

CAPÍTULO 4: ESTADÍSTICA DESCRIPTIVA

En capítulos anteriores revisamos el diseño de estudios y la obtención de datos a través del muestreo, en este capítulo aprenderemos a trabajar con los datos y a resumirlos, de manera gráfica y numérica, para convertirlos en información útil para el problema en estudio.

La estadística descriptiva trata dos aspectos: el obtener información de los datos también conocido como "análisis exploratorio de datos"* y por otro lado se preocupa de la "presentación de resultados".

En este capítulo hablaremos sobre:

Tipos de Variables

Métodos gráficos y numéricos para describir variables cualitativas

- Tablas de distribución de frecuencias.
- Gráficos para variables cualitativas: Sectorial y de Barras.

Métodos gráficos para describir variables cuantitativas

- Gráfico de Puntos.
- Diagrama de Tallo y Hojas.
- Histograma.

Métodos numéricos para describir variables cuantitativas

- Medidas de Tendencia Central: Promedio, Mediana, Moda.
- Medidas de Dispersión: Rango, Desviación Estándar, Rango entre Cuarteles.
- Medidas de Posición Relativa.

Transformaciones lineales y estandarización

* El *padre* del análisis exploratorio de datos es John W. Tukey (1915-2000) Estados Unidos

Tipos de variables

La base de datos número 1, adjunta, contiene la información de 36 alumnos de un curso de Estadística de la Universidad de Talca.

Base de datos 1:

| Número | Sexo | Edad | Estatura | Peso | Ciudad de residencia | Número de hermanos |
|--------|------|------|----------|------|----------------------|--------------------|
| 1 | M | 22 | 180 | 74 | SAN FERNANDO | 7 |
| 2 | M | 20 | 175 | 95 | CHILLAN | 2 |
| 3 | M | 20 | 178 | 68 | TALCA | 2 |
| 4 | M | 22 | 183 | 75 | TALCA | 7 |
| 5 | M | 25 | 180 | 76 | LINARES | 3 |
| 6 | M | 22 | 180 | 78 | SANTIAGO | 1 |
| 7 | M | 21 | 180 | . | TALCA | 1 |
| 8 | M | 24 | 182 | 85 | TALCA | 1 |
| 9 | M | 21 | 177 | 78 | CURICO | 1 |
| 10 | M | 21 | 184 | 85 | SANTIAGO | 0 |
| 11 | M | 20 | 172 | 70 | SAN FERNANDO | 3 |
| 12 | M | 21 | 173 | 59 | IQUIQUE | 4 |
| 13 | F | 20 | 162 | 56 | SANTIAGO | 0 |
| 14 | M | 22 | 194 | 105 | LINARES | 4 |
| 15 | M | 20 | 174 | 79 | SANTIAGO | 1 |
| 16 | F | 20 | 165 | 50 | SAN JAVIER | 1 |
| 17 | F | 22 | 167 | 58 | TALCA | 1 |
| 18 | F | 20 | 155 | 52 | PUERTO MONTT | 2 |
| 19 | M | 20 | 174 | 65 | LINARES | 2 |
| 20 | F | 20 | 160 | 48 | SANTIAGO | 2 |
| 21 | F | 22 | 155 | 58 | SANTIAGO | 1 |
| 22 | M | 19 | 174 | 80 | SAN FELIPE | 1 |
| 23 | F | 19 | 162 | 60 | MELIPILLA | 1 |
| 24 | M | 19 | 180 | 82 | TALCA | 3 |
| 25 | F | 20 | 160 | 57 | TALCA | 1 |
| 26 | F | 21 | 170 | 70 | SANTIAGO | 2 |
| 27 | F | 20 | 155 | 50 | SANTIAGO | 1 |
| 28 | F | 21 | 160 | 60 | TALCA | 1 |
| 29 | F | 22 | 166 | 61 | PUERTO IBAÑEZ | 1 |
| 30 | M | 19 | 170 | 68 | RANCAGUA | 3 |
| 31 | F | 22 | 160 | 60 | SANTIAGO | 1 |
| 32 | M | 20 | 182 | 72 | TALCA | 1 |
| 33 | F | 19 | 162 | 55 | RANCAGUA | 2 |
| 34 | F | 20 | 154 | 46 | SANTIAGO | 3 |
| 35 | F | 19 | 155 | 50 | RANCAGUA | 2 |
| 36 | M | 20 | 184 | 85 | RANCAGUA | 5 |

En esta base de datos podemos notar que los alumnos tienen distintas características, por ejemplo, no todos vienen de la misma ciudad.

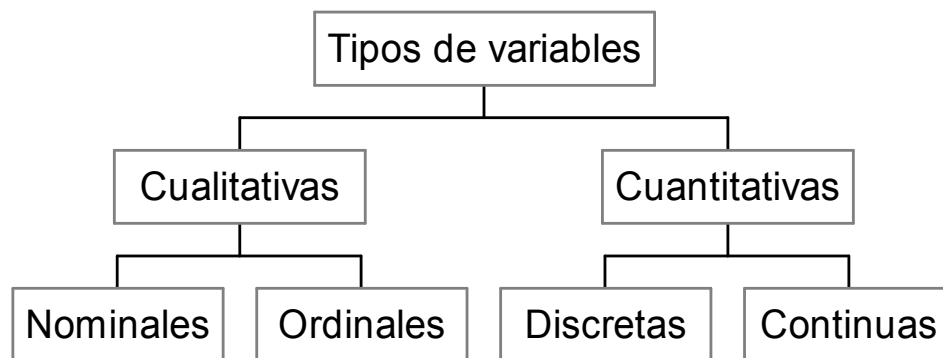
Definiciones:

Unidad es el objeto que observamos. Cuando el objeto es una persona, lo referimos como **sujeto**.

Observación es la información o característica que registramos de cada unidad.

Una característica que puede variar de unidad en unidad es llamada **variable**.

Una colección de observaciones con una o más variables se llama **base de datos**.



Variables cualitativas son aquellas que clasifican las unidades en categorías. Las categorías pueden tener un orden natural (ordinales) o no (nominales). Las variables cualitativas también se llaman variables categóricas. Con estas variables podemos contar número de casos, comparar entre categorías, pero no podemos realizar operaciones numéricas.

Variables cuantitativas tienen valores numéricos que representan medidas (largo, peso, etc.) o frecuencias (número de). Tiene sentido realizar operaciones numéricas con estas variables. Además distinguimos dentro de las variables cuantitativas las discretas y las continuas. Una variable **discreta** es aquella en la cuál se puede contar el número posible de valores. Una variable **continua** puede tomar cualquier valor en un intervalo dado.

☒ Ejemplo

Nominal: está asociada a nombres.

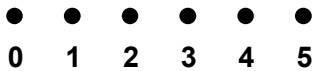
Ejemplo: Marca de auto, Sexo, Religión.

Ordinal: tiene asociado un orden.

Ejemplo: Nivel educacional, Estado nutricional, Nivel Socioeconómico.

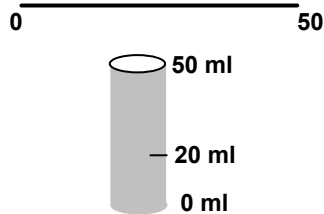
Discreta: sólo puede tomar un número finito (o contable) de posible valores.

Ejemplo: El número de respuestas correctas en una prueba de 5 preguntas de V o F.



Continua: puede tomar cualquier valor en un intervalo(s).

Ejemplo: Cantidad de agua en un vaso de 50 ml.



☒ Ejemplo

Tipo de Variable.

Determine qué tipo son las siguientes variables. Si son variables cualitativas (nominal u ordinal) o cuantitativas (discretas o continuas).

- a) Marca de automóvil.
- b) Duración de un compacto (segundos).
- c) Número de temas de un compacto.
- d) Nivel educacional (básica, media, universitaria).
- e) Temperatura al mediodía en Talca (grados Celcius).
- f) Estado civil (soltero, casado, divorciado, viudo).
- g) Cantidad de lluvia en un año en Talca (mm^3).

Métodos gráficos y numéricos para describir variables cualitativas

Definición:

La **distribución** de una variable nos da los valores posibles de la variable y cuantas veces ocurren. La distribución de una variable nos muestra la forma en que varía la variable.

Tablas de distribución de frecuencias.

Lo primero que hacemos al querer describir variables cualitativas es contar cuántas unidades caen en cada categoría de la variable. Esto lo presentamos en una tabla de distribución de frecuencias de la forma:

| Valor o categoría de la variable | Frecuencia | Porcentaje |
|----------------------------------|------------|------------|
| ... | | |
| Total | n | 100 |

☒ Ejemplo

Tabla de distribución de frecuencias del sexo de la base de datos 1

| Sexo | Número de alumnos | Porcentaje de alumnos |
|-----------|-------------------|-----------------------|
| Femenino | 16 | 44,4 |
| Masculino | 20 | 55,6 |
| Total | 36 | 100,0 |



En SPSS

Analizar > Estadísticos Descriptivos > Frecuencias.

| SEXO | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|-----------|------------|------------|-------------------|----------------------|
| Válidos F | 16 | 44.4 | 44.4 | 44.4 |
| M | 20 | 55.6 | 55.6 | 100.0 |
| Total | 36 | 100.0 | 100.0 | |

La salida de SPSS tiene columnas que no aportan información, Usted deberá editar estas tablas con la información que es relevante y borrar lo que no interesa.

Gráficos para variables cualitativas.

Una vez que conocemos la distribución de la variable, nos interesa presentarla de alguna manera gráfica, uno de los gráficos o diagramas más usados en variables cualitativas son los diagramas sectoriales o de torta y los gráficos de barra.

Un **gráfico sectorial (o de torta)** muestra la distribución de una variable cualitativa dividiendo un círculo en partes que corresponden a las categorías de la variable, tal que el tamaño (ángulo) de cada pedazo es proporcional al porcentaje de ítems en cada categoría.

Un **gráfico de barras** muestra la distribución de una variable cualitativa listando las categorías o valores de la variable en el eje X y dibujando una barra sobre cada categoría. La altura de la barra es igual al porcentaje de ítems en esa categoría. Las barras deben tener el mismo ancho.

Gráfico sectorial.

Figura 1 (a):

Diagrama sectorial con 1/4 de los ítems que comparten alguna propiedad.

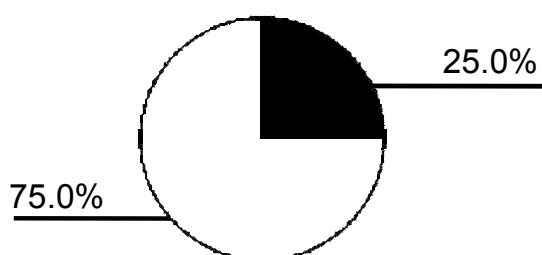


Figura 1 (b):

Diagrama sectorial con 7/8 de los ítems que comparten alguna propiedad

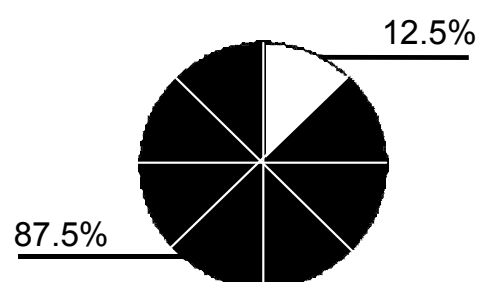


Diagrama sectorial para la variable SEXO de base de datos 1

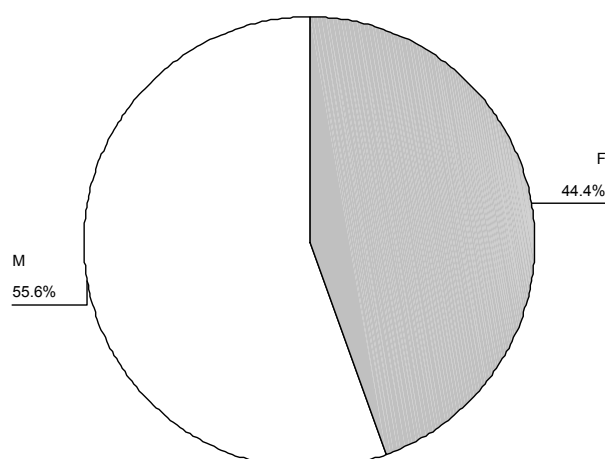
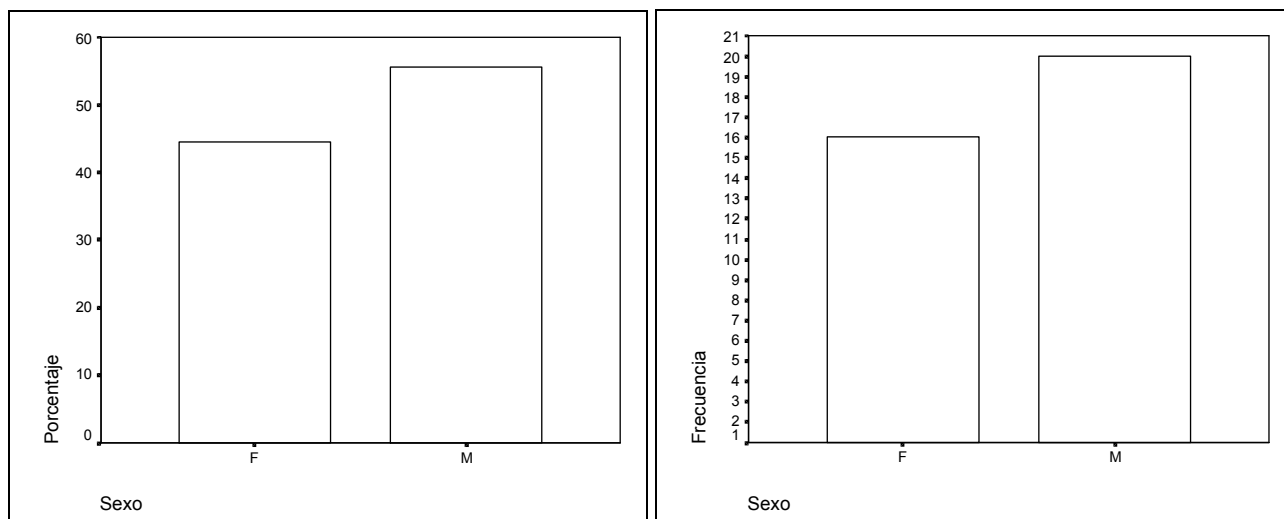


Gráfico de barras

Compare los siguientes gráficos. ¿Cuáles son las diferencias?

Gráfico de barras: Sexo en la base de datos 1.



Compare los siguientes gráficos. ¿Cuáles son las diferencias?

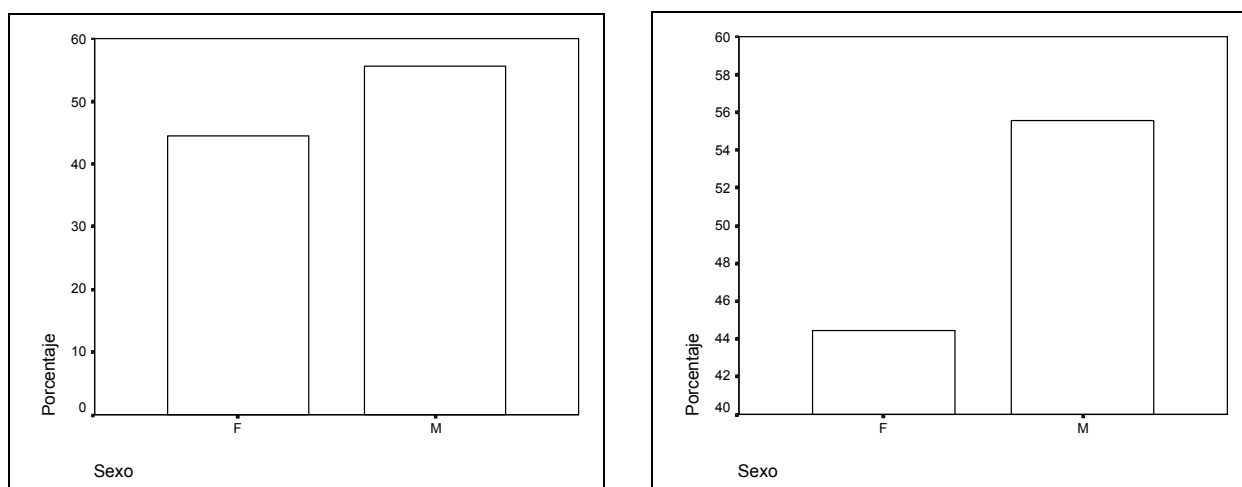
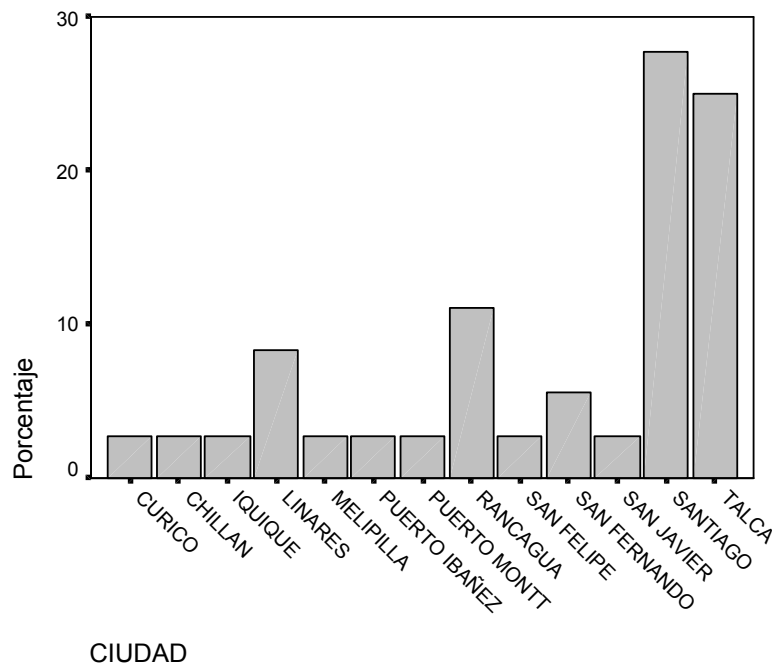


Gráfico de Barras: Ciudad de procedencia de alumnos de base de datos 1.



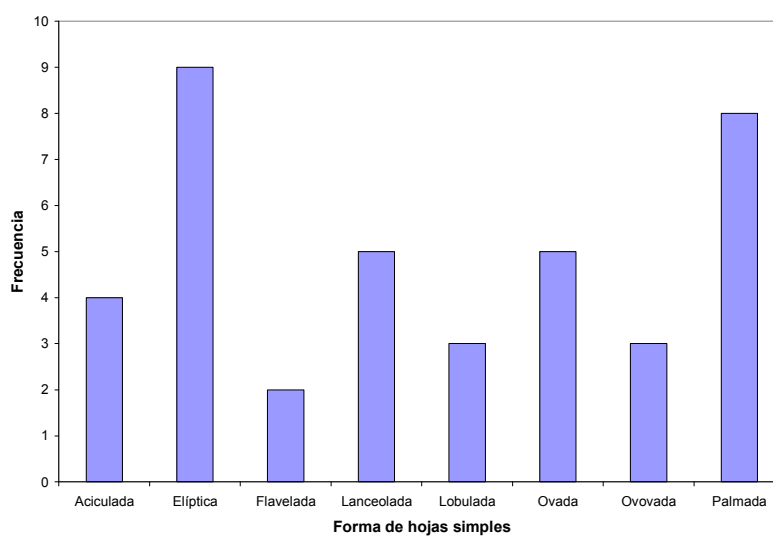
☒ Ejemplo

Métodos gráficos y numéricos para describir datos cualitativos

Tabla: Distribución de frecuencias de formas de hojas simples de una muestra de 39 hojas del parque de la Universidad de Talca, sector del edificio Prosperidad, I semestre 2001.

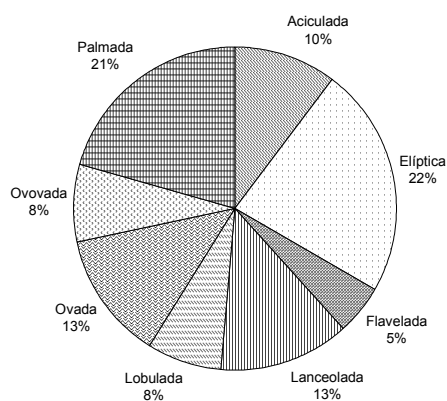
| Formas de hojas simples | Número de hojas | Porcentaje de hojas |
|-------------------------|-----------------|---------------------|
| Aciculada | 4 | 10,3 |
| Elíptica | 9 | 23,1 |
| Flagelada | 2 | 5,1 |
| Lanceolada | 5 | 12,8 |
| Lobulada | 3 | 7,7 |
| Ovada | 5 | 12,8 |
| Ovovada | 3 | 7,7 |
| Palmada | 8 | 20,5 |
| Total | 39 | 100 |

Figura 1: Gráfico de barras que muestra la frecuencia de formas de hojas simples.



Alternativamente podemos describir gráficamente con un gráfico circular como el de la figura 2.

Figura 2: Gráfico circular que muestra la frecuencia de formas de hojas simples.



Métodos gráficos para describir variables cuantitativas

En esta sección veremos de qué manera podemos describir gráficamente las variables cuantitativas. Veremos 3 tipos de gráficos:

1. Gráfico de puntos.
2. Diagrama de Tallo y Hojas.
3. Histograma.

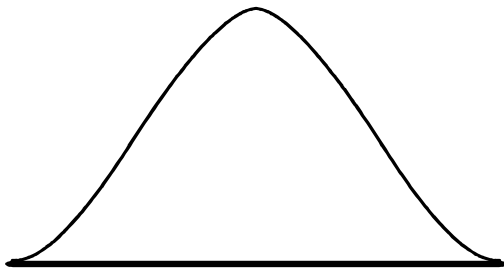
1. Gráfico de Puntos.

☒ Ejemplo

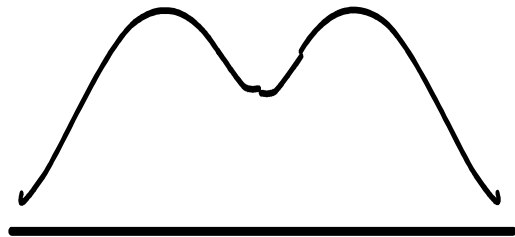
¿Cuántas llaves tiene en su bolsillo?

Haga un gráfico de frecuencias (de puntos) con el número de llaves que tienen los estudiantes que asisten hoy a clases. Describa la forma del gráfico.

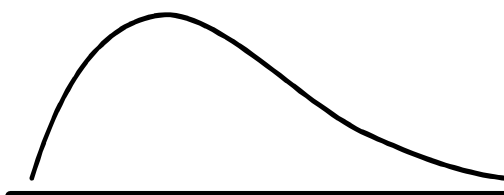
Formas de Distribuciones



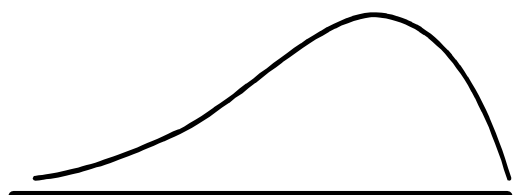
Simétrica, acampanada, unimodal



Bimodal



Sesgada a la derecha (sesgo positivo)



Sesgada a la izquierda (sesgo negativo)



Uniforme

Los términos usados para describir la forma de una distribución son:

- **Simétrica:** La distribución puede ser dividida en dos partes alrededor de un valor central y cada parte es el reflejo de la otra.
- **Sesgada:** Un lado de la distribución se alarga más que el otro. La dirección del sesgo es la dirección del lado más largo.
- **Unimodal:** La distribución tiene un único máximo que muestra el o los valores más comunes en los datos.
- **Bimodal:** La distribución tiene dos máximos. Esto resulta a menudo cuando la muestra proviene de dos poblaciones.
- **Uniforme:** Los valores posibles tienen la misma frecuencia.

☒ Ejemplo

BASE DE DATOS médica = medidas en 20 individuos que fueron parte de un estudio médico para reducir la presión sanguínea.

| Número | Sexo | Edad | N_tabletas | Presión_antes | Presión_después |
|--------|------|------|------------|---------------|-----------------|
| 1001 | M | 45 | 2 | 100.2 | 100.1 |
| 1002 | M | 41 | 1 | 98.5 | 100.0 |
| 1003 | F | 51 | 2 | 100.8 | 101.1 |
| 1004 | F | 46 | 2 | 101.1 | 100.9 |
| 1005 | F | 47 | 3 | 100.0 | 99.8 |
| 1006 | M | 42 | 2 | 99.0 | 100.2 |
| 1007 | M | 43 | 4 | 100.7 | 100.7 |
| 1008 | F | 50 | 2 | 100.3 | 100.9 |
| 1009 | M | 39 | 1 | 100.6 | 101.0 |
| 1010 | M | 32 | 1 | 99.9 | 98.5 |
| 1011 | M | 41 | 2 | 101.0 | 101.4 |
| 1012 | M | 44 | 2 | 100.9 | 100.8 |
| 1013 | F | 47 | 2 | 97.4 | 96.2 |
| 1014 | F | 49 | 3 | 98.8 | 99.6 |
| 1015 | M | 45 | 3 | 100.9 | 100.0 |
| 1016 | F | 42 | 1 | 101.1 | 100.1 |
| 1017 | M | 41 | 2 | 100.7 | 100.3 |
| 1018 | F | 40 | 1 | 97.8 | 98.1 |
| 1019 | M | 45 | 2 | 100.0 | 100.4 |
| 1020 | M | 37 | 3 | 101.5 | 100.8 |

2. Diagrama de Tallo y Hojas (Stem and Leaf).

Los gráficos o diagramas de tallo y hoja son una manera muy fácil de ordenar y mirar la distribución de los datos.

Pasos para hacer un Tallo y Hoja:

1. Separar cada medida en un tallo y una hoja.
Generalmente la hoja consiste en exactamente un dígito (el último) y el tallo consiste en uno o más dígitos.

Ejemplo: 734 => tallo=73, hoja=4 2,345 => tallo=2,34, hoja=5.

A veces se deja fuera el decimal pero se agrega una nota de cómo leer el valor.
Para 2,345 por ejemplo podremos decir que 234 | 5 se debe leer como 2,345.

2. Escribir los tallos en orden creciente de arriba abajo y dibujar una línea a la derecha de los tallos.
3. Agregar las hojas a su respectivo tallo en orden creciente.

☒ Ejemplo

Diagrama básico de Tallo y Hoja para la Edad de base de datos de un estudio médico.

Considere las edades de 20 sujetos de la base de datos médica.

1. Separamos los números en un tallo y una hoja:

| | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 45 | 41 | 51 | 46 | 47 | 42 | 43 | 50 | 39 | 32 |
| 41 | 44 | 47 | 49 | 45 | 42 | 41 | 40 | 45 | 37 |

2. Elegimos el tallo y lo anotamos en orden creciente

3. Agregamos las hojas en orden creciente:

Una modificación útil es que podemos **dividir los tallos**:

```

3 | 2
3 | 7 9
4 | 0 1 1 1 2 2 3 4
4 | 5 5 5 6 7 7 9
5 | 0 1

```

Note que el menor valor representado por 3 | 2 se lee 32 años.

Así podemos visualizar mejor que la distribución de las edades de los sujetos es aproximadamente simétrica, centrada en aproximadamente 43-44, sin valores extremos evidentes (observaciones que caen fuera del patrón general de datos).



Ejemplo

Puntajes de pruebas de dos estudiantes.

Estudiante A: 80 52 86 94 76 48 92 69 79 45

Estudiante B: 73 87 81 75 78 82 84 74 80 76

Construya un gráfico de tallo y hoja comparativo para los datos

¿Puede decir a cuál de los dos le ha ido mejor? Explique.



Pensemos

¿Qué está malo?

Explique por qué los siguientes gráficos de tallo y hojas no reflejan bien a la distribución de los datos.

Tallo y hoja 1

```

27 | 9
32 | 0 1 1 7 8
33 | 1 2 2 5 9
34 | 0 3 4
35 | 1 1

```

Nota: 27 | 9 representa 279

Tallo y Hoja 2

```

2 | 1 1 2 2 2 3 4 4 5 5 6 7 7 8 9
3 | 0 2 2 3 3 4 6 7 8
4 | 0 1 1

```

Nota: 2 | 1 representa 21

Tallo y hoja 3

```

18 | 1
19 | 0
20 |
21 | 1 2 8
22 | 0
23 | 7
24 |
25 | 5 8
26 | 2 3
27 | 0 5
28 | 1 2 9
29 | 2
30 | 7
31 | 6
32 |
33 | 0
34 |
35 | 0

```

Nota: 18 | 1 representa 181



En SPSS

Analizar > Estadísticos Descriptivos > Explorar > Gráficos > Tallo y Hojas.

TALLO1 Stem-and-Leaf Plot

| Frequency | Stem & | Leaf |
|-----------|--------|------|
| 1.00 | 27 . | 9 |
| 0.00 | 28 . | |
| 0.00 | 28 . | |
| 0.00 | 29 . | |
| 0.00 | 29 . | |
| 0.00 | 30 . | |
| 0.00 | 30 . | |
| 0.00 | 31 . | |
| 0.00 | 31 . | |
| 3.00 | 32 . | 011 |
| 2.00 | 32 . | 78 |
| 3.00 | 33 . | 122 |
| 2.00 | 33 . | 59 |
| 3.00 | 34 . | 034 |
| 0.00 | 34 . | |
| 2.00 | 35 . | 11 |

Stem width: 10
Each leaf: 1 case(s)

TALLO2 Stem-and-Leaf Plot

| Frequency | Stem & | Leaf |
|-----------|--------|------|
| 2.00 | 2 . | 11 |
| 4.00 | 2 . | 2223 |
| 4.00 | 2 . | 4455 |
| 3.00 | 2 . | 677 |
| 2.00 | 2 . | 89 |
| 1.00 | 3 . | 0 |
| 2.00 | 3 . | 22 |

Stem width: 10
Each leaf: 1 case(s)

TALLO3 Stem-and-Leaf Plot

| Frequency | Stem & | Leaf |
|-----------|--------|------------|
| 2.00 | 1 . | 89 |
| 5.00 | 2 . | 11123 |
| 10.00 | 2 . | 5566778889 |
| 1.00 | 3 . | 0 |

Stem width: 100
Each leaf: 1 case(s)



En SPSS

Analicemos ahora la salida que nos entrega el programa SPSS.

Estos diagramas contienen datos de la estatura (en cms) y de edad de los alumnos de la base de datos de 36 alumnos de Estadística:

Diagrama 1

ESTATURA Stem-and-Leaf Plot

| Frequency | Stem & Leaf |
|-----------|-----------------|
| 1.00 | 15 . 4 |
| 4.00 | 15 . 5555 |
| 7.00 | 16 . 0000222 |
| 3.00 | 16 . 567 |
| 7.00 | 17 . 0023444 |
| 3.00 | 17 . 578 |
| 10.00 | 18 . 0000022344 |
| .00 | 18 . |
| 1.00 | 19 . 4 |

Stem width: 10
Each leaf: 1 case(s)

Diagrama 2

EDAD Stem-and-Leaf Plot

| Frequency | Stem & Leaf |
|-----------|---------------------|
| 6.00 | 19 . 000000 |
| 14.00 | 20 . 00000000000000 |
| 6.00 | 21 . 000000 |
| 8.00 | 22 . 00000000 |
| .00 | 23 . |
| 1.00 | 24 . 0 |
| 1.00 | Extremes (>=25.0) |

Stem width: 1
Each leaf: 1 case(s)

3. Histograma

Los histogramas son otra manera de mostrar la distribución de una variable cuantitativa.

Pasos para hacer un histograma:

1. Dividir el rango de los datos (menor a mayor) en clases del mismo ancho. Las clases deben contener el rango posible de datos y no se deben superponer. Ej. Si los datos van de 0 a 29, comience en 0 hasta 30 de ancho 5.
2. Contar el número de observaciones (frecuencias) que caen en cada clase.
3. Dibujar en el eje horizontal y marcar las clases.
4. El eje vertical puede contener la frecuencia, la proporción, o el porcentaje.
5. Dibujar un rectángulo (una barra vertical) en cada clase con la altura igual a la frecuencia, la proporción, o el porcentaje.

☒ Ejemplo

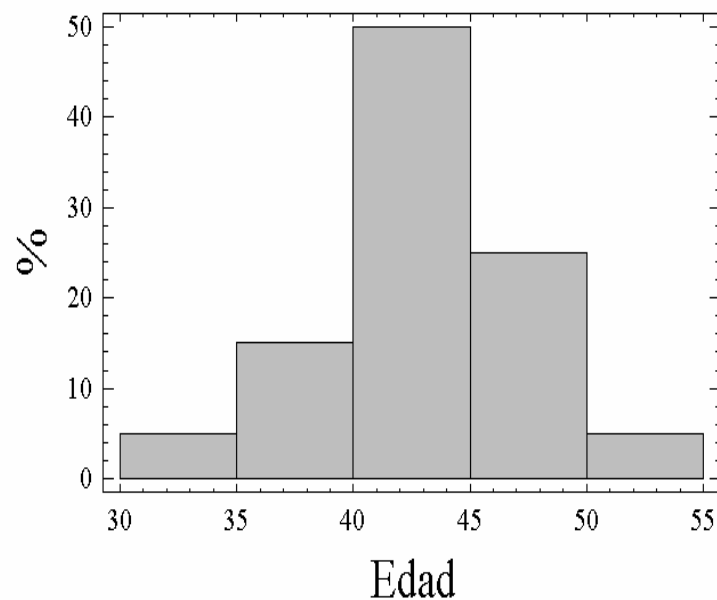
Histograma de Edad

Veamos nuevamente las edades de la base de datos médica. El rango va de 32 a 51, entonces podemos crear clases que comiencen en 30 con incrementos de 5 hasta 55. Puede intentar diferentes clases con distinto ancho hasta obtener una buena representación.

Para empezar es necesario construir una tabla de distribución de frecuencias:

| Clase | Cuenta | Número de observaciones | Porcentaje |
|---------|----------|-------------------------|---------------------------------|
| (30,35] | / | 1 | $1/20 = 0.05 \Rightarrow 5\%$ |
| (35,40] | /// | 3 | $3/20 = 0.15 \Rightarrow 15\%$ |
| (40,45] | //////// | 10 | $10/20 = 0.50 \Rightarrow 50\%$ |
| (45,50] | ///// | 5 | $5/20 = 0.25 \Rightarrow 25\%$ |
| (50,55] | / | 1 | $1/20 = 0.05 \Rightarrow 5\%$ |

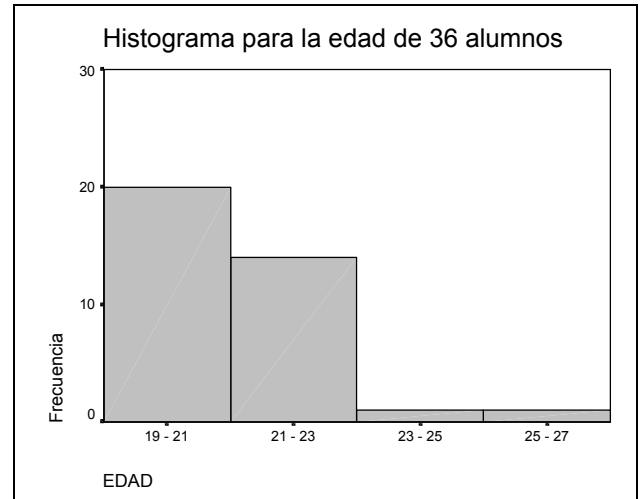
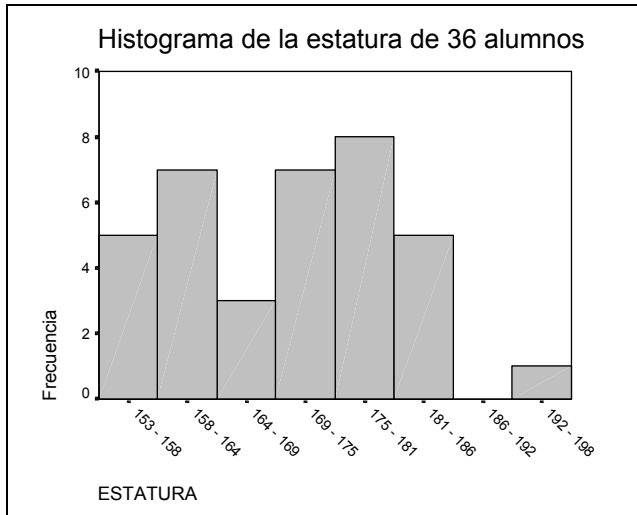
Histograma para Edad de base de datos médica:





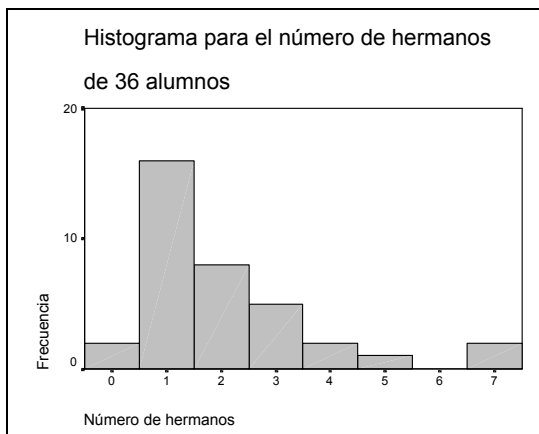
En SPSS

Gráficos > Generador de Gráficos > Histograma.



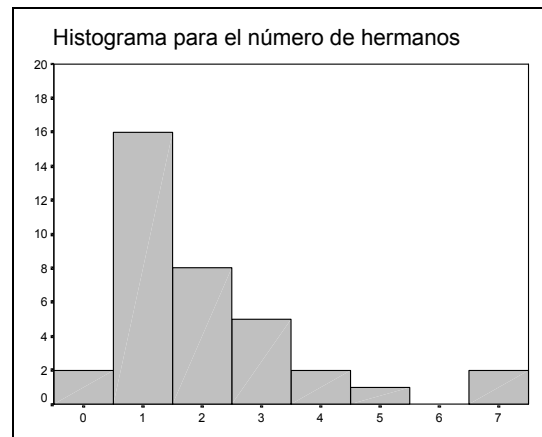
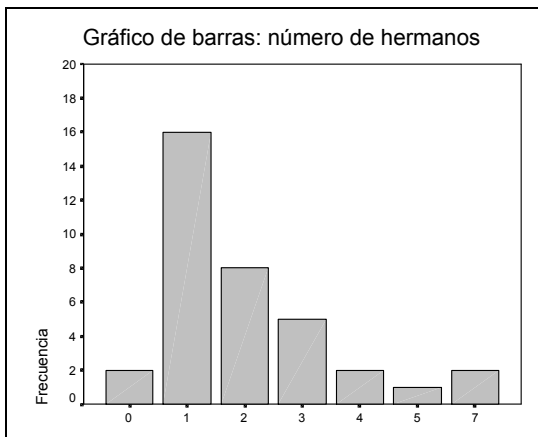
En SPSS

Comparemos histogramas con tallo y hoja.



| Número de hermanos | Stem-and-Leaf Plot |
|----------------------|----------------------|
| Frequency | Stem & Leaf |
| 2.00 | 0 . 00 |
| 16.00 | 1 . 0000000000000000 |
| 8.00 | 2 . 00000000 |
| 5.00 | 3 . 00000 |
| 2.00 | 4 . 00 |
| 1.00 | 5 . 0 |
| 2.00 | Extremes (>=7.0) |
| Stem width: 1 | |
| Each leaf: 1 case(s) | |

Cuidado con usar gráficos de barras para variables cuantitativas:



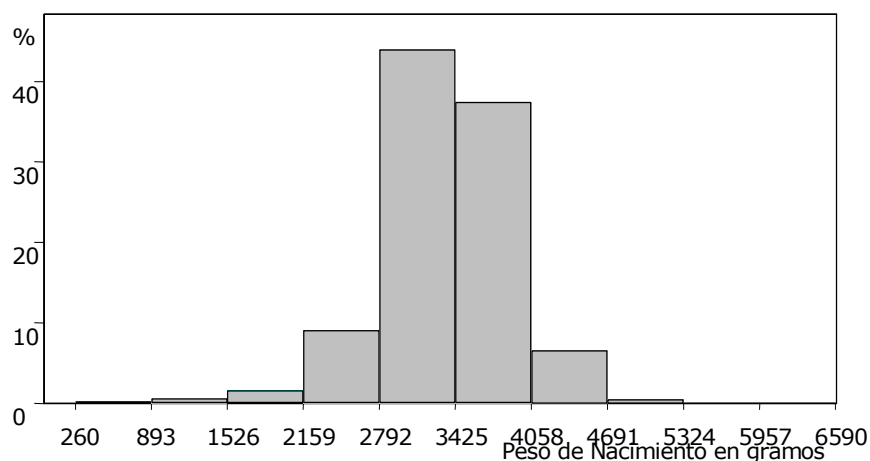
Guía para gráficos, figuras o diagramas:

Hay ciertos errores comunes que aparecen en gráficos que pueden hacer que se mal interprete la información. Cuando construya gráficos:

- Póngale un título apropiado.
- Incluya la fuente de los datos o cualquier información relevante.
- Escriba el nombre de la variable que se describe en los ejes.
- Incluya las unidades de medida de las variables.
- Verifique si el eje de la frecuencia, proporción o porcentaje comienza en cero.
- Verifique si los ejes mantienen una escala constante

☒ Ejemplo

Histograma del Peso al nacer de los recién nacidos en 1993 en Chile.



Fuente: Instituto Nacional de Estadística.

Métodos numéricos para describir variables cuantitativas

En este capítulo, empezamos a organizar y resumir los datos, primero tratamos las variables cualitativas, luego la descripción gráfica de variables cuantitativas, ahora estudiaremos cómo obtener buen resumen numérico de los datos. Específicamente estudiaremos medidas de resumen o medidas descriptivas numéricas que son de tres tipos:

- las que ayudan a encontrar el **centro** de la distribución, llamadas medidas de tendencia central.
- las que miden la **dispersión**, llamadas medidas de dispersión.
- las que describen la **posición relativa** de una observación dentro del conjunto de datos, llamadas medidas de posición relativa.

1. Medidas de Tendencia Central.

Las medidas de tendencia central son valores numéricos que quieren mostrar el centro de un conjunto de datos, nos interesan especialmente dos medidas: la **media** y la **mediana**.

Si los datos son una muestra, el promedio y la mediana se llamarán *estadísticas*. Si los datos son una población entonces estas medidas de tendencia central se llamarán *parámetros*.

Una **Estadística** es una medida descriptiva numérica calculada a partir de datos de una muestra.

Un **Parámetro** es una medida descriptiva numérica que usa la totalidad de las unidades de una población.

a) Promedio.

El **promedio** de un conjunto de n observaciones es simplemente la suma de las observaciones dividida por el número de observaciones, n .

Promedio de edad de los 20 sujetos en el estudio médico:

Sume las 20 edades y divida por 20:

$$\frac{45 + 41 + 51 + 46 + 47 + \dots + 45 + 37}{20} = 43,35 \text{ años}$$

Notación: Si X_1, X_2, \dots, X_n denota una muestra de n observaciones, entonces el *promedio de la muestra* se llama "x-barra" y se denota por:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Si se tiene TODOS los valores de una población, el promedio de la población es la suma de todos los valores dividida por cuántos son.

¹ Pueden revisar la notación de sumatorias en [Hopkins, K. Hopkins, B. Glass, G. \(1997\) Estadística básica para las ciencias sociales y del comportamiento. Tercera edición. Prentice Hall.](#)

El *promedio de la población* se denota por la letra Griega μ (mu): $\mu = \frac{\sum_{i=1}^N x_i}{N}$.

☒ Ejemplo

Número promedio de niños por hogar.

Los datos siguientes son el número de niños en una muestra aleatoria de 10 casas en un vecindario: 2, 3, 0, 2, 1, 0, 3, 0, 1, 4.

El promedio de estas 10 observaciones es: 1,6

El resultado es 1,6 aunque no sea posible observar 1,6 niños en una casa. El promedio es 1,6

Supongamos que una observación en la última casa se anotó como 40 en vez de 4, ¿Qué le pasará al promedio?

Notar que 9 de las 10 observaciones son menores que el promedio. El promedio es *sensible a las observaciones extremas*.

La mayoría de los métodos gráficos nos ayudarán de detectar observaciones extremas.

☒ Ejemplo

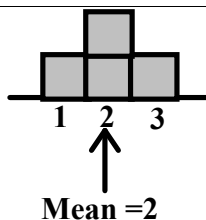
Un promedio NO es siempre representativo.

Las notas en varias pruebas de Juanita son 1,0 6,9 2,0 1,8 1,3, calcule el promedio de Juanita.

☒ Ejemplo

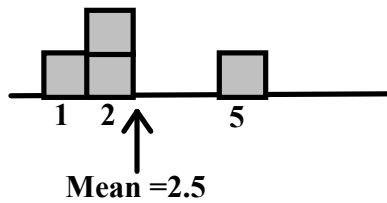
Combinando Promedios.

El promedio de 3 estudiantes es 5,4 y el promedio de otros 4 estudiantes es 6,7, ¿Cuál es el promedio de los 7 estudiantes?

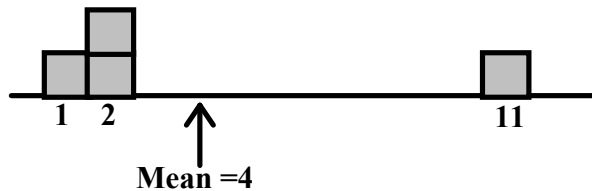


El promedio también se define como el **punto de equilibrio**, el punto donde distribución se balancea.

Si la distribución es **simétrica**, el promedio será exactamente el centro de la distribución.



Si la observación más grande se mueve a la derecha, el **promedio se mueve con la observación extrema**.



Si la distribución es sesgada, vamos a querer usar una medida que sea más **resistente** para mostrar el centro. La medida de tendencia central que es más resistente a los valores extremos es la **mediana**.

b) Mediana.

Definición:

La **mediana** de un conjunto de n observaciones, ordenadas de menor a mayor, es un valor tal que la mitad de las observaciones son menores o iguales que tal valor y la mitad de las observaciones son mayores o iguales que ese valor.

Pasos para encontrar la mediana:

1. Ordenar los datos de menor a mayor;
2. Calcular la posición de la mediana: $(n+1)/2$, donde n es el número de observaciones
3. a) Si el número de observaciones es **impar**, la mediana es un único término central.
b) Si el número de observaciones es **par**, la mediana es el promedio de los dos términos centrales.

☒ Ejemplo

Edades de $n=20$ sujetos...

Calculamos $(n+1)/2$ obtenemos $(20+1)/2 = 10,5$. Entonces los términos centrales son la décima y undécima observaciones, es decir 43 y 44. La mediana es el promedio de estos dos términos, $(43+44)/2=43,5$ años.

32 37 39 40 41 41 41 42 42 43 44 45 45 45 46 47 47 49 50 51

☒ Ejemplo

Mediana del número de niños por hogar.

Encuentre la mediana del número de niños por hogar en la muestra de 10 hogares.

Número de Niños: 2, 3, 0, 1, 4, 0, 3, 0, 1, 2.

- a) Ordenar las observaciones de menor a mayor:
- b) Calcular $(n+1)/2 =$ _____
- c) Mediana = _____
- d) ¿Qué le pasa a la mediana si la quinta observación en la lista se anota incorrectamente como 40 en vez de 4?
- e) ¿Qué le pasa a la mediana si la tercera observación en la lista se anota incorrectamente como -20 en vez de 0?

Nota: La **mediana es resistente (robusta)**, es decir, no cambia o cambia muy poco con observaciones extremas.

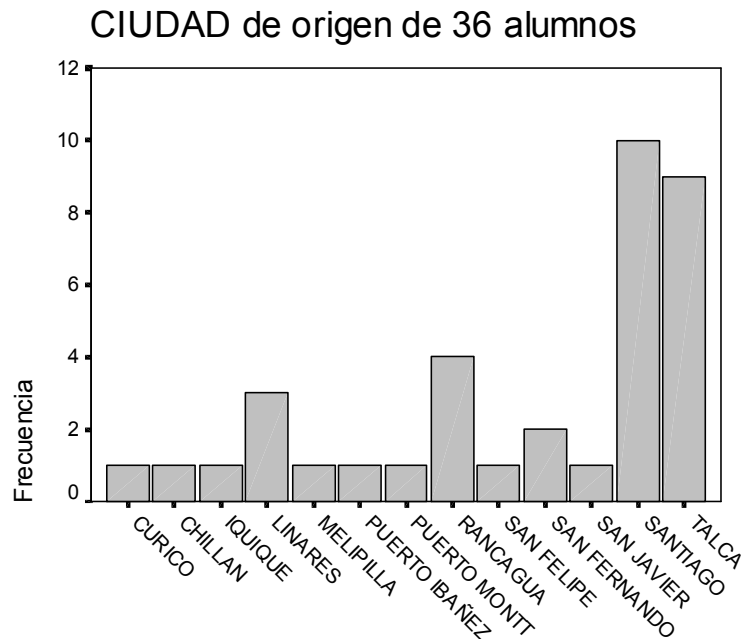
c) **Moda.**

Definición:

La **moda** de un conjunto de observaciones es el valor más frecuente.

- La moda de los valores: { 0, 0, 0, 0, 1, 1, 2, 2, 3, 4 } es 0.
- { 0, 0, 0, 1, 1, 2, 2, 2, 3, 4 } dos modas, 0 y 2 (*bimodal*).
- ¿Cuál sería la moda del siguiente conjunto de valores? { 0, 1, 2, 4, 5, 8 }.
- { 0, 0, 0, 0, 0, 1, 2, 3, 4, 4, 4, 4, 5 } ...

La Moda no se usa a menudo como medida de tendencia central para datos cuantitativos. Sin embargo la Moda es LA medida de tendencia central que puede ser calculada en datos **cualitativos**.



☒ Ejemplo

Diferentes medidas pueden dar diferentes impresiones.

El famoso trío - promedio, mediana y moda – representan tres métodos diferentes para encontrar EL valor del “centro”. Estos tres valores pueden ser un mismo valor pero a menudo son distintos. Cuando son distintos, pueden servir para diferentes interpretaciones de los datos que queremos resumir. Considere el ingreso mensual de cinco familias en un barrio:

\$120 000 \$120 000 \$300 000 \$900 000 \$1 000 000

¿Cuál es el ingreso **típico** de este grupo?

El ingreso mensual promedio es:

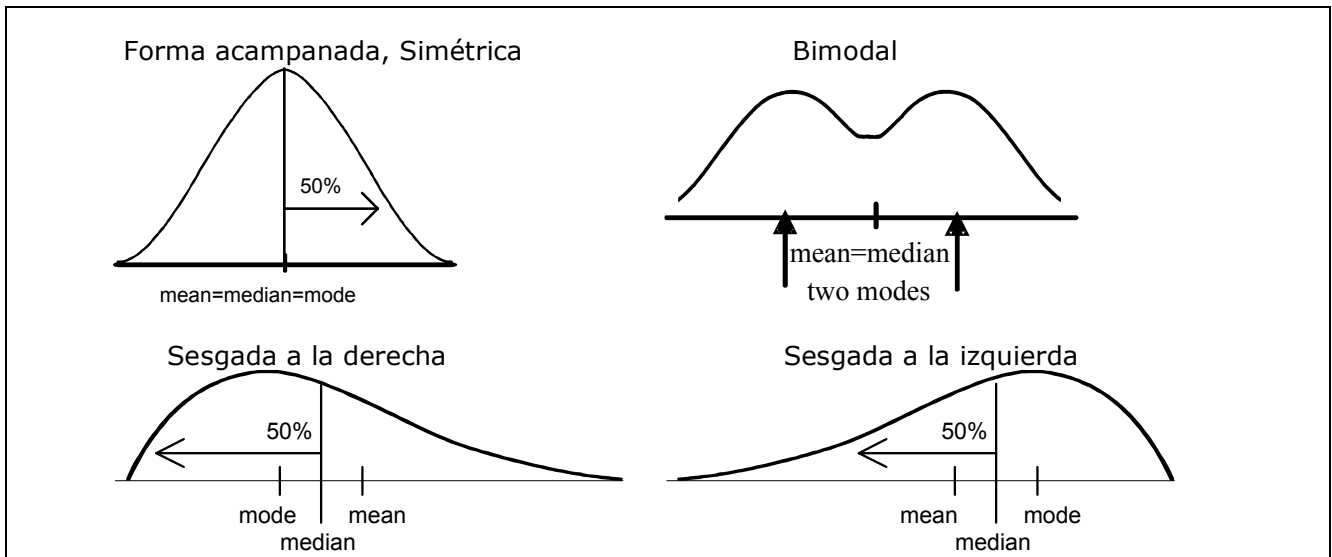
La mediana del ingreso mensual es:

La moda del ingreso mensual es:

Si Usted está tratando de promover el barrio, ¿Qué medida usaría?

Si Usted está tratando que bajen las contribuciones, ¿Qué medida usaría?

¿Cuál medida de tendencia central usar?



Pensemos

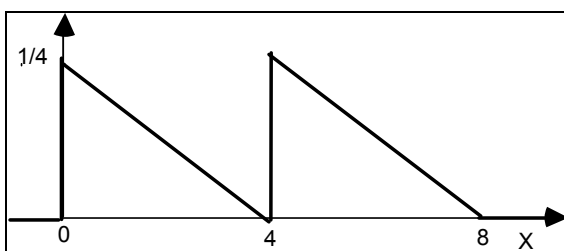
Suponga que calcula el promedio, mediana y moda de una lista de números, ¿Cuál medida es siempre un número en la lista?

Si la distribución es simétrica, ¿Cuál medida de tendencia central calcularía: el promedio o la mediana?, ¿Por qué?



Ejemplo

Una distribución diferente.



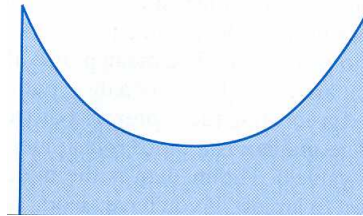
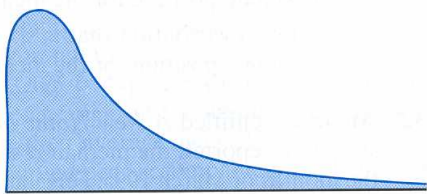
En la figura se muestra la distribución de una variable:

- ¿Es esta distribución simétrica?
- ¿Su mediana es menor, igual o mayor a 4?
- ¿Su promedio es menor, igual o mayor a 4?

☒ Ejemplo

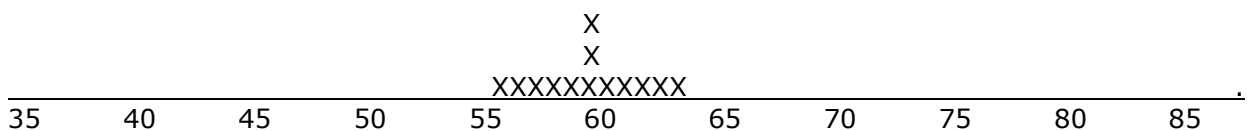
Buenas o malas medidas.

Para los siguientes gráficos describa qué tan buenas o malas son las tres medidas de tendencia central como descripción del centro de la distribución:

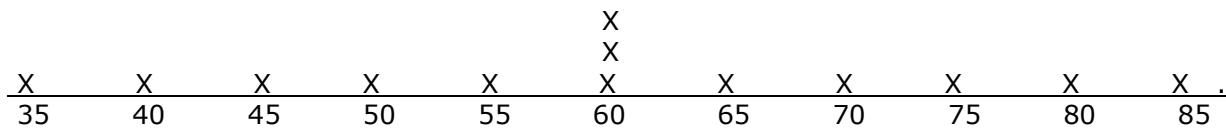

2. Medidas de Dispersión.

Las medidas de tendencia central son útiles pero nos dan una interpretación parcial de los datos. Considere los dos siguientes conjuntos de datos:

Datos 1: 55, 56, 57, 58, 59, 60, 60, 60, 61, 62, 63, 64, 65

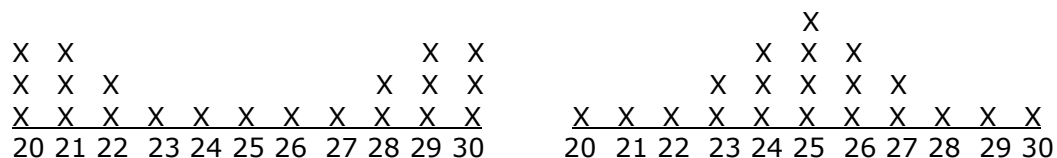


Datos 2: 35, 40, 45, 50, 55, 60, 60, 60, 65, 70, 75, 80, 85


a) Rango.

Es la medida de variabilidad o dispersión más simple. Se calcula tomando la diferencia entre el valor máximo y el mínimo observado.

$$\text{Rango} = \text{Máximo} - \text{Mínimo}.$$

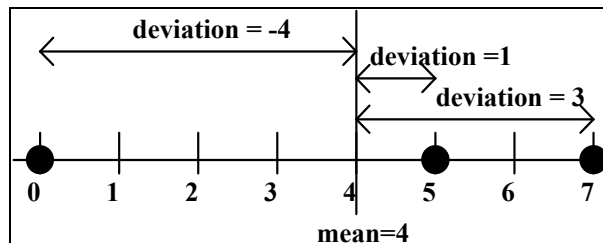


Analice cuáles podrían ser las ventajas y desventajas del rango como medida de variabilidad.

b) Desviación Estándar.

Es una medida *de la dispersión de las observaciones a la media*. Es un “promedio de la distancia de las observaciones a la media”.

☒ Ejemplo



| Observación | Desviación | Desviación al cuadrado |
|--------------|---------------|------------------------|
| x | $x - \bar{x}$ | $(x - \bar{x})^2$ |
| 0 | $0 - 4 = -4$ | 16 |
| 5 | $5 - 4 = 1$ | 1 |
| 7 | $7 - 4 = 3$ | 9 |
| Promedio = 4 | Suma = 0 | Suma = 26 |

La **varianza muestral** está definida como la suma de las desviaciones al cuadrado divididas por el tamaño muestral menos 1, es decir, divididas por $n - 1$.

$$\text{varianza muestral} = \frac{(-4)^2 + (1)^2 + (3)^2}{3 - 1} = \frac{16 + 1 + 9}{2} = \frac{26}{2} = 13$$

$$\text{desviación estándar muestral} = \sqrt{13} \approx 3,6$$

☒ Ejemplo

Desviación estándar para el número de niños por hogar.

Recordemos los datos del número de niños por hogar en una muestra de 10 casas de un barrio: 2, 3, 0, 2, 1, 0, 3, 0, 1, 4

Use su calculadora científica y compruebe el siguiente resultado:

"Los hogares tienen, **en promedio** 1,6 niños con una **variación** de alrededor de 1,43 niños".



En Resumen

Pensemos la desviación estándar como aproximadamente un *promedio de las distancias* de las observaciones a la media.

Si todas las observaciones son iguales, entonces la desviación estándar es cero.

La desviación estándar es positiva y mientras más alejados están los valores del promedio, mayor será la desviación estándar.

Si X_1, X_2, \dots, X_n denota una muestra de n observaciones, la **varianza muestral** se denota por:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

La desviación estándar muestral, denotada por s , es la raíz cuadrada de la varianza:

$$S = \sqrt{S^2}.$$

La **desviación estándar poblacional**, se denota por la letra Griega σ (sigma), es la raíz cuadrada de la varianza poblacional y se calcula como:

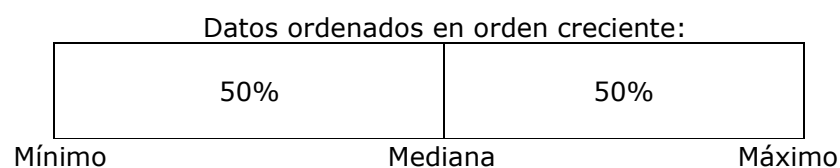
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}.$$

Notas:

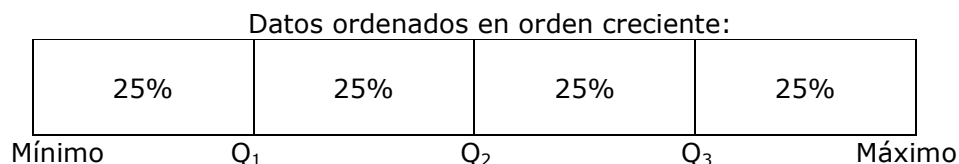
- La varianza y la desviación estándar no son medidas de variabilidad distintas, debido a que la última no puede determinarse a menos que se conozca la primera.
- A menudo se prefiere la desviación estándar en relación con la varianza, porque se expresa en las mismas unidades físicas de las observaciones.
- Así como el promedio es una medida de tendencia central que no es resistente a las observaciones extremas, la desviación estándar, que usa el promedio en su definición, tampoco es una medida de dispersión resistente a valores extremos.
- Tenemos argumentos estadísticos para demostrar por qué dividimos por $n - 1$ en vez de n en el denominador de la varianza muestral.

Cuartiles

La mediana de una distribución divide los datos en dos partes iguales:



También es posible dividir los datos en más de dos partes. Cuando se dividen un conjunto ordenado de datos en cuatro partes iguales, los puntos de división se conocen como **cuartiles** y los representamos por Q_1 , Q_2 y Q_3 .



c) Rango entre cuartiles.

La diferencia entre el tercer cuartil y el primer cuartil se llama **rango entre cuartiles**, denotado por $RQ = Q_3 - Q_1$. El rango entre cuartiles mide la variabilidad de la mitad central de los datos.

Pasos para calcular cuartiles:

1. Encontrar la mediana de todas las observaciones.
2. Encontrar el primer cuartil = Q_1 = mediana de las observaciones que son menores a la mediana.
3. Encontrar el tercer cuartil = Q_3 = mediana de las observaciones que son mayores a la mediana.

Notas:

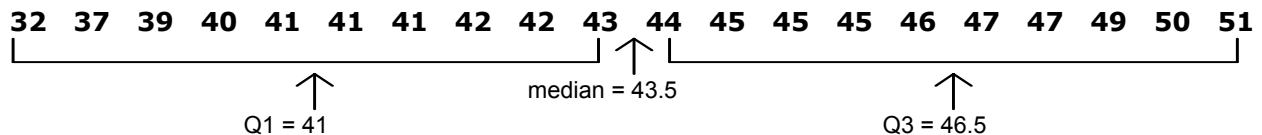
- Cuando el número de observaciones es impar, la observación del medio es la mediana. Esta observación no se incluye luego en los cálculos de Q_1 y Q_3 .
- Pueden encontrar diferentes fórmulas en libros, calculadoras o computadores, pero todas estas fórmulas se basan en el mismo concepto.
- Si la distribución es simétrica, los cuartiles deben estar a la misma distancia de la mediana.



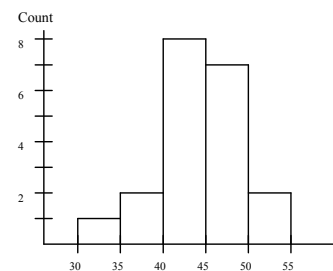
Ejemplo

Cuartiles para la Edad.

Lista ordenada de las edades de los 20 sujetos en el estudio médico:



Podemos ver que la distribución de la edad es aproximadamente simétrica y que los cuartiles están casi a la misma distancia de la mediana.



☒ Ejemplo

¿Qué es Variabilidad?

Considere los 4 conjuntos de datos siguientes y sus histogramas:

Datos I:

2 3 3 3 4 4 4 4 5 5 5
5 5

Datos II:

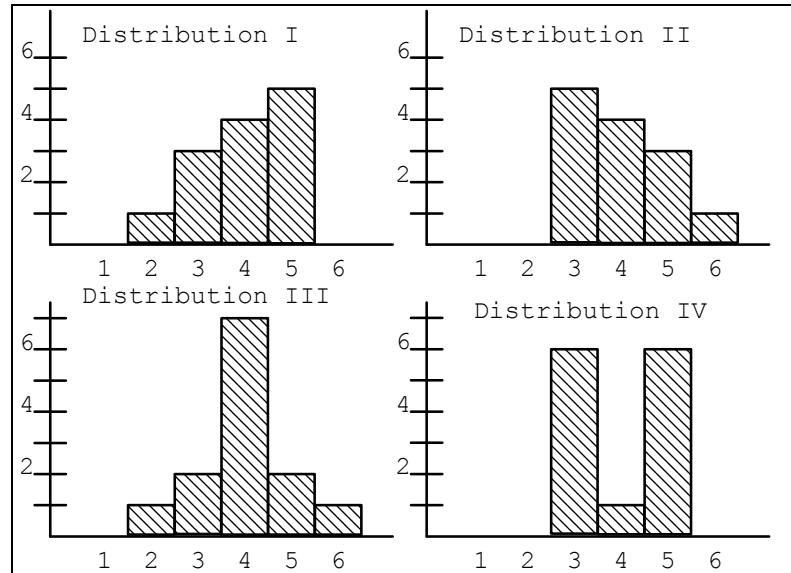
3 3 3 3 3 4 4 4 4 5 5
5 6

Datos III:

2 3 3 4 4 4 4 4 4 4 5
5 6

Datos IV:

3 3 3 3 3 3 4 5 5 5 5
5 5



| Medidas de variabilidad | I | II | III | IV |
|-------------------------|---|----|-----|----|
| Rango | | | | |
| Rango entre cuartiles | | | | |
| Desviación Estándar | | | | |

Algunas personas asocian variabilidad con rango mientras que otras asocian variabilidad con cómo difieren los valores de la media. Hay muchas medidas de variabilidad, y la desviación estándar es la más usada. Pero recuerden que una distribución con la menor desviación estándar no es necesariamente la distribución que es menos variable con respecto a otras definiciones de variabilidad².

² Referencia: Nitko, A. (1983) *Educational Tests and Measurement: An Introduction*. Harcourt.

En Resumen

Cuando queremos describir una variable usamos alguna medida de posición central y una medida de dispersión. El par de medidas más comúnmente usado es el promedio y la desviación estándar. Pero vimos que cuando la distribución de las observaciones es sesgada, el promedio no es una buena medida de posición central y preferimos la mediana. La mediana en general va acompañada del rango como medida de dispersión. Pero cuando observamos valores extraños (extremos) el rango se ve muy afectado, por lo que preferimos usar el rango entre cuartiles.

| Medida de tendencia central | Medida de dispersión | Uso en Distribuciones | Ventajas | Desventajas |
|-----------------------------|-----------------------|--------------------------------|--|--|
| Promedio | Desviación estándar | Simétricas | Buenas propiedades, muy usados. | Sensible a valores extremos. |
| Mediana | Rango | Sesgadas, sin valores extremos | Mediana robusta a valores extremos. Rango muy conocido, fácil de entender. | Rango sensible a valores extremos. |
| Mediana | Rango entre cuartiles | Sesgadas con valores extremos | Medidas robustas a valores extremos. | El rango entre cuartiles no es muy conocido. |

3. Medidas de posición relativa.

Los **cuartiles** dividen un conjunto ordenado de datos, en cuatro partes iguales:

| Datos ordenados en orden creciente: | | | | |
|-------------------------------------|-------|-------|-------|--------|
| 25% | 25% | 25% | 25% | |
| Mínimo | Q_1 | Q_2 | Q_3 | Máximo |

También podemos dividir conjuntos de datos en 100 partes iguales y los puntos de división se conocen como **percentiles**.

| Datos ordenados en orden creciente: | | | | | | | | | | | |
|-------------------------------------|-------|-------|-------|----|----|----|-----|----|----|----------|----------|
| 1% | 1% | 1% | 1% | 1% | 1% | 1% | ... | 1% | 1% | 1% | 1% |
| Mín | P_1 | P_2 | P_3 | . | . | . | . | . | . | P_{97} | P_{98} |
| | | | | | | | | | | P_{99} | Máx |

Es así como los cuartiles son en realidad los **percentiles** 25, 50 y 75, respectivamente.

En general, el **k -ésimo percentil** es un valor tal que el **$k\%$** de los datos son menores o iguales que él, y el **$(100-k)\%$** restante son mayores o iguales que él.

| Datos ordenados en orden creciente: | |
|-------------------------------------|--------------|
| $k\%$ | $(100-k)\%$ |
| Mínimo | P_k Máximo |

Por ejemplo, el 25-ésimo percentil o **percentil 25** (P_{25}) es un valor tal que el **25%** de los datos son menores o iguales que él, y el **(100-25) = 75%** restante son mayores o iguales que él.

Definición:

Las **medidas de posición relativa** son medidas que describen la posición que tiene un valor específico en relación con el resto de los datos.

☒ Ejemplo

Si su nota estuvo en el percentil 84, entonces el 84% de las notas fueron inferiores a la suya y el 16% superiores.

Además existen los quintiles y los deciles, ¿Cuáles serán?

Usos de medidas de posición relativa en:

- Calificaciones de exámenes.
- Puntajes en tests Psicológicos.
- Curvas de crecimiento en salud (<http://www.cdc.gov/growthcharts/>)

Definición

Valores extremos (outliers): son valores que se alejan del conjunto de datos.

Regla para identificar valores o datos extremos:

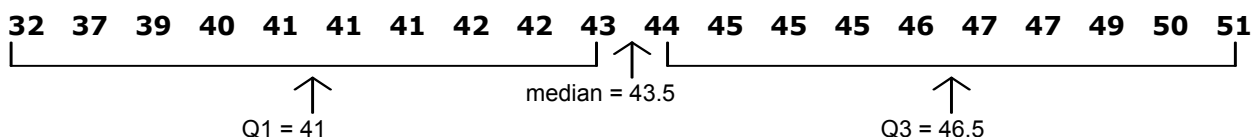
Vamos a definir una observación x_i como **extrema** si:

$$x_i < Q1 - 1,5 * (Q3 - Q1) \quad \text{o} \quad x_i > Q3 + 1,5 * (Q3 - Q1)$$

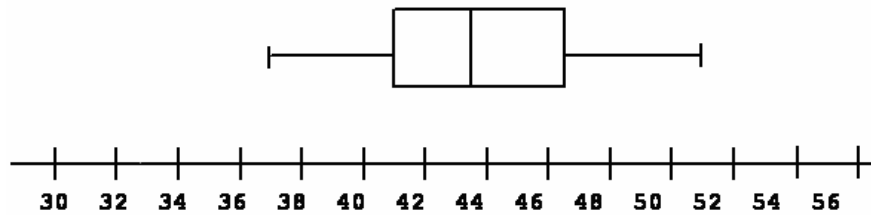
donde x_i serán las primeras y últimas observaciones en la serie ordenada de los datos.

☒ Ejemplo

¿Tiene valores extremos, la variable edad de los 20 sujetos en el estudio médico?



Diagramas de caja (boxplot):



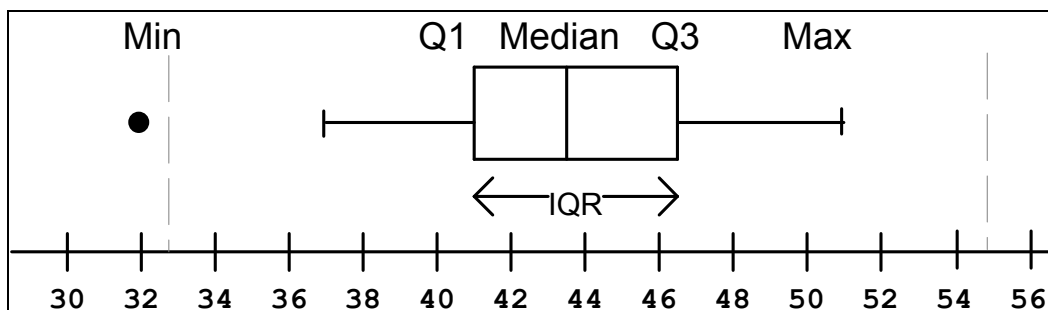
El diagrama de **caja** se construye de la siguiente manera:

1. Dibujar la caja que empieza en el primer cuartil y termina en el tercer cuartil.
2. Dibujar la mediana con una línea dentro de la caja.
3. Por último se extienden las líneas, llamadas bigotes, saliendo de la caja hasta el mínimo y el máximo (salvo en la presencia de valores extremos).

☒ Ejemplo

Gráfico de caja para la EDAD

min = 32 Q1 = 41 mediana = 43,5 Q3 = 46,5 max = 51

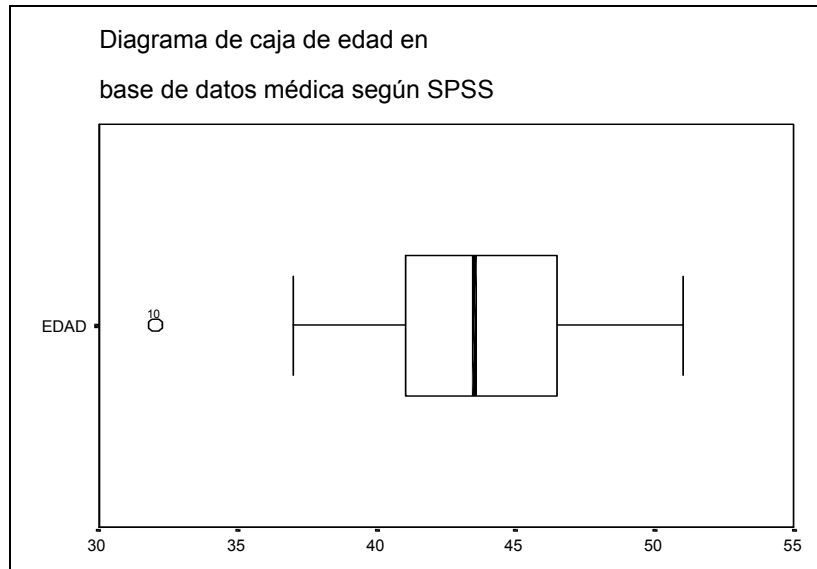


En la presencia de valores extremos, los "bigotes" se extienden hasta el valor observado anterior al valor extremo.



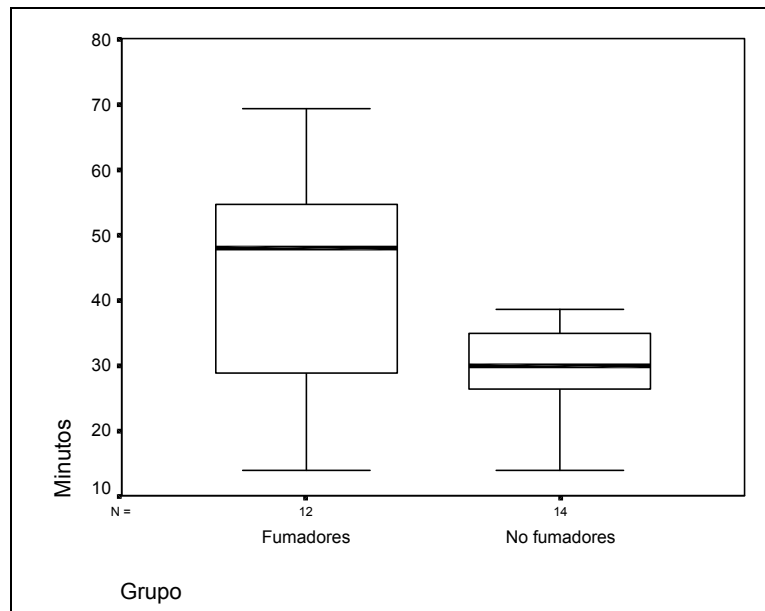
En SPSS

Gráfico > Generador de Gráficos > Diagrama de Caja.



La distancia entre la mediana y los cuartiles es aproximadamente la misma, lo que nos hace pensar que la distribución de los datos es más o menos simétrica como vimos antes en el histograma y en el tallo y hoja.

Los gráficos de caja son muy útiles para comparar distribuciones de dos o más grupos. Por ejemplo, comparar los grupos de fumadores y no fumadores (ver ejercicios propuestos).



✓ Ejemplo

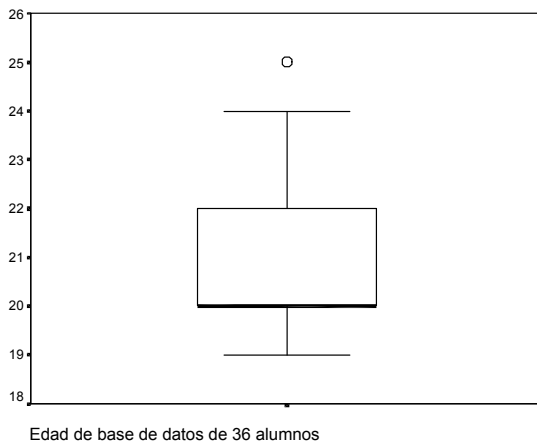
En diciembre de 2004, SERNAC realizó un estudio acerca del precio de las bicicletas en Santiago. En el siguiente gráfico de caja se presentan los precios de 5 bicicletas Bianchi Modelo Goliat 12:



- ¿Cuál es el rango aproximado del precio de las bicicletas?
- ¿Cuál es el valor aproximado del 25% de las bicicletas más caras?

✓ Ejemplo

Identifique las 5 medidas de resumen e identifique los valores extremos:



| Edad en años Stem-and-Leaf Plot | | |
|---------------------------------|----------|------------------|
| Frequency | Stem | Leaf |
| 6.00 | 19 | . 000000 |
| 14.00 | 20 | . 00000000000000 |
| 6.00 | 21 | . 000000 |
| 8.00 | 22 | . 00000000 |
| .00 | 23 | . |
| 1.00 | 24 | . 0 |
| 1.00 | Extremes | (>=25.0) |
| Stem width: 1 | | |
| Each leaf: 1 case(s) | | |

| | | | | |
|--------------------|------------|-----------------|------------|----------------|
| Mínimo = _____ | Q1 = _____ | Mediana = _____ | Q3 = _____ | Máximo = _____ |
| ¿Valores extremos? | | | | |



Pensemos

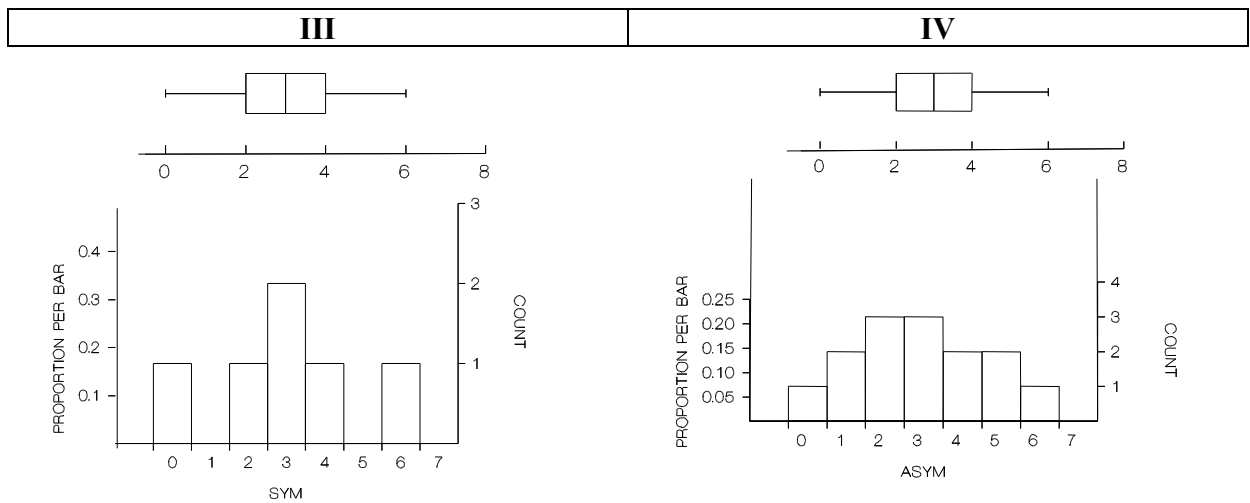
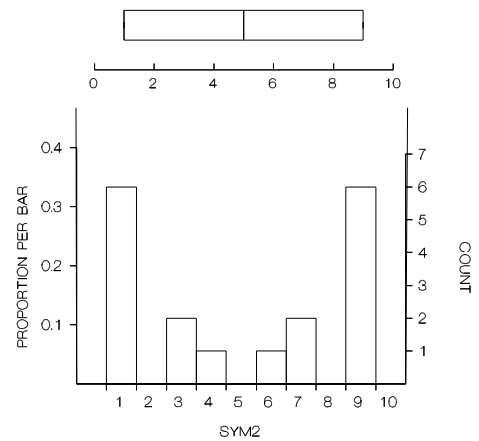
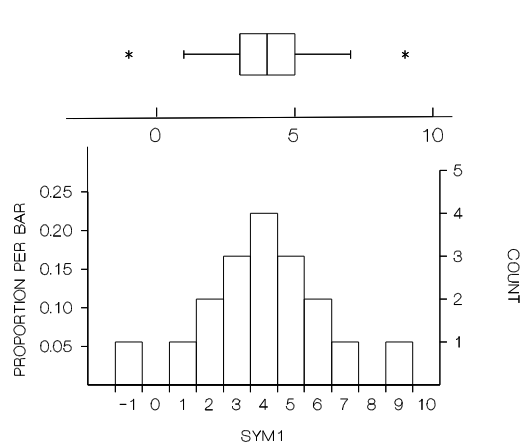
Si el gráfico de caja es simétrico, ¿Podemos concluir que la distribución de los datos es simétrica?



Ejemplo

Considere los siguientes conjuntos de datos (ordenados):

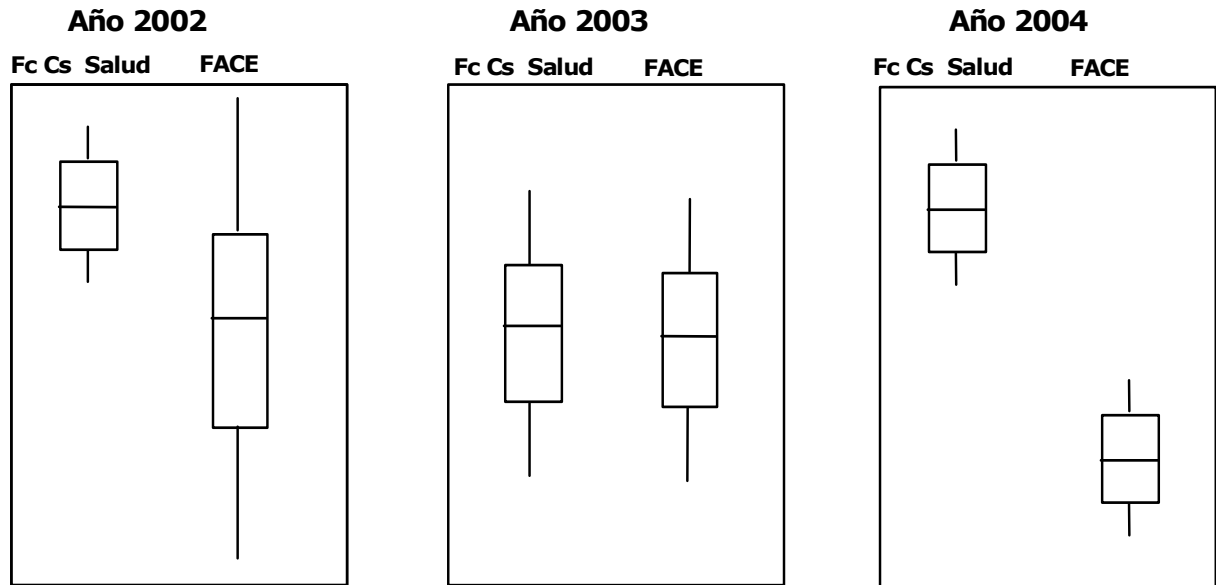
| I | II |
|--|---|
| -1 1 2 2 3 3 3 4 4 4 4 5 5 5 6 6 7 9 | 1 1 1 1 1 1 3 3 4 6 7 7 7 9 9 9 9 9 9 |
| Mínimo = -1 $Q_1 = 3$ Mediana = 4 $Q_3 = 5$ Máximo = 9 | Mínimo = 1 $Q_1 = 1$ Mediana = 5 $Q_3 = 9$ Máximo = 9 |



☒ Ejemplo

Diseño muestral.

Los gráficos representan las notas en dos cursos de Estadística de los 3 últimos años que se dictan para la Facultad de Ciencias de la Salud y la Facultad de Ciencias Económicas.



Considere tres diseños muestrales para estimar la verdadera media poblacional de las notas:

- muestreo aleatorio simple.
- muestreo aleatorio estratificado tomando muestras del mismo tamaño en cada estrato.
- muestreo aleatorio estratificado tomando más unidades de un estrato que de otro.

Asuma que el tamaño muestral total es igual en todos los diseños.

- ¿Para qué población (2002, 2003 o 2004) los diseños (i) y (ii) son igualmente efectivos?
- ¿Para qué población (2002, 2003 o 2004) el diseño (ii) será el mejor?
- ¿Para qué población (2002, 2003 o 2004) el diseño (iii) será el mejor?, ¿De cuál Facultad se debe obtener una muestra de mayor tamaño?

Anexo: Transformaciones lineales y estandarización.



Una transformación.

Se tiene datos del número de niños por hogar de 10 viviendas de un barrio:

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

El promedio es 2,0 y desviación estándar = 1,1547 niños

- Suponga que queremos describir el número de personas en cada vivienda y suponga que en cada vivienda hay 2 adultos: 4, 5, 4, 4, 3, 2, 5, 4, 3, 6
 - Encuentre el promedio y la desviación estándar de esta nueva variable y compare con las observaciones originales.
 - ¿Cómo cambia el promedio? ¿Cómo cambia la desviación estándar?
 - Describa como afecta al promedio y la desviación estándar el *sumar* una constante a cada observación.
- Suponga que cada niño recibe una mesada semanal de \$500. Describa ahora el gasto en mesadas de cada vivienda.
 - Encuentre el promedio y la desviación estándar y compare con los obtenidos de las observaciones originales.
 - ¿Cómo cambia el promedio?, ¿Cómo cambia la desviación estándar?
 - Describa cómo afecta al promedio y la desviación estándar el *multiplicar* una constante a cada observación.

Si X representa una variable, \bar{x} su promedio y s_x su desviación estándar. Sea $Y = aX + b$, una **transformación lineal** de X , entonces:

El promedio de Y es: $\bar{y} = a\bar{x} + b$

y la desviación estándar: $s_y = |a|s_x$

NOTA: $|a|$ es el valor absoluto o módulo de la constante a , donde a es cualquier valor positivo o negativo y su módulo es siempre positivo.



La temperatura mínima en Talca la semana del 14 al 20 de Mayo de 2001 fue de:

| | Lunes | Martes | Miércoles | Jueves | Viernes | Sábado | Domingo |
|------------|-------|--------|-----------|--------|---------|--------|---------|
| $X = t$ °F | 38 | 46 | 38 | 50 | 45 | 34 | 43 |

El promedio y la desviación estándar son: $\bar{x} = 42$ grados Fahrenheit y $s_x = 5,67$ grados Fahrenheit.

Sea Y = la temperatura en escala de grados Celsius, Y está relacionada con X = la temperatura en grados Fahrenheit, por la siguiente transformación lineal: $C = \frac{5}{9}(F - 32)$, o

en términos de Y y X : $Y = \frac{5}{9}X - \frac{160}{9}$.

Calcule el promedio y la desviación estándar en grados Celsius.

Si X representa una variable, \bar{x} su promedio y s_x su desviación estándar. Llamaremos z a la variable estandarizada:

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

Una variable está **estandarizada** si la variable tiene media cero y desviación estándar uno.

Note que la **variable estandarizada** $\frac{x - \bar{x}}{s_x}$ se puede expresar de la forma de una **transformación lineal**:

$$\frac{x - \bar{x}}{s_x} = \left(\frac{1}{s_x}\right)x + \left(-\frac{\bar{x}}{s_x}\right) \text{ con } a = \left(\frac{1}{s_x}\right), \text{ y } b = \left(-\frac{\bar{x}}{s_x}\right).$$

Calcule el promedio y desviación estándar de la variable número de niños estandarizada.

Transformaciones no lineales³

☒ Ejemplo

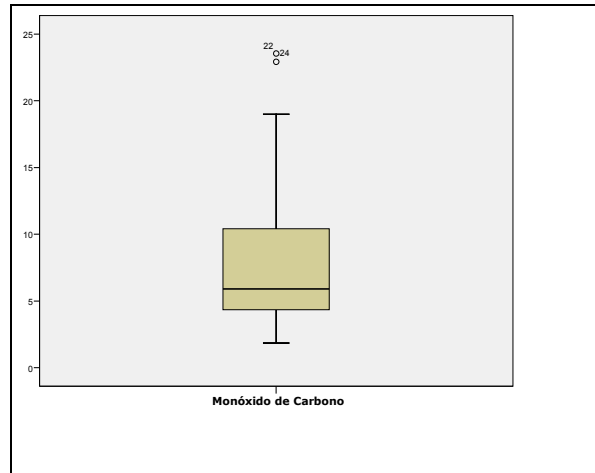
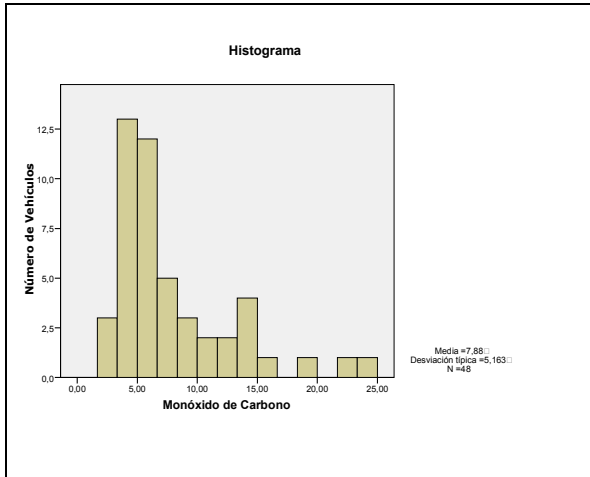
Se tienen datos sobre la emisión de monóxido de Carbono de 46 vehículos del mismo tipo (Monoxido.sav).

| EN | HC | CO | NOX |
|----|------|-------|------|
| 1 | 0,5 | 5,01 | 1,28 |
| 2 | 0,65 | 14,67 | 0,72 |
| 3 | 0,46 | 8,6 | 1,17 |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| 44 | 0,46 | 3,99 | 2,01 |
| 45 | 0,47 | 5,22 | 1,12 |
| 46 | 0,55 | 7,47 | 1,39 |

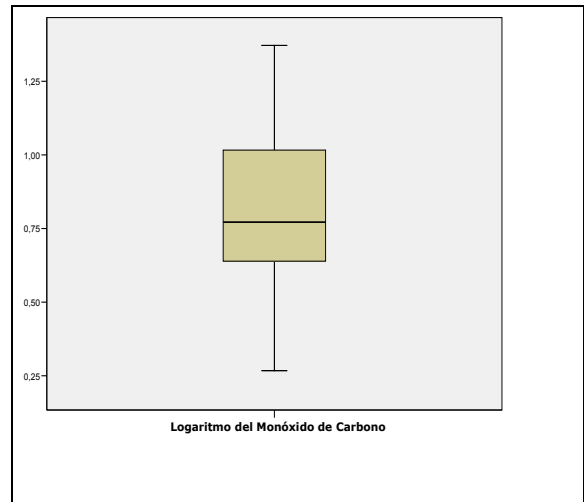
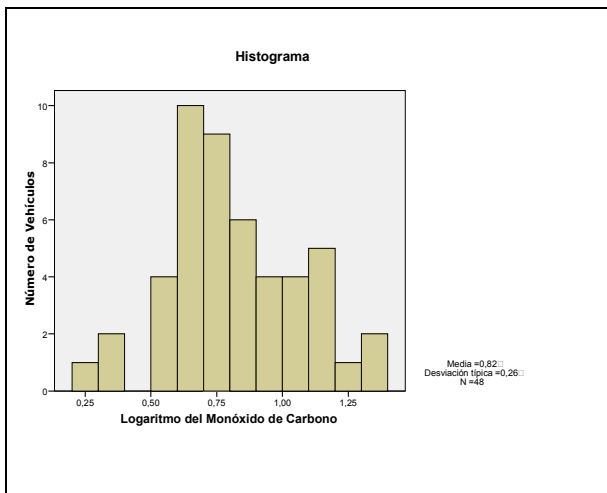
A los investigadores les interesa calcular la media del monóxido de Carbono. Si analizamos el histograma adjunto, vemos que la distribución del monóxido de Carbono es sesgada a la derecha, por lo que la media no será un buen estimador del centro de la distribución. Como solución podemos transformar la variable usando el logaritmo natural y calculamos el

³ Lectura complementaria Capítulo 6 de Peña, D. Romo, J. (1999) Introducción a la Estadística para las Ciencias Sociales. McGraw Hill.

promedio de la nueva variable. Pero al investigador le interesa conocer el valor de la media en las unidades originales de la variable, para eso convertimos a la unidad original de CO con exponencial ($e^{0,82} = 2,2705$). Esta media de la variable transformada se conoce como media geométrica.



Media = 7,88.



Media = 2,2705.