Unidad 1 / Escenario 2

Lectura Fundamental

Estadística descriptiva

Contenido

Palabras Claves: estadística descriptiva, descripción tabular, descripción gráfica, descripción numérica.

Resumen

En este documento puede encontrar una descripción de la estadística descriptiva que permitirá entender los conceptos básicos. Durante la lectura encontrará las formas básicas de cómo se describe la información de forma numérica (utilizando estadísticas), de forma tabular y de forma gráfica. Todo lo anterior implementando la herramienta estadística R.

Introducción

¿Qué es la estadística?

Stuart y Ord (1991) definen la estadística como "la rama del método científico relacionada con la recopilación de los datos que se obtienen al medir las propiedades de las poblaciones". Por otra parte, Gutiérrez Cabría (1994) explica el concepto como la ciencia que estudia cómo debe emplearse la información y cómo dar una guía de acción en situaciones prácticas que entrañan incertidumbre. En ese mismo orden de ideas, Ross (2007) determina que la estadística es el arte de aprender a partir de los datos. Es decir, la disciplina relacionada con la recopilación, descripción y análisis de datos para tomar conclusiones. Finalmente, Chao (1975) afirma lo siguiente acerca de la estadística: Es un conjunto de teorías y método que han sido desarrolladas para tratar la recolección, el análisis y la descripción de datos muestrales con el fin de extraer conclusiones útiles. Su función primordial es apoyar al investigador al decidir sobre el parámetro de la población de que procede la muestra.

¿Para que sirve la estadística?

El uso de la estadística les permite a las personas y empresas tomar decisiones correctas sobre el uso de su información y desarrollar destrezas como las siguientes: *Presentar y describir la información de forma adecuada. *Inferir conclusiones sobre poblaciones grandes basándose solamente en muestras. *Utilizar modelos para realizar pronósticos acertados.

Clasificación de la estadística

Estadística descriptiva

Se compone de aquellos métodos que incluyen técnicas para recolectar, presentar, analizar e interpretar datos.

Estadística inferencial

Abarca el conjuntos de técnicas que se utilizan para obtener información sobre el comportamiento de una población, basados en los datos de muestras tomadas a partir de ella.

Definiciones básicas

Población

Se refiere al conjunto total de objetos que tienen una característica en común. Esta característica es de interés para un problema dado, por ejemplo:

- Total de niños que están estudiando bachillerato en un País.
- Número de establecimientos educativos en un País.
- Número de hogares en un departamento.

Muestra

Hace alusión al subconjunto finito de una población. Ejemplo: si la población fuera de todos los estudiantes que estudian en un establecimiento educativo universitario, entonces el número de estudiantes en un establecimiento educativo en particular sería una muestra.

Elementos o individuos

Seres u objetos (en general, unidades experimentales) que contienen la información que se desea estudiar. Los objetos pueden ser personas, animales, productos, etc.

Variable

Representa la característica de la población observable que se analiza en el estudio estadístico.

Datos u observaciones

Son números o denominaciones que podemos asignar a un individuo o elemento de la población.

Ejemplo Respuestas a las siguiente variables: Edad de las personas, Estatura, Peso, sexo de las personas, situación laboral, etc.

Parámetro

Es cualquier característica medible de una población y sirve para sintetizar la información dada por una tabla o por un gráfico.

Ejemplo Ingreso promedio de los estudiantes universitarios que laboran.

Estadístico

Hace referencia a cualquier característica medible de una muestra.

Por ejemplo, el ingreso promedio de los estudiantes universitarios que laboran en cierta facultad de un establecimiento educativo (si la muestra son todos los estudiantes que laboran de cierta universidad).

Censo

Palabra derivada del latín censere, que significa "valuar o tasar". Es la enumeración completa de la población.

Tipos de datos

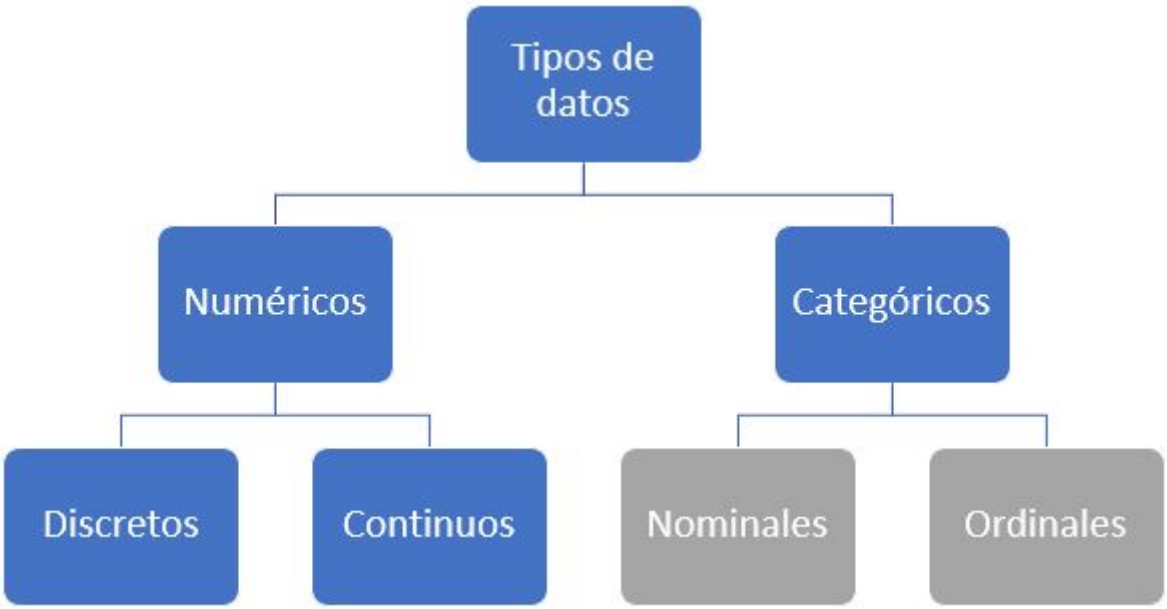


Figura 1:

Datos numéricos o cuantitativos

Producen respuestas numéricas que pueden ser discretas o continuas.

Datos discretos

Generan respuestas numéricas que surgen de un conteo.

Ejemplo Edad, número de estudiantes, número de hijos, Estatura, saldo en una cuenta de cheques, minutos que faltan para que termine la clase, número de niños en una familia.

Datos continuos

Producen respuestas numéricas que surgen de un proceso de medición, donde la característica de que se mide puede tomar cualquier valor numérico en un intervalo.

Datos categóricos o cualitativos

Representan categorías o atributos que pueden clasificarse según escala en datos nominales u ordinales.

Datos nominales

Se crean cuando se utilizan nombres para establecer categorías con la condición de que cada dato pertenezca exclusivamente a una de estas categorías. No poseen orden, distancia, origen.

Ejemplo sexo, estado civil, gusto musical, color que lugar de nacimiento, color de los ojos.

Datos ordinales

Son datos medidos en una escala nominal ordenada de alguna manera cuyas categorías, cada una de las cuales indica un nivel distinto respecto a un atributo que se está midiendo. Solo poseen la propiedad de orden.

Ejemplo Nivel de estudio, Rangos de trabajo, Evaluación docente.

Ejercicios

1. En una encuesta de *The Wall Street Journal* (13 de octubre de 2003) se les hacen a los suscriptores 46 preguntas acerca de sus características e intereses. De cada una de las preguntas siguientes, Cuáles proporcionan datos cualitativos o cuantitativos e indica el tipo de dato.
    - ¿Cuál es su edad?
    - ¿Es usted hombre o mujer?
    - ¿Cuándo empezó a leer el WSJ? Preparatoria, universidad al comienzo de la carrera, a la mitad de la carrera, al final de la carrera o ya retirado.
    - ¿Cuánto tiempo hace que tiene su trabajo o cargo actual?
    - ¿Qué tipo de automóvil piensa comprarse la próxima vez que compre uno? Ocho categorías para las respuestas, entre las que se encontraban sedán, automóvil deportivo, miniván, etcétera.
2. Diga de cada una de las variables siguientes si es cualitativa o cuantitativa e indique el tipo de dato.
    - Ventas anuales.
    - Tamaño de los refrescos (pequeño, mediano, grande).
    - Clasificación como empleado (GS 1 a GS 18).
    - Ganancia por acción.

- Modo de pago (al contado, cheque, tarjeta de crédito).

3. Se realizó una encuesta para estudiar los hábitos de fumar de los residentes del Reino Unido[1]. A continuación descargue la base de datos Smoking, para lo cual puede utilizar las siguiente librerias (recuerde instalarlas si no las tiene intaladas)

```
library(xlsx)
```

```
## Loading required package: rJava
```

```
## Loading required package: xlsxjars
```

```
db <- read.xlsx2("C:/Tareas/11263-Smoking_tcm86-13253.xls",sheetIndex = "Smoking
```

Las primeras cinco observaciones son las siguientes

```
knitr::kable(head(db))
```

Sex	Age	Marital.Status	Highest.Qualification.	Nationality	Ethnicity	Gross.Income	Region
Male	38	Divorced	No Qualification	British	White	£2600 to less than £5200	The No
Female	42	Single	No Qualification	British	White	Less than £2600	The No
Male	40	Married	Degree	English	White	£28600 to less than £36400	The No
Female	40	Married	Degree	English	White	£10400 to less than £15600	The No
Female	39	Married	GCSE/O Level	British	White	£2600 to less than £5200	The No
Female	37	Married	GCSE/O Level	British	White	£15600 to less than £20800	The No

- ¿Qué representa cada fila de los datos?
- ¿Cuántos participantes respondieron la encuesta?
- Indique si cada variable en el estudio es numérica o categórica. Si es numérico, indique si es continuo o discreto. Si es categórico, indique si es nominal u ordinal.

[1] https://www.stem.org.uk/resources/elibrary/resource/28452/large-datasets-stats4schools

Distribución de frecuencias

Es un resumen tabular de datos que muestra el número (frecuencia) de elementos en cada una de las diferentes clases disyuntas (que no se sobreponen).

Algunos tipos de Tipos de distribución de frecuencias son

- Absoluta.
- Relativa.
- Absoluta acumulada.
- Relativa acumulada.

Distribución de frecuencia (Datos Cualitativos)

A continuación se describe en forma general la forma en que se debe presentar una tabla de distribución para datos cualitativos

Variable(X)	Frecuencia absoluta (f_i)	Frecuencia relativa (h_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa acumulada (H_i)
Categoría 1	f_1	h_1	F_1	H_1
Categoría 2	f_2	h_2	F_2	H_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Categoría k	f_k	h_k	n	1
Total	n	1		

donde $h_i = \frac{f_i}{n}$, $F_i = f_1 + f_2 + \cdots + f_i$, $H_i = h_1 + h_2 + \cdots + h_i$.

Ejemplo A continuación se les pregunto a cincuenta personas que tipo de gaseosa le gusta. Las respuestas fueron las siguientes:

```
##  [1] "Coca-Cola"  "Coca-Cola"  "Postobón"   "Colombiana" "Coca-Cola"
##  [6] "Big-Cola"   "Colombiana" "Postobón"   "Postobón"   "Coca-Cola"
## [11] "Coca-Cola"  "Coca-Cola"  "Postobón"   "Coca-Cola"  "Pepsi-Cola"
## [16] "Postobón"   "Postobón"   "Sprite"     "Coca-Cola"  "Pepsi-Cola"
## [21] "Colombiana" "Coca-Cola"  "Postobón"   "Coca-Cola"  "Coca-Cola"
## [26] "Coca-Cola"  "Coca-Cola"  "Coca-Cola"  "Big-Cola"   "Coca-Cola"
## [31] "Coca-Cola"  "Postobón"   "Postobón"   "Coca-Cola"  "Pepsi-Cola"
## [36] "Postobón"   "Pepsi-Cola" "Coca-Cola"  "Postobón"   "Coca-Cola"
## [41] "Pepsi-Cola" "Postobón"   "Pepsi-Cola" "Postobón"   "Postobón"
## [46] "Pepsi-Cola" "Coca-Cola"  "Coca-Cola"  "Postobón"   "Postobón"
```

Inicialmente, hacemos un tabla con los datos anteriores, donde muestra es el objeto que contiene los datos anteriores

```
tabla <- table(muestra)
tabla
```

```
## muestra
##   Big-Cola  Coca-Cola Colombiana Pepsi-Cola   Postobón     Sprite
##          2         21          3          7         16          1
```

Mejorando el resumen podemos utilizar la función `pareto.chart` de la librería `qcc`

```
res <- qcc::pareto.chart(data = tabla,plot = F)
res
```

```
##
## Pareto chart analysis for tabla
##              Frequency Cum.Freq. Percentage Cum.Percent.
##   Coca-Cola         21        21         42           42
##   Postobón          16        37         32           74
##   Pepsi-Cola         7        44         14           88
##   Colombiana         3        47          6           94
##   Big-Cola           2        49          4           98
##   Sprite             1        50          2          100
```

Algunos análisis de los resultados anteriores tenemos lo siguiente:

- De las personas encuestadas, la gaseosa que más se consume es la coca-cola.
- El 88 % de las personas seleccionaron una de estás opciones: la Coca-Cola o la Postobón o La Pepsi-Cola.
- De las 50 personas encuestadas solo 3 escogieron la gaseosa Colombiana.

Representaciones gráficas para datos cualitativos

Gráfico de barras

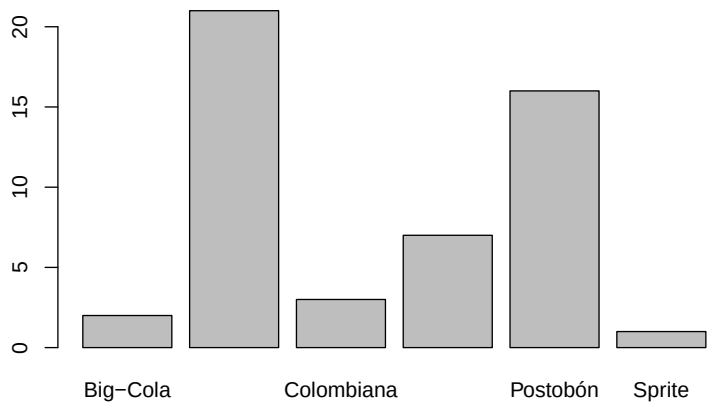
para representar datos cualitativos que hayan sido resumidos en una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual.

Gráfica circular (pastel)

datos resumidos mediante una distribución de frecuencia relativa y se basa en la subdivisión de un círculo en sectores que corresponden a la frecuencia relativa de las clases.

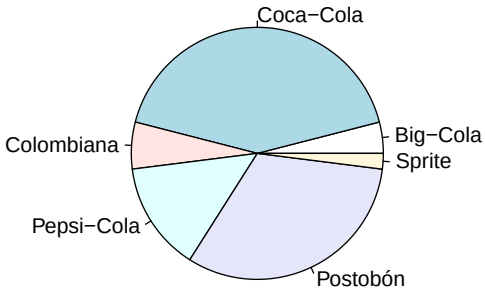
Ejemplo A continuación representaremos los resultados de la tabla númerica, utilizando la función `barplot` de la siguente manera:

```
barplot(tabla)
```



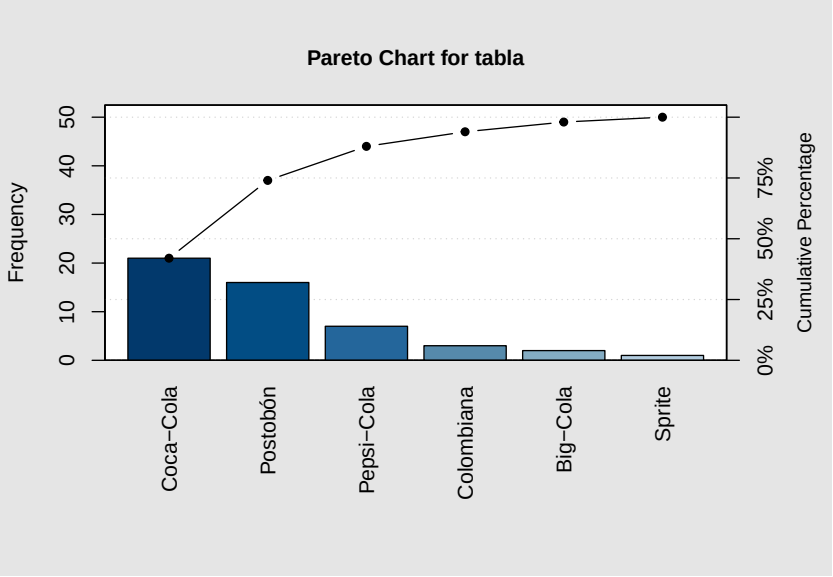
Ahora representando la misma tabla de distribución de frecuencias a partir de un gráfico circular utilizando la función `pie` tenemos lo siguiente:

```
pie(tabla)
```



Ahora representando la misma tabla de distribución de frecuencias a partir de un gráfico pareto habilitando la función de gráfica tenemos lo siguiente:

```
res <- qcc::pareto.chart(data = tabla,plot = F)
plot(res)
```



Distribución de frecuencia (Datos Cuantitativos)

Datos discretos

Los datos discretos son valores enteros, por lo cual una forma de describir la información utilizando una distribución de frecuencias es la siguiente:

Variable(X)	Frecuencia absoluta (f_i)	Frecuencia relativa (h_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa acumulada (H_i)
Categoría 1	f_1	h_1	F_1	H_1
Categoría 2	f_2	h_2	F_2	H_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Categoría k	f_k	h_k	n	1
Total	n	1		

Ejemplo A continuación se le pregunta a 40 hogares el número de personas en el hogar, los resultados son los siguientes:

```
##  [1] 2 2 3 5 2 5 5 4 4 1 2 1 4 2 4 3 4 5 2 4 5 2 4 1 2 2 1 2 5 2 3 3 3 1 5
## [36] 4 4 1 4 3
```

Contuyendo inicialmente hacemos un conteo de cada uno de los valores

```
tabla <- table(X)
```

Ahora construimos la tabla de distribución de frencuencias utilizando el siguiente codigo:

```
conteos <- c(tabla)
res <- data.frame(Frequency= conteos, Cum.Freq. = cumsum(conteos), Percentage = round
res
```

```
##   Frequency Cum.Freq. Percentage Cum.Percent.
## 1         6         6         15           15
## 2        11        17         28           43
## 3         6        23         15           57
## 4        10        33         25           82
## 5         7        40         18          100
```

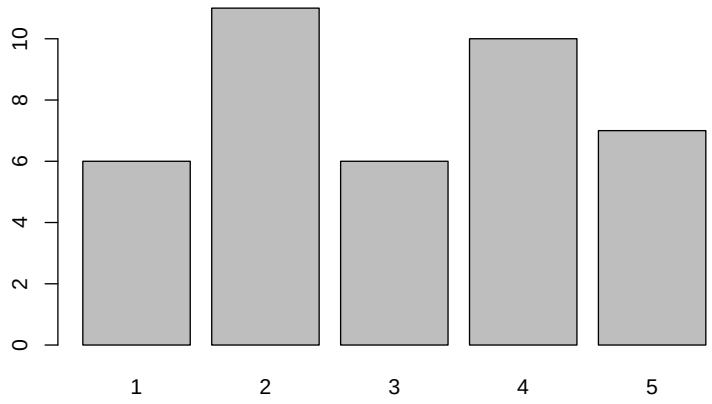
Representación gráfica para datos discretos

Gráfico de barras.

El diagrama de barras no es más que una representación gráfica de un resumen de una tabla de distribución de frecuencias, de frecuencia relativa o de frecuencia porcentual.

Ejemplo A continuación representaremos los resultados de la tabla anterior, utilizando la función `barplot` de la siguente manera:

```
barplot(tabla)
```



Datos continuos

Intervalos de clase	Marca de clase (x_i)	Frecuencia absoluta (f_i)	Frecuencia relativa (h_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa acumulada (H_i)
$[L_0 - L_1)$	x_1	f_1	h_1	F_1	H_1
$[L_1 - L_2)$	x_2	f_2	h_2	F_2	H_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_{k-1} - L_k)$	x_k	f_k	h_k	n	1
Total		n	1		

donde $x_i = \frac{\text{límite inf}_i + \text{límite sup}_i}{2}$, $h_i = \frac{f_i}{n}$, $F_i = f_1 + f_2 + \cdots + f_i$, $H_i = h_1 + h_2 + \cdots + h_i$.

Pasos para distribución de frecuencias datos cuantitativos

A continuación de describen los pasos para la construcción de una tabla de distribución de frecuencias

1. Determine el rango de los datos $R = X_{\text{máx}} - X_{\text{mín}}$.
2. Seleccione el número de clase o intervalos K.
3. Calcule la amplitud de los intervalos redondeando por encima $C = \frac{R}{K}$.
4. Realice el conteo de las observaciones en cada intervalo.

Algunas sugerencias para para determinar el número de intervalos son los siguientes:

- Criterio del investigador. Si el investigador conoce el comportamiento de los datos, será el quien defina cuantos intervalos utilizar para agrupar los datos.
- Criterio de la regla de Sturges $K = 3.322 \log_{10}(n) + 1$.
- Atendiendo al número de observaciones

Número de observaciones	K
Menos de 50	5 - 7
De 50 a 100	7 - 8
De 100 a 500	8 - 10
De 500 a 1000	10 - 11
De 1000 a 5000	11 - 14
Más de 5000	14 - 20

Ejemplo A continuación se tiene las estaturas en metros de 50 estudiantes

```
##  [1] 161 173 157 194 175 158 177 181 179 165 193 176 161 137 187 169 170
## [18] 184 182 179 184 182 171 140 179 169 168 148 163 176 190 168 176 169
## [35] 149 164 164 169 187 181 168 166 180 178 160 159 175 182 168 183
```

Calculando en Rango en R tenemos lo siguiente

```
X <- datos
max(X) # Altura máxima
```

```
## [1] 194
```

```
min(X) # Altura mínima
```

```
## [1] 137
```

```
R <- max(X)-min(X) # Rango
R
```

```
## [1] 57
```

Para determinar el número de intervalos entonces utilizamos la función `nclass.Sturges`, con el cual se determina el número de intervalos

```
K <- nclass.Sturges(X)
K
```

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

Calculamos la amplitud y redondeamos por encima está cantidad

```
C <- R/K # Amplitud
C
```

```
## [1] 8.142857
```

```
C <- ceiling(C)
C
```

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

Ahora calculamos los intervalos, los cuales seran cerrados por derecha y abiertos por izquierda y realizamos los conteos respectivos.

```
cortes <- seq(from= min(X),to = max(X)+C,by = C)
tabla <- table(cut(x = X,breaks = cortes,include.lowest = T,right = F))
tabla
```

```
##
## [137,146) [146,155) [155,164) [164,173) [173,182) [182,191) [191,200]
##         2         2         7        14        14         9         2
```

Ahora construyendo la tabla de la distribución de frecuencias

```
conteos <- c(tabla)
res <- data.frame(Frequency= conteos, Cum.Freq. = cumsum(conteos), Percentage = round
res
```

```
##           Frequency Cum.Freq. Percentage Cum.Percent.
## [137,146)         2         2          4            4
## [146,155)         2         4          4            8
## [155,164)         7        11         14           22
## [164,173)        14        25         28           50
## [173,182)        14        39         28           78
## [182,191)         9        48         18           96
## [191,200]         2        50          4          100
```

Representación gráfica para datos continuos

Histograma.

Representación gráfica de una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual que

se construye colocando los intervalos de clase sobre un eje horizontal y la frecuencia, la frecuencia relativa o la frecuencia porcentual sobre un eje vertical.

Ojiva

Gráfica de una distribución acumulada.

Diagrama de tallo y hojas

Técnica para el análisis exploratorio de datos que tanto ordena por jerarquía datos cuantitativos como proporciona claridad acerca de la forma de la distribución. La idea es expresar los datos en tallos y hojas. Para ello, los datos se expresan en fracciones decimales donde el denominador sea potencia de diez, por ejemplo, si el dato es 123 se escribe $\frac{123}{10} = 12.3$. La última cifra del noumerador se llama *hoja* y el resto *tallo*.

Ejemplo A continuación representaremos los resultados de la tabla anterior, utilizando la función `hist`, los cortes realizados anteriormente. La gráfica queda de la siguente manera:

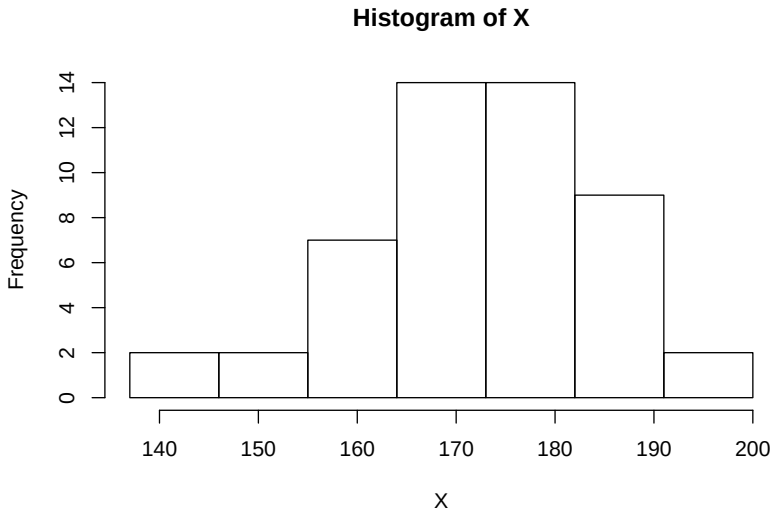
```
cortes
```

```
## [1] 137 146 155 164 173 182 191 200
```

```
tabla <- table(cut(x = X,breaks = cortes,include.lowest = T,right = F))
tabla
```

```
##
## [137,146) [146,155) [155,164) [164,173) [173,182) [182,191) [191,200]
##         2         2         7        14        14         9         2
```

```
hist(X,breaks = cortes,include.lowest = T,right = F)
```



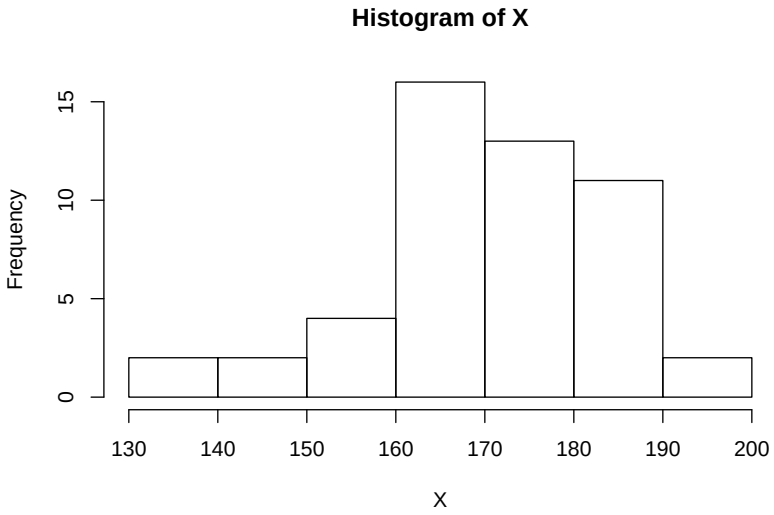
Ahora si queremos utilizar el histograma que realiza por defecto la función `hist`, debemos tener presente que los cortes lo realiza mediante la función `pretty`, la cual realiza los cortes en potencias de 1, 2, 5 o potencias de 10

dependiendo de la forma de los datos. A continuación describimos los cortes que realiza y los respectivos conteos, utilizando la función por defecto.

```
cortes <- pretty(X)
tabla <- table(cut(x = X,breaks = cortes))
tabla
```

```
##
## (130,140] (140,150] (150,160] (160,170] (170,180] (180,190] (190,200]
##         2         2         4        16        13        11         2
```

```
hist(X)
```



Para hacer un diagrama de ojiva se constuira mediante el siguiente codigo

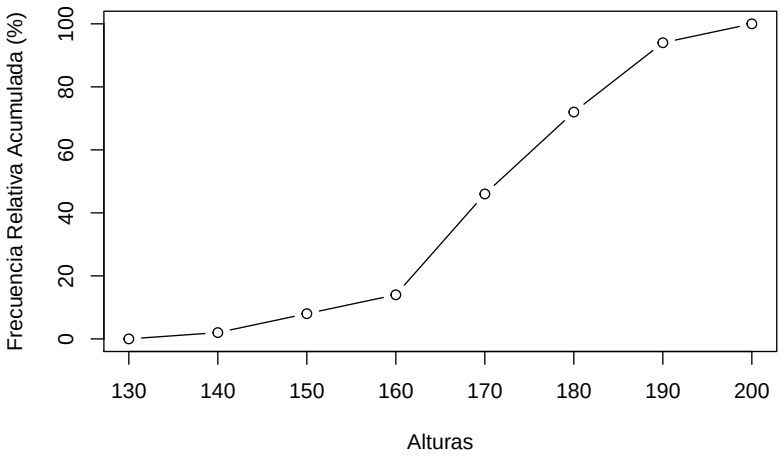
```
cortes
```

```
## [1] 130 140 150 160 170 180 190 200
```

```
tabla <- table(cut(x = X,breaks = cortes,include.lowest = T,right = F))
round(tabla/sum(tabla)*100)
```

```
##
## [130,140) [140,150) [150,160) [160,170) [170,180) [180,190) [190,200]
##         2         6         6        32        26        22         6
```

```
y <- c(0,cumsum(round(tabla/sum(tabla)*100)))
x <- cortes
plot(x,y,type="b",xaxt="n",ylab = "Frecuencia Relativa Acumulada (%)",xlab = "Alturas
        axis(1,x)
```



La función `stem` permite construir el diagrama de tallos y hojas

```
stem(X)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   13 | 7
##   14 | 089
##   15 | 789
##   16 | 0113445688889999
##   17 | 0135566678999
##   18 | 01122234477
##   19 | 034
```

Notese que esté gráfico tiene interesante resultados, entre ellos tenemos los siguientes:

- La forma del diagrama de tallos hojas es una representación de un histograma en forma vertical.
- La agrupación de los datos se hace en potencias de 10.
- Se logra determinar facilmente el valor menor y también el valor menor.
- Se puede ver facilmente cual es el dato que más repite.

Otro argumento de la función `stem` es permitir que los datos se dividan en hojas superiores e inferiores, es decir que el último decimal sea mayor o igual a 5 o menor de 5.

```
stem(X,2)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   13 | 7
```

```
##    14 | 0
##    14 | 89
##    15 |
##    15 | 789
##    16 | 011344
##    16 | 5688889999
##    17 | 013
##    17 | 5566678999
##    18 | 011222344
##    18 | 77
##    19 | 034
```

Ejercicios

Se realizó un estudio sobre satisfacción en el empleo en cuatro ocupaciones. La satisfacción en el empleo se midió mediante un cuestionario de 18 puntos en el que a cada punto había que calificarlo con una escala del 1 al 5; las puntuaciones más altas correspondían a mayor satisfacción en el empleo. La suma de las calificaciones dadas a los 18 puntos proporcionaba una medida de la satisfacción en el empleo de cada uno de los individuos de la muestra. Los datos obtenidos fueron los siguientes.

1. Realice un diagrama barras para la ocupación de los trabajadores.
2. Realice una histograma para la satisfacción de los trabajadores.
3. ¿Qué conclusiones se pueden de los gráficos anteriores?

Ocupación	Satisfacción
Terapeuta físico	77
Terapeuta físico	65
Análista de sistemas	75
Ebanista	58
Abogado	56
Ebanista	75
Ebanista	19
Análista de sistemas	52
Análista de sistemas	71
Abogado	68
Abogado	52
Abogado	80
Análista de sistemas	49
Terapeuta físico	35
Ebanista	23
Terapeuta físico	25
Análista de sistemas	41
Ebanista	55
Terapeuta físico	66
Ebanista	47

Ocupación	Satisfacción
Ebanista	84
Abogado	39
Análista de sistemas	51
Abogado	42
Terapeuta físico	65
Terapeuta físico	36
Abogado	52
Terapeuta físico	73
Ebanista	24
Terapeuta físico	81
Terapeuta físico	42
Análista de sistemas	79
Terapeuta físico	43
Abogado	42
Ebanista	52
Análista de sistemas	83
Ebanista	81
Abogado	46
Análista de sistemas	74
Terapeuta físico	88

Resúmenes numéricos

Otra forma de resumir la información es utilizando estadísticos numéricos los cuales a partir de un número se describe como están los datos. A continuación se describe los estadísticos más utilizados para resumir la información.

Estadística de tendencia central

Estos estadísticos permiten describir alrededor de que valor se están centrando los datos entre los más utilizados tenemos los siguientes:

Media

La media aritmética de cierto conjunto de datos se encuentra sumando los números y dividiendo después entre la cantidad de datos (promedio). Su formula es la siguiente:

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

Ejemplo Se registro el número de niños en 10 familias.

$$3, 2, 1, 2, 1, 2, 1, 3, 1, 4$$

La media aritmética se calcula de la siguiente manera:

$$\bar{x} = \frac{3+2+1+2+1+2+1+3+1+4}{10} = 2$$

Observación

Ventajas

- Fácil de calcular.
- Buenas ventajas como estimador.

Desventajas

- Se ve afectada por los valores extremos.

Ejemplo Se registra las edades de 30 estudiantes de un curso de estadística.

```
##  [1] 20 22 25 29 19 29 30 26 25 17 19 19 26 22 27 23 27 30 22 27 30 19 26
## [24] 18 20 22 17 22 29 21
```

En R la función `mean` permite calcular el promedio

```
X
```

```
##  [1] 20 22 25 29 19 29 30 26 25 17 19 19 26 22 27 23 27 30 22 27 30 19 26
## [24] 18 20 22 17 22 29 21
```

```
mean(X)
```

```
## [1] 23.6
```

Mediana

Es aquel valor en el cual el 50 % de los datos se encuentran por debajo del él y el otro 50 % de los datos se encuentran por encima de él. Es decir, es la observación que ocupa el lugar central de un conjunto de datos ordenados en forma ascendente.

Sea $x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[n]}$ un conjunto de datos ordenados en forma ascendente donde $x_{[i]}$ representa el valor en la posición i-ésima. La mediana se definen como:

$$\tilde{x} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{si } n \text{ es impar} \\ \frac{x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}}{2} & \text{si } n \text{ es par} \end{cases}$$

Ejemplo A continuación tenemos el siguiente conjunto de datos

$$7, 9, 0, 9, 1, 10$$

Luego de ordenarlos los datos quedan de la siguiente manera:

$$x_{[1]} = 0, x_{[2]} = 1, x_{[3]} = 7, x_{[4]} = 9, x_{[5]} = 9, x_{[6]} = 10$$

Notamos que hay un total de $n = 6$ datos, por lo cual, el número de datos es par. Utilizando la fórmula para n par se tiene lo siguiente:

$$\tilde{x} = \frac{x_{\left[\frac{6}{2}\right]} + x_{\left[\frac{6}{2}+1\right]}}{2} = \frac{x_{[3]} + x_{[4]}}{2} = \frac{7+9}{2} = 8$$

Ejemplo A continuación tenemos el siguiente conjunto de datos

$$1, 3, 2, 4, 0, 3, 0$$

Luego de ordenarlos los datos quedan de la siguiente manera:

$$x_{[1]} = 0, x_{[2]} = 0, x_{[3]} = 1, x_{[4]} = 2, x_{[5]} = 3, x_{[6]} = 3, x_{[7]} = 4$$

Notamos que hay un total de $n = 7$ datos, por lo cual, el número de datos es impar. Utilizando la fórmula para n impar se tiene lo siguiente:

$$\tilde{x} = x_{\left[\frac{7+1}{2}\right]} = x_{\left[\frac{8}{2}\right]} = x_{[4]} = 2$$

Observación Ventajas

- No se ve afectada por valores extremos.

Desventajas

- No es fácil determinar la mediana si el conjunto de datos es grande.

Ejemplo En R la función `median` permite calcular el promedio

```
                        X
```

```
##  [1] 20 22 25 29 19 29 30 26 25 17 19 19 26 22 27 23 27 30 22 27 30 19 26
## [24] 18 20 22 17 22 29 21
```

```
                        median(X)
```

```
## [1] 22.5
```

Moda

Si existe, es el valor con mayor frecuencia. Se denota por $\hat{x}$.

Observación

Ventajas

- No se ve afectada por valores extremos.
- Es útil para datos categóricos.

Desventajas

- La moda puede que no exista.
- La moda puede no ser única.

Ejemplo En R no existe una función que calcule la moda, ya que la misma definición es la que se utiliza. Por lo anterior se describe el codigo que permite determinar el valor de la moda

```
                        X
```

```
##  [1] 20 22 25 29 19 29 30 26 25 17 19 19 26 22 27 23 27 30 22 27 30 19 26
## [24] 18 20 22 17 22 29 21
```

```
                        tabla <- table(X)
                        tabla
```

```
## X
## 17 18 19 20 21 22 23 25 26 27 29 30
##  2  1  4  2  1  5  1  2  3  3  3  3
```

```
                        moda <- as.numeric(names(tabla[max(tabla)==tabla]))
                        moda
```

```
## [1] 22
```

Media ponderada

Generalmente se utiliza cuando los datos se agrupan en una tabla de distribución de frecuencias para datos discretos. Para ello se utiliza la siguiente formula:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} = \sum_{i=1}^{k} x_i \cdot h_i$$

donde:

x_i: Es el valor i-ésimo.

f_i: Es la frecuencia absoluta del i-ésimo valor.

k: Es el número de datos que se agruparon.

n: Es el total de datos.

h_i: Es la frecuencia relativa del i-ésimo valor.

Ejemplo En una investigación realizada entre los suscriptores de la revista *Fortune* se hizo la pregunta siguiente: "De los últimos números ¿cuántos ha leído?" Suponga que en la distribución de frecuencia siguiente se resumen las 500 respuestas.

Número de libros leídos	Frecuencia
0	13
1	12
2	41
3	79
4	355

Realizando el cálculo de forma manual tenemos lo siguiente:

$$\bar{x} = \frac{(0)(13) + (1)(12) + (2)(41) + (3)(79) + (4)(355)}{13 + 12 + 41 + 79 + 355} = \frac{0 + 12 + 82 + 237 + 1420}{500} = \frac{1751}{500} = 3.502$$

Para realizar el calculo en R, extraemos del objeto `data.frame` los valores de la variable y la frecuencia y realizamos el calculo de la media ponderada.

```
db # Objeto data.frame que contiene la información del ejemplo.
```

```
##   Número de libros leídos Frecuencia
## 1                       0         13
## 2                       1         12
## 3                       2         41
## 4                       3         79
## 5                       4        355
```

```
X <- as.numeric(as.character(db$`Número de libros leídos`))
X
```

```
## [1] 0 1 2 3 4
```

```
f <- as.numeric(db$Frecuencia)
f
```

```
## [1]  13  12  41  79 355
```

```
media <- sum(X*f)/sum(f)
media
```

```
## [1] 3.502
```

Una función que permite calcula la media ponderada es `weighted.mean`.

```
X
```

```
## [1] 0 1 2 3 4
```

```
f
```

```
## [1]  13  12  41  79 355
```

```
weighted.mean(X,f)
```

```
## [1] 3.502
```

Ejercicio

Los siguientes datos corresponden a los salarios mensuales, en miles de pesos, pagados por una empresa a su personal

Salarios (miles de pesos)	Número de empleados
800	10
900	16
1000	35
1200	26
1500	13

Determine en cual es el salario promedio de los trabajadores.

Media agrupada

Cuando los datos continuos están agrupados en una tabla de distribución de frecuencias para datos continuos, se realiza la aproximación de los datos utilizando la siguiente formula:

$$\bar{x} = \frac{\sum_{i=1}^{k} y_i \cdot f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} y_i \cdot f_i}{n} = \sum_{i=1}^{k} y_i \cdot h_i$$

donde:

y_i: Es la marca de clase del i-ésimo intervalo.

f_i: Es la frecuencia absoluta del i-ésimo intervalo.

k: Es el número de intervalos que se agruparon.

n: Es el total de datos.

h_i: Es la frecuencia relativa del i-ésimo intervalo.

Ejemplo Una compañía farmacéutica que vende por teléfono obtuvo la siguiente distribución de frecuencias de los tiempos de envió (número de minutos transcurridos entre el momento en el que se hace el pedido y el momento en el que se envía) de una muestra aleatoria de 40 pedidos.

Números de minutos	f_i
$4 \leq x < 10$	8
$10 \leq x < 16$	15
$16 \leq x < 22$	10
$22 \leq x < 28$	7

¿Cuál es el tiempo de envío medio aproximado?

A continuación, se calcula las marcas de clase

Números de minutos	y_i	f_i
$4 \leq x < 10$	7	8
$10 \leq x < 16$	13	15
$16 \leq x < 22$	19	10
$22 \leq x < 28$	25	7

Realizando el calculo de forma manual tenemos lo siguiente:

$$\bar{x} = \frac{(7)(8) + (13)(15) + (19)(10) + (25)(7)}{8 + 15 + 10 + 7} = \frac{56 + 195 + 190 + 175}{40} = \frac{616}{40} = 15.4$$

Para realizar los cálculos en R, se calcula la marca de clase de cada intervalo, es decir, el punto medio de cada intervalo.

```
y <- c(7, 13, 19, 25)
y
```

```
## [1]  7 13 19 25
```

```
f <- c(8,15,10,7)
f
```

```
## [1]  8 15 10  7
```

Y para calcular la media agrupada entonces utilizamos la función `weighted.mean`. El resultado es el siguiente:

```
media <- weighted.mean(y,f)
media
```

```
## [1] 15.4
```

Ejercicios

1. A continuación se muestran los puntajes de los exámenes finales de veinte estudiantes de un curso de estadística.

   4.0 4.1 4.7 4.4 3.8 4.1 3.6 4.1 3.9 4.5 3.9 3.6 2.8 4.0 4.1 4.1 3.4 3.3 3.7 3.5

   Determine el valor de la media, mediana y la moda. Interprete los resultados.

2. Parte de un estudio de control de calidad tuvo como objetivo mejorar una linea de producción, se midieron los pesos (en onzas) de 50 barras de jabón. Los resultados son los siguientes, ordenados del más pequeño al más grande.

   11.6 12.6 12.7 12.8 13.1 13.3 13.6 13.7 13.8 14.1 14.3 14.3 14.6 14.8 15.1 15.2 15.6 15.6 15.7 15.8 15.8 15.9 15.9 16.1 16.2 16.2 16.3 16.4 16.5 16.5 16.5 16.6 17.0 17.1 17.3 17.3 17.4 17.4 17.4 17.6 17.7 18.1 18.3 18.3 18.3 18.5 18.5 18.8 19.2 20.3

   - Construya un diagrama de tallos y hojas para estos datos.
   - Construya un histograma para estos datos.
   - Resalte las características mas relevantes en estos datos, en términos del contexto del problema.

Estadísticas de localización

Una estadística de localización para una distribución de frecuencias es aquel valor para el cual una porción específica de la distribución de los datos queda en cierto valor.

Entre las estadísticas de localización más utilizadas están los cuartiles, percentiles, deciles, quintiles.

Cada uno de estas divide la información en cierta cantidad de veces. Por ejemplo, si utilizamos los cuartiles, lo que hacemos es dividir la información en 4 partes iguales. Si utilizamos los deciles, en 10 partes iguales. Si utilizamos los quintiles, lo que hacemos es dividir la información en 5 partes iguales. Si utilizamos los percentiles, lo que hacemos es dividir la información en 100 partes iguales.

Actualmente éxiten muchos algoritmos que definen la manera en que se divide a información. Esto se debe a la forma en que realizan las aproximaciones a los valores. A continuación se indicara uno de las aproximaciones más utilizadas para encontrar los percentiles, ya que muchas de las estadísticas de localización anteriormente mencionadas se pueden aproximar por un percentil[2].

Percentil

Es aquel valor tal que lo más un $p\,\%$ de los datos están por debajo de esté valor y, el otro $(100-p)\,\%$ de los datos se encuentran por encima de él.

Cálculo del p-ésimo percentil

1. Ordene los datos de manera ascendente.
2. Calcule el índice $i = np/100$, siendo p el percentil de interés y n, la cantidad de datos.
3. Valide lo siguiente:
   a) Si el índice i no es entero, redondeamos al entero siguiente. El valor en la posición i indica el p-ésimo percentil.
   b) Si el índice i es entero, el p-ésimo percentil es el promedio de los valores en la posición i e $i+1$.

Ejemplo A continuación se tiene la estatura de 40 alumnos en centímetros

[1] 154 156 160 160 164 164 166 166 167 167 167 168 169 169 170 170 170
[18] 170 170 171 171 172 173 173 173 173 174 174 174 175 175 175 176 176
[35] 177 178 178 180 181 181

Notese que los estaturas está ordenadas de mayor a menor.

Si deseamos encontrar el primer cuartil en termino de percentiles, sería buscar el percentil 25, por lo cual se calcula

[2]Estos resultados pueden variar según el software que se utilice, debido al algoritmo que utilice.

el índice de posición de la siguiente manera

$$i = \frac{np}{100} = \frac{40 \cdot 25}{100} = \frac{1000}{100} = 10$$

Como el valor anterior es entero, por lo cual según la regla anterior se promedio los valores en la posición 10 y en la posición 11. Utilizando R se tiene lo siguiente.

```
X[10] # Valor en la posición 10
```

```
## [1] 167
```

```
X[11] # Valor en la posición 11
```

```
## [1] 167
```

```
(X[10] + X[11])/2 # Valor del percentil 25
```

```
## [1] 167
```

Ahora si deseamos encontrar el percentil 63, calculamos el índice de posición de la siguiente manera

$$i = \frac{np}{100} = \frac{40 \cdot 63}{100} = \frac{2520}{100} = 25.2$$

dado que el valor anterior no es entero, entonces utilizando la regla anterior se aproxima al entero siguiente que es 26. Utilizando R se tiene lo siguiente:

```
X[26] # Valor en la posición 26
```

```
## [1] 173
```

Para calcular los dos percentiles anteriores se utilizan la función `quantile`. Esta función tiene 9 algoritmos diferentes para calcular los percentiles. Para lo que estamos explicando utilizamos la opción 2, de la siguiente manera.

```
quantile(X,probs = 0.25, type = 2) # Percentil 25
```

```
## 25%
## 167
```

```
quantile(X,probs = 0.63, type = 2) # Percentil 63
```

```
## 63%
## 173
```

Resumen de cinco números

El resumen de cinco números se refiere a las cinco medidas descriptivas:

- Mínimo ($x_{\text{mín}}$).
- Primer cuartil (Q_1).
- Mediana ($\tilde{x}$).

- Tercer cuartil (Q_3).
- Máximo ($x_{\text{máx}}$).

$$x_{\text{mín}} \leq Q_1 \leq \tilde{x} \leq Q_3 \leq x_{\text{máx}}$$

Estos cinco valores son muy utilizados para construir un gráfico llamado diagrama de box-plot el cual es útil para describir la forma de los datos.

Utilizando el ejemplo de las estaturas en R se utilizan la función `summary` la cual calcula los cinco valores mencionados y la media

```
summary(X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   154.0   167.0   171.0   170.7   175.0   181.0
```

Diagrama de box-plot

Un diagrama de box-plot (cajas y bigotes) es un gráfico que describe la forma de una distribución por medio del resumen de cinco números.

Construcción un diagrama box-plot

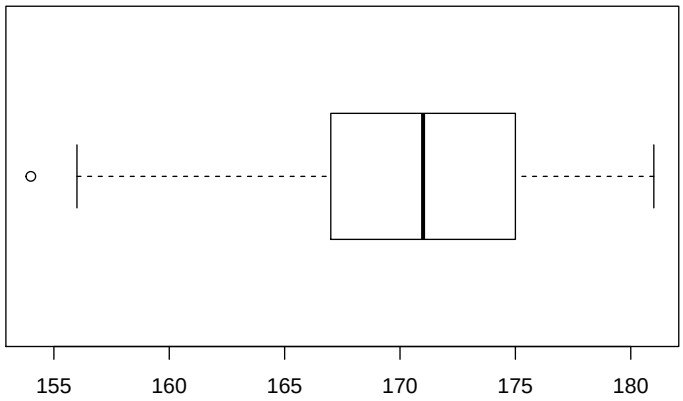
A continuación se describe los pasos a tener en cuenta para la construcción del diagrama de box-plot.

- Se ubican los valores Q_1 y Q_3 los cuales representa la caja del gráfico.
- Se localiza el valor $\tilde{x}$ como una línea vertical en la caja del gráfico.
- Se calcula el rango intercuartílico $RI = Q_3 - Q_1$[3].
- Se traza el bigote de la izquierda que va desde el Q_1 hasta el valor mayor más cercano al límite inferior $Q_1 - 1.5 * RI$.
- Se traza el bigote de la derecha que va desde el Q_3 hasta el valor menor más cercano al límite superior $Q_3 + 1.5 * RI$.
- Si hay datos que se encuentran a la izquierda del bigote izquierdo o la derecha del bigote derecho, se les denominan valores atípicos y se describen mediante un asterístico.

Utilizando el ejemplo de las estaturas en R se utiliza la función `boxplot` la cual realiza lo siguiente:

```
boxplot(X,horizontal = TRUE)
```

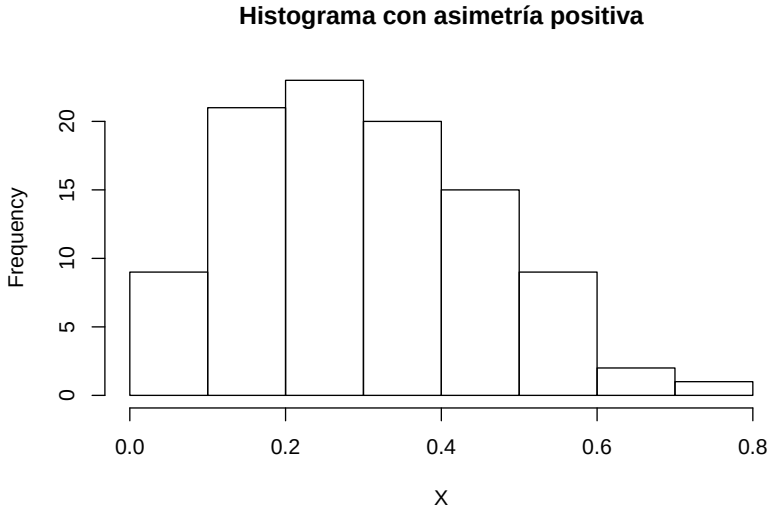
[3]El Rango intercuartílico, es una estadística de variabilidad la cual permite describir cual es el rango del 50 % de los datos.



Notese que se ubica un valor muy pequeño el cual se denota como un valor atípico, es decir, que es diferente del resto del conjunto.

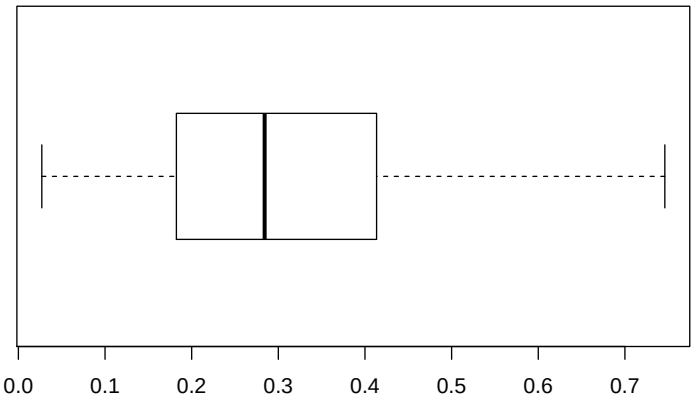
A continuación se simularan datos con algunos comportamientos, para observar como es el comportamientos del histograma y del diagrama de boxplot, según la asimetría[4] de los datos.

```
X <- rbeta(n = 100,shape1 = 2, shape2 = 5) # Datos con asimetría positiva
hist(X,main = "Histograma con asimetría positiva")
```

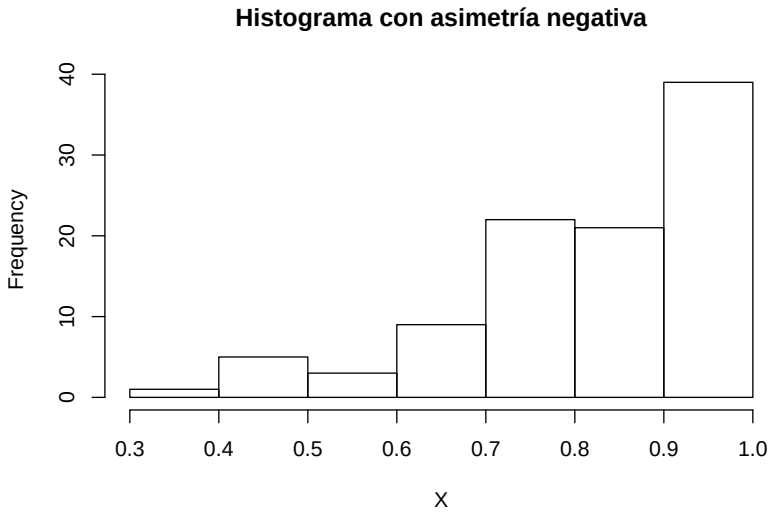



```
boxplot(X,horizontal = T)
```

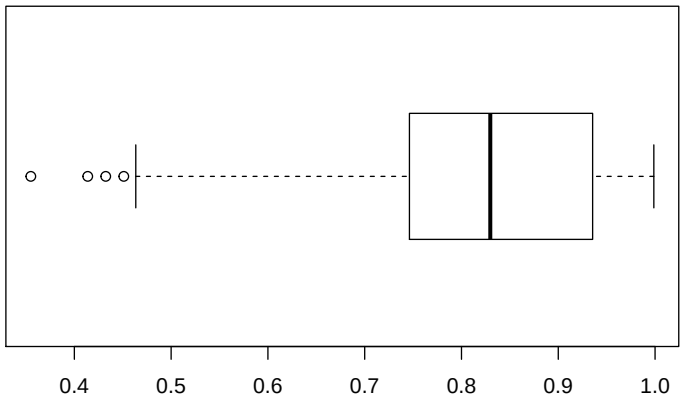
[4]Una distribución está sesgada o es asímetrica si las observaciones no están distribuidas simétricamente en ninguno de los lados del centro. Una distribución sesgada hacia la derecha (llamada a veces sesgada positivamente) tiene una cola que se extiende hacia la derecha. Una distribución sesgada hacia la izquierda (llamada a veces sesgada negativamente) tiene una cola que se extiende hacia la izquierda.



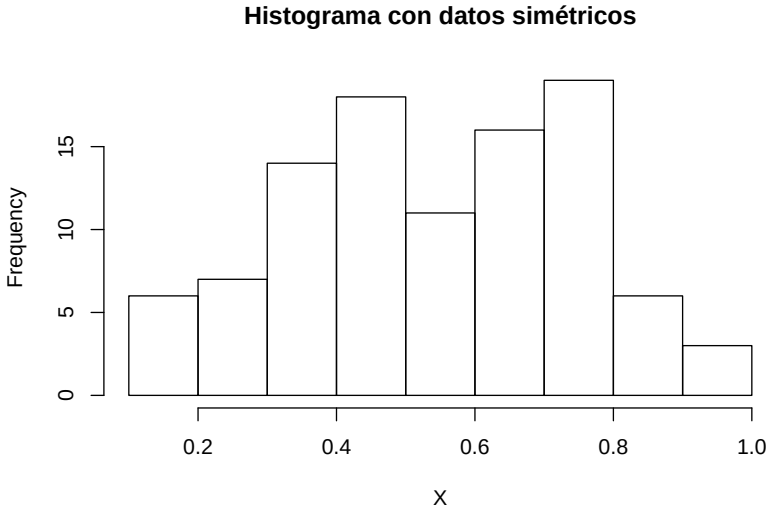

```r
X <- rbeta(n = 100,shape1 = 5, shape2 = 1) # Datos con asimetría negativa
hist(X,main = "Histograma con asimetría negativa")
```



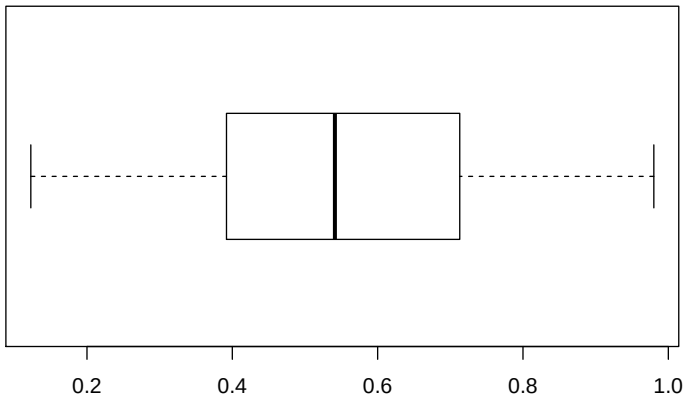

```r
boxplot(X,horizontal = T)
```




```r
X <- rbeta(n = 100,shape1 = 2, shape2 = 2) # Datos simétricos
hist(X,main = "Histograma con datos simétricos")
```

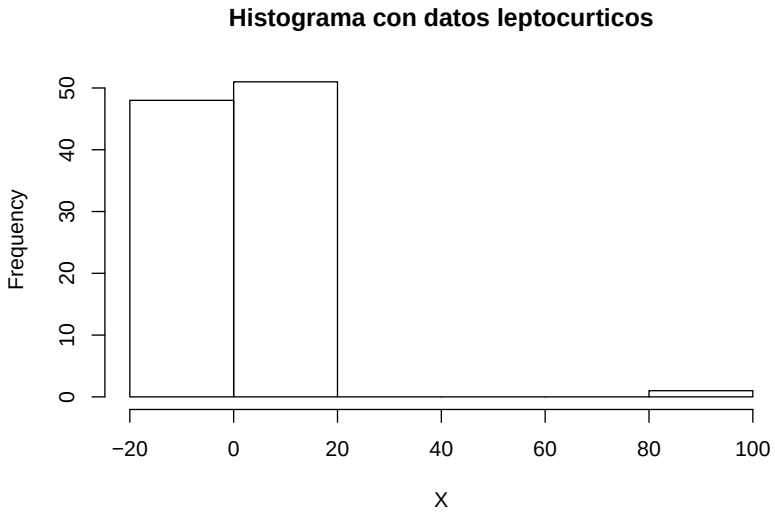



```r
boxplot(X,horizontal = T)
```



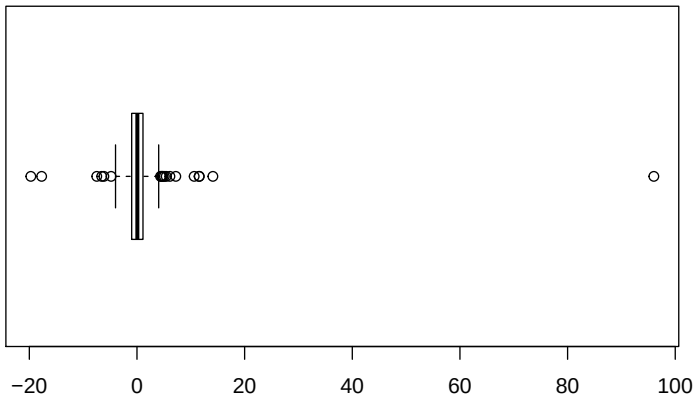
A continuación se simularan datos con algunos comportamientos, para observar como es el comportamientos del histograma y del diagrama de boxplot, según la curtosis[5] de los datos.

```
X <- rcauchy(n = 100, location = 0, scale = 1) # Datos leptocurticos
hist(X,main = "Histograma con datos leptocurticos")
```

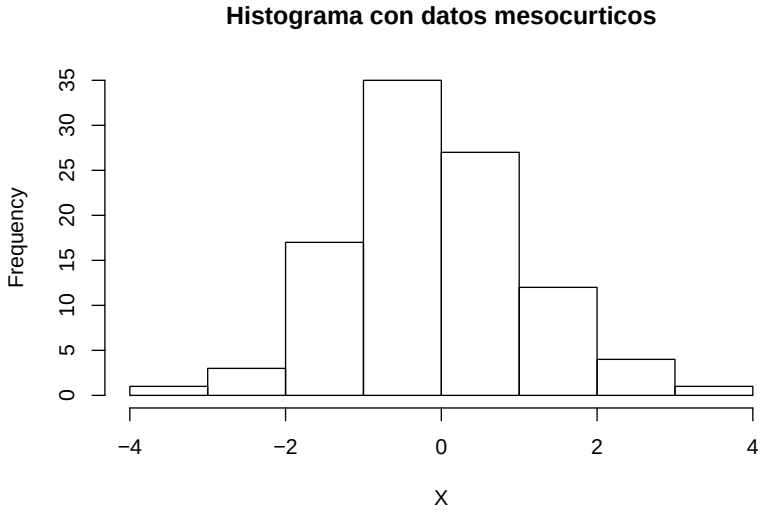



```
boxplot(X,horizontal = T)
```

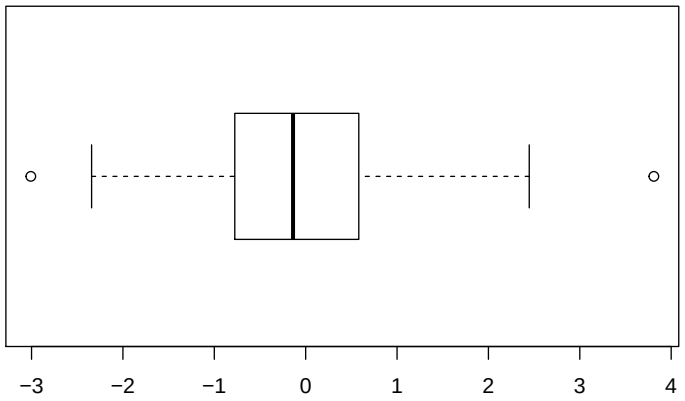
[5]La curtosis es el grado de concentración que presentan los valores alrededor de una estadística de centralidad como la media. Se dice que los datos son leptocuritocos cuando los datos se concentran mucho alrededor de la estadística de centralidad y con colas menos anchas que la normal. Se dice que los datos son mesocurticos cuando los datos tienden a tener una forma normal y tienden a ser simétricos. Se dice que los datos son platicurticos cuando los datos no se concentran alrededor de la estadística de centralidad y tiene colas más anchas que la normal.



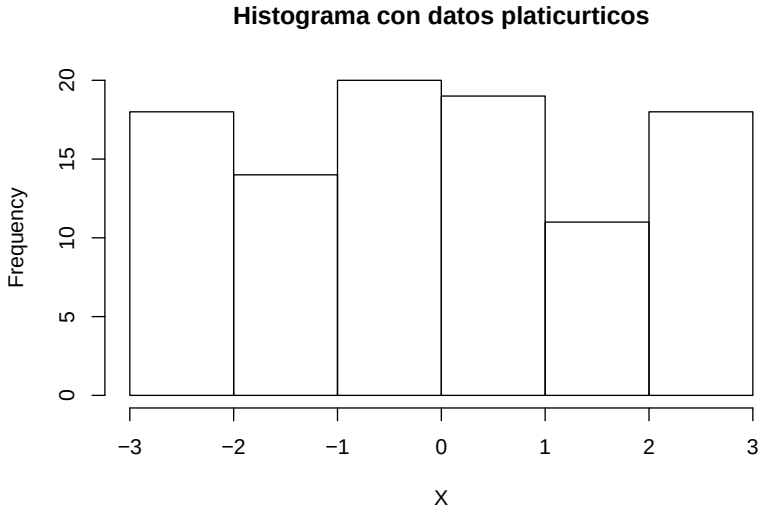

```r
X <- rnorm(n = 100, mean = 0, sd = 1) # Datos mesocurticos
hist(X,main = "Histograma con datos mesocurticos")
```



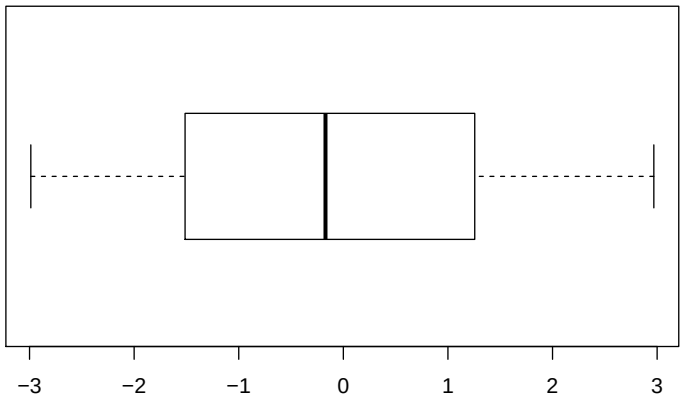

```r
boxplot(X,horizontal = T)
```




```r
X <- runif(n = 100,min = -3, max = 3) # Datos platicurticos
hist(X,main = "Histograma con datos platicurticos")
```




```r
boxplot(X,horizontal = T)
```



Ejercicios

1. Las tiendas Pelican, una división de National Clothing, es una cadena de ropa para mujer con sucursales por todo Estados Unidos. En fechas recientes la cadena realizó una promoción en la que envió cupones de descuento a clientes de otras tiendas de National Clothing. Los datos obtenidos en una muestra de 20 pagos con tarjeta de crédito en las tiendas Pelican durante un día de la promoción aparecen a continuación

Tipo de Cliente	Artículos	Ventas Netas	Forma de Pago	Género	Estado Civil	Edad
Promocional	13	198.8	Proprietary Card	Mujer	Casado(a)	42
Regular	3	54.5	Discover	Mujer	Casado(a)	40
Regular	1	44.5	Proprietary Card	Mujer	Casado(a)	54
Promocional	2	49.5	Proprietary Card	Mujer	Casado(a)	48
Promocional	4	111.14	Proprietary Card	Mujer	Casado(a)	28
Promocional	4	70.82	Proprietary Card	Mujer	Casado(a)	38
Promocional	6	117.5	Proprietary Card	Mujer	Casado(a)	50
Promocional	1	18	Proprietary Card	Mujer	Casado(a)	70
Regular	1	29.5	MasterCard	Hombre	Soltero(a)	36
Promocional	9	253	Proprietary Card	Mujer	Casado(a)	30
Promocional	1	31.6	Proprietary Card	Mujer	Soltero(a)	20
Regular	5	159.75	Proprietary Card	Mujer	Casado(a)	72
Regular	2	74	Visa	Mujer	Soltero(a)	20
Promocional	10	287.59	Proprietary Card	Mujer	Casado(a)	52
Regular	2	54	MasterCard	Mujer	Casado(a)	34
Regular	1	39.5	Discover	Hombre	Casado(a)	32
Promocional	2	80.4	Proprietary Card	Mujer	Casado(a)	48
Promocional	3	59.91	Proprietary Card	Mujer	Soltero(a)	30
Promocional	1	20.8	Proprietary Card	Mujer	Casado(a)	62
Promocional	6	123.1	Proprietary Card	Mujer	Casado(a)	54

Los directivos de Pelican desean emplear estos datos muestrales para tener información acerca de sus clientes y evaluar la promoción de los cupones de descuento. Teniendo en cuenta está información realice lo siguiente

a) Realice distribuciones de frecuencia porcentual de las variables clave y realice comentarios de los resultados obtenidos.

b) Realice un histograma sobre las edades de los clientes y comente sobre los resultados obtenidos.

c) Realice un diagrama de box-plot sobre las ventas netas realizadas y de una explicación de este.

d) Realice un diagrama de barras o un diagrama circular sobre los distintos distintos tipos de clientes. ¿La promoción con cupón si surtió efecto sobre los clientes?.

2. Una empresa de marketing cuenta con un personal de ventas de 20 Vendedores. Los datos relativos a la edad y las ventas totales obtenidos en sus territorios en un mes en particular se dan a continuación.

Vendedores	Género (M=1, F=2)	Edad	Región	Ventas (en miles de dólares)
1	1	25	1	50
2	1	22	1	75
3	1	20	2	11
4	1	27	2	77
5	1	28	3	45
6	1	24	1	52
7	1	24	2	26
8	1	23	3	24
9	2	24	3	28
10	2	30	3	31
11	2	19	2	36
12	2	24	1	72
13	2	26	1	69
14	2	26	1	51
15	2	21	2	34
16	2	24	2	40
17	2	29	3	18
18	2	27	3	35
19	2	24	1	29
20	2	25	1	68

a) Teniendo en cuenta la información anterior construya un gráfico box-plot para las ventas (realice comentarios sobre los resultados).

b) Realice un diagrama de barras correspondientes a las regiones de los trabajadores (realice comentarios sobre los resultados).

c) Construya un diagrama de pastel para el género de los trabajadores (realice comentarios sobre los resultados).

d) El gerente desea determinar cuál es el promedio recortado (al 10 %) de venta de los vendedores.

e) El dueño de la empresa comenta que las ventas de los hombres son más homogéneas que las ventas de

las mujeres. Verifique este supuesto utilizando el coeficiente de variación (realice comentarios sobre los resultados).

f) El gerente decide dar una bonificación salarial para aquellos vendedores cuyas ventas fueron mayores al noveno decíl. Por otro lado, aquellos vendedores que estén por debajo del segundo decíl, deberán realizar una lectura del sobre "*el vendedor más grande del mundo*". Considerando lo anterior, determine a partir de qué venta ganan la comisión los vendedores y cuantos las obtienen. Y por otro lado encuentre el valor de las ventas a partir del cual los vendedores tienen que leer el libro.

Estadísticas de variabilidad

Las estadísticas de variabilidad permiten determinar que tan variables o que tan poca variablidad presenta un conjunto de datos frente a una medida de centralidad como lo es la media.

Ejemplo.

Ejemplo Observemos el siguiente conjunto de datos los cuales describen los salarios anuales de siete supervisores de ventas de una empresa y los de siete de otra empresa.

Empresa 1	Empresa 2
34.5	34.9
30.7	27.5
32.9	31.6
36.0	39.7
34.1	35.3
33.8	33.8
32.5	31.7

Notese que al calcular la media y la mediana del conjunto de datos no hay diferencias ya que estás dos estadísticas son iguales

```
                        apply(db,2,mean)
```

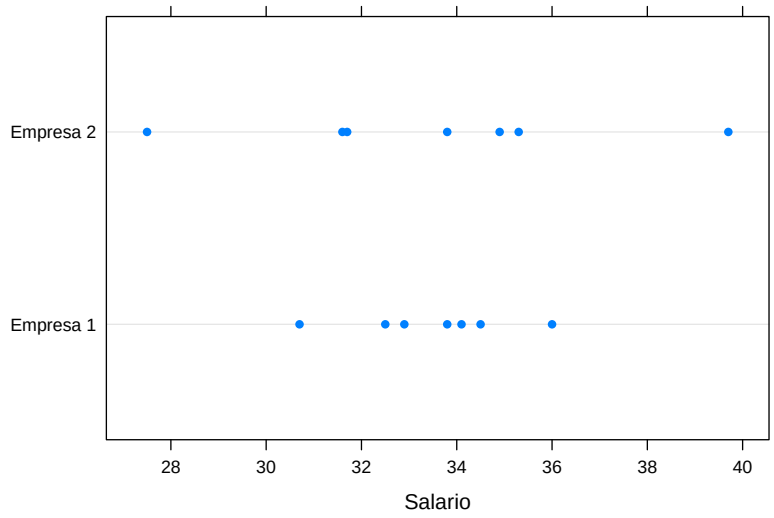
```
## Empresa 1 Empresa 2
##      33.5      33.5
```

```
                        apply(db,2,median)
```

```
## Empresa 1 Empresa 2
##      33.8      33.8
```

Pero al realizar un gráfico de puntos para ambos datos notamos que las diferencias son más de variabilidad

```
                        db.res <-reshape2::melt(db) # Reestructurando los datos
                        names(db.res) <- c("Empresas","Salario")
                        lattice::dotplot(Empresas~ Salario,data = db.res)
```



Al visualizar los salarios de los empleados de la empresa 1, etos tienen menor dispersión alrededor de la medio que los salarios de los empleados de la empresa 2. Por ello se prefieren medidas de variabilidad que permitan cuantificar la variabilidad de los datos.

A continuación se describiran las estadística de variabilidad más utilizadas.

Rango

Es la diferencia entre la observación mayor y la menor.

$$R = x_{\text{máx}} - x_{\text{mín}}$$

Rango Intercuartilico

Mide la dispersión que hay en el 50 % central de los datos; es la diferencia entre el tercer cuartil y el primer cuartil.

$$RI = Q_3 - Q_1$$

Varianza poblacional

Es el promedio de los cuadrados de las desviaciones de los valores con respecto a la media poblacional

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

Observación La anterior es utilizada cuando se tiene toda la información poblacional, sin embargo, esto en muchas ocasiones es dificil, ya que es muy común trabajar con muestras. Razón por la cual se considera trabajar con la varianza muestral o también llamada varianza corregida.

Varianza muestral

A diferencia de la varianza poblacional, está se utiliza cuando se tienen datos muestrales.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Observación 2 Aunque la varianza es una medida de dispersión muy utilizada, interpretarla es complicado, ya que las unidades de la varianza son el cuadrado de las unidades de medida, por lo cual se prefiere la desviación estándar.

Desviación estándar poblacional

No es más que la raiz cuadrada de la varianza poblacional.

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

Desviación estándar muestral

No es más que la raiz cuadrada de la varianza poblacional.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Algunas propiedades

- La varianza y la desviación estándar tiene una limitación y es que frente a presencia de datos atípicos, pueden verse afectados los cálculos.
- Otra forma de expresar la varianza muestral es dada por la siguiente ecuación

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

- La varianza siempre es positiva.
- La varianza observaciones constantes es igual a cero.

Observación 3 La relación entre la varianza muestral y la varianza poblacional es la siguiente:

$$s = \frac{n}{n-1}\sigma^2$$

Ejemplo Determine el rango, el rango intercurtilico, la varianza muestral y la desviación muestral, para los salarios de los empleados de la empresas utilizado en el ejemplo anterior.

```
                        emp1 # Salarios de los empleados de la empresa 1
```

```
## [1] 34.5 30.7 32.9 36.0 34.1 33.8 32.5
```

```
                        emp2 # Salarios de los empleados de la empresa 2
```

```
## [1] 34.9 27.5 31.6 39.7 35.3 33.8 31.7
```

```
                        R1 <- diff(range(emp1)) # Rango de los salarios de los empleados de la empres
                        R1
```

```
## [1] 5.3
```

```
                        R2 <- diff(range(emp2)) # Rango de los salarios de los empleados de la empres
                        R2
```

```
## [1] 12.2
```

```
                        RI1 <- IQR(emp1) # Rango Intercuartilico de los salarios de los empleados de
                        RI1
```

```
## [1] 1.6
```

```
                        RI2 <- IQR(emp2) # Rango Intercuartilico de los salarios de los empleados de
                        RI2
```

```
## [1] 3.45
```

```
                        var1 <- var(emp1) # Varianza muestral de los salarios de los empleados de la
                        var1
```

```
## [1] 2.816667
```

```
                        var2 <- var(emp2) # Varianza muestral de los salarios de los empleados de la
                        var2
```

```
## [1] 14.43
```

```
                        de1 <- sd(emp1) # Desviación estándar muestral de los salarios de los emplead
                        de1
```

```
## [1] 1.678293
```

```
                        de2 <- sd(emp2) # Desviación estándar muestral de los salarios de los emplead
                        de2
```

```
## [1] 3.798684
```

Ejercicios

1. A continuación se tiene la información del índice de calidad del aire en varias regiones del sur de Colombia. 30, 42, 58, 50, 45, 55, 60, 49 y 52.

a) Calcule el rango y el rango intercuartílico.

b) Calcule la varianza muestral y la desviación estándar muestral.

c) En una muestra de índices de calidad del aire en una de las localidades de Bogotá, la media muestral es 48.5 y la desviación estándar muestral es 11.66. Con base en estos estadísticos descriptivos compare la calidad del aire en está localidad con respecto al sur de Colombia.

2. A continuación se tiene los tiempos que ciertos corredores universitarios en un 500 metros y en 2000 (los tiempos están en minutos).

Tiempos en un 500 metros:	0.92	0.98	1.04	0.90	0.99
Tiempos en una 2000 metros:	4.52	4.35	4.60	4.70	4.50

Después de ver estos datos, el entrenador comentó que los corredores de 500 metros los tiempos eran más homogéneos. Usando el coeficiente de variación verifique que tan correcta es esta afirmación.

Aplicaciones de la desviación estándar

Coeficiente de variación

Es una medida de la variablidad relativa que expresa la desviación estándar en porcentaje de la media. Es igual a la desviación estándar sobre el valor absoluto de la media multiplicado por 100 %.

$$CV = \frac{s}{|\bar{x}|} \times 100\,\%$$

Esta medida es util para comparar conjunto de datos que presenten diferente escala de medida, ya que no está estadística no depende la escala de medida.

Ejemplo Compara el coeficiente de variación entre los salarios de la empleados de las empresas utilizados en el ejercicio anterior.

```
emp1 # Salarios de los empleados de la empresa 1
```

```
## [1] 34.5 30.7 32.9 36.0 34.1 33.8 32.5
```

```
emp2 # Salarios de los empleados de la empresa 2
```

```
## [1] 34.9 27.5 31.6 39.7 35.3 33.8 31.7
```

```
CV1 <- sd(emp1)/abs(mean(emp1))*100 # Coeficiente de variación de los emplead
CV1
```

```
## [1] 5.009829
```

```r
CV2 <- sd(emp2)/abs(mean(emp2))*100 # Coeficiente de variación de los emplead
CV2
```

```
## [1] 11.33936
```

Coeficiente de asimetría de Fisher

El coeficiente de asimetría es un estadístico que permite determinar la asimetría que tiene un conjunto de datos. Su calculo se define de la siguiente manera:

$$g_1 = \frac{m_3}{s^3}$$

donde m_3 se denota el momento centrado de orden 3 que se define de la siguiente manera:

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

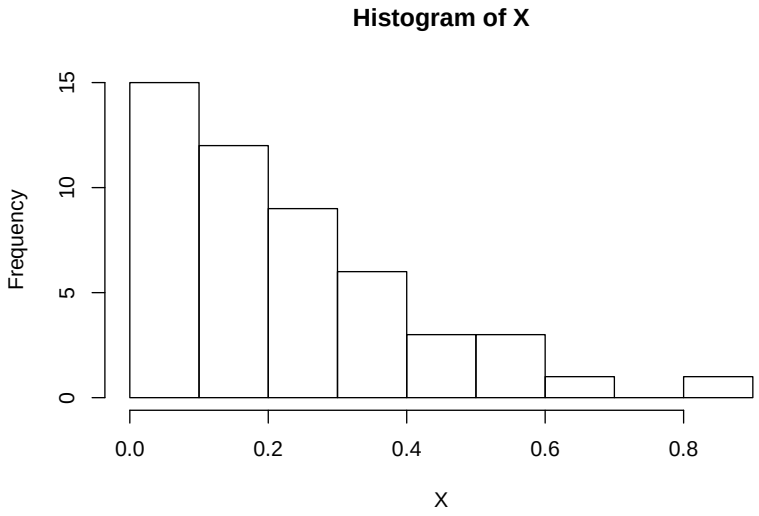
y s^3 es el cubo de la desviación estándar.

Notese que está esta estadística puede tomar valores positivos o negativos, por los cual se sugiere compara está esta estadística con cero para interpretarla de la siguiente manera:

- Si el $g_1 \approx 0$ (es cero o tienden a ser cercano a cero) la distribución es simétrica.
- Si el $g_1 < 0$ la distribución tiene asimetría negativa (es decir, la distribución es alargada a la izquierda).
- Si el $g_1 > 0$ la distribución tiene asimetría positiva (es decir, la distribución es alargada a la derecha).

Ejemplo En R se utiliza la función `skewness` del paquete `e1071`, el cual acontinuación se simulan algunos emjemplos anteriormente descritos.

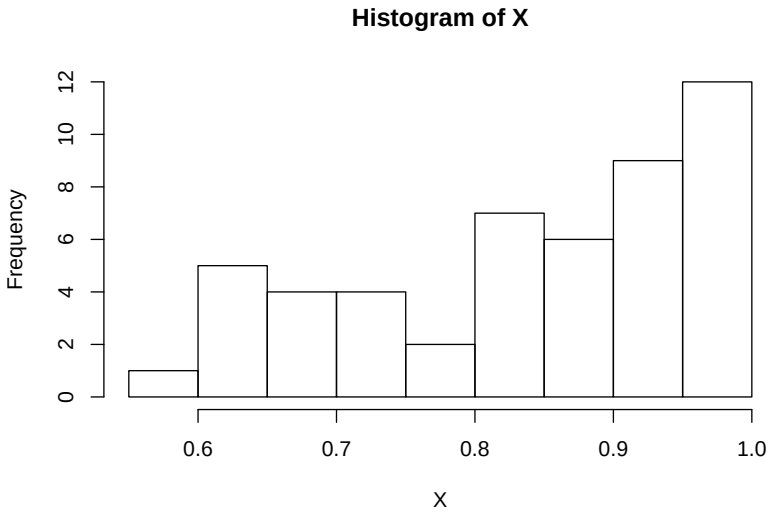
```r
library(e1071)
X <- rbeta(n = 50,shape1 = 1, shape2 = 3) # Datos con asimetría positiva
hist(X)
```



```
skewness(X,type = 1)
```

```
## [1] 1.054133
```

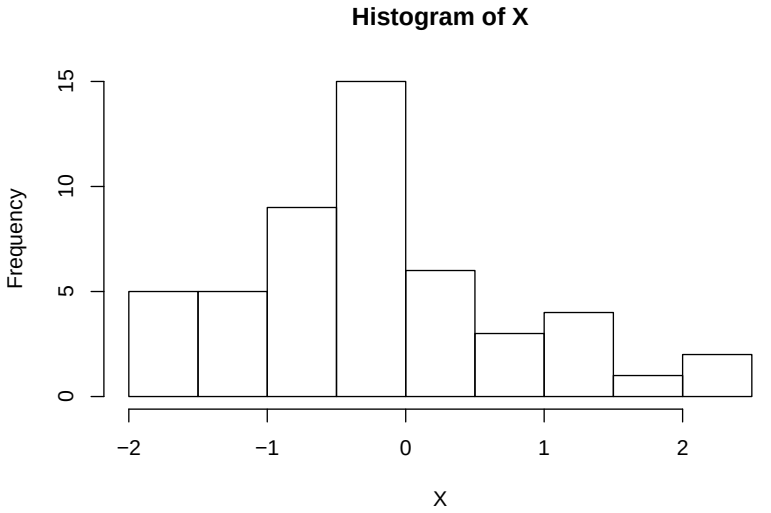
```
X <- rbeta(n = 50,shape1 = 5, shape2 = 1) # Datos con asimetría negativa
hist(X)
```




```
skewness(X,type = 1)
```

```
## [1] -0.484996
```

```
X <- rnorm(n = 50, mean = 0, sd = 1) # Datos simétricos
hist(X)
```




```
skewness(X,type = 1)
```

```
## [1] 0.5025641
```

Coeficiente de curtosis de Pearson

El coeficiente de curtosis es un estadístico que permite determinar que tanta concentración tiene la asimetría de un conjunto de datos. Su cálculo se define de la siguiente manera:

$$g_2 = \frac{m_4}{s^4} - 3$$

donde m_4 se denota el momento centro de orden 4 que se define de la siguiente manera:

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

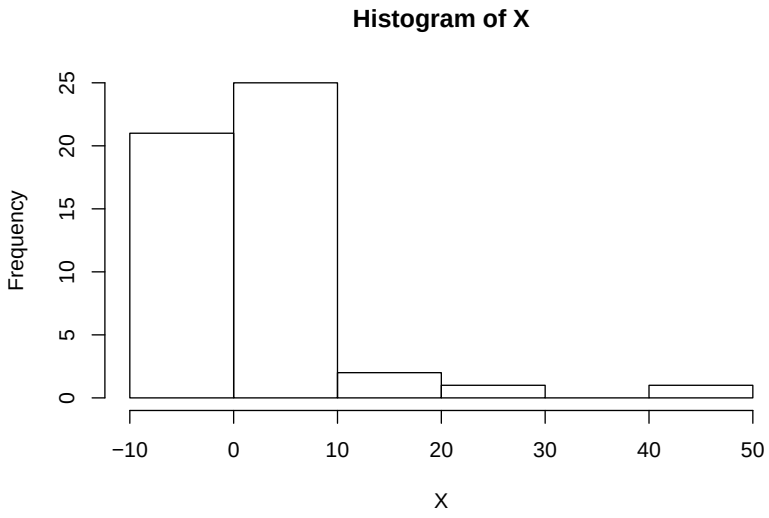
y s^4 es la cuarta potencia de la desviación estándar.

La interpretación de esté coeficiente es la siguiente:

- Si el $g_2 \approx 0$ (es cero o tienden a ser cercano a cero) la distribución de la curva es mesocurtica.
- Si el $g_2 < 0$ la distribución de la curva es platicurtica (es decir, la curva sera más plana).
- Si el $g_2 > 0$ la distribución de la curva es leptocurtica (es decir, la curva será más aguda y de colas largas).

Ejemplo En R se utiliza la función `kurtosis` del paquete `e1071`, el cual acontinuación se simulan algunos emjemplos anteriormente descritos.

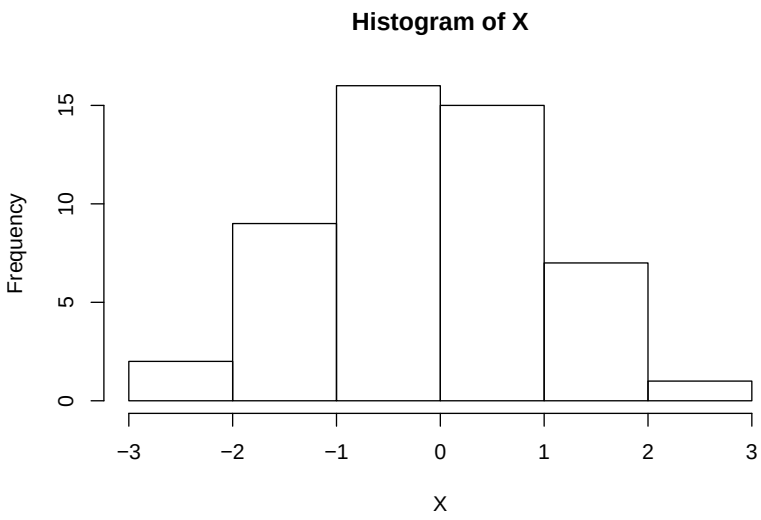
```
X <- rcauchy(n = 50, location = 0, scale = 1) # Datos leptocurticos
hist(X)
```




```
kurtosis(X,type = 1)
```

```
## [1] 19.54907
```

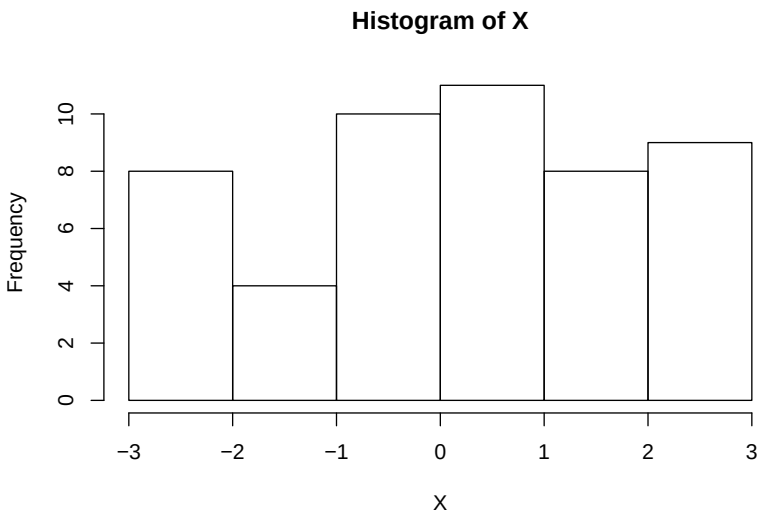
```r
X <- rnorm(n = 50, mean = 0, sd = 1) # Datos mesocurticos
hist(X)
```




```r
kurtosis(X,type = 1)
```

```
## [1] 0.01591701
```

```r
X <- runif(n = 50,min = -3, max = 3) # Datos platicurticos
hist(X)
```




```r
kurtosis(X,type = 1)
```

```
## [1] -0.9865519
```

Regla empírica

Si la distribución de un conjunto de datos poblacional tiende a ser simétrica de forma acampanada se cumple lo siguiente:

- Aproximadamente el 68 % de los valores de la población de encuentran a una distancia de más o menos una desviación estándar. $\mu \pm \sigma$.
- Aproximadamente el 95 % de los valores de la población de encuentran a una distancia de más o menos dos desviación estándar. $\mu \pm 2\sigma$.
- Aproximadamente el 99.7 % de los valores de la población de encuentran a una distancia de más o menos tres desviación estándar. $\mu \pm 3\sigma$.

Ejemplo Supongase que la distribución de los salarios de una empresa tiene una forma acampanada cuyo salario promedio poblacional es de 3'500.000 y una desviación estándar poblacional de 800.000. Según la regla empírica, se estima que el 68 % de los salarios tendran entre 2'700.000 y 4'300.000; El 95 % de los salarios estarán entre 1'900.000 y 5'100.000; El 99.7 % de los salarios estarán entre 1'100.000 y 5'900.000.

Para observar esta situación se realizar una simulación generando 100 observaciones una distribución simetríca llamada normal, con media poblacional de $\mu = 3'500.000$ y una desviación estándar poblacional $\sigma = 800.000$ y se determina cuantos valores valores están a una distancia de más o menos una, dos y tres desviaciones estándar poblacional.

```
X <- rnorm(n = 100,mean = 3500000,sd = 800000)
table(X > 2700000 & X < 4300000) # Condición a una desviación
```

```
##
## FALSE  TRUE
##    33    67
```

```
table(X > 1900000 & X < 5100000) # Condición a dos desviaciones
```

```
##
## FALSE  TRUE
##     4    96
```

```
table(X > 1100000 & X < 5900000) # Condición a tres desviaciones
```

```
##
## TRUE
##  100
```

Notese que para los tres casos los resutados se aproximan a la regla empírica.

Valores z

Además de las estadísticas de localización, variabilidad y forma, interesa conocer también la ubicación relativa de los valores de un conjunto de datos. Los valores z permiten ubicar qué tan lejos de la media se encuentra un

determinado valor en termino de las deviaciones estándar. Para ello se calcula el valor de la siguiente manera:

$$z_i = \frac{x_i - \mu}{\sigma}$$

donde μ es la media poblacional de un conjunto de datos; σ es la desviación estándar poblacional; x_i es la observación i-ésima. Al valor z también se le suele llamar *valor estándar*. El valor z_i puede ser interpretado como el *número de desviaciones estándar a las que x_i se encuentra de la media.*

Nota Cuando no se conoce la media μ y la desviación estándar poblacional σ, se suele utilizar la media muestral $\bar{x}$ y la desviación estándar muestral s, de la siguiente manera.

$$z_i = \frac{x_i - \bar{x}}{s}$$

Ejemplo Suponga que los siguientes valores son las estaturas de un curso de estadística

Calculando la media y la desviación estándardar muestral tenemos lo siguiente:

```
media <- mean(X)
media # Media Muestral
```

```
## [1] 1.787
```

```
s <- sd(X)
s # Desviación estándar muestral
```

```
## [1] 0.09108122
```

Calculando los puntos z para cada estatura tenemos lo siguiente:

```
z <- (X - media)/s
z
```

```
##  [1] -0.51602293  0.58189819 -3.15103361 -0.62581504  1.13085875
##  [6]  1.13085875 -0.62581504  0.91127453  0.03293763  0.25252186
## [11]  0.25252186 -0.40623081  0.69169030 -0.51602293 -0.07685448
## [16]  0.91127453  1.24065086 -0.40623081  0.03293763 -0.84539926
```

Por ejemplo, la persona $z_1 = -0.5$, nos indica que está persona está a -0.5 desviaciones con respecto a la media muestral.

Valores z mayores a cero indican que son observaciones cuyo valor es mayor a la media, y valores z menores que cero indican que son observaciones cuyo valor es menor a la media. Si el punto z es cero, el valor de la observación correspondiente es igual a la media.

Los valores z de cualquier observación se interpreta como una medida relativa de la posición de la observación en el conjunto de datos.

Detección de datos atípicos

Aquellos valores que tienen un comportamiento diferente al resto del conjunto se le conocen como *valores atípicos* o también llamados *outliers*. Estos valores normalmente son valores extremos que se encuentran muy por enciman

o en ocasiones muy por debajo de los valores usuales.EStos datos tienen a desviar la información por lo cual se sugiere detectar estás observaciones y revisarlas con cuidado, ya que por extraño que parezca quizá sea el valor es un dato que se escribio de forma incorrecta. De ser así se debe corregir antes de continuar con el análisis. En otras situaciones si se detecta una observación atípica pero es justficable el valor en tal caso debe conservarse.

Una forma usual de detectar estas observaciones es utilizando los *valores z* y teniendo presente la *regla empírica*. Por tanto, si usa los valores z para identificar las observaciones atípicas, es recomendable considerar cualquier dato cuyo valor $z > 3$ ó $z < -3$ son observaciones atípicas.

Ejemplo Recordando los valores z del curso anterior notamos que $z_3 = -3.2$, lo que nos indica que es una observación atípica. Al parecer es una persona muy baja con respecto al resto del grupo.

Ejercicios

1. El precio medio del galón de gasolina es de \$8500. Admita que la desviación estándar haya sido \$1000 y que el precio del galón de gasolina tenga una distribución en forma de campana.

   a) ¿Qué porcentaje de la gasolina se vendió entre \$8000 y \$9000 por galón?

   b) ¿Qué porcentaje de la gasolina se vendió entre \$8300 y \$8700 por galón?

   c) ¿Qué porcentaje de la gasolina se vendió a más de \$8900 por galón?

2. Ingrese a la siguiente dirección de internet <www.audioreview.com>. Ubique los 20 primeros artículos de la lista de Top 100. Construya una base que contenga Posición, Producto, Precio de venta al público sugerido por el fabricante(MSRP). Utilizando la información del precio MSRP determine lo siguiente:

   - Calcule la media y la mediana.
   - Aproxime el primer y tercer cuartil.
   - Estime la desviación estandar muestral.
   - Calcule el coeficiente de asimetría. Comente la forma de esta distribución.
   - Calcule los puntos z correspondientes. ¿Hay en estos datos alguna observación atípica? Explique.

Referencias

Anderson, D. R., Sweeney, D. J., y Williams, T. A. (2008). *Estadística para administración y economía.* México: Cengage.

Chao, L. L. (1975). *Estadística para las ciencias administrativas.* México: McGraw-Hill.

Esteban García, J., Bachero Nebot, J. M., Blasco Blasco, O. M., Coll Serrano, V., Díez García, R., Ivars Escortell, A., ... Ruiz Ponce, F. (2006). *Estadística descriptiva y nociones de probabilidad.* Madrid: Thomson.

Gutiérrez Cabria, S. (1994). *Filosofía de la estadística.* València: Universitat de València.

Hyndman, R. J., y Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, *50*(4), 361–365.

Llinas, H., y Rojas, C. (2006). *Estadística descriptiva y distribuciones de probabilidad.* Barranquilla: Ediciones Uninorte.

Lynch, S. M. (2013). *Using statistics in social research. A concise approach.* New York: Springer.

Mendenhall, W., y Sincich, T. (1997). *Probabilidad y estadística para ingeniería y ciencias.* México: Prentice - Hall.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., y Leisch, F. (2017). e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=e1071` (R package version 1.6-8)

Rodríguez, H. Y. (2012). *Estadística descriptiva.* Bogotá: Ediciones USTA.

Ross, S. M. (2007). *Introducción a la estadística.* Barcelona: Editorial Reverté.

Scrucca, L. (2004). qcc: an r package for quality control charting and statistical process control. *R News*, *4/1*, 11–17. Descargado de `https://cran.r-project.org/doc/Rnews/`

Stuart, A., y Ord, J. K. (1991). *Kendall's advanced theory of statistics* (Vol. 1). New York: Oxford Press.

Walpole, R. E., Myers, R. H., Myers, S. L., y Ye, K. (2008). *Probabilidad y estadística para ingeniería y ciencias.* México: Pearson.

YourDictionary. (2017, 22 de dic). *Statistics.* Descargado de `http://www.yourdictionary.com/statistics`

INFORMACIÓN TÉCNICA

Módulo: Probabilidad
Unidad 1: Estadística descriptiva
Escenario 2: Estadística descriptiva

Autor: Alex Johann Zambrano Carbonell

Asesor Pedagógico: Diana Marcela Díaz Salcedo
Diseñador Gráfico: Jully Amanda Guzman
Corrector de estilo: Jaime Posada
Asistente: Gina Quiroga