Métodos Cuantitativos

Materiales de cátedra y de consulta

Aleksander Dietrichson, PhD

01/15/2023

Indice de contenidos

Prefacio **3**
Segunda edición . . . 3
Estructura del libro . . . 3
Glosario . . . 3
R y Rstudio . . . 4
Ejemplos en R . . . 4
Edición . . . 4
Agradecimientos . . . 5

1 Conceptos fundamentales **6**
1.1 Poblaciones y muestras . . . 6
1.1.1 Muestra aleatoria . . . 7
1.1.2 Muestra cuasi-aleatoria . . . 8
1.2 Ejemplo en R: Generar una secuencia . . . 8
1.2.1 Muestra estratificada . . . 9
1.3 Representatividad . . . 9
1.4 Estadísticas descriptivas e inferenciales . . . 11
1.5 Variables y su clasificación . . . 13
Por su relación con otras variables . . . 13
Por su nivel de medición . . . 13
Por su precisión . . . 14
1.6 Glosario . . . 15

2 Distribuciones de frecuencias **16**
2.1 Explorando los datos . . . 16
2.2 Tablas de frecuencias . . . 16
2.3 Tabla de frecuencias en R . . . 17
2.4 Histogramas . . . 18
2.5 Polígono de frecuencias . . . 23
2.6 Otro ejemplo . . . 24
2.7 Perfil de la distribución . . . 24
2.7.1 Asimetría o Sesgo . . . 25
2.8 Glosario . . . 27

3 Centralización y dispersión **28**
3.1 Centralización . . . 28
3.1.1 La media . . . 28
3.1.2 La mediana . . . 30
3.1.3 La moda . . . 31
3.1.4 ¿Cuál usar? . . . 31
3.2 Medidas de dispersión . . . 33
3.2.1 Rango o amplitud . . . 34
3.2.2 El rango intercuartílico . . . 34
3.2.3 La varianza y desviación estándar . . . 35
3.2.4 Visualizar la dispersión . . . 37
3.3 Glosario . . . 38

4 La distribución normal **40**
4.1 Importancia de la distribución normal . . . 40
4.2 Propiedades de la curva normal . . . 41
Variables normalizadas . . . 43
4.3 Evaluar la normalidad . . . 47
4.4 Glosario . . . 49

5 Estimación de parámetros **50**
5.1 Distribución muestral . . . 50
5.2 El error estándar y su interpretación . . . 51
5.2.1 Intervalos de confianza . . . 52
5.3 La distribución t . . . 54
¿Dónde obtenemos los valores críticos de t? . . . 55
5.4 Glosario . . . 55

6 Diseño de proyectos y test de hipotesis **56**
6.1 El método científico . . . 56
6.2 El diseño de una investigación . . . 57
Estudios experimentales y observacionales . . . 57
Fuentes de ruido en los datos . . . 58
6.3 Tests de hipótesis . . . 58
6.3.1 Tests estadísticos de significanza . . . 58
6.3.2 La hipótesis nula y alternativa . . . 59
6.3.3 Niveles de significanza . . . 60
6.3.4 Tipos de error . . . 60
6.3.5 Tests direcionales y no direcionales . . . 60
6.3.6 ¿Qué test usar? . . . 62
6.3.7 Procedimiento . . . 63
6.4 Glosario . . . 63

7 Pruebas paramétricas **64**
7.1 Prueba t de Student para muestras independientes . . . 64
7.2 Prueba de Shapiro-Wilks . . . 67
7.3 Prueba de Fisher . . . 69
7.4 Prueba t para muestras pareadas . . . 70
7.5 Prueba de z . . . 73
7.6 Resumen de procedimiento . . . 73
7.7 Glosario . . . 74

8 Pruebas no paramétricas **75**
8.1 Prueba U de Mann-Whitney . . . 75
8.2 Prueba de los rangos con signo de Wilcoxon . . . 76
¿Y si usabamos la prueba t igual? . . . 79
8.3 Prueba de signos . . . 79
8.4 Realizar prueba de sign para N>25 . . . 80
8.5 ¿Cuál usar? . . . 80
8.6 Glosario . . . 81

9 Prueba de χ^2 **82**
9.1 Características . . . 83
9.2 Prueba de independencia o asociación . . . 84
9.2.1 Grados de libertad . . . 85
9.3 Prueba de χ^2 en R . . . 86
9.4 Glosario . . . 87

10 Correlación **88**
10.1 Visualización . . . 88
10.2 Generar diagrama de dispersión en R . . . 92
10.3 Coeficientes de correlación . . . 94
10.3.1 Coeficiente Pearson . . . 94
10.4 Coeficiente de correlación Pearson . . . 96
10.4.1 Coeficiente Spearman . . . 98
10.4.2 Coeficiente ϕ . . . 99
10.5 Interpretación de correlacciones . . . 100
10.6 Glosario . . . 100

Referencias **101**

Apéndices **101**

A Distribución t **102**

B Valores críticos del test de signo **104**

C Distribución χ^2 **106**

Prefacio

Este texto ha sido editado en respuesta a la aparente falta de un libro de texto introductorio al análisis cuantitativo y estadísticas accesible y moderno en castellano. Si bien fue concebido como material de cátedra para *Metodologías cuantitativas* materia que dicta el autor en la Escuela de Humanidades de la Universidad Nacional San Martín, se adaptará fácilmente a cursos introductorios de estadísticas en general.

Segunda edición

En la segunda edición se corrigió algunos errores ortográficos y de estilo. Optamos por actualizar los ejemplos para incorporar los paquetes del «tidyverse» ya que hemos observado que su uso y adaptación atenúa la curva de aprendizaje para quienes usan R por primera vez o con escasos conocimientos previos.

Estructura del libro

Cada capítulo desarrolla un tema y/o concepto a ser tratado en clase y la secuencia corresponde a un curso introductorio de estadísticos «clásico», por lo que conviene leerlos en orden. Sigue el orden propuesto por Butler (1985).

Glosario

Uno de los objetivos de este trabajo es dotar al lector con las herramientas necesarios para convertirse en un consumidor crítico de textos que se valen de métodos cuantitativos y/o estadísticas para su argumento. En vista de la enorme cantidad de material disponible en inglés, sobre todo en el ámbito académico, el autor ha optado por incluir terminología bilingüe español-inglés. Esta elección obedece a un criterio práctico. En cada capítulo encontrarán un glosario con los principales términos mencionados. Incluye traducción a inglés y referencias a R cuando sea relevante.

R y Rstudio

R es un lenguaje de programación especializado para análisis de datos. Es de fuente abierta (Open Source) y uso gratuito. *Rstudio* es un editor de *R* que también de uso sin cargo. Ambas herramientas están disponibles en Internet y son de amplio uso tanto en el mundo académico como la industria.

Se puede descargar e instalar R accediendo a esta URL: https://cran.r-project.org/mirrors.html.

Para Rstudio la URL es: https://www.rstudio.com/products/rstudio/download/#download.

Se recomienda siempre instalar R primero y luego Rstudio ya que este depende de aquel.

Ejemplos en R

A lo largo de este libro encontrarán ejemplos prácticos que pueden ejecutarse en R. El código se diferenciará del resto del texto por su formato, como se puede apreciar en el ejemplo siguiente:

```
1+1
```

```
[1] 2
```

Por convención no se incluye el promt (p.e. ">") de la consola de R, y los valores de retorno son comentados con "##", lo que corresponde al estándar para textos técnicos de esta índole. También se puede hacer referencia a código dentro del texto corrido con el mismo formato. Por ejemplo: `1+1`.

Edición

Este texto fue editado con *bookdown* (Xie 2018), un paquete de R (Xie 2018) que extiende las capacidades de *knitr*(Xie 2015) y R-markdown (Allaire et al. 2019) para publicaciones más voluminosas. También hace uso de los paquetes *tidyverse* (Wickham et al. 2019) y *bayestestR*(Makowski, Ben-Shachar, y Lüdecke 2019).

Agradecimientos

Agradezco a mi colega Diego Forteza por su ayuda y apoyo en durante el proceso de redacción y a Cecilia Magadán por su corrección de estilo.

Debo expresar también profunda gratitud a *Bow Street Destillery* en Dublin, Irlanda; sin cuyos productos este proyecto habría sin duda quedado inconcluso.

1 Conceptos fundamentales

En este capítulo introducimos algunos conceptos fundamentales del análisis cuantitativo y de las estadísticas. Consideramos los conceptos de población y muestra. Hacemos una brevísima introducción a la teoría de la probabilidad. Diferenciamos entre algunos de los usos importantes de la estadística: descriptiva e inferencial. Finalmente consideramos algunas maneras de clasificar variables.

1.1 Poblaciones y muestras

En su uso diario usamos *población* para designar un grupo de personas, por ejemplo *la población del Gran Buenos Aires*; o por lo menos de seres vivos como por ejemplo *la población de ratas* de la CABA. En estadísticas, en cambio, se usa el término de manera más general para significar cualquier recolección o conjunto de elementos, artículos o sujetos que gozan de características comunes con el fin de estudiarlos y de esta forma se sacar conclusiones específicas para determinar sus resultados. Así podemos hablar de la población de sustantivos en las obras de Jorge Luis Borges o de la población de notas asignadas en los cursos a nivel universitario.

Podemos distinguir entre poblaciones *finitas* e *infinitas*. La población de motocicletas vendidas en Buenos Aires en septiembre es finita. En cambio la población de temperaturas medidas en el Campus de San Martín es *infinita*, ya que, por lo menos teóricamente, podemos seguir midiendo para siempre.

Cuando una población finita no es demasiado grande podemos investigar la totalidad de ella. Pero, si la población es muy grande o potencialmente infinita tenemos que estar contentos con *muestras* extraídas de esta población. Por ejemplo: si queremos saber quién va a ganar las próximas elecciones podríamos preguntar a todo aquel que tiene derecho al voto cómo piensa votar para sacar el resultado. En la práctica esta metodología resultaría demasiado costosa, por lo que hacemos una muestra representativa de votantes, les preguntamos y generalizamos.

Resulta evidente que hay que tener cuidado al seleccionar una muestra para análisis. Los métodos estadísticos, los que nos permiten generalizar e inferir, suponen que las muestras están tomadas de manera *aleatoria* o al azar. Esto no significa que la muestra sea arbitraria, sino que cualquier unidad de la población que estamos estudiando tiene la misma probabilidad de ser seleccionada para hacer parte de la muestra.

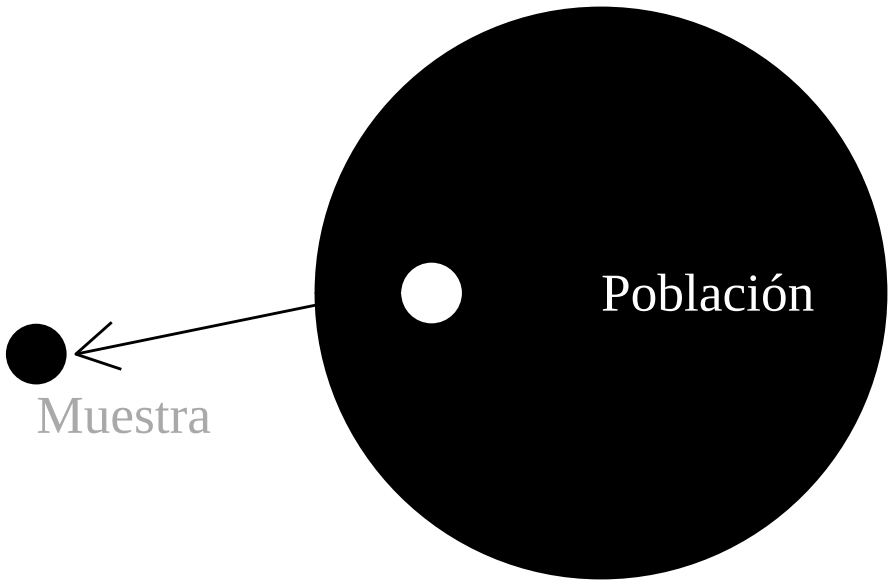


Figura 1.1: Población y muestra.

1.1.1 Muestra aleatoria

Para tener una muestra verdaderamente aleatoria de una población deberíamos asignar un número u otro identificador único a cada una de las unidades de la población –a cada persona si se trata de una población humana– escribir cada número en un papel y echarlos en una tómbola. Luego de virarla por algún tiempo y mesclar bien los papeles, podríamos de allí sacar la cantidad de papeles que corresponda al tamaño de nuestra muestra. Obviamente esto no resulta muy práctico por lo que se suele empezar con una secuencia de números aleatorios del tamaño de la muestra y extraer unidades de la población basado en ello. Por ejemplo, si quisiéramos sacar veinte libros al azar de un estante de la biblioteca que contiene doscientos libros, necesitamos veinte números aleatorios entre uno y doscientos, y sacamos los libros que desde algún punto de referencia (primer libro del primer nivel) está a esa distancia.

Ahora, ¿dónde encontramos números aleatorios? Hay secuencias en libros de estadísticas, usados principalmente antes de la existencia de computadoras. También se pueden generar esas secuencias en linea. Finalmente, R tienen un generador de números aleatorios que nos permite generar los de números de nuestra muestra con un solo comando usando la función de R *sample*.

Ejemplo 1.1 (Generar una muestra en R).

```
Ejemplo 1.1 (Generar una muestra en R).
```

```
## [1] 166  46  42 179 188 143 126 135 102  93  72 193  13 107 198 100  88  67  33  99
```

Acá le estamos pidiendo a R que nos de una muestra aleatoria (`sample`) de números entre uno y doscientos (`x = 1:200`), y que la muestra sea de veinte `size = 20`). Con estos números podemos ir al estante y sacar los libros que queremos estudiar.

Si corren este comando desde su consola de R los números deben salir diferentes, se hace una muestra aleatoria cada vez.

Ejemplo 1.2 (Ordenar los datos)**.** También es posible ordenar los números, lo cual nos ahorra un poco de tiempo al retirar los libros. Se logra con la función *sort*.

```
sort(
  sample(x = 1:200, size = 20, replace = TRUE)
)
## [1]  29  35  38  41  54  74  75  79  85  92 103 112 114 120 127 153 173 185 187 188
```

1.1.2 Muestra cuasi-aleatoria

Otra estrategia que podría emplearse para sacar veinte libros al azar del estante que describimos en la sección anterior sería decidir que vamos a sacar cada diez libros ya que $\frac{200}{20} = 10$. Este tipo de muestra lleva el epíteto cuasi-aleatoria, y funciona bien si el orden original de la población es aleatorio. Sin embargo, hay que tener en cuenta que esta estrategia puede generar una muestra no representativa si existe una estructura en ese orden. Típicamente puede resultar problemática si existe *periodicidad* en la población que estamos analizando. Si, por ejemplo, queremos tener una muestra de cuantos ómnibus pasan delante de mi casa por día sería mala idea decir que vamos a contarlos cada siete días. Si el día que empezamos es un domingo obtendremos seguramente una muestra con cantidades inferiores a la población real (en este caso definida como todos los ómnibus que pasan por mi casa en un día); y si empezamos a contar un lunes las cantidades serían superiores.

1.2 Ejemplo en R: Generar una secuencia

Si bien sacar la secuencia para sacar cada diez libros resulta trivial, existe la manera que hacerlo también con una función de R.

```r
seq( from = 10, to = 200 , by=10 )
## [1]  10  20  30  40  50  60  70  80  90 100 110 120 130 140 150 160 170 180 190 200
```

La función `seq` (de secuencia), toma tres parámetros, desde dónde empezamos (`from=10`), hasta dónde queremos llegar (`to=200`), y con qué distancia (`by=10`).

Por lo pronto se vuelve más útil si estamos trabajando con números menos redondos. Digamos que queremos sacar cada siete libros de un estante que contiene cien empezando por el número seis.

```r
seq( from = 6, to = 100 , by = 7 )
## [1]  6 13 20 27 34 41 48 55 62 69 76 83 90 97
```

1.2.1 Muestra estratificada

Cuando conocemos algunos parámetros de la población que queremos estudiar también nos podemos asegurar que nuestra muestra tenga parámetros similares. Esta estrategia puede resultar particularmente útil si suponemos que este parámetro puede tener alguna influencia en otra variable cuya distribución queremos conocer. Si por ejemplo suponemos que el *sexo* puede influir en la opinión de una persona sobre la ley del aborto podemos asegurarnos de que nuestra muestra tiene una distribución similar a la de la población en general. Se sabe que hay más o menos mitad y mitad[1] en la población general por lo que convendría que nuestra muestra tenga la misma distribución. Así podemos sacar, para una muestra de veinte, diez hombres y diez mujeres al azar[2]. Lo mismo se puede aplicar a otras variables, por ejemplo, clase social, país de origen etcétera.

1.3 Representatividad

Es importante entender que ninguna de las estrategias descritas en la sección anterior nos garantiza que la muestra que sacamos sea representativa de la población, con lo cual no está garantizado que una generalización basada en esa muestra sea válida. Lo que sí se puede calcular es la *probabilidad* de que la muestra sea representativa. Es decir, podemos tener una estimación de *en qué medida* la muestra representa la población.

Para profundizar un poco este concepto vamos a hacer un breve desvío y desarrollar un poco de teoría de la probabilidad por medio de un ejemplo sumamente sencillo. Digamos que

[1] En realidad 51 y 49%

[2] En este ejemplo hemos usado *sexo* como la variable biológica es decir ausencia o presencia de un cronozoma Y. Si queremos en cambio usar *género* obviamente también podemos incluir más categorías que las clásicas masculino y femenino si lo consideramos conveniente.

queremos hacer una muestra aleatoria de la población en Argentina. Vamos a seleccionar al azar a tan solo tres personas para nuestra muestra. Ya que sabemos que hay la misma cantidad de hombres y mujeres la probabilidad de que el/la primero/a que elijamos sea hombre es 0,5[3], lo cual también es la probabilidad de que sea mujer. Ahora, cuando seleccionamos el/la segundo/a y tercero/a las probabilidad son las mismas en todos los casos. Las leyes de probabilidad indican que la probabilidad de que dos o más eventos independientes sucedan es el producto de sus probabilidades individuales. Entonces, cuál es la probabilidad de que los tres miembros de la muestra sean mujeres?

$$0,5 \times 0,5 \times 0,5 = 0,125$$

Resulta evidente que lo mismo sucede si queremos calcular la probabilidad de que todos sean hombres.

Ahora, bien ¿cuál sería la probabilidad de que sean dos mujeres y un hombre?

Hay tres maneras que esto pueda suceder:

Cuadro 1.1: Combinaciones posibles.{#tbl-compinaciones-posibles}

Primero/a	Segundo/a	Tercero/a
Masculino	Femenino	Femenino
Femenino	Masculino	Femenino
Femenino	Femenino	Masculino

Cada una de estas posibilidades tienen la misma probabilidad y como el orden en el que fueron elegidos no es relevante para la muestra, podemos sumar las probabilidades para obtener la probabilidad total:

$$(0,5 \times 0,5 \times 0,5) + (0,5 \times 0,5 \times 0,5) + (0,5 \times 0,5 \times 0,5) = 0,375$$

Lógicamente lo mismo ocurre con el caso de dos hombres y una mujer. Entonces tenemos cuatro posibilidades con distintas probabilidades:

Cuadro 1.2: Probabilidades de las combinaciones.

Muestra	Probabilidad
Tres mujeres	0,125
Dos mujeres + un hombre	0,375
Dos hombres + una mujer	0,375
Tres hombres	0,125

[3]En estadísticas las probabilidades suelen expresarse por decimales, es decir: 0,5; en lugar de porcentajes: 50%

Observamos que las probabilidades suman 1, lo cual es matemáticamente inevitable.

Está claro que una muestra de tan solo tres personas nunca puede ser representativa de la población, sin embargo vemos que la medida en que son poco representativas varía. Cualquiera de las muestras de 2+1 sería *más representativa* que las de un solo sexo, y vemos que también son probables.

Este ejemplo es extensible a muestras más grandes con cálculos similares. Se desarrollará en más detalle en capítulos posteriores, pero para tener un ejemplo un tanto más real imaginemos que hemos decidido realizar una muestra de diez personas de la misma población (que tiene un 50 y 50 de hombres y mujeres).

Cuadro 1.3: Probabilidades de las combinaciones de una muestra de diez.

Hombres	Mujeres	Probabilidad
0	10	0,001
1	9	0,010
2	8	0,044
3	7	0,117
4	6	0,205
5	5	0,246
6	4	0,205
7	3	0,117
8	2	0,044
9	1	0,010
10	0	0,001

Obtendríamos los resultados de la tabla Tabla 1.3) y observamos que hay aproximadamente un 0,9 de probabilidad (90%) de obtener una muestra no peor que 7-3. También no es de sorprenderse que mientras más grande sea la muestra más probable es que sea representativa[4].

1.4 Estadísticas descriptivas e inferenciales

Entre los varios usos de las estadísticas este texto tratará de dos de los más importantes. Uno es el descriptivo que consiste en describir cuantitativamente un conjunto de datos y eventualmente generalizar este análisis a una población. Otro es el de inferir propiedades y diferencias entre variables.

[4]Si se lleva este argumento al extremo: si el tamaño de la muestra fuera igual al tamaño de la población, la muestra sería perfectamente representativa.

Vamos a desarrollar estas distinciones por medio de un ejemplo[5]. Supongamos que hemos hecho dos muestras aleatorias de las notas del examen final de dos cursos de la materia *Métodos cuantitativos*, uno dictado exclusivamente como curso teórico y el otro como curso teórico-práctico.

Las notas son: Curso A (teórico-práctico):

> 15, 12, 11, 18, 15, 15,9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16,17, 15, 17, 13, 14, 13, 15, 17, 19, 17, 18, 16 y 14.

Y para el Curso B (teórico):

> 11, 16, 14, 18,6,8,9, 14, 12, 12, 10, 15, 12,9, 13, 16, 17, 12,8,7, 15,5, 14, 13, 13, 12, 11, 13, 11 y 7

El examen fue idéntico para ambos grupos y se podía obtener un máximo de veinte.

```
#| include: false

myData <- data.frame(
  classA = c(15, 12, 11, 18, 15, 15,9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16,17, 15, 17,
  classB = c(11, 16, 14, 18,6,8,9, 14, 12, 12, 10, 15, 12,9, 13, 16, 17, 12,8,7, 15,5, 14,
)
saveRDS(myData,file = here::here("data","comparative_two_courses.RDS"))
```

Antes de sacar conclusiones sobre estos datos deberíamos resumirlos. Podemos construir, por ejemplo, una tabla que muestra la frecuencia de cada nota en cada curso. Esto se llama tabla de frecuencias. También nos gustaría saber cuál es la nota más típica, la nota *promedio* y cuánto varían las notas respecto a éste. Estas son estadísticas descriptivas, y los desarrollaremos en los capítulos dos y tres de este texto.

Pero seguramente también quisiéramos saber con qué nivel de confianza podemos generalizar estos datos a similares grupos de datos usando métodos similares a los mencionados. Nos gustaría saber en qué medida las dos muestras que tenemos son representativas de sus respectivas poblaciones de estudiantes tomando cursos similares. Este tipo de estimaciones se verá en detalle en el capítulo cinco.

Además quisiéramos saber si podemos afirmar que alguno de los dos grupos estuvo mejor que el otro en el examen final. Podríamos postular, por ejemplo, que el grupo que recibió el curso teórico-práctico debería sacar mejores notas en promedio que el otro. Para ello hay que construir un test de la hipótesis y someter nuestros datos a este test.

Tanto la tarea de estimación como el test de hipótesis comprenden la inferencia de relaciones a partir de medidas descriptivas y juntos constituyen el área de *estadísticas inferenciales*.

[5] Adaptado de Butler (1985)

Finalmente, podríamos juntar más datos para determinar si existe en cualquiera de los dos cursos algún sub-grupo cuyas características se relacionan con un resultado específico. Con esta información estaríamos en condiciones de predecir las notas de los estudiantes en futuras cursadas de los cursos en cuestión.

1.5 Variables y su clasificación

En estadísticas trabajamos esencialmente con cantidades *variables*. En estadística definimos *variable* como: Una característica medida u observada al hacer un experimento u observación. Si, por ejemplo, estamos investigando el clima en Buenos Aires, podemos hacer medidas de temperatura, humedad, dirección e intensidad del viento etcétera.

Las variables pueden ser clasificadas de diferentes maneras:

Por su relación con otras variables

En la mayoría de investigaciones cuantitativas *variamos* una o más conjuntos de condiciones y medimos los efectos sobre una o más propiedades que son de nuestro interés. Las condiciones que cambiamos nosotros se denominan *variables independientes*[6] y los cuya respuesta a las condiciones cambiantes medimos se llaman *variables dependientes*.

Por su nivel de medición

Cuando hacemos una medición o observación o «recogemos un dato» debemos fijarnos en su *nivel de medición*, también llamado *escala* de medición. Distinguimos cuatro niveles o escalas:

Nivel nominal

Cuando un dato identifica una etiqueta (o el nombre de un atributo) de un elemento, se considera que la escala de medición es una escala nominal. En esta carecen de sentido el orden de las etiquetas, así como la comparación y las operaciones aritméticas. La única finalidad de este tipo de datos es clasificar a las observaciones. Ejemplo:

> Una variable que indica si el visitante de este post es «hombre» o «mujer».

En esta variable se tienen dos etiquetas para clasificar a los visitantes. El orden carece de sentido, así como la comparación u operaciones aritméticas.

[6] También se conocen como *predictores* o *variables experimentales*

Nivel ordinal

Cuando los datos muestran las propiedades de los datos nominales, pero además tiene sentido el orden (o jerarquía) de estos, se dice que se mide en escala ordinal. Ejemplo:

> Una variable que mide la calidad del café en la cafetería de la universidad. Le podemos asignar de uno a cinco estrellas.

En esta variable sigue sin tener sentido las operaciones aritméticas, pero ahora sí tiene sentido el orden. Cuatro estrellas es mejor que dos.

Nivel de intervalo

En una escala de intervalo, los datos tienen las propiedades de los datos ordinales, pero a su vez la separación entre las variables tiene sentido. Este tipo de datos siempre es numérico, y el valor cero no indica la ausencia de la propiedad. Por ejemplo: La temperatura (en grados centígrados) medida de una ciudad, puede ser cero sin que tenga sentido decir que «no hay temperatura».

En este nivel de medición, los número mayores corresponden a temperaturas mayores. Es decir, el orden importa, pero a la vez la diferencias entre las temperaturas importa. La diferencia entre 10 grados y veinte grados es igual que la diferencia entre 20 y 30. El nivel de medida de intervalo también se conoce como el nivel *intervalar*.

Nivel de razón

En una escala de razón –también llamado *de ratio* o *racional*, los datos tienen todas las propiedades de los datos de intervalo, y la proporción entre ellos tiene sentido. Para esto se requiere que el valor cero de la escala indique la ausencia de la propiedad a medir. Ejemplos de este tipo de variables son el peso de una persona a el tiempo utilizado para una tarea y el salario de una persona. Si una persona gana 100, y otra 10, la primera gana más que la segunda (comparación). También tiene sentido decir que la primera gana 90 más que la segunda (diferencia), o que gana 10 veces más (proporción).

Por su precisión

Cuando hablamos de *precisión* en matemáticas y estadísticas nos referimos al *numero de decimales* que tiene una variable. Esto es distinto de *exactitud* que significaría la medida en que la medición, o predicción corresponde a la realidad. 1,000 (uno coma cero cero cero), tiene más precisión que 1 (uno) si bien miden la misma cantidad. Esto lleva a la distinción que hacemos entre variables *discretas* y *continuas*. Las discretas por su naturaleza tienen precisión cero (no lleva decimales) y las continuas pueden tener la cantidad de decimales que queramos.

Para ilustrar la diferencia consideramos dos variables: *edad* y *numero de hijos*. En cuanto a la edad se puede tener diez años, diez años y medio o si queremos agregar más precisión: 20,45 años. En cambio *numero de hijos* es una variable discreta. Se puede tener cero, uno o más, pero no se puede tener 1,45 hijo.

Por su naturaleza vemos que las variables de escala nominal y ordinal son siempre discretas. Las de escala de intervalo y de escala de razón, en cambio pueden ser tanto discretas como continuas.

La mayoría de variables de interés en las ciencias duras se miden por escala de razón o de intervalo, mientras las escalas ordinal y nominal son más importantes en ciencias humanas. El nivel de medición de una variable es de suma importancia cuando decidimos qué medidas de tendencia central, variabilidad y dispersión elegimos para nuestro análisis, y qué test de hipótesis son adecuados. Es un error muy común entre investigadores, particularmente en las ciencias sociales, asumir una escala superior a lo teóricamente sostenible.

1.6 Glosario

Aleatorio/a Al azar. También son de uso frecuente los anglicismos «random» y «randómica». Función relevante en R: `rdunif`. Equivalente en inglés: «Random».

Histograma Visualización de frecuencia de observaciones de una variable. Función relevante en R: `hist`. Equivalente en inglés: «Histogram».

Muestra aleatoria Muestra en la que todos los elementos de la población tienen igual probabilidad de ser eligidos. Función relevante en R: `sample`. Equivalente en inglés: « Random sampling».

Muestra cuasialeatoria Muestra en la que cada n número de los los elementos de la población van a ser eligidos. Equivalente en inglés: «Cuasi-Random sampling».

Muestra estadística Subconjunto de una población estadística Función relevante en R: `sample`. Equivalente en inglés: «Sample».

Muestra estratificada Muestra que conserva las proporciones conocidas de la población estadística Equivalente en inglés: «Stratified sample».

Muestreo estadístíco El hecho de generar una muestra. Función relevante en R: `sample`. Equivalente en inglés: «Sampling».

Nivel de Medida Uno de los cuatro niveles jerárquicamente definidos: Nominal, ordinal, intervalar y racional Equivalente en inglés: «Level of measurement».

Parámetro estadístico Numero que resume una characteristica de la población estadística. Equivalente en inglés: «Parameter».

Población estadística El conjunto de individuos, objetos o fenómenos de los cuales se desea estudiar una o varias características. Equivalente en inglés: «Population».

Variable Characteristica que puede tener diferentes valores. Equivalente en inglés: «Variable».

2 Distribuciones de frecuencias

En este capítulo desarrollaremos el concepto de distribución estadística. Seguiremos desarrollando el ejemplo de notas de los exámenes finales de dos grupos de estudiantes e introduciremos otros ejemplos. Exploraremos el concepto de frecuencia de observaciones, cómo visualizarlos y estimar sus algunas de sus características.

2.1 Explorando los datos

Recordemos las muestras de exámenes finales que vimos en el capitulo anterior.

Grupo A (teórico-práctico):

> 15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17, 13, 14, 13, 15, 17, 19, 17, 18, 16 y 14

Grupo B (teórico):

> 11, 16, 14, 18, 6, 8, 9, 14, 12, 12, 10, 15, 12, 9, 13, 16, 17, 12, 8, 7, 15, 5, 14, 13, 13, 12, 11, 13, 11 y 7

A simple vista no es tan fácil darse cuenta «qué pasa» con estos datos. Podemos por lo pronto darnos cuenta de que el grupo B tiene más notas de un solo dígito, pero más allá no resulta obvio cómo les fue en los distintos grupos.

2.2 Tablas de frecuencias

Para darnos cuenta mejor de las estructuras que estamos analizando podemos construir una *tabla de frecuencias*, que en este caso es un resumen de cuántos alumnos sacaron cuál nota de las posibles (sobre veinte).

Cuadro 2.1: Frecuencia de notas por grupo

Nota	Grupo A	Grupo B
1	1	2
2	2	3

Nota	Grupo A	Grupo B
3	2	5
4	3	4
5	4	3
6	6	2
7	3	2
8	4	1
9	3	1
10	2	
11		1
12		1
13		2
14		2
15		1

Ahora podemos hacer algunas observaciones adicionales. Se nota que el *rango* (distancia entre el menor y el mayor valor del conjunto) es más amplio en el grupo B que en el grupo A. Posiblemente también nos damos cuenta que el valor más frecuente del grupo A (15) es superior al más frecuente del grupo B (12).

2.3 Tabla de frecuencias en R

Si bien es posible hacer una tabla de frecuencias a mano, simplemente contando las observaciones en cada categoría y anotando el resultado en orden, también tenemos funciones en R para el propósito.

```
table(
  c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17, 13, 14,
)
```

```
#>
#>  9 11 12 13 14 15 16 17 18 19
#>  1  2  2  3  4  6  3  4  3  2
```

En este ejemplo estamos usando dos funciones, una dentro de otra. La función `c`, le pide a R que arme un conjunto de datos, y los datos que queremos usar van entre paréntesis y separados por coma. Esto, a su vez, lo estamos haciendo dentro de la función `table` que genera una tabla de frecuencias.

También es posible darle un nombre a los datos a usar o «asignarlos a una variable», lo cual puede ser útil cuando se quiere reutilizar. Esto se hace de la siguiente manera:

```
x <- c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17, 13,
```

Con esto podemos usar `x` como alias para los datos que le asignamos. Entonces:

```
table(x)
```

```
#> x
#>  9 11 12 13 14 15 16 17 18 19
#>  1  2  2  3  4  6  3  4  3  2
```

nos da el mismo resultado.

Por lo general se recomienda usar nombres de variables que tengan algún sentido, en lugar de usar genéricos como x, y, z o a, b, c. En R las variables pueden tener múltiples caracteres (pero no espacios), por lo que podríamos ingresar:

```
grupo.A <- c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17
```

y nos daría el resultado deseado:

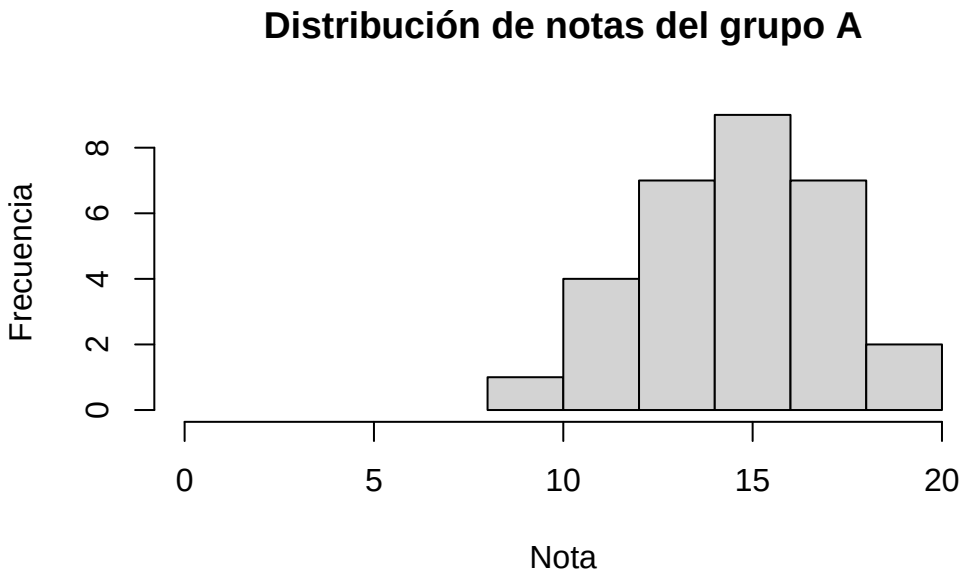
```
table(grupo.A)
```

```
#> grupo.A
#>  9 11 12 13 14 15 16 17 18 19
#>  1  2  2  3  4  6  3  4  3  2
```

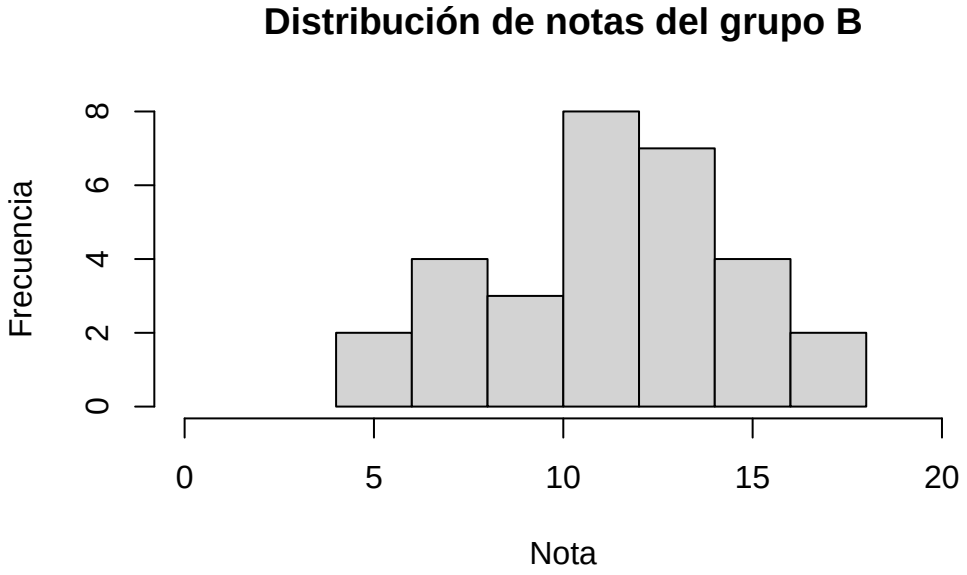
2.4 Histogramas

Para seguir explorando las tablas que hemos creado en la sección anterior se pueden visualizar con un *histograma*. El histograma resume los datos dentro de algunos rangos, por ejemplo 8-9, 10-11, 12-13 etcétera, y se cuenta el número de observaciones dentro de cada rango.

Para nuestros datos obtenemos:



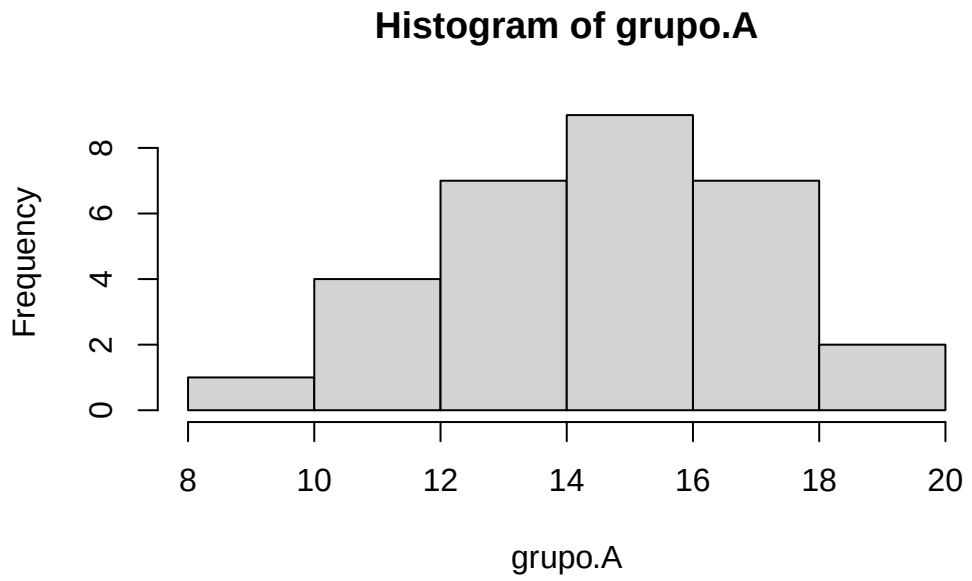
y



Comparando estos dos diagramas nos damos cuenta de que la estructura de los datos son disimilares. En el grupo A las notas se centran alrededor de quince, en cambio para el grupo B la concentración está en el rango diez-catorce, con un pico menor alrededor de siete.

Ejemplo 2.1 (Histograma en R)**.** Hacer un histograma con R es bastante sencillo. Usamos la función `hist`, de histograma y los datos que queremos visualizar. Si lo asignamos a una variable, como lo vimos en la parte de las tablas (con `table`).

```r
grupo.A <- c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17
hist(grupo.A)
```



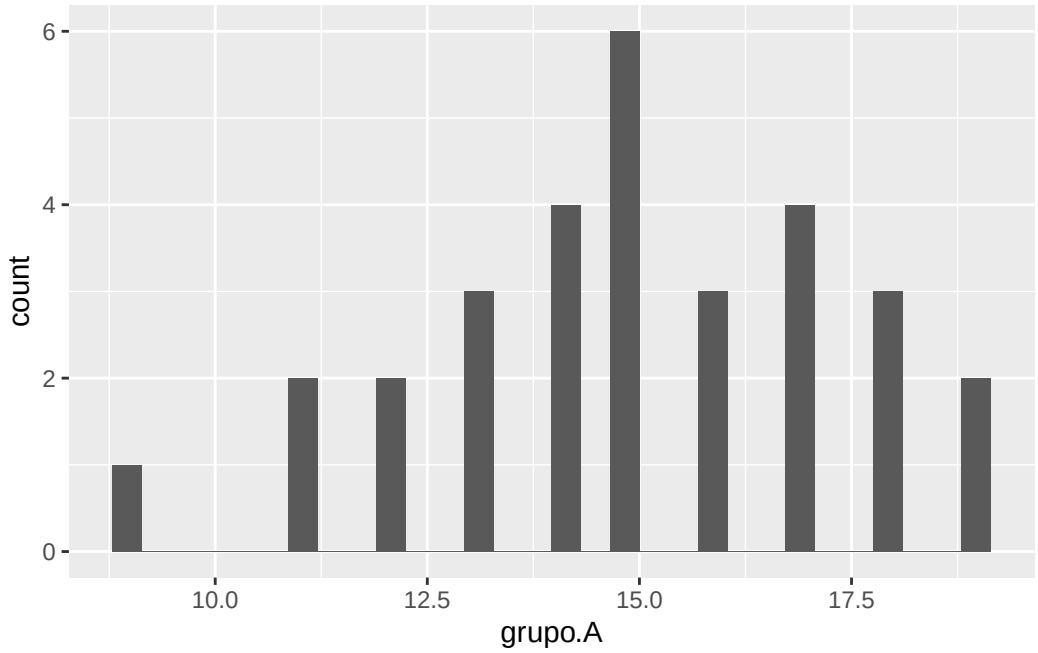
La función `hist` tiene muchas opciones adicionales. Para conocerlas se puede ingresar `?hist` (signo de interrogación y «hist») en la consola de R y aparecerá la descripción completa de ellas. Lo mismo es cierto para cualquier función de R. El mismo resultado se obtiene usando la función `help(hist)`.

Ejemplo 2.2 (Histograma en ggplot)**.** Usando los paquetes de tidyverse podemos generar un histograma con el packete ggplot2. Se carga por default junto con muchos otros paquetes. A diferencia del ejemplo anterior la función espera un `data.frame` como argumento. Para generar un histograma con los mismos datos debemos entonces proceder con crear una estructura de `data.frame` primero y luego proceder.

```r
my_data <- data.frame(
  grupo.A = c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 1
)
```

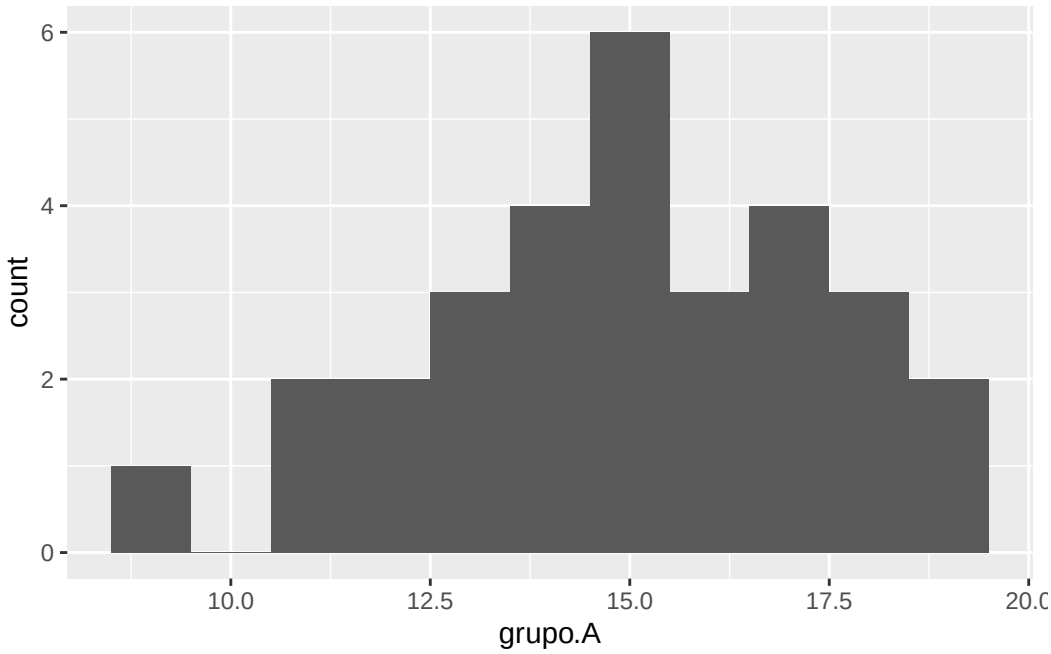
Nótese que usamos el operador `=` dentro de la definición del data.frame. Luego cargamos las funciones de tidyverse y procedemos a construir nuestro gráfico.

```r
library(tidyverse) # Carga todos los paquetes, incluso ggplot2
ggplot(my_data, aes(x=grupo.A)) +
  geom_histogram()
```



vemos que si bien los datos son los mismos las columnas parecen separados. Esto se debe a que por defecto el `geom_histogram` distribuye los datos en 30 columnas, lo cual es demasiado para el caso que tenemos. Podemos arreglar esto agregando otro parámetro a la función así:

```r
ggplot(my_data, aes(x=grupo.A)) +
  geom_histogram(binwidth = 1)
```

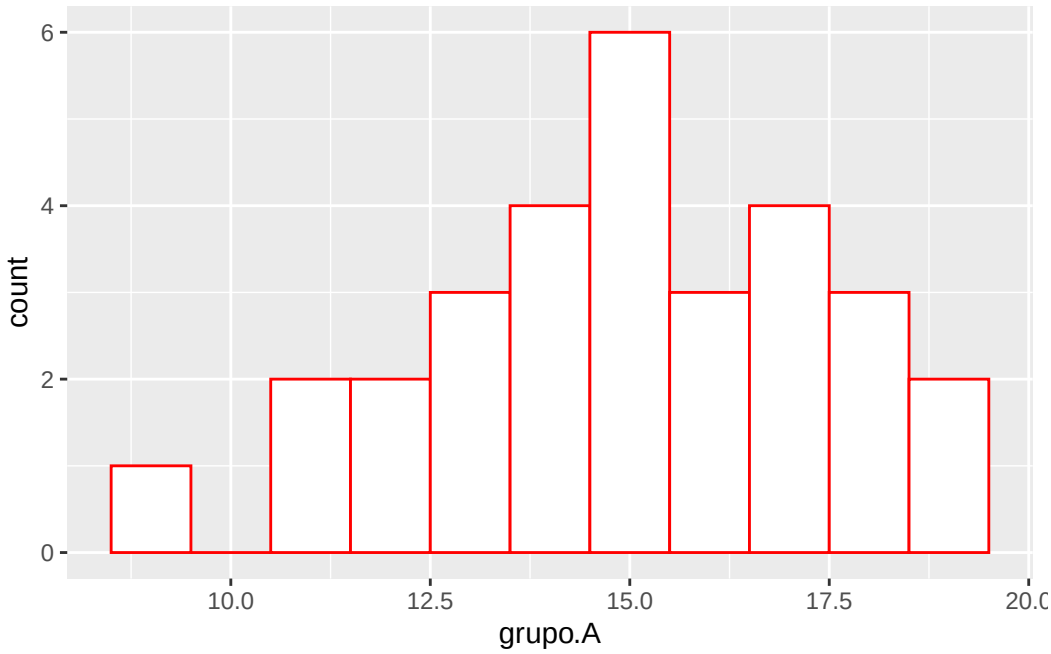


en este caso hemos especificado que el ancho de cada columna sea de 1, con lo cual se visualizan mejor estos datos.

2.4.0.0.1 Agregando un poco de color

Podemos también manipular los colores de las columnas con algunos parámetros más:

```
ggplot(my_data, aes(x=grupo.A)) +
  geom_histogram(binwidth = 1, fill="white", color='red')
```



El operador «pipe»

El uso de `%>%` es muy frecuente cuando uno trabaja con el tidyverse.

Desde R 4.0 existe también el «pipe nativo» `|>` que generalmente cumple la misma función.

2.5 Polígono de frecuencias

Los datos también de pueden visualizar con un polígono de frecuencias. En este tipo de visualización ponemos un punto en la intersección de la nota (eje horizontal) y la frecuencia (eje vertical) y trazamos una linea entre los puntos. Una de las ventajas de este tipo de visualización es que facilita la comparación entre varias distribuciones ya que los podemos desplegar en un mismo diagrama.

Apreciamos con más precisión los valores más típicos y diferencias entre los dos grupos. También podemos ver que la parte inferior de la escala de notas está sin uso, característica que comparten ambos grupos. :::{exm-otro-ejemplo}

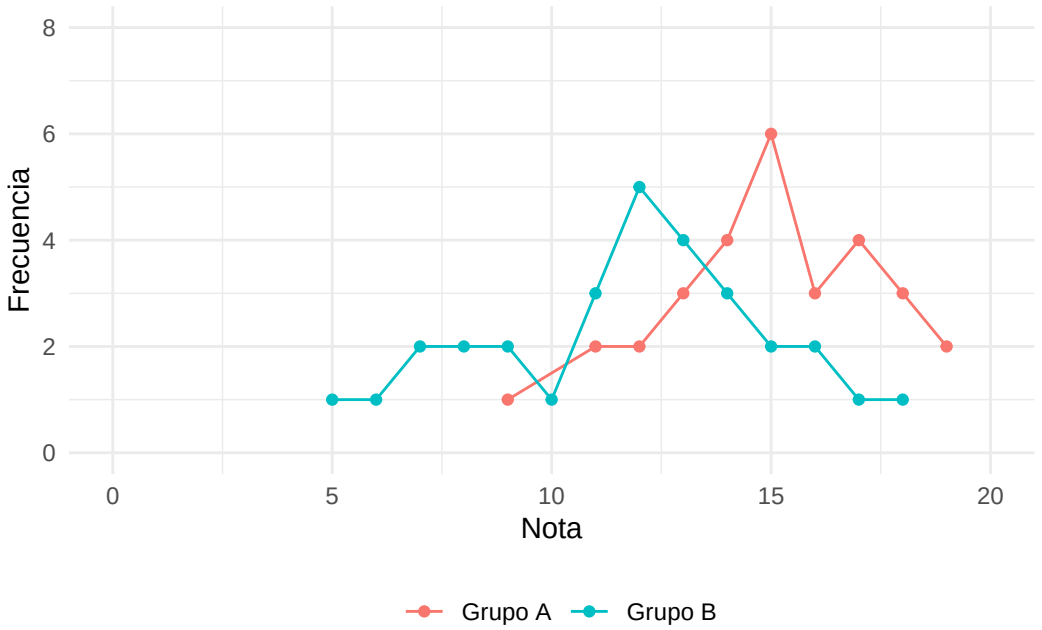


Figura 2.1: Polígono de frecuencias de notas obtenidas por dos grupos de estudiantes

2.6 Otro ejemplo

En este ejemplo vamos a considerar un libro de la literatura romántica: «Persuasion» escrito por Jane Austen [Austen (1817)][1]. Vamos a visualizar *el número de caracteres por palabra* en el texto. Obtenemos:

A diferencia de la distribución de notas, vemos acá que encontramos observaciones a lo largo del rango de uno a dieciséis, con la concentración de valores alrededor de tres. Esto tiene su interpretación bastante intuitiva ya que el uso de palabras *cortas*, como son artículos, preposiciones y conjunciones abundan en cualquier texto y las palabras muy largas son de uso menos frecuente. Resulta lógico suponer que encontraríamos un perfil similar en cualquier texto de cierta longitud.

:::

2.7 Perfil de la distribución

Las distribuciones de notas que vimos en las secciones anteriores tienen relativamente pocos datos, por lo que siempre van a parecer algo irregulares. Si tenemos muchos datos, sobre todo si con se escala de medición continua, podemos imaginarnos que en lugar de trazar una linea

[1]El texto está disponible online y en el paquete de R «tidytext» (Silge y Robinson 2016)

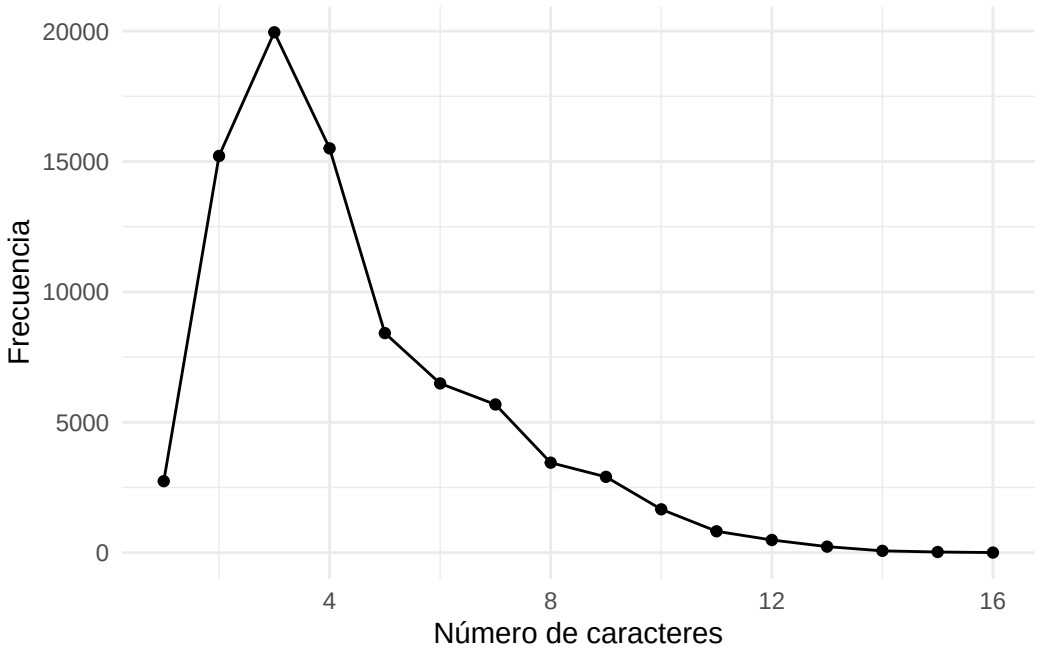


Figura 2.2: Polígono de frecuencias del largo de palabras en un texto de Austin

llegamos a trazar más bien una curva entre los puntos. Esto nos permite hacer una abstracción de las distribuciones y hablar de distribuciones teóricas. La más conocida de ellas sin duda es la *distribución normal*, también llamada de Gauss o gaussiana.

Vamos a desarrollar el tema de la distribución normal con más detalle en el capítulo 4. Por ahora simplemente vamos a considerar si los datos de nuestras muestras se asemejan a ésta o si tiene otro perfil.

2.7.1 Asimetría o Sesgo

Cuando una distribución se inclina en una dirección u otra decimos, es decir que no es simétrica, se dice que tiene un *sesgo* o que es *asimétrica*. Se habla de *sesgo negativo* y *sesgo positivo* (también: *asimetría positiva/negativa* y *a la izquierda/derecha* todos equivalentes). Es positivo o negativo según en qué dirección tiene su *cola larga*.

Vemos que nuestras distribuciones de *notas* corresponden a una distribución de *sesgo negativo*, ya que hay menos notas en la parte inferior de la escala que en la parte superior. En cambio, la distribución de *número de caracteres* en el texto de Austen tiene *sesgo positivo*.

Nótese también que la si bien la escala vertical de los dos gráficos son de muy diferente magnitud, la máxima frecuencia es veinte mil (20.000) y seis (6) respectivamente, podemos comparar las dos distribuciones.

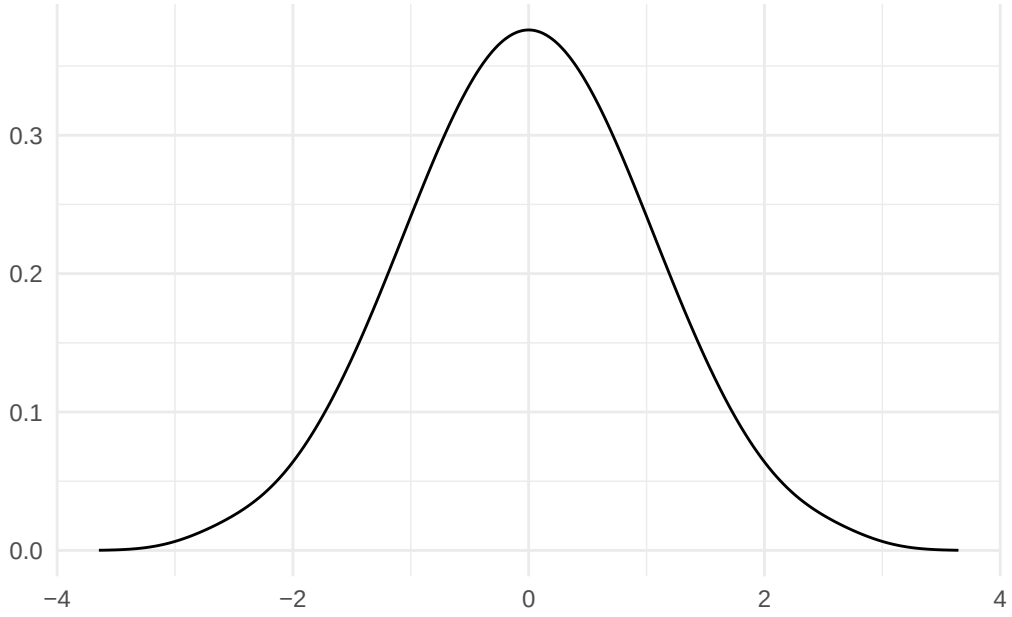


Figura 2.3: Distribución normal

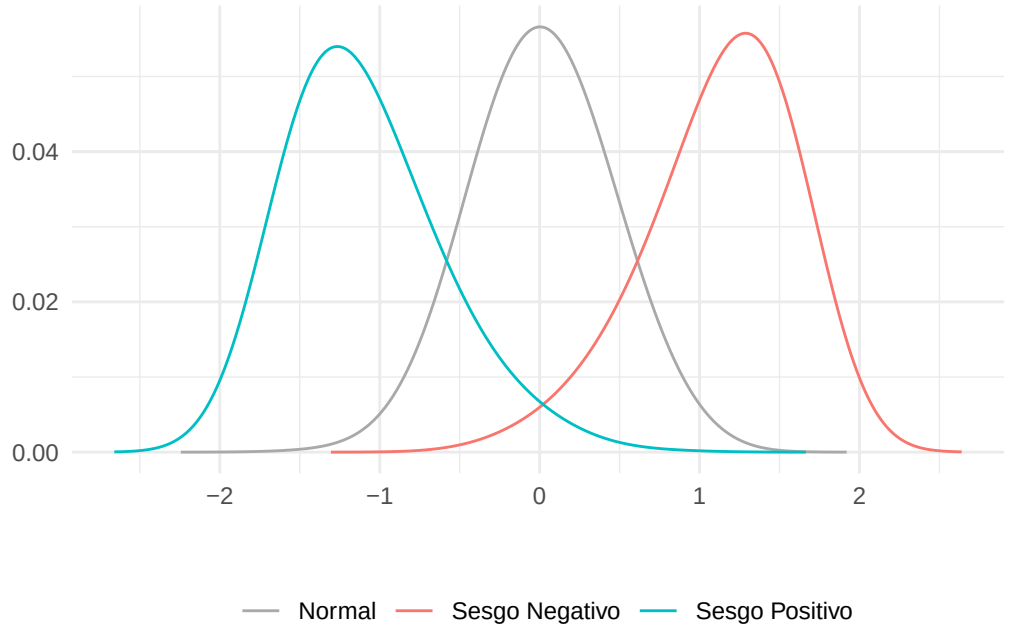


Figura 2.4: Distribuciones normal y sesgadas

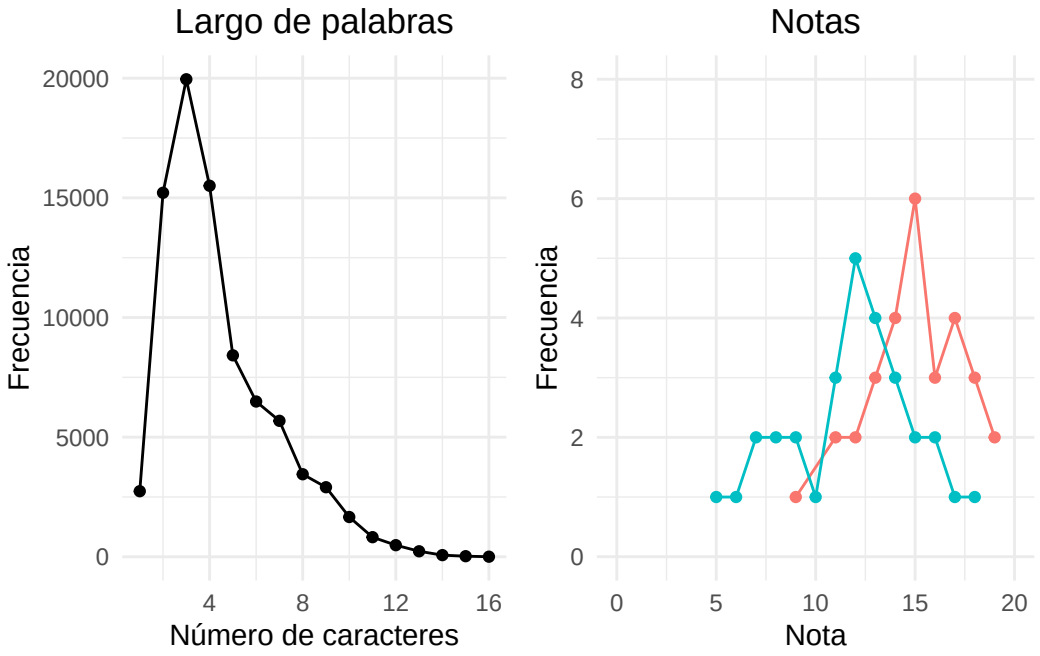


Figura 2.5: Polígonos de frecuencias

2.8 Glosario

Asimetría El hecho de que una distribución no sea simétrica. Equivalente en inglés: «Skew».

Distribución normal Distribución teórica de una variable. Es simétrica y con forma de campana. Equivalente en inglés: «Normal distribution».

Histograma Visualización de frecuencia agrupadas de observaciones de una variable. Función relevante en R: `hist`. Equivalente en inglés: «Histogram».

Polígono de frecuencias Visualización de frecuencias de observaciones de una variable. Equivalente en inglés: «Frequency poligon».

Segso El hecho de que una distribución no sea simétrica. Equivalente en inglés: «Skew».

Table de frecuencias Tabla que resume las frecuencias de las observaciones de una variable. Función relevante en R: `table`. Equivalente en inglés: «Frequency table».

3 Centralización y dispersión

En el capítulo 2 vimos que resumir los datos y generar visualizaciones nos permite entender mejor la estructura y algunas propiedades de un conjunto de datos, como son sus vales más frecuentes y rango de observaciones. En este capítulo desarrollaremos algunas medidas cuantitativas más precisas de estas propiedades. Específicamente desarrollaremos medidas de centralización o tendencia central y dispersión.

3.1 Centralización

La *centralización* o *tendencia central* de un conjunto de datos es uno o un número reducido de valores que representan todo el conjunto.

Existen tres medidas de *centralicación*: *la media*, *la mediana* y *la moda*. A continuación las vamos a definir y ver cómo se calculan y luego vamos a considerar cuándo se debe usar cada una de ellas.

3.1.1 La media

La media es seguramente la medida de centralización de uso más frecuente [1]. Se conoce también como *el promedio* y, más técnicamente, *la media arithmetica*. La media se obtiene por la suma de las observaciones dividido por el número de observaciones. Por ejemplo si queremos sacar el promedio de seis observaciones de una variable: 15, 12, 11, 18, 15 y 15; tenemos:

$$\frac{15+12+11+18+15+15}{6} = \frac{86}{6} = 14,33 \tag{3.1}$$

En el caso de nuestra muestra de notas para de capítulos anteriores tenemos:

$$\frac{15+12+11+18+15+15+9+19+14+13+11+12+18+15+16+14+16+17+15+17+13+14+}{30} \tag{3.2}$$

[1] y por ende de más uso incorrecto

Ya con el cómputo en Ecuación 3.2 nos damos cuenta de que si bien es posible hacer estos cálculos a mano puede resultar bastante engorroso. Además con tantos números dando vuelta sube la probabilidad de un error de tipeo y con lo cual sacaríamos un resultado incorrecto.

Ejemplo 3.1 (La media)**.** Por suerte es bastante sencillo sacar la media con R. Para los dos ejemplos anteriores tenemos:

```
x = c(15, 12, 11, 18, 15 , 15)
mean(x)
```

```
#> [1] 14.33333
```

y

```
notas = c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18,
          15, 16, 14, 16, 17, 15, 17, 13, 14, 13, 15, 17, 19,
          17, 18, 16, 14)
mean(notas)
```

```
#> [1] 14.93333
```

Notación matemática

En textos de matemática y estadística se usa con frecuencia *llaves* para significar un conjunto, de modo que los datos del primer conjunto se expresaría así: x = {15, 12, 11, 18, 15 , 15}.

Una notación compacta para significar la suma de las observaciones en una variable es Σ: la letra griega sigma, en mayúscula.

Para significar el número de observaciones de usa N, de **n**úmero.

Así se puede definir la media de manera compacta así:

$$\frac{\Sigma x}{N}$$

También se usa una barra vertical sobre el nombre de la variable para significar la media (o promedio aritmético): por ejemplo:

$$\bar{x} = 14,33$$

Entonces en general tenemos:

Definición 3.1 (La media)**.**

$$\bar{x} = \frac{\Sigma x}{N}$$

que se podría leer: «la media de equis es igual a la suma de las observaciones de equis sobre el número de observaciones».

3.1.2 La mediana

Otra medida de centralización es la mediana (también: valor mediano). Para obtenerla ponemos nuestros datos en orden ascendiente y sacamos el valor que está justo en la mitad. Por ejemplo: si queremos sacar la mediana de {15, 12, 11, 18, 15, 15, 9}, primero los ordenamos: {9, 11, 12, 15, 15 ,15, 18}. Vemos que hay siete observaciones con lo cual la mediana es la observación que está en cuarta posición, es decir que la mediana de estos datos es 15. Si el conjunto de datos tiene un número par de observaciones, no va a haber una observación justo en el medio. En ese caso se toman los *dos* valores del medio, se los suma y se divide por dos. Por ejemplo: {8, 8, 9, 11, 12, 15, 15 ,15}. Acá tenemos ocho observaciones (ya ordenados) tomamos los dos valores de la posición cuarta y quinta, los sumamos y dividimos por dos: $\frac{11+12}{2} = 11,5$.

Notación matemática

El valor mediano, o la mediana, se denota en notación matemática con una tilde como la que se usa en la letra ñ en español. Al igual que la barra para la media, se coloca por encima de la variable, así:

$$\tilde{x}$$

.

Ejemplo 3.2 (La mediana)**.**
Podemos sacar la mediana de forma sencilla con R con la función `median`.

```
x = c(9, 11, 12, 15, 15 ,15, 18)
median(x)
```

```
#> [1] 15
```

y

```
x = c(8, 8, 9, 11, 12, 15, 15 ,15)
median(x)
```

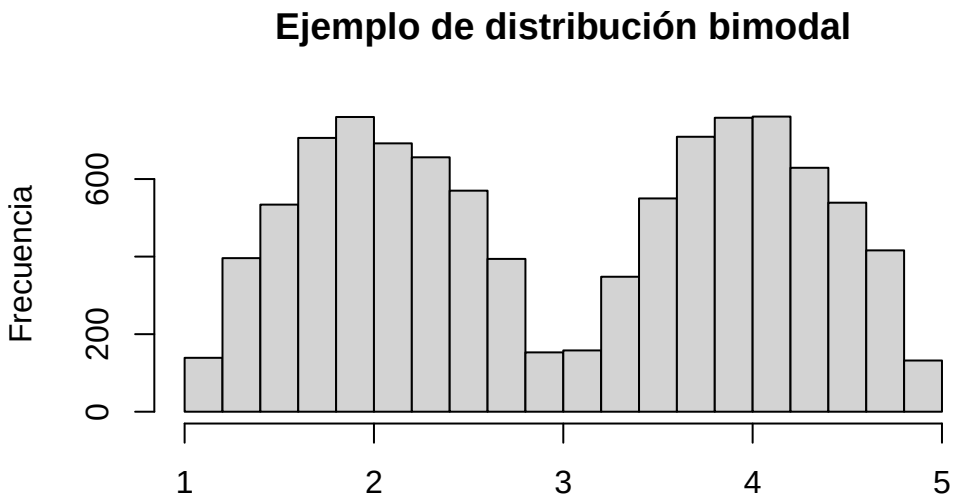
```
#> [1] 11.5
```

3.1.3 La moda

La moda es la observación más frecuente del conjunto. Por ejemplo: {9, 11, 12, 15, 15 ,15, 18}. El valor 15 es la moda de estos datos.

A diferencia de las otras medidas de centralidad la moda no necesariamente es un valor único. Si tuviéramos por ejemplo: {2, 4, 5, 7, 7, 7, 9, 11, 12, 15, 15, 15, 18} hay dos valores con la misma frecuencia máxima. Tanto 7 como 15 aparecen tres veces. En este caso hay dos modas y hablamos de una distribución *bimodal*.

Vemos un ejemplo en el gráfico que sigue.



3.1.4 ¿Cuál usar?

La selección de una medida de centralización depende de varios factores:

1. La escala de medición de la variable (nominal, ordinal, de intervalo o de razón)
2. La forma de la distribución - si hay sesgo o no
3. Para qué vamos a usar la medida.

La media debería usarse solo para variables de escala de intervalo o de razón. Si los datos son ordenables, pero sin que se pueda hablar de distancias reales entre los datos la mediana es más apropiada. Y en los casos donde ni esto es posible la moda puede ser la única medida disponible. Por ejemplo: si decimos que Italia es un país católico estamos expresando la moda de la variable nominal «religión», y si decimos que Alemania es un país católico y protestante estamos expresando una distribución bimodal de la misma variable. Podemos observar en el gráfico que en realidad se podría hablar incluso de una distribución trimodal.

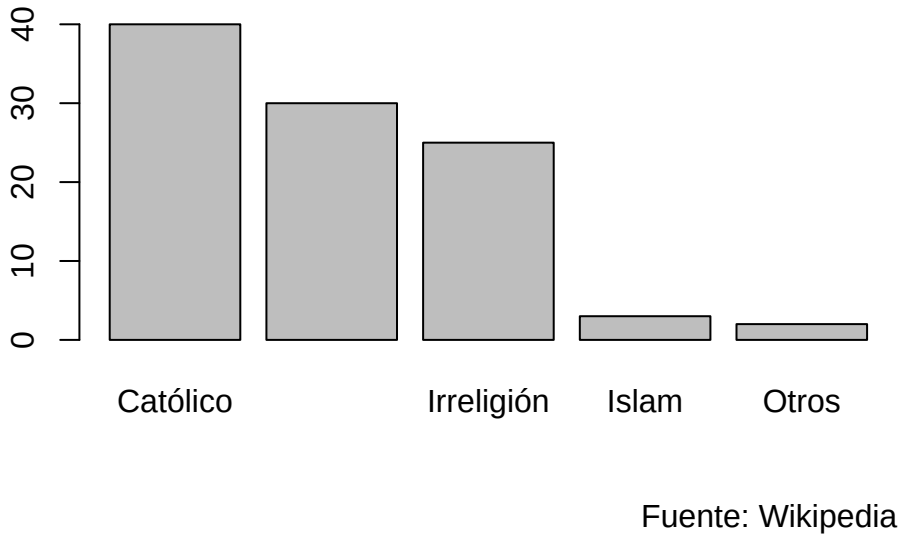


Figura 3.1: Religión en Alemania

En cuanto a la forma de la distribución se favorece la mediana por sobre la media si la distribución es muy sesgada. Esto ocurre sobre todo si hay valores extremos o atípicos. Por ejemplo si tenemos los datos: {15, 12, 11, 18, 15, 15, 200} está claro que si calculamos la media el valor extremo (200) va influir mucho más que cualquier otra observación. En este caso la media es 40,85 y el mediano 15. El primer valor (40,85) no es muy representativo de la muestra ya que no corresponde a ninguna observación y está lejos de cualquiera de ellas. El mediano, en cambio, puede resultar una mejor medida en este caso.

Para darnos cuenta de cuál de las medidas puede ser la más adecuada si tenemos datos por lo menos numéricos podemos sacar las tres medidas y ver qué tanto de asemejan unas a otras. Hay que tener en mente que *cualiér distribución de datos reales va a tener un sesgo*, la distribución perfectamente normal solo existe en teoría. Entonces debemos fijarnos si el sesgo que tenemos justifica el uso de una medida en espeficia. Por ejemplo, para nuestros datos de notas de dos grupos tenemos:

- Grupo A
  - Media: 14,93
  - Mediana: 15

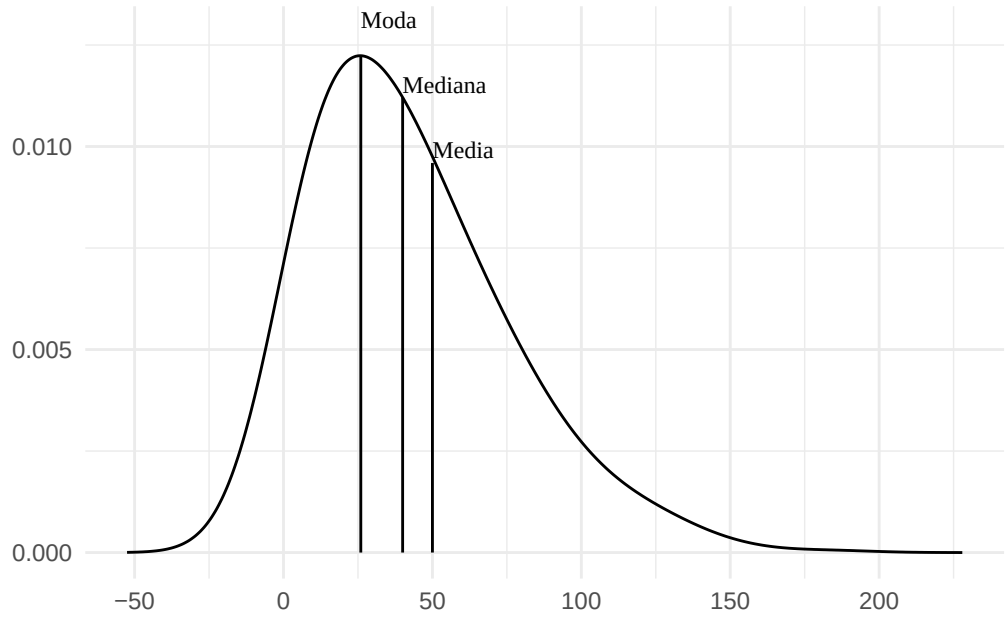


Figura 3.2: Medidas de centralización en una distribución con sesgo positivo

  - Moda: 15

- Grupo B:
  - Media: 11,76
  - Mediana: 12
  - Moda: 12

Vemos que hay muy poca diferencia entre las tres medidas por lo cual vamos a concluir que el sesgo observado no es lo suficientemente fuerte como para justificar el uso de otra medida que *la media*.

3.2 Medidas de dispersión

En la sección anterior desarrollamos varias medidas de centralización y cuál eligir para describir el valor «más típico» de los datos. Cuando calculamos medidas de dispersión estamos contestando la pregunta: ¿cuán típico es este valor?

Cuando tratamos con variables nominales, como el ejemplo de religión en Alemania de la

sección anterior, lo mejor que podemos hacer el indicar la proporción o porcentaje[2], pero si los datos son de alguna escala ya numérica tenemos algunas posibilidades que nos permiten más exactitud.

3.2.1 Rango o amplitud

El rango de un conjunto de datos son dos números: el valor mínimo y el valor máximo. Por ejemplo el conjunto de datos {9, 11, 12, 15, 15 ,15, 18} tiene un rango 9 a 18; y el conjunto {2, 4, 5, 7, 7, 7, 9, 11, 12, 15, 15 ,15, 18} tiene un rango de 2 a 18.

En castellano se usa con alguna frecuencia también el término *amplitud* como equivalente a *rango*.

Para sacar el rango de un conjunto de datos en R podemos usar la función `range`. Así:

```
x = c(2, 4, 5, 7, 7, 7, 9, 11, 12, 15, 15 ,15, 18)
range(x)
```

```
#> [1]  2 18
```

3.2.2 El rango intercuartílico

Otra medida de dispersión que tenemos a disposición es el *rango intercuartílico* o *rango intercuartíl*. Para calcularlo dividimos las observaciones en cuatro partes iguales y sacamos los valores de cada corte. Esto nos da *cinco valores*[3], le los cuales el *rango intercuartílico* es la diferencia entre el segundo y el cuarto. Este sería el rango de las observaciones del 50% de los datos que se encuentran más cerca la mediana del mismo.

El rango intercuartílico da una idea de la dispersión de los datos y es por su naturaleza menos sensitivo a valores extremos.

Ejemplo 3.3 (Sacar el rango intercuartílico en R)**.** Para sacar en rango intercuartílico podemos usar la función `quantiles`. Por defecto divide la distribución en cuartiles.

```
x = c(2, 4, 5, 7, 7, 7, 9, 11, 12, 15, 15 ,15, 18)
quantile(x)
```

[2]Las dos medidas son equivalentes ya que: 0,1 = 10%; 0,5 = 50% etcétera. En estadística y matemática se prefiere generalmente la expresión de proporción porque facilita ciertas operaciones aritméticas.

[3]Tres cortes más los valores extremos mínimo y máximo

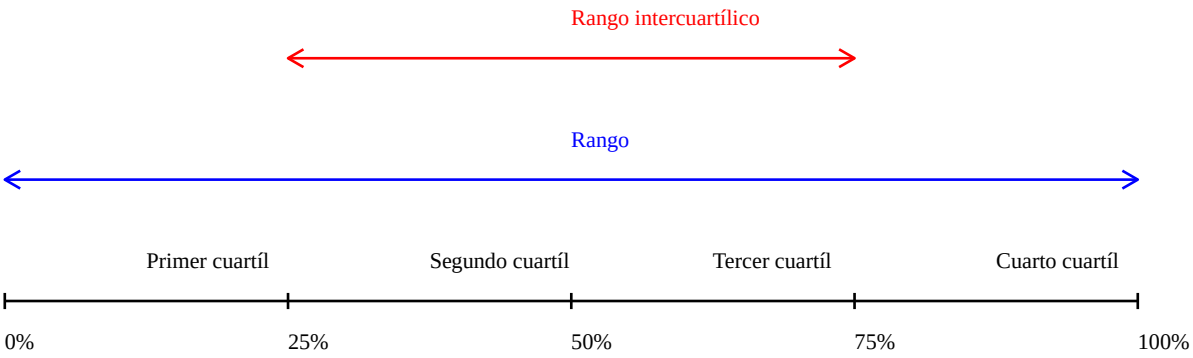


Figura 3.3: Cuartiles y rangos

```
#>   0%  25%  50%  75% 100%
#>    2    7    9   15   18
```

Vemos que en este caso el rango intercuartíl es 7 y 15, que da una amplitúd de 8 ya que $15 - 7 = 8$.

3.2.3 La varianza y desviación estándar

La medida de dispersión más usada en estadística es la *desviación estándar*, también conocida como *desviación típica*. Esta medida tiene una relación matemática muy estrecha con la *varianza* que tiene usos menos frecuentes. Ambas medidas tienen propiedades que los hacen útiles para otras técnicas estadísticas.

Para calcular la desviación estándar debemos primero calcular la varianza. Para ello tomamos la diferencia de cada observación de la media. Recordemos que la media se expresa con $\bar{x}$ (equis con barra). Entonces la diferencia entre una observación de x y la media es $x - \bar{x}$. Luego los llevamos al cuadrado $(x - \bar{x})^2$ los sumamos y dividimos por el número total de observaciones. Para expresarlo usamos la notación que ya vimos. Entonces Σ es «la suma de» y N es «el total de las observaciones». Juntando todo tenemos:

Definición 3.2 (Varianza)**.**

$$\text{varianza} = \frac{\Sigma(x - \bar{x})^2}{N}$$

Ahora para sacar la desviación estándar tomamos la raíz cuadrada de la varianza. La desviación estándar de la población se representa por la letra griega σ que es sigma pero en minúscula. Entonces tenemos:

Definición 3.3 (Desviación estándar de la población)**.**

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N}}$$

Si estamos trabajando con una muestra en lugar de la población completa, que es el caso más común cuando trabajamos con estadísticas se usa la letra «s». También se hace un ajuste en el denominador de la fórmula ya que se ha comprobado que sin el ajuste la medida puede resultar sesgada si la muestra tiene pocas observaciones. La formula para una muestra es:

Definición 3.4 (Desviación estándar de la muestra)**.**

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N - 1}}$$

Finalmente. Ya que s y σ son la raíz cuadrada de la varianza, esta también se denomina por las mismas letras, pero llevado al cuadrado: s^2 y σ^2

Por suerte es sencillo sacar tanto la varianza como la desviación estándar en R. Usamos las funciones `var` y `sd`[4].

```
x = c(2, 4, 5, 7, 7, 7, 9, 11, 12, 15, 15 ,15, 18)
var(x)
```

```
#> [1] 24.69231
```

```
sd(x)
```

```
#> [1] 4.969136
```

[4] «sd» por la abreviación del inglés «**s**tandard **d**eviation».

¿Por qué se prefiere la desviación estándar?

Hay varios motivos más bien técnicos por los que se prefiere la desviación estándar por sobre la varianza. Sin embargo tiene también algunas ventajas bastante práctica e incluso intuitivas. Una de las más importantes es que la dispersión se expresa en *la misma unidad* que los datos. Para profundizar esto vemos un ejemplo. Los salarios de una PYME son: $14.000, $14.000, $14.000, $16.000, $17.000, $18.000, $26.000 y $35.000. La media de estos es 19,250, y la desviación estándar es: 7,497. La interpretación de la desviación estándar en este caso es que los salarios en promedio tiene una diferencia de $7,497 (por arriba o abajo) del salario medio de $19,250.

3.2.4 Visualizar la dispersión

Puede resultar útil visualizar la dispersión de un conjunto de datos. Esto se logra con un diagrama de caja (box-plot). Vemos un ejemplo de ello en la figura 3.4.

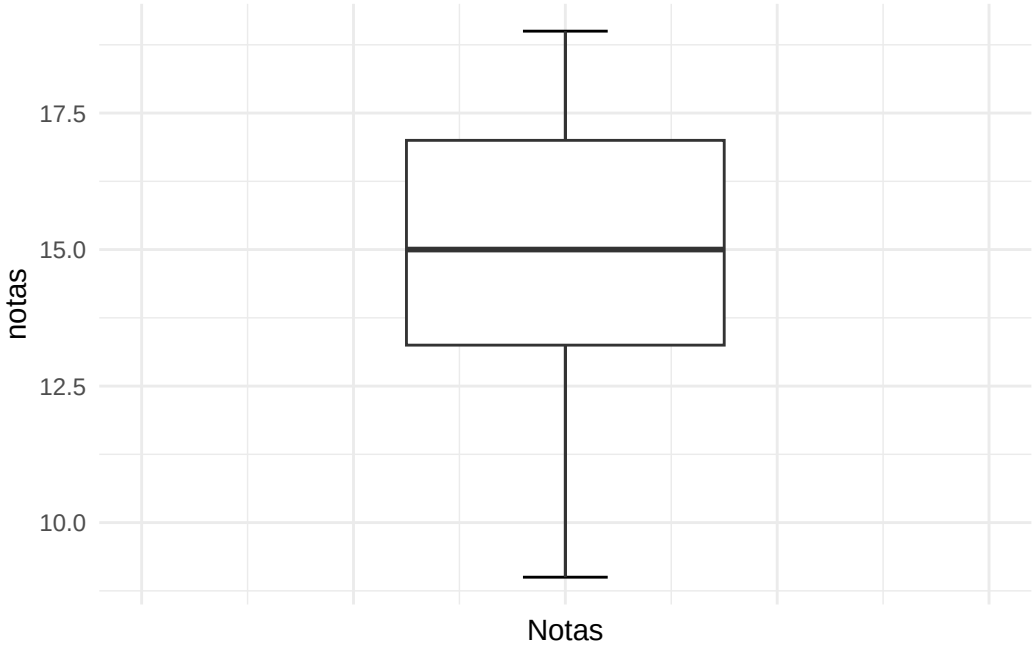


Figura 3.4: Ejemplo de box-plot

En este tipo de visualización la mediana está representada por la linea horizontal más gruesa, la caja corresponde al rango intercuartíl y los extremos de la linea horizontal representan el rango de los datos. Lo podemos apreciar en la figura @(ref:box-plot-with-explanation)

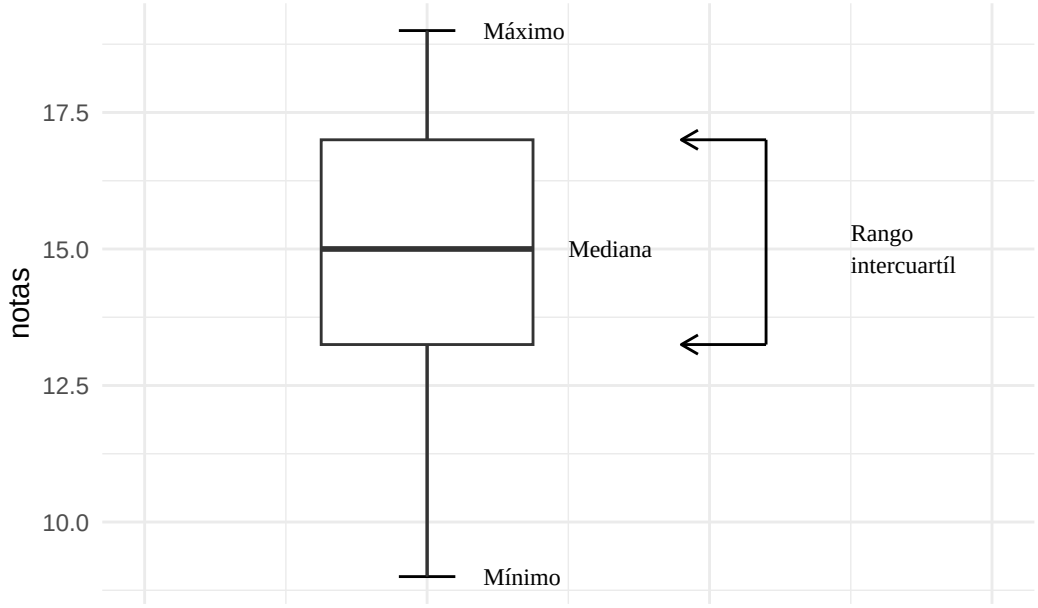
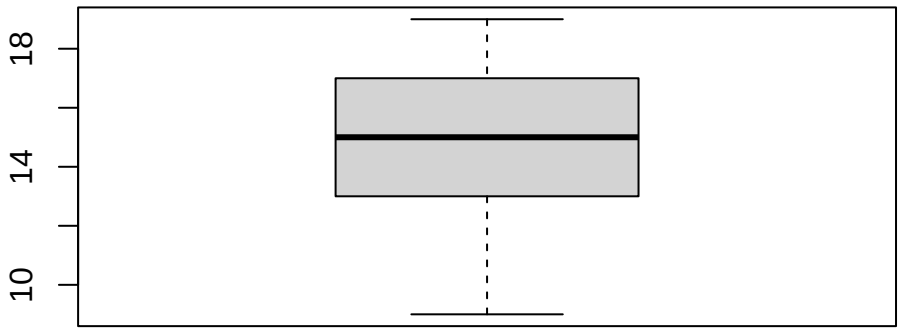


Figura 3.5: Ejemplo de box-plot con explicaciones

Ejemplo 3.4 (Boxplot)**.** La función `boxplot` nos permite generar un boxplot en R.

```
notas = c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18,
          15, 16, 14, 16, 17, 15, 17, 13, 14, 13, 15, 17, 19,
          17, 18, 16, 14)
boxplot(notas)
```



3.3 Glosario

Amplitud La diferencia entre la mínima y la máxima de una variable. También se llama *rango* Función relevante en R: `range`. Equivalente en inglés: «Range».

Centralización El hecho de que una variable puede describirse por uno o más valores. También se llama *tendiencia central*. Equivalente en inglés: «Central tendency».

Desviación estándar (de la muestra). Media de la diferencia entre la media y todas las observaciones de la muestra. Fórmula: $s = \frac{\sqrt{\Sigma(x-\bar{x})^2}}{N}$ Función relevante en R: `sd`. Equivalente en inglés: «Standard deviation».

Desviación estándar (de la población). Media de la diferencia entre la media y todas las observaciones de la población. Fórmula: $\sigma = \frac{\sqrt{\Sigma(x-\bar{x})^2}}{N}$ Función relevante en R: `sd`. Equivalente en inglés: «Standard deviation (of the population)».

Desviación típica Ver *desviación estándar*. Equivalente en inglés: «Standard deviation».

Media La suma de las observaciónes de una variable dividido por el número de las observaciones. También se conoce como *la media aritmética*. Fórmula: $\bar{x} = \frac{\Sigma x}{N}$ Función relevante en R: `mean`. Equivalente en inglés: «Mean».

Mediana El la observación de una variable que está justo en el medio cuando los valores están ordenados. Función relevante en R: `median`. Equivalente en inglés: «Median».

Moda El valor más frecuente de la observaciones de una variable. Equivalente en inglés: «Mode».

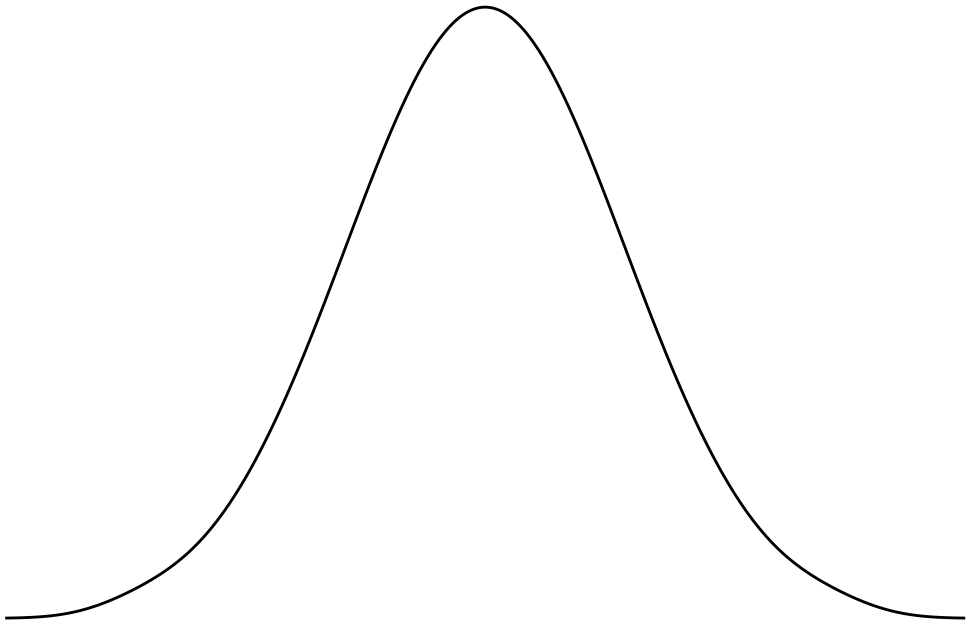
Rango La diferencia entre la mínima y la máxima de una variable. También se llama *amplitud* Función relevante en R: `range`. Equivalente en inglés: «Range».

Rango intercuartílico Rango dentro del cual se encuentras en 50% más centralizado de las variables. Función relevante en R: `quantile`. Equivalente en inglés: «Interquartile range (IQR)».

Varianza Media de la diferencia cuadrada entre la media y todas las observaciones. Fórmula: $\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{N}$ Función relevante en R: `var`. Equivalente en inglés: «Variance».

4 La distribución normal

En el Capítulo 2 tocamos brevemente la llamada *distribución normal*. En este capítulo vamos a desarrollar con más detalle esta distribución, fundamental para muchas técnicas estadísticas y cuantitativas.



4.1 Importancia de la distribución normal

Como vimos en la sección 2.7, si tenemos muchos datos y construimos un polígono de frecuencias, es posible trazar una curva entre los puntos de la distribución. También mencionamos que la llamada *distribución normal* es de particular interés para trabajo estadístico y cuantitativo. Hay varios razones de ello:

1. Muchos fenómenos que podemos medir tanto en las ciencias exactas como las sociales de asemejan en su frecuencia a esta distribución.
2. La distribución normal tiene ciertas propiedades matemáticas que nos permiten predecir qué proporción de la población (estadística) caerá dentro de cierto rango si la variable tiene distribución normal.

3. Varios tests de significanza de diferencia entre conjuntos de datos presumen que los datos del conjunto tiene una distribución normal.

4.2 Propiedades de la curva normal

Como ya vimos, la curva normal tiene forma de campana y es simétrica. Por ende, las tres medidas de centralización la media, la mediana y la moda coinciden en el punto superior de la curva, como lo podemos apreciar en la Figura 4.1.

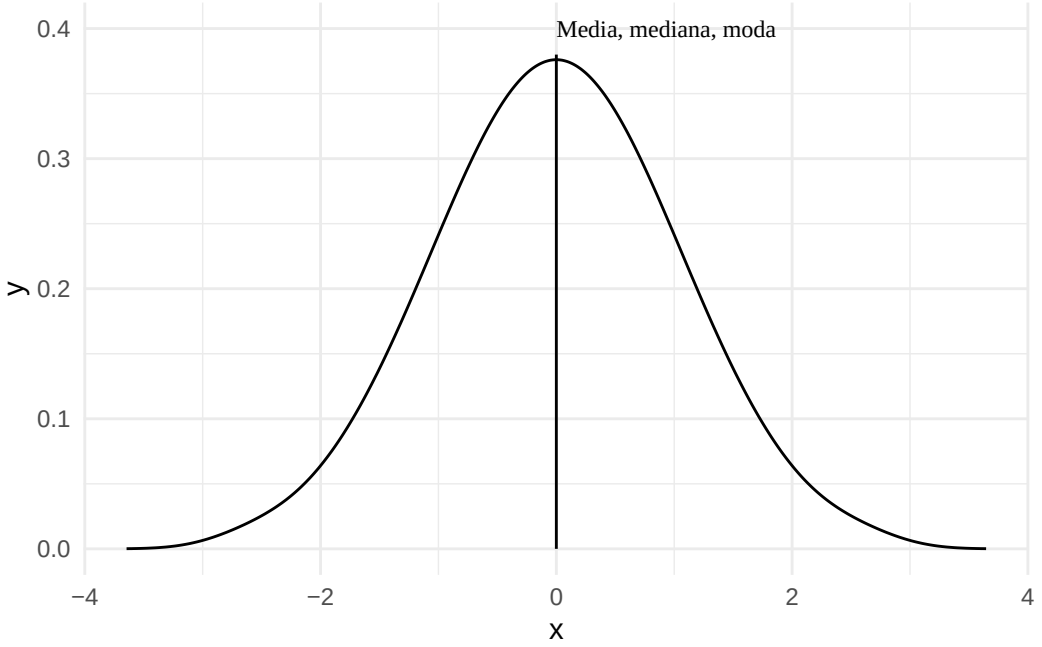


Figura 4.1: Curva normal

Ciertas propiedades importantes de esta curva se relacionan con la manera en que el área debajo de la curva de puede seccionar con lineas verticales con origen en distintos puntos del eje horizontal. Para explorar estas vamos a considerar algunos histogramas, el tipo de visualización que vimos en la sección 2.4. El alto de cada barra es proporcional a la frecuencia de observaciones y como el ancho de las barras es el mismo en todos los casos el área de cada barra también es proporcional a la frecuencia de observaciones. El ancho puede representar una sola unidad, o varias si agrupamos, por ejemplo por rango etario como lo vemos en la figura 4.2, en el que hemos sacado una muestra aleatoria de mil observaciones de un test de matemáticas a nivel nacional. Los hemos agrupado por rangos de diez, es decir de 0 a 10, de 10 a 20 y así sucesivamente. Hemos sobrepuesto una curva normal teórica para apreciar hasta qué punto se asemeja la distribución observada a la teórica.

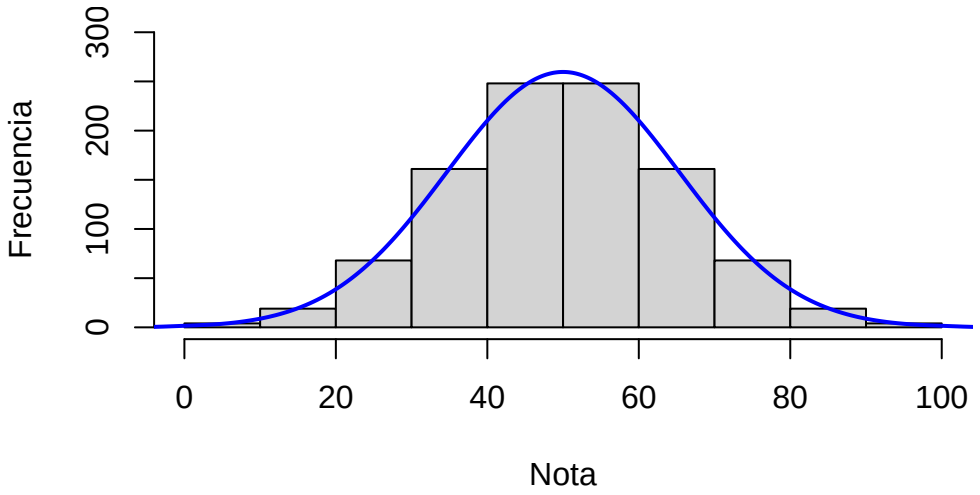


Figura 4.2: Muestra de notas de un test de matemática (N=1000)

Ahora, bien, si en lugar de agrupar las notas en grupos de diez[1] los podemos también agregar en grupos de cinco. Entonces obtenemos un histograma como el de la figura Figura 4.3 .

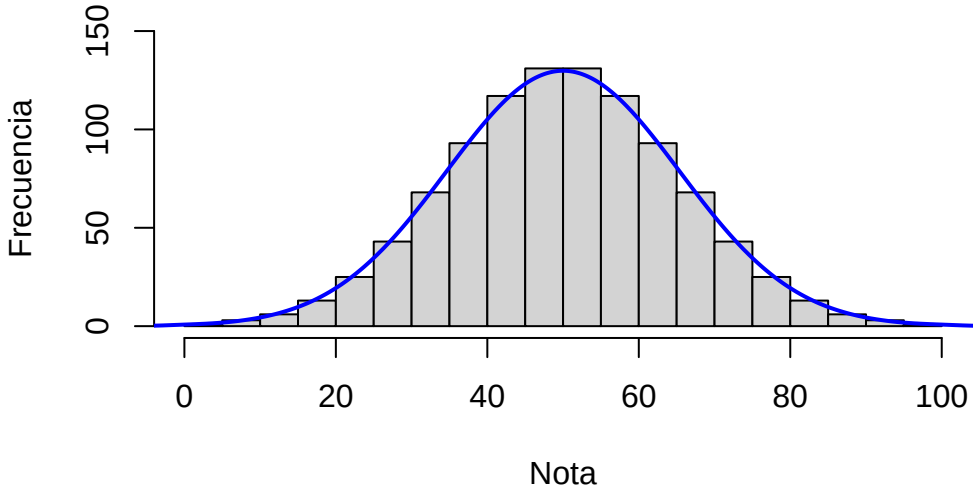


Figura 4.3: Muestra de notas de un test de matemática (N=1000)

Podemos seguir achicando el ancho de las barras, y vemos que si bien el histograma es puntudo mientras menos anchas son las barras más se aproxima a la curva. En la Figura 4.4 hemos achicado las barras para que cada una represente tan solo un valor entero, es decir tan solo una de las cien notas posibles. Se entiende que es posible seguir con más precisión si, por ejemplo, el examen fue calificado con la posibilidad de asignar notas con decimales.

La curva normal de define por dos propiedades: La media y la desviación estándar. Si conoce-

[1] también llamado deciles

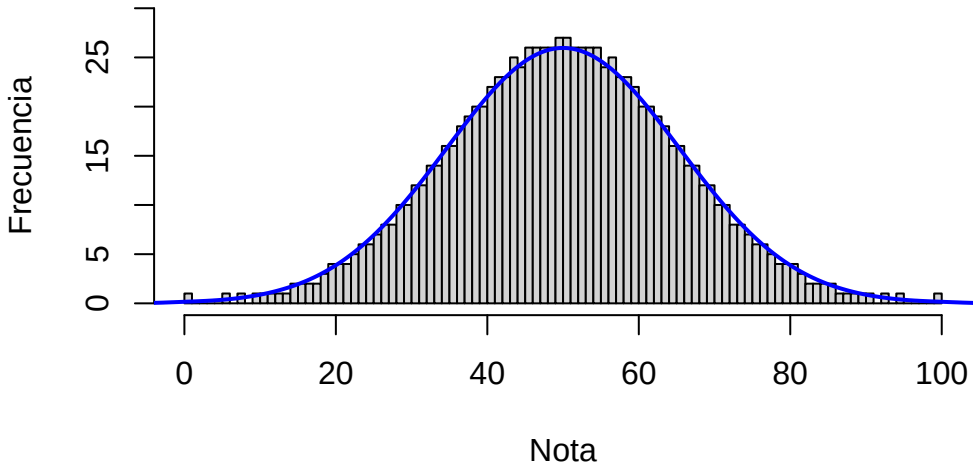


Figura 4.4: Muestra de notas de un test de matemática (N=1000)

mos estos dos valores es posible construir la curva aplicando una fórmula [2] un tanto compleja y con poca importancia fuera del ámbito plenamente teórico.

De más importancia son algunas propiedades que tiene la curva. Si graficamos la curva normal y expresamos los valores en el eje horizontal en *desviaciones estándares* (también se dice «sigmas» por su letra griega σ), el área que está de cada lado de la linea es constante y conocido. Si trazamos una linea justo en el medio ($\sigma = 0$), sabemos que un 50% de las observaciones están a la derecha y la izquierda de esa linea. Lo mismo aplica a una distribución expresado en un histograma. En la fig-normal-curve-with-cuts vemos cuales son los cortes para desviaciones estándares de menos 3 a 3.

Esta propiedad es de bastante utilidad y se puede aprovechar de varias maneras. Si tenemos una muestra de datos cuya distribución presumimos normal (en el Sección 7.2 vamos a desarrollar cómo lo podemos determinar) ya sabemos que más o menos el 68% de las observaciones va estar dentro de $\pm$ una desviación estándar de la media y más del 95% se encontrará dentro de dos desviaciones. Por último el 99% de las observaciones de encuentran dentro de tres desviaciones estándares de la media. A veces se refiere a esta propiedad como la *regla empírica* o la regla de de *68-95-99,7*.

Variables normalizadas

En textos de estadística frecuentemente se habla de *variable normalizada*, también se conoce como *unidad tipificada*, *variable centrada reducida* o *variable estandarizada*. Normalizar una variable es simplemente expresar su magnitud en unidades de desviación estándar. Para lograr ello tomamos la variable, restamos la media y dividimos por la desviación estándar. En

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

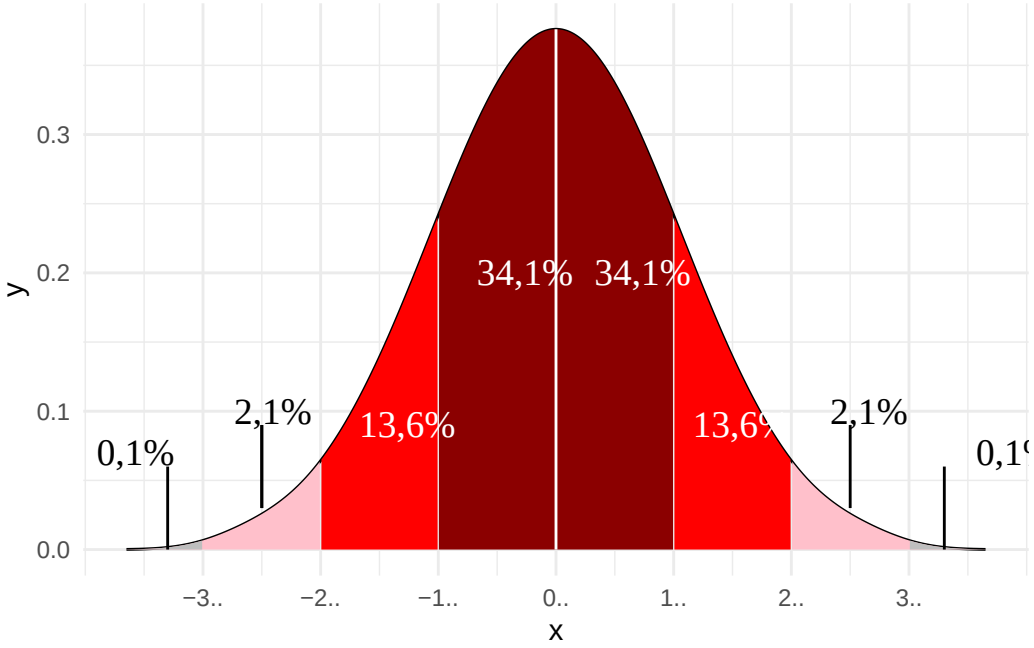


Figura 4.5: Área debajo de la curva normal

literatura en inglés es de uso frecuente el término «z-score», por lo que su definición formal (véase @def-definition-z-score)) lleva esta letra.

Definición 4.1 (Variable normalizada)**.** La variable normalizada z de un conjunto de datos X se obtiene por la fórmula siguiente:

$$z = \frac{x - \bar{x}}{\sigma}$$

donde:

- z: la variable normalizada
- x: una observación de X
- $\bar{x}$: la media de las observaciones
- σ o s: la desviación estándar de la población o muestra respectivamente.

Es importante entender que normalizar una variable no cambia su valor, solo su unidad de cuenta: El lo mismo comprar medio kilo de queso que comprar quinientos gramos.

Normalizar las variables nos permite comparar su distribución independientemente de su unidad de cuenta y amplitud, también nos permite sacar conclusiones sobre probabilidades y proporciones. Vamos a desarrollar esta idea por medio de un ejemplo.

Ejemplo 4.1 (Analizando datos del ministerio de salud)**.** En el 2007 el Ministerio de Salud de Argentina realizó un estudio (ENNyS 2007) que entre otras recopiló datos sobre la estatura de las argentinas entre 19 y 49 años. La media fue de 161,01 centímetros con una desviación estándar de 6,99. Con estos datos podemos construir nuestra curva.

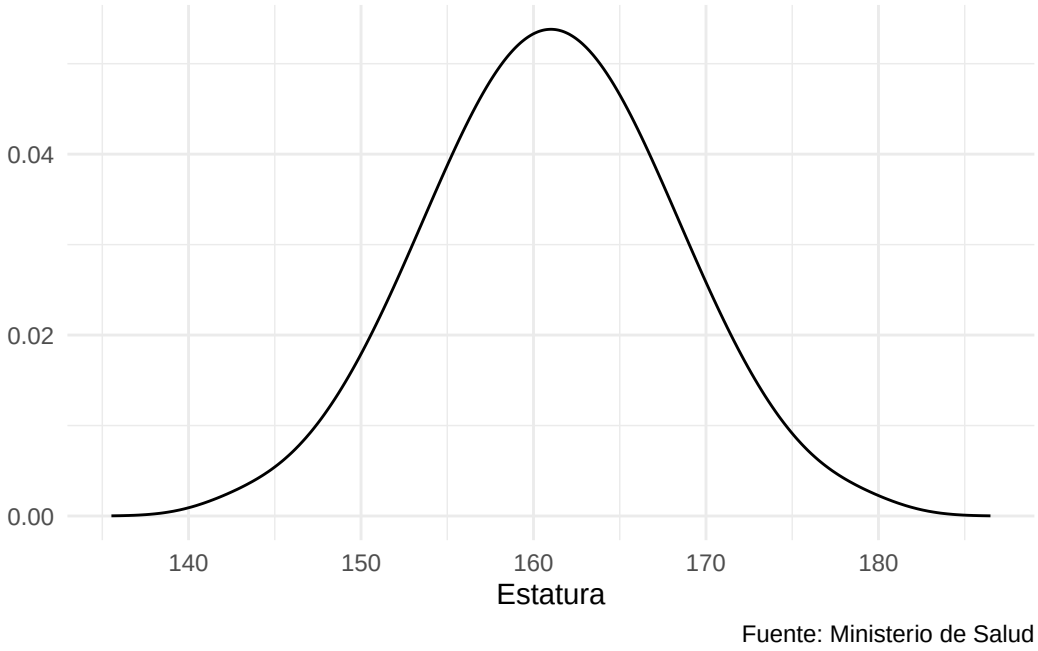


Figura 4.6: Estatura de argentinas entre 19 y 49 años

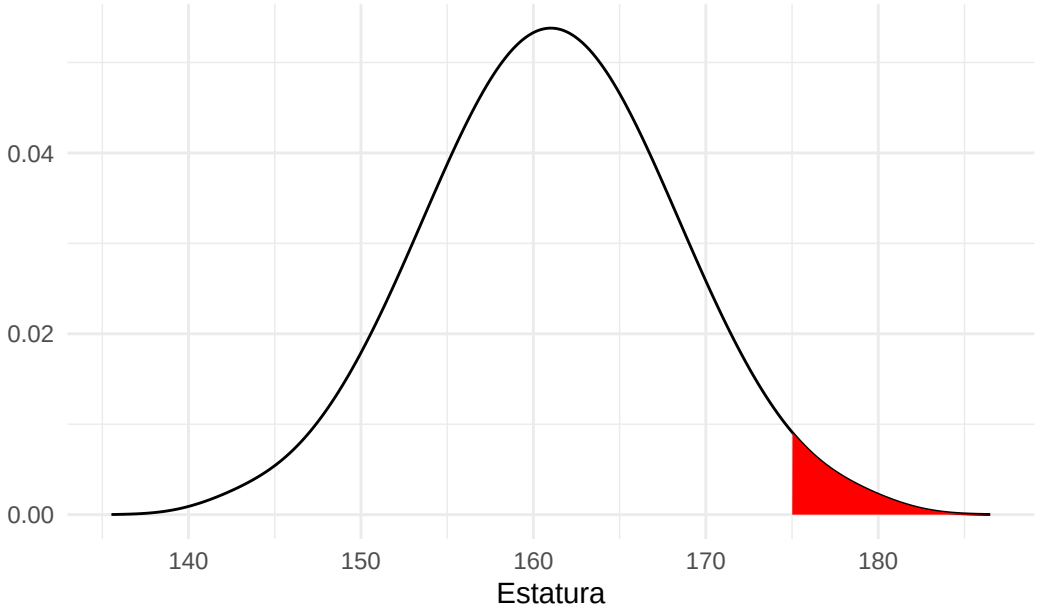
Ahora, sabiendo que esta variable tiene una distribución normal podemos saber que casi el 70% de las argentinas miden entre 154,04 y 168 centímetros. También podemos encontrar respuesta a una pregunta como: ¿qué proporción de la población femenina mide más que 175 centímetros? Para ello tenemos que normalizar el dato así:

$$z = \frac{175 - 161,01}{6.99} = \frac{13,99}{6.99} = 2,001$$

Con este número podemos volver a la @fig-normal-curve-with-cuts y fijarnos que con por arriba de 2 desviaciones estándar (o 2σ) está el 2,2% de la población. Es el área indicado en rojo en la figura @fig-curva-con-segmento.

```
#| label: fig-curva-con-segmento
#| echo: false
#| fig-cap: "Proporción de argentinas que miden más de 175 centímetros"


plot_estatura+
  geom_area(position = "identity", data = estatura %>% filter(x>175), fill='red')
```



Fuente: Ministerio de Salud

En este caso tuvimos un poco se suerte ya que la variable normalizada resultó un número redondo que era fácil encontrar en la figura Figura 4.5. Ahora digamos que queremos conocer la proporción de la población que mide menos de 150 centímetros, ¿cómo hacemos? Primero normalizamos:

$$z = \frac{150 - 161,01}{6.99} = \frac{11,01}{6.99} = -1,575$$

Con este número podemos sacar la proporción por ejemplo calculando el área debajo del segmento de la curva con cálculos integrales, lo podemos buscar en una tabla de probabilidades o podemos recurrir a la función `pnorm` (p: probabilidad, norm: normal)de R así:

```
pnorm(-1.575)
```

```
#> [1] 0.05762822
```

entonces el 5,76% de la población de argentinas entre 19 y 49 años miden menos de un metro con cincuenta.

También podemos expresar esto en términos de probabilidades: Si medimos una mujer argentina de entre 19 y 49 años seleccionada aleatoriamente de la población, la probabilidad de que mida menos de 150 centímetros es de 5,76% (p=0,0576).

4.3 Evaluar la normalidad

Hemos visto que el hecho de que una variable tenga una distribución normal nos resulta muy útil para extraer información sobre sus propiedades. También nos permite realizar algunos tests estadísticos que veremos en capítulos posteriores.

En la [sección @sec-cual-usar] decidimos usar *la media* como medida de centralización porque las tres medidas disponibles –media, mediana y moda– se aproximaban unas a otras. Si queremos saber si una variable se aproxima a la curva normal podemos generar un histograma y sobreponer una curva normal. Así podemos sacar alguna conclusión inspeccionando el gráfico.

También podemos valernos del conocimiento de la proporción de observaciones que deben estar dentro de la primera y segunda desviación estándar y verificar si nuestros datos se conforman con estas predicciones.

Ejemplo 4.2 (Notas de dos cursos)**.**
Si tomamos nuestros datos de las notas de nuestros dos cursos que vimos en la sección 1.4 y que fuimos desarrollando a lo largo de los capítulos anteriores podemos realizar este análisis.

Grupo A: {15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17, 13, 14, 13, 15, 17, 19, 17, 18, 16, 14}

Grupo B: {11, 16, 14, 18, 6, 8, 9, 14, 12, 12, 10, 15, 12, 9, 13, 16, 17, 12, 8, 7, 15, 5, 14, 13, 13, 12, 11, 13, 11, 7}

- Grupo A:
  - Media: 14.93
  - Desviación estándar: 2,49
  - Entre ± 1 desviación: 66%
  - Entre ± 2 desviaciones: 96%
- Grupo B:

- Media: 11,76
- Desviación estándar: 3,31
- Entre ± 1 desviación: 66%
- Entre ± 2 desviaciones: 96%

Observamos que nuestras notas carecen en cierta medida de valores extremos, sin embargo la muestra es relativamente pequeña con lo cual nos conformamos con estos resultados y consideramos normales las distribuciones.

Ejemplo 4.3 (Ejemplo en R)**.**
Si no queremos hacer estos cálculos a mano los podemos hacer también en R, así:

```
grupo.A = c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17,

media= mean(grupo.A)
desviacion = sd(grupo.A)
N = 30
sum(
    grupo.A <  media + desviacion
    &
    grupo.A > media - desviacion
)/N
```

```
#> [1] 0.6666667
```

```
sum(
    grupo.A <  media + desviacion * 2
    &
    grupo.A > media - desviacion  * 2
)/N
```

```
#> [1] 0.9666667
```

Existen también tests más formales de normalidad que desarrollaremos en capítulos posteriores.

4.4 Glosario

Regla empírica Cuando la distribución es normal el 68% de las observaciones se encuentran entre $\pm$ una desviación estándar de la media, el 95% entre dos desviaciones estándar y el 99,7% entre tres. Equivalente en inglés: «Empirical rule».

Variable normalizada Variable expresada en desviaciones estándar Fórmula: $z = \frac{x-\bar{x}}{\sigma}$ o $z = \frac{x-\bar{x}}{s}$ Función relevante en R: `scale`. Equivalente en inglés: «z-score».

5 Estimación de parámetros

Hemos visto que si trabajamos con poblaciones que son potencialmente infinitas o muy grandes usamos muestras para nuestro trabajo cuantitativo. Las medidas que calculamos en base a estas muestras son *estimativos* de los parámetros de la población. Si tenemos una muestra de estatura de argentinas entre 19 y 49 años de edad, como la que vimos en el ejemplo 4.1, no sabemos con certeza cuál es la media de la población. La estimamos en base a una muestra. Con ello no podemos afirmar que la media es la misma para la población, de hecho *ignoramos* cuál es la media de la población. Lo que sí podemos calcular un intervalo de valores dentro de los cuales tenemos cierta confianza de que nuestro valor estimativo sea correcto para la población.

En este capítulo desarrollaremos las técnicas que se utilizan para arribar a estos intervalos de confianza y calcular un *margen de error*.

5.1 Distribución muestral

Si suponemos que la estatura promedio de las argentinas entre 19 y 49 años es de 161 centímetros con una desviación estándar de 6,99, estos serían los *parámetros* de la población. Si sacamos cinco muestras aleatorias de veinte observaciones de esta población van a arrojar resultados distintos a estos valores. Algunas muestras van a tener una media por arriba de la media real y otras van a tener una media por debajo.

Ejemplo 5.1 (Distribucion de muestras)**.**

Como lo podemos observar en la figura Figura 5.1 la distribución de las muestras es simétrica y normal. La media de nuestras muestras es 161,42; ligeramente por arriba de la media real, y la desviación estándar es de 6,17; más de medio centímetro por debajo de la desviación estándar de la población. La distribución muestral tiene algunas propiedades que son útiles para nuestro trabajo estadístico:

1. Se aproxima a una distribución normal. Esto se conoce como el *teorema del límite central*.
2. La media de la distribución es igual (o casi igual) a la media de la población.
3. La dispersión es *menor* a la de la población general.

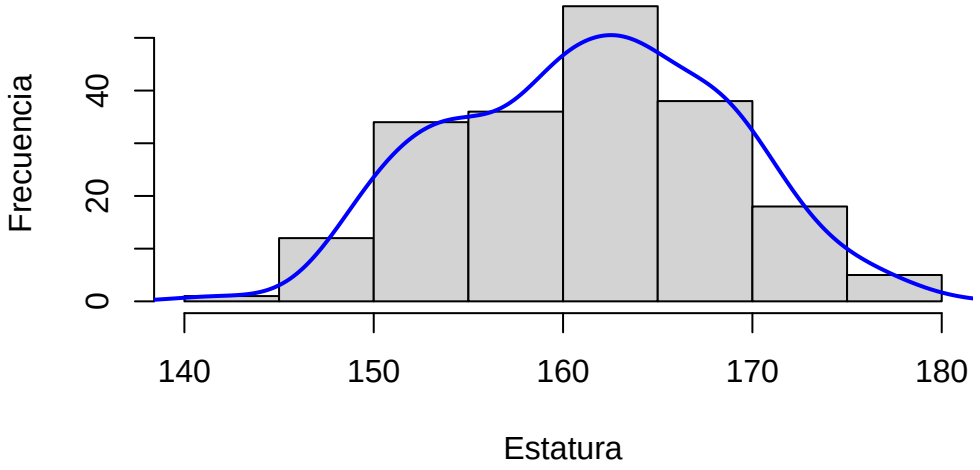


Figura 5.1: Cinco Muestras de 20 obseraciones

El número (3) de la lista tiene su lógica ya que en una muestra aleatoria un valor frecuente tiene más probabilidad de ser seleccionada que un valor extremo. La diferencia entre curva normal de la población y la curva de la distribución muestral está ilustrada en la figura 5.2.

5.2 El error estándar y su interpretación

La variabilidad de las medias muestrales se puede medir por su desviación estándar. Esta medida se conoce como el *error estándar* y tiende a disminuir cuando aumenta el tamaño de la(s) muestra(s).

Definición 5.1 (Error estandar)**.**

$$SE = \frac{\sigma}{\sqrt{N}}$$

si conocemos la desviación estándar de la población, y

$$SE = \frac{s}{\sqrt{N}}$$

si usamos la desviación estándar de la muestra.

donde:

- SE: el error estándar (por sus siglas en inglés «Standard Error»)
- σ: la desviación estándar de la población

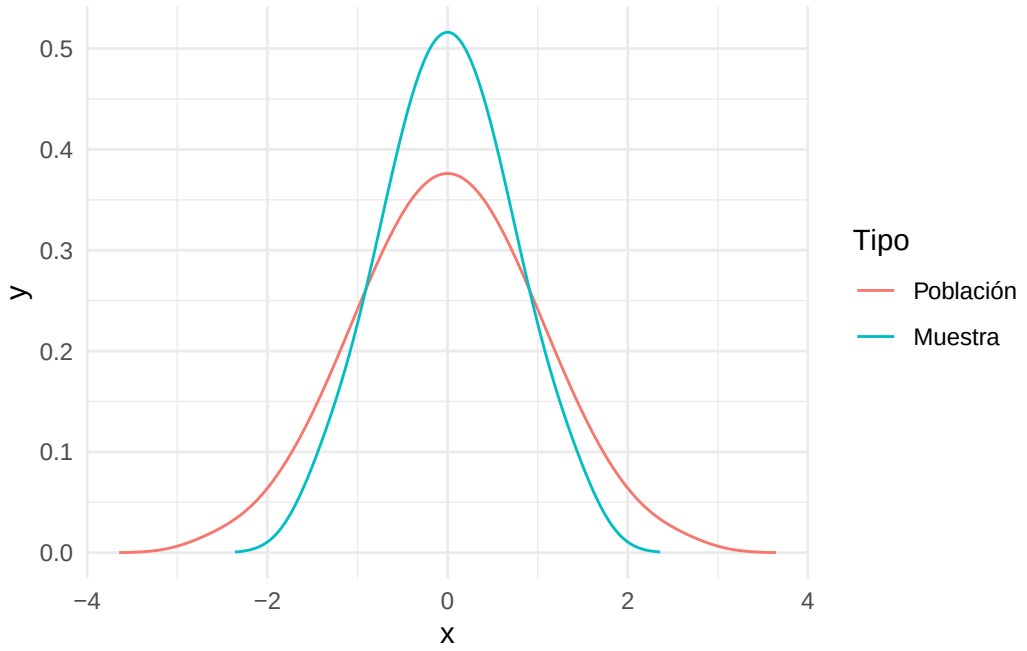


Figura 5.2: Distribución de la población y la muestra

- s: desviación estándar de la muestra
- N: número de observaciones de la muestra

Nótese que el error estándar no disminuye en relación directamente proporcional con el tamaño de la muestra. Ya que tomamos la raíz cuadrada de N, es necesario cuadruplicar el tamaño de la muestra para reducir el error estándar a la mitad.

5.2.1 Intervalos de confianza

Volvemos a nuestro ejemplo de la estatura de las argentinas entre 19 y 49 en 2007. Si sacamos una muestra aleatoria de esta población de tan solo 30 observaciones. de manera que:

Muestra = {163, 171, 171, 167, 164, 160, 153, 176, 162, 171, 166, 164, 169, 160, 151, 155, 156, 147, 162, 170, 164, 160, 158, 159, 157, 159, 156, 162, 159, 174}

podemos calcular la media y la desviación estándar de la muestra. Obtenemos $\bar{x} = 160,94$ y s = 6,89 respectivamente. Con esto podemos calcular el error estándar:

$$SE = \frac{s}{\sqrt{N}} = \frac{6,89}{\sqrt{30}} = \frac{s}{5,477} = 1,257$$

Ahora podemos estimar que la media de la población es de 160,94 ± 1,257. Hemos reportado muestra estimación con un margen de error. Pero ¿cómo se interpreta este número?

Sea μ la media real –por convención se usa la letra griega μ que corresponde a m para la $m$edia de la población. La desviación de la media de la muestra entonces es de $161,94 - \mu$. Podemos normalizar esta variable por división con la desviación estándar de la muestra:

$$z = \frac{161,94 - \mu}{1,257}$$

Recordemos que se usa z para la variable normalizada. Para muestras desde más o menos 30 observaciones, z tiene una distribución normal, con lo cual nos podemos valer de la *regla empírica* y mirar la figura 4.5 para darnos cuenta qué tan probable es que nuestro valor caiga dentro o fuera de los rangos esperados. El error estándar es, entonces, el rango de valores que caen dentro de una desviación estándar en la curva normal del error, es decir que hay un 68% de probabilidad de que el valor real esté dentro del rango reportado.

Podemos valernos de esta información para calcular rangos que nos den más confianza en nuestra estimación. La regla empírica dice que el 95% de las observaciones se encuentran entre dos desviaciones estándar de la media. Si se expresa con un poco más de precisión es de 1,96. Este *número mágico* o *valor crítico* de usa mucho en los textos con análisis cuantitativo ya que se puede demostrar matemáticamente que:

$$\text{media de la muestra} \pm (1,96 \times SE)$$

es un estimado de la media de la población con un 95% de confianza.

De la misma manera tenemos:

$$\text{media de la muestra} \pm (2,58 \times SE)$$

que nos da un rango con 99% de confianza.

Entonces, para nuestra muestra de argentinas podemos decir que estimamos que la media de la población (μ) es:

- entre 160,94 y 162,20 con un 68% de confianza
- entre 159,73 y 164,66 con un 95% de confianza
- entre 158,94 y 165,44 con un 99% de confianza

5.3 La distribución t

En la sección anterior vimos que la razón:

$$z = \frac{\bar{x}}{SE}$$

tiene una distribución normal cuando la muestra tiene un tamaño grande. Cuando la muestra es relativamente pequeña, sin embargo, tiende a otra distribución llamada *la distribución t* y a veces *distribución t de Student*[1].

El valor de t se calcula de la misma manera que el error estándar, pero debido a las características de la distribución los valores críticos son distintos dependiendo de los *grados de libertad* (que el la mayoría de los casos es igual a N-1.)

Ejemplo 5.2 (Muestra pequeña)**.** Hacemos una muestra aleatorea de 15 argentinas y medimos su estatura, esta vez con precisión milimetrica y obtenemos:

X = {153,26; 158,81; 165,73; 159,85; 160,56; 166,69; 159,85; 148,07; 160,3; 173,02; 154,55; 145,52; 159,98; 158,22; 166,12 }

La media es de 159,36 y la desviación estándar de 7,125. Por tanto:

$$SE = \frac{s}{\sqrt{N}} = \frac{7,125}{\sqrt{15}} = 1,338$$

El valor crítico de t con 14 grados de libertad (N-1) es $\pm 2,145$.

$$2,145 \times SE = 2,145 \times 1,338 = 2,869$$

Por tanto, basado en esta muestra más chica podemos estimar que la media de la población es de 159,36 $\pm 2,869$ es decir entre 156,49 y 162,23 centímetros.

Del ejemplo 5.2 vemos que si bien logramos estimar la media de la población, el margen de error es más amplio que con una muestra más grande.

[1] Por el seudónimo del matemático que primero publicó sobre este tema.

¿Dónde obtenemos los valores críticos de t?

Se pueden consultar los valores críticos de la distribución t para distintos grados de libertad en tablas estadísticas, como el del apendix **??** o en linea. También se puede sacar con una función en R llamada `qt`.

Ejemplo 5.3 (Ejemplo en R: Extraer el valor crítico de t)**.**

```
qt(p = 0.025, df = 14)
```

```
#> [1] -2.144787
```

La función toma dos argumentos `p` de qué proporción de la curva en cada lado queremos y `df` que son los grados de libertad, en este caso 15-1=14. Ponemos el valor de `0.025` porque queremos un 2,5% de arriba y un 2,5% de abajo (=5%).

5.4 Glosario

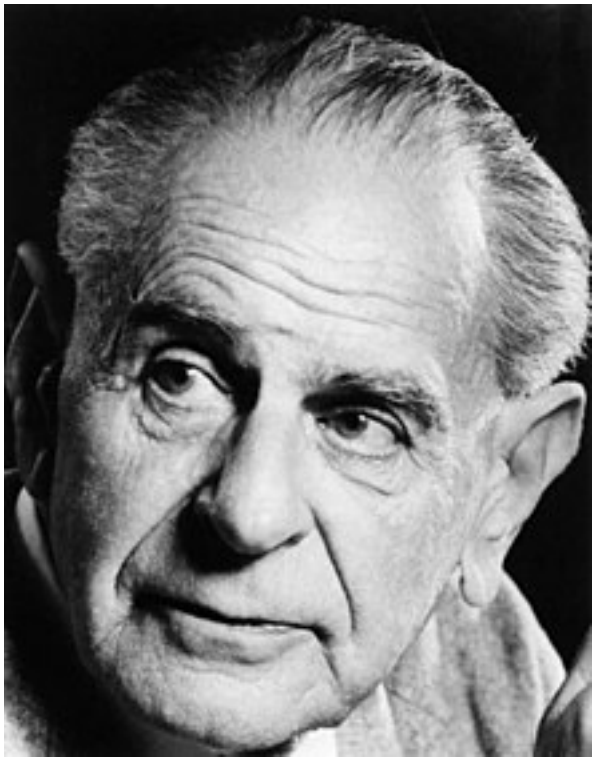
Distribución muestral El resultado de todas las muestras posibles que pueden ser tomadas de una población Equivalente en inglés: «Sample distribution».

Distribución t Distribución de probabilidad de una muestra pequeña de una distribución normal. Función relevante en R: `qt`. Equivalente en inglés: «T distribution».

Error estándar La desviación estándar de la distribución muestral. Fórmula: $SE = \frac{\sigma}{\sqrt{N}}$ o $SE = \frac{s}{\sqrt{N}}$ Equivalente en inglés: «Standard error».

Intervalo de confianza Intervalo dentro del cual estimamos que se encuentre un valor buscado, con cierto porcentaje de confianza. Equivalente en inglés: «Confidence Interval».

6 Diseño de proyectos y test de hipotesis



> Our knowledge can only be finite, while our ignorance must necessarily be infinite.
>
> —Karl Popper

6.1 El método científico

El filósofo de ciencias *Karl Popper* es considerado por muchos como el padre de la teoría científica moderna. Sostiene que la ciencia avanza proponiendo teorías e ideas que sean empíricamente refutables. Cuando una teoría es refutada por investigaciones empíricas surgen otras teorías que toman en cuenta las refutaciones de las anteriores y son sometidos al mismo proceso. Este tipo de pensamiento se conoce también como pensamiento «deductivo» y es

fundamental para las ciencias empíricas. Deductivo, en este caso es contrario a «inductivo» toma como *confirmatorias* toda observación que sostiene una teoría, sea esta falseable o no.

Por ejemplo, podemos proponer la teoría de que «en la Ciudad de Buenos Aires nunca cae nieve». Durante la segunda mitad del siglo pasado podíamos corroborar nuestra teoría día tras día al medir el nivel de nieve –que era cero–, pero bastó con una sola nevada, que ocurrió el nueve de julio del 2007, para que nuestra teoría quedara refutada.

Para que una teoría sea científica tiene que ser posible demostrar su falsedad empíricamente, es decir tenemos que poder obtener datos, por experimentos u observaciones, capazes de comprobar que la teoría no es correcta. Esto se llama el principio de falsabilidad,[1] falsación o refutabilidad.

Nuestra teoría de la falta de nieve en la ciudad de Buenos Aires es científica, ya la podemos hacer medidas para refutarla. El hecho de que la teoría resultó incorrecta no implica que no sea científica.

Cabe mencionar que existen ciencias: las ciencias «formales» como las matemáticas, la lógica formal etcétera; cuyas teorías no dependen de observación empírica ya que se concentran en el estudio abstracto de cantidades, estructuras y cambio.

6.2 El diseño de una investigación

Estudios experimentales y observacionales

Podemos distinguir entre dos tipos de investigación científica en las ciencias empíricas. Los estudios *experimentales* son estudios donde nosotros manipulamos alguna variable para darnos cuenta qué efecto tiene. En nuestro ejemplo de dos cursos con metodologías distintas, nosotros hemos manipulado la variable *metodología*. Como hemos mencionado antes en la sección 1.5 esta es la variable *independiente*. Los estudios experimentales son muy frecuentes en las ciencias naturales y también se aplican a las ciencias humanas.

En las ciencias humanas, sin embargo, a menudo nos encontramos con datos en los que no podemos manipular la variable independiente. En el caso de los datos lingüísticos de la figura Figura 2.2, no podemos *cambiar* el largo de las palabras. Nos tenemos que limitar a recoger los datos e intentar discernir alguna relación entre ellas. Igual tenemos una variable dependiente «largo de palabra» y una independiente «frecuencia», solo que no controlamos la variable independiente. Este tipo de estudios son *observacionales* y a veces se habla de estudios *correlacionales*.

[1] posiblemente por su cognado en ingles:«falsifiabilidy»

Fuentes de ruido en los datos

Cuando estamos haciendo un estudio experimental controlamos no solo la variable independiente, sino también podemos diseñar el experimento para minimizar el efecto de otras variables que puedan influir en la variable dependiente. La meta es de minimizar los efectos provinientes de de factores que no son relevantes para nuestro estudio a fin de poder afirmar con más confianza que los efectos observados en realidad tienen que ver con la variable independiente. En el caso de los dos grupos con metodologías distintos podemos, por ejemplo, asegurarnos de que los dos cursos tengan el mismo profesor, se dicte en horarios similares y que los de estudiantes que reciben el curso tengan características similares en cuanto a edad, género, promedio de notas en otras materias etcétera. En el caso de un estudio observacional no tenemos este nivel de control, lo que sí podemos hacer es intentar estimar el efecto de interferencia de otras variables y tomarlo en cuenta en nuestros análisis.

Incluso en el caso de un estudio experimental, no es realista esperar que podemos remover totalmente el efecto de variables irrelevantes. A lo que podemos aspirar y debemos intentar, sin embargo, es de remover la mayor cantidad posible de *variación sistemática*. Si en el caso del las notas de los grupos de estudiantes pusimos todos los hombres en un grupo y todas las mujeres en otro, no podemos saber si la diferencia que observamos se debe la diferencia de metodología didáctica o si es una diferencia de género. Por ello es preciso hacer lo posible para que las variables que son irrelevantes para nuestra investigación operen de manera aleatoria en nuestras muestras y, de ser posible, minimizar su efecto. Si no operan de manera aleatoria corremos el riesgo de en realidad medir otra variable –género en lugar de metodología– de la que queremos investigar, y si su efecto genera mucha varianza va a bajar la confianza que podemos tener el las conclusiones obtenidas.

6.3 Tests de hipótesis

Vimos en la sección 5.2 que podemos estimar los valores de la población en base a muestras y que podemos calcular un margen de error y niveles de confianza de estas estimaciones. Podemos valernos de los mismos conceptos para concluir algo sobre la relación entre variables: independiente y dependiente por ejemplo.

6.3.1 Tests estadísticos de significanza

En el caso de nuestros dos grupos de estudiantes (véase: 6.3.6) ya vimos que existe una diferencia entre los dos grupos en la media de la nota obtenida. De la figura 2.1 vimos que igual las dos distribuciones de solapan en gran medida. Por tanto no podemos afirmar con absoluta certeza que las diferencias observadas son el efecto de la metodología pedagógica aplicada o si son producto de la inherente variabilidad de las muestras.

El objetivo de un test estadístico de significanza es determinar si las diferencias observadas el resultado de variación aleatoria o si pueden razonablemente ser atribuidos a la variable independiente.

6.3.2 La hipótesis nula y alternativa

Para testear una hipótesis el primer paso es establecer una *hipotesis nula*. Esta hipótesis afirma que *no existe el efecto* que estamos investigando. Siguiendo los lineamientos del método cientifico, ahora nuestra labor es, a través de mediciones u observaciones, *refutar* esta hipótesis, con lo cual podemos proponer otra, llamada *hipótesis alternativa*. Una *hipótesis nula* se formula como una afirmación precisa y empiricamente refutable. En el ejemplo de los dos grupos de estudiantes la hipotesis nula podría expresarse como: «No existe diferencia entre la media de notas entre los dos grupos».

También debemos formular una o dos hipotesis alternativas. Si formulamos dos, una va a afirmar que la media de notas del grupo A es superior a la del grupo B y la otra que la media de notas del grupo B es mayor a la media de notas del grupo A. Si usamos una sola hipótesis alternativa esta simplemente plantea que la media notas de los dos grupos es desigual.

Notación formal

En notación formal, muy frecuente en textos académicos, se usa la letra H (mayuscula) para significar una hipótesis y tiene subindice «0» o «null». Las hipótesis alternativas reciben subindice numérica (1 y 2 etcétera). En el caso descrito en la sección anterior se podría expresar así:

H_0 : No hay diferencia entre los grupos

H_1 : Hay diferencia

o, includo más formal:

$H_0 : \mu_A = \mu_B$

$H_1 : \mu_A \neq \mu_B$.

La estrategia del test de hipótesis acumular evidencia empírica que nos permita *refutar* la hipótesis nula y no intentar fomentar cualquiera de las alternativas directamente. Lo que temenos que hacer es aplicar un test estadístico y calcular la probabilidad de obtener las observaciones que hemos obtenido y si esa probabilidad es muy baja, refutamos H_0 a favor de una de las alternativas.

Es preciso aclarar que nunca podemos estar absolutamente seguros de estar justificados en refutar H_0. Siempre existe la posibilidad de que las diferencias observadas de deban a la

aleatoriedad de las muestras. Lo que sí podemos mostrar es que la probabilidad de que así sea es muy baja.

6.3.3 Niveles de significanza

Dado que siempre existe la posibilidad de refutar injustificadamente nuestra H_0, tenemos que determinar un nivel debajo del cual estamos dispuestos a equivocarnos en nuestra afirmación. Este se llama el *nivel de significanza*, también se describe con la letra griega α y se llama nivel-α (nivel alfa). El nivel de significanza está coneptual y matemáticamente ligado con los [intervalos de confianza] que vimos en el capítulo @sec-estimacion-de-parametros.

Si estamos dispuestos a rechazar H_0 si la probabilidad (p) de hacerlo injustificadamente es igual o menor a 0,05, eligimos un nivel de significanza de 0,05, también llamado «nivel de 5%». Su notación a menudo se encuentra como: $p \leqslant 0,05$. Este nivel es bastante común en las ciencias humanas, en cambio en otras disciplinas de las ciencias exactas y médicas por ejemplo, a veces se opera con $p \leqslant 0,01$ o $p \leqslant 0,001$, lo que significa que se acepta rechazar injustificadamente H_0 una vez en cien o una vez en mil respectivamente.

Para cada test estadístico y cada nivel de significanza eligido existirá un valor crítico o un *rango crítico* dentro del cual el valor del cálculo estadístico tiene que encontrarse para que las diferencias observadas en las muestras se consideren estadísticamente significativos. Si el valor del test estadístico no cae en ese rango no podemos rechazar H_0 sobre la base este conjunto específico de observaciones, pero es posible que debamos repetir el estudio con muestras más grandes.

6.3.4 Tipos de error

Cuando tomamos la decisión de rechazar o aceptar la hipótesis nula hay dos errores que podemos cometer. Podemos rechazar H_0 cuando H_0 es correcta, o podemos aceptar H_0, cuando es falsa. En el primer caso estamos hablando de un *error de tipo I*, también denominado *error de tipo* α o *falso positivo*. En el segundo caso hablamos de un *error de tipo II*, *error de tipo* β (beta) o falso negativo.

6.3.5 Tests direcionales y no direcionales

En la sección **??** propusimos una hipótesis nula y su alternativa:

Ejemplo 6.1 (Hipotesis nula y una alternativa)**.**

$H_0 : \mu_A = \mu_B$

$H_1 : \mu_A \neq \mu_B$.

H_1 se leería: «la media de A es desigual a la media de B». Este ejemplo 6.1 es de una predición *no direcional*. Es decir que no hemos tomado una posición a priori sobre si esperamos que las diferencias que observemos sean positivos o negativos.

A veces tenemos razones bien fundadas en creer que las diferencias, si las observamos, van a darse en una direción u otra. Si por ejemplo estamos midiendo la estaturas de muestras aleatorias de argentinas y argentinos podemos suponer de antemano que los hombres van a ser más altos que las mujeres ya que está comprobado que es así en otros países, hay razones biológicas etcétera. En ese caso podríamos formular una predicción direccional, lo cual significa que nuestra hipótesis alternativa es una sola y va en una dirección específica:

Ejemplo 6.2 (Hipotesis nula y una alternativa direccional)**.**

$H_0 : \mu_M = \mu_F$

$H_1 : \mu_M > \mu_F$.

La diferencia entre usar un test direccional o no direccional influye en los valores críticos de los diferentes tests. Si usamos un test direccional –y está justificado su uso, claro– disminuye el riesgo de cometer un error de tipo II. Está ilustrado en las figura 6.1: para un test no-direcional necesitamos un 2,5% *en cada extremo* de la curva para que sume 5%, en el test direccional «gastamos» todo el lado positivo.

Ejemplo 6.3 (¿cara o cruz?)**.**
Para desarrollar un poco más el concepto de test de hipótesis vamos a imaginarnos que estamos jugando a *cara o cruz*. Si tiramos una moneda hay un 50 y 50 de que salga cruz o cara. Tiramos la moneda y sale cara. La tiramos dos veces y sale dos veces cara. Tres veces – tres caras… y seguimos perdiendo.

¿En qué momento empezamos a sospechar que la moneda tiene dos caras?

Aún sin conocimientos matemáticos o de la teoría de la probabilidad empieza a obrar nuestra intuición –basada en nuestra experiencia que por su naturaleza es empírica.

Podemos formalizar el problema de la siguiente manera:

H_0: La moneda es honesta

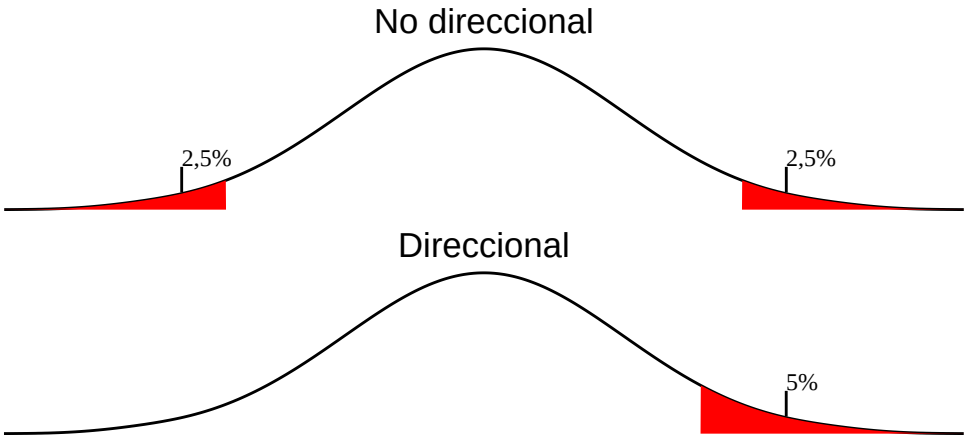


Figura 6.1: Test no direccionales y test direccionales

H_1: La moneda tiene dos caras.

Podemos también calcular las probabilidades de lo que está pasando. La probabilidad de que salga cara es 0,5 (50%) y de que salga cara dos veces es, por tanto, $0,5 \times 0,5 = 0,25$. Podemos calcular las probabilidades de varios casos más:

3 caras: $0,5 \times 0,5 \times 0,5 = 0,125$

4 caras: $0,5 \times 0,5 \times 0, \times 0,55 = 0,0625$

5 caras: $0,5 \times 0,5 \times 0, \times 0,55 = 0,03125$,

y vemos que si sale cara cinco veces de cinco ya podemos rechazar nuestra H_0 con un nivel de significanza de 0,05 ($p \leqslant 0,05$).

6.3.6 ¿Qué test usar?

En los capítulos que siguen vamos a desarrollar algunos tests de significanza estadística: el test de z, el test de t de Student, Mann-Whitney U, χ^2, Wilcoxon y sign-test. La elección de cuál de ellos usar en un caso específico dependerá de:

1. Escala de medición de las variables
2. Las características de su distribución
3. Si las muestras son correlacionadas o no,

y los iremos detallando en cada caso.

6.3.7 Procedimiento

El diseño de una investigación cuantitativa se puede resumir en estos cuatro pasos:

1. Formular hipotesis nula y alternativa(s)
2. Decidir el nivel de significanza estadística
3. Eligir un test estadístico a utilizarse
4. Aplicar la estadística y decidir si rechazamos H_0 o no.

6.4 Glosario

Error tipo I El error de rechazar H_0 cuando esta es correcta. Equivalente en inglés: «Type I error».

Error tipo II El error de no rechazar H_0 cuando esta es incorrecta. Equivalente en inglés: «Type II error».

Falsabilidad El hecho de que sea posible refutar una hipótesis, por medio de métodos empíricos. Equivalente en inglés: «Falsifiability».

Hipotesis alternativa Hipótesis a la que recurrimos si logramos refutar H_0. Fórmula: H_1 Equivalente en inglés: «Alternative hypothesis.».

Hipotesis núla La hipótesis que plantea que el patrón que estamos buscando *no existe*. A través de un estudio empírico intentaremos refutar esta hipótesis. Fórmula: H_0 Equivalente en inglés: «Null hypothesis (H_0)».

Método científico Metodología basada en la observación, medición y experimentación; y la formulación, análisis y modificación de hipótesis. Equivalente en inglés: «Scientific method».

Nivel de significanza La probabilidad de rechazar H_0 cuando esta es correcta. Fórmula: α Equivalente en inglés: «Alpha-level».

Test direccional Test estadístico en la que hypótesis alternativa de expresa en una dirección u otra. Equivalente en inglés: «Directional test».

Test no direccional Test estadístico en la que hypótesis alternativa de expresa sin dirección especificar dirección. Equivalente en inglés: «Non-directional test».

7 Pruebas paramétricas

En este capítulo vamos a desarrollar algunos técnicas estadísticos que nos permiten realizar una prueba o un test de diferencias entre medias de dos conjuntos de datos provenientes de muestras independientes o correlacionadas. Los tests que vamos a ver se llaman «paramétricos», lo cual quiere decir viene con algunas presunciones acerca de los datos:

1. Los datos son de escala de intervalo o razón
2. La población de la muestra debe aproximarse a una distribución normal
3. Las varianzas de las muestras debe aproximadamente similar[1]

Las pruebas estadísticas son las que nos permiten, a algún nivel de significanza, rechazar o aceptar la hipótesis nula (H_0), por lo que son de bastante utilidad en investigaciones cuantitativas.

7.1 Prueba t de Student para muestras independientes

Supongamos que tenemos dos muestras aleatorias e independientes con medias de $\bar{x_1}$ y $\bar{x_2}$ y que queremos saber si estas dos medias son signifacativamente distintas a un nivel de $p \leqslant 0,05$. Esto es lo mismo que decir que si afirmamos que hay una diferencia entre las muestras tenemos un 95% de probabilidad de tener razón. Lo que tenemos que calcular, entonces, es la probabilidad de que las dos muestras pueder provenir de la misma distribución y que la diferencia que vemos es por varianza en esa población. En otras palabras: queremos saber si dos muestras con la diferencia observada ($\bar{x_1} - \bar{x_2}$) podrían tener provenir de la misma población.

Si sacamos un número significativo de muestras de una misma población la media de estas muestra va a tener una diferencia con la media de la población, en algunos casos más altos y en otros más bajos. Usamos este conocimiento para calcular el error estándar:

$$SE = \frac{\sigma}{\sqrt{N}}. \tag{7.1}$$

De la misma manera existe un *error estándar de diferencias entre medias* (SED por sus siglas en ingles).

[1] Este requerimiento puede obviarse en algunos casos, sobre todo tenemos muestras grandes.

Definición 7.1 (Error estándar de diferencia entre medias)**.**

$$SED = \sqrt{\sigma_1^2/N_1 + \sigma_1^2/N_2}$$

donde:

- σ_1^2 y σ_2^2: las varianzas de las poblaciones 1 y 2
- N_1 y N_2: es el número de observaciones en cada muestra.

Al igual que con el error estándar, a menudo desconocemos la varianza de la población, por lo cual lo estimamos de la muestra y la formula es la que vemos en la definición 7.2

Definición 7.2 (Error estándar de diferencia entre medias estimado de muestras)**.**

$$SED = \sqrt{s_1^2/N_1 + s_1^2/N_2}$$

donde:

- s_1^2 y s_2^2: las varianzas de las muestras 1 y 2
- N_1 y N_2: es el número de observaciones en cada muestra.

Vimos en la sección 5.3que para muestras relativamente pequeñas (N<30) la distribución de la muestra tiende a la distribución *t de Student*. Podemos valernos de esto para calcular la probabilidad de que nuestro *SED* esté en el rango requerido aplicando la formula de la definición.

Definición 7.3 (Prueba de t)**.**

$$t = \frac{(\bar{x_1} - \bar{x_2})}{SED}.$$

Si aplicamos la fórmula de la definición 7.3 nos sale un valor que podemos comparar con los valores críticos de la tabla del apendice A para determinar si rechazamos H_0 o no.

Ejemplo 7.1 (Prueba t)**.**

Volvemos ahora a nuestros datos de notas de dos grupos de estudiantes con diferentes metodologías pedagígicos. Queremos saber con un nivel de significanza de 0,05 si existe diferencia entre la media de los dos grupos. Nuestras hipótesis nula y alternativa son entonces:

$H_0 : \mu_A = \mu_B$,

$H_1 : \mu_A \neq \mu_B$.

Los datos son:

Grupo A: {15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17, 13, 14, 13, 15, 17, 19, 17, 18, 16, 14} y

Grupo B: {11, 16, 14, 18, 6, 8, 9, 14, 12, 12, 10, 15, 12, 9, 13, 16, 17, 12, 8, 7, 15, 5, 14, 13, 13, 12, 11, 13, 11, 7}.

La media y desviación estándar:

Grupo A:
$\bar{x_A} = 14,933$
$s = 2,490$
$N = 30$

Grupo B:
$\bar{x} = 11,77$
$s = 3,308$
$N = 30$.

Aplicando la fórmula de la definición 7.2 obtenemos:

$$SED = \sqrt{s_1^2/N_1 + s_1^2/N_2} = \sqrt{2,490^2/30 + 3,308^2/N_2} = 0,756$$

y podemos calcular el valor de t aplicando la fórmula de la definición 7.3

$$t = \frac{\bar{x_1} - \bar{x_2}}{SED} = \frac{14.933 - 11,766}{0,756} = 4,188.$$

Si buscamos este valor en el Apendix A para 29 grados de libertad (N-1), vemos que debemos rechazar H_0 y concluir que existe una diferencia estadísticamente significativa entre las dos muestras. Tenemos razón de creer que el método pedagógico influye en los resultados finales de los estudiantes.

Ejemplo 7.2 (Prueba t en R)**.**

Si no queremos hacer todos estos cálculos a mano podemos hacerlos en R usando la función `t.test`. Toma como parámetros las dos muestras que queremos comparar.

```
Grupo.A = c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17,
Grupo.B = c(11, 16, 14, 18, 6, 8, 9, 14, 12, 12, 10, 15, 12, 9, 13, 16, 17, 12, 8, 7, 15,
t.test(Grupo.A,Grupo.B)
```

```
#>
#>  Welch Two Sample t-test
#>
#> data:  Grupo.A and Grupo.B
#> t = 4.1887, df = 53.88, p-value = 0.0001046
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  1.650905 4.682428
#> sample estimates:
#> mean of x mean of y
#>  14.93333  11.76667
```

Vemos que el test nos devuelve además un valor de p más preciso.

7.2 Prueba de Shapiro-Wilks

En la sección 4.3 mencionamos que existen algunas maneras de estimar si una variable tiene una distribución normal o no. Nos basamos sobre todo en la forma de los polígonos de frecuencias (figura 2.1). Ahora vamos a introducir un test más formal de normalidad.

El test de *Shapiro-Wilks* plantea la hipótesis nula que una muestra proviene de una distribución normal. Eligimos un nivel de significanza, por ejemplo 0,05, y tenemos una hipótesis alternativa que sostiene que la distribución no es normal.

Tenemos:

H_0: La distribución es normal

H_1: La distribución no es normal,

o más formalmente aún:

$H_0 : X \sim \mathcal{N}(\mu, \sigma^2)$

$H_1 : X \nsim \mathcal{N}(\mu, \sigma^2)$.

Ahora el test Shapiro-Wilks intenta rechazar la hipotesis nula a nuestro nivel de significanza. Para realizar el test usamos la función `shapiro.test` en R:

Ejemplo 7.3 (Test de Shapiro Wilks en R)**.**

```
Grupo.A = c(15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17,

shapiro.test(Grupo.A)
```

```
#>
#>  Shapiro-Wilk normality test
#>
#> data:  Grupo.A
#> W = 0.97032, p-value = 0.548
```

```
Grupo.B = c(11, 16, 14, 18, 6, 8, 9, 14, 12, 12, 10, 15, 12, 9, 13, 16, 17, 12, 8, 7, 15,


shapiro.test(Grupo.B)
```

```
#>
#>  Shapiro-Wilk normality test
#>
#> data:  Grupo.B
#> W = 0.97636, p-value = 0.7227
```

Vemos que en ambos casos el valor de probabilidad (p) es muy superios a nuestro nivel elegido (0,05), por lo que *no rechazamos la hipótesis nula.*

En el caso de los ejemplos 7.1 y 7.2 ya obramos bajo la premisa de que las variables tenían una distribución normal, pero generalmente conviene realizar el test Shapiro-Willks *antes* de decidir qué prueba estadística vamos a usar. Si rechazamos H_0, es decir si no concluimos que la distribución sea normal, no deberíamos usar un test paramétrico.

7.3 Prueba de Fisher

Al inicio del capítulo también vimos que uno de los requisitos para que una prueba estadística paramétrica sea válida es que las varianzas sean de similar magnitud. Para ello también existe un test, el *test de Fisher*[2] que plantea las hipótesis:

$H_0 : \sigma_1^2 = \sigma_2^2$,

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Sin entrar en mucho detalle teórico, en R hay una función `var.test` para este propósito. La función toma dos argumentos: los dos conjuntos de datos que queremos comparar.

Ejemplo 7.4 (Realizar la prueba de Fisher en R)**.**

```
var.test(Grupo.A, Grupo.B)
```

```
#>
#>  F test to compare two variances
#>
#> data:  Grupo.A and Grupo.B
#> F = 0.56675, num df = 29, denom df = 29, p-value =
#> 0.1321
#> alternative hypothesis: true ratio of variances is not equal to 1
#> 95 percent confidence interval:
#>  0.2697517 1.1907335
#> sample estimates:
#> ratio of variances
#>          0.5667472
```

Vemos que el valor de probabilidad p es mayor a nuestro nivel de significanza ($p \leqslant 0,05$), con lo cual *no rechazamos* H_0 y concluimos que las varianzas son relativamente similares.

[2] tambien: «F-test»

7.4 Prueba t para muestras pareadas

En los ejemplos 7.1 y 7.2 teníamos dos grupos de estudiantes de dos cursos distintos, pero en muchos tenemos observaciones *pareadas* o datos *interdependientes*. Esto es muy típico de investigaciones experimentales en los que medimos la variable dependiente antes y después[3] de cambiar la variable independiente. Si, por ejemplo, queremos investigar el efecto de la cafeína sobre el pulso sanguineo podríamos obtener una muestra de personas y tomarles el pulso antes y después de hacerles tomar una taza de café.

En este sacamos las diferencias entre las dos medidas y comparamos estas diferencias con la distribución teórica. La fórmula está en la definición 7.4.

Definición 7.4 (Prueba t para muestras dependientes)**.**

$$t = \frac{\bar{X}_D}{\frac{s_D}{\sqrt{n}}}$$

donde:

- $\bar{X}_D$: media de las diferencias
- s_D: la desviación estándar de las diferencias
- n: número de pares de observaciones.

Lo que nos va a decir la prueba t en este caso es si la diferencia es significativamente diferente a cero: Si la variable independiente no tiene efecto entonces debería dar lo mismo medir antes o después. Las hipótesis planteadas son, por tanto:

$H_0 : \bar{X}_D = 0,$

$H_1 : \bar{X}_D \neq 0.$

Ejemplo 7.5 (Prueba t dependiente)**.** En este ejemplo (Shier 2004) vamos a suponer que tenemos un grupo de veinte estudiantes y queremos investigar el efecto del uso de algún recurso didáctico, por ejemplo un video en YouTube, en su destreza para resolver cierto tipo de problemas matemáticos. Les tomamos un test inicial, pedimos que miren el video y cuando terminen tomamos otro test. Ahora tenemos dos observaciones de cada estudiante. Calculamos la diferencia entre ellos. El resultado de todo esto está resumido en la tabla 7.1.

[3] también se conoce como «medidas repetidas»

Cuadro 7.1: Resultados de dos tests de matemáticas

Nombre	Antes	Después	Diferencia
Manuel	18	22	4
Miguel	21	25	4
José	16	17	1
Antonio	22	24	2
Dolores	19	16	-3
Manuela	24	29	5
Pedro	17	20	3
Lucía	21	23	2
Cecilia	23	19	-4
Juan	18	20	2
Paula	14	15	1
Francisco	16	15	-1
Angel	16	18	2
Soledad	19	26	7
Luis	18	18	0
Cristina	20	24	4
Laura	12	18	6
Carlos	22	25	3
Carmen	15	19	4
Javier	17	16	-1

La media de las diferencias es 2.05 con una desviación estándar de 2,837. Entonces tenemos:

$$t = \frac{\bar{X}_D}{\frac{s_D}{\sqrt{n}}} = \frac{2,05}{\frac{2,837}{\sqrt{20}}} = 3,231.$$

Buscando este valor en la tabla de valores críticos con 19 (N-1) grados de libertad vemos que sí podemos rechazar la hipótesis nula y concluir que hay una diferencia estadísticamente significativa entre los resultados de los dos tests.

Ejemplo 7.6 (Ejemplo en R)**.**
Para reproducir en R lo que hicimos en el ejemplo 7.5 tenemos que tener sumo cuidado con el ingreso de los datos. Ya que hay dos observaciones por estudiante lo más conveniente es ponerlos en un `data.frame`. Vamos a incluir los nombres de los estudiantes, si bien no son necesarios para el cálculo sirve mantener la referencia para poder verificar el correcto ingreso de los datos con los tests. Vamos a ingresar los datos *a mano* aunque en la práctica seguramente

se leyera de un archivo externo de R. Usamos la función `t.test` con un paramertro adiciónal `paired=TRUE` para avisar que son datos pareados.

```
# Ingresamos los datos
 datos.pre.post = data.frame(
   Nombre = c('Luis', 'Javier', 'Pedro', 'Soledad', 'Manuel', 'Cecilia', 'Cristina', 'Ange
   Pre = c(18, 21, 16, 22, 19, 24, 17, 21, 23, 18, 14, 16, 16, 19, 18, 20, 12, 22, 15, 17)
   Post = c(22, 25, 17, 24, 16, 29, 20, 23, 19, 20, 15, 15, 18, 26, 18, 24, 18, 25, 19, 16
 )

# Verificamos la homogeneidad de varianzas
var.test(datos.pre.post$Pre,datos.pre.post$Post)
```

```
#>
#>  F test to compare two variances
#>
#> data:  datos.pre.post$Pre and datos.pre.post$Post
#> F = 0.60329, num df = 19, denom df = 19, p-value =
#> 0.2795
#> alternative hypothesis: true ratio of variances is not equal to 1
#> 95 percent confidence interval:
#>  0.238790 1.524186
#> sample estimates:
#> ratio of variances
#>          0.6032913
```

```
# Verificamos que los datos tienen distribución normal
shapiro.test(datos.pre.post$Pre)
```

```
#>
#>  Shapiro-Wilk normality test
#>
#> data:  datos.pre.post$Pre
#> W = 0.98197, p-value = 0.9569
```

```
shapiro.test(datos.pre.post$Post)
```

```
#>
#>  Shapiro-Wilk normality test
#>
#> data:  datos.pre.post$Post
#> W = 0.94235, p-value = 0.2654
```

```r
# Realizamos prueba t

t.test(datos.pre.post$Post,datos.pre.post$Pre, paired = TRUE)
```

```
#>
#>  Paired t-test
#>
#> data:  datos.pre.post$Post and datos.pre.post$Pre
#> t = 3.2313, df = 19, p-value = 0.004395
#> alternative hypothesis: true mean difference is not equal to 0
#> 95 percent confidence interval:
#>  0.7221251 3.3778749
#> sample estimates:
#> mean difference
#>            2.05
```

Vemos que el resultado tiene significanza estadística alta ($p \leqslant 0,01$). El cálculo de R también nos da un intervalo de confianza al 95%.

7.5 Prueba de z

Existe también una prueba, llamada *de z*, que se puede usar para muestras más grandes. Se basa en el hecho de que cuando las muestras son más grandes tienden a una distribución normal y no una distribución t. Aparte de eso su concepto y mecánica es similar a la de la prueba t. Puede aplicarse cuando las muestras tienen más de 30 (N>30) observaciones y la principal diferencia de que es capaz de detectar diferencias más pequeñas en los datos lo que reduce el riesgo de un error tipo II.

7.6 Resumen de procedimiento

La figura 7.1 despliega un diagrama de flujo para eligir un test estadístico inferencial.

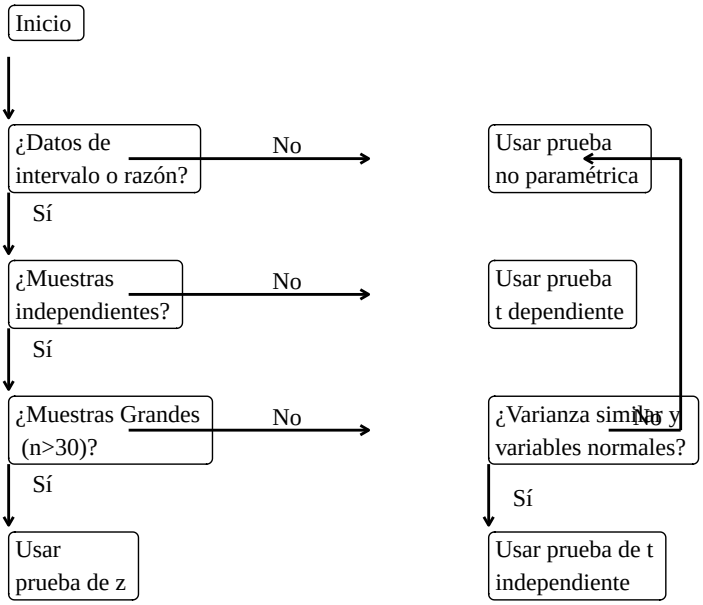


Figura 7.1: Diagrama de flujo para selección de estadística inferencial

7.7 Glosario

Error estándar de diferencias entre medias El error estándar calculado sobre la distribución de diferencias entre dos muestras. Fórmula: $SED = \sqrt{\sigma_1^2/N_1 + \sigma_1^2/N_2}$ Equivalente en inglés: «Standar error of differences (SED)».

Prueba de Fisher Prueba estadística que nos permite estimar si la varianza de dos muestras en similar. Equivalente en inglés: «Fisher test».

Prueba de Shapiro-Wilks Prueba estadística que nos permite estimar en qué medida una muestra proviene de una distribución normal. Equivalente en inglés: «Shapiro-Wilks Test».

Prueba de z Prueba estadística que nos permite estimar si dos muestras grandes (N>30 para ambas) provienen de poblaciones diferentes. Equivalente en inglés: «z-test».

Prueba t para muestras independientes Test estadístico que nos indica si la media de dos muestras tienen más diferencias de lo que se esperaría si son aleatorias. Fórmula: $t = \frac{(\bar{x_1} - \bar{x_2})}{SED}$ Función relevante en R: `t.test`. Equivalente en inglés: «T-test for independent samples».

Prueba t para muestras pareadas Test estadístico que nos indica si la media de dos muestras correlacionadas tienen más diferencias de lo que se esperaría por razones aleatorias. Fórmula: $t = \frac{\bar{X}_D}{\frac{s_D}{\sqrt{n}}}$ Función relevante en R: `t.test`. Equivalente en inglés: «T-test for dependent samples».

8 Pruebas no paramétricas

En el capítulo 7 vimos que para usar esas pruebas tenemos que cumplir con algunos requisitos sobre la distribución normal de las variables, nivel de medición y homogeneidad de las varianzas. Con alguna frecuencia, sin embargo, resulta que nuestros datos no cumplen con alguno de esos requisitos. Esto se puede dar por la naturaleza de la investigación, por ejemplo si estamos investigando un fenómeno que no se puede medir a escala de intervalo o razón; o tenemos relativamente pocos datos y luego de realizar los test de *Fisher* y *Shapiro* nos damos cuenta de que o la varianza es muy heterogénea o que las variables carecen de distribución normal.

Por suerte todavía hay esperanza. Existen algunas pruebas estadísticas, llamadas *no paramétricas* que nos pueden salvar en estos casos. En este capítulo desarrollaremos tres de ellos.

8.1 Prueba U de Mann-Whitney

La prueba U de Mann-Whitney resulta útil si tenemos dos muestras independientes y queremos si hay una diferencia en la magnitud de la variable que estamos estudiando, pero no podemos usar la prueba de t independiente o la prueba de z porque los datos no cumplen con alguno de los requisitos. Para realizar la prueba U de Mann-Whitney ponemos las observaciones de las dos muestras en orden ascendiente y asignamos un rango ordinal de manera que 1 corresponde a la observación de menor magnitud, 2 a la segunda etcétera. Luego nos fijamos en las diferencias entre las observaciones.

La prueba se basa en una comparación de cada observación de una muestra x_i con cada observación en la segunda muestra y_j. Si las muestras tienen la misma *mediana*, entones cada observación tiene un 0,5 (50%) de chance de ser mayor o menor que la observación correspondiente de la otra muestra. Por tanto plantea las hipotesis:

$H_0 : P(x_i > y_j) = \frac{1}{2}$

$H_1 : P(x_i > y_j) \neq \frac{1}{2}$

La prueba U de Mann-Whitney también se conoce con otros nombres: *Mann–Whitney–Wilcoxon*, *Wilcoxon rank-sum test* y *Wilcoxon–Mann–Whitney*. Por ello está disponible en R por medio de la función `wilcox.test`.

Ejemplo 8.1 (Prueba U de Mann-Whitney en R)**.**

En este ejemplo vamos a suponer que tenemos datos diagnósticos de cuatro mujeres y cinco hombres. Todos fueron diagnosticados con diabetes y tenemos la edad a la cual se les descubrió la enfermedad. Queremos saber si hay diferencia en la edad entre hombres y mujeres. Los datos son:

Hombres: {19, 22, 16, 29, 24},
Mujeres: {20, 11, 17, 12}.

```
Hombres = c(19, 22, 16, 29, 24)
Mujeres = c(20, 11, 17, 12)
wilcox.test(Hombres, Mujeres)
```

```
#>
#>  Wilcoxon rank sum exact test
#>
#> data:  Hombres and Mujeres
#> W = 17, p-value = 0.1111
#> alternative hypothesis: true location shift is not equal to 0
```

Vemos que no podemos rechazar H_0 en este caso.

8.2 Prueba de los rangos con signo de Wilcoxon

Vimos en la sección 8.1 que la prueba U de Mann-Whitney puede ser una alternativa a la prueba de t de Student para muestras intependiente (véase la sección 7.1) cuando los requisitos para un test paramétrico no se cumplen. Si los datos son pareados tenemos la *prueba de los rangos con signo de Wilcoxon* como alternativa a prueba t para muestras pareadas que vimos en la sección 7.4.

La lógica de la prueba de los rangos con signo de Wilcoxon es similar a la de la prueba de t pareada. Si no hay diferencia en el antes y despues, por ejemplo, las diferencias entre las observaciones deberían tender a cero.

Ejemplo 8.2 (Prueba de los rangos con signo de Wilcoxon en R)**.**

En este ejemplo vamos a suponer que tenemos un grupo de doce pacientes con artritis y les damos dos medicaciones distintas para aliviar los síntomas. Pedimos a todos que nos indiquen cuantas horas de alivio observaron con ambas drogas.

Los datos se observan en la tabla 8.1.

Cuadro 8.1: Eficiencia de dos medicamentos, reportada por los pacientes.

Paciente	Droga.A	Droga.B
1	2,0	3,5
2	3,6	5,7
3	2,6	2,9
4	2,7	2,4
5	7,3	9,9
6	3,4	3,3
7	14,9	16,7
8	6,6	6,0
9	2,3	3,8
10	2,1	4,0
11	6,8	9,1
12	8,5	20,9

En R ponemos los datos en un `data.frame`:

```
datos = data.frame(
  Paciente = c( 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12),
  Droga.A = c( 2,  3.6,  2.6,  2.7,  7.3,  3.4,  14.9,  6.6,  2.3,  2.1,  6.8,  8.5),
  Droga.B = c( 3.5,  5.7,  2.9,  2.4,  9.9,  3.3,  16.7,  6,  3.8,  4,  9.1,  20.9)
)
```

E iniciamos nuestros tests:

```
var.test(datos$Droga.A,datos$Droga.B)
```

```
#>
#>  F test to compare two variances
#>
#> data:  datos$Droga.A and datos$Droga.B
#> F = 0.41865, num df = 11, denom df = 11, p-value =
```

```
#> 0.1643
#> alternative hypothesis: true ratio of variances is not equal to 1
#> 95 percent confidence interval:
#>  0.1205199 1.4542635
#> sample estimates:
#> ratio of variances
#>          0.4186498
```

¡Bien! No tenemos problemas de varianza.

```
shapiro.test(datos$Droga.A)
```

```
#>
#>  Shapiro-Wilk normality test
#>
#> data:  datos$Droga.A
#> W = 0.80692, p-value = 0.01124
```

```
shapiro.test(datos$Droga.B)
```

```
#>
#>  Shapiro-Wilk normality test
#>
#> data:  datos$Droga.B
#> W = 0.7883, p-value = 0.006919
```

¡Ups!, las variables no tienen distribución normal. Entonces no podemos usar la prueba t pareada, tenemos que probar con *Wilcoxon*.

Usamos la función `wilcox.test` con el parametro extra de `paired = TRUE`.

```
wilcox.test(datos$Droga.A, datos$Droga.B, paired = TRUE)
```

```
#>
#>  Wilcoxon signed rank test with continuity correction
#>
#> data:  datos$Droga.A and datos$Droga.B
#> V = 8, p-value = 0.01669
#> alternative hypothesis: true location shift is not equal to 0
```

Vemos que el valor p se encuentra debajo de nuestro nivel de significanza ($\alpha = 0,05$), con lo cual rechazamos H_0 y concluimos que hay una diferencia estadísticamente significativa entre las dos mediamentos.

¿Y si usabamos la prueba t igual?

Si nos hubéramos olvidado de verificar la conformidad de los requisitos podríamos haber caído en la prueba t paramétrica, ¿qué hubiera pasado?

Veamos:

```
t.test(datos$Droga.A, datos$Droga.B, paired = TRUE)
```

```
#>
#>  Paired t-test
#>
#> data:  datos$Droga.A and datos$Droga.B
#> t = -2.1465, df = 11, p-value = 0.05498
#> alternative hypothesis: true mean difference is not equal to 0
#> 95 percent confidence interval:
#>  -4.28706458  0.05373125
#> sample estimates:
#> mean difference
#>       -2.116667
```

Podemos observar que la prueba de t es sensible a la falta de normalidad en nuestras variables y no logra rechazar H_0.

8.3 Prueba de signos

La prueba de Wilcoxon que vimos en la sección 8.2 requiere que los datos tengan una escala de medición (véase la sección 1.5) de intervalo. Pero a veces tenemos datos que solo se pueden medir a escala ordinal como por ejemplo la preferencia por alguna bebida de 1 a 5. En este caso no es razonable afirmar que la diferencia entre uno y dos es la misma que entre dos y tres, entonces no podemos tomar en cuenta la magnitud de esas diferencias.

La prueba de signos resuelve este problema convirtiendo la diferencia en una variable trinaria: puede ser cero, positiva o negativa. La lógica del test es similar a la de Wilcoxon, si no hay un patrón en las observaciones estas diferencias deberían tender a cero. Para realizar un test de signo debemos primero anotar el signo (positivo, negativo o cero) de todas los pares

de observaciones que tenemos. Cuando la diferencia es cero se excluye el par del análisis y reducimos N acorde a eso. Luego sumamos los positivos por un lado y los negativos por otro y tomamos el *menor* le los dos. Este número, a menudo significado por una W, de puede compara con la tabla de valores críticos para el N que quedó, que se puede consultar en el apendice B para N entre 5 y 25.

Cuando N es superior a 25, es decir cuando tenemos venticinco o más observaciones que no sean cero, se puede transformar W en una variable normalizada. Usando la fórmula en la definición 8.1.

Definición 8.1 (Normalizar W del test de signos)**.**

$$z = \frac{N - 2 \times W - 1}{\sqrt{N}}$$

8.4 Realizar prueba de sign para N>25

En este ejemplo vamos a suponer que hemos preguntado a 150 personas su opinión sobre el café de dos cafeterías: A y B, de la Ciudad de Buenos Aires. Les pedimos que indiquen en una escala de 1 a 5 cuánto les gusta cada producto. De ellos cincuenta dan el mismo ranking a ambos productos, con lo sus opiniones se eliminan del cálculo. De los restantes cien tenemos 39 que prefieren B y 61 que prefieren A. Tomamos el menor valor (39) y aplicamos la fórmula:

$$z = \frac{N - 2 \times W - 1}{\sqrt{N}} = \frac{100 - 2 \times 39 - 1}{\sqrt{100}} = \frac{21}{10} = 2,1$$

Recordamos que el *valor mágico* de la distribución normal –la regla empírica– es 1,96 para nuestro nivel de significanza ($p \leqslant 0,05$) y concluimos que existe una diferencia estadísticamente significativa.

8.5 ¿Cuál usar?

En la figura 8.1 podemos ver un diagrama de flujo para eligir un test no paramétrico.

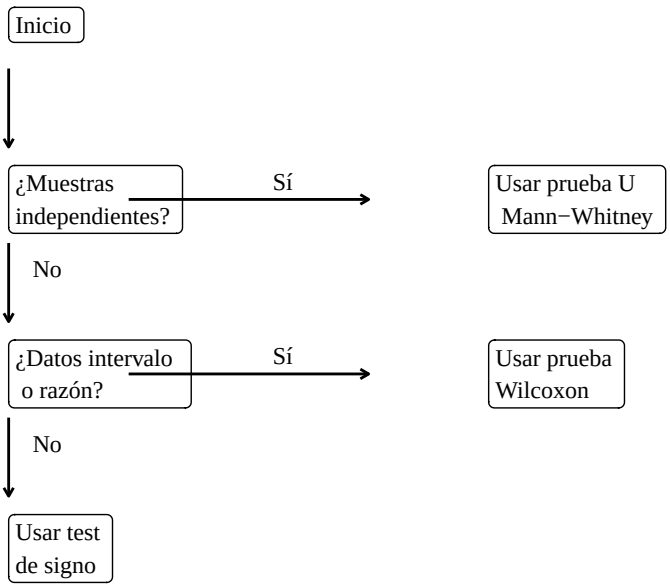


Figura 8.1: Diagrama de flujo para selección de pruebas estadísticas no paramétricas.

8.6 Glosario

Prueba de los rangos con signo de Wilcoxon Prueba estadística. Alternativa a la prueba t pareada, cuando los datos no cumplen con los requisitos para pruebas paramétricas. Función relevante en R: `wilcox.test`. Equivalente en inglés: «Wilcoxon ranked sign test».

Prueba de signos Prueba estadística. Alternativa a la prueba t pareada, cuando los datos son de escala ordinal. Equivalente en inglés: «Sign-test».

Prueba U de Mann-Whitney Prueba estadística. Alternativa a la prueba t o prueba z, cuando los datos no cumplen con los requisitos para pruebas paramétricas. Función relevante en R: `wilcox.test`. Equivalente en inglés: «Mann-Whitney U test».

9 Prueba de χ^2

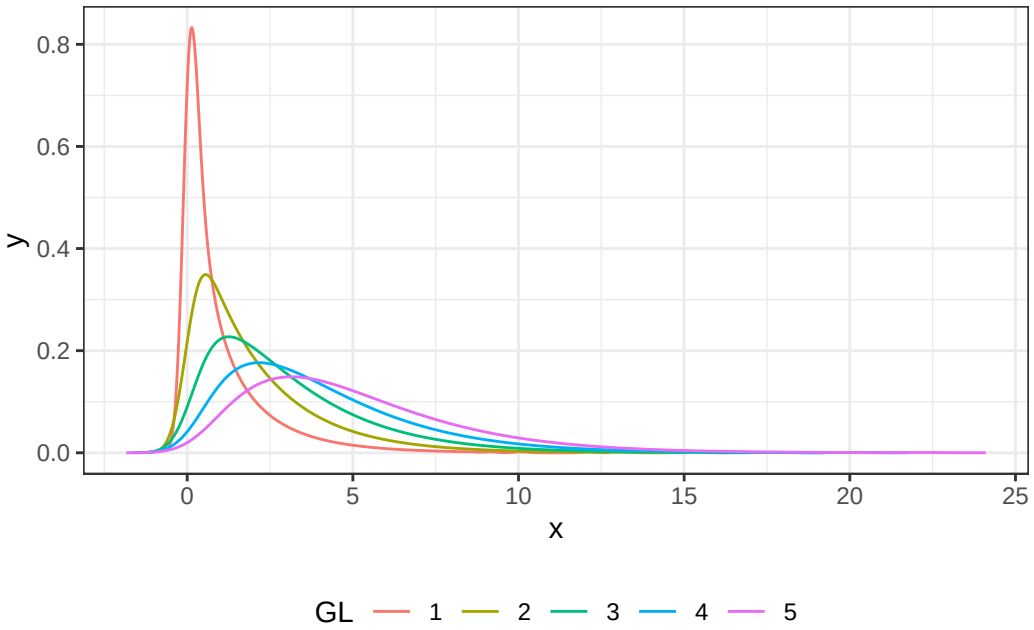
En los capítulos Capítulo 7 y Capítulo 8 vimos varios tests estadísticos que nos permiten apreciar la significanza de diferencias entre dos conjuntos de medidas cuantitativas. Las variables que vimos se medían en escala de razón, intervalo u ordinal. En este capítulo vamos a explorar algunas técnicas que nos permiten trabajar con variables que no se pueden medir en términos numéricos, sino que son de tipo «sí-o-no»; es decir que son de escala nominal.

En particular vamos a explorar la distribuciónd de χ^2 de Pearson. χ es una letra griega que suele pronunciarse «ji» (/xi/) y «chi» (/t i)/[1].

```
#| label: fig-chi-square-plot
#| message: false
#| warning: false
#| echo: false
#| fig-cap: "'Distribución «ji cuadrado» con diferentes grados de libertad."

library(bayestestR)
library(tidyverse)
tmp <- data.frame()
for(i in 1:5){
  mySeq <- seq(from=.5, to=20,by=.01)
  chi <- distribution_chisquared(n = 1000, df =  i)
  chi %>%
    density(adjust=1) %>%  # Compute density function
    as.data.frame() ->tmp2
  tmp2$GL <- i
  tmp <- tmp %>%
    bind_rows(tmp2)
}
tmp$GL <- factor(tmp$GL)
tmp %>%
  ggplot(aes(x=x, y=y, color=GL)) +
  geom_line()+theme_bw()+
  theme(legend.position="bottom")
```

[1]del inglés donde se escribe «chi» y se pronuncia /ka /



9.1 Características

El test de χ^2 nos permite comparar las frecuencias que observamos con las frecuencias que esperaríamos en base a un modelo teórico o una hipótesis sobre la distribución de la variable en cuestión. Por cada par de valores observados y esperados calculamos la diferencia y aplicamos la fórmula de la definición 9.1.

Definición 9.1 (χ^2)**.**

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

donde:

- O: la frecuencia observada
- E: la frecuencia esperada

Es importante tener en cuenta que χ^2 se calcula usando las *frecuencias* y no las proporciones.

La hipótesis nula es que no existe diferencia entre los valores observamos y los valores esperados. La alternativa es que hay tal diferencia. La forma de la distribución χ^2, al igual que la de t, depende de los grados de liberdad que desarrollaremos más adelante.

Cuadro 9.1: **?(caption)**

	A favor	En contra
Mujeres	762	468
Hombres	484	477

9.2 Prueba de independencia o asociación

Un uso muy frecuente de la prueba de χ^2 es la de probar si dos características son independientes o tienen una asociación de manera que las frecuencias elevadas en una de ellas suele ser acompañado con frecuencias altas en la otra.

Digamos que estamos haciendo una encuesta de opinión y preguntamos a 1230 argentinas y a 961 argentinos si están a favor o en contra de la ley del aborto o no. Queremos saber si en género de la persona está asociado con esa opinión. Entonces nuestros datos se pueden desplegar en una tabla 2 por 2.

La hipótesis nula es que no hay asociación entre las dos variables, es decir que el género de la persona no se asocia con su opinión política sobre este tema. Para calcular los valores esperados tenemos que calcular las sumas de las filas y las columnas y además el total de ellos.

Cuadro 9.3: Opiniones sobre la ley del aborto.

	A favor	En contra	total
Mujeres	762	468	1230
Hombres	484	477	961
total	1246	945	2191

El valor esperado es la cantidad de las observaciones que caen en cada celda si las distribuimos proporcionalmente. Esto se calcula multiplicando las sumas de la fila y columna de la celda respectiva y dividir por el total de las observaciones. Por ejemplo, el valor esperado de mujeres a favor sería:

$$E = \frac{1230 \times 1246}{2191} = 699,48$$

Si calculamos esto para todas las celdas obtenemos:

Cuadro 9.4: Valores esperados: opiniones sobre la ley del aborto.

	A favor	En contra	total
Mujeres	699,49	530,51	1230
Hombres	546,51	414,49	961
total	1246,00	945,00	2191

y con esto podemos calcular las diferencias.

Cuadro 9.5: Diferencias entre valores observados y esperados.

	A favor	En contra
Mujeres	62,51	-62,51
Hombres	-62,51	62,51

y podemos aplicar la fórmula en la definición 9.1:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{62,52^2}{699,49} + \frac{-62,52^2}{530,51} + \frac{62,52^2}{546,51} + \frac{-62,52^2}{414,49} = 29,53.$$

Podemos comparar este valor con los de la tabla en el apendice C para un grado de libertad. Vemos que rechazamos H_0 y concluimos que el género sí influye en la opinión sobre este tema.

9.2.1 Grados de libertad

A lo largo de este texto se ha mencionado en algunas ocasiones el término *grados de libertad* y hasta ahora no ha sido demasiado complejo calcularlo restando uno del número de observaciones. El concepto de grado de libertad se puede entender si consideramos la tabla 9.3 en la que tenemos una tabla de contingencia 2×2. Calculamos en ese caso las frecuencias marginales que están el la columna *suma*. Imaginemos que tenemos la misma tabla con las frecuencias marginales pero con una sola de las frecuencias observadas. Así lo hemos hecho en la tabla 9.6.

Cuadro 9.6: Tabla de contingencia con un solo valor.

	A favor	En contra	total
Mujeres	762	-	1230
Hombres	-	-	961
total	1246	945	2191

Con este único valor podemos rellenar las demás celdas, ya que su contenido está dato por la diferencia entre ese valor y los totales marginales. Esto quiere decir que en esta tabla hay un solo valor que se pueda asignar arbitrariamente, el resto está dado por este valor. Por ello decimos que tenemos un solo grado de libertad.

En capítulos anteriores hemos visto que los grados de libertad a menudo son N-1. Podemos usar un ejemplo sencillo para demostrar por qué tiene que ser así. Si hacemos un conjunto de tres números y queremos que la suma sea diez, podemos asignar cualquier número en las primeras dos posiciones, pero cuando vamos a asignar el tercero ya no tenemos libertad de

eligir. Entonces tenemos dos grados de libertad.

Para una tabla de contingencia la fórmula general para calcular los grados de libertad es: $(c-1) \times (f-1)$ es decir número de columnas menos uno por número de filas menos uno. Si la tabla es de 3×3, tendríamos 4 grados de libertad.

9.3 Prueba de χ^2 en R

En este ejemplo vamos a realizar la misma prueba de χ^2 que fuimos desarrollando en las secciones anteriores. Podemos usar la función de R `chisq.test` para realizarla. Toma como argumento una *matriz* de datos de las frecuencias.

```
M <- as.table(
  rbind(c(762, 468),
        c(484, 477))
  )
# Damos nombre a las columnas y las filas
colnames(M) <- c("A favor","En contra")
rownames(M) <- c("Mujeres","Hombres")

# Verificamos el ingreso de datos
M
```

```
#>         A favor En contra
#> Mujeres     762       468
#> Hombres     484       477
```

Como se puede observar, la sintaxis de R no es del todo intuitiva, por lo que siempre conviene verificar que tenemos los números y nombres correctos antes de proceder con el test.

```
# Realizamos test de ji-cuadrado
chisq.test(M)
```

```
#>
#>  Pearson's Chi-squared test with Yates' continuity
#>  correction
#>
#> data:  M
#> X-squared = 29.06, df = 1, p-value = 7.019e-08
```

El valor de p es tan bajo que R lo devuelve en *notación científica*. La parte *e-08* quiere decir que el número es: 0,000000007019, es decir que hay *ocho* ceros antes de los dígitos significativos.

También vemos que el valor de χ^2 que calculó R es distinto al que calculamos a mano, aunque sea por unas décimas. Esto se debe a que por defecto R hace una *corrección de Yates*. Yates descubrió que para una tabla de contingencia 2×2 hay un sesgo positivo y propuso una técnica para contrarrestar el sesgo. Si queremos usar la formula original, y la que usamos para nuestro cálculo a mano podemos agregar el parámetro `correct = FALSE` a la función así:

```
chisq.test(M, correct = FALSE)
```

```
#>
#>  Pearson's Chi-squared test
#>
#> data:  M
#> X-squared = 29.53, df = 1, p-value = 5.506e-08
```

y vemos que R coincide con nuestros cálculos.

9.4 Glosario

Grados de libertad Número de valores que se pueda asignar arbitrariamente a un conjunto manteniendo estable sus propiedades. Equivalente en inglés: «Degrees of freedom».

Prueba de χ^2 «ji cuadrado» Prueba estadística que nos permite apreciar si dos variables nominales están asociadas. Fórmula: $\chi^2 = \sum \frac{(O-E)^2}{E}$ Equivalente en inglés: «Chi-square test».

10 Correlación

La correlación es el área de las estadísticas que estudia la relación sistemática entre dos o más variables e intenta contestar a preguntas como: ¿Si sube A va a subir B también? En este capítulo desarrollaremos algunas técnicas para contestar este tipo de pregunta.

10.1 Visualización

El primer paso para estudiar posibles relaciones entre variables es visualizarlos. Si tenemos dos variables medidas por cada miembro de la población o muestra que estamos investigando podemos generar un *diagrama de disperción* también conocido como *scatterplot*. En este tipo de visualización cada miembro de la muestra/población está representado por un punto, y las coordinadas del punto corresponde a las dos variables que hemos medido, en el eje horizontal y vertical respectivamente.

El la figura 10.1, vemos que la concentración de puntos suben de la izquierda a la derecha. Es decir cuando avanzamos en el eje horizontal avanzamos en el eje vertical también. Es un ejemplo de una *correlación positiva*, como podría ser edad y estatura.

En la figura 10.2 vemos lo contrario, mientras avanzamos en el eje vertical retrocedemos (o bajamos) en el eje horizontal. Esto se conoce como *correlacion negativa*.

En la figura 10.3, también vemos correlación negativa, pero es menos fuerte que en la figura 10.2.

En la figura 10.4 vemos una correlación negativa casi perfecta entre las dos variables.

En la figura 10.5 vemos un caso de correlación inexistente entre las variables en cuestión.

En la figura 10.6 vemos que existe una relación entre las dos variables, pero que esta no es lineal.[1]

Las figuras 10.1, 10.2, 10.3, 10.4, -Figura 10.5] y 10.6 demuestran por qué es preciso graficar los datos al inicio del análisis. Nos da una indicación de si existe una correlación o no, si es positiva o negativa y que tan fuerte es. También nos podemos darnos cuenta de patrones en los datos que no son lineales, como es el caso de los datos en la figura 10.6. Asimismo, a veces nos encontramos con una correlación como la que vemos en la figura 10.4. Las correlaciones

[1] De hecho es cuadrática: $y \sim x^2$.

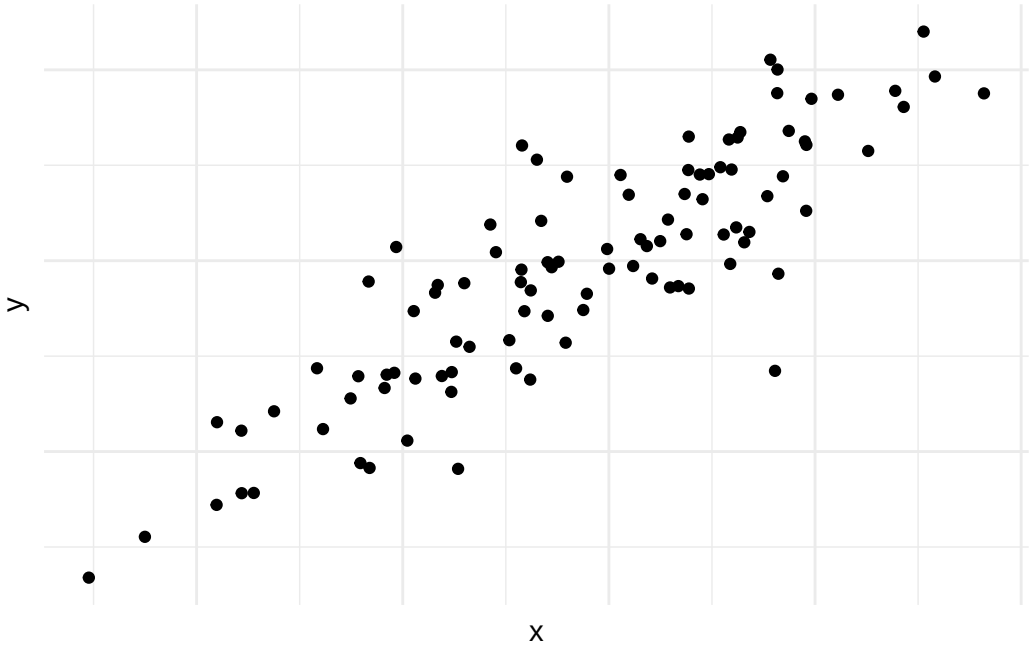


Figura 10.1: Correlación positiva

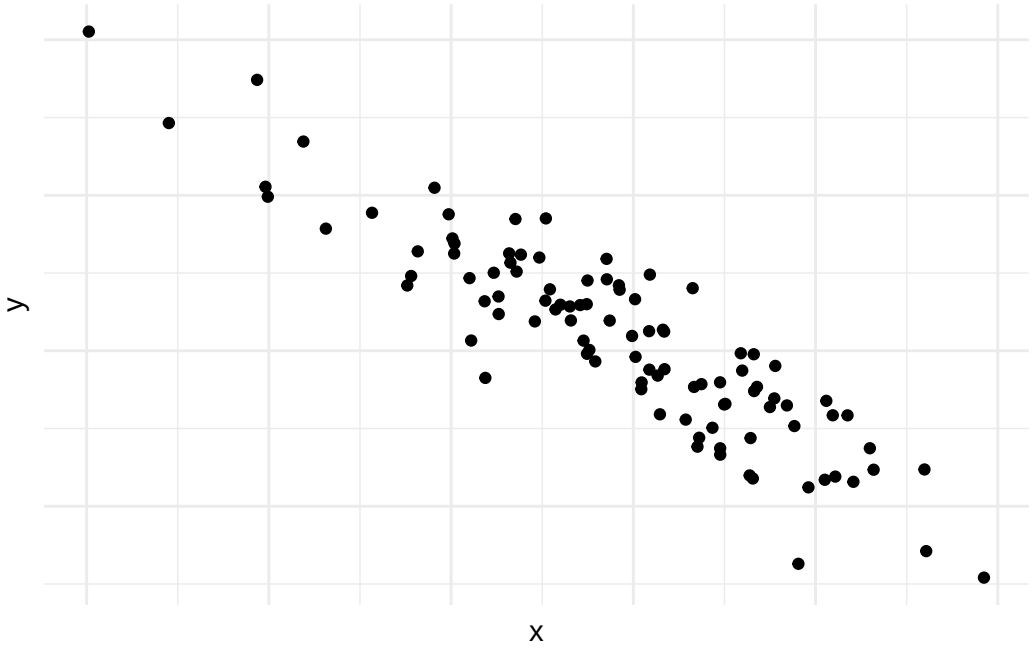


Figura 10.2: Correlación negativa

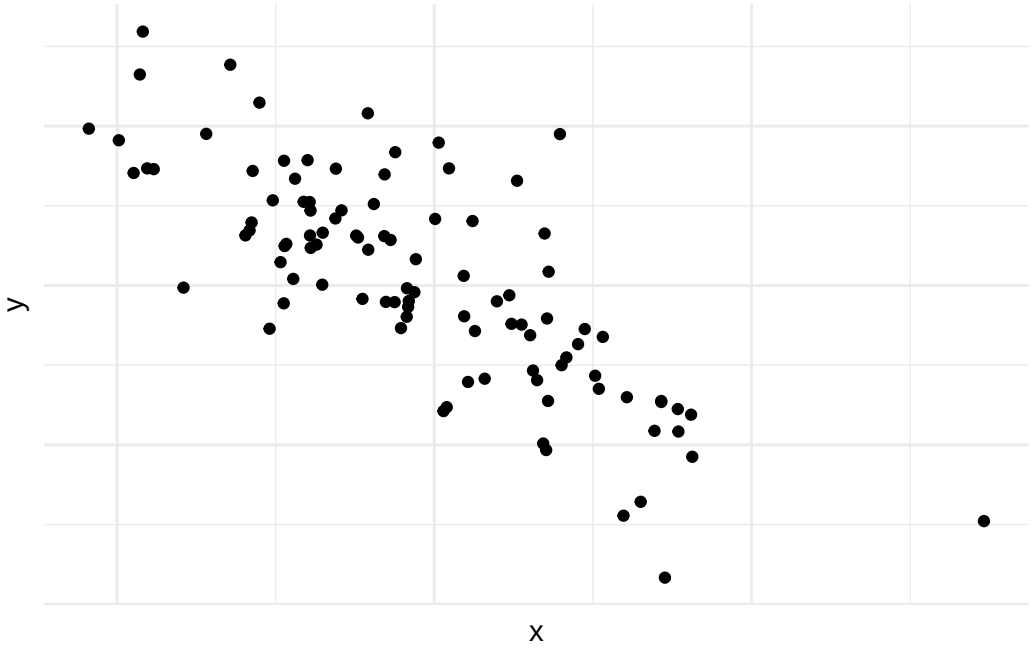


Figura 10.3: Correlación negative leve

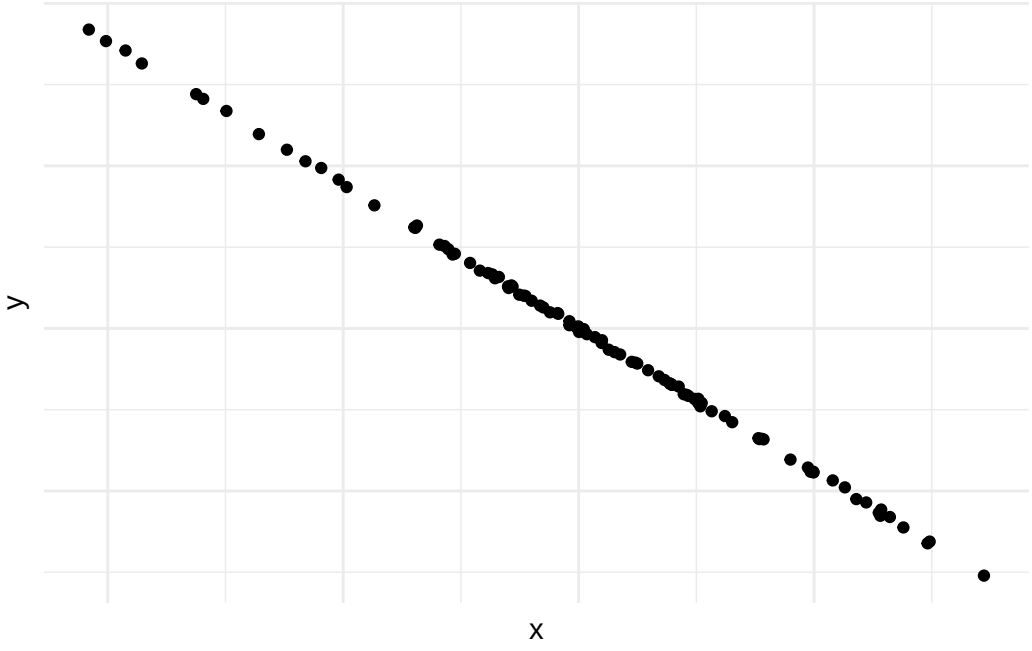


Figura 10.4: Correlación casi perfecta

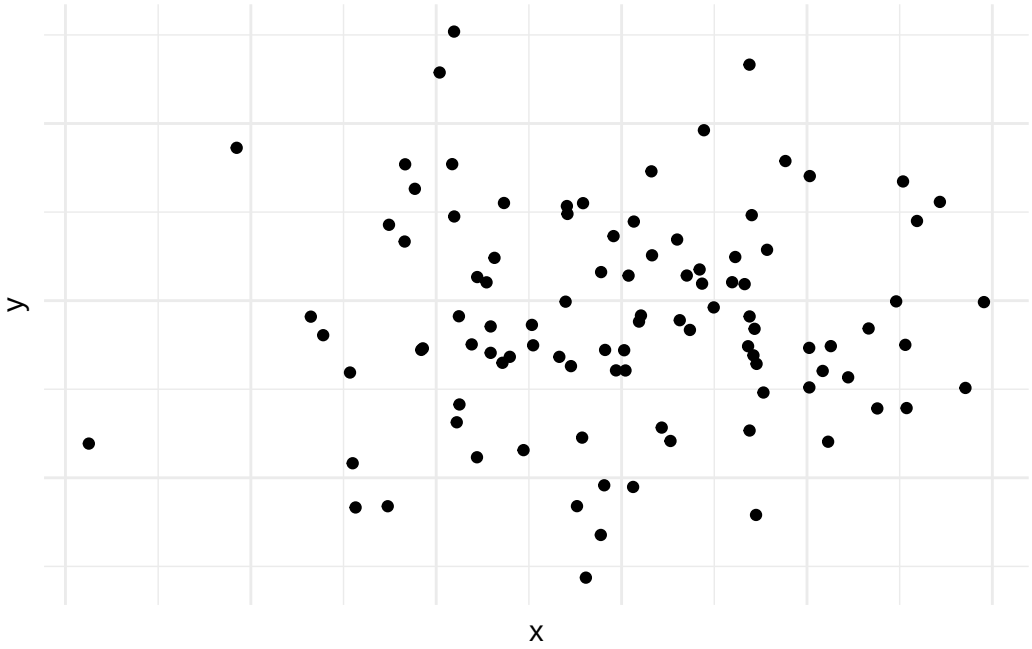


Figura 10.5: Correlación nula

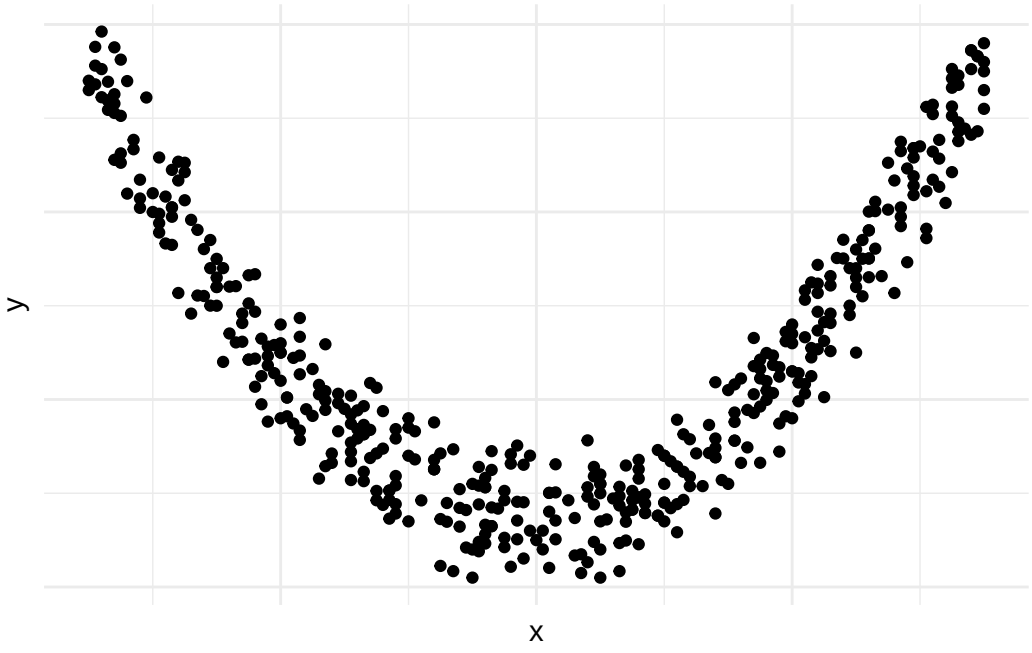


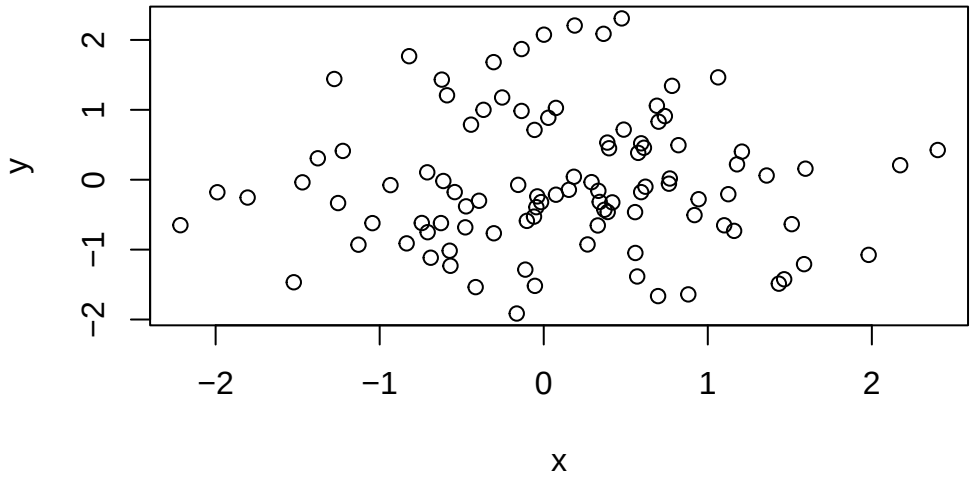
Figura 10.6: Relación no lineal

que son demasiado perfectas suelen ser un signo de advertencia y podemos preguntarnos si en realidad son dos variables distintas o si las dos están midiendo lo mismo.

Ejemplo 10.1 (Generar diagrama de dispersión en R).

```
# Generamos datos
datos = data.frame(
  x=rnorm(100),
  y=rnorm(100)
)

# Graficamos
plot(datos)
```



En el ejemplo 10.1 utilizamos la función `rnorm` para generar cien observaciones aleatorias con distribución normal y los ponemos dentro de un `data.frame`. Luego usamos la función `plot` para graficarlos. Como nuestro `data.frame` tiene solo dos columnas R entiende que estos son los datos que queremos graficar. Si el `data.frame` tiene más columnas, podemos especificar los que queremos graficar así:

```
plot(datos$x,datos$y)
```

10.2 Generar diagrama de dispersión en R

Por defecto R viene con algunos data.frames ya cargados, uno de ellos es «trees», podemos

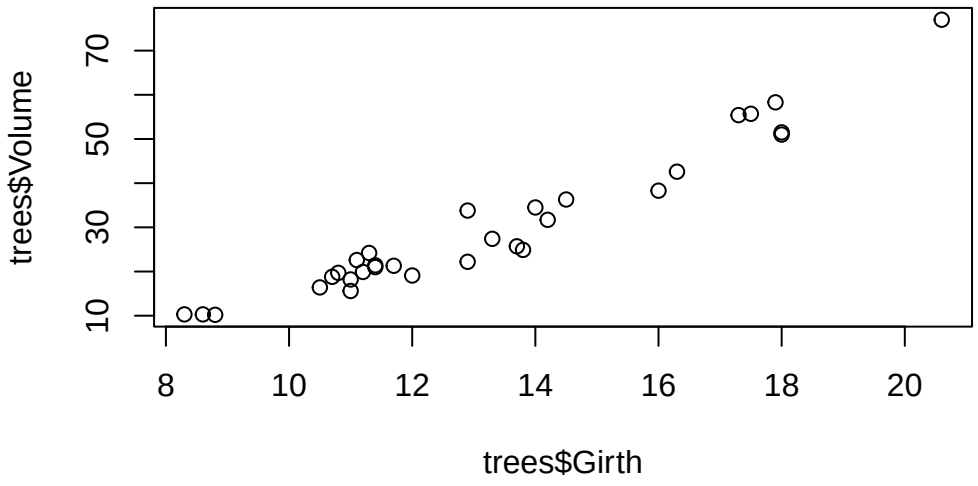
usar la función `head` para ver las primeras seis filas.

```
head(trees)
```

```
#>   Girth Height Volume
#> 1   8.3     70   10.3
#> 2   8.6     65   10.3
#> 3   8.8     63   10.2
#> 4  10.5     72   16.4
#> 5  10.7     81   18.8
#> 6  10.8     83   19.7
```

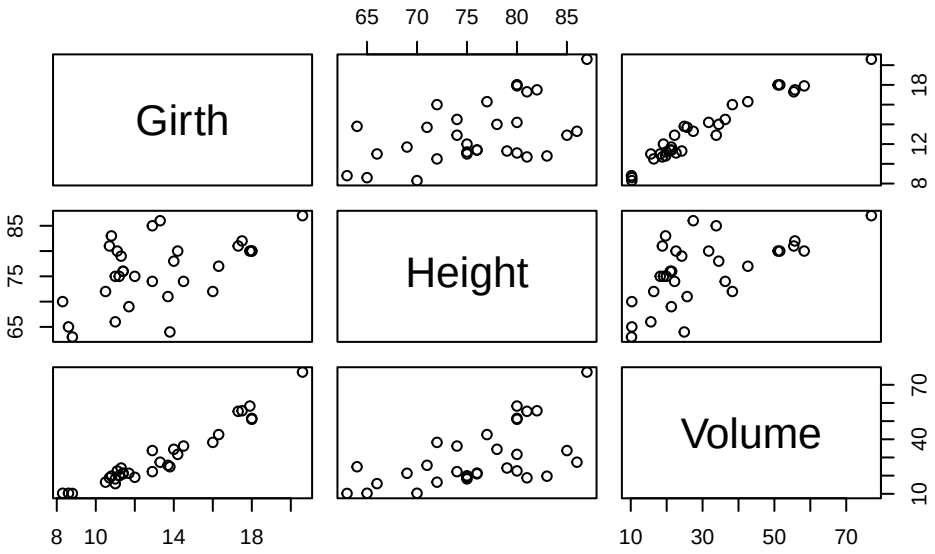
Vemos que tiene tres columnas «Girth», «Height» y «Volume» (circumferencia, alto y volumen), los que, por lógica, deben tener alta correlación. Graficamos dos de ellos.

```
plot(trees$Girth, trees$Volume)
```



Si usamos la función `plot` sin especificar columnas R entiende que queremos ver todas las combinaciones.

```
plot(trees)
```



Este tipo de visualización puede ser útil cuando tenemos algunas variables y queremos darnos cuenta qué correlaciones hay. La visualización funciona bien hasta cierto número de columnas –ocho más o menos–, luego se vuelve difícil de leer y por ende de interpretar.

10.3 Coeficientes de correlación

Para tener una medida cuantitativa precisa de la correlación entre las variables calculamos un coeficiente de correlación. A continuación vamos a tres de ellos, el de Pearson, el de Spearman y el coeficiente ϕ (de la letra griega que corresponde a f en minúscula – se pronuncia «fi». Los coeficientes de correlación se expresan por un número con varios decimales entre -1 y 1, donde -1 y 1 indican correlaciones perfectas, negativas y positivas respectivamente y 0 indica correlación nula.

El coeficiente Pearson es adecuado para datos de escala de razón o intervalo, el de Spearman para datos de escala ordinal y el coeficiente ϕ se usa para datos nominales.

10.3.1 Coeficiente Pearson

Como ya mencionamos, el coeficiente de Pearson es apropiado cuando las variables a comparar con de escala de intervalo o razón ya que toma en cuenta la magnitud relativa de las observaciones.

Si tenemos un conjunto de pares de observaciones podemos representar el primer elemento del par por x y el segundo por y. Entonces el conjunto de los x van a tener una desviación estándar se calcula según la definición 3.4, así:

$$s_x = \sqrt{\frac{(\sum (x - \bar{x})^2}{N - 1}}.$$

De la misma manera y tiene su desviación estándar:

$$s_y = \sqrt{\frac{(\sum (y - \bar{y})^2}{N - 1}}.$$

Ahora podemos normalizar las variables según la definición 4.1 asi:

$$z_x = \frac{x - \bar{x}}{s_x},$$

$$z_y = \frac{y - \bar{y}}{s_y}.$$

Y con estos datos podemos calcular el coeficiente según la definición 10.1.

Definición 10.1 (Coeficiente de correlación de Pearson)**.**

$$r = \frac{\sum z_x z_y}{N - 1}$$

donde:

- $\sum z_x z_y$: La suma de los productos[2] de las dos variables normalizadas.

Existe otra definición es matemáticamente equivalente y que se usa a veces para hacer el cálculo a mano:

[2] «Producto» en matemática es el resultado de una operación de multiplicación. Si multiplicamos $2 \times 2 = 4$, 4 es el producto.

10.4 Coeficiente de correlación Pearson

$$r = \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{\{N\Sigma x^2 - (\Sigma x)^2\} \times \{N\Sigma y^2 - (\Sigma y)^2\}}}$$

Ejemplo 10.2 (Cálculo del Coeficiente de correlación de Pearson)**.**

En este ejemplo, adaptado de (Butler 1985), vamos a suponer que hemos tomado un examen de traducción y otro de comprensión de inglés a doce estudiantes. Los resultados de estos exámenes están en la tabla 10.1.

Cuadro 10.1: Resultados de un examen de tradución (x) y de comprensión (y) de ingles.

Estudiante	x	y
1	17	15
2	13	13
3	12	8
4	14	17
5	15	16
6	8	9
7	9	14
8	13	10
9	11	16
10	14	13
11	12	14
12	16	17

Para poder aplicar la fórmula vamos a precisar los valores llevados al cuadrado, el producto de $x \times y$ y las sumas de las columnas. Calculándolos obtenemos los datos de la tabla 10.2.

Cuadro 10.2: Resultados de un examen de tradución (x) y de comprensión (y) de ingles.

Estudiante	x	y	x2	y2	xy
1	17	15	289	225	255
2	13	13	169	169	169
3	12	8	144	64	96
4	14	17	196	289	238
5	15	16	225	256	240
6	8	9	64	81	72
7	9	14	81	196	126

Estudiante	x	y	x2	y2	xy
8	13	10	169	100	130
9	11	16	121	256	176
10	14	13	196	169	182
11	12	14	144	196	168
12	16	17	256	289	272

Sabemos que N=12, pero vamos a precisar las sumas de algunas columnas:

- $\Sigma x = 154$
- $\Sigma y = 162$
- $\Sigma x^2 = 2054$
- $\Sigma y^2 = 2290$
- $\Sigma xy = 2124$.

Aplicamos la fórmula:

$$
\begin{aligned}
r &= \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{\{N\Sigma x^2 - (\Sigma x)^2\} \times \{N\Sigma y^2 - (\Sigma y)^2\}}} \\
&\Updownarrow \\
r &= \frac{12 \times 2\ 124 - 154 \times 162}{\sqrt{\{12 \times 2\ 054 - 154^2\} \times \{12 \times 2\ 290 - 162^2\}}} \\
&\Updownarrow \\
r &= \frac{25\ 488 - 24\ 948}{\sqrt{\{24\ 648 - 23\ 716\} \times \{27\ 480 - 26\ 244\}}} \\
&\Updownarrow \\
r &= \frac{540}{\sqrt{932 \times 1236}} \\
&\Updownarrow \\
r &= \frac{540}{\sqrt{1\ 151\ 932}} \\
&\Updownarrow \\
r &= \frac{540}{1\ 073,28} \\
&\Updownarrow \\
r &= 0,5031
\end{aligned}
$$

Ejemplo 10.3 (Cálculo del Coeficiente de correlación de Pearson en R)**.** Si no queremos hacer todos los cálculos del ejemplo 10.2 a mano podemos recurrir a R que con la función `cor` lo calcula.

```
# Cargamos datos
datos<- data.frame(
  Estudiante = c( 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12),
  x = c( 17, 13, 12, 14, 15, 8, 9, 13, 11, 14, 12, 16),
  y = c( 15, 13, 8, 17, 16, 9, 14, 10, 16, 13, 14, 17)
)

# Llamamos función
cor(datos$x, datos$y)
```

```
#> [1] 0.5031258
```

10.4.1 Coeficiente Spearman

Si una o ambas variables que estamos comparando son de escala ordinal, el coeficiente apropiado es el de Spearman. Para calcularlo ordenamos las observaciones de la primer variable de manera ascendiente y les damos el valor de su orden. Si dos observaciones de la misma variable tienen el mismo valor, si hay empates, se saca el promedio cual si el empate no hubiera existido. Hacemos lo mismo para la segunda variable. Calculamos la diferencia entre los rangos para cada par de observaciones. La correlación Spearman o ρ de la letra griega r se calcula según la definición 10.2.

Definición 10.2 (Coeficiente de correlación de Spearman)**.**

$$\rho = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

Imaginamos que pedimos a diez personas que ranqueen en una escala de uno a diez cuánto les gustaron dos cafeterías de Buenos Aires. Ya que los datos son de escala ordinal, tenemos que recurrir a *Spearman*. Usamos la misma función `cor` con un parámetro extra: `method = "spearman"` para indicar que queremos usar la correlación de Spearman.

```
# Cargamos datos
rankings <- data.frame(
  Cafe.A = c(7, 6, 4, 5, 8, 7, 10, 3, 9, 2),
  Cafe.B = c(5, 4, 5, 6, 10, 7, 9, 2, 8, 1)
```

```
)

# Llamamos función
cor(rankings$Cafe.A, rankings$Cafe.B, method = "spearman")
```

```
#> [1] 0.875
```

Observamos que hay alto grado de correlación entre los ranking de los dos cafés.

10.4.2 Coeficiente ϕ

Si las dos variables en cuestión son nominales la pregunta se reduce a: ¿Si observamos la propiedad A es probable que observemos también B? Si estamos trabajando con datos educativos la pregunta podría ser ¿Si el estudiante responde correctamente el 1ʳ ítem, es probable que también acierte el 2º? Esto se puede representar en una tabla 2×2 como la que vemos en la figura 10.7.

Variable 2	*Variable 1*	
	−	+
+	A	B
−	C	D

Figura 10.7: Tabla de contingencia dos por dos

En esta tabla las celdas A, B, C y D son las frecuencias de las observaciones. Por ejemplo A sería el número de estudiantes que acertaron el 1º pero no el 2º. B la frecuencia de estudiantes que pasaron ambos ítems y así sucesivamente. La correlación entre las dos variables se puede medir aplicando la formula de la definición 10.3.

Definición 10.3 (Coefficiente de ϕ)**.**

$$\phi = \frac{BC - AD}{\sqrt{(A+B) \times (C+D) \times (A+C) \times (B+D)}}$$

Debería quedar claro que el coeficiente ϕ está estrechamente relacionado con la prueba de χ^2 que vimos en la definición 9.1. De hecho se relacionan matemáticamente:

$$\phi = \sqrt{\chi^2/N} \Leftrightarrow \chi^2 = N \times \phi^2.$$

Por tanto la significanza de ϕ se puede obtener por medio de la conversión a χ^2.

10.5 Interpretación de correlacciones

Es muy importante entender que una correlación, incluso alta, entre dos variables no quiere decir que la relación entre ellas es de causa y efecto. Si hacemos una muestra en la escuela secundaria y medimos estatura, por un lado, y nivel de inglés por otra, es bastante probable que encontremos una correlación muy fuerte entre las dos variables. Pero debería estar claro que ni la estatura causa conocimientos de inglés ni tampoco lo contrario. Lo que claramente pasa es que mas *edad* los estudiantes tienen más estatura y han cursado más niveles de inglés.

Saltar a conclusiones sobre causalidad basadas en correlaciones es tal vez el error estadístico más frecuente en la literatura tanto académica como periodística. Nunca hay que olvidar que una correlación significativa solo nos dice que existe una relación *matemática* entre dos variables. No nos indica cómo interpretarla ni mucho menos sobre sus causas y efectos.

10.6 Glosario

Coeficiente ϕ «fi». Coeficiente que da cuenta de la correlación entre dos variables nominales. Fórmula: $\phi = \frac{BC-AD}{\sqrt{(A+B)\times(C+D)\times(A+C)\times(B+D)}}$ Función relevante en R: `chi.test`. Equivalente en inglés: «Phi correlation».

Coeficiente de correlación de Pearson Coeficiente que da cuenta de la correlación entre dos variables. Fórmula: $r = \frac{\sum z_x z_y}{N-1}$ Función relevante en R: `cor`. Equivalente en inglés: «Pearson coefficient».

Coeficiente Spearman Coeficiente que da cuenta de la correlación entre dos variables cuando una o ambas de ellas son de escala ordinal. Fórmula: $\rho = 1 - \frac{6\sum d^2}{N(N^2-1)}$ Función relevante en R: `cor`. Equivalente en inglés: «Spearman coefficient».

Correlación negativa Relación entre dos variables que muestra que si una aumenta la otra disminuye. Equivalente en inglés: «Positive correlation».

Correlación positiva Relación entre dos variables que muestra que ambas aumentan o disminuyen simultáneamente. Equivalente en inglés: «Positive correlation».

Referencias

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, y Richard Iannone. 2019. *rmarkdown: Dynamic Documents for R*. https://CRAN.R-project.org/package=rmarkdown.

Austen, Jane. 1817. *Persuasion*. John Murray.

Butler, Christopher. 1985. *Statistics in linguistics*. Basil Blackwell.

ENNyS. 2007. «Encuesta Nacional de Nutrición y Salud.» Ministerio de Salud de Argentina.

Makowski, Dominique, Mattan S. Ben-Shachar, y Daniel Lüdecke. 2019. «bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework.» *Journal of Open Source Software* 4 (40): 1541. https://doi.org/10.21105/joss.01541.

Shier, Rosie. 2004. «Paired t-tests». http://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf.

Silge, Julia, y David Robinson. 2016. «tidytext: Text Mining and Analysis Using Tidy Data Principles in R». *JOSS* 1 (3). https://doi.org/10.21105/joss.00037.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. «Welcome to the tidyverse». *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. http://yihui.name/knitr/.

———. 2018. *bookdown: Authoring Books and Technical Documents with R Markdown*. https://CRAN.R-project.org/package=bookdown.

A Distribución t

La tabla que sigue despliega los valores críticos de la distribución t para diferentes niveles de significanza y grados de libertad (GL). Los niveles de significanza indican un test no direccional. Para una prueba unidireccional se usará el nivel inmediatamente superior.

Cuadro A.1: Valores críticos de t por nivel de significanza y grados de libertad.

Grados	$p < 0{,}20$	$p < 0{,}10$	$p < 0{,}05$	$p < 0{,}02$	$p < 0{,}01$
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763

Grados	$p < 0{,}20$	$p < 0{,}10$	$p < 0{,}05$	$p < 0{,}02$	$p < 0{,}01$
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750

B Valores críticos del test de signo

La tabla B.1 muestra valores críticos de W para diferentes valores de N. Para que sea significativo W tiene que ser menor o igual al valor.

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0      v purrr   1.0.1
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Cuadro B.1: Valores críticos de W en el test de signo por N.

N	0,05	0,025	0,01
5	0	-	-
6	0	0	-
7	0	0	0
8	1	0	0
9	1	1	0
10	1	1	0
11	2	1	1
12	2	2	1
13	3	2	1
14	3	2	2
15	3	3	2
16	4	3	2
17	4	4	3
18	5	4	3
19	5	4	4
20	5	5	4
21	6	5	4
22	6	5	5
23	7	6	5

N	0,05	0,025	0,01
24	7	6	5
25	7	7	6

C Distribución χ^2

La tabla que sigue despliega los valores críticos de la distribución χ^2 para diferentes niveles de significanza y grados de libertad (GL). Los niveles de significanza indican un test no direccional.

Cuadro C.1: Valores críticos de 2 por nivel de significanza y grados de libertad.

GL	p < 0,20	p < 0,10	p < 0,05	p < 0,02	p < 0,01	p < 0,001
1	1,642	2,706	3,841	5,024	6,635	10,828
2	3,219	4,605	5,991	7,378	9,210	13,816
3	4,642	6,251	7,815	9,348	11,345	16,266
4	5,989	7,779	9,488	11,143	13,277	18,467
5	7,289	9,236	11,070	12,833	15,086	20,515
6	8,558	10,645	12,592	14,449	16,812	22,458
7	9,803	12,017	14,067	16,013	18,475	24,322
8	11,030	13,362	15,507	17,535	20,090	26,124
9	12,242	14,684	16,919	19,023	21,666	27,877
10	13,442	15,987	18,307	20,483	23,209	29,588
11	14,631	17,275	19,675	21,920	24,725	31,264
12	15,812	18,549	21,026	23,337	26,217	32,909
13	16,985	19,812	22,362	24,736	27,688	34,528
14	18,151	21,064	23,685	26,119	29,141	36,123
15	19,311	22,307	24,996	27,488	30,578	37,697
16	20,465	23,542	26,296	28,845	32,000	39,252
17	21,615	24,769	27,587	30,191	33,409	40,790
18	22,760	25,989	28,869	31,526	34,805	42,312
19	23,900	27,204	30,144	32,852	36,191	43,820
20	25,038	28,412	31,410	34,170	37,566	45,315
21	26,171	29,615	32,671	35,479	38,932	46,797
22	27,301	30,813	33,924	36,781	40,289	48,268
23	28,429	32,007	35,172	38,076	41,638	49,728
24	29,553	33,196	36,415	39,364	42,980	51,179
25	30,675	34,382	37,652	40,646	44,314	52,620
26	31,795	35,563	38,885	41,923	45,642	54,052
27	32,912	36,741	40,113	43,195	46,963	55,476
28	34,027	37,916	41,337	44,461	48,278	56,892

GL	$p < 0{,}20$	$p < 0{,}10$	$p < 0{,}05$	$p < 0{,}02$	$p < 0{,}01$	$p < 0{,}001$
29	35,139	39,087	42,557	45,722	49,588	58,301
30	36,250	40,256	43,773	46,979	50,892	59,703