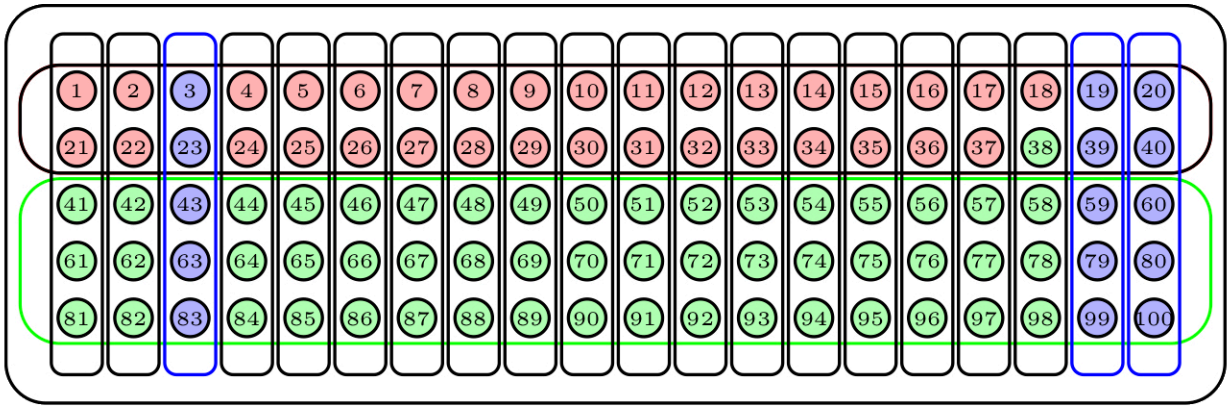


Figura 1.6: Una muestra aleatoria por conglomerados con 2 estratos y 20 conglomerados

Ejemplo `worldcup`

Consideremos la tabla de datos **worldcup** del paquete **faraway**. Esta tabla de datos nos da información sobre 595 jugadores que participaron en el Mundial de Futbol del año 2010 celebrado en Sudáfrica. La tabla nos da la información siguiente sobre cada jugador:

- Team: país del jugador.
- Position: posición en la juega el jugador: Defender (defensa), Forward (delantero), GoalKeeper (portero) y Midfielder (centrocampista)
- Time: tiempo que ha jugado el jugador en minutos.
- Shots: número de tiros a puerta.
- Passes: número de pases.
- Tackles: número de entradas.
- Saves: número de paradas.

```
library(faraway)
head(worldcup)
```

```
##                Team   Position Time Shots Passes Tackles Saves
## Abdoun      Algeria Midfielder   16     0      6       0     0
## Abe           Japan Midfielder  351     0    101      14     0
## Abidal       France   Defender  180     0     91       6     0
## Abou Diaby   France Midfielder  270     1    111       5     0
## Aboubakar  Cameroon    Forward   46     2     16       0     0
## Abreu       Uruguay    Forward   72     0     15       0     0
```

Ejemplo

Supongamos que queremos calcular una muestra de tamaño indeterminado de los jugadores por conglomerados eligiendo como conglomerados los países a los que éstos pertenecen.

En la tabla de datos hay un total de 32 países.

Elegiremos primero 4 países aleatoriamente y la muestra elegida serán los jugadores que pertenecen a dichos países:

```
set.seed(19)
números.países.elegidos = sample(1:32,4,replace=FALSE)
países.elegidos = unique(worldcup$Team)[números.países.elegidos]
```

Los países elegidos son:

```
países.elegidos
```

```
## [1] Slovakia    Mexico      New Zealand France
## 32 Levels: Algeria Argentina Australia Brazil Cameroon Chile ... Uruguay
```

La muestra elegida estará formada por los jugadores que perteneces a dichos países:

```
muestra.worldcup.con = worldcup[worldcup$Team %in% países.elegidos,]
```

Dicha muestra tiene tamaño 73. Sólo mostramos los datos de los 8 primeros jugadores:

```
head(muestra.worldcup.con,8)
```

```
##                   Team   Position Time Shots Passes Tackles Saves
## Abidal          France   Defender  180     0     91       6     0
## Abou Diaby      France Midfielder  270     1    111       5     0
## Aguilar         Mexico   Defender   55     0     31       2     0
## Alou Diarra     France Midfielder   82     0     31       0     0
## Anelka          France    Forward  117     7     37       1     0
## Barrera         Mexico Midfielder  149     4     59       2     0
## Barron      New Zealand Midfielder    1     0      0       0     0
## Bautista        Mexico    Forward   45     0      8       3     0
```

1.1.7. Muestreo polietápico

Muestreo polietápico: si una vez seleccionada la muestra aleatoria de conglomerados, tomamos de alguna manera una muestra aleatoria de cada uno de ellos, estaremos realizando un **muestreo polietápico**.

La figura muestra un ejemplo sencillo de muestreo polietápico de nuestra urna: hemos elegido al azar 5 conglomerados (marcados en azul) y de cada uno de ellos hemos elegido 3 bolas al azar sin reposición.

Ejemplo `worldcup`

Para realizar un muestreo polietápico con los datos del ejemplo anterior (tabla de datos **worldcup**), podemos elegir una submuestra de 5 jugadores para cada uno de los 4 países elegidos, obteniendo al final una muestra de tamaño 20 de todos los jugadores de la tabla de datos.

Primero definimos las 4 subtablas de datos para los jugadores de cada país elegido:

```
worldcup.pais1 = worldcup[worldcup$Team==países.elegidos[1],]
worldcup.pais2 = worldcup[worldcup$Team==países.elegidos[2],]
```

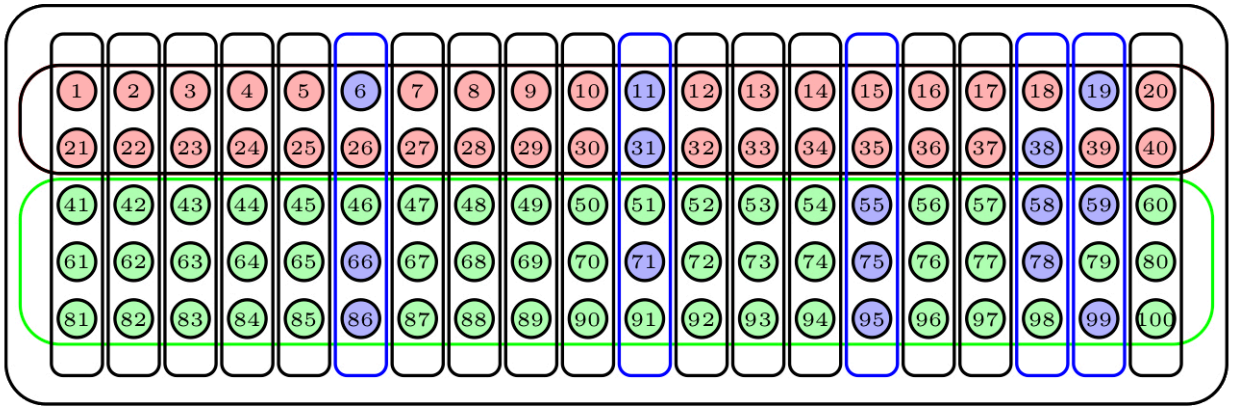


Figura 1.7: Una muestra polietápica de 5 conglomerados y 3 bolas al azar sin reposición

```
worldcup.pais3 = worldcup[worldcup$Team==países.elegidos[3],]
worldcup.pais4 = worldcup[worldcup$Team==países.elegidos[4],]
```

A continuación elegimos los 5 jugadores de cada país:

```
set.seed(28)
jugadores.pais1 = sample(1:dim(worldcup.pais1)[1],5,replace=FALSE)
jugadores.pais2 = sample(1:dim(worldcup.pais2)[1],5,replace=FALSE)
jugadores.pais3 = sample(1:dim(worldcup.pais3)[1],5,replace=FALSE)
jugadores.pais4 = sample(1:dim(worldcup.pais4)[1],5,replace=FALSE)
```

Por último juntamos las submuestras obtenidas de los jugadores de cada país:

```
muestra.worldcup.pol = rbind(worldcup.pais1[jugadores.pais1,],
                             worldcup.pais2[jugadores.pais2,],
                             worldcup.pais3[jugadores.pais3,],
                             worldcup.pais4[jugadores.pais4,])
```

Y finalmente los mostramos por pantalla: (mostramos sólo los 12 primeros)

```
head(muestra.worldcup.pol,12)
```

```
##                Team   Position Time Shots Passes Tackles Saves
## Stoch        Slovakia Midfielder  193     2     76       1     0
## Zabavnik     Slovakia   Defender  268     1     94       8     0
## Kucka        Slovakia Midfielder  181     4     71      10     0
## Weiss        Slovakia Midfielder  269     2     84       2     0
## Durica       Slovakia   Defender  360     1    159       4     0
## PerezM         Mexico Goalkeeper  360     0     58       0    13
## Moreno         Mexico   Defender  147     0     74       4     0
## Aguilar        Mexico   Defender   55     0     31       2     0
## Bautista       Mexico    Forward   45     0      8       3     0
```

```
## Hernandez      Mexico    Forward  169     6     37       1     0
## Nelsen    New Zealand   Defender  270     0     92       1     0
## Reid      New Zealand   Defender  270     2     90      10     0
```

1.2. Guía rápida en R

- `sample(x, n, replace=...)` genera una muestra aleatoria de tamaño `n` del vector `x`. Si `x` es un número natural `x`, representa el vector $1, 2, \ldots, x$. Dispone de los dos parámetros siguientes:
  - `replace` que igualado a TRUE produce muestras con reposición e igualado a FALSE (su valor por defecto) produce muestras sin reposición.
  - `prob`, que permite especificar las probabilidades de aparición de los diferentes elementos de `x` (por defecto, son todas la misma).
- `set.seed` permite fijar la semilla de aleatoriedad.

Capítulo 2

Estimación Puntual

El problema usual de la **estadística inferencial** es:

- Queremos conocer el valor de una característica en una población
- No podemos medir esta característica en todos los individuos de la población
- Extraemos una muestra aleatoria de la población, medimos la característica en los individuos de esta muestra e **inferimos** el valor de la característica para la toda la población
    - ¿Cómo lo tenemos que hacer?
    - ¿Cómo tenemos que hacer la muestra?
    - ¿Qué información podemos inferir?

Muestra aleatoria simple (m.a.s.) de tamaño n: de una población de N individuos, repetimos n veces el proceso consistente en escoger **equiprobablemente** un individuo de la población; *los individuos escogidos se pueden repetir*

Ejemplo

Escogemos al azar n estudiantes de la Universidad de las Islas Baleares (UIB) (con reposición) para medirles la estatura

De esta manera, todas las muestras posibles de n individuos (posiblemente repetidos: *multiconjuntos*) tienen la misma probabilidad

Estadístico (*Estimador puntual*): una función que aplicada a una muestra nos permite *estimar* un valor que queramos conocer sobre toda la población.

Ejemplo

La media de las estaturas de una muestra de estudiantes de la UIB nos permite estimar la media de las alturas de todos los estudiantes de la UIB.

Una m.a.s. de tamaño n (de una v.a. X) es

- un conjunto de n copias independientes de X, o
- un conjunto de n variables aleatorias independientes $X_1, \dots, X_n$, todas con la distribución de X.

Ejemplo

Sea X la v.a. "escogemos un estudiante de la UIB y le medimos la altura". Una m.a.s. de X de tamaño n serán n copias independientes $X_1, \dots, X_n$ de esta X.

Una realización de una m.a.s. son los n valores $x_1, \dots, x_n$ que toman las v.a. $X_1, \dots, X_n$.

Un *estadístico* T es una función aplicada a la muestra $X_1, \dots, X_n$:

$$T = f(X_1, \dots, X_n)$$

Este estadístico se aplica a las realizaciones de la muestra

Definición: la **media muestral** de una m.a.s. $X_1, \dots, X_n$ de tamaño n es

$$\overline{X} := \frac{X_1 + \cdots + X_n}{n}$$

y estima $E(X)$.

Ejemplo:

La **media muestral** de las alturas de una realización de una m.a.s. de las alturas de estudiantes estima la altura media de un estudiante de la UIB.

Así pues, un **estadístico** es una (otra) variable aleatoria, con distribución, esperanza, etc.

La **distribución muestral** de T es la distribución de esta variable aleatoria.

Estudiando esta distribución muestral, podremos estimar propiedades de X a partir del comportamiento de una muestra.

Error estándar de T: desviación típica de T.

Convenio: LOS ESTADÍSTICOS, EN MAYÚSCULAS; las realizaciones, en minúsculas

- $X_1, \dots, X_n$ una m.a.s. y
$$\overline{X} := \frac{X_1 + \cdots + X_n}{n},$$
el estadístico media muestral.
- $x_1, \dots, x_n$ una realización de esta m.a.s. y
$$\overline{x} := \frac{x_1 + \cdots + x_n}{n},$$
la media (muestral) de esta realización.

En la vida real, las muestras aleatorias se toman, casi siempre, sin reposición (es decir sin repetición del mismo individuo de la población).

No son muestras aleatorias simples. pero:

- Si N es mucho más grande que n, los resultados para una m.a.s. son (aproximadamente) los mismos, ya que las repeticiones son improbables y las variables aleatorias que forman la muestra son prácticamente independientes.
- En estos casos cometeremos el abuso de lenguaje de decir que es una m.a.s.
- Si n es relativamente grande, se suelen dar versiones corregidas de los estadísticos.

2.1. La media muestral

2.1.1. Definición de media muestral

Media muestral : sea $X_1, \dots, X_n$ una m.a.s. de tamaño n de una v.a. X de esperanza μ_X y desviación típica σ_X

La *media muestral* es:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

En estas condiciones,

$$E(\overline{X}) = \mu_X, \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

donde $\sigma_{\overline{X}}$ es el **error estándar** de $\overline{X}$.

2.1.2. Propiedades de la media muestral

- Es un estimador puntual de μ_X
- $E(\overline{X}) = \mu_X$: el valor esperado de $\overline{X}$ es μ_X.
- Si tomamos muchas veces una m.a.s. y calculamos la media muestral, el valor medio de estas medias tiende con mucha probabilidad a ser μ_X.
- $\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$: la variabilidad de los resultados de $\overline{X}$ tiende a 0 a medida que tomamos muestras más grandes.

Ejemplo del `iris`

Consideremos la tabla de datos `iris` (que ya vimos en el tema de *Muestreo*). Vamos a comprobar las propiedades anteriores sobre la variable **longitud del pétalo** (`Petal.Length`).

1. Generaremos 10000 muestras de tamaño 40 con reposición de las longitudes del pétalo.
2. A continuación hallaremos los valores medios de cada muestra.
3. Consideraremos la media y la desviación típica de dichos valores medios y los compararemos con los valores exactos dados por las propiedades de la media muestral.

Para generar los valores medios de las longitudes del pétalo de las 10000 muestras usaremos la función `replicate` de `R`. Fijaos en su sintaxis:

- `replicate(n,expresión)` evalúa `n` veces la `expresión`, y organiza los resultados como las columnas de una matriz (o un vector, si el resultado de cada `expresión` es unidimensional).

```
set.seed(1001)
valores.medios.long.pétalo=replicate(10000,mean(sample(iris$Petal.Length,40,
                                                       replace =TRUE)))
```

Los valores medios de las 10 primeras muestras anteriores serían

```
##  [1] 3.5975 3.5150 3.9400 3.2650 3.9125 3.9650 4.2825 3.2950 3.8500 3.7850
```

El valor medio de los valores medios de las muestras anteriores vale:

```
mean(valores.medios.long.pétalo)
```

```
## [1] 3.754478
```

Dicho valor tiene que estar cerca del valor medio de la variable longitud del pétalo:

```
mean(iris$Petal.Length)
```

```
## [1] 3.758
```

Fijaos que los dos valores están muy próximos.

La desviación típica de los valores medios de las muestras vale:

```
sd(valores.medios.long.pétalo)
```

```
## [1] 0.27965126
```

Dicho valor tiene que estar cerca de $\frac{\sigma_{lp}}{\sqrt{40}}$ (donde σ_{lp} es la desviación típica de la variable longitud del pétalo) tal como predice la propiedad de la media muestral referida a la desviación típica de la misma:

```
sd(iris$Petal.Length)/sqrt(40)
```

```
## [1] 0.27911816
```

Fijaos también en que los dos valores están muy próximos.

2.2. Poblaciones normales

2.2.1. Combinación lineal de distribuciones normales

Proposición La combinación lineal de distribuciones normales es normal. Es decir, si $Y_1, \dots, Y_n$ son v.a. normales independientes, cada $Y_i \sim N(\mu_i, \sigma_i)$, y $a_1, \dots, a_n, b \in \mathbb{R}$ entonces

$$Y = a_1 Y_1 + \cdots + a_n Y_n + b$$

es una v.a. $N(\mu, \sigma)$ con μ y σ las que correspondan:

- $E(Y) = a_1 \cdot \mu_1 + \cdots + a_n \cdot \mu_n + b$
- $\sigma(Y)^2 = a_1^2 \cdot \sigma_1^2 + \cdots + a_n^2 \cdot \sigma_n^2$

2.2.2. Distribución de la media muestral

Veamos cómo se distribuye la media muestral en el caso en que la población X sea normal.

Proposición

Sea $X_1, \dots, X_n$ una m.a.s. de una v.a. X de esperanza μ_X y desviación típica σ_X.

Si X es $N(\mu_X, \sigma_X)$, entonces

$$\overline{X} \text{ es } N\Big(\mu_X, \frac{\sigma_X}{\sqrt{n}}\Big)$$

y por lo tanto

$$Z = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \text{ es } N(0,1)$$

Z es la **expresión tipificada** de la media muestral.

2.2.3. Teorema Central del Límite

Teorema Central del Límite. Sea $X_1, \ldots, X_n$ una m.a.s. de una v.a. X **cualquiera** de esperanza μ_X y desviación típica σ_X. Cuando $n \to \infty$,

$$\overline{X} \to N\Big(\mu_X, \frac{\sigma_X}{\sqrt{n}}\Big)$$

y por lo tanto

$$Z = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \to N(0,1)$$

(estas convergencias se refieren a las distribuciones.)

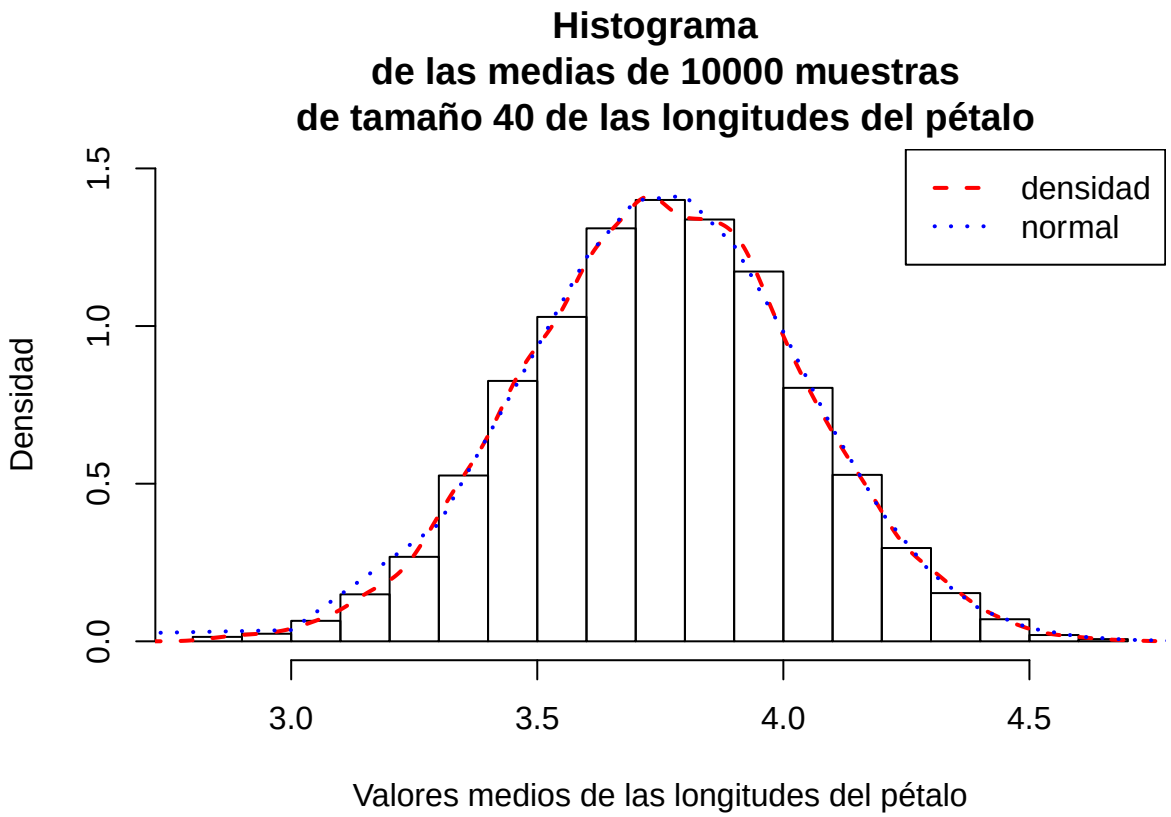
Caso n grande: Si n es grande ($n \geq 30$ **o 40**), $\overline{X}$ es aproximadamente normal, con esperanza μ_X y desviación típica $\dfrac{\sigma_X}{\sqrt{n}}$

Ejemplo

Tenemos una v.a. X de media $\mu_X = 3$ y desviación típica. $\sigma_X = 0{,}2$. Tomamos muestras aleatorias simples de tamaño 50. La distribución de la media muestral $\overline{X}$ es aproximadamente

$$N\left(3, \frac{0{,}2}{\sqrt{50}}\right) = N(3, 0{,}0283).$$

En el gráfico siguiente podemos observar el histograma de los valores medios de las longitudes del pétalo de las 10000 muestras junto con la distribución normal correspondiente:



Ejercicio

El tamaño en megabytes (MB) de un tipo de imágenes comprimidas tiene un valor medio de 115 MB, con una desviación típica de 25. Tomamos una m.a.s. de 100 imágenes de este tipo.

¿Cuál es la probabilidad de que la media muestral del tamaño de los ficheros sea ≤ 110 MB?

Sea X la variable aleatoria que nos da el tamaño en megabytes del tipo de imágenes comprimidas. La distribución de X será $X = N(\mu = 115, \sigma = 25)$

Sea $X_1, \ldots, X_{100}$ la m.a.s. La distribución aproximada de la media muestral $\overline{X}$ usando el **Teorema Central del Límite** será: $\overline{X} \approx N\left(\mu_{\overline{X}} = 115, \sigma_{\overline{X}} = \frac{25}{\sqrt{100}} = 2{,}5\right)$.

Nos piden la probabilidad siguiente: $P(\overline{X} \leq 110)$. Si estandarizamos:

$$P(\overline{X} \leq 110) = P\left(Z = \frac{\overline{X} - 115}{2{,}5} \leq \frac{110 - 115}{2{,}5}\right) = p(Z \leq -2) = 0{,}0228.$$

donde Z es la normal estándar $N(0, 1)$

2.2.4. Media muestral en muestras sin reposición

Sea $X_1, \ldots, X_n$ una m.a. **sin reposición** de tamaño n de una v.a. X de esperanza μ_X y desviación típica σ_X.

Si n es pequeño en relación al tamaño N de la población, todo lo que hemos contado funciona (aproximadamente).

Si n es grande en relación a N, entonces

$$E(\overline{X}) = \mu_X, \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

(**factor de población finita**)

El Teorema Central del Límite ya no funciona exactamente en este último caso.

2.3. Proporción muestral

2.3.1. Proporción muestral. Definición

Proporción muestral. Sea X una v.a. Bernoulli de parámetro p_X (1 éxito, 0 fracaso). Sea $X_1, \ldots, X_n$ una m.a.s. de tamaño n de X.

$S = \sum_{i=1}^{n} X_i$ es el nombre de éxitos observados es $B(n, p)$.

La **proporción muestral** es

$$\hat{p}_X = \frac{S}{n}$$

y es un estimador de p_X.

Notemos que $\hat{p}_X$ es un caso particular de $\overline{X}$, por lo que todo lo que hemos dicho para medias muestrales es cierto para proporciones muestrales.

2.3.2. Proporción muestral. Propiedades

Proposición

- Valor esperado de la proporción muestral:

$$E(\hat{p}_X) = p_X$$

- **Error estándar** de la proporción muestral:

$$\sigma_{\hat{p}_X} = \sqrt{\frac{p_X(1-p_X)}{n}}$$

- Si la muestra es sin reposición y n es relativamente grande,

$$\sigma_{\hat{p}_X} = \sqrt{\frac{p_X(1-p_X)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}.$$

Teorema: Si n es grande ($n \geq 30$ o 40) y la muestra es aleatoria simple, usando el Teorema Central del Límite,

$$\frac{\hat{p}_X - p_X}{\sqrt{\frac{p_X(1-p_X)}{n}}} \approx N(0,1)$$

Ejemplo del `iris`

Dada una muestra de 60 flores de la tabla de datos `iris`,

1. Estimar la proporción de flores de la especie `setosa`.
2. Estimar también la desviación estándar de dicha proporción.

Primero generamos la muestra de las 60 flores:

```
set.seed(1000)
flores.elegidas = sample(1:150,60,replace=TRUE)
muestra.flores = iris[flores.elegidas,]
```

A continuación miramos cuántas flores de la muestra son de la especie `setosa`:

```
table(muestra.flores$Species=="setosa")
```

```
##
## FALSE  TRUE
##    39    21
```

Tenemos entonces 21 flores de la especie `setosa`.

La estimación de la proporción de flores de especie `setosa` será:

```
(prop.setosa = table(muestra.flores$Species=="setosa")[2]/length(muestra.flores$Species))
```

```
## TRUE
## 0.35
```

valor que no está muy lejos del valor poblacional de la proporción p_{setosa} que es $p_{setosa} = \frac{50}{150} = 0{,}3333$.

Para estimar la desviación estándar de la proporción muestral de flores de tamaño 60 de la especie `setosa`, repetiremos el experimento anterior 10000 veces y hallaremos la desviación estándar de las proporciones obtenidas. Al final, compararemos dicho valor con el valor exacto dado por la propiedad correspondiente.

Para generar las proporciones de las 10000 muestras usaremos la función `replicate` de `R`:

```
set.seed(1002)
props.muestrales = replicate(10000,table(sample(iris$Species,60,
                                replace=TRUE)=="setosa")[2]/60)
```

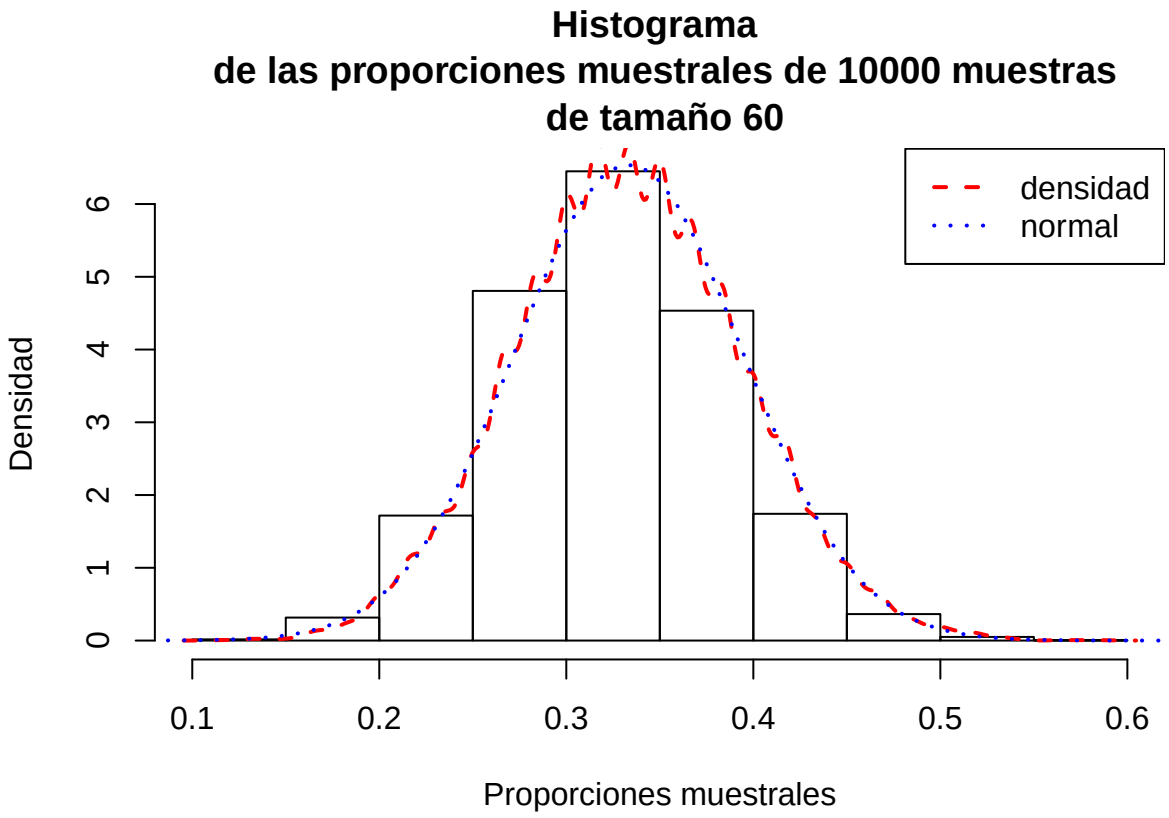
La desviación típica de las proporciones muestrales anteriores vale:

```
sd(props.muestrales)
```

```
## [1] 0.060210983
```

valor muy próximo al valor real que vale: $\sigma_{\hat{p}_X} = \sqrt{\frac{p_X(1-p_X)}{n}} = \sqrt{\frac{\frac{50}{150}\cdot\left(1-\frac{50}{150}\right)}{60}} = 0{,}0609.$

En el gráfico siguiente podemos observar el histograma de las proporciones muestrales de las 10000 muestras junto con la distribución normal correspondiente:



2.4. Varianza muestral y desviación típica muestral

Varianza muestral y desviación típica muestral. Sea $X_1, \ldots, X_n$ una m.a.s. de tamaño n de una v.a. X de esperanza μ_X y desviación típica σ_X.

La **varianza muestral** es

$$\widetilde{S}_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

La **desviación típica muestral** es

$$\widetilde{S}_X = +\sqrt{\widetilde{S}_X^2}$$

Además, escribiremos

$$S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \frac{(n-1)}{n}\widetilde{S}_X^2 \quad \text{y} \quad S_X = +\sqrt{S_X^2}$$

Propiedades

$$S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \left(\frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2 \right)$$

$$\widetilde{S}_X^2 = \frac{n}{n-1}\left(\frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2 \right)$$

Teorema. Si la v.a. X es normal, entonces $E(\widetilde{S}_X^2) = \sigma_X^2$ y la v.a.

$$\frac{(n-1)\widetilde{S}_X^2}{\sigma_X^2}$$

tiene distribución χ^2_{n-1}.

Distribución χ^2_n

La distribución χ^2_n (χ: en catalán, **khi**; en castellano, **ji**; en inglés, **chi**), donde n es un parámetro llamado **grados de libertad**:

es la de

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

donde $Z_1, Z_2, \ldots, Z_n$ son v.a. independientes $N(0,1)$.

Propiedades χ^2_n

- Su función de densidad es:

$$f_{\chi^2_n}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}, \quad \text{si } x \geq 0$$

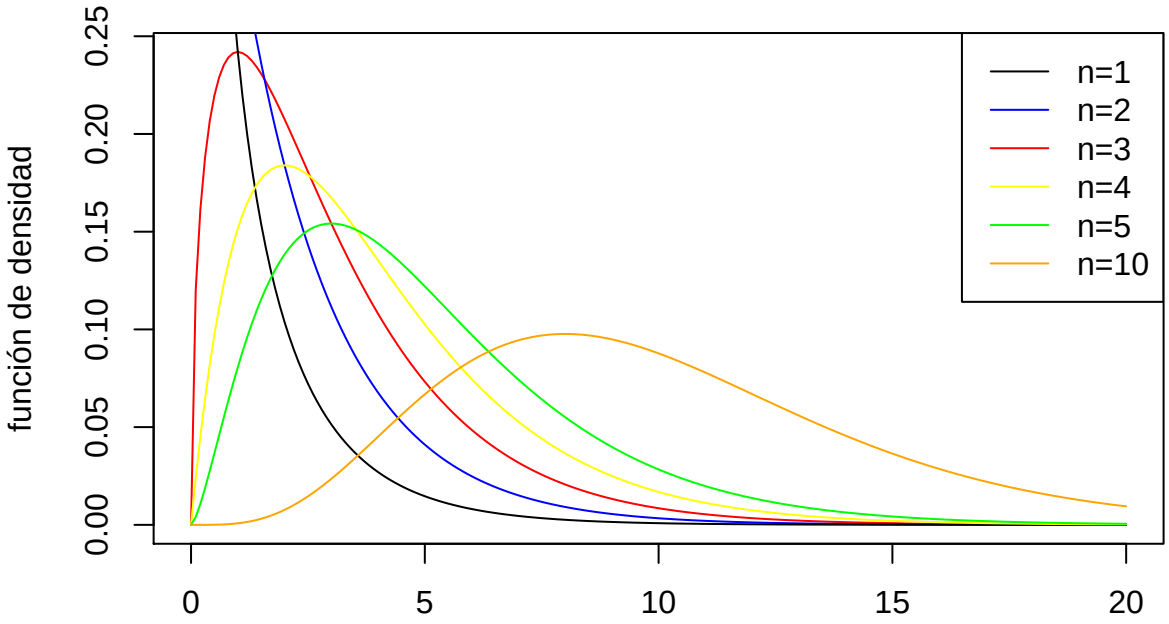
  donde $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$, si $x > 0$.

- Si $X_{\chi^2_n}$ es una v.a. con distribución χ^2_n,

$$E\left(X_{\chi^2_n}\right) = n, \quad Var\left(X_{\chi^2_n}\right) = 2n$$

- χ^2_n se aproxima a una distribución normal $N\left(n, \sqrt{2n}\right)$ para n grande ($n > 40$ o 50).

El gráfico de la función de densidad de distintas distribuciones χ^2_n para $n = 1, 2, 3, 4, 5, 10$ se puede observar en el gráfico siguiente:



Ejemplo

Supongamos que el aumento diario del la ocupación de una granja de discos duros medido en Gigas sigue una distribución normal con desviación típica 1,7. Se toma una muestras de 12 discos. Supongamos que esta muestra es pequeña respecto del total de la población de la granja de discos.

¿Cual es la probabilidad de que la desviación típica muestral sea $\leq 2{,}5$?

Sea X= aumento diario en Gigas de un disco duro elegido al azar.

Sabemos que $\sigma^2_X = (1{,}7)^2 = 2{,}89$.

Como que X es normal y $n = 12$, tenemos que

$$\frac{11 \cdot \widetilde{S}^2_X}{2{,}89} = \frac{(n-1)\widetilde{S}^2_X}{\sigma^2_X} \sim \chi^2_{11}$$

Nos piden: $P(\widetilde{S}_X < 2{,}5) = P\left(\widetilde{S}^2_X < 2{,}5^2\right)$:

$$P\left(\widetilde{S}^2_X < 2{,}5^2\right) = P\left(\frac{11 \cdot \widetilde{S}^2_X}{2{,}89} < \frac{11 \cdot 2{,}5^2}{2{,}89}\right) = P(\chi^2_{11} < 23{,}7889) = 0{,}9863.$$

2.5. Propiedades de los estimadores

2.5.1. Estimadores insesgados

¿Cuándo un estimador es bueno?

Estimadores insesgados Un estimador puntual $\hat{\theta}$ de un parámetro poblacional θ es **insesgado, no sesgado o sin sesgo** cuando su valor esperado es precisamente el valor del parámetro:

$$E(\hat{\theta}) = \theta$$

Entonces se dice que el estimador puntual es **no sesgado**.

El **sesgo** de $\hat{\theta}$ es la diferencia

$$E(\hat{\theta}) - \theta$$

Proposición

- $\overline{X}$ es estimador no sesgado de μ_X: $E(\overline{X}) = \mu_X$.
- $\hat{p}_X$ es estimador no sesgado de p_X: $E(\hat{p}_X) = p_X$.
- Si X es normal: $\widetilde{S}_X^2$ es estimador no sesgado de σ_X^2: $E(\widetilde{S}_X^2) = \sigma_X^2$
- Si X es normal: $E(S_X^2) = \dfrac{n-1}{n}\sigma_X^2$. Por lo tanto S_X^2, es sesgado, con sesgo

$$E(S_X^2) - \sigma_X^2 = \frac{n-1}{n}\sigma_X^2 - \sigma_X^2 = -\frac{\sigma_X^2}{n} \text{ que tiende a 0.}$$

2.5.2. Estimadores eficientes

¿Cuando un estimador es **bueno**?

Cuando es no segado y tiene poca variabilidad (así es más probable que aplicado a una m.a.s. dé un valor más cercano al valor esperado)

Error estándar de un estimador $\hat{\theta}$: es su desviación típica

$$\sigma_{\hat{\theta}} = \sqrt{Var(\hat{\theta})}$$

Eficiencia de un estimador Dados dos estimadores $\hat{\theta}_1$, $\hat{\theta}_2$ no sesgados (o con sesgo que tiende a 0) del mismo parámetro θ, diremos que $\hat{\theta}_1$ es **más eficiente** que $\hat{\theta}_2$ cuando

$$\sigma_{\hat{\theta}_1} < \sigma_{\hat{\theta}_2},$$

es decir, cuando

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2).$$

Ejemplo de estimador eficiente.

Sea X una v.a. con media μ_X y desviación típica σ_X

Consideremos la mediana $Me = Q_{0,5}$ de la realización de una m.a.s. de X como estimador puntual de μ_X

Si X es normal,

$$E(Me) = \mu_X,$$
$$Var(Me) \approx \frac{\pi}{2}\frac{\sigma_X^2}{n} \approx \frac{1{,}57\sigma_X^2}{n} = 1{,}57 \cdot Var(\overline{X}) > Var(\overline{X}).$$

Por lo tanto, Me es un estimador no sesgado de μ_X, pero menos eficiente que $\overline{X}$.

Proposición

- Si la población es normal, la **media muestral** $\bar{X}$ es el estimador no sesgado más eficiente de la **media poblacional** μ_X.
- Si la población es Bernoulli, la **proporción muestral** $\hat{p}_X$ es el estimador no sesgado más eficiente de la **proporción poblacional** p_X.
- Si la población es normal, la **varianza muestral** $\widetilde{S}_X^2$ es el estimador no sesgado más eficiente de la **varianza poblacional** σ_x^2.

¿S_X^2 o $\widetilde{S}_X^2$?

Como hemos visto, si la población es normal, la varianza muestral es el estimador no sesgado más eficiente de la varianza poblacional

El estimador *varianza*

$$S_X^2 = \frac{(n-1)}{n}\widetilde{S}_X^2$$

aunque sea más eficiente, tiene sesgo que tiende a 0.

Si n es pequeño (≤ 30 o 40), es mejor utilizar la varianza muestral $\widetilde{S}_X^2$ para estimar la varianza, ya que el sesgo influye, pero si n es grande, el sesgo ya no es tan importante y se puede utilizar S_X^2.

Estimación del tamaño de la población.

Tenemos una población numerada $1, 2, \ldots, N$

Tomamos una m.a.s. $x_1, \ldots, x_n$; sea $m = \text{máx}\{x_1, \ldots, x_n\}$.

Teorema. El estimador no segado más eficiente del tamaño de la población N es

$$\widehat{N} = m + \frac{m-n}{n}.$$

O sea, la manera más eficiente de estimar el número de elementos de la población a partir de una muestra es usar la fórmula anterior.

Ejemplo

Sentados en una terraza de un bar del Paseo Marítimo de Palma hemos anotado el número de licencia de los 40 primeros taxis que hemos visto pasar:

```
taxis=c(1217,600,883,1026,150,715,297,137,508,134,38,961,538,1154,
        314,1121,823,158,940,99,977,286,1006,1207,264,1183,1120,
        498,606,566,1239,860,114,701,381,836,561,494,858,187)
```

Supondremos que estas observaciones son una m.a.s. de los taxis de Palma. Vamos a estimar el número total de taxis.

Entonces, estimamos que el número de taxis de Palma es

```
(N=max(taxis)+(max(taxis)-length(taxis))/length(taxis))
```

```
## [1] 1268.975
```

En realidad, hay 1246.

2.5.3. Estimadores máximo verosímiles

¿Cómo encontramos buenos estimadores?

Antes de explicar la metodología, necesitamos una definición previa:

Función de verosimilitud de la muestra. Sea X una v.a. **discreta** con función de probabilidad $f_X(x;\theta)$ que depende de un parámetro desconocido θ.

Sea $X_1, \dots X_n$ una m.a.s. de X, y sea $x_1, x_2, \dots, x_n$ una realización de esta muestra.

La **función de verosimilitud** de la muestra es la probabilidad condicionada siguiente:

$$\begin{aligned} L(\theta|x_1, x_2, \dots, x_n) &:= P(x_1, x_2, \dots, x_n|\theta) = P(X_1 = x_1) \cdots P(X_n = x_n) \\ &= f_X(x_1;\theta) \cdots f_X(x_n;\theta). \end{aligned}$$

Dada la función de verosimilitud $L(\theta|x_1, \dots, x_n)$ de la muestra, indicaremos por $\hat{\theta}(x_1, \dots, x_n)$ el valor del parámetro θ en el que se alcanza el máximo de $L(\theta|x_1, \dots, x_n)$. Será una función de $x_1, \dots, x_n$.

Estimador máximo verosímil. Un estimador $\hat{\theta}$ de un parámetro θ es **máximo verosímil** (**MV**) cuando, para cada m.a.s, la probabilidad de observarlo es máxima, cuando el parámetro toma el valor del estimador aplicado a la muestra, es decir, si la función de verosimilitud

$$L(\theta|x_1, x_2, \dots, x_n) = P(x_1, x_2, \dots, x_n|\theta)$$

alcanza su máximo.

Ejemplo de estimador máximo verosímil.

Supongamos que tenemos una v.a. Bernoulli X de probabilidad de éxito p desconocida.

Para cada m.a.s. $x_1, \dots, x_n$ de X, sean $\hat{p}_x$ su proporción muestral y $P(x_1, \dots, x_n \mid p)$ la probabilidad de obtenerla cuando el verdadero valor del parámetro es p.

Teorema. El valor de p para el que $P(x_1, \dots, x_n \mid p)$ es máximo es $\hat{p}_x$.

Dicho en otras palabras, la proporción muestral es un estimador MV de p.

Ejercicio

Demostrar el teorema anterior.

En general, al ser ln una función creciente, en lugar de maximizar $L(\theta|x_1, \dots, x_n)$, maximizamos

$$\ln(L(\theta|x_1, \dots, x_n))$$

que suele ser más simple (ya que transforma los productos en sumas, y es más fácil derivar estas últimas).

Sea $X_1, \dots X_n$ una m.a.s. de una v.a. Bernoulli X de parámetro p (desconocido). Denotemos $q = 1 - p$

$$f_X(1;p) = P(X = 1) = p, \quad f_X(0;p) = P(X = 0) = q$$

es a decir, para $x \in \{0, 1\}$, resulta que $f_X(x;p) = P(X = x) = p^x q^{1-x}$.

La función de verosimilitud es:

$$\begin{array}{l} L(p|x_1, \dots, x_n) = f_X(x_1;p) \cdots f_X(x_n;p) = p^{x_1} q^{1-x_1} \cdots p^{x_n} q^{1-x_n} \\ = p^{\sum_{i=1}^n x_i} q^{\sum_{i=1}^n (1-x_i)} = p^{\sum_{i=1}^n x_i} q^{n-\sum_{i=1}^n x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} \end{array}$$

La función de verosimilitud es

$$L(p|x_1, \dots, x_n) = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} = p^{n\overline{x}} (1-p)^{n-n\overline{x}},$$

donde $\overline{x} = \dfrac{\sum_{i=1}^n x_i}{n}$.

Queremos encontrar el valor de p en el que se alcanza el máximo de esta función (donde $\overline{x}$ es un parámetro y la variable es p)

Maximizaremos su logaritmo:

$$\ln(L(p|x_1, \dots, x_n)) = n\overline{x} \ln(p) + n(1-\overline{x}) \ln(1-p).$$

Derivamos respecto de p:

$$\begin{array}{l} \ln(L(p|x_1, \dots, x_n))' = n\overline{x} \frac{1}{p} - n(1-\overline{x}) \frac{1}{1-p} \\ \quad = \frac{1}{p(1-p)} \Big((1-p) n\overline{x} - pn(1-\overline{x}) \Big) = \frac{1}{p(1-p)} (n\overline{x} - pn) \\ \quad = \frac{n}{p(1-p)} (\overline{x} - p) \end{array}$$

Estudiamos el signo:

$$\ln(L(p|x_1, \dots, x_n))' \geq 0 \Leftrightarrow \overline{x} - p \geq 0 \Leftrightarrow p \leq \overline{x}$$

Por lo tanto

$$\ln(L(p|x_1, \dots, x_n)) \begin{cases} \text{creciente hasta } \overline{x} \\ \text{decreciente a partir de } \overline{x} \\ \text{tiene un máximo en } \overline{x} \end{cases}$$

El resultado queda demostrado. $L(\hat{p}_X|x_1, \dots, x_n) \geq L(p|x_1, \dots, x_n)$ para cualquier p.

2.5.4. Algunos estimadores Máximo Verosímiles

Proposición

- $\hat{p}_x$ es el estimador MV del parámetro p de una v.a. Bernoulli.
- $\overline{X}$ es el estimador MV del parámetro θ de una v.a. Poisson.
- $\overline{X}$ es el estimador MV del parámetro μ de una v.a. normal.
- S_X^2 (**no** $\widetilde{S}_X^2$) es el estimador MV del parámetro σ^2 de una v.a. normal.

- El máximo (**no** $\widehat{N}$) es el estimador MV de la N en el problema de los taxis.

Ejemplo de estimadores máximo verosímiles: captura-recaptura.

En una población hay N individuos, capturamos K, los marcamos y los volvemos a soltar.

Ahora volvemos a capturar n, de los que k están marcados. A partir de estos datos, queremos estimar N.

Supongamos que N y K no han cambiado de la primera a la segunda captura.

La variable aleatoria $X =$ *Un individuo esté marcado* es $Be(p)$ con $p = \dfrac{K}{N}$.

Si $X_1, \ldots, X_n$ es la muestra capturada la segunda vez, entonces $\hat{p}_X = \frac{k}{n}$.

$\hat{p}_X$ es un estimador máximo verosímil p. Por tanto, estimamos que:

$$\frac{K}{N} = \frac{k}{n} \Rightarrow N = \frac{n \cdot K}{k}$$

Por lo tanto, el estimador

$$\widehat{N} = \frac{n \cdot K}{k}$$

maximiza la probabilidad de la observación *k marcados de n capturados*, por lo que $\hat{N}$ es el **estimador máximo verosímil** de N.

Ejercicio

Supongamos que hemos marcado 15 peces del lago, y que en una captura, de 10 peces, hay 4 marcados. ¿Cuántos peces estimamos que contiene el lago?

El número de peces estimados del lago será:

$$\widehat{N} = \frac{15 \cdot 10}{4} = 37{,}5$$

Estimamos que habrá entre 37 y 38 peces en el lago.

El estimador

$$\widehat{N} = \frac{n \cdot K}{k}$$

es sesgado, con sesgo que tiende a 0.

El **estimador de Chapman**

$$\widehat{N} = \frac{(n+1) \cdot (K+1)}{k+1} - 1,$$

es menos segado para muestras pequeñas, y no sesgado si $K + n \geq N$ (pero no máximo verosímil).

2.6. Estimación puntual con `R`

2.6.1. La función `fitdistr`

Para obtener estimaciones puntuales con `R` hay que usar la función `fitdistr` del paquete **MASS**:

```
fitdistr(x, densfun=..., start=...)
```

donde

- `x` es la muestra, un vector numérico.
- El valor de `densfun` ha de ser el nombre de la familia de distribuciones:`çhi-squared"`, `.exponential"`, `"f"`, `"geometric"`, `"lognormal"`, `"normal"` y `"poisson"`.
- Si `fitdistr` no dispone de una fórmula cerrada para el estimador máximo verosímil de algún parámetro, usa un algoritmo numérico para aproximarlo que requiere de un valor inicial para arrancar. Este valor (o valores) se puede especificar igualando el parámetro `start` a una `list` con cada parámetro a estimar igualado a un valor inicial.

Estimación del parámetro λ de una variable de Poisson.

Consideramos la muestra siguiente de tamaño 50 de una variable de Poisson de parámetro $\lambda = 5$:

```
set.seed(98)
muestra.poisson = rpois(50,lambda=5)
muestra.poisson
```

```
##  [1]  5  4  4  5  3  4  1  4  6  3  7  7  3  5  4  8  4  4  6  3  6  4  6 11  4
## [26]  7  5  2  8  3  5  4  1  5  6  4  7  7  3  4  6 10  5  4  2  9  1  5  2  2
```

Vamos a estimar el valor del parámetro λ a partir de la muestra anterior.

Para estimar λ usamos la función `fitdistr`:

```
library(MASS)
fitdistr(muestra.poisson, densfun = "poisson")
```

```
##      lambda
##   4.76000000
##  (0.30854497)
```

La función `fitdistr` nos ha dado el siguiente valor de λ: 4.76, valor que se aproxima al valor real de $\lambda = 5$, con un error típico de 0.30854497.

Recordemos que el estimador máximo verosímil de λ es $\overline{X}$ con error típico $\frac{\sqrt{\lambda}}{\sqrt{n}}$. Veamos si la función `fitdistr` nos ha mentido:

```
(estimación.lambda = mean(muestra.poisson))
```

```
## [1] 4.76
```

```
(estimación.error.típico= sqrt(estimación.lambda/50))
```

```
## [1] 0.30854497
```

Comprobamos que los valores anteriores coinciden con los dados por la función.

¿Qué estimaciones hubiésemos obtenido de la media μ y la desviación típica σ si suponemos que la muestra anterior es normal?

```
fitdistr(muestra.poisson,densfun = "normal")
```

```
##       mean          sd
##   4.76000000   2.18686991
##  (0.30927011) (0.21868699)
```

Dichos valores coinciden con la media muestral $\overline{X}$ y la desviación típica "verdadera" de la muestra considerada:

```
sd(muestra.poisson)*sqrt(49/50)
```

```
## [1] 2.1868699
```

2.7. Guía rápida

- `fitdistr` del paquete **MASS**, sirve para calcular los estimadores máximo verosímiles de los parámetros de una distribución a partir de una muestra. Parámetros principales:
  - `densfun`: el nombre de la familia de distribuciones, entre comillas.
  - `start`: permite fijar el valor inicial del algoritmo numérico para calcular el estimador, si la función lo requiere.

Capítulo 3

Intervalos de confianza

Una estimación por intervalos de un parámetro poblacional es una regla para calcular, a partir de una muestra, un intervalo en el que, con una cierta probabilidad (nivel de confianza), se encuentra el valor verdadero del parámetro.

Estas reglas definirán, a su vez, estimadores.

Ejemplo

Hemos escogido al azar 50 estudiantes de grado de la UIB, hemos calculada sus notas medias de las asignaturas del primer semestre, y la media de estas medias ha sido un 6.3, con una varianza muestral de 1.8.

Determinar un intervalo del que podamos afirmar con probabilidad 95 % que contiene la media real de las notas medias de los estudiantes de grado de la UIB este primer semestre.

Ejemplo

En un experimento en el que se ha medido la tasa oficial de alcoholemia en sangre a 40 varones (sobrios) después de tomar 3 cañas de cerveza de 330 ml. La media y la desviación típica de esta tasa han sido

$$\overline{x} = 0{,}7, \quad \tilde{s} = 0{,}1.$$

Determinar un intervalo que podamos afirmar con probabilidad 95 % que contiene la tasa de alcoholemia media en sangre de una varón después de beber 3 cañas de cerveza de 330 ml.

Intervalo de confianza. Dado un parámetro θ, el intervalo (A, B) es un intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ para al parámetro θ cuando

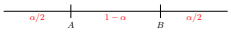
$$P(A < \theta < B) = 1 - \alpha.$$

El valor $(1-\alpha)\cdot 100\,\%$ (o contiene solo el $1-\alpha$) recibe el nombre de **nivel de confianza**.

El valor α recibe el nombre de **nivel de significación**.

Por defecto, buscaremos intervalos bilaterales tales que la **cola de probabilidad sobrante** α se reparta por igual a cada lado del intervalo:

$$P(\theta < A) = P(\theta > B) = \frac{\alpha}{2}$$


α/2
A
1 − α
B
α/2

Por ejemplo, para buscar un intervalo de confianza (A, B) del 95 %, buscaremos valores A, B de manera que

$$P(\theta < A) = 0{,}025 \quad \text{y} \quad P(\theta > B) = 0{,}025.$$

3.1. Intervalos de confianza para el parámetro μ de una población normal

3.1.1. Intervalos de confianza para el parámetro μ de una población normal con σ conocida

Sea X una v.a. normal con media poblacional μ desconocida y desviación típica poblacional σ conocida (a la práctica, usualmente, estimada en un experimento anterior)

Sea $X_1, \dots, X_n$ una m.a.s. de X, con media muestral $\overline{X}$

Queremos determinar un intervalo de confianza para a μ con un cierto nivel de confianza (digamos, 97,5 %, $\alpha = 0{,}025$): un intervalo (A, B) tal que

$$P(A < \mu < B) = 1 - \alpha = 0{,}975$$

Bajo estas condiciones, sabemos que

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

sigue una distribución normal estándar.

Comencemos calculando un intervalo centrado en 0 en el que Z tenga probabilidad 0,975:

$$0{,}975 = P(-\delta < Z < \delta) = F_Z(\delta) - F_Z(-\delta) = 2F_Z(\delta) - 1$$
$$F_Z(\delta) = \frac{1{,}975}{2} = 0{,}9875 \Rightarrow \delta = \text{qnorm}(0{,}9875) = 2{,}2414.$$

Por lo tanto

$$P(-2{,}2414 < Z < 2{,}2414) = 0{,}975$$

Substituyendo $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$:

$$P\left(-2{,}2414 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 2{,}2414\right) = 0{,}975$$
$$P\left(\overline{X} - 2{,}2414\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 2{,}2414\frac{\sigma}{\sqrt{n}}\right) = 0{,}975$$

Por lo tanto, la probabilidad que la media poblacional μ de X se encuentre dentro del intervalo

$$\left(\overline{X} - 2{,}2414\frac{\sigma}{\sqrt{n}}, \overline{X} + 2{,}2414\frac{\sigma}{\sqrt{n}}\right)$$

es 0,975: es un intervalo de confianza del 97,5 %.

Además tenemos que está centrado en $\overline{X}$: el 0.025 de probabilidad restante está repartido por igual en los dos extremos del intervalo.

Como estimador: un 97,5 % de las veces que tomemos una muestra de tamaño n de X, el verdadero valor de μ caerá dentro de este intervalo

Para una muestra concreta: la probabilidad de que, si una media μ poblacional ha producido esta muestra, entonces esté en este intervalo concreto, es del 97,5 %.

En ocasiones lo entenderemos como: *La probabilidad de que μ esté en este intervalo es del* 97,5 %.

Pero la frase anterior es mentira (*es un abuso de lenguaje*): La μ concreta es un valor fijo, por lo tanto que pertenezca o no a este intervalo concreto tiene probabilidad 1 (si pertenece) y 0 (si no pertenece).

Teorema. Sea $X \sim N(\mu, \sigma)$ con μ desconocida y σ conocida.

Tomamos una m.a.s. de X de tamaño n, con media $\overline{X}$.

Un intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ para μ es

$$\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

donde $z_{1-\frac{\alpha}{2}}$ es el $(1-\frac{\alpha}{2})$-cuantil de la normal estándar Z (es decir, $z_{1-\frac{\alpha}{2}} = F_Z^{-1}(1-\frac{\alpha}{2})$, o $P(Z \leq z_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$).

Si X es normal con σ conocida, un intervalo de confianza I.C. para μ de población normal con σ conocida μ del $(1-\alpha)\cdot 100\,\%$ es

$$\overline{X} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} := \left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.\right)$$

Observad que está centrado en $\overline{X}$.

La tabla siguiente muestra los cuantiles más usados:

Confianza $1-\alpha$	Significación α	Cuantil $z_{1-\frac{\alpha}{2}}$
0.90	0.1	1.645
0.95	0.05	1.96
0.975	0.025	2.241

3.1.2. Simulación de intervalos de confianza

Ejemplo

A continuación vamos a simular el funcionamiento de un intervalo de confianza para una distribución normal.

Consideramos una población de 10^6 valores de una distribución normal de parámetros $\mu = 1{,}5$ y $\sigma = 1$:

```
set.seed(1012)
mu=1.5; sigma=1; alpha=0.05
```

```
Población=rnorm(10^6,mu,sigma)
```

Vamos a generar 100 muestras aleatorias simples de tamaño 50 de dicha población para posteriormente generar un intervalo de confianza al 95 % de confianza para el parámetro μ para cada muestra.

Primero definiremos una función para que, dada una muestra, un valor σ y un nivel de significación α nos genere el intervalo correspondiente para el parámetro μ al nivel de confianza $100 \cdot (1-\alpha)\,\%$. La llamaremos `ICZ`:

```
ICZ=function(x,sigma,alpha){
  c(mean(x)-qnorm(1-alpha/2)*sigma/sqrt(length(x)),
  mean(x)+qnorm(1-alpha/2)*sigma/sqrt(length(x)))}
```

Usando la función `replicate` de `R` generamos las muestras y los intervalos de confianza correspondientes usando la función anterior:

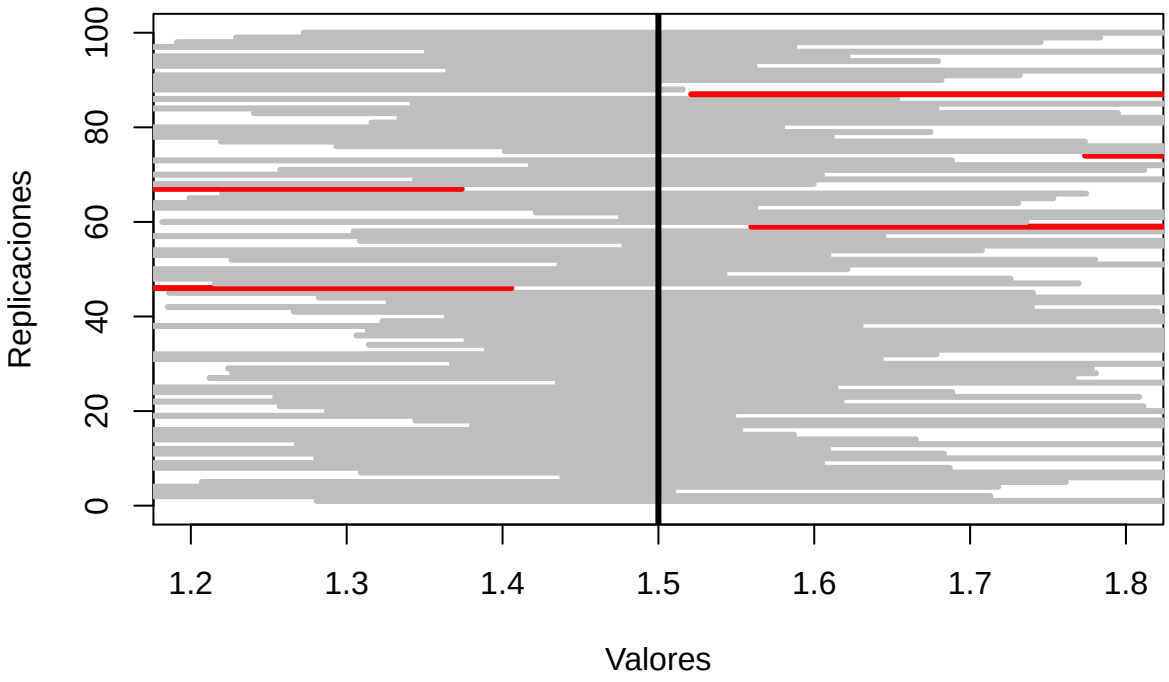
```
set.seed(2)
M=replicate(100,ICZ(sample(Población,50,replace=T),
 sigma,alpha))
```

El objeto M de `R` es una matriz de 2 filas y 100 columnas donde la columna i-ésima representa el intervalo de confianza para la muestra i-ésima generada.

Finalmente vamos a dibujar todos los intervalos anteriores y resaltaremos en color rojo aquéllos en los que el parámetro $\mu = 1{,}5$ no esté en ellos. Esperamos que haya aproximadamente 5 en los que esta condición falle.

El resultado se muestra en la figura siguiente:

```
plot(1:10,type="n",xlim=c(1.2,1.8),ylim=c(0,100),
xlab="Valores",ylab="Replicaciones")
seg.int=function(i){color="grey";
  if((mu<M[1,i]) | (mu>M[2,i])){color = "red"}
  segments(M[1,i],i,M[2,i],i,col=color,lwd=3)}
invisible(sapply(1:100,FUN=seg.int))
abline(v=mu,lwd=3)
```



3.1.3. Interpretación del intervalo de confiaza.

¡Atención! De media, un $\alpha \cdot 100\,\%$ de las veces, un intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ no contendrá el valor real del parámetro.

Por ejemplo, de media, un $5\,\%$ de las veces un intervalo de confianza del $95\,\%$ no contendrá el valor real del parámetro.

Ejercicio

Tomamos una m.a.s. de tamaño $n = 16$ de una v.a. normal con $\sigma = 4$ y μ desconocida. La media de la m.a.s. es $\overline{x} = 20$.

Calculad un intervalo de confianza del $97{,}5\,\%$ para μ de una población normal con σ conocida.

El valor de α será $\alpha = 1 - 0{,}975 = 0{,}025$. El intervalo de confianza será:

$$\begin{aligned}&\left(\overline{x} - z_{1-\frac{\alpha}{2}} \cdot \tfrac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}} \cdot \tfrac{\sigma}{\sqrt{n}}\right)\\ &= \left(20 - 2{,}241 \cdot \tfrac{4}{\sqrt{16}}, 20 + 2{,}241 \cdot \tfrac{4}{\sqrt{16}}\right)\\ &= (17{,}759, 22{,}241).\end{aligned}$$

La probabilidad de que, si una media μ poblacional ha producido esta muestra, entonces esté en este intervalo concreto, es del $97{,}5\,\%$.

Ejemplo del procesador Intel Core

Queremos analizar un sensor que mide la temperatura de un procesador en grados centígrados, en concreto un Intel Core i7-2600K, que tiene como temperatura normal de 32° a 40°. Para saber si está

bien calibrado, diseñamos un experimento en el que ponemos el procesador en las mismas condiciones y tomamos una muestra de 40 valores de su temperatura.

Los resultados son los siguientes:

```
temperatura=c(36,35,38,38,36,37,38,36,37,36,
              37,37,34,38,35,37,36,36,34,38,
              36,37,35,35,35,35,36,36,36,35,
              36,35,34,34,37,37,35,36,34,36)
```

Supongamos que las medidas de nuestro sensor siguen una distribución normal con varianza poblacional conocida $\sigma^2 = 1{,}44$. Calculad un intervalo de confianza del 90 % para al resultado medio de la temperatura del procesador.

Tenemos las siguientes condiciones:

- Población normal con $\sigma = \sqrt{1{,}44} = 1{,}2$ conocida.
- Muestra aleatoria simple de tamaño $n = 40$.
- Media de la muestra $\overline{x} = 35{,}975$.
- $1 - \alpha = 0{,}9 \Rightarrow \alpha = 0{,}1 \Rightarrow 1 - \frac{\alpha}{2} = 0{,}95$.
- $z_{0{,}95} \approx 1{,}645$.

Aplicamos la fórmula $\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ con $\overline{x} = 35{,}975, z_{0{,}95} = 1{,}645, \sigma = 1{,}2, n = 40$.

Obtenemos que el intervalo de confianza del 90 % es

$$35{,}975 \pm 1{,}645 \cdot \frac{1{,}2}{\sqrt{40}} = (35{,}663, 36{,}287).$$

Amplitud del intervalo de confianza. La amplitud A del intervalo de confianza a un nivell $100 \cdot (1-\alpha)\,\%$ de confianza será:

$$A = \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} - \left(\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 2 \cdot z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Error máximo cometido. El error máximo, al nivel de confianza $100 \cdot (1 - \alpha)\,\%$, que cometemos al estimar μ por $\overline{X}$ es la mitad de la amplitud,

$$z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Propiedades

- La **amplitud** del intervalo de confianza para el parámetro μ de una población normal con σ conocida, n y α fijos **crece**, si σ **crece**.
- La **amplitud** del intervalo de confianza para el parámetro μ de una población normal con σ conocida y α fijo **decrece**, si n **crece**.
- La **amplitud** del intervalo de confianza para el parámetro μ de una población normal con σ conocida y n fijo **crece**, si $1 - \alpha$ **crece**, o si α **decrece**.

Problema: hallar el tamaño n mínimo de la muestra para asegurar que el intervalo de confianza para μ al nivel de confianza $(1 - \alpha)$ tenga una amplitud prefijada máxima A_0 (o un error máximo $A_0/2$).

Resolución: usando que la amplitud máxima tiene que ser A_0 tenemos que se tiene que verificar $A_0 \geq 2z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

Despejando n de la expresión anterior, tendremos que el tamaño de la muestra mínimo será:

$$n \geq \left(2z_{1-\frac{\alpha}{2}}\frac{\sigma}{A_0}\right)^2.$$

Ejemplo

Recordemos que las medidas de nuestro sensor de temperatura seguían una distribución normal con varianza poblacional conocida $\sigma^2 = 1{,}44$, $\sigma = 1{,}2$.

¿Cuántas medidas tendríamos que tomar para obtener la temperatura media con un error máximo de $0{,}05^\circ$ al nivel de confianza del 90 %?

Nos dicen que $\frac{A_0}{2} = 0{,}05$, o sea, $A_0 = 0{,}1$. Usando la expresión anterior, tendremos que el número de medidas mínimo n que tendríamos que tomar será:

$$n = \left\lceil \left(2z_{1-\frac{\alpha}{2}}\frac{\sigma}{A_0}\right)^2 \right\rceil$$

donde $z_{1-\frac{\alpha}{2}} = 1{,}645,\ \sigma = 1{,}2$. Obtenemos

$$n = \left\lceil \left(2 \cdot 1{,}645 \cdot \frac{1{,}2}{0{,}1}\right)^2 \right\rceil = \lceil 1558{,}393 \rceil = 1559.$$

3.1.4. Intervalos de confianza para el parámetro μ de una población normal con σ desconocida

Recordemos que para hallar el intervalo de confianza para el parámetro μ de una población normal, era clave la variable aleatoria $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$.

El problema es que ahora no la podemos usar al no conocer σ.

Lo que haremos será sustituir la desviación típica poblacional σ por la desviación típica muestral $\widetilde{S}_X$ y nos quedará: $\frac{\overline{X}-\mu}{\widetilde{S}_X/\sqrt{n}}$, donde $\overline{X}$ es la media muestral y n, el tamaño de la muestra.

La distribución de la variable anterior $\frac{\overline{X}-\mu}{\widetilde{S}_X/\sqrt{n}}$, no será normal sino t de Student con $n-1$ grados de libertad como nos dice el teorema siguiente:

Teorema. Sea $X \sim N(\mu, \sigma)$. Sea $X_1, \ldots, X_n$ una m.a.s. de X, con media $\overline{X}$ y desviación típica muestral $\widetilde{S}_X$

En estas condiciones, la v.a. $t = \frac{\overline{X}-\mu}{\widetilde{S}_X/\sqrt{n}}$, sigue una distribución t de Student con $n-1$ grados de libertad, t_{n-1}.

Distribución t de Student

La distribución t de Student con ν grados de libertad, t_ν tiene como función de densidad

$$f_{t_\nu}(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\Big(1+\frac{x^2}{\nu}\Big)^{-\frac{\nu+1}{2}},$$

donde $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$ si $x > 0$.

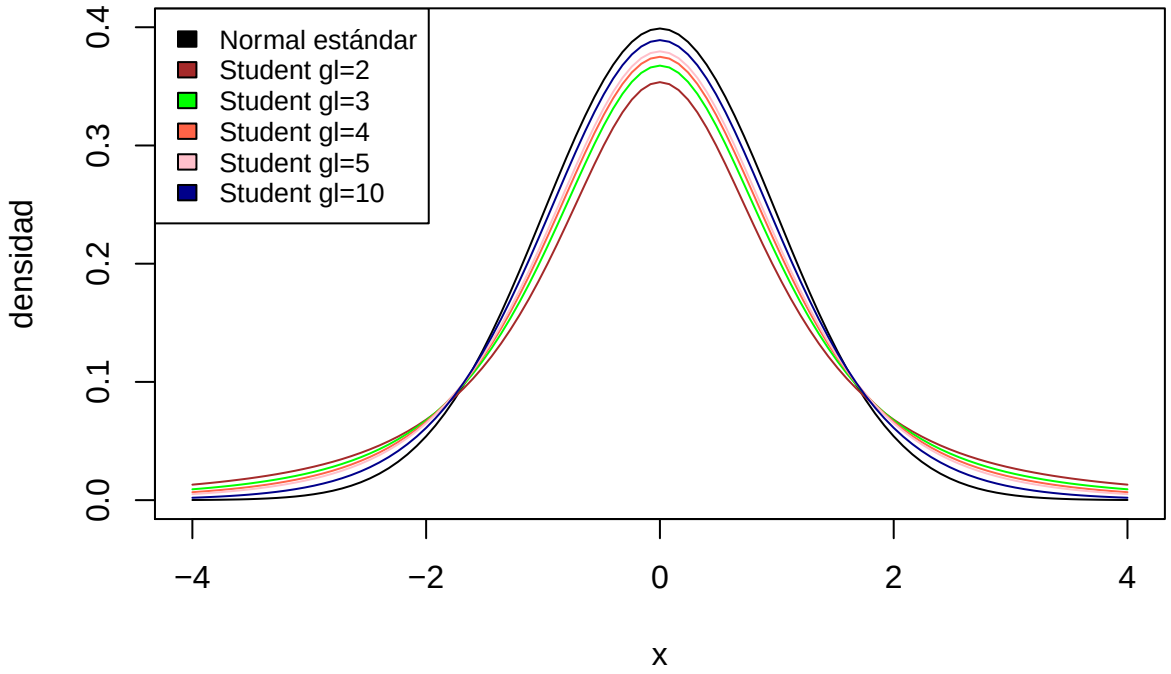
Propiedades

- $E(t_\nu) = 0$ si $\nu > 1$ y $Var(t_\nu) = \dfrac{\nu}{\nu-2}$ si $\nu > 2$.
- Su función de distribución es simétrica respecto de $E(t_\nu) = 0$ (como la de una $N(0,1)$):

$$P(t_\nu \le -x) = P(t_\nu \ge x) = 1 - P(t_\nu \le x).$$

- Si ν es grande, su distribución es aproximadamente la de $N(0,1)$ (pero con más varianza: un poco más aplastada)

Gráficas de las densidades de diferentes distribuciones t de Student junto con la densidad de la $N(0,1)$:



Indicaremos con $t_{\nu,q}$ el q-cuantil de una v.a. X que sigue una distribución t_ν:

$$P(X \le t_{\nu,q}) = q$$

Por simetría, $t_{\nu,q} = -t_{\nu,1-q}$.

Consideremos la situación siguiente:

- X una v.a. normal con μ y σ desconocidas.
- $X_1, \ldots, X_n$ una m.a.s. de X de tamaño n, con media $\overline{X}$ y varianza muestral $\widetilde{S}^2_X$.

Intervalo de confianza para el parámetro μ. En estas condiciones, un intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ para el parámetro μ de una población normal con σ desconocida es

$$\left(\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right).$$

Ejemplo 3D-print

La empresa *3D-print* ofrece una impresora industrial de papel en color de alta capacidad. En su publicidad afirma que sus cartuchos imprimen una media de 500 mil copias con la especificación:

'Ficha técnica: Muestra de tamaño n=100, población aproximadamente normal, nivel de confianza del 90 %.

La OCU (asociación de consumidores) desea comprobar estas afirmaciones y su laboratorio toma una muestra aleatoria de tamaño $n = 24$, obteniendo una media de $\overline{x} = 518$ mil impresiones y una desviación típica muestral $\tilde{s} = 40$ mil.

Con esta muestra ¿la media poblacional anunciada por fabricante cae en el intervalo de confianza del 90 %?

Hay que calcular el intervalo de confianza para la μ de una población normal con σ desconocida μ con los valores siguientes: $n = 24, \overline{x} = 518, \tilde{s} = 40, \alpha = 0{,}1$.

Dicho intervalo será el siguiente:

$$\begin{array}{l}\left(\overline{x} - t_{23,0,95}\frac{\tilde{s}}{\sqrt{n}}, \overline{x} + t_{23,0,95}\frac{\tilde{s}}{\sqrt{n}}\right)\\ = \left(518{,}000 - 1{,}714\frac{40{,}000}{\sqrt{24}}, 518{,}000 + 1{,}714\frac{40{,}000}{\sqrt{24}}\right)\\ = \left(5{,}04006298\times 10^5, \quad 5{,}31993702\times 10^5\right)\end{array}.$$

Observamos que no contiene a 500.000 (¡pero se equivoca a favor del consumidor!)

Observaciones:

- El intervalo de confianza obtenido está centrado en $\overline{X}$
- La fórmula $\left(\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right)$ nos da el intervalo de confianza para μ en una población normal con σ desconocida. La expresión anterior se puede utilizar cuando X es normal y n cualquiera
- Si n es grande $t_{n-1,1-\frac{\alpha}{2}} \approx z_{1-\frac{\alpha}{2}}$ y podemos **aproximar** el intervalo de confianza anterior mediante la expresión siguiente:

$$\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right)$$

Consideremos la situación siguiente:

- X una v.a. **cualquiera** con media poblacional μ desconocida y desviación típica σ desconocida.
- $X_1, \ldots, X_n$ una m.a.s. de X, con media $\overline{X}$.
- n **es grande** (pongamos que $n \geq 40$)

3.1.5. Intervalos de confianza para el parámetro μ de una población cualquiera con σ conocida y tamaño muestral grande

En estas condiciones, usando el **Teorema Central del Límite** $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1)$.

Teorema. En las condiciones anteriores, podemos tomar como intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ de confianza para el parámetro μ de una población cualquiera con σ conocida la expresión siguiente:

$$\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Si la σ es desconocida, podemos aplicar el Teorema anterior sustituyendo σ por $\widetilde{S}_X$:

Teorema. En las condiciones anteriores, podemos tomar como intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ de confianza para el parámetro μ de una población cualquiera con σ desconocida la expresión siguiente:

$$\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right)$$

Ejercicio

Se ha tomado una muestra del tiempo de visualización de vídeo semanal en horas de 1000 usuarios de un canal de videos por internet. Se ha obtenido una media muestral de 9.5 horas/semana con una desviació típica muestral de 0.5 horas/semana.

Calculad un intervalo de confianza del 95 % para la media poblacional del número de horas visualizadas por semana supuesto que sigue aproximadamente una población normal con σ desconocida.

Como $n = 1000$ es grande, podemos utilizar la expresión siguiente para hallar el intervalo de confianza:

$$\left(\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\tilde{s}}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\tilde{s}}{\sqrt{n}}\right),$$

donde $\overline{x} = 9{,}5,\ \tilde{s} = 0{,}5,\ \alpha = 0{,}05,\ z_{1-\frac{\alpha}{2}} = 1{,}96$. El intervalo de confianza será, pues:

```
media=9.5
sd=0.5
n=1000
alpha=0.5
cuantil=qnorm(1-alpha/2)
(extremo.izquierdo=round(media-cuantil*sd/sqrt(n),3))
```

```
## [1] 9.489
```

```
(extremo.derecho=round(media+cuantil*sd/sqrt(n),3))
```

```
## [1] 9.511
```

Podemos afirmar con un 95 % de confianza que la media poblacional de vídeo consumido en horas por semana está entre 9,489 y 9,511 horas/semana.

Amplitud del intervalo de confianza.

La amplitud de $\left(\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\widetilde{S}_X}{\sqrt{n}}\right)$ es $A = 2z_{1-\frac{\alpha}{2}} \frac{\widetilde{S}_X}{\sqrt{n}}$.

Para determinar n (grande) que dé cómo máximo una amplitud A prefijada, necesitamos $\widetilde{S}_X$, que depende de la muestra.

Soluciones:

- Si sabemos la desviación típica poblacional σ, la utilizaremos en lugar de $\widetilde{S}_X$.
- Si hemos tomado una muestra previa (piloto), emplearemos la desviación típica muestral de esta **muestra piloto** para estimar σ.

De una población X hemos tomado una m.a.s. (piloto) que ha tenido una desviación típica muestral $\tilde{s}_{piloto}$.

Estimaremos que el tamaño mínimo n de una m.a.s. de X que dé un intervalo de confianza I.C. para μ de una población normal con σ desconocida de nivell de confianza $1-\alpha$ y amplitud máxima A_0 es

$$n = \left\lceil \left(2z_{1-\frac{\alpha}{2}} \frac{\tilde{s}_{piloto}}{A_0}\right)^2 \right\rceil$$

Ejercicio

Queremos estimar la estatura media de los estudiantes de la UIB. Queremos obtener un intervalo de confianza del 99 % con una precisión máxima de 1 cm. En una muestra piloto de 25 estudiantes, obtuvimos que

$$\overline{x} = 170 \text{ cm}, \tilde{s} = 10 \text{ cm}.$$

¿Basándonos en estos datos, cuál es el tamaño necesario de la muestra para poder alcanzar nuestro objetivo?

Con una precisión de 1 cm. tendremos que la amplitud máxima será el doble: 2 cm.

Por tanto el tamaño necesario de la muestra será:

```
alpha=0.01
cuantil = qnorm(1-alpha/2)
amplitud = 2
s.piloto = 10
(n.minimo = ceiling((2*cuantil*s.piloto/amplitud)^2))
```

```
## [1] 664
```

Si queremos, podemos estudiar estos conceptos directamente utilizando la función `t.test` de `R`:

```
t.test(X,conf.level=...)$conf.int
```

donde `conf.level` es el nivel de confianza $1-\alpha$ en tanto por uno. Su valor por defecto es $1-\alpha = 0{,}95$.

Ejemplo tabla de datos `iris`

Hallemos un intervalo de confianza para la media de la longitud del pétalo para una muestra de 30 flores de la tabla de datos **iris**.

- En primer lugar elegimos las flores de la muestra:

```
set.seed(1000)
muestra.iris = sample(1:150,30,replace=TRUE)
```

A continuación calculamos las longitudes del pétalo de las flores de nuestra muestra:

```
long.pétalo.muestra = iris[muestra.iris,]$Petal.Length
```

Un intervalo de confianza al 95 % de confianza para las longitudes del pétalo sería:

```
t.test(long.pétalo.muestra,conf.level=0.95)$conf.int
```

```
## [1] 2.9865374 4.1067959
## attr(,"conf.level")
## [1] 0.95
```

3.1.6. Experimento sobre la "confianza"

Experimento

Vamos a comprobar con un experimento qué papel juega la "confianza" en los intervalos de confianza.

Vamos a generar al azar una Población de 10000000 (10^7) "individuos" con distribución normal estándard. Vamos a tomar 200 muestras aleatorias simples de tamaño 50 de esta población y calcularemos el intervalo de confianza para la media poblacional usando dicha fórmula.

Finalmente, contaremos cuántos de estos intervalos de confianza contienen la media de la población. Fijaremos la semilla de aleatoriedad para que el experimento sea reproducible y podáis comprobar que no hacemos trampa.

En primer lugar, generamos la población de valores:

```
set.seed(2020)
valores.población=rnorm(10^7)
```

Seguidamente, hallamos la media poblacional:

```
(mu=mean(valores.población))
```

```
## [1] -2.9033996e-06
```

Para hallar 200 muestras, usaremos la función `replicate` de `R` que nos permite ejecutar una misma función las veces que le indiquemos:

```
muestras=replicate(200, sample(valores.población,50,replace=TRUE))
```

De esta forma `muestras` es una matriz de 50 filas y 200 columnas donde cada fila representa una muestra.

A continuación, Vamos a aplicar a cada una de estas muestras la función `t.test` para calcular un intervalo de confianza del 95 % y luego contaremos los aciertos, es decir, cuántos de ellos contienen la media poblacional.

Primero definimos la función `IC.t` que nos da el intervalo de confianza para la media dada una muestra `X`:

```
IC.t= function(X, confianza=0.95){
  t.test(X, conf.level=confianza)$conf.int
  }
```

En segundo lugar, calculamos los 200 intervalos de confianza para nuestras 200 muestras usando la función `apply` de `R`:

```
ICs= apply(muestras,FUN=IC.t,MARGIN=2)
```

En tercer lugar, miramos cuántos de los intervalos anteriores contienen la media poblacional `mu`:

```
Aciertos=length(which((mu>=ICs[1,]) & (mu<=ICs[2,])))
Aciertos
```

```
## [1] 195
```

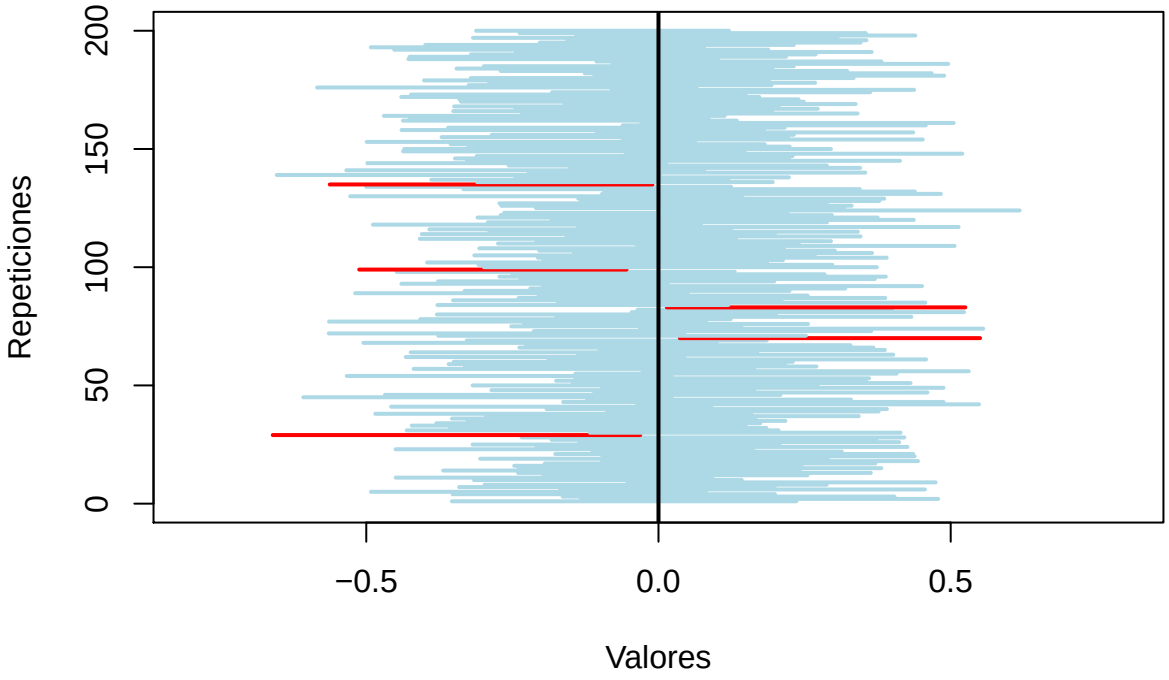
Hemos acertado 195 veces, o sea, el 97.5 % de las veces. Es una buena aproximación del valor 95 %, que era el esperado.

Para visualizar mejor los aciertos, vamos a dibujar los intervalos apilados en un gráfico, donde aparecerán en azul claro los que aciertan y en rojo los que no aciertan.

```
plot(1,type="n",xlim=c(-0.8,0.8),ylim=c(0,200),xlab="Valores",
     ylab="Repeticiones",main="")

seg.int=function(i){
  color="light blue"
  if((mu<ICs[1,i]) | (mu>ICs[2,i])){
    color = "red"
  }
  segments(ICs[1,i],i,ICs[2,i],i,col=color,lwd=2)
}

sapply(1:200,FUN=seg.int)
abline(v=mu,lwd=2)
```



3.2. Intervalos de confianza para el parámetro p de una población de Bernoulli

3.2.1. Método "exacto" o de Clopper-Pearson

Consideremos la situación siguiente:

- X una v.a. Bernoulli con p desconocido.
- $X_1, \ldots, X_n$ una m.a.s. de X, con número de éxitos x y por lo tanto la frecuencia relativa de éxitos es $\hat{p}_X = x/n$.

En este caso, la distribución de la variable Y ="número de éxitos en la muestra" es binomial de parámetros n y p, Y es $B(n,p)$

Definición. Un intervalo de confianza (p_0, p_1) del $(1-\alpha)100\,\%$ nivel de confianza para p de una población X de Bernoulli se obtiene encontrando el p_0 más grande y el p_1 más pequeño tales que

$$\sum_{k=x}^{n} \binom{n}{k} \cdot p_0^k \cdot (1-p_0)^{n-k} \leq \frac{\alpha}{2}, \quad \sum_{k=0}^{x} \binom{n}{k} \cdot p_1^k \cdot (1-p_1)^{n-k} \leq \frac{\alpha}{2}$$

Para hallar un intervalo de confianza para la proporción poblacional en `R` según el método de Clopper-Pearson, hay que usar la función `binom.exact` del paquete `epitools`:

```
binom.exact(x,n,conf.level)
```

donde `x` y `n` representan, respectivamente, el número de éxitos y el tamaño de la muestra, y `conf.level` es $1-\alpha$, el nivel de confianza en tanto por uno.

Ejemplo tabla de datos `iris`

Hallemos un intervalo de confianza para la proporción de flores con especie "setosa" dada una muestra de 60 flores.

Sabemos que la proporción real p en este caso vale $p = \frac{50}{150} = \frac{1}{3} = 0{,}333$.

Primero hallamos la muestra de las 60 flores:

```
set.seed(1000)
flores.elegidas = sample(1:150,60,replace=TRUE)
```

Las flores elegidas son: (sólo mostramos las 10 primeras)

```
muestra.flores.prop = iris[flores.elegidas,]
head(muestra.flores.prop,10)
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 68           5.8         2.7          4.1         1.0 versicolor
## 43           4.4         3.2          1.3         0.2     setosa
## 51           7.0         3.2          4.7         1.4 versicolor
## 88           6.3         2.3          4.4         1.3 versicolor
## 29           5.2         3.4          1.4         0.2     setosa
## 99           5.1         2.5          3.0         1.1 versicolor
## 61           5.0         2.0          3.5         1.0 versicolor
## 146          6.7         3.0          5.2         2.3  virginica
## 150          5.9         3.0          5.1         1.8  virginica
## 102          5.8         2.7          5.1         1.9  virginica
```

El número de flores de especie setosa será:

```
número.flores.setosa=
  table(muestra.flores.prop$Species=="setosa")[2]
número.flores.setosa
```

```
## TRUE
##   21
```

El intervalo de confianza para la proproción poblacional de flores de especie setosa al 95 % de confianza será:

```
library(epitools)
binom.exact(número.flores.setosa,60,conf.level=0.95)
```

```
##       x  n proportion      lower      upper conf.level
## TRUE 21 60       0.35 0.23132642 0.48402801       0.95
```

Según el método de Clopper-Pearson, con un 95 % de confianza podemos decir que en la tabla de datos **iris** hay entre un 23.13 % y 48.4 % de flores de especie "setosa".

3.2.2. Caso del tamaño n de la muestra grande

Consideremos la situación siguiente :

- X una v.a. Bernoulli con p desconocida.
- $X_1, \dots, X_n$ una m.a.s. de X, con n grande (por Ejemplo, $n \geq 40$) y frecuencia relativa de éxitos $\hat{p}_X$.

En estas condiciones (por el Teorema Central del Límite),

$$Z = \frac{\hat{p}_X - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$$

Por lo tanto

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{p}_X - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

El problema es que no conocemos p...

La literatura plantea entre otras soluciones:

- El método de Wilson
- La solución de Laplace (1812)

3.2.3. Método de Wilson

Definición. En estas condiciones, un intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ I.C. para p (donde $\hat{q}_X = 1 - \hat{p}_X$) es:

$$\left(\frac{\hat{p}_X + \frac{z^2_{1-\alpha/2}}{2n} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_X\hat{q}_X}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1+\frac{z^2_{1-\alpha/2}}{n}}, \frac{\hat{p}_X + \frac{z^2_{1-\alpha/2}}{2n} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_X\hat{q}_X}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1+\frac{z^2_{1-\alpha/2}}{n}}\right).$$

Para hallar un intervalo de confianza para la proporción poblacional en `R` según el método de Wilson, hay que usar la función `binom.wilson` del mismo paquete `epitools`:

```
binom.wilson(x,n,conf.level)
```

Ejemplo tabla de datos `iris`

Usando el ejemplo anterior, hallemos un intervalo de confianza para la proporción de flores de especie "setosa" según el método de Wilson al 95 % de confianza.

El intervalo será:

```
binom.wilson(número.flores.setosa,60,conf.level=0.95)
```

```
##       x  n proportion      lower      upper conf.level
## TRUE 21 60       0.35 0.24167774 0.47637381       0.95
```

Según el método de Wilson, con un 95 % de confianza podemos decir que en la tabla de datos **iris** hay entre un 24.17 % y 47.64 % de flores de especie "setosa".

3.2.4. Fórmula de Laplace.

Supongamos que la muestra aleatoria simple es considerablemente más grande que la usada en el método de **Wilson** y que, además, la proporción muestral de éxitos $\hat{p}_X$ está alejada de 0 y de 1.

- O sea, $n \geq 100$ y que $n\hat{p}_X \geq 10$ y $n(1-\hat{p}_X) \geq 10$.
- En este caso, podemos usar la fórmula de **Laplace**:

$$\hat{p}_X \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}.$$

Para hallar un intervalo de confianza para la proporción poblacional en `R` según la fórmula de Laplace, hay que usar la función `binom.approx` del mismo paquete `epitools`:

```
binom.approx(x,n,conf.level)
```

Ejemplo

En una muestra aleatoria de 500 familias con niños en edad escolar se encontró que 340 introducían fruta de forma diaria en la dieta de sus hijos

Calculad un intervalo de confianza del 95 % para conocida la proporción real de familias de esta ciudad con niños en edad escolar que incorporen fruta fresca de forma diaria en la dieta de sus hijos.

El tamaño de la muestra es $n = 500$ y la estimación de la proporción muestral, $\hat{p}_X = \dfrac{340}{500} = 0{,}68$.

Como que $n = 500 \geq 100$, $n\hat{p}_X = 340 \geq 10$ y $n \cdot (1-\hat{p}_X) = 160 \geq 10$, podemos utilizar la fórmula de Laplace.

Usando que $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$, el intervalo de confianza será:

$$\begin{aligned}&\left(0{,}68 - 1{,}96 \cdot \sqrt{\tfrac{0{,}68 \cdot 0{,}32}{500}}, 0{,}68 + 1{,}96 \cdot \sqrt{\tfrac{0{,}68 \cdot 0{,}32}{500}}\right)\\ &= (0{,}639, 0{,}721).\end{aligned}$$

Ejemplo tabla de datos `iris`

Usando el ejemplo anterior, hallemos un intervalo de confianza para la proporción de flores de especie "setosa" según la fórmula de Laplace al 95 % de confianza.

El intervalo será:

```
binom.approx(número.flores.setosa,60,conf.level=0.95)
```

```
##       x  n proportion      lower      upper conf.level
## TRUE 21 60       0.35 0.22931226 0.47068774       0.95
```

Según la fórmula de Laplace, con un 95 % de confianza podemos decir que en la tabla de datos **iris** hay entre un 22.93 % y 47.07 % de flores de especie "setosa".

Amplitud del intervalo de confianza.

Problema: hallar el tamaño de la muestra fijada la amplitud del intervalo de confianza.

La **amplitud** del intervalo de confianza usando la fórmulade Laplace es

$$A = 2z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}.$$

No podemos determinar el tamaño de la muestra para que el intervalo de confianza tenga como máximo una cierta amplitud sin conocer $\hat{p}_X$.

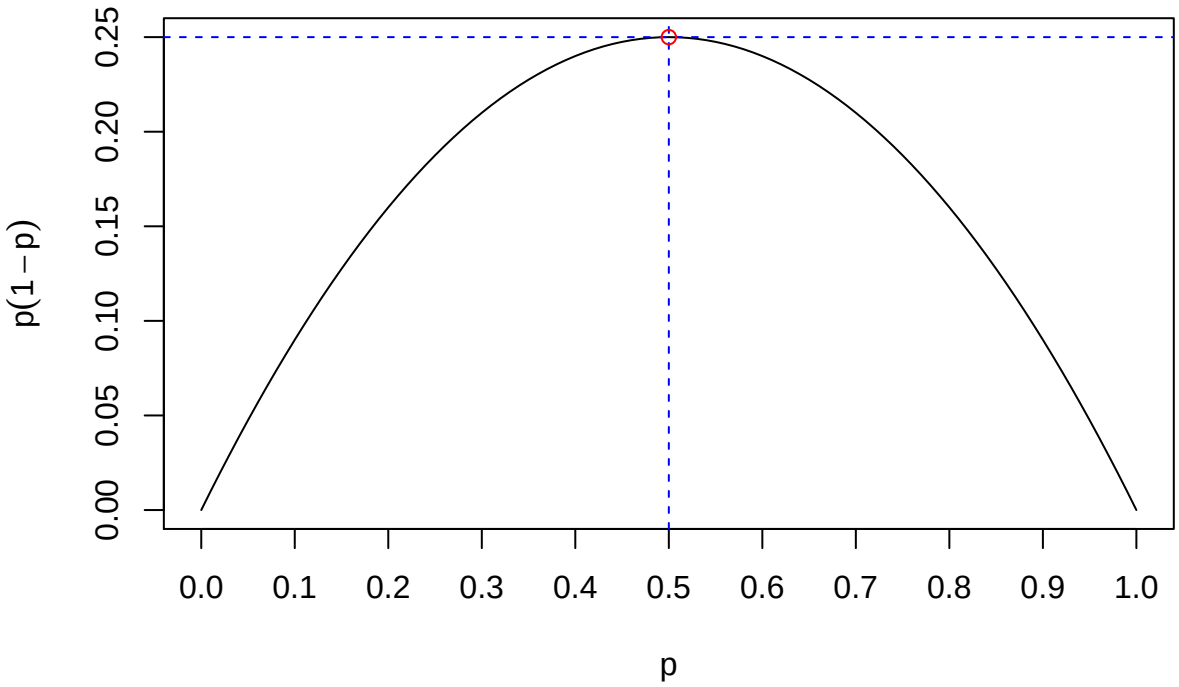
Vamos a considerar que estamos en el peor de los casos.

O sea, usando que $\hat{p}_X \in [0,1]$, nos planteamos hallar el máximo de la expresión $\hat{p}_X(1-\hat{p}_X)$ que aparece en la fórmula de la amplitud.

El máximo de la función anterior, para $\hat{p}_X \in [0,1]$ se alcanza en $\hat{p}_X = \frac{1}{2}$ y dicho máximo vale $\frac{1}{4}$:

Ejercicio

Demostrar que el máximo de la función $f(p) = p(1-p)$ se alcanza en $p = 1/2$ y vale $1/4$.



En resumen, calcularemos n para obtener una amplitud máxima A_0 suponiendo el peor de los casos ($\hat{p}_X = 0{,}5$):

$$A_0 \geq 2z_{1-\frac{\alpha}{2}}\sqrt{\frac{0{,}5^2}{n}} = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \Rightarrow n \geq \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}}{A_0^2} \right\rceil.$$

Ejemplo de los teléfonos móviles

Quemos estudiar qué fración teléfonos móviles utilizan android para determinar esta proporción con un nivel de confianza del 95 % y garantizar un error máximo de 0.05.

¿De qué tamaño ha de ser la muestra **en el peor de los casos**?

Usando la fórmula anterior, el valor de n tiene que verificar:

$$n \geq \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}}{A^2} \right\rceil$$

donde $\frac{A}{2} = 0{,}05$, $(A = 0{,}1)$ y $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$.

El tamaño de la muestra valdrá, como mínimo:

$$n \geq \left\lceil \frac{1{,}96^2}{0{,}1^2} \right\rceil = \lceil 384{,}146 \rceil = 385.$$

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

Consideremos la siguiente situación:

- Consideramos una X una v.a. normal con μ y σ desconocidas.
- Sea $X_1, \ldots, X_n$ una m.a.s. de X y varianza muestral $\widetilde{S}^2_X$.

En estas condiciones tenemos el siguiente:

Teorema.

La variable aleatoria $\frac{(n-1)\tilde{S}^2_X}{\sigma^2}$ se distribuye según una distribución χ^2_{n-1}.

Teorema.

En las condiciones anteriores, un intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ para la varianza σ^2 de la población X es

$$\left(\frac{(n-1)\widetilde{S}^2_X}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\widetilde{S}^2_X}{\chi^2_{n-1,\frac{\alpha}{2}}} \right),$$

donde $\chi^2_{\nu,q}$ es el q-cuantil de la distribución χ^2_ν.

Ejemplo

Un algoritmo probabilístico depende de la semilla de aleatorización que se genera en cada paso. Para saber si la semilla influye mucho en el resultado se ejecuta el algoritmo varias veces hasta obtener un resultado similar y se estudia la varianza de su tiempo de ejecución.

Queremos ver si la desviación típica σ es ≤ 30.

Se supone que la distribución del tiempo de ejecución del algoritmo es aproximadamente normal.

Se realizan 30 ejecuciones del algoritmo de las que se mide el tiempo de ejecución. Los resultados son:

```
tiempo=c(12, 13, 13, 14, 14, 14, 15, 15, 16, 17,
         17, 18, 18, 19, 19, 25, 25, 26, 27, 30,
         33, 34, 35, 40, 40, 51, 51, 58, 59, 83)
```

Nos piden calcular un intervalo de confianza para σ^2 del tiempo de ejecución a un 95 % de confianza.

En primer lugar calculamos la varianza muestral de nuestra muestra:

```
n=30
(var.muestral.tiempo = var(tiempo))
```

```
## [1] 301.55057
```

En segundo lugar, calculamos los cuantiles que necesitamos:

```
alpha=0.05
(cuantil.izquierda = qchisq(1-alpha/2,n-1))
```

```
## [1] 45.722286
```

```
(cuantil.derecha = qchisq(alpha/2,n-1))
```

```
## [1] 16.047072
```

El intervalo de confianza para la varianza del tiempo de ejecución será:

```
(valor.izquierdo= (n-1)*var.muestral.tiempo/cuantil.izquierda)
```

```
## [1] 191.26267
```

```
(valor.derecho= (n-1)*var.muestral.tiempo/cuantil.derecha)
```

```
## [1] 544.95716
```

El intervalo de confianza para la desviación típica σ del tiempo de ejecución será:

```
c(sqrt(valor.izquierdo),sqrt(valor.derecho))
```

```
## [1] 13.829775 23.344318
```

Vemos que el valor 30 está a la derecha del intervalo de confianza.

Por tanto, podemos afirmar con un 95 % de confianza que $\sigma \leq 30$.

3.3.1. Intervalo de confianza para la varianza de una población normal en R

Para hallar un intervalo de confianza para la varianza poblacional en `R` hay que usar la función `varTest` del paquete `EnvStats`:

```
varTest(X,conf.level)$conf.int
```

donde `X` es el vector que contiene la muestra y `conf.level` el nivel de confianza, que por defecto es igual a 0.95.

Ejemplo

Hallemos un intervalo de confianza para la varianza de la amplitud del sépalo de la tabla de datos **iris** a partir de la muestra anterior. Suponemos que dicha variable es normal. Veremos en temas posteriores cómo se puede comprobar la normalidad de una variable.

Hallemos los valores de la amplitud del sépalo para las flores de nuestra muestra:

```
(amplitud.sépalo.muestra = iris[flores.elegidas,]$Sepal.Width)
```

```
##  [1] 2.7 3.2 3.2 2.3 3.4 2.5 2.0 3.0 3.0 2.7 3.8 3.4 2.7 3.5 3.5 3.0 2.3 3.2 2.0
## [20] 2.8 2.5 2.3 3.1 3.0 2.3 2.9 3.4 3.1 3.4 2.5 2.6 3.7 2.5 4.4 3.0 3.4 2.5 2.5
## [39] 3.2 3.0 3.3 3.2 3.8 3.6 3.0 3.4 2.8 2.6 2.9 3.1 3.1 3.3 3.2 3.2 3.5 3.5 3.0
## [58] 3.4 3.1 2.3
```

Un intervalo de confianza para la varianza de las amplitudes del sépalo para la tabla de datos **iris** al 95 % de confianza será:

```
library(EnvStats)
varTest(amplitud.sépalo.muestra,conf.level=0.95)$conf.int
```

```
##        LCL        UCL
## 0.16256399 0.33657861
## attr(,"conf.level")
## [1] 0.95
```

3.4. Bootstrap o remuestreo

Cuando no se satisfacen las condiciones teóricas que garantizan que el intervalo obtenido contiene el 95 % de las veces el parámetro poblacional deseado, podemos recurrir a un método no paramétrico. El más utilizado es el **bootstrap**, que básicamente consiste en:

1. **Remuestrear** la muestra: tomar muchas muestras aleatorias simples de la muestra de la que disponemos, cada una de ellas del mismo tamaño que la muestra original (pero simples, es decir, con reposición).
2. Calcular el estimador sobre cada una de estas submuestras.
3. Organizar los resultados en un vector.
4. Usar este vector para calcular un intervalo de confianza.

3.4.1. Bootstrap: método de los percentiles

La manera más sencilla de llevar a cabo el cálculo final del intervalo de confianza es el llamado **método de los percentiles** , en el que se toman como extremos del intervalo de confianza del $(1-\alpha)\cdot 100\,\%$ los cuantiles de orden $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$ del vector de estimadores.

Ejemplo

Como aplicación del **método de los percentiles** hallemos un intervalo de confianza para la varianza de la longitud del pétalo de la tabla de datos **iris**.

No podemos usar la fórmula vista anteriormente ya que la variable considerada no puede considerarse normal.

Tomaremos la muestra de la tabla de datos **iris** que hemos calculado anteriormente en la sección de intervalo de confianza para proporciones.

Usaremos la función `replicate` de `R` para calcular las varianzas de 1000 muestras "remuestradas" de nuestra muestra original:

```
set.seed(42)
X=replicate(1000,
            var(sample(iris[flores.elegidas,]$Petal.Length,
                       replace=TRUE)
                )
            )
```

A continuación hallamos el intervalo de confianza al 95 % $(1-\alpha = 0{,}95)$ calculando los cuantiles del método: (cuantiles de orden $\frac{\alpha}{2} = 0{,}025$ y $1-\frac{\alpha}{2} = 0{,}975$)

```
alpha = 0.05
IC.boot=c(quantile(X,alpha/2),quantile(X,1-alpha/2))
round(IC.boot,2)
```

```
##  2.5% 97.5%
##  2.41  3.53
```

Para aplicar el **método de los percentiles** en `R`, podemos usar la función `boot` del paquete `boot`:

```
boot(X,estadístico,R)
```

donde:

- `X` es el vector que forma la muestra de la que disponemos
- `R` es el número de muestras que queremos extraer de la muestra original
- El `estadístico` es la función que calcula el estadístico deseado de la submuestra, y tiene que tener dos parámetros: el primero representa la muestra original `X` y el segundo representa el vector de índices de una m.a.s. de `X`.

Ejemplo anterior

Vamos a aplicar la función `boot` al ejemplo anterior definiendo primero el estadístico a usar que sería la varianza en nuestro caso.

```
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, melanoma
```

```
var.boot=function(X,índices){var(X[índices])}
simulación=boot(iris[flores.elegidas,]$Petal.Length,var.boot,1000)
```

El intervalo de confianza viene dado por la función `boot.ci`:

```
boot.ci(simulación)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = simulación)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 2.531,  3.567 )   ( 2.529,  3.603 )
##
## Level     Percentile            BCa
## 95%   ( 2.409,  3.483 )   ( 2.488,  3.535 )
## Calculations and Intervals on Original Scale
```

Obtenemos cuatro intervalos de confianza para σ^2, calculados con cuatro métodos a partir de la simulación realizada.

El intervalo `Percentile` es el calculado con el método de los percentiles que hemos explicado antes, y se obtiene con el sufijo `$percent[4:5]`.

Vemos que los valores son parecidos a los obtenidos en la simulación *hecha a mano*.

Ejercicio

Mirando la documentación e investigando un poco, elaborad un pequeño resumen de cómo se obtienen y qué significan cada uno de los otros tres intervalos de confianza de la función `boot.ci`.

3.5. Guía rápida

- `t.test(X, conf.level=...)$conf.int` calcula el intervalo de confianza del `conf.level`$\times 100\,\%$ para la media poblacional usando la fórmula basada en la t de Student aplicada a la muestra `X`.
- `binom.exact(x,n,conf.level=...)` del paquete **epitools**, calcula el intervalo de confianza del `conf.level`$\times 100\,\%$ para la proporción poblacional aplicando el método de Clopper-Pearson a una muestra de tamaño `n` con `x` éxitos.
- `binom.wilson(x,n,conf.level=...)` del paquete **epitools**, calcula el intervalo de confianza del `conf.level`$\times 100\,\%$ para la proporción poblacional aplicando el método de Wilson a una muestra de tamaño `n` con `x` éxitos.
- `binom.approx(x,n,conf.level=...)` del paquete **epitools**, calcula el intervalo de confianza del `conf.level`$\times 100\,\%$ para la proporción poblacional aplicando la fórmula de Laplace a una muestra de tamaño `n` con `x` éxitos.
- `varTest(X,conf.level=...)$conf.int` del paquete `EnvStats`, calcula el intervalo de confianza del `conf.level`$\times 100\,\%$ para la varianza poblacional usando la fórmula basada en la khi cuadrado aplicada a la muestra `X`.

- `boot(X,E,R)` del paquete `boot`, lleva a cabo una simulación *bootstrap*, tomando `R` submuestras del vector `X` y calculando sobre ellas el estadístico representado por la función `E`.
- `boot.ci` del paquete `boot`, aplicado al resultado de una función `boot`, calcula diversos intervalos de confianza a partir del resultado de la simulación efectuada con `boot`. El nivel de confianza se especifica con el parámetro `conf`.

Capítulo 4

Contrastes de hipótesis paramétricos

Para que la estadística inferencial sea útil no solo necesitamos estimar un valor sino que además tendremos que tomar una *decisión* apoyada en los datos (muestras) que acepte o rechace alguna afirmación relativa al valor de un parámetro.

Ejemplo moluscos

Los responsables de salud pública del gobierno han determinado que el número medio de bacterias por cc en las aguas en las que se practica la recogida de moluscos para el consumo humano tiene que ser ≤ 70.

Tomamos una serie de muestras de agua de una zona, y hemos de decidir si podemos recoger moluscos.

Ejemplo routers

Una empresa de telecomunicaciones recibe una partida de 100 routers cada mes. El técnico que se encarga de la recepción del material tiene la orden de rechazar entera las partidas que contengan más de un 5 % de unidades defectuosas.

El técnico, al no disponer de tiempo material para revisar todos los routers, toma la decisión de aceptar o rechazar la partida basándose en el análisis de una muestra aleatoria de unidades.

Estas afirmaciones reciben el nombre de *hipótesis* y el método estadístico de toma de una decisión sobre una hipótesis recibe el nombre de **contraste de hipótesis**.

En un contraste de hipótesis, se contrastan dos hipótesis alternativas: la **hipótesis nula** H_0 y la **hipótesis alternativa** H_1.

La hipótesis alternativa H_1 es de la que buscamos evidencia.

La hipótesis nula H_0 es la que rechazaremos si obtenemos evidencia de la hipótesis alternativa.

Si no obtenemos evidencia a favor de H_1, *no podemos rechazar* H_0 (diremos que aceptamos H_0, pero es un **abuso de lenguaje**).

Ejemplo moluscos

Sea μ el número medio de bacterias por cc de agua.

El **contraste** que nos planteamos es el siguiente:

$$\begin{cases} H_0 : \mu \leq 70 \\ H_1 : \mu > 70 \end{cases}$$

La **decisión** que tomaremos se basará en algunas muestras de las que calcularemos la media muestral del número de bacterias por cc.

Si es bastante grande, lo consideraremos como una evidencia de H_1, y si no, aceptaremos H_0.

Ejemplo routers

Sea p la proporción de unidades defectuosas.

El **contraste** que nos planteamos es el siguiente:

$$\begin{cases} H_0 : p \leq 0{,}05 \\ H_1 : p > 0{,}05 \end{cases}$$

La **decisión** que tomemos se basará en las comprobaciones que realice el encargado de algunas unidades.

Calculará la proporción muestral de routers defectuosos. Si es bastante grande, lo consideraremos una evidencia de H_1, y si no, aceptaremos H_0.

4.1. Los contrastes de hipótesis

Definición. Un *contraste de hipótesis*

$$\begin{cases} H_0 : \text{hipótesis nula} \\ H_1 : \text{hipótesis alternativa} \end{cases}$$

consiste en plantear dos hipótesis:

- *Hipótesis nula* H_0: es la hipótesis que "por defecto" aceptamos como verdadera, y que rechazamos si hay pruebas en contra,
- *Hipótesis alternativa* H_1: es la hipótesis contra la que contrastamos la hipótesis nula y que aceptamos cuando rechazamos la nula,

y generar una **regla de decisión** para **rechazar** o no la hipótesis nula a partir de la información contenida en una muestra.

En un juicio, tenemos que declarar a un acusado inocente o culpable.

O sea, se plantea el **contraste** siguiente:

$$\begin{cases} H_0 : \text{El acusado es inocente.} \\ H_1 : \text{El acusado es culpable.} \end{cases}$$

Las pruebas serían los elementos de la muestra.

Si el jurado encuentra pruebas suficientemente incriminatorias, declara **culpable** al acusado (rechaza H_0 en favor de H_1).

En caso contrario, si no las encuentra suficientemente incriminatorias, le declara **no culpable** (no rechaza H_0)

Considerar no culpable $\neq$ declarar inocente.

Las pruebas tienen que aportar evidencia de H_1, lo que nos permitirá rechazar H_0.

Es imposible encontrar evidencias de que μ sea igual a un cierto valor μ_0. En cambio, sí que es puede hallar evidencias de que $\mu > \mu_0$, o de que $\mu < \mu_0$, o que $\mu \neq \mu_0$.

En este contexto:

- H_1 se define con $>$, $<$, o $\neq$.
- H_0 se define con $=$, $\leq$, o $\geq$.
- H_1 es la hipótesis de la que podemos hallar pruebas incriminatorias, H_0 la que estamos dispuestos a aceptar si no hay pruebas en contra.

Ejemplo

Queremos decidir si la media es más pequeña que 2 o no:

$$\begin{cases} H_0 : \mu = 2 \text{ (o } \mu \geq 2), \\ H_1 : \mu < 2. \end{cases}$$

Ejemplo

Queremos decidir si la media es igual o diferente de 5

$$\begin{cases} H_0 : \mu = 5 \\ H_1 : \mu \neq 5 \end{cases}$$

Ejemplo

Queremos dar la alerta de **desastre natural inminente** y poner a la población a salvo si la media de cierta variable meteorológica (temperatura, presión,...) toma el valor μ_0:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

4.1.1. Tipos de hipótesis alternativas

- **Hipótesis unilateral** (*one-sided*, también *de una cola*, *one-tailed*): $H : \theta > \theta_0$, $H : \theta < \theta_0$.
- **Hipótesis bilateral** (*two-sided*, también *de dos colas*, *two-tailed*): $H : \theta \neq \theta_0$

Los tests suelen tomar el nombre de la hipótesis alternativa: **test unilateral**, **test de dos colas**, etc.

4.1.2. Tipos de errores

La tabla siguiente resume los 4 casos que se pueden dar dependiendo de la decisión tomada:

Decisión/Realidad	H_0 cierta	H_0 falsa
Aceptar H_0	Decisión correcta Probabilidad=$1-\alpha$	Error Tipo II Probabilidad=β
Rechazar H_0	Error Tipo I Probabilidad=α	Decisión correcta Probabilidad=$1-\beta$

- **Error de Tipo I**: rechazar H_0 cuando es cierta. La probabilidad de cometerlo es:

$$P(\text{Error Tipo I}) = P(\text{Rechazar } H_0 \mid H_0 \text{ cierta}) = \alpha,$$

  donde α es el **nivel de significación del contraste**.

- **Error de Tipo II**: aceptar H_0 cuando es falsa. La probabilidad de cometerlo es:

$$P(\text{Error Tipo II}) = P(\text{Aceptar } H_0 | H_0 \text{ falsa}) = \beta,$$

  donde $1-\beta = P(\text{Rechazar } H_0 | H_0 \text{ falsa})$ es la **potencia del contraste**.

En un juicio, se declarar un acusado inocente o culpable.

- El **error de Tipo I** sería declarar culpable a un inocente.
- El **Error de Tipo II** sería declarar no culpable a un culpable.

Es más grave desde el punto de vista *ético* cometer un error tipo I ya que es peor castigar a un inocente que perdonar a un culpable. Por tanto, conviene minimizarlo.

En el desastre natural, damos la alerta si μ se acerca a cierto valor μ_0.

- El **error de Tipo I** sería no dar la alarma cuando el desastre natural ocurre (muertes varias).
- El **Error de Tipo II** sería dar la alarma a pesar de que no haya desastre natural (falsa alarma).

Lo más conveniente es encontrar una regla de rechazo de H_0 que tenga poca probabilidad de error de tipo I, α.

Pero también querríamos minimizar la probabilidad de error de tipo II, β.

Observación: cuando hacemos disminuir α, suele aumentar β.

¿Qué se suele hacer?

- Encontrar una regla de decisión para a un α máximo fijado.
- Después, si es posible, controlar la tamaño n de la muestra para minimizar β.

4.1.3. Terminología

En un contraste de hipótesis, tenemos los siguientes conceptos:

- **Estadístico de contraste**: es una variable aleatoria función de la muestra que nos permite definir una regla de rechazo de H_0.
- **Nivel de significación** α: la probabilidad de error de tipo I.

- **Región crítica o de rechazo**: zona o región de números reales donde se verifica que si el **estadístico de contraste** pertenece a la **región crítica**, entonces rechazamos H_0.
- **Región de aceptación**: zona o región complementaria de la **región crítica**.
- **Intervalo de confianza del** $(1-\alpha)\cdot 100\,\%$: intervalo de confianza para el parámetro poblacional del contraste. Es equivalente afirmar que el estadístico de contraste pertenece a la **región de aceptación** que afirmar que el parámetro del contraste pertenece al **intervalo de confianza del contraste**.

4.2. Contrastes de hipótesis para el parámetro μ de una variable normal con σ conocida

Sea X una variable aleatoria $N(\mu, \sigma)$ con μ desconocida y σ conocida.

Sea $X_1, \dots, X_n$ una m.a.s. de X de tamaño n.

Nos planteamos el contraste siguiente:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

De cara a hallar la región de rechazo, pensemos que tenemos que rechazar H_0 en favor de H_1 si $\overline{X}$ es "bastante más grande" que μ_0.

Si H_0 es verdadera,

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Entonces, la regla consistirá en rechazar H_0 si el **estadístico de contraste** Z es mayor que un cierto umbral, que determinaremos con α, el **nivell de significación del contraste** o **el error tipo I**.

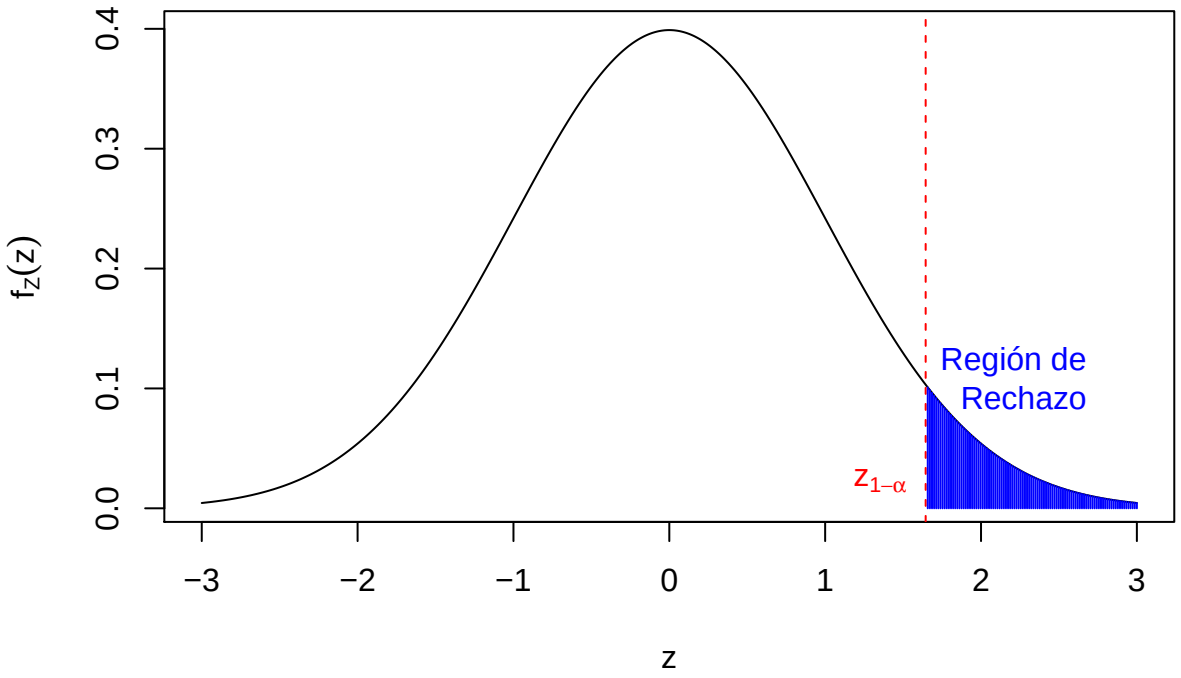
De cara a hallar la región de rechazo, queremos que se cumpla lo siguiente:

$$\begin{aligned} \alpha = P(\text{rechazar } H_0 | H_0 \text{ cierta }) = P(Z > \text{umbral }) \\ \Longrightarrow 1-\alpha = P(Z \leq \text{umbral }) \Longrightarrow \text{umbral } = z_{1-\alpha}. \end{aligned}$$

Por tanto, para que el **nivel de significación del contraste** sea α, la regla de rechazo tiene que ser: $Z > z_{1-\alpha}$

En resumen, **rechazamos** H_0 si $\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$.

Gráfico de la región de rechazo. Las abscisas o coordenadas x de la zona en azul serían los valores z para los que rechazaríamos la hipótesis nula H_0:



El contraste anterior tiene como:

- **Estadístico de contraste**: $Z = \dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$.
- **Región crítica**: $(z_{1-\alpha}, \infty)$.
- **Región de aceptación**: $(-\infty, z_{1-\alpha}]$.
- **Regla de decisión**: rechazar H_0 si $Z > z_{1-\alpha}$.
- **Intervalo de confianza**:

$$Z < z_{1-\alpha} \iff \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha} \iff \mu_0 > \overline{X} - z_{1-\alpha} \cdot \tfrac{\sigma}{\sqrt{n}}$$
$$\iff \mu_0 \in \left(\overline{X} - z_{1-\alpha} \cdot \tfrac{\sigma}{\sqrt{n}}, \infty\right)$$

- **Regla de decisión II**: rechazar H_0 si el μ_0 contrastado no pertenece al intervalo de confianza.

Ejercicio

Sea X una población normal con $\sigma = 1{,}8$. Queremos hacer el contraste

$$\begin{cases} H_0 : \mu = 20 \\ H_1 : \mu > 20 \end{cases}$$

con un nivel de significación de 0,05.

Tomamos una m.a.s. de $n = 25$ observaciones y obtenemos $\overline{x} = 20{,}25$.

¿Qué decidimos?

Tenemos los siguientes valores: $\alpha = 0{,}05$, $\sigma = 1{,}8$, $n = 25$, $\overline{x} = 20{,}25$.

El **Estadístico de contraste** valdrá $Z = \dfrac{\overline{X} - 20}{\frac{1{,}8}{\sqrt{25}}} = 0{,}694$.

La **Región crítica** será $(z_{1-0{,}05}, \infty) = (1{,}645, \infty)$.

Decisión: Como que $0{,}694 < 1{,}645$, no pertenece a la región crítica y por tanto no tenemos suficientes evidencias para rechazar H_0.

El **Intervalo de confianza** será:

$$\left(\overline{X} - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \infty\right) = (19{,}658, \infty)$$

Decisión II: Como $\mu_0 = 20$ pertenece al intervalo de confianza, no podemos rechazar H_0.

Sea X una v.a. $N(\mu, \sigma)$ con μ desconocida y σ conocida

Sea $X_1, \ldots, X_n$ una m.a.s. de X de tamaño n

Nos planteamos el contraste

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \ < \ \mu_0 \end{cases}$$

donde vamos a rechazar H_0 si $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ es *inferior a* un cierto umbral, que determinaremos con α.

Queremos que el **Error Tipo I** sea α:

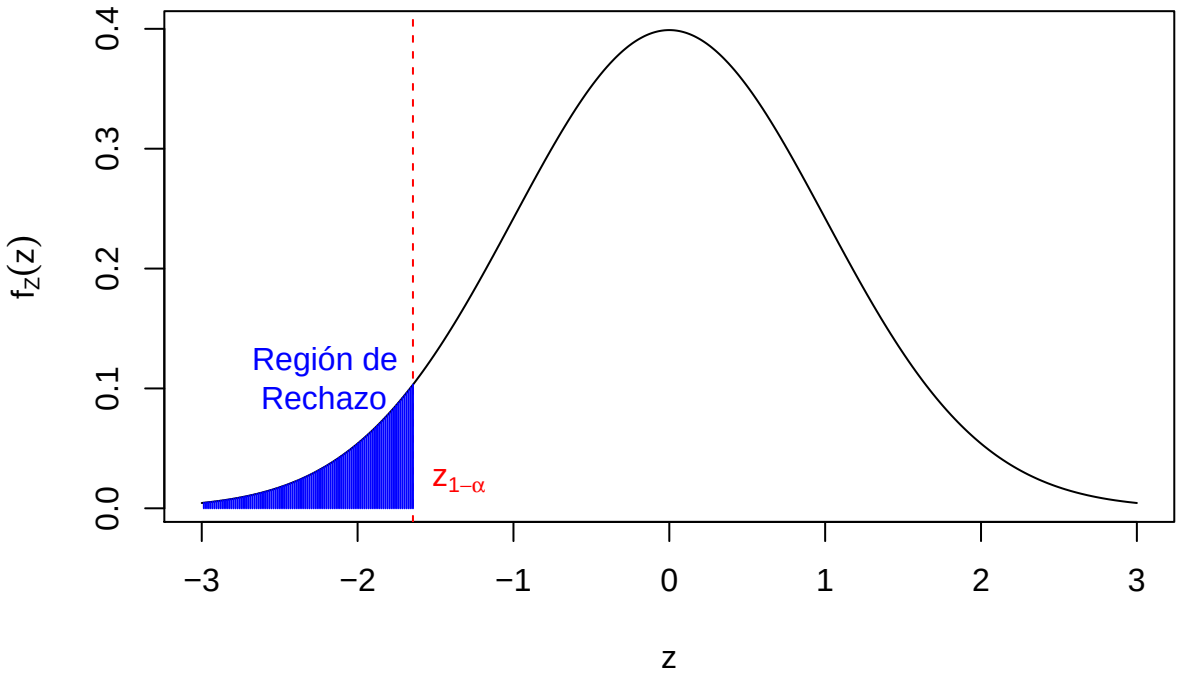
$$\alpha = P(\text{rechazar } H_0 | H_0 \text{ cierta}) = P(Z < \text{umbral }) \Longrightarrow \text{umbral } = z_\alpha,$$

por lo tanto, para que el nivel de significación del contraste Sea α, la regla de rechazo tiene que ser $Z < z_\alpha$.

La Región crítica es $(-\infty, z_\alpha)$.

En resumen, **rechazamos** H_0 si $\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha = -z_{1-\alpha}$.

Gráfico de la región de rechazo. Las abscisas o coordenadas x de la zona en azul serían los valores z para los que rechazaríamos la hipótesis nula H_0:



El contraste anterior tiene como:

- **Estadístico de contraste**: $Z = \dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$.
- **Región crítica**: $(-\infty, -z_{1-\alpha})$.
- **Región de aceptación**: $[-z_{1-\alpha}, \infty)$.
- **Regla de decisión**: rechazar H_0 si $Z < -z_{1-\alpha}$.
- **Intervalo de confianza**:

$$Z > -z_{1-\alpha} \iff \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > -z_{1-\alpha} \iff \mu_0 < \overline{X} + z_{1-\alpha} \cdot \tfrac{\sigma}{\sqrt{n}}$$
$$\iff \mu_0 \in \left(-\infty, \overline{X} + z_{1-\alpha} \cdot \tfrac{\sigma}{\sqrt{n}} \right)$$

- **Regla de decisión II**: rechazar H_0 si el μ_0 contrastado no pertenece al intervalo de confianza.

Sea X una v.a. $N(\mu, \sigma)$ con μ desconocida y σ conocida

Sea $X_1, \ldots, X_n$ una m.a.s. de X de tamaño n

Consideremos ahora el contraste

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Rechazar H_0 si $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ está a *bastante lejos de* de 0, y la determinaremos con el valor de α

Queremos como antes que el **Error Tipo I** sea α:

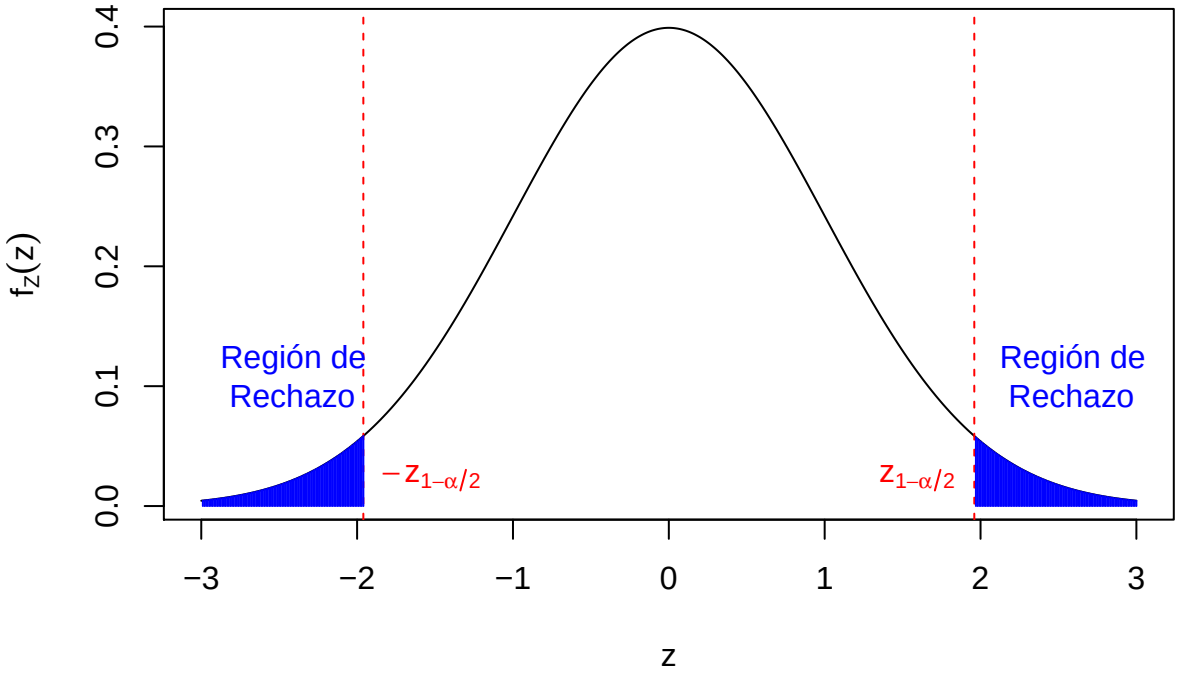
$$\begin{aligned}\alpha &= P(\text{rechazar } H_0 | H_0 \text{ cierta }) = P(Z < -\text{umbral} \text{ o } Z > \text{umbral }) \\ &= P(Z < -\text{umbral }) + P(Z > \text{umbral }) = 2P(Z > \text{umbral }) \\ &= 2(1 - P(Z < \text{umbral })) \Longrightarrow P(Z < \text{umbral }) = 1 - \frac{\alpha}{2}, \\ &\Longrightarrow \text{umbral } = z_{1-\frac{\alpha}{2}}.\end{aligned}$$

Ahora para que el nivel de significación del contraste sea α, la **regla de rechazo** tiene que ser

$$Z < -z_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}} \text{ o } Z > z_{1-\frac{\alpha}{2}}.$$

La región crítica es $(-\infty, z_{\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty)$.

Gráfico de la región de rechazo. Las abscisas o coordenadas x de la zona en azul serían los valores z para los que rechazaríamos la hipótesis nula H_0:



Seguidamente, calculemos el **Intervalo de confianza** para el contraste anterior:

$$\begin{aligned}-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}} &\Longleftrightarrow -z_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \\ &\Longleftrightarrow -z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu_0 < z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \\ &\Longleftrightarrow \overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu_0 < \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \\ &\Longleftrightarrow \mu_0 \in \left(\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

Ejercicio

Sea X una población normal con $\sigma = 1{,}8$. Queremos realizar el contraste

$$\begin{cases} H_0 : \mu = 20 \\ H_1 : \mu \neq 20 \end{cases}$$

con un nivel de significación de 0,05.

Tomamos una m.a.s. de $n = 25$ observaciones y obtenemos $\overline{x} = 20{,}5$.

¿Qué decidimos?

Tenemos los valores siguientes: $\alpha = 0{,}05$, $\sigma = 1{,}8$, $n = 25$, $\overline{x} = 20{,}5$.

El **Estadístico de contraste** vale $Z = \dfrac{\overline{X} - 20}{\frac{1{,}8}{\sqrt{25}}} = 1{,}389$.

La **Región crítica** será: $(-\infty, z_{0{,}025}[\cup] z_{0{,}975}, \infty)$=$(-\infty, -1{,}96) \cup (1{,}96, \infty)$.

El **Intervalo de confianza** será: $\left(20{,}5 - 1{,}96\frac{1{,}8}{\sqrt{25}}, 20{,}5 + 1{,}96\frac{1{,}8}{\sqrt{25}}\right) = (19{,}794, 21{,}206)$.

Decisión: No tenemos evidencias suficientes para rechazar H_0 ya que, por un lado, el estadístico de contraste no pertenece a la región crítica y, por otro, el valor $\mu_0 = 20$ pertenece al intervalo de confianza.

4.2.1. El p-valor

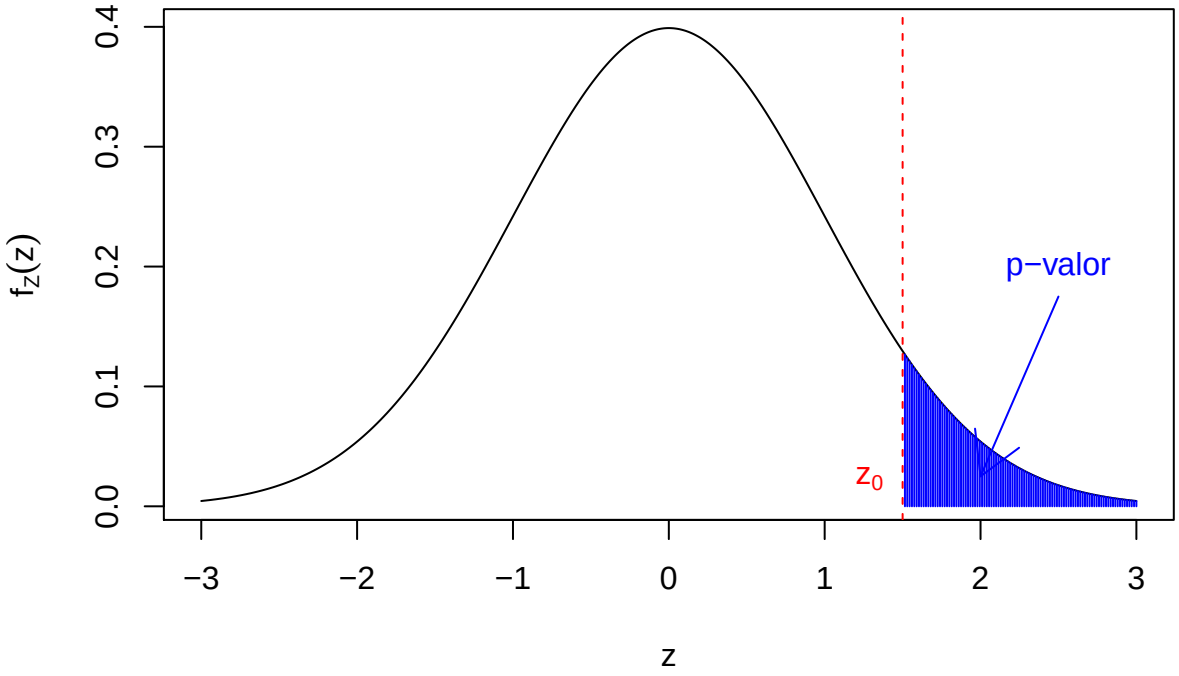
El *p-valor* o *valor crítico* (*p-value*) de un contraste es la probabilidad que, si H_0 es verdadera, el estadístico de contraste tome un valor tan extremo o más que el que se ha observado.

Consideremos por ejemplo un contraste del tipo:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0. \end{cases}$$

Si el estadístico Z tiene el valor z_0, el p-valor será:

$$p\text{-valor} = P(Z \geq z_0).$$

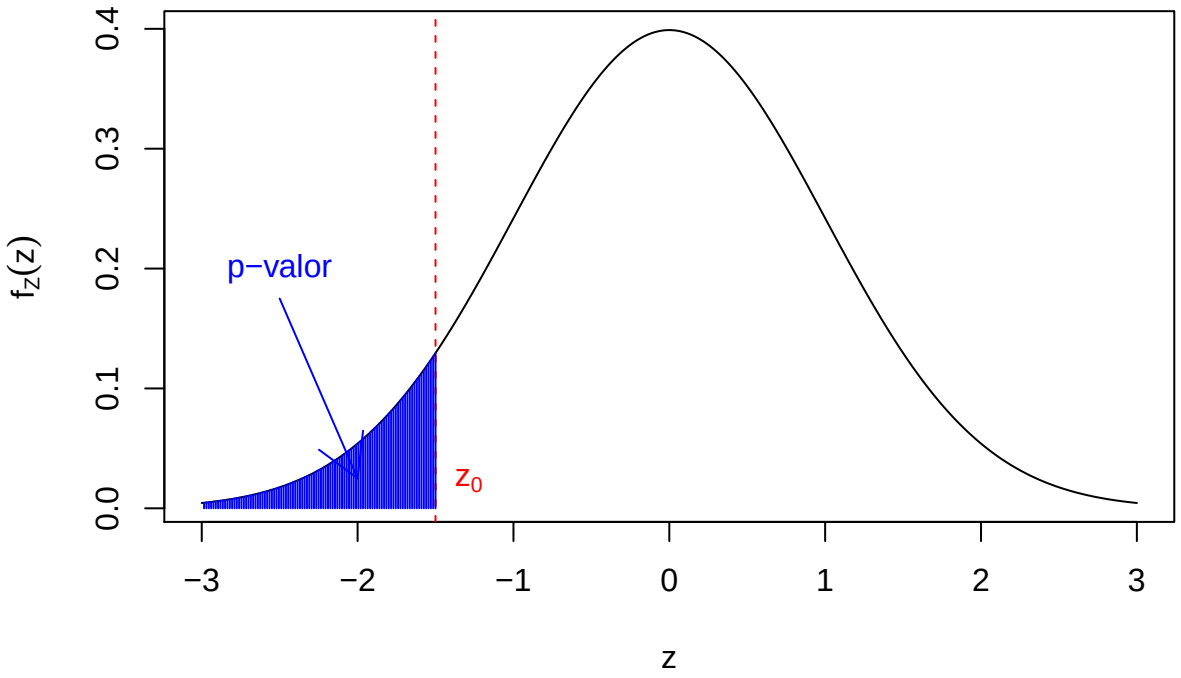


Para el contraste:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0. \end{cases}$$

Si el estadístico Z tiene el valor z_0, el p-valor será:

$$p\text{-valor} = P(Z \leq z_0).$$

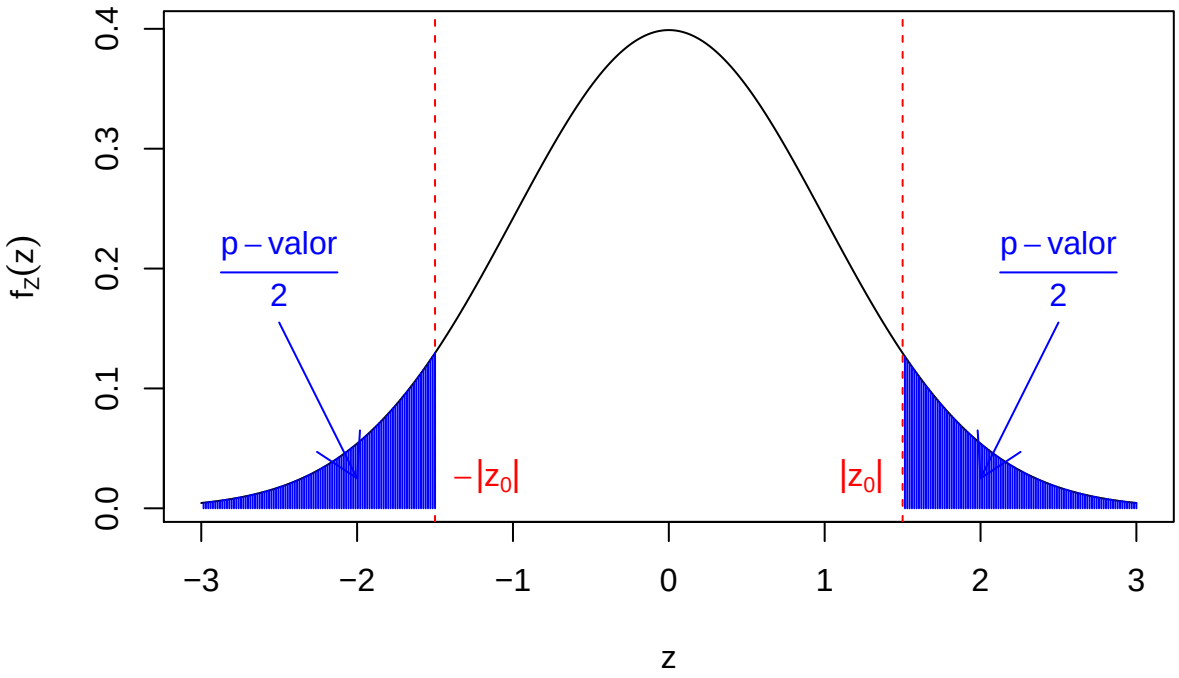


Si ahora consideramos el contraste

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

y si el estadístico Z ha dado z_0, el p-valor será:

$$p\text{-valor} = 2 \cdot \text{mín}\{P(Z \leq -|z_0|), P(Z \geq |z_0|) = 2 \cdot P(Z \geq |z_0|)$$



El *p-valor* o *valor crítico* (*p-value*) de un contraste es la probabilidad que, si H_0 es verdadera, el estadístico de contraste tome un valor tan extremo o más que el que se ha observado.

Es una *medida inversa de la fuerza de las pruebas o evidencias que hay en contra de* H_1: si H_0 es verdadera, cuanto más pequeño sea el p-valor, más improbable es observar lo que hemos observado.

En consecuencia, cuanto más pequeño sea el p-valor, con más fuerza podemos rechazar H_0.

Supongamos, por ejemplo, que hemos obtenido un p-valor de 0,03:

- *Significa* que la probabilidad de que, si H_0 es verdadera, el estadístico de contraste tome un valor tan extremo o más que el que ha tomado, es 0.03 (**pequeño: evidencia de que H_0 es falsa.**)
- *No significa*:
  - La probabilidad que H_0 Sea verdadera es 0,03
  - H_0 es verdadera un 3 % de les veces

Importante:

En un contraste con nivel de significación α,

- rechazamos H_0 si p-valor $< \alpha$.
- aceptamos H_0 si $\alpha \leq p$-valor.

Si consideramos por ejemplo un contraste del tipo:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

y suponemos que el estadístico Z vale z_0. El p-valor es $P(Z \geq z_0)$. Entonces:

- Rechazamos $H_0 \iff z_0 > z_{1-\alpha}$,,
- O, dicho de otra forma,

$$p\text{-valor} = P(Z \geq z_0) < P(Z \geq z_{1-\alpha}) = 1 - (1-\alpha) = \alpha.$$

Si ahora consideramos un contraste del tipo:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

y suponemos que el estadístico Z vale z_0. El p-valor es $P(Z \leq z_0)$. Entonces:

- Rechazamos $H_0 \iff z_0 < z_\alpha$,
- O, dicho de otra forma,

$$p\text{-valor} = P(Z \leq z_0) < P(Z \leq z_\alpha) = \alpha.$$

Por último, supongamos que el contraste es del tipo:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

y que el estadístico Z vale $z_0 > 0$. El p-valor es $2P(Z \geq |z_0|)$. Entonces:

- Rechazamos $H_0 \iff |z_0| > z_{1-\frac{\alpha}{2}}$,
- O, dicho de otra forma,

$$p\text{-valor} = 2P(Z \geq |z_0|) < 2P(Z \geq z_{1-\frac{\alpha}{2}}) = 2\left(1 - \left(1 - \frac{\alpha}{2}\right)\right) = \alpha.$$

El *p-valor* de un contraste es:

- El nivel de significación α más pequeño para el que rechazamos la hipótesis nula.
- El nivel de significación α más grande para el que aceptaríamos la hipótesis nula.
- La probabilidad mínima de error de Tipo I que permitimos si rechazamos la hipótesis nula con el valor del estadístico de contraste obtenido.
- La probabilidad máxima de error de Tipo I que permitimos si aceptamos la hipótesis nula con el valor del estadístico de contraste obtenido.

Importante:

Si no establecemos un nivel de significación α, entonces

- Aceptamos H_0 si el p-valor es "grande" ($\geq 0{,}1$).
- Rechazamos H_0 si el p-valor es "pequeño" ($< 0{,}05$). En este caso, el p-valor es:
  - *Significativo* si es $< 0{,}05$ (En `R`, se simboliza con un asterisco, `*`).
  - *Fuertemente significativo* si es $< 0{,}01$ (En `R`, se simboliza con dos asteriscos, `**`).
  - *Muy significativo* si es $< 0{,}001$ (En `R`, se simboliza con tres asteriscos, `***`).

Si el p-valor está entre 0,05 y 0,1 y no tenemos nivel de significación, se requieren estudios posteriores para tomar una decisión.

Es la denominada **zona crepuscular**, o *twilight zone*.

Ejercicio

Sea X una población normal con $\sigma = 1{,}8$. Queremos hacer el contraste

$$\begin{cases} H_0 : \mu = 20, \\ H_1 : \mu > 20. \end{cases}$$

Tomamos una m.a.s. de $n = 25$ observaciones y obtenemos $\overline{x} = 20{,}25$.

¿Qué decidimos?

Como no nos dan el nivel de significación α, calcularemos el p-valor.

Si calculamos el **estadístico de contraste**, obtenemos $z_0 = \dfrac{\overline{X} - 20}{\frac{1{,}8}{\sqrt{25}}} = \dfrac{20{,}25 - 20}{\frac{1{,}8}{\sqrt{25}}} = 0{,}694$.

El p**-valor** valdrá: $p = P(Z \geq 0{,}694) = 0{,}244 > 0{,}1$ grande.

La **decisión** que tomamos por consiguiente es que no tenemos evidencias suficientes para rechazar H_0.

Ejercicio

Sea X una población normal con $\sigma = 1{,}8$. Queremos hacer el contraste

$$\begin{cases} H_0 : \mu = 20 \\ H_1 : \mu > 20 \end{cases}$$

Tomamos una m.a.s. de $n = 25$ observaciones y obtenemos $\overline{x} = 20{,}75$.

¿Qué decidimos?

El **estadístico de contraste** será $Z = \dfrac{\overline{X} - 20}{\frac{1{,}8}{\sqrt{25}}} = \dfrac{20{,}75 - 20}{\frac{1{,}8}{\sqrt{25}}} = 2{,}083$.

El p**-valor** será: $P(Z \geq 2{,}083) = 0{,}019$ pequeño.

En este caso la **decisión** será rechazar H_0 ya que tenemos suficientes evidencias para hacerlo.

Si conocemos el **nivel de significación** α, la decisión que tomemos en un contraste se puede basar en:

- **la región crítica:** si el estadístico de contraste cae dentro de la **región crítica** para al nivel de significación α, rechazamos H_0.
- **el intervalo de confianza:** si el **parámetro poblacional** a contrastar cae dentro del **intervalo de confianza** para el nivel $(1 - \alpha) \cdot 100\,\%$ de confianza, aceptamos H_0.
- **el p-valor:** si el p-valor es más pequeño que el nivel de significación α, rechazamos H_0.

Si desconocemos el **nivel de significación** α, la decisión que tomemos en un contraste se puede basar en:

- **el p-valor:** Si el p-valor es pequeño, rechazamos H_0, y si es grande, la aceptamos.

4.2.2. El método de los *seis* pasos (caso de conocer α)

1) Establecer la hipótesis nula H_0 y la hipótesis alternativa H_1.
2) Fijar un nivel de significación α.
3) Seleccionar el estadístico de contraste apropiado.
4) Calcular el valor del estadístico de contraste a partir de les datos muestrales.
5) Calcular el p-valor del contraste.
6) **Decisión:** rechazar H_0 en favor de H_1 si el p-valor es más pequeño que α; en caso contrario, aceptar H_0.

4.2.3. El método de los *cinco* pasos (caso de no conocer α)

1) Establecer la hipótesis nula H_0 y la hipótesis alternativa H_1.
2) Seleccionar el estadístico de contraste apropiado.
3) Calcular el valor del estadístico de contraste a partir de los valores de la muestra.
4) Calcular el p-valor del contraste.
5) **Decisión:** rechazar H_0 en favor de H_1 si el p-valor es pequeño ($< 0{,}05$), aceptar H_0 si el p-valor es grande ($\geq 0{,}1$), y ampliar el estudio si el p-valor está entre 0.05 y 0.1.

Ejercicio

Los años de vida de un router sigue aproximadamente una ley de distribución normal con $\sigma = 0{,}89$ años.

Una muestra aleatoria de la duración de 100 aparatos ha dado una vida media de 7.18 años.

Queremos decidir si la vida media en de estos routers es superior a 7 años:

$$\begin{cases} H_0 : \mu = 7, \\ H_1 : \mu > 7. \end{cases}$$

Tomamos un nivel de significación $\alpha = 0{,}05$.

EL estadístico de contraste es

$$z_0 = \frac{\overline{X} - 7}{0{,}89/\sqrt{100}} = \frac{\overline{X} - 7}{0{,}0089} = \frac{7{,}18 - 7}{0{,}089} = 2{,}022.$$

El p-valor es $p = P(Z \geq 2{,}022) = 0{,}022$.

Como $0{,}022 < \alpha$, rechazamos H_0.

Concluimos que tenemos suficientes evidencias para aceptar que la vida media de los routers es superior a los 7 años: $\mu > 7$.

Supongamos ahora que tomamos un nivel de significación $\alpha = 0{,}01$.

Como el p-valor $0{,}022 > \alpha$, no podemos rechazar H_0.

En este caso, concluimos que no tenemos evidencias suficientes para rechazar que la vida media de los routers sea de 7 años o menor: $\mu \leq 7$.

Como el p-valor obtenido, 0,022, es pequeño ($< 0{,}05$), rechazamos H_0.

Concluimos que tenemos suficientes evidencias para aceptar que la vida media de los routers es superior a los 7 años: $\mu > 7$.

4.2.4. Un último consejo

Como una regla recomendaríamos en un informe:

- *Si conocemos α,* encontrar el p-valor y el intervalo de confianza del contraste para α dado (nivel de confianza $(1-\alpha)\cdot 100\,\%$).
- *Si no tenemos fijado (no conocemos) α,* encontrar el p-valor, y el intervalo de confianza del contraste al nivel de confianza $95\,\%$.

4.3. Contrastes de hipótesis para el parámetro μ de una variable normal con σ desconocida

4.3.1. Contraste para μ cuando n es grande: Z-test

Si el tamaño n de la muestra es grande (pongamos $n \geq 40$), podemos aplicar las reglas anteriores aunque la población no sea normal.

Si además σ es desconocida, ésta se puede sustituir por la desviación típica muestral $\widetilde{S}_X$ en la expresión de Z:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\widetilde{S}_X}{\sqrt{n}}}$$

Ejemplo

Una organización ecologista afirma que el peso medio de los individuos adultos de una especie marina ha disminuido drásticamente.

Se sabe por los datos históricos que el peso medio poblacional era de 460 g.

Una muestra aleatoria de 40 individuos de esta especie ha dado una media muestral de 420 g. y una desviación típica muestral de 119 g.

Con estos datos, ¿podemos afirmar con un nivel de significación del $5\,\%$ que el peso mediano es inferior a 460 g?

Ejemplo

El contraste que nos planteamos es el siguiente:

$$\begin{cases} H_0 : \mu = 460, \\ H_1 : \mu < 460, \end{cases}$$

donde μ representa el peso medio de todos los individuos de la especie.

Consideramos un **nivel de significación** $\alpha = 0{,}05$.

Podemos usar como **estadístico de contraste**, como $n = 40$ es grande, la expresión:

$$Z = \frac{\overline{X} - \mu_0}{\widetilde{S}_X/\sqrt{n}},$$

cuyo valor es: $z_0 = \dfrac{420 - 460}{119/\sqrt{40}} = -2{,}126$.

El p**-valor** será:

$$P(Z \leq -2{,}126) = 0{,}017.$$

Decisión: como $\alpha > p$-valor, rechazamos (al nivel de significación $\alpha = 0{,}05$) que el peso medio sea de 460 g. (H_0) en contra que sea menor de 460 g. (H_1).

Concluimos que tenemos suficientes evidencias para afirmar que el peso medio es menor que 460 g. y por tanto, ha menguado en los últimos años.

El **intervalo de confianza** será:

$$\left(-\infty, \overline{X} - z_\alpha \cdot \frac{\widetilde{S}_X}{\sqrt{n}}\right) =]-\infty, 450{,}949].$$

Informe: el p**-valor** de este contraste es 0,017, y el intervalo de confianza al nivel de significación $\alpha = 0{,}05$ para la media poblacional μ es $]-\infty, 450{,}949]$.

Como $460 \notin (-\infty, 450{,}949)$, hay evidencia significativa para rechazar la hipótesis nula en favor de $\mu < 460$.

4.3.2. Contraste para μ de normal con σ desconocida: T-test

Las reglas de decisión son similares al caso con σ conocida, excepto que ahora **sustituimos σ por $\widetilde{S}_X$** y empleamos la distribución t de Student.

Recordemos que si $X_1, \ldots, X_n$ es una m.a.s. de una población normal X con mediana μ_0, la variable $T = \frac{\overline{X} - \mu_0}{\frac{\widetilde{S}_X}{\sqrt{n}}}$ sigue una distribución t de Student con $n - 1$ grados de libertad.

Los p-valores se calculan con esta distribución.

Condiciones: supongamos que disponemos de una m.a.s. de tamaño n de una población $N(\mu, \sigma)$ con μ y σ desconocidas.

Nos planteamos los contrastes siguientes:

$$\begin{cases} H_0 : \mu = \mu_0 & (\text{ o } H_0 : \mu \leq \mu_0) \\ H_1 : \mu > \mu_0 \end{cases}$$

$$\begin{cases} H_0 : \mu = \mu_0 & (\text{ o } H_0 : \mu \geq \mu_0) \\ H_1 : \mu < \mu_0 \end{cases}$$

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Para los contrastes anteriores, usaremos como **estadístico de contraste**:

$$T = \frac{\overline{X} - \mu_0}{\frac{\widetilde{S}_X}{\sqrt{n}}}$$

y calcularemos su valor t_0 sobre la muestra.

Los **p-valores** serán los siguientes:

p-valor: $P(t_{n-1} \geq t_0)$.

p-valor: $P(t_{n-1} \leq t_0)$.

p-valor: $2P(t_{n-1} \geq |t_0|)$.

Ejercicio

Se espera que el nivel de colesterol en plasma de unos enfermos bajo un determinado tratamiento se distribuya normalmente con media 220 mg/dl.

Se toma una muestra de 9 enfermos, y se miden sus niveles:

$$203, 229, 215, 220, 223, 233, 208, 228, 209.$$

Contrastar la hipótesis que esta muestra efectivamente proviene de una población con media 220 mg/dl.

El contraste planteado es el siguiente:

$$\begin{cases} H_0 : \mu = 220, \\ H_1 : \mu \neq 220, \end{cases}$$

donde μ representa la media del colesterol en plasma de la población.

Bajo estas condiciones (población normal, σ desconocida, muestra pequeña de $n = 9$) usaremos como **estadístico de contraste**: $T = \frac{\overline{X}-\mu_0}{\widetilde{S}_X/\sqrt{9}}$ cuya distribución es t_8.

El valor de dicho estadístico será:

```
colesterol=c(203,229,215,220,223,233,208,228,209)
media.muestral = mean(colesterol)
desv.típica.muestral = sd(colesterol)
(estadístico.contraste = (media.muestral-220)/
    (desv.típica.muestral/sqrt(length(colesterol))))
```

```
## [1] -0.38009147
```

El **p-valor** del contraste será:

```
(p=round(2*pt(abs(estadístico.contraste),lower.tail=FALSE,df=8),4))
```

```
## [1] 0.7138
```

Decisión: Como que el p-valor es muy grande, no podemos rechazar que el nivel mediano de colesterol en plasma sea igual a 220 mg/dl.

Por tanto, aceptamos que el nivel de colesterol en plasma en esta población tiene media 220 mg/dl.

El **intervalo de confianza** al 95 % será:

$$\begin{aligned}\left(\overline{X} - t_{8,0,975}\frac{\widetilde{S}_X}{\sqrt{n}},\ \overline{X} + t_{8,0,975}\frac{\widetilde{S}_X}{\sqrt{n}}\right) &= \left(218{,}667 - 2{,}306 \cdot \frac{10{,}524}{\sqrt{9}}, 218{,}667 + 2{,}306 \cdot \frac{10{,}524}{\sqrt{9}}\right) \\ &= (210{,}577, 226{,}756)\end{aligned}$$

Informe: El p-valor de este contraste es 0,7138 y el intervalo de confianza del 95 % para el nivel medio de colesterol μ es $(210{,}577, 226{,}756)$.

Como el p-valor es grande y $220 \in (210{,}577, 226{,}756)$, no hay evidencia que nos permita rechazar que $\mu = 220$.

4.3.3. Contraste de μ de normal con σ desconocida en R: función `t.test`

La sintaxis básica de la función `t.test` es

```
t.test(x, y, mu=..., alternative=..., conf.level=..., paired=...,
       var.equal=..., na.omit=...)
```

donde los parámetros necesarios para realizar un contraste de una muestra son los siguientes:

- `x` es el vector de datos que forma la muestra que analizamos.
- `mu` es el valor μ_0 de la hipótesis nula: $H_0 : \mu = \mu_0$.
- El parámetro `alternative` puede tomar tres valores: `"two.sided"`, para contrastes bilaterales, y `"less"` y `"greater"`, para contrastes unilaterales. En esta función, y en todas las que explicamos en esta lección, su valor por defecto, que no hace falta especificar, es `"two.sided"`. El significado de estos valores depende del tipo de test que efectuemos:
- `"two.sided"` representa la hipótesis alternativa $H_1 : \mu \neq \mu_0$, `"less"` corresponde a $H_1 : \mu < \mu_0$, y `"greater"` corresponde a $H_1 : \mu > \mu_0$.
- El valor del parámetro `conf.level` es el nivel de confianza $1 - \alpha$. Su valor por defecto es 0.95, que corresponde a un nivel de confianza del 95 %, es decir, a un nivel de significación $\alpha = 0{,}05$.
- El parámetro `na.action` sirve para especificar qué queremos hacer con los valores NA. Es un parámetro genérico que se puede usar en casi todas las funciones de estadística inferencial y análisis de datos. Sus valores más útiles son:
  - `na.omit`, su valor por defecto, elimina las entradas NA de los vectores (o los pares que contengan algún NA, en el caso de muestras emparejadas). Por ahora, esta opción por defecto es la adecuada, por lo que no hace falta usar este parámetro, pero conviene saber que hay alternativas.
  - `na.fail` hace que la ejecución pare si hay algún NA en los vectores.
  - `na.pass` no hace nada con los NA y permite que las operaciones internas de la función sigan su curso y los manejen como les corresponda.

El ejemplo anterior se resolvería de la forma siguiente:

```
t.test(colesterol,mu=220,alternative="two.sided",conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  colesterol
## t = -0.380091, df = 8, p-value = 0.71377
## alternative hypothesis: true mean is not equal to 220
## 95 percent confidence interval:
##  210.57737 226.75596
## sample estimates:
## mean of x
## 218.66667
```

Ejercicio

Veamos si, dada una muestra de tamaño 40 de flores de la tabla de datos iris, podemos considerar que la media de la longitud del sépalo es mayor que 5,7.

Para ello, primero obtenemos la muestra correspondiente fijando la semilla de aleatoriedad:

```
set.seed(230)
flores.elegidas=sample(1:150,40,replace=TRUE)
```

Seguidamente, hallamos las longitudes del sépalo de las flores de la muestra:

```
(long.sépalo.muestra=iris[flores.elegidas,]$Sepal.Length)
```

```
##  [1] 5.0 4.9 6.0 4.6 4.7 5.1 5.8 4.4 4.6 7.0 7.7 4.8 4.9 7.2 6.5 4.8 7.7 6.2 5.1
## [20] 6.8 7.2 5.0 6.7 6.9 4.6 5.7 6.4 6.1 6.4 4.7 5.0 7.7 6.2 5.0 5.1 4.9 6.3 5.0
## [39] 5.6 5.2
```

Por último, realizamos el contraste requerido:

```
t.test(long.sépalo.muestra,mu=5.7,alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  long.sépalo.muestra
## t = 0.236644, df = 39, p-value = 0.40709
## alternative hypothesis: true mean is greater than 5.7
## 95 percent confidence interval:
##  5.4705051       Inf
## sample estimates:
## mean of x
##    5.7375
```

Fijémonos que se trata de un contraste de una muestra, por tanto, no ha sido necesario especificar el vector y.

El contraste que hemos realizado ha sido el siguiente:

$$\left.\begin{array}{ll} H_0: & \mu = 5{,}7, \\ H_1: & \mu > 5{,}7, \end{array}\right\}$$

donde μ representa la media de la longitud del sépalo de todas las flores de la tabla de datos **iris**.

El p-valor obtenido ha sido 0.4071, valor superior a 0,1.

Por tanto, podemos concluir que no tenemos evidencias suficientes para rechazar la hipótesis nula y concluir que la media de la longitud del sépalo de las flores de la tabla de datos **iris** no es mayor que 5,7. De hecho, podemos observar en el "output" del `t.test` que la media de la muestra considerada vale 5.737, valor no significativamente mayor que 5,7.

Observamos que el `t.test` nos dice que el valor del estadístico de contraste es 1.499 y que dicho estadístico se distribuye según una t de Student con 39 grados de libertad (tamaño de la muestra, 40 menos 1).

El "output" del `t.test` también nos da el intervalo de confianza al 95 % de confianza asociado al contraste:

```
t.test(long.sépalo.muestra,mu=5.7,alternative = "greater")$conf.int
```

```
## [1] 5.4705051       Inf
## attr(,"conf.level")
## [1] 0.95
```

intervalo que contiene el valor de $\mu_0 = 5{,}7$, razón por la cual hemos aceptado la hipótesis nula H_0.

4.3.4. Z-test contra T-test

En el caso de una población con σ desconocida:

- Si la muestra es pequeña y la población es normal, tenemos que usar el T-test.
- Si la muestra es grande y la población cualquiera, podemos usar el Z-test.
- Si la muestra es grande y la población es normal, podemos usar ambos. En este último caso, os recomendamos que uséis el T-test debido a que es más preciso.

4.4. Contrastes de hipótesis para el parámetro p de una variable de Bernoulli

Supongamos que tenemos una m.a.s. de tamaño n de una población Bernoulli de parámetro p.

Obtenemos x_0 éxitos, de forma que la proporción muestral de éxitos será: $\widehat{p}_X = x_0/n$

Consideramos un contraste con hipótesis nula: $H_0 : p = p_0$

Si H_0 es verdadera, el número de éxitos sigue una distribución $B(n, p_0)$.

Nos planteamos los contrastes siguientes:

$$\begin{cases} H_0 : p = p_0, & (\text{ o } H_0 : p \leq p_0), \\ H_1 : p > p_0. \end{cases}$$

$$\begin{cases} H_0 : p = p_0, & (\text{ o } H_0 : p \geq p_0), \\ H_1 : p < p_0. \end{cases}$$

$$\begin{cases} H_0 : p = p_0, \\ H_1 : p \neq p_0. \end{cases}$$

Los **p-valores** serán los siguientes:

p-valor: $P(B(n, p_0) \geq x_0)$.

p-valor: $P(B(n, p_0) \leq x_0)$.

p-valor: $2\,\text{mín}\{P(B(n, p_0) \leq x_0), P(B(n, p_0) \geq x_0)\}$.

Ejemplo

Tenemos un test para detectar un determinado microorganismo. En una muestra de 25 cultivos con este microorganismo, el test lo detectó en 21 casos. Hay evidencia que la sensibilidad del test sea superior al 80 %?

El contraste planteado es el siguiente:

$$\begin{cases} H_0 : p = 0{,}8, \\ H_1 : p > 0{,}8, \end{cases}$$

donde p representa la probabilidad de que el test detecte el microorganismo.

Com **estadístico de contraste** usaremos el número de éxitos x_0, que bajo la hipótesis nula H_0, se distribuye según una $B(25, 0{,}8)$.

El valor del **estadístico de contraste** es: $x_0 = 21$

El p**-valor** será:

$$P(B(25, 0{,}8) \geq 21) = 1 - \texttt{pbinom}(20, 25, 0{,}8) = 0{,}421.$$

Decisión: como el p-valor es muy grande, no podemos rechazar la hipótesis nula.

No hay evidencia que la sensibilidad de la test sea superior al 80 %.

4.4.1. Contrastes para proporciones en R

Este test está implementado en la función `binom.test`, cuya sintaxis es

```
binom.test(x, n, p=..., alternative=..., conf.level=...)
```

donde

- `x` y `n` son números naturales: el número de éxitos y el tamaño de la muestra.
- `p` es la probabilidad de éxito que queremos contrastar.

Puede ser útil saber que el intervalo de confianza para la p que da `binom.test` en un contraste bilateral es el de Clopper-Pearson.

El contraste anterior sería en R:

```
binom.test(21,25,p=0.8,alternative="greater",conf.level=0.95)
```

```
##
##  Exact binomial test
##
## data:  21 and 25
## number of successes = 21, number of trials = 25, p-value = 0.42067
## alternative hypothesis: true probability of success is greater than 0.8
## 95 percent confidence interval:
##  0.6703917 1.0000000
## sample estimates:
## probability of success
##                   0.84
```

Ejercicio

Consideremos la tabla de datos **birthwt** del paquete **MASS**. Dicha tabla de datos contiene información acerca de 189 recién nacidos en un hospital de Springfield en el año 1986.

Las variables consideradas son las siguientes:

- low: indicador de si el peso del recién nacido ha sido menor que 2.5 kg.
- age: edad de la madre en años.
- lwt: peso de la madre en libras durante el último período.
- race: raza de la madre (1: blanca, 2: negra, 3: otra)
- smoke: indicador de si la madre fumaba durante el embarazo.
- ptl: número de embarazos previos de la madre.
- ht: indicador de si la madre es hipertensa.
- ui: indicador de irritabilidad uterina en la madre.
- ftw: número de visitas médicas realizadas durante el primer trimestre.
- bwt: peso del recién nacido en gramos.

Vamos a contrastar si la proporción de madres fumadoras supera el 30 %:

$$\left.\begin{array}{ll} H_0: & p = 0{,}3, \\ H_1: & p > 0{,}3, \end{array}\right\}$$

donde p representa la proporción de madres fumadoras.

En primer lugar consideramos una muestra de tamaño 30:

```
library(MASS)
set.seed(1001)
madres.elegidas=sample(1:189,30,replace=TRUE)
muestra.madres.elegidas=birthwt[madres.elegidas,]
```

A continuación vemos cuál es el número de "éxitos" o número de madres fumadoras:

```
table(muestra.madres.elegidas$smoke)
```

```
##
##  0  1
## 14 16
```

Tenemos un total de 16 madres fumadoras en nuestra muestra de 30 madres.

Por último realizamos el contraste planteado:

```
número.madres.fumadoras=table(muestra.madres.elegidas$smoke)[2]
binom.test(número.madres.fumadoras,30,p=0.3,alternative="greater")
```

```
##
##  Exact binomial test
##
## data:  número.madres.fumadoras and 30
## number of successes = 16, number of trials = 30, p-value = 0.0063703
## alternative hypothesis: true probability of success is greater than 0.3
## 95 percent confidence interval:
##  0.36994756 1.00000000
## sample estimates:
## probability of success
##             0.53333333
```

Como el p-valor del contraste es prácticamente nulo, concluimos que tenemos evidencias suficientes para afirmar que la proporción de madres fumadoras supera el 30 %.

Si nos fijamos en el intervalo de confianza para la proporción asociado al contraste:

```
binom.test(número.madres.fumadoras,30,p=0.3,alternative="greater")$conf.int
```

```
## [1] 0.36994756 1.00000000
## attr(,"conf.level")
## [1] 0.95
```

vemos que no contiene la proporción 0.3, hecho que nos reafirma la conclusión anterior.

4.4.2. Contrastes para proporciones cuando n es grande

Si indicamos con p la proporción poblacional y $\hat{p}_X$ la proporción muestral, sabemos que si la muestra es grande ($n \geq 40$) $Z = \frac{\hat{p}_X - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$.

Si la hipótesis nula $H_0 : p = p_0$ es verdadera, $Z = \frac{\hat{p}_X - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1)$.

Podemos usar los mismos p-valores que en el Z-test.

Se tiene que ir alerta con el intervalo de confianza. Si tenemos $n \geq 100$, $n\hat{p}_X \geq 10$ y $n(1-\hat{p}_X) \geq 10$, se puede usar el de Laplace. En caso contrario, se tiene que usar el de Wilson.

Ejercicio

Una asociación ganadera afirma que, en las matanzas caseras en las Baleares, como mínimo el 70 % de los cerdos han sido analizados de triquinosis.

En una investigación, se visita una muestra aleatoria de 100 matanzas y resulta que en 53 de éstas se ha realizado el análisis de triquinosis.

¿Podemos aceptar la afirmación de los ganaderos?

El contraste planteado es el siguiente:

$$\begin{cases} H_0 : p \geq 0{,}7, \\ H_1 : p < 0{,}7, \end{cases}$$

donde p representa la probabilidad de que en una matanza elegida al azar, ésta sea analizada de triquinosis.

El **estadístico de contraste** será:

$$Z = \frac{\hat{p}_X - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

cuyo valor es:

$$\hat{p}_X = \frac{53}{100} = 0{,}53 \Longrightarrow z_0 = \frac{0{,}53 - 0{,}7}{\sqrt{\frac{0{,}7 \cdot 0{,}3}{100}}} = -3{,}71.$$

El p**-valor** del contraste será:

$$P(Z \leq -3{,}71) = 0.$$

Decisión: como el p**-valor** es muy pequeño, rechazamos la hipótesis nula en favor de la alternativa.

¡Podemos afirmar con contundencia que la afirmación de los ganaderos es falsa!

El **intervalo de confianza** al 95 % de confianza será en este caso:

$$\left(-\infty, \hat{p}_X - z_{0{,}05}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}\right) = \left(-\infty, 0{,}53 - (-1{,}645) \cdot \sqrt{\frac{0{,}53 \cdot 0{,}47}{100}}\right) = (-\infty, 0{,}612)\,.$$

Informe: El p-valor de este contraste es prácticamente nulo y el intervalo de confianza del 95 % para la proporción p de matanzas donde se han hecho análisis de triquinosi es $(-\infty, 0{,}612)$.

Como el p-valor es muy pequeño y $0{,}7 \notin (-\infty, 0{,}612)$, hay evidencia muy significativa para rechazar que $p = 0{,}7$.

En `R` está implementado en la función `prop.test`, que además también sirve para contrastar dos proporciones por medio de muestras independientes grandes. Su sintaxis es

```
prop.test(x, n, p =..., alternative=..., conf.level=...)
```

donde:

- `x` puede ser dos cosas:
  - Un número natural: en este caso, `R` entiende que es el número de éxitos en una muestra.
  - Un vector de dos números naturales: en este caso, `R` entiende que es un contraste de dos proporciones y que éstos son los números de éxitos en las muestras.

- Cuando trabajamos con una sola muestra, `n` es su tamaño. Cuando estamos trabajando con dos muestras, `n` es el vector de dos entradas de sus tamaños.
- Cuando trabajamos con una sola muestra, `p` es la proporción poblacional que contrastamos. En el caso de un contraste de dos muestras, no hay que especificarlo.
- El significado de `alternative` y `conf.level`, y sus posibles valores, son los usuales.

La resolución del ejemplo anterior con `R` es la siguiente:

```
prop.test(53,100,p=0.7,alternative="less",conf.level=0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  53 out of 100, null probability 0.7
## X-squared = 12.9643, df = 1, p-value = 0.00015874
## alternative hypothesis: true p is less than 0.7
## 95 percent confidence interval:
##  0.00000000 0.61503639
## sample estimates:
##    p
## 0.53
```

`R` usa como estadístico de contraste Z^2 donde Z recordemos que es: $Z = \frac{\hat{p}_X - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

Si hacemos z_0^2 obtenemos:

```
z0=(0.53-0.7)/sqrt(0.7*(1-0.7)/100)
z0^2
```

```
## [1] 13.761905
```

No da exactamente el mismo valor en la salida de `R` de la función `prop.test` debido a que `R` hace una pequeña corrección a la continuidad.

Este hecho también se manifiesta en la pequeña diferencia que hay en los intervalos de confianza calculados a mano y en la salida de `R`.

4.5. Contrastes de hipótesis para el parámetro σ de una variable con distribución normal

Recordamos que si $X_1, \dots, X_n$ es una m.a.s. de una v.a. $X \sim N(\mu, \sigma)$, entonces el **estadístico** $\chi^2_{n-1} = \frac{(n-1)\widetilde{S}^2_X}{\sigma^2}$ sigue una distribución χ^2 con $n-1$ grados de libertad

Por lo tanto, si la hipótesis nula $H_0 : \sigma = \sigma_0$ es verdadera, $\chi^2_{n-1} = \frac{(n-1)\widetilde{S}^2_X}{\sigma_0^2}$ tendrá una distribución χ^2 con $n-1$ grados de libertad.

Calculamos su valor χ^2_0 sobre la muestra.

Nos planteamos los contrastes siguientes:

$$\begin{cases} H_0 : \sigma = \sigma_0, & (\text{ o } H_0 : \sigma \leq \sigma_0), \\ H_1 : \sigma > \sigma_0. \end{cases}$$

$$\begin{cases} H_0 : \sigma = \sigma_0, & (\text{ o } H_0 : \sigma \geq \sigma_0), \\ H_1 : \sigma < \sigma_0. \end{cases}$$

$$\begin{cases} H_0 : \sigma = \sigma_0, \\ H_1 : \sigma \neq \sigma_0. \end{cases}$$

Los **p-valores** serán los siguientes:

p-valor: $P(\chi^2_{n-1} \geq \chi^2_0)$.

p-valor: $P(\chi^2_{n-1} \leq \chi^2_0)$.

p-valor: $2\,\text{mín}\left\{P(\chi^2_{n-1} \leq \chi^2_0), P(\chi^2_{n-1} \geq \chi^2_0)\right\}$.

Ejercicio

Se han medido los siguientes valores en miles de personas para la audiencia de un programa de radio en $n = 10$ días:

$$521, 742, 593, 635, 788, 717, 606, 639, 666, 624$$

Contrastar si la varianza de la audiencia es 6400 al nivel de significación del 5 %, suponiendo que la población es normal.

El contraste de hipótesis planteado es el siguiente:

$$\begin{cases} H_0 : \sigma = \sqrt{6400} = 80, \\ H_1 : \sigma \neq 80. \end{cases}$$

El nivel de significación será: $\alpha = 0{,}05$

El **estadístico de contraste** es:

$$\chi^2_{n-1} = \frac{(n-1)\widetilde{S}^2_X}{\sigma^2_0}.$$

Su valor será:

```
x=c(521,742,593,635,788,717,606,639,666,624)
(chi02= (length(x)-1)*var(x)/6400)
```

```
## [1] 8.5945156
```

El p**-valor** será:

$$\begin{aligned} 2 \cdot P(\chi^2_9 \geq 8{,}595) &= 0{,}951, \\ 2 \cdot P(\chi^2_9 \leq 8{,}595) &= 1{,}049. \end{aligned}$$

Tomamos como p-valor el más pequeño: 0,951

Decisión: No podemos rechazar la hipótesis que la varianza sea 6400 al nivel de significación del 5 %.

El **intervalo de confianza** del 95 % de confianza será:

$$\left(\frac{(n-1)\widetilde{S}^2_X}{\chi^2_{n-1,0{,}975}}, \frac{(n-1)\widetilde{S}^2_X}{\chi^2_{n-1,0{,}025}}\right) = (2891{,}53, 20369{,}247)$$

Informe: El p-valor de este contraste es 0,951, y el intervalo de confianza del 95 % para la varianza σ^2 de la audiencia es $(2891{,}53, 20369{,}247)$.

Como el p-valor es muy grande y $6400 \in (2891{,}53, 20369{,}247)$, no hay evidencia que nos permita rechazar que $\sigma^2 = 6400$.

Dicho test está convenientemente implementado en la función `sigma.test` del paquete **TeachingDemos**.

Su sintaxis es la misma que la de la función `t.test` para una muestra, substituyendo el parámetro `mu` de `t.test` por el parámetro `sigma` (para especificar el valor de la desviación típica que contrastamos, σ_0) o `sigmasq` (por "sigma al cuadrado", para especificar el valor de la varianza que contrastamos, σ_0^2).

El ejemplo anterior se resolvería de la forma siguiente:

```
library(TeachingDemos)
sigma.test(x,sigma=80,alternative="two.sided",conf.level=0.95)
```

```
##
##  One sample Chi-squared test for variance
##
## data:  x
## X-squared = 8.6, df = 9, p-value = 1
## alternative hypothesis: true variance is not equal to 6400
## 95 percent confidence interval:
##   2892 20369
## sample estimates:
## var of x
##     6112
```

Ejemplo

Vamos a contrastar si la varianza de la amplitud del sépalo de las flores de la tabla de datos **iris** es menor que 0,2.

En primer lugar consideremos una muestra de 40 flores:

```
set.seed(2019)
flores.elegidas=sample(1:150,40,replace=TRUE)
muestra.flores.elegidas = iris[flores.elegidas,]
```

A continuación realizamos el contraste:

$$\left.\begin{array}{ll} H_0: & \sigma^2 = 0{,}2, \\ H_1: & \sigma^2 < 0{,}2, \end{array}\right\}$$

donde σ^2 representa la varianza de la amplitud del sépalo de las flores de la tabla de datos **iris**.

El contraste anterior, en `R`, se realiza de la forma siguiente:

```
library(TeachingDemos)
sigma.test(muestra.flores.elegidas$Sepal.Width,sigmasq = 0.2,alternative = "less")
```

```
##
##  One sample Chi-squared test for variance
```

```
##
## data:  muestra.flores.elegidas$Sepal.Width
## X-squared = 45, df = 39, p-value = 0.8
## alternative hypothesis: true variance is less than 0.2
## 95 percent confidence interval:
##  0.0000 0.3531
## sample estimates:
## var of muestra.flores.elegidas$Sepal.Width
##                                     0.2327
```

El p-valor del contraste ha sido 0.7763, valor muy superior a 0.1.

Concluimos por tanto, que no tenemos evidencias suficientes para aceptar que la varianza de la amplitud del sépalo sea menor que 0.2.

Si observamos el intervalo de confianza,

```
sigma.test(muestra.flores.elegidas$Sepal.Width,sigmasq = 0.2,
           alternative = "less")$conf.int
```

```
## [1] 0.0000 0.3531
## attr(,"conf.level")
## [1] 0.95
```

vemos que el valor 0.2 está en él, hecho que nos reafirma nuestra conclusión.

4.6. Contrastes de hipótesis para dos muestras

Queremos comparar el valor de un mismo parámetro en dos poblaciones.

Para ello dispondremos de una muestra para cada población.

Hay que tener en cuenta que las muestras pueden ser de dos tipos:

- **Muestras independientes:** las dos muestras se han obtenido de manera independiente.

Ejemplo

Probamos un medicamento sobre dos muestras de enfermos de características diferentes

- **Muestras emparejadas:** las dos muestras corresponden a los mismos individuos, o a individuos aparejados de alguna manera.

Ejemplo

Probamos dos medicamentos sobre los mismos enfermos.

4.6.1. Muestras independientes

Tenemos dos variables aleatorias (que representan los valores de la característica a estudiar sobre dos **poblaciones**).

Ejemplo

Poblaciones: Hombres y Mujeres. Característica a estudiar: estatura.

Queremos comparar el valor de un parámetro a las dos poblaciones

Ejemplo

¿Son, de media, los hombres más altos que las mujeres?

Lo haremos a partir de una m.a.s. de cada v.a., escogidas además de manera independiente.

4.7. Contrastes para dos medias poblacionales independientes μ_1 y μ_2

Tenemos dos v.a. X_1 y X_2, de medias μ_1 y μ_2

Tomamos una m.a.s. de cada variable:

$$\begin{array}{l} X_{1,1}, X_{1,2}, \dots, X_{1,n_1}, \text{ de } X_1 \\ X_{2,1}, X_{2,2}, \dots, X_{2,n_2}, \text{ de } X_2 \end{array}$$

Sean $\overline{X}_1$ y $\overline{X}_2$ sus medias, respectivamente.

La hipótesis nula será del tipo:

$$H_0 : \mu_1 = \mu_2, \text{ o, equivalentemente, } H_0 : \mu_1 - \mu_2 = 0.$$

Las hipótesis alternativas que nos plantearemos serán del tipo:

$$\begin{array}{l} \mu_1 < \mu_2, \text{ o, equivalentemente, } \mu_1 - \mu_2 < 0, \\ \mu_1 > \mu_2, \text{ o, equivalentemente, } \mu_1 - \mu_2 > 0, \\ \mu_1 \neq \mu_2, \text{ o, equivalentemente, } \mu_1 - \mu_2 \neq 0. \end{array}$$

4.7.1. Poblaciones normales o n grandes: σ conocidas

Suponemos una de las dos situaciones siguientes:

- X_1 y X_2 son normales, o n_1 y n_2 son grandes ($n_1, n_2 \geq 30$ o 40)

Suponemos que conocemos además las desviaciones típicas σ_1 y σ_2 de X_1 y X_2, respectivamente.

En este caso el **estadístico de contraste** es $Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$, que, si la hipótesis nula es cierta ($\mu_1 = \mu_2$), se distribuye según una $N(0,1)$.

Sea z_0 el valor del estadístico de contraste sobre la muestra. Los p-valores dependiendo de la hipótesis alternativa son:

- $H_1 : \mu_1 > \mu_2$: $p = P(Z \geq z_0)$.
- $H_1 : \mu_1 < \mu_2$: $p = P(Z \leq z_0)$.
- $H_1 : \mu_1 \neq \mu_2$: $p = 2 \cdot P(Z \geq |z_0|)$.

Ejemplo

Queremos comparar los tiempos de realización de una tarea entre estudiantes de dos grados G_1 y G_2, y contrastar si es verdad que los estudiantes de G_1 emplean menos tiempo que los de G_2

Suponemos que las desviaciones típicas son conocidas: $\sigma_1 = 1$ y $\sigma_2 = 2$

Disponemos de dos muestras independientes de tiempos realizados por estudiantes de cada grado, de tamaños $n_1 = n_2 = 40$. Calculamos las medias de los tiempos empleados en cada muestra (en minutos):

$$\overline{X}_1 = 9{,}789, \quad \overline{X}_2 = 11{,}385$$

El contraste planteado es el siguiente:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases} \iff \begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{cases}$$

El **estadístico de contraste** toma el valor: $z_0 = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{9{,}789-11{,}385}{\sqrt{\frac{1^2}{40}+\frac{2^2}{40}}} = -4{,}514.$

El p-valor será: $P(Z \leq -4{,}514) \approx 0$ muy pequeño.

Decisión: rechazamos la hipótesis de que son iguales, en favor de que los alumnos del grado G_1 tardan menos que los del grado G_2.

Si calculamos un intervalo de confianza del 95 % para la diferencia de medias $\mu_1 - \mu_2$ asociado al contraste anterior, obtenemos:

$$\begin{aligned} \left(-\infty, \overline{X}_1 - \overline{X}_2 - z_{0,05}\sqrt{\tfrac{\sigma_1^2}{n_1} + \tfrac{\sigma_2^2}{n_2}}\right) &= \left(-\infty, 9{,}789 - 11{,}385 + 1{,}645 \cdot \sqrt{\tfrac{1^2}{40} + \tfrac{2^2}{40}}\right) \\ &= (-\infty, -1{,}014). \end{aligned}$$

Observamos que el valor 0 no pertenece al intervalo de confianza anterior, hecho que nos hace reafirmar la decisión de rechazar $H_0 : \mu_1 - \mu_2 = 0$.

4.7.2. Poblaciones normales o n grandes: σ_1 o σ_2 desconocidas

Suponemos otra vez que estamos en una de las dos situaciones siguientes, pero ahora no conocemos σ_1 o σ_2:

- X_1 y X_2 son normales, o
- n_1 y n_2 son grandes ($n_1, n_2 \geq 40$).

Recordemos que disponemos de una m.a.s. de cada variable:

$$\begin{array}{l} X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}, \text{ de } X_1, \\ X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}, \text{ de } X_2. \end{array}$$

En este caso, tenemos que distinguir dos subcasos:

- Suponemos que $\sigma_1 = \sigma_2$.

- Suponemos que $\sigma_1 \neq \sigma_2$.

¿Como decidimos en qué caso estamos? Dos posibilidades:

- Realizamos los dos casos, y si dan lo mismo, es lo que contestamos.
- En caso de poblaciones normales, realizamos un contraste de igualdad de varianzas para decidir cuál es el caso.

Si suponemos que $\sigma_1 = \sigma_2$, el estadístico de contraste es

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \cdot \frac{((n_1-1)\widetilde{S}_1^2 + (n_2-1)\widetilde{S}_2^2)}{(n_1+n_2-2)}}},$$

que, cuando $\mu_1 = \mu_2$, tiene distribución (aproximadamente, en caso de muestras grandes) $t_{n_1+n_2-2}$.

Si suponemos que $\sigma_1 \neq \sigma_2$, el estadístico de contraste es $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\widetilde{S}_1^2}{n_1} + \frac{\widetilde{S}_2^2}{n_2}}} \sim t_f$, que, cuando $\mu_1 = \mu_2$, tiene distribución (aproximadamente, en caso de muestras grandes) t_f con

$$f = \left\lfloor \frac{\left(\frac{\widetilde{S}_1^2}{n_1} + \frac{\widetilde{S}_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{\widetilde{S}_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{\widetilde{S}_2^2}{n_2}\right)^2} \right\rfloor - 2.$$

Los p**-valores** usando las mismas expresiones que en el caso en que σ_1 y σ_2 conocidas sustituyendo el **estadístico de contraste** Z por el **estadístico de contraste** correspondiente.

Ejemplo

Queremos comparar los tiempos de realización de una tarea entre estudiantes de dos grados G_1 y G_2, y determinar si es verdad que los estudiantes de G_1 emplean menos tiempo que los de G_2 suponiendo que desconocemos una o las dos desviaciones típicas poblaciones σ_1 y σ_2.

Disponemos de dos muestras independientes de tiempos de tareas realizadas por estudiantes de cada grado de tamaños $n_1 = 40$ y $n_2 = 60$. Las medias y las desviaciones típicas muestrales de los tiempos empleados para cada muestra son:

$$\overline{X}_1 = 9{,}789,\ \overline{X}_2 = 11{,}385,\ \widetilde{S}_1 = 1{,}201,\ \widetilde{S}_2 = 1{,}579.$$

El contraste a realizar es el siguiente:

$$\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 < \mu_2, \end{cases} \iff \begin{cases} H_0 : \mu_1 - \mu_2 = 0, \\ H_1 : \mu_1 - \mu_2 < 0, \end{cases}$$

donde μ_1 y μ_2 representan los tiempos medios que tardan los estudiantes de los grados G_1 y G_2 para realizar la tarea, respectivamente.

Consideremos los dos casos anteriores:

- Caso 1: Suponemos $\sigma_1 = \sigma_2$.

El **estadístico de contraste** es: $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\cdot\frac{((n_1-1)\widetilde{S}_1^2+(n_2-1)\widetilde{S}_2^2)}{(n_1+n_2-2)}}} \sim t_{40+60-2} = t_{98}$, cuyo valor, usando los valores correspondientes de las muestras, será: $t_0 = \frac{9{,}789-11{,}385}{\sqrt{(\frac{1}{40}+\frac{1}{60})\frac{(39\cdot 1{,}201^2+59\cdot 1{,}579^2)}{98}}} = -5{,}428$.

El p-valor será, en este caso: $P(t_{78} < -5{,}428) \approx 0$, valor muy pequeño.

La decisión que tomamos, por tanto, es rechazar la hipótesis de que son iguales, en favor de que los estudiantes del grado G_1 tardan menos tiempo en realizar la tarea que los estudiantes del grado G_2.

Consideremos ahora el otro caso:

- Caso 2: Suponemos $\sigma_1 \neq \sigma_2$.

El **estadístico de contraste** será, en este caso: $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\widetilde{S}_1^2}{n_1}+\frac{\widetilde{S}_2^2}{n_2}}} \sim t_f$ donde

$$f = \left\lfloor \frac{\left(\frac{1{,}201^2}{40}+\frac{1{,}201^2}{60}\right)^2}{\frac{1}{39}\left(\frac{1{,}201^2}{40}\right)^2+\frac{1}{59}\left(\frac{1{,}579^2}{60}\right)^2} \right\rfloor - 2 = \lfloor 96{,}22 \rfloor - 2 = 94.$$

El valor que toma el estadístico anterior será:

$$t_0 = \frac{9{,}789-11{,}385}{\sqrt{\frac{1{,}201^2}{40}+\frac{1{,}579^2}{60}}} = -5{,}729.$$

El p**-valor** del contraste será: $P(t_{94} \leq -5{,}729) = 0$, valor muy pequeño.

La decisión que tomamos en este caso es la misma que en el caso anterior: rechazar la hipótesis de que los tiempos de ejecución son iguales, en favor de que los alumnos del grado G_1 tardan menos tiempo en realizar la tarea que los alumnos del grado G_2.

La decisión final, al haber decidido lo mismo en los dos casos, será concluir que los alumnos del grado G_1 tardan menos tiempo en realizar la tarea que los alumnos del grado G_2.

4.7.3. Contrastes para dos medias independientes en R: función `t.test`

Recordemos la sintaxis básica de la función `t.test` es

```
t.test(x, y, mu=..., alternative=..., conf.level=..., paired=...,
       var.equal=..., na.omit=...)
```

donde los nuevos parámetros para realizar un contraste de dos medias independientes son:

- `x` es el vector de datos de la primera muestra.
- `y` es el vector de datos de la segunda muestra.
- Podemos sustituir los vectores `x` e `y` por una fórmula `variable1~variable2` que indique que separamos la variable numérica `variable1` en dos vectores definidos por los niveles de un factor `variable2` de dos niveles (o de otra variable asimilable a un factor de dos niveles, como por ejemplo una variable numérica que solo tome dos valores diferentes).

- Parámetro `alternative`:
  - Si llamamos μ_x y μ_y a las medias de las poblaciones de las que hemos extraído las muestras x e y, respectivamente, entonces `"two.sided"` representa la hipótesis alternativa $H_1 : \mu_x \neq \mu_y$; `"less"` indica que la hipótesis alternativa es $H_1 : \mu_x < \mu_y$; y `"greater"`, que la hipótesis alternativa es $H_1 : \mu_x > \mu_y$.
- El parámetro `var.equal` solo lo tenemos que especificar si llevamos a cabo un contraste de dos medias usando muestras independientes, y en este caso sirve para indicar si queremos considerar las dos varianzas poblacionales iguales (igualándolo a TRUE) o diferentes (igualándolo a FALSE, que es su valor por defecto).

Ejercicio

Imaginemos ahora que nos planteamos si la media de la longitud del pétalo es la misma para las flores de las especies setosa y versicolor.

Para ello seleccionamos una muestra de tamaño 40 flores para cada especie:

```
set.seed(45)
flores.elegidas.setosa = sample(1:50,40,replace=TRUE)
flores.elegidas.versicolor = sample(51:100,40,replace=TRUE)
```

Las muestras serán las siguientes:

```
muestra.setosa = iris[flores.elegidas.setosa,]
muestra.versicolor = iris[flores.elegidas.versicolor,]
```

El contraste planteado se realiza de la forma siguiente:

```
t.test(muestra.setosa$Petal.Length,muestra.versicolor$Petal.Length,
       alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  muestra.setosa$Petal.Length and muestra.versicolor$Petal.Length
## t = -43, df = 50, p-value <0.0000000000000002
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.913 -2.652
## sample estimates:
## mean of x mean of y
##     1.407     4.190
```

El contraste realizado es de dos muestras independientes:

$$\left.\begin{array}{ll} H_0: & \mu_{setosa} = \mu_{versicolor}, \\ H_1: & \mu_{setosa} \neq \mu_{versicolor}, \end{array}\right\}$$

donde μ_{setosa} representa la media de la longitud del pétalo de las flores de la especie setosa y $\mu_{versicolor}$, la media de la longitud del pétalo de las flores de la especie versicolor.

El p-valor del contraste ha sido pràcticamente cero, lo que nos hace concluir que tenemos evidencias suficientes para concluir que las medias de la longitud del pétalo son diferentes para las dos especies.

De hecho, las medias de cada una de la dos muestras son 1.4075 y 4.19, valores muy diferentes.

El intervalo de confianza al 95 % de confianza para la diferencia de medias $\mu_{setosa} - \mu_{versicolor}$ asociado al contraste anterior vale, si nos fijamos en el "output" del `t.test`:

```
t.test(muestra.setosa$Petal.Length,muestra.versicolor$Petal.Length,
       alternative="two.sided")$conf.int
```

```
## [1] -2.913 -2.652
## attr(,"conf.level")
## [1] 0.95
```

intervalo que no contiene el valor cero y está totalmente a la izquierda de cero. Por tanto, debemos rechazar la hipótesis nula.

Fijémonos que hemos considerado que las varianzas de las dos variables son diferentes. Si las hubiésemos considerado iguales, tendríamos que hacer:

```
t.test(muestra.setosa$Petal.Length,muestra.versicolor$Petal.Length,
       alternative="two.sided",var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  muestra.setosa$Petal.Length and muestra.versicolor$Petal.Length
## t = -43, df = 78, p-value <0.0000000000000002
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.912 -2.653
## sample estimates:
## mean of x mean of y
##     1.407     4.190
```

En este caso, el p-valor también es despreciable, por lo que llegamos a la misma conclusión anterior: las medias son diferentes.

Más adelante veremos cómo realizar un contraste de varianzas para comprobar si éstas son iguales o no y por tanto, actuar en consecuencia con el parámetro `var.equal`.

4.8. Contrastes para dos proporciones p_1 y p_2

4.8.1. Test de Fisher

Tenemos dos variables aleatorias X_1 y X_2 Bernoulli de proporciones p_1 y p_2

Tomamos m.a.s. de cada una y obtenemos la tabla siguiente:

	X_1	X_2	Total
Éxitos	n_{11}	n_{12}	$n_{1\bullet}$
Fracasos	n_{21}	n_{22}	$n_{2\bullet}$

	X_1	X_2	Total
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

donde n_{11} es la cantidad de éxitos en la primera muestra, n_{12}, la cantidad de éxitos en la segunda muestra, n_{21}, la cantidad de fracasos en la primera muestra y n_{22}, la cantidad de fracasos en la segunda muestra.

De la misma forma, $n_{1\bullet}$, es la cantidad total de éxitos en las dos muestras y $n_{2\bullet}$ la cantidad total de fracasos en las dos muestras.

Por último, $n_{\bullet 1}$ es el tamaño de la primera muestra, $n_{\bullet 2}$, el tamaño de la segunda muestra y $n_{\bullet\bullet} = n_{\bullet 1} + n_{\bullet 2}$ es la suma de los dos tamaños.

Supongamos $p_1 = p_2$.

Para hallar la probabilidad de obtener n_{11} éxitos para la variable X_1 podemos razonar de la forma siguiente:

En una bolsa tenemos $n_{1\bullet}$ bolas E y $n_{2\bullet}$ bolas F. La probabilidad anterior sería la probabilidad de obtener n_{11} bolas E si escogemos $n_{\bullet 1}$ de golpe.

Sea X una variable hipergeométrica de parámetros $H(n_{1\bullet}, n_{2\bullet}, n_{\bullet 1})$. La probabilidad anterior sería: $P(X = n_{11})$.

Usaremos la variable anterior X como estadístico de contraste.

Nos planteamos los contrastes siguientes:

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 > p_2. \end{cases}$$

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 < p_2. \end{cases}$$

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 \neq p_2. \end{cases}$$

Los **p-valores** serán los siguientes:

p-valor: $P(H(n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}) \geq n_{11})$.

p-valor: $P(H(n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}) \leq n_{11})$.

p-valor: $2\,\text{mín}\{P(H \leq n_{11}), P(H \geq n_{11})\}$.

Ejemplo

Para determinar si el Síndrome de Muerte Repentina del Bebé (SIDS) tiene componiendo genético, se consideran los casos de SIDS en parejas de gemelos monocigóticos y dicigóticos. Sea:

- p_1: proporción de parejas de gemelos monocigóticos con algún caso de SIDS donde solo un hermano la sufrió.
- p_2: proporción de parejas de gemelos dicigóticos con algún caso de SIDS donde solo un hermano la sufrió.

Si el SIDS tiene componiendo genético, es de esperar que $p_1 < p_2$.

Nos piden realizar el contraste siguiente:

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 < p_2. \end{cases}$$

En un estudio (*Peterson et al, 1980*), se obtuvieron los datos siguientes:

Casos de SIDS	Monocigóticos	Dicigóticos	Total
Uno	23	35	58
Dos	1	2	3
Total	24	37	61

El **p-valor** del contraste anterior sería: $P(H(58, 3, 24) \leq 23)$:

```
phyper(23,58,3,24)
```

```
## [1] 0.7841
```

Al obtener un p-valor grande, podemos concluir que no tenemos evidencias suficientes para rechazar la hipótesis nula y por tanto, el SID no tiene componente genética.

- El test exacto de Fisher está implementado en la función `fisher.test`. Su sintaxis es

```
fisher.test(x, alternative=..., conf.level=...)
```

donde

- `x` es la matriz anterior, donde recordemos que los números de éxitos van en la primera fila y los de fracasos en la segunda, y las poblaciones se ordenan por columnas.

Ejercicio

Realicemos el contraste anterior de igualdad de proporciones de madres fumadores de raza blanca y negra usando el test de Fisher.

En primer lugar calculamos las etiquetas de las madres de cada raza:

```
madres.raza.blanca = rownames(birthwt[birthwt$race==1,])
madres.raza.negra = rownames(birthwt[birthwt$race==2,])
```

Seguidamente, elegimos las muestras de tamaño 50 de cada raza y creamos las muestras correspondientes:

```
set.seed(2000)
madres.elegidas.blanca=sample(madres.raza.blanca,50,replace=TRUE)
madres.elegidas.negra = sample(madres.raza.negra,50, replace=TRUE)
muestra.madres.raza.blanca = birthwt[madres.elegidas.blanca,]
muestra.madres.raza.negra = birthwt[madres.elegidas.negra,]
```

Definimos ahora una nueva tabla de datos que contenga la información de las dos muestras consideradas:

```
muestra.madres = rbind(muestra.madres.raza.blanca,muestra.madres.raza.negra)
```

A continuación calculamos la matriz para usar en el test de Fisher:

```
(matriz.fisher=table(muestra.madres$smoke,muestra.madres$race))
```

```
##
##      1  2
##   0 24 33
##   1 26 17
```

La matriz anterior no es correcta ya que la primera fila debería ser la fila de "éxitos" y es la fila de "fracasos".

Lo arreglamos permutando las filas:

```
(matriz.fisher = rbind(matriz.fisher[2,],matriz.fisher[1,]))
```

```
##       1  2
## [1,] 26 17
## [2,] 24 33
```

Por último realizamos el contraste:

```
fisher.test(matriz.fisher)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  matriz.fisher
## p-value = 0.1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8723 5.1038
## sample estimates:
## odds ratio
##      2.087
```

El p-valor del contraste ha sido 0.1056, valor mayor que 0.1. Concluimos que no tenemos evidencias para rechazar que las proporciones de madres fumadoras de razas blanca y negra sean iguales.

O, dicho de otra manera, no rechazamos la hipótesis nula de igualdad de proporciones.

Ejercicio

Como el test de Fisher es exacto, dejamos como ejercicio repetir el experimento anterior pero en lugar de tomando muestras de tamaño 50, tomando muestras de tamaño más pequeño como por ejemplo 10.

¡Atención!

Hay que ir con cuidado con la interpretación del intervalo de confianza que da esta función: no es ni para la diferencia de las proporciones ni para su cociente, sino para su **odds ratio**: el cociente

$$\left(\frac{p_b}{1-p_b}\right)\Big/\left(\frac{p_n}{1-p_n}\right).$$

4.8.2. Introducción a las odds

Odds

El **odds** de un suceso A es el cociente

$$\text{Odds}(A) = \frac{P(A)}{1-P(A)},$$

donde $P(A)$ es la probabilidad que suceda A y mide cuántas veces es más probable A que su contrario.

Las *odds* son una función creciente de la probabilidad, y por lo tanto

$$\text{Odds}(A) < \text{Odds}(B) \iff P(A) < P(B).$$

Esto permite comparar *odds* en vez de probabilidades, con la misma conclusión.

Por ejemplo, en nuestro caso, como el intervalo de confianza para la *odds ratio* va de 0.8723 a 5.1038. En particular, contiene el 1, por lo que no podemos rechazar que

$$\Big(\frac{p_b}{1-p_b}\Big)\Big/\Big(\frac{p_n}{1-p_n}\Big) = 1,$$

es decir, no podemos rechazar que

$$\frac{p_b}{1-p_b} = \frac{p_n}{1-p_n}$$

y esto es equivalente a $p_b = p_n$.

Si, por ejemplo, el intervalo de confianza hubiera ido de 0 a 0.8, entonces la conclusión a este nivel de confianza hubiera sido que

$$\Big(\frac{p_b}{1-p_b}\Big)\Big/\Big(\frac{p_n}{1-p_n}\Big) < 1$$

es decir, que

$$\frac{p_b}{1-p_b} < \frac{p_n}{1-p_n}$$

y esto es equivalente a $p_b < p_n$.

4.8.3. Contraste para dos proporciones: muestras grandes

Supongamos ahora que tenemos dos variables aleatorias X_1 y X_2 de Bernoulli de parámetros p_1 y p_2.

Consideremos una m.a.s. de cada variable aleatoria de tamaños n_1 y n_2, respectivamente, grandes ($n_1, n_2 \geq 50$ o 100):

$$X_{1,1}, X_{1,2}, \dots, X_{1,n_1}, \text{ de } X_1,$$
$$X_{2,1}, X_{2,2}, \dots, X_{2,n_2}, \text{ de } X_2.$$

Sean $\hat{p}_1$ y $\hat{p}_2$ sus proporciones muestrales.

Suponemos que los números de éxitos y de fracasos en cada muestra son ≥ 5 o 10).

Nos planteamos los contrastes siguientes como en el caso del test de Fisher:

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 > p_2. \end{cases}$$

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 < p_2. \end{cases}$$

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 \neq p_2. \end{cases}$$

El **estadístico de contraste** para los contrastes anteriores es:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\left(\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(1-\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

que, usando el *Teorema Central del Límite* y suponiendo cierta la hipótesis nula $H_0 : p_1 = p_2$, tiene aproximadamente una distribución $N(0, 1)$.

Sea z_0 el valor del **estadístico de contraste** usando las proporciones muestrales $\widehat{p}_1$ y $\widehat{p}_2$.

Los **p-valores** serán los siguientes:

p-valor: $P(Z \geq z_0)$.

p-valor: $P(Z \leq z_0)$.

p-valor: $2P(Z \geq |z_0|)$.

Ejercicio

Se toman una muestra de ADN de 100 individuos con al menos tres generaciones familiares en la isla de Mallorca, y otra de 50 individuos con al menos tres generaciones familiares en la isla de Menorca.

Se quiere saber si un determinado alelo de un gen es presente con la misma proporción en las dos poblaciones.

En la muestra mallorquina, 20 individuos lo tienen, y en la muestra menorquina, 12.

Contrastar la hipótesis de igualdad de proporciones al nivel de significación 0,05, y calcular el intervalo de confianza para la diferencia de proporciones para este α.

Fijémonos que los tamaños de las muestras (100 y 50) son bastante grandes

El contraste pedido es el siguiente:

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 \neq p_2, \end{cases}$$

donde p_1 y p_2 representan las proporciones de individuos que tienen el alelo en el gen para los individuos de la isla de Mallorca y Menora, respectivamente.

El **estadístico de contraste** será: $Z = \frac{\widehat{p}_1-\widehat{p}_2}{\sqrt{\left(\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(1-\frac{n_1\widehat{p}_1+n_2\widehat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$.

Las proporciones muestrales serán: $\widehat{p}_1 = \frac{20}{100} = 0{,}2$, $\widehat{p}_2 = \frac{12}{50} = 0{,}24$.

Si hallamos el valor que toma el **estadístico de contraste** para las proporciones muestrales anteriores, obtenemos:

$$z_0 = \frac{0{,}2 - 0{,}24}{\sqrt{\left(\frac{20+12}{100+50}\right)\left(1 - \frac{20+12}{100+50}\right)\left(\frac{1}{100} + \frac{1}{50}\right)}} = -0{,}564.$$

El p**-valor** será: $2 \cdot P(Z \geq |-0{,}564|) = 0{,}573$.

Decisión: como el p-valor es grande y mayor que $\alpha = 0{,}05$, aceptamos la hipótesis que las dos proporciones son la misma al no tener evidencias suficientes para rechazarla.

El intervalo de confianza para $p_1 - p_2$ al nivel de confianza $(1-\alpha)\cdot 100\,\%$ en un contraste bilateral es

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

que, en nuestro caso será:

$$(0{,}2 - 0{,}24 - 1{,}96 \cdot 0{,}071, 0{,}2 - 0{,}24 + 1{,}96 \cdot 0{,}071) = (-0{,}179, 0{,}099).$$

Observemos que contiene el 0. Por tanto no podemos rechazar que $p_1 - p_2 = 0$ llegando a la misma conclusión que con el p**-valor**.

- En `R` está implementado en la función `prop.test`, que además también sirve para contrastar dos proporciones por medio de muestras independientes grandes. Su sintaxis es

```
prop.test(x, n, p =..., alternative=..., conf.level=...)
```

donde:

- `x` en el caso de un contraste de dos proporciones es un vector de dos números naturales cuyas componentes son los números de éxitos en las dos muestras.
- Cuando estamos trabajando con dos muestras, `n` es el vector de dos entradas de sus tamaños.
- El significado de `alternative` y `conf.level`, y sus posibles valores, son los usuales.

Ejemplo

Siguiendo el ejemplo anterior, contrastemos otra vez si la proporción de madres fumadoras de raza blanca es la misma que la proporción de madres fumadoras de raza negra pero usando ahora la función `prop.test`.

En primer lugar, calculamos cuántas madres fumadores hay de cada muestra:

```
table(muestra.madres.raza.blanca$smoke)
```

```
##
##  0  1
## 24 26
```

```
table(muestra.madres.raza.negra$smoke)
```

```
##
##  0  1
## 33 17
```

```
n.blanca = table(muestra.madres.raza.blanca$smoke)[2] ## número de madres fumadoras
## de raza blanca
n.negra = table(muestra.madres.raza.negra$smoke)[2] ## número de madres fumadoras
## de raza negra
```

Tenemos un total de 26 madres fumadoras de raza blanca entre las 50 de la muestra y 17 madres fumadores de raza negra entre las 50 de la muestra.

Finalmente, realizamos el contraste planteado:

$$\left.\begin{array}{ll} H_0: & p_b = p_n, \\ H_1: & p_b \neq p_n, \end{array}\right\}$$

donde p_b y p_n representan las proporciones de madres fumadoras de raza blanca y negra, respectivamente.

El contraste en R se realizaría de la forma siguiente:

```
prop.test(c(n.blanca,n.negra),c(50,50))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(n.blanca, n.negra) out of c(50, 50)
## X-squared = 2.6, df = 1, p-value = 0.1
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.03083  0.39083
## sample estimates:
## prop 1 prop 2
##   0.52   0.34
```

El p-valor del contraste ha sido 0.1061, muy parecido al del test de Fisher, y mayor que 0.1. Concluimos otra vez que no tenemos evidencias para rechazar que las proporciones de madres fumadoras de razas blanca y negra sean iguales.

Si nos fijamos en el intervalo de confianza para la diferencia de proporciones:

```
prop.test(c(n.blanca,n.negra),c(50,50))$conf.int
```

```
## [1] -0.03083  0.39083
## attr(,"conf.level")
## [1] 0.95
```

vemos que el 0 está dentro de dicho intervalo, hecho que reafirma nuestra conclusión.

4.9. Contrastes de dos muestras más generales

Dado un parámetro θ (θ puede ser la media μ, la proporción p, etc.) y dadas dos poblaciones X_1 y X_2 cuyas distribuciones dependen de parámetros θ_1 y θ_2, hemos realizado contrastes en los que la hipótesis nula era de la forma $H_0 : \theta_1 = \theta_2$, o $H_0 : \theta_1 - \theta_2 = 0$.

Existen contrastes más generales del tipo:

$$\begin{cases} H_0 : \theta_1 - \theta_2 = \Delta \\ H_1 : \theta_1 - \theta_2 < \Delta \text{ o } \theta_1 - \theta_2 > \Delta \text{ o } \theta_1 - \theta_2 \neq \Delta \end{cases}$$

con $\Delta \in \mathbb{R}$.

4.9.1. Cambios en los estadísticos de contraste

Para realizar los contrastes anteriores, se pueden usar los mismos **estadísticos** que en el caso en que $H_0 : \theta_1 - \theta_2 = 0$ realizando los cambios siguientes:

- Si $\theta = \mu$, la media, hay que sustituir $\overline{X}_1 - \overline{X}_2$ en el numerador del **estadístico** por $\overline{X}_1 - \overline{X}_2 - \Delta$.
- Si $\theta = p$, proporción muestral, hay que sustituir $\hat{p}_1 - \hat{p}_2$ en el numerdor del **estadístico** por $\hat{p}_1 - \hat{p}_2 - \Delta$.

Ejemplo

Tenemos dos tratamientos, A y B, de una dolencia. Tratamos 50 enfermos con A y 100 con B. 20 enfermos tratados con A y 25 tratados con B manifiestan haber sentido malestar general durante los 7 días posteriores a iniciar el tratamiento.

¿Podemos concluir, a un nivel de significación del 5 %, que A produce malestar general en una proporción de los enfermos que es 5 puntos porcentuales superior a la proporción de los enfermos en que lo produce B?

Sean p_1 la proporción de enfermos en que A produce malestar general y p_2, la proporción de enfermos en que B produce malestar general.

El contraste a realizar es el siguiente:

$$\begin{cases} H_0 : p_1 \leq p_2 + 0{,}05, \\ H_1 : p_1 > p_2 + 0{,}05. \end{cases}$$

El **estadístico de contraste** es el siguiente:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - \Delta}{\sqrt{\left(\frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}\right)\left(1 - \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

que, si la hipótesis nula es cierta, sigue aproximadamente la distribución $N(0, 1)$.

Las proporciones y los tamaños muestrales son: $\hat{p}_1 = 0{,}4$, $\hat{p}_2 = 0{,}25$, $n_1 = 50$, $n_2 = 100$ y el valor de Δ será $\Delta = 0{,}05$.

El valor que toma el **estadístico de contraste** es:

$$z_0 = \frac{0{,}4 - 0{,}25 - 0{,}05}{\sqrt{\left(\frac{20+25}{50+100}\right)\left(1 - \frac{20+25}{50+100}\right)\left(\frac{1}{50} + \frac{1}{100}\right)}} = 1{,}26.$$

El p**-valor** del contraste será: $P(Z \geq 1{,}26) = 0{,}104$.

Decisión: como el p**-valor** es relativamente grande y mayor que $\alpha = 0{,}05$, no tenemos indicios para rechazar la hipótesis que $p_1 - p_2$ es inferior o igual a un 5 %.

Si hallamos el **intervalo de confianza** para $p_1 - p_2$ al nivel de confianza $(1-\alpha)\cdot 100\,\%$, obtenemos:

$$\left(\hat{p}_1 - \hat{p}_2 + z_\alpha \sqrt{\left(\frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}\right)\left(1 - \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \infty\right) =$$
$$\left(0{,}4 - 0{,}25 - 1{,}645\sqrt{\left(\frac{50\cdot 0{,}4 + 100\cdot 0{,}25}{50+100}\right)\left(1 - \frac{50\cdot 0{,}4 + 100\cdot 0{,}25}{50+100}\right)\left(\frac{1}{50} + \frac{1}{100}\right)}, \infty\right) =$$
$$(0{,}019, \infty)$$

Nos fijamos que el intervalo anterior contiene el valor $\Delta = 0{,}05$, razón que nos reafirma la decisión tomada de no rechazar que $p_1 \leq p_2 + 0{,}05$ pero, en cambio, no contiene el valor 0 y por tanto, podríamos rechazar que $p_1 = p_2$.

4.10. Contrastes para dos varianzas

Dadas dos poblaciones de distribución normal e indpendientes, nos planteamos si las varianzas de dichas poblaciones son iguales o diferentes.

Una aplicación del contraste de varianzas es decidir qué opción elegir en el marco de una comparación de medias de muestras independientes.

Tenemos dos variables aleatorias X_1 y X_2 normales de desviaciones típicas σ_1, σ_2 desconocidas

Suponemos que tenemos una m.a.s de cada variable:

$$X_{1,1}, X_{1,2}, \ldots, X_{1,n_1} \text{ de } X_1$$
$$X_{2,1}, X_{2,2}, \ldots, X_{2,n_2} \text{ de } X_2$$

Sean $\widetilde{S}_1^2$ y $\widetilde{S}_2^2$ sus varianzas muestrales.

Nos planteamos los contrastes siguientes:

$$\begin{cases} H_0 : \sigma_1 = \sigma_2, & \left(\text{ o } H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1\right), \\ H_1 : \sigma_1 > \sigma_2. \end{cases}$$

$$\begin{cases} H_0 : \sigma_1 = \sigma_2, & \left(\text{ o } H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1\right), \\ H_1 : \sigma_1 < \sigma_2. \end{cases}$$

$$\begin{cases} H_0 : \sigma_1 = \sigma_2, \\ H_1 : \sigma_1 \neq \sigma_2. \end{cases} \quad \left(\text{o } H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \right),$$

Se emplea el siguiente **estadístico de contraste**:

$$F = \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2}$$

que, si las dos poblaciones son normales y la hipótesis nula $H_0 : \sigma_1 = \sigma_2$ es cierta, tiene distribución F de Fisher con grados de libertad $n_1 - 1$ y $n_2 - 1$.

Sea f_0 el valor que toma usando las desviaciones típicas muestrales.

La distribución F de Fisher

La distribución $F_{n,m}$ de Fisher, donde n, m son los grados de libertad se define como el cociente de dos variables chi2 independientes de n y m grados de libertad, respectivamente: χ_n^2/χ_m^2.

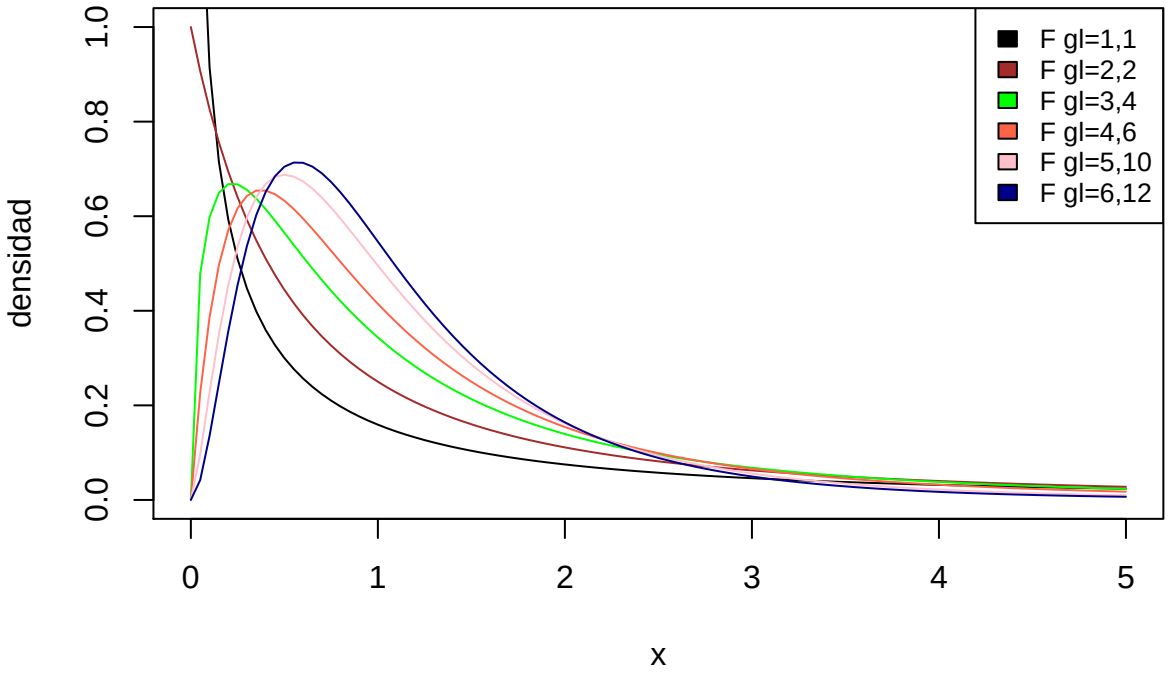
Su función de densidad tiene la siguiente expresión:

$$f_{F_{n,m}}(x) = \frac{\Gamma\left(\frac{n+m}{2}\right) \cdot \left(\frac{m}{n}\right)^{m/2} x^{(m-2)/2}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)\left(1 + \frac{m}{n}x\right)^{(m+n)/2}}, \text{ si } x \geq 0,$$

donde $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$, si $x > 0$.

Se trata de una distribución no simétrica.

Gráfica de la función de densidad de algunas distribuciones F de Fisher.



Los p**-valores** asociados a los contrastes anteriores son:

p-valor: $P(F_{n_1-1,n_2-1} \geq f_0)$.

p-valor: $P(F_{n_1-1,n_2-1} \leq f_0)$.

p-valor: mín$\{2 \cdot P(F_{n_1-1,n_2-1} \leq f_0), 2 \cdot P(F_{n_1-1,n_2-1} \geq f_0)\}$.

Ejercicio

Consideramos el ejemplo donde queríamos comparar los tiempos de realización de una tarea entre estudiantes de dos grados G_1 y G_2. Suponemos que estos tiempos siguen distribuciones normales.

Disponemos de dos muestras independientes de los tiempos usados por los estudiantes de cada grado para realizar la tarea. Los tamaños de cada muestra son $n_1 = n_2 = 40$.

Las desviaciones típicas muestrales de los tiempos empleados para cada muestra son:

$$\widetilde{S}_1 = 1{,}201, \quad \widetilde{S}_2 = 1{,}579$$

Contrastar la hipótesis de igualdad de varianzas al nivel de significación 0,05.

El contraste planteado es el siguiente:

$$\begin{cases} H_0 : \sigma_1 = \sigma_2, \\ H_1 : \sigma_1 \neq \sigma_2, \end{cases}$$

donde σ_1 y σ_2 son las desviaciones típicas de los tiempos empleados para realizar la tarea por los estudiantes de los grados G_1 y G_2, respectivamente.

El **estadístico de contraste** para el contraste anterior es: $F = \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2} \sim F_{39,39}$.

Dicho estadístico toma el siguiente valor: $f_0 = \frac{1{,}201^2}{1{,}579^2} = 0{,}579$.

El p**-valor** para el contraste anterior será:

$$\begin{array}{l} \text{mín}\{2 \cdot P(F_{n_1-1,n_2-1} \leq f_0), 2 \cdot P(F_{n_1-1,n_2-1} \geq f_0)\} = \\ \text{mín}\{2 \cdot P(F_{n_1-1,n_2-1} \leq 0{,}579), 2 \cdot P(F_{n_1-1,n_2-1} \geq 0{,}579)\} = \\ \text{mín}\{0{,}091, 1{,}909\} = 0{,}091. \end{array}$$

Decisión: como que el p-valor es moderado pero mayor que $\alpha = 0{,}05$, no podemos rechazar la hipótesis que las dos varianzas sean iguales.

Concluimos que no tenemos evidencias suficientes para rechazar que $\sigma_1 = \sigma_2$.

Por tanto, en el contraste de las dos medias, tendríamos que suponer que las varianzas de las dos poblaciones son la misma.

El **intervalo de confianza** para $\frac{\sigma_1^2}{\sigma_2^2}$ al nivel de confianza $(1-\alpha) \cdot 100\,\%$ es

$$\left(\frac{\widetilde{S}_1^2}{\widetilde{S}_2^2} \cdot F_{n_1-1,n_2-1,\frac{\alpha}{2}}, \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2} \cdot F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} \right) = \left(\frac{1{,}201^2}{1{,}579^2} \cdot F_{39,39,0{,}025}, \frac{1{,}201^2}{1{,}579^2} \cdot F_{39,39,0{,}975} \right) = (0{,}306, 1{,}094)$$

Observemos que el intervalo de confianza anterior contiene el valor 1, hecho que reafirma la decisión tomada de no rechazar la hipótesis de igualdad de varianzas.

Ejemplo Se desea comparar la actividad motora espontánea de un grupo de 25 ratas control y otro de 36 ratas desnutridas. Se midió el número de veces que pasaban ante una célula fotoeléctrica durante 24 horas. Los datos obtenidos fueron los siguientes:

	n	$\overline{X}$	$\widetilde{S}$
1. Control	25	869,8	106,7
2. Desnutridas	36	665	133,7

¿Se observan diferencias significativas entre el grupo de control y el grupo desnutrido?

Supondremos que los datos anteriores provienen de poblaciones normales.

El contraste a realizar es el siguiente:

$$\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 \neq \mu_2, \end{cases}$$

donde μ_1 y μ_2 representan los valores medios del número de veces que las ratas de control y desnutridas pasan ante la célula fotoeléctrica, respectivamente.

Antes de nada, tenemos que averiguar si las varianzas de los dos grupos son iguales o no ya que es un parámetro a usar en el contraste a realizar.

Por tanto, en primer lugar, realizaremos el contraste:

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{cases}$$

donde σ_1 y σ_2 representan las desviaciones típicas del número de veces que las ratas de control y desnutridas pasan ante la célula fotoeléctrica, respectivamente.

El **Estadístico de contraste** para el contraste anterior vale: $F = \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2} \sim F_{24,35}$.

El valor que toma es el siguiente: $f_0 = \frac{106,7^2}{133,7^2} = 0{,}637$.

El p**-valor** para el contraste anterior vale:

$$\begin{array}{l} \text{mín}\{2 \cdot P(F_{n_1-1,n_2-1} \leq f_0), 2 \cdot P(F_{n_1-1,n_2-1} \geq f_0)\} = \\ \text{mín}\{2 \cdot P(F_{n_1-1,n_2-1} \leq 0{,}637), 2 \cdot P(F_{n_1-1,n_2-1} \geq 0{,}637)\} = \\ \text{mín}\{0{,}251, 1{,}749\} = 0{,}251. \end{array}$$

El p**-valor** es un valor grande, por tanto, concluimos que no podemos rechazar la hipótesis nula y decidimos que las varianzas de las dos poblaciones son iguales.

Realicemos a continuación el contraste pedido:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

El **estadístico de contraste** al suponer que $\sigma_1 = \sigma_2$, será: $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \cdot \frac{(n_1-1)\widetilde{S}_1^2 + (n_2-1)\widetilde{S}_2^2}{n_1+n_2-2}}} \sim t_{59}$.

El valor que toma dicho estadístico en los valores muestrales vale: $t_0 = \frac{869{,}8-665}{\sqrt{(\frac{1}{25} + \frac{1}{36}) \cdot \frac{24 \cdot 106,7^2 + 35 \cdot 133,7^2}{25+36-2}}} = 6{,}373$.

El p**-valor** del contraste será: $p = 2 \cdot P(t_{59} \geq 6{,}373) \approx 0$.

Decisión: como el p**-valor** es prácticamente nulo, concluimos que tenemos evidencias suficientes para rechazar la hipótesis nula y por tanto hay diferencias entre las ratas de control y las desnutridas entre el número de veces que pasan ante la célula fotoeléctrica.

4.10.1. Contrastes para varianzas en R

La función para efectuar este test en R es `var.testy` su sintaxis básica es la misma que la de `t.test` para dos muestras:

```
var.test(x, y, alternative=..., conf.level=...)
```

donde `x` e `y` son los dos vectores de datos, que se pueden especificar mediante una fórmula como en el caso de `t.test`, y el parámetro `alternative` puede tomar los tres mismos valores que en los tests anteriores.

Ejercicio

Recordemos que cuando explicábamos el contraste para dos medias independientes, contrastamos si las medias de las longitudes del pétalo para las especies setosa y versicolor eran iguales o no pero necesitábamos saber si las varianzas eran iguales o no para poder tenerlo en cuenta en la función `t.test`.

Veamos ahora si podemos considerar las varianzas iguales o no.

Las muestras eran `muestra.setosa` y `muestra.versicolor`.

Realicemos el contraste de igualdad de varianzas:

```
var.test(muestra.setosa$Petal.Length,muestra.versicolor$Petal.Length)
```

```
##
##  F test to compare two variances
##
## data:  muestra.setosa$Petal.Length and muestra.versicolor$Petal.Length
## F = 0.14, num df = 39, denom df = 39, p-value = 0.00000002
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.0758 0.2710
## sample estimates:
## ratio of variances
##             0.1433
```

El p-valor del contraste ha sido prácticamente cero. Por tanto, concluimos que tenemos evidencias suficientes para afirmar que las varianzas de las longitudes del pétalo de las flores de las especies setosa y versicolor son diferentes.

Si nos fijamos en el intervalo de confianza en el cociente de varianzas $\frac{\sigma^2_{setosa}}{\sigma^2_{versicolor}}$,

```
var.test(muestra.setosa$Petal.Length,muestra.versicolor$Petal.Length)$conf.int
```

```
## [1] 0.0758 0.2710
## attr(,"conf.level")
## [1] 0.95
```

vemos que no contiene el valor 1, de hecho está a la izquierda de él. Este hecho nos hace reafirmar la conclusión anterior.

Para que el contraste anterior tenga sentido, hemos de suponer que las longitudes del pétalo de las flores de las especies setosa y versicolor siguen distribuciones normales.

- Hemos insistido en que el test F solo es válido si las dos poblaciones cuyas varianzas comparamos son normales.
- ¿Qué podemos hacer si dudamos de su normalidad? Usar un test no paramétrico que no presuponga esta hipótesis.
- Hay diversos tests no paramétricos para realizar contrastes bilaterales de dos varianzas. Aquí os recomendamos el **test de Fligner-Killeen**, implementado en la función `fligner.test`.
    - Se aplica o bien a una `list` formada por las dos muestras, o bien a una fórmula que separe un vector numérico en dos muestras por medio de un factor de dos niveles.

Realicemos el contraste previo de igualdad de varianzas usando el test no paramétrico anterior para ver si llegamos a la misma conclusión:

```
fligner.test(list(muestra.setosa$Petal.Length,muestra.versicolor$Petal.Length))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(muestra.setosa$Petal.Length, muestra.versicolor$Petal.Length)
## Fligner-Killeen:med chi-squared = 22, df = 1, p-value = 0.000002
```

Como el p-valor vuelve a ser insignificante, llegamos a la misma conclusión anterior: tenemos evidencias suficientes para afirmar que las varianzas de las longitudes del pétalo de las flores de las especies setosa y versicolor son diferentes.

La ventaja de este test es que no necesitamos la normalidad de las muestras, aunque su potencia, que explicaremos más adelante, sea inferior.

4.11. Muestras emparejadas

Las muestras consideradas hasta el momento se han supuesto **independientes**.

Un caso completamente diferente es cuando las dos muestras corresponden a los mismos individuos o a individuos emparejados por algún factor.

Ejemplos:

- Se estudia el estado de una dolencia a los mismos individuos antes y después de un tratamiento.
- Se mide la incidencia de cáncer en parejas de hermanos gemelos.

En estos casos, se habla de **muestras emparejadas**, o **paired samples** en inglés.

Para decidir si hay diferencias entre los valores de dos **muestras emparejadas**, el contraste más común consiste a calcular las diferencias de los valores de cada una de las parejas de muestras y realizar un contraste para averiguar si la media de las diferencias es 0.

Observación: El **diseño experimental** para realizar un contraste de **muestras emparejadas** se tiene que fijar **antes** de la **recogida de datos**.

4.11.1. Contrastes de medias de muestras emparejadas

En el caso de un contraste de muestras emparejadas, sean X_1 y X_2 las variables correspondientes y sean

$$X_{1,1}, X_{1,2}, \dots, X_{1,n}, \text{ de } X_1$$
$$X_{2,1}, X_{2,2}, \dots, X_{2,n}, \text{ de } X_2$$

las m.a.s. de cada una de las variables correspondientes a las dos muestras.

Fijémonos que, al ser la muestras emparejadas, los tamaños de las mismas deben ser iguales.

Consideramos la variable diferencia $D = X_1 - X_2$. La m.a.s. de D construida a partir de las muestras anteriores será:

$$D_1 = X_{1,1} - X_{2,1},\ D_2 = X_{1,2} - X_{2,2}, \dots, D_n = X_{1,n} - X_{2,n}.$$

Los contastes planteados son los siguientes:

$$\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 > \mu_2. \end{cases}$$

$$\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 < \mu_2. \end{cases}$$

$$\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 \neq \mu_2. \end{cases}$$

que, escritos en términos de la media de la variable diferencia D, μ_d, serán:

$$\begin{cases} H_0 : \mu_d = 0, \\ H_1 : \mu_d > 0. \end{cases}$$

$$\begin{cases} H_0 : \mu_d = 0, \\ H_1 : \mu_d < 0. \end{cases}$$

$$\begin{cases} H_0 : \mu_d = 0, \\ H_1 : \mu_d \neq 0. \end{cases}$$

O sea, hemos reducido un contraste de medias de dos muestras dependientes a un contraste de una sola media de una sola muestra.

A partir de aquí, podemos calcular los **p-valores** y los **intervalos de confianza** de los contrates anteriores usando las expresiones de los contrastes de una media de una sola media vistos anteriormente.

Ejemplo de medias emparejadas

Disponemos de dos algoritmos de alineamiento de proteínas. Los dos producen resultados de la misma calidad.

Estamos interesados en saber cuál de los dos algoritmos es *más eficiente*, en el sentido de tener un tiempo de ejecución más corto. Suponemos que dichos tiempos de ejecución siguen leyes normales.

Tomamos una muestra de proteínas y les aplicamos los dos algoritmos, anotando los tiempos de ejecución sobre cada proteína.

Los resultados obtenidos son:

	1	2	3	4	5	6	7	8	9	10
algoritmo 1	8.1	11.9	11.4	12.9	9.0	7.2	12.4	6.9	8.9	8.3
algoritmo 2	6.9	6.7	8.3	8.6	18.9	7.9	7.4	8.7	7.9	12.4
diferencias	1.2	5.2	3.1	4.3	-9.9	-0.7	5.0	-1.8	1.0	-4.1

La media y la desviación típica muestrales de las difencias son $\overline{d} = 0{,}33$, $\tilde{s}_d = 4{,}715$.

Queremos contrastar la igualdad de medias con el test que corresponda. Y si son diferentes, decidir cuál tiene mayor tiempo de ejecución.

O sea, queremos realizar el contraste siguiente:

$$\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 \neq \mu_2, \end{cases}$$

donde μ_1 y μ_2 son los tiempos de ejecución de los algoritmos 1 y 2, respectivamente.

Escribimos el contraste anterior en función de μ_d, la media de las diferencias de los tiempos de ejecución entre los dos algoritmos:

$$\begin{cases} H_0 : \mu_d = 0, \\ H_1 : \mu_d \neq 0. \end{cases}$$

El **estadístico de contraste** para el contraste anterior es $T = \frac{\overline{d}}{\widetilde{S}_d/\sqrt{n}}$, que tiene distribución $t_{n-1} = t_9$.

Dicho estadístico toma el siguiente valor usando los valores muestrales: $t_0 = \frac{0{,}33}{4{,}715/\sqrt{10}} = 0{,}221$.

El ***p*-valor** del contraste anterior será: $p = 2 \cdot p(t_9 > |0{,}221|) = 0{,}83$.

Es un valor grande. Por tanto, no tenemos evidencias suficientes para rechazar la hipótesis nula y concluimos que los tiempos de ejecución de los dos algoritmos es el mismo.

4.11.2. Contrastes para medias emparejadas en R. El test t

Recordemos la sintaxis básica del test t en `R`

```
t.test(x, y, mu=..., alternative=..., conf.level=..., paired=...,
       var.equal=..., na.omit=...)
```

donde el único parámetro para indicarle si las muestras son emparejadas o independientes es el parámetro `paired`: con `paired=TRUE` indicamos que las muestras son emparejadas, y con `paired=FALSE` (que es su valor por defecto) que son independientes.

Ejercicio

Nos planteamos si la longitud del sépalo supera la longitud del pétalo para las flores de la especie virginica en la tabla de datos `iris`.

En este caso se trataría de un contraste de medias dependientes:

$$\left.\begin{array}{ll} H_0 : & \mu_{spalo,virginica} = \mu_{ptalo,virginica}, \\ H_1 : & \mu_{spalo,virginica} > \mu_{ptalo,virginica}, \end{array}\right\}$$

donde $\mu_{spalo,virginica}$ y $\mu_{ptalo,virginica}$ son las longitudes del sépalo y del pétalo de las flores de la especie virginica.

Para realizar dicho contraste, vamos a considerar una muestra de 40 flores de la especie virgínica y sobre **las mismas flores** calcular las longitudes del sépalo y del pétalo.

En primer lugar seleccionamos las flores de la muestra:

```
set.seed(100)
flores.elegidas.virginica=sample(101:150,40,replace=TRUE)
```

La muestra elegida será:

```
muestra.virginica = iris[flores.elegidas.virginica,]
```

El contraste a realizar es el siguiente:

```
t.test(muestra.virginica$Sepal.Length,muestra.virginica$Petal.Length,
       paired=TRUE,alternative="greater")
```

```
##
##  Paired t-test
##
## data:  muestra.virginica$Sepal.Length and muestra.virginica$Petal.Length
## t = 22, df = 39, p-value <0.0000000000000002
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.9051    Inf
## sample estimates:
## mean of the differences
##                    0.98
```

Vemos que el p-valor del contraste es prácticamente nulo, lo que nos hace concluir que tenemos evidencias suficientes para afirmar que la longitud del sépalo es superior a la longitud del pétalo para las flores de la especie virginica.

Fijémonos que la media de la diferencia entre las medias de las longitudes del sépalo y del pétalo vale 0.98, valor suficientemente alejado del cero para poder afirmar que la media de la longitud del sépalo es superior a la media de la longitud del pétalo.

El intervalo de confianza al 95 % de confianza para la diferencia de medias asociado al contraste anterior vale:

```
t.test(muestra.virginica$Sepal.Length,muestra.virginica$Petal.Length,
       paired=TRUE,alternative="greater")$conf.int
```

```
## [1] 0.9051    Inf
## attr(,"conf.level")
## [1] 0.95
```

intervalo que no contiene el cero y que está a la derecha del mismo, lo que nos hace reafirmar que tenemos evidencias suficientes para rechazar la hipótesis nula H_0.

4.11.3. Contrastes de proporciones de muestras emparejadas

Supongamos que evaluamos dos características dicotómicas sobre una misma muestra de n sujetos. Resumimos los resultados obtenidos en la tabla siguiente:

	Característica 1	
Característica 2	Sí	No
Sí	a	b
No	c	d

Se cumple $a+b+c+d=n$. Esta tabla quiere decir, naturalmente, que a sujetos de la muestra tuvieron la característica 1 y la característica 2, que b sujetos de la muestra tuvieron la característica 2 y pero no tuvieron la característica 2, etc.

Vamos a llamar p_1 a la proporción poblacional de individuos con la característica 1, y p_2 a la proporción poblacional de individuos con la característica 2.

Queremos contrastar la hipótesis nula $H_0 : p_1 = p_2$ contra alguna hipótesis alternativa. En este caso, no pueden usarse las funciones `prop.test` o `fisher.test`.

La solución es realizar el contraste bilateral: (o los unilaterales asociados)

$$\begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 \neq p_2. \end{cases}$$

Dicho contraste tiene sentido cuando n es grande y el número $b+c$ de **casos discordantes** (en los que una característica da Sí y la otra da No) es razonablemente grande, pongamos ≥ 20.

El **estadístico de contraste** para el contraste anterior es $Z = \frac{\frac{b}{n}-\frac{c}{n}}{\sqrt{\frac{b+c}{n^2}}}$, cuya distribución aproximada es una $N(0,1)$. Sea z_0 el valor que toma sobre los valores muestrales.

Por tanto el p**-valor** será: $p = 2 \cdot p(Z > |z_0|)$.

Ejercicio

Hallar los p**-valores** para los contrastes unilaterales.

Ejemplo de proporciones emparejadas

Se toma una muestra de 1000 personas afectadas por migraña. Se les facilita un fármaco porque aligere los síntomas.

Después de la administración se les pregunta si han notado alivio en el dolor.

Al cabo de un tiempo se suministra a los mismos individuos un placebo y se les vuelve a preguntar si han notado o no mejora.

Nos preguntamos si es más efectivo el fármaco que el placebo en base a los resultados del estudio:

Fármaco/Placebo	Si	No
Si	300	62
No	38	600

El contraste que nos piden realizar es el siguiente:

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 > p_2 \end{cases}$$

donde p_1 y p_2 representan las proporciones de gente que encuentra mejora con el fármaco y el placebo, respectivamente.

El estadístico de contraste para el contraste anterior es: $Z = \frac{\frac{b}{n} - \frac{c}{n}}{\sqrt{\frac{b+c}{n^2}}}$, cuya distribución aproximada es una $N(0,1)$, donde $a = 300$, $b = 62$, $c = 38$ y $d = 600$ en nuestro caso.

El valor que toma dicho estadístico es: $z_0 = \frac{\frac{62}{1000} - \frac{38}{1000}}{\sqrt{\frac{62+38}{1000^2}}} = 2{,}4$.

Este contraste solo es válido cuando la muestra es grande y el número de *casos discordantes* $b + d$ (100 en nuestro caso) es "bastante grande", ≥ 20.

El p-**valor** para el contraste considerado es $P(Z > 2{,}4) = 0{,}008$, pequeño.

Por lo tanto, concluimos que tenemos evidencias suficientes para rechazar la hipótesis nula y poder afirmar que el fármaco es más efectivo que el placebo.

4.11.4. Contrastes para proporciones de muestras emparejadas en R

En R podemos usar el **test de McNemar**, que se lleva a cabo con la instrucción `mcnemar.test`. Su sintaxis básica es

```
mcnemar.test(X)
```

donde `X` es la matriz $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ que corresponde a la tabla anterior.

Ejercicio

Usando la tabla de datos **birthw** del paquete **MASS**, vamos a ver si la proporción de madres fumadoras es la misma que la proporción de madres hipertensas.

Para ello, vamos a considerar una muestra de 30 madres y vamos a realizar el contraste correspondiente.

En primer lugar elegimos las madres y consideramos la muestra correspondiente:

```
set.seed(333)
madres.elegidas.prop.empar = sample(1:189,30,replace=TRUE)
muestra.madres.prop.empar = birthwt[madres.elegidas.prop.empar,]
```

Seguidamente, calculamos la matriz para usar en el contraste:

```
(matriz.prop.empar = table(muestra.madres.prop.empar$smoke,muestra.madres.prop.empar$ht))
```

```
##
##      0  1
##   0 16  3
##   1 10  1
```

Fijémonos que dicha matriz no es correcta ya que $a = 1$, $b = 10$, $c = 3$ y $d = 16$. Arreglamos la matriz:

```
matriz.prop.empar = rbind(matriz.prop.empar[2,],matriz.prop.empar[1,])
matriz.prop.empar = cbind(matriz.prop.empar[,2],matriz.prop.empar[,1])
```

Comprobamos que es correcta:

```
matriz.prop.empar
```

```
##      [,1] [,2]
## [1,]    1   10
## [2,]    3   16
```

Por último, realizamos el contraste planteado:

```
mcnemar.test(matriz.prop.empar)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  matriz.prop.empar
## McNemar's chi-squared = 2.8, df = 1, p-value = 0.1
```

Hemos obtenido un p-valor de 0.0961, valor que está entre 0.05 y 0.1, la llamada zona de penumbra donde no se puede tomar una decisión clara.

Podemos decir, si consideramos que el p-valor es suficientemente grande, que no tenemos evidencias suficientes para aceptar que la proporción de madres fumadoras y con hipertensión sea diferente.

En otras palabras, no rechazamos la hipótesis nula H_0.

Ahora bien, hay que tener en cuenta que el p-valor no es demasiado grande para tal conclusión.

Otra posibilidad para realizar un contraste de dos proporciones usando muestras emparejadas, que no requiere de ninguna hipótesis sobre los tamaños de las muestras, es usar de manera adecuada la función `binom.test`.

Para explicar este método, consideremos la tabla siguiente, donde ahora damos las probabilidades poblacionales de las cuatro combinaciones de resultados:

	Característica 1	
Característica 2	Sí	No
Sí	p_{11}	p_{01}
No	p_{10}	p_{00}

De esta manera $p_1 = p_{11} + p_{10}$ y $p_2 = p_{11} + p_{01}$.

Entonces, $p_1 = p_2$ es equivalente a $p_{10} = p_{01}$ y cualquier hipótesis alternativa se traduce en la misma desigualdad, pero para p_{10} y p_{01}:

- $p_1 \neq p_2$ es equivalente a $p_{10} \neq p_{01}$;
- $p_1 < p_2$ es equivalente a $p_{10} < p_{01}$; y
- $p_1 > p_2$ es equivalente a $p_{10} > p_{01}$.

Por lo tanto podemos traducir el contraste sobre p_1 y p_2 al mismo contraste sobre p_{10} y p_{01}.

La gracia ahora está en que si la hipótesis nula $p_{10} = p_{01}$ es cierta, entonces, en el total de casos discordantes, el número de sujetos en los que la característica 1 da Sí y la característica 2 da No sigue una ley binomial con $p = 0{,}5$.

Por lo tanto, podemos efectuar el contraste usando un test binomial exacto tomando

- como muestra los casos discordantes de nuestra muestra, de tamaño $b + c$,
- como éxitos los sujetos que han dado Sí en la característica 1 y No en la característica 2, de tamaño c,
- con proporción a contrastar $p = 0{,}5$ y con hipótesis alternativa la que corresponda.

La ventaja de este test es que su validez no requiere de ninguna hipótesis sobre los tamaños de las muestras. El inconveniente es que el intervalo de confianza que nos dará será para $p_{10}/(p_{10} + p_{01})$, y no permite obtener un intervalo de confianza para la diferencia o el cociente de las probabilidades p_1 y p_2 de interés.

Ejercicio

Volvamos a realizar el contraste anterior usando este método.

Recordemos que la matriz de proporciones era:

```
matriz.prop.empar
```

```
##      [,1] [,2]
## [1,]    1   10
## [2,]    3   16
```

Por tanto, el tamaño de nuestra muestra será:

```
(n=matriz.prop.empar[1,2]+matriz.prop.empar[2,1])
```

```
## [1] 13
```

El número de éxitos será:

```
(éxitos=matriz.prop.empar[2,1])
```

```
## [1] 3
```

El contraste a realizar será:

```
binom.test(éxitos,n,p=0.5)
```

```
##
##  Exact binomial test
##
## data:  éxitos and n
## number of successes = 3, number of trials = 13, p-value = 0.09
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.05038 0.53813
## sample estimates:
```

```
## probability of success
##                  0.2308
```

Vemos que el p-valor es parecido usando el método anterior y por tanto, las conclusiones son las mismas.

4.12. Guía rápida

Excepto en las que decimos lo contrario, todas las funciones para realizar contrastes que damos a continuación admiten los parámetros `alternative`, que sirve para especificar el tipo de contraste (unilateral en un sentido u otro o bilateral), y `conf.level`, que sirve para indicar el nivel de confianza $1-\alpha$. Sus valores por defecto son contraste bilateral y nivel de confianza 0.95.

- `t.test` realiza tests t para contrastar una o dos medias (tanto usando muestras independientes como emparejadas). Aparte de `alternative` y `conf.level`, sus parámetros principales son:
  - `mu` para especificar el valor de la media que queremos contrastar en un test de una media.
  - `paired` para indicar si en un contraste de dos medias usamos muestras independientes o emparejadas.
  - `var.equal` para indicar en un contraste de dos medias usando muestras independientes si las varianzas poblacionales son iguales o diferentes.
- `sigma.test`, para realizar tests χ^2 para contrastar una varianza (o una desviación típica). Dispone de los parámetros `sigma` y `sigmasq` para indicar, respectivamente, la desviación típica o la varianza a contrastar.
- `var.test`, para realizar tests F para contrastar dos varianzas (o dos desviaciones típicas).
- `fligner.test`, para realizar tests no paramétricos de Fligner-Killeen para contrastar dos varianzas (o dos desviaciones típicas). No dispone de los parámetros `alternative` (solo sirve para contastes bilaterales) ni `conf.level` (no calcula intervalos de confianza).
- `binom.test`, para realizar tests binomiales exactos para contrastar una proporción. Dispone del parámetro `p` para indicar la proporción a contrastar.
- `prop.test`, para realizar tests aproximados para contrastar una proporción o dos proporciones de poblaciones usando muestras independientes. También dispone del parámetro `p` para indicar la proporción a contrastar en un contraste de una proporción.
- `fisher.test`, para realizar tests exactos de Fisher para contrastar dos proporciones usando muestras independientes.
- `mcnemar.test`, para realizar tests bilaterales de McNemar para contrastar dos proporciones usando muestras emparejadas. No dispone de los parámetros `alternative` ni `conf.level`.

Capítulo 5

Bondad de Ajuste

En el capítulo anterior hemos visto toda una batería de contrastes de hipótesis basados en parámetros de poblaciones como por ejemplo μ, media de la población, p, proporción de éxitos de la población, σ, desviación típica de la población, etc.

En este tipo de contrastes, suponemos que conocemos el tipo de distribución de la población. O sea, sabemos que la variable X, que nos da los valores de la población, es normal, binomial, o de otro tipo. Lo que no conocemos, y ésta es la razón de por qué realizamos este tipo de contrastes, es uno o más parámetros de los que depende la distribución de la variable X.

Por ejemplo, si suponemos que X es normal y hacemos contrastes de hipótesis sobre su media, tenemos, por un lado, el caso en que conocemos la desviación típica σ y el caso en que no la conocemos.

A todo este tipo de contrastes se les conoce como **contrastes paramétricos**.

Ahora bien, suponer de qué tipo es la distribución de la variable X que nos da los valores de la población es de hecho "un brindis al sol". ¿En qué nos basamos en decir que la distribución de X es normal por ejemplo? ¿Qué evidencias basadas en información sobre los valores de una muestra de dicha población tenemos de la normalidad de X?

Este tipo de preguntas son las que intentan responder los **contrastes no paramétricos**. Son contrastes en los que la hipótesis nula no consiste en averiguar si un determinado parámetro vale un cierto valor sino que la distribución de la variable X es de un tipo u otro. Una vez que tengamos evidencias suficientes de la normalidad de X, podemos pasar a una "segunda fase" e intentar saber alguna información sobre los parámetros de los que depende la distribución de X usando los **contrastes paramétricos**.

5.1. Contrastes de bondad de ajuste

Uno de las técnicas más conocidas para estudiar los **contrastes no paramétricos** son los **tests de bondad de ajuste** o **tests** χ^2.

El contraste que intentamos estudiar es del tipo siguiente:

$$\left.\begin{array}{l} H_0: \text{ La distribución de } X \text{ es del tipo } X_0, \\ H_1: \text{ La distribución de } X \text{ no es del tipo } X_0, \end{array}\right\}$$

donde X_0 es un tipo de distribución conocida.

Observación: en la distribución de X_0 no hace falta indicar los parámetros de los que ésta depende. Por ejemplo, X_0 puede ser normal, binomial, Poisson, etc. pero no indicamos los parámetros de los que depende. Veremos más adelante cómo estimar o aproximar dichos parámetros a partir de los valores de la muestra.

5.1.1. Pruebas gráficas: histogramas

Supongamos que nuestra variable X_0 es continua.

Para comprobar si una determinada muestra proviene de la variable X_0, lo primero que podemos hacer es realizar distintas pruebas gráficas como por ejemplo histogramas.

A partir de dichos histogramas, podemos "estimar" la función de densidad de la muestra y ver si dicha función de densidad se parece o no a la función de densidad de la variable X_0.

La estimación de la función de densidad a partir del histograma de la muestra se sale de los objetivos del curso pero vamos a ver con un ejemplo cómo podemos usar dicha función. Si queréis detalles, consultad https://en.wikipedia.org/wiki/Kernel_density_estimation.

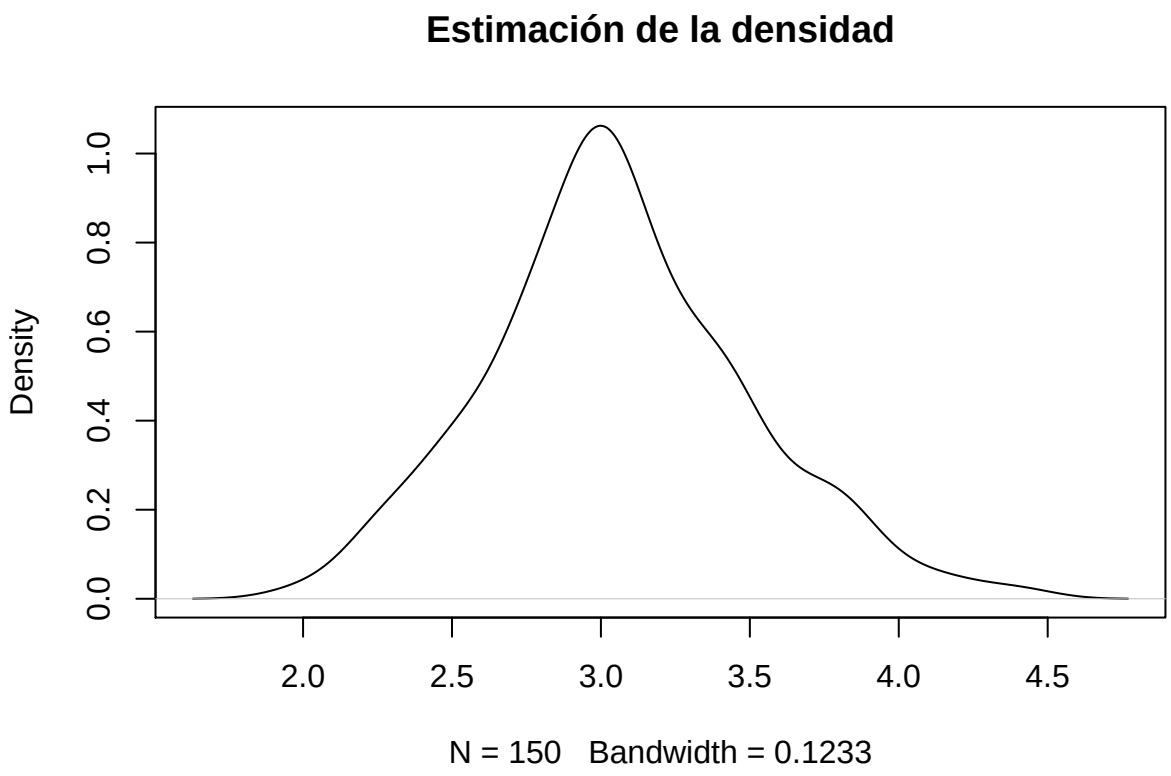
Ejemplo

Consideremos la tabla de datos `iris`, concretamente la variable anchura del sépalo (`Sepal.Width`).

Queremos ver a qué se puede aproximar dicha variable.

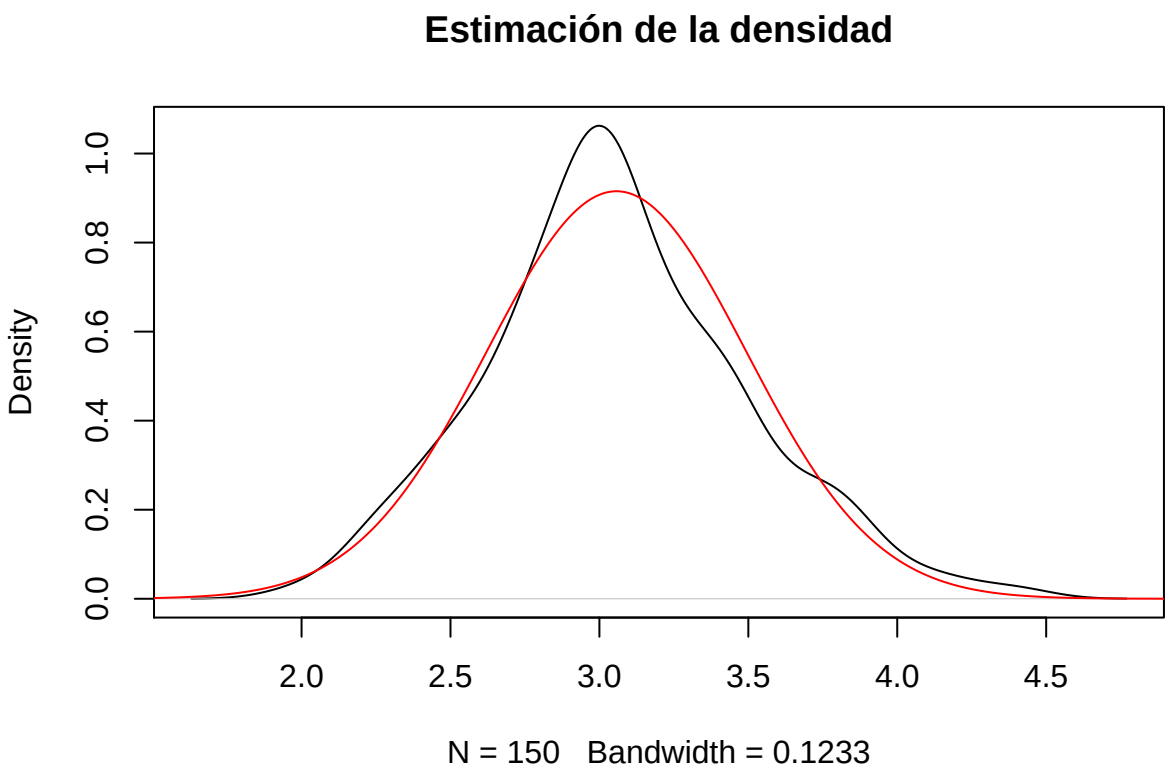
Vamos a realizar un gráfico de la estimación de la función de densidad de la muestra usando la función `density` de `R`:

```
muestra=iris$Sepal.Width
plot(density(muestra),main="Estimación de la densidad")
```



A "ojo de buen cubero", parece una campana de Gauss. Para comprobarlo, dibujemos además la función de densidad de la distribución normal con parámetros μ igualado a la media de la muestra y σ, a la desviación típica:

```
muestra=iris$Sepal.Width
plot(density(muestra),main="Estimación de la densidad")
x=seq(from=1,to=5,by=0.01)
mu=mean(iris$Sepal.Width)
sigma=sd(iris$Sepal.Width)
lines(x,dnorm(x,mean=mu,sd=sigma),col="red")
```



Vemos que la campana de Gauss se parece bastante a la estimación de la densidad.

La cuestión ahora es: ¿podemos aceptar que el parecido es suficiente para aceptar que la distribución de la anchura del sépalo es normal?

La respuesta será contestada en secciones posteriores.

5.1.2. Pruebas gráficas: Q-Q-plots

Otro tipo de prueba gráfica que podemos realizar son los llamados **gráficos cuantil-cuantil o Q-Q-plots**.

Este tipo de gráficos compara los **cuantiles observados de la muestra** con los **cuantiles teóricos de la distribución teórica.**

La función de `R` que realiza un gráfico de este tipo es la función `qqPlot` del paquete `car`.

Su sintaxis básica es

```
qqPlot(x, distribution=...,  parámetros, id=FALSE, ...)
```

donde:

- `x` es el vector con la muestra.
- El parámetro `distribution` se ha de igualar al nombre de la familia de distribuciones entre comillas, y puede tomar como valor cualquier familia de distribuciones de la que R sepa calcular

la densidad y los cuantiles: esto incluye las distribuciones que hemos estudiado hasta el momento: `"norm"`, `"binom"`, `"poisson"`, `"t"`, etc.

- A continuación, se tienen que entrar los parámetros de la distribución, igualando su nombre habitual (`mean` para la media, `sd` para la desviación típica, `df` para los grados de libertad, etc.) a su valor.

- Por defecto, el gráfico obtenido con la función `qqPlot` identifica los dos Q-Q-puntos con ordenadas más extremas. Para omitirlos, usad el parámetro `id=FALSE`.

Otros parámetros a tener en cuenta:

- `qqPlot` añade por defecto una rejilla al gráfico, que podéis eliminar con `grid=FALSE`.

- `qqPlot` añade por defecto una línea recta que une los Q-Q-puntos correspondientes al primer y tercer cuartil: se la llama **recta cuartil-cuartil**. Un buen ajuste de los Q-Q-puntos a esta recta significa que la muestra se ajusta a la distribución teórica, pero posiblemente con parámetros diferentes a los especificados. Os recomendamos mantenerla, pero si queréis eliminarla por ejemplo para substituirla por la diagonal $y = x$, podéis usar el parámetro `line="none"`.

- `qqPlot` también añade dos curvas discontinuas que abrazan una "región de confianza del 95 %" para el Q-Q-plot. Sin entrar en detalles, esta región contendría todos los Q-Q-puntos en un 95 % de las ocasiones que tomáramos una muestra de la distribución teórica del mismo tamaño que la nuestra. Por lo tanto, si todos los Q-Q-puntos caen dentro de esta franja, no hay evidencia para rechazar que la muestra provenga de la distribución teórica. Esta franja de confianza es muy útil para interpretar el Q-Q-plot, pero la podéis eliminar con `envelope=FALSE`.

- Se pueden usar los parámetros usuales de `plot` para poner nombres a los ejes, título, modificar el estilo de los puntos, etc., y otros parámetros específicos para modificar el aspecto del gráfico. Por ejemplo, `col.lines` sirve para especificar el color de las líneas que añade.

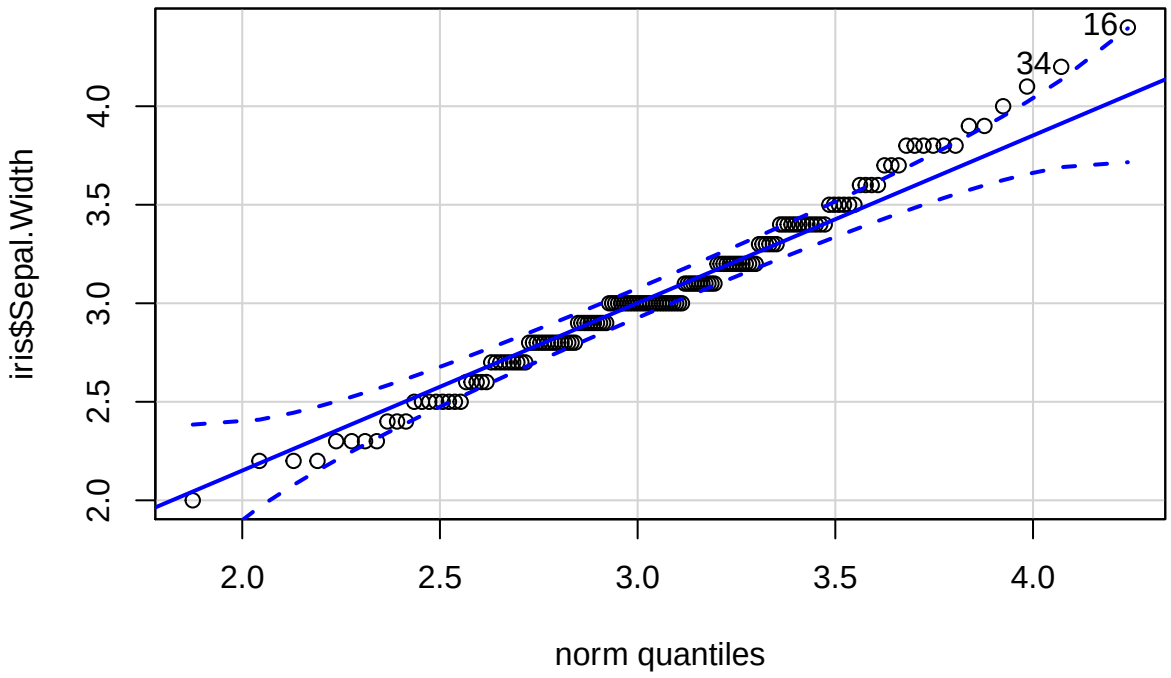
Hagamos un gráfico Q-Q-plot del ejemplo anterior.

Ejemplo

Recordemos que considerábamos la variable anchura del sépalo de la tabla de datos `iris`.

El Q-Q-plot para ver si la variable anterior sigue la ley normal es el siguiente:

```
library(car)
qqPlot(iris$Sepal.Width,distribution = "norm", mean=mu,sd=sigma)
```




```
## [1] 16 34
```

Observamos que la mayoría de los valores de la muestra caen dentro de la franja del 95 % de confianza.

Sin embargo, fijáos que hay unos pocos puntos que se salen de dicha franja.

En resumen, según esta prueba gráfica, "parece" que la distribución es normal pero tenemos dudas.

5.1.3. Contraste χ^2 de Pearson

Veamos a continuación cómo se realiza un **contraste χ^2 de Pearson**.

Suponemos que disponemos de los valores de una muestra de tamaño n de la variable X que nos da los valores de la población: $x_1, x_2, \dots, x_n$.

A continuación, clasificamos los valores x_i, $i = 1, \dots, n$ en k clases. La elección de estas clases depende del problema estudiado y del contexto del mismo.

Sean $n_1, \dots, n_k$, el número de valores de la muestra que están en cada una de las clases: n_1 sería el número de valores de la muestra que están en la clase 1, n_2, el número de valores de la muestra que están en la clase 2 y así sucesivamente hasta n_k.

Obtendríamos lo que se conoce como **tabla de frecuencias empíricas**:

Clases	Clase 1	Clase 2	...	Clase k	Total
Frecuencias empíricas	n_1	n_2	...	n_k	n

El siguiente paso es obtener la tabla de la función de probabilidad de la variable discreta X_k de valores

$\{1, \dots, k\}$ y con función de probabilidad $p_i = P(X_k = i) = P(X_0 \in \text{Clase i})$, $i = 1, \dots, k$.

Esta función de probabilidad tiene que hallarse a partir del conocimiento de X_0. Si desconocemos alguno(s) del (de los) parámetro(s) de (los) que depende X_0, los tendremos que estimar usando las técnicas vistas en el capítulo de estimación de parámetros.

La tabla de la función de probabilidad X_k quedaría de la forma siguiente:

X_k	1	2	...	k	Total
$P(X_k = i)$	$P(X_k = 1)$	$P(X_k = 2)$	...	$P(X_k = k)$	1

A partir de dicha tabla, calculamos la **tabla de frecuencias teóricas**:

Clases	Clase 1	Clase 2	...	Clase k	Total
Frecuencias teóricas	$n \cdot P(X_k = 1)$	$n \cdot P(X_k = 2)$	...	$n \cdot P(X_k = k)$	n
Frecuencias teóricas	e_1	e_2	...	e_k	n

Llamaremos $e_i = n \cdot P(X_k = i)$ a la **frecuencia teórica** de la clase i-ésima. (e de "esperada")

El **test** χ^2 o **test de bondad de ajuste** se basa en que, si la hipótesis nula es cierta, las **frecuencias empíricas** y las **frecuencias teóricas** son "parecidas".

Más concretamente, si la hipótesis nula es cierta, el estadístico siguiente:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{frec. empíricas}_i - \text{frec. teóricas}_i)^2}{\text{frec. teóricas}_i} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i},$$

sigue aproximadamente al distribución χ^2_{k-1} grados de libertad.

Sea χ_0 el valor del estadístico de contraste anterior para nuestra muestra. El p-valor del contraste vale:

$$p = P(\chi^2_{k-1} > \chi_0),$$

con el significado usual:

- si $p < 0{,}05$, concluimos que tenemos evidencias suficientes para rechazar la independencia de los criterios,
- si $p > 0{,}1$, concluimos que no tenemos evidencias suficientes para rechazar la independencia de los criterios y,
- si $0{,}05 \leq p \leq 0{,}1$, estamos en la zona de penumbra. Necesitamos más datos para tomar una decisión clara.

Ejemplo del lanzamiento de un dado

Imaginemos que queremos ver si un dado está trucado o no.

Si no está trucado, cuando tiramos el dado y miramos el resultado X, cada resultado $i = 1, \dots, 6$ tiene probabilidad $P(X = i) = \frac{1}{6}$. Ésta sería, por tanto, la función de distribución de la variable X_k.

Las clases serían posibles valores que puede tener el dado al lanzarse. La distribución de la variable X_k sería en este caso:

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

Nos dicen que han lanzado el dado 120 veces y se han obtenido los resultados siguientes:

Clases	Clase 1	Clase 2	Clase 3	Clase 4	Clase 5	Clase 6	Total
Frecuencias empíricas	20	22	17	18	19	24	120

¿Hay bastante evidencia que el dado esté trucado?

Resolución

La tabla de **frecuencias teóricas** sería:

Clases	Clase 1	Clase 2	Clase 3	Clase 4	Clase 5	Clase 6	Total
Frecuencias teóricas	$\frac{120}{6} = 20$	$\frac{120}{6} = 20$	$\frac{120}{6} = 20$	$\frac{120}{6} = 20$	$\frac{120}{6} = 20$	$\frac{120}{6} = 20$	120

El valor del estadístico χ^2 sería:

$$\chi_0 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1{,}7.$$

El p-valor del contraste sería $P(\chi^2_5 > 1{,}7)$:

```
pchisq((22-20)^2/20+(17-20)^2/20+(18-20)^2/20+(19-20)^2/20+(24-20)^2/20,5,
       lower.tail=FALSE)
```

```
## [1] 0.8889
```

Como el p-valor es grande, concluímos que no tenemos evidencias suficientes para rechazar que el dado esté trucado.

5.1.4. Condiciones para poder aplicar el test χ^2 de Pearson

El test de **bondad de ajuste** está basado en el **estadístico** χ^2 que recordemos sigue aproximadamente una distribución χ^2_{k-1} grados de libertad.

Al estar basado en un **Teorema Límite**, para que dicha aproximación sea efectiva, las condiciones siguientes se tienen que verificar:

- el tamaño de la muestra tiene que ser grande: $n \geq 25$ o mejor $n \geq 30$,
- las clases tienen que cubrir todos los resultados posibles, (en la práctica: $n = \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} e_i$)

- las **frecuencias teóricas** tienen que ser mayores o iguales que 5: $e_i \geq 5$, para todo $i = 1, \dots, k$.

Ejemplo del lanzamiento de un dado

En el ejemplo del lanzamiento del dado, observamos que se verifican las condiciones anteriores:

- $n = 120 \geq 30$,
- $$\begin{array}{rl} \sum_{i=1}^{6} n_i = & 20 + 22 + 17 + 18 + 19 + 24 = 120 \\ = & \sum_{i=1}^{6} e_i = 20 + 20 + 20 + 20 + 20 + 20 \end{array},$$
- $e_1 = e_2 = e_3 = e_4 = e_5 = e_6 = 20 \geq 5$.

5.1.5. Resolución de un test χ^2 de Pearson en R

Para resolver un **contraste de bondad de ajuste** en R, hemos de usar la función `chisq.test`.

Su sintaxis básica es

```
chisq.test(x, p=..., rescale.p=..., simulate.p.value=...)
```

donde:

- `x` es el vector (o la tabla, calculada con `table`) de frecuencias **absolutas observadas** de las clases en la muestra, recordemos que las hemos llamado n_i, $i = 1, \dots, k$.
- `p` es el vector de probabilidades teóricas de las clases para la distribución que queremos contrastar. O sea, es el vector de la función de probabilidad de la variable X_k: $p_i = P(X_k = i)$, $i = 1, \dots k$.

Si no lo especificamos, se entiende que la probabilidad es la misma para todas las clases. Obviamente, estas probabilidades se tienen que especificar en el mismo orden que las frecuencias de `x` y, como son las probabilidades de todos los resultados posibles, en principio tienen que sumar 1; esta condición se puede relajar con el siguiente parámetro.

- `rescale.p` es un parámetro lógico que, si se iguala a `TRUE`, indica que los valores de `p` no son probabilidades, sino solo proporcionales a las probabilidades; esto hace que `R` tome como probabilidades teóricas los valores de `p` partidos por su suma, para que sumen 1.

Por defecto vale `FALSE`, es decir, se supone que el vector que se entra como `p` son probabilidades y por lo tanto debe sumar 1, y si esto no pasa se genera un mensaje de error indicándolo. Igualarlo a `TRUE` puede ser útil, porque nos permite especificar las probabilidades mediante las frecuencias esperadas o mediante porcentajes. Pero también es peligroso, porque si nos hemos equivocado y hemos entrado un vector en `p` que no corresponda a una probabilidad, `R` no nos avisará.

- `simulate.p.value` es un parámetro lógico que indica a la función si debe optar por una simulación para el cálculo del p-valor del contraste.

Por defecto vale `FALSE`, en cuyo caso este p-valor no se simula sino que se calcula mediante la distribución χ^2 correspondiente.

Si falla una o más condiciones para que se aplique el **test de bondad de ajuste**, tendremos que especificarlo como `TRUE` y `R` realizará una serie de replicaciones aleatorias de la situación teórica: por defecto, 2000, pero su número se puede especificar mediante el parámetro `B`. Es decir, generará un conjunto de vectores aleatorios de frecuencias con la distribución que queremos contrastar, cada uno

de suma total la de `x`. A continuación, calculará la proporción de estas repeticiones en las que el estadístico de contraste es mayor o igual que el obtenido para `x`, y éste será el p-valor que dará.

Ejemplo de la tabla de datos `iris`

Consideremos la tabla de datos `iris`. Imaginemos que queremos ver si en una muestra de tamaño 10, hay la misma cantidad de flores de las tres especies: setosa, versicolor y virgínica.

En primer lugar, elegimos una muestra de 10 flores:

```
set.seed(2020)  ## fijamos la semilla de aleatorización
muestra.flores = sample(iris$Species,10)
```

A continuación, realizamos el contraste de bondad de ajuste:

```
chisq.test(table(muestra.flores))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(muestra.flores)
## X-squared = 0.8, df = 2, p-value = 0.7
```

Fijaos que `R` nos avisa que las aproximaciones pueden ser incorrectas. La razón es que las **frecuencias observadas** no son mayores que 5 ya que éstas valen: $e_{setosa} = e_{versicolor} = e_{virginica} = \frac{10}{3} \approx 3{,}333$.

Para solventar este problema, vamos a simular el p-valor:

```
chisq.test(table(muestra.flores), simulate.p.value = TRUE, B=2000)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 2000 replicates)
##
## data:  table(muestra.flores)
## X-squared = 0.8, df = NA, p-value = 0.8
```

Vemos que con 2000 replicaciones, al obtener un p-valor grande, podemos concluir que no tenemos evidencias suficientes para rechazar que la proporción de especies en la muestra no sea la misma.

Ejemplo para comprobar si cierta distribución es normal usando el test χ^2 de Pearson en `R`

Un técnico de medio ambiente quiere estudiar el aumento de temperatura del agua a dos kilómetros de los vertidos de agua autorizados de una planta industrial.

El responsable de la empresa afirma que *estos aumentos de temperatura siguen una ley normal con* $\mu = 3{,}5$ *décimas de grado C y* $\sigma = 0{,}7$ *décimas de grado C.*

El técnico lo posa en entredicho. Para decidirlo, toma una muestra aleatoria de 40 observaciones del aumento de las temperaturas (en décimas de grado) y se obtienen los resultados siguientes:

Rango de temperaturas	Frecuencias
1.45-1.95	2
1.95-2.45	1

Rango de temperaturas	Frecuencias
2.45-2.95	4
2.95-3.45	15
3.45-3.95	10
3.95-4.45	5
4.45-4.95	3

¿Hay evidencia que la sospecha del técnico sea verdadera?

El contraste a realizar es el siguiente:

$$\begin{cases} H_0 : \text{La distribución de los aumentos de temperatura es } N(3{,}5, 0{,}7), \\ H_1 : \text{La distribución de los aumentos de temperatura no es } N(3{,}5, 0{,}7). \end{cases}$$

Para poder aplicar el **test χ^2 de Pearson**, como la distribución teórica es normal y su dominio es todo $\mathbb{R}$, tendremos que ampliar los intervalos de la tabla anterior para asegurarnos que los valores de la tabla pueden alcanzar todos los valores de la distribución teórica:

Rango de temperaturas	Frecuencias
$(-\infty, 1{,}95]$	2
$(1{,}95, 2{,}45]$	1
$(2{,}45, 2{,}95]$	4
$(2{,}95, 3{,}45]$	15
$(3{,}45, 3{,}95]$	10
$(3{,}95, 4{,}45]$	5
$(4{,}45, \infty)$	3

La tabla anterior nos determina las clases a considerar que corresponderían a los intervalos. Las frecuencias empíricas serían las que nos da la segunda columna de la tabla.

A continuación, vamos a calcular las frecuencias teóricas. Para ello, en primer lugar, hay que hallar la función de probabilidad de la variable X_k:

- Cálculo de $p_1 = P(X_0 \in \text{Clase 1})$:

$$p_1 = P(X_0 \leq 1{,}95) = P\left(Z \leq \frac{1{,}95 - 3{,}5}{0{,}7}\right) = P(Z \leq -2{,}214) = 0{,}013,$$

  donde $Z = N(0, 1)$. Por tanto, $e_1 = n \cdot 0{,}013 = 40 \cdot 0{,}013 = 0{,}54$.

- Cálculo de $p_2 = P(X_0 \in \text{Clase 2})$:

$$\begin{aligned} p_2 &= P(1{,}95 < X_0 \leq 2{,}45) = P\left(\tfrac{1{,}95-3{,}5}{0{,}7} < Z \leq \tfrac{2{,}45-3{,}5}{0{,}7}\right) = P(-2{,}214 < Z \leq -1{,}5) \\ &= P(Z \leq -1{,}5) - P(Z \leq -2{,}214) = 0{,}067 - 0{,}013 = 0{,}053. \end{aligned}$$

  Por tanto, $e_2 = n \cdot 0{,}053 = 40 \cdot 0{,}053 = 2{,}14$.

- Cálculo de $p_3 = P(X_0 \in \text{Clase 3})$:

$$\begin{aligned} p_3 &= P(2{,}45 < X_0 \leq 2{,}95) = P\left(\tfrac{2{,}45-3{,}5}{0{,}7} < Z \leq \tfrac{2{,}95-3{,}5}{0{,}7}\right) = P(-1{,}5 < Z \leq -0{,}786) \\ &= P(Z \leq -0{,}786) - P(Z \leq -1{,}5) = 0{,}216 - 0{,}067 = 0{,}149. \end{aligned}$$

Por tanto, $e_3 = n \cdot 0{,}149 = 40 \cdot 0{,}149 = 5{,}97$.

- Cálculo de $p_4 = P(X_0 \in \text{Clase } 4)$:

$$\begin{aligned} p_4 &= P(2{,}95 < X_0 \leq 3{,}45) = P\left(\tfrac{2{,}95-3{,}5}{0{,}7} < Z \leq \tfrac{3{,}45-3{,}5}{0{,}7}\right) = P(-0{,}786 < Z \leq -0{,}071) \\ &= P(Z \leq -0{,}071) - P(Z \leq -0{,}786) = 0{,}472 - 0{,}216 = 0{,}256. \end{aligned}$$

Por tanto, $e_4 = n \cdot 0{,}256 = 40 \cdot 0{,}256 = 10{,}22$.

- Cálculo de $p_5 = P(X_0 \in \text{Clase } 5)$:

$$\begin{aligned} p_5 &= P(3{,}45 < X_0 \leq 3{,}95) = P\left(\tfrac{3{,}45-3{,}5}{0{,}7} < Z \leq \tfrac{3{,}95-3{,}5}{0{,}7}\right) = P(-0{,}071 < Z \leq 0{,}643) \\ &= P(Z \leq 0{,}643) - P(Z \leq -0{,}071) = 0{,}74 - 0{,}472 = 0{,}268. \end{aligned}$$

Por tanto, $e_5 = n \cdot 0{,}268 = 40 \cdot 0{,}268 = 10{,}73$.

- Cálculo de $p_6 = P(X_0 \in \text{Clase } 6)$:

$$\begin{aligned} p_6 &= P(3{,}95 < X_0 \leq 4{,}45) = P\left(\tfrac{3{,}95-3{,}5}{0{,}7} < Z \leq \tfrac{4{,}45-3{,}5}{0{,}7}\right) = P(0{,}643 < Z \leq 1{,}357) \\ &= P(Z \leq 1{,}357) - P(Z \leq 0{,}643) = 0{,}913 - 0{,}74 = 0{,}173. \end{aligned}$$

Por tanto, $e_6 = n \cdot 0{,}173 = 40 \cdot 0{,}173 = 6{,}91$.

- Cálculo de $p_7 = P(X_0 \in \text{Clase } 7)$:

$$p_7 = P(X_0 \geq 4{,}45) = P\left(Z \geq \frac{4{,}45 - 3{,}5}{0{,}7}\right) = P(Z \geq 1{,}357) = 0{,}087,$$

donde $Z = N(0,1)$. Por tanto, $e_7 = n \cdot 0{,}087 = 40 \cdot 0{,}087 = 3{,}49$.

Observamos que las **frecuencias teóricas** e_1, e_2 y e_7 son menores que 5. Por tanto, no se cumplen las condiciones para poder aplicar el **contraste de bondad de ajuste**.

Para solventar este problema, agruparemos los intervalos $(-\infty, 1{,}95]$, $(1{,}95, 2{,}45]$, $(2{,}45, 2{,}95]$ en el intervalo $(-\infty, 2{,}95)$ y los intervalos $(3{,}95, 4{,}45]$ y $(4{,}45, \infty)$ en el intervalo $(3{,}95, \infty)$. De esta forma, la tabla de frecuencias empíricas quedará de la forma siguiente:

Rango de temperaturas	Frecuencias
$(-\infty, 2{,}95]$	$2 + 1 + 4 = 7$
$(2{,}95, 3{,}45]$	15
$(3{,}45, 3{,}95]$	10
$(3{,}95, \infty]$	$5 + 3 = 8$

Las **frecuencias teóricas** de la nueva tabla serán las siguientes:

- $e_1 = 0{,}54 + 2{,}14 + 5{,}97 = 8{,}64$.
- $e_2 = n \cdot 0{,}256 = 40 \cdot 0{,}256 = 10{,}22$.
- $e_3 = n \cdot 0{,}268 = 40 \cdot 0{,}268 = 10{,}73$.
- $e_4 = 6{,}91 + 3{,}49 = 10{,}41$.

Ahora vemos que las **frecuencias teóricas** son mayores o iguales que 5 y, por tanto, se verifican las condiciones para poder aplicar el **test de bondad de ajuste**.

La tabla siguiente resume los cálculos realizados:

Rango de temperaturas	Frecuencias empíricas	Frecuencias teóricas
$(-\infty, 2{,}95]$	7	8,64
$(2{,}95, 3{,}45]$	15	10,22
$(3{,}45, 3{,}95]$	10	10,73
$(3{,}95, \infty]$	8	10,41

El valor del **estadístico de contraste** χ^2 vale:

$$\chi_0 = \frac{(7-8{,}64)^2}{8{,}64} + \frac{(15-10{,}22)^2}{10{,}22} + \frac{(10-10{,}73)^2}{10{,}73} + \frac{(8-10{,}41)^2}{10{,}41} = 3{,}153.$$

El p-valor del contraste será:

$$p = P(\chi_3 > 3{,}153) = 0{,}369.$$

Como el p-valor es grande, concluimos que no tenemos evidencias suficientes para rechazar que el aumento de temperatura no siga una distribución normal de parámetros $\mu = 3{,}5$ décimas de grado y $\sigma = 0{,}7$ décimas de grado.

Vamos a realizar el ejemplo anterior usando R.

En primer lugar, definimos las clases definiendo los extremos de los intervalos y las **frecuencias empíricas**:

```
extremos.izquierdos = c(-Inf,1.95,2.45,2.95,3.45,3.95,4.45)
extremos.derechos = c(1.95,2.45,2.95,3.45,3.95,4.45,Inf)
frecuencias.empíricas = c(2,1,4,15,10,5,3)
n=sum(frecuencias.empíricas)
```

Para hallar las **frecuencias teóricas** usamos la función `pnorm` de R:

```
mu=3.5; sigma=0.7;
probabilidades.teóricas = pnorm(extremos.derechos,mu,sigma)-
  pnorm(extremos.izquierdos,mu,sigma)
frecuencias.teóricas = n*probabilidades.teóricas
round(frecuencias.teóricas,2)
```

```
## [1]  0.54  2.14  5.97 10.22 10.73  6.91  3.49
```

Por último, aplicamos el test de la χ^2 usando la función `chisq.test`:

```
chisq.test(frecuencias.empíricas,p=probabilidades.teóricas)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  frecuencias.empíricas
## X-squared = 8.1, df = 6, p-value = 0.2
```

R nos avisa que la aproximación χ^2 puede no ser correcta. Nosotros sabemos la razón: hay tres probabilidades teóricas que son menores que 5.

Llegados a este punto, podemos actuar de dos formas: juntamos intervalos o simulamos el p-valor.

Si optamos por la primera opción, tendremos que hacer:

```
extremos.izquierdos2=extremos.izquierdos[c(1,4,5,6)]
extremos.derechos2 = extremos.derechos[c(3,4,5,7)]
frecuencias.empíricas2 = c(sum(frecuencias.empíricas[1:3]),
                           frecuencias.empíricas[4:5],sum(frecuencias.empíricas[6:7]))
probabilidades.teóricas2 =pnorm(extremos.derechos2,mu,sigma)-
  pnorm(extremos.izquierdos2,mu,sigma)
frecuencias.teóricas2 = n*probabilidades.teóricas2
chisq.test(frecuencias.empíricas2,p=probabilidades.teóricas2)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  frecuencias.empíricas2
## X-squared = 3.2, df = 3, p-value = 0.4
```

Vemos que obtenemos los mismos valores que los cálculos realizados a mano.

Si optamos por la segunda opción, hemos de hacer:

```
chisq.test(frecuencias.empíricas,p=probabilidades.teóricas,simulate.p.value = TRUE,
           B=2000)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 2000 replicates)
##
## data:  frecuencias.empíricas
## X-squared = 8.1, df = NA, p-value = 0.2
```

Aunque obtengamos un p-valor distinto, llegamos a la misma conclusión anterior.

5.1.6. Test χ^2 de Pearson con parámetros poblacionales desconocidos

El ejemplo visto anteriormente de normalidad no es realista en el sentido que la mayoría de las veces desconoceremos la media μ y la desviación típica σ de la variable de la población X_0 a la que se refiere la hipòtesis nula H_0.

Cuando la variable X_0 de la población de contraste dependa de algún(os) parámetro(s) desconocido(s), necesitamos conocer su valor de cara a calcular las **frecuencias esperadas** o **frecuencias teóricas** e_i.

La manera de resolver este inconveniente, es estimar dichos parámetros usando el **estimador máximo verosímil** correspondiente.

Una vez estimados el(los) parámetro(s) del(de los) que depende la variable de la población X_0, podemos realizar el **test de bondad de ajuste** tal como hemos visto pero ahora los grados de libertad de la distribución χ^2 disminuyen. En concreto, valen: $k-1-$ número de parámetros estimados.

Ejemplo

Se quiere determinar si el número de veces que aparece la secuencia GATACA en una cadena de ADN de longitud 1000 sigue una ley Poisson.

Se toman varias muestras de cadenas de ADN de longitud 1000 y se cuentan los números de GATACA

número x_i de veces queaparece GATACA	0	1	2	3	4	5
frecuencia empírica n_i	229	211	93	35	7	1

Hemos realizado en total $n = 229 + 211 + 93 + 35 + 7 + 1 = 576$ observaciones.

El contraste a realizar es el siguiente:

$$\begin{cases} H_0 : \text{La muestra proviene de una distribución } Po(\lambda), \\ H_1 : \text{La muestra no proviene de esta distribución.} \end{cases}$$

Al no conocer el parámetro λ, hemos de estimarlo.

Recordemos que el **estimador máximo verosímil** del parámetro λ de una distribución de Poisson es su **media muestral**: $\hat{\lambda} = \overline{X}$:

$$\hat{\lambda} = \frac{229 \cdot 0 + 211 \cdot 1 + 93 \cdot 2 + 35 \cdot 3 + 7 \cdot 4 + 1 \cdot 5}{229 + 211 + 93 + 35 + 7 + 1} = \frac{535}{576} = 0{,}929.$$

Las clases deben cubrir todo el conjunto de valores de la variable X_0 que, en nuestro caso, al ser una variable de Poisson, serían todos los enteros positivos: $D_{X_0} = \{0, 1, ...\}$.

Usando la tabla de **frecuencias empíricas**, consideramos las clases siguientes:

$$\begin{array}{l} \text{Clase } 0 \quad : X_0 = 0, \ \text{Clase } 1 : X_0 = 1, \ \text{Clase } 2 : X_0 = 2, \ \text{Clase } 3 : X_0 = 3, \\ \text{Clase } 4 \quad : X_0 = 4, \ \text{Clase } 5 : X_0 \geq 5. \end{array}$$

Recordemos que la función de probabilidad de una variable de Poisson de parámetro λ vale: $P(X_0 = i) = \frac{\lambda^i}{i!}\mathrm{e}^{-\lambda}$, donde $i \in D_{X_0}$.

Las **frecuencias esperadas o teóricas** e_i se calculan de la forma siguiente:

- $e_0 = n \cdot P(X_0 \in \text{Clase } 0) = n \cdot P(X_0 = 0) = 576 \cdot \frac{0{,}929^0}{0!}\mathrm{e}^{-0{,}929} = 576 \cdot \mathrm{e}^{-0{,}929} = 227{,}53.$
- $e_1 = n \cdot P(X_0 \in \text{Clase } 1) = n \cdot P(X_0 = 1) = 576 \cdot \frac{0{,}929^1}{1!}\mathrm{e}^{-0{,}929} = 211{,}34.$
- $e_2 = n \cdot P(X_0 \in \text{Clase } 2) = n \cdot P(X_0 = 2) = 576 \cdot \frac{0{,}929^2}{2!}\mathrm{e}^{-0{,}929} = 98{,}15.$
- $e_3 = n \cdot P(X_0 \in \text{Clase } 3) = n \cdot P(X_0 = 3) = 576 \cdot \frac{0{,}929^3}{3!}\mathrm{e}^{-0{,}929} = 30{,}39.$
- $e_4 = n \cdot P(X_0 \in \text{Clase } 4) = n \cdot P(X_0 = 4) = 576 \cdot \frac{0{,}929^4}{4!}\mathrm{e}^{-0{,}929} = 7{,}06.$
- El cálculo de e_5 se realiza de forma ligeramente diferente de los demás:

$$\begin{array}{rl} e_5 & = n \cdot P(X_0 \in \text{Clase } 5) = n \cdot P(X_0 \geq 5) = n \cdot (1 - P(X_0 \leq 4)) \\ & = n \cdot (1 - (P(X_0 = 0) + P(X_0 = 1) + P(X_0 = 2) + P(X_0 = 3) + P(X_0 = 4))) \\ & = 576 - (227{,}53 + 211{,}34 + 98{,}15 + 30{,}39 + 7{,}06) = 1{,}54. \end{array}$$

Como la clase 5 no verifica las condiciones de aplicación del **contraste de bondad de ajuste** al tener una **frecuencia esperada** menor que 5, juntaremos las dos últimas clases y las nuevas clases serán:

$$\text{Clase } 0: X_0 = 0,\ \text{Clase } 1: X_0 = 1,\ \text{Clase } 2: X_0 = 2,\ \text{Clase } 3: X_0 = 3,\ \text{Clase } 4: X_0 \geq 4.$$

Las **frecuencias esperadas** de las nuevas clases serán:

- $e_0 = n \cdot P(X_0 \in \text{Clase } 0) = n \cdot P(X_0 = 0) = 227{,}53$.
- $e_1 = n \cdot P(X_0 \in \text{Clase } 1) = n \cdot P(X_0 = 1) = 211{,}34$.
- $e_2 = n \cdot P(X_0 \in \text{Clase } 2) = n \cdot P(X_0 = 2) = 98{,}15$.
- $e_3 = n \cdot P(X_0 \in \text{Clase } 3) = n \cdot P(X_0 = 3) = 30{,}39$.
- El cálculo de e_4 se realiza de forma parecida al cálculo de la clase 5 anterior:

$$\begin{aligned} e_4 &= n \cdot P(X_0 \in \text{Clase } 4) = n \cdot P(X_0 \geq 4) = n \cdot (1 - P(X_0 \leq 3)) \\ &= n \cdot (1 - (P(X_0 = 0) + P(X_0 = 1) + P(X_0 = 2) + P(X_0 = 3)) \\ &= 576 - (227{,}53 + 211{,}34 + 98{,}15 + 30{,}39) = 8{,}6. \end{aligned}$$

Resúmamos en una tabla las **frecuencias empíricas** y las **frecuencias esperadas**:

Clase	Frecuencias empíricas	Frecuencias teóricas
$\{X_0 = 0\}$	229	227,53
$\{X_0 = 1\}$	211	211,34
$\{X_0 = 2\}$	93	98,15
$\{X_0 = 3\}$	35	30,39
$\{X_0 \geq 4\}$	$7 + 1 = 8$	8,6

El valor del **estadístico de contraste** será:

$$\begin{aligned} \chi_0 &= \tfrac{(229-227{,}53)^2}{227{,}53} + \tfrac{(211-211{,}34)^2}{211{,}34} + \tfrac{(93-98{,}15)^2}{98{,}15} + \tfrac{(35-30{,}39)^2}{30{,}39} + \tfrac{(8-8{,}6)^2}{8{,}6} \\ &= 1{,}022. \end{aligned}$$

Antes de calcular el p-valor del contraste, calculemos los grados de libertad de la variable χ^2. Éstos serán: $g.l. = 5 - 1 - 1 = 3$ ya que hemos estimado un parámetro, λ.

El p-valor del contraste será:

$$p = P(\chi^2_3 > 1{,}022) = 0{,}796.$$

Como el p-valor es muy grande, concluimos que no tenemos evidencias suficientes para rechazar que el número de veces que aparece la secuencia GATACA en una cadena ADN de longitud 1000 sigue una ley de Poisson.

Resolvamos el ejemplo anterior con ayuda de `R`.

En primer lugar, definimos las **frecuencias empíricas** y las **probabilidades esperadas y teóricas**:

```
frecuencias.empíricas = c(229,211,93,35,8);
estimación.lambda = (211+93*2+35*3+7*4+1*5)/(229+211+93+35+7+1);
probabilidades.esperadas = c(dpois(0,estimación.lambda),dpois(1,estimación.lambda),
                             dpois(2,estimación.lambda),dpois(3,estimación.lambda),
                             1-ppois(3,estimación.lambda));
```

A continuación, realizamos el **test de bondad de ajuste** usando la función `chisq.test`:

```
chisq.test(frecuencias.empíricas,p=probabilidades.esperadas)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  frecuencias.empíricas
## X-squared = 1, df = 4, p-value = 0.9
```

Nos fijamos que el valor del **estadístico de contraste** coincide con el valor obtenido haciendo los cálculos a mano. Sin embargo, el p-valor no coincide. La razón es que `R` no tiene forma de saber si hemos estimado algún parámetro o no. Por este motivo, considera que los grados de libertad del **estadístico de contraste** son 4 en lugar de 3.

Entonces para hallar el p-valor correcto, hemos de usar la función `pchisq`:

```
test.chi2=chisq.test(frecuencias.empíricas,p=probabilidades.esperadas)
pchisq(test.chi2[[1]],3,lower.tail=FALSE)
```

```
## X-squared
##    0.7959
```

Comprobamos que ahora sí tenemos el p-valor correcto.

5.2. Contrastes donde la variable X_0 es continua

En esta sección, vamos a ver contrastes específicos para comprobar si una variable sigue una determinada distribución continua.

El **test de bondad de ajuste** se puede aplicar en estos casos pero exige que el tamaño de la muestra n tiene que ser grande por dos razones: en primer lugar, n debe ser mayor que 30 y en segundo lugar, las **frecuencias esperadas** deben ser mayores que 5, condición que raramente se consigue para valores de n pequeños.

5.2.1. Test de Kolmogorov-Smirnov (K-S test)

El **test de Kolgomorov-Smirnov** (**K-S**) es un test genérico para contrastar la bondad de ajuste a distribuciones continuas.

Se puede usar con muestras pequeñas (se suele recomendar 5 elementos como el tamaño mínimo para que el resultado sea significativo), pero la muestra no puede contener valores repetidos: si los contiene, la distribución del estadístico de contraste bajo la hipótesis nula no es la que predice la teoría sino que solo se aproxima a ella, y por lo tanto los p-valores que se obtienen son aproximados.

El test K-S realiza un contraste en el que la hipótesis nula es que la muestra proviene de una distribución continua completamente especificada. Es decir, no sirve por ejemplo para contrastar si la muestra proviene de "alguna" distribución normal, sino solo para contrastar si proviene de una distribución normal con una media y una desviación típica concretas.

Por tanto, si queremos contrastar que la muestra proviene de alguna distribución de una familia concreta y estimamos sus parámetros a partir de la muestra, el test K-S solo nos permite rechazar o no la hipótesis de que la muestra proviene de la distribución de esa familia con exactamente esos parámetros.

Si el resultado es rechazar la hipótesis nula, significa que tenemos indicios suficientes para rechazar que la muestra proviene de la distribución de la familia con los parámetros que hemos especificado pero podría ser que la muestra siguiese una distribución de la misma familia con otros parámetros.

Sean $x_1, x_2, \dots, x_n$, con todos los valores diferentes tal como hemos comentado, y queremos contrastar si ha sido producida por una variable X con distribución F_X.

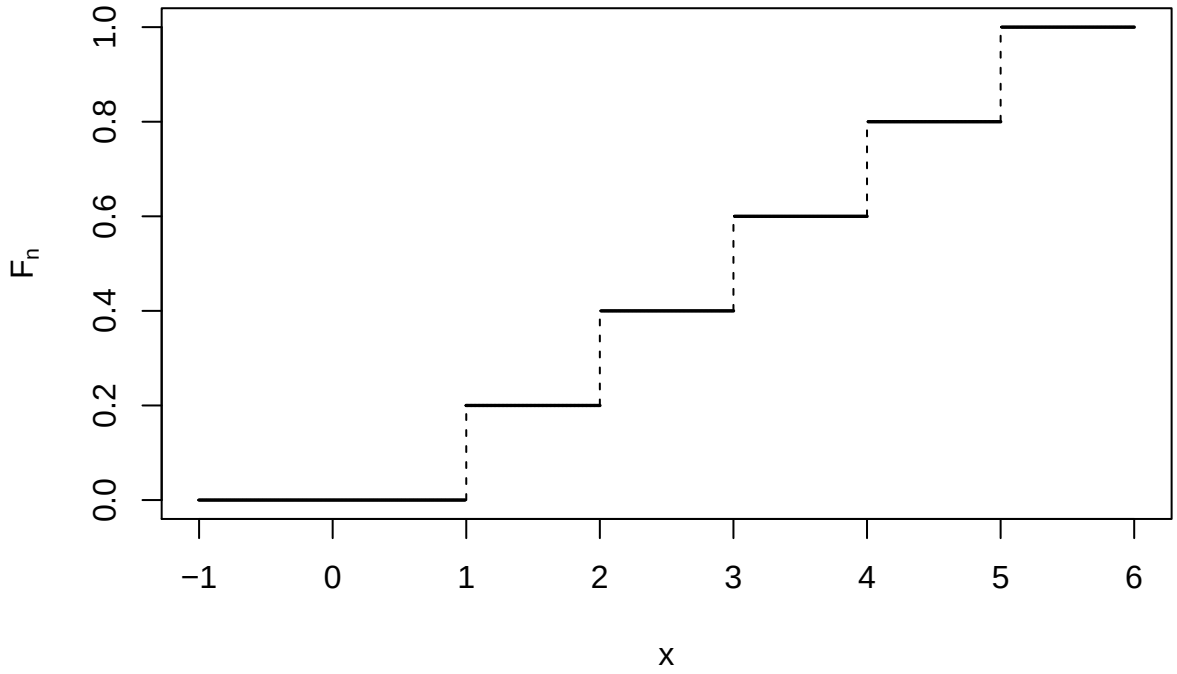
Para aplicar el test **K-S** seguimos los pasos siguientes:

- Ordenamos la muestra de menor a mayor: $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$.
- Calculamos la **función de distribución muestral** F_n de esta muestra, definida de la forma siguiente:

$$F_n(x) = \begin{cases} 0, & \text{si } x < x_{(1)}, \\ \dfrac{k}{n}, & \text{si } x_{(k)} \leq x < x_{(k+1)}, \\ 1, & \text{si } x_{(n)} \leq x. \end{cases}$$

  O sea, sería la función de distribución que corresponderia a la variable aleatoria discreta X_n con dominio $D_X = \{x_{(1)} < x_{(2)} < \cdots < x_{(n)}\}$ y función de probabilidad $P(X_n = x_{(i)}) = \frac{1}{n}$, $i = 1, \dots, n$.

El gráfico siguiente muestra la **función de distribución muestral** F_n para el caso en que la muestra sean los valores $1, 2, 3, 4, 5$:



- El siguiente paso es comparar $F_n(x)$ con $F_X(x)$. Si son muy diferentes, concluimos que tenemos indicios suficientes para rechazar que la muestra proviene de la variable X. ¿Cómo realizamos

dicha comparación?

Calculamos $\sup\{|F_n(x) - F_X(x)| \mid x \in \mathbb{R}\}$. Como F_X es creciente, este supremo se alcanza en algún extremo del "escalón" de la función F_n.

Para calcular dicho supremo, calculamos la denominada discrepancia de cada valor $x_{(i)}$:

$$\begin{aligned} D_n(x_{(i)}) &= \text{máx}\left\{|F_X(x_{(i)}) - F_n(x_{(i)}^-)|, |F_X(x_{(i)}) - F_n(x_{(i)})|\right\} \\ &= \text{máx}\left\{\left|F_X(x_{(i)}) - \frac{i-1}{n}\right|, \left|F_X(x_{(i)}) - \frac{i}{n}\right|\right\}, \end{aligned}$$

para todos los $i = 1, \dots, n$.

- El último paso es calcular la discrepancia máxima:

$$D_n = \text{máx}\left\{D_n(x_{(i)}) \mid i = 1, \dots, n\right\}.$$

El resultado interesante viene ahora:

Teorema de Kolmogorov-Smirnov Si la hipótesis nula es cierta, la distribución límite de la variable aleatoria $\sqrt{n} \cdot D_n$ no depende de la variable X y converge a la denominada **distribución de Kolmogorov** K. Aunque n tiene que ser grande, se puede hacer un cambio de variable para que el error de aproximar $\sqrt{n} \cdot D_n$ por la **distribución de Kolmogorov** K sea despreciable. Para más detalles, consultar https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

En la práctica, se aproxima $D_n \approx \frac{1}{\sqrt{n}} \cdot K$ y se calcular el p-valor como:

$$p = P\left(D_n > d_n\right),$$

donde d_n representa el valor obtenido de la variable D_n usando nuestra muestra. La probabilidad anterior se calcula usando la función de distribución de la variable $\frac{1}{\sqrt{n}} \cdot K$. Dicha función de distribución está tabulada o hay que usar `R`.

Ejemplo

Queremos decidir si los valores

$$5{,}84, 4{,}57, 1{,}34, 3{,}58, 1{,}54, 2{,}25$$

provienen de una distribución normal con $\mu = 3$ y $\sigma = 1{,}5$.

El contraste a realizar es el siguiente:

$$\begin{cases} H_0: \text{ la muestra proviene de una } X \sim N(3, 1{,}5), \\ H_0: \text{ la muestra no proviene de una } X \sim N(3, 1{,}5). \end{cases}$$

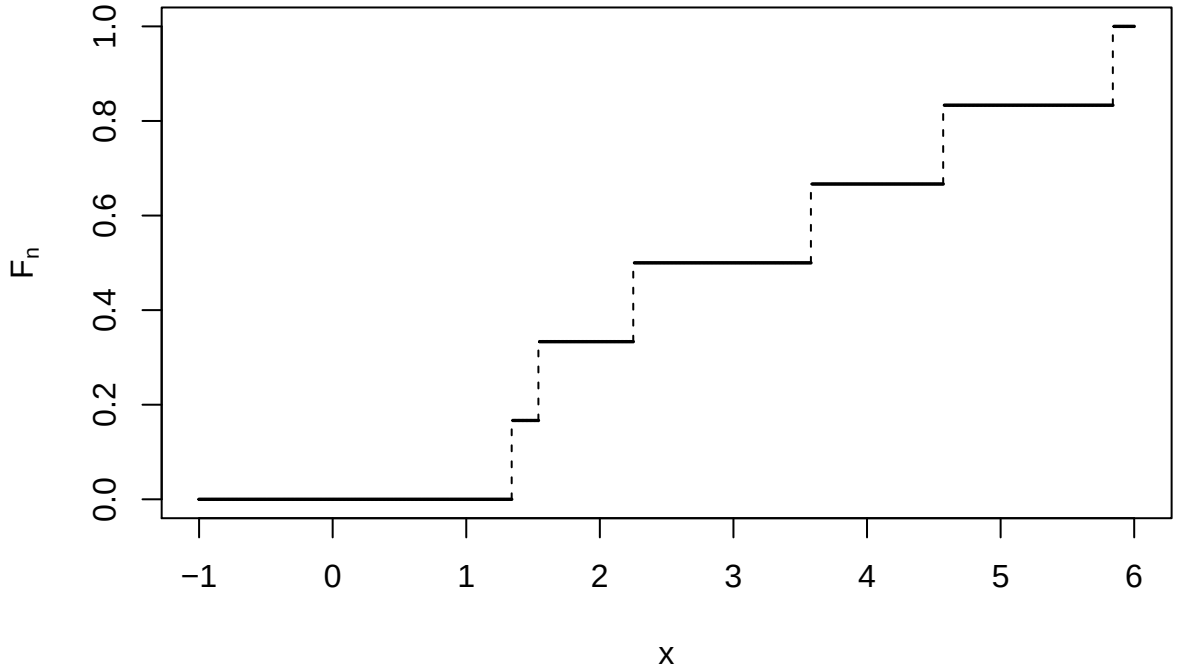
Para aplicar el test K-S, primero ordenamos la muestra usando `R` "como calculadora":

```
muestra=c(5.84,4.57,1.34,3.58,1.54,2.25)
(muestra.ordenada = sort(muestra))
```

```
## [1] 1.34 1.54 2.25 3.58 4.57 5.84
```

A continuación, calculamos la **función de distribución muestral** F_n:

$$F_n(x) = \begin{cases} 0, & \text{si } x < 1{,}34, \\ \frac{1}{6}, & \text{si } 1{,}34 \leq x < 1{,}54, \\ \frac{2}{6}, & \text{si } 1{,}54 \leq x < 2{,}25, \\ \frac{3}{6}, & \text{si } 2{,}25 \leq x < 3{,}58, \\ \frac{4}{6}, & \text{si } 3{,}58 \leq x < 4{,}57, \\ \frac{5}{6}, & \text{si } 4{,}57 \leq x < 5{,}84, \\ 1, & \text{si } 5{,}84 \leq x, \end{cases}$$



Calculamos la **discrepancia** de cada observación $x_{(i)}$ que recordemos que vale:

$$D_n(x_{(i)}) = \text{máx}\left\{\left|F_X(x_{(i)}) - \frac{i-1}{n}\right|, \left|F_X(x_{(i)}) - \frac{i}{n}\right|\right\}.$$

Esquematizemos los resultados en la tabla siguiente donde Z representa la normal estándar $N(0,1)$:

i	$x_{(i)}$	$F_X(x_{(i)})$	$\left\|F_X(x_{(i)}) - \frac{i-1}{6}\right\|$	$\left\|F_X(x_{(i)}) - \frac{i}{6}\right\|$	$D_6(x_{(i)})$
1	1,34	$F_X(1{,}34) = F_Z\left(\frac{1{,}34-3}{1{,}5}\right) = 0{,}134$	0,134	$\left\|\frac{1}{6} - 0{,}134\right\| = 0{,}032$	0,134
2	1,54	$F_X(1{,}54) = F_Z\left(\frac{1{,}54-3}{1{,}5}\right) = 0{,}165$	$\left\|0{,}165 - \frac{1}{6}\right\| = 0{,}001$	$\left\|0{,}165 - \frac{2}{6}\right\| = 0{,}168$	0,168

i	$x_{(i)}$	$F_X(x_{(i)})$	$\left\lvert F_X(x_{(i)}) - \frac{i-1}{6} \right\rvert$	$\left\lvert F_X(x_{(i)}) - \frac{i}{6} \right\rvert$	$D_6(x_{(i)})$
3	2,25	$F_X(2{,}25) = F_Z\left(\frac{2{,}25-3}{1{,}5}\right) = 0{,}309$	$\lvert 0{,}309 - \frac{2}{6} \rvert = 0{,}025$	$\lvert 0{,}309 - \frac{3}{6} \rvert = 0{,}191$	0,191
4	3,58	$F_X(3{,}58) = F_Z\left(\frac{3{,}58-3}{1{,}5}\right) = 0{,}65$	$\lvert 0{,}65 - \frac{3}{6} \rvert = 0{,}15$	$\lvert 0{,}65 - \frac{4}{6} \rvert = 0{,}016$	0,15
5	4,57	$F_X(4{,}57) = F_Z\left(\frac{4{,}57-3}{1{,}5}\right) = 0{,}852$	$\lvert 0{,}852 - \frac{4}{6} \rvert = 0{,}186$	$\lvert 0{,}852 - \frac{5}{6} \rvert = 0{,}019$	0,186
6	5,84	$F_X(5{,}84) = F_Z\left(\frac{5{,}84-3}{1{,}5}\right) = 0{,}971$	$\lvert 0{,}971 - \frac{5}{6} \rvert = 0{,}138$	$\lvert 0{,}971 - 1 \rvert = 0{,}029$	0,138

El valor del **estadístico** D_6 será el máximo de la última columna de la tabla anterior: $D_6 = 0{,}191$.

De cara a hallar el p-valor tenemos que consultar **la tabla del test de Kolmogorov-Smirnov**. Si vais a http://images.google.com y escribís "tabla de dn Kolmogorov" en la casilla de búsqueda, encontraréis un montón de tablas de la variable D_n.

La primera fila de dichas tablas corresponde al **error tipo I del contraste** y la primera columna, al tamaño de la muestra n.

Si os fijáis en la fila correspondiente a $n = 6$, veréis que el valor de $D_6 = 0{,}191$ no sale, lo que significa que el valor α debe ser mayor que 0.2. Por tanto, nuestro p-valor serà mayor que 0.2, hecho que nos hace concluir que no tenemos indicios suficientes para rechazar que la muestra se distribuya según una distribución normal $N(\mu = 3, \sigma = 1{,}5)$.

5.2.2. Contraste K-S en R

La función básica para realizar el test K-S es `ks.test`. Su sintaxis básica para una muestra es

```
ks.test(x, y, parámetros)
```

donde:

- `x` es la muestra de una variable continua.
- `y` puede ser un segundo vector, y entonces se contrasta si ambos vectores han sido generados por la misma distribución continua, o el nombre de la función de distribución (empezando con `p`) que queremos contrastar, entre comillas; por ejemplo `"pnorm"` para la distribución normal.
- Los `parámetros` de la función de distribución si se ha especificado una; por ejemplo `mean=0, sd=1` para una distribución normal estándar.

Ejemplo anterior con R

Para realizar el contraste K-S para nuestro ejemplo, tenemos que hacer lo siguiente:

```
ks.test(muestra,"pnorm",mean=3,sd=1.5)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  muestra
## D = 0.19, p-value = 0.9
## alternative hypothesis: two-sided
```

Observamos que nos da el mismo valor de **discrepancia máxima** que hemos obtenido anteriormente con un p-valor altísimo, hecho que corrobora la conlusión que hemos escrito.

5.3. Tests de normalidad

5.3.1. Test de Kolmogorov-Smirnov-Lilliefors (K-S-L)

Para contrastar si una muestra proviene de una distribución normal con los parámetros μ y σ desconocidos, el **test K-S** nos "obliga" a darles un valor.

Los valores "óptimos" para dichos parámetros serían las estimaciones dadas por los **estimadores de máxima verosimilitud**: la **media muestral** para μ y la **desviación típica muestral** para sigma.

La **prueba de Kolmogorov-Smirnov-Lilliefors** consiste en estimar dichos parámetros, calcular la **discrepancia máxima** tal como hemos explicado pero a la hora de calcular el p-valor, se usa otra distribución, llamada **distribución de Lilliefors** en lugar de usar la **distribución de Kolmogorov** ya que con la **distribución de Lilliefors**, el contraste es más robusto.

Vamos a aplicar el **test K-S-L** a nuestra muestra anterior para ver si se distribuye según la ley normal.

El **test K-S-L** test en `R` se aplica usando la función `lillie.test` del paquete `nortest`:

```
library(nortest)
lillie.test(muestra)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  muestra
## D = 0.2, p-value = 0.6
```

Vemos que el p-valor no es tan grande como en el caso anterior pero aún así, es suficientemente grande para concluir que no tenemos indicios suficientes para rechazar que nuestra muestra siga la distribución normal.

El test K-S-L tiene un inconveniente: aunque es muy sensible a las diferencias entre la muestra y la distribución teórica alrededor de sus valores medios, le cuesta detectar diferencias prominentes en un extremo u otro de la distribución.

Su potencia se ve afectada por dicho inconveniente.

Veamos un ejemplo de este hecho intentando ver si una muestra de una distribución t de Student nos acepta que es normal o no:

```
set.seed(100)
x=rt(50,3)
lillie.test(x)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.1, p-value = 0.2
```

Nos dice que no podemos rechazar que la muestra `x` sea normal.

Esto es debido a que la función de densidad de la distribución t de Student es algo más aplanada que la distribución normal, donde en los dos extremos está por encima de la de la normal.

Como el test K-S-L no detecta las diferencias en los extremos, acepta que `x` es normal.

5.3.2. Test de normalidad de Anderson-Darling (A-D)

El **test de normalidad de Anderson-Darling** resuelve el inconveniente del **test de K-S-L**.

Este test está implementado en la función `ad.test` del paquete `nortest`.

Si ahora, aplicamos el **test A-D** a la muestra anterior de la distribución t de Student, la normalidad queda rechazada:

```
ad.test(x)
```

```
##
##  Anderson-Darling normality test
##
## data:  x
## A = 1.2, p-value = 0.004
```

Un inconveniente común a los **tests K-S-L y A-D** es que, si bien pueden usarse con muestras pequeñas (pongamos de más de 5 elementos), se comportan mal con muestras grandes, de varios miles de elementos.

En muestras de este tamaño, cualquier pequeña divergencia de la normalidad se magnifica y en estos dos tests aumenta la probabilidad de errores de tipo I.

5.3.3. Test de Shapiro-Wilks (S-W)

Un test que resuelve este problema es el de **Shapiro-Wilk** (**S-W**), implementado en la función `shapiro.test` de la instalación básica de `R`.

Apliquemos el **test S-W** a las dos muestras anteriores: la muestra del ejemplo y la muestra de la distribución t de Student:

```
shapiro.test(muestra)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  muestra
## W = 0.93, p-value = 0.5
```

```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.89, p-value = 0.0003
```

Vemos que acepta que la muestra del ejemplo anterior sea normal y rechaza la normalidad en el caso de la muestra de la t de Student.

Si nuestra muestra de valores tiene empates, los p-valores de los contrastes calculados a partir de las distribuciones de los estadísticos usados en los tests **K-S-L**, **A-D** y **S-W** se pueden ver afectados hasta el punto de que, si hay muchos empates, su significado no tenga ningún sentido.

Hay que decir que el menos afectado por los empates es el test de **S-W**.

5.3.4. Test omnibus de D'Agostino-Pearson

Un test que no es sensible a los empates es el test de normalidad de **D'Agostino-Pearson**.

Este test se encuentra implementado en la función `dagoTest` del paquete **fBasics**, y lo que hace es cuantificar lo diferentes que son la asimetría y la curtosis de la muestra (dos parámetros estadísticos relacionados con la forma de la gráfica de la función de densidad muestral) respecto de los esperados en una distribución normal, y resume esta discrepancia en un p-valor con el significado usual.

Para poder aplicar dicho test, el tamaño de la muestra debe ser 20 como mínimo.

Por tanto, sólo podemos aplicar dicho test a la muestra de datos correspondiente a la distribución t de Student:

```
library(fBasics)
dagoTest(x)
```

```
##
## Title:
##  D'Agostino Normality Test
##
## Test Results:
##   STATISTIC:
##     Chi2 | Omnibus: 21.8125
##     Z3   | Skewness: 2.8069
##     Z4   | Kurtosis: 3.7328
```

```
##   P VALUE:
##     Omnibus  Test: 0.00001834
##     Skewness Test: 0.005001
##     Kurtosis Test: 0.0001894
##
## Description:
##  Thu Oct 14 12:44:43 2021 by user:
```

Si nos fijamos en el resultado, el test calcula tres estadísticos de contraste: el test **Omnibus** basado en la distribución χ^2, el test de asimetría y el test de curtosis con sus correspondientes p-valores. Para más información, id a https://en.wikipedia.org/wiki/D%27Agostino%27s_K-squared_test

Vemos que según el test de **D'Agostino-Pearson**, la muestra `x` correspondiente a la distribución t de Student no sigue la ley normal.

5.4. Guía rápida

- `qqPlot` del paquete **car**, sirve para dibujar un Q-Q-plot de una muestra contra una distribución teórica. Sus parámetros principales son:
  - `distribution`: el nombre de la familia de distribuciones, entre comillas.
  - Los parámetros de la distribución: `mean` para la media, `sd` para la desviación típica, `df` para los grados de libertad, etc.
  - Los parámetros usuales de `plot`.
- `chisq.test` sirve para realizar tests χ^2 de bondad de ajuste. Sus parámetros principales son:
  - `p`: el vector de probabilidades teóricas.
  - `rescale.p`: igualado a `TRUE`, indica que los valores de `p` no son probabilidades, sino sólo proporcionales a las probabilidades.
  - `simulate.p.value`: igualado a `TRUE`, R calcula el p-valor mediante simulaciones.
  - `B`: en este último caso, permite especificar el número de simulaciones.
- `ks.test` realiza el test de Kolmogorov-Smirnov. Tiene dos tipos de uso:
  - `ks.test(x,y)`: contrasta si los vectores `x` e `y` han sido generados por la misma distribución continua.
  - `ks.test(x, "distribución", parámetros)`: contrasta si el vector `x` ha sido generado por la distribución especificada, que se ha de indicar con el nombre de la función de distribución de R (la que empieza con p).
- `lillie.test` del paquete **nortest**, realiza el test de normalidad de Kolmogorov-Smirnov-Lilliefors.
- `ad.test` del paquete **nortest**, realiza el test de normalidad de Anderson-Darling.
- `shapiro.test`, realiza el test de normalidad de Shapiro-Wilk.
- `dagoTest` del paquete **fBasics**, realiza el test ómnibus de D'Agostino-Pearson.

Capítulo 6

Contrastes de independencia y homogeneidad

Una de las aplicaciones más usadas del test de bondad de ajuste es contrastar si dos maneras de clasificar n objetos son **independientes** o no.

Veamos un ejemplo ilustrativo:

Ejemplo

En un estudio de una vacuna de hepatitis participan 1083 voluntarios. De éstos, se eligen aleatoriamente 549 y son vacunados. Los otros, 534, no son vacunados. Después de un cierto tiempo, se observa que 70 de los 534 no vacunados han contraído la hepatitis, mientras que sólo 11 de los 549 vacunados la han contraído.

Esquematicemos los resultados en lo que se llama una tabla de contingencia:

¿Enfermó?/¿Vacunado?	Sí	No	Total
Sí	11	70	81
No	538	464	1002
Total	549	534	1083

¿Es el hecho de contraer hepatitis independiente de haber sido vacunado contra la dolencia?

En este ejemplo, contrastar si la manera de clasificar a los voluntarios entre vacunados y no vacunados y la manera de clasificarlos entre enfermos por hepatitis y no enfermos es equivalente a contrastar si la vacuna es efectiva contra la hepatitis.

Decir que la vacuna no es efectiva sería equivalente a decir que vacunar a un individuo es independiente de que contraiga la hepatitis.

6.1. Tablas de contingencia

La situación en general sería la siguiente:

Tenemos n individuos y los clasificamos según dos criterios: X e Y. Sean $x_1, \dots, x_I$ los distintos **niveles** del criterio X e $y_1, \dots, y_J$, los distintos **niveles** del criterio Y.

En el ejemplo anterior $I = J = 2$, X sería el criterio de clasificación por vacunación con x_1 :"vacunados" y x_2 :"no vacunados" e Y sería el criterio de clasificación por contracción de la hepatitis con y_1 :"enfermo de hepatitis" e y_2 :"no enfermo de hepatitis".

Definimos n_{ij} como el número de individuos clasificados en el nivel x_i según el criterio X y clasificados en el nivel y_j según el criterio Y. A partir de dichos valores construimos la denominada **tabla de contingencia**:

X/Y	y_1	...	y_j	...	y_J	$n_{i\bullet}$
x_1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet J}$	n

En la tabla anterior, $n_{i\bullet}$ sería el número total de individuos clasificados en el nivel x_i según el criterio X y $n_{\bullet j}$, el número total de individuos clasificados en el nivel y_j según el criterio Y.

El contraste que nos planteamos es el siguiente:

$$\left.\begin{array}{l} H_0: \text{ Los criterios de clasificación } X \text{ e } Y \text{ son independientes,} \\ H_1: \text{ Los criterios de clasificación } X \text{ y } Y \text{ no son independientes.} \end{array}\right\}$$

Para poder realizar el contraste anterior, lo plantearemos como un contraste de **bondad de ajuste.**

6.2. Contraste de independencia como un contraste de bondad de ajuste

Para poder modelar el contraste de independencia como un contraste de **bondad de ajuste**, en primer lugar necesitamos definir una variable "modelo".

A partir de nuestros datos empíricos, contrastaremos si dichos datos siguen la variable "modelo" usando el test de la χ^2 de bondad de ajuste.

Nuestra variable "modelo" será una variable aleatoria discreta **bidimensional** (X, Y) con dominio $\{(x_1, y_1), \dots, (x_I, y_J)\}$, o, si se quiere $\{(x_i, y_j) \mid i = 1, \dots, I,\ j = 1, \dots, J\}$. Sería una variable con $I \cdot J$ valores.

Para calcular la función de probabilidad de la variable (X, Y), hay que suponer que la hipótesis nula H_0 es cierta o que los criterios X e Y son **independientes**. Por tanto:

$$P((X,Y)=(x_i,y_j)) = P(X=x_i)\cdot P(Y=y_j) = \frac{n_{i\bullet}}{n}\cdot\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}\cdot n_{\bullet j}}{n^2},$$

$i = 1,\dots,I$, $j = 1,\dots,J$.

Los valores n_{ij} serían los **valores empíricos** con los que hay que contrastar si siguen la distribución de la variable (X,Y).

6.3. Test χ^2 de independencia

Una vez planteado el contraste de independencia como una contraste de **bondad de ajuste**, recordemos que el **test** χ^2 asociado al contraste es:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - n\cdot P((X,Y)=(x_i,y_j))\right)^2}{n\cdot P((X,Y)=(x_i,y_j))} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - \frac{n_{i\bullet}\cdot n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\cdot n_{\bullet j}}{n}},$$

donde las frecuencias n_{ij} serían las **frecuencias observadas** y las frecuencias $\frac{n_{i\bullet}\cdot n_{\bullet j}}{n}$ serían las **frecuencias esperadas**.

Recordemos que si n es grande y cada frecuencia esperada $\frac{n_{i\bullet}\cdot n_{\bullet j}}{n}$ es ≥ 5, este estadístico sigue aproximadamente una ley χ^2 con $(I-1)\cdot(J-1)$ **grados de libertad**.

6.4. Test χ^2 de independencia

Observación.

¿Por qué los **grados de libertad** del **estadístico de contraste** son $(I-1)\cdot(J-1)$?

La razón es la siguiente: recordemos que al realizar un **test de bondad de ajuste**, los **grados de libertad** del estadístico χ^2 era: $g.l.$ = número de clases − número de parámetros estimados − 1.

Fijémonos que en el contraste de independencia hemos estimado $I+J-2$ **parámetros** que corresponden a las **medias de las variables** X **e** Y: $\frac{n_{i\bullet}}{n}$ y $\frac{n_{\bullet j}}{n}$, $i = 1,\dots,I$, $j = 1,\dots,J$.

En nuestro caso, nos queda:

$$g.l. = I\cdot J - (I+J-2) - 1 = (I-1)\cdot(J-1).$$

Como siempre, sea χ_0 el valor que toma el **estadístico de contraste**. El **p-valor** del contraste es:

$$p = P(\chi^2_{(I-1)\cdot(J-1)} \geq \chi_0),$$

con el significado usual:

- si $p < 0{,}05$, concluimos que tenemos evidencias suficientes para rechazar la independencia de los criterios,
- si $p > 0{,}1$, concluimos que no tenemos evidencias suficientes para rechazar la independencia de los criterios y,

- si $0{,}05 \leq p \leq 0{,}1$, estamos en la zona de penumbra. Necesitamos más datos para tomar una decisión clara.

Ejemplo del estudio de la vacuna de hepatitis

Recordemos que la tabla de contingencia de dicho estudio era:

¿Enfermó?/¿Vacunado?	Sí	No	Total
Sí	11	70	81
No	538	464	1002
Total	549	534	1083

A continuación, calculemos la tabla de las frecuencias esperadas $\frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$ a partir de las sumas de las filas y las columnas anteriores:

¿Enfermó?/¿Vacunado?	Sí	No	Total
Sí	$\frac{81 \cdot 549}{1083} = 41{,}061$	$\frac{81 \cdot 534}{1083} = 39{,}939$	81
No	$\frac{1002 \cdot 549}{1083} = 507{,}939$	$\frac{1002 \cdot 534}{1083} = 494{,}061$	1002
Total	549	534	1083

El valor del estadístico de contraste será:

$$\chi_0 = \frac{(11-41{,}061)^2}{41{,}061} + \frac{(70-39{,}939)^2}{39{,}939} + \frac{(538-507{,}939)^2}{507{,}939} + \frac{(464-494{,}061)^2}{494{,}061} = 48{,}242.$$

El p-valor del contraste será:

$$p = P(\chi^2_1 > 48{,}242) = 0.$$

Como el p-valor es muy pequeño, de hecho despreciable, concluimos que tenemos suficientes evidencias para rechazar que la vacuna y la hepatitis son independientes. O sea, vacunarse afecta al hecho de contraer la enfermedad.

Ejemplo: Lobas

Un investigador quiere saber si el número de crías por loba es independiente de la zona donde viva. Para ello, considera 3 zonas (X): X_1 =“Norte”, X_2 =“Centro” y X_3 =“Sur”.

Clasifica los números de crías (Y) en Y_1 =”Dos o menos“, Y_2 =” Entre tres y cinco“, Y_3 =”Entre seis y ocho“ Y_4 =“Nueve o más”.

Queremos hacer el contraste:

$$\left.\begin{array}{ll} H_0: & \text{El número de crías por loba es independiente de la zona,} \\ H_1: & \text{El número de crías por loba no es independiente de la zona.} \end{array}\right\}$$

Toma una muestra de 200 lobas y obtiene la tabla siguiente:

X/Y	Y_1	Y_2	Y_3	Y_4	$n_{i\bullet}$
X_1	5	8	15	22	50
X_2	20	26	46	8	100
X_3	15	10	15	10	50
$n_{\bullet j}$	40	44	76	40	200

Las frecuencias esperadas si las variables son independientes son:

X/Y	Y_1	Y_2	Y_3	Y_4	$n_{i\bullet}$
X_1	10	11	19	10	50
X_2	20	22	38	20	100
X_3	10	11	19	10	50
$n_{\bullet j}$	40	44	76	40	200

El valor del estadístico de contraste será:

$$\begin{aligned}\chi_0 = & \frac{(5-10)^2}{10}+\frac{(8-11)^2}{11}+\frac{(15-19)^2}{19}+\frac{(22-10)^2}{10}+\frac{(20-20)^2}{20}+\frac{(26-22)^2}{22}\\ & +\frac{(46-38)^2}{38}+\frac{(8-20)^2}{20}+\frac{(15-10)^2}{10}+\frac{(10-11)^2}{11}+\frac{(15-19)^2}{19}+\frac{(10-10)^2}{10}\\ = & \ 31{,}605.\end{aligned}$$

El p-valor del contraste vale:

$$p = P(\chi^2_6 > 31{,}605) = 0.$$

Como el p-valor es muy pequeño, de hecho despreciable, concluimos que tenemos suficientes evidencias para rechazar que el número de crías y la zona donde viven las lobas son independientes.

6.5. Contraste de independencia en R

Para realizar un contraste de independencia en R hay que usar la función `chisq.test(tabla.contingencia,correct)` con los parámetros siguientes:

- `tabla.contingencia`: es la tabla de las frecuencias empíricas.
- `correct`: es un parámetro lógico. Si su valor es `FALSE`, hará los cálculos como hemos explicado. Si su valor es `TRUE`, aplica la corrección a la continuidad sólo para tablas de contingencia 2×2. (Ver https://es.wikipedia.org/wiki/Correcci%C3%B3n_de_Yates)

6.6. Contraste de independencia en R

Ejemplo del estudio de la vacuna de hepatitis

Para realizar el contraste de independencia en R hacemos lo siguiente:

```
chisq.test(matrix(c(11,538,70,464),2,2),correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  matrix(c(11, 538, 70, 464), 2, 2)
## X-squared = 48, df = 1, p-value = 0.000000000004
```

Observamos que obtenemos los mismos valores que en los cálculos hechos a mano.

6.6.1. Ejemplo

Ejemplo del estudio del número de crías de una loba y la zona donde vive

En este ejemplo, para realizar el contraste en R, basta hacer:

```
chisq.test(matrix(c(5,20,15,8,26,10,15,46,15,22,8,10),3,4))
```

```
##
##  Pearson's Chi-squared test
##
## data:  matrix(c(5, 20, 15, 8, 26, 10, 15, 46, 15, 22, 8, 10), 3, 4)
## X-squared = 32, df = 6, p-value = 0.00002
```

En este caso, también se obtienen los mismos valores que en los cálculos realizados anteriormente.

6.7. Contraste de independencia en R

Ejemplo: WorldPhones

La tabla de datos `WorldPhones` de R nos da el número de teléfonos (en miles de unidades) que había en distintas regiones del mundo en los años 1951, 1956, 1957, 1958, 1959, 1960 y 1961:

```
WorldPhones
```

```
##      N.Amer Europe Asia S.Amer Oceania Africa Mid.Amer
## 1951  45939  21574 2876   1815    1646     89      555
## 1956  60423  29990 4708   2568    2366   1411      733
## 1957  64721  32510 5230   2695    2526   1546      773
## 1958  68484  35218 6662   2845    2691   1663      836
## 1959  71799  37598 6856   3000    2868   1769      911
## 1960  76036  40341 8220   3145    3054   1905     1008
## 1961  79831  43173 9053   3338    3224   2005     1076
```

Como puede observarse, las regiones son: Norte América, Europa, Asia, Sudamérica, Oceanía, África y América Central.

Vamos a contrastar si el año es independiente de la región para el hecho de tener teléfono o no.

Para hallar las distribuciones marginales de las variables "Año" y "Región" hemos de usar la función `addmargins`:

```
(Tabla.con.marginales = addmargins(WorldPhones))
```

```
##      N.Amer Europe  Asia S.Amer Oceania Africa Mid.Amer    Sum
## 1951  45939  21574  2876   1815    1646     89      555  74494
## 1956  60423  29990  4708   2568    2366   1411      733 102199
## 1957  64721  32510  5230   2695    2526   1546      773 110001
## 1958  68484  35218  6662   2845    2691   1663      836 118399
## 1959  71799  37598  6856   3000    2868   1769      911 124801
## 1960  76036  40341  8220   3145    3054   1905     1008 133709
## 1961  79831  43173  9053   3338    3224   2005     1076 141700
## Sum  467233 240404 43605  19406   18375  10388     5892 805303
```

La tabla de frecuencias esperadas sería:

```
(tabla.frec.esperadas = rowSums(WorldPhones)%*%
     t(colSums(WorldPhones))/sum(WorldPhones))
```

```
##      N.Amer Europe Asia S.Amer Oceania Africa Mid.Amer
## [1,]  43221  22238 4034   1795    1700  960.9    545.0
## [2,]  59295  30509 5534   2463    2332 1318.3    747.7
## [3,]  63822  32838 5956   2651    2510 1419.0    804.8
## [4,]  68695  35345 6411   2853    2702 1527.3    866.3
## [5,]  72409  37256 6758   3007    2848 1609.9    913.1
## [6,]  77577  39916 7240   3222    3051 1724.8    978.3
## [7,]  82214  42301 7673   3415    3233 1827.9   1036.7
```

Para realizar el contraste, usamos tal como hemos indicado la función `chisq.test`:

```
chisq.test(WorldPhones)
```

```
##
##  Pearson's Chi-squared test
##
## data:  WorldPhones
## X-squared = 2194, df = 36, p-value <0.0000000000000002
```

Vemos que el p-valor es depreciable. Concluimos que tenemos evidencias suficientes para rechazar que el año y la zona son independientes respecte del hecho de tener teléfono o no.

6.7.1. Caso en que las frecuencias esperadas son inferiores a 5

Si algunas frecuencias absolutas esperadas son inferiores a 5, la aproximación del p-valor por una distribución χ^2 podría no ser adecuada.

Si se da esta situación, lo mejor es recurrir a simular el p-valor usando el parámetro `simulate.p.value=TRUE`.

Veamos un ejemplo en el que se da esta situación:

6.7.2. Ejemplo

Consideremos la tabla de datos `iris`. Nos planteamos si la especie de la flor es independiente de la longitud del pétalo.

En primer lugar, hallamos la tabla de contingencia de las dos variables (especie y longitud del pétalo), agrupando en cuatro clases la variable continua "Longitud del pétalo":

```
(tabla.contingencia =table(cut(iris$Petal.Length,4),iris$Species))
```

```
##
##                setosa versicolor virginica
##   (0.994,2.48]     50          0         0
##   (2.48,3.95]       0         11         0
##   (3.95,5.43]       0         39        22
##   (5.43,6.91]       0          0        28
```

Calculemos a continuación la tabla de frecuencias esperadas:

```
(tabla.frec.esperadas = rowSums(tabla.contingencia)%*%t(colSums(tabla.contingencia))
 /sum(tabla.contingencia))
```

```
##      setosa versicolor virginica
## [1,] 16.667     16.667    16.667
## [2,]  3.667      3.667     3.667
## [3,] 20.333     20.333    20.333
## [4,]  9.333      9.333     9.333
```

¡Uy! Observamos que hay frecuencias esperadas menores que 5.

Veamos que pasa si usamos la función `chisq.test`:

```
chisq.test(tabla.contingencia)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla.contingencia
## X-squared = 216, df = 6, p-value <0.0000000000000002
```

`R` nos avisa que la aproximación puede ser incorrecta debido al hecho de que tenemos frecuencias esperadas menores que 5.

Para resolver este inconveniente, simularemos el valor del p-valor reiniciando la semilla a `NULL`:

```
set.seed(NULL)
chisq.test(tabla.contingencia,simulate.p.value = TRUE, B=5000)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 5000
##  replicates)
##
## data:  tabla.contingencia
## X-squared = 216, df = NA, p-value = 0.0002
```

El parámetro `B` indica el número de replicaciones que hemos de realizar en nuestra simulación.

Observamos que el p-valor es pequeño. Por tanto, concluimos que tenemos evidencias suficientes para rechazar que la longitud del pétalo y la especie sean independientes.

6.7.3. Diseño experimental en el contraste de independencia

En un **contraste de independencia** se toma una **muestra transversal** de la población, es decir, se selecciona al azar una cierta cantidad de individuos de la población, se observan las dos variables sobre cada uno de ellos, y se contrasta si las probabilidades conjuntas son iguales al producto de las probabilidades marginales de cada variable.

En el ejemplo de la vacuna y la hepatitis, seleccionamos una muestra de 1083 individuos de toda la población, y los clasificamos en 4 subgrupos: los que han sido vacunados y han contraído la enfermedad (ha habido 11), los que no han sido vacunados y han contraído la enfermedad (ha habido 70), los que han sido vacunados y no han contraído la enfermedad (ha habido 538) y, por último, los que no han sido vacunados y no han contraído la enfermedad (ha habido 534).

6.8. Contrastes de homogeneidad

En un contraste de homogeneidad, tenemos dos variables discretas X e Y y nos planteamos si la distribución de la variable Y depende de los valores de la variable X.

El contraste planteado es el siguiente:

$$\left.\begin{array}{ll} H_0: & \text{La distribución de la variable condicional } Y|X=x \\ & \text{es la misma para cualquier valor de } x \text{ del dominio de la variable } X, \\ H_1: & \text{La distribución de la variable condicional } Y|X=x \\ & \text{no es la misma para cualquier valor de } x. \end{array}\right\}$$

6.8.1. Tablas de contingencia

Para realizar un contraste de homogeneidad, tenemos al igual que en el contraste de independencia n individuos clasificados según la variables X e Y.

Recordemos que $x_1, \ldots, x_I$ los distintos **valores** de la variable X e $y_1, \ldots, y_J$, los distintos **valores** de la variable Y.

Recordemos también que n_{ij} es el número de individuos clasificados en valor x_i según la variable X y clasificados en el valor y_j según la variable Y. A partir de dichos valores construimos lo que hemos definido como la **tabla de contingencia**:

X/Y	y_1	...	y_j	...	y_J	$n_{i\bullet}$
x_1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

X/Y	y_1	...	y_j	...	y_J	$n_{i\bullet}$
x_I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet J}$	n

En la tabla anterior, $n_{i\bullet}$ sería el número total de individuos clasificados en el nivel x_i según el criterio X y $n_{\bullet j}$, el número total de individuos clasificados en el nivel y_j según el criterio Y.

6.8.2. Contraste de homogeneidad como un contraste de bondad de ajuste

Para poder modelar el contraste de homogeneidad como un contraste de **bondad de ajuste**, en primer lugar necesitamos definir una variable "modelo".

La fila i-ésima de la **tabla de contingencia** anterior representa la tabla de **frecuencias empíricas** de la variable $Y|X = x_i$:

$Y\|X = x_i$	y_1	...	y_j	...	y_J	Total
	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i\bullet}$

Si la hipótesis nula es cierta, o, si la distribución de la variable condicionada $Y|X = x_i$ es la misma para cualquier x_i, significa que la distribución de cada fila "coincidirá" con la distribución de la última fila de la tabla de contingencia:

$Y\|X = x$	y_1	...	y_j	...	y_J	Total
	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet J}$	n

Dicha variable será nuestra **variable "modelo"**. Nuestro problema será, pues, chequear si la distribución de las **frecuencias empíricas** (fila i-ésima de la **tabla de contingencia**) coincide con la distribución de las **frecuencias teóricas** (última fila de la **tabla de contingencia**). ¿Pero cómo podemos compararlas si, en principio tendrán tamaño diferente?

Antes de poder aplicar toda la maquinaría del contraste χ^2 de bondad de ajuste, necesitamos que la suma de las **frecuencias empíricas** coincida con la suma de las **frecuencias teóricas**.

La suma de las **frecuencias empíricas** vale $n_{i\bullet}$ y la suma de las **frecuencias teóricas**, n.

Para "forzar" que las dos sumas sean iguales, multiplicaremos las **frecuencias teóricas** por $\frac{n_{i\bullet}}{n}$. Así, la tabla de **frecuencias teóricas** quedará:

$Y\|X = x$	y_1	...	y_j	...	y_J	Total
	$\frac{n_{i\bullet}n_{\bullet 1}}{n}$	...	$\frac{n_{i\bullet}n_{\bullet j}}{n}$	...	$\frac{n_{i\bullet}n_{\bullet J}}{n}$	$n_{i\bullet}$

Si aplicamos el test χ^2 de bondad de ajuste, tenemos que el estadístico de contraste será:

$$\chi_i^2 = \sum_{j=1}^{J} \frac{(\text{frec. empíricas} - \text{frec. teóricas})^2}{\text{frec. teóricas}} = \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet J}}{n}}.$$

Ahora bien, tenemos en total i filas. Por tanto, tenemos que sumar el estadístico anterior para todas las filas:

$$\chi^2 = \sum_{i=1}^{I} \chi_i^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet J}}{n}}.$$

Observemos que nos queda la misma expresión que el **estadístico de contraste** que en el **contraste de independencia.**

Por tanto, desde el punto de vista de cómputo o de cálculo, realizar un **contraste de independencia** o de **homogeneidad** sería lo mismo.

Ahora bien, desde el punto de vista de diseño de experimentos, no.

Recordemos que en un **contraste de independencia** se toma una **muestra transversal** de la población. En cambio, en un **contraste de homogeneidad** se escoge una de las variables (para todo el estudio realizado, sería la variable X) y para cada uno de sus posibles valores se toma una muestra aleatoria, de tamaño prefijado, de individuos con ese valor para esa variable; su unión forma una **muestra estratificada** en el sentido que hemos visto en la sección de Muestreo.

Ejemplo del estudio del número de crías de una loba y la zona donde vive

Recordemos que la tabla de contingencia en este ejemplo era:

X/Y	Y_1	Y_2	Y_3	Y_4	$n_{i\bullet}$
X_1	5	8	15	22	50
X_2	20	26	46	8	100
X_3	15	10	15	10	50
$n_{\bullet j}$	40	44	76	40	200

Recordemos que la variable X representaba la zona donde vive la loba y la variable Y, el número de crías.

Los valores de X eran: X_1 =“Norte”, X_2 =“Centro” y X_3 =“Sur”.

Los valores de Y eran: Y_1 =”Dos o menos“, Y_2 =” Entre tres y cinco“, Y_3 =”Entre seis y ocho“ Y_4 =“Nueve o más”.

El contraste de homogeneidad consiste en este caso en testear si la distribución de la variable Y, número de crías, es la misma para los tres valores de la variable X.

Para que dicho estudio tenga sentido (desde el punto de vista de la homogeneidad), tendremos que **estratificar** primero la población de las lobas según los valores de la variable X. Tendremos que elegir un número prefijado de lobas que vivan en el norte, un número prefijado de lobas que vivan en el centro y un número prefijado de lobas que vivan en el sur.

Según los datos de la tabla anterior, se eligen aleatoriamente 50 lobas que viven en el norte, 100 lobas

que viven en el centro y 50 lobas que viven en el sur. Fijaos que estos tres números se fijan antes de realizar los tres muestreos, cosa que no pasaba en el contraste de independencia. En dicho contraste, se fijaba un sólo número que sería el número total de lobas a clasificar.

El contraste realizado en `R` era el siguiente:

```
chisq.test(matrix(c(5,20,15,8,26,10,15,46,15,22,8,10),3,4))
```

```
##
##  Pearson's Chi-squared test
##
## data:  matrix(c(5, 20, 15, 8, 26, 10, 15, 46, 15, 22, 8, 10), 3, 4)
## X-squared = 32, df = 6, p-value = 0.00002
```

Como el p-valor es muy pequeño, concluimos que tenemos indicios suficientes para rechazar que la distribución del número de crías de las lobas es la misma para las tres zonas donde viven.

6.8.3. Guía rápida

- `table` calcula tablas de contingencia de frecuencias absolutas.
- `prop.table` calcula tablas de contingencia de frecuencias relativas.
- `addmargins` sirve para añadir a una `table` una fila o una columna obtenidas aplicando una función a todas las columnas o a todas las filas de la tabla, respectivamente. Sus parámetros principales son:
  - `margin`: igualado a `1`, se aplica la función por columnas, añadiendo una nueva fila; igualado a `2`, se aplica la función por filas, añadiendo una nueva columna; igualado a `c(1,2)`, que es su valor por defecto, hace ambas cosas.
  - `FUN`: la función que se aplica a las filas o columnas; su valor por defecto es `sum`.
- `colSums` calcula un vector con las sumas de las columnas de una matriz o una tabla.
- `rowSums` calcula un vector con las sumas de las filas de una matriz o una tabla.
- `chisq.test` sirve para realizar tests χ^2 de independencia y homogeneidad. El resultado es una `list` formada, entre otros, por los objetos siguientes: `statistic` (el valor del estadístico X^2), `parameter` (los grados de libertad) y `p.value` (el p-valor). Sus parámetros principales en el contexto de esta lección son:
  - `simulate.p.value`: igualado a `TRUE`, calcula el p-valor mediante simulaciones.
  - `B`: en este último caso, permite especificar el número de simulaciones.

Capítulo 7

Análisis de la Varianza

El problema que intentamos resolver es un **contraste de igualdad de medias** cuando tenemos **más de dos poblaciones.**

Concretamente, supongamos que $k > 2$ poblaciones.

Algunas veces, segregaremos la población por k subpoblaciones definidas por los niveles de un factor.

Por ejemplo, si queremos estudiar el peso de una población de una determinada edad, pongamos entre 30 y 40 años, podemos segregar por tipo de sangre, A, B, O y AB ($k = 4$) y preguntarnos si el peso medio de cada subpoblación segredada por el tipo de sangre es el mismo.

Más concretamente, sean $\mu_1, \dots, \mu_k$ las medias de esta magnitud (peso en el ejemplo anterior) en cada una de las subpoblaciones o poblaciones. Nos planteamos el contraste siguiente:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \\ H_1 : \exists i, j \mid \mu_i \neq \mu_j. \end{cases}$$

Como es habitual, la decisión que tomaremos será en función de una muestra aleatoria de cada población.

La técnica que resuelve el contraste anterior se llama **ANOVA** de *ANalysis Of VAriance* en inglés, **Análisis de la varianza**, es castellano.

Esta técnica se puede aplicar bajo diferentes diseños de experimentos:

- según cuántos factores usamos para separar la población en subpoblaciones,
- según cómo escogemos los niveles de los factores,
- según cómo escogemos las muestras.

Veremos los diseños más básicos. En un problema concreto, se tiene que decidir primero el tipo de experimento que se tiene que realizar.

Recordemos que en el caso de $k = 2$ poblaciones, para realizar la **comparación de sus medias**, calculábamos las medias de dos muestras y las comparábamos usando el **estadístico de contraste** correspondiente.

Si quisiéramos hacer lo mismo para $k \geq 3$ poblaciones, tendríamos que comparar en total $\binom{k}{2}$ pares de medias.

Para ver si hay diferencias, tenemos que realizar todas las comparaciones ya que podría pasar lo siguiente:

$$\mu_1 \approx \mu_2, \mu_2 \approx \mu_3, \mu_3 \approx \mu_4 \text{ pero } \mu_1 \not\approx \mu_4$$

Queremos un test que nos diga en un solo paso si todas son iguales, o si hay alguna diferente. Si hubiera diferencias, en una segunda fase, buscaríamos en qué pares de poblaciones hay medias diferentes.

El **test ANOVA** realiza la comparación de las medias de 3 o más poblaciones basándose en la variabilidad de los datos por grupos:

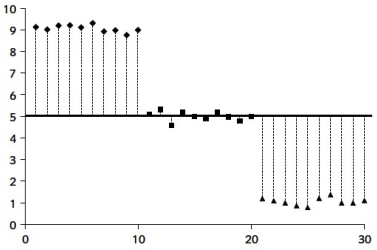
- **variabilidad de los datos (respecto de la media global)**,
- **variabilidad dentro de cada población** (respecto de la media dentro de la población),
- **variabilidad de las medias por poblaciones (respecto de la media global)**.

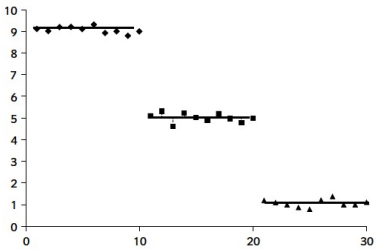
La idea del **test ANOVA** es la siguiente: si la **variabilidad total de los datos** es explicada por la **variabilidad de las medias de las poblaciones** y la **poca "variabilidad" dentro de cada población**, es indicio que las medias son diferentes.

Las dos figuras siguientes muestra un ejemplo de tres muestras de tres poblaciones donde hay **mucha variabilidad** entre las **medias de las muestras**.

En cambio, hay **poca variabilidad** entre los valores dentro de cada muestra.

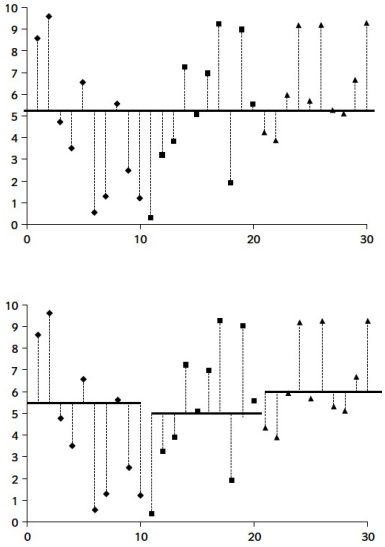
La media de la muestra de "rombos" está entorno al valor 9, la media de la muestra de "cuadrados", entorno al valor 5 y la media de la muestra de "triángulos", entorno al valor 1. Hay mucha diferencia entre las **medias de las tres muestras**.





En cambio en las dos figuras siguientes se muestra un ejemplo de tres muestras de tres poblaciones donde los valores tienen mucha **variabilidad** que no es explicada por la diferencia entre las **medias de las muestras**.

La media de la muestra de "rombos" está ahora entorno al valor 5.5, la media de la muestra de "cuadrados", entorno al valor 5 y la media de la muestra de "triángulos", entorno al valor 6. Vemos que en este caso, hay poca diferencia entre las **medias de las tres muestras.**



7.1. ANOVA de un factor

7.1.1. Clasificación simple, efectos fijos, diseño completamente aleatorio

Supongamos que estamos en los casos siguientes:

- usamos un solo factor para clasificar la población en subpoblaciones (**clasificación simple**),
- el investigador decide qué niveles (o tratamientos) del factor usará (**efectos fijos**),
- se toma una m.a.s. de cada subpoblación, de manera independiente unas de las otras (**completamente aleatorio**)

Ejemplo: Pseudomonas fragi

Se realizó un estudio para investigar el efecto del CO_2 sobre la tasa de crecimiento de *Pseudomonas fragi* (un corruptor de alimentos). Se cree que el crecimiento se ve afectado por la cantidad de CO_2 en el aire.

Para contrastarlo, en un experimento se administró CO_2 a 5 presiones atmosféricas diferentes a 10 cultivos diferentes por cada nivel, y se anotó el cambio (en %) de la masa celular al cabo de una hora:

Presión de CO_2 (en atmósferas)

0.0	0.083	0.29	0.50	0.86
62.6	50.9	45.5	29.5	24.9
59.6	44.3	41.1	22.8	17.2
64.5	47.5	29.8	19.2	7.8
59.3	49.5	38.3	20.6	10.5
58.6	48.5	40.2	29.2	17.8
64.6	50.4	38.5	24.1	22.1
50.9	35.2	30.2	22.6	22.6
56.2	49.9	27.0	32.7	16.8
52.3	42.6	40.0	24.4	15.9
62.8	41.6	33.9	29.6	8.8

7.1.2. Almacenamiento de datos en ANOVA

Los datos se suelen dar en forma de tabla donde las columnas suelen ser los niveles del factor.

Por tanto, en la columna i-ésima habrá la muestra correspondiente al factor i-ésimo.

Más concretamente, supongamos que los datos vienen con la estructura siguiente:

Niveles del factor

F_1	F_2	$\cdots$	F_k
X_{11}	X_{21}	$\cdots$	X_{k1}
X_{12}	X_{22}	$\cdots$	X_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}

donde

- n_i es el tamaño de la muestra del nivel i,
- X_{ij} es el valor de la característica bajo estudio correspondiente al individuo j del nivel i.

En el ejemplo anterior,

- $k = 5$,
- $F_1 = 0{,}0$, $F_2 = 0{,}083$, $F_3 = 0{,}29$, $F_4 = 0{,}50$ y $F_5 = 0{,}86$,
- $n_1 = n_2 = n_3 = n_4 = n_5 = 10$.

Para poder aplicar la **técnica ANOVA**, el paso previo es almacenar los datos en dos variables:

- variable característica, la llamaremos X,
- variable factor, la llamaremos F.

La variable X tendrá como valores los valores de la tabla anterior X_{ij} y la variable F valdrá i si el valor de la variable X corresponde al nivel i-ésimo del factor.

De esta manera **transformaremos** la tabla anterior en una tabla de $N = n_1 + \cdots + n_k$ filas y 2 columnas donde la primera columna serán los valores de la variable X y la segunda columna, los valores de la

variable F.

Ejemplo (continuación)

Realicemos la transformación anterior con los datos del ejemplo.

Llamaremos a la variable X `Incremento.celular` y a la variable F, `nivel.Presión`.

Los valores de las variables anteriores serán:

```
Incremento.celular=c(
  62.6,50.9,45.5,29.5,24.9,59.6,44.3,41.1,22.8,17.2,
  64.5,47.5,29.8,19.2,7.8,59.3, 49.5,38.3,20.6,10.5,
  58.6,48.5,40.2,29.2,17.8,64.6,50.4,38.5,24.1,22.1,
  50.9,35.2,30.2,22.6,22.6,56.2,49.9,27.0,32.7,16.8,
  52.3,42.6,40.0, 24.4,15.9,62.8,41.6,33.9,29.6,8.8)
nivel.Presión=rep(c("0.0","0.083","0.29","0.50","0.86"),times=10)
```

Los primeros elementos de nuestra **tabla transformada** son:

```
tabla.datos.ANOVA = data.frame(Incremento.celular,nivel.Presión)
head(tabla.datos.ANOVA)
```

```
##   Incremento.celular nivel.Presión
## 1               62.6           0.0
## 2               50.9         0.083
## 3               45.5          0.29
## 4               29.5          0.50
## 5               24.9          0.86
## 6               59.6           0.0
```

Fijémonos que hemos transformado la tabla original de 10 filas y 5 columnas en otra tabla de 50 filas y 2 columnas.

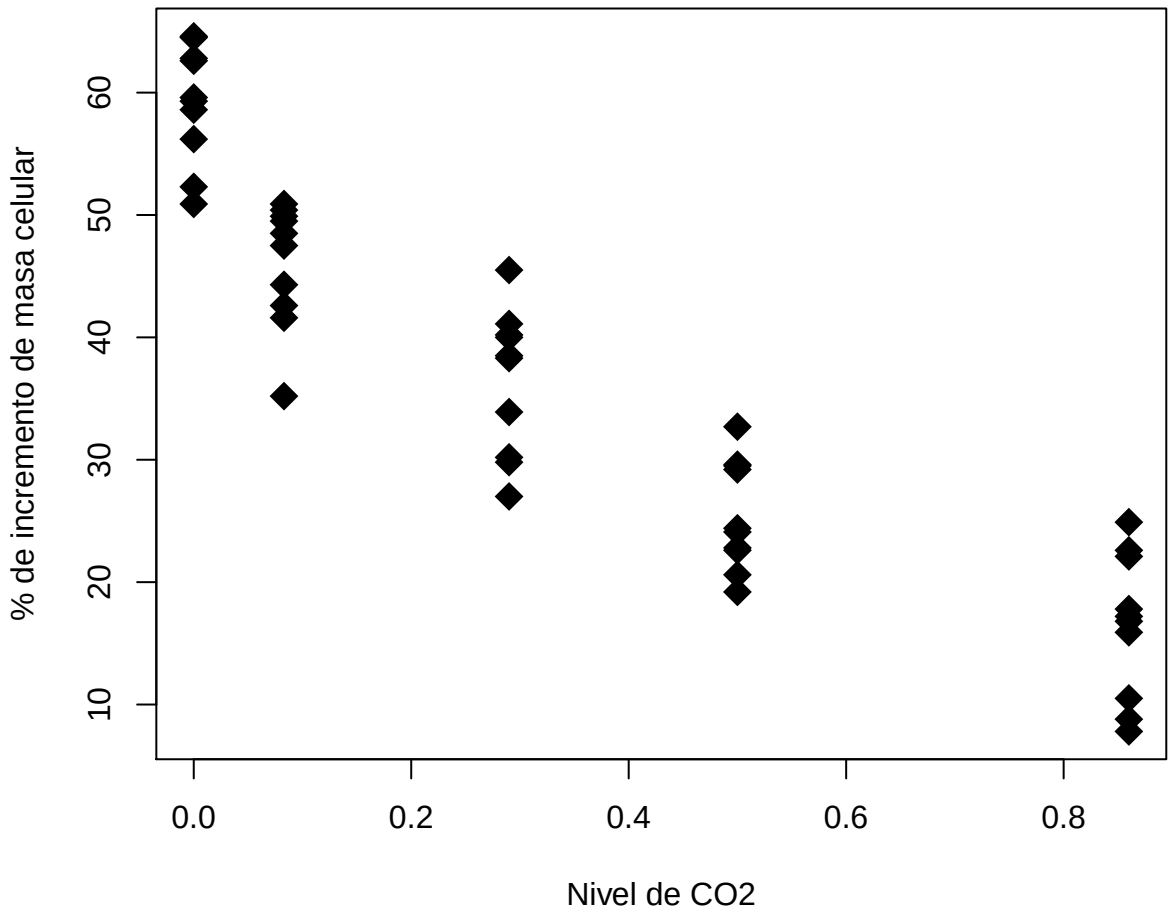
En general, si los datos vienen dados en una tabla donde las **columnas** representan los **niveles** de la variable **factor**, la tabla transformada tendrá $N = n_1 + \cdots + n_k$ filas y 2 columnas.

Ésta será la tabla sobre la que trabajaremos para realizar el **contraste ANOVA**.

Si realizamos un gráfico del porcentaje de aumento de masa celular según la variable presión del nivel de CO_2, obtenemos lo siguiente:

```
presión = as.numeric(as.character(tabla.datos.ANOVA$nivel.Presión))
plot(presión,tabla.datos.ANOVA$Incremento.celular,type="p",
     pch=18,cex=2,xlab="Nivel de CO2",ylab="% de masa celular")
```

Observamos gráficamente que las medias del porcentaje del aumento de masa celular de las muestras correspondientes a los 5 niveles de presión parecen diferentes.



En el gráfico siguiente, dibujamos los 50 porcentajes de aumentos de masa celular correspondientes a los 50 microorganismos separándolos en 5 grupos correspondientes a los niveles de presión del nivel de CO_2.

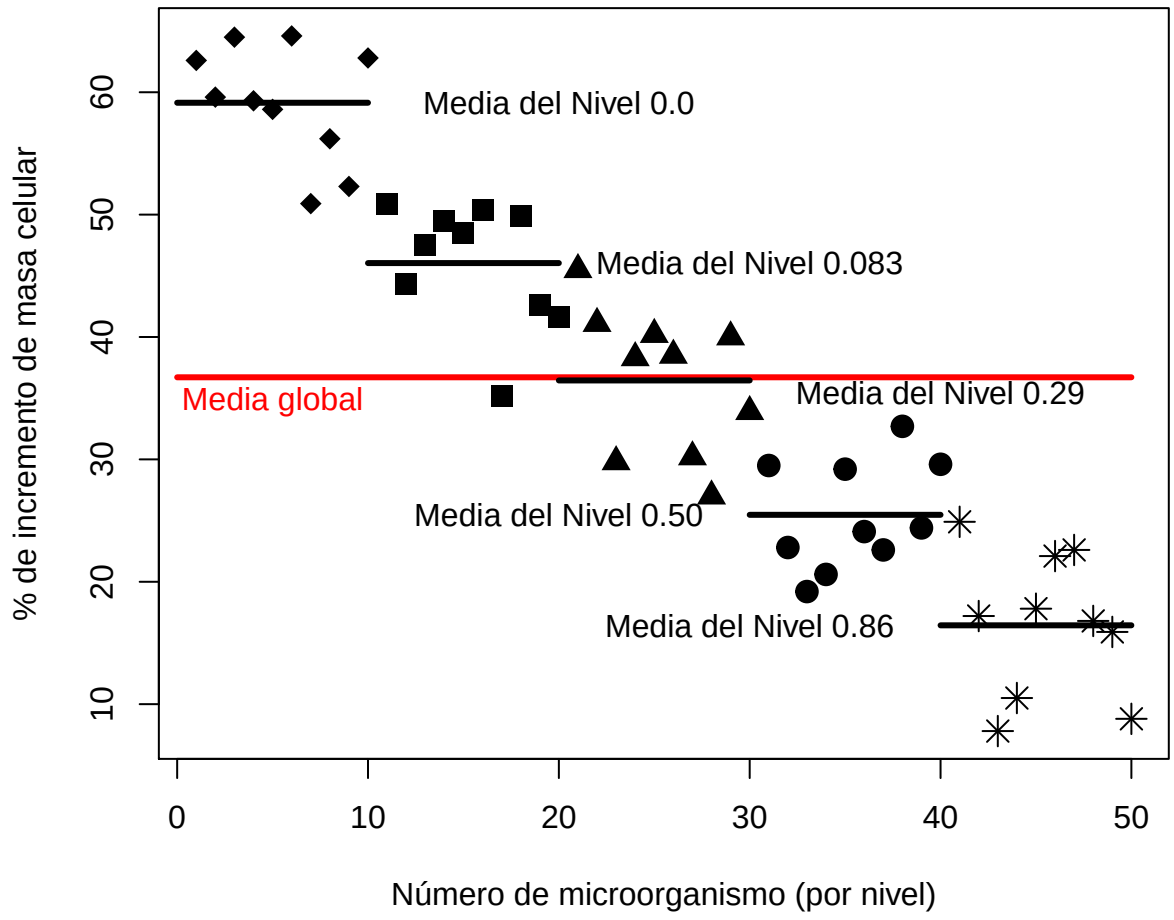
También aparecen las medias de cada grupo junto con la media global.

Este gráfico corrobora lo que hemos dicho anteriormente: parecen que hay diferencias entre las medias de los porcentajes de aumentos de masa celular entre los 5 grupos.

```
plot(1:50,tabla.datos.ANOVA$Incremento.celular,type="p",pch=17,
     xlab="Número de microorganismo (por nivel)",ylab="% de incremento de masa celular",cex=0)
points(1:10,tabla.datos.ANOVA$Incremento.celular[seq(from=1,to=46,by=5)],pch=18,cex=1.5)
points(11:20,tabla.datos.ANOVA$Incremento.celular[seq(from=2,to=47,by=5)],pch=15,cex=1.5)
points(21:30,tabla.datos.ANOVA$Incremento.celular[seq(from=3,to=48,by=5)],pch=17,cex=1.5)
points(31:40,tabla.datos.ANOVA$Incremento.celular[seq(from=4,to=49,by=5)],pch=19,cex=1.5)
points(41:50,tabla.datos.ANOVA$Incremento.celular[seq(from=5,to=50,by=5)],pch=8,cex=1.5)

lines(c(0,50),c(mean(tabla.datos.ANOVA$Incremento.celular),
                mean(mean(c(tabla.datos.ANOVA$Incremento.celular)))),lwd=3,col="red")
text(5,mean(mean(tabla.datos.ANOVA$Incremento.celular)-2),"Media global",col="red")
```

```
for (i in 1:5){lines(c(0+10*(i-1),10*i),
          c(mean(tabla.datos.ANOVA$Incremento.celular[seq(from=i,to=46+i,by=5)]),
          mean(tabla.datos.ANOVA$Incremento.celular[seq(from=i,to=46+i,by=5)])),lwd=3)}
text(20,mean(tabla.datos.ANOVA$Incremento.celular[seq(from=1,to=46,by=5)]),
     "Media del Nivel 0.0")
text(30,mean(tabla.datos.ANOVA$Incremento.celular[seq(from=2,to=47,by=5)]),
     "Media del Nivel 0.083")
text(40,mean(tabla.datos.ANOVA$Incremento.celular[seq(from=3,to=48,by=5)]-1),
     "Media del Nivel 0.29")
text(20,mean(tabla.datos.ANOVA$Incremento.celular[seq(from=4,to=49,by=5)]),
     "Media del Nivel 0.50")
text(30,mean(tabla.datos.ANOVA$Incremento.celular[seq(from=5,to=50,by=5)]),
     "Media del Nivel 0.86")
```



7.1.3. El contraste ANOVA

Recordemos que los datos vienen dados según la estructura siguiente:

Niveles del factor

F_1	F_2	$\cdots$	F_k
X_{11}	X_{21}	$\cdots$	X_{k1}
X_{12}	X_{22}	$\cdots$	X_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}

donde

- n_i es el tamaño de la muestra del nivel i,
- X_{ij} es el valor de la característica bajo estudio correspondiente al individuo j del nivel i.

7.1.4. Estadísticos

A partir de los datos de las muestras, definimos los **estadísticos** siguientes:

- Suma total de los datos del nivel i-ésimo: $T_{i\bullet} = \sum_{j=1}^{n_i} X_{ij}$.
- Media muestral para el nivel i-ésimo: $\overline{X}_{i\bullet} = \dfrac{T_{i\bullet}}{n_i}$.
- Suma total de los datos: $T_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^{k} T_{i\bullet}$.
- Media muestral de todos los datos: $\overline{X}_{\bullet\bullet} = \dfrac{T_{\bullet\bullet}}{N}$, donde $N = n_1 + \cdots + n_k$.

Calculemos los **estadísticos** anteriores para los datos de nuestro ejemplo.

- Para calcular la suma total de los datos del i-ésimo nivel, usaremos la función `aggregate` de `R`:

```
sumas.niveles = aggregate(Incremento.celular ~ nivel.Presión,
                          data = tabla.datos.ANOVA, FUN="sum")
sumas.niveles
```

```
##   nivel.Presión Incremento.celular
## 1           0.0              591.4
## 2         0.083              460.4
## 3          0.29              364.5
## 4          0.50              254.7
## 5          0.86              164.4
```

- Para calcular la media total de los datos del i-ésimo nivel, podemos usar otra vez la función `aggregate` de `R`:

```
medias.niveles = aggregate(Incremento.celular ~  nivel.Presión,
                           data=tabla.datos.ANOVA, FUN="mean")
medias.niveles
```

```
##   nivel.Presión Incremento.celular
## 1           0.0              59.14
## 2         0.083              46.04
## 3          0.29              36.45
## 4          0.50              25.47
## 5          0.86              16.44
```

- La suma total de los datos será:

```
suma.total = sum(tabla.datos.ANOVA$Incremento.celular)
suma.total
```

```
## [1] 1835
```

- La media muestral de los datos será:

```
media.muestral = mean(tabla.datos.ANOVA$Incremento.celular)
media.muestral
```

```
## [1] 36.71
```

7.1.5. El modelo

Los parámetros que intervendrán en el contraste son:

- μ: **media poblacional** del conjunto de la población (ignorando los niveles).
- μ_i: **media poblacional dentro del nivel i-ésimo**, $i = 1, \dots, k$.

Los estimadores de los parámetros son los siguientes:

- De: μ, $\overline{X}_{\bullet\bullet}$.
- De cada μ_i, $\overline{X}_{i\bullet}$.

Las suposiciones del modelo son:

- Las k muestras son m.a.s. **independientes** extraídas de k poblaciones específicas con medias $\mu_1, \dots, \mu_k$.
- Cada una de las k poblaciones sigue una **ley normal.**
- Todas estas poblaciones tienen la **misma varianza** σ^2 (**homocedasticidad**).

La expresión matemática del modelo a estudiar consiste en separar las diferencias de los datos respecto la **media global** en dos sumandos: las diferencias de los datos respecto las **medias de cada nivel** y las diferencias de las **medias de cada nivel** respecto la **media global**:

$$X_{ij} - \mu = (X_{ij} - \mu_i) + (\mu_i - \mu),\ i = 1, \dots, k,\ j = 1, \dots, n_i,$$

dónde

- X_{ij}: valor del j-ésimo individuo dentro del nivel i-ésimo,
- $X_{ij} - \mu$: **desviación** del individuo respecto de la **media global**,
- $X_{ij} - \mu_i$: **desviación** del individuo respecto de la **media de su grupo**,

- $\mu_i - \mu$: **desviación** de la **media del grupo** i**-ésimo** respecto de la **media global**.

7.1.6. Identidad de la suma de cuadrados

El teorema siguiente es la clave del contraste ANOVA.

Separa la **variabilidad de los datos** respecto la **media global** denominada **Suma Total de Cuadrados** en dos variabilidades:

- La **variabilidad de las medias de cada grupo** respecto la **media global**. A dicha **variabilidad** la llamaremos **Suma de Cuadrados de los Tratamientos**. Interesa que dicha **variabilidad** sea pequeña para que la hipótesis nula de **igualdad de medias** sea cierta.
- La **variabilidad de los datos** respecto **la media de cada grupo**. A dicha **variabilidad** la llamaremos **Suma de Cuadrados de los Residuos o Errores**. Interesa que dicha **variabilidad** sea grande para tener grupos lo más **heterogéneos** posibles de cara a aumentar la **potencia** del contraste ANOVA.

Teorema.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{\bullet\bullet})^2=\sum_{i=1}^{k}n_i(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\bullet})^2$$

- $SS_{Total}=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}(X_{ij}-\overline{X}_{\bullet\bullet})^2$ (Suma Total de Cuadrados)
- $SS_{Tr}=\sum\limits_{i=1}^{k}n_i(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet})^2$ (Suma de Cuadrados de los Tratamientos)
- $SS_E=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\bullet})^2$ (Suma de Cuadrados de los Residuos o Errores)

De cara a calcular las **variabilidades** o las **sumas** anteriores podemos usar las siguientes fórmulas equivalentes:

- $SS_{Total}=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}X_{ij}^2-\dfrac{T_{\bullet\bullet}^2}{N}=T_{\bullet\bullet}^{(2)}-\dfrac{T_{\bullet\bullet}^2}{N}$.
- $SS_{Tr}=\sum\limits_{i=1}^{k}\dfrac{T_{i\bullet}^2}{n_i}-\dfrac{T_{\bullet\bullet}^2}{N}$.
- $SS_E=SS_{Total}-SS_{Tr}$.

Ejercicio

Demostrar las fórmulas anteriores.

Usualmente escribiremos, para abreviar,

$$T_{\bullet\bullet}^{(2)}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2.$$

Calculemos las **variabilidades** o **sumas** anteriores para los datos de nuestro ejemplo de dos maneras distintas tal como queda indicado en las expresiones. Nos vamos a ayudar de las variables siguientes:

```
(ni=table(tabla.datos.ANOVA$nivel.Presión))
```

```
##
##   0.0 0.083  0.29  0.50  0.86
##    10    10    10    10    10
```

```
(N=sum(ni))
```

```
## [1] 50
```

- $SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\bullet\bullet}^2}{N}$:

```
(SSTotal1 =sum((tabla.datos.ANOVA$Incremento.celular-media.muestral)^2))
```

```
## [1] 12522
```

```
(SSTotal = sum(tabla.datos.ANOVA$Incremento.celular^2)-suma.total^2/N)
```

```
## [1] 12522
```

- $SS_{Tr} = \sum_{i=1}^{k} n_i(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{N}$:

```
(SSTr1=sum(ni*(medias.niveles[,2]-media.muestral)^2))
```

```
## [1] 11274
```

```
(SSTr=sum(sumas.niveles[,2]^2/ni)-(suma.total^2)/N)
```

```
## [1] 11274
```

- $SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\bullet})^2 = SS_{Total} - SS_{Tr}$:

```
(SSE1=sum((tabla.datos.ANOVA$Incremento.celular-medias.niveles[,2])^2))
```

```
## [1] 1248
```

```
(SSE=SSTotal-SSTr)
```

```
## [1] 1248
```

7.1.7. Estadísticos del contraste

Usaremos los **estadísticos de contraste** siguientes:

- Cuadrado medio de los tratamientos:

$$MS_{Tr} = \frac{SS_{Tr}}{k-1}.$$

- Cuadrado medio residual:

$$MS_E = \frac{SS_E}{N-k}.$$

Estos **estadísticos** son variables aleatorias, y se tiene que

- $E(MS_{Tr}) = \sigma^2 + \sum_{i=1}^{k} \frac{n_i(\mu_i - \mu)^2}{k-1}$.
- $E(MS_E) = \sigma^2$.

En particular, se puede usar MS_E para estimar la varianza común σ^2.

Si la hipótesis nula $H_0 : \mu_1 = \cdots = \mu_k = \mu$ es cierta, tenemos la siguiente condición:

$$\sum_{i=1}^{k} \frac{n_i(\mu_i - \mu)^2}{k-1} = 0,$$

y si H_0 no es cierta, esta cantidad es estrictamente positiva.

Por lo tanto

- si la hipótesis nula H_0 es cierta, $E(MS_E) = E(MS_{Tr})$ y consecuentemente tendríamos que esperar que estos dos estadísticos tuvieran valores parecidos, es decir
$$\frac{MS_{Tr}}{MS_E} \approx 1.$$
- si la hipótesis nula H_0 es falsa, $E(MS_E) < E(MS_{Tr})$ y consecuentemente tendríamos que esperar que
$$\frac{MS_{Tr}}{MS_E} > 1.$$

Basándonos en las consideraciones anteriores, consideraremos como **estadístico de contraste** el cociente

$$F = \frac{MS_{Tr}}{MS_E},$$

que, si la hipótesis nula H_0 es cierta, se distribuye según una $F_{k-1,N-k}$ (F de Fisher con $k-1$ y grados $N-k$ de libertad).

Su valor será cercano a 1. Por tanto, rechazaremos la hipótesis nula si F es "bastante más grande" que 1.

Sinteticemos cómo realizar el contraste ANOVA en 4 pasos:

- Primer paso: calculamos las **sumas de cuadrados**: $SS_{Total}, SS_{Tr}, SS_E$.
- Segundo paso: calculamos los **cuadrados medios**: $MS_{Tr} = \frac{SS_{Tr}}{k-1}$, $MS_E = \frac{SS_E}{N-k}$.
- Tercer paso: calculamos el **estadístico de contraste** F: $F = \frac{MS_{Tr}}{MS_E}$.
- Cuarto paso: calculamos el p-valor del contraste: $P(F_{k-1,N-k} \geq F)$. Si dicho p-valor es más pequeño que el nivel de significación α, rechazamos H_0 y concluimos que no todas las medias son iguales. En caso contrario, aceptamos H_0.

Realicemos el contraste ANOVA para los datos de nuestro ejemplo.

- Sumas de cuadrados:

```
SSTotal
```

```
## [1] 12522
```

```
SSTr
```

```
## [1] 11274
```

```
SSE
```

```
## [1] 1248
```

- Cuadrados medios:

```
k=5
(MStr = SSTr/(k-1))
```

```
## [1] 2819
```

```
(MSE = SSE/(N-k))
```

```
## [1] 27.73
```

- Estadístico de contraste:

```
(F = MStr/MSE)
```

```
## [1] 101.6
```

- p-valor:

```
(p = pf(F,k-1,N-k,lower.tail = FALSE))
```

```
## [1] 0.000000000000000000000006233
```

Comprobamos que el p-valor es muy pequeño. Concluimos, por tanto, que tenemos evidencias suficientes para concluir que las medias del aumento de masa celular no son iguales para los diferentes niveles de presión de CO_2.

Y, consecuentemente, el nivel de presión de CO_2 puede influir en el crecimiento del microorganismo *Pseudomonas fragi*.

Observación: sólo concluimos que no todas las medias son iguales, no que todas las medias son diferentes. No es lo mismo.

7.1.8. Tabla ANOVA

El contraste ANOVA se resume en la tabla siguiente:

Origen variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios	Estadístico de contraste	p-valor
Nivel	$k-1$	SS_{Tr}	$MS_{Tr} = \frac{SS_{Tr}}{k-1}$	$F = \frac{MS_{Tr}}{MS_E}$	p-valor
Residuo	$N-k$	SS_E	$MS_E = \frac{SS_E}{N-k}$		

La tabla anterior para los datos de nuestro ejemplo es la siguiente:

Origen variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios	Estadístico de contraste	p-valor
Nivel	4	11274.32	2818.58	101.63	0
Residuo	45	1248.04	27.73		

7.1.9. Contraste ANOVA con R

Para realizar un contraste ANOVA en R hay que usar la función aov.

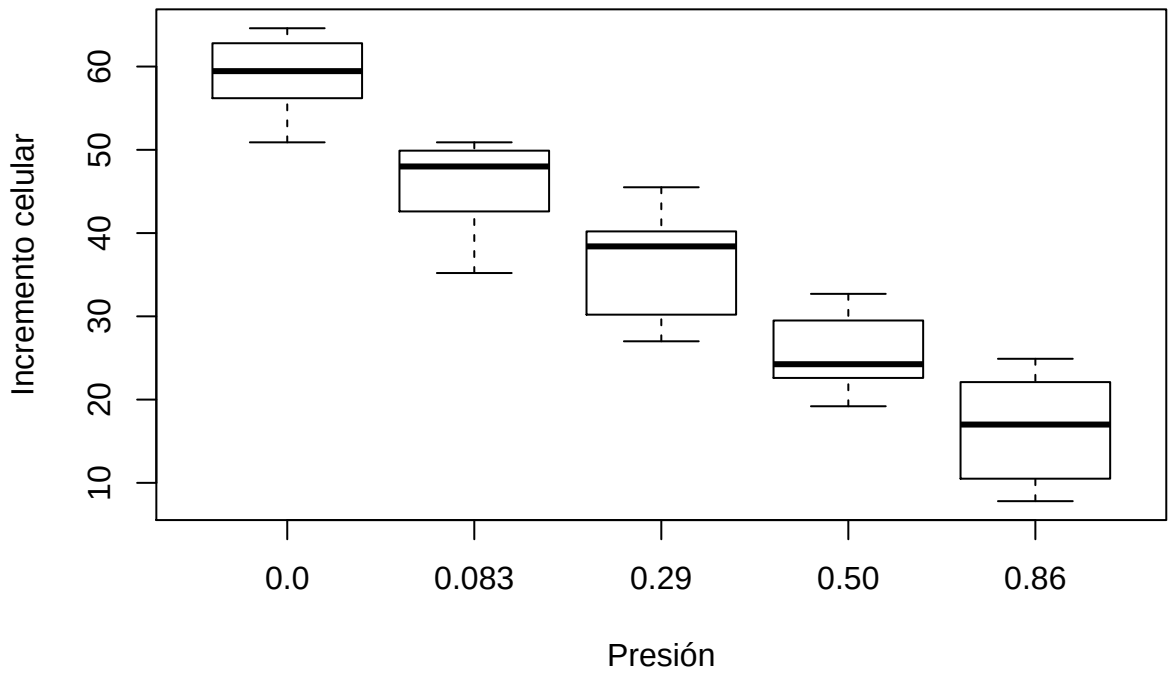
Dicha función se aplica a la tabla de datos modificada que hemos explicado:

```
summary(aov(X ~ F))
```

donde recordemos que en la variable X se almacenan los valores X_{ij} y en la variable factor F, los niveles del factor.

Lo primero que podríamos hacer es visualizar los datos del ejemplo con un boxplot:

```
boxplot(Incremento.celular ~ nivel.Presión, data = tabla.datos.ANOVA,
        xlab="Presión",ylab="Incremento celular")
```



Vemos que gráficamente se observan diferencias para los distintos niveles de presión.

El contraste ANOVA en R se realizaría de la forma siguiente:

```
X=tabla.datos.ANOVA$Incremento.celular
F=tabla.datos.ANOVA$nivel.Presión
summary(aov(X~F))
```

```
##             Df Sum Sq Mean Sq F value              Pr(>F)
## F            4  11274    2819     102 <0.0000000000000002 ***
## Residuals   45   1248      28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que tenemos la misma tabla ANOVA obtenida haciendo los cálculos "a mano".

7.2. Comparaciones por parejas

Si hemos rechazado la hipótesis nula $H_0 : \mu_1 = \cdots = \mu_k$, el siguiente paso es averiguar cuáles son los niveles diferentes.

Es decir, hallar aquellas parejas (μ_i, μ_j) para las que podamos decir que $\mu_i \neq \mu_j$.

Aunque hay diferentes formas de hacerlo vamos a ver las más usuales.

7.2.1. Test T de Bonferroni

Hay que tener en cuenta que hay que realizar en total $\binom{k}{2}$ contrastes del tipo:

$$\left.\begin{array}{ll} H_0 & : \mu_i = \mu_j, \\ H_1 & : \mu_i \neq \mu_j. \end{array}\right\}$$

El estadístico de cada contraste es el siguiente:

$$T = \frac{\overline{X}_{i\bullet} - \overline{X}_{j\bullet}}{\sqrt{MS_E \cdot (\frac{1}{n_i} + \frac{1}{n_j})}},$$

que, si la hipótesis nula H_0 es cierta, sigue una distribución t de Student con $N-k$ grados de libertad, t_{N-k}.

El p-valor de cada contraste es $2P(t_{N-k} \geq |t_{i,j}|)$, donde $t_{i,j}$ es el valor que toma el estadístico.

Observación: si se realizan c contrastes a un nivel de significación α, la probabilidad de Error de Tipo I en al menos uno de ellos es mayor que α. Si la calculamos, sería uno menos la probabilidad de no equivocarnos en ninguno de los contrastes, es decir $1-(1-\alpha)^c$.

En el ejemplo del aumento de la masa corporal del microorganismo, si realizamos $c = \binom{5}{2} = 10$ contrastes con nivel de significación $\alpha = 0{,}05$, ¡la probabilidad de Error de Tipo I en al menos uno de ellos es $1-(1-0{,}05)^{10} \approx 0{,}4$!

Por tanto, tendremos que reducir el nivel de significación de cada contraste para que la probabilidad final de Error de Tipo I sea α.

Usaremos la aproximación $1-(1-x)^c \approx cx$ y entonces, si queremos efectuar c contrastes con nivel de significación (global) α, los haremos con nivel de significación α/c.

En el ejemplo del aumento de la masa corporal del microorganismo, si realizamos los 10 contrastes, para obtener un nivel de significación global $\alpha = 0{,}05$, realizamos cada contraste con nivel de significación $\frac{0{,}05}{10} = 0{,}005$.

Realicemos el test de Bonferroni para los datos de nuestro ejemplo.

En primer lugar creamos una matriz cuyas filas son las parejas de niveles las medias de los cuales contrastaremos:

```
(pares=rbind(c(1,2),c(1,3),c(1,4),c(1,5),c(2,3),c(2,4),c(2,5),c(3,4),c(3,5),c(4,5)))
```

```
##       [,1] [,2]
##  [1,]    1    2
##  [2,]    1    3
##  [3,]    1    4
##  [4,]    1    5
##  [5,]    2    3
##  [6,]    2    4
##  [7,]    2    5
##  [8,]    3    4
##  [9,]    3    5
## [10,]    4    5
```

A continuación calculamos los valores de todos los estadísticos de contraste:

$$T = \frac{\overline{X}_{i\bullet} - \overline{X}_{j\bullet}}{\sqrt{MS_E \cdot (\frac{1}{n_i} + \frac{1}{n_j})}},$$

```
(est.contraste.pares = (medias.niveles[pares[,1],2]-medias.niveles[pares[,2],2])/
  (sqrt(MSE*(1/10+1/10))))
```

```
##  [1]  5.562  9.634 14.296 18.130  4.072  8.734 12.568  4.662  8.496  3.834
```

Los añadimos como columna a la matriz de parejas de niveles:

```
(pares=cbind(pares,est.contraste.pares))
```

```
##           est.contraste.pares
##  [1,] 1 2               5.562
##  [2,] 1 3               9.634
##  [3,] 1 4              14.296
##  [4,] 1 5              18.130
##  [5,] 2 3               4.072
##  [6,] 2 4               8.734
##  [7,] 2 5              12.568
##  [8,] 3 4               4.662
##  [9,] 3 5               8.496
## [10,] 4 5               3.834
```

Calculamos los p-valores:

```
calculo.p.valor=function(x){2*(1-pt(abs(x),N-k))}
(p.valores=sapply(est.contraste.pares,calculo.p.valor))
```

```
##  [1] 0.000001387522490681 0.000000000001649791 0.000000000000000000
##  [4] 0.000000000000000000 0.000186374413800872 0.000000000030210501
##  [7] 0.000000000000000222 0.000028080321327284 0.000000000066096462
## [10] 0.000389221804025119
```

Lo añadimos como columna a la matriz de parejas de niveles y estadísticos:

```
(pares=cbind(pares,p.valores))
```

```
##           est.contraste.pares            p.valores
##  [1,] 1 2               5.562 0.000001387522490681
##  [2,] 1 3               9.634 0.000000000001649791
##  [3,] 1 4              14.296 0.000000000000000000
##  [4,] 1 5              18.130 0.000000000000000000
##  [5,] 2 3               4.072 0.000186374413800872
##  [6,] 2 4               8.734 0.000000000030210501
##  [7,] 2 5              12.568 0.000000000000000222
##  [8,] 3 4               4.662 0.000028080321327284
##  [9,] 3 5               8.496 0.000000000066096462
## [10,] 4 5               3.834 0.000389221804025119
```

A continuación nos preguntamos qué p-valores son menores que 0.05/10 para saber entre qué pares podemos aceptar que hay diferencias:

```
pares[which(p.valores<0.005),]
```

```
##           est.contraste.pares            p.valores
##  [1,] 1 2               5.562 0.000001387522490681
##  [2,] 1 3               9.634 0.000000000001649791
##  [3,] 1 4              14.296 0.000000000000000000
##  [4,] 1 5              18.130 0.000000000000000000
##  [5,] 2 3               4.072 0.000186374413800872
##  [6,] 2 4               8.734 0.000000000030210501
##  [7,] 2 5              12.568 0.000000000000000222
##  [8,] 3 4               4.662 0.000028080321327284
##  [9,] 3 5               8.496 0.000000000066096462
## [10,] 4 5               3.834 0.000389221804025119
```

Vemos que en todos los pares se verifica que el p-valor es menor que 0.005.

Concluimos que hay evidencias suficientes para rechazar la igualdad de medias de aumento de masa celular del microorganismo entre dos niveles cualquiera de presión de CO_2. Por tanto, $\mu_i \neq \mu_j$ para todo $i \neq j$.

Para realizar la comparación por parejas con `R`, usamos la función `pairwise.t.test`. El parámetro `p.adjust.method` puede tomar el valor "none" (que no hace ajuste alguno):

```
pairwise.t.test(X,F,p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  X and F
##
##       0.0                   0.083              0.29
## 0.083 0.0000013875224908    -                  -
## 0.29  0.0000000000016497    0.0001863744138008 -
## 0.50  < 0.0000000000000002 0.0000000000302105 0.0000280803213273
## 0.86  < 0.0000000000000002 0.0000000000000003 0.0000000000660966
##       0.50
## 0.083 -
## 0.29  -
## 0.50  -
## 0.86  0.0003892218040251
##
## P value adjustment method: none
```

O bien el parámetro `p.adjust.method` puede tomar el valor "bonferroni" (que multiplicará el p-valor del contraste por el número de comparaciones llevadas a cabo, $\binom{k}{2}$):

```
pairwise.t.test(X,F,p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  X and F
##
##       0.0                   0.083              0.29               0.50
## 0.083 0.000013875224908     -                  -                  -
## 0.29  0.000000000016497     0.002              -                  -
## 0.50  < 0.0000000000000002 0.000000000302105 0.000280803213273  -
## 0.86  < 0.0000000000000002 0.000000000000003 0.000000000660966 0.004
##
## P value adjustment method: bonferroni
```

`R` nos muestra una tabla donde las filas y las columnas son los niveles del factor y cuyos valores son los p-valores de los contrastes por parejas.

Observamos que, en el primer caso todos los p-valores son menores que $\alpha/c = 0{,}05/10$ (sin hacer ningún ajuste en el método) o bien que en el segundo caso, todos los p-valores son menores que $\alpha = 0{,}05$, llegando en ambos casos a la misma conclusión que antes.

7.2.2. Test T de Holm (más potente)

El test de Holm es otro método que nos permite realizar las comparaciones entre las parejas de los distintos niveles del factor.

Dicho método es más usado ya que tiene más potencia que el método de Bonferroni.

Consta de los pasos siguientes:

- Sean $C_1, \dots, C_c$ los contrastes y los $P_1, \dots, P_c$ p-valores correspondientes
- Ordenamos estos p-valores en orden creciente $P_{(1)} \leq \cdots \leq P_{(c)}$ y reenumeramos consistentemente los contrastes $C_{(1)}, \dots, C_{(c)}$.
- Para cada $j = 1, \dots, c$, calculamos el p-valor ajustado $\widetilde{P}_{(j)} = (c + 1 - j)P_{(j)}$.
- Entonces rechazamos la hipótesis nula del contraste $C_{(j)}$ si $\widetilde{P}_{(j)} < \alpha$.

Vamos a aplicar el método de Holm al ejemplo que hemos estado desarrollando.

En primer lugar, ordenamos las filas de la tabla de datos `pares` ordenando los p-valores de menor a mayor:

```
(pares.ord=pares[order(pares[,4]),])
```

```
##           est.contraste.pares            p.valores
##  [1,] 1 4              14.296 0.000000000000000000
##  [2,] 1 5              18.130 0.000000000000000000
##  [3,] 2 5              12.568 0.000000000000000222
##  [4,] 1 3               9.634 0.000000000001649791
##  [5,] 2 4               8.734 0.000000000030210501
##  [6,] 3 5               8.496 0.000000000066096462
##  [7,] 1 2               5.562 0.000001387522490681
##  [8,] 3 4               4.662 0.000028080321327284
##  [9,] 2 3               4.072 0.000186374413800872
## [10,] 4 5               3.834 0.000389221804025119
```

Calculamos los p-valores ajustados y los añadimos como columna a `pares.ord`:

```
p.valores.ajust=pares.ord[,4]*(10+1-1:10)
pares.ord=cbind(pares.ord,p.valores.ajust)
round(pares.ord,12)
```

```
##           est.contraste.pares      p.valores p.valores.ajust
##  [1,] 1 4              14.296 0.000000000000  0.000000000000
##  [2,] 1 5              18.130 0.000000000000  0.000000000000
##  [3,] 2 5              12.568 0.000000000000  0.000000000000
##  [4,] 1 3               9.634 0.000000000002  0.000000000012
##  [5,] 2 4               8.734 0.000000000030  0.000000000181
##  [6,] 3 5               8.496 0.000000000066  0.000000000330
##  [7,] 1 2               5.562 0.000001387522  0.000005550090
##  [8,] 3 4               4.662 0.000028080321  0.000084240964
##  [9,] 2 3               4.072 0.000186374414  0.000372748828
## [10,] 4 5               3.834 0.000389221804  0.000389221804
```

A continuación, averiguamos qué contrastes verifican $\widetilde{P}_j \leq 0{,}05$:

```
round(pares.ord[which(pares.ord[,5]<=0.05),],6)
```

```
##          est.contraste.pares p.valores p.valores.ajust
##  [1,] 1 4              14.296  0.000000        0.000000
##  [2,] 1 5              18.130  0.000000        0.000000
##  [3,] 2 5              12.568  0.000000        0.000000
##  [4,] 1 3               9.634  0.000000        0.000000
##  [5,] 2 4               8.734  0.000000        0.000000
##  [6,] 3 5               8.496  0.000000        0.000000
##  [7,] 1 2               5.562  0.000001        0.000006
##  [8,] 3 4               4.662  0.000028        0.000084
##  [9,] 2 3               4.072  0.000186        0.000373
## [10,] 4 5               3.834  0.000389        0.000389
```

En este caso, también vemos que en todos los pares se verifica que el p-valor ajustado es menor que 0.05.

Concluimos lo mismo que en los contrastes realizados usando el método de Bonferroni: tenemos indicios suficientes para rechazar la igualdad de medias de aumento de masa celular del microorganismo para dos niveles cualquiera de presión de CO_2.

Para realizar la comparación de parejas con `R`, tenemos que usar la misma función `pairwise.t.test` que usábamos en los contrastes por parejas usando el método de Bonferroni pero cambiando el parámetro `p.adjust.method="holm"`:

```
pairwise.t.test(X,F,p.adjust.method = "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  X and F
##
##       0.0                   0.083             0.29
## 0.083 0.000005550089963     -                 -
## 0.29  0.000000000011548     0.000372748827602 -
## 0.50  < 0.0000000000000002 0.000000000181263 0.000084240963982
## 0.86  < 0.0000000000000002 0.000000000000002 0.000000000330483
##       0.50
## 0.083 -
## 0.29  -
## 0.50  -
## 0.86  0.000389221804025
##
## P value adjustment method: holm
```

`R` nos vuelve a mostrar una tabla donde las filas y las columnas son los niveles del factor y cuyos valores son los p-valores de los contrastes por parejas pero ajustados usando el método de Holm.

Observamos que todos los p-valores ajustados son menores que 0.05 llegando a la misma conclusión que antes.

7.2.3. Contraste de Duncan

El **contraste de Duncan** es otro método para ver en qué niveles hay diferencias.

Los pasos a realizar para llevarlo a cabo son los siguientes:

- Se ordenan en forma ascendente las k medias muestrales.
- Se considera cada par Y de medias muestrales y se calcula el valor absoluto D_Y de la diferencia entre las dos medias y el número p de medias que hay entre las dos (incluyendo las dos medias que comparamos).
- Decidimos que existe diferencia entre estos dos niveles cuándo $D_Y > SSR_p = r_p\sqrt{\frac{MS_E(n_i+n_j)}{2n_in_j}}$, dónde n_i y n_j son los tamaños de las subpoblaciones correspondientes a los dos niveles que comparamos y r_p es el **menor rango significativo** con $N-k$ grados de libertad, que encontraréis en la tabla del test de Duncan. Si vais a http://images.google.com y escribís "tabla del test de Duncan" en la casilla de búsqueda, encontraréis un montón de tablas del test de Duncan.

La primera columna de la tabla del test de Duncan corresponde a los grados de libertad del error $N-k$, y la primera fila, al valor p. Para calcular el valor r_p, tenéis que buscar el valor $N-k$ en la primera columna y el valor p en la primera fila y ver dónde se intersecan.

Vamos a aplicar el test de Duncan a los datos del ejemplo que hemos desarrollado.

Las medias de los cinco niveles eran:

```
medias.niveles
```

```
##   nivel.Presión Incremento.celular
## 1           0.0              59.14
## 2         0.083              46.04
## 3          0.29              36.45
## 4          0.50              25.47
## 5          0.86              16.44
```

Si los ordenamos de menor a mayor, obtenemos:

$$\overline{X}_{5\bullet} < \overline{X}_{4\bullet} < \overline{X}_{3\bullet} < \overline{X}_{2\bullet} < \overline{X}_{1\bullet}.$$

A continuación, calculamos las medias siguientes:

- $\overline{X}_{1\bullet} - \overline{X}_{5\bullet}$ $(p = 5)$
- $\overline{X}_{1\bullet} - \overline{X}_{4\bullet}$ $(p = 4)$
- $\overline{X}_{1\bullet} - \overline{X}_{3\bullet}$ $(p = 3)$
- $\overline{X}_{1\bullet} - \overline{X}_{2\bullet}$ $(p = 2)$
- $\overline{X}_{2\bullet} - \overline{X}_{5\bullet}$ $(p = 4)$
- $\overline{X}_{2\bullet} - \overline{X}_{4\bullet}$ $(p = 3)$
- $\overline{X}_{2\bullet} - \overline{X}_{3\bullet}$ $(p = 2)$
- $\overline{X}_{3\bullet} - \overline{X}_{5\bullet}$ $(p = 3)$
- $\overline{X}_{3\bullet} - \overline{X}_{4\bullet}$ $(p = 2)$
- $\overline{X}_{4\bullet} - \overline{X}_{5\bullet}$ $(p = 2)$

Tenemos que $n_i = 10$ para cada $i = 1, 2, 3, 4, 5$. El valor SSR_p será, por tanto:

$$SSR_p = r_p \sqrt{\frac{MS_E}{10}}$$

Calculamos los valores de r_p cada $p = 2, 3, 4, 5$ y con el nivel de significación $\alpha = 0{,}05$: (usamos 40 como grados de libertad del error para consultar en la tabla, puesto que es el valor más próximo a $N - k = 45$)

p	2	3	4	5
$r_p(con\ N-k=40)$	2.858	3	3.102	3.171
SSR_p	4.76	4.996	5.166	5.281

Resumimos todo el cálculo realizado:

Diferencias	d	p	SSR_p	$d > SSR_p$?	Conclusión
$\overline{X}_{1\bullet} - \overline{X}_{5\bullet}$	42,7	5	5,281	Sí	$\mu_1 \neq \mu_5$
$\overline{X}_{1\bullet} - \overline{X}_{4\bullet}$	33,67	4	5,166	Sí	$\mu_1 \neq \mu_4$
$\overline{X}_{1\bullet} - \overline{X}_{3\bullet}$	22,69	3	4,996	Sí	$\mu_1 \neq \mu_3$
$\overline{X}_{1\bullet} - \overline{X}_{2\bullet}$	13,1	2	4,760	Sí	$\mu_1 \neq \mu_2$
$\overline{X}_{2\bullet} - \overline{X}_{5\bullet}$	29,6	4	5,166	Sí	$\mu_2 \neq \mu_5$

Diferencias	d	p	SSR_p	$d > SSR_p$?	Conclusión
$\overline{X}_{2\bullet} - \overline{X}_{4\bullet}$	20,57	3	4,996	Sí	$\mu_2 \neq \mu_4$
$\overline{X}_{2\bullet} - \overline{X}_{3\bullet}$	9,59	2	4,760	Sí	$\mu_2 \neq \mu_3$
$\overline{X}_{3\bullet} - \overline{X}_{5\bullet}$	20,01	3	4,996	Sí	$\mu_3 \neq \mu_5$
$\overline{X}_{3\bullet} - \overline{X}_{4\bullet}$	10,98	2	4,760	Sí	$\mu_3 \neq \mu_4$
$\overline{X}_{4\bullet} - \overline{X}_{5\bullet}$	9,03	2	4,760	Sí	$\mu_4 \neq \mu_5$

Concluimos que, a un nivel de significación de $\alpha = 0{,}05$, todos los niveles tienen medias diferentes.

7.2.4. Contraste de Duncan en R

En R, el contraste de Duncan se realiza con la función `duncan.test` del paquete `agricolae`. La sintaxis es:

```
duncan.test(aov,"factor",group=...)$sufijo
```

donde

- `aov` es el resultado del ANOVA de partida,
- el `factor` es el factor del ANOVA,
- `group` puede ser `TRUE` o `FALSE` dependiendo de cómo queremos ver el resultado,

- el `sufix` es `group` si `group=TRUE` y `comparison` si `group=FALSE`.

Ejemplo

Vamos a aplicar el test de Duncan a los datos de nuestro ejemplo:

```
library(agricolae)
```

```
##
## Attaching package: 'agricolae'
```

```
## The following objects are masked from 'package:timeDate':
##
##     kurtosis, skewness
```

```
## The following objects are masked from 'package:EnvStats':
##
##     kurtosis, skewness
```

```
resultado.anova=aov(X~F)
duncan.test(resultado.anova,"F",group=FALSE)$comparison
```

```
##              difference pvalue signif.    LCL   UCL
## 0.0 - 0.083       13.10 0.0000     ***  8.356 17.84
## 0.0 - 0.29        22.69 0.0000     *** 17.702 27.68
## 0.0 - 0.50        33.67 0.0000     *** 28.521 38.82
## 0.0 - 0.86        42.70 0.0000     *** 37.435 47.97
## 0.083 - 0.29       9.59 0.0002     ***  4.846 14.33
## 0.083 - 0.50      20.57 0.0000     *** 15.582 25.56
## 0.083 - 0.86      29.60 0.0000     *** 24.451 34.75
## 0.29 - 0.50       10.98 0.0000     ***  6.236 15.72
## 0.29 - 0.86       20.01 0.0000     *** 15.022 25.00
## 0.50 - 0.86        9.03 0.0004     ***  4.286 13.77
```

Nos da una tabla donde las filas son las comparaciones entre los distintos pares del niveles del factor.

La tabla contiene 5 columnas:

- la primera nos da la diferencia entre las dos medias de los dos niveles que se comparan,
- la segunda nos da el p-valor que indica si hay o no diferencias entre las dos medias,
- la tercera indica si la diferencia es significativa o no. Cuantos más asteriscos (`*`) aparezcan, más significativa es la diferencia y
- las dos últimas representan un intervalo de confianza para la diferencia de medias.

Si en lugar de asignar el valor `FALSE` al parámetro `group`, le asignamos el valor `TRUE`, obtenemos lo siguiente:

```
library(agricolae)
resultado.anova=aov(X~F)
duncan.test(resultado.anova,"F",group=TRUE)$group
```

```
##           X groups
## 0.0   59.14      a
## 0.083 46.04      b
```

```
## 0.29  36.45      c
## 0.50  25.47      d
## 0.86  16.44      e
```

Si hubiese dos niveles donde las medias no fuesen significativamente diferentes aparecerían en la columna `groups` de la tabla anterior.

Por ejemplo, si las medias en los niveles `0.0` y `0.50` no fuesen significativamente diferentes, observaríamos el valor `ad` en la columna `groups`.

Como no vemos ningún caso en que aparezcan dos letras, podemos concluir que las medias de cualquier par de niveles son significativamente diferentes.

7.2.5. Contraste de Tukey

Si la tabla de datos es balanceada, es decir, todas las submuestras correspondientes a cada uno de los niveles del factor tienen el mismo tamaño, el método más preciso de comparación de medias es el llamado **método de Tukey**.

Es un test similar al **t-test** pero el **estadístico de contraste** tiene una distribución diferente. Podéis consultar el test en el link de la Wikipedia.

Par aplicar el **contraste de Tukey** en `R` hay que usar la función `TukeyHSD` (*Honestly Significative Difference*) y aplicarla al resultado de haber aplicado la función `aov`:

```
TukeyHSD(aov(X~F))
```

Ejemplo

Apliquemos el **test de Tukey** a los datos de nuestro ejemplo al ser la tabla de datos balanceada:

```
TukeyHSD(aov(X~F))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = X ~ F)
##
## $F
##              diff    lwr     upr  p adj
## 0.083-0.0  -13.10 -19.79  -6.408 0.0000
## 0.29-0.0   -22.69 -29.38 -15.998 0.0000
## 0.50-0.0   -33.67 -40.36 -26.978 0.0000
## 0.86-0.0   -42.70 -49.39 -36.008 0.0000
## 0.29-0.083  -9.59 -16.28  -2.898 0.0017
## 0.50-0.083 -20.57 -27.26 -13.878 0.0000
## 0.86-0.083 -29.60 -36.29 -22.908 0.0000
## 0.50-0.29  -10.98 -17.67  -4.288 0.0003
## 0.86-0.29  -20.01 -26.70 -13.318 0.0000
## 0.86-0.50   -9.03 -15.72  -2.338 0.0034
```

Observando los p-valores, concluimos que hay diferencias entre todas las submuestras correspondientes a los 5 niveles de la presión.

7.3. Efectos aleatorios

En el modelo de efectos fijos que es el que hemos visto hasta ahora, el experimentador elige los niveles a estudiar.

Cuando el número de niveles es muy grande, y se quiere averiguar si los niveles del factor tienen influencia en el valor medio del parámetro con el contraste:

$$\left.\begin{array}{l} H_0 : \text{Las medias de todos los niveles son iguales,} \\ H_1 : \text{No es cierto que todos los niveles tengan la misma media.} \end{array}\right\}$$

una posibilidad es elegir una m.a.s. de niveles, k, y aplicar la técnica ANOVA a estos niveles.

Este es el **modelo de efectos aleatorios**.

Las suposiciones del modelo son:

- Los k niveles elegidos forman una m.a.s. del conjunto de niveles.
- Las medias μ_i de todos los niveles siguen una distribución normal con valor medio μ (el valor medio de toda la población) y desviación típica σ_{Tr}.
- Todas las poblaciones, para todos los niveles, siguen leyes normales.
- Todas las poblaciones, para todos los niveles, tienen la misma varianza σ^2. (**homocedasticidad**)
- Las k muestras son m.a.s. independientes extraídas de las k poblaciones elegidas.

Una vez elegidos los k niveles, calculamos MS_{Tr} y cómo MS_E antes. Con las hipótesis anteriores, en este caso

- $E(MS_{Tr}) = \sigma^2 + \dfrac{N - \sum_{i=1}^{k} \frac{n_i^2}{N}}{k-1} \cdot \sigma_{Tr}^2$,
- $E(MS_E) = \sigma^2$.

Si la hipótesis nula H_0 es cierta, todas las medias de todos los niveles son iguales, es decir, $\sigma_{Tr}^2 = 0$, y por lo tanto $F = \frac{MS_{Tr}}{MS_E} \approx 1$.

Si la hipótesis nula H_0 es cierta, este estadístico F tiene distribución $F_{k-1,N-k}$.

Por lo tanto, el test ANOVA es el mismo que en el caso de efectos fijos, pero usando los niveles seleccionados.

7.3.1. Condiciones del ANOVA

Recordemos que para poder realizar el contraste ANOVA se debían de cumplir las condiciones siguientes:

- Las k muestras son m.a.s. **independientes** extraídas de poblaciones k específicas con medias $\mu_1, \cdots, \mu_k$.

- Cada una de las k poblaciones sigue una **ley normal.**
- Todas estas poblaciones tienen la **misma varianza σ^2.** (**homocedasticidad**)

En esta sección vamos a aprender cómo comprobar la normalidad y la igualdad de varianzas de la tabla de datos tratada.

7.3.2. Normalidad

Para verificar la normalidad de cada submuestra podemos usar todos los contrastes de normalidad aprendidos en el tema de **bondad de ajuste**:

- Test de Kolmogorov-Smirnov-Lilliefors.
- Test de normalidad de Anderson-Darling.
- Test de Shapiro-Wilks.
- Test omnibus de D'Agostino-Pearson.

Veamos si los datos de las submuestras para cada nivel de presión son normales usando el test de Kolmogorov-Smirnov-Lilliefors:

```
library(nortest)
lillie.test(X[F=="0.0"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[F == "0.0"]
## D = 0.16, p-value = 0.6
```

Ejemplo: condiciones ANOVA

```
lillie.test(X[F=="0.083"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[F == "0.083"]
## D = 0.21, p-value = 0.2
```

```
lillie.test(X[F=="0.29"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[F == "0.29"]
## D = 0.22, p-value = 0.2
```

```
lillie.test(X[F=="0.50"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[F == "0.50"]
```

```
## D = 0.2, p-value = 0.3
```

```
lillie.test(X[F=="0.86"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X[F == "0.86"]
## D = 0.16, p-value = 0.6
```

Vemos que según el test de Kolmogorov-Smirnov-Lilliefors, no podemos rechazar que las 5 submuestras correspondientes a cada nivel de presión sigan la distribución normal.

Ejercicio

Comprueba la normalidad de cada una de las poblaciones anteriores usando los otros tests vistos en el tema de Bondad de Ajuste.

7.3.3. Igualdad de varianzas u homocedasticidad

Para contrastar si las k submuestras tienen las mismas varianzas, se usa el **test de Bartlett**.

Veamos en qué consiste.

Sean $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in_i})$, $i = 1, \dots, k$, k muestras aleatorias simples de tamaño n_i, para $i = 1, \dots, k$ de k variables aleatorias normales X_i de varianza σ_i^2, para $i = 1, \dots, k$.

Nos planteamos el contraste de igualdad de varianzas:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2, \\ H_1 : \exists i, j \mid \sigma_i^2 \neq \sigma_j^2. \end{cases}$$

Para realizar el contraste anterior, usaremos el **estadístico de Bartlett**:

$$K^2 = \frac{(N-k)\ln(\tilde{s}_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(\tilde{s}_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{N-k}\right)},$$

donde:

$$N = \sum_{i=1}^{k} n_i, \ \tilde{s}_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)\tilde{s}_i^2}{N-k}, \ \tilde{s}_i^2 = \frac{\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2}{n_i - 1}.$$

El estadístico anterior sigue aproximadamente la distribución χ^2_{k-1} si la hipótesis nula es cierta.

Intuitivamente, si la hipótesis nula es cierta, o, si las varianzas son iguales ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$), tendremos que

$$\tilde{s}_1^2 \approx \tilde{s}_2^2 \approx \cdots \approx \tilde{s}_k^2 \approx \tilde{s}_p^2,$$

y el valor del estadístico K^2 será pequeño.

El p-valor del contraste se calcula como:

$$p = P(\chi^2_{k-1} > K^2).$$

Observación: Para realizar el **test de Bartlett**, primero hay que comprobar que las submuestras siguen la distribución normal. Es decir, no tiene sentido aplicar el **test de Bartlett** a muestras no normales.

Realizar el test de Bartlett "a mano" es bastante tedioso. Por dicho motivo, explicaremos cómo realizarlo en `R`.

En `R` hay que usar la función `bartlett.test`:

```
bartlett.test(X ~ F)
```

Ejemplo: Test de Bartlett

Apliquemos el test de Bartlett para los datos de nuestro ejemplo:

```
bartlett.test(X ~ F)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  X by F
## Bartlett's K-squared = 1.1, df = 4, p-value = 0.9
```

Como el p-valor es muy grande, concluimos que no tenemos evidencias para rechazar la igualdad de varianzas para las 5 submuestras consideradas.

7.4. Bloques completos aleatorios

El contraste ANOVA de **efectos fijos** o **efectos aleatorios** generaliza el contraste de **dos medias independientes** al contraste de k **medias independientes.**

Nos podemos preguntar si existe una generalización de contraste de **dos medias dependientes** a k **medias dependientes**.

En esta sección, veremos un contraste "ANOVA" que hace dicha generalización: el contraste de **bloques completos aleatorios**.

Más concretamente, supongamos que tenemos una tabla de datos como en el caso del contraste ANOVA de un factor.

O sea, queremos estudiar si las medias de una variable X segmentada en k muestras definidas por los niveles de otra variable factor F son iguales o no.

La diferencia fundamental con respecto al ANOVA de un factor es que sospechamos que hay otra **variable extraña** que nos puede distorsionar los resultados.

Por dicho motivo, creamos bloques a partir de dicha variable extraña para reducir su efecto. Veamos cómo.

Suponemos que tenemos k tratamientos que queremos comparar.

Escogemos como **bloques** conjuntos de individuos k relacionados (por ejemplo, k copias del mismo individuo).

Dentro de cada bloque, asignamos aleatoriamente a cada individuo un tratamiento.

Estos bloques vienen a ser los emparejamientos de los datos en los **contrastes de medias dependientes**.

En un contraste de **bloques completos aleatorios**,

- Se han emparejado los individuos en **bloques.** (**bloques**)
- Los tratamientos se asignan de manera aleatoria dentro de los **bloques**. (**aleatorios**)
- Cada tratamiento se usa exactamente una vez dentro de cada **bloque**. (**completos**)
- En cuanto a los tratamientos, es de **efectos fijos**. (la inferencia será válida sólo para los tratamientos usados)
- En cuanto a los bloques, puede ser de **efectos fijos** (se eligen todos los bloques adecuados) o **aleatorio**, en este último caso el modelo es **mixto**.

7.4.1. Tabla de datos

Los datos se presentan en una tabla de la forma:

Bloques/Tratamientos	Tratamiento 1	Tratamiento 2	...	Tratamiento k
1	X_{11}	X_{21}	...	X_{k1}
2	X_{12}	X_{22}	...	X_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
b	X_{1b}	X_{2b}	...	X_{kb}

Fijaos que ahora la fila j-ésima de la tabla de datos corresponde a los datos de la variable X para los individuos del bloque j-ésimo y la columna i-ésima, a los datos de la variable X para los individuos tratados con el tratamiento i-ésimo.

O sea, X_{ij} es el valor del tratamiento i-ésimo en el individuo correspondiente del bloque j-ésimo.

7.4.2. Contraste a realizar

El contraste que se quiere realizar es el siguiente:

$$\left.\begin{array}{l} H_0 : \mu_{1\bullet} = \mu_{2\bullet} = \cdots = \mu_{k\bullet}, \\ H_1 : \exists i, j \mid \mu_{i\bullet} \neq \mu_{j\bullet}. \end{array}\right\}$$

donde cada $\mu_{i\bullet}$ representa la media del tratamiento i-ésimo.

Ejemplo

Queremos determinar si la energía que se requiere para llevar a cabo tres actividades físicas (correr, pasear y montar en bicicleta) es la misma o no. Para cuantificar esta energía, medimos el número de Kcal. consumidas por Km. recorrido.

Las diferencias metabólicas entre los individuos pueden afectar la energía requerida para llevar a cabo una determinada actividad. Ésta sería la variable extraña que nos puede distorsionar los resultados.

Por lo tanto, no es aconsejable elegir tres grupos de individuos y a cada uno hacerle hacer una de las tres actividades físicas: las diferencias metabólicas entre los individuos elegidos podrían afectar los resultados y dar demasiada variación.

Lo que hacemos es seleccionar algunos individuos al azar (los **bloques aleatorios**), pedir a cada uno que corra, ande y recorra en bicicleta una distancia fijada, y determinar para cada individuo el número de Kcal. consumidas por Km. durante cada actividad.

Cada individuo es utilizado como un bloque. Las actividades se realizan en orden aleatorio, con tiempo de recuperación entre una y otra.

Al emparejar cada individuo con él mismo, eliminamos el efecto de la variación individual.

Por tanto, vamos a realizar un **Diseño de bloques completos aleatorios mixto**.

En la tabla siguiente se muestran los resultados obtenidos para 8 individuos:

Bloque/Tratamiento	1 (corriendo)	2 (andando)	3 (pedaleando)
1	1.4	1.1	0.7
2	1.5	1.2	0.8
3	1.8	1.3	0.7
4	1.7	1.3	0.8
5	1.6	0.7	0.1
6	1.5	1.2	0.7
7	1.7	1.1	0.4
8	2.0	1.3	0.6

El contraste que queremos realizar es el siguiente:

$$\left.\begin{array}{l} H_0 : \mu_{1\bullet} = \mu_{2\bullet} = \mu_{3\bullet}, \\ H_1 : \exists i, j \mid \mu_{i\bullet} \neq \mu_{j\bullet}, \end{array}\right\}$$

donde $\mu_{i\bullet}$, $i = 1, 2, 3$ representa la media de Kcal. consumidas por Km. mientras se corre, se pasea o se monta en bicicleta, respectivamente.

7.4.3. Modelo

La expresión matemática del modelo a estudiar consiste ahora en expresar los datos en cuatro sumandos : la **media global**, las diferencias de las **medias de cada nivel** respecto de la **media global**, las diferencias de las **medias de cada bloque** respecto la **media global** y los errores residuales:

$$X_{ij} = \mu + (\mu_{i\bullet} - \mu) + (\mu_{\bullet j} - \mu) + E_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, b,$$

dónde:

- X_{ij}: valor del tratamiento i-ésimo en el bloque j-ésimo.

- μ: media global.
- $\mu_{i\bullet}$: media del tratamiento i-ésimo.
- $\mu_{\bullet j}$: media del bloque j-ésimo.
- $\mu_{i\bullet} - \mu$: efecto del tratamiento i-ésimo. (**efecto tratamiento**)
- $\mu_{\bullet j} - \mu$: efecto de pertenecer al bloque j-ésimo. (**efecto bloque**)
- E_{ij}: error residual o aleatorio.

7.4.4. El modelo

Las suposiciones del modelo son:

- Las $k \cdot b$ observaciones constituyen muestras aleatorias independientes, cada una de tamaño 1, de $k \cdot b$ poblaciones.
- Estas $k \cdot b$ poblaciones son todas normales y con la misma varianza σ^2.
- El efecto de los bloques y los tratamientos es **aditivo**: no hay **interacción** entre los bloques y los tratamientos:
  - La diferencia de las medias poblacionales de cada pareja concreta de bloques es la misma para cada tratamiento.
  - La diferencia de las medias poblacionales de cada pareja concreta de tratamientos es la misma para cada bloque.

7.4.5. No interacción

Expliquemos con más detalle qué entendemos cuando no hay **interacción** entre **bloques** y **tratamientos**.

Consideremos dos ejemplos en que tenemos dos bloques: hombres y mujeres y tres tratamientos o niveles.

Las tablas siguientes nos dan las medias poblacionales de cada bloque/tratamiento dentro de cada grupo para los dos ejemplos.

En el primer caso, no tenemos interacción. En cambio, en el segundo caso, sí hay interacción.

- Caso de no interacción.

Bloque/Tratamiento	1 (corriendo)	2 (andando)	3 (pedaleando)
Hombres	$\mu_{11} = 4$	$\mu_{21} = 5$	$\mu_{31} = 7$
Mujeres	$\mu_{12} = 3$	$\mu_{22} = 4$	$\mu_{32} = 6$

- Caso de interacción.

Bloque/Tratamiento	1 (corriendo)	2 (andando)	3 (pedaleando)
Hombres	$\mu_{11} = 4$	$\mu_{21} = 5$	$\mu_{31} = 7$
Mujeres	$\mu_{12} = 3$	$\mu_{22} = 4$	$\mu_{32} = 2$

- En el primer caso no hay interacción ya que la diferencia de las medias poblacionales de cada pareja concreta de bloques es la misma para cada tratamiento:

$$\mu_{11} - \mu_{12} = \mu_{21} - \mu_{22} = \mu_{31} - \mu_{32} = 1.$$

  De la misma manera, la diferencia de las medias poblacionales de cada pareja concreta de tratamientos es la misma para cada bloque:

$$\begin{array}{l} \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = -1,\ \mu_{21} - \mu_{31} = \mu_{22} - \mu_{32} = -2, \\ \mu_{11} - \mu_{31} = \mu_{12} - \mu_{32} = -3. \end{array}$$

- En cambio, en el segundo caso sí hay interacción ya que la diferencia de las medias poblacionales de cada pareja concreta de bloques no es la misma para cada tratamiento:

$$\mu_{11} - \mu_{12} = \mu_{21} - \mu_{22} = 1 \neq \mu_{31} - \mu_{32} = 5.$$

  De la misma manera, la diferencia de las medias poblacionales de cada pareja concreta de tratamientos tampoco es la misma para cada bloque:

$$\begin{array}{l} \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = -1,\ \mu_{21} - \mu_{31} = -2 \neq \mu_{22} - \mu_{32} = 2, \\ \mu_{11} - \mu_{31} = -3 \neq \mu_{12} - \mu_{32} = 1. \end{array}$$

7.4.6. Estadísticos

Definimos los siguientes estadísticos de cara a realizar el estudio:

- $T_{i\bullet} = \sum_{j=1}^{b} X_{ij}$, suma total del tratamiento i-ésimo, $i = 1, 2, \dots, k$.
- $\overline{X}_{i\bullet} = \dfrac{T_{i\bullet}}{b}$, media muestral del tratamiento i-ésimo, $i = 1, 2, \dots, k$.
- $T_{\bullet j} = \sum_{i=1}^{k} X_{ij}$, suma total del bloque j-ésimo, $j = 1, 2, \dots, b$.
- $\overline{X}_{\bullet j} = \dfrac{T_{\bullet j}}{k}$, media muestral del bloque j-ésimo, $j = 1, 2, \dots, b$.
- $T_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{b} X_{ij} = \sum_{i=1}^{k} T_{i\bullet} = \sum_{j=1}^{b} T_{\bullet j}$, suma total.
- $\overline{X}_{\bullet\bullet} = \dfrac{T_{\bullet\bullet}}{k \cdot b}$, media muestral global.
- $T_{\bullet\bullet}^{(2)} = \sum_{i=1}^{k} \sum_{j=1}^{b} X_{ij}^2$, suma total de cuadrados.

Ejemplo: Cuantificación Energía (continuación) Vamos a calcular los estadísticos anteriores para nuestro ejemplo de cuantificación de la energía consumida al realizar distintos ejercicios físicos.

En primer lugar, tal como hicimos con el ejemplo de ANOVA de un factor, transformamos nuestra tabla de datos en una tabla de datos de 3 columnas: en la primera habrá el número de Kcal. consumidas, en la segunda el tipo de ejercicio físico (tratamiento) y en la tercera el individuo que ha realizado el ejercicio físico (bloque):

```
kilocal = c(1.4,1.1,0.7,1.5,1.2,0.8,1.8,1.3,
            0.7,1.7,1.3,0.8,1.6,0.7,0.1,1.5,
            1.2,0.7,1.7,1.1,0.4,2.0,1.3,0.6)
tratamiento = rep(1:3,8)
bloque = rep(1:8,each=3)
tabla.datos.ANOVA.BLOQUES = data.frame(kilocal,tratamiento,bloque)
```

```
head(tabla.datos.ANOVA.BLOQUES)
```

```
##   kilocal tratamiento bloque
## 1     1.4           1      1
## 2     1.1           2      1
## 3     0.7           3      1
## 4     1.5           1      2
## 5     1.2           2      2
## 6     0.8           3      2
```

Calculemos a continuación los estadísticos definidos:

- Suma total del tratamiento i-ésimo:

```
(sumas.tratamientos = aggregate(kilocal ~ tratamiento,
                                data = tabla.datos.ANOVA.BLOQUES, FUN="sum"))
```

```
##   tratamiento kilocal
## 1           1    13.2
## 2           2     9.2
## 3           3     4.8
```

- Media muestral del tratamiento i-ésimo:

```
(medias.tratamientos = aggregate(kilocal ~ tratamiento,
                                 data = tabla.datos.ANOVA.BLOQUES, FUN="mean"))
```

```
##   tratamiento kilocal
## 1           1    1.65
## 2           2    1.15
## 3           3    0.60
```

- Suma total del bloque j-ésimo:

```
(sumas.bloques = aggregate(kilocal ~ bloque,
                                 data = tabla.datos.ANOVA.BLOQUES, FUN="sum"))
```

```
##   bloque kilocal
## 1      1     3.2
```

```
## 2      2     3.5
## 3      3     3.8
## 4      4     3.8
## 5      5     2.4
## 6      6     3.4
## 7      7     3.2
## 8      8     3.9
```

- Media muestral del bloque j-ésimo:

```
(medias.bloques = aggregate(kilocal ~ bloque,
                            data = tabla.datos.ANOVA.BLOQUES, FUN="mean"))
```

```
##   bloque kilocal
## 1      1   1.067
## 2      2   1.167
## 3      3   1.267
## 4      4   1.267
## 5      5   0.800
## 6      6   1.133
## 7      7   1.067
## 8      8   1.300
```

- Suma total:

```
(suma.total = sum(tabla.datos.ANOVA.BLOQUES$kilocal))
```

```
## [1] 27.2
```

- Media muestral global:

```
(media.muestral = suma.total/nrow(tabla.datos.ANOVA.BLOQUES))
```

```
## [1] 1.133
```

- Suma total de cuadrados:

```
(suma.total.cuadrados = sum(tabla.datos.ANOVA.BLOQUES$kilocal^2))
```

```
## [1] 36.18
```

7.4.7. Identidad de la suma de cuadrados

En el diseño de **ANOVA por bloques**, la **variabilidad de los datos** respecto la **media global**, lo que llamamos **Suma Total de Cuadrados** se descompone de tres variabilidades:

- La **variabilidad** debida a los **tratamientos**.
- La **variabilidad** debida a los **bloques**.
- La **variabilidad** debida a **factores aleatorios**.

El siguiente teorema nos da dicha descomposición.

7.4.8. Identidad de la suma de cuadrados

Teorema.

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{\bullet\bullet})^2 &= b\sum_{i=1}^{k}(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet})^2 \\ &+k\sum_{j=1}^{b}(\overline{X}_{\bullet j}-\overline{X}_{\bullet\bullet})^2 \\ &+\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{i\bullet}-\overline{X}_{\bullet j}+\overline{X}_{\bullet\bullet})^2,\end{aligned}$$

donde:

- $SS_{Total}=\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{\bullet\bullet})^2$, variabilidad total.
- $SS_{Tr}=b\sum_{i=1}^{k}(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet})^2$, variabilidad debida a los tratamientos.
- $SS_{Bl}=k\sum_{j=1}^{b}(\overline{X}_{\bullet j}-\overline{X}_{\bullet\bullet})^2$, variabilidad debida a los bloques.

7.4.9. Identidad de la suma de cuadrados

- $SS_E=\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{i\bullet}-\overline{X}_{\bullet j}+\overline{X}_{\bullet\bullet})^2$, variabilidad debida a factores aleatorios.

Usando la notación definida, el teorema anterior puede escribirse de la forma siguiente:

$$SS_{Total}=SS_{Tr}+SS_{Bl}+SS_E.$$

Para calcular las variabilidades definidas anteriormente, hay que usar las expresiones siguientes:

- $SS_{Total}=T_{\bullet\bullet}^{(2)}-\dfrac{T_{\bullet\bullet}^2}{k\cdot b}$.
- $SS_{Tr}=\sum_{i=1}^{k}\dfrac{T_{i\bullet}^2}{b}-\dfrac{T_{\bullet\bullet}^2}{k\cdot b}$.
- $SS_{Bl}=\sum_{j=1}^{b}\dfrac{T_{\bullet j}^2}{k}-\dfrac{T_{\bullet\bullet}^2}{k\cdot b}$.
- $SS_E=SS_{Total}-SS_{Tr}-SS_{Bl}$.

Calculemos las variabilidades definidas para los datos del ejemplo que estamos desarrollando:

- Variabilidad total:

```
(SST = suma.total.cuadrados-suma.total^2/nrow(tabla.datos.ANOVA.BLOQUES))
```

```
## [1] 5.353
```

- Variabilidad debida a los tratamientos:

```
num.bloques = 8
num.tratamientos = 3
(SS.Tr = (1/num.bloques)*sum(sumas.tratamientos[,2]^2)-
    suma.total^2/nrow(tabla.datos.ANOVA.BLOQUES))
```

```
## [1] 4.413
```

- Variabilidad debida a los bloques:

```
(SS.Bl = (1/num.tratamientos)*sum(sumas.bloques[,2]^2)-
   suma.total^2/nrow(tabla.datos.ANOVA.BLOQUES))
```

```
## [1] 0.5533
```

- Variabilidad debida a factores aleatorios:

```
(SSE = SST-SS.Tr-SS.Bl)
```

```
## [1] 0.3867
```

7.4.10. Contraste

Recordemos que el contraste a realizar es el siguiente:

$$\left.\begin{array}{l} H_0 : \mu_{1\bullet} = \cdots = \mu_{k\bullet}, \\ H_1 : \exists i, j = 1, \ldots, k \mid \mu_{i\bullet} \neq \mu_{j\bullet}. \end{array}\right\}$$

Para realizar dicho contraste, usaremos los estadísticos siguientes:

- Cuadrado medio de los tratamientos: $MS_{Tr} = \dfrac{SS_{Tr}}{k-1}$.
- Cuadrado medio del error: $MS_E = \dfrac{SS_E}{(b-1)(k-1)}$.
- Cuadrado medio de los bloques: $MS_{Bl} = \dfrac{SS_{Bl}}{b-1}$.

7.4.11. Contraste

El valor medio o la esperanza de los estadísticos MS_{Tr} y MS_E es el siguiente:

$$\begin{aligned} E(MS_{Tr}) &= \sigma^2 + \frac{b}{k-1}\sum_{i=1}^{k}(\mu_{i\bullet} - \mu)^2, \\ E(MS_E) &= \sigma^2. \end{aligned}$$

Entonces, si $H_0 : \mu_{1\bullet} = \cdots = \mu_{k\bullet} = \mu$ es cierta, se verificará que la cantidad siguiente será nula:

$$\sum_{i=1}^{k}(\mu_{i\bullet}-\mu)^2=0,$$

y si H_0 no es cierta, dicha cantidad sería positiva.

7.4.12. Estadísticos del contraste

Basándonos en la apreciación anterior, se considera como **estadístico de contraste** el cociente siguiente:

$$F=\frac{MS_{Tr}}{MS_E},$$

que, si H_0 es cierta, se distribuye según una $F_{k-1,(k-1)(b-1)}$ (F de Fisher con $k-1$ y grados $(k-1)(b-1)$ de libertad).

Por tanto, su valor será próximo a 1.

Entonces, rechazaremos la hipótesis nula si F es bastante más grande que 1 basándonos en el p-valor:

$$p=P(F_{k-1,(k-1)(b-1)}\geq F),$$

con el significado usual: si el p-valor es más pequeño que el nivel de significación α, rechazamos H_0 y concluimos que no todas las medias son iguales. En caso contrario, aceptamos H_0.

Realicemos el contraste ANOVA por bloques en el ejemplo que vamos desarrollando.

Los cuadrados medios serán:

- Cuadrado medio de los tratamientos, $MS_{Tr}=\frac{SS_{Tr}}{k-1}$:

```
(MS.Tr = SS.Tr/(num.tratamientos-1))
```

```
## [1] 2.207
```

- Cuadrado medio del error, $MS_E=\frac{SS_E}{(b-1)(k-1)}$:

```
(MSE=SSE/((num.bloques-1)*(num.tratamientos-1)))
```

```
## [1] 0.02762
```

- Cuadrado medio de los bloques, $MS_{Bl}=\frac{SS_{Bl}}{b-1}$:

```
(MS.Bl = SS.Bl/(num.bloques-1))
```

```
## [1] 0.07905
```

- Estadístico de contraste, $F=\frac{MS_{Tr}}{MS_E}$:

```
(est.contraste = MS.Tr/MSE)
```

```
## [1] 79.9
```

- p-valor, $p = P(F_{k-1,(k-1)(b-1)} \geq F)$:

```
(p=pf(est.contraste,num.tratamientos-1,(num.tratamientos-1)*(num.bloques-1),
      lower.tail = FALSE))
```

```
## [1] 0.00000002201
```

Como el p-valor es muy pequeño, concluimos que tenemos indicios suficientes para rechazar que las medias de los tratamientos sean iguales. Es decir, la energía consumida al correr, pasear y montar en bicicleta no es la misma para los tres casos.

7.4.13. Tabla del contraste

El contraste realizado se resume en la tabla siguiente:

Origen variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios	Estadístico de contraste	p-valor
Tratamiento	$k-1$	SS_{Tr}	$MS_{Tr} = \frac{SS_{Tr}}{k-1}$	$F = \frac{MS_{Tr}}{MS_E}$	p-valor
Bloque	$b-1$	SS_{Bl}	$MS_{Bl} = \frac{SS_{Bl}}{b-1}$		
Error	$(k-1)\cdot(b-1)$	SS_E	$MS_E = \frac{SS_E}{(k-1)\cdot(b-1)}$		

En el ejemplo que estamos estudiando, la tabla ANOVA por bloques es la siguiente:

Origen variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios	Estadístico de contraste	p-valor
Tratamiento	2	4,413	2,207	79,897	0
Bloque	7	0,553	0,079		
Error	14	0,387	0,028		

7.5. ANOVA por bloques en R

Para realizar un contraste ANOVA por bloques en R, hay que usar la misma función aov que hemos usado cuando explicamos el ANOVA de un factor.

Recordemos que la función aov se aplica a la tabla de datos modificada que hemos explicado:

```
summary(aov(X ~ Tr + Bl))
```

donde recordemos que en la variable X se almacenan los valores de la variable a estudiar, la variable factor Tr nos dice el tratamiento que se aplica y la variable factor Bl, el bloque correspondiente.

Ejemplo: ANOVA por bloques

El contraste ANOVA por bloques en R para nuestro ejemplo se realizaría de la forma siguiente:

```
summary(aov(kilocal ~ tratamiento + bloque,
            data = tabla.datos.ANOVA.BLOQUES))
```

```
##             Df Sum Sq Mean Sq F value       Pr(>F)
## tratamiento  1   4.41    4.41   98.31 0.0000000022 ***
## bloque       1   0.00    0.00    0.03         0.87
## Residuals   21   0.94    0.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fijaos que la tabla obtenida no sería correcta ya que los grados de libertad de los tratamientos y los grados de libertad de los bloques (columna `Df`) son 1 y deberían ser 2 y 7, respectivamente.

Éste es uno de los errores más comunes que pueden ocurrir cuando realizamos un contraste ANOVA en R.

La razón es que R no ha considerado como factor ni la columna de los tratamientos ni la columna de los bloques en la tabla de datos `tabla.datos.ANOVA.BLOQUES`.

Para solucionar este problema, le decimos a R que dichas columnas son factores:

```
summary(aov(kilocal ~  as.factor(tratamiento) + as.factor(bloque),
            data = tabla.datos.ANOVA.BLOQUES))
```

```
##                         Df Sum Sq Mean Sq F value      Pr(>F)
## as.factor(tratamiento)  2   4.41   2.207   79.90 0.000000022 ***
## as.factor(bloque)       7   0.55   0.079    2.86       0.045 *
## Residuals              14   0.39   0.028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que hemos obtenido la misma tabla que la tabla obtenida haciendo los cálculos "a mano".

Notemos que no nos interesa ni el segundo estadístico de contraste ni el segundo p-valor, puesto que queremos evaluar si los tratamientos tienen o no la misma efectividad (pero R no sabe quien es el bloque ni quien es el tratamiento).

7.5.1. Efectividad en la construcción de los bloques

Recordemos que el objetivo del diseño por bloques es la reducción de algunas variables extrañas que podrían distorsionarnos los resultados si hubiésemos realizado el contraste ANOVA de un factor sin tener en cuenta los bloques.

Llegados a este punto, nos podemos preguntar si el diseño por bloques ha sido efectivo, es decir, si dicha construcción realmente ha reducido el efecto de las variables extrañas que no controlamos.

Expresado en términos de la variabilidad, la efectividad del diseño por bloques significa que la variabilidad debida a los bloques, SS_{Bl}, explica una parte importante de la variabilidad total, SST.

En este caso, el valor de SSE disminuiría aumentando el valor del estadístico de contraste F, lo que haría más "difícil" aceptar la hipótesis nula, mejorando la potencia del contraste.

La efectividad en la construcción de los bloques se estima con la **eficiencia relativa**, RE.

Se interpreta cómo la relación entre el número de observaciones de un experimento completamente aleatorio (CA) y el número de observaciones de un experimento de bloques completo aleatorio (BCA) necesaria para obtener tests equivalentes.

Por ejemplo, si $RE = 3$ significa que el diseño CA requiere tres veces tantas observaciones cómo el diseño de BCA. En este caso, ha merecido la pena el uso de bloques.

En cambio, un valor de $RE = 0{,}5$ significa que con un diseño CA hubiera bastado la mitad de observaciones que al diseño BCA. En este caso, no era aconsejable el uso de bloques.

Para la estimación de la **eficiencia relativa** RE, se usa el estadístico siguiente:

$$\widehat{RE} = c + (1-c)\frac{MS_{Bl}}{MS_E},$$

dónde $c = \dfrac{b(k-1)}{(bk-1)}$.

Por convenio, si $\widehat{RE} > 1{,}25$, se entiende que la construcción de los bloques ha sido efectiva.

Calculemos la **eficiencia relativa** en nuestro ejemplo:

```
c=num.bloques*(num.tratamientos-1)/(num.bloques*num.tratamientos-1)
(RE=c+(1-c)*MS.Bl/MSE)
```

```
## [1] 1.567
```

Hemos obtenido un valor mayor que 1.25. Por tanto, la construcción de bloques ha sido efectiva en nuestro caso.

7.6. ANOVA de dos vías

En la sección ANOVA de un factor, realizábamos un contraste de k medias ($k \geq 3$) para k poblaciones o subpoblaciones clasificadas según un factor.

Cada nivel del factor, daba lugar a una subpoblación.

En esta sección, supondremos que tendremos dos factores que nos clasifican los valores de la población en las correspondientes subpoblaciones.

Como tenemos dos factores, hay tres cuestiones sobre la mesa que habrá que resolver:

- ¿Existen diferencias entre las medias de las subpoblaciones debido al factor 1?
- ¿Existen diferencias entre las medias de las subpoblaciones debido al factor 2?
- ¿Existe interacción entre los dos factores?

Concretamente, para llevar a cabo el **ANOVA de dos vías**, consideraremos el caso más sencillo: **diseño completamente aleatorio** con **efectos fijos**:

- Usaremos dos factores (*dos vías*).
- Usaremos todos los niveles de cada factor (*efectos fijos*).

- Tomaremos muestras aleatorias independientes del mismo tamaño de cada combinación de niveles de los dos factores (**completamente aleatorio y balanceado**).

Llamaremos A y B a los factores a partir de los cuales vamos a segregar los datos de la variable cuyas medias de las subpoblaciones segregadas queremos comparar.

Supondremos que el factor A tiene a niveles y el factor B, b niveles.

Tomamos n observaciones para cada combinación de tratamientos. (**estudio balanceado**)

Por tanto, el número total de observaciones será $n \cdot a \cdot b$.

La variable aleatoria X_{ijk}, $i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, n$, nos da la respuesta de la k-ésima unidad experimental al nivel i-ésimo del factor A y el nivel j-ésimo del factor B.

La tabla de los datos tendrá la estructura siguiente:

Factor B/Factor A	1	2	$\cdots$	a
1	X_{111}	X_{211}	$\cdots$	X_{a11}
	X_{112}	X_{212}	$\cdots$	X_{a12}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	X_{11n}	X_{21n}	$\cdots$	X_{a1n}

Factor B/Factor A	1	2	$\cdots$	a
2	X_{121}	X_{221}	$\cdots$	X_{a21}
	X_{122}	X_{222}	$\cdots$	X_{a22}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	X_{12n}	X_{22n}	$\cdots$	X_{a2n}

Factor B/Factor A	1	2	$\cdots$	a
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
b	X_{1b1}	X_{2b1}	$\cdots$	X_{ab1}
	X_{1b2}	X_{2b2}	$\cdots$	X_{ab2}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	X_{1bn}	X_{2bn}	$\cdots$	X_{abn}

Ejemplo

En un experimento para determinar el efecto de la luz y la temperatura sobre el índice gonadosomático (GSI, una medida de crecimiento del ovario) de una especie de pez, se utilizaron dos fotoperiodos (14 horas de luz/10 horas de oscuridad, y 9 horas de luz/15 horas de oscuridad) y dos temperaturas (16° y 27 °C).

El experimento se realizó sobre 20 hembras. Se dividieron aleatoriamente en 4 subgrupos de 5 ejemplares cada uno. Cada grupo recibió una combinación diferente de luz y temperatura.

A los 3 meses se midieron los GSI de los peces y se obtuvieron los resultados siguientes:

7.6.1. Ejemplo

Factor B (temperatura)/Factor A	9 horas	14 horas
27°C	0,90	0,83
	1,06	0,67
	0,98	0,57
	1,29	0,47
	1,12	0,66
16°C	1,30	1,01
	2,88	1,52
	2,42	1,02
	2,66	1,32
	2,94	1,63

7.6.2. Almacenamiento de datos en ANOVA de dos vías

Vamos a almacenar los datos de una forma parecida a la usada en ANOVA de un factor.

Sea X la variable característica de la que comparamos las medias de las subpoblaciones. Sean A y B los factores.

Vamos a transformar la tabla anterior de los datos en una tabla de datos con $N = n \cdot a \cdot b$ filas y tres columnas.

La primera columna serán los valores de la variable X, la segunda los valores o niveles de la variable factor A y la tercera, los valores o niveles de la variable factor B.

La transformación de la tabla de datos para el ejemplo anterior se realizaría de la forma siguiente:

```
GSI = c(0.90,0.83,1.06,0.67,0.98,0.57,1.29,0.47,1.12,0.66,
        1.30,1.01,2.88,1.52,2.42,1.02,2.66,1.32,2.94,1.63)
temperatura = factor(rep(c(27,16),each=10))
fotoperiodos = factor(rep(c(9,14),times=10))
tabla.datos.GSI = data.frame(GSI,temperatura,fotoperiodos)
head(tabla.datos.GSI)
```

```
##    GSI temperatura fotoperiodos
## 1 0.90          27            9
## 2 0.83          27           14
## 3 1.06          27            9
## 4 0.67          27           14
## 5 0.98          27            9
## 6 0.57          27           14
```

7.6.3. El modelo

Para poder realizar un ANOVA de dos factores, supondremos que los datos verifican las suposiciones siguientes:

- Las observaciones para cada combinación de niveles constituyen **muestras aleatorias simples independientes**, cada una de tamaño n, de poblaciones $a \cdot b$,
- Cada una de las $a \cdot b$ poblaciones es normal.
- Todas las $a \cdot b$ poblaciones tienen la misma varianza, σ^2.

Los parámetros que intervendrán en el contraste son:

- μ: media poblacional global.
- $\mu_{i\bullet\bullet}$: media poblacional del nivel i-ésimo del factor A.
- $\mu_{\bullet j\bullet}$: media poblacional del nivel j-ésimo del factor B.
- $\mu_{ij\bullet}$: media poblacional de la combinación (i, j) de niveles A de y B.

En este caso la expresión matemática del modelo consiste en separar los valores de la variable X en 5 sumandos:

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}, \quad i = 1, \dots, a,\ j = 1, \dots, b,\ k = 1, \dots, n$$

donde

- μ: es la **media global**,
- $\alpha_i = \mu_{i\bullet\bullet} - \mu$: efecto al pertenecer al nivel i-ésimo del factor A,
- $\beta_j = \mu_{\bullet j\bullet} - \mu$: efecto al pertenecer al nivel j-ésimo del factor B,
- $(\alpha\beta)_{ij} = \mu_{ij\bullet} - \mu_{i\bullet\bullet} - \mu_{\bullet j\bullet} + \mu$: efecto de la **interacción** entre el nivel i-ésimo del factor A y el nivel j-ésimo del factor B,
- $E_{ijk} = X_{ijk} - \mu_{ij\bullet}$: error residual o aleatorio.

7.6.4. Sumas y medias

Definimos los estadísticos siguientes:

- Suma y media de los datos para la combinación de niveles i y j:

$$T_{ij\bullet} = \sum_{k=1}^{n} X_{ijk}, \quad \overline{X}_{ij\bullet} = \frac{T_{ij\bullet}}{n}.$$

- Suma y media de los datos para el nivel i-ésimo:

$$T_{i\bullet\bullet} = \sum_{j=1}^{b}\sum_{k=1}^{n} X_{ijk} = \sum_{j=1}^{b} T_{ij\bullet}, \quad \overline{X}_{i\bullet\bullet} = \frac{T_{i\bullet\bullet}}{bn}.$$

- Suma y media de los datos para el nivel j-ésimo:

$$T_{\bullet j\bullet} = \sum_{i=1}^{a}\sum_{k=1}^{n} X_{ijk} = \sum_{i=1}^{a} T_{ij\bullet}, \quad \overline{X}_{\bullet j\bullet} = \frac{T_{\bullet j\bullet}}{an}.$$

- Suma total de los datos:

$$T_{\bullet\bullet\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_{ijk} = \sum_{i=1}^{a} T_{i\bullet\bullet} = \sum_{j=1}^{b} T_{\bullet j\bullet}.$$

- Media muestral de todos los datos:

$$\overline{X}_{\bullet\bullet\bullet} = \frac{T_{\bullet\bullet\bullet}}{abn}.$$

- Suma de los cuadrados de los datos:

$$T^{(2)}_{\bullet\bullet\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X^2_{ijk}.$$

Los valores de los estadísticos anteriores para los datos de nuestro ejemplo son los siguientes:

- Suma de los datos para la combinación de niveles i y j:

```
(suma.combinación.niveles = aggregate(GSI ~ temperatura+fotoperiodos,
                                      data = tabla.datos.GSI, FUN="sum"))
```

```
##   temperatura fotoperiodos   GSI
## 1          16            9 12.20
## 2          27            9  5.35
## 3          16           14  6.50
## 4          27           14  3.20
```

Los valores de los estadísticos anteriores para los datos de nuestro ejemplo son los siguientes:

- Media de los datos para la combinación de niveles i y j:

```
(media.combinación.niveles = aggregate(GSI ~ temperatura+fotoperiodos,
                                       data=tabla.datos.GSI, FUN="mean"))
```

```
##   temperatura fotoperiodos  GSI
## 1          16            9 2.44
## 2          27            9 1.07
## 3          16           14 1.30
## 4          27           14 0.64
```

- Suma y media de los datos para el nivel i-ésimo:

```
(suma.fotoperiodos = aggregate(GSI ~ fotoperiodos, data=tabla.datos.GSI, FUN="sum"))
```

```
##   fotoperiodos   GSI
## 1            9 17.55
## 2           14  9.70
```

```
(media.fotoperiodos = aggregate(GSI ~ fotoperiodos, data = tabla.datos.GSI, FUN="mean"))
```

```
##   fotoperiodos   GSI
## 1            9 1.755
## 2           14 0.970
```

- Suma y media de los datos para el nivel j-ésimo:

```
(suma.temperatura = aggregate(GSI ~ temperatura, data = tabla.datos.GSI, FUN="sum"))
```

```
##   temperatura   GSI
## 1          16 18.70
## 2          27  8.55
```

```
(media.temperatura = aggregate(GSI ~ temperatura, data = tabla.datos.GSI, FUN="mean"))
```

```
##   temperatura   GSI
## 1          16 1.870
## 2          27 0.855
```

- Suma total de los datos:

```
(suma.total = sum(tabla.datos.GSI$GSI))
```

```
## [1] 27.25
```

- Media muestral de todos los datos:

```
(media.muestral = mean(tabla.datos.GSI$GSI))
```

```
## [1] 1.363
```

- Suma de los cuadrados de los datos:

```
(suma.cuadrados = sum(tabla.datos.GSI$GSI^2))
```

```
## [1] 48.26
```

7.6.5. Identidades de sumas de cuadrados

En el caso de ANOVA de dos factores, la **variabilidad total de los datos** respecto la **media global**, (**Suma Total de Cuadrados**) se separa en 4 variabilidades:

- la **variabilidad de las medias de cada grupo del factor** A respecto la **media global**,
- la **variabilidad de las medias de cada grupo del factor** B respecto la **media global**,
- la **variabilidad de las medias de cada combinación de grupos de los factores** A **y** B respecto la **media global**,
- la **variabilidad debida a factores aleatorios**.

Para estimar dichas variabilidades, se introducen las sumas de cuadrados siguientes:

- Estimación de la **variabilidad total de los datos** respecto la **media global**: $SS_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \overline{X}_{\bullet\bullet\bullet})^2$.
- Estimación de la **variabilidad de las medias de cada grupo del factor** A respecto la **media global**: $SS_A = bn\sum_{i=1}^{a}(\overline{X}_{i\bullet\bullet} - \overline{X}_{\bullet\bullet\bullet})^2$.

- Estimación de la **variabilidad de las medias de cada grupo del factor** B respecto la **media global**: $SS_B = an\sum_{j=1}^{b}(\overline{X}_{\bullet j\bullet} - \overline{X}_{\bullet\bullet\bullet})^2$.
- Estimación de la **variabilidad de las medias de cada combinación de grupos de los factores** A **y** B respecto la **media global** o **variabilidad debida a la interacción de los factores** A **y** B: $SS_{AB} = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{ij\bullet} - \overline{X}_{i\bullet\bullet} - \overline{X}_{\bullet j\bullet} + \overline{X}_{\bullet\bullet\bullet})^2$.
- Estimación de la **variabilidad que tendríamos si consideramos la combinación de factores** A **y** B **como si fuese un sólo factor**: $SS_{Tr} = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{ij\bullet} - \overline{X}_{\bullet\bullet\bullet})^2$.
- Estimación de la la **variabilidad debida a factores aleatorios**: $SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \overline{X}_{ij\bullet})^2$.

El teorema siguiente nos da la descomposición que comentamos antes:

Teorema. La **variabilidad total de los datos** se descompone de la forma siguiente en las variabilidades definidas anteriormente:

$$SS_{Total} = SS_{Tr} + SS_E, \quad \text{con } SS_{Tr} = SS_A + SS_B + SS_{AB}.$$

7.6.6. Cálculo de las sumas de cuadrados

Para calcular las **variabilidades**, se usan las fórmulas equivalentes siguientes:

- $SS_{Total} = T^{(2)}_{\bullet\bullet\bullet} - \dfrac{T^2_{\bullet\bullet\bullet}}{abn}$.
- $SS_A = \dfrac{1}{bn}\sum_{i=1}^{a} T^2_{i\bullet\bullet} - \dfrac{T^2_{\bullet\bullet\bullet}}{abn}$.
- $SS_B = \dfrac{1}{an}\sum_{j=1}^{b} T^2_{\bullet j\bullet} - \dfrac{T^2_{\bullet\bullet\bullet}}{abn}$.
- $SS_{Tr} = \dfrac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} T^2_{ij\bullet} - \dfrac{T^2_{\bullet\bullet\bullet}}{abn}$.
- $SS_{AB} = SS_{Tr} - SS_A - SS_B$.
- $SS_E = SS_{Total} - SS_{Tr}$.

Calculemos las variabilidades para los datos de nuestro ejemplo:

- **Variabilidad total de los datos** respecto la **media global**:

```
a=2; b=2; n=5;
(SST = suma.cuadrados - suma.total^2/(a*b*n))
```

```
## [1] 11.13
```

- **Variabilidad de las medias de cada grupo del factor** A respecto la **media global**:

```
(SSA = (1/(b*n))*sum(suma.fotoperiodos[,2]^2)-suma.total^2/(a*b*n))
```

```
## [1] 3.081
```

- **Variabilidad de las medias de cada grupo del factor** B respecto la **media global**:

```
(SSB = (1/(a*n))*sum(suma.temperatura[,2]^2)-suma.total^2/(a*b*n))
```

```
## [1] 5.151
```

- **Variabilidad que tendríamos si consideramos la combinación de factores** A **y** B **como si fuese un sólo factor**:

```
(SSTr = (1/n)*sum(suma.combinación.niveles[,3]^2)-suma.total^2/(a*b*n))
```

```
## [1] 8.862
```

- **Variabilidad debida a la interacción de los factores** A **y** B:

```
(SSAB = SSTr-SSA-SSB)
```

```
## [1] 0.6301
```

- **Variabilidad debida a factores aleatorios**:

```
(SSE=SST-SSTr)
```

```
## [1] 2.271
```

7.6.7. Cuadrados medios

Para realizar el ANOVA de dos factores, usaremos los siguientes **cuadrados medios**:

- **Cuadrado medio del factor** A: $MS_A = \dfrac{SS_A}{a-1}$.
- **Cuadrado medio del factor** B: $MS_B = \dfrac{SS_B}{b-1}$.
- **Cuadrado medio de la interacción entre los factores** A **y** B: $MS_{AB} = \dfrac{SS_{AB}}{(a-1)(b-1)}$.

7.6.8. Cuadrados medios

- **Cuadrado medio de los tratamientos**: $MS_{Tr} = \dfrac{SS_{Tr}}{ab-1}$.
- **Cuadrado medio residual**: $MS_E = \dfrac{SS_E}{ab(n-1)}$.

Calculemos los cuadrados medios para los datos del ejemplo que vamos desarrollando:

- **Cuadrado medio del factor** A (fotoperíodo):

```
(MSA = SSA/(a-1))
```

```
## [1] 3.081
```

- **Cuadrado medio del factor** B (temperatura):

```
(MSB = SSB/(b-1))
```

```
## [1] 5.151
```

- **Cuadrado medio de la interacción entre los factores** A **y** B:

```
(MSAB = SSAB/((a-1)*(b-1)))
```

```
## [1] 0.6301
```

- **Cuadrado medio de los tratamientos**:

```
(MSTr = (SSTr/(a*b-1)))
```

```
## [1] 2.954
```

- **Cuadrado medio residual**:

```
(MSE = (SSE/(a*b*(n-1))))
```

```
## [1] 0.142
```

7.6.9. Contrastes a realizar

En una ANOVA de dos vías, nos pueden interesar los cuatro contrastes siguientes:

- **Contraste de medias del factor** A: contrastamos si hay diferencias entre los niveles del factor A:

$$\begin{cases} H_0 : \mu_{1\bullet\bullet} = \mu_{2\bullet\bullet} = \cdots = \mu_{a\bullet\bullet}, \\ H_1 : \exists i, i' \mid \mu_{i\bullet\bullet} \neq \mu_{i'\bullet\bullet}. \end{cases}$$

El **estadístico de contraste** es $F = \frac{MS_A}{MS_E}$, que, si la hipótesis nula H_0 es cierta, tiene una distribución F de Fisher con $a-1$ y grados $ab(n-1)$ de libertad y valor próximo a 1.

- **Contraste de medias del factor** B: contrastamos si hay diferencias entre los niveles del factor B:

$$\begin{cases} H_0 : \mu_{\bullet 1\bullet} = \mu_{\bullet 2\bullet} = \cdots = \mu_{\bullet b\bullet}, \\ H_1 : \exists j, j' \mid \mu_{\bullet j\bullet} \neq \mu_{\bullet j'\bullet}. \end{cases}$$

El **estadístico de contraste** es $F = \frac{MS_B}{MS_E}$, que, si la hipótesis nula H_0 es cierta, tiene una distribución F de Fisher con $b-1$ y grados $ab(n-1)$ de libertad y valor próximo a 1.

- **Contraste de los tratamientos**: contrastamos si hay diferencias entre las parejas (nivel i de A, nivel j de B):

$$\begin{cases} H_0 : \forall i, j, i', j' \mid \mu_{ij\bullet} = \mu_{i'j'\bullet}, \\ H_1 : \exists i, j, i', j' \mid \mu_{ij\bullet} \neq \mu_{i'j'\bullet}. \end{cases}$$

El **estadístico de contraste** es $F = \frac{MS_{Tr}}{MS_E}$, que, si la hipótesis nula H_0 es cierta, tiene una distribución F de Fisher con $ab-1$ y grados $ab(n-1)$ de libertad y valor próximo a 1.

- **Contraste de no interacción**: contrastamos si hay interacción entre los factores A y B

$$\begin{cases} H_0 : \forall i, j \mid (\alpha\beta)_{ij} = 0, \\ H_1 : \exists i, j \mid (\alpha\beta)_{ij}. \neq 0 \end{cases}$$

El **estadístico de contraste** es $F = \frac{MS_{AB}}{MS_E}$, que, si la hipótesis nula H_0 es cierta, tiene distribución F de Fisher con $(a-1)(b-1)$ y grados $ab(n-1)$ de libertad y valor próximo a 1.

En los cuatro casos, el p-valor es

$$P(F_{x,y} \geq \text{valor del estadístico}),$$

donde $F_{x,y}$ representa la distribución F de Fisher con los grados de libertad que correspondan:

- **Contraste de medias del factor** A: $x = a-1$, $y = ab(n-1)$.
- **Contraste de medias del factor** B: $x = b-1$, $y = ab(n-1)$.
- **Contraste de los tratamientos**: $x = ab-1$, $y = ab(n-1)$.
- **Contraste de no interacción**: $x = (a-1)(b-1)$, $y = ab(n-1)$.

Los contrastes anteriores se resumen en la tabla siguiente:

7.6.10. Tabla ANOVA

Variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Estadístico F	p-valores
Tratamientos	$ab-1$	SS_{Tr}	MS_{Tr}	$\frac{MS_{Tr}}{MS_E}$	p-valor
A	$a-1$	SS_A	MS_A	$\frac{MS_A}{MS_E}$	p-valor
B	$b-1$	SS_B	MS_B	$\frac{MS_B}{MS_E}$	p-valor
AB	$(a-1)(b-1)$	SS_{AB}	MS_{AB}	$\frac{MS_{AB}}{MS_E}$	p-valor
Error	$ab(n-1)$	SS_E	MS_E		

La tabla ANOVA para los datos de nuestro ejemplo es la siguiente:

Variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Estadístico F	p-valores
Tratamientos	3	8.862	2.954	$\frac{2{,}954}{0{,}142} = 20{,}809$	$P(F_{3,16} > 20{,}809) = 0$
A	1	3.081	3.081	$\frac{3{,}081}{0{,}142} = 21{,}704$	$P(F_{1,16} > 21{,}704) = 0$
B	1	5.151	5.151	$\frac{5{,}151}{0{,}142} = 36{,}285$	$P(F_{1,16} > 36{,}285) = 0$

Variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Estadístico F	p-valores
AB	1	0.63	0.63	$\frac{0,63}{0,142} = 4,439$	$P(F_{1,16} > 4,439) = 0,051$
Error	16	2.271	0.142		

Las conclusiones que se obtienen de los 4 contrastes a partir de la tabla anterior son las siguientes:

- **Contraste de medias del factor** A: como el p-valor es prácticamente 0, concluimos que tenemos indicios suficientes para rechazar que no hay diferencias entre los niveles del factor A. Es decir, el índice GSI se ve afectado por el fotoperíodo.
- **Contraste de medias del factor** B: como el p-valor es prácticamente 0, concluimos también que tenemos indicios suficientes para rechazar que no hay diferencias entre los niveles del factor B. Es decir, el índice GSI se ve afectado por la temperatura.
- **Contraste de los tratamientos**: como el p-valor es prácticamente 0, concluimos también que tenemos indicios suficientes para rechazar que no hay diferencias entre los niveles de la combinación de factores A y B. Es decir, el índice GSI se ve afectado por los cuatro niveles de la combinación fotoperíodo/temperatura.
- **Contraste de no interacción**: el p-valor está entre 0.05 y 0.1. Por tanto, al estar en la zona de penumbra, no podemos concluir si hay o no interacción entre los factores A y B. Dicho de otra manera, no queda claro si hay interacción entre el fotoperíodo y la temperatura.

7.6.11. Contraste ANOVA de dos vías en R

Para realizar un contraste ANOVA de dos vías en R hay que usar la función `aov` que usábamos en los casos de ANOVA de un factor y ANOVA por bloques.

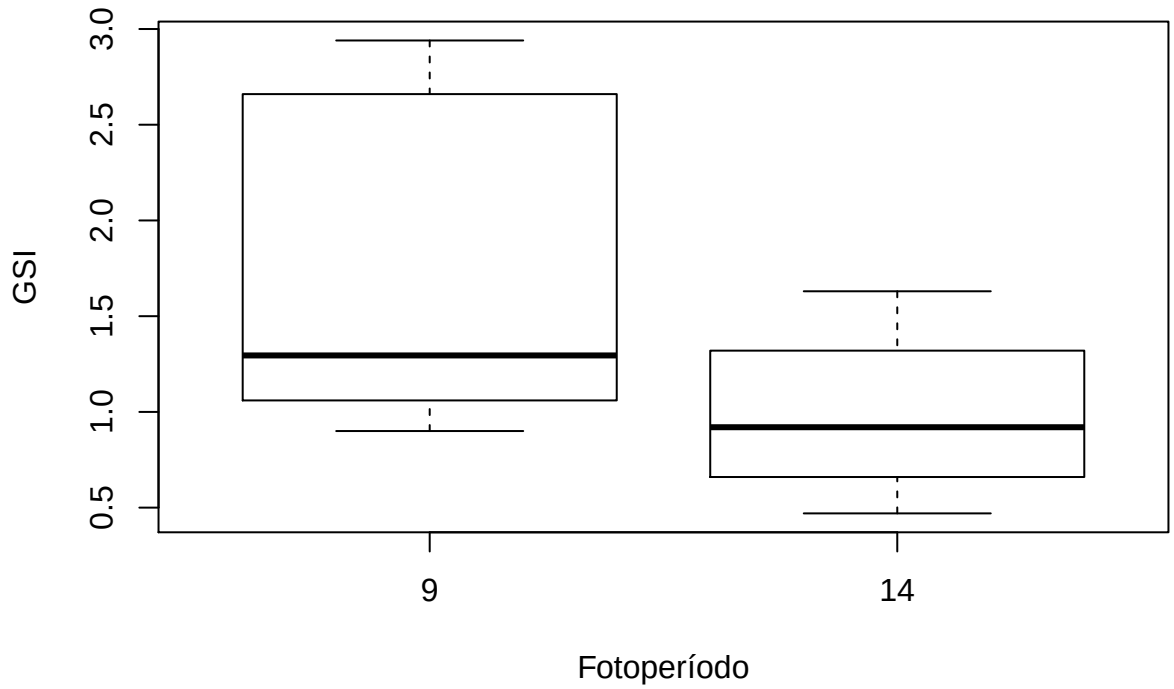
Dicha función se aplica a la tabla de datos modificada que hemos explicado:

```
summary(aov(X ~ A*B))
```

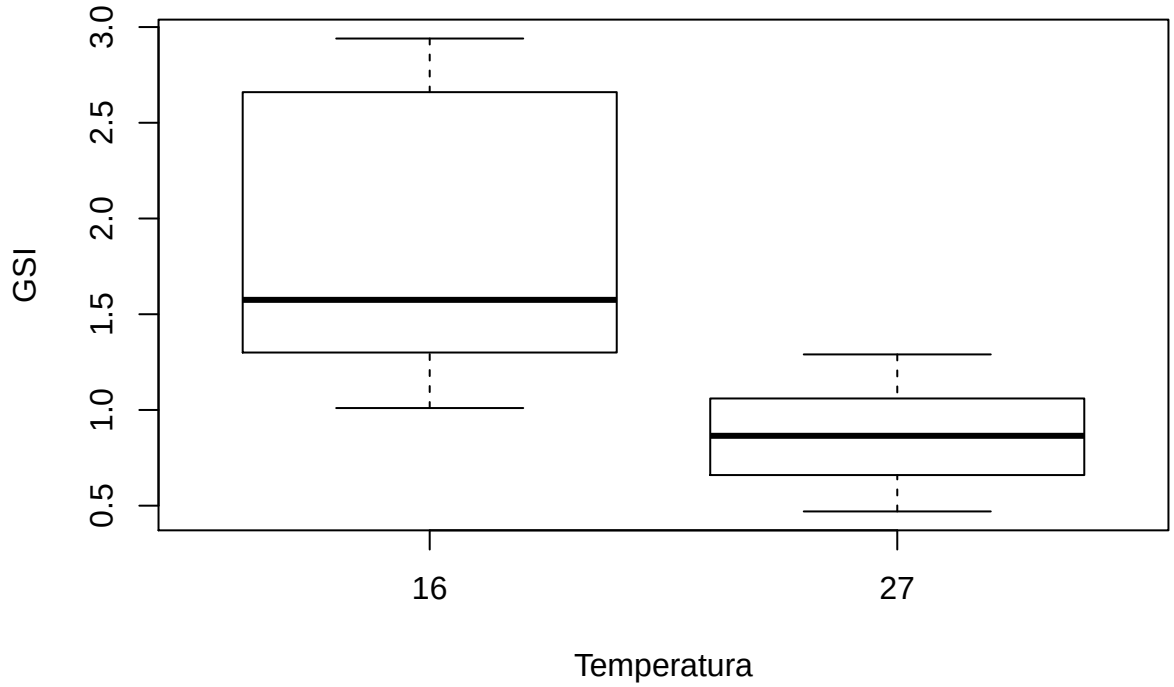
donde `X` es la variable donde se almacenan los valores X_{ijk} y en las variables factor `A` y `B`, los niveles de los factores `A` y `B`, respectivamente.

Hagamos un boxplot de la variable GSI según el fotoperíodo y según la temperatura para observar gráficamente si hay diferencias:

```
boxplot(GSI ~ fotoperiodos, data = tabla.datos.GSI, xlab="Fotoperíodo",ylab="GSI")
```

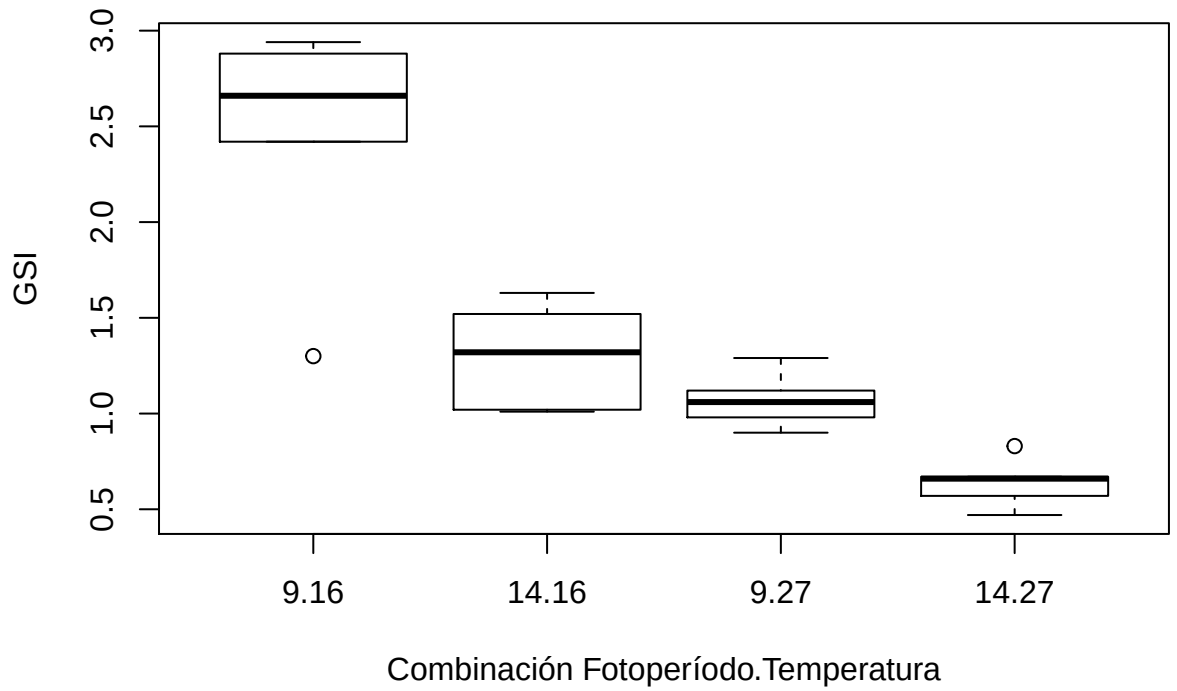



```
boxplot(GSI ~ temperatura, data = tabla.datos.GSI, xlab="Temperatura",ylab="GSI")
```



Si hacemos el boxplot de la variable GSI según la combinación de los dos factores, fotoperíodo y temperatura, obtenemos:

```
boxplot(GSI ~ fotoperiodos+temperatura, data = tabla.datos.GSI,
        xlab="Combinación Fotoperíodo.Temperatura",ylab="GSI")
```



Parece que sí hay diferencias en la variable GSI según el fotoperíodo, según la temperatura y según la combinación fotoperíodo/temperatura.

El contraste ANOVA de dos vías para los datos de nuestro ejemplo se realiza de la forma siguiente en R:

```
summary(aov(GSI ~ fotoperiodos*temperatura, data = tabla.datos.GSI))
```

```
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## fotoperiodos              1   3.08    3.08   21.70  0.00026 ***
## temperatura               1   5.15    5.15   36.29 0.000018 ***
## fotoperiodos:temperatura  1   0.63    0.63    4.44  0.05127 .
## Residuals                16   2.27    0.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtenemos los mismos resultados que hemos obtenido anteriormente.

Observamos que falta la fila de los tratamientos. Para realizar el contraste de los tratamientos en R, hacemos lo siguiente:

```
summary(aov(GSI ~ fotoperiodos:temperatura, data = tabla.datos.GSI))
```

```
##                          Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## fotoperiodos:temperatura  3   8.86   2.954    20.8 0.0000091 ***
## Residuals                16   2.27   0.142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.6.12. Gráficos de interacción

El gráfico de interacción entre los dos factores consiste en unir mediante segmentos los valores medios que toma la variable que comparamos `X` para cada factor en los que hemos segregado dicha variable.

Si no hay ninguna interacción entre los factores, los segmentos anteriores son paralelos. Cuanto más alejados del paralelismo estén dichos segmentos, más evidencia de interacción existe entre estos dos factores.

Para realizar un gráfico de interacción en `R` se usa la función `interaction.plot`:

```
interaction.plot(F1, F2, X)
```

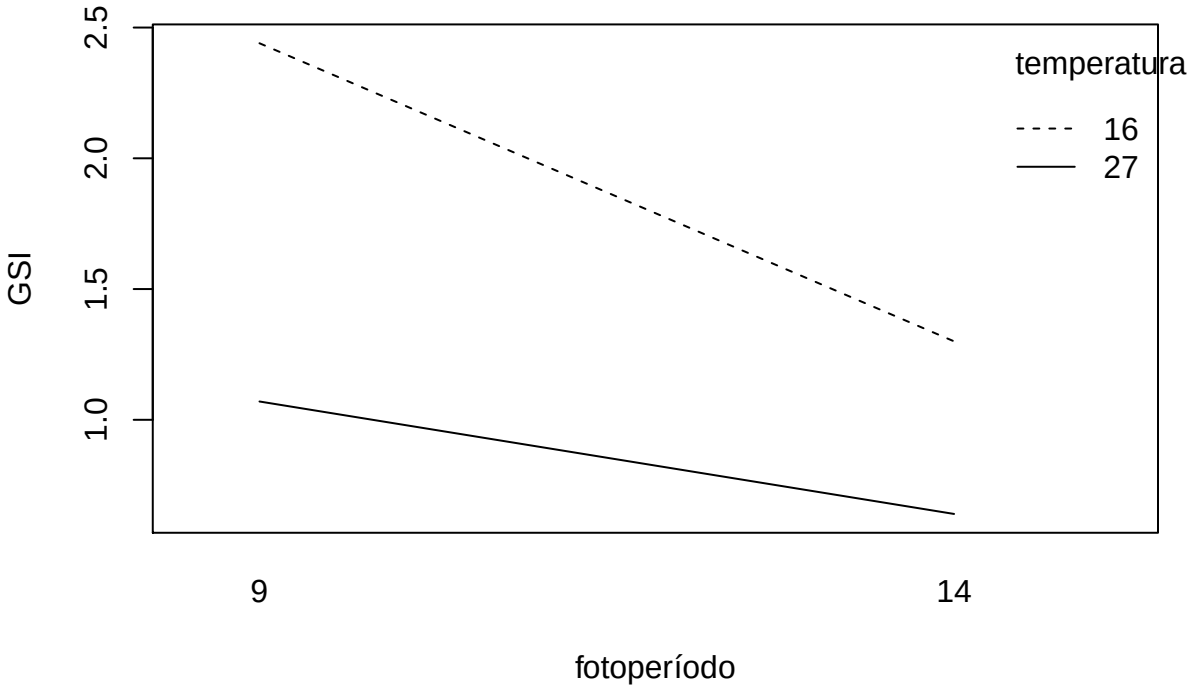
donde `F1` es el factor que dibujamos en el eje de abscisas o eje X y `F2` es el otro factor usado para dibujar los segmentos.

Para dibujar el gráfico de interacción entre factores, vamos a crear primero las tres variables siguientes:

```
GSI=tabla.datos.GSI$GSI
fotoperiodos=tabla.datos.GSI$fotoperiodos
temperatura=tabla.datos.GSI$temperatura
```

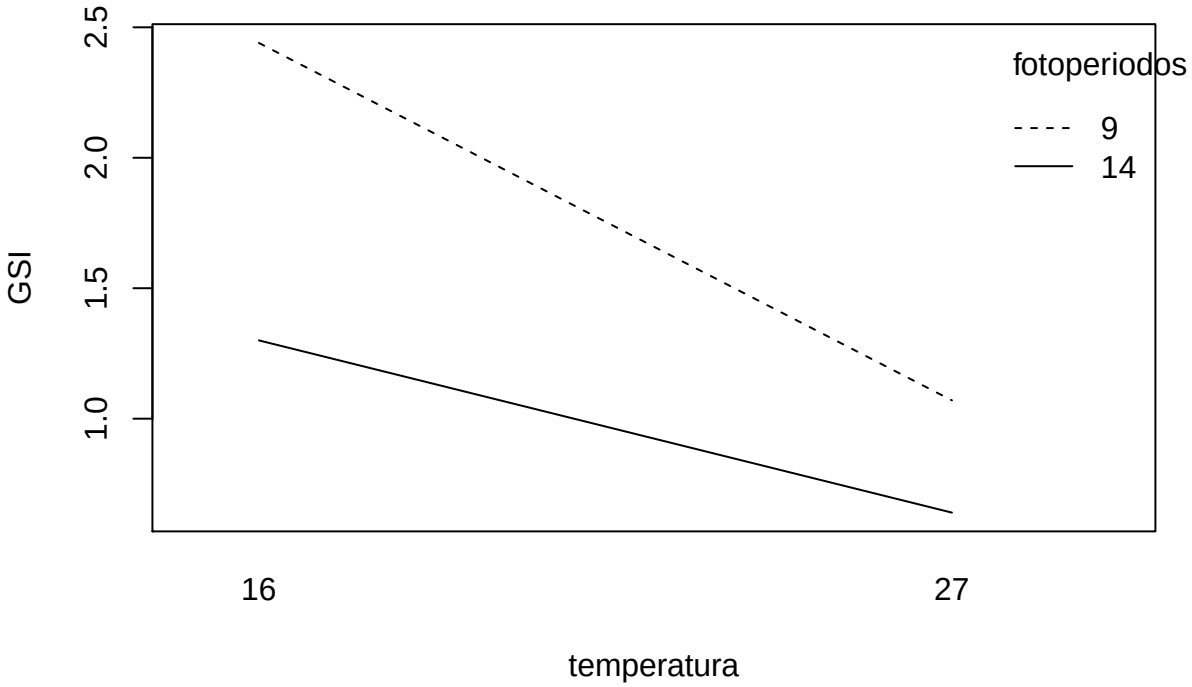
Para dibujar el gráfico de interacción del fotoperíodo según la temperatura, hacemos lo siguiente:

```
interaction.plot(fotoperiodos,temperatura,GSI, xlab="fotoperíodo",ylab="GSI")
```



Para dibujar el gráfico de interacción de la temperatura según el fotoperíodo, hacemos lo siguiente:

```
interaction.plot(temperatura,fotoperiodos,GSI,xlab="temperatura",ylab="GSI")
```



Vemos que las rectas anteriores no son paralelas pero tampoco "parecen" estar demasiado lejos del paralelismo.

De ahí el p-valor obtenido en la zona de penumbra.

7.7. Guía rápida

- `summary(aov(X ~ F))` nos la tabla ANOVA de un factor cuando contrastamos las medias de las subpoblaciones de la variable `X` segregada según los niveles de la variable factor `F`.
- `pairwise.t.test(X,F,p.adjust.method = ...)` nos realiza la comparación por parejas de niveles del factor `F` entre las medias de las subpoblaciones de la variable `X`. El parámetro `p.adjust.method` puede ser:
  - `none`: no hace ajuste alguno,
  - `bonferroni`: realiza el ajuste del método de Bonferroni: multiplicar el p-valor del contraste por el número de comparaciones llevadas a cabo, $\binom{k}{2}$.
  - `holm`: realiza el ajuste del método de Holm.
- `duncan.test(aov,"factor",group=...)$sufijo` del paquete `agricolae`. Realiza el test de Duncan de comparación por parejas de niveles del factor `F` entre las medias de las poblaciones de la variable `X`, donde:
  - `aov` es el resultado del ANOVA de partida,
  - el `factor` es el factor del ANOVA,
  - `group` puede ser `TRUE` o `FALSE` dependiendo de cómo queremos ver el resultado,
  - el `sufix` es `group` si `group=TRUE` y `comparison` si `group=FALSE`.
- `TukeyHSD(aov(X~F))`: realiza el test de Tukey de comparación por parejas de niveles del factor `F` entre las medias de las subpoblaciones de la variable `X`.
- `bartlett.test(X~F)`: realiza el test de Bartlett para ver si las varianzas de las subpoblaciones de la variable `X` segregada según los niveles de la variable factor `F` son iguales o no.
- `summary(aov(X ~ Tr+Bl))`: nos la tabla ANOVA de bloques completos aleatorios cuando contrastamos las medias de las subpoblaciones de la variable `X` segregada según los niveles de la variable factor `Tr` (tratamientos) donde los valores de cada vector de tratamientos está emparejada según la variable factor `Bl` (bloques).
- `summary(aov(X ~ A*B))`: nos da la tabla ANOVA de dos factores cuando cuando contrastamos las medias de las subpoblaciones de la variable `X` segregada según los niveles de las variables factor `A` y `B`.
- `interaction.plot(F1,F2,X)`: nos da el gráfico de interacción de los niveles del factor `F1` según los niveles de factor `F2` y `X` es la variable cuyas medias contrastamos en el ANOVA.

Capítulo 8

Regresión Lineal

8.1. Regresión lineal simple

El problema de **regresión** consiste en hallar la mejor **relación funcional** entre dos variables X e Y.

Más concretamente, dada una muestra de las dos variables X, Y, $(x_i, y_i)_{i=1,2,\ldots,n}$, queremos estudiar cómo depende el valor de Y en función del valor de X.

La variable aleatoria Y es la variable **dependiente** o **de respuesta**.

La variable (no necesariamente aleatoria) X es la variable **de control**, **independiente** o **de regresión**. Pensemos por ejemplo, en un experimento donde la variable X es la que controla el experimentador y la variable Y es el valor que se obtiene del experimento.

El problema de **regresión** es encontrar la mejor **relación funcional** que explique la variable Y conocidas las observaciones de la variable X.

Si dicha **relación funcional** es una recta, $Y = \beta_0 + \beta_1 x$, la **regresión** se denomina **regresión lineal**.

En la **regresión lineal**, se hace la suposición siguiente:

$$\mu_{Y|x} = \beta_0 + \beta_1 x,$$

dónde $\mu_{Y|x}$ es el valor esperado de la variable aleatoria Y cuando la variable X vale x. Dicho valor esperado es una función lineal de X con **término independiente** β_0 y **pendiente** β_1. Dichos valores son dos parámetros que tendremos que estimar.

Las estimaciones de β_0 y β_1 se llaman b_0 y b_1, respectivamente y se tienen que realizar a partir de la muestra $(x_i, y_i)_{i=1,2,\ldots,n}$.

Una vez halladas las estimaciones b_0 y b_1, obtendremos la **recta de regresión** para nuestra muestra:

$$\hat{y} = b_0 + b_1 x,$$

que dado un valor x_0 de X, estimará el valor $\hat{y}_0 = b_0 + b_1 x_0$ de la variable Y.

8.1.1. Mínimos cuadrados

Vamos a explicar el método para hallar las estimaciones b_0 y b_1.

Dicho método se denomina **método de los mínimos cuadrados**.

Dada una observación cualquiera de la muestra, (x_i, y_i), podremos separar la componente y_i como la suma de su **valor predicho por el modelo** y el error cometido:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i).$$

Llamamos **error cuadrático teórico** de este modelo a la suma al cuadrado de todos los errores cometidos por los valores de la muestra:

$$SS_\varepsilon = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

La **regresión lineal por mínimos cuadrados** consiste en hallar los estimadores b_0 y b_1 de β_0 y β_1 que minimicen dicho **error cuadrático teórico** SS_ε.

Observación: los errores cometidos pueden ser positivos o negativos. Entonces, para asegurarse de penalizar siempre los errores, se elevan éstos al cuadrado y de esta forma, siempre se suman y no pueden anularse.

Para hallar el mínimo del **error cuadrático teórico**, (b_0, b_1), hay que derivar respecto las variables β_0 y β_1 e igualar a 0 dichas derivadas:

$$\begin{aligned} \frac{\partial SS_\varepsilon}{\partial \beta_0}_{|\beta_0=b_0,\beta_1=b_1} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0, \\ \frac{\partial SS_\varepsilon}{\partial \beta_1}_{|\beta_0=b_0,\beta_1=b_1} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i = 0. \end{aligned}$$

La solución del sistema anterior es:

$$\begin{aligned} b_1 &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \\ b_0 &= \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}. \end{aligned}$$

Vamos a escribir las estimaciones b_0 y b_1 obtenidas anteriormente en función de las medias, varianzas y covarianza de la muestra de valores (x_i, y_i).

Sean entonces

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

las medias de las componentes x e y de los valores de la muestra y sean

$$\begin{aligned}
\tilde{s}_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{n}{n-1}\left(\frac{1}{n}\Big(\sum_{i=1}^{n}x_i^2\Big)-\overline{x}^2\right),\\
\tilde{s}_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{n}{n-1}\left(\frac{1}{n}\Big(\sum_{i=1}^{n}y_i^2\Big)-\overline{y}^2\right),\\
\tilde{s}_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \frac{n}{n-1}\left(\frac{1}{n}\Big(\sum_{i=1}^{n}x_iy_i\Big)-\overline{x}\cdot\overline{y}\right),
\end{aligned}$$

las varianzas y la covarianza de la muestra (x_i, y_i), $i = 1, \dots, n$.

Las estimaciones b_0 y b_1 se pueden reescribir de la forma siguiente:

Teorema. Los estimadores b_0 y b_1 de β_0 y β_1, respectivamente, hallados por el **método de los mínimos cuadrados** son los siguientes: $b_1 = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}$, $\quad b_0 = \overline{y} - b_1\overline{x}$.

Ejercicio

Se deja como ejercicio la demostración del teorema anterior a partir de la expresión hallada anteriormente.

Dado un valor x de la variable X, llamaremos $\hat{y}$ a la expresión $\hat{y} = b_0 + b_1 x$ al **valor estimado** de Y cuando $X = x$.

Dada una observación (x_i, y_i), llamaremos **error** de la observación e_i a la expresión $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$.

Ejemplo

En un experimento donde se quería estudiar la asociación entre consumo de sal y presión arterial, se asignó aleatoriamente a algunos individuos una cantidad diaria constante de sal en su dieta, y al cabo de un mes se los midió la tensión arterial media. Algunos resultados fueron los siguientes:

X (sal, en g)	Y (Presión, en mm de Hg)
1.8	100
2.2	98
3.5	110
4.0	110
4.3	112
5.0	120

Vamos a hallar la **recta de regresión lineal por mínimos cuadrados** de Y en función de X.

En primer lugar calculamos las medias, varianzas y covarianza de la muestra de datos:

```
sal=c(1.8,2.2,3.5,4.0,4.3,5.0)
tensión=c(100,98,110,110,112,120)
(media.sal = mean(sal))
```

```
## [1] 3.467
```

```
(media.tensión = mean(tensión))
```

```
## [1] 108.3
```

```
(var.sal = var(sal))
```

```
## [1] 1.543
```

```
(var.tensión = var(tensión))
```

```
## [1] 66.27
```

```
(cov.sal.tensión = cov(sal,tensión))
```

```
## [1] 9.773
```

Los estimadores b_0 y b_1 serán:

```
(b1 = cov.sal.tensión/var.sal)
```

```
## [1] 6.335
```

```
(b0 = media.tensión-b1*media.sal)
```

```
## [1] 86.37
```

La recta de regresión será, en este caso: $\hat{y} = 86{,}3708 + 6{,}3354 \cdot x$.

8.1.2. Recta de regresión en R

Para hallar la recta de regresión en R hay que usar la función `lm`:

```
lm(tensión ~sal)
```

```
##
## Call:
## lm(formula = tensión ~ sal)
##
## Coefficients:
## (Intercept)          sal
##       86.37         6.34
```

Propiedades de la recta de regresión.

La **recta de regresión** hallada por el **método de los mínimos cuadrados** verifica las propiedades siguientes:

- La **recta de regresión** pasa por el vector medio $(\overline{x}, \overline{y})$ de nuestra muestra de datos (x_i, y_i), $i = 1, \ldots, n$:
$$\overline{y} = b_0 + b_1 \overline{x}.$$
- La media de los valores estimados a partir de la **recta de regresión** es igual a la media de los

observados y_i, $\overline{y}$. Es decir:

$$\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1 x_i) = b_0 + b_1\overline{x} = \overline{y}.$$

- Los errores $(e_i)_{i=1,\ldots,n}$ tienen media 0:

$$\overline{e} = \frac{1}{n}\sum_{i=1}^{n}e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0.$$

Llamaremos **suma de cuadrados de los errores** a la cantidad siguiente: $SS_E = \sum_{i=1}^{n} e_i^2$.

Usando que los errores $(e_i)_{i=1,\ldots,n}$ tienen media 0, su varianza será:

$$s_e^2 = \frac{1}{n}\Big(\sum_{i=1}^{n}e_i^2\Big) - \overline{e}^2 = \frac{SS_E}{n} - 0 = \frac{SS_E}{n}.$$

Definimos las variables aleatorias E_{x_i} como $E_{x_i} = y_i - b_0 - b_1 \cdot x_i$ donde (x_i, y_i) es un valor de la muestra y b_0 y b_1 son los estimadores obtenidos por el **método de los mínimos cuadrados**. Entonces,

Teorema. Si las variables aleatorias error E_{x_i} tienen todas media 0 y la misma varianza σ_E^2 y, dos a dos, tienen covarianza 0, entonces

- b_0 y son b_1 los estimadores lineales no sesgados óptimos (más eficientes) de β_0 y β_1, respectivamente.

y un **estimador no sesgado de** σ_E^2 es el siguiente: $S^2 = \frac{SS_E}{n-2}$.

Si, además, las variables aleatorias error E_{x_i} son **normales**, entonces b_0 y son b_1 los estimadores máximo verosímiles de β_0 y β_1, respectivamente.

Ejemplo

Comprobemos las propiedades para los datos del ejemplo anterior:

- La **recta de regresión** pasa por el vector medio $(\overline{x}, \overline{y})$:

```
(round(media.tensión-b0-b1*media.sal,6))
```

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

- La media de los valores estimados a partir de la **recta de regresión** es igual a la media de los observados y_i, $\overline{y}$.

```
tensión.estimada = b0+b1*sal
(mean(tensión.estimada)-mean(tensión))
```

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

La estimación de la varianza para los datos del ejemplo anterior es la siguiente:

```
errores=tensión.estimada-tensión
SSE = sum(errores^2)
n=length(sal)
(estimación.varianza = SSE/(n-2))
```

```
## [1] 5.436
```

Entonces tenemos que el valor aproximado o estimado de σ_E^2 es 5.4365.

8.1.3. Coeficiente de determinación

Llegados a este punto, nos preguntamos lo efectiva que es la **recta de regresión**.

Es decir, cómo medir si la aproximación hallada $\widehat{y} = b_0 + b_1 x$ a la nube de puntos $(x_i, y_i),\ i = 1, \dots, n$ ha sido suficientemente buena.

Una forma de realizar dicha medición es a través del **coeficiente de determinación** R^2 que estima cuánta **variabilidad** de los valores y_i heredan los valores estimados $\widehat{y}_i$.

Para ver su definición, necesitamos introducir las **variabilidades** siguientes:

- **Variabilidad total** o suma total de cuadrados: $SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 = (n-1) \cdot \tilde{s}_y^2$.
- **Variabilidad de la regresión** o suma de cuadrados de la regresión: $SS_R = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = (n-1) \cdot \tilde{s}_{\widehat{y}}^2$.
- **Variabilidad del error** o suma de cuadrados del error: $SS_E = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = (n-1) \cdot \tilde{s}_e^2$.

El teorema siguiente nos relaciona las variabilidades anteriores:

Teorema. En una regresión lineal usando el método de los mínimos cuadrados, se cumple la siguiente relación entre las **variabilidades**:

$$SS_T = SS_R + SS_E,$$

o equivalentemente,

$$\tilde{s}_y^2 = \tilde{s}_{\widehat{y}}^2 + \tilde{s}_e^2.$$

Entonces, cuántas más "próximas" estén las **variabilidades** SS_T y SS_R, o, si se quiere, $\tilde{s}_y^2$ y $\tilde{s}_{\widehat{y}}^2$, más efectiva habrá sida la regresión, ya que la regresión habrá heredado mucha variabilidad de los datos y_i, $i = 1, \dots, n$ y la variabilidad del error, SS_E será pequeña.

El comentario anterior motiva la definición siguiente del **coeficiente de determinación** para medir la efectividad de la recta de regresión:

Definición: se define el **coeficiente de determinación** R^2 en la regresión por el método de los mínimos cuadrados como: $R^2 = \frac{SS_R}{SS_T} = \frac{\tilde{s}_{\widehat{y}}^2}{\tilde{s}_y^2}$.

Observación: el **coeficiente de determinación** R^2 es la fracción de la variabilidad de las componentes y que queda explicada por la variabilidad de las estimaciones correspondientes $\widehat{y}$.

Propiedades del coeficiente de determinación.

- El **coeficiente de determinación** es una cantidad entre 0 y 1: $0 \leq R^2 \leq 1$. Entonces, cuánto más próximo a 1 esté dicho coeficiente, más precisa será la recta de regresión.
- El **coeficiente de determinación** se puede expresar en función de la **variabilidad del error** de la forma siguiente:

$$R^2 = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T} = 1 - \frac{\tilde{s}_e^2}{\tilde{s}_y^2}.$$

- Se define el **coeficiente de correlación lineal** r_{xy} como $r_{xy} = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y}$. Entonces, el **coeficiente de determinación** R^2 es el cuadrado del **coeficiente de correlación lineal**: $R^2 = r_{xy}^2$.

Veamos la demostración de la última propiedad:

$$\begin{aligned} R^2 &= \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n}(b_1 x_i + b_0 - \overline{y})^2}{(n-1)\tilde{s}_y^2} = \frac{\sum_{i=1}^{n}\left(\frac{\tilde{s}_{xy}}{\tilde{s}_x^2}x_i - \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\overline{x}\right)^2}{(n-1)\tilde{s}_y^2} \\ &= \frac{\frac{\tilde{s}_{xy}^2}{\tilde{s}_x^4}\sum_{i=1}^{n}(x_i - \overline{x})^2}{(n-1)\tilde{s}_y^2} = \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^4} \cdot \frac{\tilde{s}_x^2}{\tilde{s}_y^2} = \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2 \cdot \tilde{s}_y^2} = r_{xy}^2 \end{aligned}$$

Ejemplo

Calculemos las variabilidades anteriores y el **coeficiente de determinación** para los datos de nuestro ejemplo:

- **Variabilidad total**:

```
(SST = sum((tensión-media.tensión)^2))
```

```
## [1] 331.3
```

- **Variabilidad de la regresión**:

```
(SSR = sum((tensión.estimada-media.tensión)^2))
```

```
## [1] 309.6
```

- **Variabilidad del error**:

```
(SSE = sum((tensión-tensión.estimada)^2))
```

```
## [1] 21.75
```

Comprobemos que se cumple $SST = SSR + SSE$:

```
(round(SST-SSR-SSE,6))
```

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

El coeficiente de determinación R^2 será:

```
(R2=SSR/SST)
```

```
## [1] 0.9344
```

Otra manera de calcularlo es:

```
(R2=var(tensión.estimada)/var(tensión))
```

```
## [1] 0.9344
```

En este caso, la regresión explica un 93.44 % de la variabilidad de los datos.

Coeficiente de determinación en R.

Para hallar el **coeficiente de determinación** en R hemos de usar las funciones `lm` y `summary` junto con el parámetro `r.squared`:

```
summary(lm(y ~ x))$r.squared
```

Si aplicamos las funciones anteriores a los datos de nuestro ejemplo, obtenemos:

```
summary(lm(tensión ~ sal))$r.squared
```

```
## [1] 0.9344
```

Usar solamente el **coeficiente de determinación** para medir la calidad de la regresión es un error.

Tenemos que observar más información para poder afirmar que la regresión obtenida es adecuada y se ajusta bien a nuestros datos.

En R existe una tabla de datos denominada `anscombe` que pone de manifiesto este hecho. Vamos a echarle un vistazo:

```
data(anscombe)
str(anscombe)
```

```
## 'data.frame':    11 obs. of  8 variables:
##  $ x1: num  10 8 13 9 11 14 6 4 12 7 ...
##  $ x2: num  10 8 13 9 11 14 6 4 12 7 ...
##  $ x3: num  10 8 13 9 11 14 6 4 12 7 ...
##  $ x4: num  8 8 8 8 8 8 8 19 8 8 ...
##  $ y1: num  8.04 6.95 7.58 8.81 8.33 ...
##  $ y2: num  9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 9.13 7.26 ...
##  $ y3: num  7.46 6.77 12.74 7.11 7.81 ...
##  $ y4: num  6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 5.56 7.91 ...
```

Vemos que tiene 4 parejas de valores (x_i, y_i) de tamaño 11:

```
anscombe
```

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
```

```
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

Si calculamos los **coeficientes de determinación** para las 4 parejas, obtenemos un resultado similar:

```
summary(lm(y1~x1,data=anscombe))$r.squared
```

```
## [1] 0.6665
```

```
summary(lm(y2~x2,data=anscombe))$r.squared
```

```
## [1] 0.6662
```

```
summary(lm(y3~x3,data=anscombe))$r.squared
```

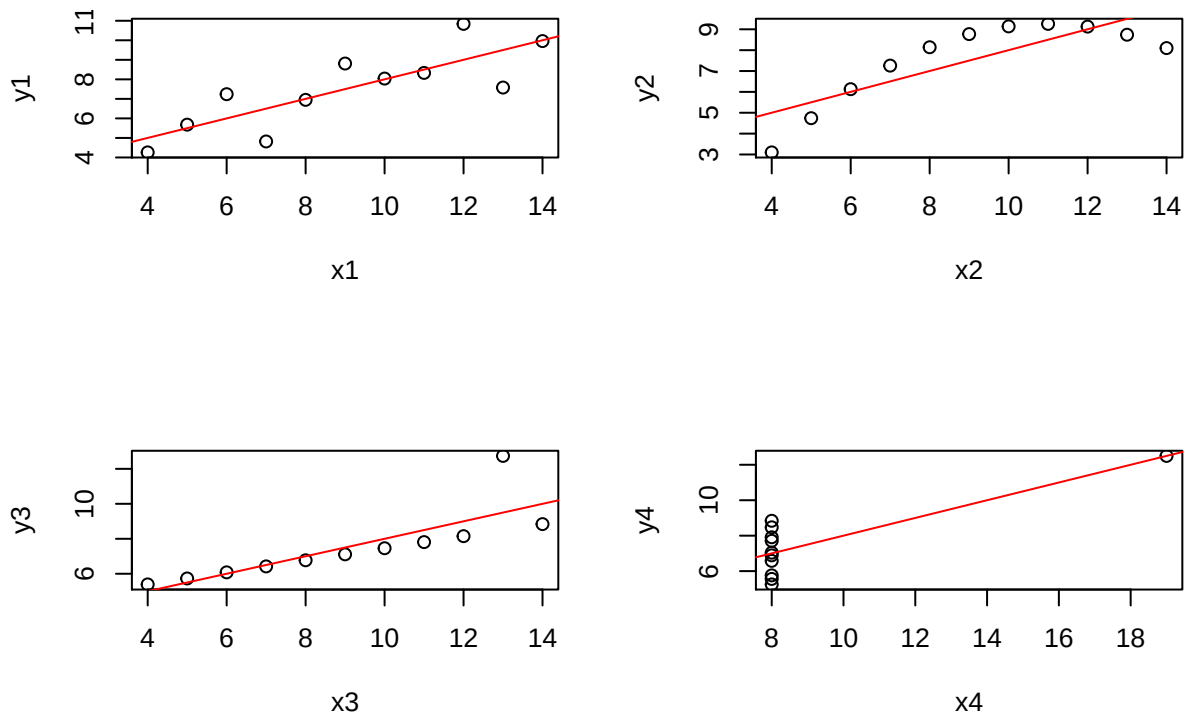
```
## [1] 0.6663
```

```
summary(lm(y4~x4,data=anscombe))$r.squared
```

```
## [1] 0.6667
```

En cambio, si vemos su representación gráfica, su aspecto es muy distinto:

```
par(mfrow=c(2,2))
plot(y1~x1,data=anscombe)
abline(lm(y1~x1,data=anscombe),col=2)
plot(y2~x2,data=anscombe)
abline(lm(y2~x2,data=anscombe),col=2)
plot(y3~x3,data=anscombe)
abline(lm(y3~x3,data=anscombe),col=2)
plot(y4~x4,data=anscombe)
abline(lm(y4~x4,data=anscombe),col=2)
```



Observamos que en el caso de la tabla de datos `(x3,y3)`, la recta de regresión ha sido efectiva pero en los demás hemos obtenido un error considerable donde el peor caso es el de la tabla de datos `(x4,y4)`.

Por tanto, considerar sólo el valor del **coeficiente de determinación** para medir el ajuste de la recta de regresión a nuestros datos no es conveniente.

8.1.4. Intervalos de confianza

Para poder hallar los intervalos de confianza al $100 \cdot (1-\alpha)\,\%$ de confianza sobre los parámetros β_0 y β_1, necesitamos el supuestos siguiente:

Para cada valor x_i de la variable X, las variables aleatorias E_{x_i} sigue una distribución normal con media $\mu_{E_{x_i}} = 0$ y varianza σ_E^2 constante independiente del valor x_i. También supondremos que dados x_i y x_j dos valores distintos de la variable X, la covarianza entre las variables E_{x_i} y E_{x_j} es nula: $\sigma(E_{x_i}, E_{x_j}) = 0$.

Ejemplo

Comprobemos para los datos de nuestro ejemplo si los errores siguen una distribución normal usando el test de **Kolmogorov-Smirnov-Lilliefors**:

```
library(nortest)
lillie.test(errores)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
```

```
## data:  errores
## D = 0.28, p-value = 0.1
```

Como el p-valor es grande, podemos concluir que no tenemos evidencias suficientes para rechazar que los errores siguen una distribución normal.

Bajo las hipótesis anteriores tenemos los dos resultados siguientes:

Teorema. Los errores estándar de los estimadores b_1 y b_0 son, respectivamente,

$$\frac{\sigma_E}{\tilde{s}_x\sqrt{n-1}} \quad \text{y} \quad \sigma_E\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{(n-1)\tilde{s}_x^2}},$$

donde recordemos que para estimar σ_E, usamos el estimador $S=\sqrt{S^2}$.

Teorema. Las variables aleatorias

$$\frac{b_1-\beta_1}{\frac{S}{\tilde{s}_x\sqrt{n-1}}} \quad \text{y} \quad \frac{b_0-\beta_0}{S\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{(n-1)\tilde{s}_x^2}}},$$

siguen leyes t de Student con $n-2$ grados de libertad.

Entonces bajo el supuesto anterior, los intervalos de confianza para los parámetros β_1 y β_0 al $100\cdot(1-\alpha)\,\%$ de confianza son los siguientes:

- β_1: $\left]b_1-t_{n-2,1-\frac{\alpha}{2}}\frac{S}{\tilde{s}_x\sqrt{n-1}}, b_1+t_{n-2,1-\frac{\alpha}{2}}\frac{S}{\tilde{s}_x\sqrt{n-1}}\right[$. Lo escribiremos para simplificar de la forma siguiente: $\beta_1=b_1\pm t_{n-2,1-\frac{\alpha}{2}}\frac{S}{\tilde{s}_x\sqrt{n-1}}$.
- β_0: $b_0\pm t_{n-2,1-\frac{\alpha}{2}}S\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{(n-1)\tilde{s}_x^2}}$.

Ejemplo

Hallemos los intervalos de confianza para el 95 % de confianza para los parámetros β_1 y β_0 usando los datos del ejemplo que hemos ido desarrollando:

```
alpha=0.05
S=sqrt(estimación.varianza)
extremo.izquierda.b1 = b1-qt(1-alpha/2,n-2)*S/(sd(sal)*sqrt(n-1))
extremo.derecha.b1 = b1+qt(1-alpha/2,n-2)*S/(sd(sal)*sqrt(n-1))
print('Intervalo confianza para b1:')
```

```
## [1] "Intervalo confianza para b1:"
```

```
print(c(extremo.izquierda.b1,extremo.derecha.b1))
```

```
## [1] 4.004 8.666
```

```
extremo.izquierda.b0 = b0-qt(1-alpha/2,n-2)*S*sqrt(1/n+media.sal^2/((n-1)*var(sal)))
extremo.derecha.b0 = b0+qt(1-alpha/2,n-2)*S*sqrt(1/n+media.sal^2/((n-1)*var(sal)))
print('Intervalo confianza para b0:')
```

```
## [1] "Intervalo confianza para b0:"
```

```
print(c(extremo.izquierda.b0,extremo.derecha.b0))
```

```
## [1] 77.87 94.87
```

Intervalos de confianza en R.

Para hallar los intervalos de confianza de los parámetros β_1 y β_0 en R hay que aplicar la función `confint` al objeto `lm(...)`. El parámetro `level` nos da el nivel de confianza cuyo valor por defecto es 0.95.

Para los datos de nuestro ejemplo, los intervalos de confianza para los parámetros β_0 y β_1 serían los siguientes en R:

```
confint(lm(tensión~sal),level=0.95)
```

```
##               2.5 % 97.5 %
## (Intercept) 77.869 94.873
## sal          4.004  8.666
```

Fijado un valor x_0 de la variable X, podemos considerar dos parámetros a estudiar: el valor medio de la variable aleatoria $Y|x_0$, $\mu_{Y|x_0}$ y el valor estimado y_0 por la recta de regresión.

Dichos intervalos nos ayudan a estudiar cómo se comporta la regresión cuando el valor de la variable X vale un determinado valor x_0.

El estimador de los parámetros anteriores, $\mu_{Y|x_0}$ y y_0 es el mismo: $\hat{y}_0 = b_0 + b_1 \cdot x_0$ pero los errores estándares cambian dependiendo del parámetro que consideremos como indican los dos resultados siguientes:

Bajo el supuesto anterior, sea x_0 un valor concreto de la variable X. Entonces:

Teorema. El error estándar del estimador $\hat{y}_0$ del parámetro $\mu_{Y|x_0}$ es el siguiente: $\sigma_E\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{(n-1)\tilde{s}_x^2}}$.

Usando el resultado anterior, se tiene que la variable aleatoria: $\frac{\hat{y}_0-\mu_{Y/x_0}}{S\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{(n-1)\tilde{s}_x^2}}}$ sigue una ley t de Student con $n-2$ grados de libertad.

Teorema. El error estándar del estimador $\hat{y}_0$ del parámetro y_0 es el siguiente: $\sigma_E\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{(n-1)\tilde{s}_x^2}}$.

Usando el resultado anterior, se tiene que la variable aleatoria: $\frac{\hat{y}_0-y_0}{S\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{(n-1)\tilde{s}_x^2}}}$ sigue una ley t de Student con $n-2$ grados de libertad.

A partir de los teoremas anteriores, hallamos los intervalos de confianza para los parámetros $\mu_{Y|x_0}$ y y_0 al $100\cdot(1-\alpha)\,\%$ de confianza:

- $\mu_{Y|x_0}$: $\hat{y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} S\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{(n-1)\tilde{s}_x^2}}$.
- y_0: $\hat{y}_0 \pm t_{n-2,1-\frac{\alpha}{2}} S\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{(n-1)\tilde{s}_x^2}}$.

Ejemplo

Hallemos un intervalo de confianza para un nivel de sal de $x_0 = 4{,}5$ g. para los parámetros $\mu_{Y|4,5}$ y y_0 al 95 % de confianza:

```
alpha=0.05
x0=4.5
y0.estimado = b0+b1*x0
extremo.izquierda.mu.x0 = y0.estimado-qt(1-alpha/2,n-2)*S*
  sqrt(1/n+(x0-media.sal)^2/((n-1)*var(sal)))
extremo.derecha.mu.x0 = y0.estimado+qt(1-alpha/2,n-2)*S*
  sqrt(1/n+(x0-media.sal)^2/((n-1)*var(sal)))
print(paste("Intervalo de confianza para mu de Y para x0=",x0))
```

```
## [1] "Intervalo de confianza para mu de Y para x0= 4.5"
```

```
print(c(extremo.izquierda.mu.x0,extremo.derecha.mu.x0))
```

```
## [1] 111.3 118.5
```

```
extremo.izquierda.y0 = y0.estimado-qt(1-alpha/2,n-2)*S*
  sqrt(1+1/n+(x0-media.sal)^2/((n-1)*var(sal)))
extremo.derecha.y0 = y0.estimado+qt(1-alpha/2,n-2)*S*
  sqrt(1+1/n+(x0-media.sal)^2/((n-1)*var(sal)))
print(paste("Intervalo de confianza para y0 para x0=",x0))
```

```
## [1] "Intervalo de confianza para y0 para x0= 4.5"
```

```
print(c(extremo.izquierda.y0,extremo.derecha.y0))
```

```
## [1] 107.5 122.3
```

Para hallar el intervalo de confianza para el parámetros $\mu_{Y|x_0}$ al $100 \cdot (1-\alpha)\,\%$ de confianza hay que usar la función `predict.lm` de la forma siguiente:

```
newdata=data.frame(x=x0)
predict.lm(lm(y~x),newdata,interval="confidence",level= nivel.confianza)
```

Para el parámetro y_0 hay que usar la misma función anterior pero cambiando el parámetro `interval` al valor `prediction`:

```
newdata=data.frame(x=x0)
predict.lm(lm(y~x),newdata,interval="prediction",level= nivel.confianza)
```

Ejemplo

Hallemos los intervalos de confianza para los parámetros $\mu_{Y|4,5}$ y y_0 al $95\,\%$ de confianza usando `R`:

```
newdata=data.frame(sal=4.5)
predict.lm(lm(tensión~sal),newdata,interval="confidence",level= 0.95)
```

```
##     fit   lwr   upr
## 1 114.9 111.3 118.5
```

```
predict.lm(lm(tensión~sal),newdata,interval="prediction",level= 0.95)
```

```
##     fit   lwr   upr
## 1 114.9 107.5 122.3
```

8.1.5. Contraste de hipótesis sobre la pendiente de la recta β_1

Cuando nos planteamos hacer una regresión de la variable Y sobre la variable X, estamos suponiendo que la variable X influye en la variable Y. Es decir, que si cambiamos el valor de la variable X, habrá un cambio en la variable Y.

Decir que la variable X influye en la variable Y en el contexto de la **regresión lineal** es equivalente a decir que $\beta_1 \neq 0$ ya que si $\beta_1 = 0$, tendríamos que la variación de la variable Y sólo se debería a fluctuaciones aleatorias.

Por tanto, es interesante plantearse el **contraste de hipótesis** siguiente sobre el parámetro **pendiente de la recta de regresión**: β_1:

$$\begin{cases} H_0 : \beta_1 = 0, \\ H_1 : \beta_1 \neq 0. \end{cases}$$

En el supuesto de que las variables aleatorias E_{x_i} son normales $N(0, \sigma_E^2)$, el **estadístico de contraste** para realizar el contraste anterior es el siguiente: $T = \frac{b_1}{\frac{S}{\tilde{s}_x\sqrt{n-1}}}$, que, suponiendo que la hipótesis nula es cierta, se distribuye según una t de Student con $n-2$ grados de libertad.

Si t_0 es el valor del estadístico de contraste para los valores de nuestra muestra, el p-valor del contraste anterior es el siguiente:

$$p = 2 \cdot P(t_{n-2} > |t_0|),$$

con el significado usual:

- si $p < 0{,}05$, concluimos que tenemos evidencias suficientes para rechazar la hipótesis nula y, por tanto, tiene sentido la regresión al no rechazar que $\beta_1 \neq 0$,
- si $p > 0{,}1$, concluimos que no tenemos evidencias suficientes para rechazar la hipótesis nula. En este caso, la regresión no tendría sentido al no rechazar que $\beta_1 = 0$ y,
- si $0{,}05 \leq p \leq 0{,}1$, estamos en la zona de penumbra. Necesitamos más datos para tomar una decisión clara.

Otra forma de realizar el contraste anterior es observar el intervalo de confianza para β_1 y ver si contiene el valor 0.

En caso de contener el valor 0, la regresión no tendría sentido para el nivel de confianza de $100 \cdot (1-\alpha)\,\%$ ya que no rechazamos que $\beta_1 = 0$ y en caso de no contener el valor 0, la regresión sí tendría sentido al nivel de confianza anterior.

Ejemplo

Realicemos el contraste anterior para los datos de nuestro ejemplo. El valor del estadístico de contraste será:

```
(t0 = b1/(S/(sd(sal)*sqrt(n-1))))
```

```
## [1] 7.546
```

y el p-valor vale:

```
(p=2*pt(abs(t0),n-2,lower.tail = FALSE))
```

```
## [1] 0.001652
```

Como el p-valor es pequeño, podemos concluir que la regresión tiene sentido en este caso.

Para realizar el contraste anterior en R, hay que estudiar la salida de la función `summary` aplicada a la función `lm`:

```
summary(lm(y~x))
```

La salida anterior para los datos de nuestro ejemplo es la siguiente:

```
summary(lm(tensión ~ sal))
```

```
##
## Call:
## lm(formula = tensión ~ sal)
##
## Residuals:
##     1     2     3     4     5     6
##  2.23 -2.31  1.46 -1.71 -1.61  1.95
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    86.37       3.06   28.21 0.0000094 ***
## sal             6.33       0.84    7.55    0.0017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 4 degrees of freedom
## Multiple R-squared:  0.934,  Adjusted R-squared:  0.918
## F-statistic: 56.9 on 1 and 4 DF,  p-value: 0.00165
```

En primer lugar R nos da los errores de los datos cometidos al estimar los valores y_i por $\hat{y}_i$.

A continuación, en una tabla, nos da las estimaciones de los parámetros β_0 y β_1, b_0 y b_1, en la columna `Estimate`, los errores estándar de dichos estimadores en la columna `Std. Error` y el valor del estadístico t en la columna `t value`. En la última columna nos da el p-valor para el contraste anterior.

Observamos que hay dos valores de los estadísticos de contraste, uno corresponde al contraste para β_1 (la segunda fila) y otro corresponde al contraste para β_0 que no tiene ningún interés para ver si la regresión tiene sentido. Pensemos que el hecho de que el parámetro β_0 sea nulo no contradice el hecho de que la variable X tenga efecto en la variable Y.

Por último, en el último párrafo de la salida, nos da el valor del error residual o la estimación de S, el valor del coeficiente de determinación R^2, el `Multiple R-squared`, y el valor de R^2 ajustado del que hablaremos más adelante.

En la última fila, nos habla del estadístico F que es otra manera de realizar el contraste sobre el parámetro β_1 que hemos explicado anteriormente pero en lugar de usar el estadístico T, se usa el estadístico T^2 que, si la hipótesis nula es cierta, se distribuye según una F de Fisher con 1 y $n-2=4$ grados de libertad. Podéis comprobar por ejemplo que t_0^2 vale el valor del `F-statistic`:

```
t0^2
```

```
## [1] 56.95
```

8.2. Regresión lineal múltiple

En la **regresión lineal simple**, estudiábamos si una **variable dependiente** Y dependía linealmente de una **variable independiente o de control** X.

En la práctica, dicha situación raramente se da ya que la **variable dependiente o de respuesta** Y suele depender de más de una **variable de control**.

Por tanto, en esta sección vamos a generalizar todo el estudio que hemos hecho para la **regresión lineal simple** donde sólo hay una **variable de control** al caso en que tengamos k variables de control $X_1, \dots, X_k$.

En resumen, suponemos que tenemos una **variable dependiente** Y y k **variables independientes o de control** $X_1, \dots, X_k$.

8.2.1. Modelo de regresión lineal múltiple

De forma similar a la **regresión lineal simple**, suponemos que la relación de la media de la variable aleatoria Y, fijados k valores de las variables $X_1, \dots, X_k$, $x_1, \dots, x_k$, es la siguiente:

$$\mu_{Y|x_1,\dots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Es decir, la media la media de la variable aleatoria $Y_{x_1,\dots,x_k}$ es una función lineal de los valores $x_1, \dots, x_k$.

8.2.2. Modelo de regresión lineal múltiple

Los valores $\beta_0, \beta_1, \dots, \beta_k$ son los llamados **parámetros de regresión** y se tienen que estimar a partir de una muestra de las variables consideradas:

$$(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)_{i=1,\dots,n}$$

Para que dichas estimaciones se puedan realizar, hay que suponer que $n > k$ ya que en caso contrario tendríamos un problema *subestimado*: tendríamos más parámetros que valores en la muestra.

Denotaremos el vector $\underline{x}_i$ al conjunto de los k valores del individuo i-ésimo: $\underline{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$.

Escribimos el modelo de **regresión lineal múltiple** de la forma siguiente:

$$Y|x_1, \dots, x_k = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + E_{x_1,\dots,x_k},$$

donde

- $Y|x_1, \dots, x_k$ es la v.a. que da el valor de Y cuando cada X_i vale x_i, $X_i = x_i$,
- $E_{x_1,\dots,x_k}$ son las v.a. error, o residuales, y representan el error aleatorio del modelo asociado a $(x_1, \dots, x_k)$.

A partir de una muestra

$$(\underline{x}_i, y_i)_{i=1,2,\dots,n}$$

vamos a obtener estimaciones $b_0, b_1, \dots, b_k$ de los **parámetros de regresión** $\beta_0, \beta_1, \dots, \beta_k$.

Una vez obtenidas las estimaciones $b_0, b_1, \dots, b_k$, podemos definir los valores siguientes:

$$\begin{aligned} \hat{y}_i &= b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}, \\ y_i &= b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + e_i, \end{aligned}$$

donde

- $\hat{y}_i$ es el valor predicho de a partir de y_i y $\underline{x}_i$ y
- e_i estima el error $E_{\underline{x}_i}$: $e_i = y_i - \hat{y}_i$.

Para simplificar la notación, escribimos los datos de la muestra en forma matricial.

En primer lugar, definimos los vectores siguientes:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}, \ \hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}, \ \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Definimos la matriz $\mathbf{X}$ a partir de los datos de la muestra de las variables X_i, $i = 1, \dots, k$:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}.$$

Las ecuaciones

$$\begin{aligned} \hat{y}_i &= b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}, \\ y_i &= b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + e_i, \end{aligned}$$

se escriben en forma matricial de la forma siguiente:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X} \cdot \mathbf{b}, \\ \mathbf{y} &= \mathbf{X} \cdot \mathbf{b} + \mathbf{e}. \end{aligned}$$

8.2.3. Método de los mínimos cuadrados

Definimos el **error cuadrático** SS_E cómo:

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2.$$

Los **estimadores** de los **parámetros** $\beta_0, \beta_1, \dots, \beta_k$ por el método de **mínimos cuadrados** serán los valores $b_0, b_1, \dots, b_k$ que minimicen SS_E.

Para calcularlos, calculamos las derivadas parciales del error cuadrático SS_E respecto cada b_i, las igualamos a 0, las resolvemos, y comprobamos que la solución $(b_0, \dots, b_k)$ encontrada corresponde a un mínimo.

El teorema siguiente nos da la expresión de dichos **estimadores**:

Teorema. Los **estimadores** por el método de los **mínimos cuadrados** de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ a partir de la muestra $(\underline{x}_i, y_i)_{i=1,2,\ldots,n}$ son los siguientes:

$$\mathbf{b} = \left(\mathbf{X}^\top \cdot \mathbf{X}\right)^{-1} \cdot \left(\mathbf{X}^\top \cdot \mathbf{y}\right).$$

Ejemplo

Se postula que la altura de un bebé (y) tiene una relación lineal con su edad en días (x_1), su altura al nacer en cm. (x_2), su peso en kg. al nacer (x_3) y el aumento en tanto por ciento de su peso actual respecto de su peso al nacer (x_4)

El modelo es:

$$\mu_{Y|x_1,x_2,x_3,x_4} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

En una muestra de $n = 9$ niños, se obtuvieron los resultados siguientes:

y	x_1	x_2	x_3	x_4
57.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67	88	49	5.52	36.5
53.5	67	43	3.21	27.2
62.7	80	48	4.32	27.7
56.2	74	48	2.31	28.3
68.5	94	53	4.3	30.3
69.2	102	58	3.71	28.7

Hallemos los estimadores de los parámetros $\beta_0, \beta_1, \beta_2, \beta_3$ y β_4.

En primer lugar, hallamos la matriz $\mathbf{X}$ y el vector $\mathbf{y}$:

$$\mathbf{X} = \begin{pmatrix} 1 & 78 & 48{,}2 & 2{,}75 & 29{,}5 \\ 1 & 69 & 45{,}5 & 2{,}15 & 26{,}3 \\ 1 & 77 & 46{,}3 & 4{,}41 & 32{,}2 \\ 1 & 88 & 49 & 5{,}52 & 36{,}5 \\ 1 & 67 & 43 & 3{,}21 & 27{,}2 \\ 1 & 80 & 48 & 4{,}32 & 27{,}7 \\ 1 & 74 & 48 & 2{,}31 & 28{,}3 \\ 1 & 94 & 53 & 4{,}3 & 30{,}3 \\ 1 & 102 & 58 & 3{,}71 & 28{,}7 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 57{,}5 \\ 52{,}8 \\ 61{,}3 \\ 67 \\ 53{,}5 \\ 62{,}7 \\ 56{,}2 \\ 68{,}5 \\ 69{,}2 \end{pmatrix}.$$

El vector de los valores estimados $\mathbf{b}$ valdrá:

$$\mathbf{b} = \left(\mathbf{X}^t \cdot \mathbf{X}\right)^{-1} \cdot \left(\mathbf{X}^t \cdot \mathbf{y}\right).$$

Realicemos los cálculos anteriores en R:

```
X=matrix(c(1,78,48.2,2.75,29.5,1,69,45.5,2.15,26.3,
 1,77,46.3,4.41,32.2,1,88,49,5.52,36.5,
```

```
 1,67,43,3.21,27.2,1,80,48,4.32,27.7,
 1,74,48,2.31,28.3,1,94,53,4.3,30.3,
 1,102,58,3.71,28.7),nrow=9,byrow=TRUE)
y.bebes=cbind(c(57.5,52.8,61.3,67,53.5,62.7,56.2,68.5,69.2))
(estimaciones.b = solve(t(X)%*%X)%*%(t(X)%*%y.bebes))
```

```
##          [,1]
## [1,]  7.14753
## [2,]  0.10009
## [3,]  0.72642
## [4,]  3.07584
## [5,] -0.03004
```

La función lineal de regresión buscada se:

$$\widehat{y} = 7{,}1475 + 0{,}1001x_1 + 0{,}7264x_2 + 3{,}0758x_3 - 0{,}03x_4.$$

Podemos observar, por ejemplo, que cuánto más edad tenga el niño, más altura tiene, cuánto más peso al nacer, más altura tiene pero cuánto más haya aumentado su peso respecto de su peso al nacer, su altura disminuye aunque dicha disminución es minúscula.

8.2.4. Cálculo de la función de regresión en R

Para calcular la función de regresión en R hay que usar la función lm:

```
lm(y ~ x1+x2+...+xk)
```

En nuestro ejemplo, se calcula de la forma siguiente:

```
lm(y.bebes ~ X[,2]+X[,3]+X[,4]+X[,5])
```

```
##
## Call:
## lm(formula = y.bebes ~ X[, 2] + X[, 3] + X[, 4] + X[, 5])
##
## Coefficients:
## (Intercept)       X[, 2]       X[, 3]       X[, 4]       X[, 5]
##       7.148        0.100        0.726        3.076       -0.030
```

Propiedades de la función de regresión.

- La **función de regresión** pasa por el vector medio $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k, \overline{y})$:

$$\overline{y} = b_0 + b_1\overline{x}_1 + \cdots + b_k\overline{x}_k.$$

- La media de los valores estimados se igual a la media de los observados:

$$\overline{\widehat{y}} = \overline{y}.$$

- Los errores $(e_i)_{i=1,\ldots,n}$ tienen media 0 y varianza:

$$\widetilde{s}_e^2 = \frac{SS_E}{n-1}.$$

Verifiquemos las propiedades anteriores para los datos de nuestro ejemplo:

- La **función de regresión** pasa por el vector medio $(\overline{x}_1, \overline{x}_2, \dots, \overline{x}_k, \overline{y})$:

```
vectores.medios = apply(X[,1:5],2,mean)
round(mean(y.bebes)-t(estimaciones.b)%*%vectores.medios,6)
```

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

- La media de los valores estimados se igual a la media de los observados:

```
valores.estimados = X%*%estimaciones.b
round(mean(y.bebes)-mean(valores.estimados),6)
```

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

- Los errores $(e_i)_{i=1,\dots,n}$ tienen media 0 y varianza $\tilde{s}_e^2 = \frac{SS_E}{n-1}$.

```
errores=y.bebes-valores.estimados
round(mean(errores))
```

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

```
SSE=sum(errores^2)
n=dim(X)[1]
var(errores)-SSE/(n-1)
```

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

8.2.5. Coeficiente de determinación

Al igual que hicimos que la **regresión lineal simple**, vamos a definir el **coeficiente de determinación** que es una manera (no la única como ya hemos comentado) de medir lo efectiva que es la regresión.

Introducimos las **variabilidades** siguientes tal como hicimos en la **regresión lineal simple**:

- **Variabilidad total** o suma total de cuadrados: $SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1)\cdot \tilde{s}_y^2$.
- **Variabilidad de la regresión** o suma de cuadrados de la regresión: $SS_R = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = (n-1)\cdot \tilde{s}_{\widehat{y}}^2$.
- **Variabilidad del error** o suma de cuadrados del error: $SS_E = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = (n-1)\cdot \tilde{s}_e^2$.

Recordemos que tal como pasaba en la **regresión lineal simple**, la **variabilidad total** se puede descomponer como la suma de la **variabilidad de la regresión** y la **variabilidad del error**.

Teorema. En una regresión lineal múltiple usando el método de los mínimos cuadrados, se cumple la siguiente relación entre las **variabilidades**:

$$SS_T = SS_R + SS_E,$$

o equivalentemente,

$$\tilde{s}_y^2 = \tilde{s}_{\hat{y}}^2 + \tilde{s}_e^2.$$

En este caso, tal como comentábamos en la **regresión lineal simple**, cuántas más "próximas" estén las **variabilidades** SS_T y SS_R, o, si se quiere, $\tilde{s}_y^2$ y $\tilde{s}_{\hat{y}}^2$, más efectiva habrá sida la regresión, ya que la regresión habrá heredado mucha variabilidad de los datos y_i, $i = 1, \dots, n$ y la variabilidad del error, SS_E será pequeña.

Definimos el **coeficiente de determinación** en una **regresión lineal múltiple** como: $R^2 = \frac{SS_R}{SS_T} = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2}$, y representa la fracción de la variabilidad de y que queda explicada por la variabilidad del modelo de regresión lineal.

De la misma manera, definimos el **coeficiente de correlación múltiple** de y respecto de $x_1, \dots, x_k$ como $R = \sqrt{R^2}$.

Propiedades del coeficiente de determinación.

El **coeficiente de determinación** verifica las dos primeras propiedades que verificaba el **coeficiente de determinación** en el caso de la **regresión lineal simple**:

- El **coeficiente de determinación** es una cantidad entre 0 y 1: $0 \leq R^2 \leq 1$. Entonces, cuánto más próximo a 1 esté dicho coeficiente, más precisa será la recta de regresión.
- El **coeficiente de determinación** se puede expresar en función de la **variabilidad del error** de la forma siguiente:

$$R^2 = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T} = 1 - \frac{\tilde{s}_e^2}{\tilde{s}_y^2}.$$

Vamos a calcular las variabilidades y el coeficiente de determinación para los datos del ejemplo que estamos trabajando:

- **Variabilidad total**:

```
(SST=sum((y.bebes-mean(y.bebes))^2))
```

```
## [1] 321.2
```

- **Variabilidad de la regresión**:

```
(SSR=sum((valores.estimados-mean(y.bebes))^2))
```

```
## [1] 318.3
```

- **Variabilidad del error**

```
(SSE = sum(errores^2))
```

```
## [1] 2.966
```

Comprobemos que la **variabilidad total** se descompone en la suma de la **variabilidad de la regresión** y la **variabilidad del error**:

```
round(SST-SSR-SSE,6)
```

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

El coeficiente de determinación será:

```
(R2=SSR/SST)
```

```
## [1] 0.9908
```

o, si se quiere,

```
(R2 = var(valores.estimados)/var(y.bebes))
```

```
##         [,1]
## [1,] 0.9908
```

Coeficiente de determinación en R.

Para hallar el **coeficiente de determinación** en R podemos usar las mismas funciones que usábamos en la **regresión lineal simple**:

```
summary(lm(y~x1+...+xk))$r.squared
```

Para los datos de nuestro ejemplo, hemos de hacer lo siguiente:

```
summary(lm(y.bebes~X[,2]+X[,3]+X[,4]+X[,5]))$r.squared
```

```
## [1] 0.9908
```

8.2.6. Coeficiente de determinación ajustado

El **coeficiente de determinación** definido anteriormente aumenta si aumentamos el número de variables independientes k, incluso si éstas aportan información redundante o poca información. Por ejemplo, si añadimos variables que son linealmente dependientes de las demás.

Para evitar este problema o para penalizar el aumento de variables independientes se usa en su lugar el **coeficiente de regresión ajustado**:

$$R^2_{adj} = \frac{MS_T - MS_E}{MS_T},$$

donde $MS_T = \frac{SS_T}{n-1}$, $\quad MS_E = \frac{SS_E}{n-k-1}$.

Fijarse ahora que si aumentamos k, el valor de MS_E aumenta y el valor de R^2_{adj} disminuirá. Por tanto, con el R^2_{adj} penalizamos el aumento de variables independientes.

Si aumentamos el número de variables independientes con variables explicativas que aporten mucha información a la regresión, el valor de SS_E disminuirá haciendo que se mantenga estable el valor de MS_E y el valor de R^2_{adj} no disminuirá.

La relación entre el **coeficiente de determinación ajustado** y el **coeficiente de determinación** es la siguiente:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

Calculemos el **coeficiente de determinación ajustado** para los datos de nuestro ejemplo:

```
k=dim(X)[2]-1
(R2.adj = 1-(1-R2)*(n-1)/(n-k-1))
```

```
##         [,1]
## [1,] 0.9815
```

Observamos que obtenemos un **coeficiente de determinación ajustado** menor que el **coeficiente de determinación**.

En general, lo que hemos observado en el ejemplo anterior, ocurre siempre, es decir $0 \leq R^2_{adj} < R^2 \leq 1$, lo que nos permite concluir que para el **coeficiente de determinación ajustado**, es más difícil obtener un valor cercano a 1 que para el **coeficiente de determinación**.

Coeficiente de determinación ajustado en `R`.

El cálculo del **coeficiente de determinación ajustado** en `R` es parecido al cálculo del **coeficiente de determinación**: sólo hemos de cambiar el parámetro `r.squared` por `adj.r.squared`

```
summary(lm(y~x1+...+xk))$adj.r.squared
```

Para los datos de nuestro ejemplo, hemos de hacer lo siguiente:

```
summary(lm(y.bebes~X[,2]+X[,3]+X[,4]+X[,5]))$adj.r.squared
```

```
## [1] 0.9815
```

8.2.7. Comparación de modelos

Imaginemos que tenemos el modelo

$$Y_{|x_1,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + E_{x_1,\ldots,x_k},$$

y añadimos una nueva variable independiente x_{k+1}. Entonces, tendremos otro modelo:

$$Y_{|x_1,\ldots,x_k,x_{k+1}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} + E_{x_1,\ldots,x_k,x_{k+1}}.$$

¿Cómo podemos comparar los dos modelos anteriores?

O, dicho en otras palabras, ¿cómo decidir si la nueva variable introducida x_{k+1} aporta información relevante?

Una manera de realizar dicha comparación es usando el **coeficiente de regresión ajustado** R^2_{adj}: el modelo que tenga mayor R^2_{adj} es el mejor.

Por tanto, si el valor del **coeficiente de regresión ajustado** del segundo modelo supera el **coeficiente de regresión ajustado** del primero, diremos que la nueva variable x_{k+1} aporta información relevante y hay que tenerla en cuenta.

Con los datos de nuestro ejemplo, consideremos sólo la variable independiente correspondiente a la segunda columna de la matriz **X**, variable que corresponde a la edad del niño en días.

El **coeficiente de correlación ajustado** de este modelo que, por cierto, sería de **regresión lineal simple** vale:

```
summary(lm(y.bebes~X[,2]))$adj.r.squared
```

```
## [1] 0.8823
```

Consideremos ahora las dos primeras variables correspondientes a las columnas segunda y tercera de la matriz **X**, variables que corresponden a la edad del niño en días y su altura en nacer.

El **coeficiente de correlación ajustado** de este modelo vale:

```
summary(lm(y.bebes~X[,2]+X[,3]))$adj.r.squared
```

```
## [1] 0.9617
```

Hemos obtenido un **coeficiente de regresión ajustado** superior. Por tanto, ha valido la pena añadir la variable correspondiente a la altura al nacer del niño ya que nos ha aportado información relevante al modelo.

Consideremos ahora las tres primeras variables correspondientes a las columnas segunda, tercera y cuarta de la matriz **X**, variables que corresponden a la edad del niño en días, su altura en nacer y su peso al nacer.

El **coeficiente de correlación ajustado** de este modelo vale:

```
summary(lm(y.bebes~X[,2]+X[,3]+X[,4]))$adj.r.squared
```

```
## [1] 0.9851
```

Hemos vuelto a obtener un **coeficiente de regresión ajustado** superior al anterior. Por tanto, también ha valido la pena añadir la variable peso del niño al nacer al aportar dicha variable información relevante al modelo.

Recordemos que el **coeficiente de regresión ajustado** del modelo completo era 0.9815, valor inferior al obtenido anteriormente.

Por tanto, la última variable, el aumento del peso actual del niño respecto de su peso al nacer, no aporta información relevante y no debe ser considerada en el modelo de regresión lineal múltiple.

Comparar dos modelos usando el **coeficiente de regresión ajustado** es uno de los métodos que hay en la literatura para ver si un método es más adecuado que otro.

Existen más métodos como son el **AIC** (*Akaike's Information Criterion*) o el método **BIC** (*Bayesian Information Criterion*).

El método **AIC** cuantifica cuánta información de Y se pierde con el modelo y cuántas variables usamos: el mejor modelo es el que tiene un valor de **AIC** más pequeño. Concretamente, se calcula la cantidad siguiente: $AIC = n\ln(SS_E/n) + 2k$ y el modelo con menor **AIC** es el más adecuado.

Para usar el método **AIC** en `R`, hay que usar la función `AIC`:

```
AIC(lm(y~x1+...+xk))
```

Veamos usando el método **AIC** cuál de los cuatro modelos vistos anteriormente para los datos de nuestro ejemplo es el más adecuado:

```
AIC(lm(y.bebes~X[,2]))
```

```
## [1] 43.26
```

```
AIC(lm(y.bebes~X[,2]+X[,3]))
```

```
## [1] 33.76
```

```
AIC(lm(y.bebes~X[,2]+X[,3]+X[,4]))
```

```
## [1] 25.62
```

```
AIC(lm(y.bebes~X[,2]+X[,3]+X[,4]+X[,5]))
```

```
## [1] 27.55
```

El método **AIC** concuerda con el método del **coeficiente de determinación ajustado**. El mejor modelo es el que incluye las variables correspondientes a las columnas 2, 3 y 4 de la matriz **X** que, recordemos, son las variables la edad del niño en días, su altura en nacer y su peso al nacer.

El método **BIC** cuantifica cuánta información de Y se pierde con el modelo y cuántas variables y datos usamos: el mejor modelo es el que tiene un valor de **BIC** más pequeño. Para saber si un modelo es mejor que otro, hay que calcular la cantidad siguiente: $BIC = n\ln(SS_E/n) + k\ln(n)$ y, como ya hemos comentado, el modelo con menor **BIC** es el más adecuado.

Para usar el método **BIC** en `R`, hay que usar la función `BIC`:

```
BIC(lm(y~x1+...+xk))
```

Veamos usando el método **BIC** cuál de los cuatro modelos vistos anteriormente para los datos de nuestro ejemplo es el más adecuado:

```
BIC(lm(y.bebes~X[,2]))
```

```
## [1] 43.85
```

```
BIC(lm(y.bebes~X[,2]+X[,3]))
```

```
## [1] 34.55
```

```
BIC(lm(y.bebes~X[,2]+X[,3]+X[,4]))
```

```
## [1] 26.61
```

```
BIC(lm(y.bebes~X[,2]+X[,3]+X[,4]+X[,5]))
```

```
## [1] 28.73
```

El método **BIC** concuerda con los dos métodos usados anteriormente: el método del **coeficiente de determinación ajustado** y el método **AIC**. El mejor modelo es el que incluye las variables correspondientes a las columnas 2, 3 y 4 de la matriz **X** que, recordemos, son las variables la edad del niño en días, su altura en nacer y su peso al nacer.

8.2.8. Intervalos de confianza

Vamos a calcular los intervalos de confianza para los $k+1$ parámetros de modelo $\beta_0, \beta_1, \dots, \beta_k$.

Para ello, al igual que hicimos en la **regresión lineal simple**, supondremos que las variables aleatorias error $E_i = E_{\underline{x}_i}$ son incorreladas, es decir, la covarianza entre un par de ellas cualesquiera es cero y todas normales de media 0 y de misma varianza σ_E^2.

Antes de dar los intervalos de confianza, necesitamos conocer las propiedades sobre los estimadores de los parámetros anteriores, es decir, si son insesgados, cómo estimar la varianza común σ_E^2 y sus errores estándar.

Dichas propiedades vienen dadas en los teoremas siguientes:

Teorema. Bajo las hipótesis anteriores, los estimadores $b_0, \dots, b_k$ de los parámetros $\beta_0, \dots, \beta_k$ son máximo verosímiles y además no sesgados.

Teorema. Bajo las hipótesis anteriores,

$$\mathrm{Cov}(b_0, b_1, \dots, b_k) = \sigma_E^2 \cdot (\mathbf{X}^\top \cdot \mathbf{X})^{-1},$$

donde $\mathrm{Cov}(b_0, b_1, \dots, b_k)$ es la matriz de covarianzas de los estimadores $b_0, b_1, \dots b_k$ de los parámetros $\beta_0, \beta_1, \dots, \beta_k$ de componentes $\mathrm{Cov}(b_0, b_1, \dots, b_k)_{ij} = \mathrm{Cov}(b_i, b_j)$, $i, j = 0, 1 \dots, k$ y un estimador no sesgado de la varianza común σ_E^2 es $S^2 = \frac{SS_E}{n-k-1}$.

Teorema. Bajo las hipótesis anteriores, el error estándar de cada estimador b_i vale $\sqrt{(\sigma_E^2 \cdot (\mathbf{X}^\top \mathbf{X})^{-1})_{ii}}$, (la raíz cuadrada de la i-ésima entrada de la diagonal de la matriz $\sigma_E^2 \cdot (\mathbf{X}^\top \mathbf{X})^{-1}$ empezando por $i = 0$.)

Teorema. Bajo las hipótesis anteriores,

- la variable aleatoria $\frac{\beta_i - b_i}{\sqrt{(S^2 \cdot (\mathbf{X}^\top \mathbf{X})^{-1})_{ii}}}$, sigue un ley t de Student con $n-k-1$ grados de libertad,
- un intervalo de confianza del $(1-\alpha) \cdot 100\,\%$ de confianza para el parámetro β_i es $b_i \pm t_{n-k-1, 1-\frac{\alpha}{2}} \cdot \sqrt{(S^2 \cdot (\mathbf{X}^\top \mathbf{X})^{-1})_{ii}}$.

Veamos si los errores de los datos del ejemplo que vamos desarrollando se distribuyen normalmente usando el test de **Kolmogorov-Smirnov-Lilliefors**:

```
lillie.test(errores)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  errores
## D = 0.14, p-value = 0.9
```

Como el p-valor obtenido es muy grande, concluimos que no tenemos evidencias suficientes para rechazar la normalidad de los errores.

La estimación de la varianza común S^2 será:

```
(S2 = SSE/(n-k-1))
```

```
## [1] 0.7414
```

La estimación de la matriz de covarianzas de los estimadores b_0, b_1, b_2, b_3, b_4 es la siguiente:

```
S2*solve(t(X)%*%X)
```

```
##          [,1]     [,2]      [,3]       [,4]      [,5]
## [1,] 270.919  5.32456 -12.52088 -13.743055 -1.399872
## [2,]   5.325  0.11540  -0.26585  -0.325518 -0.017627
## [3,] -12.521 -0.26585   0.61764   0.742418  0.041580
## [4,] -13.743 -0.32552   0.74242   1.121860 -0.005976
## [5,]  -1.400 -0.01763   0.04158  -0.005976  0.027710
```

En la matriz anterior, podemos observar que el estimador con más varianza sería b_0 seguido de b_3 que corresponde a la varianza del peso del niño al nacer.

También podemos observar que entre las variables x_2 (altura del niño al nacer) y x_3 (peso del niño al nacer) existe una gran correlación lineal: 0.7424.

Las estimaciones de los errores estándar de los estimadores b_0, b_1, b_2, b_3, b_4 son las siguientes:

```
(errores.estandar = sqrt(S2*diag(solve(t(X)%*%X))))
```

```
## [1] 16.4596  0.3397  0.7859  1.0592  0.1665
```

Un intervalo de confianza para los estimadores b_0, b_1, b_2, b_3, b_4 al 95 % de confianza es el siguiente:

- Intervalo de confianza para b_0 al 95 % de confianza:

```
alpha=0.05
c(estimaciones.b[1]-qt(1-alpha/2,n-k-1)*errores.estandar[1],
  estimaciones.b[1]+qt(1-alpha/2,n-k-1)*errores.estandar[1])
```

```
## [1] -38.55  52.85
```

- Intervalo de confianza para b_1 al 95 % de confianza:

```
c(estimaciones.b[2]-qt(1-alpha/2,n-k-1)*errores.estandar[2],
  estimaciones.b[2]+qt(1-alpha/2,n-k-1)*errores.estandar[2])
```

```
## [1] -0.8431  1.0433
```

- Intervalo de confianza para b_2 al 95 % de confianza:

```
c(estimaciones.b[3]-qt(1-alpha/2,n-k-1)*errores.estandar[3],
  estimaciones.b[3]+qt(1-alpha/2,n-k-1)*errores.estandar[3])
```

```
## [1] -1.456  2.908
```

- Intervalo de confianza para b_3 al 95 % de confianza:

```
c(estimaciones.b[4]-qt(1-alpha/2,n-k-1)*errores.estandar[4],
  estimaciones.b[4]+qt(1-alpha/2,n-k-1)*errores.estandar[4])
```

```
## [1] 0.1351 6.0166
```

- Intervalo de confianza para b_4 al 95 % de confianza:

```
c(estimaciones.b[5]-qt(1-alpha/2,n-k-1)*errores.estandar[5],
  estimaciones.b[5]+qt(1-alpha/2,n-k-1)*errores.estandar[5])
```

```
## [1] -0.4922  0.4321
```

Intervalos de confianza en R.

Recordemos que para hallar los intervalos de confianza en R había que usar la función `confint` aplicado al objeto `lm(...)`. El parámetro `level` nos da el nivel de confianza con valor por defecto 0.95:

```
confint(lm(y~x1+...+xk))
```

Si aplicamos la función anterior a los datos de nuestro ejemplo, obtenemos los intervalos de confianza para los estimadores b_0, b_1, b_2, b_3, b_4 que ya hemos obtenido antes:

```
confint(lm(y.bebes~X[,2]+X[,3]+X[,4]+X[,5]),level=0.95)
```

```
##                 2.5 %  97.5 %
## (Intercept) -38.5517 52.8467
## X[, 2]       -0.8431  1.0433
## X[, 3]       -1.4556  2.9084
## X[, 4]        0.1351  6.0166
## X[, 5]       -0.4922  0.4321
```

Tal como hemos hecho en la **regresión lineal simple**, fijados unos valores concretos de las variables independientes $(x_{10}, \dots, x_{k0})$ podemos considerar dos parámetros más a estudiar: el valor medio de la variable aleatoria $Y|x_{10}, \dots, x_{k0}$, $\mu_{Y|x_{10},\dots,x_{k0}}$ y el valor estimado $y_0 = b_0 + b_1 \cdot x_{10} + \cdots + b_k \cdot x_{k0}$ por la función de regresión.

Recordemos que los dos intervalos de confianza anteriores nos ayudan a interpretar la regresión cuando el valor de las variables $(X_1, \dots, X_k)$ valen un determinado valor $(x_{10}, \dots, x_{k0})$.

Para realizar la estimación puntual de los dos parámetros anteriores usaremos el mismo estimador $\widehat{y}_0 = b_0 + b_1 x_1 + \cdots + b_k x_k$ pero tal como pasaba en la **regresión lineal simple**, los errores estándar no serán los mismos tal como nos dicen el teorema siguiente.

Teorema. Sean $\underline{x}_0 = (x_{01}, \dots, x_{0k})$ unos valores concretos de las variables $(X_1, \dots, X_k)$.

Bajo las hipótesis anteriores,

- el error estándar de $\widehat{y}_0$ como estimador del parámetro $\mu_{Y|\underline{x}_0}$ es el siguiente:

$$S\sqrt{\mathbf{x}_0 \cdot (\mathbf{X}^\top \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_0^\top},$$

- el error estándar de $\widehat{y}_0$ como estimador de y_0 es el siguiente:

$$S\sqrt{1 + \mathbf{x}_0 \cdot (\mathbf{X}^\top \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_0^\top},$$

  donde $\mathbf{x}_0 = (1, \underline{x}_0) = (1, x_{01}, \dots, x_{0k})$.

Los estimadores para hallar los intervalos de confianza para los parámetros $\mu_{Y|x_{10},\dots,x_{k0}}$ y el valor estimado y_0 vienen dados por el teorema siguiente:

Teorema. Sean $\underline{x}_0 = (x_{01}, \dots, x_{0k})$ unos valores concretos de las variables $(X_1, \dots, X_k)$. Bajo nuestras hipótesis, las variables aleatorias $\dfrac{\mu_{Y|\underline{x}_0} - \widehat{y}_0}{S\sqrt{\mathbf{x}_0 \cdot (\mathbf{X}^\top \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_0^\top}}$, $\dfrac{y_0 - \widehat{y}_0}{S\sqrt{1 + \mathbf{x}_0 \cdot (\mathbf{X}^\top \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_0^\top}}$, siguen la ley t de Student con $n-k-1$ grados de libertad.

Por último, bajo las hipótesis anteriores, un intervalo de confianza al $100 \cdot (1-\alpha)\,\%$ de confianza para cada uno de los parámetros $\mu_{Y|x_{10},\dots,x_{k0}}$ y el valor estimado y_0 es el siguiente:

- Parámetro $\mu_{Y|x_{10},\dots,x_{k0}}$: $\widehat{y}_0 \pm t_{n-k-1,1-\frac{\alpha}{2}} \cdot S\sqrt{\mathbf{x}_0 \cdot (\mathbf{X}^\top \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_0^\top}$.
- Parámetro y_0: $\widehat{y}_0 \pm t_{n-k-1,1-\frac{\alpha}{2}} \cdot S\sqrt{1 + \mathbf{x}_0 \cdot (\mathbf{X}^\top \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_0^\top}$.

Hallemos un intervalo de confianza para los datos de nuestro ejemplo suponiendo un niño con $x_{10} = 75$ días de edad, de altura $x_{20} = 50$ cm. al nacer, de $x_{30} = 4$ kg. al nacer y con un $x_{40} = 30\,\%$ de aumento de peso con respecto su peso al nacer.

El intervalo de confianza para el parámetro $\mu_{Y|x_{10}=75,x_{20}=50,x_{30}=4,x_{40}=30}$ al $95\,\%$ de confianza es el siguiente:

```
alpha=0.05
x0 = c(1,75,50,4,30)
y0.estimado = sum(estimaciones.b*x0)
c(y0.estimado-qt(1-alpha/2,n-k-1)*sqrt(S2*(t(x0)%*%solve(t(X)%*%X)%*%x0)),
  y0.estimado+qt(1-alpha/2,n-k-1)*sqrt(S2*(t(x0)%*%solve(t(X)%*%X)%*%x0)))
```

```
## [1] 52.99 71.77
```

El intervalo de confianza para el parámetro y_0 al $95\,\%$ de confianza es el siguiente:

```
c(y0.estimado-qt(1-alpha/2,n-k-1)*sqrt(S2*(1+(t(x0)%*%solve(t(X)%*%X)%*%x0))),
  y0.estimado+qt(1-alpha/2,n-k-1)*sqrt(S2*(1+(t(x0)%*%solve(t(X)%*%X)%*%x0))))
```

```
## [1] 52.69 72.07
```

Intervalos de confianza en R.

Para hallar el intervalo de confianza para el parámetros $\mu_{Y|x_{10},\dots,x_{k0}}$ al $100 \cdot (1-\alpha)\,\%$ de confianza hay que usar la función `predict.lm` de la forma siguiente:

```
newdata=data.frame(x1=x10,...,xk=xk0)
predict.lm(lm(y~x1+...+xk),newdata,interval="confidence",level= nivel.confianza)
```

Para el parámetro y_0 hay que usar la misma función anterior pero cambiando el parámetro `interval` al valor `prediction`:

```
newdata=data.frame(x1=x10,...,xk=xk0)
predict.lm(lm(y~x1+...+xk),newdata,interval="prediction",level= nivel.confianza)
```

Para los datos del ejemplo anterior, los intervalos de confianza para los parámetros $\mu_{Y|x_{10}=75,x_{20}=50,x_{30}=4,x_{40}=30}$ y y_0 al $95\,\%$ de confianza se calcularían en R de la forma siguiente:

```
newdata=data.frame(x1=75,x2=50,x3=4,x4=30)
x1=X[,2]
```

```
x2=X[,3]
x3=X[,4]
x4=X[,5]
predict.lm(lm(y.bebes~x1+x2+x3+x4),newdata,interval="confidence",level= 0.95)
```

```
##     fit   lwr   upr
## 1 62.38 52.99 71.77
```

```
predict.lm(lm(y.bebes~x1+x2+x3+x4),newdata,interval="prediction",level= 0.95)
```

```
##     fit   lwr   upr
## 1 62.38 52.69 72.07
```

8.2.9. Contrastes de hipótesis sobre los parámetros β_i

En el caso de la **regresión lineal múltiple** podemos plantearnos el contraste de hipótesis siguiente:

$$\left.\begin{array}{l} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0, \\ H_1 : \text{existe algún } i \text{ tal que } \beta_i \neq 0. \end{array}\right\}$$

Es decir, queremos contrastar si la **regresión lineal múltiple** realizada ha tenido sentido ya que la hipótesis nula equivaldría a decir que ninguna variable independiente X_i, $i = 1, \dots, k$ ha tenido efecto sobre la variable Y y como consecuencia, la regresión no ha tenido sentido.

Para realizar el contraste de hipótesis anterior, podríamos plantear los k contrastes siguientes:

$$\left.\begin{array}{l} H_0 : \beta_i = 0, \\ H_1 : \beta_i \neq 0, \end{array}\right\}$$

usando como **estadístico de contraste** $\frac{\beta_i - b_i}{\sqrt{(S^2 \cdot (\mathbf{X}^\top \mathbf{X})^{-1})_{ii}}}$, que sabemos que sigue una ley t de Student con $n - k - 1$ grados de libertad.

El problema es que los k contrastes anteriores no son **independientes**, es decir, los **estadísticos de contraste** son variables aleatorias **no independientes** entre otras cosas porque su matriz de covarianzas no tiene por qué ser diagonal.

Por tanto, si realizamos el contraste de hipótesis original a partir de los k contrastes anteriores, es muy complicado hallar el **error tipo I** α del contraste original a partir de los **errores tipo I** de cada uno de los k contrastes al fallar la independencia de los mismos.

Otra posibilidad para evitar el problema comentado anteriormente es plantear el contraste original como un contraste **ANOVA** donde las subpoblaciones consideradas serían las variables $Y|\underline{x}_1, \dots, Y|\underline{x}_n$ siendo $\underline{x}_1, \dots, \underline{x}_n$, n valores concretos de las variables independientes $(X_1, \dots, X_k)$.

Fijarse que si la hipótesis nula H_0 es cierta, es decir $\beta_1 = \cdots = \beta_k = 0$, entonces los valores medios de las variables anteriores serían los mismos ya que recordemos que suponemos que:

$$\mu_{Y|\underline{x}} = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k = \beta_0,$$

para todo valor de $\underline{x}$.

Por tanto, el contraste original sería equivalente a realizar el contraste **ANOVA** siguiente:

$$\left.\begin{array}{l} H_0 : \mu_{Y|\underline{x}_1} = \cdots = \mu_{Y|\underline{x}_n}, \\ H_1 : \text{existen } i \text{ y } j \text{ tales que } \mu_{Y|\underline{x}_i} \neq \mu_{Y|\underline{x}_j}. \end{array}\right\}$$

Para que el modelo de **regresión lineal múltiple** tenga sentido hemos de rechazar la hipótesis nula anterior.

La **tabla ANOVA** es la siguiente:

Origen variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios	Estadístico de contraste	p-valor
Regresión	k	SS_R	$MS_R = \frac{SS_R}{k}$	$F = \frac{MS_R}{MS_E}$	p-valor
Error	$n-k-1$	SS_E	$MS_E = \frac{SS_E}{n-k-1}$		

donde el **estadístico de contraste** $F = \frac{MS_R}{MS_E}$, suponiendo la hipótesis nula cierta, sigue la distribución F de k y $n-k-1$ grados de libertad.

El p-valor del contraste anterior vale $p = P(F_{k,n-k-1} \geq f_0)$, siendo f_0 el valor del estadístico de contraste F para nuestros datos.

Tabla ANOVA en `R`.

Para calcular la tabla ANOVA en `R` de una **regresión lineal múltiple** hay que usar la función `anova` de la forma siguiente:

```
anova(lm(y ~ Xd))
```

donde `Xd` es una matriz cuyas columnas son los valores de las variables independientes $x_1, \ldots, x_k$.

Para hallar la tabla ANOVA para los datos de nuestro ejemplo, hemos de hacer lo siguiente:

```
anova(lm(y.bebes~X[,2:5]))
```

```
## Analysis of Variance Table
##
## Response: y.bebes
##           Df Sum Sq Mean Sq F value  Pr(>F)
## X[, 2:5]   4    318    79.6     107 0.00025 ***
## Residuals  4      3     0.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que el p-valor es muy pequeño. Por tanto, rechazamos la hipótesis nula y concluimos que la regresión ha tenido sentido en este caso.

Otro tipo de contrastes que podemos plantearnos sobre los parámetros del modelo β_i es, si fijado i, β_i aporta algo al modelo, o, si la variable x_i es **significativa**.

Concretamente, fijado i, nos planteamos el contraste siguiente:

$$\left.\begin{array}{l} H_0 : \beta_i = 0, \\ H_1 : \beta_i \neq 0, \end{array}\right\}$$

usando como **estadístico de contraste** $\frac{\beta_i - b_i}{\sqrt{(S^2 \cdot (\mathbf{X}^\top \mathbf{X})^{-1})_{ii}}}$, que sabemos que sigue una ley t de Student con $n - k - 1$ grados de libertad.

El p-valor del contraste anterior vale $p = 2 \cdot P(t_{n-k-1} > |t_0$, donde t_0 es el valor obtenido por el estadístico de contraste usando nuestros datos.

Para que la variable x_i sea **significativa** o para que aporte información relevante al modelo de **regresión lineal múltiple**, debemos rechazar la hipótesis nula en el contraste anterior u obtener un p-valor pequeño.

Contrastes de hipótesis sobre los parámetros β_i en `R`.

Para realizar todos los contrastes anteriores en `R` hay que usar la función `summary` aplicada al objecto `lm(...)`:

```
summary(lm(y~x1+...+xk))
```

y `R` nos da toda la información sobre la **regresión múltiple** realizada.

Si aplicamos la función `summary` a los datos de nuestro ejemplo, obtenemos la salida siguiente:

```
summary(lm(y.bebes~x1+x2+x3+x4))
```

```
##
## Call:
## lm(formula = y.bebes ~ x1 + x2 + x3 + x4)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9
## -0.0405 -0.1290  0.2150 -0.4324 -0.6461  0.2214  0.5225  1.1276 -0.8385
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.147     16.460    0.43    0.687
## x1             0.100      0.340    0.29    0.783
## x2             0.726      0.786    0.92    0.408
## x3             3.076      1.059    2.90    0.044 *
## x4            -0.030      0.167   -0.18    0.866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.861 on 4 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.982
## F-statistic:  107 on 4 and 4 DF,  p-value: 0.000254
```

En primer lugar `R` nos da la lista de los errores cometidos en las estimaciones.

En segundo lugar, `R` nos muestra una tabla con las columnas siguientes:

- `Estimate`: los valores estimados de los parámetros $\beta_0, \beta_1, \beta_2, \beta_3$ y β_4. Es decir, los valores b_0, b_1, b_2, b_3 y b_4.
- `Std. Error`: los errores estándares de los estimadores b_0, b_1, b_2, b_3 y b_4.
- `t value`: el valor del estadístico de contraste cuando realizamos el contraste $\left.\begin{array}{l} H_0 : \beta_i = 0, \\ H_1 : \beta_i \neq 0, \end{array}\right\}$ sobre cada parámetro β_i, $i = 0, 1, 2, 3, 4$.
- `Pr(>|t|)`: los p-valores del contraste anterior.

Observamos que todas las variables excepto la variable `x3`, peso del niño al nacer, son no significativas para el modelo.

A continuación nos da el error residual que es la estimación de $\sqrt{S^2}$, el valor de coeficiente de determinación R^2 y el coeficiente de determinación ajustado R^2_{adj}.

Para finalizar nos da el valor del estadístico de contraste ANOVA comentado anteriormente junto con el p-valor del mismo.

8.2.10. Diagnósticos. Estudio de los residuos

Para que el **modelo de regresión lineal** tanto **simple** como **múltiple** sea fiable en las conclusiones derivadas de las estimaciones e inferencias (intervalos de confianzas, contrastes de hipótesis) que realizamos a partir de dicho modelo, se tienen que verificar unas hipótesis.

Las tareas que realizan dichas verificaciones se denominan **diagnósticos de regresión**.

Los **diagnósticos de regresión** se clasifican en tres categorías:

- *Errores*: los errores tienen que seguir una $N(0, \sigma)$, con la misma varianza, y ser incorrelados.
- *Modelo*: los puntos se tienen que ajustar a la estructura lineal considerada.
- *Observaciones anómalas*: a veces unas cuántas observaciones no se ajustan al modelo y hay que detectarlas.

8.2.11. Tipos de diagnósticos de regresión

Hay dos métodos usados en los **diagnósticos de regresión**:

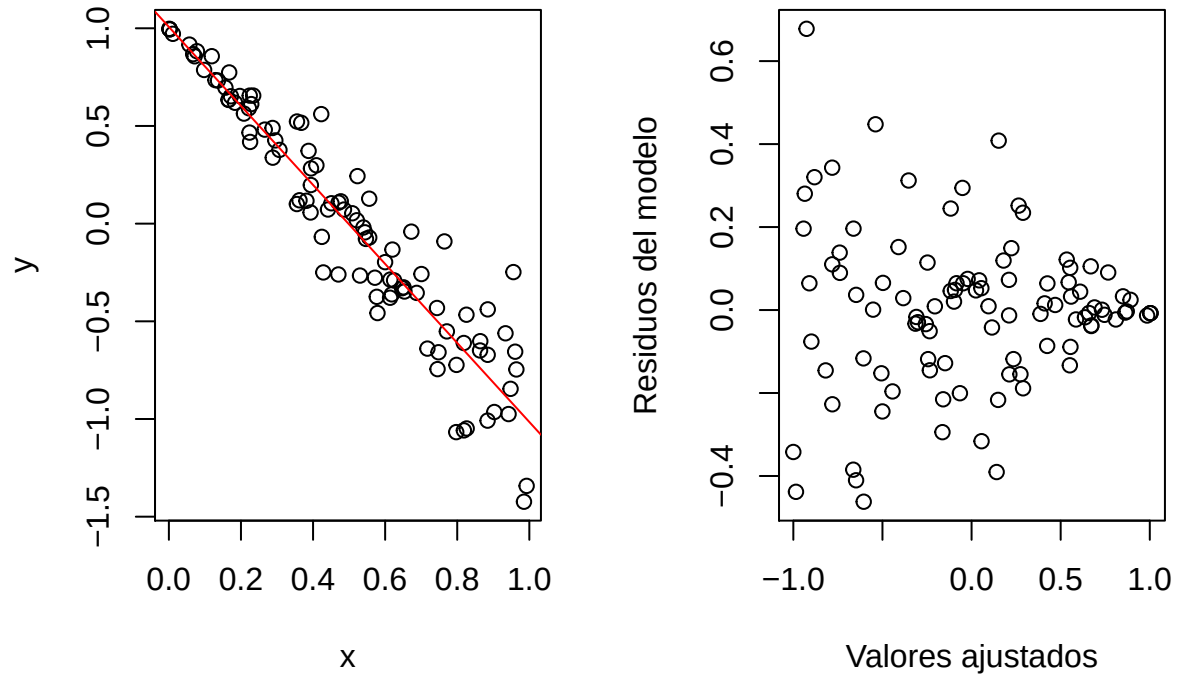
- Métodos gráficos: son métodos muy flexibles pero difíciles de interpretar.
- Métodos numéricos: son métodos de utilidad más limitada con respecto a los métodos gráficos pero con interpretación inmediata.

8.2.12. Distribución de los errores

Uno de los problemas que puede sufrir nuestro modelo es que la varianza de los residuos no sea constante. Veamos uno ejemplo

```
set.seed(2020)
x<-runif(100)
y<-1-2*x+0.3*x*rnorm(100)
```

```
par(mfrow=c(1,2))
plot(x,y)
rlm=lm(y~x)
abline(rlm,col="red")
plot(rlm$res~rlm$fitted.values,xlab="Valores ajustados",ylab="Residuos del modelo")
```



Como podemos observar, hemos realizado una regresión simple de 100 puntos cuyas coordenadas x son valores aleatorios distribuidos uniformemente en el intervalo $(0, 1)$ y cuyas coordenadas y son valores ajustados a la recta $y = 1 - 2x$ añadiendo un "ruido" normal estándar amortiguado con un coeficiente de 0.3.

El gráfico de la izquierda muestra los 100 puntos junto con la correspondiente recta de regresión en color rojo.

El gráfico de la derecha muestra la distribución de los pares de los errores del modelo vs. los valores ajustados, $(\widehat{y}_i, e_i)$, $i = 1, \ldots, n = 100$. Si el modelo fuese **homocedástico**, es decir, que la varianza de los residuos fuese la misma para cualquier valor x, observaríamos una distribución de puntos uniforme, o lo que coloquialmente se llama un "cielo estrellado".

En cambio, observamos una distribución "triangular" o "en forma de cuña" donde a medida que aumenta el valor ajustado de los puntos, disminuye la dispersión o la variación de los errores, hecho que nos detecta que dicho modelo es anómalo en el sentido de no ser **homocedástico**.

El método gráfico anterior, es decir, el gráfico de los errores e_i en función de los valores estimados $\widehat{y}_i$ es un método para "observar" gráficamente si el modelo es **homocedástico**.

Existe un método numérico para detectar la **homocedasticidad** del modelo llamado **Test de White**.

8.2.13. Test de White

Para aplicar el **Test de White**, hay que seguir los pasos siguientes:

- Obtener los $\{e_i\}_{i=1,\ldots,n}$ residuos de la regresión lineal inicial.
- Calcular el coeficiente de determinación R^2 de la regresión lineal de los e_i^2 respecto de las variables iniciales, sus cuadrados y los productos cruzados dos a dos. Es decir, calcular el coeficiente de determinación del modelo de regresión siguiente:

$$\begin{aligned}\mu_{E^2|x_1,\ldots,x_k,x_1^2,\ldots,x_k^2,x_ix_j(i,j=1,\ldots,n,i<j)} = \quad & \beta_0+\beta_1x_1+\cdots+\beta_kx_k+\\ & \beta_1^{(2)}x_1^2+\cdots+\beta_k^{(2)}x_k^2+\\ & \beta_{12}x_1x_2+\cdots+\beta_{k-1,k}x_{k-1}x_k.\end{aligned}$$

- Calcular el **estadístico** $X_0 = nR^2$, el cual, suponiendo que la varianza es constante, sigue una χ_q^2, donde q es el número de variables independientes de la regresión del paso anterior.
- Calculamos el p-valor $P(\chi_q^2 \geq X_0)$ con el significado usual.

Ejemplo

Apliquemos el **Test de White** a los datos del ejemplo anterior:

```
residuos=rlm$res
(X0=length(residuos)*summary(lm(residuos^2~x+I(x^2)))$r.squared)
```

```
## [1] 21.17
```

```
(p.valor=pchisq(X0,2,lower.tail = FALSE))
```

```
## [1] 0.00002536
```

Obtenemos un valor muy pequeño. Concluimos consecuentemente que tenemos evidencias suficientes para rechazar que las varianzas de los residuos no son iguales para cualquier valor de x o, dicho en otras palabras, el modelo no es **homocedástico**, es **heterocedástico**.

Test de White en `R`.

Para usar el **Test de White** en `R`, tenemos que usar la función `bptest` del paquete `lmtest`.

```
library(lmtest)
bptest(lm1, ~ X + I(X^2))
```

donde `lm1` es el objeto de `R` donde hemos guardado la información de la regresión original y `X` es la matriz que contiene los valores de la muestra de las variables independientes.

Si realizamos el **Test de White** para los datos del ejemplo anterior en `R` obtenemos lo siguiente:

```
library(lmtest)
#bptest(rlm,~x+I(x^2)) #TODO: error, revisar
```

Ejemplo de los bebés

Recordemos que en este ejemplo teníamos $k = 4$ variables independientes:

- x_1: la edad del bebé en días.

- x_2: su altura al nacer.
- x_3: su peso al nacer.
- x_4: su aumento en tanto por ciento de su peso actual respecto de su peso al nacer.

La tabla de datos constaba de $n = 9$ bebés.

En este caso, realizar el test de White no tendría sentido ya que tendríamos más variables independientes que valores en la muestra en la regresión auxiliar de e_i^2.

Las variables independientes en dichas regresión auxiliar tal como hemos explicado serían: $x_1, x_2, x_3, x_4, x_1^2, x_2^2, x_3^2, x_4^2, x_1 \cdot x_2, x_1 \cdot x_3, x_1 \cdot x_4, x_2 \cdot x_3, x_2 \cdot x_4, x_3 \cdot x_4$.

Por tanto, tendríamos 14 variables independientes pero sólo $n = 9$ valores de e_i^2. Se violaría la hipótesis en la regresión lineal múltiple de que el número de variables independientes debe ser menor que el número de valores en la muestra.

8.2.14. Test de Breusch-Pagan

Cuando nos encontramos en una situación como la del ejemplo de los bebés, en lugar de aplicar el **Test de White** para contrastar la **homocedasticidad** de los residuos, podemos aplicar el **Test de Breuch-Pagan**.

Su aplicación es bastante parecida al **Test de White** pero evita los términos de segundo orden en la regresión auxiliar.

Para aplicar el **Test de Breusch-Pagan**, hay que seguir los pasos siguientes:

- Obtener los $\{e_i\}_{i=1,\dots,n}$ residuos de la regresión lineal inicial.
- Calcular el coeficiente de determinación R^2 de la regresión lineal de los e_i^2 respecto de las variables iniciales. Es decir, calcular el coeficiente de determinación del modelo de regresión siguiente:

$$\mu_{E^2|x_1,\dots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

- Calcular el **estadístico** $X_0 = nR^2$, el cual, suponiendo que la varianza es constante, sigue una χ_k^2.
- Calculamos el p-valor $P(\chi_k^2 \geq X_0)$ con el significado usual.

Ejemplo de los bebés

Apliquemos el **Test de Breusch-Pagan** al ejemplo de los bebés.

Los pasos a seguir son:

- Los residuos e_i, $i = 1, \dots, n = 9$ ya se habían calculado anteriormente

```
errores
```

```
##           [,1]
##  [1,] -0.04053
##  [2,] -0.12898
##  [3,]  0.21498
##  [4,] -0.43238
```

```
##  [5,] -0.64610
##  [6,]  0.22143
##  [7,]  0.52245
##  [8,]  1.12764
##  [9,] -0.83852
```

- Calculamos el coeficiente de determinación del modelo de regresión auxiliar:

```
(coef.deter.modelo.auxiliar = summary(lm(errores^2 ~ X[,2]+X[,3]+X[,4]+X[,5]))$r.squared)
```

```
## [1] 0.6176
```

- Calculamos el valor del estadístico de contraste $X_0 = nR^2$ y el p-valor correspondiente:

```
(X0 = n*coef.deter.modelo.auxiliar)
```

```
## [1] 5.558
```

```
(p.valor = pchisq(X0,k,lower.tail=FALSE))
```

```
## [1] 0.2347
```

Hemos obtenido un p-valor bastante grande. Concluimos consecuentemente que no tenemos indicios suficientes para rechazar la **homocedasticidad** de los errores en este caso.

Test de Breusch-Pagan en `R`.

Para usar el **Test de Breusch-Pagan** en `R`, tenemos que usar la misma función `bptest` del paquete `lmtest`.

```
library(lmtest)
bptest(lm1)
```

donde `lm1` es el objeto de `R` donde hemos guardado la información de la regresión original.

Ejemplo de los bebés

Si aplicamos el **Test de Breusch-Pagan** en `R` en este ejemplo obtenemos:

```
y.bebes=c(57.5,52.8,61.3,67,53.5,62.7,56.2,68.5,69.2)
reg.mul.original = lm(y.bebes~X[,2]+X[,3]+X[,4]+X[,5])
bptest(reg.mul.original)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg.mul.original
## BP = 5.6, df = 4, p-value = 0.2
```

obteniendo los mismos resultados que hemos hallado anteriormente realizando los cálculos "a mano".

8.2.15. Normalidad de los residuos

Para detectar la normalidad de los residuos, podemos aplicar todos los **tests de normalidad** vistos en el tema de **Bondad de Ajuste**:

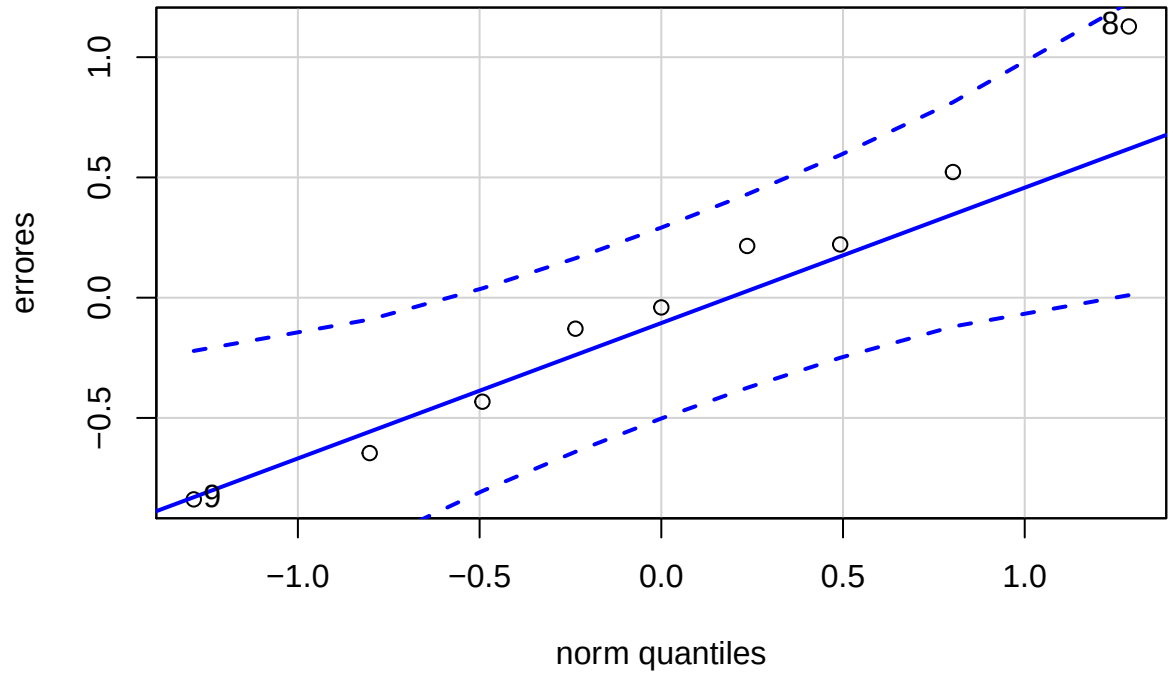
- Test de Kolmogorov-Smirnov. Usar como parámetros la media y de la desviación típica los valores 0 y el valor de la estimación de σ: $\sqrt{S^2}$.
- Test de Kolmogorov-Smirnov-Lilliefors.
- Test de normalidad de Anderson-Darling.
- Test de Shapiro-Wilks (S-W).
- Test omnibus de D'Agostino-Pearson.

Es aconsejable también realizar las pruebas gráficas de normalidad vistos también en el tema de **Bondad de ajuste** como los **Q-Q-plots**.

Ejemplo de los bebés

Realicemos un **Q-Q-plot** de los residuos en el ejemplo considerado:

```
library(car)
estimación.sigma2 = sum(errores^2)/(n-k-1)
qqPlot(errores,distribution = "norm", mean=0,sd=sqrt(estimación.sigma2))
```




```
## [1] 8 9
```

Vemos que los $n = 9$ valores de la muestra están dentro de las bandas de confianza, lo que nos permite concluir que gráficamente se observa que los residuos parece que siguen la normalidad.

Recordemos que ya hemos aplicado el test de **Kolmogorov-Smirnov-Lilliefors** a los residuos, obteniendo:

```
lillie.test(errores)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  errores
## D = 0.14, p-value = 0.9
```

un p-valor muy grande, lo que nos permite concluir también que no tenemos indicios suficientes para rechazar la normalidad de los residuos en este caso.

8.2.16. Correlación de los residuos: Test de Durbin-Watson

Otra de las hipótesis que se deben verificar para que el análisis de regresión sea correcto es la incorrelación de los residuos.

La autocorrelación de los residuos puede ser de dos tipos:

- Autocorrelación positiva: un valor positivo (negativo) de un error genera una sucesión de residuos positivos (negativos).
- Autocorrelación negativa: los residuos van alternando de signo.

Para comprobar si se satisface que los residuos no presenten correlación, se puede aplicar el **Test de Durbin-Watson**.

Expliquemos cómo aplicar dicho test.

Sean $\{e_i\}_{i=1,\dots,n}$ los residuos de la regresión.

Sean E_i y las E_{i-1} variables aleatorias error (trasladadas en un índice) y la recta de regresión de E_i con respecto a E_{i-1}: $E_i = \beta_1 E_{i-1} + \beta_0$.

Se plantea el siguiente contraste:

$$\left.\begin{array}{l} H_0 : \beta_1 = 0, \\ H_1 : \beta_1 \neq 0, \end{array}\right\}$$

con el siguiente **estadístico de contraste**: $d = \dfrac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$.

El valor de este estadístico es aproximadamente $2(1 - b_1)$ donde b_1 se una estimación de β_1.

Si la hipótesis nula H_0 es cierta, su distribución es la de una cierta combinación lineal de χ^2.

El test necesita de una tabla de valores críticos para tomar la decisión final.
Dicha tabla puede encontrarse en Images Google y escribiendo `test de Durbin Watson` en la casilla de búsqueda.

Concretamente, d se tiene que comparar con dos valoras críticos $d_{L,\alpha}$ y $d_{U,\alpha}$, donde α es el nivel de significación que depende de n y de k.

Decidimos si hay autocorrelación positiva:

- Si $d < d_{L,\alpha}$, hay autocorrelación positiva.
- Si $d > d_{U,\alpha}$, no hay autocorrelación positiva.
- De lo contrario, nos encontramos en la zona de penumbra.

Decidimos si hay autocorrelación negativa:

- Si $4 - d < d_{L,\alpha}$, hay autocorrelación negativa.

- Si $4 - d > d_{U,\alpha}$, no hay autocorrelación negativa.
- De lo contrario, nos encontramos en la zona de penumbra.

Ejemplo de los bebés

Vamos a ver si hay autocorrelación para la tabla de datos de los bebés.

El valor del estadístico de contraste d valdrá en este caso:

```
diferencias = errores[2:n]-errores[1:(n-1)]
(estadístico.d = sum(diferencias^2)/sum(errores^2))
```

```
## [1] 1.911
```

Si miramos los valores críticos para $\alpha = 0{,}05$, $n = 9$ y $k = 4$ en la Tabla del estadístico de Durbin-Watson, obtenemos los valores siguientes: $d_{L,0,05} = 0{,}3$ y $d_{U,0,05} = 2{,}59$.

A continuación, miramos si hay autocorrelación positiva: como 1,9106 está entre $d_{L,0,05}$ y $d_{U,0,05}$, estamos en la zona de penumbra y no podemos tomar una decisión clara.

Finalmente, testeamos si hay autocorrelación negativa. El valor de $4 - d$ será: 2.0894. Observamos también que $4 - d$ está entre $d_{L,0,05}$ y $d_{U,0,05}$, por tanto, volvemos a estar estamos en la zona de penumbra y no podemos tomar una decisión clara.

En resumen, en este ejemplo no podemos decidir de forma clara a partir del estadístico de Durbin-Watson si los errores están incorrelados .

Test de Durbin-Watson en `R`.

El test de Durbin-Watson está implementado en la función `dwtest` del paquete `lmtest`.

El parámetro `alternative` nos indica si se testea que la autocorrelación es positiva (`greater`) o negativa (`less`):

```
dwtest(r,alternative=...)
```

donde en `r` se ha guardado el objecto de la regresión `lm(y~x1+...+xk)` y, como hemos indicado, con `alternative=greater` testeamos si los residuos tienen autocorrelación positiva y con `alternative=less`, si tienen autocorrelación negativa.

Ejemplo de los bebés

Para testear si los errores tienen autocorrelación positiva, hacemos lo siguiente:

```
library(lmtest)
dwtest(reg.mul.original,alternative = 'greater')
```

```
##
##  Durbin-Watson test
##
## data:  reg.mul.original
## DW = 1.9, p-value = 0.3
## alternative hypothesis: true autocorrelation is greater than 0
```

Observamos que obtenemos el mismo valor del estadístico de contraste que en los cálculos realizados "a mano".

Sin embargo, R nos da un p-valor del que podemos concluir que no hay autocorrelación positiva en los residuos. Recordemos que en los cálculos realizados "a mano", no podíamos tomar una decisión clara.

Esto es debido a que R es capaz de hallar los p-valores a partir de la matriz de los datos **X**. Es decir, R usa información de nuestra muestra para hallar los p-valores. Por tanto, si os encontráis una discrepancia de este tipo, la conclusión que os dé R prevalece sobre los cálculos realizados a partir de la tabla del **Test de Durbin-Watson**.

Si testeamos si los errores tienen autocorrelación negativa, obtenemos:

```
dwtest(reg.mul.original,alternative = 'less')
```

```
##
##  Durbin-Watson test
##
## data:  reg.mul.original
## DW = 1.9, p-value = 0.7
## alternative hypothesis: true autocorrelation is less than 0
```

Obtenemos un p-valor grande, concluyendo a partir de R que no tenemos indicios suficientes para rechazar que los errores no tengan autocorrelación negativa. Dicha conclusión prevalecería sobre la conclusión que hemos realizado antes haciendo los cálculos "a mano".

8.2.17. Aditividad y linealidad

Cuando se plantea un modelo lineal, se supone implícitamente las condiciones siguientes:

- Aditividad: para cada variable independiente X_i, la variación de asociada $\mu_{Y|x_1,\dots,x_k}$ con un aumento en X_i (manteniendo las otras variables constantes) es la misma sean cuales sean los valores de las otras variables independientes.

Dicho de otra forma:

$$\begin{aligned}\mu_{Y|x_1,\dots,x_i+\Delta x_i,\dots,x_k} - \quad & \mu_{Y|x_1,\dots,x_i,\dots,x_k} = \\ & \beta_0+\beta_1x_1+\cdots+\beta_i(x_i+\Delta x_i)+\cdots+\beta_kx_k - \\ & (\beta_0+\beta_1x_1+\cdots+\beta_ix_i+\cdots+\beta_kx_k) = \beta_i\Delta x_i.\end{aligned}$$

Fijémonos que la variación en el parámetro $\mu_{\dots}$ es independiente de los valores de $x_1,\dots,x_{i-1},x_{i+1},\dots,x_k$.

- Linealidad: para cada variable independiente X_i, la variación de asociada $\mu_{Y|x_1,\dots,x_k}$ con un aumento en X_i (manteniendo las otras variables constantes) es la misma sea cual sea el valor de X_i.

Dicho de otra forma, si miramos la expresión anterior, la variación en el parámetro $\mu_{\dots}$ es independiente de los valores de x_i.

8.2.18. Aditividad: test de Tukey

Podemos comprobar la **aditividad** con el **test de Tukey**, usando los llamados **gráficos de residuos parciales para la linealidad**.

La idea principal es verificar que no haya **interacción** entre las variables independientes y así, cada una tendrá un efecto aditivo en el modelo.

Si existe la **interacción**, algunos términos cuadráticos tendrán peso en el modelo. Esta es la base del **Test de Tukey**:

- Se obtienen los valores ajustados $\{\widehat{y}_i\}$ por la regresión lineal inicial.
- Se lleva a cabo una segunda regresión lineal incluyendo como nueva variable independiente los $\widehat{y}_i^2$. Sea β el coeficiente de esta nueva variable.
- Se testea si la variable $\widehat{y}^2$ es significativa en la segunda regresión. Es decir, se realiza el contraste

$$\left.\begin{array}{l} H_0 : \beta = 0, \\ H_1 : \beta \neq 0. \end{array}\right\}$$

  Si no podemos descartar la hipótesis nula, la variable de los $\widehat{y}_i^2$ no es significativa y el modelo es aditivo.

Ejemplo de los bebés

Testeemos la aditividad para los datos del ejemplo de los bebés:

- Valores ajustados:

```
reg.mul.original$fitted.values
```

```
##     1     2     3     4     5     6     7     8     9
## 57.54 52.93 61.09 67.43 54.15 62.48 55.68 67.37 70.04
```

- Segunda regresión lineal incluyendo los valores $\widehat{y}_i^2$:

```
valores.ajustados2 = reg.mul.original$fitted.values^2
summary(lm(y.bebes~X[,2]+X[,3]+X[,4]+X[,5]+valores.ajustados2))[[4]]
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -35.0865   35.50437 -0.9882   0.3959
## X[, 2]               0.4581    0.41441  1.1055   0.3496
## X[, 3]               1.9176    1.15881  1.6548   0.1965
## X[, 4]               8.8188    4.47436  1.9710   0.1433
## X[, 5]              -0.1208    0.16794 -0.7195   0.5238
## valores.ajustados2  -0.0168    0.01277 -1.3151   0.2800
```

Vemos que la variable $\widehat{y}_i^2$ no es significativa del modelo con un p-valor de 0.28. Concluimos por tanto que no tenemos evidencias suficientes para rechazar la aditividad del modelo.

Test de Tukey en `R`.

Para realizar el **Test de Tukey** en `R` hay que usar la función `residualPlots` del paquete `car`:

```
residualPlots(r,plot=...)
```

donde `r` es el objeto de `R` donde hemos guardado la información de la regresión original y `plot` es un parámetro que si vale `TRUE` (valor por defecto) nos dibuja los gráficos de los residuos frente a las variables regresoras y frente a los valores estimados junto con una curva de color azul indicando su tendencia y si vale `FALSE` simplemente da el valor del estadístico de Tukey y su p-valor.

Si optamos por ver los gráficos, no debe mostrarse ningún tipo de estructura, todos ellos deben parecer "cielos estrellados".

Ejemplo de los bebés

Apliquemos el **Test de Tukey** para los datos del ejemplo de los bebés:

```
library(car)
residualPlots(reg.mul.original,plot=FALSE)
```

```
##            Test stat Pr(>|Test stat|)
## X[, 2]         -1.73             0.18
## X[, 3]         -2.24             0.11
## X[, 4]         -0.14             0.90
## X[, 5]         -1.04             0.37
## Tukey test     -1.32             0.19
```

Obtenemos un p-valor grande, llegando a la misma conclusión que los cálculos realizados "a mano".

8.2.19. Linealidad: gráficos de residuos parciales

Los gráficos de residuos parciales son una herramienta útil para detectar la no linealidad en una regresión.

Se definen los **residuos parciales** e_{ij} para la variable independiente X_j como

$$e_{ij} = e_i + b_j x_{ij},$$

donde e_i es el residuo i-ésimo de la regresión lineal, b_j es el coeficiente de X_j en la regresión original y x_{ij} es la observación j-ésima del individuo i-ésimo.

Los residuos parciales se dibujan contra los valores de x_j y se hace su recta de regresión.

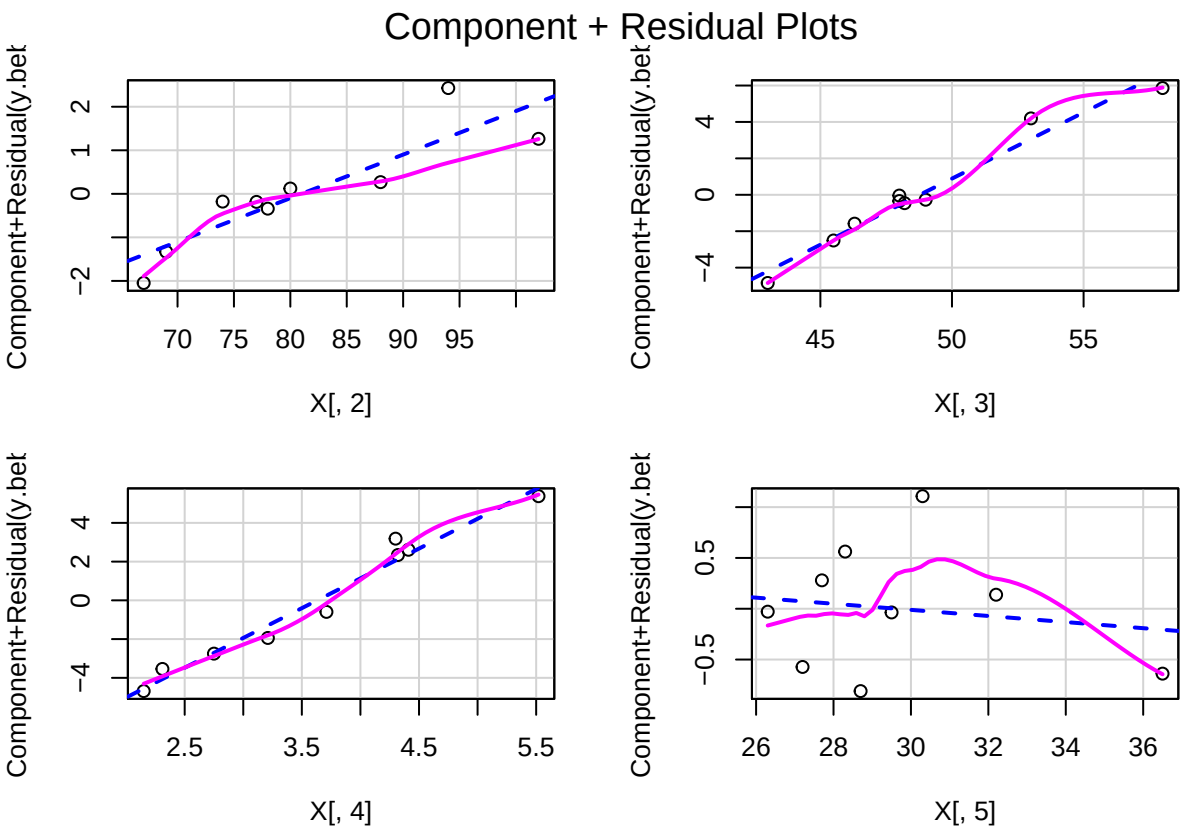
Si ésta no se ajusta a la curva dada por una regresión no paramétrica suave (las variables independientes no están predeterminadas y se construyen con los datos), el modelo no es lineal.

La función de `R` para representar estos gráficos es `crPlots` del paquete `car`.

Ejemplo de los bebés

Realicemos los **gráficos de residuos parciales** para los datos del ejemplo de los bebés para ver gráficamente la linealidad:

```
library(car)
crPlots(reg.mul.original)
```



Observamos que la única variable que no se ajusta el modelo lineal es la variable x_4: aumento en tanto por ciento de su peso actual respecto de su peso al nacer.

Todas las demás presentan un ajuste bastante aceptable al modelo lineal.

8.2.20. Observaciones anómalas

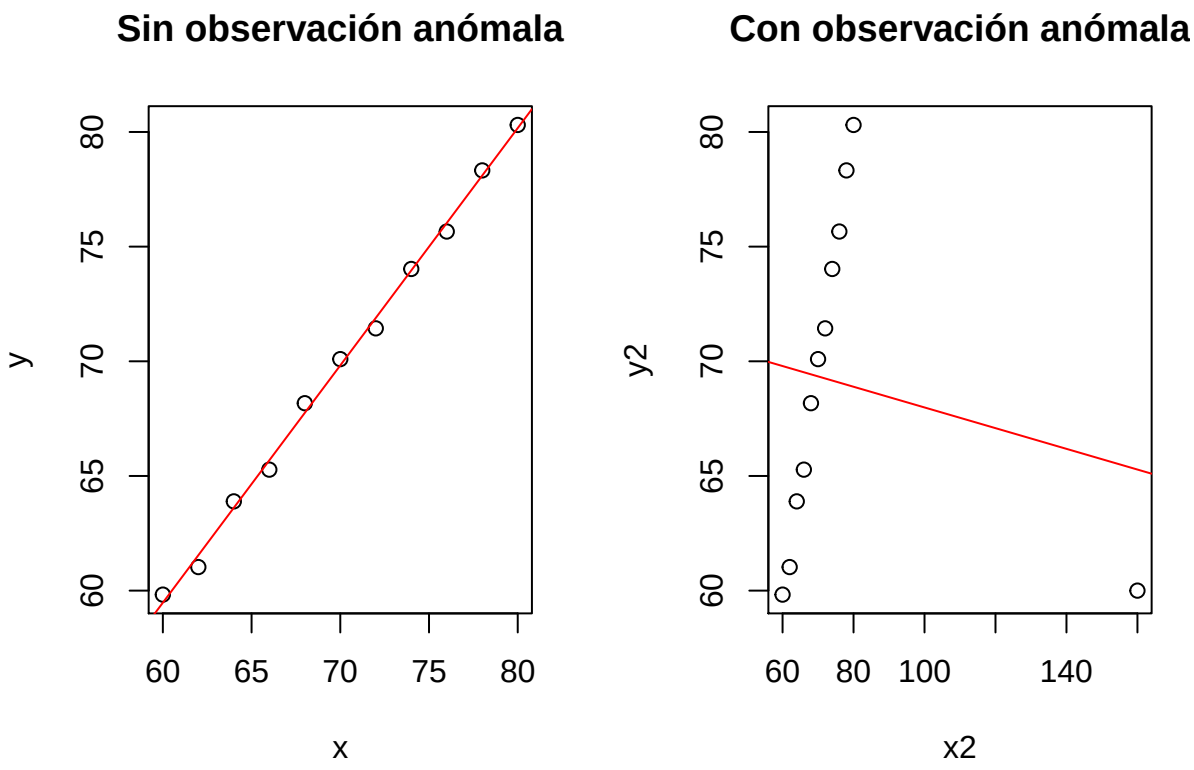
Las observaciones anómalas pueden provocar que se malinterpreten patrones en el conjunto de datos.

Además, puntos aislados pueden tener una gran influencia en el modelo de regresión dando resultados completamente diferentes.

Por ejemplo, pueden provocar que nuestro modelo no capture características importantes de los datos.

Por dichos motivos, es importante detectarlas.

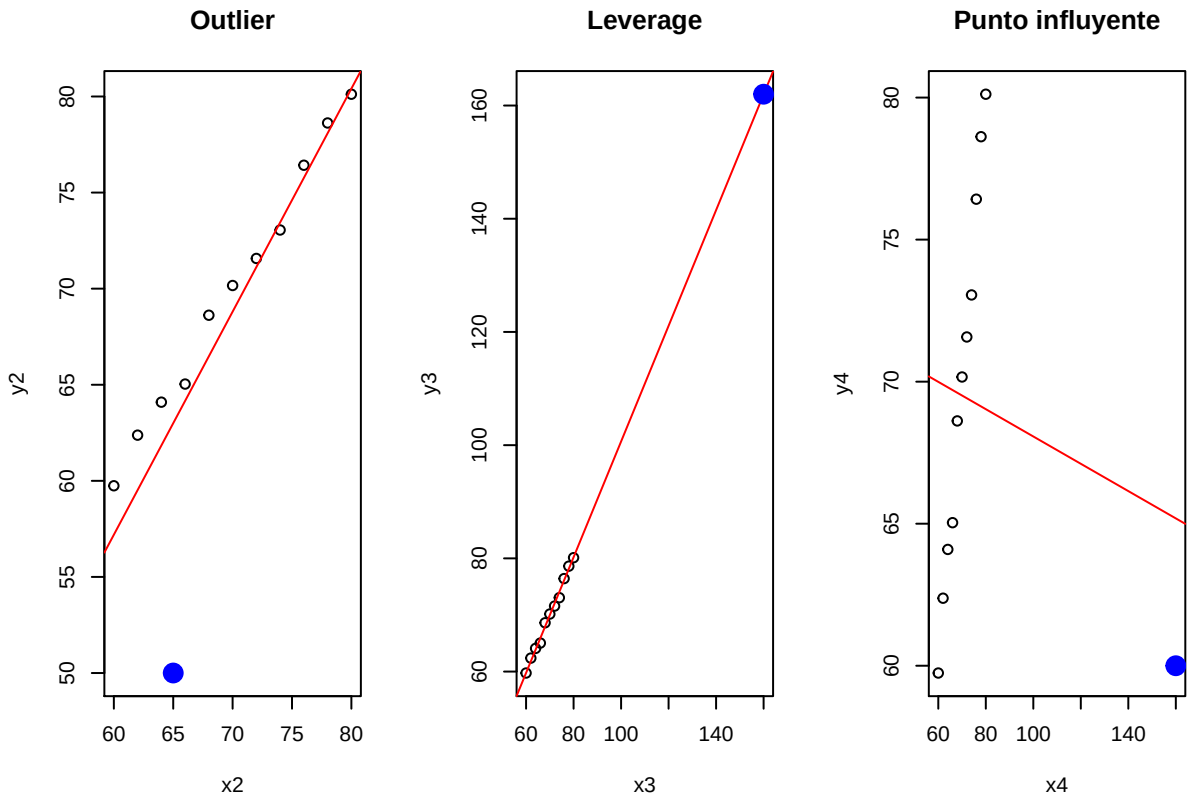
Como se puede observar en el gráfico siguiente, la presencia de un valor anómalo distorsiona completamente el modelo.



Existen tres tipos de observaciones anómalas:

- **Outliers de regresión**: son observaciones que tiene un valor anómalo de la variable dependiente Y, condicionado a los valores de sus variables independientes X_i. Tendrá un residuo muy alto pero puede no afectar demasiado a los coeficientes de la regresión.
- **Leverages**: son observaciones con un valor anómalo de las variables independientes X_i. No tiene porqué afectar los coeficientes de la regresión.
- **Observaciones influyentes**: son aquellas que tienen un **leverage** alto, son **outliers de regresión** y afectan fuertemente a la regresión.

En el gráfico siguiente vemos un ejemplo de un **outlier**, un **leverage** y una **observación anómala**.



8.2.21. Leverages

Para hallar las observaciones que son **leverages**, en primer lugar, necesitamos definir la **matriz Hat** como:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

Esta matriz es simétrica ($\mathbf{H}^\top = \mathbf{H}$) e idempotente ($\mathbf{H}^2 = \mathbf{H}$).

Además, es fácil comprobar que

$$\hat{y} = \mathbf{X}b = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y = \mathbf{H}y,$$

usando la expresión dada anteriormente que nos hallaba los valores estimados $\hat{y}$, y así tenemos que $\hat{y}_j = h_{1j}y_1 + h_{2j}y_2 + ... + h_{nj}y_n = \sum_{i=1}^{n} h_{ij}y_i$, para $j = 1, ..., n$.

Si la componente (i, j) de la **matriz Hat**, h_{ij} es grande, la observación i-ésima tiene un impacto sustancial en el valor predicho j-ésimo.

Se define el **leverage** de la observación i-ésima h_i como su **valor hat** $h_i = \sum_{j=1}^{n} h_{ij}^2$, y así, el **valor hat** h_i mide el **leverage potencial** de y_i en todos los valores predichos.

Propiedades de los leverages.

- El **valor hat** medio es $\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_i = \frac{k+1}{n}$.
- Los **valores hat** satisfacen $\frac{1}{n} \leq h_i \leq 1$.
- En la regresión lineal simple, los **valores hat** miden la distancia de x_i a la media de X: $h_i = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}$.
- En la regresión múltiple, h_i mide la distancia de una observación al vector medio de X de una forma parecida a la anterior.

Basándonos en las propiedades anteriores, la regla de decisión que nos dirá si una observación tiene leverage grande (y por lo tanto, tiene que ser considerada con cuidado) es cuando su **valor hat** cumpla: $h_i > 2\frac{k+1}{n}$.

La función `hatvalues` de `R` calcula los valores hat dado un modelo de regresión basándose en la regla anterior:

```
hatvalues(r)
```

donde `r` es el objeto de `R` donde hemos guardado la información de la regresión original.

Ejemplo de los bebés

Hallemos los valores **leverages** para la regresión realizada usando el ejemplo de los bebés:

```
(valores.hat=hatvalues(reg.mul.original))
```

```
##      1      2      3      4      5      6      7      8      9
## 0.4086 0.3615 0.3857 0.7313 0.5447 0.7008 0.6264 0.5139 0.7272
```

```
which(valores.hat > 2*(k+1)/n)
```

```
## named integer(0)
```

Vemos que en nuestro ejemplo no hay ninguna observación que puede considerarse con **leverage** alto.

8.2.22. Outliers

La estrategia para determinar qué observaciones son susceptibles de ser outliers se basan en los llamados **residuos estunderizados**.

Se basan en recalcular el modelo después de eliminar la observación i-ésima y hallar el correspondiente $(MSE)_i$.

Los **residuos estunderizados** se definen como:

$$E_i^\star = \frac{e_i}{\sqrt{(MSE)_i(1-h_i)}},$$

y, si el modelo es correcto, la variable anterior sigue una distribución t de Student con $n-k-2$ grados de libertad.

Para detectar los **outliers** se siguen los pasos siguientes:

- Para cada observación i-ésima, se calcula el **residuo estunderizado** $E_i^\star$.
- A continuación, se realiza una corrección de Bonferroni al p-valor multiplicándolo por n y así, el p-valor ajustado es $2nP(t_{n-k-2} \geq E_i^\star)$.
- Se van considerando por orden decreciente los p-valores de las observaciones hasta que se encuentra una observación que ya no sea un **outlier**.

La función de `R` que realiza este test de detección de **outliers** es la función `outlierTest` del paquete `car`.

```
outlierTest(r)
```

donde `r` es el objeto de `R` donde hemos guardado la información de la regresión original.

Ejemplo de los bebés

Veamos si tenemos **outliers** para la regresión realizada usando el ejemplo de los bebés:

```
outlierTest(reg.mul.original)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 8    4.737           0.01784       0.1606
```

Observamos que el bebé número 8 es el que tiene un **residuo estunderizado** más alto pero el p-valor ajustado de Bonferroni nos permite rechazar que sea un **outlier**.

8.2.23. Observaciones influyentes: Distancia de Cook

Como ya hemos indicado, una observación influyente es aquella que combina **discrepancia** con **leverage**.

Una forma de determinarlas es examinar cómo cambian los coeficientes de la regresión si se elimina una observación en concreto.

La medida para evaluar este cambio es la llamada **distancia de Cook**:

$$D_i = \frac{e_{S_i}^2}{k+1} \cdot \frac{h_i}{1-h_i},$$

dónde h_i es el **leverage** y es e_{S_i} el llamado **residuo estandarizado**, dado por

$$e_{S_i} = \frac{e_i}{\sqrt{MSE(1-h_i)}}.$$

El primer factor en la expresión de la **distancia de Cook** ($\frac{e_{S_i}^2}{k+1}$) mide el grado de ser **outlier** mientras que el segundo ($\frac{h_i}{1-h_i}$) mide el grado de **leverage**.

Una regla para determinar qué observaciones son influyentes es $D_i > \frac{4}{n-k-1}$.

En `R` la **distancia de Cook** se calcula con la función `cooks.distance` del paquete `car`:

```
cooks.distance(r)
```

donde `r` es el objeto de `R` donde hemos guardado la información de la regresión original.

Ejemplo de los bebés

Calculemos las observaciones influyentes usando la **distancia de Cook** para la regresión realizada usando el ejemplo de los bebés:

```
(distancias.cook=cooks.distance(reg.mul.original))
```

```
##         1         2         3         4         5         6         7         8
## 0.0005175 0.0039792 0.0127434 0.5110093 0.2958675 0.1035268 0.3303470 0.7459674
##         9
## 1.8532517
```

```
which(distancias.cook > 4/(n-k-1))
```

```
## 9
## 9
```

Observamos que el bebé número 9 es una observación influyente según la **distancia de Cook**.

8.2.24. Tratamiento de las observaciones anómalas

El tratamiento de las observaciones anómalas es bastante complejo.

Nos debemos preguntar si se deben a errores en la entrada o recogida de los datos y si éste es el caso, se debían de eliminar.

Pero también pueden explicar que no se ha considerado alguna variable independiente que afecta al conjunto de observaciones anómalas.

Las más peligrosas son las influyentes.

En el supuesto de que se determine que se pueden eliminar, se tienen que eliminar de una a una, actualizando el modelo cada vez.

8.2.25. Algunas consideraciones finales: selección del modelo

El modelo de regresión lineal no es el único que podemos usar. Existen otros modelos como los polinómicos o los logarítmicos que podrían ajustar mejor nuestra tabla de datos.

El modelo puede ser más eficaz si añadimos otras variables, o puede ser igual de eficaz si eliminamos variables redundantes.

Puede haber dependencias lineales entre las variables que las haga redundantes. Podemos detectar dichas dependencias con la matriz de covarianzas entre las variables regresoras o independientes.

En `R` existe la función `step` que, a partir del método `AIC` nos da el mejor modelo desde en el sentido de buscar un equilibrio entre la simplicidad y la adecuación:

```
step(r)
```

donde `r` es el objeto de `R` donde hemos guardado la información de la regresión original.

Ejemplo de los bebés

Si aplicamos la función `step` a la tabla de datos de los bebés, veamos cuál es el mejor modelo:

```
step(reg.mul.original)
```

Con un `AIC` de -3.78, `R` nos dice que el mejor modelo es considerar las variables regresoras x_2 (altura del bebé al nacer) y x_3 (peso del bebé al nacer).

```
## Start:  AIC=0.01
## y.bebes ~ X[, 2] + X[, 3] + X[, 4] + X[, 5]
##
##          Df Sum of Sq  RSS   AIC
## - X[, 5]  1      0.02 2.99 -1.92
## - X[, 2]  1      0.06 3.03 -1.80
## - X[, 3]  1      0.63 3.60 -0.25
## <none>                2.97  0.01
## - X[, 4]  1      6.25 9.22  8.22
##
## Step:  AIC=-1.92
## y.bebes ~ X[, 2] + X[, 3] + X[, 4]
##
##          Df Sum of Sq  RSS   AIC
## - X[, 2]  1      0.05 3.04 -3.78
## <none>                2.99 -1.92
## - X[, 3]  1      0.79 3.78 -1.80
## - X[, 4]  1      6.23 9.22  6.22
##
## Step:  AIC=-3.78
## y.bebes ~ X[, 3] + X[, 4]
##
##          Df Sum of Sq RSS   AIC
## <none>                  3 -3.78
## - X[, 4]  1       103 106 26.19
## - X[, 3]  1       132 135 28.38

##
## Call:
## lm(formula = y.bebes ~ X[, 3] + X[, 4])
##
## Coefficients:
## (Intercept)       X[, 3]       X[, 4]
##       2.183        0.958        3.325
```

8.3. Guía rápida. Regresión lineal simple

- `lm(y~x)`: objeto donde `R` guarda la información de la recta de regresión de la variable `y` en función de la variable `x`.
- `summary(lm(y~x))`: información detallada de la recta de regresión de la variable `y` en función de la variable `x`:
  - `summary(lm(y~x))$r.squared`: nos da el coeficiente de determinación.
  - `summary(lm(y~x))$coefficients`: nos da los coeficientes (b_0 y b_1) de la recta de regresión, las desviaciones estándar del error de dichos estimadores, el valor t al hacer los contrastes de hipótesis para contrastar si son nulos y los p-valores correspondientes de dichos contrastes.
- `abline(lm(y~x))`: dibuja la recta de regresión de la variable `y` en función de la variable `x` una vez que se ha realizado un plot de la nube de puntos (x_i, y_i), $i = 1, \dots, n$.
- `confint(lm(y~x),level=q)`: nos da el intervalo de confianza de los parámetros β_0 y β_1 al 100·`q` % de confianza. Valor por defecto: `q`=0.95.
- `predict.lm(lm(y~x),newdata,interval=...,level=q)`: da el intervalo de confianza para el parámetro $\mu_{Y|x_0}$ o y_0 al 100·`q` % de confianza dependiendo de los valores del parámetro `interval`: (`newdata` debe ser un `dataframe` de una fila con el valor de la variable x_0)
  - `interval=çonfidence"`: intervalo de confianza para $\mu_{Y|x_0}$.
  - `interval="prediction"`: intervalo de confianza para y_0.

8.4. Guía rápida. Regresión lineal múltiple

- `lm(y~x1+...xk)`: objeto donde `R` guarda la información de la función de regresión de la variable `y` en función de las variables regresoras `x1`,...,`xk`.
- `summary(lm(y~x1+...xk))`: información detallada de la recta de regresión de la variable `y` en función de las variables regresoras `x1`,...,`xk`:
  - `summary(lm(y~x))$r.squared`: nos da el coeficiente de determinación.
  - `summary(lm(y~x))$coefficients`: nos da los coeficientes $(b_0, b_1, \dots, b_k)$ de la función de regresión, las desviaciones estándar del error de dichos estimadores, el valor t al hacer los contrastes de hipótesis para contrastar si son nulos y los p-valores correspondientes de dichos contrastes.
  - `summary(lm(y~x))$adj.r.squared`: nos da el coeficiente de determinación ajustado.
- `AIC(lm(y~x1+...+xk))`: nos da el valor `AIC` del modelo de regresión múltiple `lm(y~x1+...xk)`.
- `BIC(lm(y~x1+...+xk))`: nos da el valor `BIC` del modelo de regresión múltiple `lm(y~x1+...xk)`.
- `confint(lm(y~x1+...+xk),level=q)`: nos da el intervalo de confianza de los parámetros $b_0, b_1, \dots, b_k$ al 100·`q` % de confianza. Valor por defecto: `q`=0.95.
- `predict.lm(lm(y~x1+...+xk),newdata,interval=...,level=q)`: da el intervalo de confianza para el parámetro $\mu_{Y|x_{01},\dots,x_{0k}}$ o y_0 al 100·`q` % de confianza dependiendo de los valores del parámetro `interval`: (`newdata` debe ser un `dataframe` de una fila con el valor de las variables $x_{01}, \dots, x_{0k}$)

  - `interval=çonfidence"`: intervalo de confianza para $\mu_{Y|x_{01},\ldots,x_{0k}}$.
  - `interval="prediction"`: intervalo de confianza para y_0.

- `anova(lm(y~Xd))`: calcula la tabla ANOVA de una regresión lineal múltiple de `y` en función de las columnas de la matriz `Xd`, es decir, `Xd` es una matriz cuyas columnas son los valores de las variables independientes $x_1, \ldots, x_k$.

- `bptest(lm(y~x1+...+xk),~X+I(X^2))` del paquete `lmtest`: realiza el test de White para verificar la homocedasticidad de los residuos en la regresión lineal múltiple `lm(y~x1+...+xk)` donde `X` es la matriz que contiene los valores de la muestra de las variables independientes.

- `bptest(lm(y~x1+...+xk)` del paquete `lmtest`: realiza el test de Breuch-Pagan para verificar la homocedasticidad de los residuos en la regresión lineal múltiple `lm(y~x1+...+xk)`.

- `dwtest(lm(y~x1+...+xk,alternative=))` del paquete `lmtest`: realiza el test de Durbin-Watson para comprobar la autocorrelación de los residuos donde si el parámetro `alternative` vale:

  - `greater`: comprueba si los residuos tienen autocorrelación positiva.
  - `less`: comprueba si los residuos tienen autocorrelación negativa.

- `residualPlot(lm(y~x1+...+xk),plot=...)` del paquete `car`: realiza el test de Tukey para verificar la aditividad del modelo lineal múltiple `lm(y~x1+...+xk)`. Los valores del parámetro `plot` pueden ser:

  - `FALSE`: da el valor del estadístico de contraste y el correspondiente p-valor.
  - `TRUE`: dibuja los gráficos de los residuos frente a las variables regresoras $x_1, \ldots, x_k$ y frente a los valores estimados $\widehat{y}_i$, $i = 1, \ldots, n$.

- `crPlots(lm(y~x1+...+xk))` del paquete `car`: realiza los gráficos de los residuos parciales para testear la linealidad del modelo de regresión múltiple `lm(y~x1+...+xk)`.

- `hatvalues(lm(y~x1+...+xk))`: cálculos los valores hat para el modelo de regresión múltiple `lm(y~x1+...+xk)` con el objetivo de hallar las posibles observaciones *leverages*.

- `outlierTest(lm(y~x1+...+xk))`: realiza el test de detección de outliers para el modelo de regresión múltiple `lm(y~x1+...+xk)`.

- `cooks.distance(lm(y~x1+...+xk))`: calcula las distancias de Cook de las observaciones para el modelo de regresión múltiple `lm(y~x1+...+xk)` con el fin de hallar las observaciones influyentes.

- `step(lm(y~x1+...+xk))`: da el mejor modelo de regresión múltiple en el sentido de buscar un equilibrio entre la simplicidad y la adecuación a partir del modelo completo `lm(y~x1+...+xk)` usando el valor `AIC`.

Capítulo 9

Introducción al Clustering

Uno de los problemas que más se presentan en el ámbito del **machine learning** es la clasificación de objetos.

Más concretamente, nos planteamos el problema siguiente: dado un conjunto de objetos, clasificarlos en grupos (*clusters*) basándonos en sus parecidos y diferencias.

9.1. ¿Qué es el clustering?

Algunas aplicaciones del **clustering** son las siguientes:

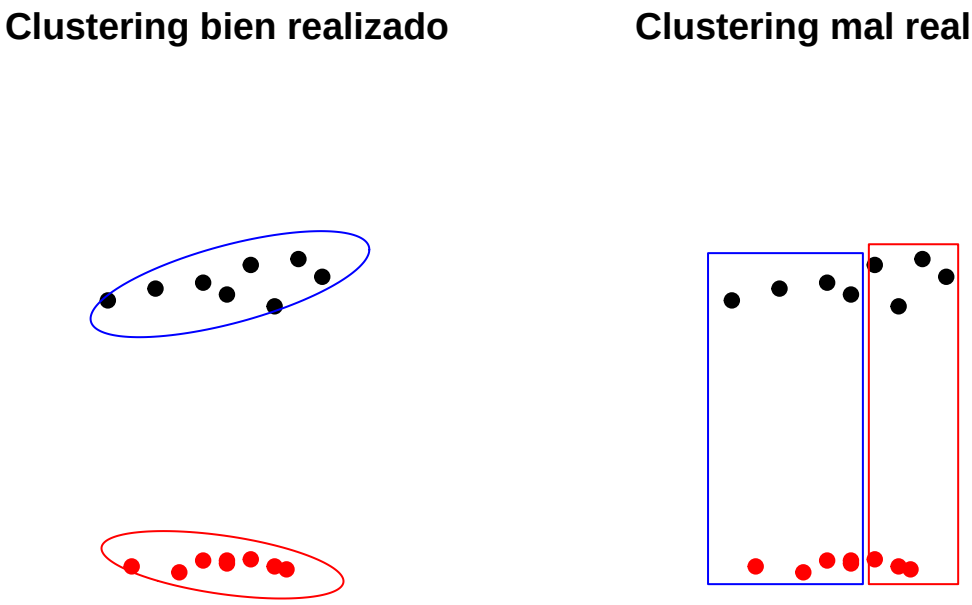
- En Biología: clasificación jerárquica de organismos (relacionada con una filogenia), agrupamiento de genes y agrupamiento de proteínas por parecido estructural.
- En Marketing: identificación de individuos por comportamientos similares (de compras, ocio, etc.)
- En Tratamiento de imágenes (en particular imágenes médicas): eliminación de ruido, detección de bordes, etc.
- En Biometría: identificación de individuos a partir de sus caras, huellas dactilares, etc.

9.1.1. Principios básicos

Los **algoritmos de clasificación o clustering** deben verificar dos principios básicos:

- **Homogeneidad**: los **clusters** deben ser homogéneos en el sentido de que objetos dentro de un mismo **cluster** tienen que ser **próximos o parecidos**.
- **Separación**: los objectos que pertenezcan a **clusters** diferentes tienen que estar **alejados**.

En los dos gráficos de la figura siguiente vemos un ejemplo de dos algoritmos de clustering donde en el de la izquierda, los clusters cumplen los principios anteriores pero en el de la derecha se violan los dos principios.



9.1.2. Tipos de algoritmos de clustering

Vamos a intentar **formalizar** los principios anteriores de cara a definir **algoritmos de clustering** que los verifiquen.

Existen dos tipos de **algoritmos de clustering**:

- **De partición:** el número de clusters con los que vamos a clasificar nuestro conjunto de objetos es un valor conocido y prefijado de entrada.

Sin embargo, veremos métodos que nos dirán como calcular el número **óptimo de cantidad de clusters** con los que dividir o clasificar nuestros objectos de nuestra tabla de datos. Por ejemplo, si consideramos los objetos de la figura anterior, se observa gráficamente que el número óptimo de clusters con los que clasificar dichos puntos es 2.

- **Jerárquico**: el **algoritmo de clustering** se compone de un número finito de pasos donde usualmente dicho número coincide con el número de objetos menos uno.

Los métodos **jerárquicos** a su vez se subclasifican en dos tipos más:

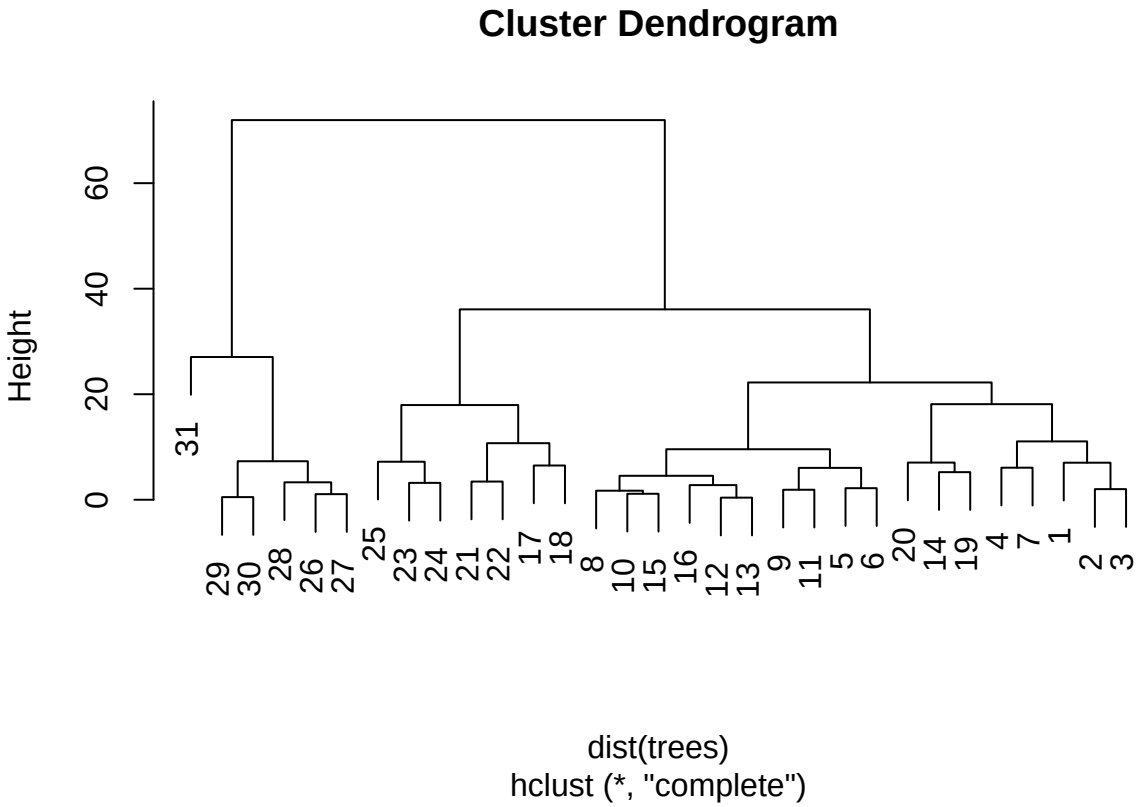
- **métodos aglomerativos**, donde en el paso inicial todos los objetos están separados y forman un cluster de un sólo objeto y en cada paso, se van agrupando aquellos objetos o clusters más **próximos** hasta llegar a un único cluster formado por todos los objetos.
- **métodos divisivos**, donde en el paso inicial existe un único cluster formado por todos los objetos y en cada paso se van dividiendo aquellos clusters más heterogéneos hasta llegar a tantos clusters como objetos existían inicialmente.

Los métodos **jerárquicos** tanto **aglomerativos** como **divisivos** producen un árbol binario de clasificación donde cada nodo de dicho árbol indica una **agrupación** de un cluster en dos en el caso de los métodos **aglomerativos** y una **partición** de un cluster en dos en el caso de los métodos **divisivos**.

En el siguiente gráfico, se observa cómo se han clasificado los 31 árboles de la tabla de datos `trees` de `R` a partir de los valores de las tres variables de dicha tabla de datos usando un método **jerárquico aglomerativo**:

- `Girth`: el valor del contorno del diámetro del árbol en pulgadas.
- `Height`: la altura del árbol en pies.
- `Volume`: el volumen del tronco del árbol en pies cúbicos.

Fijaos cómo al principio en la base del árbol todos los árboles forman un cluster y en la cima del árbol hay un único cluster formado por todos los árboles.



9.2. Métodos de partición

9.2.1. Algoritmo de las k-medias (k-*means*)

El **algoritmos de las k-medias** o *k-means* en inglés es el algoritmo de **partición** más conocido y más usado.

Recordemos que, al ser un **algoritmo de partición**, el número de clusters k se ha prefijado de entrada.

Dicho algoritmo busca una **partición** del conjunto de objetos, donde suponemos que conocemos un conjunto de características o variables que tienen valores continuos.

Concretamente, tenemos una tabla de datos de n filas y m columnas, donde cada fila representa un objeto u individuo y cada columna representa una característica de dicho individuo.

Individuos/Variables	X_1	X_2	...	X_m
1	x_{11}	x_{12}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	...	x_{im}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	...	x_{nm}

Por tanto, podemos identificar el individuo i-ésimo con un vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$ de $\mathbb{R}^m$.

El **algoritmo de las k-medias** va a clasificar los n individuos usando la información de la tabla anterior, es decir, la información de las m variables continuas.

Para realizar dicha clasificación, necesitamos definir cuándo dos objetos están **próximos**.

Una manera de definir la proximidad entre dos individuos, (no es la única) es a partir de la **distancia euclídea**.

Dados dos objetos $\mathbf{x}$ y $\mathbf{y}$ en $\mathbb{R}^m$, se define la **distancia euclídea** entre los dos como:

$$\|\mathbf{x}-\mathbf{y}\| = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}.$$

Para $m = 2$ o $m = 3$, la **distancia euclídea** es la longitud del segmento que une los puntos $\mathbf{x}$ e $\mathbf{y}$.

Por tanto, dos objectos estarán más **próximos**, cuánto más pequeña sea la **distancia euclídea** entre ambos.

Una vez establecidos las bases para el algoritmo, enunciemos el objetivo del mismo:

El objetivo del **algoritmo de las k-medias** es, a partir de la tabla de datos anterior de n filas (los individuos u objetos) y m columnas (las variables), hallar k puntos $\mathbf{c}_1, \dots, \mathbf{c}_k \in \mathbb{R}^n$ que minimicen

$$SS_C(\mathbf{x}_1, \dots, \mathbf{x}_n; k) = \sum_{i=1}^{n} \min_{j=1,\dots,k} \|\mathbf{x}_i - \mathbf{c}_j\|^2.$$

La cantidad SS_C se denomina suma de cuadrados dentro de los clusters.

Estos k puntos $\mathbf{c}_1, \dots, \mathbf{c}_k$ serán los centros de los clusters $C_1, \dots, C_k$ que queremos hallar.

Fijaos que, una vez hallados dichos centros $\mathbf{c}_1, \dots, \mathbf{c}_k$, los clusters quedan definidos por:

$$C_j = \{\mathbf{x}_i \text{ tal que} \|\mathbf{x}_i - \mathbf{c}_j\| < \|\mathbf{x}_i - \mathbf{c}_l\| \text{ para todo } l \neq j\}$$

Es decir, el cluster i-ésimo estará formado por los objetos $\mathbf{x}_l$ más próximos al centro $\mathbf{c}_i$.

Desgraciadamente, el problema anterior es un **problema abierto**, es decir, no se sabe hallar la solución para cualquier tabla de datos.

El **algoritmo de las k-medias** es un intento de hallar una solución local del mismo. Es decir, halla unos centros $\mathbf{c}_1, \dots, \mathbf{c}_k$ que solucionan el problema parcialmente pero no tenemos asegurado que los centros que halla el algoritmo minimicen globalmente SS_C.

Una vez establecidos las bases y el objetivo del algoritmo, vamos a explicar los pasos de los que consta.

Existen bastantes variantes del **algoritmo de las k-medias**, básicamente se diferencian en cómo iniciamos el algoritmo. En este curso, explicaremos el **algoritmo de Lloyd**:

- Paso 1: escogemos aleatoriamente los centros $\mathbf{c}_1, \dots, \mathbf{c}_k$.
- Paso 2: para cada $i = 1, \dots, n$, asignamos el individuo i-ésimo, $\mathbf{x}_i$, al cluster C_j definido por el centro $\mathbf{c}_j$ más próximo. Dicho en otras palabras, definimos los clusters a partir de los centros como hemos explicado antes.
- Paso 3: una vez hallados los clusters, hallamos los nuevos centros $\mathbf{c}_j$ calculando el valor medio de los objetos que forman el cluster C_j:

$$\mathbf{c}_j = \Big(\sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i \Big) / |C_j|.$$

- Paso 4: se repiten los pasos 2 y 3 hasta que los clusters estabilizan, o se llega a un número prefijado de iteraciones ya que el algoritmo anterior puede entrar en un "bucle infinito".

Observación: el resultado final, es decir, los clusters obtenidos, depende de cómo inicializemos el algoritmo, es decir, de cómo definamos los centros iniciales $\mathbf{c}_1, \dots, \mathbf{c}_k$.

Como ya hemos comentado, el algoritmo anterior no tiene porque dar un clustering óptimo, es decir, los centros obtenidos $\mathbf{c}_1, \dots, \mathbf{c}_k$ no tienen por qué minimizar la suma de cuadrados de los clusters SS_C. Por este motivo, conviene repetirlo varias veces con diferentes inicializaciones.

Ejemplo

Veamos un ejemplo de aplicación del algoritmo de k-medias.

La figura siguiente nos muestra los valores de un conjunto de puntos en el plano y queremos clasificarlos en 3 clusters usando el **algoritmo de k-medias**.

En la figura se observa el paso 1: una elección inicial de los centros $\mathbf{c}_1, \mathbf{c}_2$ y $\mathbf{c}_3$ representados con tres cuadrados de color verde, azul y rojo.

Por tanto, las dos coordenadas de los puntos serían las variables. Como consecuencia, tendríamos que $m = 2$.

En la figura siguiente, observamos el paso 2 del algoritmo, es decir, la creación de los clusters correspondientes a los tres centros iniciales.

Los puntos de color verde son los puntos más cercanos al centro de color verde, los puntos de color azul son los puntos más cercanos al centro de color azul y los puntos de color rojo son los puntos más cercanos al centro de color rojo.

En el tercer paso de aplicación del algoritmo, recalculamos los centros a partir del valor medios de los puntos que forman cada cluster C_j, $j = 1, 2, 3$ en la siguiente figura.

Por ejemplo, el centro de los dos puntos que forman el cluster de color verde está situado en medio de dichos puntos.

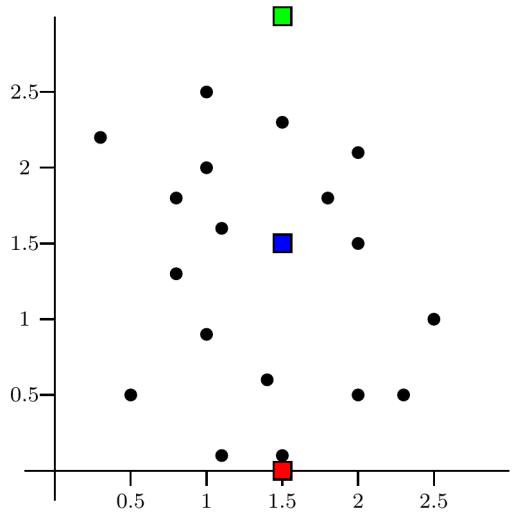


Figura 9.1: Configuración inicial o paso 1

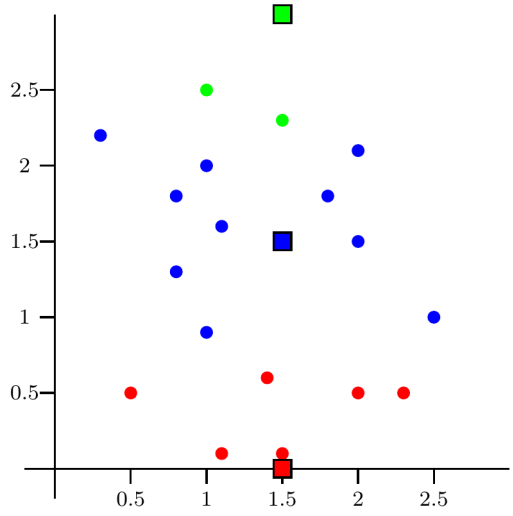


Figura 9.2: Paso 2

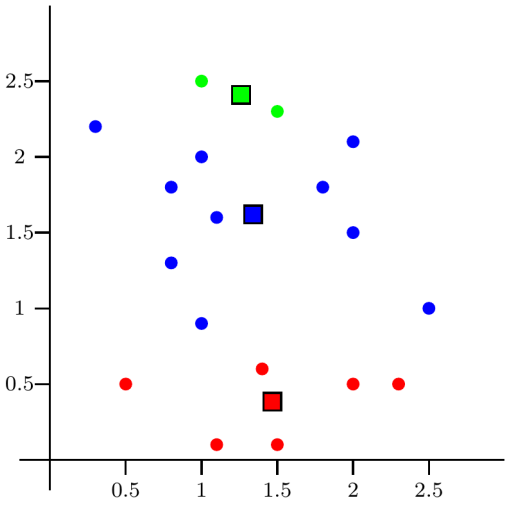


Figura 9.3: Paso 3

En el paso 4 repetimos los pasos 2 y 3 hasta que el algoritmo se estabiliza.

En la figuras siguientes repetimos los pasos 2 y 3.

Recordemos que en los pasos 2, recalculamos los clusters para los nuevos centros recalculados, es decir, asignamos a cada punto al cluster cuyo centro sea más cercano a dicho punto.

En los pasos 3, calculamos los nuevos centros para los clusters calculados en el paso 2.

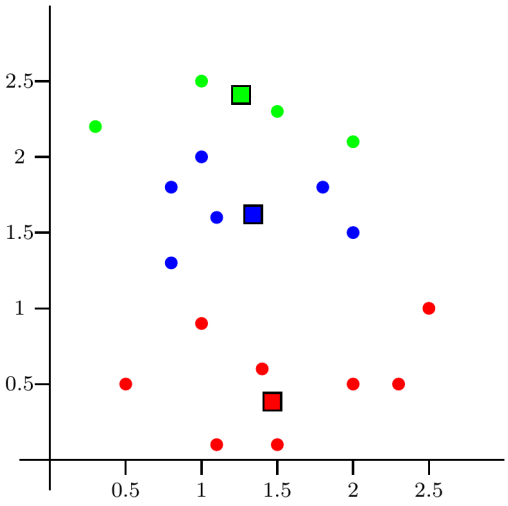


Figura 9.4: Paso 4: repetimos paso 2

En la última figura el algoritmo se ha estabilizado con un valor de la suma de cuadrados de los clusters de $SS_C = 7{,}25375$.

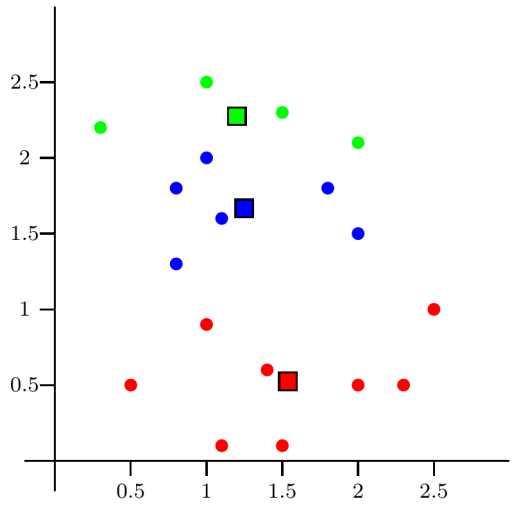


Figura 9.5: Repetimos paso 3

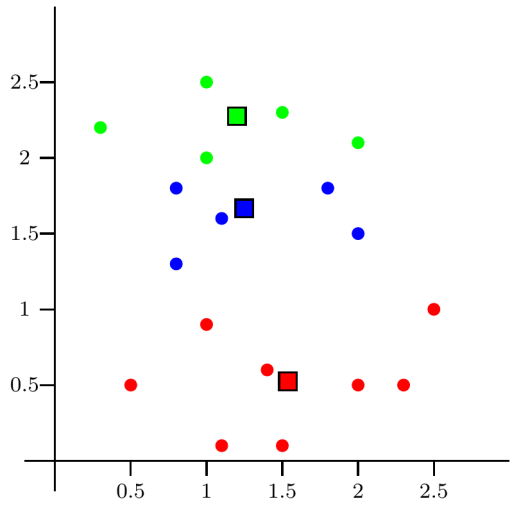


Figura 9.6: Repetimos paso 2

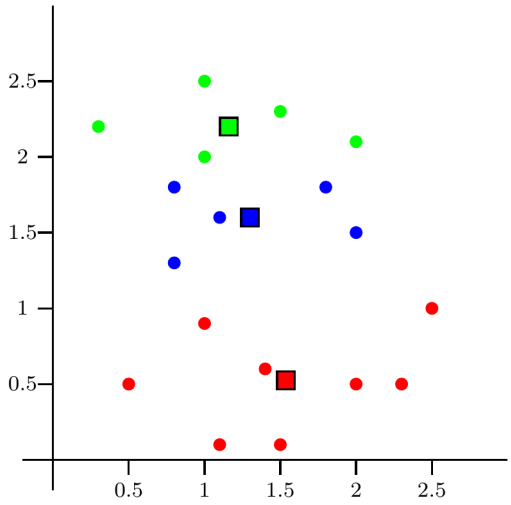


Figura 9.7: Repetimos paso 3

9.2.2. Limitación del algoritmo de k-medias

El **algoritmo de k-medias** tiene las limitaciones siguientes:

- No existe un método **eficiente y universal** de elegir los **centros de partida**.
- Como ya hemos comentado anteriormente, no se puede garantizar un **óptimo global**.
- No se puede determinar de manera efectiva el número k de clusters **a priori** aunque existe algún método como veremos más adelante.
- Si no queremos que la variación de los datos intervenga en el análisis, conviene estandarizar dichos datos ya que el algoritmo **no es invariante** a **cambio de escala**. De esta manera todas las variables tendrán media 0 y varianza 1.
- El algoritmo es sensible a **outliers** o datos atípicos.
- Si nuestra tabla de datos contiene variables no numéricas tipo **factor, ordinales**, etc., el algoritmo no se puede aplicar. Es decir, sólo es aplicable para variables continuas que tengan valores en $\mathbb{R}^n$ usando la **distancia euclídea**.
- Los clusters que encuentra son esféricos debido a que usa la **distancia euclídea**.

9.2.3. Métodos para hallar el número de clusters k. Método del codo

Vamos a ver el método más usado para determinar el número óptimo de clusters a calcular con el **algoritmo de las k medias**.

Uno de dichos métodos se denomina el **método del codo**. Dicho método se basa en observar como varía la **suma de cuadrados de los clusters** SS_C como función del número de clusters k y escoger aquel valor k que haga que la variación de dicha función sea lo más pronunciada posible.

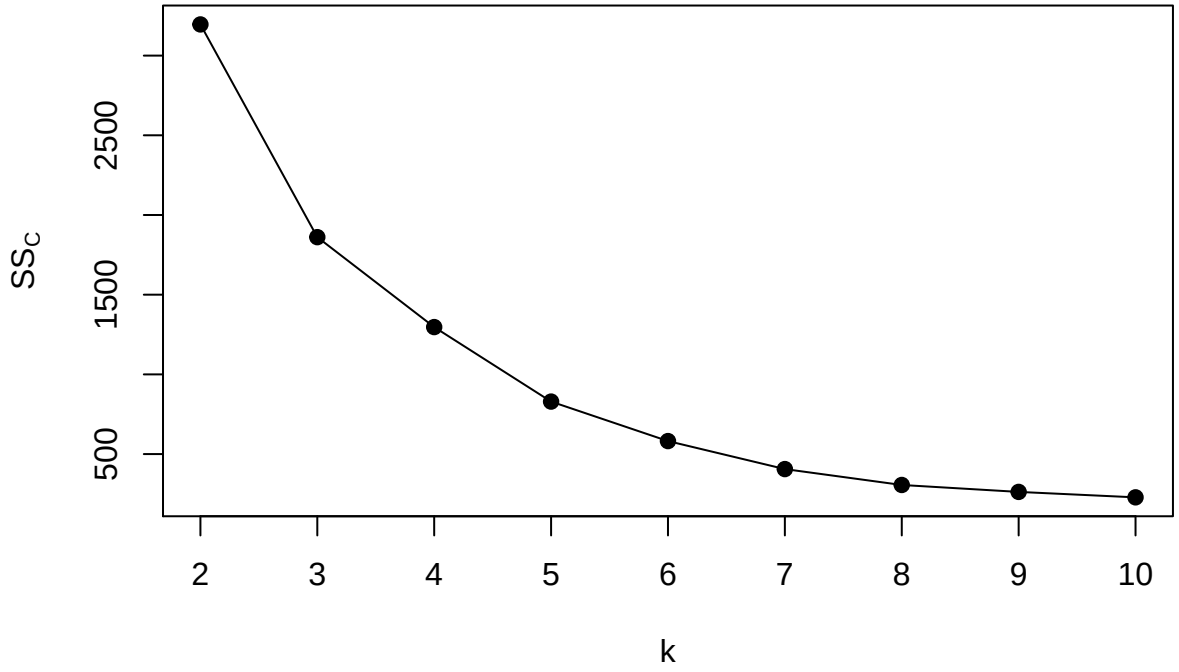
La idea es que dicho k es el valor óptimo en el sentido que hace que la disminución de SS_C sea lo más "óptima" posible.

Veamos un ejemplo para ver cómo aplicar el **método del codo** para hallar el número óptimo de clusters k.

Ejemplo

Consideremos la tabla de datos `trees` de `R` que ya hemos comentado anteriormente.

En la figura siguiente observamos como evoluciona el valor de la suma de cuadrados de los clusters SS_C con el número de clusters k:



Vemos que la función $SS_C(k)$ es una función cóncava.

El método del codo elige el valor k a partir del cual SS_C disminuye mucho más lentamente que antes de él.

Dicho matemáticamente, el valor k óptimo es aquel en el que la variación de la pendiente antes de k y después de k es más pronunciada.

Los valores de las pendientes en el gráfico anterior son los siguientes:

k's	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
	-1334.531	-564.209	-467.337	-247.838	-175.649	-99.766	-43.484	-34.082

Entonces el k óptimo es $k = 3$ en este caso.

9.2.4. Algoritmo de k-medias con R

Para ejecutar el algoritmo de k-medias con R hay que usar la función `kmeans`:

```
kmeans(x, centers=..., iter.max=..., algorithm =...)
```

donde:

- `x` es la matriz o el data frame cuyas filas representan los objetos; en ambos casos, todas las variables han de ser numéricas como ya hemos indicado.
- `centers` sirve para especificar los centros iniciales, y se puede usar de dos maneras: igualado a un número k, R escoge aleatoriamente los k centros iniciales, mientras que igualado a una matriz de k filas y el mismo número de columnas que `x`, R toma las filas de esta matriz como centros de partida.
- `iter.max` permite especificar el número máximo de iteraciones a realizar; su valor por defecto es 10. Al llegar a este número máximo de iteraciones, si el algoritmo aún no ha acabado porque los clusters aún no hayan estabilizado, se para y da como resultado los clusters que se han obtenido en la última iteración.
- `algorithm` indica el algoritmo a usar. Los algoritmos pueden ser los siguientes:
  - **Lloyd**: es el que hemos explicado.
  - **Hartigan-Wong**: este algoritmo empieza igual que el de Lloyd: se escogen k centros, se calculan las distancias euclídeas de cada punto a cada centro, y se asigna a cada centro el cluster de puntos que están más cerca de él que de los otros centros. A continuación, en pasos sucesivos se itera el bucle 1–5 siguiente hasta que en una iteración del mismo los clusters no cambian:

  (1) Se sustituye cada centro por el punto medio de los puntos asignados a su cluster.

  (2) Se calculan las distancias euclídeas de cada punto a cada centro.

  (3) Se asigna (temporalmente) a cada centro el cluster formado por los puntos que están más cerca de él que de los otros centros.

  (4) Si en esta asignación algún punto ha cambiado de cluster, digamos que el punto $\mathbf{x}_i$ se ha incorporado al cluster C_j de centro $\mathbf{c}_j$, entonces:

  - Se calcula el valor SSE_j que se obtiene multiplicando la suma de cuadrados de cada cluster $SS_{C_j} = \sum_{l=1}^{n_j} \|\mathbf{x}_{jl} - \mathbf{c}_j\|^2$, donde $\mathbf{x}_{jl}$ son los puntos del cluster C_j, para $l = 1, \dots, n_j$, por $n_j/(n_j - 1)$ (donde n_j indica el número de elementos del cluster C_j).
  - Se calcula, para todo otro cluster C_k, el correspondiente valor $SSE_{i,k}$ como si $\mathbf{x}_i$ hubiera ido a parar a C_k.
  - Si algún $SSE_{i,k}$ resulta menor que SSE_j, $\mathbf{x}_i$ se asigna definitivamente al cluster C_k que da valor mínimo de $SSE_{i,k}$.

5.Una vez realizado el procedimiento anterior para todos los puntos que han cambiado de cluster, éstos se asignan a sus clusters definitivos y se da el bucle por completado.

- **MacQueen**: es el mismo método que el de Lloyd salvo por el hecho de que no se recalculan todos los clusters y sus centros de golpe, sino elemento a elemento. Es decir, se empieza igual que en los dos algoritmos anteriores: se escogen k centros, se calculan las distancias euclídeas de cada punto a cada centro, se asigna a cada centro el cluster de puntos que están más cerca de él que de los otros centros, y se sustituye cada centro por el punto medio de los puntos asignados a su cluster. A partir de aquí, en pasos sucesivos se itera el bucle siguiente (recordemos que los puntos a clasificar son $\mathbf{x}_1, \dots, \mathbf{x}_n$, y los supondremos ordenados por su fila en la tabla de datos):
  - Para cada $i = 1, \dots, n$, se mira si el punto $\mathbf{x}_i$ está más cerca del centro de otro cluster que del centro del cluster al que está asignado.
  - Si no lo está, se mantiene en su cluster y se pasa al punto siguiente, $\mathbf{x}_{i+1}$. Si se llega al final de la lista de puntos y todos se mantienen en sus clusters, el algoritmo se para.
  - Si $\mathbf{x}_i$ está más cerca de otro centro, se traslada al cluster definido por este centro, se recalculan los centros de los dos clusters afectados (el que ha abandonado $\mathbf{x}_i$ y aquél al que se ha incorporado), y se reinicia el bucle, empezando de nuevo con $\mathbf{x}_1$.

El clustering resultante está formado por los clusters existentes en el momento de parar.

La salida del algoritmo de k-medias con `R` tiene las componentes siguientes: (supongamos que hemos guardado en el objeto `resultado.km` la salida del algoritmo de k-medias aplicado a nuestra tabla de datos)

- `resultado.km$size`: nos da los números de objetos, es decir, los tamaños de cada cluster.
- `resultado.km$cluster`: nos dice qué cluster pertenece cada uno de los objetos de nuestra tabla de datos.
- `resultado.km$centers`: nos da los centros de cada cluster en filas. Es decir, la fila 1 sería el centro del cluster 1, la fila 2, del cluster 2 y así sucesivamente.
- `resultado.km$withinss`: nos da las sumas de cuadrados de cada cluster, lo que antes hemos denominado SS_{C_j}, para $j = 1, \dots, k$.
- `resultado.km$tot.withinss`: la suma de cuadrados de todos los clusters, lo que antes hemos denominado SS_C. También se puede calcular sumando las sumas de los cuadrados de cada cluster: `sum(resultado.km$withinss)`.
- `resultado.km$totss`: es la suma de los cuadrados de las distancias de los puntos en su punto medio de todos estos puntos. Es decir, seria la suma de los cuadrados SS_C pero suponiendo que sólo hubiera un sólo cluster.
- `resultado.km$betweenss` es la diferencia entre `resultado.km$totss` y `resultado.km$tot.withinss` y puede demostrarse (es un cálculo bastante tedioso) que es igual a la suma, ponderada por el número de objetos del cluster correspondiente, de los cuadrados de las distancias de los centros de los clusters al punto medio de todos los puntos.

Es decir:

- sean $n_1, \dots, n_k$ los números de objectos de los clusters $C_1, \dots, C_k$,
- sean $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$ los objectos pertenecientes al cluster C_i,
- sean $\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_k$, los centros de los clusters $C_1, \dots, C_k$: $\overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij},\ i = 1, \dots, k$,

- sea $\mathbf{x}$ el punto medio de todos los puntos: $\overline{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$.

Entonces:

$$\texttt{resultado.km\$betweenss} = \sum_{i=1}^{k} n_i \|\overline{\mathbf{x}}_i - \mathbf{x}\|^2.$$

Podríamos considerar la medida anterior como una medida de dispersión de los centros o una medida de cuan separados están los clusters ya que cuanto mayor sea `resultado.km$betweenss` más separados estarán los centros del punto medio global de todos los puntos y mayor separación habrá entre los clusters.

Entonces, nos interesa el cociente `resultado.km$betweenss/resultado.km$totss`, que mide la fracción de la variabilidad de los datos que explican los clusters. Cuanto mayor mejor.

Ejemplo

Apliquemos el algoritmo de k-medias en `R` a la tabla de datos `trees` usando la variante del algoritmo de **McQueen** con $k = 3$ clusters

```
resultado.km.trees=kmeans(trees,centers=3,algorithm = 'MacQueen')
resultado.km.trees
```

```
## K-means clustering with 3 clusters of sizes 19, 7, 5
##
## Cluster means:
##   Girth Height Volume
## 1 12.52  76.47  25.32
## 2 17.94  81.00  55.93
## 3  9.44  67.20  12.56
##
## Clustering vector:
##  [1] 3 3 3 3 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 1335.0  737.9  100.5
##  (between_SS / total_SS =  77.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
resultado.km.trees=kmeans(trees,centers=3,algorithm = 'MacQueen')
resultado.km.trees
```

```
## K-means clustering with 3 clusters of sizes 23, 7, 1
##
## Cluster means:
##   Girth Height Volume
## 1 11.70  74.65  21.98
## 2 17.29  78.86  50.40
```

```
## 3 20.60  87.00  77.00
##
## Clustering vector:
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3
##
## Within cluster sum of squares by cluster:
## [1] 2177.8  397.2    0.0
##  (between_SS / total_SS =  73.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Vemos que R nos ha dado 3 clusters de 23, 7 y 1 elementos cada uno.

Recordemos que el vector `resultado.km.trees$cluster` nos dice a qué cluster pertenece cada uno de los 31 árboles de nuestra tabla de datos.

Por ejemplo los árboles enumerados como

```
which(resultado.km.trees$cluster==1)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
```

están en el cluster número 1, los árboles enumerados como

```
which(resultado.km.trees$cluster==2)
```

```
## [1] 24 25 26 27 28 29 30
```

están en el cluster número 2

y los árboles enumerados como

```
which(resultado.km.trees$cluster==3)
```

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

están en el cluster número 3.

El centro del cluster número 1 ha sido:

```
resultado.km.trees$centers[1,]
```

```
##  Girth Height Volume
##  11.70  74.65  21.98
```

El centro del cluster número 2 ha sido:

```
resultado.km.trees$centers[2,]
```

```
##  Girth Height Volume
##  17.29  78.86  50.40
```

y el del cluster 3,

```
resultado.km.trees$centers[3,]
```

```
##  Girth Height Volume
##   20.6   87.0   77.0
```

Las sumas de cuadrados de cada cluster SS_{C_1}, SS_{C_2} y SS_{C_3} son las siguientes:

```
resultado.km.trees$withinss
```

```
## [1] 2177.8  397.2    0.0
```

La suma de cuadrados de todos los clusters vale, en nuestro caso:

```
sum(resultado.km.trees$withinss)
```

```
## [1] 2575
```

```
resultado.km.trees$tot.withinss
```

```
## [1] 2575
```

El valor `resultado.km.trees$totss` nos da la suma de los cuadrados SS_C pero suponiendo que sólo hubiera un cluster:

```
resultado.km.trees$totss
```

```
## [1] 9620
```

La dispersión de los centros de los clusters respecto del punto medio de todos los árboles vale:

```
resultado.km.trees$betweenss
```

```
## [1] 7044
```

Comprobemos la expresión vista anteriormente:

```
(centro.global=apply(trees,2,mean))
```

```
##  Girth Height Volume
##  13.25  76.00  30.17
```

```
sum(resultado.km.trees$size*apply((t(resultado.km.trees$centers)-centro.global)^2,2,sum))
```

```
## [1] 7044
```

En este caso, los clusters han explicado un 73.23 % de la variabilidad total de los puntos.

Tal como hemos indicado anteriormente, ejecutar sólo una vez el algoritmo de k-means no es aconsejable ya que no tenemos asegurado que hemos llegado al mínimo.

Por dicho motivo, ejecutemos el algoritmo de k-means sobre la tabla de datos `trees` 50 veces a ver si podemos obtener un mínimo más óptimo:

```
veces=50
SSCs=c()
for (i in 1:veces){
  SSCs=c(SSCs,kmeans(trees,3,algorithm = "MacQueen")$tot.withinss)
```

```
}
(min(SSCs))
```

```
## [1] 1861
```

¡Uy!, no habíamos llegado al mínimo. Ahora sabemos que hay un mínimo más óptimo que el hallado anteriormente.

9.3. Clustering jerárquico

En esta parte vamos a introducir un tipo de clustering que, en lugar de darnos los clusters de la partición de objetos, va a darnos un árbol binario que nos va a indicar cómo se van agrupando los objetos de nuestra tabla de datos.

De cara a no extendernos demasiado, vamos a describir con todo detalle los **métodos aglomerativos** al ser los más usados tanto en técnicas de **machine learning** como en la literatura.

Otra diferencia a tener en cuenta con respecto a los **métodos de partición** es que los métodos de **clustering jerárquico** dan el árbol binario de clasificación a partir de una matriz de distancias entre los objetos de nuestra tabla de datos, no a partir de la tabla de datos misma, tal como hacíamos con el **algoritmo de k-medias**.

Por tanto, antes de empezar a aplicar las técnicas del **clustering jerárquico**, hemos de hallar una matriz de distancias entre los objectos de nuestra tabla de datos.

En resumen, los **métodos jerárquicos** parten de una matriz $\mathbf{D}$ de distancias entre los n objetos de nuestra tabla de datos.

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & ... & d_{1n} \\ d_{21} & d_{22} & ... & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & ... & d_{nn} \end{pmatrix},$$

dónde cada d_{ij} es la distancia o el parecido entre el objeto i y lo objeto j.

9.3.1. Distancias

A continuación, especifiquemos la definición general de **distancia** en un conjunto X:

Definición. Una **distancia** sobre un conjunto X de objetos es una aplicación $d : X \times X \to [0, \infty[$ que satisface las condiciones siguientes:

- **Separación**: $d(x, y) = 0$ si, y sólo si, $x = y$.
- **Simetría**: dados dos objectos cualesquiera $x, y \in X$, $d(x, y) = d(y, x)$.
- **Desigualdad triangular**: dados tres objectos cualesquiera $x, y, z \in X$, se verifica $d(x, z) \leq d(x, y) + d(y, z)$.

Diremos que dos objetos x, y son más parecidos o más cercanos cuanto más pequeña es $d(x, y)$, la distancias entre ellos.

Si tenemos una tabla de datos, el conjunto X será un conjunto finito formado por los n objetos o las n filas de la tabla de datos: $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$.

Por tanto, dar una distancia sobre X es equivalente a dar una matriz $\mathbf{D}$ de n filas y n columnas, donde la componente i, j de dicha matriz, $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$, sería la distancia entre el objeto i y el objeto j-ésimo.

Usando la definición de distancia vista anteriormente, tendremos que dicha matriz $\mathbf{D}$ debe ser

- simétrica: $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$, para cualquier i y j,
- cumplir la desigualdad triangular: $d(\mathbf{x}_i, \mathbf{x}_k) \leq d(\mathbf{x}_i, \mathbf{x}_j) + d(\mathbf{x}_j, \mathbf{x}_k)$, para cualquier i, j, k.

La primera decisión a tomar cuando queramos realizar un **clustering jerárquico** es elegir la **distancia** que vamos a usar.

Es una decisión muy importante ya que dicha decisión determinará el **árbol binario** que vamos a obtener y dará un significado u otro a los clusters que se deriven del mismo.

La **distancia** a elegir dependerá del tipo de datos con los que trabajemos. Más concretamente, vamos a distinguir dos tipos de datos: **binarios** y **continuos**.

El significado de las distancias dependerá del tipo de datos con los que trabajemos.

9.3.2. Datos binarios

Supongamos que las variables de nuestra matriz de datos sólo contiene datos **binarios**, es decir, datos con sólo dos posibles valores: 0/1, ausencia/presencia, éxito/fracaso, etc.

Usar por ejemplo la distancia euclídea en este tipo de datos no tendría ningún sentido ya que los valores 0 y 1 no tienen ningún significado numérico.

Para fijar ideas, supongamos que nuestra tabla de datos tiene n filas (objetos) y m columnas (variables), donde los valores x_{ij} valen 0 o 1, $i = 1 \dots, n$, $j = 1, \dots, m$.

Consideramos los valores de dos objectos cualesquiera $\mathbf{x}_i = (x_{i1}, \dots, x_{im})$, $\mathbf{x}_j = (x_{j1}, \dots, x_{jm})$ y queremos **medir** de alguna manera lo parecidos que son dichos objetos.

Para ello, definimos primeramente las cantidades siguientes:

$$\begin{aligned} a_0 &= \left|\{k \mid x_{ik} = x_{jk} = 0\}\right|, \\ a_1 &= \left|\{k \mid x_{ik} = x_{jk} = 1\}\right|, \\ a_2 &= \left|\{k \mid x_{ik} \neq x_{jk}\}\right|. \end{aligned}$$

Es decir, a_0 serían el número de variables en los que los objetos i y j tienen el valor 0, a_1, el número de variables en los que los objetos i y j tienen el valor 1 y a_2, el número de variables en los que los objetos i y j difieren.

Intuitivamente, cuanto mayores sean a_0 y a_1 y cuanto menor sea a_2, más parecidos deben ser los objetos i y j.

Basándonos en la consideración anterior, definimos el "parecido" o la **distancia** entre los objetos i y j como:

$$\sigma_{ij} = \frac{a_1 + \delta a_0}{\alpha a_1 + \beta a_0 + \lambda a_2},$$

donde los parámetros α, β, δ y λ son valores a elegir que nos determinarán la distancia entre los objetos i y j.

Fijarse que el numerador de la distancia σ_{ij} sólo depende del número de variables coincidentes a_0 y a_1 y el denominador depende de todas, tanto las coincidentes a_0 y a_1 como la disidente a_2.

Los valores más comunes de los parámetros anteriores junto con sus nombres se presentan en la tabla siguiente:

Nombre	δ	λ	α	β	Definición
Hamming	1	1	1	1	$\frac{a_1+a_0}{a_0+a_1+a_2}$
Jaccard	0	1	1	0	$\frac{a_1}{a_1+a_2}$
Tanimoto	1	2	1	1	$\frac{a_1+a_0}{a_1+2a_2+a_0}$
Rusell–Rao	0	1	1	1	$\frac{a_1}{a_0+a_1+a_2}$
Dice	0	0,5	1	0	$\frac{2a_1}{2a_1+a_2}$
Kulczynski	0	1	0	0	$\frac{a_1}{a_2}$

Ejemplo

La tabla de datos `Arrests` de la librería `car` nos da la información de 5226 arrestos en Toronto por posesión de cantidades pequeñas de marihuana.

En este ejemplo nos interesan las siguientes variables binarias:

- `colour`: la raza del arrestado/a con niveles `Black` y `White`.
- `sex`: el sexo del arrestado/a con niveles `Female` y `Male`.
- `employed`: si el arrestado/a tenía empleo con niveles `No` y `Yes`.
- `citizen`: si el arrestado/a era ciudadano/a de Toronto con niveles `No` y `Yes`.

Como dicha tabla de datos contiene muchos individuos, vamos a considerar sólo una muestra de 25 individuos:

```
library(car)
set.seed(888)  ## Fijamos la semilla
individuos.elegidos = sample(1:5226,25)
tabla.arrestados = Arrests[individuos.elegidos,c("colour","sex","employed","citizen")]
rownames(tabla.arrestados)=1:25
```

Los 10 primeros individuos de nuestra tabla de datos son:

```
head(tabla.arrestados,10)
```

```
##    colour    sex employed citizen
## 1   White   Male       No     Yes
## 2   White   Male      Yes     Yes
## 3   White   Male      Yes     Yes
```

```
## 4   White   Male      Yes     Yes
## 5   White   Male      Yes     Yes
## 6   White   Male      Yes     Yes
## 7   White   Male      Yes     Yes
## 8   White Female       No     Yes
## 9   Black   Male      Yes     Yes
## 10  Black   Male      Yes     Yes
```

Vamos a hallar la matriz de **distancias de Hamming** entre los 25 individuos.

En primer lugar, codificamos los valores de los niveles de las variables de nuestra tabla de datos de la forma siguiente:

- `colour`: `White`=0, `Black`=1.
- `sex`: `Male`=0, `Female`=1.
- `employed`: `No`=0, `Yes`=1.
- `citizen`: `No`=0, `Yes`=1.

```
tabla.arrestados$colour = ifelse(tabla.arrestados$colour=="White",0,1)
tabla.arrestados$sex = ifelse(tabla.arrestados$sex=="Male",0,1)
tabla.arrestados$employed = ifelse(tabla.arrestados$employed=="No",0,1)
tabla.arrestados$citizen = ifelse(tabla.arrestados$citizen=="No",0,1)
```

En segundo lugar, definimos la función siguientes para que nos dé los valores a_0, a_1 y a_2 dados un par de individuos i y j:

```
as = function(xi,xj){
  n=length(xi)
  a0 = length(which(xi==xj & xi==0))
  a1 = length(which(xi==xj & xi==1))
  a2 = length(which(xi!=xj))
  return(c(a0,a1,a2))
  }
```

Recordemos que la distancia de Hamming vale $\sigma_{ij} = \frac{a_1+a_0}{a_0+a_1+a_2}$:

```
n=dim(tabla.arrestados)[1]
matriz.dist.hamming=matrix(1,n,n)
for (i in 1:(n-1)){
  for (j in (i+1):n){
    aux=as(tabla.arrestados[i,],tabla.arrestados[j,])
    matriz.dist.hamming[i,j]=(aux[1]+aux[2])/sum(aux)
    matriz.dist.hamming[j,i]=matriz.dist.hamming[i,j]
  }
}
```

La distancia de Hamming de los 10 primeros individuos vale:

```
matriz.dist.hamming[1:10,1:10]
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,] 1.00 0.75 0.75 0.75 0.75 0.75 0.75 0.75 0.50  0.50
##  [2,] 0.75 1.00 1.00 1.00 1.00 1.00 1.00 0.50 0.75  0.75
```

```
##  [3,] 0.75 1.00 1.00 1.00 1.00 1.00 1.00 0.50 0.75  0.75
##  [4,] 0.75 1.00 1.00 1.00 1.00 1.00 1.00 0.50 0.75  0.75
##  [5,] 0.75 1.00 1.00 1.00 1.00 1.00 1.00 0.50 0.75  0.75
##  [6,] 0.75 1.00 1.00 1.00 1.00 1.00 1.00 0.50 0.75  0.75
##  [7,] 0.75 1.00 1.00 1.00 1.00 1.00 1.00 0.50 0.75  0.75
##  [8,] 0.75 0.50 0.50 0.50 0.50 0.50 0.50 1.00 0.25  0.25
##  [9,] 0.50 0.75 0.75 0.75 0.75 0.75 0.75 0.25 1.00  1.00
## [10,] 0.50 0.75 0.75 0.75 0.75 0.75 0.75 0.25 1.00  1.00
```

9.3.3. Datos continuos

Supongamos ahora que nuestra matriz de datos contiene datos continuos, es decir, los objectos a clasificar se pueden considerar elementos de $\mathbb{R}^m$, donde recordemos que tenemos n individuos u objetos y m variables.

El significado de **distancia** en este caso es el opuesto al significado de **distancia** en el caso de datos **binarios**.

Ahora en el caso de datos continuos, cuanto mayor sea la distancia entre dos objetos, más disimilares son y cuanto menor sea, más similares serán.

En este caso, las **distancias** que se definen entre los objectos se basan en las llamadas **normas** L_r.

Más concretamente, dados dos objectos $\mathbf{x}_i$ y $\mathbf{x}_j$ de componentes,

$$\mathbf{x}_i = (x_{i1}, \dots, x_{im}),\ \mathbf{x}_j = (x_{j1}, \dots, x_{jm}),$$

la **distancia** L_r entre dichos objetos se define como:

$$d_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\|_r = \left(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^r \right)^{1/r}$$

Cuando $r = 1$, la distancia anterior: $d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$, se denomina **distancia de Manhattan** y cuando $r = 2$, $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$, **distancia euclídea**.

Ejemplo

Para ilustrar el cálculo de distancias en caso de variables continuas, vamos a considerar la tabla de datos `iris` vista ya anteriormente y vamos a elegir aleatoriamente 10 flores al azar:

```
set.seed(2020)
flores.elegidas = sample(1:150,10)
tabla.iris = iris[flores.elegidas,]
rownames(tabla.iris)=1:10
```

Vamos a calcular la matriz de distancias euclídeas entre las 30 flores anteriores usando las 4 variables continuas de la tabla de datos: la longitud y la amplitud del sépalo y del pétalo.

En primer lugar definimos una función que nos calcula la distancia euclídea entre dos vectores:

```
dist.euclidea = function(x,y){
  n=length(x)
  d = sqrt(sum((x-y)^2))
  return(d)
}
```

La matriz de distancias será la siguiente:

```
n=dim(tabla.iris)[1]
matriz.dist.iris = matrix(0,n,n)
for (i in 1:(n-1)){
  for (j in (i+1):n){
    matriz.dist.iris[i,j]=dist.euclidea(tabla.iris[i,1:4],tabla.iris[j,1:4])
    matriz.dist.iris[j,i]=matriz.dist.iris[i,j]
  }
}
```

Las distancias entre las 10 primeras flores vale:

```
round(matriz.dist.iris,2)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,] 0.00 3.79 1.58 3.89 1.29 0.71 1.08 4.30 1.71  3.86
##  [2,] 3.79 0.00 2.47 0.41 4.94 3.87 3.31 1.54 2.44  0.39
##  [3,] 1.58 2.47 0.00 2.67 2.54 1.48 0.91 2.80 0.45  2.54
##  [4,] 3.89 0.41 2.67 0.00 5.09 4.03 3.51 1.84 2.64  0.58
##  [5,] 1.29 4.94 2.54 5.09 0.00 1.19 1.74 5.23 2.63  5.00
##  [6,] 0.71 3.87 1.48 4.03 1.19 0.00 0.78 4.27 1.71  3.97
##  [7,] 1.08 3.31 0.91 3.51 1.74 0.78 0.00 3.60 1.06  3.38
##  [8,] 4.30 1.54 2.80 1.84 5.23 4.27 3.60 0.00 2.62  1.31
##  [9,] 1.71 2.44 0.45 2.64 2.63 1.71 1.06 2.62 0.00  2.44
## [10,] 3.86 0.39 2.54 0.58 5.00 3.97 3.38 1.31 2.44  0.00
```

9.3.4. Cálculo de distancias en R

La matriz de distancias **D** entre los objetos de nuestra tabla de datos se puede calcular con la función `dist`, cuya sintaxis básica es:

```
dist(x, method=...)
```

donde:

- `x` es nuestra tabla de datos (una matriz o un data frame de variables cuantitativas).
- `method` sirve para indicar la distancia que queremos usar, cuyo nombre se ha de entrar entrecomillado. La distancia por defecto es la euclídea que hemos venido usando hasta ahora. Otros posibles valores son (en lo que sigue, $\mathbf{x} = (x_1, \dots, x_m)$ e $y = (y_1, \dots, y_m)$ son dos vectores de $\mathbb{R}^m$):
  - La distancia de Manhattan, `manhattan`, que recordemos que vale $\sum_{i=1}^{m} |x_i - y_i|$.

- La distancia del máximo, `maximum`, que vale: $\max_{i=1,\ldots,m} |x_i - y_i|$.
- La distancia de Canberra, `canberra`, que vale $\sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i| + |y_i|}$.
- La distancia de Minkowski, `minkowski`, que depende de un parámetro $p > 0$ (que se ha de especificar en la función `dist` con `p` igual a su valor), y que vale: $\left(\sum_{i=1}^{m} |x_i - y_i|^p\right)^{1/p}$. Observad que cuando $p = 1$ se obtiene la distancia de Manhattan y cuando $p = 2$, la distancia euclídea usual.
- La distancia binaria, `binary`, que sirve básicamente para comparar vectores binarios (si los vectores no son binarios, `R` los entiende como binarios sustituyendo cada entrada diferente de 0 por 1). La distancia binaria entre `x` e `y` binarios es el número de posiciones en las que estos vectores tienen entradas diferentes, dividido por el número de posiciones en las que alguno de los dos vectores tiene un 1.

La salida de aplicar la función `dist` a nuestra tabla de datos es un objeto `dist` de `R`, no es una matriz de distancias usual.

Calculemos la matriz de distancias en el ejemplo anterior.

La matriz de distancias con la distancia euclídea de la tabla de datos de las 30 flores anteriores vale:

```
round(dist(tabla.iris[,1:4]),2)
```

```
##       1    2    3    4    5    6    7    8    9
## 2  3.79
## 3  1.58 2.47
## 4  3.89 0.41 2.67
## 5  1.29 4.94 2.54 5.09
## 6  0.71 3.87 1.48 4.03 1.19
## 7  1.08 3.31 0.91 3.51 1.74 0.78
## 8  4.30 1.54 2.80 1.84 5.23 4.27 3.60
## 9  1.71 2.44 0.45 2.64 2.63 1.71 1.06 2.62
## 10 3.86 0.39 2.54 0.58 5.00 3.97 3.38 1.31 2.44
```

Si queremos transformar la salida de la matriz de distancias anterior a una matriz de distancias "usual" hay que aplicarle la función `as.matrix`:

```
round(as.matrix(dist(tabla.iris[,1:4])),2)
```

```
##       1    2    3    4    5    6    7    8    9   10
## 1  0.00 3.79 1.58 3.89 1.29 0.71 1.08 4.30 1.71 3.86
## 2  3.79 0.00 2.47 0.41 4.94 3.87 3.31 1.54 2.44 0.39
## 3  1.58 2.47 0.00 2.67 2.54 1.48 0.91 2.80 0.45 2.54
## 4  3.89 0.41 2.67 0.00 5.09 4.03 3.51 1.84 2.64 0.58
## 5  1.29 4.94 2.54 5.09 0.00 1.19 1.74 5.23 2.63 5.00
## 6  0.71 3.87 1.48 4.03 1.19 0.00 0.78 4.27 1.71 3.97
## 7  1.08 3.31 0.91 3.51 1.74 0.78 0.00 3.60 1.06 3.38
## 8  4.30 1.54 2.80 1.84 5.23 4.27 3.60 0.00 2.62 1.31
## 9  1.71 2.44 0.45 2.64 2.63 1.71 1.06 2.62 0.00 2.44
## 10 3.86 0.39 2.54 0.58 5.00 3.97 3.38 1.31 2.44 0.00
```

9.3.5. Escalado de los datos

En el caso de que no queramos que la variación de las variables no influya en el posterior análisis del cálculo del **árbol binario** debido a que los datos están en escalas diferentes o para que la contribución a la matriz de distancias de cada una de las variables sea parecida, es conveniente que los datos estén en la misma escala.

En este caso, se escalan los datos de la forma siguiente: a los datos de la variable k o columna k, x_{ik}, $i = 1, \dots, n$ se les resta la media de dicha variable $\overline{x}_k = \frac{\sum_{i=1}^{n} x_{ik}}{n}$ y se dividen por su desviación estándar $\tilde{s}_k = \sqrt{\frac{n}{n-1}\left(\frac{\sum_{i=1}^{n} x_{ik}^2}{n} - \overline{x}_k^2\right)}$:

$$\tilde{x}_{ik} = \frac{x_{ik} - \overline{x}_k}{\tilde{s}_k},$$

donde $\tilde{x}_{ik}$ serían los nuevos valores de la tabla de datos escalada.

De esta forma, todas las m variables o columnas de nuestra tabla de datos tendrán media 0 y desviación estándar 1.

Observación: Cuando usamos la distancia euclídea, la distancia entre los objetos i y j escalados puede calcularse a partir de los valores iniciales x_{ik} y y_{jk}, $k = 1, \dots, m$ de la forma siguiente:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(x_{ik} - x_{jk})^2}{\tilde{s}_k^2}}.$$

Para escalar los datos en `R` hemos de usar la función `scale`:

```
scale(x, center = TRUE, scale = TRUE)
```

donde:

- `x`: nuestra tabla de datos.
- `center`: es un parámetro lógico o un vector numérico de longitud el número de columnas de `x` indicando la cantidad que queremos restar a los valores de cada variable o columna. Si vale `TRUE` que es su valor por defecto, a cada columna se le resta la media de dicha columna.
- `scale`: es un parámetro lógico o un vector numérico de longitud el número de columnas de `x` indicando la cantidad por la que queremos dividir los valores de cada variable o columna. Si vale `TRUE` que es su valor por defecto, a los valores de cada columna se les divide por la desviación estándar de dicha columna.

Las desviaciones estándar de las variables de la tabla de datos de la muestra de flores iris que hemos trabajado son bastante diferentes:

```
apply(tabla.iris[,1:4],2,sd)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##       0.6913       0.5181       1.7373       0.6289
```

Si no queremos que la variación de los datos intervenga en el análisis posterior que vamos a hacer, tenemos que escalar los datos:

```
tabla.iris.escalada = scale(tabla.iris[,1:4])
```

Comprobemos que las variables de la tabla de datos nueva tienen media 0 y desviación estándar 1:

```
apply(tabla.iris.escalada,2,mean)
```

```
##            Sepal.Length             Sepal.Width            Petal.Length
##  0.00000000000000013601 -0.00000000000000006665  0.00000000000000002776
##             Petal.Width
##  0.00000000000000002219
```

```
apply(tabla.iris.escalada,2,sd)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##            1            1            1            1
```

9.3.6. Clustering jerárquico aglomerativo

Una vez explicado con todo detalle cómo realizar el cálculo de una matriz de distancias a partir de una tabla de datos, vamos a desarrollar los pasos de los que consta el **clustering jerárquico aglomerativo**.

Recordemos que la salida del **algoritmo del clustering jerárquico aglomerativo** es un árbol binario donde se va indicando cómo se van agrupando los clusters partiendo de una configuración inicial en la que cada objeto forma un cluster.

Los pasos del algoritmo son los siguientes:

- Suponemos que partimos de n objetos, y de una matriz de **distancias D** de dimensiones $n \times n$, es decir, n filas y n columnas.
- En el primer paso, suponemos que cada objeto forma un cluster inicial.
- En el segundo paso, hallamos los dos clusters C_1 y C_2 cuya distancia sea la más pequeña de entre todos los pares de clusters. Si hay pares de clusters que empatan en la distancia mínima, escogemos un par al azar. Seguidamente, unimos C_1 y C_2 en un cluster nuevo que denotamos $C_1 + C_2$. ¡Ojo, la suma no significa sumar en el sentido clásico sino unir!
- Los clusters C_1 y C_2 que hemos unido en el paso anterior ya no existen. Por tanto, en el tercer paso, los eliminamos de la lista de clusters.
- En el cuarto paso, recalculamos la distancia del nuevo cluster $C_1 + C_2$ a los demás clusters tal como indicaremos más adelante. Tendremos una matriz de distancias $\mathbf{D}'$ con una fila menos y una columna menos.
- Repetimos los pasos 2, 3 y 4 hasta que sólo queda un único cluster y una matriz de distancias que sea un único número.

Tal como hemos comentado en el algoritmo anterior, falta indicar en el cuarto paso cómo se calcula la distancia del nuevo cluster $C_1 + C_2$ a los demás clusters. Dicha distancia se puede calcular de varias maneras, dando lugar a resultados diferentes o a tipos de clusters diferentes:

- Por **enlace simple**: la distancia entre dos clusters C y C' cualquiera se calcula como: $d(C, C') = \text{mín}\{d(a, b) \mid a \in C, b \in C'\}$.

Por tanto, la distancia del nuevo cluster $C_1 + C_2$ a un cluster C sería:

$$d(C, C_1 + C_2) = \text{mín}\{d(C, C_1), d(C, C_2)\}.$$

- Por **enlace completo**: la distancia entre dos clusters C y C' cualquiera se calcula como: $d(C, C') = \text{máx}\{d(a, b) \mid a \in C, b \in C'\}$.

Por tanto, la distancia del nuevo cluster $C_1 + C_2$ a un cluster C sería:

$$d(C, C_1 + C_2) = \text{máx}\{d(C, C_1), d(C, C_2)\}.$$

- Por enlace **medio**: la distancia entre dos clusters C y C' cualquiera se calcula como: $d(C, C') = \dfrac{\sum\limits_{a \in C, b \in C'} d(a, b)}{|C| \cdot |C'|}$, donde $|C|$ indica el tamaño del cluster C o el número de elementos del mismo.

Por tanto, la distancia del nuevo cluster $C_1 + C_2$ a un cluster C sería:

$$d(C, C_1 + C_2) = \frac{|C_1|}{|C_1| + |C_2|} d(C, C_1) + \frac{|C_2|}{|C_1| + |C_2|} d(C, C_2).$$

Las expresiones anteriores son casos particulares de la siguiente regla general: suponiendo conocidas las distancias siguientes:

$$d(C, C_1), \quad d(C, C_2), \quad d(C_1, C_2),$$

la formula genérica para hallar $d(C, C_1 + C_2)$ es la siguiente:

$$\begin{aligned} d(C, C_1 + C_2) = \quad & \delta_1 d(C, C_1) + \delta_2 d(C, C_2) + \delta_3 d(C_1, C_2) \\ & + \delta_0 |d(C, C_1) - d(C, C_2)|, \end{aligned}$$

donde los δ_i son parámetros a elegir. Cada elección da lugar a un algoritmo diferente, con resultados posiblemente diferentes como se muestra en la tabla siguiente:

Nombre	δ_1	δ_2	δ_3	δ_0
Enlace simple	$1/2$	$1/2$	0	$-1/2$
Enlace completo	$1/2$	$1/2$	0	$1/2$
Enlace promedio	$\frac{\lvert C_1\rvert}{\lvert C_1\rvert+\lvert C_2\rvert}$	$\frac{\lvert C_2\rvert}{\lvert C_1\rvert+\lvert C_2\rvert}$	0	0
Centroide	$\frac{\lvert C_1\rvert}{\lvert C_1\rvert+\lvert C_2\rvert}$	$\frac{\lvert C_2\rvert}{\lvert C_1\rvert+\lvert C_2\rvert}$	$-\frac{\lvert C_1\rvert\lvert C_2\rvert}{(\lvert C_1\rvert+\lvert C_2\rvert)^2}$	0
Mediana	$1/2$	$1/2$	$-1/4$	0
Ward	$\frac{\lvert C\rvert+\lvert C_1\rvert}{\lvert C\rvert+\lvert C_1\rvert+\lvert C_2\rvert}$	$\frac{\lvert C\rvert+\lvert C_2\rvert}{\lvert C\rvert+\lvert C_1\rvert+\lvert C_2\rvert}$	$-\frac{\lvert C\rvert}{\lvert C\rvert+\lvert C_1\rvert+\lvert C_2\rvert}$	0

Ejercicio

Demostrar que para el **enlace simple** y para el **enlace completo** las fórmulas dadas anteriormente y la fórmulas dadas por la tabla anterior son equivalentes.

Más concretamente, demostrar que dados C_1, C_2 y C clusters,

- **Enlace simple**:

$$d(C, C_1 + C_2) = \text{mín}\{d(C, C_1), d(C, C_2)\} = \frac{1}{2}d(C, C_1) + \frac{1}{2}d(C, C_2) - \frac{1}{2}|d(C, C_1) - d(C, C_2)|.$$

- **Enlace completo**:

$$d(C, C_1 + C_2) = \text{máx}\{d(C, C_1), d(C, C_2)\} = \frac{1}{2}d(C, C_1) + \frac{1}{2}d(C, C_2) + \frac{1}{2}|d(C, C_1) - d(C, C_2)|.$$

Ejemplo de la muestra de flores iris

Vamos a aplicar el método del enlace simple a la muestra de flores de la tabla de datos `iris`.

Usaremos los datos sin escalar. Dejamos como ejercicio repetir este ejemplo con los datos escalados.

Recordemos que la matriz de distancias estaba en la matriz `matriz.dist.iris`.

Hallemos las dos flores distintas que tienen distancia mínima que no estén en la diagonal ya que los valores de la diagonal en la matriz de distancias valen 0 y éstas no nos interesan ya que corresponden a las distancias entre todas las flores iguales.

Una forma de hacerlo (no la única) es definir otra matriz asignando el valor máximo de nuestra matriz de distancias a los valores de la diagonal y hallar el mínimo de la nueva matriz. De esta manera, no nos saldrán los valores de la diagonal:

```
matriz.nueva=matriz.dist.iris
diag(matriz.nueva)=max(matriz.dist.iris)
(flores.min=which(matriz.nueva == min(matriz.nueva), arr.ind = TRUE))
```

```
##      row col
## [1,]  10   2
## [2,]   2  10
```

Vemos que las flores 10 y flores 2 son las más cercanas con una distancia de 0.3873.

Definimos el cluster nuevo $\{10, 2\}$ y eliminamos las flores 10 y flores 2.

Los nuevos clusters serán:

```
## [1] 1
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
```

```
## row col
##  10   2
```

A continuación tenemos que hallar la distancia del nuevo cluster $\{10, 2\}$ a los demás clusters que en este caso serán clusters con una sola flor usando la expresión del **enlace simple**: $d(C, C_1 + C_2) = \text{mín}\{d(C, C_1), d(C, C_2)\}$:

$$
\begin{aligned}
d(\{1\},\{10,2\}) &= \text{mín}\{d(\{1\},\{10\}), d(\{1\},\{2\})\} = \text{mín}\{3{,}8626, 3{,}7908\} = 3{,}7908,\\
d(\{3\},\{10,2\}) &= \text{mín}\{d(\{3\},\{10\}), d(\{3\},\{2\})\} = \text{mín}\{2{,}5417, 2{,}4718\} = 2{,}4718,\\
d(\{4\},\{10,2\}) &= \text{mín}\{d(\{4\},\{10\}), d(\{4\},\{2\})\} = \text{mín}\{0{,}5831, 0{,}4123\} = 0{,}4123,\\
d(\{5\},\{10,2\}) &= \text{mín}\{d(\{5\},\{10\}), d(\{5\},\{2\})\} = \text{mín}\{4{,}998, 4{,}9447\} = 4{,}9447,\\
d(\{6\},\{10,2\}) &= \text{mín}\{d(\{6\},\{10\}), d(\{6\},\{2\})\} = \text{mín}\{3{,}9724, 3{,}8743\} = 3{,}8743,\\
d(\{7\},\{10,2\}) &= \text{mín}\{d(\{7\},\{10\}), d(\{7\},\{2\})\} = \text{mín}\{3{,}3808, 3{,}3136\} = 3{,}3136,\\
d(\{8\},\{10,2\}) &= \text{mín}\{d(\{8\},\{10\}), d(\{8\},\{2\})\} = \text{mín}\{1{,}3115, 1{,}5395\} = 1{,}3115,\\
d(\{9\},\{10,2\}) &= \text{mín}\{d(\{9\},\{10\}), d(\{9\},\{2\})\} = \text{mín}\{2{,}4372, 2{,}4352\} = 2{,}4352.
\end{aligned}
$$

La nueva matriz de distancias entre los clusters:

$$\{1\},\ \{3\},\ \{4\},\ \{5\},\ \{6\},\ \{7\},\ \{8\},\ \{9\},\ \{10,2\},$$

es la siguiente:

```
##        [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
##  [1,] 0.000 1.581 3.886 1.288 0.707 1.082 4.299 1.715 3.791
##  [2,] 1.581 0.000 2.672 2.542 1.483 0.911 2.804 0.447 2.472
##  [3,] 3.886 2.672 0.000 5.085 4.027 3.514 1.838 2.642 0.412
##  [4,] 1.288 2.542 5.085 0.000 1.192 1.741 5.233 2.634 4.945
##  [5,] 0.707 1.483 4.027 1.192 0.000 0.781 4.273 1.709 3.874
##  [6,] 1.082 0.911 3.514 1.741 0.781 0.000 3.596 1.063 3.314
##  [7,] 4.299 2.804 1.838 5.233 4.273 3.596 0.000 2.619 1.311
##  [8,] 1.715 0.447 2.642 2.634 1.709 1.063 2.619 0.000 2.435
##  [9,] 3.791 2.472 0.412 4.945 3.874 3.314 1.311 2.435 0.000
```

Observamos que tiene una dimensión menor y es una matriz de 9 filas y 9 columnas.

El siguiente paso es hallar los dos clusters siguientes con distancia mínima.

Haremos el mismo truco que hemos hecho anteriormente: (la nueva matriz de distancias la hemos guardado en `matriz.dist.iris2`)

```
matriz.nueva=matriz.dist.iris2
diag(matriz.nueva)=max(matriz.dist.iris2)
(flores.min2 = which(matriz.nueva == min(matriz.nueva),arr.ind=TRUE))
```

```
##      row col
## [1,]   9   3
## [2,]   3   9
```

Vemos que los clusters $\{10,2\}$ y $\{4\}$ son las más cercanos con una distancia de 0.4123. Acordarse que la fila 9 correspondía al cluster que hemos unido en el paso anterior $\{10,2\}$.

Definimos el cluster nuevo $\{10,2,4\}$ y eliminamos los clusters $\{10,2\}$ y $\{4\}$.

Los nuevos clusters serán:

```
## [1] 1
## [1] 3
## [1] 5
## [1] 6
## [1] 7
```

```
## [1] 8
## [1] 9
```

```
## col     row
##   2   4  10
```

A continuación tenemos que hallar la distancia del nuevo cluster $\{10, 2, 4\}$ a los demás clusters que en este caso serán clusters con una sola flor usando la expresión del **enlace simple**: $d(C, C_1 + C_2) = \text{mín}\{d(C, C_1), d(C, C_2)\}$:

$$\begin{aligned}
d(\{1\},\{10,2,4\}) &= \text{mín}\{d(\{1\},\{10,2\}), d(\{1\},\{4\})\} = \text{mín}\{3{,}7908, 3{,}8859\} = 3{,}7908,\\
d(\{3\},\{10,2,4\}) &= \text{mín}\{d(\{3\},\{10,2\}), d(\{3\},\{4\})\} = \text{mín}\{2{,}4718, 2{,}6721\} = 2{,}4718,\\
d(\{5\},\{10,2,4\}) &= \text{mín}\{d(\{5\},\{10,2\}), d(\{5\},\{4\})\} = \text{mín}\{4{,}9447, 5{,}0853\} = 4{,}9447,\\
d(\{6\},\{10,2,4\}) &= \text{mín}\{d(\{6\},\{10,2\}), d(6\},\{4\})\} = \text{mín}\{3{,}8743, 4{,}0274\} = 3{,}8743,\\
d(\{7\},\{10,2,4\}) &= \text{mín}\{d(\{7\},\{10,2\}), d(\{7\},\{4\})\} = \text{mín}\{3{,}3136, 3{,}5143\} = 3{,}3136,\\
d(\{8\},\{10,2,4\}) &= \text{mín}\{d(\{8\},\{10,2\}), d(\{8\},\{4\})\} = \text{mín}\{1{,}3115, 1{,}8385\} = 1{,}3115,\\
d(\{9\},\{10,2,4\}) &= \text{mín}\{d(\{9\},\{10,2\}), d(\{9\},\{4\})\} = \text{mín}\{2{,}4352, 2{,}642\} = 2{,}4352.
\end{aligned}$$

La nueva matriz de distancias entre los clusters:

$$\{1\},\ \{3\},\ \{5\},\ \{6\},\ \{7\},\ \{8\},\ \{9\},\ \{10, 2, 4\},$$

es la siguiente:

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
## [1,] 0.000 1.581 1.288 0.707 1.082 4.299 1.715 3.791
## [2,] 1.581 0.000 2.542 1.483 0.911 2.804 0.447 2.472
## [3,] 1.288 2.542 0.000 1.192 1.741 5.233 2.634 4.945
## [4,] 0.707 1.483 1.192 0.000 0.781 4.273 1.709 3.874
## [5,] 1.082 0.911 1.741 0.781 0.000 3.596 1.063 3.314
## [6,] 4.299 2.804 5.233 4.273 3.596 0.000 2.619 1.311
## [7,] 1.715 0.447 2.634 1.709 1.063 2.619 0.000 2.435
## [8,] 3.791 2.472 4.945 3.874 3.314 1.311 2.435 0.000
```

Observamos que tiene una dimensión menor y es una matriz de 8 filas y 8 columnas.

El siguiente paso es hallar los dos clusters siguientes con distancia mínima.

Haremos el mismo truco que hemos hecho en los dos pasos anteriores: (la nueva matriz de distancias la hemos guardado en `matriz.dist.iris3`)

```
matriz.nueva=matriz.dist.iris3
diag(matriz.nueva)=max(matriz.dist.iris3)
(flores.min3 = which(matriz.nueva == min(matriz.nueva),arr.ind=TRUE))
```

```
##      row col
## [1,]   7   2
## [2,]   2   7
```

Vemos que los clusters $\{9\}$ y $\{3\}$ son las más cercanos con una distancia de 0.4472.

Definimos el cluster nuevo $\{3, 9\}$ y eliminamos los clusters $\{3\}$ y $\{9\}$

Los nuevos clusters serán:

```
## [1] 1
## [1] 5
## [1] 6
## [1] 7
## [1] 8
```

```
## col     row
##   2   4  10
```

```
## [1] 3 9
```

A continuación tenemos que hallar la distancia del nuevo cluster $\{3,9\}$ a los demás clusters que en este caso serán clusters con una sola flor usando la expresión del **enlace simple**: $d(C, C_1 + C_2) = \text{mín}\{d(C, C_1), d(C, C_2)\}$:

$$\begin{aligned}
d(\{1\},\{3,9\}) &= \text{mín}\{d(\{1\},\{3\}), d(\{1\},\{9\})\} = \text{mín}\{1{,}7146, 1{,}5811\} = 1{,}5811,\\
d(\{5\},\{3,9\}) &= \text{mín}\{d(\{5\},\{3\}), d(\{5\},\{9\})\} = \text{mín}\{2{,}6344, 2{,}5417\} = 2{,}5417,\\
d(\{6\},\{3,9\}) &= \text{mín}\{d(\{6\},\{3\}), d(\{6\},\{9\})\} = \text{mín}\{1{,}7088, 1{,}4832\} = 1{,}4832,\\
d(\{7\},\{3,9\}) &= \text{mín}\{d(\{7\},\{3\}), d(\{7\},\{9\})\} = \text{mín}\{1{,}063, 0{,}911\} = 0{,}911,\\
d(\{8\},\{3,9\}) &= \text{mín}\{d(\{8\},\{3\}), d(\{8\},\{9\})\} = \text{mín}\{2{,}6192, 2{,}8036\} = 2{,}6192,\\
d(\{2,4,10\},\{3,9\}) &= \text{mín}\{d(\{2,4,10\},\{3\}), d(\{2,4,10\},\{9\})\} = \text{mín}\{2{,}4352, 2{,}4718\}\\
&= 2{,}4352.
\end{aligned}$$

La nueva matriz de distancias entre los clusters:

$$\{1\},\ \{5\},\ \{6\},\ \{7\},\ \{8\},\ \{2,4,10\},\ \{3,9\},$$

es la siguiente:

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
## [1,] 0.000 1.288 0.707 1.082 4.299 3.791 1.581
## [2,] 1.288 0.000 1.192 1.741 5.233 4.945 2.542
## [3,] 0.707 1.192 0.000 0.781 4.273 3.874 1.483
## [4,] 1.082 1.741 0.781 0.000 3.596 3.314 0.911
## [5,] 4.299 5.233 4.273 3.596 0.000 1.311 2.619
## [6,] 3.791 4.945 3.874 3.314 1.311 0.000 2.435
## [7,] 1.581 2.542 1.483 0.911 2.619 2.435 0.000
```

Observamos que tiene una dimensión menor y es una matriz de 7 filas y 7 columnas.

Y así sucesivamente hasta tener un solo cluster.

9.3.7. Clustering jerárquico aglomerativo con R

La función de R para realizar un clustering jerárquico aglomerativo es `hclust`:

```
hclust(d, method = ...)
```

donde:

- `d` es la matriz de distancias entre nuestros objetos calculada con la función `dist`.

- `method` sirve para especificar cómo se define la distancia de la unión de dos clusters al resto de los clusters. El nombre del método se ha de entrar entrecomillado. Los más populares son los siguientes:
  - Método de enlace completo, `complete`,
  - Método de enlace simple, `single`,
  - Método de enlace promedio, `average`,
  - Método de la mediana, `median`,
  - Método del centroide, `centroid`,
  - Método de Ward clásico, `ward.D`.

Para la interpretación de la salida, imaginemos que hemos guardado en el objeto `estudio.clustering` el resultado de la función `hclust`:

```
estudio.clustering = hclust(d,method=...)
```

entonces:

- `estudio.clustering$merge` nos indica cómo ha ido agrupando los clusters. Es una matriz de $n-1$ filas y 2 columnas, donde n es el número de objetos. En esta matriz, los objetos originales se representan con números negativos, y los nuevos clusters con números positivos que indican el paso en el que se han creado.
- `estudio.clustering$height` es un vector que contiene las distancias a las que se han ido agrupando los pares de clusters.

Ejemplo de la muestra de flores iris

Para realizar el estudio de la muestra de flores de la tabla de datos `iris` hacemos los siguiente:

```
estudio.clustering.iris = hclust(dist(tabla.iris[,1:4]),method="single")
```

Veamos cómo ha realizado los agrupamientos:

```
estudio.clustering.iris$merge
```

```
##       [,1] [,2]
##  [1,]   -2  -10
##  [2,]   -4    1
##  [3,]   -3   -9
##  [4,]   -1   -6
##  [5,]   -7    4
##  [6,]    3    5
##  [7,]   -5    6
##  [8,]   -8    2
##  [9,]    7    8
```

`R` primero agrupa las flores 2 y 10 tal como hemos hecho anteriormente a mano.

Seguidamente agrupa la flor número 4 con el cluster creado en el paso 1 ($\{2,10\}$) creando el nuevo cluster $\{2,4,10\}$.

Seguidamente agrupa las flores 3 y 9 creando el cluster $\{3,9\}$.

Fijarse que los tres pasos coinciden con los pasos hechos a mano.

En el cuarto paso agrupa las flores 1 y 6 creando el cluster $\{1, 6\}$.

En el quinto, la flor 7 con el cluster creado en el paso 4 ($\{1, 6\}$) creando el cluster $\{1, 6, 7\}$.

En el sexto paso agrupa los clusters creados en el paso 3 $\{3, 9\}$ y en el paso 5 $\{1, 6, 7\}$ creando el cluster $\{1, 3, 6, 7, 9\}$.

En el séptimo paso agrupa la flor número 5 con el cluster creado en el sexto paso ($\{1, 3, 6, 7, 9\}$) creando el cluster $\{1, 3, 5, 6, 7, 9\}$.

En el octavo paso agrupa la flor número 8 con el cluster creado en el segundo paso ($\{2, 10, 4\}$) creando el cluster $\{2, 4, 8, 10\}$.

Por último, en el noveno paso, agrupa los clusters creados en el paso 7 ($\{1, 3, 5, 6, 7, 9\}$) y en el paso 8 ($\{2, 4, 8, 10\}$) creando el cluster formado por todas las flores de la muestra.

Para ver a qué distancias ha realizado los agrupamientos, hacemos lo siguiente:

```
estudio.clustering.iris$height
```

```
## [1] 0.3873 0.4123 0.4472 0.7071 0.7810 0.9110 1.1916 1.3115 2.4352
```

Fijarse que las tres primeras distancias corresponden a las distancias con las que hemos creado los nuevos clusters en los tres primeros pasos realizados a mano.

9.3.8. Gráfico del árbol binario o dendrograma

La visualización de los pasos del **clustering aglomerativo** se realiza mediante un árbol binario denominado **dendrograma**.

Para visualizarlo en `R` basta usar la función `plot`:

```
plot(estudio.clustering,hang=...,labels=...)
```

donde

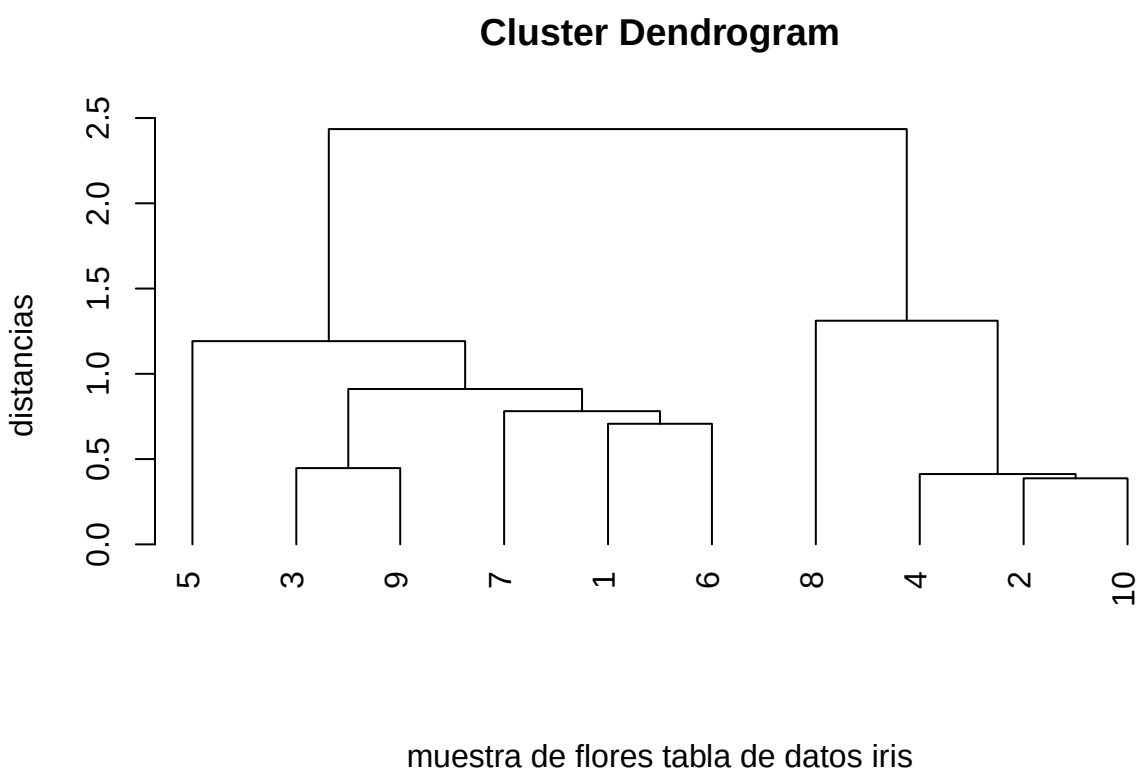
- `hang` es un parámetro que controla la situación de las hojas del dendrograma respecto del margen inferior.
- `labels` es un vector de caracteres que permite poner nombres a los objetos; por defecto, se identifican en la representación gráfica por medio de sus números de fila en la matriz o el data frame que contiene los datos.

Ejemplo de la muestra de flores iris

El **dendrograma** para estudio del clustering aglomerativo de los datos de la muestra de flores iris es el siguiente:

```
plot(estudio.clustering.iris,hang=-1,xlab="muestra de flores tabla de datos iris",
     sub="", ylab="distancias",labels=1:10)
```

Fijaos como los agrupamientos comentados anteriormente se visualizan en el **dendrograma**.



9.3.9. ¿Cómo calcular los clusters definidos por un clustering jerárquico?

Un **clustering jerárquico** puede usarse para definir un clustering ordinario, es decir, una clasificación de los objetos bajo estudio dando los clusters correspondientes.

Dichos clusters se pueden hallar de dos maneras: indicando cuántos clusters deseamos, o indicando a qué altura queremos cortar el dendrograma, de manera que clusters que se unan a una distancia mayor que dicha altura queden separados.

En el dendrograma del ejemplo anterior, si sólo queremos tres clusters, éstos deben ser:

$$\{1,3,5,6,7,9\},\ \{8\},\ \{2,4,10\}.$$

En cambio, si cortamos por una distancia de 1.5, sólo aparecerán dos clusters:

$$\{1,3,5,6,7,9\},\ \{2,4,8,10\}.$$

Para calcular los clusters en `R` debemos usar la función `cutree`:

```
cutree(estudio.clustering,k=...,h=...)
```

donde:

- `k` es un parámetro que indica el número de clusters deseado

- `h` es un parámetro que indica la altura a la que queremos cortar.

Ejemplo de la muestra de flores iris

En el ejemplo de la muestra de flores iris, si sólo queremos tres clusters para el estudio de **clustering aglomerativo**, hemos de hacer lo siguiente:

```
cutree(estudio.clustering.iris,k=3)
```

```
##  1  2  3  4  5  6  7  8  9 10
##  1  2  1  2  1  1  1  3  1  2
```

Observamos que `R` nos ha dado los mismos clusters que hemos indicado anteriormente.

Si cortamos a una distancia de 1.5, los clusters serán:

```
cutree(estudio.clustering.iris,h=1.5)
```

```
##  1  2  3  4  5  6  7  8  9 10
##  1  2  1  2  1  1  1  2  1  2
```

dando los mismos clusters vistos anteriormente.

Para visualizar los clusters en el **dendrograma** hemos de usar la función `rect.hclust`:

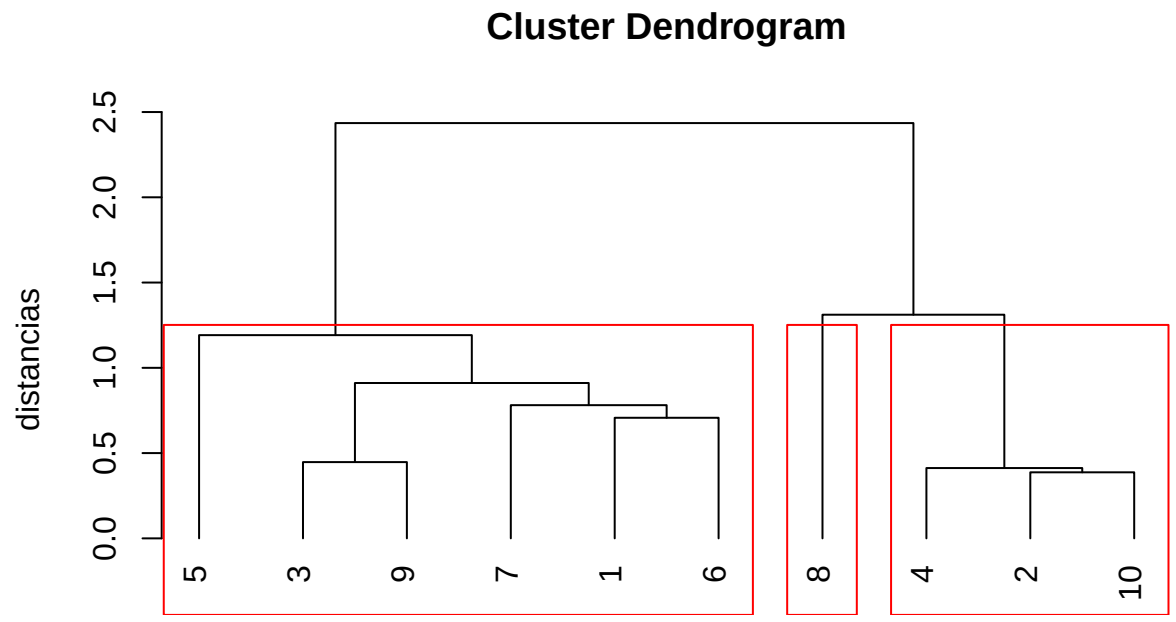
```
plot(estudio.clustering,...)
rect.hclust(estudio.clustering,h=...,k=...)
```

Fijarse que antes de llamar a la función `rect.hclust` hemos de realizar el **dendrograma** usando la función `plot`.

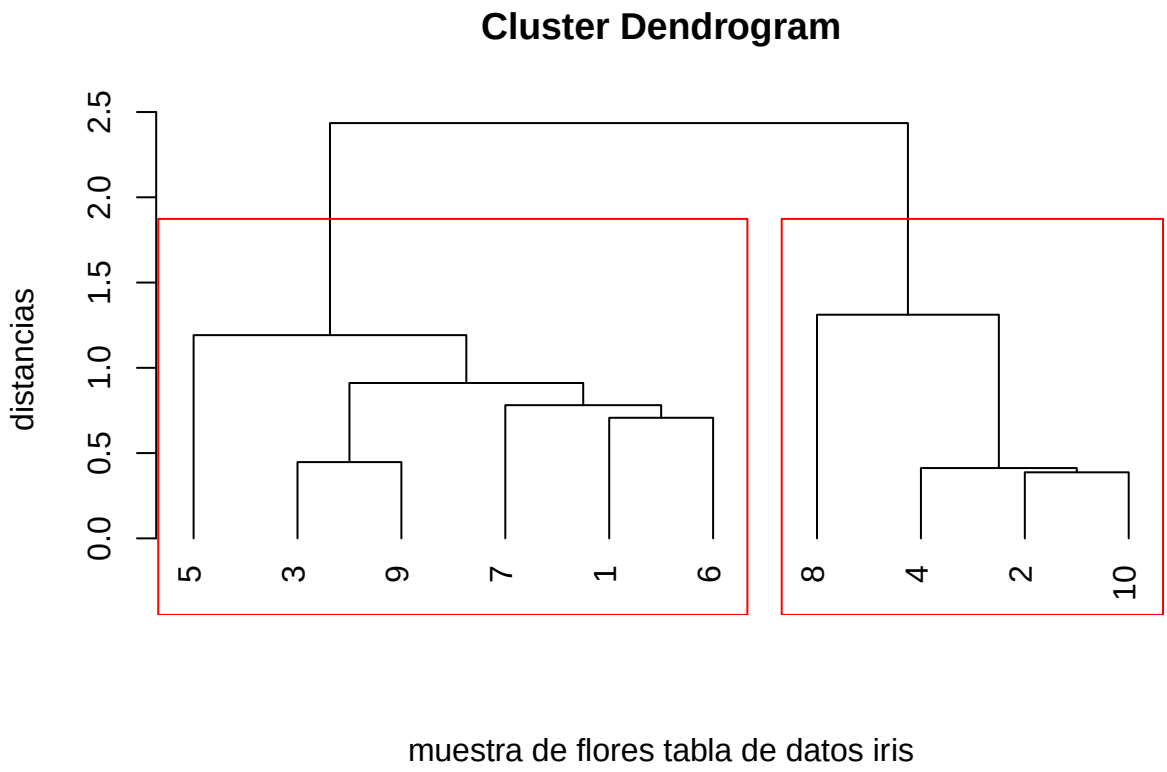
Ejemplo de la muestra de flores iris

Visualicemos los clusters anteriores:

```
plot(estudio.clustering.iris,hang=-1,xlab="muestra de flores tabla de datos iris",
     sub="", ylab="distancias",labels=1:10)
rect.hclust(estudio.clustering.iris,k=3)
```




```
plot(estudio.clustering.iris,hang=-1,xlab="muestra de flores tabla de datos iris",
     sub="", ylab="distancias",labels=1:10)
rect.hclust(estudio.clustering.iris,h=1.5)
```



9.3.10. Propiedades de los métodos en el clustering jerárquico

Si usamos el método del **enlace simple**, donde recordemos que

$$d(C, C_1 + C_2) = \text{mín}\left(d(C, C_1), d(C, C_2)\right),$$

tiende a construir clusters grandes: clusters que tendrían que ser diferentes pero que tienen dos individuos próximos se unen en un único cluster.

Si usamos el método del **enlace completo**, donde recordemos que

$$d(C, C_1 + C_2) = \text{máx}\left(d(C, C_1), d(C, C_2)\right),$$

se comporta de forma totalmente diferente agrupando clusters solo cuando todos los puntos están próximos.

Si usamos el método de **enlace promedio**, donde recordemos que

$$d(C, C_1 + C_2) = \frac{|C_1|}{|C_1| + |C_2|} d(C, C_1) + \frac{|C_2|}{|C_1| + |C_2|} d(C, C_2),$$

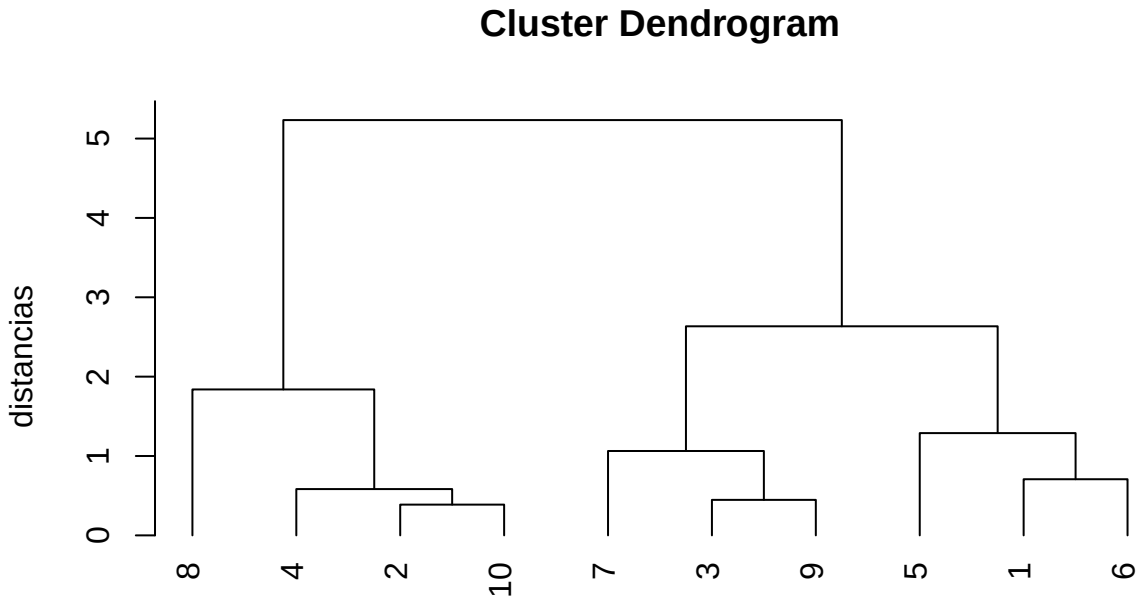
se comportaría como una solución intermedia entre los dos métodos anteriores.

Dicho método es muy usado en la reconstrucción de árboles filogenéticos a partir de matrices de distancias (método *UPGMA*, *Unweighted Pair Group Method Using Arithmetic Averages*)

Ejemplo de la muestra de flores iris

Si en lugar de usar el método del **enlace simple** usamos el método del **enlace completo** en la muestra de la tabla de datos de flores iris, obtenemos el dendrograma siguiente:

```
estudio.clustering.iris.completo=hclust(dist(tabla.iris[,1:4]),method="complete")
plot(estudio.clustering.iris.completo,hang=-1,xlab="muestra de flores tabla de datos iris",
     sub="", ylab="distancias",labels=1:10)
```



Si cortamos a la misma distancia que antes, `h=1.5` obtenemos los clusters siguientes:

```
cutree(estudio.clustering.iris.completo,h=1.5)
```

```
##  1  2  3  4  5  6  7  8  9 10
##  1  2  3  2  1  1  3  4  3  2
```

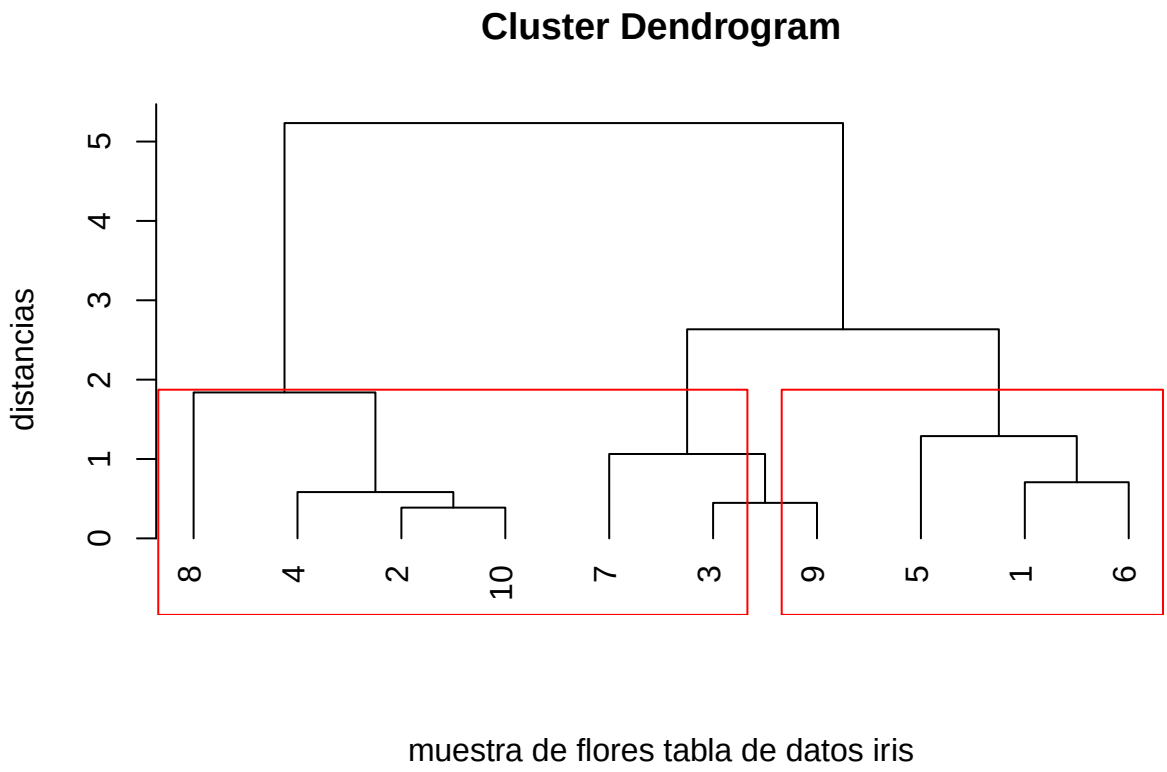
Observamos que ahora nos ha creado 4 clusters:

$$\{1,5,6\},\ \{2,4,10\},\ \{7,9\},\ \{8\},$$

comportándose tal como hemos indicado, es decir, "le cuesta" hacer clusters grandes.

Visualicemos el resultado:

```
plot(estudio.clustering.iris.completo,hang=-1,
     xlab="muestra de flores tabla de datos iris",
     sub="", ylab="distancias",labels=1:10)
rect.hclust(estudio.clustering.iris,h=1.5)
```



Ejemplo de los arrestados por posesión de cantidades pequeñas de marihuana

Realizemos el análisis de clustering aglomerativo usando el algoritmo de Ward para los datos binarios de este ejemplo.

Como la matriz de distancias es una matriz de **similaridad** y la matriz de distancias que tenemos que considerar es una matriz de **disimilaridad** al usar la función `hclust`, hacemos `1-matriz.dist.hamming` para tener una matriz de **disimilaridad**.

Otra cosa a tener en cuenta es que la función `hclust` sólo admite objetos tipo `dist`, por tanto, tenemos que transformar la matriz de distancias en un objeto `dist`.

Sin más preámbulos, guardamos en el objeto `estudio.clustering.arrestados` el análisis del clustering jerárquico:

```
estudio.clustering.arrestados=hclust(as.dist(1-matriz.dist.hamming),method="ward.D")
```

La visualización de los resultados es la siguiente:

```
plot(estudio.clustering.arrestados,hang=-1,xlab="muestra de arrestados",
     sub="", ylab="distancias")
```

Observamos que, como hay muchas distancias con valores 0, tenemos 12 arrestados en un cluster inicial que se unen a distancia 0: $\{2, 3, 4, 5, 6, 7, 11, 13, 21, 23, 24, 25\}$. También se unen a dicha distancia los arrestados 20 y 22 en el cluster $\{20, 22\}$, los arrestados 8 y 17 en el cluster $\{8, 17\}$, los arrestados 1, 14 y 19 en el cluster $\{1, 14, 19\}$ y los arrestados 9, 10 y 15 en el cluster $\{9, 10, 15\}$.

A continuación a distancia $\frac{1}{3}$, se unen los clusters $\{18\}$, y $\{8, 17\}$ en el cluster $\{8, 17, 18\}$.

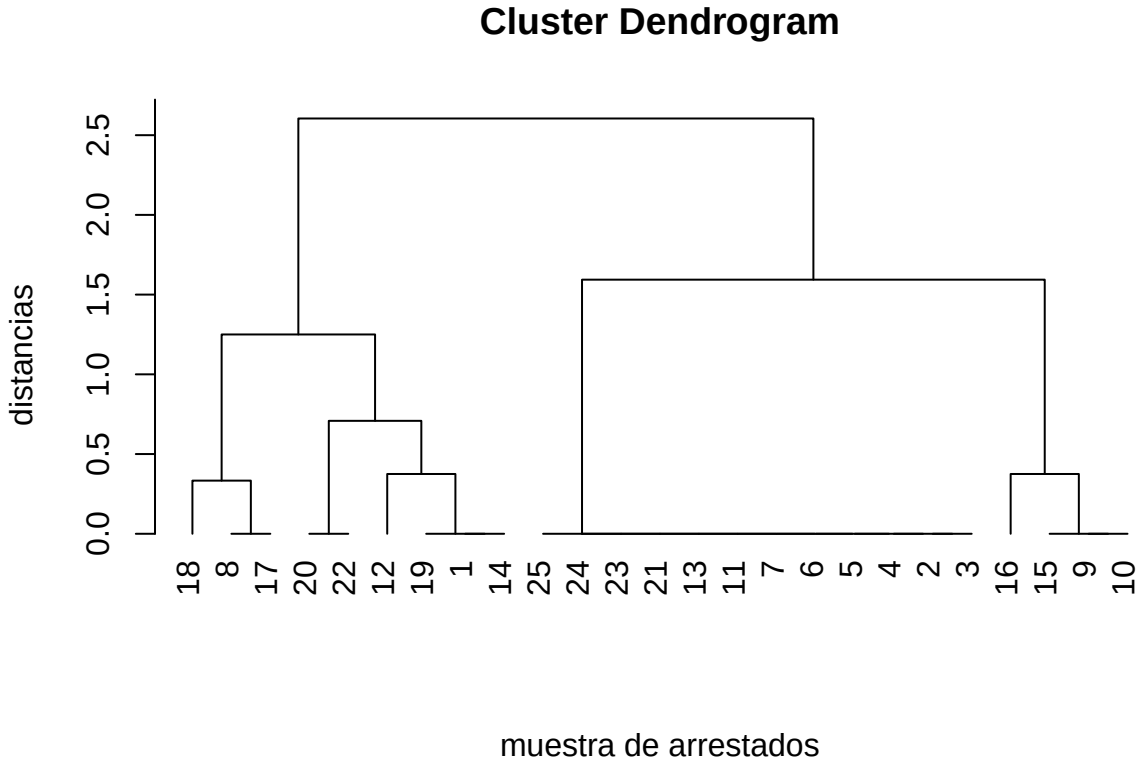
Después, a distancia 0.375, se unen los clusters $\{12\}$ y $\{1, 14, 19\}$ en el cluster $\{1, 12, 14, 19\}$ y los clusters $\{16\}$ y $\{9, 10, 15\}$ en el cluster $\{9, 10, 15, 16\}$.

El siguiente paso es la unión a distancia 0.7083333 de los clusters $\{20, 22\}$ y $\{1, 12, 14, 19\}$ en el cluster $\{1, 12, 14, 19, 20, 22\}$.

El siguiente paso es la unión a distancia 1.25 de los clusters $\{8, 17, 18\}$ y $\{1, 12, 14, 19, 20, 22\}$ en el cluster $\{1, 8, 12, 14, 18, 19, 20, 22\}$.

El penúltimo paso es la unión a distancia 1.59375 de los clusters $\{2, 3, 4, 5, 6, 7, 11, 13, 21, 23, 24, 25\}$ y $\{9, 10, 15, 16\}$ en el cluster $\{2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 15, 16, 21, 23, 24, 25\}$.

El último paso es la unión a distancia 2.6045833 de los clusters $\{1, 8, 12, 14, 18, 19, 20, 22\}$ y $\{2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 15, 16, 21, 23, 24, 25\}$ en el cluster formado por todos los arrestados.



9.4. Guía rápida

- `kmeans(x, centers=..., iter.max=..., algorithm =...)`: aplica el algoritmo de k-medias a una tabla de datos, donde:
  - `x` es la matriz o el data frame cuyas filas representan los objetos; en ambos casos, todas las variables han de ser numéricas como ya hemos indicado.

  - `centers` sirve para especificar los centros iniciales, y se puede usar de dos maneras: igualado a un número k, `R` escoge aleatoriamente los k centros iniciales, mientras que igualado a una matriz de k filas y el mismo número de columnas que `x`, `R` toma las filas de esta matriz como centros de partida.
  - `iter.max` permite especificar el número máximo de iteraciones a realizar; su valor por defecto es 10. Al llegar a este número máximo de iteraciones, si el algoritmo aún no ha acabado porque los clusters aún no hayan estabilizado, se para y da como resultado los clusters que se han obtenido en la última iteración.
  - `algorithm` indica el algoritmo a usar. Los algoritmos pueden ser los siguientes: `Hartigan-Wong`, `Lloyd`, `MacQueen`.

- Salida de `kmeans` (suponemos que hemos guardado en el objeto `resultado.km` la salida del algoritmo):
  - `resultado.km$size`: nos da los números de objetos, es decir, los tamaños de cada cluster.
  - `resultado.km$cluster`: nos dice qué cluster pertenece cada uno de los objetos de nuestra tabla de datos.
  - `resultado.km$centers`: nos da los centros de cada cluster en filas. Es decir, la fila 1 sería el centro del cluster 1, la fila 2, del cluster 2 y así sucesivamente.
  - `resultado.km$withinss`: nos da las sumas de cuadrados de cada cluster, lo que antes hemos denominado SS_{C_j}, para $j = 1, \dots, k$.
  - `resultado.km$tot.withinss`: la suma de cuadrados de todos los clusters, lo que antes hemos denominado SS_C. También se puede calcular sumando las sumas de los cuadrados de cada cluster: `sum(resultado.km$withinss)`.
  - `resultado.km$totss`: es la suma de los cuadrados de las distancias de los puntos en su punto medio de todos estos puntos. Es decir, seria la suma de los cuadrados SS_C pero suponiendo que sólo hubiera un sólo cluster.
  - `resultado.km$betweenss` es la diferencia entre `resultado.km$totss` y `resultado.km$tot.withinss` y puede demostrarse (es un cálculo bastante tedioso) que es igual a la suma, ponderada por el número de objetos del cluster correspondiente, de los cuadrados de las distancias de los centros de los clusters al punto medio de todos los puntos.

- `dist(x, method=...)`: calcula la matriz de distancias de una tabla de datos, donde:
  - `x` es nuestra tabla de datos (una matriz o un data frame de variables cuantitativas).
  - `method` sirve para indicar la distancia que queremos usar, cuyo nombre se ha de entrar entrecomillado. La distancia por defecto es la euclídea que hemos venido usando hasta ahora. Otros posibles valores son (en lo que sigue, $\mathbf{x} = (x_1, \dots, x_m)$ e $y = (y_1, \dots, y_m)$ son dos vectores de $\mathbb{R}^m$):
  - La distancia de Manhattan, `manhattan`, que recordemos que vale $\sum_{i=1}^{m} |x_i - y_i|$.
  - La distancia del máximo, `maximum`, que vale: $\text{máx}_{i=1,\dots,m} |x_i - y_i|$.
  - La distancia de Canberra, `canberra`, que vale $\sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i| + |y_i|}$.

  - La distancia de Minkowski, `minkowski`, que depende de un parámetro $p > 0$ (que se ha de especificar en la función `dist` con `p` igual a su valor), y que vale: $\left(\sum_{i=1}^{m} |x_i - y_i|^p\right)^{1/p}$. Observad que cuando $p = 1$ se obtiene la distancia de Manhattan y cuando $p = 2$, la distancia euclídea usual.
  - La distancia binaria, `binary`, que sirve básicamente para comparar vectores binarios (si los vectores no son binarios, `R` los entiende como binarios sustituyendo cada entrada diferente de 0 por 1). La distancia binaria entre `x` e `y` binarios es el número de posiciones en las que estos vectores tienen entradas diferentes, dividido por el número de posiciones en las que alguno de los dos vectores tiene un 1.

La salida de aplicar la función `dist` a nuestra tabla de datos es un objeto `dist` de `R`, no es una matriz de distancias usual.

- `scale(x, center = TRUE, scale = TRUE)`: escala los datos de una tabla de datos, donde:
  - `x`: nuestra tabla de datos.
  - `center`: es un parámetro lógico o un vector numérico de longitud el número de columnas de `x` indicando la cantidad que queremos restar a los valores de cada variable o columna. Si vale `TRUE` que es su valor por defecto, a cada columna se le resta la media de dicha columna.
  - `scale`: es un parámetro lógico o un vector numérico de longitud el número de columnas de `x` indicando la cantidad por la que queremos dividir los valores de cada variable o columna. Si vale `TRUE` que es su valor por defecto, a los valores de cada columna se les divide por la desviación estándar de dicha columna.
- `hclust(d, method = ...)`: aplica el algoritmo de clustering aglomerativo a una tabla de datos, donde:
  - `d` es la matriz de distancias entre nuestros objetos calculada con la función `dist`.
  - `method` sirve para especificar cómo se define la distancia de la unión de dos clusters al resto de los clusters. El nombre del método se ha de entrar entrecomillado. Los más populares son los siguientes:
  - Método de enlace completo, `complete`,
  - Método de enlace simple, `single`,
  - Método de enlace promedio, `average`,
  - Método de la mediana, `median`,
  - Método del centroide, `centroid`,
  - Método de Ward clásico, `ward.D`.
- Salida de `hclust`: imaginemos que hemos guardado en el objeto `estudio.clustering` la salida del algoritmo:
  - `estudio.clustering$merge` nos indica cómo ha ido agrupando los clusters. Es una matriz de $n - 1$ filas y 2 columnas, donde n es el número de objetos. En esta matriz, los objetos originales se representan con números negativos, y los nuevos clusters con números positivos que indican el paso en el que se han creado.
  - `estudio.clustering$height` es un vector que contiene las distancias a las que se han ido agrupando los pares de clusters.

- `plot(estudio.clustering,hang=...,labels=...)`: dibuja el dendrograma o el árbol binario del clustering aglomerativo realizado, donde:
  - `hang` es un parámetro que controla la situación de las hojas del dendrograma respecto del margen inferior.
  - `labels` es un vector de caracteres que permite poner nombres a los objetos; por defecto, se identifican en la representación gráfica por medio de sus números de fila en la matriz o el data frame que contiene los datos.
- `cutree(estudio.clustering,k=...,h=...)`: calcula los clusters del clustering aglomerativo realizado, donde:
  - `k` es un parámetro que indica el número de clusters deseado
  - `h` es un parámetro que indica la altura a la que queremos cortar.

- `rect.hclust(estudio.clustering,h=...,k=...)`: dibuja los clusters hallados en el algoritmo de clustering aglomerativo realizado según el valor del parámetro `h` o el parámetro `k`. Antes de llamar a esta función, se tiene que haber llamado a la función `plot` para dibujar el dendrograma. En caso contrario, `R` da error.