Facultad de Ciencias Exactas, Ingeniería y Agrimensura

FCEIA

UNIVERSIDAD NACIONAL DE ROSARIO

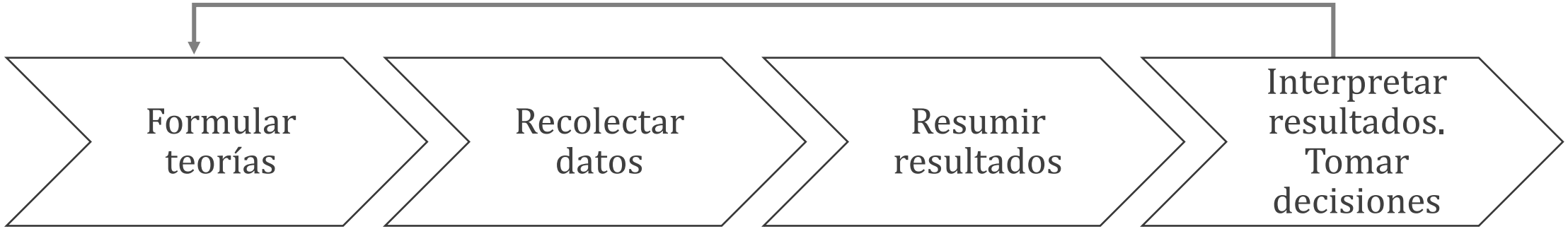
Probabilidad y Estadística

Unidad 2: Resumen de datos en forma gráfica y numérica

Lic. Maite San Martín

Marzo 2024

Método científico



Estamos aquí

Sistematización de la información

Resumen, análisis y presentación de la información.

Estadística descriptiva
Introducción

Posiblemente un investigador que ha recopilado datos desee resumir y describir características importantes de los mismos. Esto implica utilizar métodos de **estadística descriptiva**. Algunos de estos son de naturaleza gráfica, como la construcción de histogramas, diagramas de caja y gráficos de barras o de sectores, entre tantos otros. Otros métodos descriptivos implican calcular medidas numéricas, tales como los promedios, desviaciones estándar y coeficientes de correlación.

Si bien en otros tiempos todos estos resúmenes se realizaban manualmente, hoy en día el soporte informático hace todo más sencillo. Desde computadoras hasta calculadoras y teléfonos celulares ofrecen aplicaciones y programas para construir gráficos y hacer resúmenes numéricos.

Estadística descriptiva
Tipos de análisis

Podemos clasificar la estadística descriptiva desde distintas aristas. Ya dijimos que podíamos obtener resúmenes gráficos o numéricos, en función del tipo de resultado que deseamos. Asimismo, podemos también pensar en la cantidad de variables que deseamos resumir de forma simultánea. Por ejemplo, de un grupo de personas podemos determinar qué porcentaje de ellas fuman y cuál no, así como también la distribución según franja etaria (**análisis univariado**, de a una variable por vez), a la vez que podríamos analizar de forma simultánea estas dos variables para determinar si el porcentaje de fumadores es mayor en personas de menor o mayor edad (**análisis bivariado**, de a pares de variables simultáneamente), y también realizar este análisis según género (**análisis multivariado**, tres o más variables a la vez).

Estadística descriptiva
Tipos de variables

Asimismo, el tipo de análisis descriptivo a realizar (sea este numérico o gráfico, o bien uni, bi o multivariado) depende de un aspecto fundamental: cuál es el tipo de variable/s que se desea analizar.

Ya mencionamos que una variable es una característica observable que varía entre los diferentes individuos de una población.

Cuando la característica que medimos clasifica los elementos en clases o categorías hablamos de **variables cualitativas o categóricas**. Sus valores (categorías o niveles) no se pueden asociar naturalmente a un número, y no se pueden hacer operaciones algebraicas con ellos. Ejemplos de variables categóricas pueden ser el género de una persona, la marca de un auto, la opinión respecto a alguna postura (a favor o en contra).

Estadística descriptiva
Tipos de variables

Cuando la característica que medimos es una cantidad, entonces hablamos de **variables cuantitativas**. Sus valores son numéricos y en este caso sí tiene sentido hacer operaciones algebraicas con ellos.

Dentro de las variables cuantitativas pueden a su vez diferenciarse dos tipos:

- **Variables cuantitativas discretas**: pueden tomar valores enteros, y en general provienen de procesos de conteo. Por ejemplo, cantidad de hijos de una persona, cantidad de materias aprobadas por un estudiante, cantidad de canciones en una lista de Spotify.
- **Variables cuantitativas continuas**: pueden tomar infinita cantidad de valores distintos dentro de un intervalo dado, y suelen provenir de procesos de medición. Por ejemplo, la altura de una persona, el peso de una encomienda, el tiempo que se tarda en llegar a una meta.

Estadística descriptiva
Escalas de medida

Un carácter adicional que puede analizarse de las variables es la escala de medida. Cuando se "mide" (utilizando este concepto en un sentido amplio) un atributo de una unidad en particular, se realiza una correspondencia entre el atributo medido y la cualidad o número que se registra como medición. Dado que pueden existir distintas reglas para realizar esta correspondencia es que existen distintas **escalas de medición**.

Las distintas escalas se diferencian básicamente según tres cuestiones:

a) Las reglas de asignación de números a los atributos,

b) Las propiedades matemáticas de la escala resultante, y

c) Las estadísticas admisibles de cálculo en cada escala.

Estadística descriptiva
Escalas de medida

La **escala nominal** es la más débil de todas. Se trata de la asignación de una etiqueta a cada clase (o a cada individuo) de forma tal que no se puede establecer una jerarquía entre los valores que toma la variable, simplemente son distintas modalidades. Cabe destacar que las variables medidas en escala nominal siempre serán de tipo categóricas o cualitativas, con niveles exhaustivos y mutuamente excluyentes.

Ejemplos de variables medidas en escala nominal pueden ser género de una persona, grupo sanguíneo (0/A/B/AB), religión, o nacionalidad.

En casos de este tipo, solo puede establecerse una relación de equivalencia entre objetos: puede identificarse si dos unidades tienen el mismo valor para la variable (por ejemplo, si dos personas tienen la misma religión) pero no puede establecerse una relación de orden entre esos valores.

Estadística descriptiva
Escalas de medida

La **escala ordinal** posee las propiedades enunciadas para la escala nominal, pero además de la relación de equivalencia puede establecerse una relación de orden entre los aspectos de las unidades. En este caso también se tratará de variables de tipo categóricas o cualitativas.

Ejemplos de variables medidas en escala ordinal pueden ser intensidad del dolor ante un estímulo (baja, media, alta) o el nivel de instrucción de una persona (sin estudios, primario, secundario, superior).

En casos de este tipo, puede establecerse una relación de equivalencia entre objetos y también de orden: puede identificarse si alguno de los dos valores es mayor al otro (por ejemplo, si la intensidad del dolor sentida por una persona es menor, igual o mayor que la intensidad sentida por otra).

Estadística descriptiva
Escalas de medida

La **escala de intervalos** constituye formalmente una escala cuantitativa, por lo que las variables medidas en esta escala serán de este tipo. Posee todas las propiedades de la escala ordinal pero además de la relación de equivalencia y de orden puede conocerse la distancia exacta entre dos unidades. El punto cero sobre una escala de este tipo es un tema de convención o de conveniencia, por lo que el valor cero no implica la ausencia del atributo. En esta escala no se puede decir que un valor es (por ejemplo) el doble que otro, como consecuencia del origen arbitrario que posee.

Ejemplos de variables medidas en escala de intervalo pueden ser el año calendario (discreta), la temperatura en grados celcius o la hora del día (continuas).

Estadística descriptiva
Escalas de medida

Las **escalas de razón** solo son posibles de alcanzar cuando se pueden establecer todas las propiedades: equivalencia, orden, igualdad de intervalos e igualdad de razones. En estas escalas para variables numéricas el creo absoluto indica ausencia del atributo, aunque en ocasiones pueda no existir (como por ejemplo, una persona con altura de 0 cm o peso igual a 0 kg).

Ejemplos de variables medidas en escala de razón pueden ser el monto cobrado a fin de mes, el tiempo transcurrido hasta cierto evento (variables continuas), la cantidad de hijos/as de una persona (variable discreta).

Para entrar en calor... clasifiquemos variables

a. Antecedentes de hipertensión (Sí, No, No sabe) Categórica - Nominal
b. Diámetro de la arteria lesionada Cuantitativa - Razón
c. Cantidad de arterias lesionadas Cuantitativa - Razón
d. Género musical preferido por lxs estudiantes Categórica - Nominal
e. Fecha de cumpleaños de lxs asistentes a una clase Cuantitativa - Intervalo
f. Marca de un automóvil Categórica - Nominal
g. Cantidad de lluvia caída en verano en una determinada ciudad Cuantitativa – Razón
h. Razones para no consumir productos con jarabe de maíz de alta fructosa Categórica - Nominal
i. Capacidad de ahorro (nula, baja, media, alta, muy alta) Categórica – Ordinal

Estadística descriptiva

Información a partir de datos

Una característica de un libro es la cantidad de páginas que contiene. Inclusive dentro de los distintos géneros literarios, la cantidad de páginas puede ser muy variable. Se obtuvo una muestra de 40 libros de misterio de una libreria y se registró el número de páginas.

229	247	347	246	307	181	198	214	234	340
314	260	202	320	360	320	200	414	262	248
376	211	214	218	276	628	255	352	197	308
203	371	203	406	261	378	223	181	284	196

Estadística descriptiva
Información a partir de datos

Algunas de las preguntas que se podrían plantear acerca de los datos podrían ser:

- ¿Cuál es el número de páginas más común o representativo?
- ¿Se concentran las observaciones alrededor del valor más común o están dispersas?
- ¿Predominan los libros cortos o los libros largos, o se observaron aproximadamente la misma cantidad de ambos tipos de libros?
- ¿Existe algún libro cuya cantidad de páginas es inusual con respecto al resto de los libros?
- ¿Qué proporción de los libros observados tienen al menos 500 páginas o menos de 200 páginas?

Para poder responder estas preguntas es necesario organizar los datos de una manera adecuada mediante presentaciones tabulares y gráficas

Estadística descriptiva

Información a partir de datos

Descripción gráfica

- Tablas de distribución de frecuencias
  - Conteos/Frecuencias absolutas
  - Porcentajes/Frecuencias relativas
- Tablas de contingencia
- Gráficos de barras
- Gráficos de sectores circulares
- Histogramas
- Gráficos de caja/Boxplot
- Diagramas de tallo y hoja
- Diagramas de dispersión
- Gráficos de evolución/Series de tiempo

Descripción numérica

- Medidas de posición
  - Media aritmética/Promedio
  - Percentiles/Cuartiles
  - Modo
- Medidas de dispersión
  - Rango
  - Desvío estándar
  - Desviación cuartílica/Rango intercuartílico
  - Coeficiente de variación
- Medidas de relación
  - Coeficiente de asociación
  - Coeficiente de correlación

Análisis descriptivo de datos
Análisis univariado gráfico

Estadística descriptiva
Descripción gráfica

Un **gráfico** es una representación visual de los datos estadísticos, en el que los datos están representados por símbolos como barras o líneas.

Todos hemos oído el viejo dicho: "una imagen vale más que mil palabras". Una de las mejores técnicas para hacer comprensibles los datos es la representación de los números mediante imágenes. Esto puede hacer mucho más fácil apreciar un patrón o exponer ciertos patrones que de otro modo podrían quedar ocultos.

Se pueden mostrar los datos de muchas maneras diferentes, desde sencillos gráficos de barras a diagramas de dispersión más complejos, mapas temáticos y pirámides de población animadas. Es mucho más fácil entender las estadísticas presentadas mediante un gráfico o mapa, que en largas listas de números -asumiendo, por supuesto-, que las representaciones gráficas están correctamente realizadas.

Estadística descriptiva
Descripción gráfica

Descripción gráfica: Pautas generales para la presentación

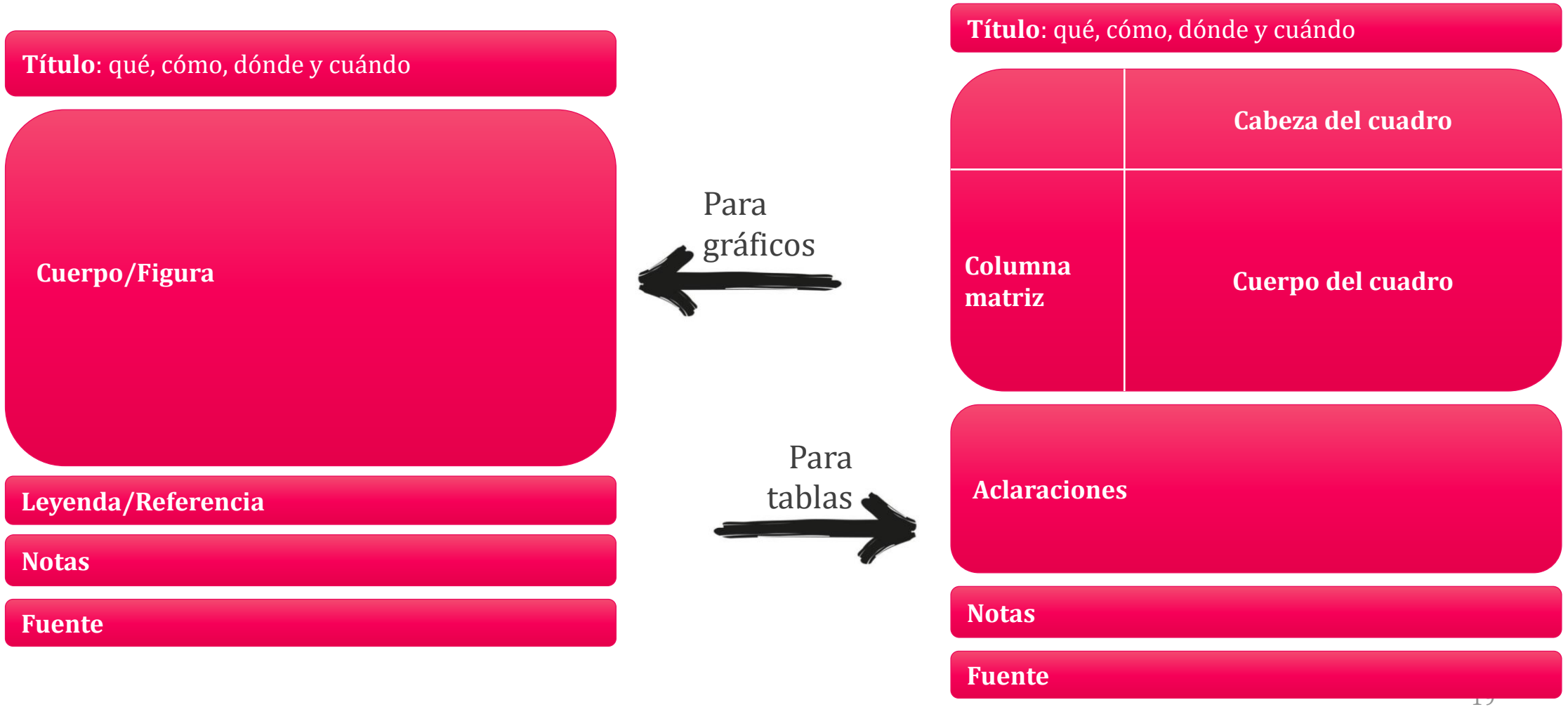
- Fácil comprensión
- Información necesaria para la interpretación
- Formato y numeración simplifican el análisis

Descripción gráfica: Pautas específicas para gráficos y figuras

- Información resumida
- Información útil y significativa
- Representación fidedigna
- Autónomos
- Soporte visual que facilita la interpretación
- Formato: menos es más

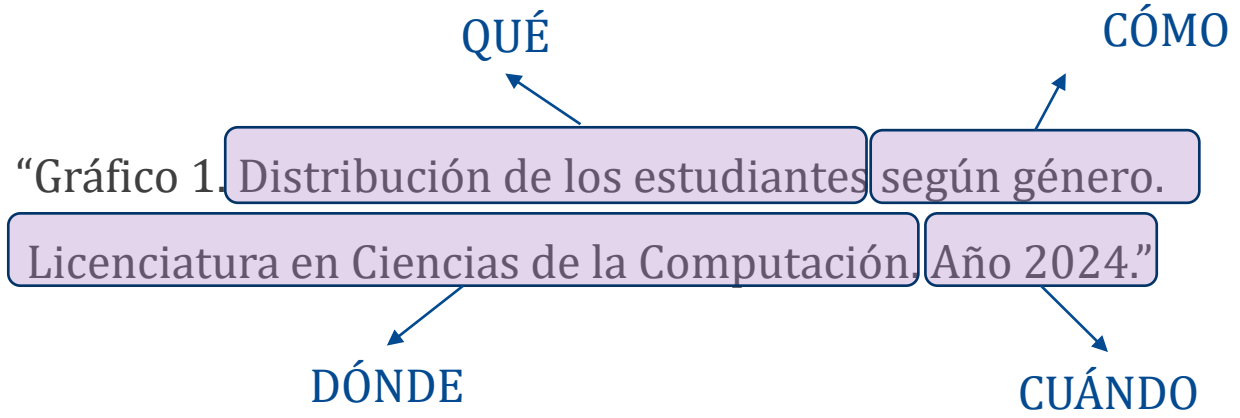
Estadística descriptiva

Descripción gráfica: elementos de las figuras



Estadística descriptiva
Descripción gráfica

- El **título** debe ser lo suficientemente informativo de manera de responder a las preguntas: ¿Qué se está representando? ¿Cómo se encuentra resumida la información? ¿Cuándo: a qué período corresponden los datos (o cuándo fueron relevados)? ¿Y a dónde (espacio, ciudad, lugar, etc.) pertenece dicha información? Por ejemplo:



Estadística descriptiva
Descripción gráfica

- En el **cuerpo** se presenta la tabla o el gráfico.
- La **nota** es una aclaración sobre alguna particularidad que pudieran presentar los datos (por ejemplo, “los datos del año 2008 son parciales”), así como cualquier información adicional necesaria para comprender y utilizar correctamente los datos (por ejemplo, definiciones del tipo “el índice de masculinidad es la razón de hombres por cada 100 mujeres: $\frac{Cantidad\ de\ hombres}{Cantidad\ de\ mujeres} \times 100$”). Es opcional.
- La **fuente** hace mención al origen de los datos que se están representando, es decir la organización que elaboró los datos y el método de recogida de datos (por ejemplo, “INDEC - Censo de Población, Hogares y Vivienda, 2010”).

Estadística descriptiva
Descripción gráfica

- La **referencia** o **leyenda** debe identifica los símbolos, patrones o colores utilizados para representar los datos en el gráfico. La leyenda no se debe mostrar cuando sólo una serie de valores está representada en el gráfico (análisis univariado). Cuando es necesario identificar distintos colores, tramas o líneas, la leyenda se vuelve una componente imprescindible, ya que es fundamental para poder interpretar de forma íntegra la figura.

  En alguna ocasiones las leyendas pueden reemplazarse por etiquetas de datos, por ejemplo en los gráficos de sectores circulares.

- Para los gráficos que utilicen un eje cartesiano para su representación, es necesario incluir los **nombres de los ejes**, con la unidad de medida en caso de ser necesario. Por ejemplo: “Velocidad (km/h)” o “Cantidad de estudiantes”.

DESCRIPCIÓN GRÁFICA

TIPOS DE FIGURAS: TABLAS

Tabla 1. Distribución de los niños encuestados según sexo

Sexo	Cantidad de niños	Porcentaje de niños
Mujer	31	35,2%
Varón	57	64,8%
Total	**88**	**100,0%**

Tabla de distribución de frecuencias

Tabla de contingencia

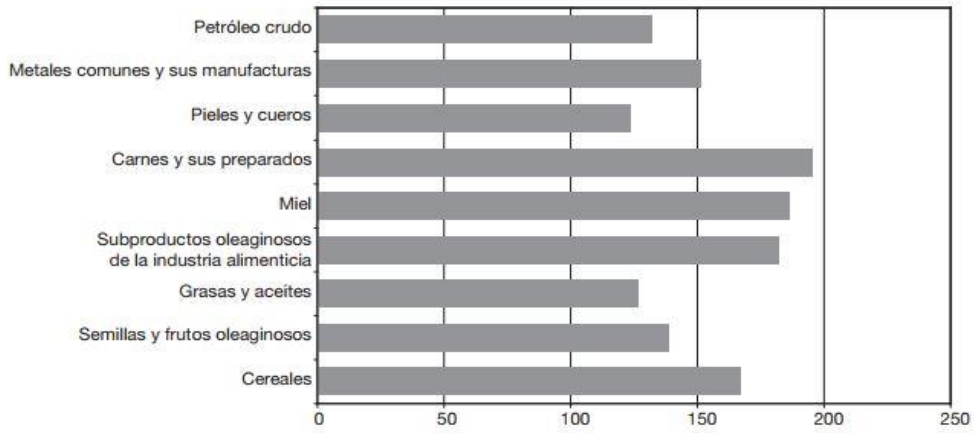
Tabla 4. Distribución de los encuestados según origen de la recomendación de suplementos y sexo

	Varón	Mujer	Total
Entrenador	65	14	79
Amigo	23	10	33
Nutricionista	19	5	24
Médico	12	3	15
Otro	6	3	9
Publicidad	6	2	8
Farmacéutico	3	0	3
Total	**134**	**37**	**171**

DESCRIPCIÓN GRÁFICA

TIPOS DE FIGURAS: GRÁFICOS

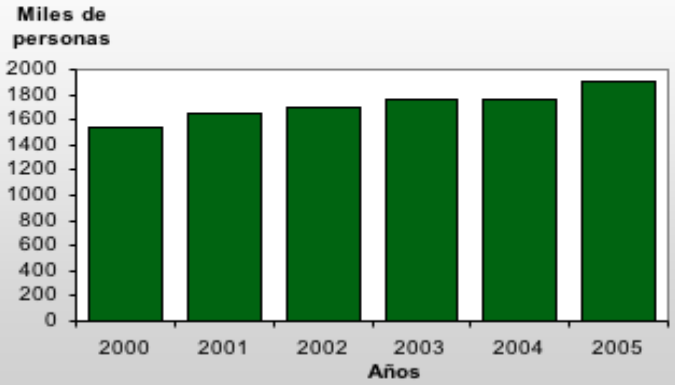
Gráfico 44. **Índice de precios de exportación por rubros seleccionados, base 2004=100. Año 2015**



Fuente: INDEC. Dirección Nacional de Estadísticas del Sector Externo.

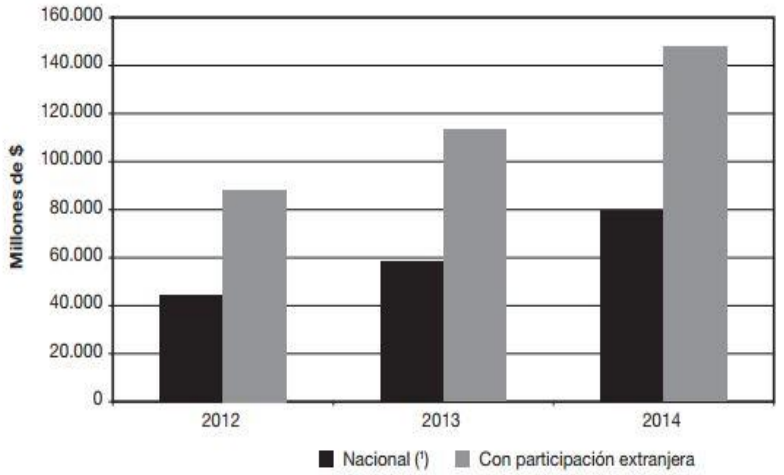
Gráficos de barras

Gráfico N°13
COSTA RICA: POBLACIÓN DE 12 AÑOS Y MÁS ECONOMICAMENTE ACTIVA
Julio 2000 - 2005



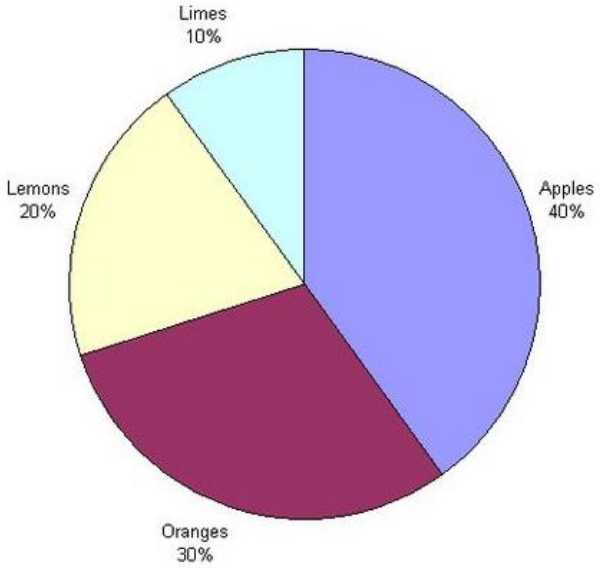
FUENTE: elaborado a partir del Cuadro, N°8 con los datos del Instituto Nacional de Estadística y Censos.

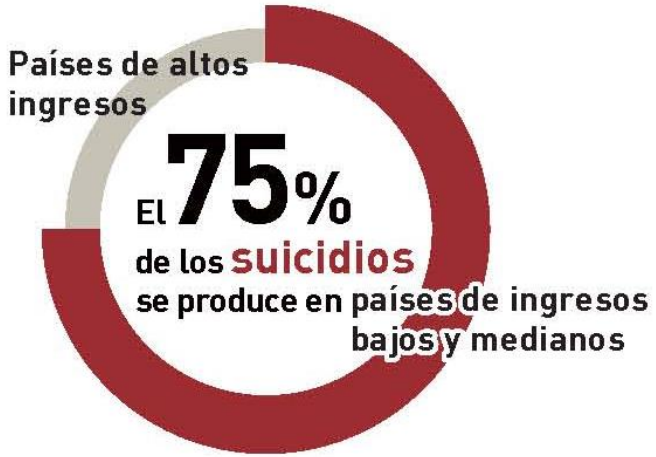
Gráfico 37. **Grandes empresas. Salarios devengados por origen del capital. Años 2012-2014**

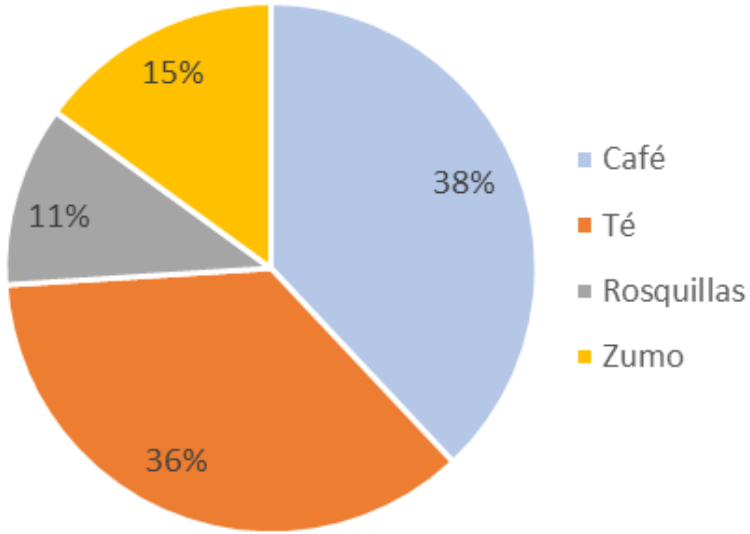


DESCRIPCIÓN GRÁFICA

TIPOS DE FIGURAS: GRÁFICOS





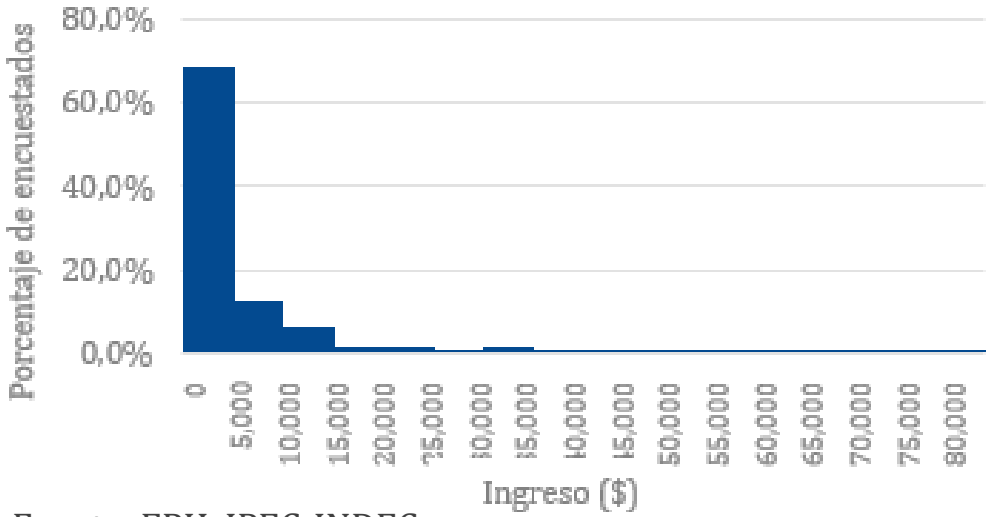


Gráficos de sectores circulares

DESCRIPCIÓN GRÁFICA

TIPOS DE FIGURAS: GRÁFICOS

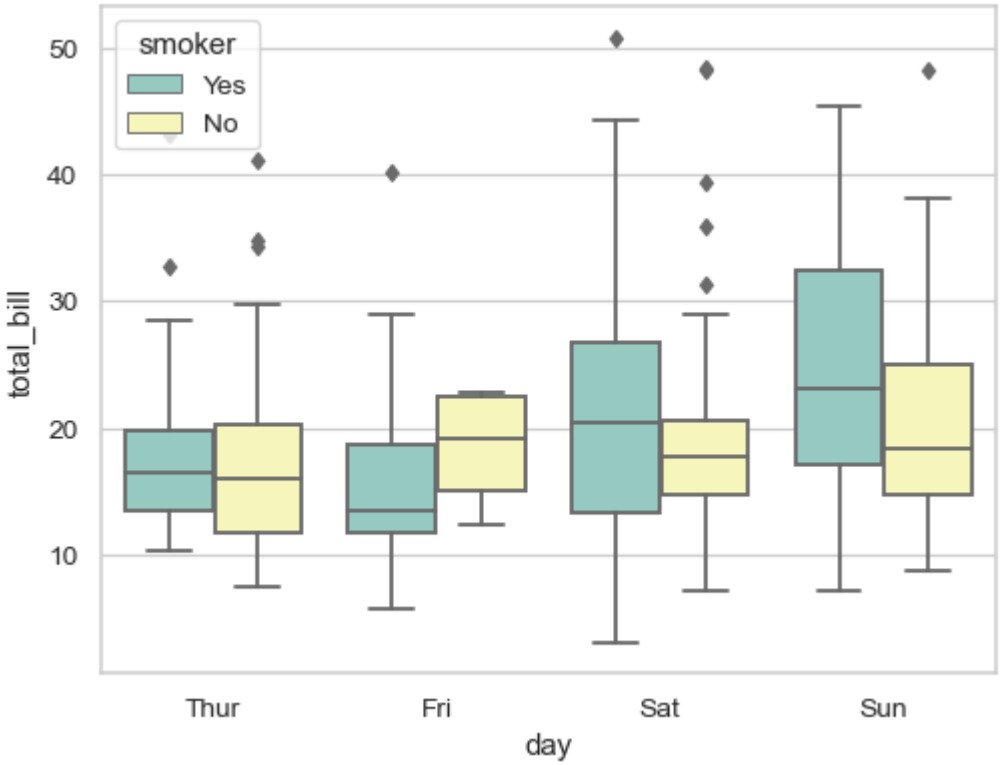
Figura 3. Ingreso de la población santafesina. Año 2003.



Fuente: EPH, IPEC-INDEC.

Histograma

Diagrama de caja / Boxplot



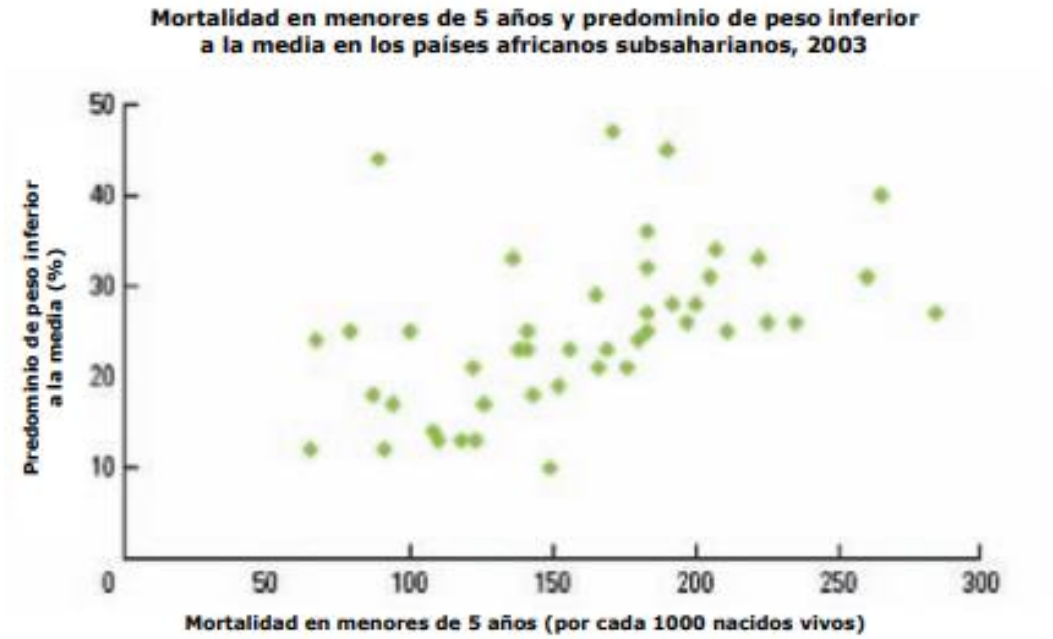
DESCRIPCIÓN GRÁFICA

TIPOS DE FIGURAS: GRÁFICOS

Diagrama de tallo y hoja

Tallos	Hojas
0	5 8
1	2 3 5 7
2	0 0 0 5 8 8 9
3	0 0 1 3 3 3 6 6 7 7 7 7 7 8 8 8 8 9 9
4	1 3 5 5 5 6 7 7 8 8 8 8 9 9
5	0 0 0 1 1 1 1 2 6 8
6	0 0 1 1 2 4 4 4 4 4 8 8 9
7	0 5 5 5 5 7
8	3 4 4 5 6 6 6 7 8 9
9	0 1 2 2 2 2 5 5 6 8 9 9
10	2 2 2 5 7

Datos brutos en fila:
102, 102, 102, 105, 107



Fuente: Jamison et al. (2006) ***Disease and Mortality in Sub-Saharan Africa, 2nd edition***, Washington D.C., The World Bank[7].

Diagrama de dispersión

DESCRIPCIÓN GRÁFICA

TIPOS DE FIGURAS: GRÁFICOS

Gráfico de líneas / Series temporales

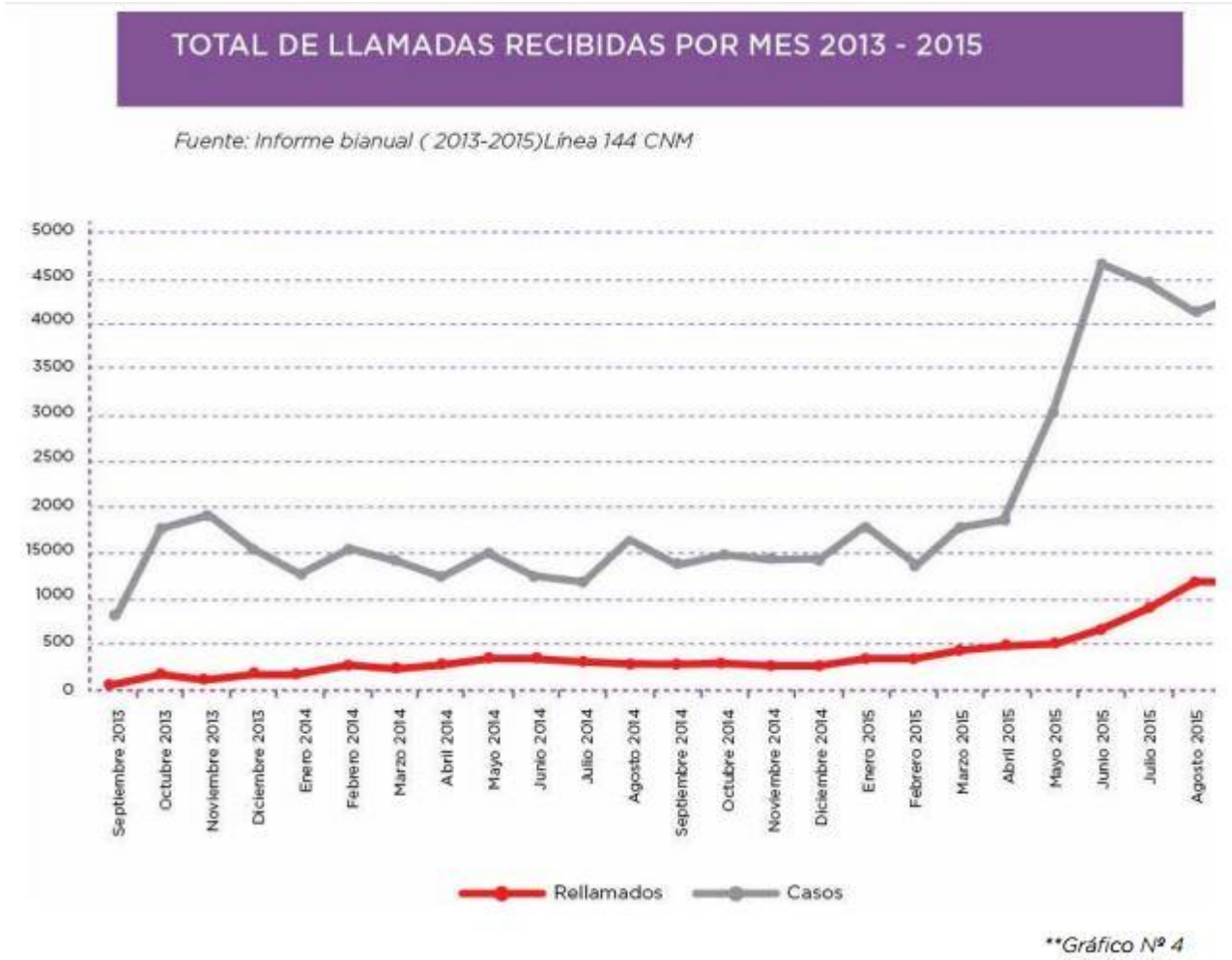


**Gráfico Nº 4

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Una tabla de distribución de frecuencias es una tabla que muestra los distintos niveles de la variable con sus respectivas frecuencias. Definimos dos tipos de **frecuencias simples**: absolutas (conteos) y relativas (proporciones).

Para una variable categórica con k categorías $c_1, c_2, \dots c_k$, y siendo n_i la cantidad de observaciones en cada una de las $i = 1, \dots k$ categorías:

Categoría	Frecuencia absoluta	Frecuencia relativa
c_1	$f_1 = n_1$	$h_1 = \frac{n_1}{n}$
c_2	$f_2 = n_2$	$h_2 = \frac{n_2}{n}$
$\vdots$	$\vdots$	$\vdots$
c_k	$f_k = n_k$	$h_k = \frac{n_k}{n}$
Total	$n = \sum_{i=1}^{k} n_i$	1

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Frecuencias absolutas f_i

Es la cantidad de veces que se repite el valor i de la variable. La suma de todas las frecuencias absolutas es igual a la cantidad de individuos bajo estudio (n si se trata de una muestra, N si se trata de una población.

Frecuencias relativas h_i

Es la cantidad de veces que se repite el valor i de la variable relativo al total. Se calcula como la frecuencia absoluta dividida por la cantidad total de individuos bajo estudio (n o N, según se trate de una muestra o una población respectivamente). Dado que se trata de proporciones, la suma de todas las frecuencias relativas es igual a la unidad (1).

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Supongamos que recolectamos datos sobre un grupo de mujeres e interesa obtener información sobre su estado civil.

Los datos obtenidos son los siguientes:

D	Ca	Ca	Ca	D	Ca	Ca	Ca
Ca	Co	Ca	Ca	Ca	Ca	Ca	Ca
Ca	Ca	Ca	S	D	Ca	Ca	Ca
D	S	Ca	Co	S	V	Ca	Ca
Ca	Ca	Ca	Ca	D	S	Ca	Ca
Ca	Ca	S	Ca	V	Co	Ca	Ca
Ca	S	Ca	Ca	Ca	Ca	Ca	D
S	Ca	Ca	Ca	S	Ca	Ca	Ca
Ca	Co	Ca	Ca	Ca	V	Ca	Ca
D	S	V	V	Ca	Ca	Co	D
Ca	Ca	Ca	D	Ca	D	Ca	Ca
		D	Ca	Ca			

¿Cómo podríamos armar una tabla de distribución de frecuencias?

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Tabla 2. Distribución de las encuestadas de acuerdo al estado civil

Estado civil	Cantidad de mujeres	Porcentaje de mujeres
Casada	61	67,0%
Divorciada	11	12,1%
Soltera	9	9,9%
Concubina	5	5,5%
Viuda	5	5,5%
Total	**91**	**100,0%**

Variable

Categorías de la variable

Frecuencias absolutas

Frecuencias relativas porcentuales

¿Cómo interpretamos una fila?

De las 91 mujeres bajo estudio, 11 son divorciadas, lo que representa al 12,1% del total.

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

La tabla de distribución de frecuencias tendrá tantas filas como niveles de la variable se necesiten representar. Si se trata de variables categóricas, entonces habrá tantas filas como categorías en la variable. Si se trata de variables cuantitativa discretas, entonces habrá tantas filas como valores enteros distintos tome la variable en ese conjunto de datos. Finalmente, si se trata de variables cuantitativas continuas, debe tenerse en cuenta que no tiene sentido expresar cada uno de los distintos valores observados en una fila (dado que pueden tomar infinitos valores posibles estas variables); basta con pensar en la altura de un grupo de personas medida en milímetros: probablemente pocos individuos midan exactamente lo mismo. En estos casos, el rango de valores se parte en intervalos, y la tabla de distribución de frecuencias tendrá tantas filas como intervalos se formen.

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Para las variables medidas en una escala superior a la nominal (es decir, para todas las variables en las que se pueda establecer al menos una relación de orden en los niveles, incluyendo tanto a las variables categóricas ordinales como a todas las variables cuantitativas), pueden además calcularse las **frecuencias acumuladas**.

Las frecuencias acumuladas pueden ser tanto absolutas como relativas, y se las llama así porque la frecuencia de cada categoría acumula también las frecuencias de todas las categorías previas. Es por esto que es necesario que la escala de la variable sea al menos ordinal: es posible pensar en la proporción de autos modelo 2018 o anteriores, o en la cantidad de personas que tienen a lo sumo nivel socioeconómico medio, pero no tiene sentido por ejemplo el porcentaje de mujeres con el pelo al menos colorado.

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Para una variable cuantitativa discreta que toma k valores distintos $y_1, y_2, \dots y_k$, y siendo n_i la cantidad de observaciones que toma cada uno de los $i = 1, \dots k$ valores:

Valor	Frecuencia absoluta ($\boldsymbol{f_i}$)	Frecuencia relativa ($\boldsymbol{h_i}$)	Frecuencia absoluta acumulada ($\boldsymbol{F_i}$)	Frecuencia relativa acumulada ($\boldsymbol{H_i}$)
y_1	$f_1 = n_1$	$h_1 = \frac{n_1}{n}$	$F_1 = f_1$	$H_1 = h_1 = \frac{f_1}{n}$
y_2	$f_2 = n_2$	$h_2 = \frac{n_2}{n}$	$F_2 = f_1 + f_2$	$H_2 = h_1 + h_2 = \frac{f_1 + f_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{k-1}	$f_{k-1} = n_{k-1}$	$h_{k-1} = \frac{n_{k-1}}{n}$	$F_{k-1} = f_1 + \cdots + f_{k-1}$	$H_{k-1} = h_1 + \cdots + h_{k-1}$
y_k	$f_k = n_k$	$h_k = \frac{n_k}{n}$	$F_k = f_1 + \cdots + f_k = n$	$H_k = h_1 + \cdots + h_k = 1$
Total	$n = \sum_{i=1}^{k} n_i$	1	-	-

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Para una variable cuantitativa continua que toma muchos valores distintos por lo que se la agrupa en k clases o intervalos $v_1, v_2, \dots v_k$, donde $v_i = [l_{inf}; l_{sup})$ y siendo n_i la cantidad de observaciones que caen en cada intervalo:

Intervalo	Frecuencia absoluta $(\boldsymbol{f_i})$	Frecuencia relativa $(\boldsymbol{h_i})$	Frecuencia absoluta acumulada $(\boldsymbol{F_i})$	Frecuencia relativa acumulada $(\boldsymbol{H_i})$
$v_1 = [l_{\text{inf } 1}; l_{\text{sup } 1})$	$f_1 = n_1$	$h_1 = \frac{n_1}{n}$	$F_1 = f_1$	$H_1 = h_1 = \frac{f_1}{n}$
$v_2 = [l_{\text{inf } 2}; l_{\text{sup } 2})$	$f_2 = n_2$	$h_2 = \frac{n_2}{n}$	$F_2 = f_1 + f_2$	$H_2 = h_1 + h_2 = \frac{f_1 + f_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$v_{k-1} = [l_{\text{inf } k-1}; l_{\text{sup } k-1})$	$f_{k-1} = n_{k-1}$	$h_{k-1} = \frac{n_{k-1}}{n}$	$F_{k-1} = f_1 + \cdots + f_{k-1}$	$H_{k-1} = h_1 + \cdots + h_{k-1}$
$v_k = [l_{\text{inf } k}; l_{\text{sup } k})$	$f_k = n_k$	$h_k = \frac{n_k}{n}$	$F_k = f_1 + \cdots + f_k = n$	$H_k = h_1 + \cdots + h_k = 1$
Total	$n = \sum_{i=1}^{k} n_i$	1	-	-

DESCRIPCIÓN GRÁFICA

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Ejemplo *En la Encuesta de Presupuestos Familiares nos interesa estudiar la distribución del número de individuos en los hogares.*

Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	6	0.0800	6	0.0800
2	11	0.1467	17	0.2267
3	11	0.1467	28	0.3733
4	20	0.2667	48	0.6400
5	15	0.2000	63	0.8400
6	8	0.1067	71	0.9467
7	3	0.0400	74	0.9867
9	1	0.0133	75	1.0000
Total			75	1.00000

¿Cómo interpretamos una fila?

De los 75 hogares bajo estudio, 11 se encontraban compuestos por 3 individuos, lo que representa una proporción igual a 0,1467 sobre el total (o bien, el 14,67%). Asimismo, 28 hogares se encontraban compuestos por 3 personas o menos, lo que representa el 37,33% del total de hogares analizados.

DESCRIPCIÓN GRÁFICA

GRÁFICOS PARA VARIABLES CATEGÓRICAS

Un **gráfico de sectores circulares** muestra la distribución de una variable cualitativa dividiendo un círculo (que representa todo el conjunto de datos) en porciones o sectores de acuerdo al porcentaje de individuos que se contabilizan en cada categoría. El área de la porción es proporcional a la frecuencia observada. Los gráficos de sectores son útiles para presentar resultados de variables categóricas con pocas categorías.

Un **gráfico de barras** muestra la distribución de una variable cualitativa enumerando las categorías de la variable en un eje y dibujando una barra sobre cada categoría con un área igual a la frecuencia de individuos contabilizados en esa categoría. Las barras deben ser de igual ancho.

DESCRIPCIÓN GRÁFICA
VARIABLE CUALITATIVA DICOTÓMICA

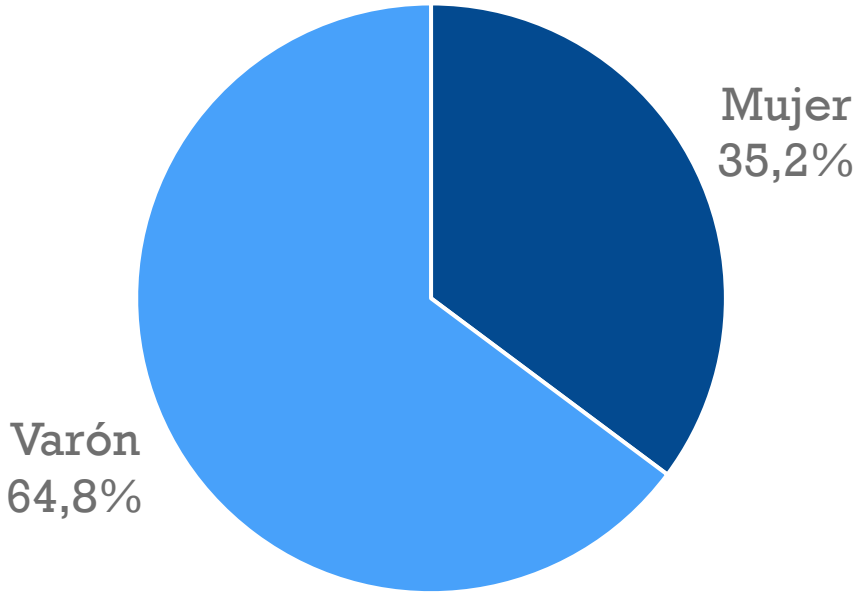


Gráfico 1. Distribución de los niños encuestados según sexo

Tabla 1. Distribución de los niños encuestados según sexo

Sexo	Cantidad de niños	Porcentaje de niños
Mujer	31	35,2%
Varón	57	64,8%
Total	**88**	**100,0%**

Opciones alternativas para interpretar

- La muestra está compuesta en su mayoría por varones (64,8%).
- El 64,8% de los niños estudiados son varones.
- Del total de niños estudiados 57 son varones, representando el 64,8% de la muestra.
- Sólo se observaron 31 niñas en la muestra (35,2%).

DESCRIPCIÓN GRÁFICA

VARIABLE CUALITATIVA POLITÓMICA

Opciones alternativas para interpretar

- La mayoría de las mujeres bajo estudio están casadas (67,0%). Con menor frecuencia se encuentran mujeres divorciadas (12,1%) o solteras (9,9%), e incluso una porción menor aún de las mujeres se encuentra en concubinato (5,5%) o viuda (5,5%).
- Al analizar la distribución de las mujeres bajo estudio de acuerdo a su estado civil, puede afirmarse que la muestra está compuesta en mayor medida por mujeres casadas (67,0%). Le siguen en frecuencia las mujeres divorciadas (12,1%) y mujeres solteras (9,9%), hallándose en menor proporción en la muestra mujeres en concubinato (5,5%) y mujeres viudas (5,5%).

Tabla 2. Distribución de las encuestadas de acuerdo al estado civil

Estado civil	Cantidad de mujeres	Porcentaje de mujeres
Casada	61	67,0%
Divorciada	11	12,1%
Soltera	9	9,9%
Concubina	5	5,5%
Viuda	5	5,5%
Total	**91**	**100,0%**

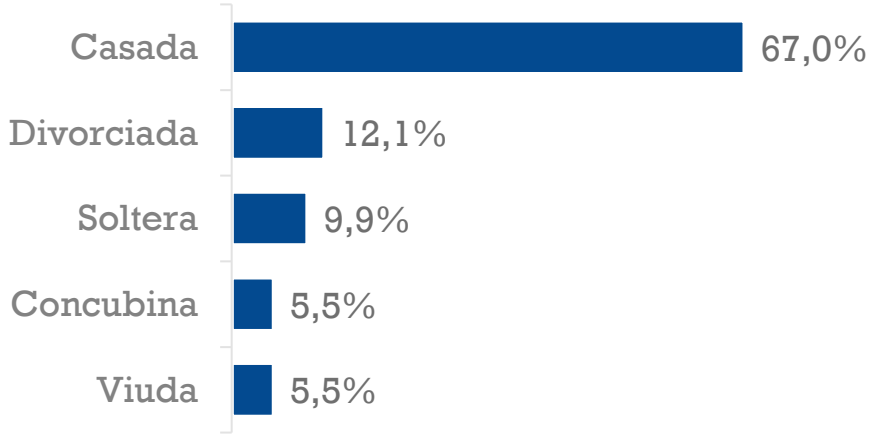


Gráfico 2. Distribución de las encuestadas de acuerdo al estado civil

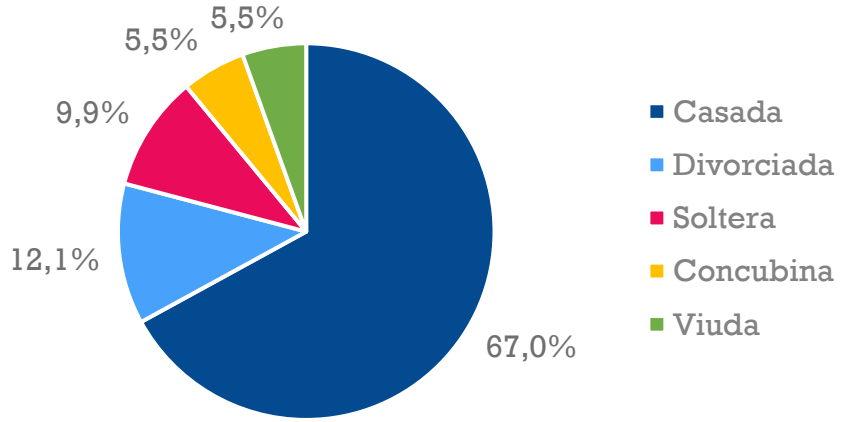
DESCRIPCIÓN GRÁFICA

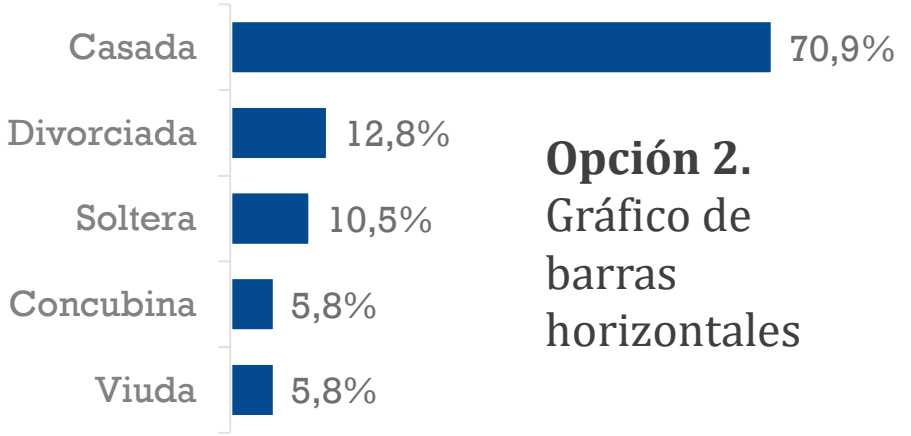
VARIABLE CUALITATIVA POLITÓMICA

Para tener en cuenta

- Si la variable tiene muchas categorías es preferible el gráfico de barras al de sectores circulares.
- Las barras horizontales suelen ser más fáciles de leer que las verticales, sobre todo si la variable tiene muchas categorías.
- Orden de las categorías/barras:
  - Si la variable es nominal: según frecuencia
  - Si la variable es ordinal: según orden natural

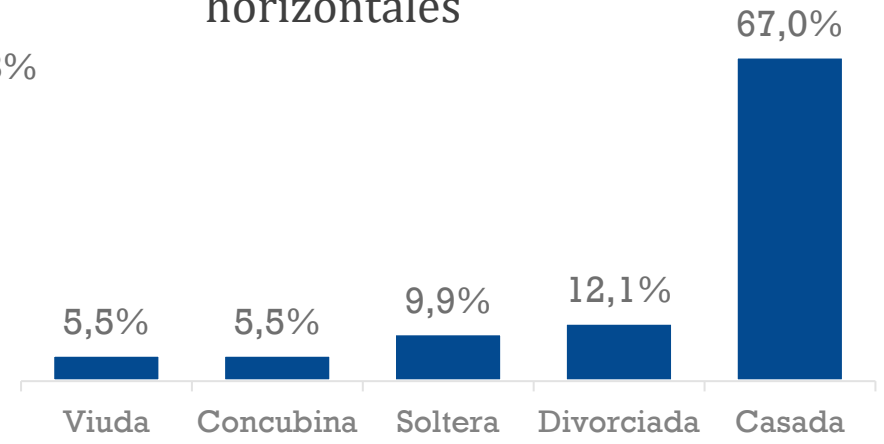
Opción 1. Gráfico de sectores circulares





Opción 2. Gráfico de barras horizontales

Opción 3. Gráfico de barras verticales



DESCRIPCIÓN GRÁFICA

VARIABLE CUALITATIVA ORDINAL

Opciones alternativas para interpretar

- En la empresa con Entorno Laboral Saludable, casi el 80% de los trabajadores se caracteriza por tener un estado nutricional normal (40,0%) o a lo sumo con sobrepeso (38,2%). Aun así, el 21,8% de los trabajadores presenta algún grado de obesidad.
- Aun cuando los trabajadores de la empresa con Entorno Laboral Saludable son en su mayoría personas con estado nutricional normal o con sobrepeso hay diez empleados con algún nivel de obesidad, que representan el 21,8% del total. Incluso dos de ellos presentan Obesidad I y otros dos Obesidad II.

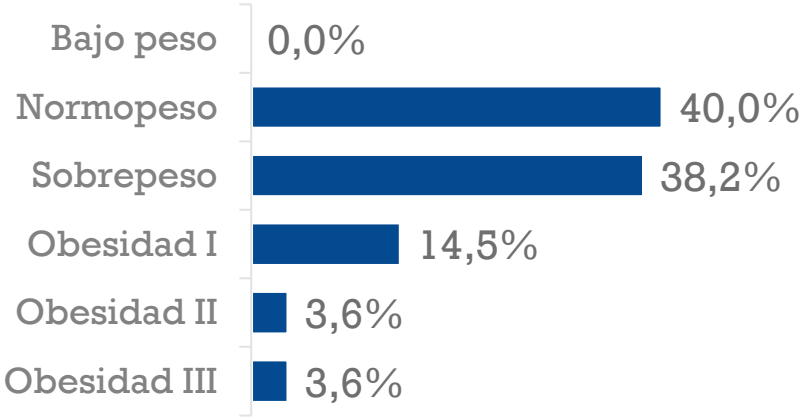


Gráfico 3. Distribución de los empleados con Entorno Laboral Saludable según su estado nutricional

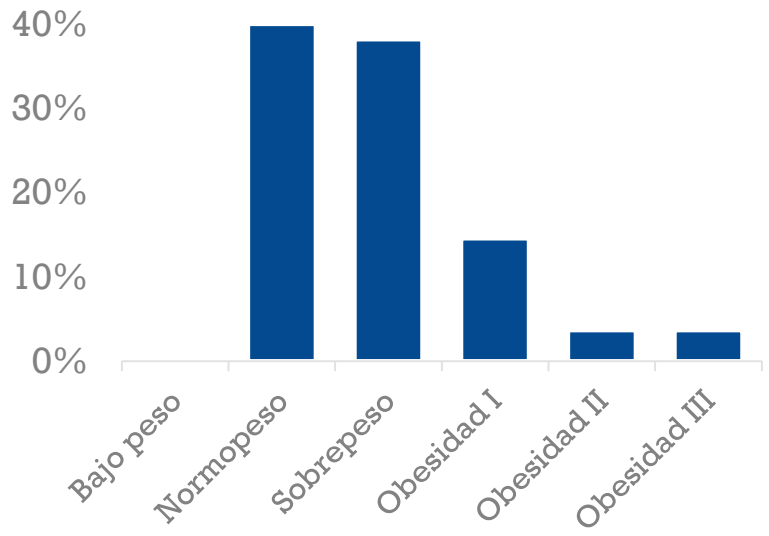


Gráfico 3. Distribución de los empleados con Entorno Laboral Saludable según su estado nutricional

DESCRIPCIÓN GRÁFICA

VARIABLE CUALITATIVA DE RESPUESTA MÚLTIPLE

Descripción del ejemplo

En la Encuesta Nacional de Victimización del año 2017 se relevó la siguiente pregunta para aquellos individuos que habían sufrido una situación delictiva, la habían denunciado pero no habían quedado satisfechos con la denuncia:

¿Por qué motivo no quedó satisfecho con la denuncia? (marcar tantas opciones como consideres necesarias)

Para tener en cuenta

Dado que el total de respuestas suele ser superior al total de individuos en la muestra el total no suma 100% y, por lo tanto:

- Es incorrecto usar un gráfico de sectores circulares
- No se suele reportar la fila de "Total" en la tabla

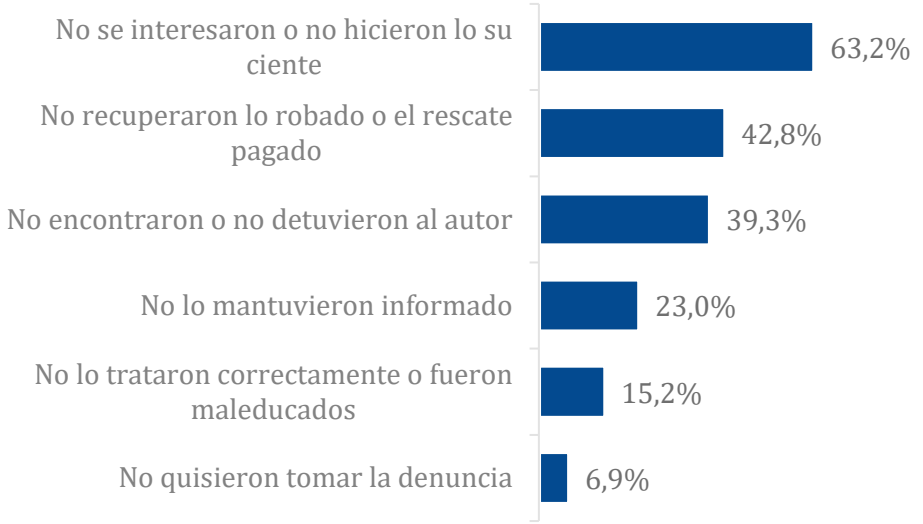


Gráfico 4. Distribución de los encuestados de acuerdo al motivo de insatisfacción al denunciar.

Tabla 3. Distribución de los encuestados de acuerdo al motivo de insatisfacción al denunciar.

Respuesta	Porcentaje de personas
No se interesaron o no hicieron suficiente	63,2%
No recuperaron lo robado o el rescate pagado	42,8%
No encontraron o no detuvieron al autor	39,3%
No lo mantuvieron informado	23,0%
No lo trataron correctamente o fueron maleducados	15,2%
No quisieron tomar la denuncia	6,9%

Para entrar en calor... ¿cómo lo analizaríamos?

En un estudio para evaluar ciertas características y patrones de los accidentes viales se recogieron datos de los siniestros ocurridos durante algunos días y se observaron las siguientes frecuencias:

Velocidad	Frecuencia absoluta	Frecuencia relativa	Porcentaje
Baja	36	0,50	50%
Media	17	0,24	24%
Alta	19	0,26	26%
Total	**72**	**1,00**	**100%**

- Indicar población, muestra y unidad de análisis.
- ¿Qué opinás respecto al orden que tienen las filas de la tabla?
- ¿Qué tipo de gráfico te parece más adecuado para representar esta información?
- ¿Te animás a construir ese gráfico?
- ¿Qué título le pondrías? ¿Te parece que está completa la información proporcionada?

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CON POCOS VALORES

Un gráfico de puntos es una manera simple de presentar datos numéricos cuando el conjunto de datos es razonablemente pequeño. Cada observación se representa por un punto sobre la ubicación correspondiente a su valor en una escala horizontal. Cuando un valor se presenta en más de una ocasión, los puntos se apilan verticalmente.

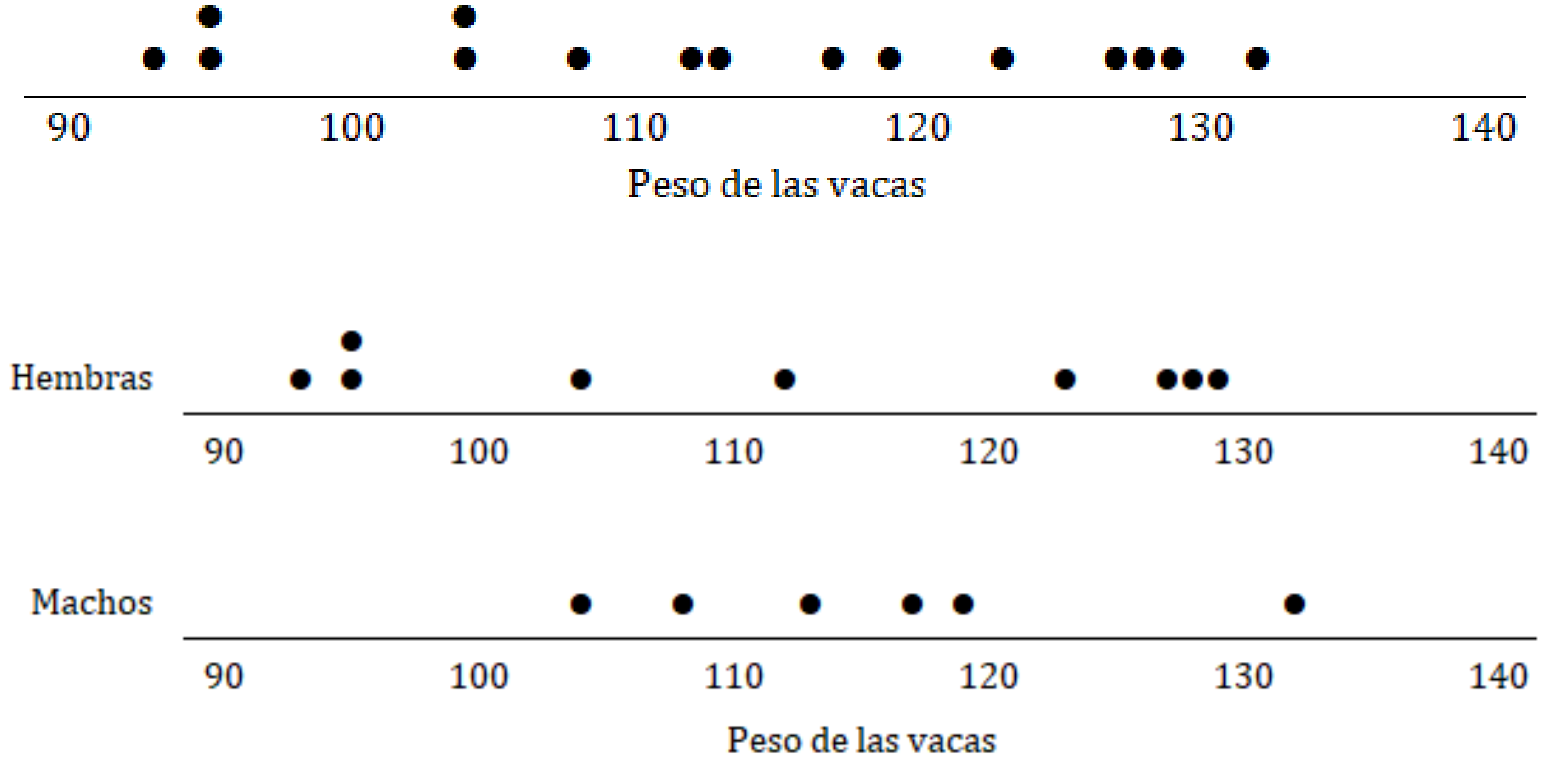
Peso y sexo de 15 potrillos recién nacidos

Potrillo	Sexo	Peso	Potrillo	Sexo	Peso	Potrillo	Sexo	Peso
1	H	129	6	M	113	11	M	108
2	M	119	7	H	95	12	H	95
3	M	132	8	H	104	13	M	117
4	H	123	9	M	104	14	H	128
5	H	112	10	H	93	15	H	127

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CON POCOS VALORES

Los gráficos de puntos permiten ver la distribución del conjunto de datos de un vistazo rápido, a la vez que permite comparar distribuciones de distintos grupos.



DESCRIPCIÓN GRÁFICA

GRÁFICOS PARA VARIABLES CUANTITATIVAS

Un **gráfico de bastones** muestra la distribución de una variable cuantitativa discreta. En un sistema de coordenadas cartesianas se representan en el eje de las abscisas u horizontal los distintos valores que asume la variable discreta en estudio y en el eje de las ordenadas o vertical se construye una escala adecuada para representar la frecuencia correspondiente a cada uno de esos valores. Sobre cada valor de la variable, se levanta una línea o bastón igual a la frecuencia de la categoría en cuestión.

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA DISCRETA

Tabla 4. Distribución de frecuencias de los antecedentes en la justicia provincial de los imputados por robo durante agosto de 2019.

Causas	Cantidad de imputados	Porcentaje de imputados	Cantidad acumulada de imputados	Proporción acumulada de imputados
0	40	29,0%	40	29,0%
1	50	36,2%	90	65,2%
2	20	14,5%	110	79,7%
3	12	8,7%	122	88,4%
4	8	5,8%	130	94,2%
5	6	4,3%	136	98,6%
6	2	1,4%	138	100,0%
Totales	**138**	**100,0%**	**-**	**-**

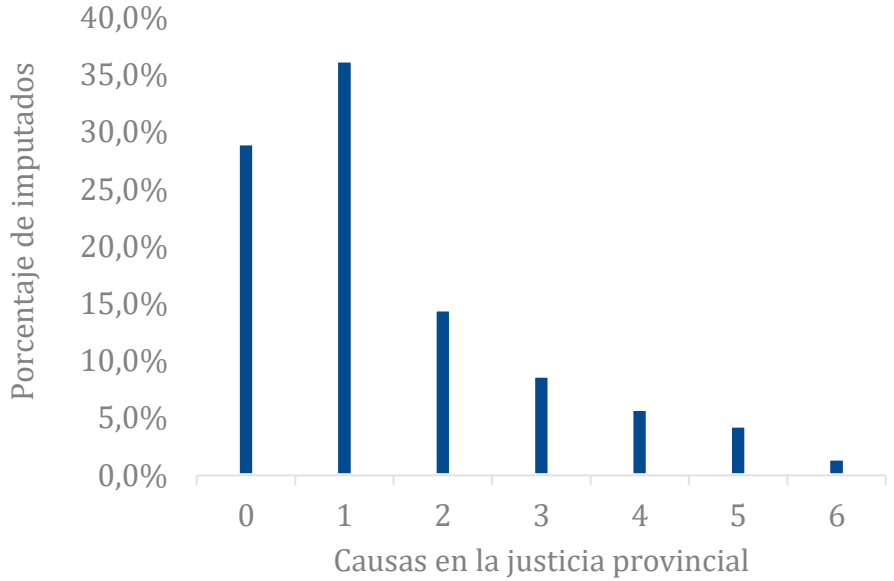


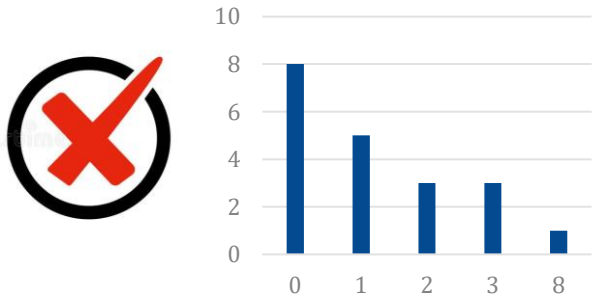
Gráfico 5. Distribución de los antecedentes en la justicia provincial de los imputados por robo durante el mes de agosto de 2019.

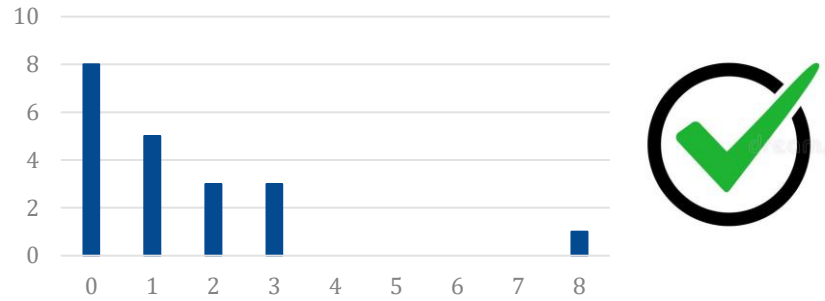
Los imputados por robo tienen entre 0 y 6 antecedentes en la justicia provincial. El 65% de los individuos tiene a lo sumo 1 antecedente.

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA DISCRETA

Los gráficos de bastones tienen los valores de la variable en el eje horizontal y la frecuencia con que las unidades toman ese valor en el eje vertical. Cuando hablamos de frecuencia nos referimos a las frecuencias absolutas, relativa o relativas porcentuales simples (nunca a las frecuencias acumuladas). Los bastones son líneas finas (o barras muy angostas) verticales, de una altura proporcional a la frecuencia de cada valor. En los gráficos de este tipo es muy importante mantener completo el rango de valores de la variable. Por ejemplo, si al analizar la cantidad de materias a rendir en diciembre por los estudiantes de 5to año de una escuela se observaran los siguientes valores: 0 0 0 0 0 0 0 0 1 1 1 1 1 2 2 2 3 3 3 8





DESCRIPCIÓN GRÁFICA

GRÁFICOS PARA VARIABLES CUANTITATIVAS

En algunos casos un conjunto de datos numéricos discretos contiene un gran número de valores posibles y también pueden existir algunos valores muy grandes o muy pequeños que están muy alejados del resto de los datos. En este caso, en lugar de construir una distribución de frecuencias con una larga lista de valores posibles, es común agrupar los valores observados en intervalos o rangos.

Ejemplo: Uso de alcohol en una muestra de 176 estudiantes universitarios.

Tragos por Semana	Frecuencia
0 a 1	52
2 a 5	38
6 a 9	17
10 a 15	35
16 o más	34

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA

La edad de los individuos encuestados varía entre18 y 60 años aproximadamente. En el histograma que muestra el Gráfico 5 puede apreciarse que la muestra está compuesta en mayor medida por sujetos más jóvenes (entre 18 y 36 años) y en menor medida por personas de más de 36 años.

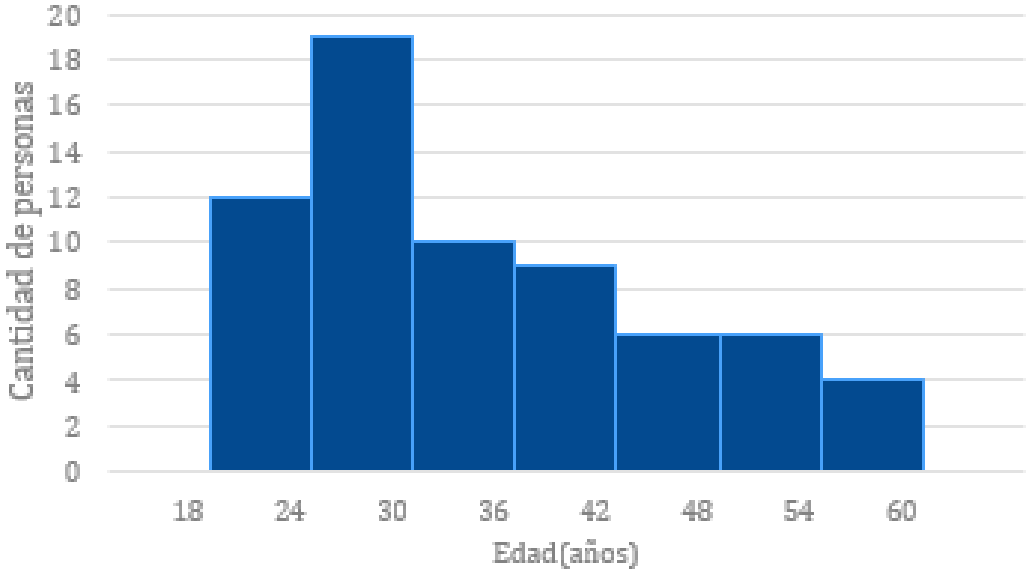


Gráfico 5. Distribución de los encuestados según su edad

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA

Los histogramas tienen los valores de la variable en el eje horizontal, en general con marcas en el eje correspondientes a los límites de cada intervalo. Al igual que en el diagrama de bastones, en el eje vertical se grafican las frecuencias absolutas, relativa o relativas porcentuales simples (nunca a las frecuencias acumuladas). Aquí también y debe respetarse la escala completa del eje (aunque existan intervalos con frecuencia nula) y los intervalos deben ser de igual amplitud. Las barras del histograma siempre van pegadas unas a las otras, marcando la continuidad de la variable bajo estudio.

¿Cuántas clases/intervalos usar?

Si n no es demasiado grande, se suele tomar $\sqrt{n}$. El objetivo siempre es mostrar apropiadamente el patrón de los datos, logrando una distribución ni demasiado dentada ni demasiado en bloques

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA

Para k $= \sqrt{n}$ intervalos, la amplitud de cada intervalo será igual a:

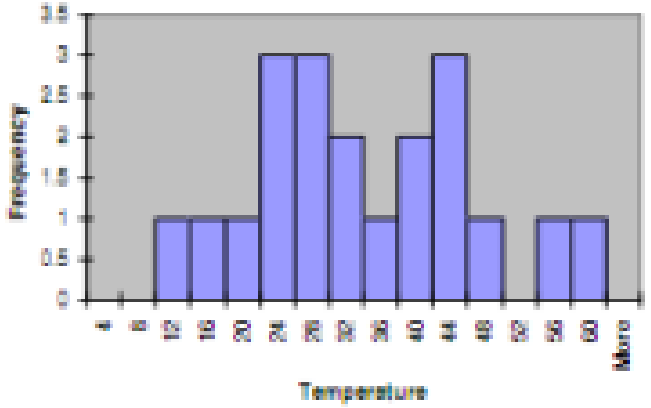
$$amplitud = h = \frac{valor\ máximo - valor\ mínimo}{cantidad\ de\ intervalos}$$

En general se redondea la amplitud óptima, y por lo tanto también los límites de los intervalos para facilitar la lectura de los datos. Por ejemplo, si $h = 4{,}87$ bien podría usarse una amplitud de 5 para los intervalos.

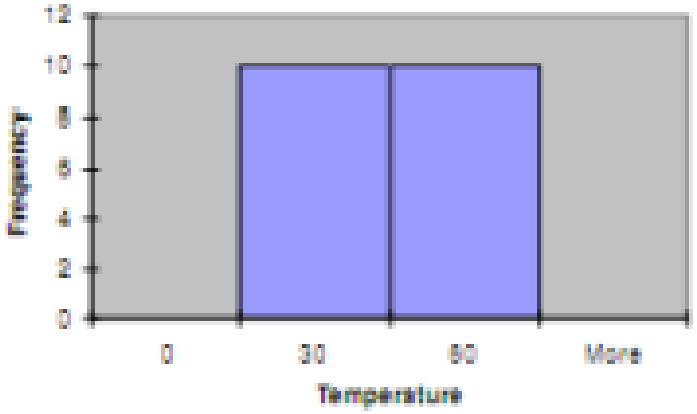
Como regla general, se usan al menos 5 intervalos, pero no más de 15 o 20. Los intervalos deben construirse de forma tal que sean exhaustivos respecto a los valores observados y mutuamente excluyentes. Por ejemplo, los intervalos podrían ser: [0, 5), [5, 10), [10, 15), etc, todos de igual amplitud.

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA



Utilizar muchos intervalos de pequeña amplitud puede dar lugar a una distribución dentada con huecos de clases vacías.



Utilizar pocos intervalos de gran amplitud puede comprimir en exceso el patrón de los datos y dar lugar a una distribución en bloque.

En ambos casos se dificulta la visualización del patrón de variabilidad de los datos.

DESCRIPCIÓN GRÁFICA

DISTRIBUCIÓN DE UNA VARIABLE

Cuando hablamos de la distribución de una variable, término que venimos mencionando prácticamente desde el comienzo de la unidad, hacemos referencia a los valores/niveles posibles que puede tomar una variable y a la frecuencia con que los individuos bajo análisis toman dichos valores/niveles.

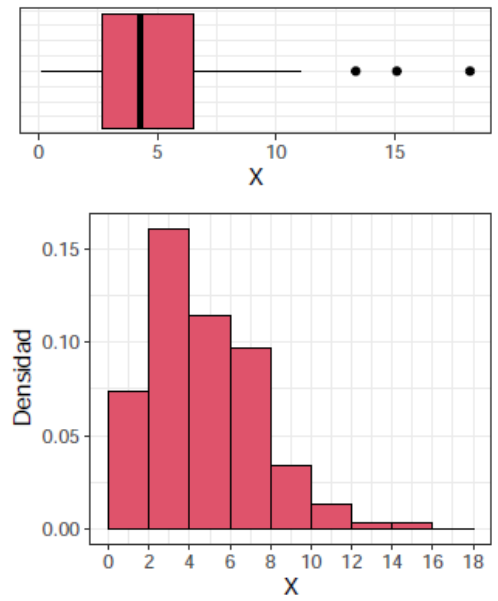
Bajo este concepto, tanto las tablas como los gráficos son herramientas para analizar la distribución de una variable, eligiendo siempre el tipo de figura más apropiada según el tipo de variable y su escala de medición.

Al analizar la distribución de una variable es posible identificar el aspecto general, el rango de valores posibles/observados, el centro del conjunto de datos, la variabilidad de los datos, así como también la existencia de observaciones atípicas (observaciones que se distinguen del patrón general de los datos).

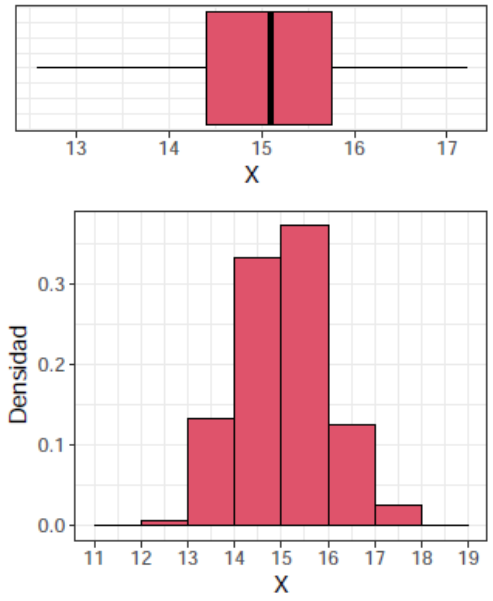
DESCRIPCIÓN GRÁFICA

DISTRIBUCIÓN DE UNA VARIABLE

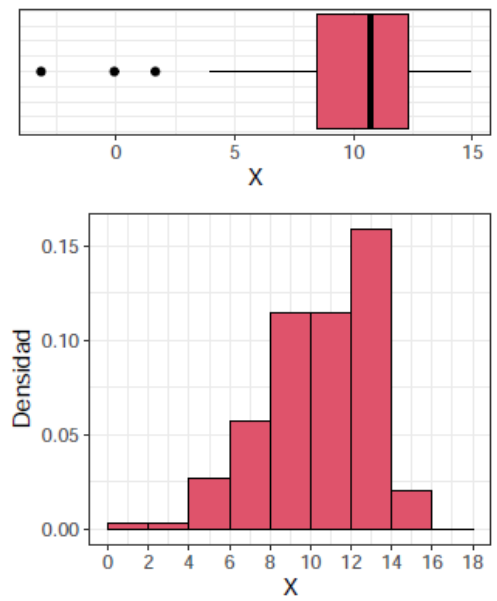
La simetría de una distribución de frecuencias hace referencia al grado en que valores de la variable, equidistantes a un valor que se considere centro de la distribución, poseen frecuencias más o menos iguales. Cuanto más similares sean, más simétrica será la distribución; cuanto más distintas, más asimétrica.



Distribución asimétrica por derecha



Distribución simétrica

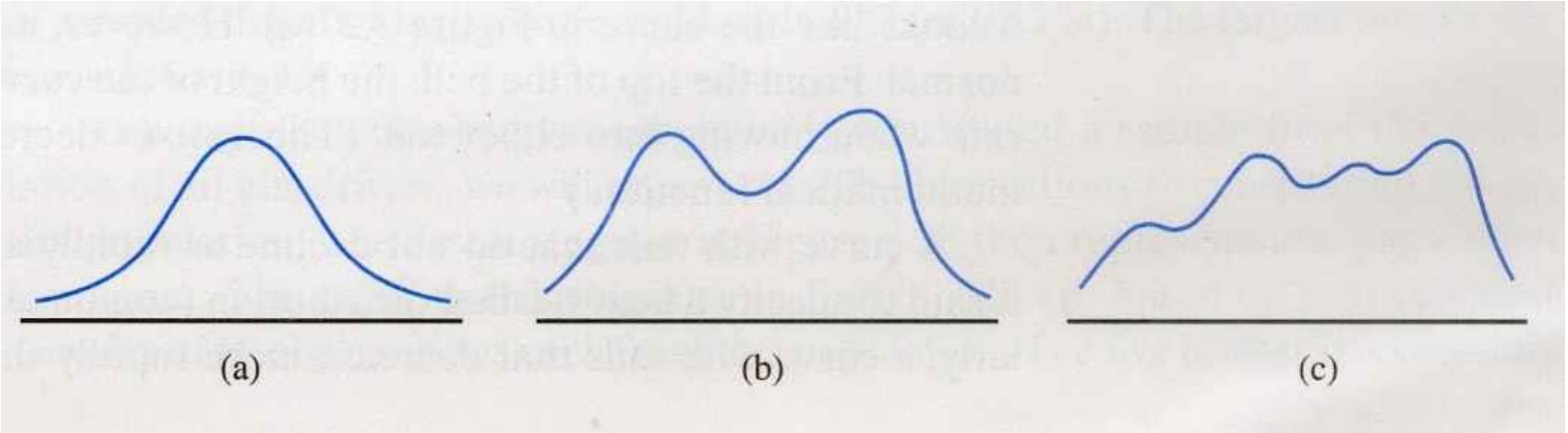


Distribución asimétrica por izquierda

DESCRIPCIÓN GRÁFICA

DISTRIBUCIÓN DE UNA VARIABLE

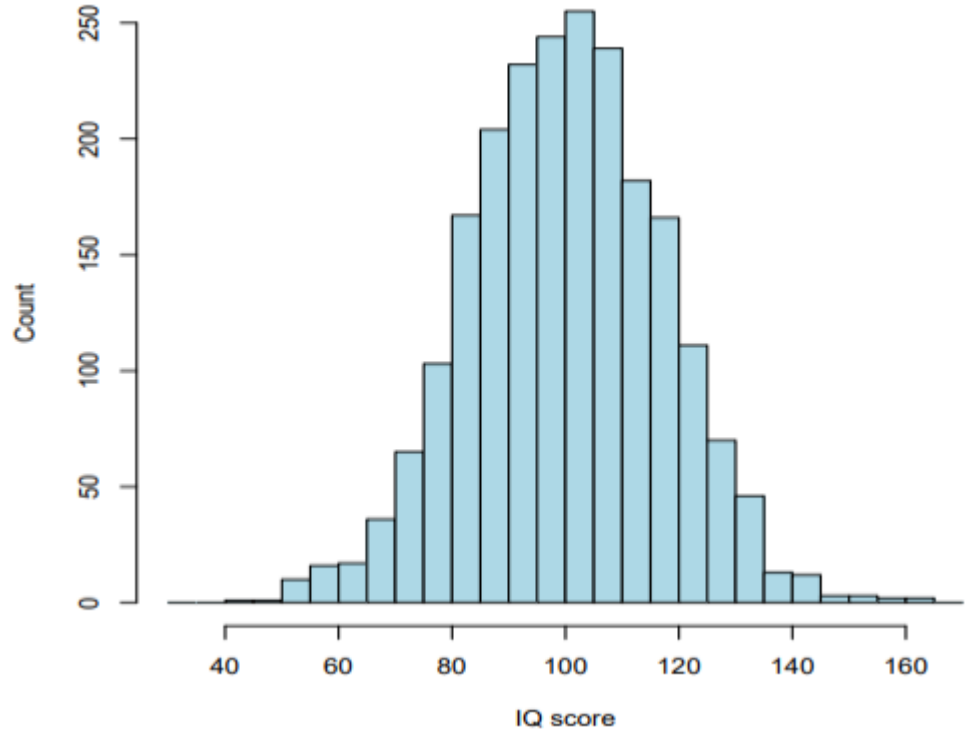
Otra caracterización de la forma general se relaciona con el número de "picos" o "modos". Se dice que un histograma es "unimodal" si tiene un único pico, "bimodal" si tiene dos picos y "multimodal" si tiene más de dos picos.



Distribuciones bimodales se suelen presentar cuando los datos son observaciones realizadas en dos grupos diferentes de individuos u objetos.

Para movernos un poco… ¿cómo lo analizaríamos?

Wechsler Adult Intelligence Scale (WAIS) es una prueba de coeficiente intelectual. Se administró una versión reciente de esta prueba a una muestra de n = 2200 individuos en los Estados Unidos. Estados (de 16 a 90 años). El histograma de las puntuaciones se muestra en la figura siguiente:



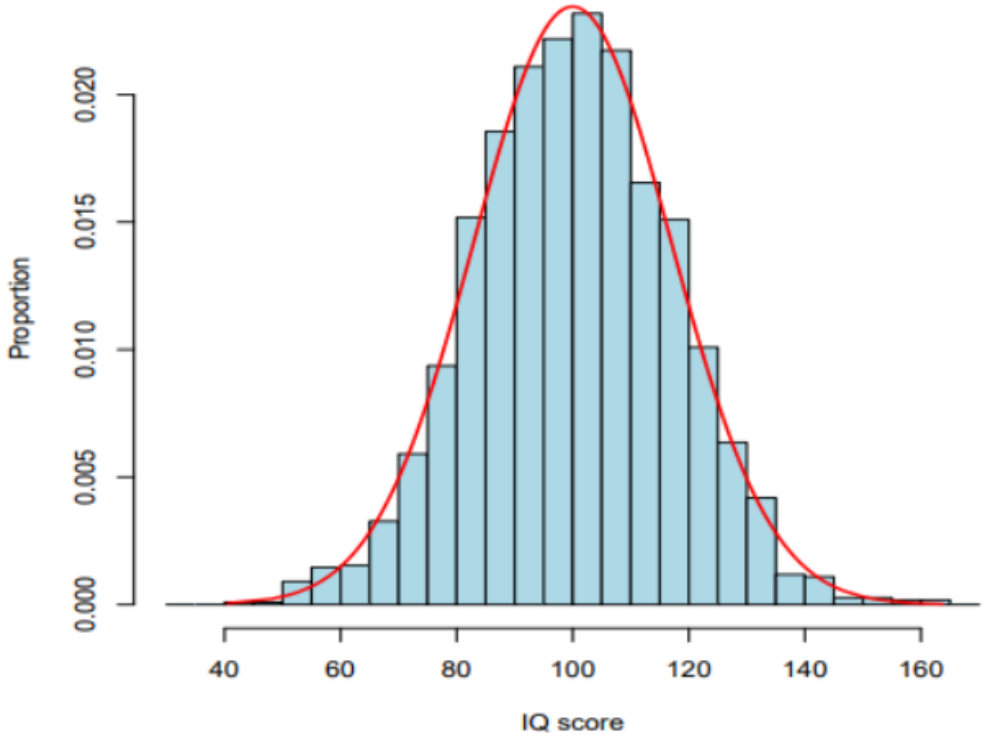
Posibles lecturas:

• El centro de la distribución es de alrededor de 100.

• Dispersión: la mayoría de los puntajes de IQ están entre 60 y 140, pero hay algunos fuera de este rango. No hay valores atípicos llamativos.

• Forma: La distribución de la variable tiene un pico único (alrededor de 100) y es aproximadamente simétrica

Para movernos un poco… ¿cómo lo analizaríamos?



Se superpuso una curva suave sobre los datos de muestra de IQ del grafico anterior. Esta es una estimación de la curva de densidad de población.

En este ejemplo, la curva de densidad de población describe la distribución de los puntajes de IQ para toda la población de individuos (es decir, todos los estadounidenses).

La distribución Normal es la curva de densidad de población más común. Estudiaremos distribuciones Normales en la Unidad 5.

Para movernos un poco… ¿cómo lo analizaríamos?

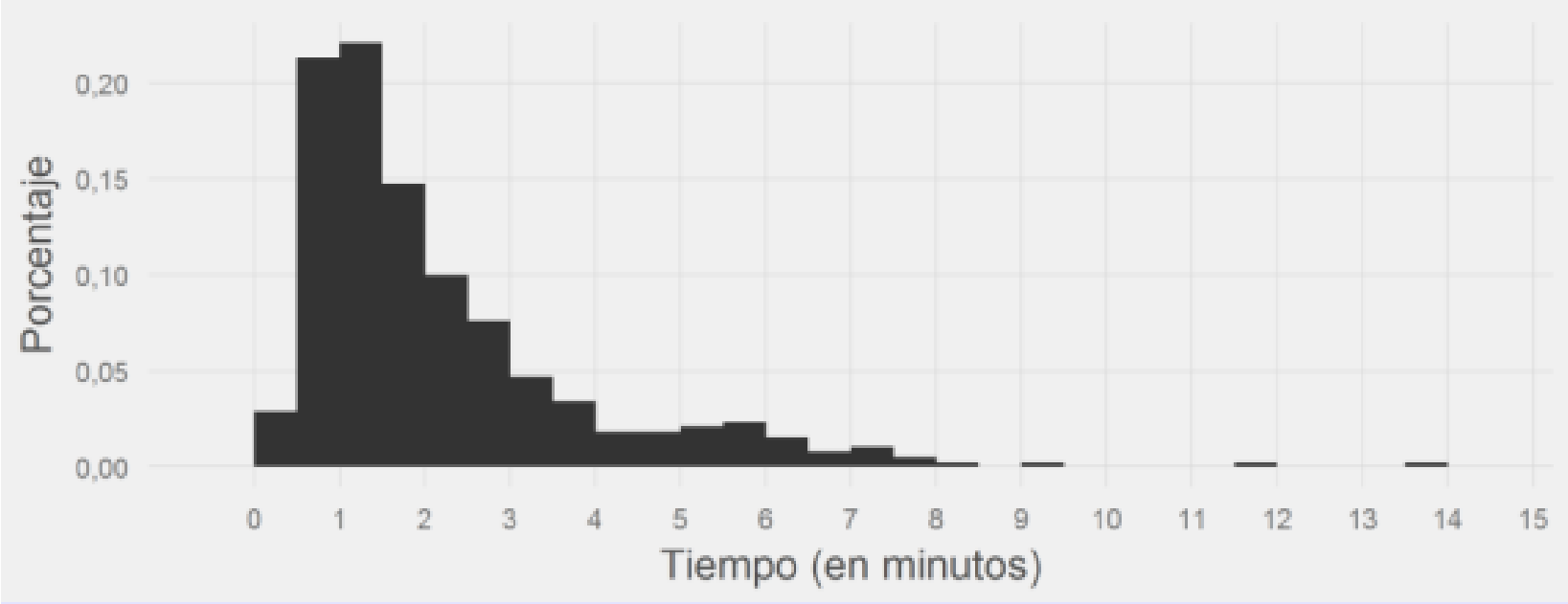
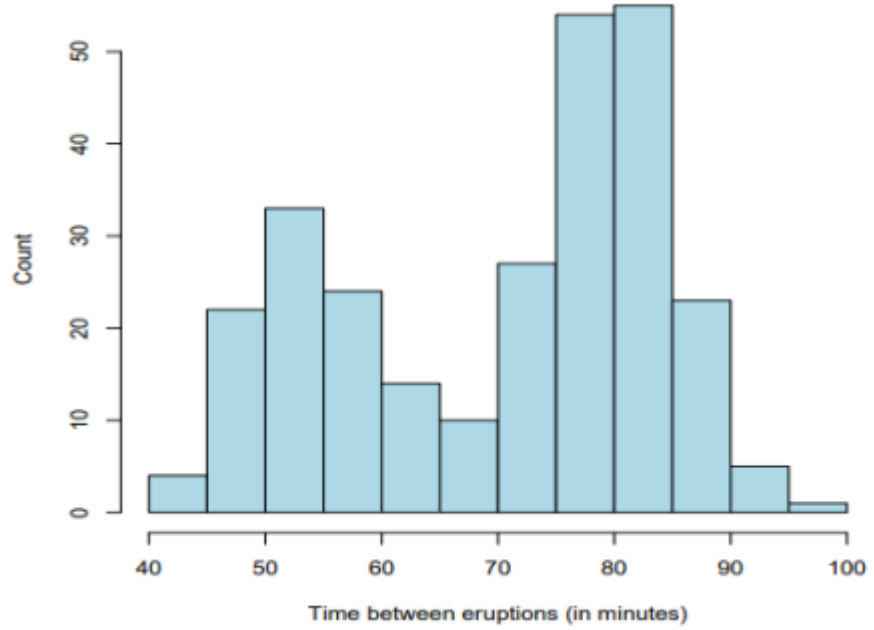


Figura. Distribución del tiempo de realización de un examen de tomografía computada.

La figura muestra la distribución del tiempo en minutos que tardan los examinadores de distintas tomografías computadas. Esta distribución es asimétrica hacia la derecha. Es decir, hay muchos exámenes cortos (de 30 segundos a 2 minutos) y muy pocos largos (más de 5 minutos), de manera que la cola de la derecha del histograma se extiende mucho más lejos que la cola de la izquierda.

Para movernos un poco… ¿cómo lo analizaríamos?



El histograma muestra los tiempos de espera entre inicios de sucesivas erupciones del géiser “Old Faithful” en el Parque Nacional Yellowstone, Wyoming, Estados Unidos. Los datos corresponden a mediciones recolectadas en la década de 1980, resultando un total 272 observaciones.

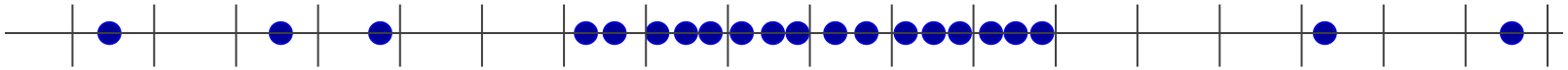
- ¿Forma? Esta distribución tiene dos picos, uno alrededor de 50 minutos y otro alrededor de 80 minutos. Este es un ejemplo de una distribución bimodal.
- El centro de la distribución es de unos 70 minutos, pero esta cifra puede ser engañosa porque la distribución es bimodal.
- Todos los tiempos se encuentran entre 40 minutos y 100 minutos. No hay valores atípicos.

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA: Intervalos de clase de ancho desigual

¿Qué sucede cuando gran parte de los datos se encuentra concentrada en una zona del eje (supongamos el centro de la distribución) y se observan valores muy grandes y/o muy chicos? Empieza a resultarnos difícil encontrar una partición en intervalos que nos sirva completamente:

a) Muchos intervalos angostos: unos pocos intervalos incluiran todas las observaciones, y muchos contendrán 0 observaciones.

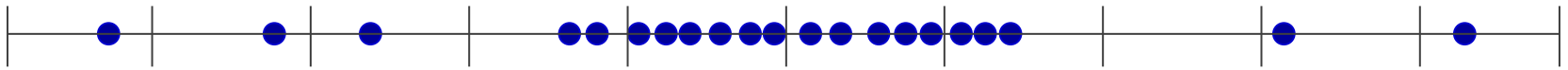


DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA: Intervalos de clase de ancho desigual

¿Qué sucede cuando gran parte de los datos se encuentra concentrada en una zona del eje (supongamos el centro de la distribución) y se observan valores muy grandes y/o muy chicos? Empieza a resultarnos difícil encontrar una partición en intervalos que nos sirva completamente:

b) Pocos intervalos anchos: unos pocos intervalos incluirán la mayoría de las observaciones.

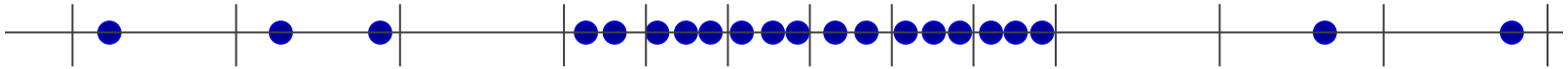


DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA: Intervalos de clase de ancho desigual

¿Qué sucede cuando gran parte de los datos se encuentra concentrada en una zona del eje (supongamos el centro de la distribución) y se observan valores muy grandes y/o muy chicos? Empieza a resultarnos difícil encontrar una partición en intervalos que nos sirva completamente:

c) La mejor alternativa es utilizar pocas clases (relativamente anchas) en los extremos e intervalos angostos en el centro de la distribución



DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA: Intervalos de clase de ancho desigual

En este caso, las frecuencias o frecuencias relativas no deben usarse en los ejes verticales. En su lugar se utiliza la "**densidad**" de la clase, definida del siguiente modo:

$$\text{densidad} = \text{altura del rectángulo} = \frac{\text{frecuencia relativa}}{\text{ancho de la clase}}$$

El uso de la escala de densidad al construir el histograma asegura que el área de cada rectángulo en el histograma es proporcional a la frecuencia relativa correspondiente.

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA: Intervalos de clase de ancho desigual

La tabla siguiente presenta la distribución de la diferencia entre los notas promedio reportadas por una muestra de estudiantes y las verdaderas notas. El objetivo fue evaluar la confiabilidad de utilizar notas auto-reportadas en una investigación sobre técnicas educativas.

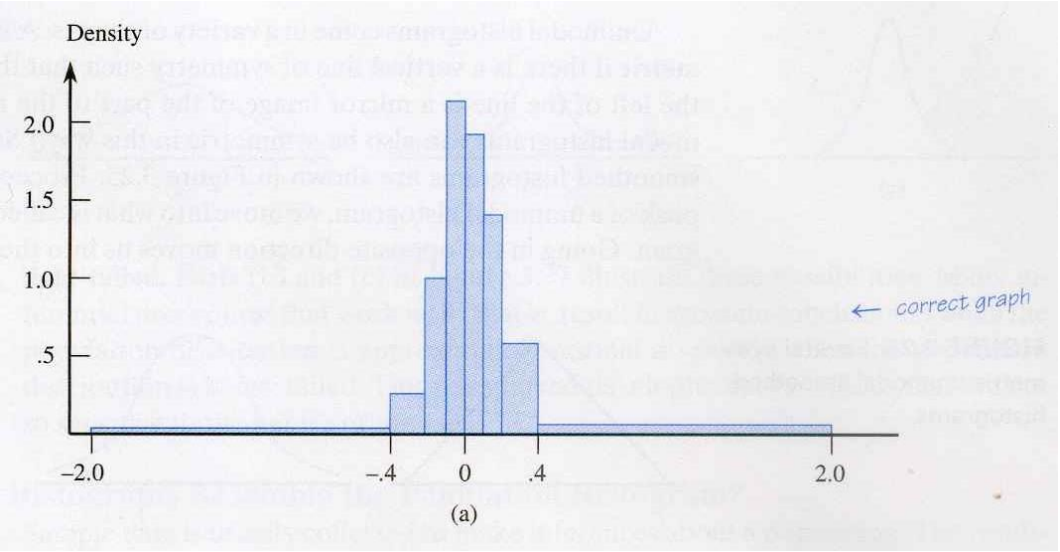
Valores positivos resultan de individuos que reportan promedios mayores a los valores correctos. La mayoría de las diferencias son cercanas a 0, pero se observaron algunas diferencias bastantes grandes.
Como consecuencia de esto, una distribución con intervalos desiguales brindará un resumen conciso y a la vez informativo.

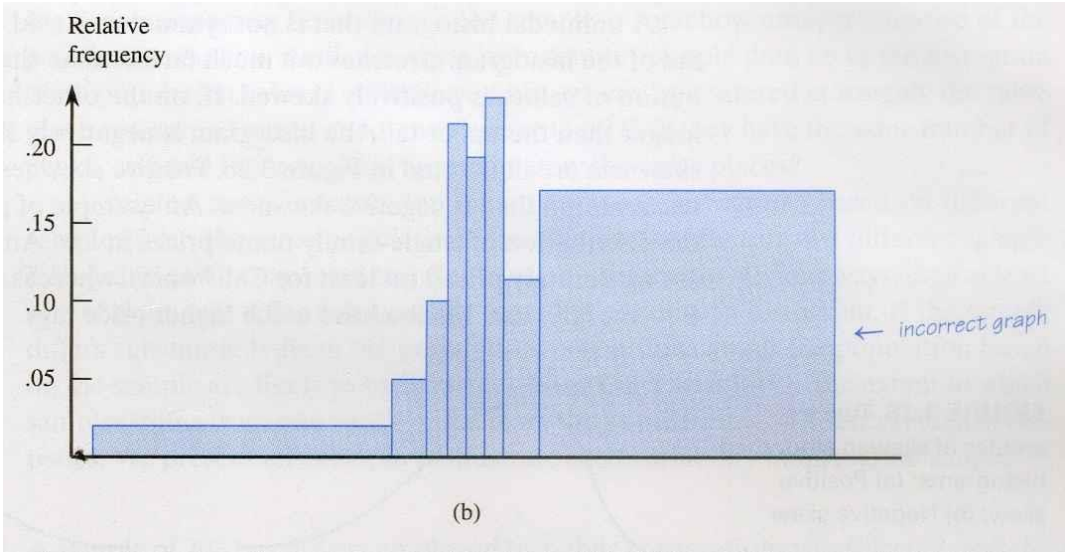
Intervalo de Clase	Frecuencia Relativa	Ancho	Densidad
-2.0 a < -0.4	0.023	1.6	0.014
-0.4 a < -0.2	0.055	0.2	0.275
-0.2 a < -0.1	0.097	0.1	0.970
-0.1 a < 0.0	0.210	0.1	2.100
0.0 a < 0.1	0.189	0.1	1.890
0.1 a < 0.2	0.139	0.1	1.390
0.2 a < 0.4	0.116	0.2	0.580
0.4 a < 2.0	0.171	1.6	0.107

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA CONTINUA: Intervalos de clase de ancho desigual

El gráfico de la izquierda muestra la construcción correcta para el histograma con intervalos de distinto ancho: la representación de la densidad en el eje vertical permite ver la mayor concentración de estudiantes en el centro, cerca del cero, y la fracuencia baja hacia los extremos. El gráfico de la derecha representa la frecuencia relativa en el eje vertical, distorsionando la información que se desea transmitir.





DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA

Los histogramas y gráficos de bastones no son la única forma de representar gráficamente a variables cuantitativas. Existen otros dos diagramas clásicos: el diagrama de caja (o boxplot) y el diagrama de tallo y hoja. El primero de los casos lo veremos cuando abordemos el resumen numérico de una variable, dado que para su construcción se basa en cinco medidas resumen.

El diagrama de tallo y hoja puede ser una buena opción para conjuntos de datos no tan grandes: es sencillo de construir y presenta información más detallada que un histograma. Es una herramienta de análisis exploratorio de datos que muestra el rango de los datos, dónde están más concentrados, su simetría y la presencia de datos atípicos.

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA

Para hacer un diagrama de tallos y hojas:

- Separá cada observación en un tallo que contenga todos los dígitos menos el del final (el situado más a la derecha) y una hoja, con el dígito final. Los tallos pueden tener tantos dígitos como se desee, pero las hojas solo poseen uno.
- Situá los tallos de forma vertical en orden de arriba hacia abajo. Trazá una línea vertical a la derecha de los tallos.
- Situá cada hoja a la derecha de su tallo, en orden creciente dentro de cada tallo.

Los siguientes datos corresponden a los puntajes obtenidos por 66 estudiantes en el examen final de un curso de posgrado en esta universidad.

95 98 93 91 95 90 90 96 98 89 93 92
88 79 91 83 85 90 81 87 79 87 88 83
86 80 77 81 81 78 79 82 78 76 79 76
73 81 78 84 70 71 77 65 69 70 63 77
74 82 69 74 67 44 63 70 57 63 51 52
22 57 47 54 52 76

Para estos datos, los tallos podrían ser las decenas y las hojas las unidades. Por ejemplo, para 95: el tallo es el 9 y la hoja es 5.

DESCRIPCIÓN GRÁFICA

VARIABLE CUANTITATIVA

Interpretación:

• El centro de la distribución se encuentra entre 70 y 80.

• La mayoría de los puntajes están entre 44 y 98. Hay un valor atípico obvio en 22.

• La variable presenta una distribución asimétrica hacia la izquierda.

```
2 | 2
3 |
4 | 47
5 | 122477
6 | 3335799
7 | 00013446667778889999
8 | 01111223345677889
9 | 000112335568 8
```

Importante: la escala de valores de los tallos siempre debe estar completa, como los ejes X en los gráficos de bastones y en los histogramas

DESCRIPCIÓN GRÁFICA

VARIABLE TEMPORAL

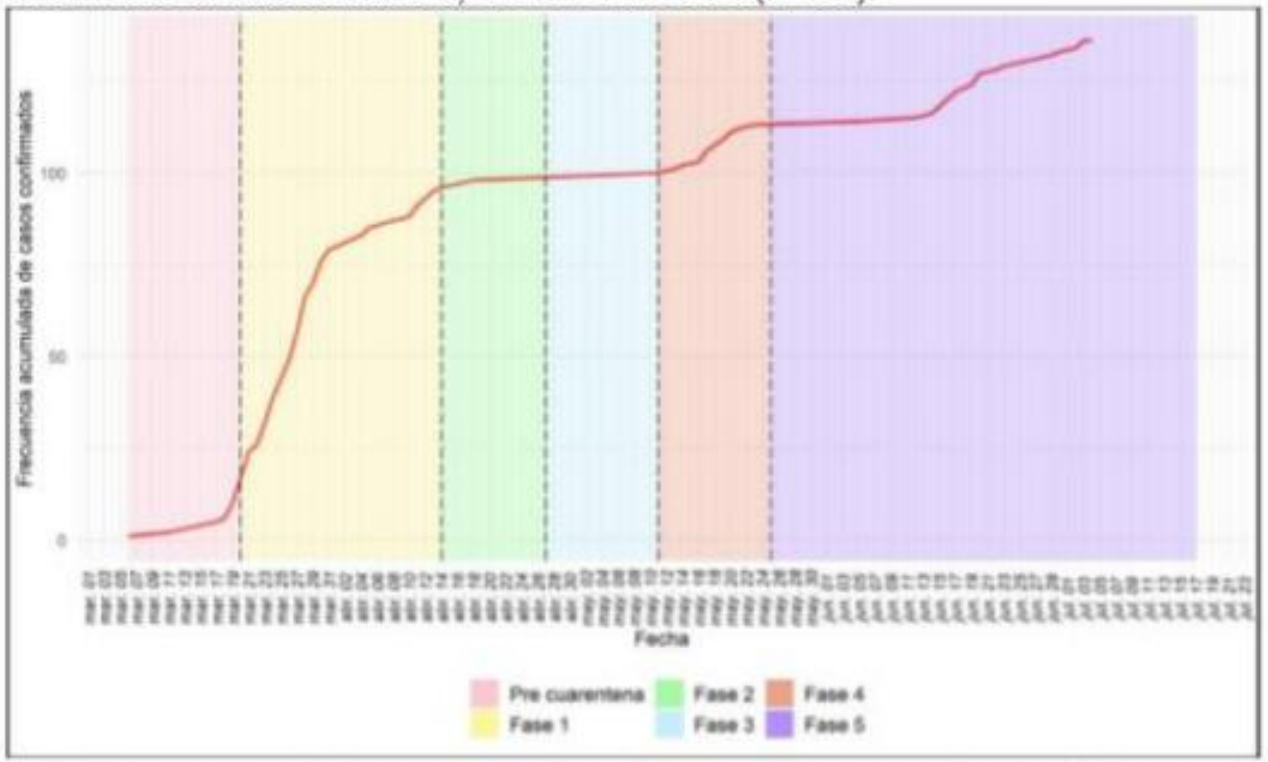
Hay ocasiones en que no se cuenta con un conjunto de datos en las que se mide la misma variable a distintas unidades, sino que se trata de distintas mediciones de la misma variable a través del tiempo.

Un gráfico temporal de una variable representa cada observación en relación al momento en que se midió. Sitúa siempre el tiempo en el eje de las abscisas, y el valor de la variable se ubica en el eje de las ordenadas. Cada punto indica el valor de la variable medido en un momento en particular del tiempo, y la unión de puntos contiguos mediante segmentos facilita la visualización de los cambios a lo largo del tiempo.

DESCRIPCIÓN GRÁFICA

VARIABLE TEMPORAL

Gráfico 3: Frecuencia acumulada de casos confirmados de COVID-19, según fases de aislamiento. Ciudad de Rosario, SE 1 a 27 de 2020 (n=136).



Fuente: A partir de datos disponibles en Sistema Integrado de Información Sanitaria Argentino (SISA). Sistema Municipal de Epidemiología. Secretaría de Salud Pública. Municipalidad de Rosario. Rosario, 5 de julio de 2020.

Nota: Las sucesivas fases de aislamiento se describen en ANEXO II.

DESCRIPCIÓN GRÁFICA

FRECUENCIAS ACUMULADAS

Hay aplicaciones en las que las frecuencias acumuladas resultan de mayor interés que las frecuencias simples. Los gráficos que representan este tipo de frecuencias básicamente se basan en identificar con puntos las frecuencias acumuladas y trazar líneas con algún criterio. Los gráficos de frecuencias acumuladas sirven particularmente para determinar el número (o proporción) de datos que se encuentran por debajo de un valor determinado.

Para poder construir un gráfico de frecuencias acumuladas es necesario tener en cuenta la naturaleza de la variable a representar: si es una variable discreta se construirá un gráfico escalonado y si es una variable continua el gráfico recibe el nombre de ojiva.

DESCRIPCIÓN GRÁFICA

FRECUENCIAS ACUMULADAS

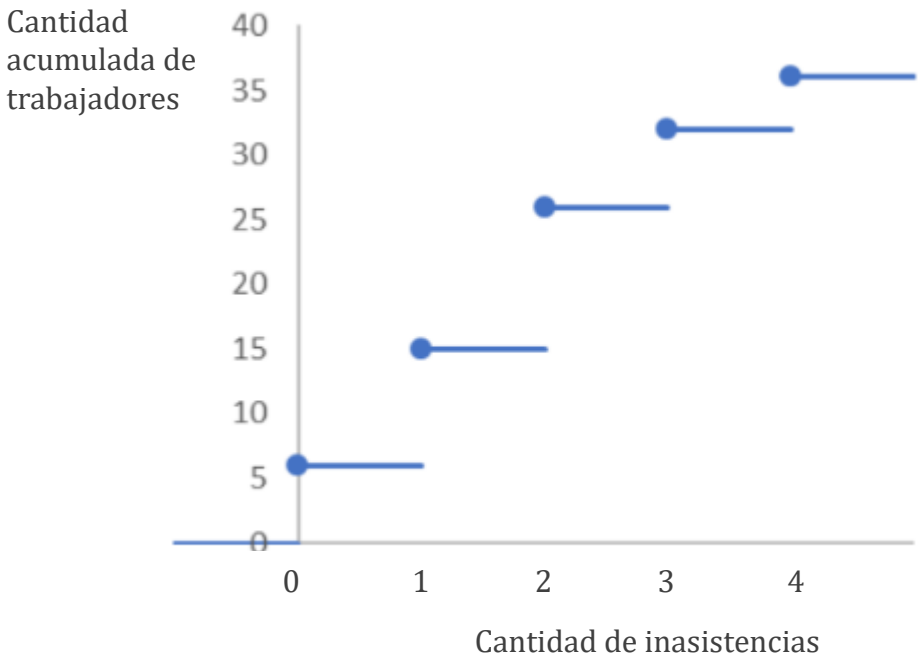
En los gráficos escalonados se ubican los valores de la variable (x_i) en el eje horizontal y los de las frecuencias acumuladas (F_i o H_i) en el eje vertical. Para cada valor de x se traza una línea horizontal a la altura de la frecuencia acumulada correspondiente a dicho valor. Esta línea horizontal se extiende hasta el siguiente valor de x. Antes del primer valor observado la frecuencia acumulada es igual a cero, y esto se representa mediante una línea horizontal sobre el eje horizontal. Para el último valor observado, y todos los valores mayores a él, ya se acumuló el 100% del conjunto de datos, y esto se representa trazando una línea horizontal a partir de este valor.

DESCRIPCIÓN GRÁFICA

FRECUENCIAS ACUMULADAS

x_i: número de inasistencias	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	

Gráfico N°1. Inasistencias de los empleados de un importante laboratorio. Primer bimestre, año 2024.



DESCRIPCIÓN GRÁFICA

FRECUENCIAS ACUMULADAS

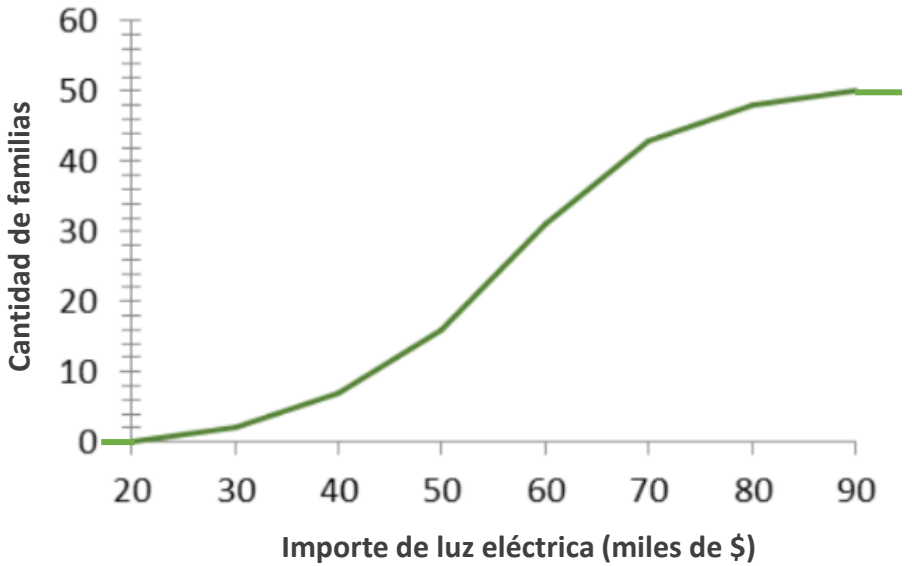
En la ojiva los intervalos de la variable x se representan en el eje horizontal y las frecuencias acumuladas (F_i o H_i) en el eje vertical. Para cada intervalo, se indica un punto de altura igual al valor de su frecuencia acumulada correspondiente. Dicho punto se marca a la altura del límite superior de cada intervalo. Para el límite inferior del primer intervalo se indica un punto de altura 0, es decir que toda ojiva comienza desde el eje horizontal. Para el límite superior del último intervalo, y todos los valores mayores a él, ya se acumuló el 100% del conjunto de datos, y esto se representa trazando una línea horizontal a partir de este valor.

DESCRIPCIÓN GRÁFICA

FRECUENCIAS ACUMULADAS

x_i: importe de luz (miles de $)	f_i	Fi	h_i	H_i
[20, 30)	2	2	0,04	**0,04**
[30, 40)	5	7	0,10	**0,14**
[40, 50)	9	16	0,18	**0,32**
[50, 60)	15	31	0,30	**0,62**
[60, 70)	12	43	0,24	**0,46**
[70, 80)	5	48	0,10	**0,96**
[80, 90)	2	50	0,04	**1,00**
Total	**50**		**1**	

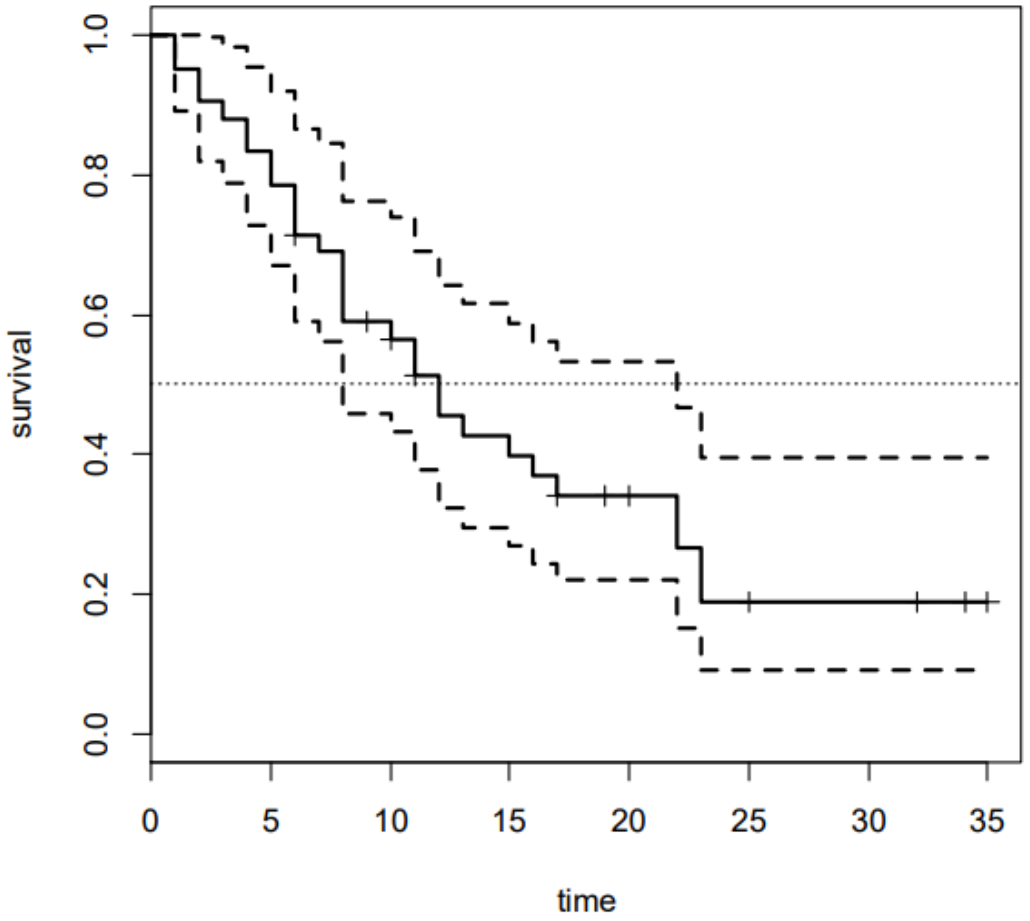
Gráfico N°2. Importe total de luz abonado por familia durante el último semestre de 2023. Ciudad de Alvear.



Fuente: Encuesta de gastos de hogares (IPEC).

DESCRIPCIÓN GRÁFICA

FRECUENCIAS ACUMULADAS



Hay aplicaciones en las que se hace uso de las frecuencias acumuladas pero de no de forma directa. El gráfico que se muestra a la izquierda es una figura típica del análisis de datos de supervivencia: se observa el tiempo que transcurre en cada unidad de análisis hasta que ocurre un evento de interés. Lo que se suele graficar en estos casos es la función de supervivencia, que no es más que $1 - H_i$ para representar la proporción de unidades que aún no presentaron dicho evento.

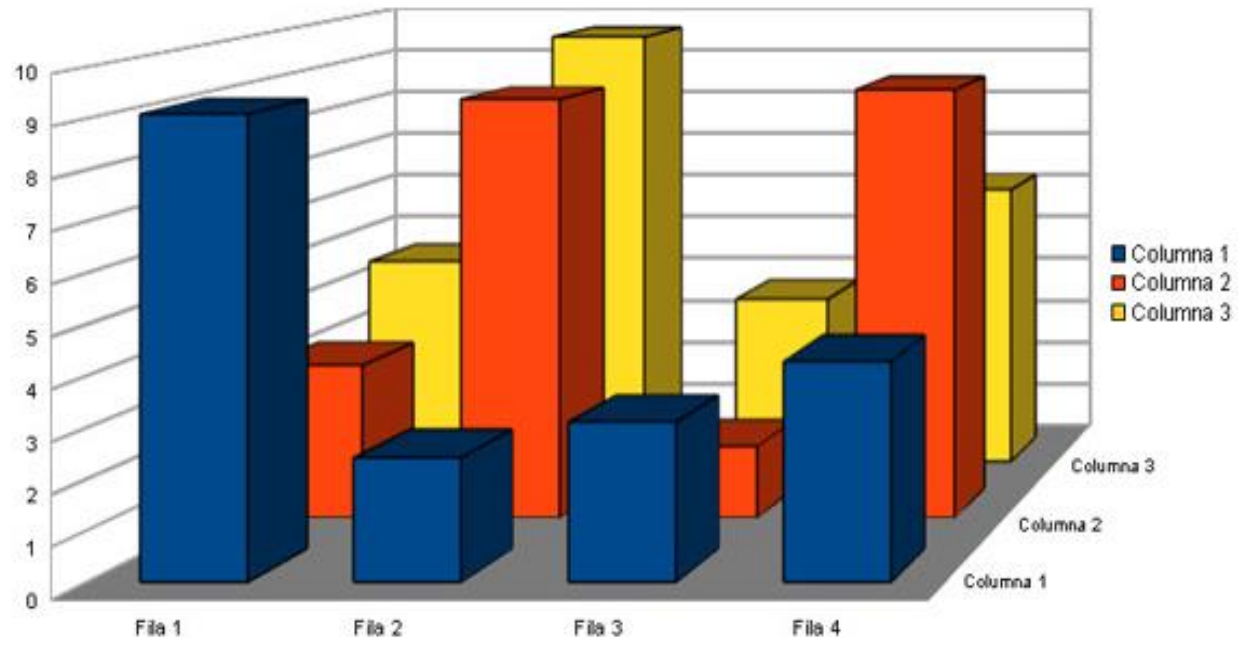
DESCRIPCIÓN GRÁFICA

ERRORES FRECUENTES

Los cuadros y gráficos estadísticos son herramientas poderosas que permiten la transmisión de información de forma sintética y rápida. Sin embargo, gráficos construidos de forma incorrecta pueden llevar a la mal interpretación de los datos. En esta sección se presentan algunos ejemplos.

DESCRIPCIÓN GRÁFICA

ERRORES FRECUENTES

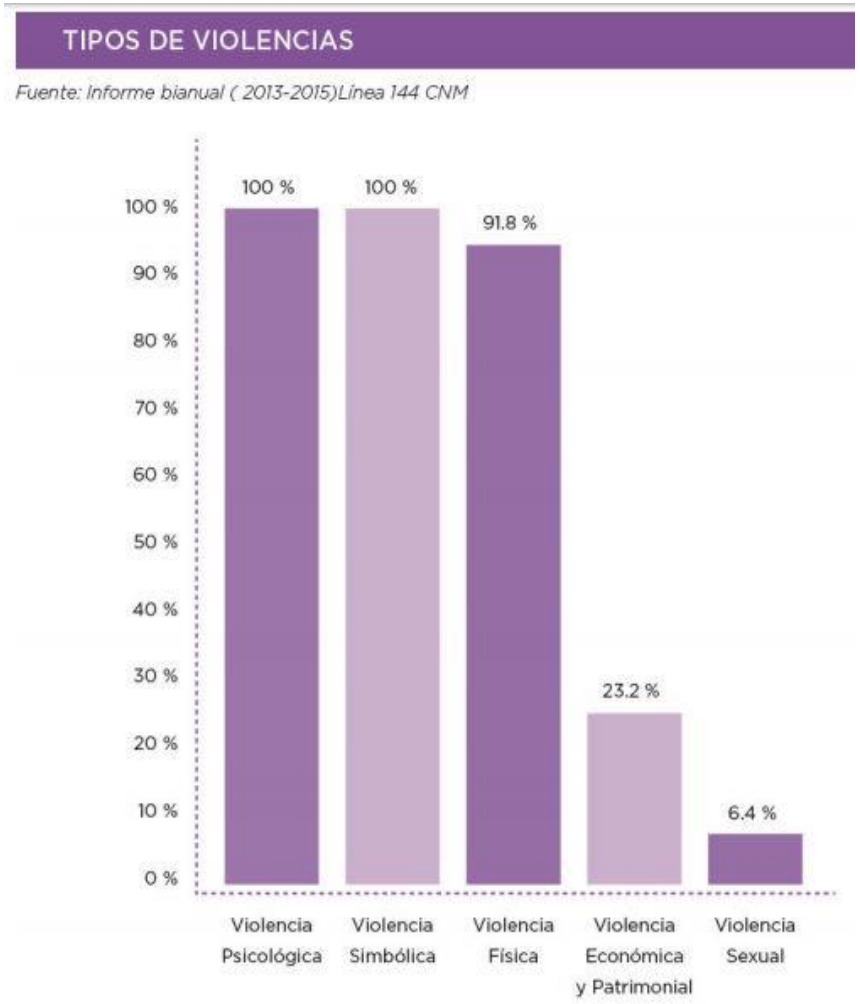


Gráficos en 3D

Distorsionan la visualización de la información, principalmente de las magnitudes al comparar categorías.

DESCRIPCIÓN GRÁFICA

ERRORES FRECUENTES

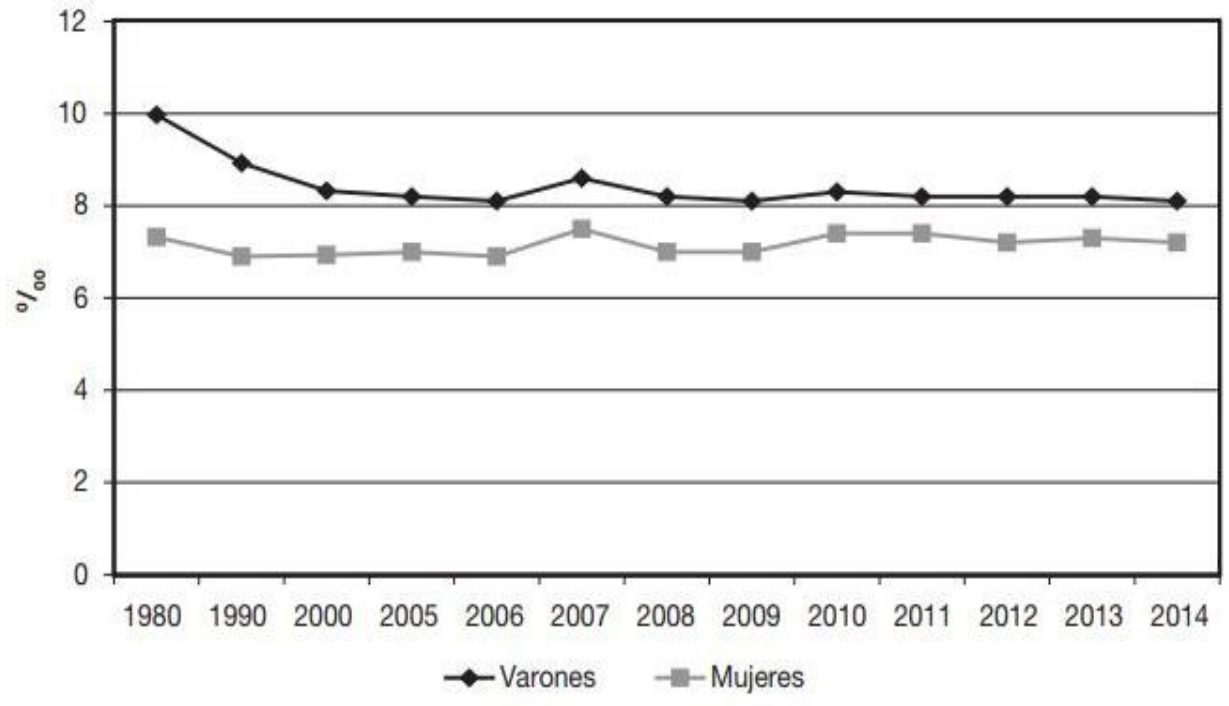


Gráficos de barras simples con diferentes colores

Diferenciar las barras con colores puede dar la sensación de que se trata de grupos distintos, cuando en realidad todas las barras son categorías de un mismo atributo.

DESCRIPCIÓN GRÁFICA

ERRORES FRECUENTES



Escalas incorrectas en los ejes

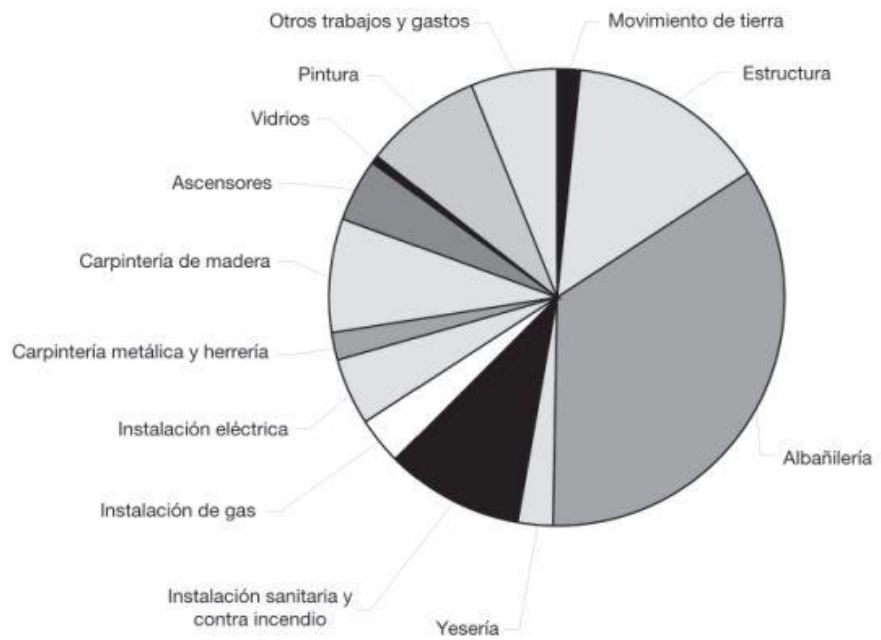
Debe mantenerse una distancia entre períodos que represente de forma fiel el tiempo transcurrido entre uno y otro.

Entre 1980 y 1990 no transcurrió el mismo tiempo que entre 2000 y 2005 ni que entre 2006 y 2007.

DESCRIPCIÓN GRÁFICA

ERRORES FRECUENTES

Gráfico 24. **Estructura de ponderaciones por ítem de obra. Índice del Costo de la Construcción (ICC)**



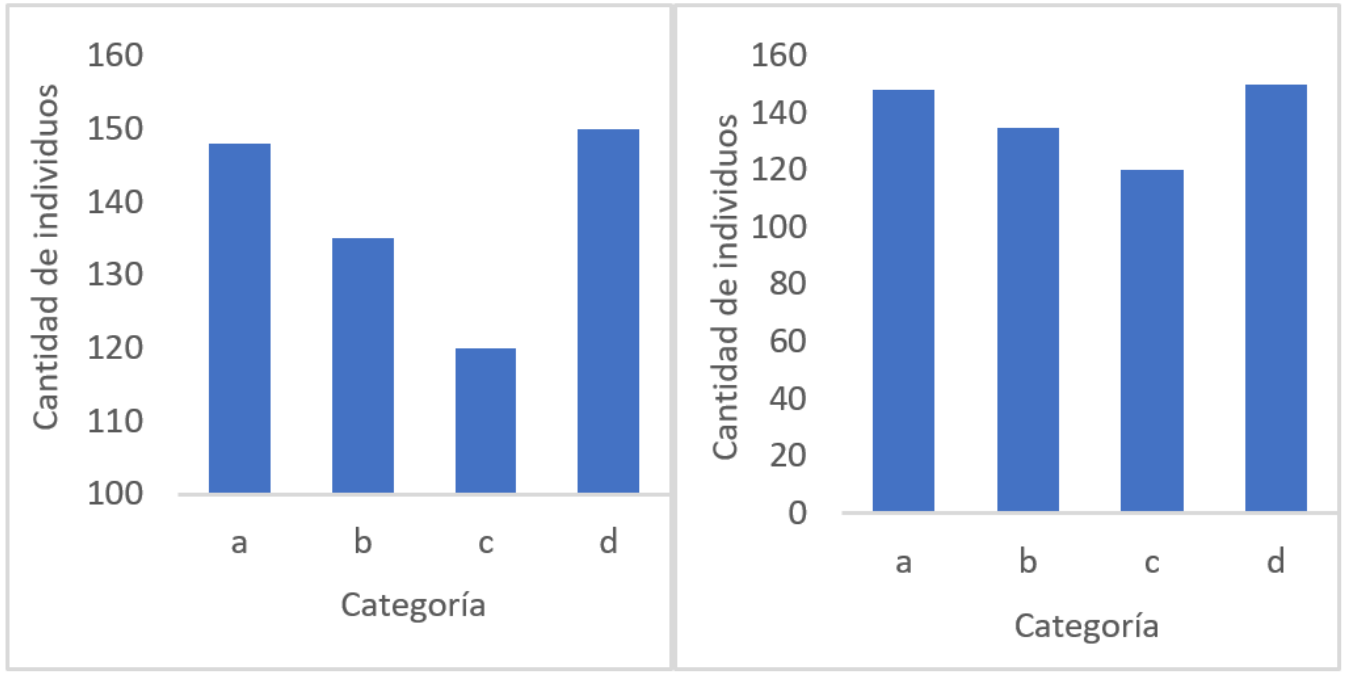
Fuente: INDEC. Índice del Costo de la Construcción (ICC).

Gráfico de sectores circulares con muchos niveles

Dificultan la interpretación y la comparación entre valores de distintas categorías.

DESCRIPCIÓN GRÁFICA

ERRORES FRECUENTES

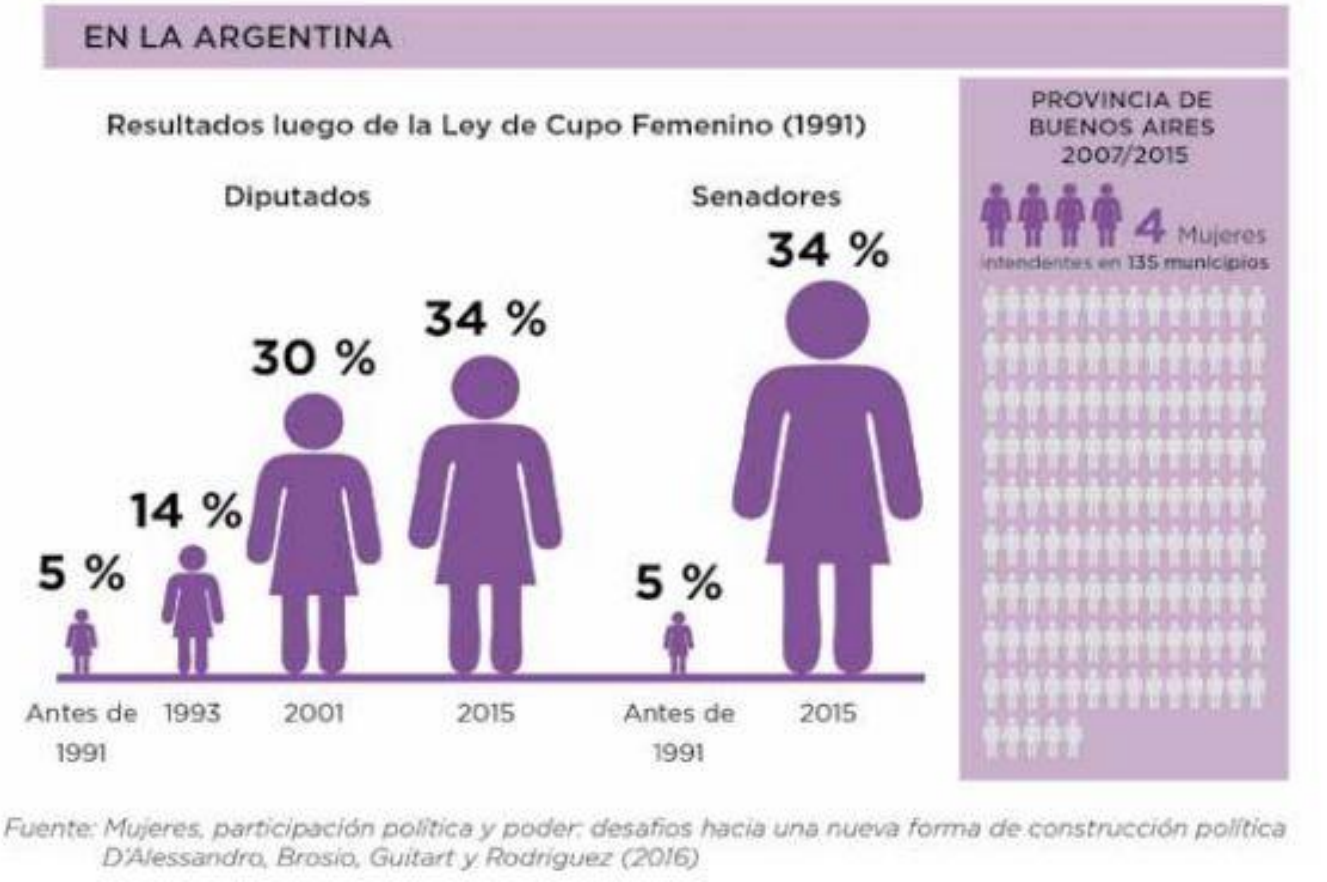


Ejes cortados

Modifican la percepción al comparar categorías. La primera figura muestra el eje vertical de 100 a 160, mostrando amplias diferencias entre las barras. Estas diferencias se reducen al colocar el eje correctamente.

DESCRIPCIÓN GRÁFICA

ERRORES FRECUENTES



Fuente: Mujeres, participación política y poder: desafíos hacia una nueva forma de construcción política D'Alessandro, Brosio, Guitart y Rodríguez (2016)

***Infografía 1*

Pictogramas e infografías

Se obtienen al reemplazar las barras o gráficos tradicionales por otros recursos gráficos. Se debe ser cuidadoso con la construcción dado que lo que el ojo humano percibe para comparar categorías es la superficie de las figuras, y no la altura como se suele creer.

Ejemplo… para retomar lo ya visto!

Ejemplo: Un importante laboratorio, preocupado por la inasistencia de sus empleados, decidió investigar el tema. Para ello, seleccionó una muestra aleatoria de 36 empleados y registró el número de inasistencias en el primer bimestre del corriente año.
Los datos obtenidos fueron los siguientes:

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

La variable investigada es…

número de inasistencias en el primer bimestre del corriente año.

Es una variable de tipo…

Dado que sólo puede asumir valores enteros, se trata de una variable discreta.
Dado que el valor 0 indica la ausencia de inasistencias (ausencia del atributo)
está medida en escala de razón.

La cantidad total de elementos de la muestra es…

36 (n=36).

Ejemplo… para retomar lo ya visto!

¿Cómo procederíamos para armar una tabla de distribución de frecuencias?

x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	

Tabla N°1. Inasistencias de los empleados de un importante laboratorio. Primer bimestre, año 2024.

Ejemplo… para retomar lo ya visto!

¿Cómo interpretaríamos una fila de la tabla de distribución de frecuencias?

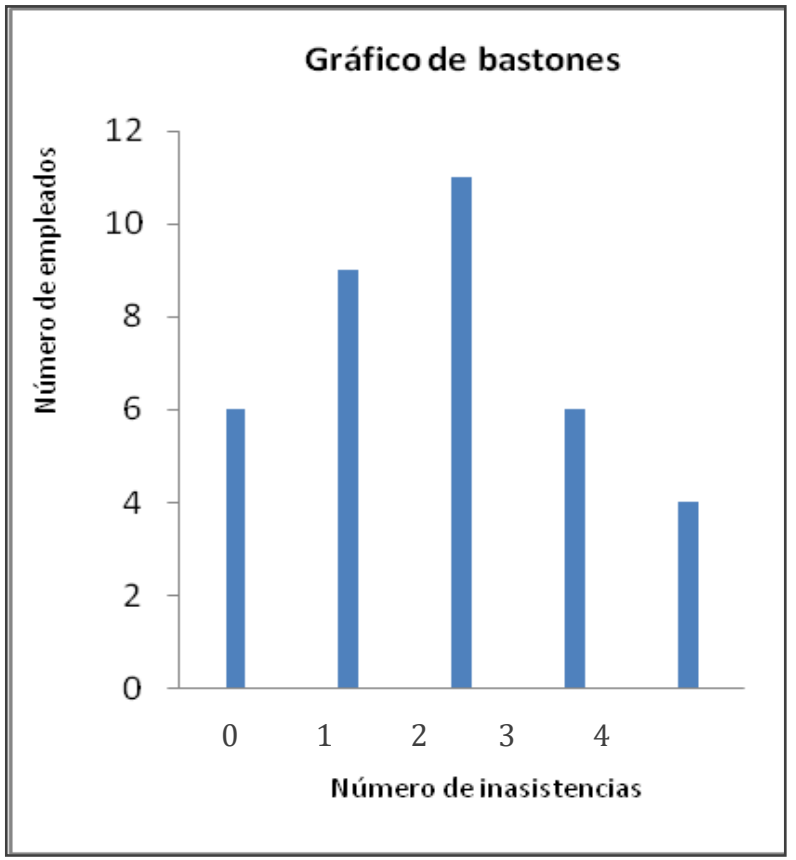
x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	

- 11 de los empleados investigados tuvieron 2 inasistencias en el primer bimestre del corriente año.
- 26 de los empleados tuvieron hasta 2 inasistencias en el primer bimestre del corriente año.
- El 30% de los empleados tuvo 2 inasistencias en el primer bimestre del corriente año.
- El 72% de los empleados investigados tuvo 2 inasistencias o menos en el primer bimestre del corriente año.

Ejemplo… para retomar lo ya visto!

¿Cómo graficaríamos esta distribución de frecuencias?

x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	



Análisis descriptivo de datos
Análisis univariado numérico

DESCRIPCIÓN NUMÉRICA

Es relativamente sencillo tener en mente y captar la información que aportan diez números; cien es difícil y con mil, estamos perdidos. Por esa razón, es muy importante contar con medidas resumen, que de alguna manera describan las características más sobresalientes del conjunto de datos que se está analizando.

Una **medida resumen** es un número. Se obtiene a partir de un conjunto de datos y, en cierta forma, lo caracteriza. Si se trata de una muestra, es el valor de un estadístico; si se cuenta con datos de una población entonces será un parámetro. Por ejemplo, un porcentaje o una proporción son medidas resumen.

Las medidas resumen permiten tener una idea rápida de cómo son los datos. Pero, un estadístico mal utilizado puede dar una idea equivocada respecto de las características generales que interesa mostrar.

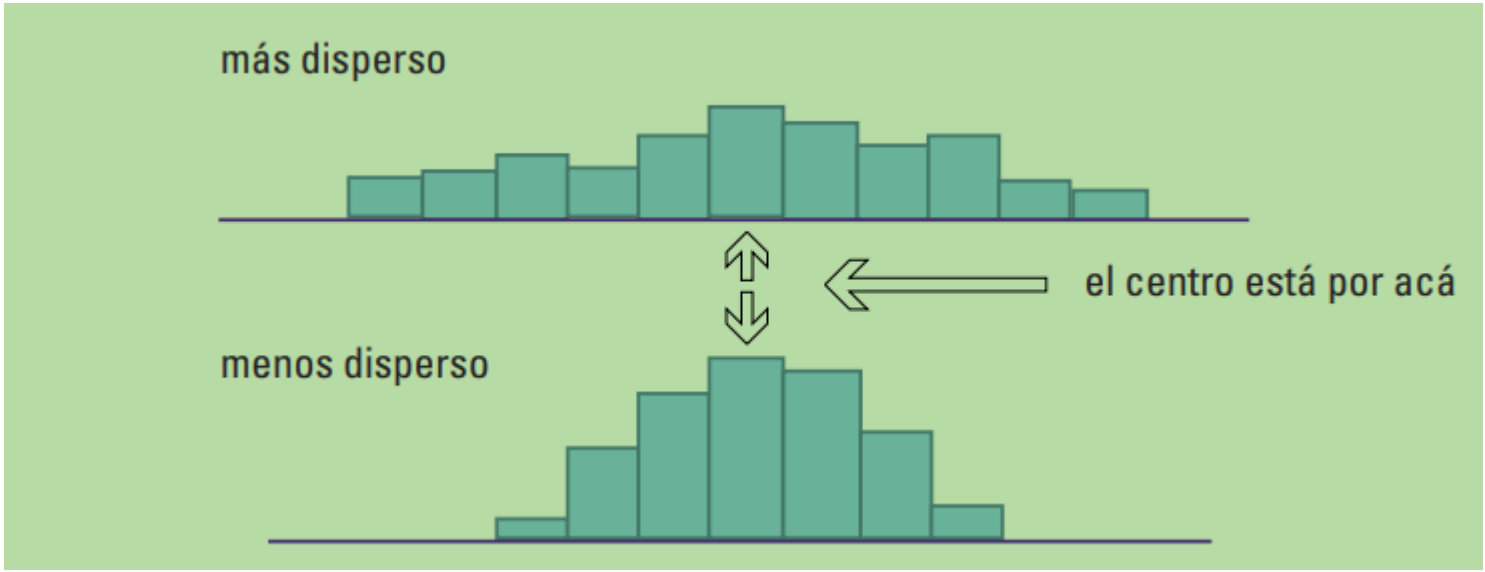
DESCRIPCIÓN NUMÉRICA

El cálculo de medidas resumen es el primer paso, junto con la descripción gráfica, que se realiza cuando se recolectan los datos en un estudio para tener una idea de qué está pasando. Posteriormente, los investigadores pondrán a prueba sus hipótesis respecto a algún parámetro poblacional, estimarán características de la población y estudiarán posibles relaciones entre las variables. Cuando presentan sus conclusiones al público en general, las medidas resumen son útiles para mostrar los resultados en forma concisa y clara.

En principio, se pueden obtener muchísimas formas de resumir los valores de un conjunto de datos numéricos. Es importante que sean fáciles de interpretar.

DESCRIPCIÓN NUMÉRICA

Cualquier conjunto de datos numéricos tiene dos propiedades importantes: un **valor central** y la **dispersión** alrededor de ese valor. Vemos esta idea en los siguientes histogramas hipotéticos:



DESCRIPCIÓN NUMÉRICA

Medidas de posición

Resumen numéricamente en qué lugar del eje de valores se encuentran los datos observados.

Medidas de tendencia central

Reflejan valores centrales de la distribución. Las más usuales son la media aritmética, la mediana y la moda.

Medidas de dispersión

Resumen numéricamente la homogeneidad o heterogeneidad existente en los datos. Valores más grandes indican mayor dispersión.

DESCRIPCIÓN NUMÉRICA

Variables categóricas

- Proporción ($\hat{p}$)
- Modo/Moda (Mo)

Variables cuantitativas

Medidas de posición

- Media aritmética ($\bar{x}$)
- Mediana ($Mna = Q_2$)
- Modo/Moda (Mo)
- Cuartiles (Q_1, Q_2, Q_3)
- Percentiles

Medidas de dispersión

- Rango (R)
- Variancia ($Var = S^2$)
- Desvío estándar (S)
- Rango intercuartílico (RI)
- Coeficiente de variación

Estadística descriptiva

DATOS CATEGÓRICOS

Las medidas resumen para un conjunto de datos categórico son las frecuencias relativas para las distintas categorías. Cada frecuencia relativa es la **proporción** (fracción) de respuestas que caen en la categoría correspondiente.

La proporción muestral de éxitos, simbolizada con $\hat{p}$, es

$$\hat{p} = \frac{cantidad\ de\ observaciones\ en\ la\ categoría}{n}$$

Estadística descriptiva

DATOS CATEGÓRICOS: Ejemplo

El uso de equipamiento contra la contaminación en automóviles ha mejorado la calidad del aire en ciertas áreas. Desafortunadamente, muchos propietarios de automóviles modifican los mecanismos contra la contaminación para obtener mejor rendimiento en sus automóviles. Suponer que se selecciona una muestrea de 15 automóviles y se los clasifica según si los mecanismos han sido modificados (M) o si permanecen originales (O).

M	**O**	**M**	**M**	**M**	**O**	**O**	**M**	**M**	**O**	**M**	**M**	**M**	**O**	**O**

La muestra contiene 9 autos con sus mecanismos modificados, por lo tanto

$$\hat{p} = \frac{9}{15} = 0{,}60$$

Para representar la proporción poblacional se utiliza la letra griega π o la letra p.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Media aritmética

La media aritmética es una de las medidas de tendencia central más conocidas. Se obtiene sumando todos los datos y dividiendo por la cantidad total de observaciones.

Si contamos con una muestra de tamaño n, con observaciones: $x_1, x_2, \dots, x_n$. La media aritmética de esos datos se define como

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

Es una estadística, dado que se calcula con las observaciones de la muestra.

Si contamos con una población de tamaño N, con observaciones: $x_1, x_2, \dots, x_N$. La media aritmética de esos datos se define como

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Es un parámetro, dado que se calcula con las observaciones de la población.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

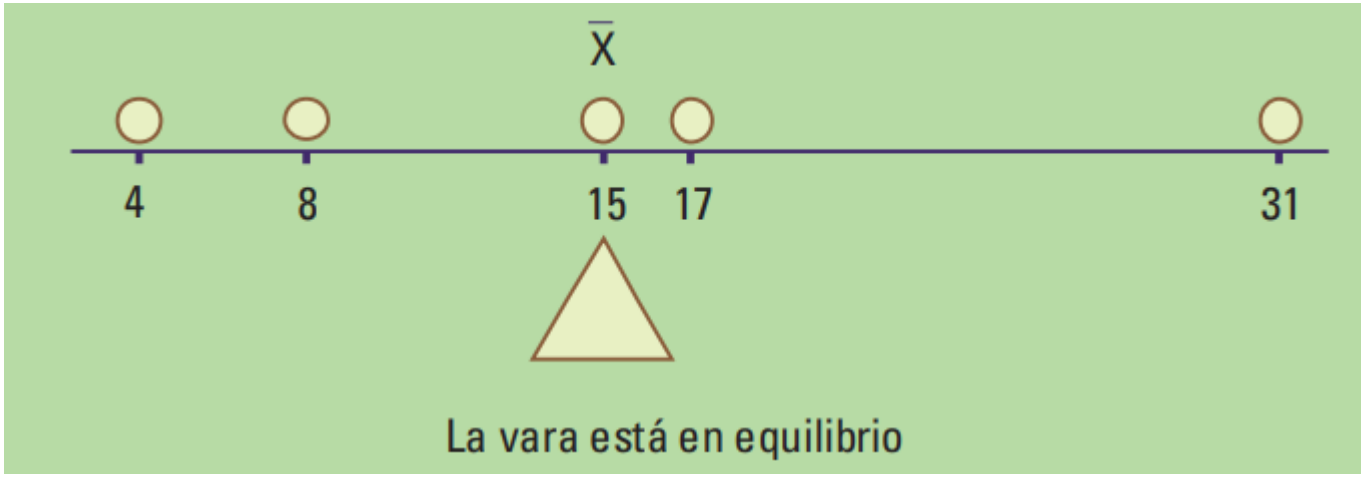
Media aritmética

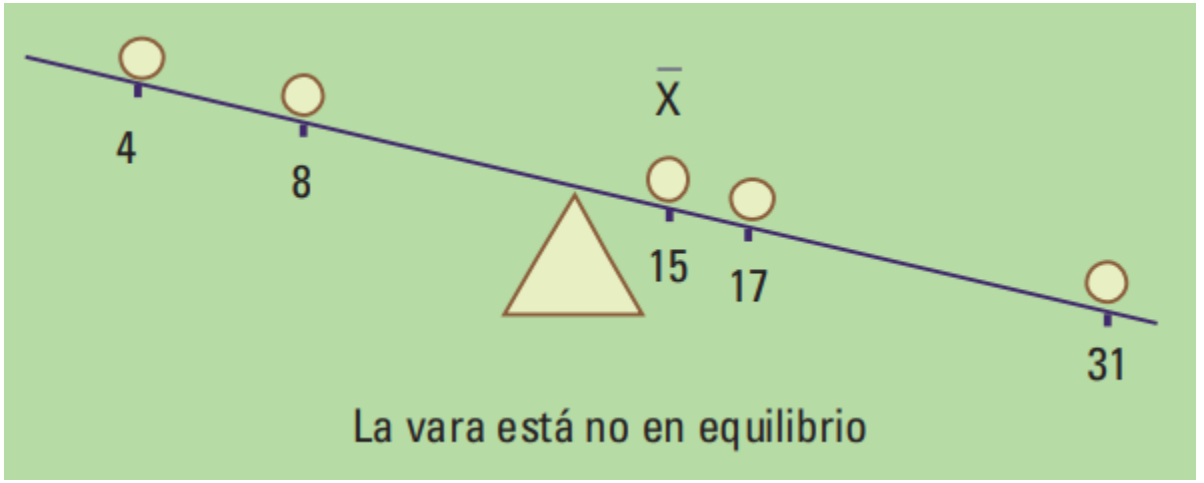
De manera coloquial, solemos llamar a la media aritmética como “promedio” o simplemente como “media”. El promedio representa el valor característico o central de un conjunto de números.

También es considerada como el punto de equilibrio del conjunto de datos: si sobre una vara numerada sin peso se colocan pesos idénticos sobre el valor de cada dato, la vara queda en equilibrio cuando se la apoya en el punto correspondiente a la media aritmética; la vara no queda en equilibrio si se la apoya en cualquier otro punto distinto al de la media.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL





Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Media aritmética

La expresión $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ es válida cuando se tienen datos sin agrupar.

Si en cambio, los datos están organizados en una distribución de frecuencias, es decir, se sabe para cada valor de x_i cuántas veces se repite (f_i), el promedio se obtiene de la siguiente manera:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{n} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$$

donde k es la cantidad de clases en las que se agrupa la variable.

Ejemplo… para ir calentando y retomar lo ya visto!

¿Cómo calcularíamos la media aritmética?

x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	

$$\bar{x} = \frac{0 \times 6 + 1 \times 9 + 2 \times 11 + 3 \times 6 + 4 \times 4}{36} = 1{,}8055$$

"En promedio, los empleados del laboratorio faltaron 1,8 veces durante el segundo bimestre del año", "El número promedio de inasistencias en el segundo bimestre del año de los empleados del laboratorio, es de 1,8 días" y "Los empleados del laboratorio tuvieron un promedio de 1,8 inasistencias durante el segundo bimestre del año" son interpretaciones alternativas.

Los valores en la muestra eran todos enteros, sin embargo la media se reportó como: 1,8. Es común utilizar más dígitos de precisión decimal cuando se reporta la media.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

x_i: tiempo hasta terminar cierta tarea	f_i	Fi	h_i	H_i
[5; 10)	6	6	6/36=0,17	**0,17**
[10; 15)	9	15	0,25	**0,42**
[15; 20)	11	26	0,30	**0,72**
[20; 25)	6	32	0,17	**0,89**
[25; 30)	4	36	0,11	**1**
Total	**36**		**1**	

En este caso, no tenemos un único valor x_i para la variable bajo análisis en cada fila de la tabla, sino un intervalo de valores. En este caso, se construirá una columna auxiliar con el punto medio del intervalo (al que llamaremos x'_i) y el promedio se obtiene de la siguiente manera:

$$\bar{x} = \frac{x'_1 f_1 + x'_2 f_2 + \cdots + x'_k f_k}{n} = \frac{1}{n}\sum_{i=1}^{k} x'_i f_i$$

En este caso, el cálculo de la media aritmética es aproximado, no exacto como cuando contamos con los datos originales sin tabular.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Características de la media aritmética

- La media aritmética representa con un solo número al conjunto de datos, y su concepto es familiar, es decir, que la mayoría de las personas tiene una idea intuitiva del promedio. Es, además única para un conjunto de datos.

- Sin embargo, en el caso de una variable continua, si la distribución de frecuencias no tiene definido el extremo inferior del primer intervalo de clase y/o el extremo superior del último, es imposible obtener para dicho intervalo, su punto medio. Por lo tanto, no puede obtenerse el promedio, ya que se necesita de dicho valor.

Peso	Punto medio	Cantidad de personas
[40; 50)	45	13
[50; 60)	55	22
[60; 70)	65	35
[70 y más)	¿?	28
Total	-	98

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Características de la media aritmética

- Además, dado que en su cálculo intervienen todos los valores, se ve afectada por los valores extremos, es decir, valores muy pequeños o muy grandes de la variable. En este caso, puede calcularse la media, pero deja de ser representativa del conjunto de datos.

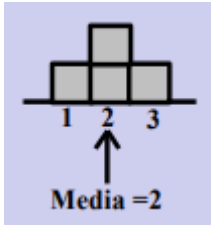
Ejemplo: se tienen las edades de 10 estudiantes: 20, 21, 20, 22, 23, 21, 20, 21, 20, 68.
La edad promedio de ese grupo es 25,6 años. Este valor, como se ve, no es representativo del conjunto.

Estadística descriptiva

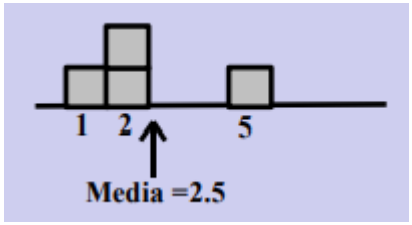
MEDIDAS DE TENDENCIA CENTRAL

Características de la media aritmética

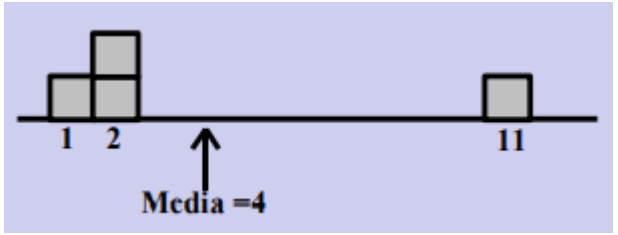
Retomando el concepto de la media como punto de equilibrio de los datos:



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



Si la observación mayor es corrida a la derecha, haciendo que ésta sea levemente extrema, la media se acerca más hacia la observación extrema.



Si una distribución es asimétrica, desearíamos además informar una medida de centralización más robusta.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Mediana

La mediana es el valor de la variable que separa al conjunto de datos ordenados en forma creciente, en dos partes iguales. De esta manera queda el 50% de los elementos que toman valores menores o iguales a la mediana y el 50% restante que toman valores mayores o iguales que la mediana.

También puede definirse de la siguiente forma: dada una serie ordenada de datos, la mediana es el valor de la variable que acumula el 50% de los mismos.

La determinación de la mediana requiere de dos pasos:

1- La determinación del orden o lugar que ocupa el valor de la mediana en el conjunto ordenado de datos (mediana de orden): $Mna^0 = \frac{n+1}{2}$

2- La determinación del valor de la mediana, es decir, del valor que se encuentra en el lugar determinado en el punto anterior.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Mediana

Si los datos están sin agrupar, se los debe ordenar de menor a mayor. Luego se calcula el orden de la mediana y finalmente se determina su valor.

Ejemplo:

a) número impar de datos: se tiene el siguiente conjunto de datos, correspondiente a las edades de 5 estudiantes: 14 4 8 9 6

Primero los ordenamos de menor a mayor: 4 6 8 9 14

Luego calculamos la mediana de orden: $Mna^0 = \frac{n+1}{2} = \frac{6}{2} = 3$

Es decir, que la mediana es el valor que ocupa el tercer lugar en la serie ordenada: $Mna = 8\ años$

Este valor se interpreta de la siguiente manera: el 50% de los alumnos tiene entre 4 y 8 años y el 50% restante tiene entre 8 y 14 años

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Mediana

Si los datos están sin agrupar, se los debe ordenar de menor a mayor. Luego se calcula el orden de la mediana y finalmente se determina su valor.

Ejemplo:
b) número par de datos: se tiene el siguiente conjunto de datos, correspondiente a las edades de 4 alumnos ordenadas de menor a mayor: 4 6 8 9

La mediana de orden: $Mna^0 = \frac{n+1}{2} = \frac{5}{2} = 2{,}5$

Es decir, que la mediana está entre el segundo y tercer lugar en la serie ordenada. En este caso, se obtiene un promedio de los valores que ocupan los lugares mencionados: $Mna = \frac{6+8}{2} = 7\ años$

O sea que, el 50% de los alumnos tiene entre 4 y 7 años y el 50% restante tiene entre 7 y 14 años.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Mediana

Cuando los datos están agrupados y la variable es discreta, se obtiene el orden de la mediana, luego se ubica dicho valor en la columna de frecuencias acumuladas y se determina el valor de la variable que lo corresponde.

Si no está el valor exacto de la Mna^0 en la columna Fi se toma el primer valor que lo supere.

Ejemplo… para ir calentando y retomar lo ya visto!

¿Cómo obtendríamos la mediana?

x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	

- En primer lugar determinamos el orden de la mediana: $Mna^0 = \frac{36+1}{2} = 18{,}5$.
- Luego se busca dicho valor en la columna Fi. En este caso se toma el valor 26, dado que es el primero que supera al 18,5, indicando que las observaciones de orden 18 y 19 presentan ambas valor 2; es decir, si se alineara a todos los empleados ordenándolos según su cantidad de inasistencias, los empleados ubicados en las posiciones 18 y 19 habrían faltado 2 veces en el segundo bimestre.
- Finalmente se toma como valor mediana al $x_i = 2$, al que le corresponde la frecuencia 26.

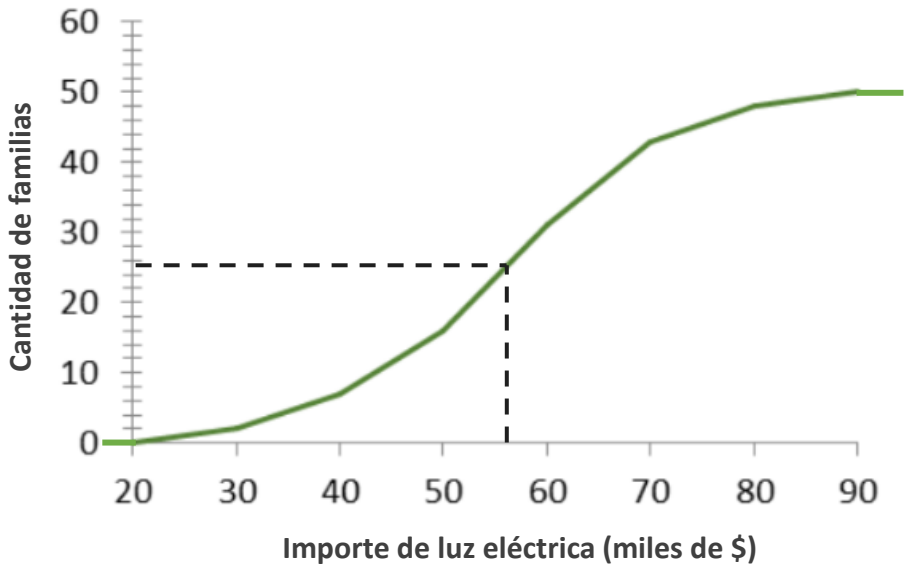
$$\boldsymbol{Mna = 2\ inasistencias}$$

El 50% de los empleados tuvo 2 inasistencias o menos en el segundo bimestre y el 50% restante tuvo 2 inasistencias o más en el mismo periodo.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Gráfico N°2. Importe total de luz abonado por familia durante el último semestre de 2023. Ciudad de Alvear.



Fuente: Encuesta de gastos de hogares (IPEC).

Cuando la variable es cuantitativa y los datos ya están agrupados en clases, la forma más práctica de obtener la mediana es mediante el gráfico de frecuencias acumuladas:

- En primer lugar determinamos el orden de la mediana: $Mna^0 = \frac{50+1}{2} = 25{,}5$.
- Luego buscamos dicho valor en el eje vertical y trazamos una línea horizontal hasta toparnos con el polígono. En esa intersección se traza una línea vertical hasta el eje X: ese será el valor de la mediana.

$$Mna \cong \$56.000$$

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Propiedades de la mediana

Ventajas de la mediana: la mediana ofrece varias ventajas sobre la media. La más importante, consiste en que los valores extremos no la afectan como a la media. La mediana es fácil de entender, y puede ser calculada con cualquier clase de datos (aún a partir de datos agrupados con clases abiertas), a menos que la mediana caiga dentro de una clase abierta.

Desventaja de la mediana: la mediana tiene también ciertas desventajas. Ciertos procedimientos estadísticos que se sirven de ella son más complejos que los que usan la media.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Moda/o

Se simboliza con Mo y es el valor de la variable que está en correspondencia con la mayor frecuencia. Es decir que, el modo es el valor de la variable que se repite más frecuentemente en el conjunto de datos.

Cuando los datos están agrupados el valor del modo puede obtenerse analítica o gráficamente de la siguiente manera:

- En forma analítica, se señala en primer lugar el mayor valor de f_i. Luego se obtiene el modo de la siguiente manera: $Mo = x_i \ tal\ que\ fi\ es\ máxima$

- En forma gráfica, en el gráfico de bastones/barras se identifica al valor x_i al cual le corresponde el bastón/barra de mayor altura.

-En el caso de las variables cuantitativas continuas hablaremos de intervalo modal y no de moda.

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Características del modo

Ventajas del Modo: a semejanza de la mediana, tampoco al modo le afectan los valores extremos. Aun cuando los valores superiores sean muy altos y los valores inferiores muy bajos, decidimos que el valor más frecuente del conjunto de datos sea el valor modal. Podemos servirnos del modo sin importar la magnitud o la dispersión de los valores en el conjunto de datos.

Desventaja del modo: a pesar de sus ventajas, la moda se usa menos para medir la tendencia central que la media y la mediana. Algunas veces no hay un valor modal porque el conjunto de datos no contiene valores que ocurran más de una vez. Otras veces todos varios valores son modo ya que ocurren el mismo número de veces. Sin duda, el modo es una medida inútil en tales situaciones.

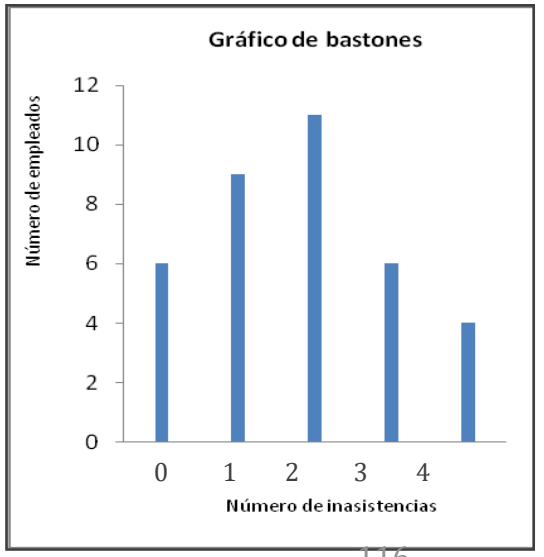
Ejemplo… para ir calentando y retomar lo ya visto!

¿Cómo obtendríamos el modo?

x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	

- En primer lugar, se determina la máxima frecuencia absoluta: 11.
- Luego se busca el valor de x que le corresponde, por lo tanto, $Mo = 2$ inasistencias; es decir que, más frecuentemente, los empleados tuvieron 2 inasistencias en el segundo bimestre del corriente año.

- Si se observa el gráfico de bastones, está claro que el modo es el 2, ya que es el valor de la variable al que le corresponde el bastón más alto.



Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Diferentes medidas pueden dar diferentes impresiones

El famoso trío -la media, la mediana, y el modo- representan tres métodos diferentes para encontrar el "valor central" de una distribución. Estos tres valores pueden ser iguales pero generalmente tomarán valores diferentes. Cuando son diferentes, van a llevar a interpretaciones distintas de los datos que están siendo resumidos.

Considere los ingresos diarios de cinco familias de un barrio:

$12.000 $12.000 $30.000 $90.000 $100.000

Cuál es el ingreso típico para este grupo?

- El ingreso medio es: $48.800
- El ingreso mediano es: $30.000
- El ingreso más frecuente (modo) es: $12.000

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Diferentes medidas pueden dar diferentes impresiones

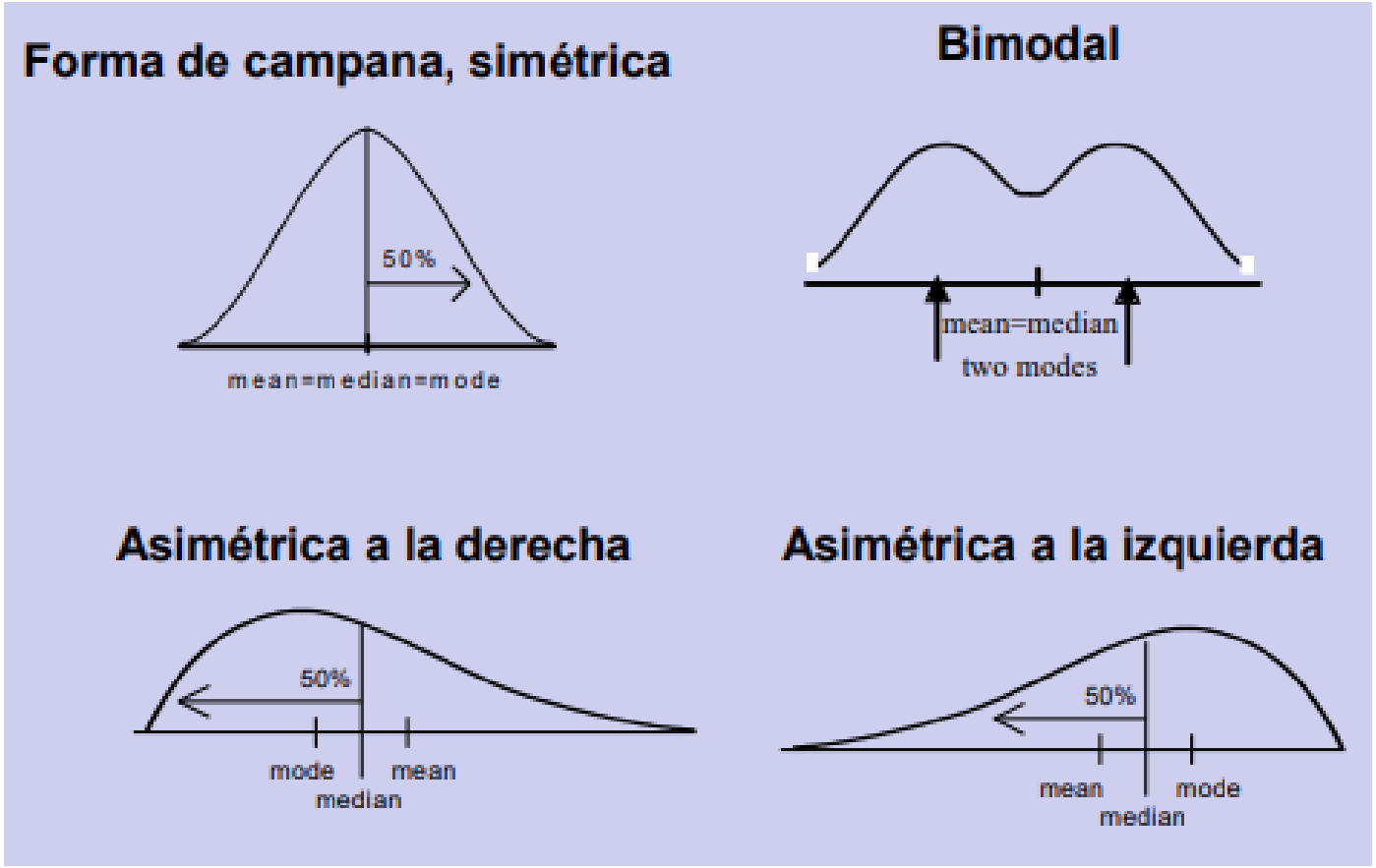
- El ingreso medio es: $48.800
- El ingreso mediano es: $30.000
- El ingreso más frecuente (modo) es: $12.000

Si estás intentado probar que este es un vecindario con un nivel económico alto, preferirás informar el ingreso medio. Si estás intentando discutir contra un aumento de los impuestos en el vecindario, podrías argumentar que los ingresos son bajos para un aumento de los impuestos e informarías el modo. Tres medidas diferentes, cada una es válida e informativa a su manera

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

¿Qué medida de tendencia central usar?



Pensemos!

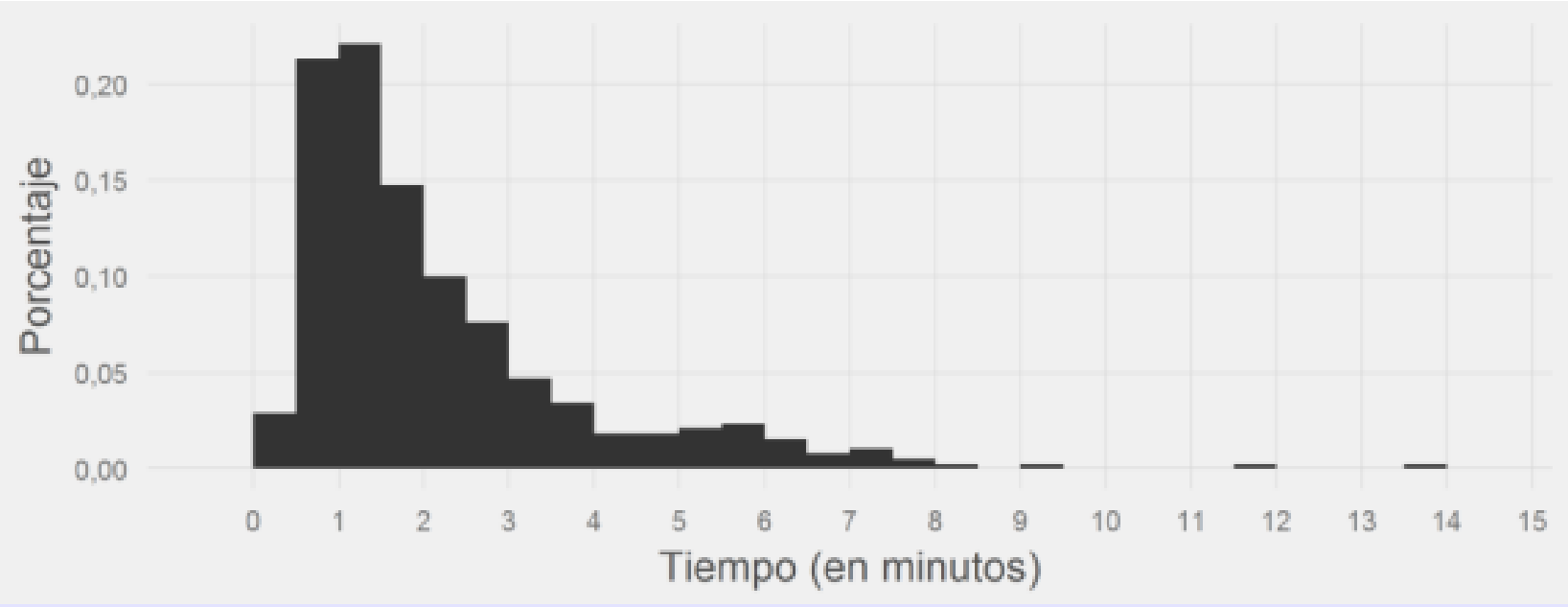
Supongamos que calculás el promedio, la mediana y el modo para una lista de números. ¿Cuál de las tres medidas debe ser siempre alguno de los números de la lista?

Si la distribución es simétrica, ¿qué medida de tendencia central calcularías? ¿Por qué?

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

Figura. Distribución del tiempo de realización de un examen de tomografía computada.



Mínimo	0,3
Máximo	13,7
Promedio	2,2
Mediana	1,6
Intervalo modal	[1,5 – 2)

¿Se puede intuir la forma de la distribución en función de las medidas resúmenes?

Estadística descriptiva

MEDIDAS DE TENDENCIA CENTRAL

A tener en cuenta

La media aritmética se ve afectada por observaciones extremas (outliers y valores que están alejados en la cola de la distribución que es asimétrica). Entonces la media parece ser una buena elección para ubicar el centro de una distribución que es unimodal y simétrica, y no tiene observaciones atípicas.

La mediana es una medida de centralización más robusta, esto es, ésta no es influenciada por los valores extremos. Para distribuciones asimétricas o distribuciones con observaciones atípicas, la mediana parece ser la mejor elección para ubicar el centro de la distribución.

El modo es el valor(es) que ocurre con mayor frecuencia. El valor más frecuente puede estar lejos del centro de la distribución. Sin embargo, el modo es la única medida de estas tres que puede ser usada para datos cualitativos.

Estadística descriptiva

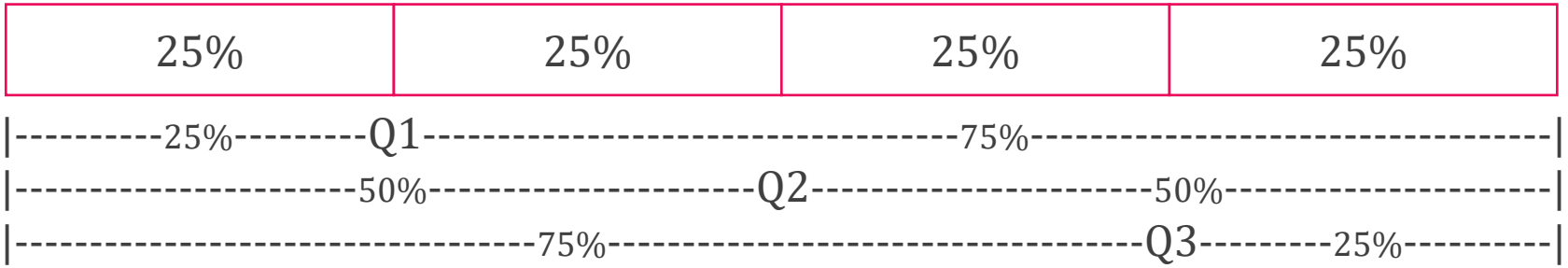
OTRAS MEDIDAS DE POSICIÓN

Cuartiles: Hay tres cuartiles que dividen al conjunto de datos en 4 partes con igual número de elementos.

- Q1: deja un cuarto de valores menores o iguales que él y tres cuartos mayores o iguales que él

- Q2: deja un medio de valores menores o iguales que él y un medio mayores o iguales que él (coincide con la mediana).

- Q3: deja tres cuartos de valores menores o iguales que él y un cuarto mayores o iguales que él.



Estadística descriptiva

OTRAS MEDIDAS DE POSICIÓN

Cuartiles

La determinación de los cuartiles se realiza de manera análoga a la de la mediana:

1. se determina el orden o lugar que ocupan en el conjunto ordenado de datos (cuartiles de orden):

$$Q_1^o = (n+1) \times \frac{1}{4} \qquad Q_2^o = Mna^0 = (n+1) \times \frac{2}{4} \qquad Q_3^o = (n+1) \times \frac{3}{4}$$

2. Se determina el valor que se encuentra en dicho lugar.

Ejemplo… para ir calentando y retomar lo ya visto!

¿Cómo obtendríamos los cuartiles?

x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	

Para Q1

- En primer lugar, se determina el orden del primer cuartil: $Q1^0 = \frac{(36+1)}{4} = 9{,}25$.
- Luego se busca dicho valor en la columna F_i. En este caso se toma el valor 15, en el cual está contenido el valor 9,25.
- Finalmente se toma como valor del primer cuartil al $x_i = 1$, al que corresponde la frecuencia 15.

$$\boldsymbol{Q1 = 1\ inasistencia}$$

El 25% de los empleados tuvo entre 0 y 1 inasistencia en el segundo bimestre y el 75% restante tuvo entre 1 y 4 inasistencias en el mismo periodo

Ejemplo… para ir calentando y retomar lo ya visto!

¿Cómo obtendríamos los cuartiles?

x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	**0,17**
1	9	15	0,25	**0,42**
2	11	26	0,30	**0,72**
3	6	32	0,17	**0,89**
4	4	36	0,11	**1**
Total	**36**		**1**	

Para Q3

- En primer lugar, se determina el orden del primer cuartil: $Q3^0 = \frac{(36+1)\times 3}{4} = 27{,}75$.
- Luego se busca dicho valor en la columna F_i. En este caso se toma el valor 32, en el cual está contenido el valor 27,75.
- Finalmente se toma como valor del tercer cuartil al $x_i = 3$, al que corresponde la frecuencia 32.

$$\boldsymbol{Q3 = 3\ inasistencias}$$

El 75% de los empleados tuvo entre 0 y 3 inasistencias en el segundo bimestre y el 25% restante tuvo entre 3 y 4 inasistencias en el mismo periodo

Estadística descriptiva

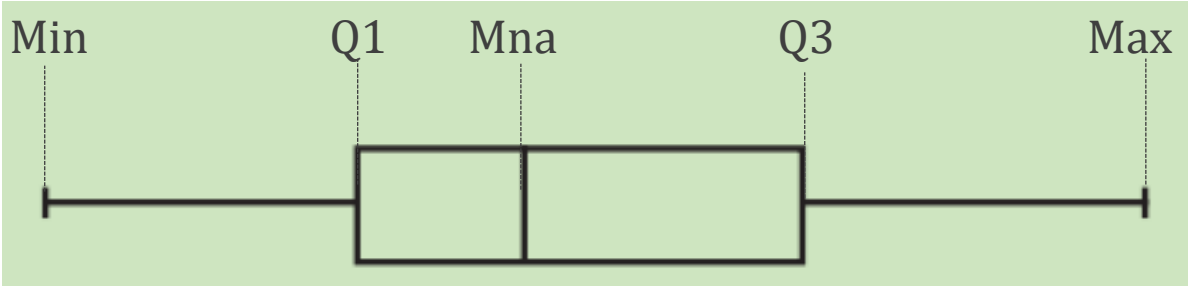
OTRAS MEDIDAS DE POSICIÓN

Diagrama de caja (Boxplot)

El mínimo, el cuartil inferior, la mediana, el cuartil superior y el máximo son cinco números. Dan una idea de cómo está distribuido un conjunto de datos. Se los llama los cinco números resumen y se los representa por:

Min Q1 Mna Q3 Max

Estos cinco números resumen se representan gráficamente en el diagrama de caja (Boxplot).

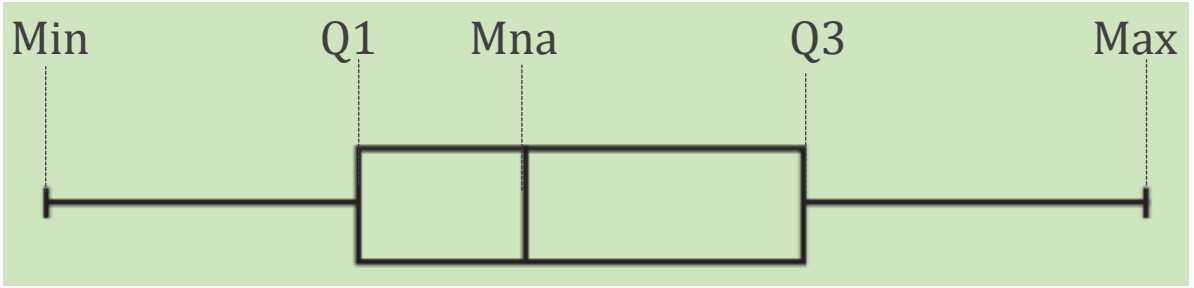


Estadística descriptiva

OTRAS MEDIDAS DE POSICIÓN

Diagrama de caja (Boxplot)

Los cuartiles forman los bordes de la caja y la mediana está dentro de la caja. Dos líneas - los brazos- se extienden, una desde cada borde de la caja, hasta el dato con valor máximo y mínimo respectivamente



|----------50%----------|

|----------50%----------|

|----------50%----------|

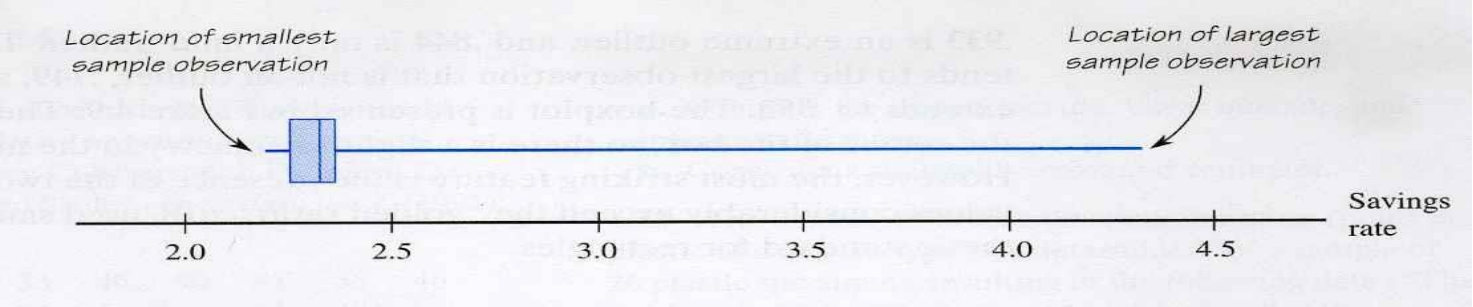
Estadística descriptiva

OTRAS MEDIDAS DE POSICIÓN

Diagrama de caja (Boxplot)

Supongamos que tenemos datos sobre la tasa de interés cobrada por una entidad financiera según el tipo de servicio prestado:

2.15	2.20	2.20	2.22	2.22	2.23	2.25	2.25
2.28	2.28	2.29	2.30	2.30	2.32	2.32	2.33
2.33	2.33	2.33	2.38	2.43	2.55	2.79	3.05
2.28	2.33	4.35	2.27	2.33	3.68		



Para construir el boxplot se necesita la siguiente información

la menor observación = 2.15

Q1 = mediana de la mitad inferior = 2.25

Mna = promedios de las observaciones 15 y 16 en la lista ordenada = 2.31

Q3 = mediana de la mitad superior = 2.33

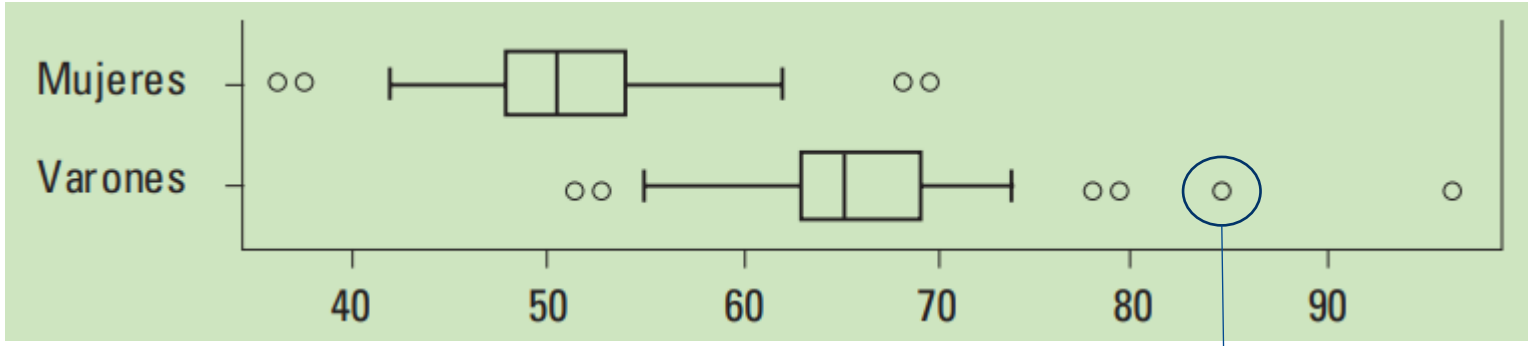
la mayor observación = 4.35

Estadística descriptiva

OTRAS MEDIDAS DE POSICIÓN

Diagrama de caja (Boxplot)

Los gráficos caja sirven especialmente cuando queremos comparar varios conjuntos de datos. Por ejemplo, si quisiéramos comparar el peso de varones y mujeres que cursan el 4to año de una escuela secundaria:



Esto lo vamos a retomar en análisis bivariado

Valores atípicos: valores que se encuentren a una distancia de 1,5 RI de los cuartiles, donde RI representa al rango intercuartílico (Q3-Q1).

¿Cuál es el diagrama de caja correcto?

Los datos presentados a continuación corresponden al número de respuestas correctas en un examen de ingreso para 50 estudiantes. Se realizó un diagrama de tallo y hoja para estos datos:

6	89
7	233
7	566
8	011234
8	56
9	12224
9	556788
10	0024
10	66678
11	23
11	55899
12	4
12	678
13	24
13	
14	1

Nota: 7|2 representa 72 respuestas correctas.

Basándote en el diagrama de tallo y hoja, cuál de los tres diagramas de caja mostrados a la derecha es el correspondiente a estos datos? Diagrama de Caja **b** Explica como realizaste tu selección.

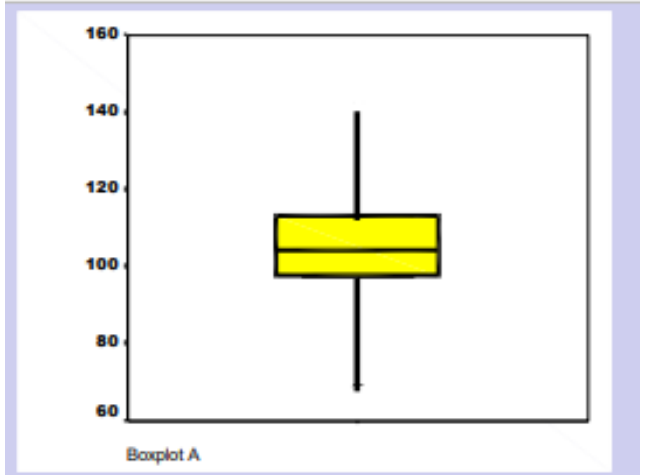
$$X_{min} = 68$$

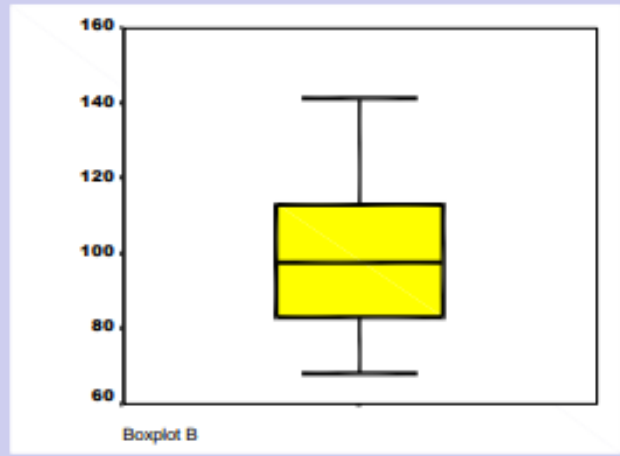
$$Q1^0 = \frac{(50+1)\times 1}{4} = 12{,}75 \rightarrow Q1 = \frac{82+83}{2} = 82{,}5$$

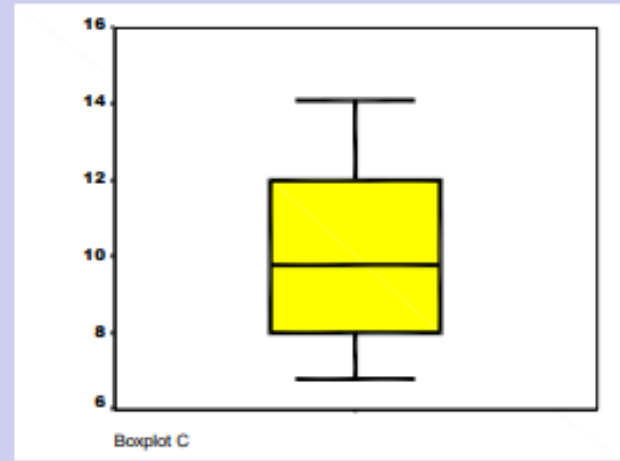
$$Q1^0 = \frac{(50+1)\times 2}{4} = 25{,}5 \rightarrow Q2 = Mna = \frac{97+98}{2} = 97{,}5$$

$$Q1^0 = \frac{(50+1)\times 3}{4} = 38{,}25 \rightarrow Q3 = \frac{113+115}{2} = 114$$

$$X_{max} = 141$$







Estadística descriptiva

OTRAS MEDIDAS DE POSICIÓN

Diagrama de caja (Boxplot) modificado: Utilizando la Regla 1.5 x RI para Identificar Observaciones Atípicas y para Construir un Diagrama de Caja Modificado

- Ordenar los datos observados de menor a mayor.
- Calcular las cinco medidas resumen: mínimo, Q1, mediana, Q3, y máximo.
- Ubicar los valores de Q1, la mediana y Q3 en la escala. Estos valores determinan la parte de la "caja" del diagrama. Los cuartiles determinan los bordes de la caja y se dibuja una línea dentro de la caja para identificar el valor de la mediana.
- Hallar el RI = Q3 – Q1.
- Calcular la cantidad 1.5 x (RI)

Estadística descriptiva

OTRAS MEDIDAS DE POSICIÓN

Diagrama de caja (Boxplot): Utilizando la Regla 1.5 x RI para Identificar Observaciones Atípicas y para Construir un Diagrama de Caja Modificado

- Hallar las barreras internas de la siguiente forma:
  - barrera interna inferior = Q1 – 1.5 x RI;
  - barrera interna superior = Q3 + 1.5 x RI.
- Dibujar dos líneas (bigotes) desde los extremos de la caja hasta los valores mínimo y máximo DENTRO de las barreras internas.
- Las observaciones que caen FUERA de las barreras internas son consideradas como potenciales outliers.

Estadística descriptiva

OTRAS MEDIDAS DE POSICIÓN

Cuantiles

Los percentiles son una generalización de la idea de la mediana: mientras que la mediana divide un conjunto de datos en dos partes iguales, los cuantiles dividen el conjunto de datos en otras proporciones.

Los cuartiles, por ejemplo, parten de la idea de dividir los datos en cuatro partes iguales. Podríamos llamarlos como $Q_1=P_{0,25}$, $Q_2=P_{0,50}$ y $Q_3=P_{0,75}$

Genéricamente, si consideramos un número α entre 0 y 1, entonces podemos definir el cuantil $(\alpha \times 100)\%$, llamado también percentil P_α, como aquel valor de la variable que divide el conjunto de datos en dos partes de forma tal que de un lado queda el $(\alpha \times 100)\%$ y del otro lado el $(1-\alpha) \times 100\%$.

Si α toma los valores 0,10, 0,20, etc. se llamarán deciles. Si toma los valores 0,20, 0,40, 0,60, etc. entonces serán quintiles. Hablamos de percentiles cuando α toma los valores 0,01, 0,02, 0,03, etc., partiendo el conjunto de datos en 100 partes iguales.

Estadística descriptiva

OTRAS MEDIDAS DE POSICIÓN

Percentiles

Percentiles

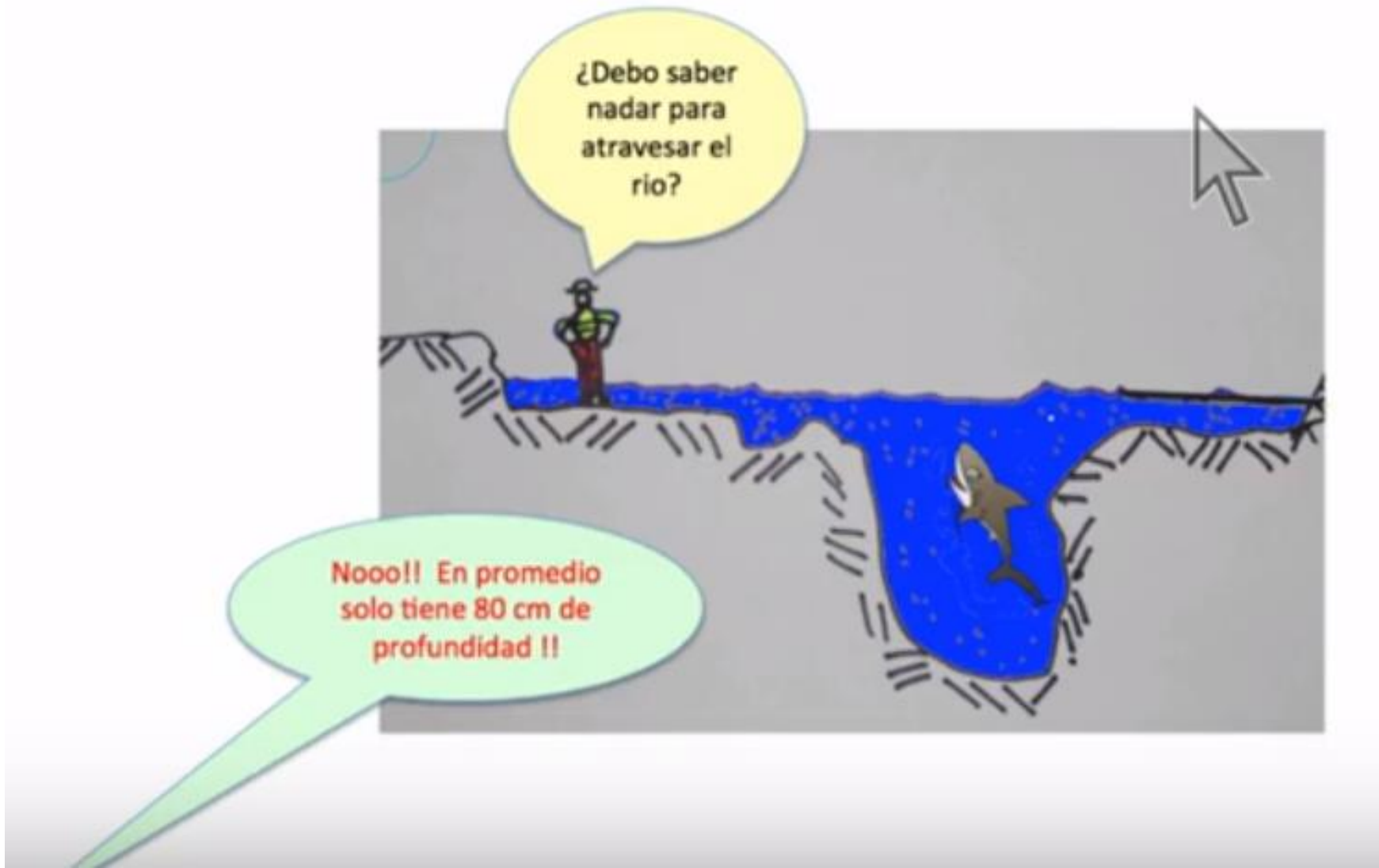
El percentil P_α representa aquel valor de la variable tal que el α% de los datos resulta menor o igual a él, dejando un $(1 - \alpha)$% de datos mayor o igual a él.

Casos particulares: quintiles, deciles, cuartiles.

P_α	Valor	Interpretación
P_1	100	El 1% de los encuestados percibe ingresos mensuales de $100 o menos
P_{10}	600	El 10%, percibe $600 o menos al mes
P_{25}	1.335	(Q_1: cuartil 1) El 25% de los encuestados registra ingresos mensuales de $1.335 o menos.
P_{50}	2.400	(Q_2: mediana) El 50% de los encuestados registra ingresos mensuales de $2.400 o menos.
…	…	…

Estadística descriptiva

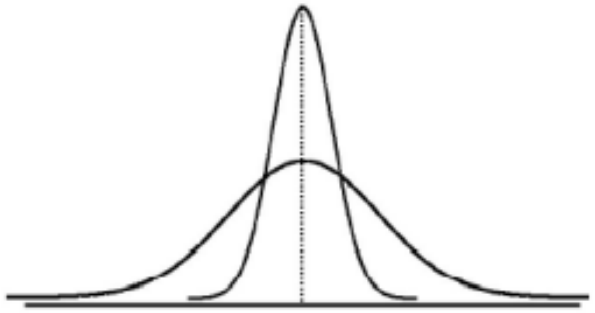
MEDIDAS DE DISPERSIÓN



Estadística descriptiva
MEDIDAS DE DISPERSIÓN

Las medidas de tendencia central nos dan una idea de dónde se concentran los datos bajo análisis. Sin embargo, dos conjuntos pueden tener valores muy similares en las medidas de tendencia central, por ejemplo misma media aritmética, y sin embargo provenir de poblaciones con distinta función de densidad, concentrándose de formas diversas alrededor de la media.

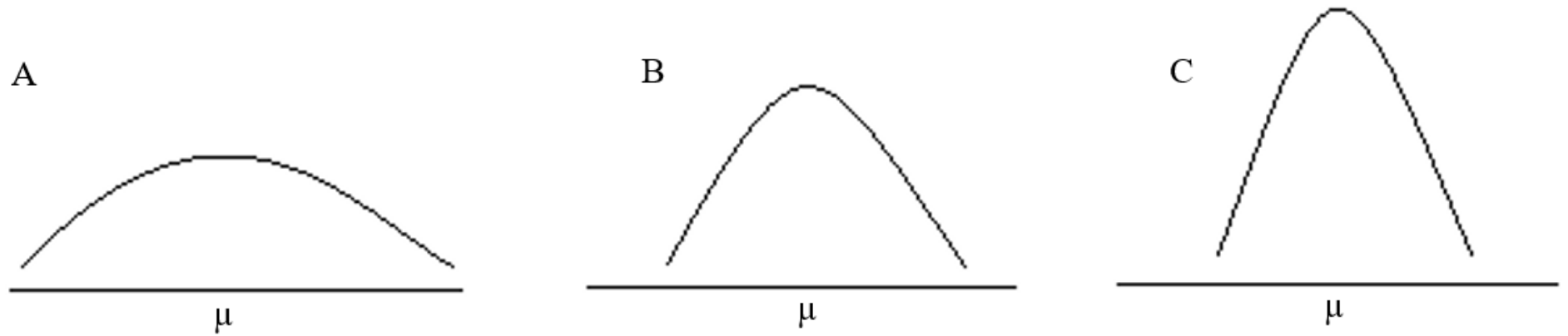
En estos casos, las medidas de tendencia central no son suficientes para caracterizar los datos: la media, la mediana, el modo y los cuartiles nos indican sólo parte de lo que necesitamos conocer en torno a las características de los datos. La concentración o dispersión de las observaciones alrededor de algún valor en particular es otra propiedad que caracteriza los datos y su distribución.



Estadística descriptiva

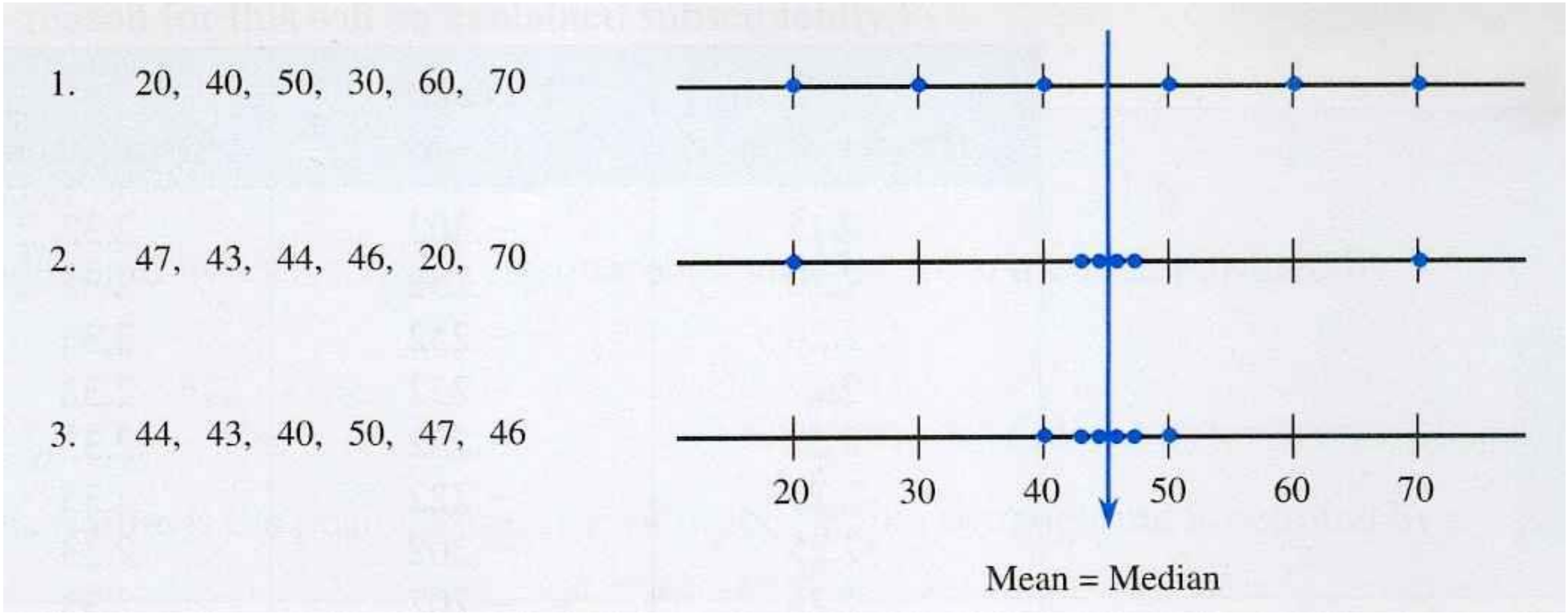
MEDIDAS DE DISPERSIÓN

¿Por qué la dispersión de la distribución es una característica tan importante de entender y medir? Nos suministra información complementaria que nos permite juzgar la confiabilidad de nuestra medida de tendencia central. Si los datos están ampliamente dispersos, como los de la curva A , la localización central será menos representativa de los datos en su conjunto de lo que sería en el caso de datos que se acumulasen más alrededor de la media, como sucede en la curva C.



Estadística descriptiva
MEDIDAS DE DISPERSIÓN

Miremos sino este ejemplo más extremo donde los tres conjuntos de datos tienen exactamente la misma media y mediana, pero distintos patrones de variabilidad:



Estadística descriptiva
MEDIDAS DE DISPERSIÓN

Rango

Se simboliza con R, y es la diferencia entre el mayor y menor valor del conjunto de datos.

$$R = x_{máx} - x_{mín}$$

Propiedades: El rango es fácil de entender y calcular, pero es escasa su utilidad como medida de dispersión. El rango incluye únicamente los valores máximo y mínimo de una distribución, sin tener en cuenta ninguna otra observación dentro del conjunto de datos. De ahí que ignore la naturaleza de la variabilidad entre todas las demás observaciones, siendo afectado profundamente por los valores extremos. Las distribuciones con clases abiertas no tienen rango, dado que en las clases abiertas no existe valor máximo y/o mínimo.

Estadística descriptiva

MEDIDAS DE DISPERSIÓN

Rango

El rango puede dar una idea distorsionada del verdadero patrón de variabilidad.

Dos distribuciones: el mismo rango pero diferentes patrones de dispersión. La primer distribución tiene la mayoría de sus valores alejados del centro, mientras que la segunda tiene la mayoría de sus valores cercanos al centro.

Ejemplo… para ir calentando y retomar lo ya visto!

¿Cómo calcularíamos el rango?

x_i: número de inasistencias en el 2do. bimestre	f_i	Fi	h_i	H_i
0	6	6	6/36=0,17	0,17
1	9	15	0,25	0,42
2	11	26	0,30	0,72
3	6	32	0,17	0,89
4	4	36	0,11	1
Total	**36**		**1**	

$R = x_{máx} - x_{mín} = 4 - 0 = 4$ inasistencias.

La máxima diferencia que existe entre quien más inasistencias tuvo y quien tuvo menos es de 4 inasistencias.

Observación: Para variables continuas agrupadas en clases no es posible conocer exactamente el rango. Puede obtenerse de forma aproximada utilizando el límite inferior del menor intervalo y el límite superior del último intervalo. En caso de contar con los datos originales sin agrupar, situación óptima, sí será posible encontrarlo.

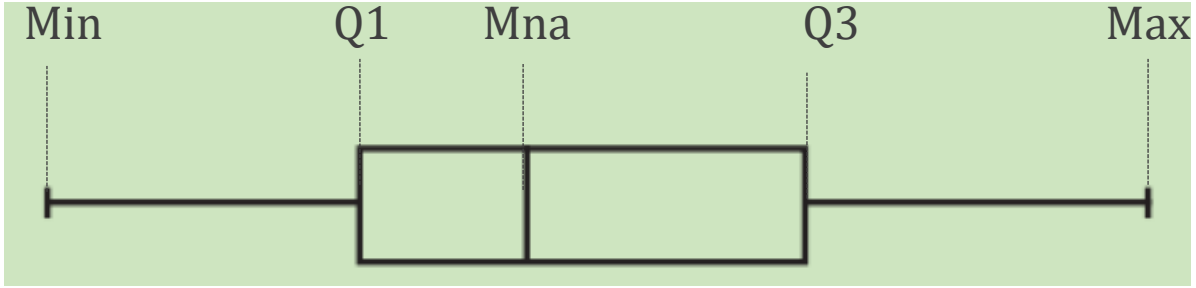
Estadística descriptiva
MEDIDAS DE DISPERSIÓN

Rango intercuartílico

Se simboliza con RI, y mide la dispersión del 50% de los valores centrales.

$$RI = Q3 - Q1$$

Es la medida de dispersión que suele acompañar a la mediana. Al igual que esta, es una medida robusta, no influenciada por valores atípicos.



|----------RI----------|

Estadística descriptiva
MEDIDAS DE DISPERSIÓN

Variancia y desvío estándar

Cuando la media aritmética ha sido elegida como la medida de localización de un conjunto de valores, una medida para la variabilidad deberá medir el grado en que los valores se alejan o desvían de su media. Si pensamos en n valores de la variable X, siendo $\bar{x}$ su promedio tendremos n desvíos de tipo $(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$, donde los valores de X mayores que $\bar{x}$ producirán desvíos positivos y los menores, desvíos negativos, resultando:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

Estadística descriptiva

MEDIDAS DE DISPERSIÓN

Variancia y desvío estándar

Dado que interesa la magnitud de los desvíos, deberían usarse los valores absolutos, aunque por facilidad se acostumbra tomar los cuadrados de las desviaciones y promediarlos. En esta forma, una medida de la variación de un conjunto de valores alrededor de la media del conjunto, está dada por la variancia, que toma distintos nombres y fórmulas según se esté analizando una muestra o la población completa.

La variancia poblacional se simboliza con σ^2, y se obtiene como:

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

La variancia muestral se simboliza con S^2, y se obtiene como:

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

Estadística descriptiva

MEDIDAS DE DISPERSIÓN

Variancia y desvío estándar

Dada la dificultad práctica del cálculo, de la fórmula mencionada se derivó una fórmula de trabajo, que conduce al mismo resultado de manera más simple cuando los datos se encuentran agrupados, es decir, acomodados en una tabla de distribución de frecuencias.

$$s^2 = \frac{\sum(x_i - \bar{x})^2 f_i}{n-1} = \frac{\sum({x_i}^2 f_i - 2\, x_i \bar{x} f_i + \bar{x}^2) f_i}{n-1} = \frac{\sum {x_i}^2 f_i - 2\,\bar{x} \sum x_i\, f_i + \bar{x}^2 \sum f_i}{n-1} =$$

$$S^2 = \frac{\sum {x_i}^2 f_i - 2\,\bar{x} n + \bar{x}^2 n}{n-1} = \frac{\sum {x_i}^2 f_i - \, n\,\bar{x}^2}{n-1}$$

Estadística descriptiva
MEDIDAS DE DISPERSIÓN

Variancia y desvío estándar

El desvío estándar se simboliza con S y se define como la raíz cuadrada de la variancia.

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

A diferencia de la variancia, la desviación estándar se da en unidades que son las mismas que los datos originales, e indica cuánto se alejan en promedio los datos del conjunto respecto a su media aritmética.

Si pensamos a la desviación estándar como una distancia promedio de las observaciones respecto de la media, entonces la desviación estándar será 0 si todas las observaciones son iguales. Además la desviación estándar es positiva y cuanto más dispersas están las observaciones respecto del promedio mayor es el valor de la desviación estándar.

Estadística descriptiva

MEDIDAS DE DISPERSIÓN

Variancia y desvío estándar

La variancia está medida en unidades al cuadrado. Calculando la raíz cuadrada de la variancia llevamos esta medida de dispersión a las unidades originales.

Al igual que la media, la desviación estándar está fuertemente influenciada por las observaciones extremas. Unas pocas observaciones atípicas puede hacer que S sea muy grande.

Estadística descriptiva
MEDIDAS DE DISPERSIÓN

¿Cuál elegir?

El rango intercuartílico, RI, es la distancia entre el primer y tercer cuartil (Q3 - Q1), y mide la dispersión del 50% de los datos centrales. Cuando la mediana es usada como una medida de centralización, el RI es utilizado a menudo como una medida de dispersión. Para distribuciones asimétricas, o distribuciones con observaciones atípicas, el RI puede ser una mejor medida de dispersión si el objetivo es resumir la distribución.

La desviación estándar es la distancia promedio de los valores observados a la media. La media y la desviación estándar son más útiles para distribuciones aproximadamente simétricas y que no presentan observaciones atípicas.

Estadística descriptiva

MEDIDAS DE DISPERSIÓN

Las medidas numéricas presentadas en esta unidad brindan información sobre el centro y la dispersión de la distribución, pero un gráfico, tal como un histograma o un diagrama de tallo y hoja, proporciona el mejor cuadro de la forma de la distribución.

Primero grafica tus datos!

Estadística descriptiva
MEDIDAS DE DISPERSIÓN

Coeficiente de variación

En algunas ocasiones puede ser de interés comparar las características de dos conjuntos de datos. Si estas variables fueron medidas en distintas escalas y/o unidades de medida, dicha comparación será compleja.

Imaginemos que queremos saber si el peso de una población de hormigas es más homogéneo o heterogéneo que el de una población de elefantes. Si podemos suponer que ambas poblaciones tienen distribuciones relativamente simétricas, entonces la media aritmética y el desvío estándar serían medidas resumen apropiadas para datos de este tipo.

Estadística descriptiva
MEDIDAS DE DISPERSIÓN

Coeficiente de variación

<u>Hormigas</u>

Peso promedio: $\bar{x} = 3g = 0{,}003\ kg$

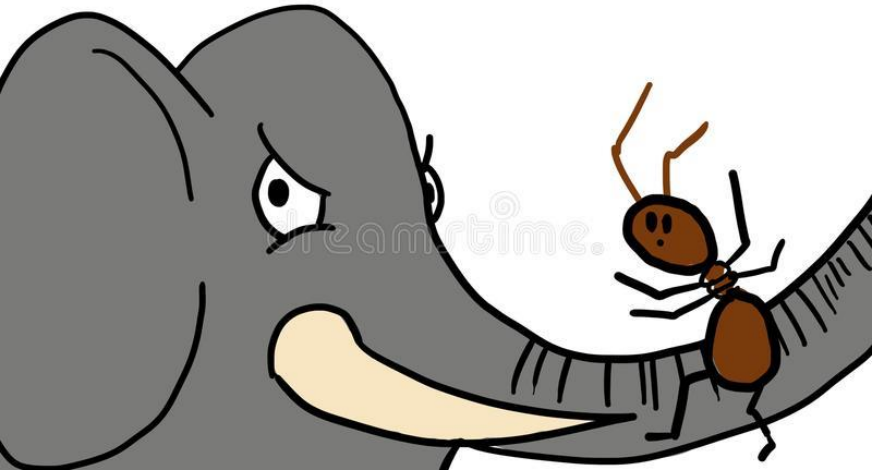
Desviación estándar: $S = 1{,}5\ g = 0{,}0015\ kg$

<u>Elefantes</u>

Peso promedio: $\bar{x} = 4200\ kg$

Desviación estándar: $S = 720\ kg$

Claramente las magnitudes son incomparables!



El coeficiente de variación es una medida de dispersión relativa: especifica cuánto representa la variabilidad de los datos (desvío estándar) en función de la media aritmética, expresado en porcentaje.

$$CV = \frac{S}{\bar{x}} \times 100$$

Estadística descriptiva

MEDIDAS DE DISPERSIÓN

Coeficiente de variación

Hormigas

Peso promedio: $\bar{x} = 3g = 0{,}003\ kg$

Desviación estándar: $S = 1{,}5\ g = 0{,}0015\ kg$

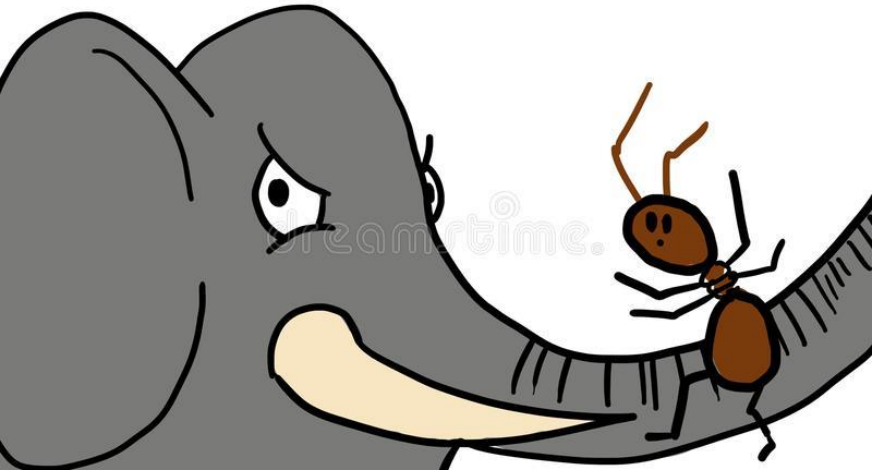
$$CV_{hormiga} = \frac{0{,}0015}{0{,}003} \times 100 = 50\%$$

Elefantes

Peso promedio: $\bar{x} = 4200\ kg$

Desviación estándar: $S = 720\ kg$

$$CV_{elefante} = \frac{720}{4200} \times 100 = 17\%$$



Mientras que la dispersión promedio del peso de las hormigas representa el 50% de su media aritmética, en el caso de los elefantes el desvío estándar representa apenas el 17%. En este sentido, puede afirmarse que el peso de los elefantes es mas homogéneo que el de las hormigas.

IMPORTANTE: El CV solo puede calcularse para variables medidas en escala de razón, donde exista un cero absoluto que indique ausencia del atributo y tenga sentido realizar cocientes entre valores.

Variable cuantitativa

Análisis descriptivo numérico

Promedio	+	Desvío std	Para variables cuantitativas con distribución simétrica.
Mediana	+	Rango intercuartílico	Para variables cuantitativas con distribución asimétrica o presencia de valores atípicos.
Moda			Es menos informativa que el promedio o la mediana. No suele usarse a menos que sea una tendencia muy marcada hacia un valor (o unos pocos valores).
Percentiles			Suelen utilizarse como complemento de otras medidas si los datos lo ameritan.
Rango			Para variables de escala de intervalo o de razón, como complemento de otras medidas para completar la descripción.

Variable cualitativa

Análisis descriptivo numérico

Promedio	+	Desvío std	No es posible su cálculo, requiere al menos escala de intervalos.
Mediana	+	Rango intercuartílico	La mediana puede calcularse para variables medidas en escala ordinal, dado que solo requiere valores ordenados. El RI no es aplicable (ver Rango)
Moda			La medida más usada para variables cualitativas, dado que es la única medida de posición que admiten variables de escala nominal.
Percentiles			Siguiendo la lógica de la mediana, pueden ser calculados para variables medidas en escala ordinal dado que solo requieren datos ordenados. Pueden perder sentido si son pocas categorías.
Rango			No es posible su cálculo, requiere al menos escala de intervalos ya que se basa en la distancia entre dos valores (X_{min} y X_{max}, o Q1 y Q3 en el caso del rango intercuartílico).

VARIABLE CUANTITATIVA

ANÁLISIS DESCRIPTIVO NUMÉRICO

Media aritmética y desvío estándar

Mínimo y máximo

Opciones alternativas para interpretar

- Los individuos encuestados tenían entre 18 años y 59 años. La edad promedio fue de 33,8 años con un desvío estándar de 11,2 años.

n

- La muestra quedó constituida por 66 personas. La edad promedio del grupo fue de 33,8 años (DE 11,2 años), siendo la edad mínima observada de 18 años y la máxima de 59 años. Si se seleccionan al azar dos encuestados, la máxima diferencia de edad posible de observar es de 49 años.

Rango

- La edad de los encuestados varía entre los 18 y los 59 años. El 25% de los encuestados tenía 24 años o menos mientras que otro 25% tenía 40 años o más. La edad mediana es de 29 años, y la dispersión del 50% central de los datos se alza a 16 años.

Q1

Q3

Mediana y RI

Tabla 4. Medidas resumen de la edad de los encuestados

Tamaño muestral	66 personas
Promedio	33,8 años
Desvio estándar	11,2 años
Mínimo	18 años
1er cuartil	24 años
Mediana	29 años
3er cuartil	40 años
Máximo	59 años

Análisis descriptivo de datos
Análisis bivariado

Estadística descriptiva

ANÁLISIS BIVARIADO

Muchos estudios estadísticos se llevan a cabo para detectar relaciones entre dos variables. Hasta ahora vimos técnicas para estudiar una sola variable mostrándola gráficamente o numéricamente. Ahora discutiremos métodos para descubrir y describir relaciones entre dos o más variables.

Los datos utilizados para estudiar la relación entre dos variables son llamados datos bivariados. Los datos bivariados se obtienen por la medición de dos variables sobre el mismo individuo u objeto. Si se registra el puntaje obtenido en el examen de mediados de año y el obtenido en el examen final para una muestra de estudiantes, estos datos nos ayudarán a estudiar la asociación entre las dos variables. Así, estaremos en condiciones de ver si cierto puntaje obtenido en el examen de mediados de año tiende a ocurrir más a menudo con cierto puntaje del examen final que con otros puntajes del examen final.

Estadística descriptiva

ANÁLISIS BIVARIADO

Comencemos con un ejemplo…

En una ciudad con graves problemas de obesidad en la población, se solicitó a un grupo de 60 adolescentes que registrara durante un mes la cantidad de horas que dedicaban cada día a actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) y las promediaran. La Tabla 1 presenta la edad en años (Edad), el género (Varón, Mujer), el promedio de horas por día dedicadas a actividades sedentarias (Horas), el tipo de domicilio en el que vive(Casa, Departamento) y un número (Id) para identificar a cada participante:

Estadística descriptiva

ANÁLISIS BIVARIADO

Tabla 1. Promedio de horas dedicadas a actividades sedentarias

id	edad	horas	genero	vivienda
1	11,2	5,5	Varón	Departamento
2	11,4	5,4	Varón	Casa
3	11,4	4,5	Varón	Casa
4	11,5	4,8	Varón	Casa
5	11,6	5	Varón	Casa
6	11,7	5,5	Varón	Departamento
7	11,9	4,3	Varón	Casa
8	12,6	5,7	Varón	Casa
9	12,8	4,7	Varón	Casa
10	13,2	5,4	Varón	Casa
11	13,8	5,6	Varón	Casa
12	13,8	5,5	Varón	Casa
13	14	6,6	Varón	Casa
14	14,3	5,5	Varón	Casa
15	14,5	5,4	Varón	Departamento
16	14,6	5,3	Varón	Departamento
17	15	5,2	Varón	Casa
18	15,4	7	Varón	Departamento
19	15,6	5,9	Varón	Departamento
20	15,9	6,6	Varón	Departamento
21	16,2	6,3	Varón	Departamento
22	16,5	5,8	Varón	Casa
23	17	6,9	Varón	Casa
24	17,3	6,9	Varón	Departamento
25	17,4	6,2	Varón	Departamento
26	17,5	5,5	Varón	Departamento
27	17,8	6	Varón	Casa
28	17,9	6,5	Varón	Departamento
29	18,2	6,4	Varón	Departamento
30	18,3	5,7	Varón	Departamento

id	edad	horas	genero	vivienda
31	11,1	4,3	Mujer	Departamento
32	11,2	5,1	Mujer	Casa
33	11,2	4,7	Mujer	Casa
34	11,5	4,5	Mujer	Casa
35	11,6	4,7	Mujer	Casa
36	11,6	4,8	Mujer	Casa
37	11,8	4,4	Mujer	Casa
38	11,9	4,7	Mujer	Casa
39	12,3	5	Mujer	Departamento
40	12,8	4,7	Mujer	Casa
41	12,8	5,1	Mujer	Departamento
42	12,9	5,2	Mujer	Departamento
43	13,1	5,8	Mujer	Departamento
44	13,5	5,2	Mujer	Departamento
45	13,6	5,1	Mujer	Casa
46	13,9	5	Mujer	Casa
47	14,2	4,4	Mujer	Departamento
48	14,4	5,6	Mujer	Casa
49	14,9	4,4	Mujer	Casa
50	15,1	5,2	Mujer	Departamento
51	15,4	5,1	Mujer	Departamento
52	15,6	5,1	Mujer	Departamento
53	15,9	5,3	Mujer	Casa
54	16,2	4,7	Mujer	Departamento
55	16,4	4,9	Mujer	Casa
56	16,6	6,7	Mujer	Departamento
57	17,2	5	Mujer	Casa
58	17,4	5,8	Mujer	Casa
59	17,9	5,6	Mujer	Casa
60	18,1	5,8	Mujer	Casa

Estadística descriptiva

ANÁLISIS BIVARIADO

¿Qué podría interesarnos analizar en estos datos?

- Composición de la muestra según género, edad, horas que pasa en actividades sedentarias o tipo de vivienda.
- Relación entre la cantidad de horas en actividades sedentarias y la edad de lxs adolescentes (si los más jóvenes tienden a pasar más horas en actividades de este tipo que los más grandes o viceversa).
- Relación entre la cantidad de horas en actividades sedentarias y el género (si las mujeres tienden a pasar más horas en actividades de este tipo que los varones o viceversa).
- Relación entre el género y el tipo de vivienda (si varones y mujeres viven por igual en casas o departamentos o si hay una tendencia a que adolescentes de un género vivan en un tipo de vivienda).

Estadística descriptiva

ANÁLISIS BIVARIADO

Análisis gráfico bivariado:

Si las dos variables son cuantitativas	DIAGRAMA DE DISPERSIÓN
Si una variable es cuantitativa y la otra es cualitativa: BOXPLOT COMPARATIVOS DE LA VARIABLE CUANTITATIVA PARA CADA NIVEL DE LA VARIABLE CUALITATIVA	
Si las dos variables con cualitativas	TABLA DE CONTINGENCIA GRÁFICOS DE BARRAS SUBDIVIDIDAS GRÁFICOS DE BARRAS AGRUPADAS

Estadística descriptiva

ANÁLISIS BIVARIADO

Diagrama de dispersión

La forma gráfica más habitual de describir la relación entre dos variables cuantitativas es utilizando un diagrama de dispersión. Cada punto corresponde a un par de valores (uno para cada variable), medidos sobre el mismo individuo.

En general, si una de las variables puede pensarse como explicativa de la otra (variable explicativa), siempre se la grafica en el eje horizontal (eje x) y la otra (variable respuesta) en el eje vertical (eje y).

Estadística descriptiva

ANÁLISIS BIVARIADO

En el ejemplo de la figura, la edad es la variable explicativa. Pensamos que la edad puede explicar, aunque sea en parte, la cantidad de horas diarias dedicadas a actividades sedentarias (variable respuesta, graficada en el eje y). Cada punto representa a un varón. Está determinado por su edad y la cantidad de horas diarias dedicadas a las actividades sedentarias relevadas (mirar televisión, estudiar o utilizar la computadora). Para ilustrar como se realiza el gráfico, se destaca el punto correspondiente a Id = 13, Edad =14, Horas = 6,6.

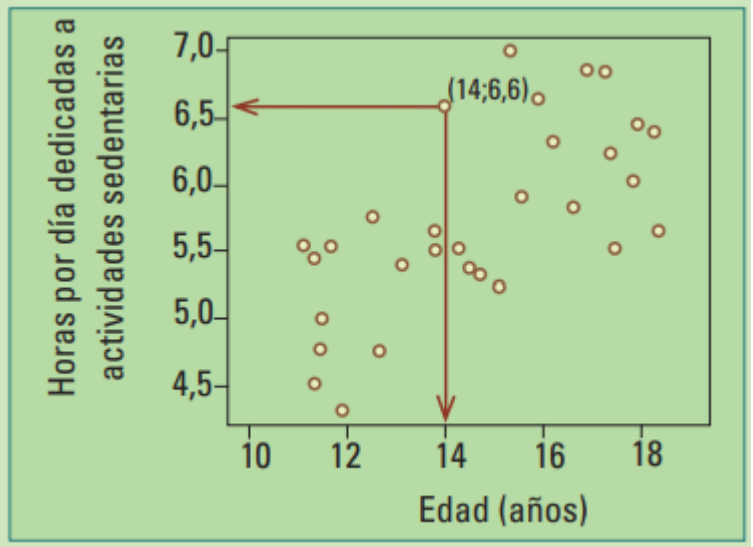


Diagrama de dispersión de las horas en función de la edad para los varones. Se destaca el punto correspondiente a Id = 13, Edad =14, Horas = 6,6.

Estadística descriptiva

ANÁLISIS BIVARIADO

En un diagrama de dispersión observamos el patrón general de la relación entre las variables mirándolo de izquierda a derecha. Si a medida que x aumenta (es decir, nos corremos hacia la derecha del gráfico):

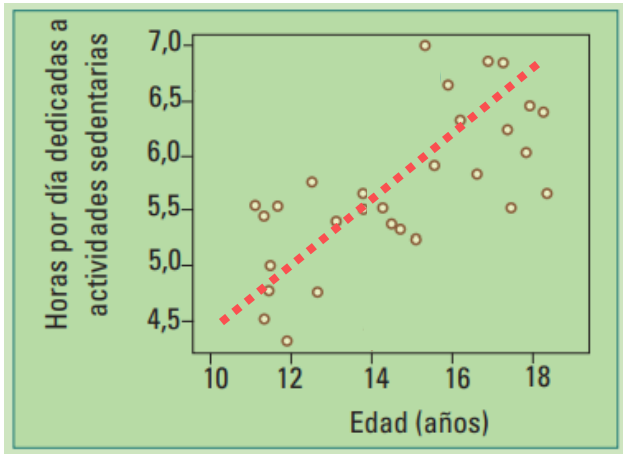
• en promedio también lo hace y (los valores de y se encuentran más arriba); esto indica una **asociación lineal positiva** entre las variables.

• en promedio y decrece (los valores de y se encuentran más abajo, esto indica una **asociación lineal negativa** entre las variables.

• no puede determinarse una tendencia de crecimiento o decrecimiento en los valores de y; esto significa que **no hay una asociación lineal entre las variables**.

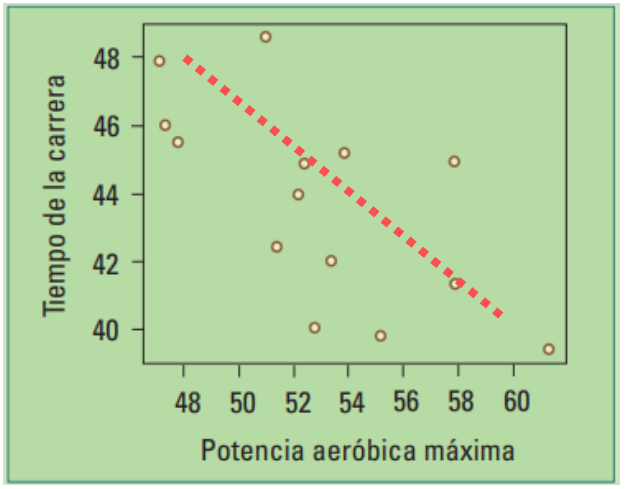
Estadística descriptiva

ANÁLISIS BIVARIADO

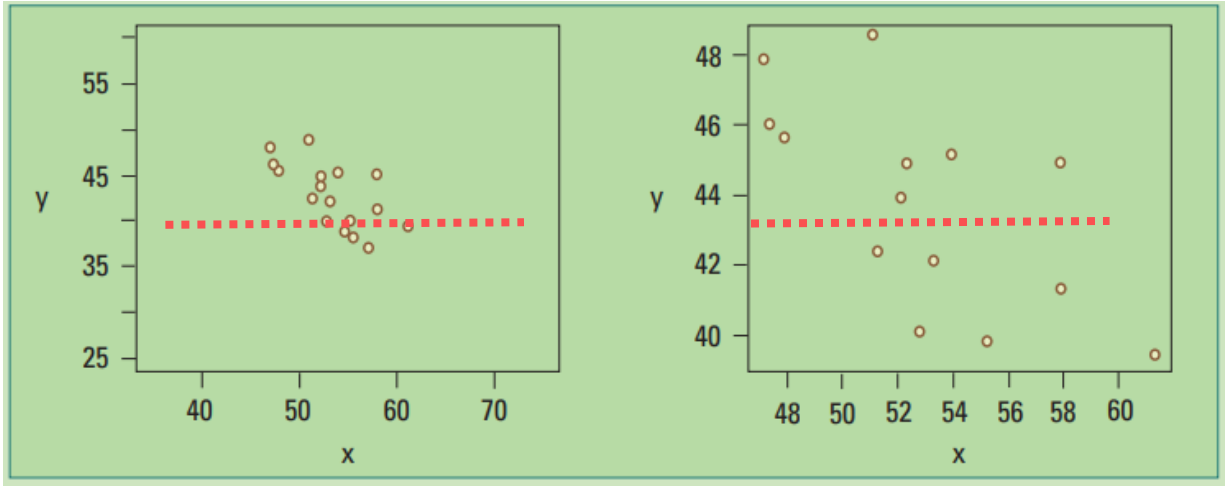


Correlación lineal positiva

Correlación lineal negativa



Sin correlación lineal

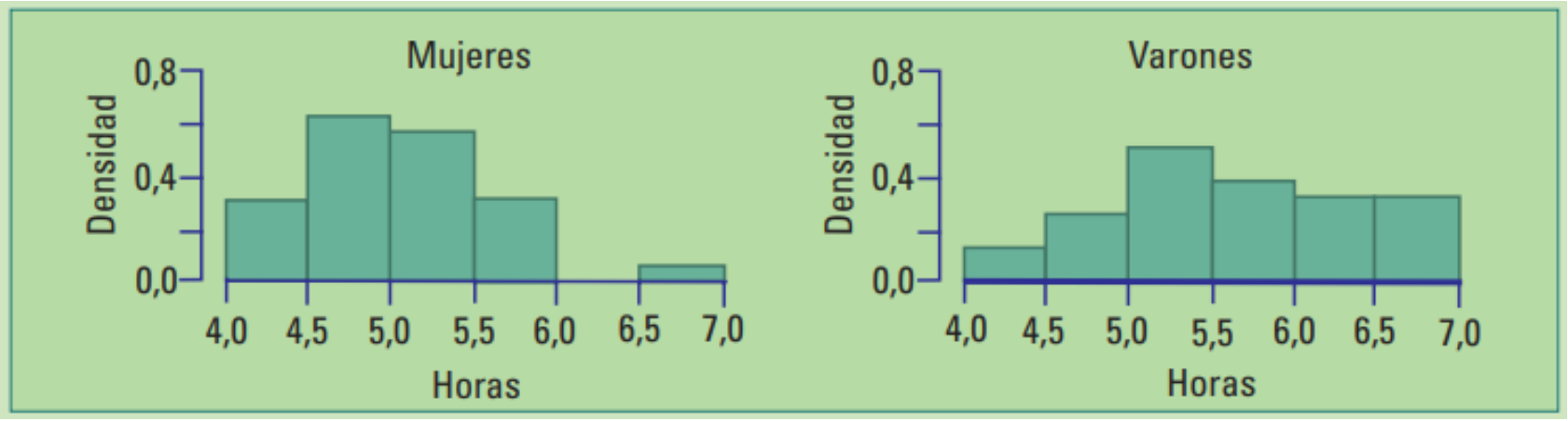


Estadística descriptiva

ANÁLISIS BIVARIADO

Boxplot comparativo

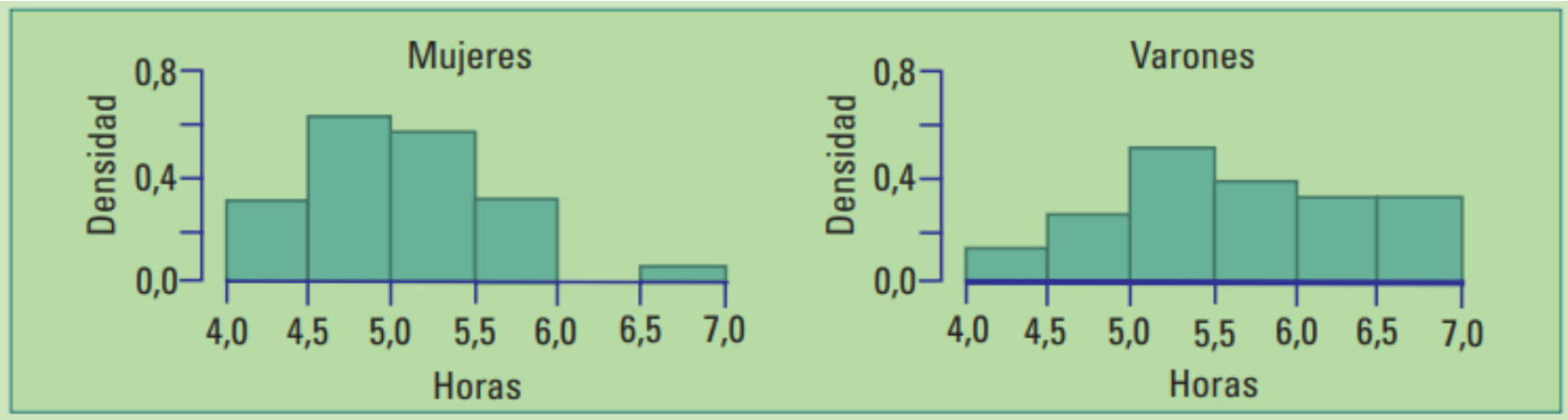
La forma gráfica más habitual de describir la relación entre dos variables cuantitativas es utilizando diagramas de caja comparativos. Si bien podríamos representar las distribuciones de una variable (por ejemplo, horas de actividades sedentarias) para los distintos grupos que interesa comparar (mujeres y varones), el análisis no sería tan valioso:



Estadística descriptiva

ANÁLISIS BIVARIADO

Para las mujeres la media muestral de la cantidad de horas por día dedicadas a actividades sedentarias es 5,06. En los varones es 5,73 horas, aproximadamente 3/4 de hora más. La figura refuerza esta situación. Los intervalos correspondientes a la mayor cantidad de horas (de 6 a 7) presentan mayor densidad de datos para los varones que para las mujeres. El intervalo más poblado para las mujeres es entre 4,5 y 5,0 horas y en los varones entre 5 y 5,5 horas.



Histogramas de las horas por día dedicadas a actividades sedentarias de mujeres y varones

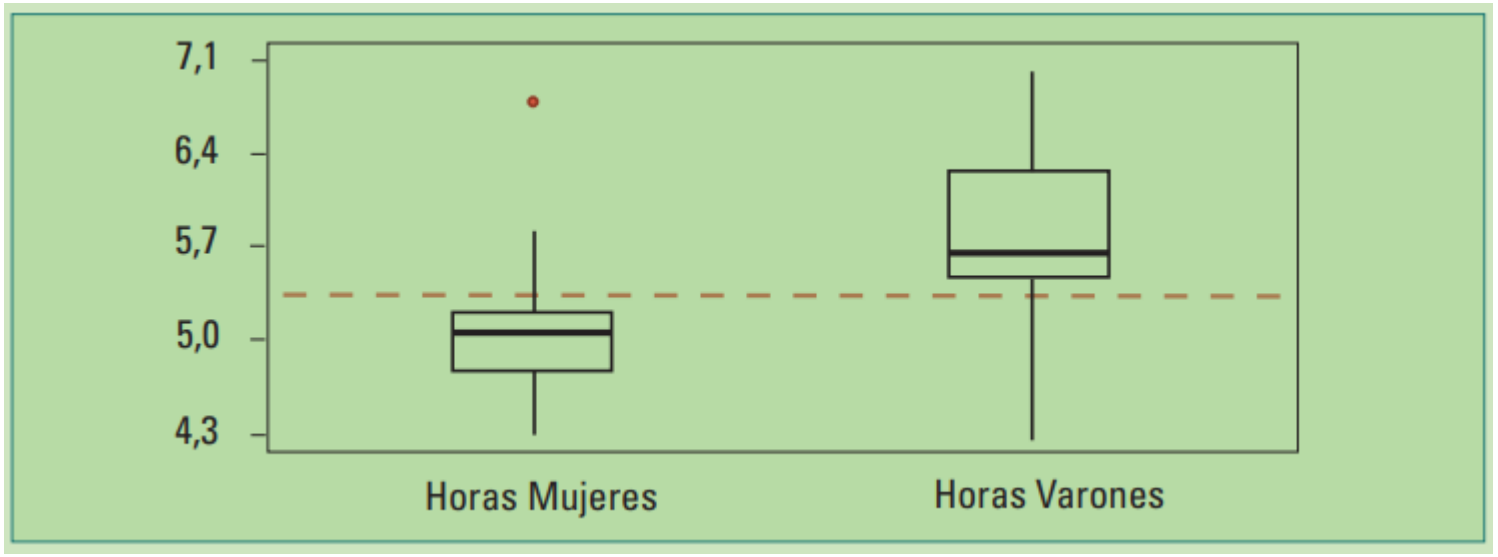
Estadística descriptiva

ANÁLISIS BIVARIADO

Por este motivo se construyen los gráficos de caja y bigote comparativos, donde se grafican tantos boxplot como grupos interese comparar.

El gráfico caja de la figura para mujeres muestra algo no percibido en el análisis de los hisogramas. un valor atípico; se trata de un valor muy alejado del resto. Se destaca también que la caja correspondiente a los varones se encuentra desplazada hacia arriba (hacia los valores mayores) en comparación con las mujeres; por lo tanto,

más del 75% de los valores de horas para los varones son mayores que más del 75% de los valores menores de las horas para mujeres.



Estadística descriptiva

ANÁLISIS BIVARIADO

Tablas de contingencia

Las tablas de contingencia muestran la distribución conjunta de los individuos bajo análisis de acuerdo a dos variables simultáneamente:

Vivienda / Género	Casa	Departamento	Total
Mujer	18	12	**30**
Varón	16	14	**30**
Total	**34**	**26**	**60**

En la muestra bajo estudio, se registran 12 adolescentes mujeres que viven en departamentos.

Estadística descriptiva

ANÁLISIS BIVARIADO

Tablas de contingencia

Las tablas de contingencia muestran la distribución conjunta de los individuos bajo análisis de acuerdo a dos variables simultáneamente:

Vivienda / Género	Casa	Departamento	Total
Mujer	30,0%	20,0%	**50,0%**
Varón	26,7%	23,3%	**50,0%**
Total	**56,7%**	**43,3%**	**100,0%**

Distribución conjunta

En la muestra bajo estudio, se registran 12 adolescentes mujeres que viven en departamentos. Esto representa el 20% del total de encuestados.

Estadística descriptiva

ANÁLISIS BIVARIADO

Tablas de contingencia

Vivienda / Género	Casa	Departamento	Total
Mujer	52,9%	46,2%	**50,0%**
Varón	47,1%	53,8%	**50,0%**
Total	**100,0%**	**100,0%**	**100,0%**

Vivienda / Género	Casa	Departamento	Total
Mujer	60,0%	40,0%	**100,0%**
Varón	53,3%	46,7%	**100,0%**
Total	**56,7%**	**43,3%**	**100,0%**

Distribuciones condicionales

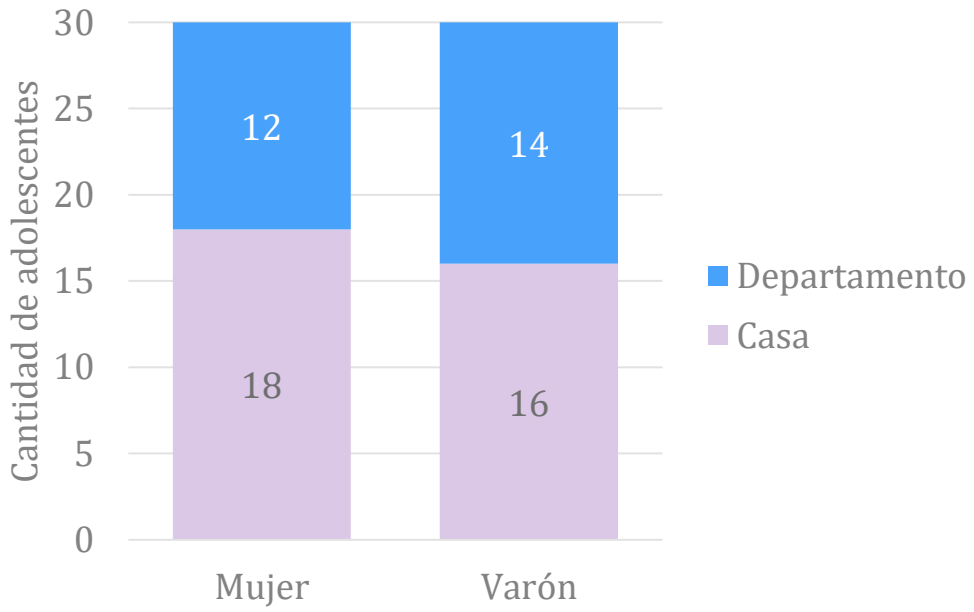
Del total de adolescentes que habitan en casa, el 52,9% son mujeres, mientras que dentro de quienes habitan en departamentos este porcentaje asciende al 46,2%.

Asimismo, el 60% de las mujeres y el 53,3% de los varones habitan en casas.

Estadística descriptiva

ANÁLISIS BIVARIADO

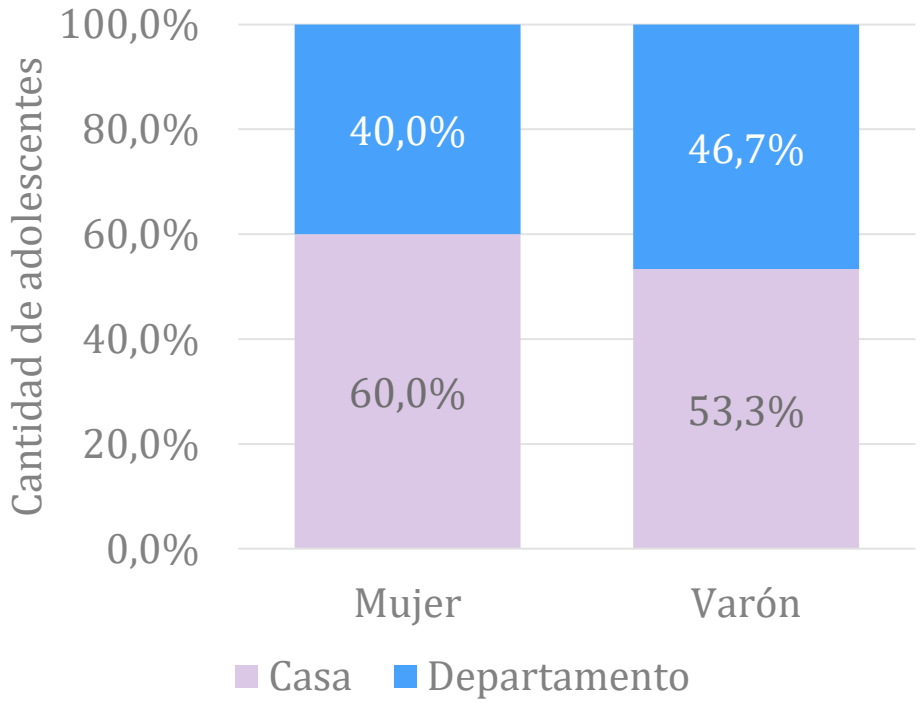
Gráficos de barras subdivididas

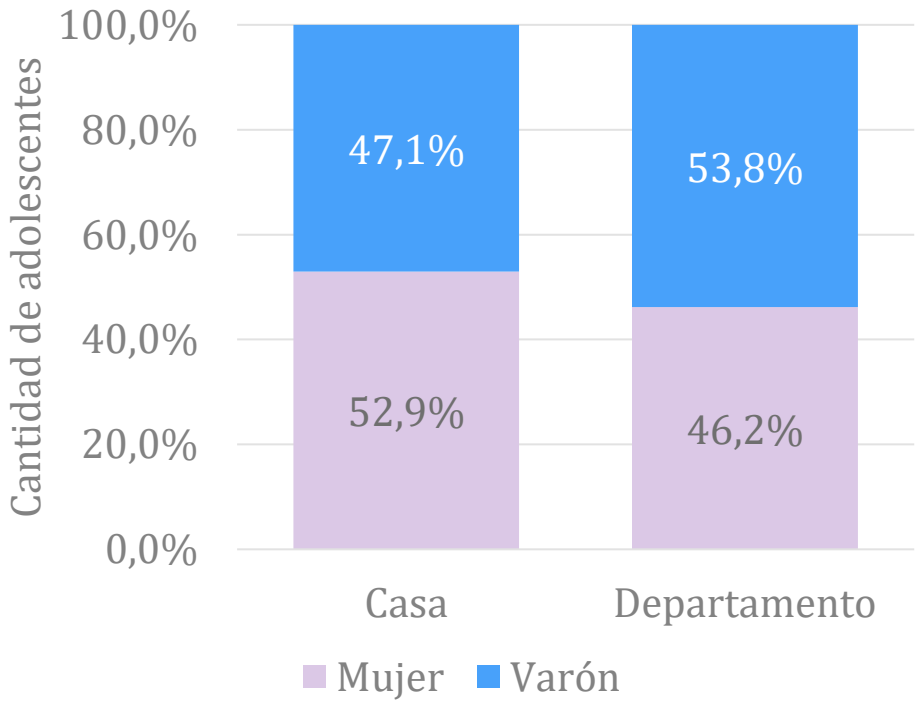


Los gráficos de barras subdivididas permiten analizar la relación entre dos variables categóricas. En primer lugar se decide qué grupos se desean comparar (en este caso, mujeres y varones) y se los grafica con ambas barras alcanzando el 100% (o el tamaño de cada grupo, en este caso ambos iguales a 30) y se identifica al interior de cada barra cuántos individuos presentan cada valor de la otra variable (tipo de vivienda en este caso).

Estadística descriptiva

ANÁLISIS BIVARIADO



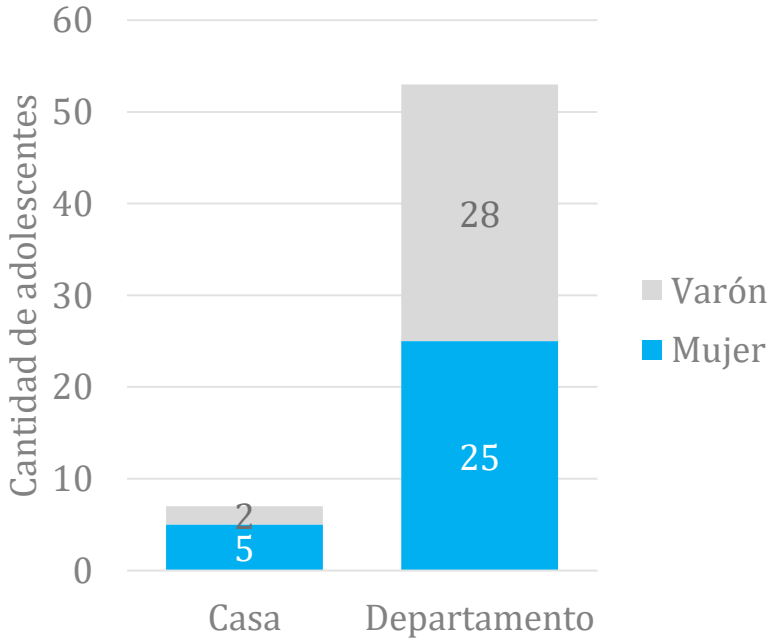


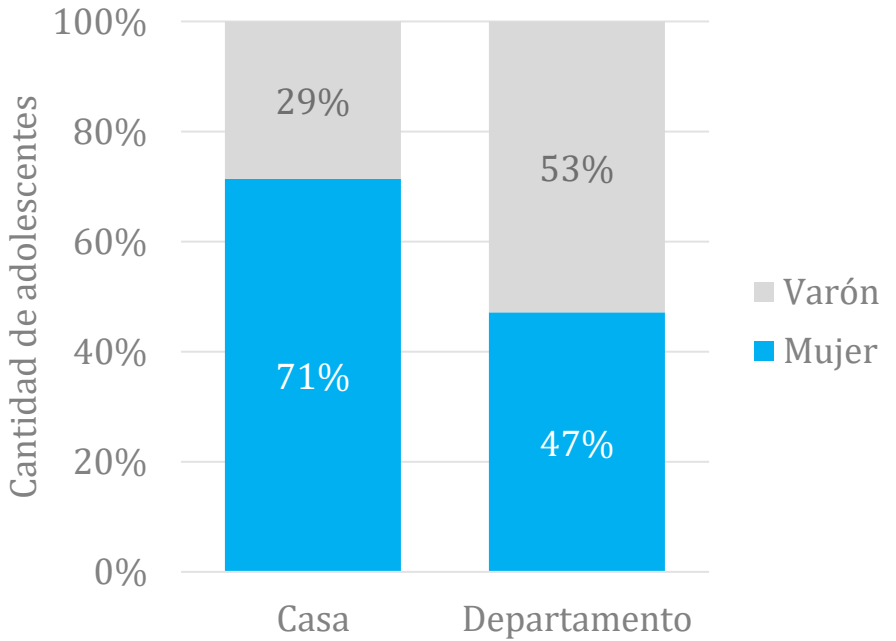
Estadística descriptiva

ANÁLISIS BIVARIADO

Gráficos de barras subdivididas

Podría incurrirse en un problema en aquellos casos en los que se desee representar la frecuencia absoluta en grupos de tamaño muy desigual: es difícil identificar diferencias o similitudes entre los grupos (en este caso, casa o departamento). Para situaciones de este tipo, será preferible utilizar gráficos de barras subdivididas porcentuales, llevando la barra de cada grupo al 100% y representando las distribuciones de frecuencia condicionales.

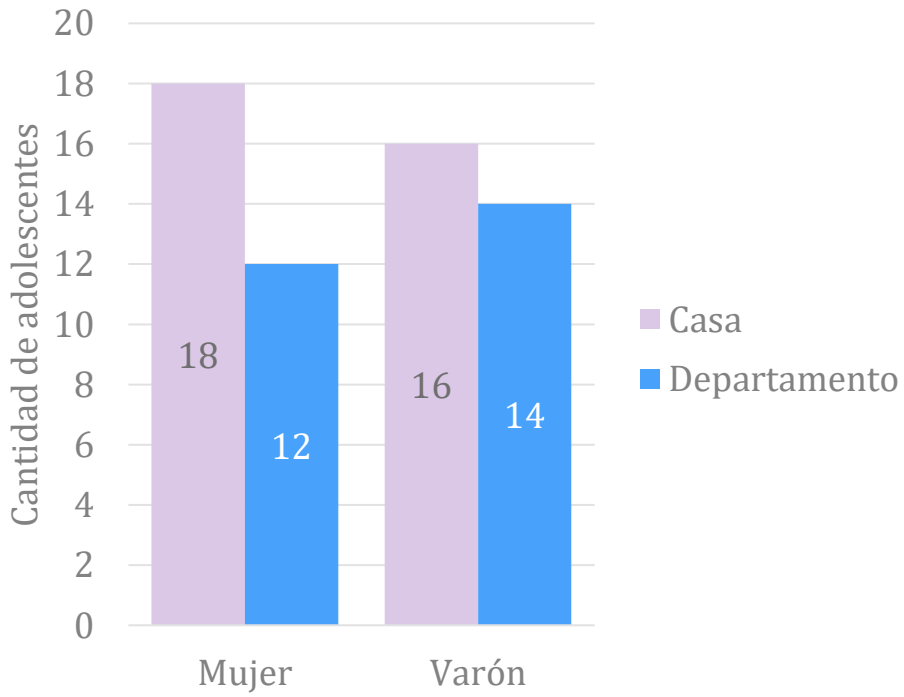




Estadística descriptiva

ANÁLISIS BIVARIADO

Gráficos de barras subdivididas



Los gráficos de barras agrupadas también permiten analizar la relación entre dos variables categóricas. En primer lugar se decide qué grupos se desean comparar (en este caso, mujeres y varones) y se grafican –en este caso- cuatro barras, dos en cada grupo correspondiéndose estas con las categorías de la otra variable (tipo de vivienda).

Estadística descriptiva

ANÁLISIS BIVARIADO