

Técnicas cuantitativas

Enric Aguilar y Benito Zaragozí

14-Apr-2021

Índice general

Antes de empezar	5
0.1 Prerrequisitos	5
0.2 Tipos de ficheros	6
0.3 Un ejemplo muy sencillo	6
0.4 Estructura básica de un documento Markdown	6
0.5 Ejercicio	7
1 Introducción	9
1.1 R como lenguaje de programación	10
1.2 Utilizando RStudio	11
1.3 Interpretación de valores	11
1.4 Variables	12
1.5 Tipos básicos	13
1.6 Estructuras de datos	16
1.7 Funciones	17
1.8 Tidyverse	17
1.9 Ejercicios	20
2 Estadística descriptiva	23
2.1 Tendencia central	24
2.2 Dispersión	26
2.3 Correlación	28
2.4 Ejercicios	35
3 Estadística inferencial	37
3.1 Diferencia entre probabilidad y estadística	37
3.2 Probabilidad frecuentista vs Bayesiana	39
3.3 Introducción a las distribuciones de probabilidad	43
3.4 Introducción a los test de hipótesis	54
3.5 Test de hipótesis en R: cálculo e informes	72
3.6 Un ejemplo sobre brecha salarial entre géneros	73
3.7 ANOVA	73
3.8 Algunas consideraciones finales	73

3.9	Ejercicios	75
4	El muestreo estadístico	77
4.1	Example one	77
4.2	Example two	77
5	Regresión lineal	79
5.1	¿Qué es un modelo de regresión lineal?	79
5.2	Estimación de un modelo de regresión lineal	82
5.3	Regresión lineal múltiple	85
6	Estadística multivariante	87

Antes de empezar

En este breve capítulo explicamos los aspectos básicos para que podáis reproducir los ejercicios y entregar las actividades utilizando R, Markdown y Rstudio. Aquí planteamos los conceptos básicos para preparar un documento que incluye explicaciones, código y los resultados de análisis R. Se trata de una estrategia de *programación literaria* que en los últimos años está siendo cada vez más utilizada en el análisis de datos (Knuth, 1984; Xie, 2015).

0.1 Prerrequisitos

Para seguir las explicaciones de este curso será necesario instalar primero R y RStudio con los paquetes `knitr` y `rmarkdown`. A continuación mostramos la configuración general del sistema R utilizado.

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04 LTS
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblas-r0.3.8.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=C
##  [7] LC_PAPER=en_US.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
```

```
## [1] pander_0.6.3
##
## loaded via a namespace (and not attached):
## [1] compiler_4.0.2 magrittr_1.5 bookdown_0.21 htmltools_0.5.0
## [5] tools_4.0.2 rstudioapi_0.11 yaml_2.2.1 Rcpp_1.0.5
## [9] stringi_1.5.3 rmarkdown_2.4 knitr_1.30 stringr_1.4.0
## [13] digest_0.6.25 xfun_0.22 rlang_0.4.7 evaluate_0.14
```

0.2 Tipos de ficheros

- Los archivos para producir documentos de RMarkdown tienen la extensión `.Rmd`.
- Los archivos deben abrirse con RStudio y se compilan haciendo clic en el botón `knitr`.
- El resultado es un documento en formato `.pdf`, `.html` o `.doc`.

0.3 Un ejemplo muy sencillo

Cread un fichero con el siguiente contenido

Hola, soy **R Markdown**

Aprende más sobre mi [aquí](<http://rmarkdown.rstudio.com/>).

al hacer clic en el botón `knitr` a HTML de Rstudio se crea un archivo `.html` con este contenido.

0.4 Estructura básica de un documento Markdown

Revisemos las partes más importantes del documento `Rmd`.

0.4.1 Cabecera

La cabecera está en la parte superior del documento dentro de estas dos líneas

```
---
```

```
---
```

```
title: "Write your title here"
```

```
author: "Write your name here"
```

```
date: "Write the date here"
```

```
output:
```

```
  pdf_document: default
```

```
  html_document: default
```

```
  word_document: default
```

```
---
```

En el encabezado del archivo debe escribir el título del documento, su nombre y la fecha. La declaración de salida se utiliza para la clase del documento final, `pdf`, `html` o `doc`. Para producir un PDF es necesario tener instalado un motor de LaTeX.

0.4.2 Insertar código R

A continuación, configuramos las opciones necesarias para imprimir el código R y la salida R en el documento final.

La primera línea es ocultar este fragmento de código en el documento final.

La segunda línea es para imprimir el código R y la salida R en el documento final.

0.4.3 Formatos de texto

El texto sin formato se escribe como en cualquier otro documento, como un documento de Word. Debe tener cuidado con las letras en cursiva o negrita y algunos caracteres especiales. Por ejemplo

- **Negrita:** escriba el texto entre `**Negrita**` o `__Negrita__`
- *Cursiva:* escriba su texto entre `*cursiva*` o `_Italica_`
- Encabezados de sección

```
# Título 1
## Título 2
### Título 3
```

Cuanto más símbolos `#` escriba antes de su texto, menor será el tamaño de su título

Puede encontrar más información sobre cómo escribir en el siguiente archivo.

0.4.4 Generando el documento

Para compilar el archivo `.Rmd` y obtener su documento final, simplemente haga clic en el botón `Knit` y seleccione `Knit a PDF` para producir un archivo `.pdf`.

0.5 Ejercicio

- Utilizando Rstudio, crea tu primer documento con Rmarkdown. El documento debe mostrar los metadatos básicos (título, autor y fecha), un título de primer nivel (por ejemplo, 'Información de la sesión') y un recuadro con la información básica de vuestra sesión de R.

Capítulo 1

Introducción

Para cuando llega el momento de analizar los datos, la mayor parte del trabajo ya está hecho. Antes de tener un conjunto de datos se ha tenido que definir el problema de investigación, desarrollar e implementar un plan de muestreo, decidir las escalas de medidas y desarrollar un diseño de investigación. Si se ha hecho bien este trabajo, el análisis de los datos puede ser bastante sencillo.

En la mayoría de las investigaciones de ciencias sociales, el análisis de datos implica tres pasos principales, realizados aproximadamente en este orden:

1. Limpieza y organización de los datos para su análisis (preparación de datos)
2. Describir los datos (estadística descriptiva)
3. Prueba de hipótesis y modelos (estadística inferencial)

La **preparación de datos** implica verificar o registrar los datos, comprobar su exactitud, cargar los datos en el software de análisis, transformar los datos y crear una base de datos adecuada para el análisis que se vaya a realizar.

Las **estadísticas descriptivas** se utilizan para describir las características básicas de los datos en un estudio. Proporcionan resúmenes sencillos sobre la muestra y las medidas. Junto con el análisis gráfico simple, forman la base de prácticamente todos los análisis cuantitativos de datos. Las estadísticas descriptivas simplemente se *describen los datos*.

La **estadística inferencial** investiga preguntas, modelos e hipótesis. En muchos casos, las conclusiones de la estadística inferencial se extienden más allá de los datos inmediatos por sí solos. Por ejemplo, usamos estadísticas inferenciales para tratar de inferir de los datos de la muestra lo que piensa la población total. También utilizamos estadística inferencial para hacer juicios sobre la probabilidad de que una diferencia observada entre grupos sea confiable o si podría haber ocurrido por casualidad en este estudio. Por lo tanto, usamos estadísticas inferenciales para *inferir lo que sucede más allá de nuestros datos*.

En la mayoría de los estudios de investigación, la sección de análisis sigue estas tres fases de análisis. Las descripciones de cómo se prepararon los datos tienden a ser breves y a centrarse solo en los aspectos más exclusivos de su estudio, como las transformaciones de datos específicas que se realizan. Las estadísticas descriptivas que observa en realidad pueden ser voluminosas. En la mayoría de los estudios, las estadísticas descriptivas se seleccionan cuidadosamente y se organizan en tablas de resumen y gráficos que solo muestran la información más relevante o importante. Por lo general, el investigador vincula cada uno de los análisis inferenciales con preguntas o hipótesis de investigación específicas que se plantearon en la introducción, o toma nota de los modelos que se probaron y que surgieron como parte del análisis. En la mayoría de los informes de análisis, es especialmente importante *“no dejar que los árboles nos impidan ver el bosque”*. Si se presentan demasiados detalles sobre el análisis, es posible que se pierda de vista el problema de la realidad que estamos estudiando.

Todos estos pasos del análisis se pueden realizar habitualmente en todos los paquetes estadísticos. En este curso nos centraremos en el uso de R, por lo que a continuación haremos una breve introducción.

1.1 R como lenguaje de programación

R fue creado en 1992 por Ross Ihaka y Robert Gentleman en la Universidad de Auckland, Nueva Zelanda. R es una implementación gratuita de código abierto del lenguaje de programación estadística **S** creado inicialmente en Bell Labs. En esencia, R es un lenguaje de programación funcional (sus principales funcionalidades giran en torno a la definición y ejecución de funciones). Sin embargo, ahora es compatible, y se usa comúnmente como un lenguaje de programación imperativo (enfocado en instrucciones sobre variables y estructuras de control de programación) y orientado a objetos (que involucra estructuras de objetos complejas).

En términos simples, hoy en día, la programación en R se enfoca principalmente en diseñar una serie de instrucciones para ejecutar una tarea, más comúnmente, cargar y analizar un conjunto de datos (Wickham and Grolemund, 2017).

Como tal, R se puede usar para programar creando secuencias de **instrucciones** que involucren **variables**, que son entidades con nombre que pueden almacenar valores. Ese será el tema principal de esta sesión práctica. Las instrucciones pueden incluir estructuras de flujo de control, como puntos de decisión (*if/else*) y bucles, que serán el tema de la próxima sesión práctica. Las instrucciones también se pueden agrupar en **funciones**, que también veremos en la próxima sesión práctica.

R es **interpretado**, no compilado. Lo que significa que un intérprete de R (si está utilizando R Studio, el intérprete de R simplemente está oculto en el backend y R Studio es el frontend que le permite interactuar con el intérprete) recibe una instrucción que escribe en R, la interpreta y la ejecuta .

Otros lenguajes de programación requieren que su código sea compilado en un ejecutable para ser ejecutado en un ordenador.

1.2 Utilizando RStudio

La interfaz de RStudio se divide en dos secciones principales. En el lado izquierdo, encontrará la *Consola*, así como el editor de secuencias de comandos R, cuando se está editando una secuencia de comandos. La *Consola* es una ventana de entrada/salida en el intérprete de R, donde se pueden escribir instrucciones y se muestra la salida calculada.

Por ejemplo, si escribís en la *Consola*

```
1 + 1
```

el intérprete de R entiende eso como una instrucción para sumar uno más uno, y produce el resultado (dado que los materiales para este módulo se crean en RMarkdown, la salida del cálculo siempre está precedida por ‘##’).

```
## [1] 2
```

Fíjate cómo el valor de salida 2 está precedido por [1], lo que indica que la salida está constituida por un solo elemento. Si la salida está constituida por más de un elemento, como la lista de números a continuación, cada fila de la salida está precedida por el índice del primer elemento de la salida.

```
## [1] 1 4 9 16 25 36 49 64 81 100 121 144 169 196 225 256 289 324 361
## [20] 400
```

En el lado derecho, encontrarás dos grupos de paneles. En la parte superior derecha, el elemento principal es el panel *Entorno*, que es una representación del estado actual de la memoria del intérprete y, como tal, muestra todas las variables, conjuntos de datos y funciones almacenados. En la parte inferior derecha, encontrarás el panel *Archivos*, que muestra el sistema de archivos (archivos y carpetas del ordenador), así como el panel *Ayuda*, que le muestra las páginas de ayuda cuando sea necesario. Discutiremos los otros paneles más adelante.

1.3 Interpretación de valores

Cuando se escribe un valor en la *Consola*, el intérprete simplemente devuelve el mismo valor. En los ejemplos siguientes, 2 es un valor numérico simple, mientras que "Valor de cadena" es un valor textual, que en R se conoce como un valor de *carácter* y en programación también se conoce comúnmente como una *cadena* (abreviatura de *una cadena de caracteres*).

Ejemplo numérico

2

```
## [1] 2
```

Ejemplo de cadena

```
"String value"
```

```
## [1] "String value"
```

Tened en cuenta que los valores de los caracteres deben comenzar y terminar con comillas simples o dobles ('o"), que no forman parte de la información en sí. La Guía de estilo de Tidyverse sugiere usar siempre comillas dobles ("), así que serán las que usaremos en este curso.

Todo lo que sigue a un símbolo # se considera un *comentario* y el intérprete lo ignora. Cada lenguaje de programación puede usar sus propios símbolos para identificar los comentarios. Por ejemplo cabe destacar la diferencia entre # en Markdown que identifica un título de primer nivel, mientras que los comentarios en un fichero de Rmarkdown se identifican entre < y >.

```
# Este comentario es ignorado por R. Solo sirve para documentar el código.
```

Como se ha mencionado anteriormente, el intérprete también comprende operaciones aritméticas simples sobre valores numéricos.

```
1 + 1
```

```
## [1] 2
```

Además, también hay una gran cantidad de funciones predefinidas, por ejemplo, raíz cuadrada: `sqrt`.

```
sqrt(2)
```

```
## [1] 1.414214
```

Las funciones se recopilan y almacenan en *bibliotecas* (a veces denominadas *paquetes*), que contienen funciones relacionadas. Las bibliotecas pueden variar desde la biblioteca `base`, que incluye la función `sqrt` anterior, hasta la biblioteca `rgdal`, que funciona como un puente hacia las funciones de la GDAL (Biblioteca de abstracción de datos geoespaciales), que es una importante librería en el mundo de los Sistemas de Información Geográfica. Por lo tanto, es mediante la creación de librerías que podemos extender las capacidades de R.

1.4 Variables

Una variable se puede definir usando un **identificador** (por ejemplo, `una_variable`) a la izquierda de un **operador de asignación** `<-`, seguido del *objeto* que se vinculará al identificador, como un **valor** (por ejemplo, 1) que se

asignará a la derecha. Una vez realizada la asignación, el valor de la variable se puede probar/invocar simplemente especificando el **identificador**.

```
una_variable <- 1
una_variable
```

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

Si escribes `una_variable <- 1` en la *Consola* de RStudio, aparece un nuevo elemento en el panel *Environment* (derecha-arriba), que representa la nueva variable en la memoria. La parte izquierda de la entrada contiene el identificador `una_variable`, y la parte derecha contiene el valor asignado a la variable `una_variable`, es decir, 1.

No es necesario aportar un valor directamente. La parte derecha de la tarea puede ser una **llamada a una función**. En ese caso, la función se **ejecuta** en la entrada proporcionada y **el resultado se asigna a la variable**.

```
una_variable <- sqrt(4)
una_variable
```

```
## [1] 2
```

Observa cómo, al escribir `una_variable <- sqrt(4)` en la *Consola* de RStudio, el elemento en el panel *Environment* cambia para reflejar el nuevo valor asignado a `una_variable`, que ahora es el resultado de `sqrt(4)`, es decir 2.

En el siguiente ejemplo, se crea otra variable llamada `otra_variable` y se suma a `una_variable`, guardando el resultado en `suma_de_dos_variables`. La raíz cuadrada de esa suma se almacena en la variable `raiz_cuadrada_de_suma`.

```
otra_variable <- 4
otra_variable
```

```
## [1] 4
```

```
suma_de_dos_variables <- una_variable + otra_variable
```

```
raiz_cuadrada_de_suma <- sqrt(suma_de_dos_variables)
raiz_cuadrada_de_suma
```

```
## [1] 2.44949
```

1.5 Tipos basicos

1.5.1 Números

El tipo *numeric* representa números en general (tanto enteros como reales), pero R es capaz de distinguirlos utilizando funciones.

```

un_numero <- 1.41
is.numeric(un_numero)

## [1] TRUE
is.integer(un_numero)

## [1] FALSE
is.double(un_numero) # i.e., es real

## [1] TRUE

```

Operadores numéricos básicos.

Operador	Significado	Ejemplo	Resultado
+	Suma	5+2	7
-	Resta	5-2	3
*	Multiplicación	5*2	10
/	División	5/2	2.5
%%	Div. de enteros	5%%2	2
%%	Módulo	5%%2	1
^	Potencia	5^2	25

Algunas funciones predefinidas en R son:

```

abs(-2) # Valor absoluto

## [1] 2
ceiling(3.475) # Redondeo al alza

## [1] 4
floor(3.475) # Redondeo a la baja

## [1] 3
trunc(5.99) # Truncar

## [1] 5
log10(100) # Logaritmo en base 10

## [1] 2
log(exp(2)) # Logaritmo natural y exponencial

## [1] 2

```

Como en cualquier otro entorno, podéis utilizar paréntesis simples para

especificar el orden de ejecución. Si no se especifica, el orden predeterminado es: potencia, multiplicación y división, suma y resta al final.

```
un_numero <- 1
(un_numero + 2) * 3
```

```
## [1] 9
```

```
un_numero + (2 * 3)
```

```
## [1] 7
```

```
un_numero + 2 * 3
```

```
## [1] 7
```

R devuelve el resultado NaN (*Not a number*) cuando el resultado de una operación no es un número.

```
0/0
```

```
## [1] NaN
```

```
is.nan(0/0)
```

```
## [1] TRUE
```

No hay que confundir NaN con NA (*No Available*), que sirve para identificar datos faltantes.

1.5.2 Lógicos o booleanos

El tipo *lógico* codifica dos valores dicotómicos: Verdadero y Falso.

```
valor_logico <- TRUE
is.logical(valor_logico)
```

```
## [1] TRUE
```

```
isTRUE(valor_logico)
```

```
## [1] TRUE
```

```
as.logical(0) # se pueden convertir 1/0 a TRUE/FALSE
```

```
## [1] FALSE
```

Operadores lógicos básicos

Operador	Significado	Ejemplo	Resultado
==	Igual	5==2	FALSE
!=	No igual	5!=2	TRUE
>	Mayor que	5>2	TRUE

Operador	Significado	Ejemplo	Resultado
<	Menor que	5<2	FALSE
>=	Mayor o igual	5>=2	TRUE
<=	Menor o igual	5<=2	FALSE
!	No	!TRUE	FALSE
&	Y	TRUE & FALSE	FALSE
	O	TRUE FALSE	TRUE

1.5.3 Cadenas de caracteres

El tipo *character* representa objetos de texto, incluidos caracteres individuales y cadenas de caracteres (es decir, objetos de texto de más de un carácter, comúnmente denominados simplemente *cadenas* o *strings* en informática).

```
una_cadena <- "¡Hola!"
is.character(una_cadena)

## [1] TRUE
is.numeric(una_cadena)

## [1] FALSE
as.character(2) # Conversión de número a cadena (en inglés, hacer un 'cast')

## [1] "2"
as.numeric("2")

## [1] 2
as.numeric("¡Hasta luego!")

## Warning: NAs introduced by coercion
## [1] NA
```

1.6 Estructuras de datos

Estos tipos de datos más básicos se gestionan habitualmente dentro de estructuras de datos más complejas. R tiene muchas otras estructuras de datos pero las más básicas comprenden:

- Vectores
- Listas
- Matrices
- *data frames* (se usa el término en inglés por comodidad).
- Factores

1.6.1 Vectores

1.6.2 Listas

1.6.3 Matrices

1.6.4 Data frames

1.6.5 Factores

1.7 Funciones

Hasta este punto ya habéis visto y usado varias funciones. Por ejemplo, la función `c()` se puede usar para combinar objetos en un vector. En general, todas las llamadas a funciones tienen el mismo aspecto: el nombre de una función siempre va seguido de paréntesis. A veces, los paréntesis incluyen argumentos como sucede también en el siguiente ejemplo:

```
# Crea el vector `z` a partir de la función seq (secuenciar).  
z <- seq(from = 1, to = 5, by = 1)
```

En este ejemplo usamos una función llamada `seq()` para crear una secuencia que progresa por unidades, desde 1 a 5 (probad de modificar los parámetros para obtener otras secuencias distintas).

Si no estáis seguros de qué argumentos acepta una función, siempre se puede consultar la documentación de dicha función (todas las librerías *oficiales* suelen tener una buena documentación). Por ejemplo, supongamos que no estamos seguros de cómo funcionan los argumentos necesarios para `seq()`. Podemos escribir `?seq` en la consola y, al ejecutar este comando, la página de documentación para esa función aparece en el panel inferior derecho de RStudio. En la sección *Argumentos* está toda la información que buscamos. En la parte inferior de casi todas las páginas de ayuda, suele haber ejemplos sobre cómo utilizar las funciones correspondientes.

1.8 Tidyverse

Como se ha mencionado anteriormente, las librerías o *paquetes* son colecciones de funciones y/o conjuntos de datos. Las librerías se pueden instalar en R usando la función `install.packages()` o usando el menú **Herramientas > Instalar Librerías ...** en RStudio. Algunas librerías de R están relacionadas entre sí o forman parte de flujos de trabajo mayores. A día de hoy la meta-librería Tidyverse contiene algunas de las librerías más utilizadas en el análisis de datos (Wickham et al., 2019). Solo por mencionar algunas:

- `ggplot2` para crear gráficos.
- `dplyr` para manipular datos (filtrar, seleccionar, agregar, sumarizar, etc).

- **tidyr** para organizar los datos de un modo que las otras librerías del **Tidyverse** puedan trabajar mejor.
- **readr** para importar tablas de datos a partir de formatos habituales (csv, tsv, o fwf). Permite minimizar la introducción de errores en la importación.
- **purrr** se utiliza para facilitar la automatización de tareas en R, mientras se escribe menos código.
- **stringr** facilita la manipulación de cadenas de texto (unir, separar, filtrar palabras en un documento, etc).
- **forcats** para trabajar con factores.
- **tibble**...

Se puede cargar una librería usando la función `library()`, como se muestra a continuación (tened en cuenta que el nombre de la biblioteca no está entrecomillado). Una vez que una librería está instalada en un ordenador, no es necesario que la instale nuevamente, pero cada secuencia de comandos o *script* debe cargar todas las librerías que utiliza. Una vez que se carga una librería, se pueden utilizar todas sus funciones.

```
library(tidyverse)
```

1.8.1 stringr

El siguiente código presenta un mínimo ejemplo del uso de las funciones de la librería **stringr**.

```
str_length("Tarragona")
```

```
## [1] 9
```

```
str_detect("Tarragona", "a")
```

```
## [1] TRUE
```

```
str_replace_all("Tarragona", "r", "R")
```

```
## [1] "TaRRagona"
```

1.8.2 El operador pipe

El operador **pipe** (tubería) es útil para reducir el número de asignaciones en operaciones más complejas. Un *pipe* (`%>%`) toma el resultado de una función y lo pasa a la siguiente función como **primer argumento**, de este modo ya no hace falta repetir el resultado de la primera función en el código.

El siguiente código muestra un ejemplo sencillo. El número 2 se toma como entrada para el primer **pipe** que lo pasa como primer argumento a la función `sqrt`. El valor de salida 1.4142136 se toma como entrada para el segundo **pipe**, que lo pasa como primer argumento a la función `trunc`. Finalmente se devuelve la salida final "1".

```
2 %>%
  sqrt() %>%
  trunc()
```

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

```
sqrt(2) %>%
  round(digits = 2)
```

El primer paso de una secuencia de *pipes* puede ser un valor, una variable o una función que incluya argumentos. El siguiente código muestra una serie de ejemplos de diferentes formas de lograr el mismo resultado. Los ejemplos usan la función `round`, que también permite un segundo argumento `digits = 2`. Tened en cuenta que, cuando se utiliza el operador `%>%`, solo se proporciona el segundo argumento nominalmente a la función `round`, es decir, `round(digits = 2)`

```
# R básico, sin utilizar '%>%', pero en varios pasos
variable_temporal_a <- 2
variable_temporal_b <- sqrt(variable_temporal_a)
round(variable_temporal_b, digits = 2)
```

```
# R básico, sin utilizar '%>%', pero sin asignaciones
round(sqrt(2), digits = 2)
```

```
# Pipe a partir de un valor
2 %>%
  sqrt() %>%
  round(digits = 2)
```

```
# Pipe a partir de una variable
el_numero_dos <- 2
el_numero_dos %>%
  sqrt() %>%
  round(digits = 2)
```

```
# Pipe empezando por una función
sqrt(2) %>%
  round(digits = 2)
```

Una operación compleja creada mediante el uso de `%>%` se puede usar en el lado derecho de una asignación `<-`, de modo que se guarda resultado de toda la operación de la derecha.

```
raiz_cuadrada_de_dos <- 2 %>%
  sqrt() %>%
  round(digits = 2)
```

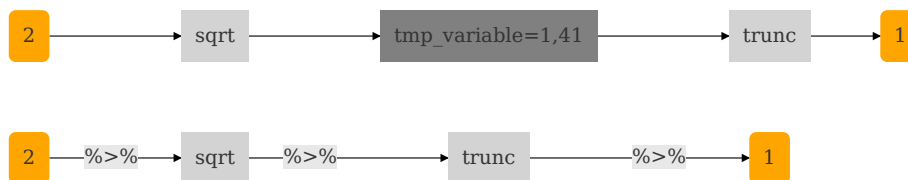


Figura 1.1: Ejecución del código sin o con `%>%`. Sin *pipes* los resultados intermedios tienen que guardarse en una variable en memoria.

1.9 Ejercicios

Crea un *script* para responder a cada pregunta. En las preguntas de la 4 a la 8 es posible que tengas que consultar la referencia de la librería **stringr** (stringr.tidyverse.org/reference) o a otras funciones de R **base**.

1. Escribe un fragmento de código utilizando el operador `%>%` que haga lo siguiente:
 - tomar como entrada el número 1632
 - calcular el logaritmo en base 10
 - redondear el número a la baja
 - truncar la parte entera y
 - verificar si se trata de un número entero.
2. Escribe un fragmento de código utilizando el operador `%>%` que haga lo siguiente:
 - tomar como entrada el número 1632
 - calcular la raíz cuadrada,
 - redondear al alza y quedarse con la parte entera, y
 - verificar que se trata de un entero.
3. Escribe un fragmento de código utilizando el operador `%>%` que haga lo siguiente:
 - tomar como entrada la cadena 1632
 - convertirla a número
 - comprobar si el resultado es o no un número.
4. Escribe un fragmento de código utilizando el operador `%>%` que haga lo siguiente:
 - tomar como entrada la cadena "-16.32"
 - transformarla en un número
 - calcular el valor absoluto y truncarlo
 - comprobar si el resultado es o no un número.
5. Escribe un fragmento de código utilizando el operador `%>%` y la librería **stringr** que haga lo siguiente:
 - tomar la cadena "Siempre r que r" como entrada
 - transformar la cadena a mayúsculas.
6. Escribe un fragmento de código utilizando el operador `%>%` y la librería

and the `stringr` que haga lo siguiente:

- tomar la cadena "Siempre r que r" como entrada
- truncarla para dejar solamente 'Siempre R'.

Capítulo 2

Estadística descriptiva

La estadística descriptiva es una rama de la estadística cuyo objetivo es resumir, describir y presentar una serie de valores o un conjunto de datos. Estas estadísticas pueden ser realmente útiles al analizar largas series de datos en las que resulte difícil reconocer algún patrón. En éste capítulo utilizaremos una muestra de datos sobre salarios (en euros) de una *población* de 100 trabajadores:

2186, 1218, 1682, 1816, 1702, 1447, 2256, 1453, 2509, 1469, 2152, 2643, 806, 1361, 1433, 1818, 1358, 172, 280, 2160, 1347, 609, 1414, 2107, 2448, 1285, 1371, 618, 1730, 1180, 1728, 1852, 2018, 1196, 1752, 642, 1108, 1074, 293, 1518, 1603, 1320, 1879, 1137, 816, 1716, 1094, 2222, 1284, 1828, 1661, 1108, 2288, 1821, 1545, 1638, 1840, 1545, 4, 1642, 1316, 1593, 1791, 2200, 1136, 2151, 1668, 2019, 1960, 1860, 978, 1455, 1812, 1023, 1229, 1790, 1884, 1732, 1057, 950, 2256, 1629, 1544, 1440, 903, 1806, 1391, 1409, 1967, 1911, 2196, 1262, 1825, 2196, 945, 1070, 934, 770, 1540 y 1827

De un primer vistazo es difícil (por no decir imposible) que podamos comprender los datos y tener una visión clara de los salarios de este grupo de personas. Las estadísticas descriptivas permiten resumir y así tener una mejor visión general de los datos. Por supuesto, al resumir los datos a través de una o varias medidas, inevitablemente se perderá parte de la información. Sin embargo, en muchos casos es mejor perder algo de información pero, a cambio, obtener una visión general. Podríamos decir que se trata de *ganar perspectiva*.

La estadística descriptiva es a menudo el primer paso y una parte importante en cualquier análisis estadístico. Permite comprobar la calidad de los datos detectando posibles valores atípicos (*outliers*), es decir, datos que parecen ser significativamente distintos del resto. También se puede utilizar estadística descriptiva para detectar errores de recopilación o codificación, determinar si están bien presentados, entre otras posibles aplicaciones.

Podemos distinguir dos tipos básicos de estadísticos para describir un conjunto

de datos: de centralidad y de dispersión. Habitualmente, ambos tipos de medidas se utilizan juntos para resumir los datos de la forma más concisa.

2.1 Tendencia central

Las medidas de tendencia central permiten ver “dónde” se ubican los datos, alrededor de qué valores. En otras palabras, las medidas de ubicación permiten comprender cuál es la tendencia central o la “posición” de los datos en su conjunto. Entre las estadísticas más habituales de este tipo podemos distinguir:

- Mínimo y máximo
- Media
- Mediana
- Primer y tercer cuartil
- Moda

2.1.1 Mínimo y máximo

Mínimo (*min*) y máximo (*max*) son simplemente los valores más bajo y más alto de la muestra. Dada una muestra de 6 de estos salarios:

2186, 1218, 1682, 1816, 1702 y 1447

El mínimo es 1217.7 euros y el máximo es 2185.5 euros. Estas dos estadísticas básicas dan una idea clara sobre los extremos de la muestra y su cálculo con R es bastante sencillo:

```
min(salarios_sel)
```

```
## [1] 1217.7
```

```
max(salarios_sel)
```

```
## [1] 2185.5
```

2.1.2 Media

La media o promedio, es probablemente la estadística más habitual. Da una idea de cuál es el valor medio, es decir, el valor central de los datos o, en otras palabras, su centro de gravedad. La media se encuentra sumando todos los valores y dividiendo esta suma por el número de observaciones (*n*):

$$Media = \bar{x} = \frac{\text{suma de todos los valores}}{\text{número de valores}} = \frac{1}{n} \sum_{i=1}^n x_i$$

Dada nuestra muestra de 6 salarios presentada anteriormente, la media es:

$$\bar{x} = \frac{2186 + 1218 + 1682 + 1816 + 1702 + 1447}{6} = 1675,033$$

En conclusión, el tamaño medio de la nuestra muestra es 1675.03 euros (redondeado a 2 decimales). El cálculo con R también resulta bastante sencillo:

```
mean(salarios_sel)
```

```
## [1] 1675.033
```

2.1.3 Mediana

La mediana es otra medida de centralidad. La interpretación de la mediana es que hay tantas observaciones por debajo como por encima de la mediana. En otras palabras, el 50% de las observaciones se encuentran por debajo de la mediana y el 50% de las observaciones están por encima de la mediana.

La forma más fácil de calcular la mediana es primero ordenar los datos de menor a mayor (es decir, en orden ascendente) y luego tomar el punto medio como la mediana. A partir de los valores ordenados, para un número impar de observaciones, el punto medio es fácil de encontrar: es el valor con tantas observaciones abajo como arriba. Aún a partir de los valores ordenados, para un número par de observaciones, el punto medio está exactamente entre los dos valores medios. Formalmente, después de ordenar, la mediana es:

- si n (número de observaciones) es impar:

$$\text{mediana}(x) = x_{\frac{n+1}{2}}$$

- si n es par:

$$\text{mediana}(x) = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$$

donde el subíndice de x denota la numeración de los datos ordenados. El cálculo en R es de nuevo bastante sencillo:

```
median(salarios_sel)
```

```
## [1] 1691.85
```

2.1.4 Primer y tercer cuartil

El primer y tercer cuartil son similares a la mediana en el sentido de que también dividen las observaciones en dos partes, solo que estas partes no son iguales. Recordad que la mediana divide los datos en dos partes iguales (con el 50% de las observaciones por debajo y el 50% por encima de la mediana). El primer cuartil divide las observaciones de modo que haya un 25% de las observaciones *debajo de* este punto y un 75% **por encima** del primer cuartil. El tercer cuartil se calcula del mismo modo pero representa el valor con el 75% de las observaciones por debajo y el 25% de las observaciones por encima. Existen varios métodos para calcular el primer y tercer cuartil, pero por ejemplo se puede calcular siguiendo los siguientes pasos:

1. Ordenar los datos en orden ascendente

2. Calcular $0.25 \cdot n$ y $0.75 \cdot n$ (es decir, 0.25 y 0.75 veces el número de observaciones)
3. Redondear estos dos números al siguiente número entero `ceil`

Los pasos son los mismos para un número par e impar de observaciones. A continuación se muestra un ejemplo para el cálculo de estos valores sobre los salarios de nueve trabajadores:

```
quantile(salarios_sel)
```

```
##          0%          25%          50%          75%         100%
## 1217.700 1505.575 1691.850 1787.825 2185.500
```

Dado que obtenemos un vector de resultados, podemos seleccionar por posición para obtener el que necesitamos en cada caso.

2.1.5 Moda

La moda de una serie es el valor que aparece con mayor frecuencia. En otras palabras, es el valor que tiene el mayor número de ocurrencias. Dados los siguientes salarios:

1700, 1680, 1710, 1700, 1820, 1650, 1700, 1890 y 1670

La moda es 1700 porque es el valor más común con 3 apariciones. Todos los demás valores aparecen solo una vez. En conclusión, la mayoría de los trabajadores de esta muestra perciben exactamente 1700 euros al mes.

Hay que tener en cuenta que es posible que una serie no tenga moda (p. Ej., 4, 7, 2 y 10) o más de una moda (p. Ej., 4, 2, 2, 8, 11 y 11). Los datos con dos modas a menudo se denominan bimodales y los datos con más de dos modos a menudo se denominan multimodales, a diferencia de las series con una moda, que se denominan unimodales.

A diferencia de las anteriores estadísticas (mínimo, máximo, media, mediana, primer y tercer cuartil) que solo se pueden calcular para variables cuantitativas, **la moda se puede calcular para variables cuantitativas y cualitativas**. Atendiendo al tipo de empleo que desarrollan los 9 trabajadores presentados anteriormente:

ingeniero, ingeniero, ingeniero, ingeniero, asistente, asistente, asistente, ingeniero y camarero

La moda es “ingeniero”, por lo que la mayoría de los trabajadores de esta muestra son de uso ingenieros.

2.2 Dispersión

2.2.1 Rango

El rango es la diferencia entre el máximo y el mínimo:

$$\text{rango} = \max - \min$$

Dada nuestra muestra de salarios:

2186, 1218, 1682, 1816, 1702 y 1447

El rango es $2185.5 - 1217.7 = 967.8$ euros. El rango es muy sencillo de calcular y en algunos casos puede dar una buena idea de lo que podemos esperar de un conjunto de datos. En R se puede utilizar la función `range`

```
range(salarios_sel)
```

```
## [1] 1217.7 2185.5
```

No obstante, esta métrica no da ninguna información de la distribución interna del resto de medidas.

2.2.2 Desviación estándar

La desviación estándar es la medida de dispersión más común en estadística. Si tenemos que presentar una estadística que resuma la distribución de los datos, suele ser la desviación estándar. Como sugiere su nombre, la desviación estándar indica cuál es la desviación *normal* de los datos. De hecho, calcula la desviación promedio de la **media**. Cuanto mayor sea la desviación estándar, más dispersos estarán los datos. Por el contrario, cuanto menor es la desviación estándar, más se centran los datos alrededor de la media.

Existen dos fórmulas para calcular la desviación estándar en función de si nos enfrentamos a una muestra o a una población. Una población incluye a todos los miembros de un grupo específico, mientras que una muestra contiene algunas observaciones extraídas de la población, es decir, una parte o un subconjunto de la población. Por ejemplo, la población puede ser “**todas** personas que viven en España” y la muestra puede ser “**algunas** personas que viven en España”.

La desviación estándar para una población, a partir de ahora σ , es:

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

Como se puede ver en la fórmula, la desviación estándar es en realidad la desviación promedio de los datos de su media μ . Hay que calcular primero el cuadrado de la diferencia entre las observaciones y la media para evitar que las diferencias negativas se compensen con diferencias positivas.

Por ejemplo, imagine una población de solo 3 trabajadores:

770, 1540 y 1827

La media es 1379 (redondeado a 1 decimal). La desviación estándar es entonces:

$$\sigma = \sqrt{\frac{1}{3}[(770.4 - 1379)^2 + (1540 - 1379)^2 + (1827 - 1379)^2]}$$

$$\sigma = 546,2$$

Por lo tanto, la desviación estándar para los salarios de estos trabajadores es de 546,2 euros. Esto significa que los salarios de los trabajadores de esta población se desvían de la media en 546,2 euros.

...

2.2.3 Varianza

2.2.4 Coeficiente de variación

2.3 Correlación

Las correlaciones entre variables juegan un papel importante en un **análisis descriptivo**. Una correlación mide la **relación entre dos variables**, es decir, cómo están vinculadas entre sí. En este sentido, una correlación permite saber si dos variables evolucionan en la misma dirección, en sentido contrario y si son independientes.

En este capítulo, se muestra cómo calcular **coeficientes de correlación**, cómo realizar **pruebas de correlación** y cómo **visualizar** correlaciones entre variables usando R.

La correlación generalmente se calcula en dos variables *cuantitativas*, pero también se puede calcular en dos variables *cualitativas ordinales* (ver prueba de independencia de chi-cuadrado).

2.3.1 Datos

Usaremos el conjunto de datos `mtcars`. Este conjunto de datos viene cargado por defecto en R, de modo que se utiliza en numerosas demostraciones.

```
# mostrar las primeras cinco filas
head(mtcars, 5)
```

##	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
## Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
## Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
## Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
## Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
## Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

Las variables `vs` y `am` son variables categóricas, por lo que se eliminan para este artículo:

```
# Eliminar las variables vs y am
library(tidyverse)
dat <- mtcars %>%
select(-vs, -am)

# mostrar las primeras cinco filas
head(dat, 5)
```

```
##           mpg cyl disp  hp drat   wt  qsec gear carb
## Mazda RX4      21.0   6  160 110 3.90 2.620 16.46   4    4
## Mazda RX4 Wag  21.0   6  160 110 3.90 2.875 17.02   4    4
## Datsun 710     22.8   4  108  93 3.85 2.320 18.61   4    1
## Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44   3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02   3    2
```

2.3.2 Coeficiente de correlación

La correlación entre 2 variables se calcula con la función `cor()`. Supón que queremos calcular la correlación entre caballos de potencia (`hp`) y millas por galón (`mpg`):

```
# Correlación de Pearson entre 2 variables
cor(dat$hp, dat$mpg)
```

```
## [1] -0.7761684
```

Hay que fijarse en que la correlación entre las variables X e Y es igual a la correlación entre las variables Y y X , por lo que el orden de las variables en la función `cor()` no importa.

La función `cor()` calcula por defecto la correlación de Pearson, por lo que si se quiere calcular la correlación por otro método, se puede agregar el argumento `method = "spearman"` a la función `cor()`:

```
# Correlación de Spearman entre 2 variables
cor(dat$hp, dat$mpg,
    method = "spearman")
```

```
## [1] -0.8946646
```

Hay varios métodos de correlación (se puede consultar la ayuda de la la función para saber más `?cor`):

- **Pearson** se usa a menudo para variables *cuantitativas continuas* que tienen una relación lineal.
- **Spearman** (es similar a Pearson pero se basa en los valores ordenados para cada variable en lugar de en los datos brutos) se usa a menudo para

evaluar relaciones que involucren variables cualitativas ordinales en las que la relación sea parcialmente lineal.

- **Kendall** se calcula a partir del número de pares concordantes y discordantes, se utiliza a menudo para variables ordinales cualitativas.

La función `cor()` también permite calcular correlaciones para varios pares de variables a la vez:

```
# Correlaciones entre todas las variables
round(cor(dat),
      digits = 2 # redondeo a dos decimales
    )
```

```
##      mpg   cyl  disp    hp  drat    wt  qsec  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00 -0.21 -0.66
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66  0.27  1.00
```

La correlación varía de **-1 a 1**, de modo que el signo nos indica la dirección de la relación (aumentan a la vez o son opuestas) y el valor nos indica la fuerza de la relación (más fuerte cuanto más alejado de 0).

Una **correlación negativa** implica que las dos variables consideradas varían en **direcciones opuestas**, es decir, si una variable aumenta la otra disminuye y viceversa. Por otro lado, una **correlación positiva** implica que las dos variables consideradas varían en la **misma dirección**, es decir, si una variable aumenta, la otra aumenta y si una disminuye, la otra también disminuye.

En cuanto a la fuerza de la relación: cuanto **más extremo** es el coeficiente de correlación (cuanto más cerca de -1 o 1), **más fuerte es la relación**. Esto también significa que una **correlación cercana a 0** indica que las dos variables son **independientes**, es decir, a medida que una variable aumenta, no hay tendencia en la otra variable a disminuir o aumentar.

Por ejemplo, la correlación de Pearson entre caballos de potencia (**hp**) y millas por galón (**mpg**) encontrada es -0.78, lo que significa que las 2 variables varían en dirección opuesta. Esto tiene sentido, los automoviles con más caballos de potencia suelen a consumir más combustible (hacen menos millas con el mismo combustible que los automóviles más potentes). Por el contrario, de la matriz de correlación vemos que la correlación entre millas por galón (**mpg**) y el tiempo para conducir un cuarto de milla (**qsec**) es 0.42, lo que significa que los automóviles rápidos (con un menor **qsec**) tienden a tener un peor rendimiento por galón (bajo **mpg**). De nuevo, esto tiene sentido, ya que los coches rápidos tienden a

consumir más combustible.

2.3.3 Test de correlación

Volver una vez leída la sección sobre test de hipótesis.

Hay que tener en cuenta que el valor p se basa en el coeficiente de correlación y en el tamaño de la muestra. Cuanto mayor sea el tamaño de la muestra y más extrema será la correlación (más cercana a -1 o 1). Con un tamaño de muestra pequeño, es posible obtener una correlación *relativamente* grande en la muestra (según el coeficiente de correlación), pero aún así encontrar una correlación no significativamente diferente de 0 en la población (según la prueba de correlación). Por este motivo, se recomienda realizar siempre un test de correlación antes de interpretar un coeficiente de correlación para evitar conclusiones erróneas.

A diferencia de una matriz de correlación que indica los coeficientes de correlación entre algunos pares de variables en la muestra, se utiliza un test de correlación para probar si la correlación (ρ) entre dos variables es significativamente diferente de 0 o no en la *población*.

En realidad, un coeficiente de correlación diferente de 0 en la muestra no significa que la correlación sea **significativamente** diferente de 0 en la población. Esto debe probarse con un **test de hipótesis**.

Las hipótesis (nula y alternativa) para el test de correlación son las siguientes:

- $H_0: \rho = 0$ (si no existe una relación lineal entre las dos variables)
- $H_1: \rho \neq 0$ (si existe una relación lineal entre las dos variables)

A través de esta prueba de correlación, lo que realmente estamos probando es si:

- La muestra contiene evidencia suficiente para rechazar la hipótesis nula y concluir que el coeficiente de correlación no es igual a 0, por lo que la relación existe en la población.
- La muestra no contiene suficiente evidencia de que el coeficiente de correlación no sea igual a 0, por lo que en este caso no rechazamos la hipótesis nula de no-relación entre las variables de la población.

Supongamos que queremos probar si el ratio del eje trasero (**drat**) está correlacionado con el tiempo necesario para conducir 1/4 de milla (**qsec**):

```
# Test de correlación de Pearson
test <- cor.test(dat$drat, dat$qsec)
test

##
## Pearson's product-moment correlation
##
## data:  dat$drat and dat$qsec
## t = 0.50164, df = 30, p-value = 0.6196
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.265947  0.426340
## sample estimates:
##          cor
## 0.09120476
```

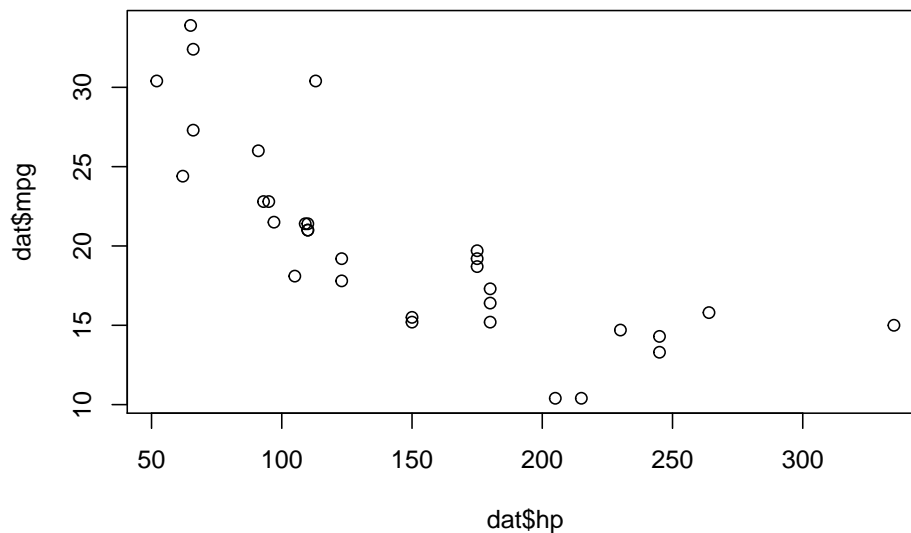
El p -valor de la prueba de correlación entre estas 2 variables es 0.62. Al nivel de significancia del 5%, no se rechaza la hipótesis nula de no correlación. Por lo tanto, concluimos que no rechazamos la hipótesis de que no existe una relación lineal entre las 2 variables.¹

Esta prueba demuestra que incluso si el coeficiente de correlación es diferente de 0 (la correlación es 0.09 en la muestra), en realidad no es significativamente diferente de 0 en la población.

2.3.4 Visualizando correlaciones

Una buena forma de visualizar una correlación entre 2 variables es mediante un diagrama de dispersión. Por ejemplo:

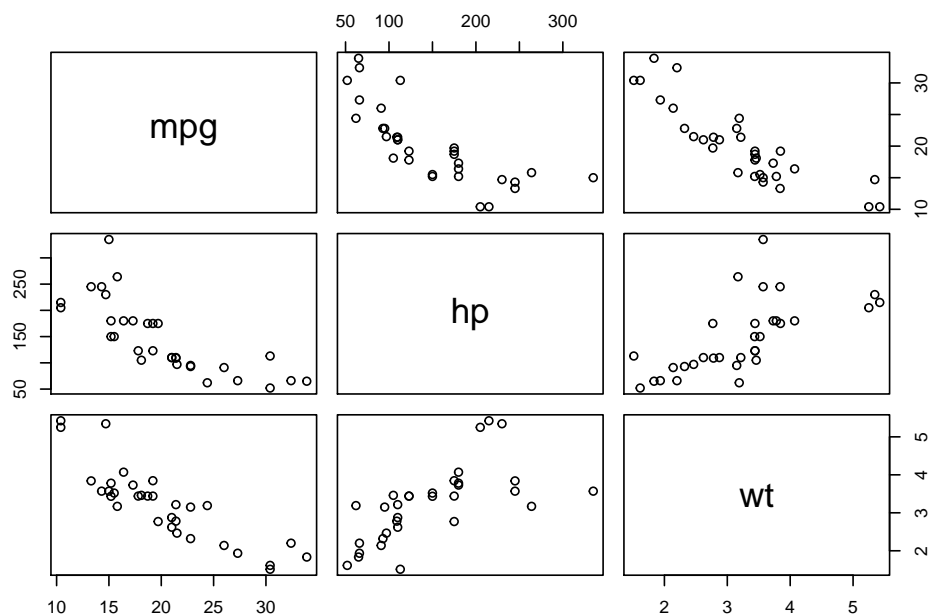
```
# Diagrama de dispersión con R base
plot(dat$hp, dat$mpg)
```



Para visualizar la relación entre más de 2 variables se puede usar la función `pair()`. En este caso limitamos el ejemplo a tres variables:

¹Es importante recordar que probamos una relación *lineal* entre las dos variables ya que usamos la correlación de Pearson. Puede darse el caso de que exista una relación entre las dos variables en la población, pero esta relación puede no ser lineal.

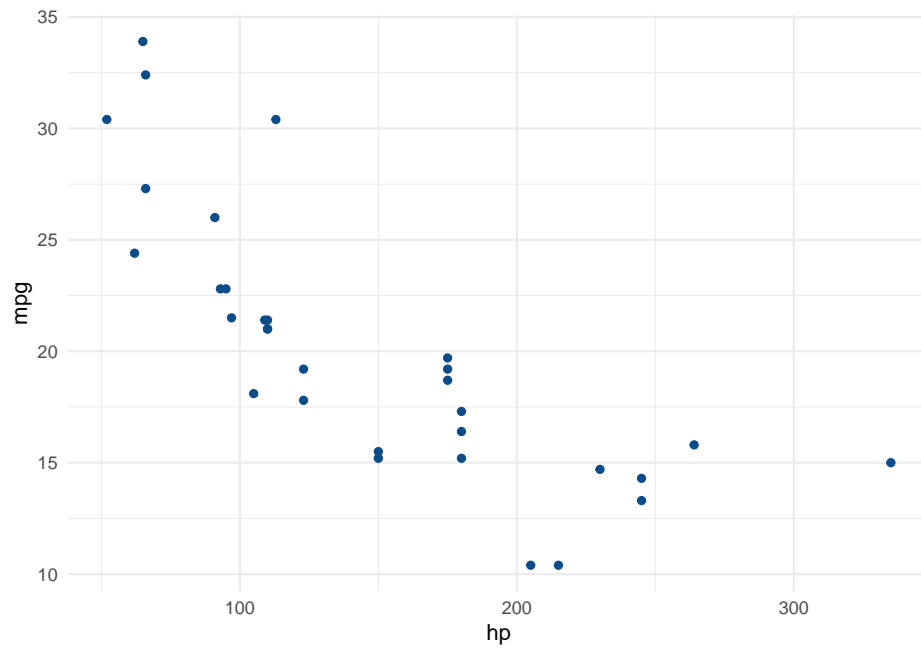

```
# Múltiples diagramas de dispersión
pairs(dat[, c(1, 4, 6)])
```



Por otra parte, existen numerosas librerías de R que permiten generar este tipo de gráficos con distintas opciones. Por ejemplo con `ggplot2`:

```
# Diagrama de dispersión con ggplot2
library(ggplot2)

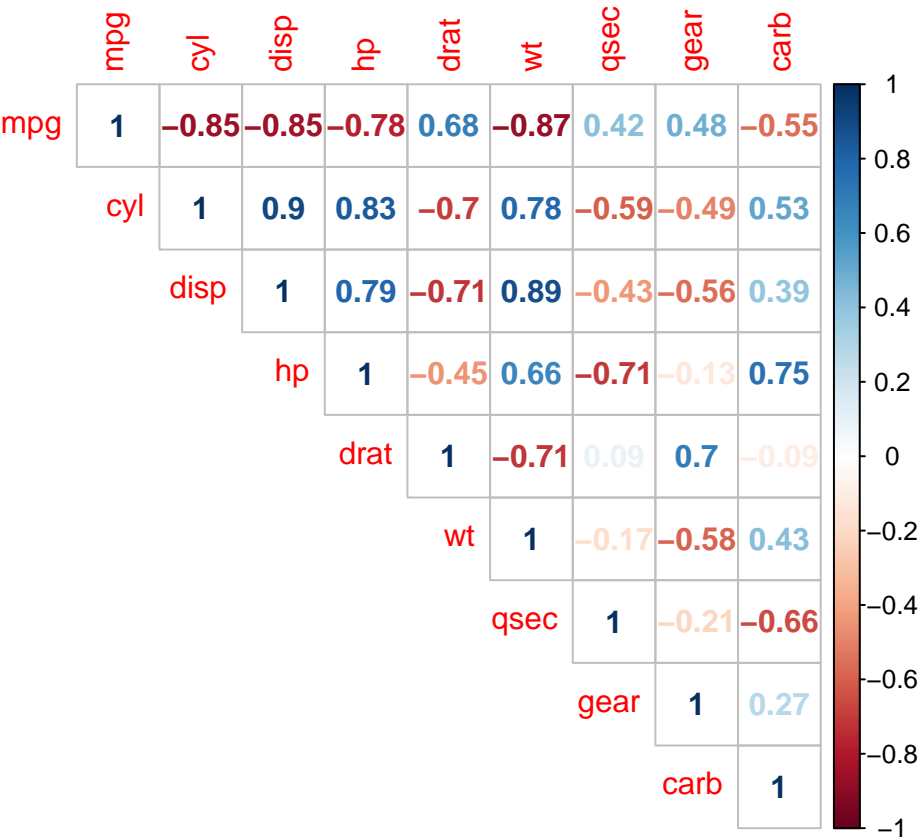
ggplot(dat) +
  aes(x = hp, y = mpg) +
  geom_point(colour = "#0c4c8a") +
  theme_minimal()
```



O representaciones más modernas como con la librería `corrplot`:

```
# improved correlation matrix
library(corrplot)

corrplot(cor(dat),
  method = "number",
  type = "upper" # show only upper side
)
```



2.4 Ejercicios

1. Estableced en clase un debate sobre variables de carácter territorial que consideréis que pueden estar correlacionadas y en qué medida.

Capítulo 3

Estadística inferencial

La **estadística descriptiva** es la rama de la estadística que tiene como objetivo **describir y resumir un conjunto de datos** de la mejor manera posible, es decir, con la menor pérdida de información posible. Con la estadística descriptiva no hay incertidumbre, porque describimos solo el grupo de observaciones en las que decidimos trabajar y no se intenta generalizar las características observadas o estudiar un grupo más grande a partir de un conjunto de datos limitado.

Por otro lado, **La estadística inferencial** es la rama de la estadística que utiliza una muestra aleatoria de datos tomados de una población para hacer inferencias, es decir, **sacar conclusiones sobre la población de interés**. En otras palabras, la información de la muestra se utiliza para hacer generalizaciones sobre el *parámetro* de interés en la población.

Las dos herramientas más importantes utilizadas en estadística inferencial son los test de hipótesis y los intervalos de confianza.

Ver caso de bulos o noticias *fake*

La estadística inferencial proporciona las herramientas que necesitamos para responder a este tipo de preguntas, y dado que este tipo de preguntas es una pieza fundamental de lo que podríamos denominar **lenguaje científico**. Sin embargo, la inferencia estadística se basa en la teoría de la probabilidad. Aquí no vamos a hablar de probabilidad pero, ya que la teoría de la probabilidad sustenta gran parte de las estadísticas, vale la pena cubrir algunos de los conceptos básicos.

3.1 Diferencia entre probabilidad y estadística

Probabilidad y estadística son dos disciplinas que están estrechamente relacionadas pero no son idénticas. La teoría de la probabilidad es *“la doctrina*

de las posibilidades”. Es una rama de las matemáticas que estiaud con qué frecuencia ocurrirán diferentes tipos de eventos. Por ejemplo, todas estas preguntas son cosas que puede responder usando la teoría de la probabilidad:

- ¿Cuáles son las posibilidades de que una moneda corriente salga cara 10 veces seguidas?
- Si tiro dos dados de seis caras, ¿qué probabilidad hay de que saque dos seises?
- ¿Qué posibilidades hay de que cinco cartas extraídas de una baraja perfectamente mezclada sean corazones?
- ¿Cuáles son las posibilidades de que me toque la lotería?

Todas estas preguntas tienen algo en común. En cada caso, la “verdad” se conoce de antemano, y cada pregunta se relaciona con “qué tipo de eventos” sucederán. En la primera pregunta, sabiendo que no se trata de una moneda trucada, hay un 50% de posibilidades de que cualquier lanzamiento de moneda individual salga cara. En la segunda pregunta, sabemos que la posibilidad de sacar un 6 en un solo dado es de 1 entre 6. En la tercera pregunta, conocemos también el número de cartas y que han sido barajadas *perfectamente*. En la cuarta pregunta, también se conocen las reglas específicas de cada juego (Euromillones, Primitiva, Lotería de Navidad, etc). El punto crítico es que las preguntas probabilísticas comienzan con un modelo conocido del mundo, y usamos ese modelo para hacer algunos cálculos. El modelo subyacente puede ser bastante simple. Por ejemplo, en el ejemplo del lanzamiento de una moneda, podemos escribir el modelo de esta manera:

$$P(\text{caras}) = 0.5$$

que se puede leer como “la probabilidad de que salga cara es 0,5 sobre 1” (las probabilidades son solo números que van del 0 al 1). Utilizamos este modelo pero, no se sabe exactamente lo que va a pasar. Todo es posible. Tal vez salgan diez caras, como dice la pregunta, pero tal vez consiga tres caras. Dicho de otro modo, en la teoría de la probabilidad, el modelo es conocido, pero los datos no.

En cambio, las preguntas estadísticas funcionan al revés. En estadística, no conocemos la “verdad” pero tenemos algunos datos, y es a partir de los datos que queremos aprender la “verdad”. Las preguntas estadísticas tienden a parecerse más a estas:

- Si alguien lanza una moneda 10 veces y obtiene 10 caras, ¿me están haciendo trampas?
- Si cinco cartas de la parte superior de la baraja son todos corazones, ¿qué probabilidad hay de que la baraja se haya barajado?
- Si un político gana n veces seguidas a la lotería, ¿qué probabilidades hay de que nos ensté engañando?

Esta vez, lo único que tenemos son datos. Se sabe lo que ha sucedido y se infiere si todo ha sucedido de un modo *normal* o si hay alguna regla que no se ha

respetado. Los datos que tenemos se ven así:

Cara Cara Cara Cara Cara Cara Cara Cara Cara Cara

y lo que tratamos de averiguar es si debemos confiar en que esto sea “verdad”. Si la moneda es una moneda común, entonces el modelo que debo adoptar es uno que diga que la probabilidad de que salga cara es 0.5; es decir, $P(\text{caras}) = 0.5$. Si la moneda está trucada, entonces debería concluir que la probabilidad de que salga cara es *no* 0.5, que escribiríamos como $P(\text{caras}) \neq 0.5$. En otras palabras, el problema de la inferencia estadística es averiguar cuál de estos dos modelos de la realidad es el correcto. Así pues, la pregunta estadística no es la misma que la pregunta de probabilidad, pero están profundamente conectadas entre sí. Debido a esto, una buena introducción a la teoría estadística comenzará con una discusión sobre qué es la probabilidad y cómo funciona.

3.2 Probabilidad frecuentista vs Bayesiana

3.2.1 Enfoque frecuentista

El enfoque predominante para el estudio de la probabilidad en estadística, se conoce como *punto de vista frecuentista*, y define la probabilidad como una *frecuencia a largo plazo*. Lanzando una moneda que tiene $P(\text{caras}) = 0.5$ podría suceder lo siguiente:

Cruz, Cara, Cara, Cara, Cara, Cruz, Cruz, Cara, Cara, Cara, Cara, Cruz, Cara, Cara, Cruz, Cruz, Cruz, Cruz, Cruz, Ca

En este caso, 11 de estas 20 monedas (55%) salieron cara. Ahora supongamos que he estado llevando un recuento continuo del número de caras (que llamaré N_{caras}) que he visto, en los primeros N volteos, y calculo la proporción de caras N_{caras}/N cada vez. Esto es lo que obtendría (¡literalmente lancé monedas para producir esto!):

Lanzamientos	Caras	Proporción
1	0	0.00
2	1	0.50
3	2	0.67
4	3	0.75
5	4	0.80
6	4	0.67
7	4	0.57
8	5	0.63
9	6	0.67
10	7	0.70
11	8	0.73
12	8	0.67
13	9	0.69
14	10	0.71
15	10	0.67
16	10	0.63
17	10	0.59
18	10	0.56
19	10	0.53
20	11	0.55

Al comienzo de la secuencia, la *proporción* de caras fluctúa mucho, comenzando en .00 y subiendo hasta .80. Después de un cierto número de lanzamientos, da la impresión de que se la proporción disminuye un poco, y que cada vez más valores se acercan bastante a la respuesta que sabemos que es la “correcta” (.50). Esta es la definición **frecuentista** de probabilidad.

Lanza una moneda común una y otra vez, y a medida que N crece (se acerca al infinito, denotado $N \rightarrow \infty$), la proporción de caras se acercará al 50%.

Simulando con un ordenador es posible lanzar una moneda virtual 1000 veces para ver lo que sucede con la proporción N_{caras}/N a medida que aumenta N . Los resultados se muestran en la Figura 3.1. La *proporción de caras observadas* finalmente deja de fluctuar y se estabiliza; cuando lo hace, el número en el que finalmente se asienta es la verdadera probabilidad de que salga cara.

La definición frecuentista de probabilidad resulta bastante interesante puesto que: 1. Es objetiva. La probabilidad de un evento está *necesariamente* basada en el la realidad. Los enunciados de probabilidad pueden tener sentido si se refieren a (una secuencia de) eventos que ocurren en el universo físico. 2. Es inequívoca. Dos personas cualesquiera que observen el desarrollo de la misma secuencia de eventos, tratando de calcular la probabilidad de un evento, inevitablemente deben llegar a la misma respuesta.

Sin embargo, también hay que tener ciertas precauciones ya que: 1. Las

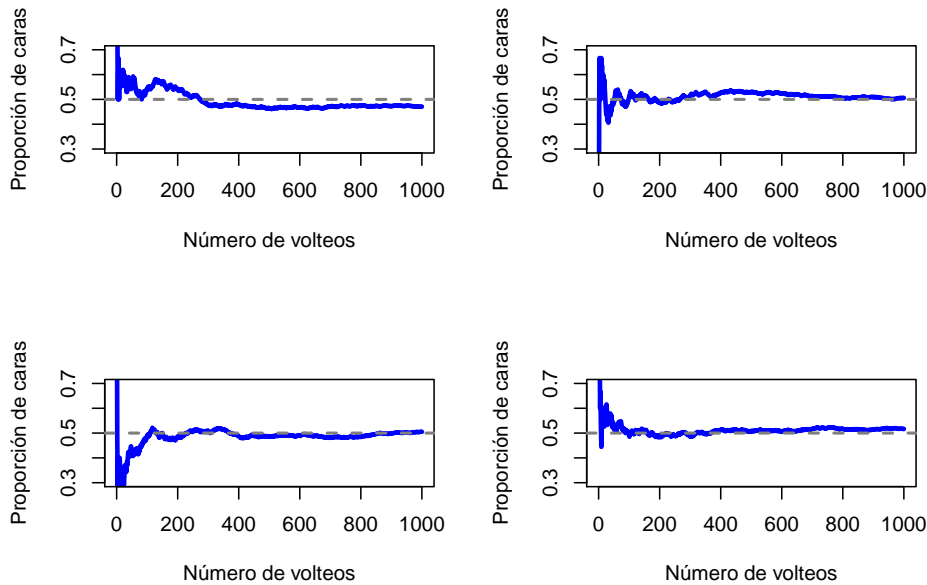


Figura 3.1: Una ilustración de cómo funciona la probabilidad frecuentista. Si lanzamos una moneda una y otra vez, la proporción de caras que ha visto finalmente se estabiliza y converge a la probabilidad real de 0.5. Cada panel muestra cuatro experimentos simulados diferentes: en cada caso, simulamos que lanzamos una moneda 1000 veces y seguimos la pista de la proporción de lanzamientos que fueron caras a medida que avanzamos. Aunque ninguna de estas secuencias en realidad terminó con un valor exacto de .5, si hubiéramos extendido el experimento para un número infinito de lanzamientos de monedas, lo hubieran hecho.

secuencias infinitas no existen en el mundo físico. Por ejemplo, si se toma una moneda y se comienza a lanzar al suelo infinitas veces, cada vez que aterriza, impacta contra el suelo. Cada impacto desgasta un poco la moneda; eventualmente, la moneda quedará modificada y ya no volverá a ser la misma.

2. La definición frecuentista tiene un alcance limitado. Por ejemplo, si un meteorólogo aparece en la televisión y dice: “la probabilidad de que llueva en Adelaida el 2 de noviembre de 2022 es del 60%”, no está claro cómo definir esto en términos frecuentistas. Solo hay una ciudad de Adelaida, y solo habrá un 2 de noviembre de 2022. No hay una secuencia infinita de eventos aquí, solo una vez. La probabilidad frecuentista no contempla hacer enunciados de probabilidad sobre un solo evento. Desde la perspectiva frecuentista, mañana lloverá o no; no hay “probabilidad” que se adhiera a un solo evento no repetible.

3.2.2 Enfoque Bayesiano

El enfoque bayesiano de la probabilidad a menudo se denomina visión subjetivista, y es una visión minoritaria entre los estadísticos, pero que ha ido ganando terreno de manera constante durante las últimas décadas. La forma más común de pensar sobre la probabilidad subjetiva es definir la probabilidad de un evento como el grado de creencia que alguien inteligente y racional asigna a la probabilidad de ese evento. Según esto, las probabilidades no existen en el mundo, sino en los pensamientos y suposiciones de las personas. Sin embargo, para que este enfoque funcione, necesitamos alguna forma de operacionalizar “grado de creencia”. Supongamos que creo que hay un 60% de probabilidad de que llueva mañana. Si alguien me ofrece una apuesta: si mañana llueve, gano 5 euros, pero si no llueve, pierdo 5 euros. Claramente, desde mi perspectiva, esta es una apuesta bastante buena. Por otro lado, si creo que la probabilidad de lluvia es solo del 40%, entonces es una mala apuesta. Por lo tanto, podemos operacionalizar la noción de una “probabilidad subjetiva” en términos de las apuestas que estoy dispuesto a aceptar.

¿Cuáles son las ventajas y desventajas del enfoque bayesiano? La principal ventaja es que te permite asignar probabilidades a cualquier evento. No es necesario que se limite a los eventos que se pueden repetir. La principal desventaja es que no podemos ser puramente objetivos: especificar una probabilidad requiere que especifiquemos una entidad que tenga el grado de creencia relevante. Esta entidad puede ser un humano, un extraterrestre, un robot o incluso un estadístico, pero tiene que haber un agente inteligente que crea en las cosas. Para mucha gente esto es incómodo: parece hacer que la probabilidad sea arbitraria. Si bien el enfoque bayesiano requiere que el agente en cuestión sea racional (es decir, obedezca las reglas de probabilidad), sí permite que todos tengan sus propias creencias; Puedo creer que la moneda es común y tú no tienes que hacerlo, aunque ambos seamos racionales. La visión frecuentista no permite que dos observadores atribuyan diferentes probabilidades al mismo evento: cuando eso sucede, al menos uno de ellos debe estar equivocado. La visión bayesiana no evita que esto ocurra. Dos

observadores con conocimientos previos diferentes pueden tener legítimamente creencias diferentes sobre el mismo evento. En resumen, donde la visión frecuentista a veces se considera demasiado estrecha (prohíbe muchas cosas a las que queremos asignar probabilidades), la visión bayesiana a veces se piensa que es demasiado amplia (permite demasiadas diferencias entre observadores).

3.3 Introducción a las distribuciones de probabilidad

Una distribución de probabilidad es una función que describe la probabilidad de obtener los posibles valores que puede asumir una variable aleatoria. En otras palabras, los valores de la variable varían según la distribución de probabilidad subyacente.

Supongamos que seleccionamos una muestra aleatoria de personas y medimos la altura de los sujetos. A medida que vamos midiendo las alturas, podemos crear una distribución de alturas. Este tipo de distribución es útil cuando necesita saber qué resultados son más probables, la dispersión de los valores potenciales y la probabilidad de resultados diferentes. Por lo tanto se puede utilizar distribuciones de probabilidad para realizar inferencias.

Ejemplo a partir de la película “El sargento de hierro” de Clint Eastwood.

Supongamos que el profesor solo tiene 5 jerseys (X_1, X_2, X_3, X_4 y X_5). Cada prenda (es decir, cada X) sería un **evento elemental**. La característica clave de los eventos elementales es que cada vez que hacemos una observación (por ejemplo, cada vez que me pongo un jersey), el resultado será uno y solo uno de estos eventos. De manera similar, el conjunto de todos los eventos posibles se denomina **espacio muestral**.

Definido el espacio muestral, que se construye a partir de muchos posibles eventos elementales (jerseys), lo que queremos hacer es asignar una **probabilidad** de uno de estos eventos elementales. Para un evento X , la probabilidad de ese evento $P(X)$ es un número que se encuentra entre 0 y 1. Cuanto mayor sea el valor de $P(X)$, es más probable que ocurra el evento. Entonces, por ejemplo, si $P(X) = 0$, significa que el evento X es imposible (es decir, nunca uso ese jersey). Por otro lado, si $P(X) = 1$ significa que el evento X seguramente ocurrirá (es decir, siempre uso ese jersey). Todos los demás valores, entre 0 y 1, significarían que unas veces uso un jersey y otras veces otros. Por ejemplo, si $P(X) = 0.5$ significa que uso ese jersey la mitad del tiempo.

Las probabilidades de los todos los eventos elementales deben sumar 1. Esto se conoce como la **ley de la probabilidad total**. Si se satisfacen estos requisitos, entonces lo que tenemos es una **distribución de probabilidad**. Por ejemplo, este es un ejemplo de distribución de probabilidad

Jersey	Azul	Gris	Naranja	Amarillo	Marrón
Etiqueta	X_1	X_2	X_3	X_4	X_5
Probabilidad	$P(X_1) = .5$	$P(X_2) = .3$	$P(X_3) = .1$	$P(X_4) = 0$	$P(X_5) = .1$

Cada uno de los eventos tiene una probabilidad que se encuentra entre 0 y 1, y si sumamos la probabilidad de todos los eventos, suman 1. Impresionante. Incluso podemos dibujar un bonito gráfico de barras para visualizar esta distribución, como se muestra en la Figura ??.

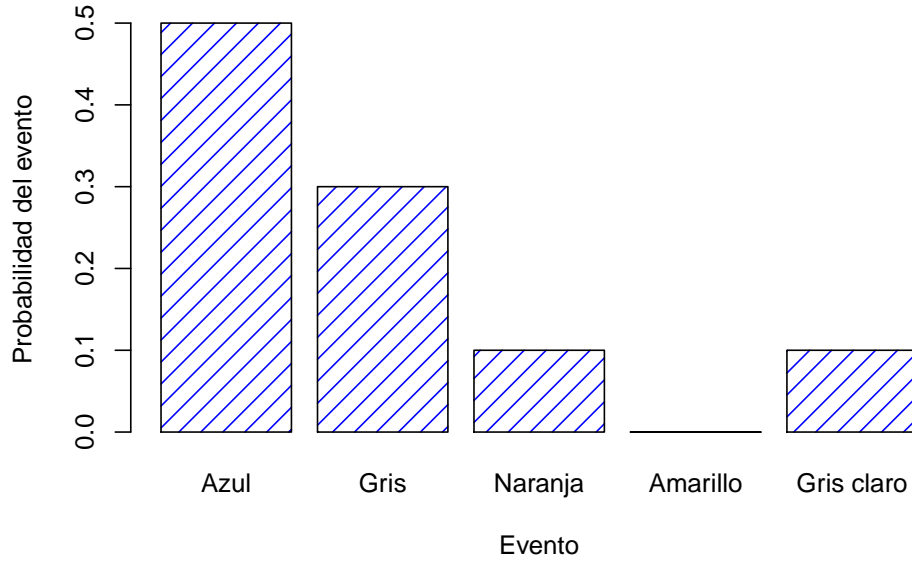


Figura 3.2: Representación visual de la distribución de probabilidad de jerseys del profesor. Hay cinco eventos elementales, correspondientes a los cinco jerseys. Cada evento tiene alguna probabilidad de ocurrir: esta probabilidad es un número entre 0 y 1. La suma de estas probabilidades es 1.

Cabe ser señalado es que la teoría de la probabilidad permite hablar de eventos elementales y también de **eventos no elementales**. En el ejemplo de los jerseys, es perfectamente legítimo referirse a la probabilidad de que vista un color claro. En términos matemáticos, definimos el evento “jersey de color claro” E para que corresponda al conjunto de eventos elementales (X_1, X_2, X_3) . Si ocurre alguno de estos eventos elementales, también se dice que ha ocurrido E . Habiendo decidido escribir la definición de E de esta manera, es bastante sencillo establecer cuál es la probabilidad $P(E)$: simplemente sumamos todo. En este caso particular

$$P(E) = P(X_1) + P(X_2) + P(X_3)$$

y, dado que las probabilidades de los jerseys amarillo, naranja y gris claro, respectivamente, son .1, 0 y .1, la probabilidad de que use un jersey de color claro es igual a .2.

Tabla 3.1: Algunas reglas básicas que deben cumplir las probabilidades.

Expresión	Notación	NANA	Formula
No \$A\$	$P(\neg A)$	=	$1-P(A)$
\$A\$ o \$B\$	$P(A \cup B)$	=	$P(A) + P(B) - P(A \cap B)$
\$A\$ y \$B\$	$P(A \cap B)$	=	$P(A B) P(B)$

A partir de estos principios tan simples es posible construir algunas herramientas matemáticas extremadamente poderosas. En la Tabla 3.1 aparecen algunas de las otras reglas que satisfacen las probabilidades.

Las distribuciones de probabilidad varían enormemente. Sin embargo, no todas son igualmente importantes. Las más utilizadas serían: la distribución binomial, la distribución normal, la distribución t , la distribución χ^2 (“chi-cuadrado”) y la distribución de F . Aquí prestaremos especial atención a la binomial y a la normal.

3.3.1 La distribución binomial

La teoría de la probabilidad se originó en el intento de describir cómo funcionan los juegos de azar, por lo que parece apropiado que nuestra discusión sobre la **distribución binomial** incluya una discusión sobre el lanzamiento de dados y monedas. Imaginemos un “experimento” simple: en un cubilete hay 20 dados idénticos de seis caras. En una cara de cada dado hay una imagen de un bufón (*joker*) y las otras cinco caras están todas en blanco. Si lanzamos los 20 dados, ¿cuál es la probabilidad de que obtenga exactamente 4 *jokers*? Suponiendo que los dados no estén trucados, sabemos que la probabilidad de que se obtenga un *joker* es de 1 en 6; Para decir esto de otra manera, la probabilidad de *joker* para un solo dado es aproximadamente .167.

Si N es el número de tiradas de dados en nuestro experimento; que a menudo se denomina **parámetro de tamaño** de nuestra distribución binomial. Mientras que θ es la probabilidad de que un solo dado produzca un *joker*, una cantidad que generalmente se llama **probabilidad de éxito** del binomio. Finalmente, X serán resultados de nuestro experimento, es decir, el número de *jokers* obtenido al tirar los dados. Dado que el valor real de X se debe al azar, nos referimos a él como **variable aleatoria**. La cantidad que queremos calcular es la probabilidad de que $X = 4$ dado que sabemos que $\theta = .167$ y $N = 20$. La “forma” general de lo que me interesa calcular podría escribirse como

$$P(X | \theta, N)$$

y estamos interesados en el caso especial donde $X = 4$, $\theta = .167$ y $N = 20$. Si quiero decir que X se genera aleatoriamente a partir de una distribución

binomial con los parámetros θ y N , la notación que usaría es la siguiente:

$$X \sim \text{Binomial}(\theta, N)$$

La distribución binomial tiene el aspecto que muestran en la Figura 3.3 y traza las probabilidades binomiales para todos los valores posibles de X . Si se lanzan los dados, desde $X = 0$ (sin *jokers*) hasta $X = 20$ (todos los *jokers*). Esto es básicamente un gráfico de barras, y no es diferente del gráfico de “probabilidad de jerseys” de la Figura 3.2. En el eje horizontal tenemos todos los eventos posibles y en el eje vertical podemos leer la probabilidad de cada uno de esos eventos. Entonces, la probabilidad de sacar 4 *jokers* de 20 veces es de aproximadamente 0.20. En otras palabras, esperaríamos que eso sucediera aproximadamente el 20% de las veces que se lancen los dados.

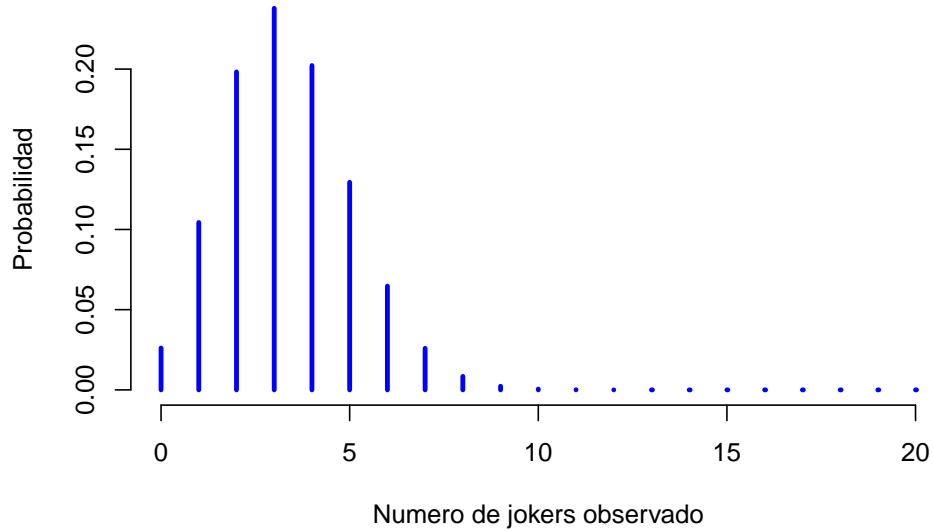


Figura 3.3: La distribución binomial con parámetro de tamaño de $N = 20$ y una probabilidad de éxito subyacente de $\theta = 1/6$. Cada barra vertical representa la probabilidad de un resultado específico (es decir, un valor posible de X). Debido a que esta es una distribución de probabilidad, cada una de las probabilidades debe ser un número entre 0 y 1, y las alturas de las barras también deben sumar 1.

Como se ha visto en la tabla ??, R tiene una función llamada `dbinom()` que calcula probabilidades binomiales. Los principales argumentos de la función son:

- **x.** Éste es un número o vector, que especifica los resultados cuya probabilidad se está tratando de calcular.
- **size.** Este es un número que le dice a R el tamaño del experimento.
- **prob.** Ésta es la *probabilidad de éxito* de cualquier ensayo del experimento.

3.3. INTRODUCCIÓN A LAS DISTRIBUCIONES DE PROBABILIDAD 47

Entonces, para calcular la probabilidad de obtener $x = 4$ *jokers*, a partir de un experimento de `size = 20` ensayos, en el que la probabilidad de obtener un *joker* en cualquier ensayo es `prob = 1/6`, el comando que usaríamos es:

```
dbinom( x = 4, size = 20, prob = 1/6 )
```

```
## [1] 0.2022036
```

Para ver cómo cambia la distribución binomial cuando modificamos los valores de θ y N , cambiemos los dados por monedas. De este modo, la probabilidad de éxito ahora es $\theta = 1/2$. Suponiendo que se lanzara la moneda $N = 20$ veces. Es decir que estamos cambiando la probabilidad de éxito, pero manteniendo el tamaño del experimento. Como muestra la Figura 3.4, el efecto principal de esto es cambiar toda la distribución. ¿Y si lanzamos una moneda $N = 100$ veces? Bueno, en ese caso, se obtiene la distribución que aparece en la Figura 3.5. La distribución se mantiene aproximadamente en el medio, pero hay un poco más de variabilidad en los posibles resultados.

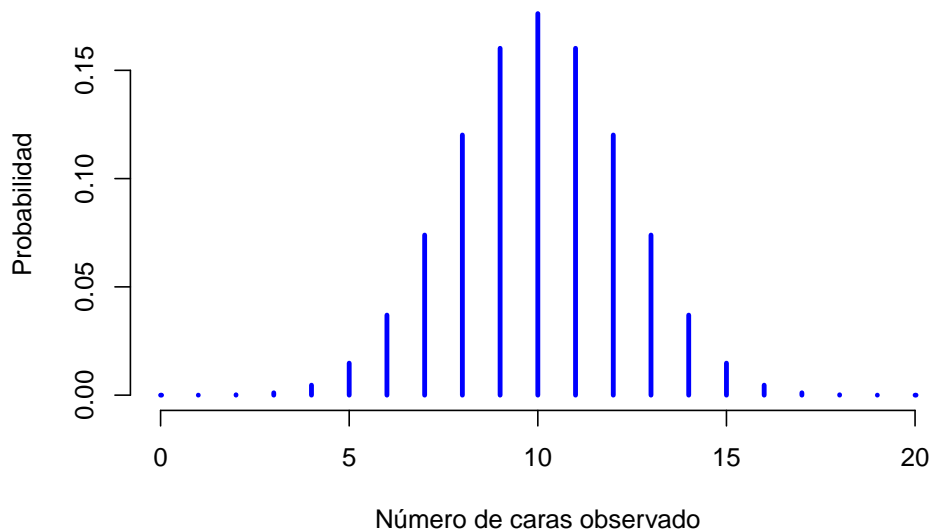


Figura 3.4: Dos distribuciones binomiales, que involucran un escenario en el que estoy lanzando una moneda, por lo que la probabilidad de éxito subyacente es $\theta = 1/2$. Aquí asumimos que estoy lanzando la moneda $N = 20$ veces.

La fórmula de la distribución binomial que calcula R es la siguiente: $P(X|\theta, N) = \frac{N!}{X!(N-X)!} \theta^X (1-\theta)^{N-X}$

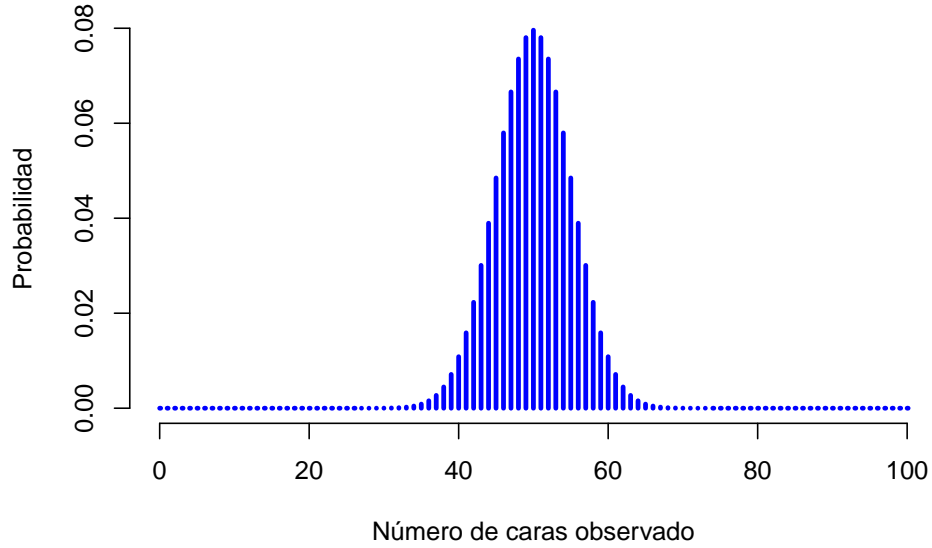


Figura 3.5: Dos distribuciones binomiales, que involucran un escenario en el que se lanza una moneda, por lo que la probabilidad de éxito subyacente es $\theta = 1/2$. Aquí asumimos que la moneda se lanza $N = 100$ veces.

3.3.2 La distribución normal

La fórmula de la distribución normal que calcula R es la siguiente: $p(X|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X - \mu)^2}{2\sigma^2}\right)$

La **distribución normal**, que también se conoce como “de campana” o una “distribución gaussiana” es la distribución más utilizada. Una distribución normal se describe usando dos parámetros, la media de la distribución μ y la desviación estándar de la distribución σ . La notación que a veces usamos para decir que una variable X se distribuye normalmente es la siguiente:

$$X \sim \text{Normal}(\mu, \sigma)$$

3.3.3 Funciones de R para distribuciones de probabilidad

R tiene varias funciones para trabajar con cada distribución de probabilidad. Hay un nombre de raíz, por ejemplo, el nombre de raíz para la distribución normal es `norm`. Esta raíz tiene como prefijo una de las letras:

- `p` para “probabilidad”, la función de distribución acumulativa.
- `q` para “cuantil”, el inverso de la función de distribución acumulativa.
- `d` para “densidad”, la función de densidad.

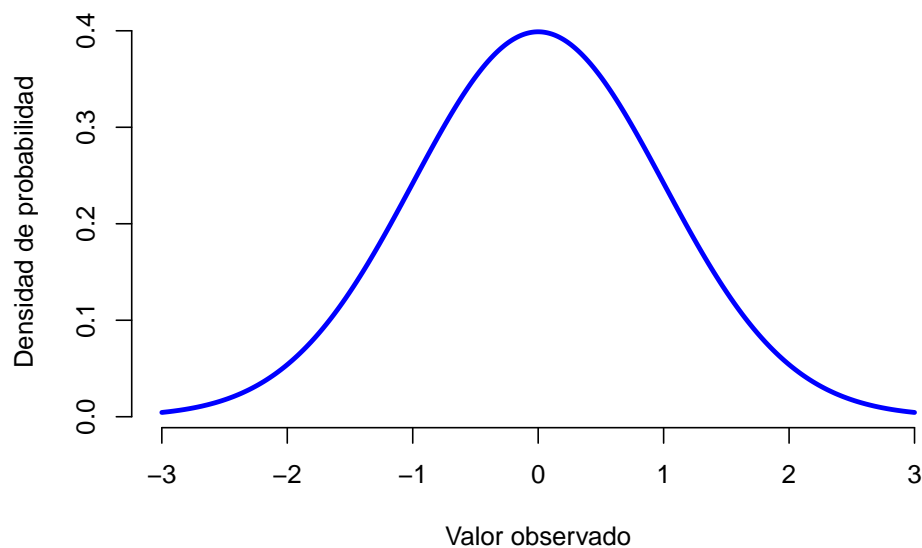


Figura 3.6: {La distribución normal con media $\mu = 0$ y desviación estándar $\sigma = 1$. El eje x corresponde al valor de alguna variable, y el eje y nos dice algo sobre la probabilidad de que observemos ese valor. Sin embargo, observe que el eje y está etiquetado como “Densidad de probabilidad” y no como “Probabilidad”. Existe una característica sutil y algo frustrante de las distribuciones continuas que hace que el eje y se comporte un poco extraño: la altura de la curva aquí no es en realidad la probabilidad de observar un valor particular de x . Por otro lado, *es* cierto que las alturas de la curva indican qué valores de x son más probables.

- **r** para “aleatorio”, una variable aleatoria que tiene la distribución especificada.

Las cuatro versiones de la cada función requieren que especifiquen los argumentos **size** y **prob**. Sin embargo, difieren en términos de cuál es el otro argumento y cuál es el resultado:

- La forma **d** requiere un resultado particular **x**, y el devuelve la probabilidad de obtener exactamente ese resultado.
- La forma **p** calcula la **probabilidad acumulada**. Se le da un cuantil particular **q**, y se le dice la probabilidad de obtener un resultado *menor o igual que q*.
- La forma **q** calcula los **cuantiles** de la distribución. Se especifica un valor de probabilidad **p** y devuelve el percentil correspondiente. Es decir, el valor de la variable para el que existe una probabilidad **p** de obtener un resultado menor que ese valor.
- La forma **r** es un **generador de números aleatorios**: específicamente, genera **n** resultados aleatorios de la distribución.

Para la distribución normal, estas funciones son **pnorm**, **qnorm**, **dnorm** y **rnorm**, mientras que para la distribución binomial, estas funciones son **pbinom**, **qbinom**, **dbinom** y **rbinom**.

Para una distribución continua (como la normal), las funciones más útiles para resolver problemas que involucran cálculos de probabilidad son las funciones **p** y **q**, porque la densidad por la función **d** solo se puede usar para calcular probabilidades a través de integrales.

Para una distribución discreta (como la binomial), la función **d** calcula la densidad, Que en este caso es una probabilidad

$$f(x) = P(X = x)$$

y por lo tanto es útil para calcular probabilidades.

R tiene funciones para manejar muchas distribuciones de probabilidad. La siguiente tabla proporciona los nombres de las funciones para cada distribución y un enlace a la documentación en línea que es la referencia autorizada sobre cómo se utilizan las funciones. Pero no lea la documentación en línea todavía. Primero, pruebe los ejemplos de las secciones que siguen a la tabla.

3.3.4 Ejemplos con la distribución Binomial

De nuevo, si lanzamos dados, y cada dado tiene una probabilidad de 1 en 6 de obtener *jokers*, supongamos, que queremos saber la probabilidad de sacar 4 o menos jockers. Podríamos usar la función **dbinom ()** para calcular la probabilidad exacta de obtener 0 *jokers*, 1 *joker*, 2 *jokers*, 3 *jokers* y 4 *jokers* y luego sumarmos, pero hay una manera más rápida. En su lugar, se puede usar la función **pbinom ()**:

Tabla 3.2: (#tab:table:probability-functions)Funciones para distribución de probabilidades in R.

Distribución	p	q	d	r
Beta	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Exponential	pxp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Wilcoxon Rank Sum Statistic	pnwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank Statistic	psignrank	qsignrank	dsignrank	rsignrank

```
pbinom(q= 4, size = 20, prob = 1/6)
```

```
## [1] 0.7687492
```

En otras palabras, hay un 76,9% de posibilidades de que saque 4 *jokers* o menos. R dice que un valor de 4 es en realidad el percentil 76,9 de esta distribución binomial.

A continuación, consideremos la función `qbinom()`. Digamos que queremos calcular el percentil 75 de la distribución binomial. Siguiendo con el ejemplo de los dados:

```
qbinom( p = 0.75, size = 20, prob = 1/6)
```

```
## [1] 4
```

Lo que la función `qbinom()` parece estar diciendo es que el percentil 75 de la distribución binomial es 4, aunque según en la función `pbinom()` se sabe que 4 es *en realidad* el percentil 76,9. La rareza aquí proviene del hecho de que nuestra distribución binomial realmente no *tiene* un percentil 75. Hay un 56,7% de posibilidades de sacar 3 *jokers* o menos (ver `pbinom(3, 20, 1/6)`) y un 76,9% de posibilidades de sacar 4 calaveras o menos. Entonces, en cierto sentido el percentil 75 debería estar “entre” 3 y 4 *jokers*. Pero aquí los decimales no tienen sentido. Este problema se puede manejar de diferentes maneras:

1. Se puede informar un valor intermedio (o un valor *interpolado*) como 3.9,
2. Se puede redondear a la baja (a 3) o hacia arriba (a 4).

La función `qbinom()` redondea hacia arriba si se solicita un percentil que en realidad no existe (como el 75 en este ejemplo), R encuentra el valor más pequeño para el cual el rango percentil es *al menos* lo que se pidió. En este caso, dado que el percentil 75 “verdadero” se encuentra entre 3 y 4 *jokers*, R redondea y devuelve un valor de 4. Esto solo es un problema para distribuciones discretas como la binomial.

Finalmente, tenemos el generador de números aleatorios (`rbinom()`). Hay que especificar cuántas veces R debe “simular” el experimento usando el argumento `n`, y generará resultados aleatorios a partir de la distribución binomial. Entonces, por ejemplo, supongamos que se tuviera que repetir el experimento de lanzamiento de dados 100 veces. Podría hacer que R simule los resultados de estos experimentos usando el siguiente comando:

```
rbinom( n = 100, size = 20, prob = 1/6 )
```

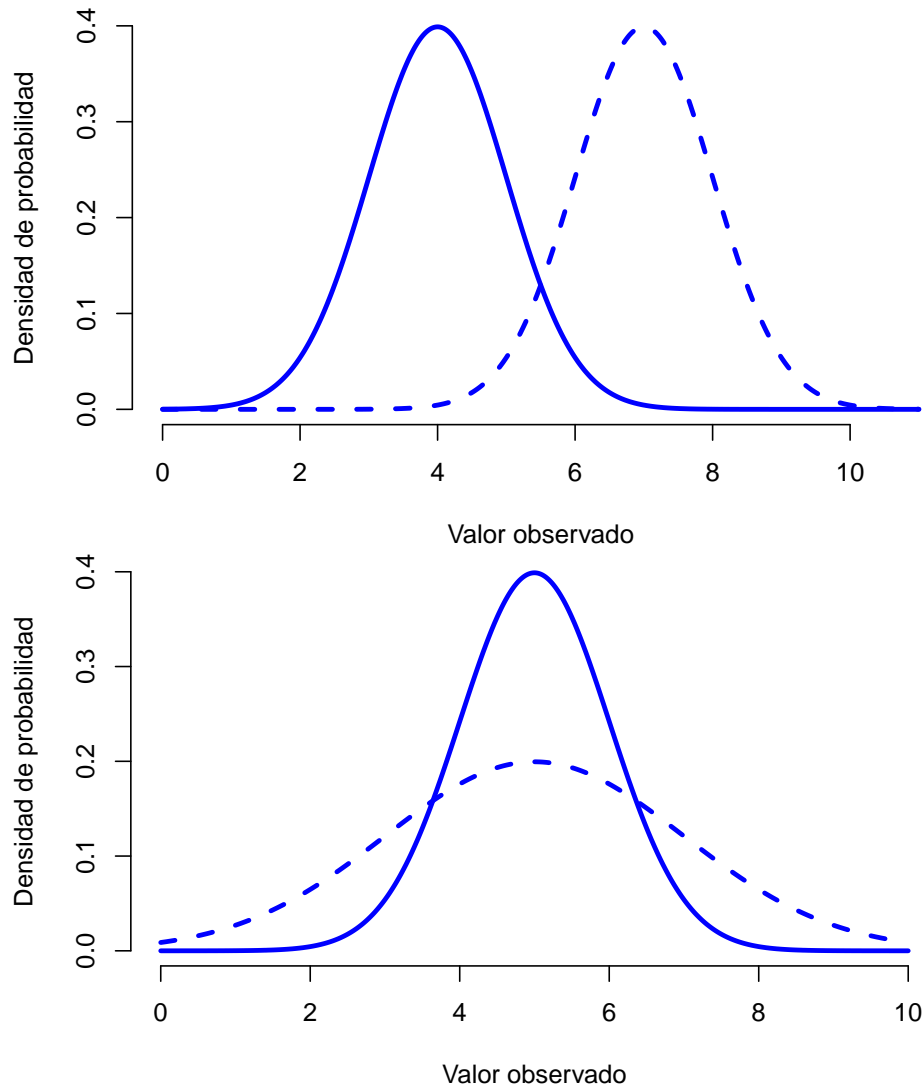
```
## [1] 3 1 4 5 3 4 1 2 4 3 3 5 6 6 6 6 3 4 1 3 3 4 3 3 2 2 3 5 3 4 4 5 3 6 4 6 3
## [38] 5 2 3 4 3 6 3 0 2 1 3 1 4 4 4 5 1 2 5 7 6 4 6 4 1 3 3 4 3 4 2 2 2 2 2 2
## [75] 2 2 3 3 6 4 3 4 2 5 4 2 3 3 6 8 2 2 4 5 2 2 2 5 1 4
```

Como puede ver, estos números son más o menos los que se puede ver en la Figura @ref(fig: binomial1). La mayoría de las veces se obtienen entre 1 y 5

jokers.

3.3.5 Ejemplos con la distribución normal

Como ya se ha dicho, las funciones R para la distribución normal son `dnorm()`, `pnorm()`, `qnorm()` y `rnorm()`. Sin embargo, se comportan prácticamente de la misma manera que las funciones correspondientes para la distribución binomial, por lo que no hay mucho que deba saber. Únicamente, cabe mencionar, que los nombres de los argumentos para los parámetros son `mean` y `sd`.



En la Figura 3.6, se traza una distribución normal con media $\mu = 0$ y desviación

estándar $\sigma = 1$. En vez de un histograma, la imagen de la distribución normal en la Figura 3.6 muestra una curva suave. Esta no es una elección arbitraria: la distribución normal es continua, mientras que la binomial es discreta. Las escalas continuas no tienen esta restricción. Por ejemplo, la temperatura de un día de primavera podría ser de 23 grados, 24 grados, 23,9 grados o cualquier punto intermedio, ya que la temperatura es una variable continua, por lo que una distribución normal podría ser muy apropiada para describir las temperaturas de primavera.

Mencionar el caso de las escalas Likert

La Figura ?? traza distribuciones normales que tienen diferentes medias, pero tienen la misma desviación estándar. Como era de esperar, todas estas distribuciones tienen el mismo “ancho”. La única diferencia entre ellos es que se han desplazado hacia la izquierda o hacia la derecha. En todos los demás aspectos, son idénticos. Por el contrario, si aumentamos la desviación estándar mientras mantenemos la media constante, el pico de la distribución permanece en el mismo lugar, pero la distribución se ensancha, como puede ver en la Figura ?. Sin embargo, cuando ampliamos la distribución, la altura del pico se reduce. Esto tiene que suceder: de la misma manera que las alturas de las barras que usamos para dibujar una distribución binomial discreta tienen que *sumar* 1, el área total *bajo la curva* para la distribución normal debe ser igual a 1. No obstante, independientemente de cuál sea la media real y la desviación estándar, el 68,3% del área se encuentra dentro de 1 desviación estándar de la media. Del mismo modo, el 95,4% de la distribución se encuentra dentro de 2 desviaciones estándar de la media y el 99,7% de la distribución está dentro de 3 desviaciones estándar. Esta idea se ilustra en la Figura ?.

3.4 Introduccion a los test de hipótesis

Primero cabe preguntarse por qué intentaríamos hacer inferencias sobre un parámetro de una población basándonos en una muestra, en lugar de simplemente recopilar datos para toda la población, calcular estadísticas que nos interesan y tomar decisiones basadas en eso. La principal razón por la que utilizamos una muestra en lugar de toda la población es porque recopilar datos sobre toda la población es más complicado o en ocasiones impracticable por varios motivos (complejidad, coste, limitación de tiempo, entre muchos otros motivos).¹

El objetivo general de una prueba de hipótesis es sacar conclusiones para confirmar o refutar una creencia sobre una población basándonos en un grupo más pequeño de observaciones.

¹Por ejemplo, una investigación podría consistir en conocer si la población de la provincia de Tarragona está satisfecha con el nuevo plan de movilidad. Si pudiéramos preguntar a toda la población en un período de tiempo razonable, no haríamos ninguna estadística inferencial. No obstante, aún habría que decidir qué preguntas se les hace para entender mejor el motivo de su grado de satisfacción, complicando y encareciendo aún más la encuesta.

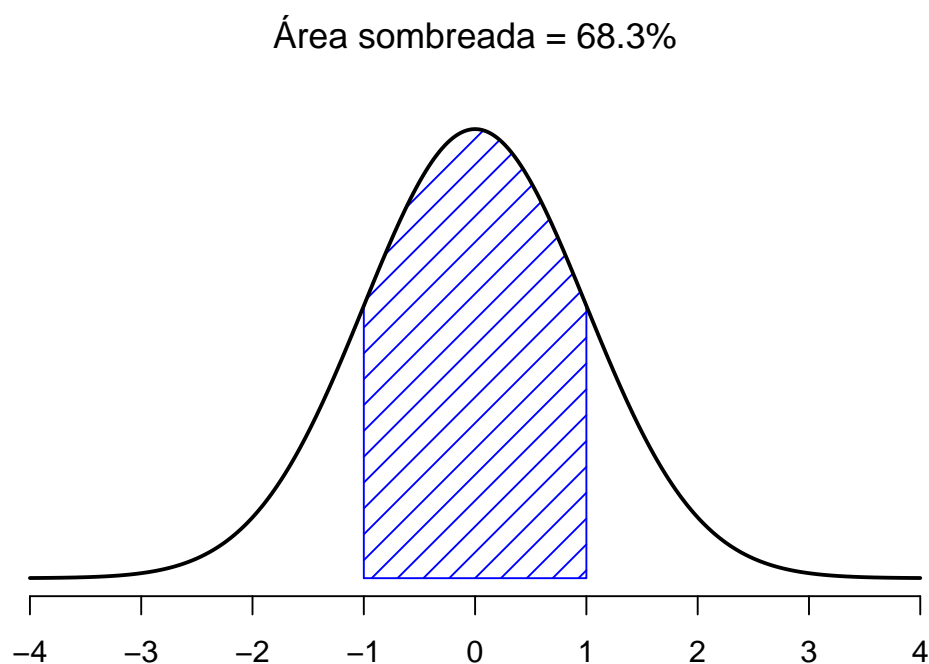


Figura 3.7: El área debajo de la curva indica la probabilidad de que una observación se encuentre dentro de un rango particular. Las líneas continuas trazan distribuciones normales con media $\mu = 0$ y desviación estándar $\sigma = 1$. Las áreas sombreadas ilustran “áreas bajo la curva” para dos casos importantes. Aquí podemos ver que hay un 68,3% de probabilidad de que una observación caiga dentro de una desviación estándar de la media.

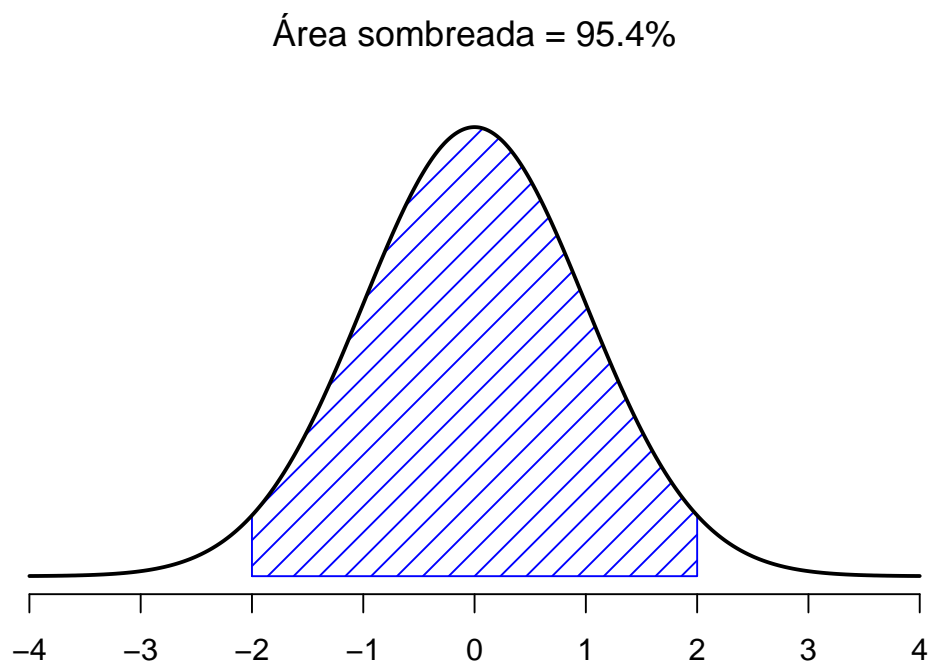


Figura 3.8: El área debajo de la curva indica la probabilidad de que una observación se encuentre dentro de un rango particular. Las líneas continuas trazan distribuciones normales con media $\mu = 0$ y desviación estándar $\sigma = 1$. Las áreas sombreadas ilustran “áreas bajo la curva” para dos casos importantes. Aquí vemos que hay un 95,4% de probabilidad de que una observación caiga dentro de dos desviaciones estándar de la media.

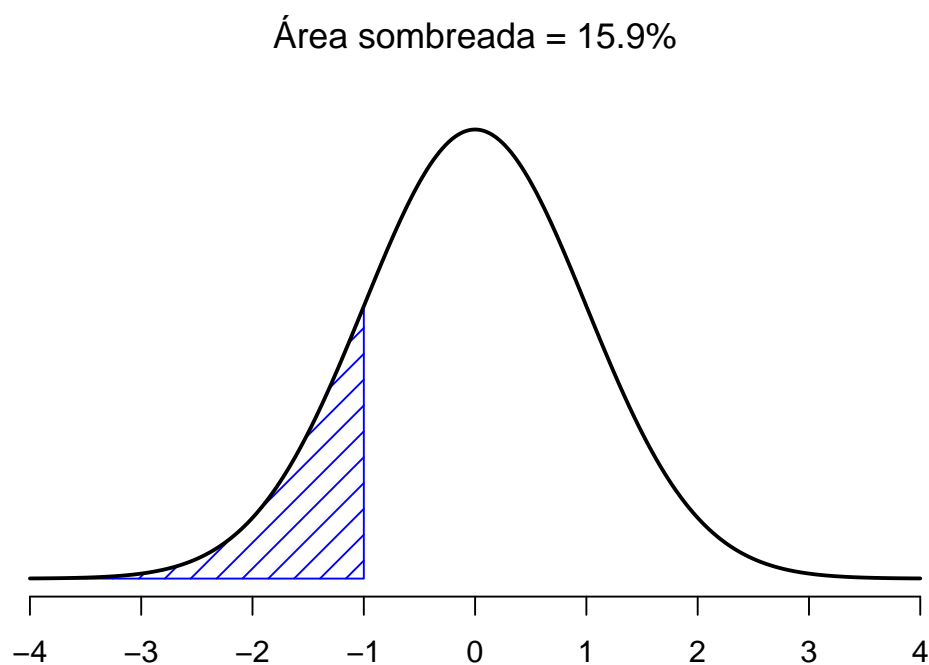


Figura 3.9: Dos ejemplos más de la “idea del área bajo la curva”. Hay un 15.9% de probabilidad de que una observación esté una desviación estándar por debajo de la media o menor.

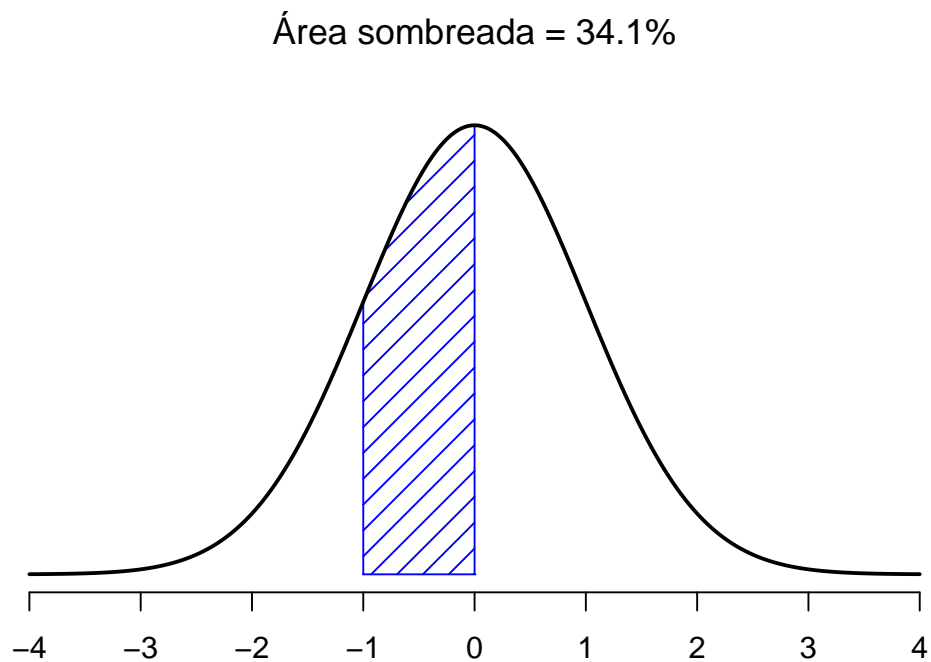


Figura 3.10: Existe una probabilidad del 34,1% de que la observación sea mayor que una desviación estándar por debajo de la media pero aún por debajo de la media. Si suma estos dos números, se obtiene $15.9 + 34.1 = 50$. Para datos distribuidos normalmente, existe un 50% de probabilidad de que una observación caiga por debajo de la media. Y, por supuesto, eso también implica que hay un 50% de probabilidad de que esté por encima de la media.

Las pruebas de hipótesis tienen muchas aplicaciones prácticas. Aquí ponemos algunos ejemplos:

1. Una media: supongamos que a un político le gustaría probar si el salario medio de los trabajadores españoles es diferente de 1200 euros.
2. Dos medias:
 - Muestras independientes: supongamos que a un profesor le gustaría probar la idoneidad de un nuevo método docente midiendo la nota media para los alumnos de un grupo de control y los del grupo en el que se ha introducido la novedad.
 - Muestras pareadas o relacionadas: supongamos que el mismo profesor quisiera probar la idoneidad de este nuevo método de enseñanza pero lo hiciera midiendo los conocimientos de los alumnos antes y después de la explicación.
3. Una proporción: supongamos que a un analista político quisiera comprobar si la proporción de ciudadanos que van a votar por un candidato específico es inferior al 25%.
4. Dos proporciones: supongamos que a un geógrafo le gustaría probar si la proporción de visitantes a una playa es diferente entre jóvenes y personas mayores.
5. Una variación: supongamos que un ingeniero quisiera probar si una batería tiene una variabilidad en el tiempo de carga menor que la indicada en la descripción técnica.
6. Dos variaciones: supongamos que, en una fábrica, dos líneas de producción funcionan independientemente una de la otra. El gerente querría probar si los costes del mantenimiento semanal de estas dos cadenas de producción tienen la misma variación.

Por supuesto, hay muchísimas más aplicaciones potenciales y muchas preguntas de investigación pueden responderse gracias a una prueba de hipótesis.

Por lo general, **las pruebas de hipótesis se utilizan para responder preguntas de investigación en análisis confirmatorios**. Los análisis confirmatorios se refieren a análisis estadísticos donde las hipótesis — deducidas de la teoría — se definen de antemano (preferiblemente antes de la recopilación de datos). En este enfoque, la investigadora tiene una idea específica sobre las variables en consideración y está tratando de ver si su idea, especificada como hipótesis, está respaldada por datos.

Podemos utilizar al menos tres métodos diferentes para realizar una prueba de hipótesis comparando:

1. la estadística de prueba con el **valor crítico**.
2. el **p-valor** con el nivel de significancia α .
3. el parámetro objetivo con el **intervalo de confianza**.

Estos enfoques pueden diferir en algunos aspectos pero tienen muchos puntos en común. El uso de uno u otro método es a menudo una cuestión de elección personal o de contexto.

Para los tres métodos, explicaré los pasos necesarios para realizar una prueba de hipótesis desde un punto de vista general y los ilustraré con la siguiente situación:².

Supongamos que un político quisiera comprobar si el salario medio de los trabajadores españoles es diferente de 1.200 euros.

En la mayoría de las pruebas de hipótesis, la prueba que vamos a utilizar como ejemplo a continuación requiere algunas condiciones. En esta sección asumimos que se cumplen todos los supuestos pero más adelante hablaremos de esto.

3.4.1 Comparando la estadística de prueba con el valor crítico.

Este metodo consiste en reproducir los siguientes 4 pasos:

1. Establecer la **hipótesis nula y alternativa**
2. Calcular la **estadística de prueba**
3. Encontrar el **valor crítico**
4. **Concluir** e interpretar los resultados

3.4.1.1 Estableciendo la hipótesis nula y alternativa

Una prueba de hipótesis primero requiere una suposición sobre un fenómeno o hipótesis, que se deriva de la teoría y la pregunta de investigación.

Dado que una prueba de hipótesis se utiliza para confirmar o refutar una creencia previa, necesitamos **formular nuestra creencia de modo que haya una hipótesis nula y una alternativa**. Esas hipótesis deben ser **mútuamente excluyentes**, lo que significa que no pueden ser verdaderas al mismo tiempo. En el contexto, las hipótesis nula y alternativa son así:

- Hipótesis nula $H_0 : \mu = 1200$
- Hipótesis alternativa $H_1 : \mu \neq 1200$

Al plantear la hipótesis nula y alternativa, tenga en cuenta los siguientes tres puntos:

1. *Siempre estamos interesados en la población y no en la muestra.* Esta es la razón por la que H_0 y H_1 siempre se escribirán en términos de población y no en términos de muestra (en este caso, μ y no \bar{x}).
2. *La suposición que nos gustaría probar es a menudo la hipótesis alternativa.* Si quisieramos probar si el salario medio de los trabajadores españoles es inferior a 1200 euros, habríamos establecido que $H_0 : \mu = 1200$ (o equivalentemente, $H_0 : \mu \geq 1200$) y $H_1 : \mu < 1200$. **No hay que**

²Puede ver más o menos pasos en otros artículos o libros de texto, dependiendo de si estos pasos son detallados o concisos. Sin embargo, la prueba de hipótesis debe seguir el mismo proceso independientemente del número de pasos

confundir la hipótesis nula con la alternativa, o las conclusiones serán diametralmente opuestas.

3. La hipótesis nula es a menudo el *status quo*. Por ejemplo, suponiendo que un empresario quiere probar si el nuevo logo de su marca es mejor valorado que el logo anterior. El *status quo* es que los dos logos sean igualmente valorados. Suponiendo que un valor mayor es mejor, entonces se escribirá $H_0 : \mu_{nuevo} = \mu_{viejo}$ (o equivalentemente, $H_0 : \mu_{nuevo} - \mu_{viejo} = 0$) y $H_1 : \mu_{nuevo} > \mu_{viejo}$ (o equivalentemente, $H_0 : \mu_{nuevo} - \mu_{viejo} > 0$). Por el contrario, si cuanto más bajo mejor, habríamos escrito $H_0 : \mu_{nuevo} = \mu_{viejo}$ (o equivalentemente, $H_0 : \mu_{nuevo} - \mu_{viejo} = 0$) y $H_1 : \mu_{nuevo} < \mu_{viejo}$ (o equivalentemente, $H_0 : \mu_{nuevo} - \mu_{viejo} < 0$).

3.4.1.2 Calcular la estadística de prueba

La **estadística de prueba** (o **t-stat**) es una métrica que indica **qué tan extremas son las observaciones en comparación con la hipótesis nula**. Cuanto mayor sea el *t*-stat (en valor absoluto), más extremas serán las observaciones.

Hay varias fórmulas para calcular el t-stat, con una fórmula para cada tipo de prueba de hipótesis: una o dos medias, una o dos proporciones, una o dos varianzas. Esto significa que hay una fórmula para calcular el t-stat para una prueba de hipótesis en una media, otra fórmula para una prueba en dos medias, otra para una prueba en una proporción, etc.³ La única dificultad en este segundo paso es elegir la fórmula adecuada. Tan pronto como se sepa qué fórmula utilizar según el tipo de prueba, simplemente debe aplicársele a los datos. Afortunadamente, las fórmulas para las pruebas de hipótesis en una y dos medias, y una y dos proporciones siguen la misma estructura.

Calcular la estadística de prueba para estas pruebas es similar a *escalar* una variable aleatoria (un proceso también conocido como “estandarización” o “normalización”) que consiste en restar la media de esa variable aleatoria y dividir el resultado por la desviación estándar:

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

Para estas 4 pruebas de hipótesis (una/dos medias y una/dos proporciones), calcular el estadístico de prueba es como escalar el estimador (calculado a partir de la muestra) correspondiente al parámetro de interés (en la población). Así que básicamente restamos el parámetro objetivo del estimador puntual y luego dividimos el resultado por el error estándar (que es equivalente a la desviación estándar, pero para un estimador).

Si esto no está claro, así es como se calcula la estadística de prueba (t_{obs}) en nuestro ejemplo (asumiendo que se desconoce la varianza de la población):

³Incluso hay diferentes fórmulas dentro de cada tipo de prueba, dependiendo de si se cumplen o no algunos supuestos.

$$t_{obs} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

dónde:

- \bar{x} es la media de la muestra (es decir, el estimador)
- μ es la media bajo la hipótesis nula (es decir, el parámetro objetivo)
- s es la desviación estándar de la muestra
- n es el tamaño de la muestra
- $(\frac{s}{\sqrt{n}})$ es el error estándar)

Suponiendo que en nuestro caso tenemos una media muestral de 1150 euros ($\bar{x} = 1150$), una desviación estándar muestral de 200 euros ($s = 200$) y un tamaño de muestra de 30 trabajadores ($n = 30$) y, teniendo en cuenta que la media poblacional (la media bajo la hipótesis nula) es 1200 euros ($\mu = 1200$), el t-stat quedaría así:

$$t_{obs} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{1150 - 1200}{\frac{200}{\sqrt{30}}} = -1.369306$$

Aunque las fórmulas son diferentes según el parámetro que esté probando, el valor encontrado para la estadística de prueba nos da una indicación de cuán extremas son nuestras observaciones.

Recordemos este valor de -1.369306 porque se volverá a utilizar al final de este test para compararlo con el valor crítico.

3.4.1.3 Encontrando el valor crítico

Aunque el **t-stat** nos da una indicación como de extremas son nuestras observaciones, necesitamos comparar este valor con un umbral o **valor crítico**, que viene dado por una **distribución de probabilidad**.

De la misma manera que la fórmula para calcular el **t-stat** es diferente para cada parámetro de interés, la distribución de probabilidad subyacente en la que se basa el *valor crítico* también es diferente para cada parámetro objetivo. Esto significa que, además de elegir la fórmula apropiada para calcular el **t-stat**, también necesitamos seleccionar la distribución de probabilidad apropiada dependiendo del parámetro que estemos probando.

Afortunadamente, solo hay 4 distribuciones de probabilidad diferentes para las pruebas de hipótesis cubiertas aquí (recordemos que son una/dos medias, una/dos proporciones y una/dos varianzas):

1. Distribución normal estándar:

- prueba en una y dos medias con varianzas de población conocidas.
- prueba en dos muestras donde se conoce la varianza de la diferencia entre las 2 muestras σ_D^2

- prueba en una y dos proporciones (dado que se cumplen algunos supuestos).
- 2. Distribución de Student:
 - prueba en una y dos medias con *varianza(s) de población desconocida(s).
 - prueba en dos muestras donde la varianza de la diferencia entre las 2 muestras σ_D^2 es *desconocida*.
- 3. Distribución Chi-cuadrado:
 - prueba en una varianza.
- 4. Distribution de fisher:
 - prueba en dos varianzas.

Cada distribución de probabilidad tiene sus propios parámetros, definiendo su forma y/o ubicación. Los parámetros de una distribución de probabilidad pueden verse como si fuesen marcadores de ADN; lo que significa que la distribución está completamente definida por su(s) parámetro(s).

Volviendo a nuestra investigación, la distribución de probabilidad subyacente de una prueba en una media es la distribución Normal estándar o de Student, dependiendo de si la varianza de la *población* (no la varianza de la muestra) es conocida o no:

- Si se conoce la varianza de la población \rightarrow , se usa la distribución Normal estándar
- Si la varianza de la población es *desconocida* \rightarrow , se utiliza la distribución de Student

Si no se proporciona explícitamente la varianza de la población, se puede suponer que es desconocida, ya que no se puede calcular basándonos en una muestra. Si pudiera calcularlo, eso significaría que tiene acceso a toda la población y, en este caso, no tiene sentido realizar una prueba de hipótesis (simplemente podría usar algunas estadísticas descriptivas para confirmar o refutar su creencia dicha hipótesis. En nuestro ejemplo, no se especifica la varianza de la población, por lo que se supone que es desconocida. Por lo tanto, usaremos la distribución de Student.

La distribución Student tiene un parámetro que la define: el número de grados de libertad. El número de grados de libertad depende del tipo de prueba de hipótesis. Por ejemplo, el número de grados de libertad para una prueba en una media es igual al número de observaciones menos uno ($n - 1$). Sin ir demasiado lejos en los detalles, el -1 proviene del hecho de que hay una cantidad que se estima (es decir, la media). Siendo el tamaño de la muestra igual a 30 en nuestro ejemplo, los grados de libertad son iguales a $n - 1 = 30 - 1 = 29$.

Por último, para encontrar el valor crítico también es necesario conocer el **nivel de significancia** α , que es la **probabilidad de rechazar erróneamente la hipótesis nula aunque en realidad sea verdadera**. En este sentido, es un error de tipo I (en contraposición al error de tipo II) que aceptamos para poder sacar conclusiones sobre una población a partir de un subconjunto de ella.

En muchas aplicaciones el nivel de significancia se suele establecer en el 5%. En cambio, en algunos campos (como la medicina o la ingeniería, entre otros), el nivel de significancia también se establece a veces en el 1% para disminuir la tasa de error. Es mejor especificar el nivel de significancia *antes* de realizar una prueba de hipótesis para evitar la tentación de establecer el nivel de significancia de acuerdo con los resultados (la tentación es aún mayor cuando los resultados están al borde de ser significativos). En nuestro caso, tomamos $\alpha = 5\% = 0.05$.

Además, queremos probar si el salario medio de los trabajadores españoles es **diferente** de 1200 euros. Si quisiéramos probar que el salario medio fuera inferior a 1200 euros ($H_1 : \mu < 1200$) o superior a 1200 ($H_1 : \mu > 1200$), habríamos realizado una prueba unilateral. Asegúrese de realizar la prueba correcta (bilateral o unilateral) porque tiene un impacto en cómo encontrar el valor crítico.

Ahora que conocemos la distribución apropiada (distribución de Student), su parámetro (grados de libertad (gl) = 29), el nivel de significancia ($\alpha = 0.05$) y la dirección (bilateral), tenemos todo lo que necesitamos para calcular el valor crítico. Podríamos localizar este valor en la tabla estadística correspondiente o directamente lo podríamos calcular con R.

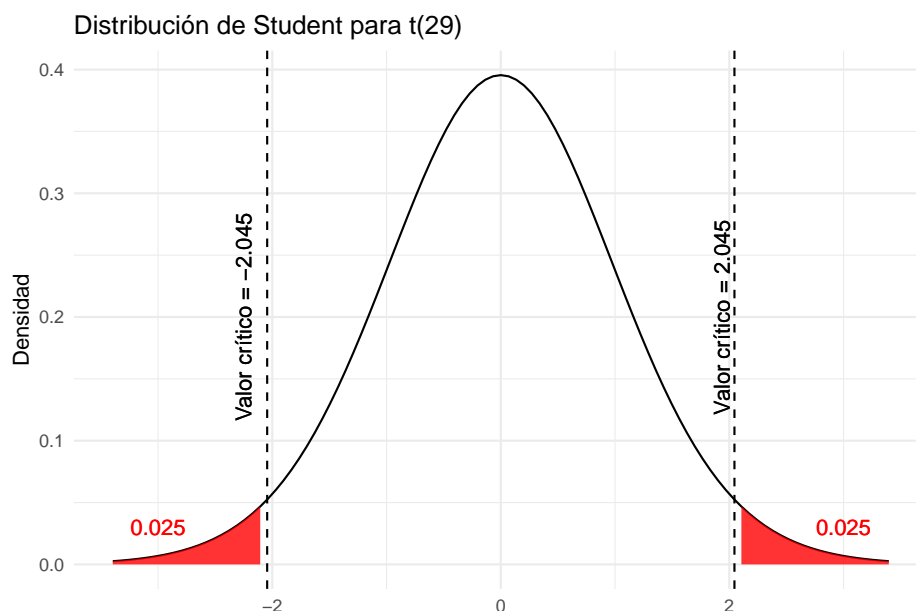
Al observar la fila $df = 29$ y la columna $t_{.025}$ en la tabla de distribución de Student, encontramos un valor crítico de:

$$t_{n-1;\alpha/2} = t_{29;0.025} = 2.04523$$

Tomamos $t_{\alpha/2} = t_{.025}$ y no $t_{\alpha} = t_{.05}$ ya que el nivel de significancia es 0.05 y estamos haciendo una prueba bilateral (de dos lados; $H_1 : \mu \neq 1200$), por lo que la tasa de error de 0.05 debe dividirse en 2 para encontrar el valor crítico a la derecha de la distribución. Dado que la distribución de Student es simétrica, el valor crítico a la izquierda de la distribución es simplemente: -2.04523.

Visualmente, la tasa de error de 0.05 se divide en dos partes:

- 0,025 a la izquierda de -2,04523 y
- 0,025 a la derecha de 2,04523



Al igual que en el apartado anterior, cabe recordar estos valores críticos de -2.045 y 2.045 el último paso.

Las áreas sombreadas en rojo en el gráfico anterior también se conocen como regiones de rechazo.

Estos valores críticos también se pueden encontrar en R, gracias a la función `qt()`:

```
qt(0.025, df = 29, lower.tail = TRUE)
```

```
## [1] -2.04523
```

```
qt(0.025, df = 29, lower.tail = FALSE)
```

```
## [1] 2.04523
```

Como se ha visto en el tema sobre distribuciones de probabilidad, la función `qt()` se usa para la distribución de Student (`q` significa cuantil `yt` para Student). Cabe recordar que hay otras funciones que acompañan a las diferentes distribuciones:

- `qnorm()` para la distribución Normal
- `qchisq()` para la distribución Chi-cuadrado
- `qf()` para la distribución de Fisher

3.4.1.4 Conclusión e interpretación de los resultados

Las únicas dos posibilidades al concluir una prueba de hipótesis son:

1. Rechazo de la hipótesis nula, o
2. No rechazo de la hipótesis nula

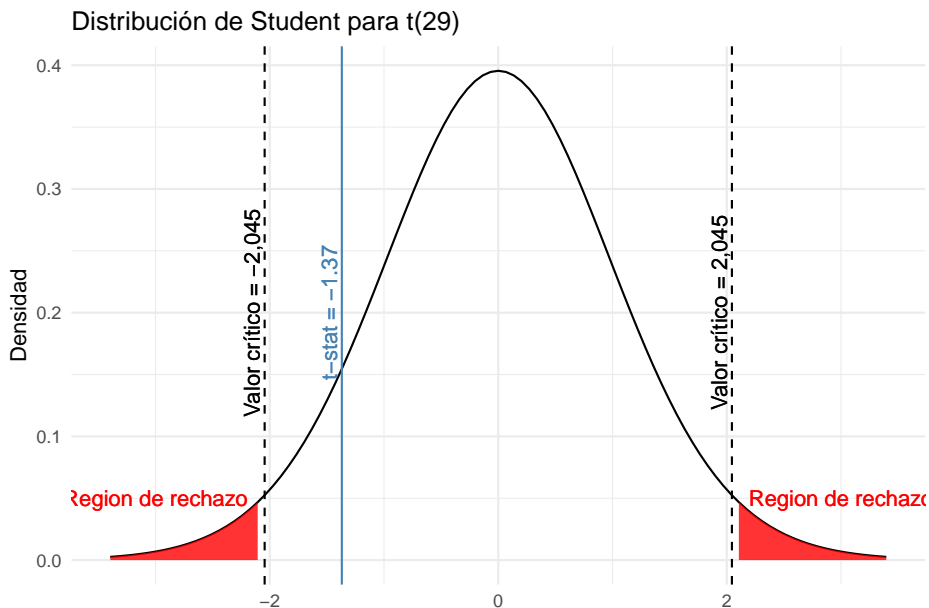
En nuestro ejemplo sobre los salarios de los españoles, recordamos que hemos determinado que el:

- el t -stat es -1.369306, y
- los valores críticos son -2.04523 y 2.04523

Recordemos que:

- el t -stat da una indicación de cuán extrema es nuestra muestra en comparación con la hipótesis nula
- los valores críticos son el umbral a partir del cual el t -stat se considera *demasiado* extremo

Para comparar el t -stat con los valores críticos de manera gráfica:



Los dos valores críticos forman las regiones de rechazo (las áreas sombreadas en rojo):

- de $-\infty$ a -2.045, y
- de 2.045 a ∞

Si el t -stat se encuentra dentro de una de estas regiones, rechazamos la hipótesis nula. Por el contrario, si t -stat *no* se encuentra dentro de ninguna de las regiones, no rechazamos la hipótesis nula.

Como podemos ver en el gráfico anterior, el t -stat es menos extremo que el valor crítico. En conclusión, no rechazamos la hipótesis nula de que $\mu = 1200$.

Esta es la conclusión en términos estadísticos, pero no tienen sentido sin una interpretación adecuada. Por tanto, es una buena práctica interpretar también el resultado en el contexto del problema:

Con un nivel de significancia del 5%, no rechazamos la hipótesis de que el salario medio de los trabajadores españoles es de 1200 euros.

¿Qué significa esto realmente? Dicho de otro modo:

“nosotros *no rechazamos* la hipótesis nula” y “nosotros *no rechazamos* la hipótesis de que el salario medio de los trabajadores españoles es igual a 1200 euros”. No escribimos “*aceptamos o estamos de acuerdo con* la hipótesis nula” ni “el salario medio de los trabajadores españoles es de 1200 euros”.

En los test de hipótesis, llegamos a una conclusión sobre la población a partir de una muestra. Por tanto, siempre existe cierta incertidumbre y no podemos decir que estemos seguros al 100% de que nuestra conclusión sea correcta.

Quizás sea el caso de que el salario medio de los trabajadores españoles sea en realidad diferente a 1200 euros, pero **no lo pudimos demostrar** con los datos disponibles. Si tuviéramos más observaciones hubiéramos rechazado la hipótesis nula (dado que todo lo demás es igual, un tamaño de muestra más grande implica un **t-stat** más extremo). O puede darse el caso de que, incluso con más observaciones, no hubiéramos rechazado la hipótesis nula porque el salario de los trabajadores españoles en realidad se acerca a los 1200 euros. Con los datos disponibles no podemos distinguir entre estas dos posibilidades. Simplemente debemos admitir que no encontramos suficiente evidencia en contra de la hipótesis de partida, pero tampoco concluimos que la media sea igual a 1200 euros.

3.4.2 Comparando el p -valor con el nivel de significancia α

Este método consiste en los siguientes pasos:

1. Enunciar las **hipótesis nula y alternativa**
2. Calcular la **estadística de prueba** (t-stat).
3. Calcular el **p -valor**
4. **Concluir** e interpretar los resultados

En este segundo método que utiliza el valor p , los dos primeros pasos son similares a los del primer método, mientras que la interpretación de los resultados tiene algunos puntos en común.

3.4.2.1 Establecer las hipótesis

Las hipótesis de investigación (nula y alternativa) siguen siendo las mismas:

- $H_0 : \mu = 1200$

- $H_1 : \mu \neq 1200$

3.4.2.2 Calcular la estadística de prueba

Cabe recordar que la fórmula del estadístico t es diferente según el tipo de prueba de hipótesis (una o dos medias, una o dos proporciones, una o dos varianzas). En nuestro caso de una sola media con varianza desconocida, tenemos que:

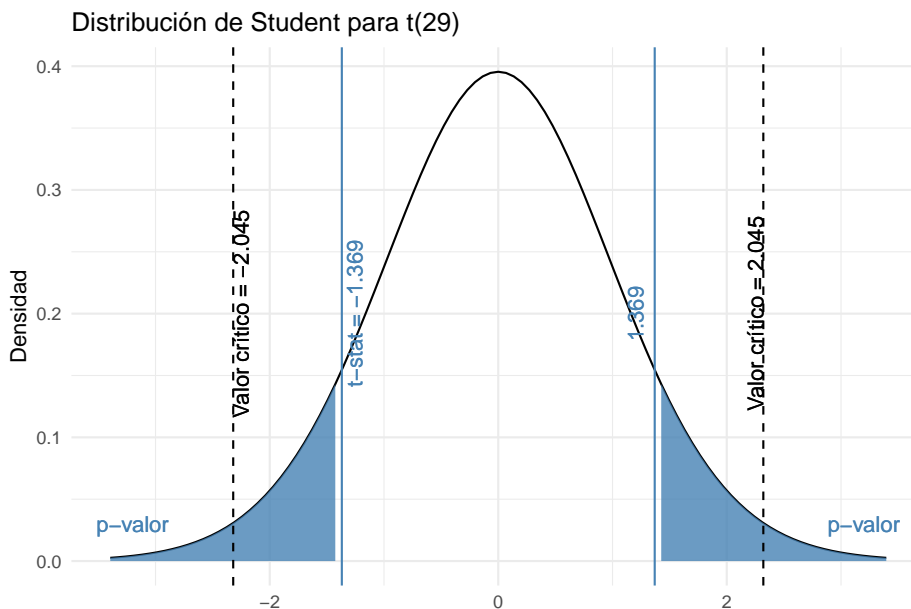
$$t_{obs} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{1150 - 1200}{\frac{200}{\sqrt{30}}} = -1.369306$$

3.4.2.3 Cálculo del valor p

El **p -valor** es la probabilidad (de 0 a 1) de observar una muestra al menos tan extrema como la que observamos si la hipótesis nula fuera cierta. Dicho de otro modo: **¿cómo de probable es la hipótesis nula?**. También se define como el nivel de significancia más pequeño para el cual los datos indican el rechazo de la hipótesis nula.

Formalmente, el valor p es el área más allá del estadístico de prueba. Como estamos haciendo una prueba bidireccional, el valor p es, por lo tanto, la suma del área por encima de 1,369306 y por debajo de -1,369306.

Visualmente, el valor p es la suma de las dos áreas sombreadas en azul en la siguiente gráfica:



El valor p se puede obtener también con tablas estadísticas o es posible calcularlo con precisión en R con la función `pt()`:

```
p_val <- pt(-1.369306, df = 29, lower.tail = TRUE) + pt(1.369306, df = 29, lower.tail = FALSE)
p_val
```

```
## [1] 0.1814156
```

```
# Que es lo mismo que...
```

```
p_val <- 2 * pt(1.369306, df = 29, lower.tail = FALSE)
p_val
```

```
## [1] 0.1814156
```

El valor p es 0.1814, que indica que hay un 18.14% de probabilidad de observar una muestra al menos tan extrema como la observada si el hipótesis nula eran verdaderas. Esto ya nos da una pista sobre si nuestro t-stat es demasiado extremo o no (y, por lo tanto, si nuestra hipótesis nula es probable o no).

Como la función `qt()` para encontrar el valor crítico, usamos `pt()` para encontrar el valor p porque la distribución subyacente es la distribución de Student. En otros casos se utilizarían las funciones `pnorm()`, `pchisq()` y `pf()` para las otras distribuciones mencionadas anteriormente (Normal, Chi-cuadrado y Fisher).

3.4.2.4 Concluir e interpretar los resultados

Finalmente, hay que comparar el valor p que acabamos de calcular con el nivel de significancia α . Como para todas las pruebas estadísticas:

- Si el **p -valor es menor** que α ($p\text{-valor} < 0.05$), entonces H_0 es poco probable \rightarrow **rechazamos** la hipótesis nula.
- Si el **p -valor es mayor que o igual** a α ($p\text{-valor} \geq 0.05$), entonces H_0 es probable \rightarrow **no podemos rechazar** la hipótesis nula.

No importa si tomamos en consideración el p -valor exacto (es decir, 0.1814) o el acotado ($0.05 < p\text{-valor} < 0.10$), es mayor que 0.05, entonces no rechazamos la hipótesis nula. En el contexto del problema, no rechazamos la hipótesis nula de que el salario medio de los trabajadores españoles es igual a 1200 euros.

El resultado obtenido ha sido el mismo que en el primer método. Evidentemente, debería dar lo mismo si se usan los mismos datos y con el mismo nivel de significancia.

3.4.3 Comparación del parámetro objetivo con el intervalo de confianza

Este método consiste en calcular primero el intervalo de confianza y comparar sobre éste el parámetro objetivo (el parámetro bajo la hipótesis nula). Podemos distinguir tres pasos:

1. Enunciar las **hipótesis nula y alternativa**

2. Calcular el **intervalo de confianza**
3. **Concluir** e interpretar los resultados

También se pueden apreciar varias similitudes con los métodos anteriores.

3.4.3.1 Enunciar las hipótesis

Nuevamente, las hipótesis nula y alternativa siguen siendo las mismas:

- $H_0 : \mu = 1200$
- $H_1 : \mu \neq 1200$

3.4.3.2 Calcular el intervalo de confianza

Al igual que los test de hipótesis, los intervalos de confianza son una herramienta bien conocida en la estadística inferencial. El **intervalo de confianza** es un procedimiento de estimación que produce un **intervalo que contiene el parámetro verdadero con una cierta probabilidad**.

De la misma manera que existe una fórmula para cada tipo de prueba de hipótesis al calcular las estadísticas de la prueba, existe una fórmula para cada tipo de intervalo de confianza. La fórmula para calcular un intervalo de confianza en una media μ (con varianza poblacional desconocida):

$$(1 - \alpha)\% \text{ IC para } \mu = \bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

donde $t_{\alpha/2, n-1}$ se encuentra en la tabla de distribución de Student o se puede calcular con R (y es similar al valor crítico encontrado en el primer método).

Dados nuestros datos y con $\alpha = 0.05$, tenemos que:

$$\begin{aligned} 95\% \text{ IC para } \mu &= \bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \\ &= 1150 \pm 2.045 \frac{200}{\sqrt{30}} \\ &= [1075, 33; 1224, 67] \end{aligned}$$

El intervalo de confianza del 95% para μ es [1075,33; 1224,67] euros. **¿Qué significa este intervalo de confianza del 95%?**

Sabemos que este procedimiento de estimación tiene una probabilidad del 95% de producir un intervalo que contenga la media verdadera μ . En otras palabras, **si construimos muchos intervalos de confianza** (con diferentes muestras del mismo tamaño), **el 95% de ellos incluirá la media de la población** (el verdadero parámetro). Del mismo modo el 5% de estos intervalos de confianza no cubrirán la media real.

Si desea disminuir este último porcentaje, puede disminuir el nivel de significancia (por ejemplo $\alpha = 0.01$). En igualdad de condiciones, esto

disminuirá el rango del intervalo de confianza y, por lo tanto, aumentará la probabilidad de que incluya el parámetro verdadero.

3.4.3.3 Conclusión e interpretación de los resultados

Finalmente, hay comparar el intervalo de confianza con el valor del parámetro objetivo (el valor cuestionado por la hipótesis nula):

- Si el **intervalo de confianza no incluye** el valor hipotético, H_0 es poco probable \rightarrow **rechazamos** la hipótesis nula.
- Si el **intervalo de confianza incluye** el valor hipotético, H_0 es probable \rightarrow , **no rechazamos** la hipótesis nula

En nuestro ejemplo:

- el valor hipotético es 1200 (desde $H_0 : \mu = 1200$)
- 1200 se incluye en el intervalo de confianza del 95%, ya que va de 1075,33 a 1224,67 euros
- Entonces **no rechazamos** la hipótesis nula de que el salario medio de los trabajadores españoles sea de 1200 euros.

Por supuesto, la conclusión es equivalente a la que se había llegado por los otros dos métodos. Esto debe ser así, ya que usamos los mismos datos y el mismo nivel de significancia α para los tres métodos.

3.4.4 Dos tipos de errores

Antes de entrar en detalles sobre cómo se construye una prueba estadística, es útil comprender la filosofía detrás de ella. Idealmente, nos gustaría construir nuestra prueba de modo que nunca cometamos errores. Desafortunadamente, dado que el mundo es caótico, esto no es posible. A veces simplemente tienes mala suerte: por ejemplo, supongamos que lanzas una moneda 10 veces seguidas y sale cara las 10 veces. Eso se parece una evidencia muy fuerte de que la moneda está trucada, pero, por supuesto, hay una probabilidad de 1 entre 1024 de que esto suceda incluso si la moneda fuera completamente normal. En otras palabras, en la vida real *siempre* tenemos que aceptar que existe la posibilidad... Como consecuencia, el objetivo detrás de las pruebas de hipótesis estadísticas no es *eliminar* errores, sino *minimizarlos*.

Debemos ser un poco más precisos sobre lo que entendemos por “errores”. En primer lugar se da el caso de que la hipótesis nula es verdadera o es falsa; y nuestra prueba rechazará la hipótesis nula o la retendrá. [^][Un comentario aparte con respecto al lenguaje que usa para hablar sobre la prueba de hipótesis. En primer lugar, una cosa que realmente desea evitar es la palabra “probar”: una prueba estadística realmente no *prueba* que una hipótesis sea verdadera o falsa. La prueba implica certeza pero las estadísticas significan nunca tener que decir que estás seguro. Pueden pasar una de cuatro cosas:

	Mantener H_0	Mantener H_0
H_0 es verdadero	Decisión correcta	Error (Tipo I)
H_0 es falso	Error (tipo II)	Decisión correcta

Como consecuencia, en realidad hay *dos* tipos diferentes de error. Si rechazamos una hipótesis nula que sea realmente cierta, entonces hemos cometido un **error de tipo I**. Por otro lado, si mantenemos la hipótesis nula cuando en realidad es falsa, entonces hemos cometido un **error tipo II**. En juicio penal donde se busca establecer “más allá de toda duda razonable” que el acusado es culpable. Todas las reglas probatorias están (en teoría, al menos) diseñadas para garantizar que (casi) no haya posibilidad de condenar injustamente a un acusado inocente. El juicio está diseñado para proteger los derechos de un acusado: como dijo el famoso jurista inglés William Blackstone, es “mejor que escapen diez culpables que sufra un inocente”. En otras palabras, un juicio penal no trata los dos tipos de error de la misma manera ~ ... castigar al inocente se considera mucho peor que dejar en libertad al culpable. Una prueba estadística es prácticamente lo mismo: el principio de diseño más importante de la prueba es *controlar* la probabilidad de un error de tipo I, para mantenerla por debajo de una probabilidad fija. Esta probabilidad, que se denota α , se denomina **nivel de significancia** de la prueba (o, a veces, el *tamaño* de la prueba). Y lo diré de nuevo, porque es tan fundamental para toda la configuración ~ ... se dice que una prueba de hipótesis tiene un nivel de significancia α si la tasa de error de tipo I no es mayor que α .

Entonces, ¿qué pasa con la tasa de error de tipo II? Bueno, también nos gustaría mantenerlos bajo control, y denotamos esta probabilidad por β . Sin embargo, es mucho más común referirse a la **potencia** de la prueba, que es la probabilidad con la que rechazamos una hipótesis nula cuando realmente es falsa, que es $1 - \beta$. Para ayudar a mantener esto claro, aquí está la misma tabla nuevamente, pero con los números relevantes agregados:

	no descartar H_0	descartar H_0
H_0 es verdadera	$1 - \alpha$ (probabilidad de mantener)	α (tasa de error tipo I)
H_0 es falsa	β (tasa de error tipo II)	$1 - \beta$ (potencia del test)

Una prueba de hipótesis “potente” es aquella que tiene un valor pequeño de β , mientras mantiene α fijo en algún (pequeño) nivel deseado. Por convención, los científicos utilizan tres niveles α diferentes: .05, .01 y .001. Observad la asimetría aquí ~ ... las pruebas están diseñadas para *asegurar* que el nivel α se mantenga pequeño, pero no hay garantía correspondiente con respecto a β . Ciertamente nos gustaría que la tasa de error de tipo II fuera pequeña, y tratamos de diseñar pruebas que la mantengan pequeña, pero esto es muy secundario a la abrumadora necesidad de controlar la tasa de error de tipo I.

3.5 Test de hipótesis en R: cálculo e informes

3.6 Un ejemplo sobre brecha salarial entre géneros

3.7 ANOVA

3.8 Algunas consideraciones finales

3.8.1 ¿Cuándo no se necesita inferencia?

Hemos analizado varios ejemplos sobre cómo realizar inferencias estadísticas: realización de test de hipótesis y construcción de intervalos de confianza. Antes de empezar a realizar un experimento, siempre es necesario realizar un análisis exploratorio de los datos. Este primer vistazo siempre puede ayudar a intuir sobre lo que los métodos estadísticos como los intervalos de confianza y las pruebas de hipótesis pueden decirnos (y lo que no pueden). En los apartados anteriores hemos querido explicar cómo funciona la inferencia pero no nos hemos preguntado si era realmente necesaria.

Consideremos un ejemplo. Supongamos que estamos interesados en la siguiente pregunta: De *todos* los vuelos que salen de un aeropuerto de la ciudad de Nueva York, ¿los vuelos de Hawaiian Airlines están en el aire por más tiempo que los vuelos de Alaska Airlines? Además, supongamos que los vuelos de 2013 son una muestra representativa de todos esos vuelos. Entonces podemos usar el dataframe `flights` disponible en el paquete `nycflights13` para responder nuestra pregunta. Filtremos este dataframe para incluir solo a Hawaiian y Alaska Airlines usando sus códigos de “operador” “HA” y “AS”:

```
library(tidyverse)
library(nycflights13)
flights_sample <- flights %>%
  filter(carrier %in% c("HA", "AS"))
```

Hay dos posibles métodos de inferencia estadística que podríamos utilizar para responder a estas preguntas. Primero, podríamos construir un intervalo de confianza del 95% para la diferencia en las medias poblacionales $\mu_{HA} - \mu_{AS}$, donde μ_{HA} es el tiempo de vuelo medio de todos los vuelos de Hawaiian Airlines y μ_{AS} es el tiempo medio de vuelo de los vuelos de Alaska Airlines. Luego podríamos verificar si la totalidad del intervalo es mayor que 0, sugiriendo que $\mu_{HA} - \mu_{AS} > 0$, o, en otras palabras, sugiriendo que $\mu_{HA} > \mu_{AS}$. En segundo lugar, podríamos realizar una prueba de hipótesis de la hipótesis nula $H_0 : \mu_{HA} - \mu_{AS} = 0$ frente a la hipótesis alternativa $H_A : \mu_{HA} - \mu_{AS} > 0$.

Construyamos primero una visualización exploratoria como acabamos de sugerir. Dado que `air_time` es numérico y `carrier` es categórico, un diagrama de caja (*boxplot*) puede mostrar la relación entre estas dos variables (ver la Figura 3.11).

```
ggplot(data = flights_sample, mapping = aes(x = carrier, y = air_time)) +
  geom_boxplot() +
  labs(x = "Carrier", y = "Air Time")
```

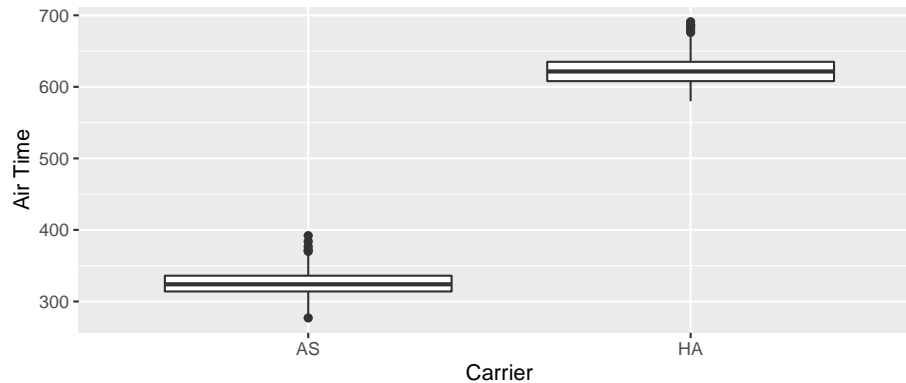


Figura 3.11: Air time for Hawaiian and Alaska Airlines flights departing NYC in 2013.

No es necesario ser un experto en estadísticas para saber que Alaska Airlines y Hawaiian Airlines tienen horarios aéreos *significativamente* diferentes. ¡Los dos diagramas de caja ni siquiera se superponen! La construcción de un intervalo de confianza o la realización de una prueba de hipótesis, francamente, no proporcionaría mucha más información que la Figura 3.11.

Investiguemos por qué observamos una diferencia tan clara entre estas dos aerolíneas que utilizan la manipulación de datos. Primero agrupemos por las filas de `vuelos_muestra` no solo por `transportista` sino también por destino `dest`. Posteriormente, calcularemos dos estadísticas resumidas: el número de observaciones usando `n()` y el tiempo medio de transmisión:

```
flights_sample %>%
  group_by(carrier, dest) %>%
  summarize(n = n(), mean_time = mean(air_time, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
## # Groups:   carrier [2]
##   carrier dest      n mean_time
##   <chr>   <chr> <int>     <dbl>
## 1 AS     SEA     714     326.
## 2 HA     HNL     342     623.
```

Resulta que desde la ciudad de Nueva York en 2013, Alaska solo voló a “SEA” (Seattle) desde la ciudad de Nueva York (NYC) mientras que Hawaiian solo voló a “HNL” (Honolulu) desde Nueva York. Dada la clara diferencia en la

distancia entre la ciudad de Nueva York y Seattle y la ciudad de Nueva York a Honolulu, no es sorprendente que observemos tiempos de vuelo tan diferentes (__ estadísticamente significativamente diferentes__, de hecho) en los vuelos.

Este es un claro ejemplo de que no es necesario hacer nada más que un simple análisis exploratorio de datos utilizando visualización de datos y estadísticas descriptivas para llegar a una conclusión adecuada. Por lo tanto, es recomendable empezar siempre por realizar un análisis exploratorio con estadísticas descriptivas antes de aplicar inferencia estadística.

3.8.2 Problemas con los p -valores

Además de los muchos malentendidos comunes sobre las pruebas de hipótesis y los valores de p que hemos comentado al explicar la interpretación de las pruebas de hipótesis, otra consecuencia desafortunada del uso ampliado de los valores de p y las pruebas de hipótesis es un fenómeno conocido como “ p -hacking”, que es el acto de “seleccionar” sólo los resultados que son “estadísticamente significativos” y descartar los que no lo son, aunque sea a expensas de las ideas científicas. Hay muchos artículos escritos recientemente sobre malentendidos y problemas con los valores de p . Le recomendamos que consulte algunos de ellos:

1. Malentendidos de los valores de p
2. Qué debate más nerd sobre los valores de p sobre la ciencia y cómo solucionarlo
3. Los estadísticos emiten una advertencia sobre el uso indebido de los valores de p
4. No puede confiar en lo que lee sobre nutrición
5. Una letanía de problemas con valores p

Tales problemas se estaban volviendo tan recurrentes que la Asociación Estadounidense de Estadística (ASA) emitió una declaración en 2016 titulada, “Declaración de la ASA sobre la importancia estadística y los valores de p ” con seis principios subyacentes al uso e interpretación adecuados de los valores de p . La ASA publicó esta guía sobre los valores de p para mejorar la conducta y la interpretación de la ciencia cuantitativa y para informar el creciente énfasis en la reproducibilidad de la investigación científica.

Quizás el uso de intervalos de confianza para la inferencia estadística permita evitar ciertos malentendidos. Sin embargo, en muchos campos todavía se usan exclusivamente valores de p para la inferencia estadística y esta es una razón para incluirlos en este texto.

3.9 Ejercicios

3.9.1 Distribución Binomial

En nuestro municipio hay 500 hombres de la misma edad y con buena salud. Según las estadísticas actuales, la probabilidad de que estas personas vivan 30 años o más es de $2/3$. Hay que hallar las probabilidades de que dentro de esos 30 años vivan

1. Los 500 hombres.
2. Al menos 300 de ellos.
3. 200 hombres

3.9.2 Distribución Normal

El ayuntamiento consume una media de 80 kWh/hab/a , con una desviación estándar de 14 kWh/hab/a . Hay que calcular:

1. La probabilidad de que este año hagan falta entre 75 y 90 ($p(75 \leq x \leq 90)$).
2. La probabilidad de que hagan falta 75 o menos ($p(75 \leq x)$).
3. Describid en un párrafo alguna variable municipal que se pueda ajustar a una distribución normal y haced la comprobación.

Capítulo 4

El muestreo estadístico

Some *significant* applications are demonstrated in this chapter.

4.1 Example one

4.2 Example two

Capítulo 5

Regresión lineal

El objetivo de este capítulo es presentar la *regresión lineal*, la herramienta estándar en la que confían los estadísticos al analizar la relación entre los predictores y los resultados. Los modelos de regresión lineal son básicamente una versión un poco más elegante de la correlación de Pearson (Sección @ref(correlation)) aunque, como veremos, los modelos de regresión son herramientas mucho más poderosas.

5.1 ¿Qué es un modelo de regresión lineal?

El diagrama de dispersión real que dibujamos es el que se muestra en la Figura @ref(fig: regression0), y como vimos anteriormente, esto corresponde a una correlación de $r = -.90$, pero lo que nos encontramos imaginando en secreto es algo que se ve más cerca de la Figura ? (fig: regression1a). Es decir, dibujamos mentalmente una línea recta a través del medio de los datos. En estadística, esta línea que estamos dibujando se llama *línea de regresión*. Tened en cuenta que la línea de regresión pasa por el centro de los datos. No hay nada parecido en la gráfica que se muestra en la Figura @ref(fig: regression1b). La línea dibujada en la Figura 5.3 no “se ajusta” muy bien a los datos, por lo que no tiene mucho sentido proponerla como una forma de resumir los datos, ¿verdad? Esta es una observación muy simple de hacer, pero resulta ser muy poderosa.

La fórmula para una línea recta generalmente se escribe así:

$$y = mx + c$$

Las dos *variables* son x y y , y tenemos dos *coeficientes*, m y c . El coeficiente m representa la *pendiente* de la línea y el coeficiente c representa la intersección y de la línea. Profundizando, la intersección se interpreta como “el valor de y que obtienes cuando $x = 0$ ”. De manera similar, una pendiente de m significa que

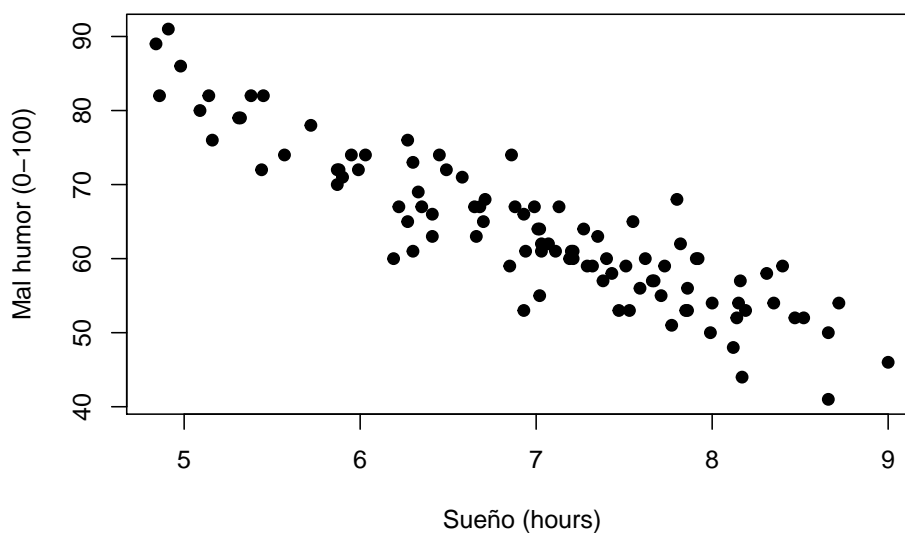


Figura 5.1: Diagrama de dispersión que muestra el mal humor en función de las horas de sueño.

Linea de regresión que mejor se ajusta

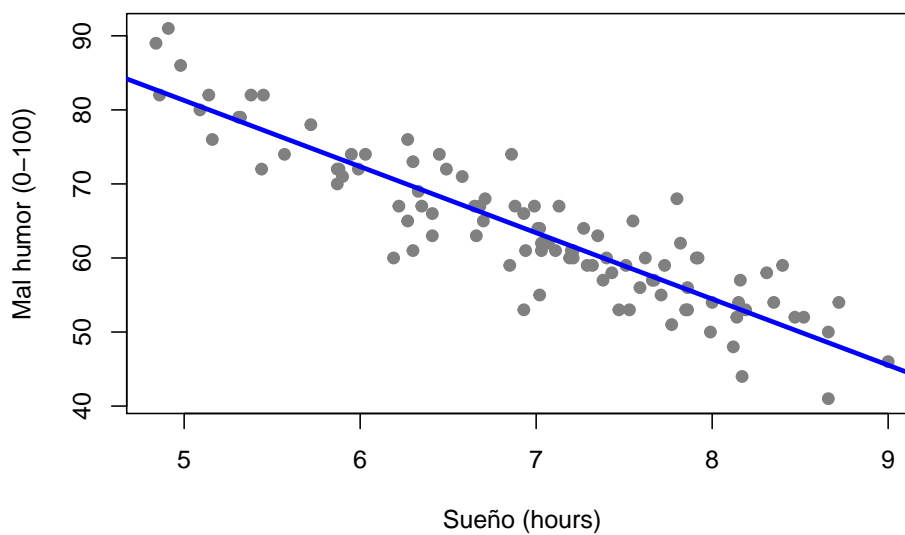


Figura 5.2: El panel a muestra el diagrama de dispersión del mal humor del sueño desde arriba con la línea de regresión que mejor se ajusta dibujada en la parte superior. No es sorprendente que la línea pase por el medio de los datos.

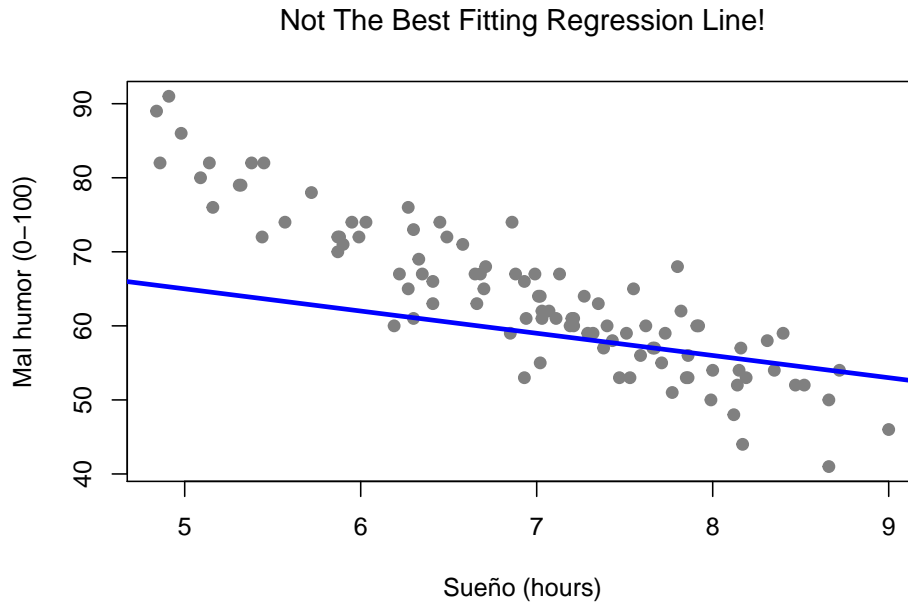


Figura 5.3: Por el contrario, este gráfico muestra los mismos datos, pero con una elección muy pobre de la línea de regresión dibujada en la parte superior.

si aumenta el valor x en 1 unidad, entonces el valor y aumenta en m unidades; una pendiente negativa significa que el valor y bajaría en lugar de subir.

Usamos exactamente la misma fórmula para describir una línea de regresión. Si Y es la variable de resultado (variable dependiente) y X es la variable predictora (variable independiente), entonces la fórmula que describe nuestra regresión se escribe así:

$$\hat{Y}_i = b_1 X_i + b_0$$

Parece la misma fórmula, pero hay algunos detalles extra en esta versión. En primer lugar, observad que tenemos X_i y Y_i en lugar de simplemente los viejos X y Y . Esto se debe a que queremos recordar que estamos tratando con datos reales. En esta ecuación, X_i es el valor de la variable predictora para la i ésima observación (es decir, el número de horas de sueño que obtuvo el día i de este pequeño estudio), y Y_i es el correspondiente valor de la variable de resultado (es decir, el mal humor de ese día). Asumimos que esta fórmula funciona para todas las observaciones en el conjunto de datos (es decir, para todos los i). En segundo lugar, observad se escribe \hat{Y}_i y no Y_i . Esto se debe a que queremos hacer la distinción entre los *datos reales* Y_i y la *estimación* \hat{Y}_i (es decir, la predicción que hace nuestra línea de regresión). En tercer lugar, las letras utilizadas para describir los coeficientes de m y c cambian a b_1 y b_0 . Esa es la forma en que a los estadísticos les gusta referirse a los coeficientes en un modelo de regresión.

En cualquier caso, b_0 siempre se refiere al término de intersección y b_1 se refiere a la pendiente.

Independientemente de si estamos hablando de una línea de regresión *buen*a o *mal*a, los datos no encajan perfectamente en la línea. O, para decirlo de otra manera, los datos Y_i no son idénticos a las predicciones del modelo de regresión \hat{Y}_i . Dado que a los estadísticos les encanta adjuntar letras, nombres y números a todo, hagamos referencia a la diferencia entre la predicción del modelo y ese punto de datos real como un *residual* o residuos, y lo denominaremos ϵ_i .¹ Escritos usando matemáticas, los residuos se definen como:

$$\epsilon_i = Y_i - \hat{Y}_i$$

lo que a su vez significa que podemos escribir el modelo de regresión lineal completo como:

$$Y_i = b_1 X_i + b_0 + \epsilon_i$$

5.2 Estimación de un modelo de regresión lineal

Bien, ahora redibujemos nuestras imágenes, pero esta vez agregaré algunas líneas para mostrar el tamaño del residual para todas las observaciones. Cuando la línea de regresión es buena, nuestros residuos (las longitudes de las líneas negras continuas) se ven bastante pequeños, como se muestra en la Figura 5.4, pero cuando la línea de regresión es mala, los residuos son un mucho más grandes, como puede ver al mirar la Figura 5.5. Lo que buscamos en un modelo de regresión son residuos *pequeños*. La línea de regresión de “mejor ajuste” es la que tiene los residuos más pequeños.

Los coeficientes de regresión estimados, \hat{b}_0 y \hat{b}_1 son los que minimizan la suma de los residuos al cuadrado, que podríamos escribir como $\sum_i (Y_i - \hat{Y}_i)^2$ o como $\sum_i \epsilon_i^2$.

Los coeficientes de regresión son *estimaciones* (estamos tratando de adivinar los parámetros que describen una población), razón por la cual hay unos pequeños sombreros sobre la fórmula, de modo que obtenemos \hat{b}_0 y \hat{b}_1 en lugar de b_0 y b_1 . Finalmente, dado que en realidad hay más de una forma de estimar un modelo de regresión, el nombre más técnico para este proceso de estimación es ***regresión de mínimos cuadrados ordinarios (MCO)***.

En este punto, ahora tenemos una definición concreta de lo que cuenta como nuestra “mejor” elección de coeficientes de regresión, \hat{b}_0 y \hat{b}_1 . Si los coeficientes de regresión óptimos son aquellos que minimizan los residuos de la suma al cuadrado, ¿cómo *encontramos* estos números? La respuesta real a esta pregunta

¹El símbolo ϵ es la letra griega epsilon. Es tradicional usar ϵ_i o e_i para denotar un residual.

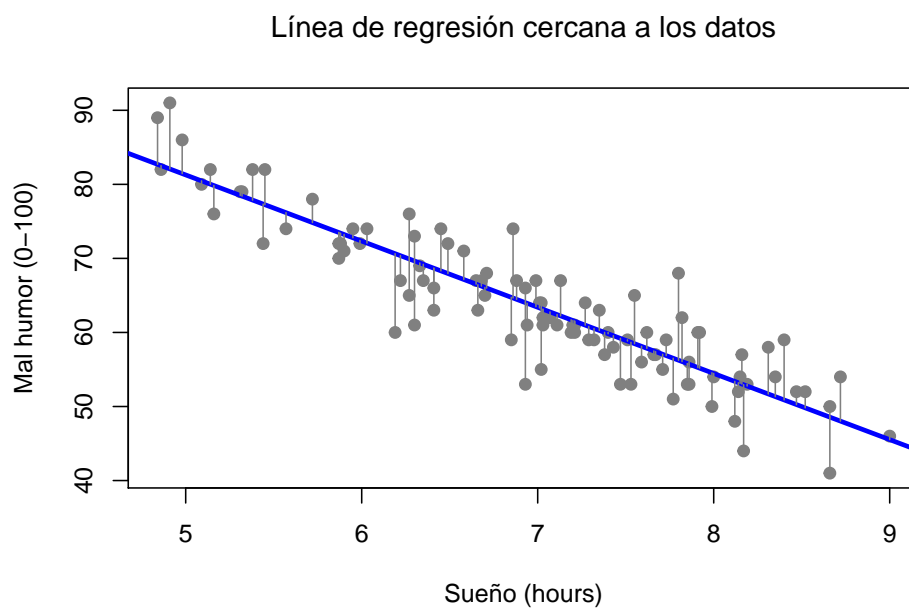


Figura 5.4: Una descripción de los residuos asociados con la línea de regresión que mejor se ajusta

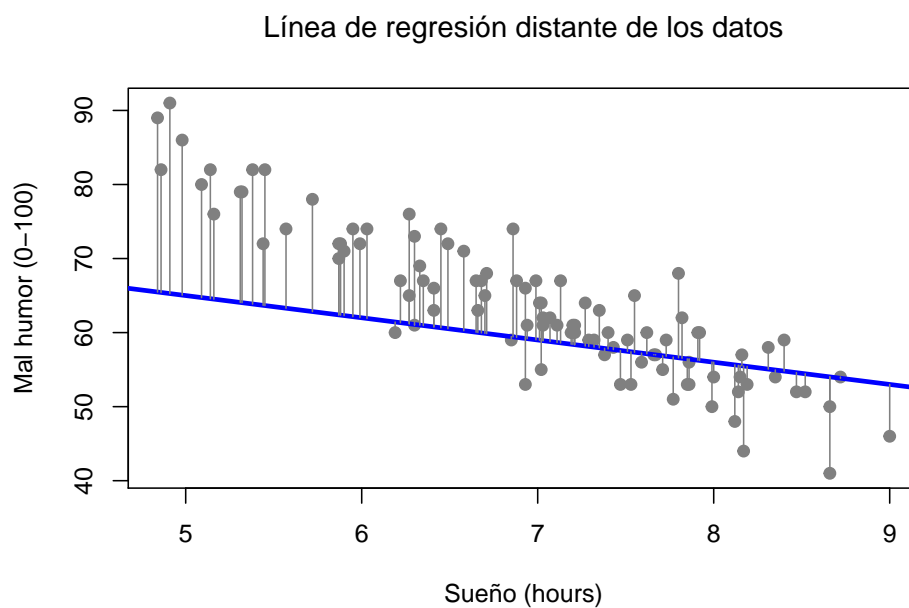


Figura 5.5: Los residuos asociados con una línea de regresión pobre

es complicada y no le ayuda a entender la lógica de la regresión.² R proporciona una manera más simple de calcular la regresión, utilizando la función `lm()` (abreviatura de “modelo lineal”).

5.2.1 Usando la función `lm()`

La función `lm()` es bastante complicada: si escribes `?lm`, los archivos de ayuda revelarán que hay muchos argumentos que puede especificar. Sin embargo, en este momento, solo hay dos de ellos que nos interesan, y resulta que los hemos visto antes:

- **fórmula.** Una fórmula que especifica el modelo de regresión. Para los modelos de regresión lineal simple de los que hemos hablado hasta ahora, en los que tiene una única variable predictora, así como un término de intersección, esta fórmula tiene la forma **resultado ~ predictor**. Sin embargo, se permiten fórmulas más complicadas y las discutiremos más adelante.
- **datos.** El dataframe que contiene las variables.

La salida de la función `lm()` es un objeto bastante complicado, con bastante información técnica. Debido a que otras funciones utilizan esta información técnica, generalmente es una buena idea crear una variable que almacene los resultados de una regresión. Con esto en mente, para ejecutar la regresión lineal, hacemos:

```
regression.1 <- lm( formula = dan.grump ~ dan.sleep,
                    data = parenthood )
```

La fórmula utilizada es `dan.grump ~ dan.sleep`: en el modelo que estoy tratando de estimar, `dan.grump` es la variable *resultado* y `dan.sleep` es la variable predictora. De todos modos, lo que esto hace es crear un “objeto `lm`” (es decir, una variable cuya clase es “`lm`”) llamada `regression.1`. Echemos un vistazo a lo que sucede cuando lo `print()`:

```
print( regression.1 )

##
## Call:
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)      dan.sleep
##      125.956         -8.937
```

²O al menos, asumo que no ayuda a la mayoría de la gente. Pero en la remota posibilidad de que alguien que lea esto, *ayudará* a *saber eso la solución al problema de estimación resulta ser $\hat{b} = (X^T X)^{-1} X^T y$, donde \hat{b} es un vector que contiene los coeficientes de regresión estimados, X es la “matriz de diseño” que contiene las variables predictoras (más una columna adicional que contiene todas unas; estrictamente X es una matriz de los regresores, pero aún no he discutido la distinción), y y es un vector que contiene la variable de resultado.

Hay dos piezas de información separadas. En primer lugar, R nos recuerda cuál fue el comando que usamos para especificar el modelo, lo que puede ser útil. Más importante desde nuestra perspectiva, sin embargo, es la segunda parte, en la que R nos da la intersección $\hat{b}_0 = 125.96$ y la pendiente $\hat{b}_1 = -8.94$. En otras palabras, la línea de regresión que mejor se ajusta (ver la Figura 5.2 tiene esta fórmula:

$$\hat{Y}_i = -8.94 X_i + 125.96$$

5.2.2 Interpretación del modelo estimado

Lo más importante para poder entender es cómo interpretar estos coeficientes. Comencemos con \hat{b}_1 , la pendiente. Si recordamos la definición de la pendiente, un coeficiente de regresión de $\hat{b}_1 = -8.94$ significa que si aumento X_i en 1, entonces estoy disminuyendo Y_i en 8.94. Es decir, cada hora adicional de sueño que gano mejorará mi estado de ánimo, reduciendo mi mal humor en 8,94 puntos de mal humor. ¿Qué pasa con la intersección? Bueno, dado que \hat{b}_0 corresponde al “valor esperado de Y_i cuando X_i es igual a 0”, es bastante sencillo. Implica que si dormimos cero horas ($X_i = 0$) entonces mi mal humor se saldrá de la escala, a un valor insano de ($Y_i = 125.96$).

5.3 Regresión lineal múltiple

El modelo de regresión lineal simple que hemos discutido hasta este punto supone que hay una única variable predictora que le interesa, en este caso, `dan.sleep`. De hecho, hasta este punto, *todas* las herramientas estadísticas de las que hemos hablado han asumido que su análisis utiliza una variable predictora y una variable de resultado. Sin embargo, en muchos (quizás la mayoría) de los proyectos de investigación, en realidad tiene varios predictores que desea examinar. Si es así, sería bueno poder extender el marco de regresión lineal para poder incluir múltiples predictores.

La regresión múltiple es conceptualmente muy simple. Todo lo que hacemos es agregar más términos a nuestra ecuación de regresión. Supongamos que tenemos dos variables que nos interesan; tal vez queramos usar tanto `dan.sleep` como `baby.sleep` para predecir la variable `dan.grump`. Como antes, dejamos que Y_i se refiera a mi mal humor el día i -th. Pero ahora tenemos dos variables X : la primera corresponde a la cantidad de horas que dormí y la segunda que corresponde a la cantidad de horas que durmió un bebé. Así que dejaremos que X_{i1} se refiera a las horas que durmió Dan el día i -ésimo, y X_{i2} se refiere a las horas que durmió su bebé ese día. Si es así, entonces podemos escribir nuestro modelo de regresión así:

$$Y_i = b_2 X_{i2} + b_1 X_{i1} + b_0 + \epsilon_i$$

Como antes, ϵ_i es el residuo asociado con la i -ésima observación, $\epsilon_i = Y_i - \hat{Y}_i$. En este modelo, ahora tenemos tres coeficientes que deben estimarse: b_0 es la intersección, b_1 es el coeficiente asociado con el sueño de Dan y b_2 es el coeficiente asociado con el sueño de su hijo. Sin embargo, aunque ha cambiado el número de coeficientes que deben estimarse, la idea básica de cómo funciona la estimación no ha cambiado: los coeficientes estimados \hat{b}_0 , \hat{b}_1 y \hat{b}_2 son los que minimizan la suma de los residuos al cuadrado.

5.3.1 Haciéndolo en R

La regresión múltiple en R no es diferente de la regresión simple: todo lo que tenemos que hacer es especificar una “fórmula” más complicada cuando usamos la función “`lm()`”. Por ejemplo, si queremos usar tanto `dan.sleep` como `baby.sleep` como predictores en nuestro intento de explicar por qué estoy tan malhumorado, entonces la fórmula que necesitamos es la siguiente:

```
dan.grump ~ dan.sleep + baby.sleep
```

Observe que, al igual que la última vez, no he incluido explícitamente ninguna referencia al término de intersección en esta fórmula; sólo las dos variables predictoras y el resultado. Por defecto, la función `lm()` asume que el modelo debe incluir una intersección (aunque puede deshacerse de ella si lo desea). En cualquier caso, puedo crear un nuevo modelo de regresión, al que llamaré `regression.2`, usando el siguiente comando:

```
regression.2 <- lm( formula = dan.grump ~ dan.sleep + baby.sleep,
                    data = parenthood )
```

Y al igual que la última vez, si `print()` este modelo de regresión, podemos ver cuáles son los coeficientes de regresión estimados:

```
print( regression.2 )

##
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)    dan.sleep    baby.sleep
##   125.96557      -8.95025       0.01052
```

El coeficiente asociado con `dan.sleep` es bastante grande, lo que sugiere que cada hora de sueño que pierde le pone mucho más gruñón. Sin embargo, el coeficiente de `baby.sleep` es muy pequeño, lo que sugiere que realmente no importa cuánto duerma su hijo. Para tener una idea de cómo se ve este modelo de regresión múltiple, la Figura @ref(fig:regresión múltiple) muestra un gráfico 3D que traza las tres variables, junto con el modelo de regresión en sí.

Capítulo 6

Estadística multivariante

We have finished a nice book.

Bibliografía

- Knuth, D. E. (1984). Literate Programming. *The Computer Journal*, 27(2):97–111.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.
- Wickham, H. and Golemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O’Reilly Media, Inc., 1st edition.
- Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.