Métodos Cuantitativos

Materiales de cátedra y de consulta

Aleksander Dietrichson, PhD

01/15/2023

Indice de contenidos

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 estadístocos «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 acadédico, 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 leguaje 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

```
source("_common.R")
```

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 los algunos usos importantes de la estadística: descriptiva e inferencial. Finalmente consideramos algunas maneras de clasificar variables.

1.1 Poblaciones y muestras

En su uso diaro 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 de un conjunto, 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 *finintas* 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 fininta no es demasiado grande podemos investigar la totalidad de la población. 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 a votar 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 selecionar 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 selecionada para hacer parte de la muestra.

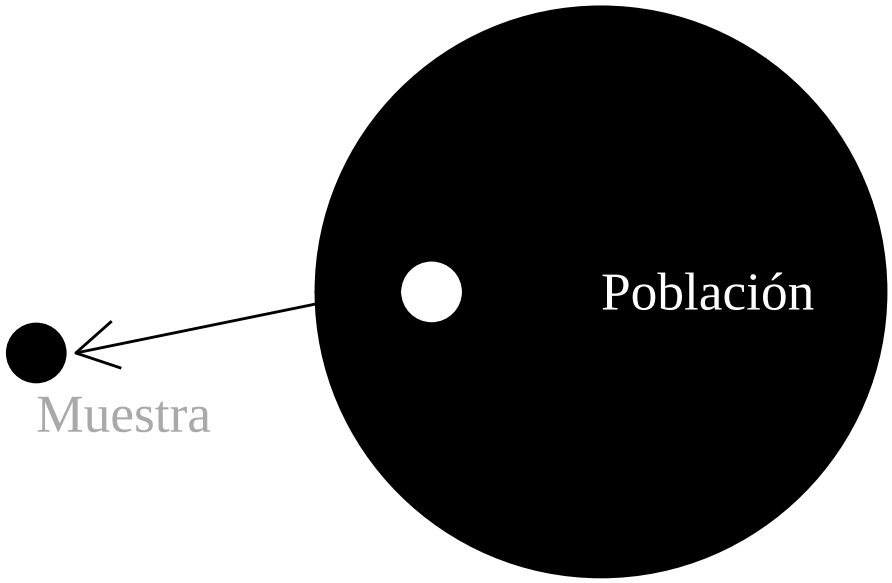


Figura 1.1: Población y muestra.

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 unidaded 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 quisieramos sacar veinte libros al hazar 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 en R: Generar muestra

```
sample(x = 1:200, size = 20)

## [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 en R: 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 cuasialeatoria

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 *cuasialeatoria*, 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.

1.1.2.1 Ejemplo en R: Generar una sequencia

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

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

1.1.3 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.

1.2 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

[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.

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 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: (#tab:compinaciones-posibles) Combinaciones 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:

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

Cuadro 1.2: (#tab:probabilidades-combinaciones-posibles) 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

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: (#tab:probabilidades-combinaciones-posibles-diez) 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 @ref(tab:probabilidades-combinaciones-posibles-diez) 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].

1.3 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.

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 identico para ambos grupos y se podía obtener un máximo de veinte.

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.

[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.

[5] Adaptado de Butler (1985)

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*.

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.

1.4 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».

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

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

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.

1.5 Glosario

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).

Ejemplo 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 **c**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:

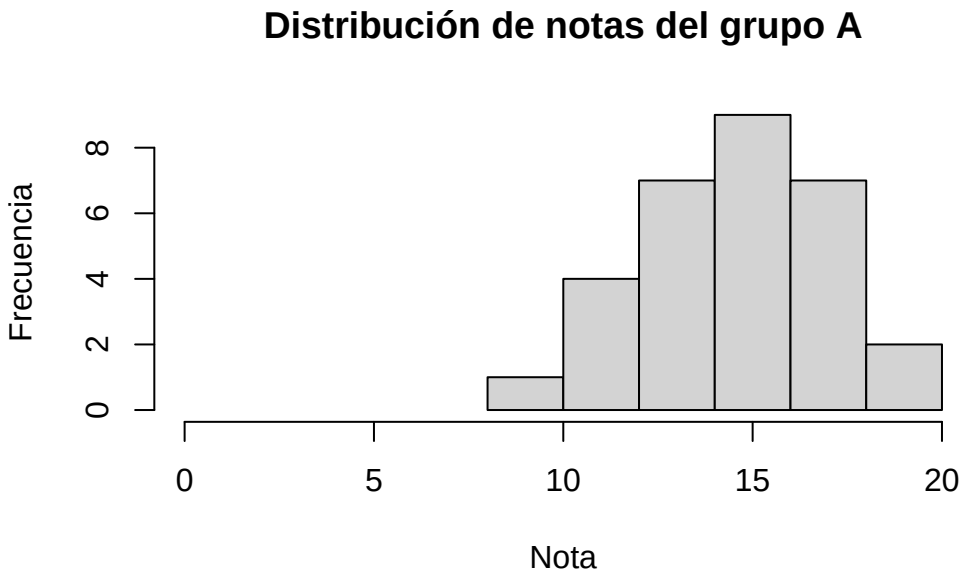
```
table(grupo.A)
```

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

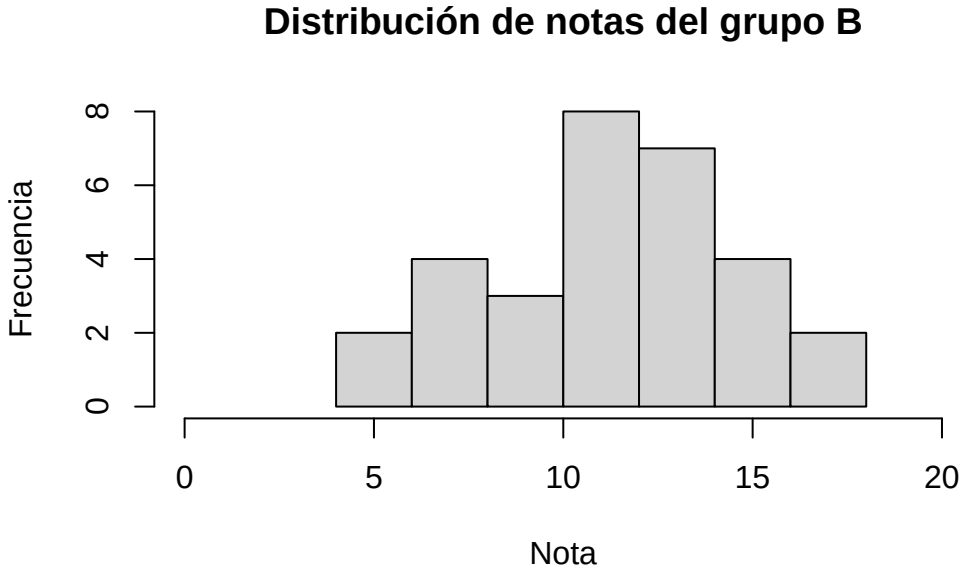
2.3 Histogramas

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



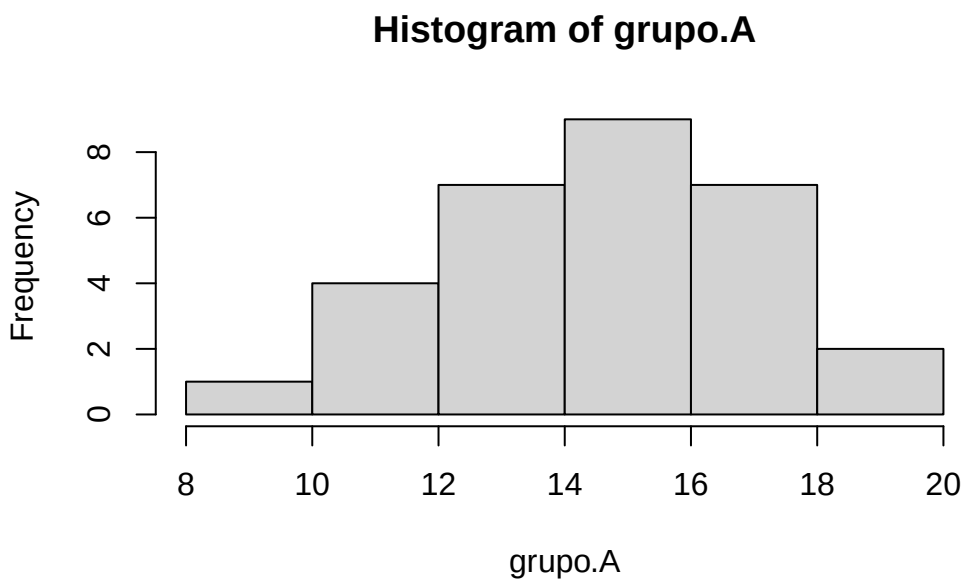
Comparando estos dos diagramas nos damos cuenta de que la estructura de los datos son disimilares. En el grupo A las notas se centran alrededor de quince, en cambio para el grupo B la concentración está en el rango diez-catorce, con un pico menor alrededor de siete.

2.3.0.1 Ejemplo en R: Histograma

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`).

```
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)`.

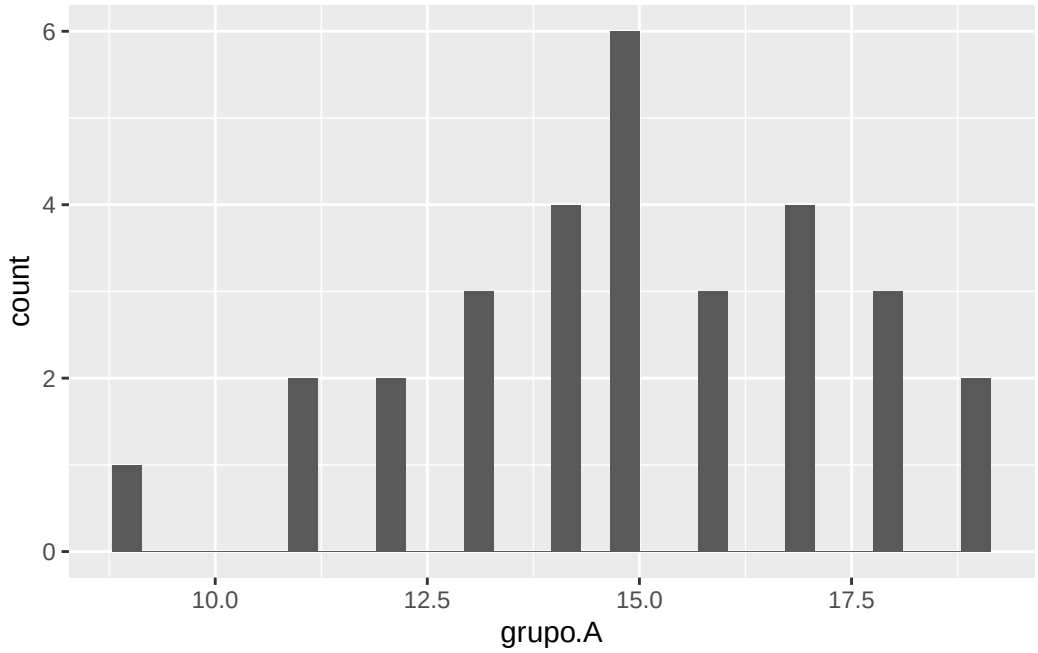
2.3.0.2 Ejemplo en R: Histograma en ggplot

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.

```
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
)
```

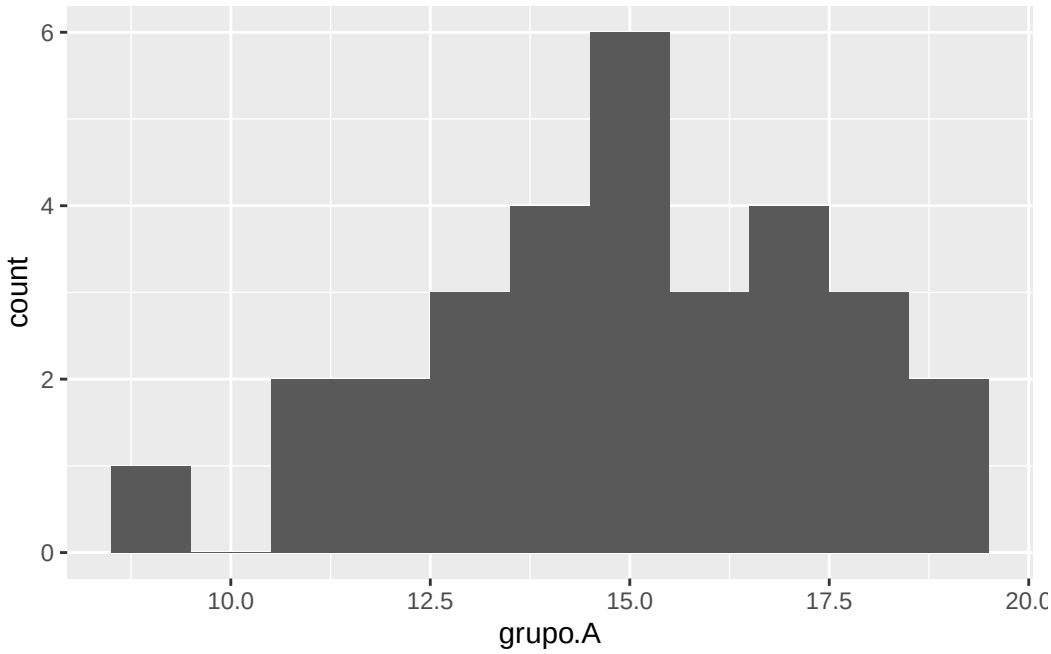
Nótese que usamos el operador `=` dentro de la definición del data.frame. Luego cargamos las funciones de tidyverse y procedemos a construir nuestro gráfico.

```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 agragando otro parametro a la función asi:

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

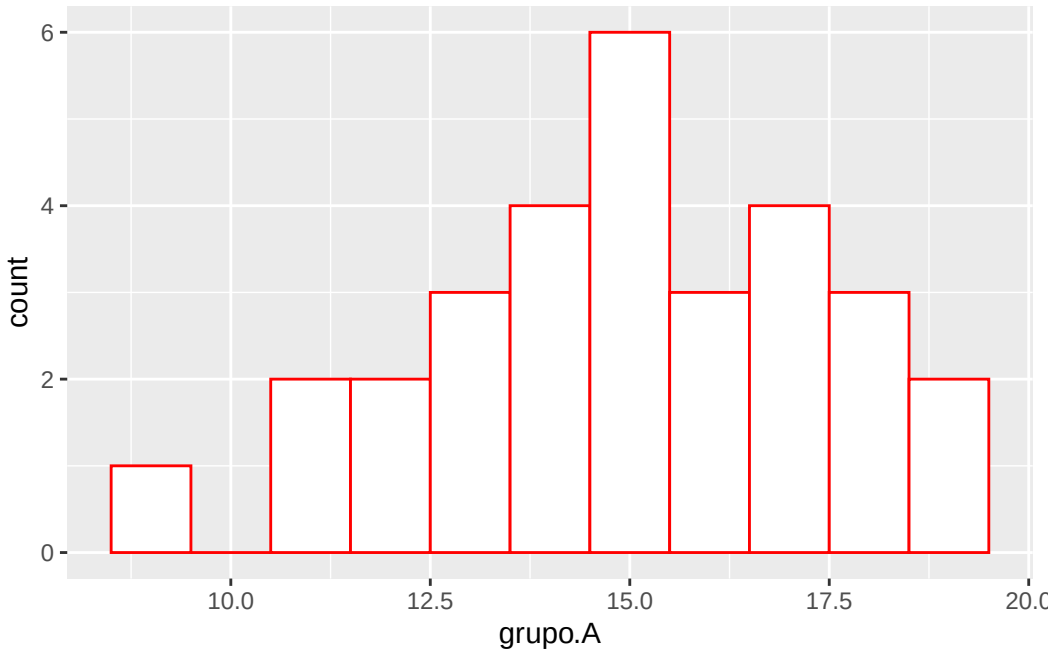


en este caso hemos especificado que el ancho de cada columna sea de 1, con lo cual se visualizan mejor estos datos.

2.3.0.2.1 Agregando un poco de color

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

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



2.3.0.3 El operador «pipe»

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

2.4 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.

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

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

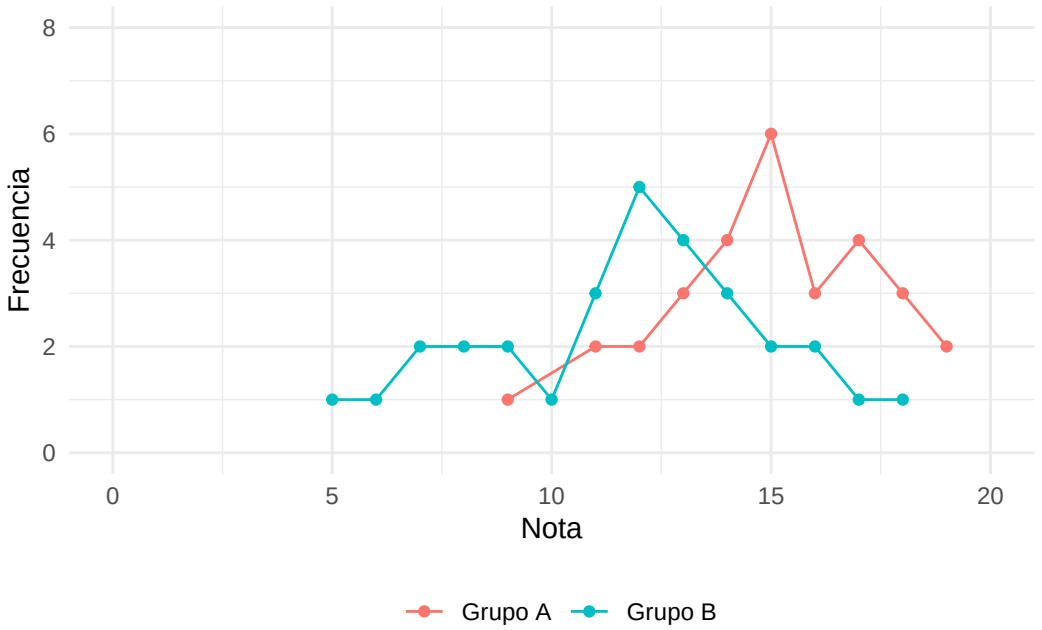


Figura 2.1: Polígono de frecuencias de notas obtenidas por dos grupos de estudiantes

el texto. Obtenemos:

A differencia de la distribución de notas, vemos acá que encontramos observaciones a lo largo del rango de uno a deciseis, 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 cuanquier 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.

2.5 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 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 @ref(la-distribucion-normal). Por ahora simplemente vamos a considerar si los datos de nuestras muestras se asemejan a ésta o si tiene otro perfil.

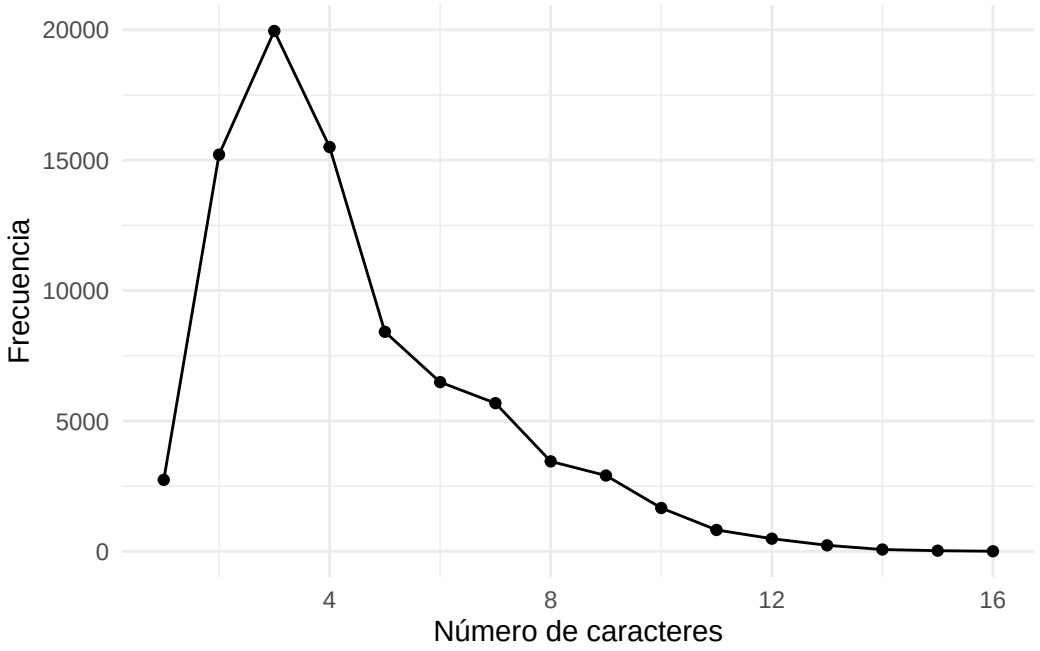


Figura 2.2: Polígono de frecuencias del largo de palabras en un texto de Austin

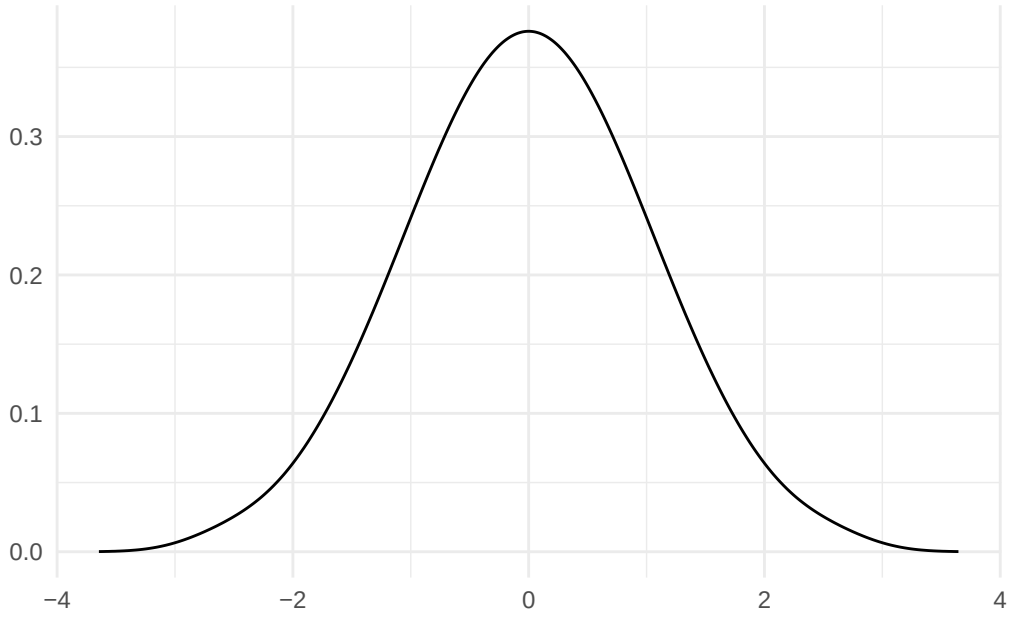


Figura 2.3: Distribución normal

2.5.1 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*.

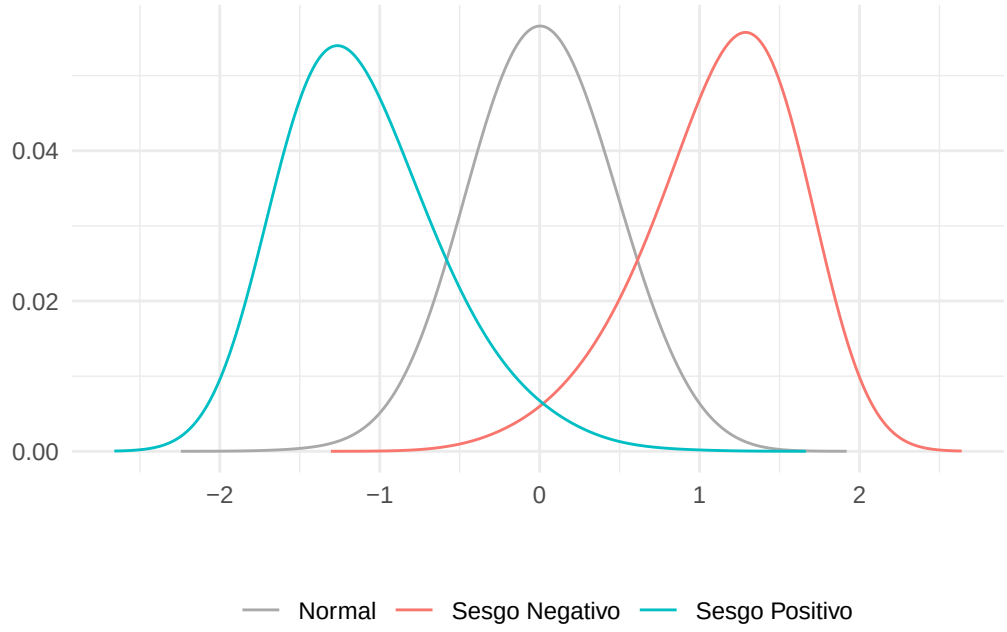


Figura 2.4: Distribuciónes normal y sesgadas

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 characteres* en el texto de Austen tiene *sesgo positivo*.

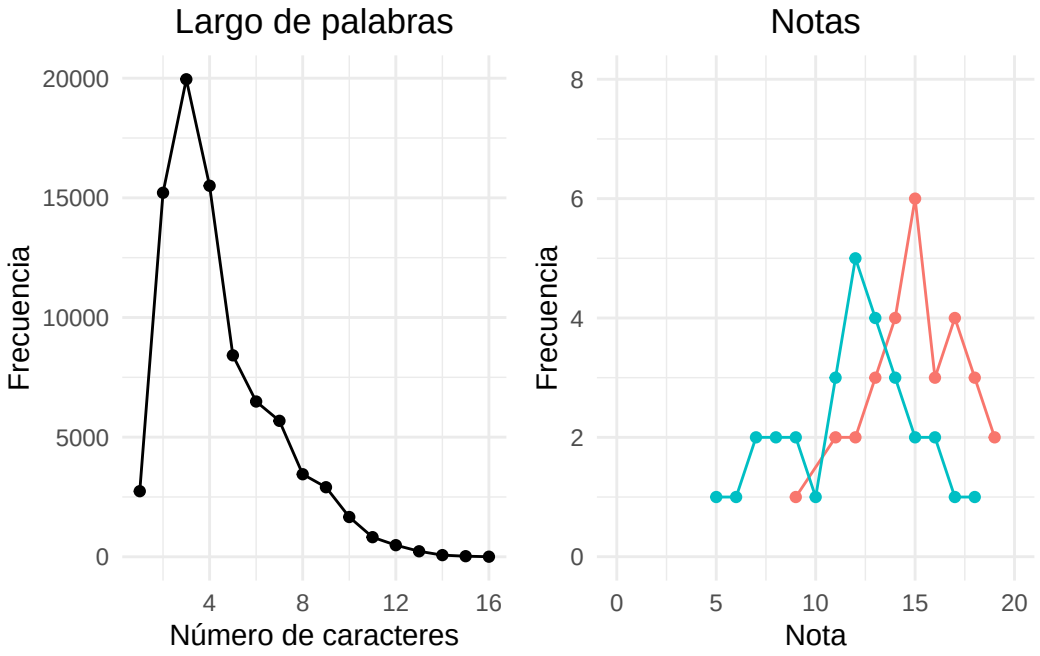
Nótese también que la si bien la escala vertical de los dos gráficos son de muy diferente magnitud, la máxima frecuencia es veinte mil (20.000) y seis (6) respectivamente, podemos comparar las dos distribuciones.

2.6 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:

$$\frac{15+12+11+18+15+15}{6} = \frac{86}{6} = 14,33(\#eq : media - de - seis - observaciones) \quad (3.1)$$

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

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

[1] y por ende de más uso incorrecto

Ya con el cómputo en @ref(eq:media-de-treinta-observaciones) 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)
```

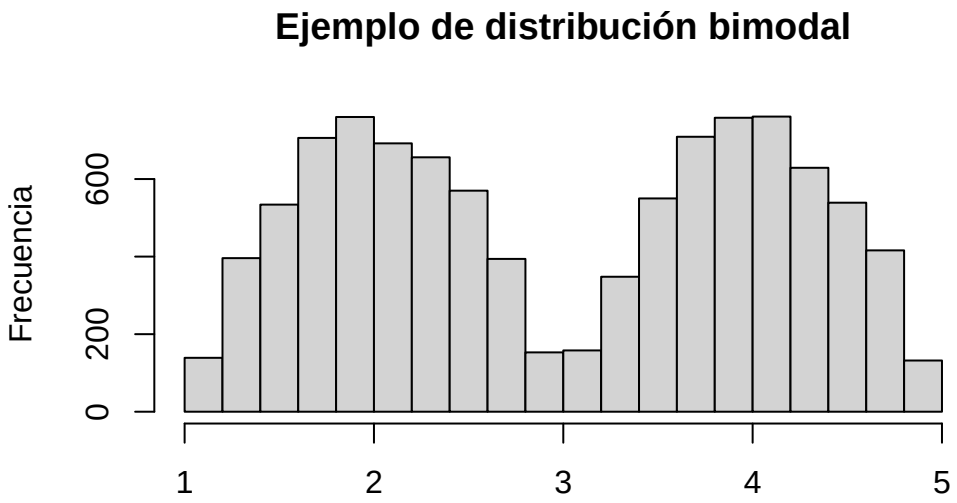
```
#> [1] 11.5
```

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.

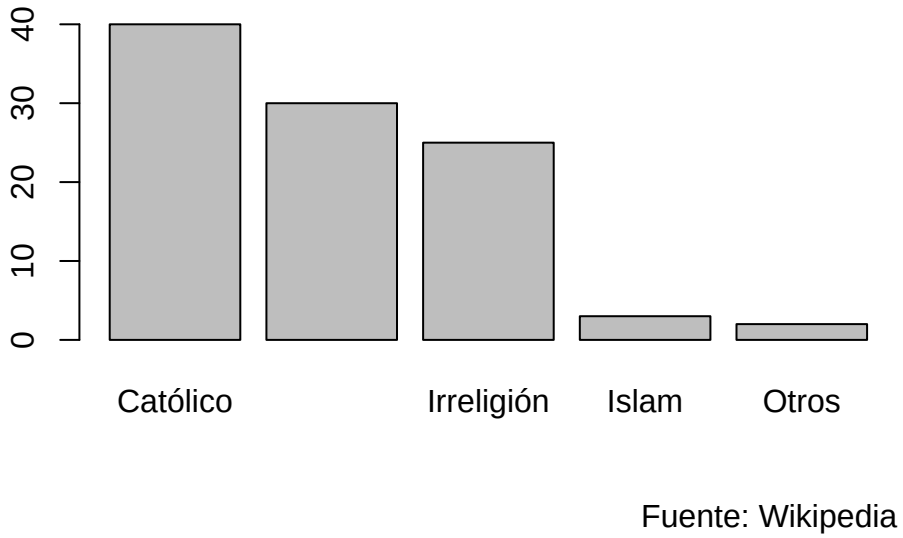


Figura 3.1: Religión en Alemania

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

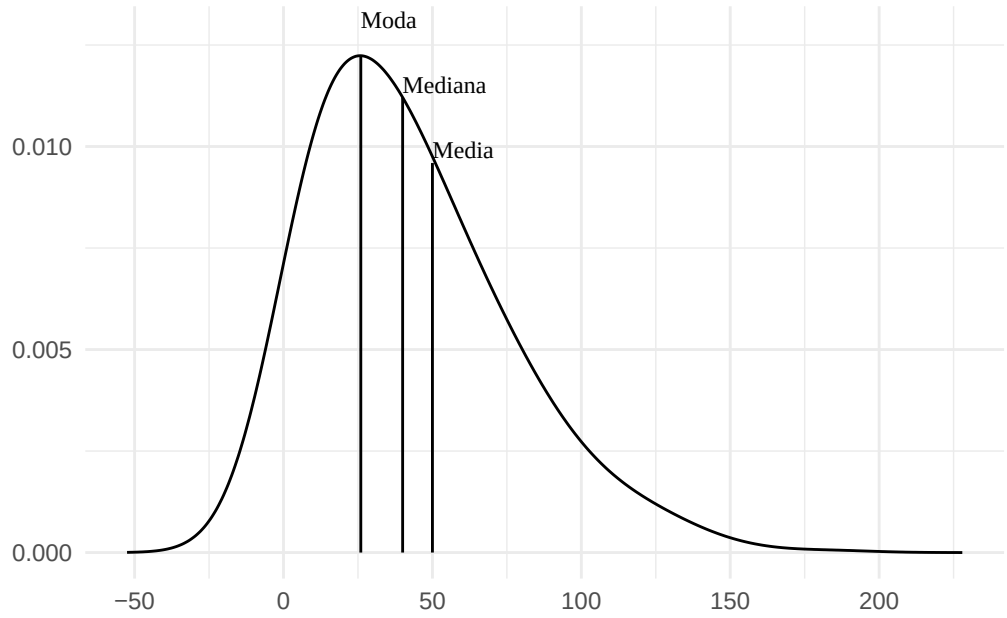


Figura 3.2: Medidas de centralización en una distribución con sesgo positivo

  - 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.

```
:
```

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

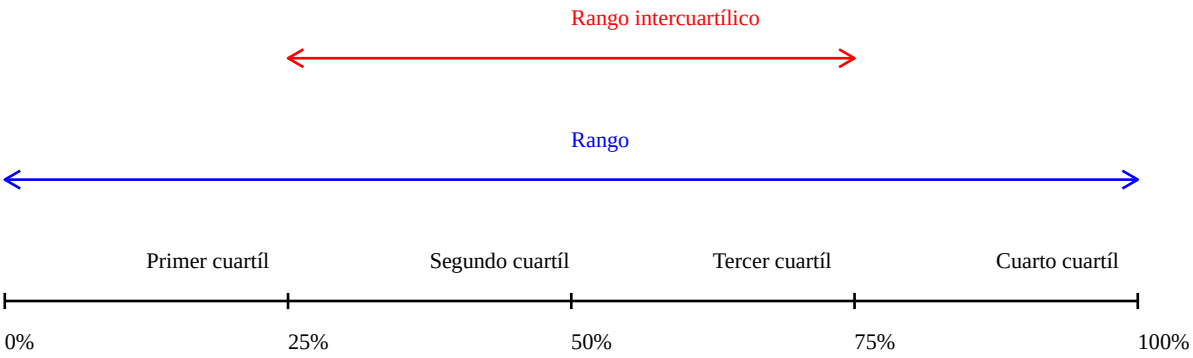


Figura 3.3: Cuartiles y rangos

```
#>    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:

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

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 la definición:

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

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:

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

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 qué llevamos todo al cuadrado?**
>
> Puede parecer enredado llevar todo al cuadrado para luego volver a sacar la raíz cuadrada. La razón es que si se resta todas las obvervaciones de la media, gran parte de estas diferencias van a ser negativas. Sabemos que un número negativo llevado al cuadrado se vuelve positivo igual que un número positivo, entonces esta parte del procedimiento sirve para que todos los valores que sumamos tengan el mismo signo positivo.

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 @ref(fig:boxplot-example).

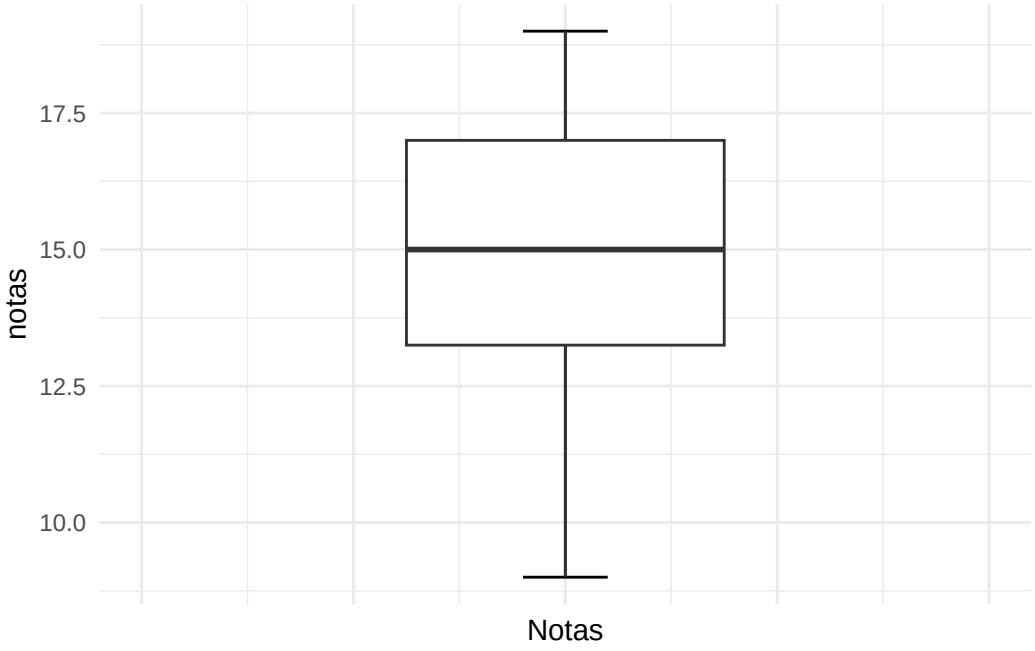


Figura 3.4: Ejemplo de box-plot

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)

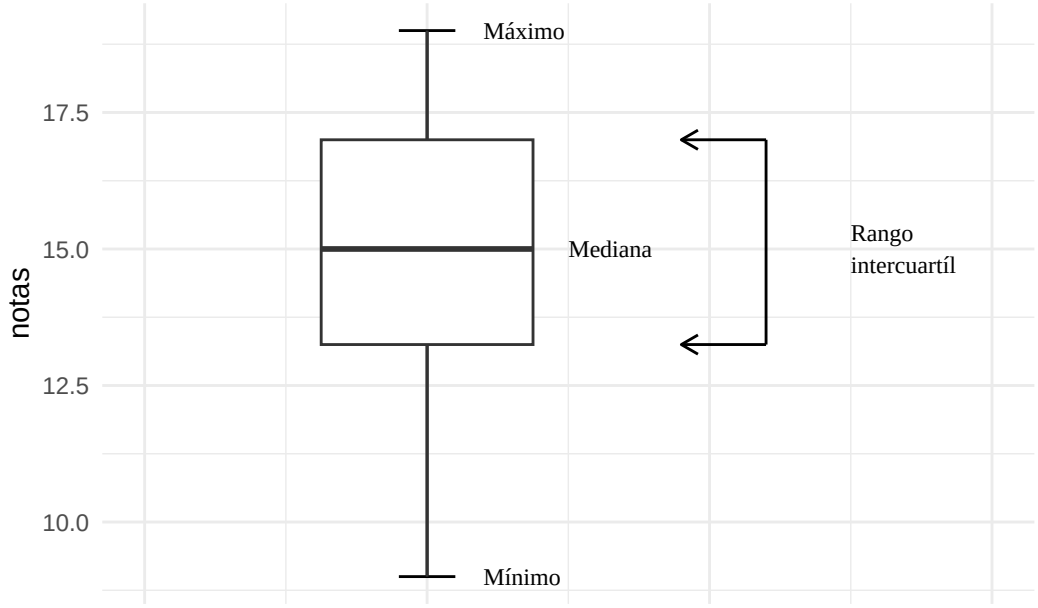
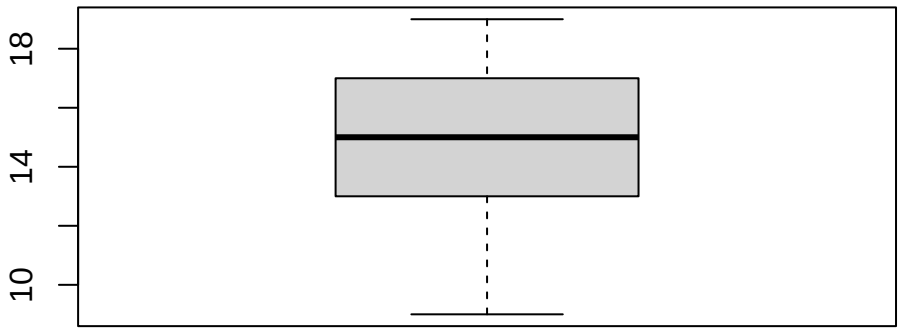


Figura 3.5: Ejemplo de box-plot con explicaciones

Ejemplo 3.3 (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».

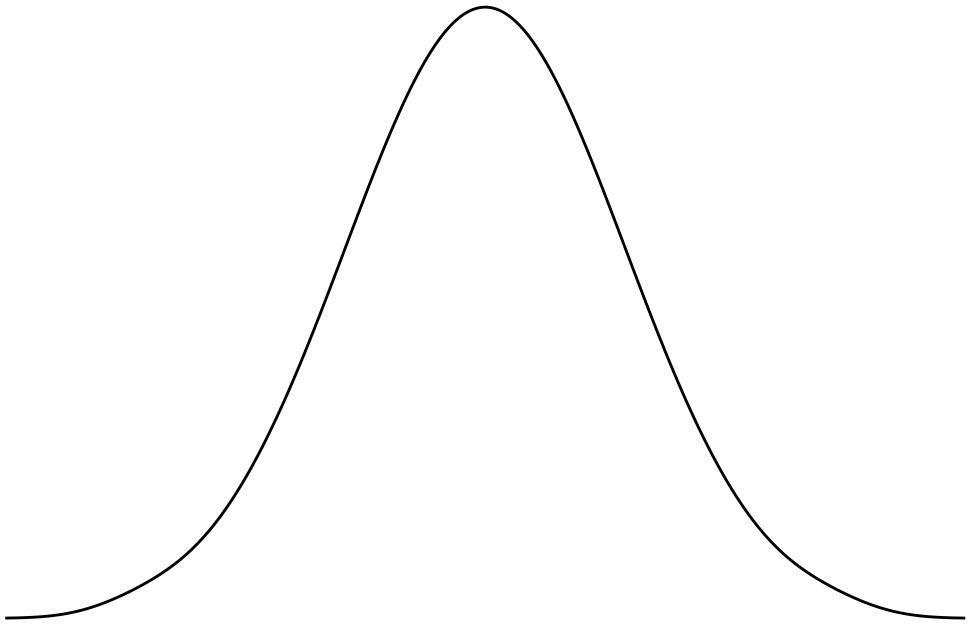
Rango La diferencia entre la mínima y la máxima de una variable. También se llama *amplitud* Función relevante en R: `range`. Equivalente en inglés: «Range».

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 @ref(perfil-de-la-distribucion), 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.

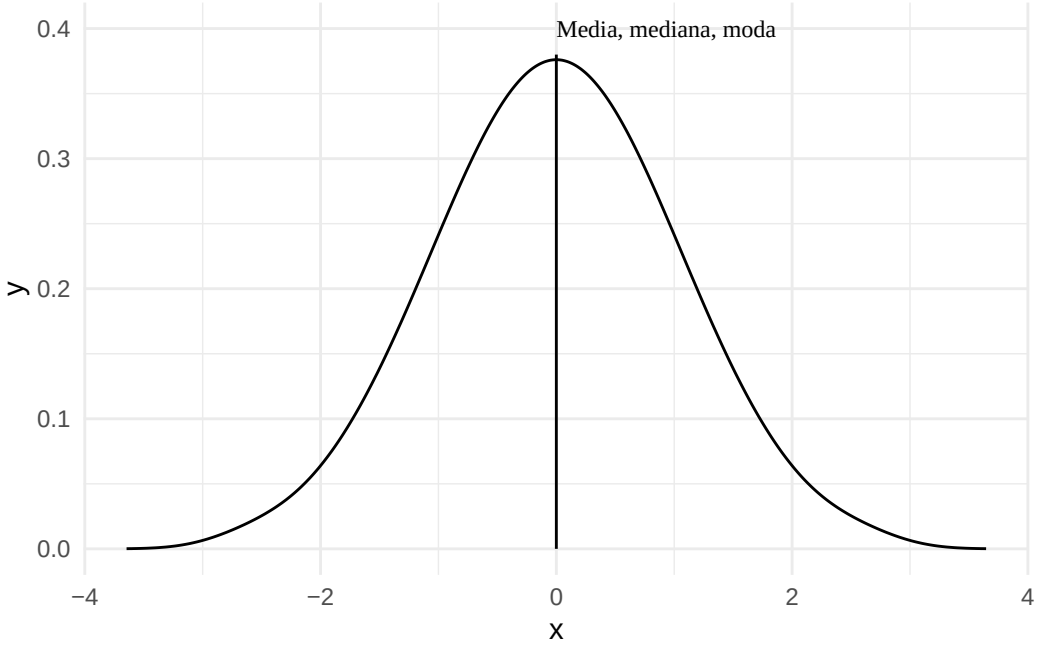


Figura 4.1: Curva normal

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 @ref(histogramas). 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.

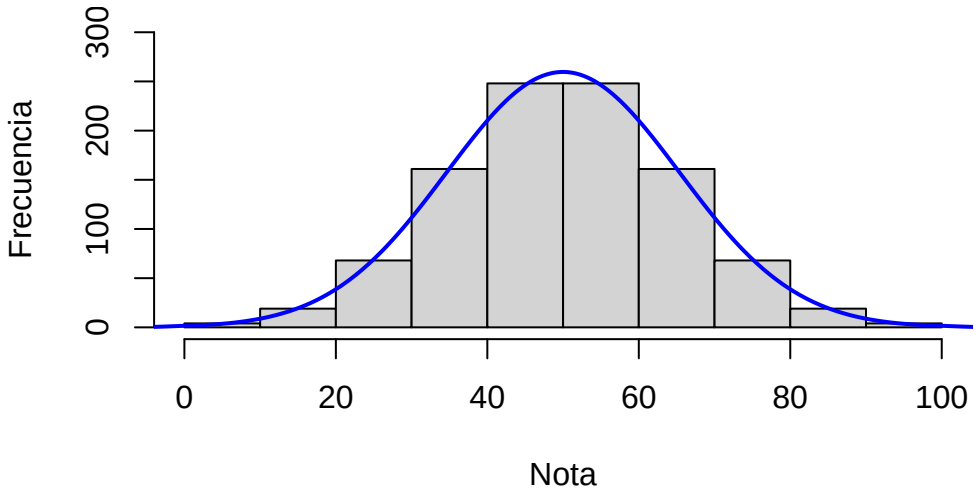


Figura 4.2: Muestra de notas de un test de matemática (N=1000)

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 .

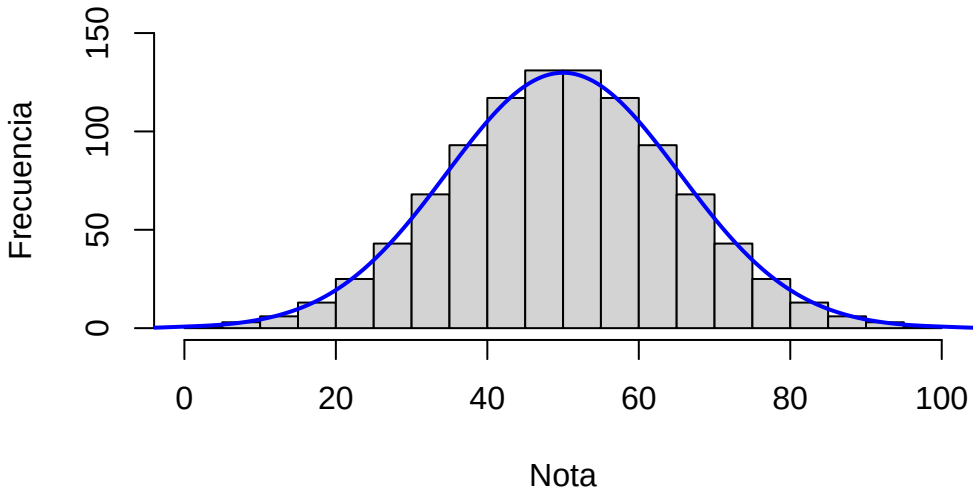


Figura 4.3: Muestra de notas de un test de matemática (N=1000)

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

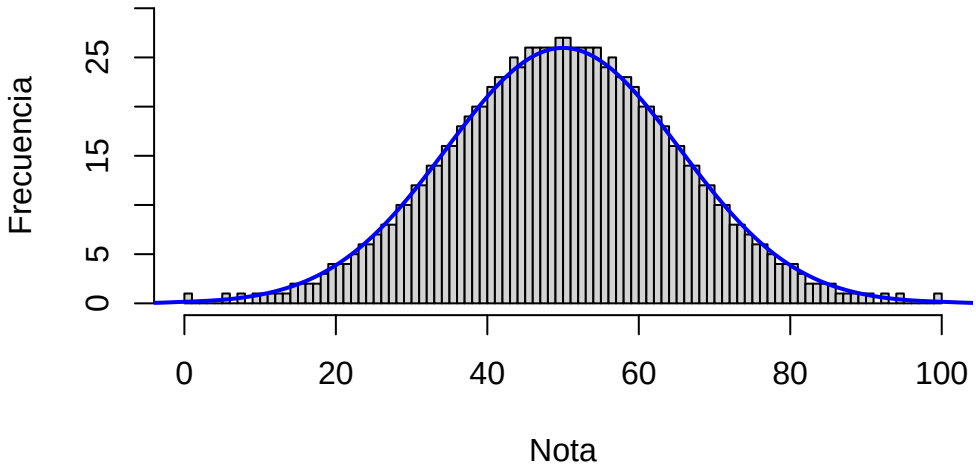


Figura 4.4: Muestra de notas de un test de matemática (N=1000)

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 **?@sec-test-de-normalidad** 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] $\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

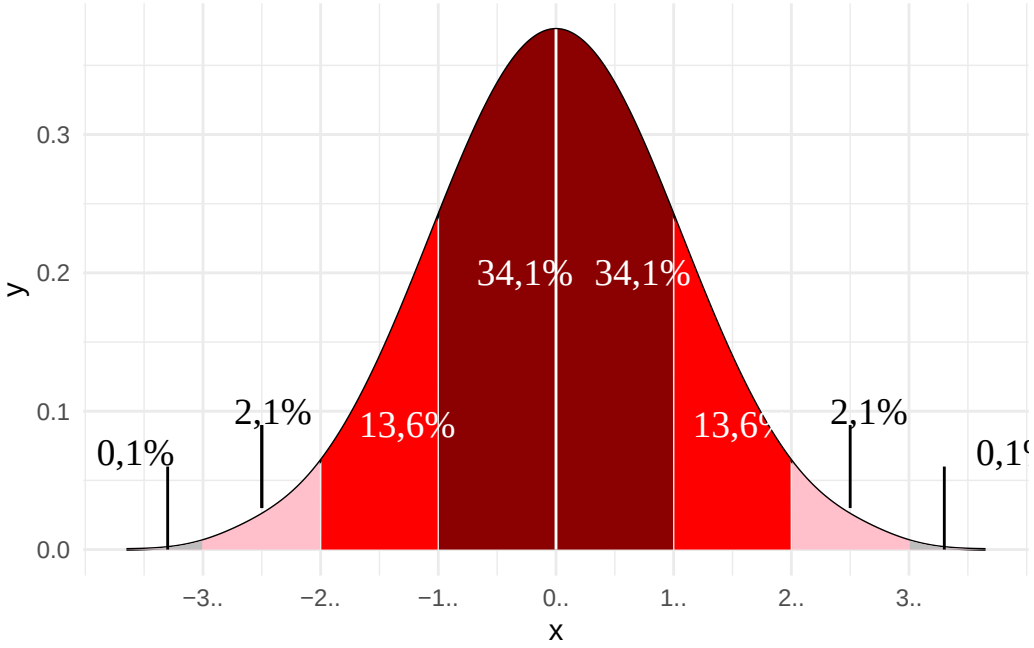


Figura 4.5: Área debajo de la curva normal

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.

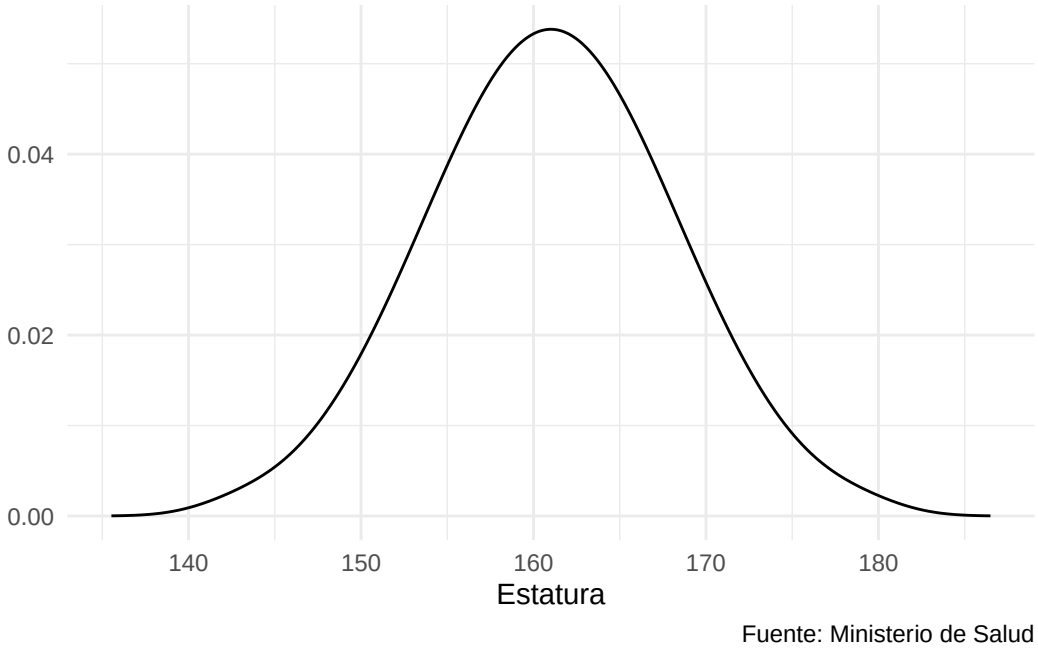


Figura 4.6: Estatura de argentinas entre 19 y 49 años

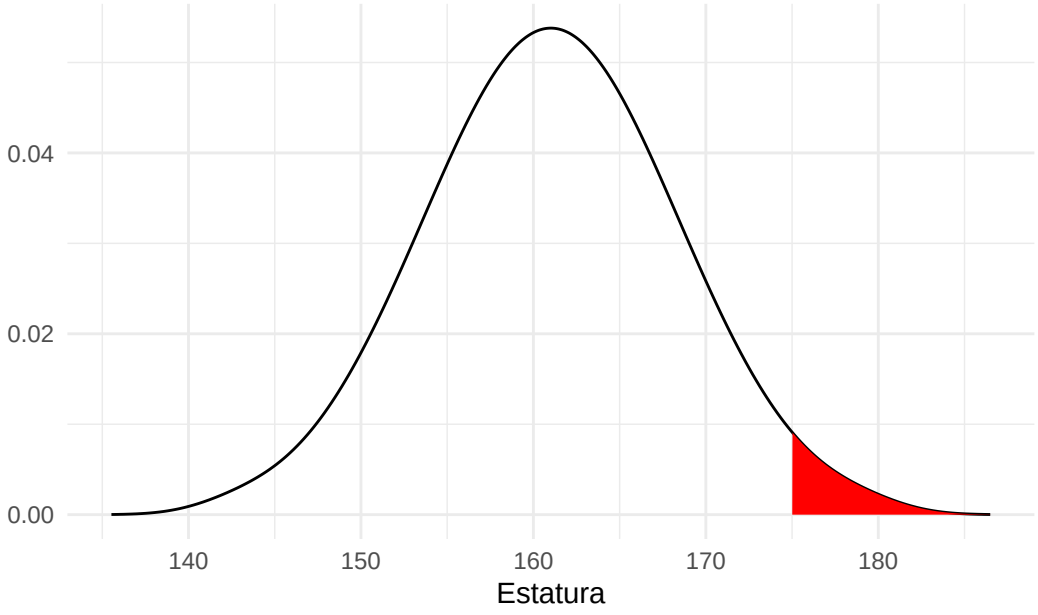
Ahora, sabiendo que esta variable tiene una distribución normal podemos saber que casi el 70% de las argentinas miden entre 154,04 y 168 centímetros. También podemos encontrar respuesta a una pregunta como: ¿qué proporción de la población femenina mide más que 175 centímetros? Para ello tenemos que normalizar 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 @ref(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. Si tomamos nuestros datos de las notas de nuestros dos cursos que vimos en la sección @ref(estadisticas-descriptivas-e-inferenciales) 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».

```
Como vimos en la sección @ref(perfil-de-la-distribucion), si tenemos muchos datos y constr

Muchos fenómenos que podemos medir tanto en las ciencias exactas como las sociales de asem

La distribución normal tiene ciertas propiedades matemáticas que nos permiten predecir qué

Varios tests de significanza de diferencia entre conjuntos de datos presumen que los datos

Propiedades de la curva normal

Como ya vimos, la curva normal tiene forma de campana y es simétrica. Por ende, las tres m

{r}
#| echo: false
#| fig.cap: 'Curva normal'
#| label: fig-curva-normal-IV


x <- distribution_normal(n = 100, mean = 0, sd = 1)
x %>%
  density() %>%  # Compute density function
  as.data.frame() %>%
  ggplot(aes(x=x, y=y)) +
  geom_line()+
  theme_minimal()+
  annotate("text",x=0,y=.4,label="Media, mediana, moda",hjust=0)+
  annotate("segment",x=0,xend=0,y=0,yend=.38)


Ciertas propiedades importantes de esta curva se relacionan con la manera en que el área d

{r histograma-ejemplo-i}
#| echo: false
```

```
#| label: fig-histograma-ejemplo-i
#| fig-cap: 'Muestra de notas de un test de matemática (N=1000)'

x <- distribution_normal(1000,mean=50,sd=15)
x <- x[which(x>=0)]
x <- x[which(x<=100)]
x %>% hist(main=NULL, ylab = "Frecuencia", xlab = 'Nota', breaks = 10, ylim=c(0,300)) ->h

multiplier <- h$counts/h$density
d <- density(x)
d$y <- d$y*multiplier[1]
lines(d, col = "blue", lwd = 2)

Ahora, bien, si en lugar de agrupar las notas en grupos de diez los podemos también agrega

{r}
#| label: fig-histograma-ejemplo-ii
#| echo: false
#| fig.cap: 'Muestra de notas de un test de matemática (N=1000)'


x %>% hist(main=NULL, ylab = "Frecuencia", xlab = 'Nota', breaks = 20, ylim=c(0,150)) ->h

multiplier <- h$counts/h$density
d <- density(x)
d$y <- d$y*multiplier[1]
lines(d, col = "blue", lwd = 2)


Podemos seguir achicando el ancho de las barras, y vemos que si bien el histograma es punt

{r}
#| label: fig-histograma-ejemplo-iii
#| echo: false
#| fig.cap: 'Muestra de notas de un test de matemática (N=1000)'


x %>% hist(main=NULL, ylab = "Frecuencia", xlab = 'Nota', breaks = 100, ylim=c(0,30)) ->h

multiplier <- h$counts/h$density
d <- density(x)
```

```
d$y <- d$y*multiplier[1]
lines(d, col = "blue", lwd = 2)


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

De más importancia son algunas propiedades que tiene la curva. Si graficamos la curva norm

{r}
#| label: fig-normal-curve-with-cuts
#| echo: false
#| warning: false
#| fig.cap: 'Área debajo de la curva normal'

x <- distribution_normal(n = 100, mean = 0, sd = 1)
x %>%
  density() %>%  # Compute density function
  as.data.frame() ->tmp
myCuts <- sd(x)*seq(-3,3,1)
names(myCuts) <- seq(-3,3,1)

tmp %>%
  ggplot(aes(x=x, y=y)) +
  geom_line()+
  geom_area(position = "identity", data=tmp %>%  filter(between(x,myCuts[3],myCuts[5])),fi
  geom_area(position = "identity",
            data = tmp %>%
              filter(between(x,myCuts[2],myCuts[3])),
            fill='red')+
  geom_area(position = "identity", data = tmp %>% filter(between(x,myCuts[5],myCuts[6])),
  geom_area(position = "identity", data = tmp %>% filter(x<myCuts[2]), fill='pink')+
  geom_area(position = "identity", data = tmp %>% filter(x>myCuts[6]), fill='pink')+
  geom_area(position = "identity", data = tmp %>% filter(x>myCuts[7]), fill='gray')+
  geom_area(position = "identity", data = tmp %>% filter(x<myCuts[1]), fill='gray')+
  ## Add vertical line for mean
  annotate("segment",x=0,xend=0,y=0,yend=max(tmp$y), color='white')+
  annotate("text",x=c(myCuts[3],myCuts[4]),y=c(.2,.2),label="34,1%",hjust=-.4, color="whit
  annotate("text",x=c(myCuts[2],myCuts[5]),y=c(.09,.09),label="13,6%",hjust=c(-.4,-.2), co
  annotate("text",x=c(myCuts[1],myCuts[6]),y=c(.1,.1),label="2,1%",hjust=c(-.4,-.5), cex=5
  annotate("text",x=c(min(x),max(x)),y=c(.07,.07),label="0,1%",hjust=c(2,-1.5), cex=5)+
  theme_minimal()+
```

```
  annotate("segment",x=sd(x)*-2.5,xend=sd(x)*-2.5,y=0.03,yend=.09)+
  annotate("segment",x=sd(x)*2.5,xend=sd(x)*2.5,y=0.03,yend=.09)+
  annotate("segment",x=sd(x)*-2.5,xend=sd(x)*-2.5,y=0.03,yend=.09)+
  annotate("segment",x=sd(x)*-3.3,xend=sd(x)* -3.3,y=0,yend=.06)+
  annotate("segment",x=sd(x)* 3.3,xend=sd(x)*  3.3,y=0,yend=.06)+
  scale_x_continuous(breaks=myCuts,labels=paste0(names(myCuts),"\u03C3"))
#annotate("text",x=0,y=.4,label="Media, mediana, moda",hjust=0)+
#annotate("segment",x=0,xend=0,y=0,yend=.38)
```

Esta propiedad es de bastante utilidad y se puede aprovechar de varias maneras. Si tenemos

Variables normalizadas

En textos de estadística frecuentemente se habla de variable normalizada, también se conoc

Variable normalizada

La variable normalizada z de un conjunto de datos X se obtiene por la fórmula siguiente:

```
$$
  z = {x-\bar{x}\over{\sigma}}
$$
```

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 c

Normalizar las variables nos permite comparar su distribución independientemente de su uni

Analizando datos del ministerio de salud

En el 2007 el Ministerio de Salud de Argentina realizó un estudio [@ministerioDeSalud:2007

```
{r}
#| label: fig-curva-estatura-argentinas
#| echo: false
#| fig.cap: "Estatura de argentinas entre 19 y 49 años"


estatura <- distribution_normal(n = 100, mean = 161.01, sd = 6.99) %>%
  density() %>%  # Compute density function
  as.data.frame()
estatura %>%
  ggplot(aes(x=x, y=y)) +
  geom_line()+
  theme_minimal() +
   labs(caption="Fuente: Ministerio de Salud",y=NULL,x="Estatura") -> plot_estatura
plot_estatura

Ahora, sabiendo que esta variable tiene una distribución normal podemos saber que casi el

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

Con este número podemos volver a la @fig-normal-curve-with-cuts y fijarnos que con por arr

{r}
#| 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')

En este caso tuvimos un poco se suerte ya que la variable normalizada resultó un número re

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

Con este número podemos sacar la proporción por ejemplo calculando el área debajo del segm
```

```{r}
pnorm(-1.575)
```

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

También podemos expresar esto en términos de probabilidades: Si medimos una mujer argentin

Evaluar la normalidad

Hemos visto que el hecho de que una variable tenga una distribución normal nos resulta muy

En la sección @ref(cual-usar) decidimos usar la media como medida de centralización porque

También podemos valernos del conocimiento de la proporción de observaciones que deben esta

Si tomamos nuestros datos de las notas de nuestros dos cursos que vimos en la sección @ref

Grupo A: {15, 12, 11, 18, 15, 15, 9, 19, 14, 13, 11, 12, 18, 15, 16, 14, 16, 17, 15, 17, 1

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

Grupo A:

Media: 14.93

Desviación estándar: 2,49

Entre $\pm{1}$ desviación: 66%

Entre $\pm{2}$ desviaciones: 96%

Grupo B:

Media: 11,76

Desviación estándar: 3,31

Entre $\pm{1}$ desviación: 66%

Entre $\pm{2}$ desviaciones: 96%

```
Observamos que nuestras notas carecen en cierta medida de valores extremos, sin embargo la

Ejemplo en R

Si no queremos hacer estos cálculos a mano los podemos hacer también en R, así:

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


Existen también tests más formales de normalidad que desarrollaremos en capítulos posterio

Glosario

{r, echo=FALSE, results='asis'}
.format_dataframe_as_def_list(chapter = 4)
```

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 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.

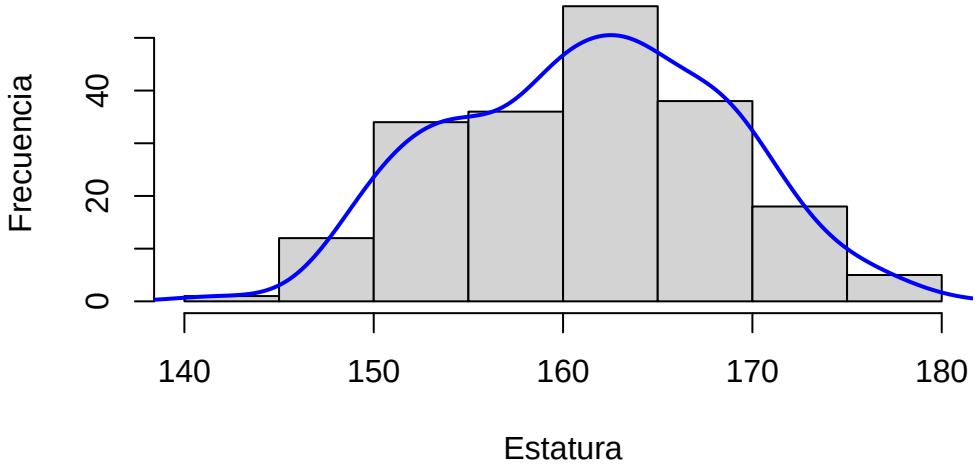


Figura 5.1: Cinco Muestras de 20 obseraciones

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 @ref(fig:curva-normal-poblacion-muestral).

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»)

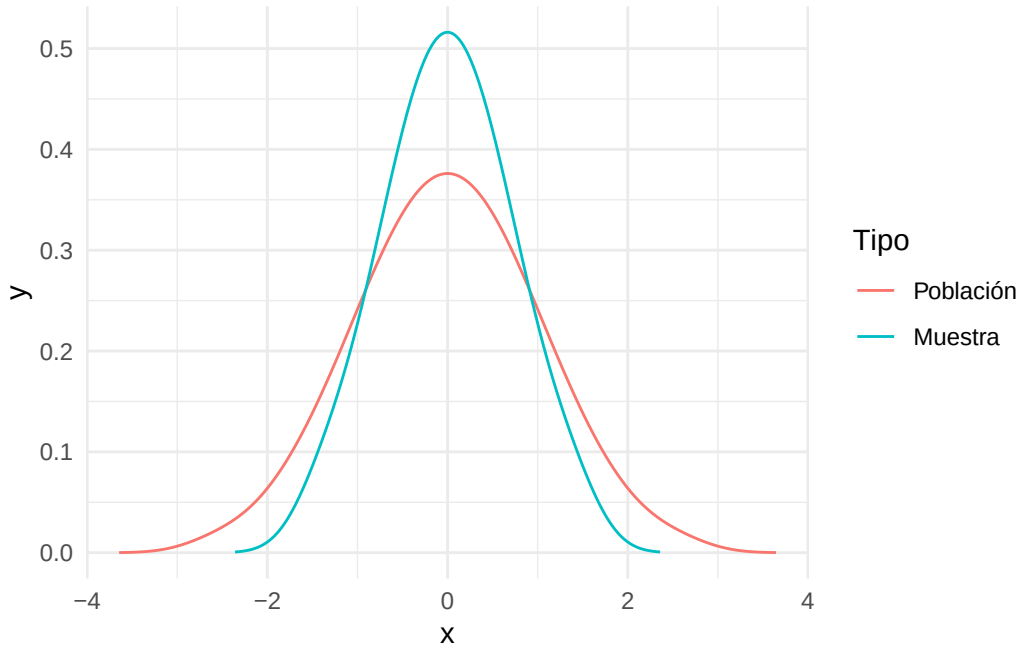


Figura 5.2: Distribución de la población y la muestra

- σ: la desviación estándar de la población
- 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 @ref(fig:normal-curve-with-cuts) 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 @ref(exm:small-sample) 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 A][Apendix A: distribución t] 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».

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.

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.