Deporte favorito	Personas	Frecuencia (%)	Frecuencia acumulada (%)
Fútbol	43	39,4	39,4
Básquet	27	24,8	64,2
Tenis	13	11,9	76,1
Hockey	8	7,3	83,5
Balonmano	7	6,4	89,9
Ciclismo	5	4,6	94,5
Béisbol	4	3,7	98,2
Volley	2	1,8	100,0
TOTAL	**109**	**100,0**	**100,0**

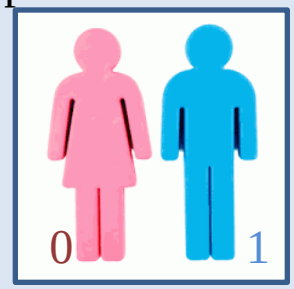
CONCEPTOS BÁSICOS DE

ESTADISTICA DESCRIPTIVA

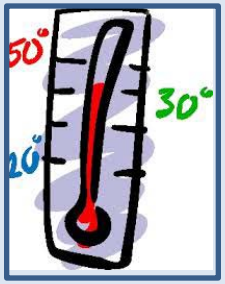
Procesamiento de datos

Codificación

Consiste en proporcionar códigos numéricos o alfanuméricos a diversos procesos



Cualitativos



Cuantitativos

Numero de decimales

Tabulación: Expresar valores, magnitudes u otros datos por medio de tablas.

	A	B	C	D
1	No.	Género	Lateralidad	
2	1	Hombre	Zurdo	
3	2	Mujer	Diestro	
4	3	Hombre	Diestro	
5	4	Mujer	Zurdo	
6	5	Hombre	Diestro	
7	6	Hombre	Diestro	
8	7	Mujer	Ambidiestro	
9	8	Mujer	Diestro	
10	9	Hombre	Zurdo	
11	10	Mujer	Diestro	
12				

Filas, columnas

- ✓Proporcionan una primera idea de las tendencias de los resultados.
- ✓Presentan los valores encontrados para cada variable
- ✓Primer paso para su presentación gráfica y para el tratamiento estadístico .
- ✓Su organización depende del número y de las características de las variables

Organización: Planificar o estructurar, poner orden

Diseños transversales o estáticos (los datos representan observaciones realizadas en un solo momento temporal).

Diseños longitudinales (los datos son registrados a lo largo de intervalos temporales):

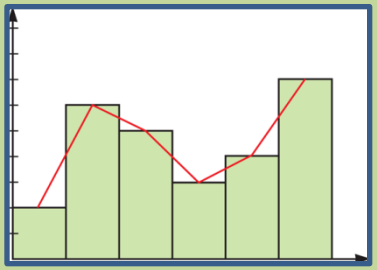
Análisis de datos

Es la técnica que consiste en el estudio de los hechos y el usos de sus expresiones en cifras para lograr información valida y confiable.

Del análisis	Técnica estadística
1. Hallar que hay en los datos.	1. Media, mediana, moda
2. Conocer que tanto varían los datos.	2. Desviación estándar, varianza, etc.
3. ¿Cómo están distribuidos los datos?.	3. Frecuencia
4. ¿Qué relación existe entre variable?.	4. Correlación y medidas de asociación
5. Estimaciones y predicciones.	5. Estimación de punto e intervalos y regresión
6. Describir las diferencias entre grupos y variables.	6. Prueba T, Z y análisis de varianza
7. Demostrar causalidad.	7. Prueba T, Z y análisis de varianza

Estadística descriptiva

Deporte favorito	Personas	Frecuencia (%)	Frecuencia acumulada (%)
Fútbol	43	39,4	39,4
Básquet	27	24,8	64,2
Tenis	13	11,9	76,1
Hockey	8	7,3	83,5
Balonmano	7	6,4	89,9
Ciclismo	5	4,6	94,5
Béisbol	4	3,7	98,2
Volley	2	1,8	100,0
TOTAL	109	100,0	100,0



✓Cuando se dispone de datos de una población, y antes de abordar análisis estadísticos mas complejos, un primer paso consiste en presentar esa información de forma que ésta se pueda visualizar de una manera mas sistemática y reducida

✓Para variables categóricas, se requiere conocer la frecuencia y el porcentaje del total de caso que ”caen” en cada categoría.

✓Representar resultados mediante diagrama de barras o diagrama de sectores.

Analiza metódicamente los datos, simplificándolos y presentándolos en forma clara; eliminando la confusión característica de los datos preliminares. Permite la elaboración de cuadros, gráficos e índices bien calculados

Estadística Inferencial

Provee conclusiones o inferencias, basándose en los datos simplificados y analizados; detectando las interrelaciones que pueden unirlos, las leyes que los rigen y eliminando las influencias del azar; llegando más allá de las verificaciones físicas posibles

Análisis multivariante

❖*Regresión múltiple:* Permite evaluar la influencia simultánea de varias variables independientes sobre una variable dependiente

❖*Análisis discriminante:* Se utiliza cuando en la variable dependiente existe más de una categoría y quiere averiguarse cómo se relaciona esta división con las variables independiente.

❖*Análisis factorial:* Herramienta para obtener información sobre las características subyacentes a un conjunto de datos.

Coeficientes de correlación

Prueba de hipótesis

Frecuencia absoluta n_i

Se define como el **número de veces que aparece repetido el valor** en cuestión de la variable estadística en el conjunto de las observaciones realizadas.

Las frecuencias absolutas cumplen las propiedades

$$0 \le n_i \le N \quad ; \quad \sum_{i=1}^{k} n_i = N.$$

$N \rightarrow$ número de observaciones (o tamaño de la muestra)

La frecuencia absoluta, aunque nos dice el número de veces que se repite un dato, no nos informa de la importancia de éste.

Frecuencia relativa f_i

Cociente entre la frecuencia absoluta y el número de observaciones realizadas N. Es decir

$$f_i = \frac{n_i}{N},$$

cumpliéndose las propiedades

$$0 \le f_i \le N \quad ; \quad \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{N}. = \frac{\sum_{i=1}^{k} n_i}{N} = 1.$$

Esta frecuencia relativa se puede expresar también en tantos por cientos del tamaño de la muestra

$$(\%)_{x_i} = 100 \times f_i$$

Frecuencia absoluta acumulada N_i

Suma de las frecuencias absolutas de los valores inferiores o igual a x_i , o número de medidas por debajo, o igual, que x_i . Evidentemente la frecuencia absoluta acumulada de un valor se puede calcular a partir de la correspondiente al anterior como

$$N_i = N_{i-1} + n_i \qquad N_1 = n_1.$$

La frecuencia absoluta acumulada del último valor será $N_k = N.$

Frecuencia relativa acumulada F_i

Cociente entre la frecuencia absoluta acumulada y el número de observaciones. Coincide además con la suma de las frecuencias relativas de los valores inferiores o iguales a x_i .

$$F_i = \frac{N_i}{N} = \frac{\sum_{i=1}^{k} n_j}{N} = \sum_{j=1}^{i} \frac{n_j}{N} = \sum_{j=1}^{i} f_i,$$

y la frecuencia relativa acumulada del último valor es 1

$$F_k = 1.$$

Se puede expresar asimismo como un porcentaje (multiplicando por 100) y su significado será el tanto por ciento de medidas con valores por debajo o igual que x_i .

Tablas de frecuencias

Supongamos que tenemos una muestra de tamaño N , donde la variable estadística x toma los valores distintos $x_1, x_2, \ldots, x_k$.

En primer lugar hay que ordenar los diferentes valores que toma la variable estadística en orden (normalmente creciente).

Valores de la variable estadística x_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k

Es posible hacer también una tabla de frecuencias de una variable cualitativa. En ese caso, en la primera columna se escribirán las diferentes cualidades o atributos que puede tomar la variable

Ejercicio. Supongamos que el numero de hijos de una muestra de 20 familias es la siguiente

4	1	1	3	1	2	5	1	2	3
2	2	3	2	1	4	2	3	2	1

El tamaño de la muestra es N=20, el numero de valores posibles k=5

x_i	n_i	f_i $n_i/20$	N_i $\sum_1^i n_j$	F_i $\sum_1^i f_j$
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

Agrupamiento en intervalos de clase

Cuando el número de valores distintos que toma la variable estadística es demasiado grande o la variable es continua no es útil elaborar una tabla de frecuencias como la vista anteriormente.

Se realiza un agrupamiento de los datos en intervalos y se hace un recuento del número de observaciones que caen dentro de cada uno de ellos. Dichos intervalos se denominan **intervalos de clase**, y al valor de la variable en el centro de cada intervalo se le llama **marca de clase**. A la diferencia entre el extremo superior e inferior de cada intervalo se le llama **amplitud del intervalo**.

Intervalos de la clase	**Marcas de la clase**	**Frecuencias absolutas**	**Frecuencias relativas**	**Frecuencias absolutas acumuladas**	**Frecuencias relativas acumuladas**
$a_i - a_{i+1}$	c_i	n_i	$f_i = n_i / N$	N_i	$F_i = N_i / N$
$a_1 - a_2$	c_1	n_1	f_1	N_1	F_1
$a_2 - a_3$	c_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_k - a_{k+1}$	c_k	n_k	f_k	N_k	F_k

La tabla de frecuencias resultante es similar a la vista anteriormente

El realizar el estudio mediante el agrupamiento en intervalos de clase simplifica el trabajo, pero también supone una pérdida de información, ya que no se tiene en cuenta cómo se distribuyen los datos dentro de cada intervalo. Para que dicha pérdida sea mínima es necesario elegir con cuidado los intervalos.

1. Determinar el recorrido, o rango, de los datos
2. Decidir el número k de intervalos de clase en que se van a agrupar los datos. Dicho número se debe situar normalmente entre 5 y 20, dependiendo del caso. $\sqrt{N}$
3. Dividir el recorrido entre el número de intervalos para determinar la amplitud (constante) de cada intervalo.
4. Determinar los extremos de los intervalos de clase.
5. Calcular las marcas de clase de cada intervalo como el valor medio entre los límites inferior y superior de cada intervalo de clase.

Una vez determinados los intervalos se debe hacer un recuento cuidadoso del número de observaciones que caen dentro de cada intervalo, para construir así la tabla de frecuencias.

En la tabla siguiente se listan los datos medidos por **James Short en 1763** sobre la paralaje del Sol en segundos de arco. La paralaje es el ángulo subtendido por la Tierra vista desde el Sol. Se midió observando tránsitos de Venus desde diferentes posiciones y permitió la primera medida de la distancia Tierra-Sol, que es la unidad básica de la escala de distancias en el Sistema Solar (la unidad astronómica).

Datos en segundos de arco

8.63	**10.16**	**8.50**	**8.31**	**10.80**	**7.50**	**8.12**
8.42	9.20	8.16	8.36	9.77	7.52	7.96
7.83	8.62	7.54	8.28	9.32	7.96	**7.47**

Recorrido: 10.80-7.47=3.33

Número de intervalos: $k = \sqrt{21} = 4.53 \Rightarrow k = 5$

Amplitud del intervalo: $3.33/5 = 0.666 \Rightarrow 0.7$

Extremos de los intervalos $a_1 = 7.405 < 7.470$ mínimo

$7.405 + 5 \times 0.7 = 10.905 > 10.80$ máximo

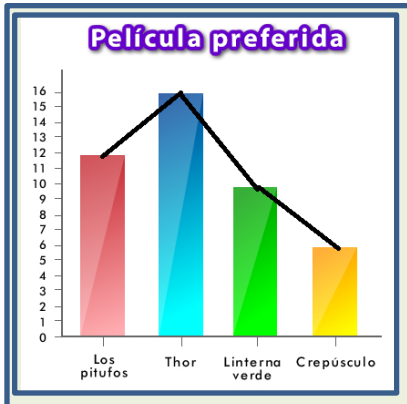
Recuento y construcción de la tabla

$a_i - a_i + 1$	c_i	n_i	f_i	$\frac{N_i}{\sum_1^i n_j}$	$\frac{F_i}{\sum_1^i f_j}$
7.405 - 8.105	7.755	7	0.333	7	0.333
8.105 - 8.805	8.455	9	0.429	16	0.762
8.805 - 9.505	9.155	2	0.095	18	0.857
9.505 - 10.205	9.855	2	0.095	20	0.952
10.205 - 10.905	10.555	1	0.048	21	1.00
		21	1.000		

Representaciones gráficas

Representaciones gráficas para datos sin agrupar

Diagrama de barras: En éste se representan en el eje de abscisas los distintos valores de la variable y sobre cada uno de ellos se levanta una barra de longitud igual a la frecuencia correspondiente.



Polígono de frecuencias: Este se obtiene uniendo con rectas los extremos superiores de las barras del diagrama anterior. De la misma forma, pueden representarse frecuencias absolutas relativas, o ambas a la vez

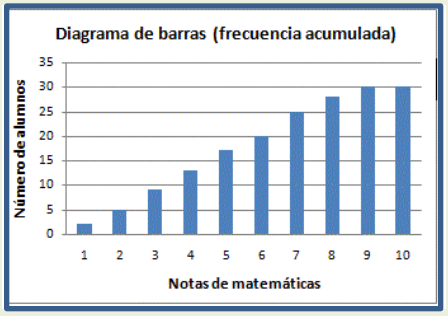
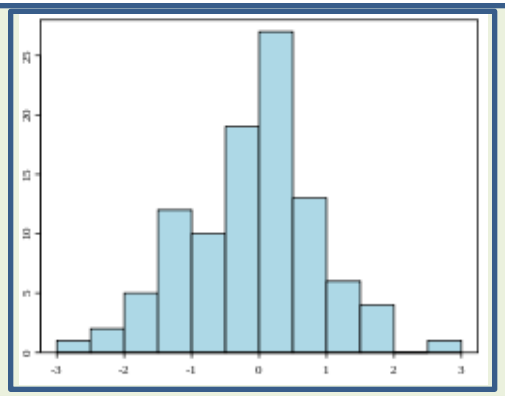
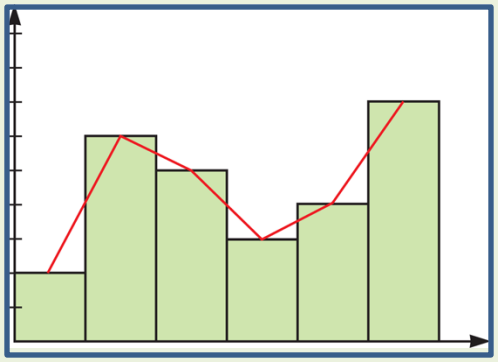


Diagrama de frecuencias acumuladas: Este gráfico, en forma de escalera, se construye representando en abscisas los distintos valores de la variable y levantando sobre cada x_i una perpendicular cuya longitud será la frecuencia acumulada (N_i o F_i) de ese valor.

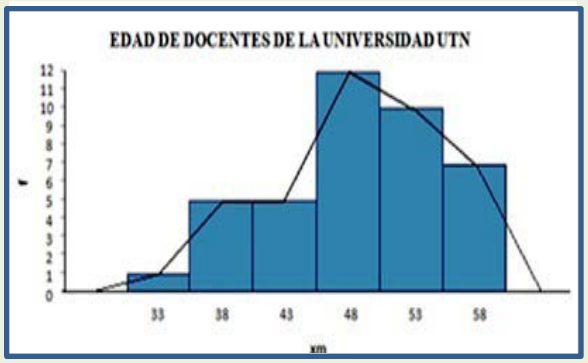
Representaciones gráficas para datos agrupados



Histograma: Es un conjunto de rectángulos adyacentes, cada uno de los cuales representa un intervalo de clase. Las base de cada rectángulo es proporcional a la amplitud del intervalo. La altura se suele determinar para que el área de cada rectángulo sea igual a la frecuencia de la **marca de clase** correspondiente.



Polígono de frecuencias: Este se obtiene uniendo por líneas rectas los puntos medios de cada segmento superior de los rectángulos en el histograma.



Polígono de frecuencias acumuladas: Sirve para representar las frecuencias acumuladas de datos agrupados por intervalos. En abscisas se representan los diferentes intervalos de clase.

Representaciones gráficas para variables cualitativas

Existe una gran variedad de representaciones para variables cualitativas

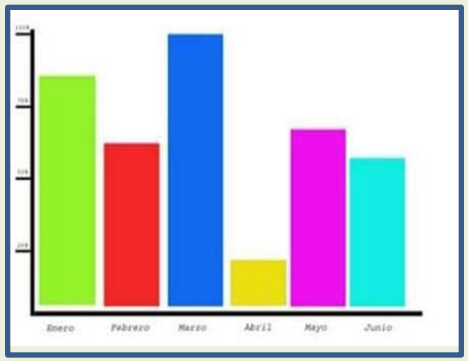


Diagrama de rectángulos: Es similar al diagrama de barras y el histograma para las variables cuantitativas. Consiste en representar en el eje de abscisas los diferentes caracteres cualitativos y levantar sobre cada uno de ellos un rectángulo (de forma no solapada) cuya altura sea la frecuencia (absoluta o relativa) de dicho carácter.

Diagrama de sectores (también llamado diagrama de torta): En él se representa el valor de cada carácter cualitativo como un sector de un círculo completo, siendo el área de cada sector, o, lo que es lo mismo, el arco subtendido, proporcional a la frecuencia del carácter en cuestión. Este tipo de diagrama proporciona una idea visual muy clara de cuáles son los caracteres que más se repiten.

Caracterización de un conjunto de mediciones

1. Medidas de Centralización o localización

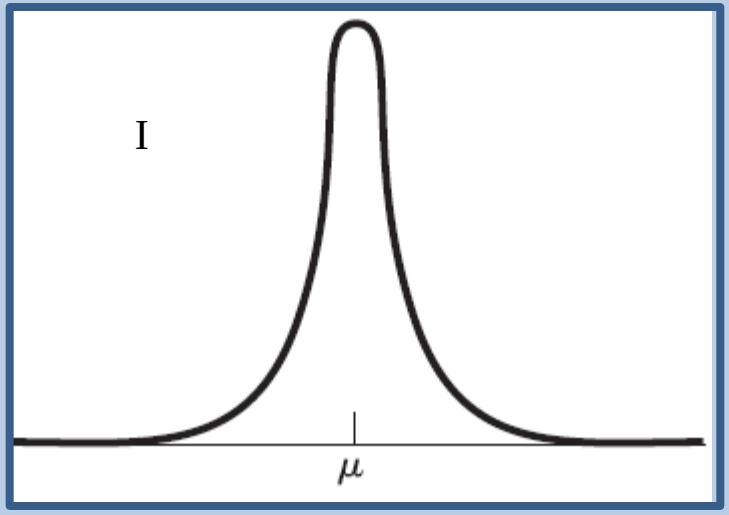
1. 1 Media aritmética

La media de una muestra de n respuestas medidas $x_1, x_2, \ldots, x_n$ está dada por

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

La media poblacional correspondiente se denota como μ .

Dos conjuntos de mediciones podrían tener distribuciones de frecuencia muy diferentes pero iguales medias



La diferencia entre las distribuciones I y II de la figura se encuentra en la variación o dispersión de las mediciones que están a lado y lado de la media.

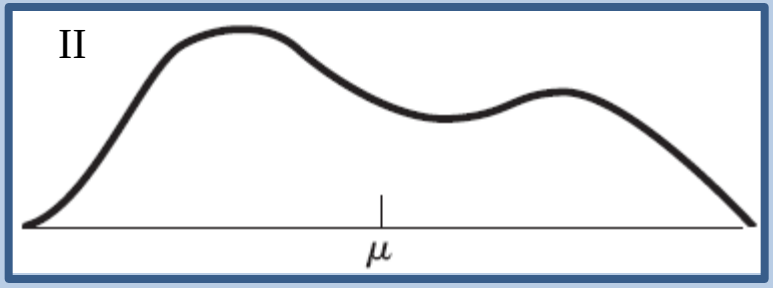


Fig. Distribuciones de frecuencia con iguales medias pero con diferentes cantidades de variación

Una característica importante de la media como medida de tendencia central es que es muy poco robusta, es decir depende mucho de valores particulares de los datos.

Para diferentes valores de x que aparezcan repetidos, con frecuencias $n_1, n_2, \ldots, n_k$

Promedio

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

En función de las frecuencias relativas

$$\bar{x} = \sum_{i=1}^{k} x_i f_i$$

Para una muestra agrupada en k intervalos de clase, la media se puede calcular, a partir de las marcas de clase c_i y el número n_i de datos en cada intervalo,

$$\bar{x} = \frac{\sum_{i=1}^{k} c_i n_i}{N}$$

hay que indicar que es solamente aproximada.

Una propiedad importante de la media aritmética es que la suma de las desviaciones de un conjunto de datos respecto a su media es cero.

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

La media representa entonces una especie de centro de gravedad, o centro geométrico, del conjunto de medidas

Es decir, la media equilibra las desviaciones positivas y negativas respecto a su valor

Ejercicio. Supongamos que el numero de hijos de una muestra de 20 familias es la siguiente

4	1	1	3	1	2	5	1	2	3
2	2	3	2	1	4	2	3	2	1

x_i	n_i	f_i $n_i/20$	$x_i \times n_i$	$x_i \times f_i$
1	6	0.30	6	0.30
2	7	0.35	14	0.70
3	4	0.20	12	0.60
4	2	0.10	8	0.40
5	1	0.05	5	0.25

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i n_i}{N} = \frac{45}{20} = 2.25$$

$$\bar{x} = \sum_{i=1}^{5} x_i f_i = 2.25$$

En la tabla siguiente se listan los datos medidos por **James Short en 1763** sobre la paralaje del Sol en segundos de arco. La paralaje es el ángulo subtendido por la Tierra vista desde el Sol. Se midió observando tránsitos de Venus desde diferentes posiciones y permitió la primera medida de la distancia Tierra-Sol, que es la unidad básica de la escala de distancias en el Sistema Solar (la unidad astronómica).

Datos en segundos de arco

8.63	**10.16**	**8.50**	**8.31**	**10.80**	**7.50**	**8.12**
8.42	9.20	8.16	8.36	9.77	7.52	7.96
7.83	8.62	7.54	8.28	9.32	7.96	7.47

$a_i - a_i + 1$	c_i	n_i	$c_i \times n_i$
7.405 - 8.105	7.755	7	54.285
8.105 - 8.805	8.455	9	76.095
8.805 - 9.505	9.155	2	18.310
9.505 - 10.205	9.855	2	19.710
10.205 - 10.905	10.555	1	10.555
		21	178.955

$$\bar{x} = \frac{\sum_{i=1}^{5} c_i n_i}{N} = \frac{178.955}{21} = 8.522$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{21} x_i = \frac{178.43}{21} = 8.497$$

1.2 Medias geométrica x_G

En el caso de una muestra con valores diferentes de la variable se define como la raíz enésima (N es el tamaño de la muestra) del producto de los valores de la variable

$$\bar{x}_G = \sqrt[N]{x_1 x_2 \dots x_N}\,.$$

Si los datos aparecen agrupados en k valores distintos la definición será

$$\bar{x}_G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}\,.$$

- Esta media tiene la característica negativa de que si uno de los valores es nulo, la media sería asimismo cero, y por lo tanto sería poco representativa del valor central.
- Además si existen valores negativos es posible que no se pueda calcular

De poca utilidad

El logaritmo de la media geométrica es la media aritmética del logaritmo de los datos

$$\log \bar{x}_G = \frac{\sum_{i=1}^{k} n_i \log x_i}{N}$$

1.3 La media armónica x_A

Se define como la inversa de la media aritmética de las inversas de los valores de la variable. Es decir, para variables no agrupadas y agrupadas, sería

$$\bar{x}_A = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \qquad \bar{x}_A = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}.$$

Es evidente que si una de las medidas es 0, la media armónica no tiene sentido.

1.3 Media cuadrática x_Q

Se define ésta como la raíz cuadrada de la media aritmética de los cuadrados de los valores

$$\bar{x}_Q = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \qquad \bar{x}_Q = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{N}}$$

Esta media tiene su utilidad con frecuencia en la aplicación a fenómenos físicos

Relación con la media aritmética $\boxed{\bar{x}_A \leq \bar{x}_G \leq \bar{x} \leq \bar{x}_Q.}$

Ninguna de estas medias es muy robusta en general

Por ejemplo, la media armónica es muy poco sensible a valores muy altos de x , mientras que a la media cuadrática apenas le afectan los valores muy bajos de la variable.

Supongamos una serie de medidas experimentales con un péndulo simple para obtener el valor de la aceleración de la gravedad (en m/s2).

9.77 9.78 9.80 9.81 9.83 10.25

Media aritmética

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{9.77 + 9.78 + 9.80 + 9.81 + 9.83 + 10.25}{6} \Rightarrow \bar{x} = 9.873$$

Media geométrica

$$\bar{x}_G = \sqrt[6]{x_1 x_2 \ldots x_6} = \sqrt[6]{9.77 \times 9.78 \times \ldots \times 10.25} \Rightarrow \bar{x}_G = 9.872$$

Media armónica

$$\bar{x}_A = \frac{6}{\sum_{i=1}^{6} \frac{1}{x_i}} = \frac{6}{\frac{1}{9.77} + \frac{1}{9.78} + \ldots \frac{1}{10.25}} \Rightarrow \bar{x}_A = 9.871$$

Media cuadrática

$$\bar{x}_Q = \sqrt{\frac{\sum_{i=1}^{6} x_i^2}{6}} = \sqrt{\frac{9.77^2 + 9.78^2 + \ldots + 10.25^2}{6}} \Rightarrow \bar{x}_Q = 9.875$$

$$9.871 \leq 9.872 \leq 9.873 \leq 9.875$$

$$\bar{x}_A \leq \bar{x}_G \leq \bar{x} \leq \bar{x}_Q .$$

1.4 Mediana

Una medida de centralización importante es la mediana M_e. Se define como una medida central tal que, con los datos ordenados de ***menor a mayor***, el 50% de los datos son inferiores a su valor y el 50% de los datos tienen valores superiores. Es decir, la mediana divide en **dos partes iguales** la distribución de frecuencias o, gráficamente, divide el histograma en dos partes de áreas iguales.

Casos

Supongamos en primer lugar que los diferentes valores de la variable no aparecen, en general, repetidos.

N = impar La mediana será el valor central

N = par La mediana es la media aritmética de los dos valores centrales

No depende tanto del valor extremo. Es una medida más robusta

Ejercicio: Supongamos una serie de medidas experimentales con un péndulo simple para obtener el valor de la aceleración de la gravedad (en m/s2).

9.77 9.78 9.80 9.81 9.83 10.25

$N = 6 = par$

$$M_e = \frac{9.80 + 9.81}{2} \Rightarrow M_e = 9.805$$

15.25 M_e ← igual

$\bar{x}$ ← cambia

En el caso de que tengamos una variable discreta con valores repetidos sobre la cual hemos elaborado una tabla de frecuencias se calcula en primer lugar el número de observaciones N dividido entre 2.

Podemos distinguir entonces dos casos.

El primero de ellos es cuando dicho valor $\frac{N}{2}$ coincide con la frecuencia absoluta acumulada N_j de un valor de la variable x_j . En este caso la mediana se ha de situar entre este valor de la variable y el siguiente ya que de esta forma dividirá la distribución de frecuencias en 2.

$$M_e = \frac{x_j + x_{j+1}}{2}$$

Si $\frac{N}{2}$ no coincidiese con ningún valor de la columna de **frecuencias acumuladas** (**como suele ocurrir**) la mediana sería el primer valor de x_j con frecuencia absoluta acumulada N_j mayor que $\frac{N}{2}$, ya que el valor central de la distribución **correspondería a una de las medidas englobadas en ese** x_j .

Ejemplo1: 1–1–1–1–1–1–2–2–2–2–3–3–3–3–3–4–4–5–5–5

x_i	N_i
1	6
2	10
3	15
4	17
5	20

$$\frac{N}{2} = 10 = N_2$$

$$M_e = \frac{x_2 + x_{2+1}}{2} = \frac{2+3}{2} = 2.5$$

Ejemplo2: 1–1–1–1–1–1–2–2–2–2–2–2–2–3–3–3–3–4–4–5

x_i	N_i
1	6
2	13
3	17
4	19
5	20

$$\frac{N}{2} = 10$$

La mediana será el primer valor de x_i con frecuencia absoluta acumulada $N_i > 10$

$$M_e = x_2 = 2$$

Muestra de una variable continua cuyos valores están agrupados en intervalos de clase.

En este caso pueden ocurrir dos situaciones

En primer lugar, si $\frac{N}{2}$ coincide con la frecuencia absoluta acumulada N_j de un intervalo (a_j, a_{j+1}) (con marca de clase c_j), la mediana será sencillamente el extremo superior a_{j+1} de ese intervalo.

Supongamos que el valor $\frac{N}{2}$ se encuentra entre las frecuencias N_{j-1} y N_j, correspondientes a los intervalos (a_{j-1}, a_j) y (a_j, a_{j+1}) respectivamente, la mediana se situará en algún lugar del intervalo superior (a_j, a_{j+1}).

$$\frac{a_{j+1} - a_j}{N_j - N_{j-1}} = \frac{M_e - a_j}{N/2 - N_{j-1}} =$$

$$M_e = a_j + \frac{\frac{N}{2} - N_{j-1}}{N_j - N_{j-1}}\left(a_{j+1} - a_j\right) = a_j + \frac{\frac{N}{2} - N_{j-1}}{n_j}\left(a_{j+1} - a_j\right) \qquad N_j - N_{j-1} = n_j$$

- La mayor ventaja de la media es que se utiliza toda la información de la distribución de frecuencias (todos los valores particulares de la variable), en contraste con la mediana, que solo utiliza el orden en que se distribuyen los valores.

- La mediana, es una medida robusta, siendo muy insensible a valores que se desvíen mucho.

- En general, lo mejor es considerar media aritmética y mediana como medidas complementarias

Ejemplo 3

$a_i - (a_i + 1)$	n_i	N_i
7.405 - 8.105	7	7
8.105 - 8.805	9	16
8.805 - 9.505	2	18
9.505 - 10.205	2	20
10.205 - 10.905	1	21

21

$$\frac{N}{2} = \frac{21}{2} = 10.5 \neq N_i$$

$$(N_1 = 7) < \left(\frac{N}{2} = 10.5\right) < (N_2 = 16)$$

M_e Intervalo $8.105 - 8.805$

$$8.105 < M_e < 8.805$$

$$M_e = a_j + \frac{\frac{N}{2} - N_{j-1}}{N_j - N_{j-1}}(a_{j+1} - a_j) \Rightarrow \qquad M_e = a_j + \frac{\frac{N}{2} - N_{j-1}}{n_j}(a_{j+1} - a_j)$$

$$j = 2$$

$$M_e = a_2 + \frac{\frac{N}{2} - N_1}{n_2}(a_3 - a_2) \qquad \Rightarrow M_e = 8.105 + \frac{10.5 - 7}{9}(8.805 - 8.105)$$

$$M_e = 8.38$$

1.5 Moda

Se define la moda M_o de una muestra como aquel valor de la variable que tiene una frecuencia máxima. En otras palabras, es el valor que más se repite. Hay que indicar que puede suceder que la moda no sea única, es decir que aparezcan varios máximos en la distribución de frecuencias. En ese caso diremos que tenemos una distribución bimodal, trimodal, etc.

Evidentemente, en el caso de una variable discreta que no toma valores repetidos, la moda no tiene sentido. Cuando sí existen valores repetidos su cálculo es directo ya que puede leerse directamente de la tabla de distribución de frecuencias

Caso de variables continuas agrupadas en intervalos de clase

Existe un intervalo en el que la frecuencia sea máxima, llamado **intervalo modal**. Es posible asociar la moda a un valor determinado de la variable dentro de dicho intervalo modal.

supongamos que

(a_j, a_{j+1}) *el intervalo con frecuencia máxima* n_j

n_{j+1} *y* n_{j-1} *son las frecuencias de los intervalos anterior y posterior al modal*

Definimos $\delta_1 = n_j - n_{j-1} \quad \delta_2 = n_j - n_{j+1}$

$$M_o = a_j + \frac{\delta_1}{\delta_1 + \delta_2}\left(a_{j+1} - a_j\right)$$

Caso particular

$$n_j = n_{j-1} \Rightarrow \delta_1 = 0 \ \wedge \ M_o = a_j$$

$$n_{j+1} = n_j \Rightarrow \delta_2 = 0 \ \wedge \ M_o = a_{j+1}$$

La moda estará más próxima a a_j cuanto menor sea la diferencia de frecuencias con el intervalo anterior, y al revés.

Ejercicio. Supongamos que el numero de hijos de una muestra de 20 familias es la siguiente

x_i	n_i	f_i $n_i/20$	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1

El valor que más se repite es 2 hijos, que ocurre en siete familias de la muestra ($n_i = 7$).

La moda es por tanto $M_0 = 2$ ⟸

$\frac{N}{2} = 10$ La mediana será el primer valor de x_i con frecuencia absoluta acumulada $N_i > 10$

$M_e = x_2 = 2$ ⟸

Ejercicio. Tabla de frecuencias del paralaje del sol medidos por James Short en 1763

$a_i - a_i + 1$	c_i	n_i
7.405 - 8.105	7.755	7
8.105 - 8.805	8.455	9
8.805 - 9.505	9.155	2
9.505 - 10.205	9.855	2
10.205 - 10.905	10.555	1

Intervalo con frecuencia máxima $(a_j, a_{j+|}) = 8.105, 8.805$

$j = 2; \quad n_{j-1} = 7; \quad n_j = 9; \quad n_{j+1} = 2$

$$\delta_1 = n_j - n_{j-1} = 9 - 7 = 2$$
$$\delta_2 = n_j - n_{j+1} = 9 - 2 = 7$$

$$M_o = a_j + \frac{\delta_1}{\delta_1 + \delta_2}(a_{j+1} - a_j) \Rightarrow \qquad M_o = 8.805 + \frac{2}{2+7}(8.805 - 8.105)$$

$$M_o = 8.26$$

1.6 Medidas de posición no centrales. Percentiles, Cuartiles, y Deciles

1.5.5 Percentiles

El percentil ***p*** *es un valor tal que por lo menos* ***p*** *por ciento de las observaciones son* menores o iguales que este valor y por lo menos (***100 - p***) *por ciento de las observaciones* son mayores o iguales que este valor.

Cálculo del percentil *p*

Paso 1. Ordenar los datos de menor a mayor (colocar los datos en orden ascendente).

Paso 2. Calcular el índice *i*

$$i = \left(\frac{p}{100}\right) n$$

donde *p es el percentil deseado y n es el número de observaciones.*

Paso 3. (a) Si *i no es un numero entero, debe redondearlo. El primer entero mayor que i denota la posición del percentil p.* **(b)** Si *i es un numero entero, el percentil p es el promedio de los valores en las* posiciones *i e i + 1.*

Ejercicio. Determine el percentil 85 en los sueldos mensuales

Profesor	1	2	3	4	5	6	7	8	9	10	11	12
Sueldo	3450	3550	3650	3480	3355	3310	3490	3730	3540	3925	3520	3480

Paso 1. 3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 **3730** 3925

Paso 1. $i = \left(\frac{85}{100}\right) \times 12 \Rightarrow i = 10.2$

Paso 1. La posición del percentil 85 es el primer entero mayor que 10.2, es la posición 11. **3730**

Generalización de la definición de Mediana

Se define como Cuartiles a los tres valores que dividen la muestra en cuatro partes iguales.

1.6.1 Primer cuartil $Q_{1/4}$ o percentil 25

Será la medida tal que, el 25% de los datos sean inferiores a su valor y el 75% de los datos sean superiores

1.6.2 Segundo cuartil $Q_{1/2}$ o percentil 50

Coincide con la mediana

1.6.3 Tercer cuartil $Q_{3/4}$ o percentil 75

Marcará el valor tal que las tres cuartas partes de las observaciones sean inferiores a él y una cuarta parte sea superior

La forma de calcular los Cuartiles es igual a la vista para la mediana pero sustituyendo $N/2$ por $N/4$ y $3N/4$ para $Q_{1/4}$ y $Q_{3/4}$ respectivamente

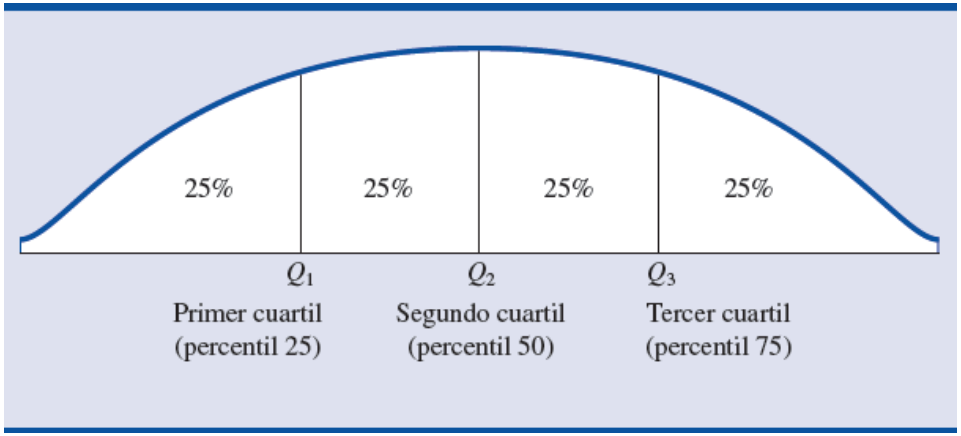


Fig. Localización de los Cuartiles

1.6 Deciles

Se definen como aquellos valores de la variable que dividen la muestra, ordenada, en 10 partes iguales.

Estos valores, denotados por D_k , con $k = 1,2,\ldots,9$, tienen entonces un valor tal que el decil k-ésimo deja por debajo de él al $10 \times k$ por ciento de los datos de la muestra.

Algunos valores de Cuartiles, deciles y centiles coinciden, cumpliéndose por ejemplo,

$$P_{50} = D_5 = Q_{1/2} = M_e$$

Ejercicio. Supongamos que el numero de hijos de una muestra de 20 familias es la siguiente

x_i	N_i
1	6
2	13
3	17
4	19
5	20

$M_e = x_2 = 2$

1-1-1-1-1- | 1-2-2-2-2- | 2-2-2-2-3-3- | 3-3-4-4-5

Primer cuartil $\frac{N}{4} = \frac{20}{4} = 5 \Rightarrow \quad Q_{1/4} = 1$

Segundo cuartil $\frac{N}{2} = \frac{20}{2} = 10 \Rightarrow \quad Q_{1/2} = M_e = 2$

Tercer cuartil $3 \times \frac{N}{4} = 15 \Rightarrow \quad Q_{3/4} = 3$

En el caso de las medidas agrupadas en intervalos de clase se trabaja igual que para determinar la mediana

Ejercicio. Tabla de frecuencias del paralaje del sol medidos por James Short en 1763

$a_i - a_i + 1$	n_i	N_i
7.405 - 8.105	7	7
8.105 - 8.805	9	16
8.805 - 9.505	2	18
9.505 - 10.205	2	20
10.205 - 10.905	1	21

$\frac{N}{4} = 5.25 < 7 \qquad 3 \times \frac{N}{4} = 15.75 < 16$

$Q_{1/4} \Rightarrow$ Se sitúa en el primer intervalo 7.405 - 8.105

$Q_{3/4} \Rightarrow$ Se sitúa en el primer intervalo 8.105 - 8.805

$j = 1 \qquad a_1 = 7.405 \qquad a_2 = 8.105$

$$Q_{1/4} = a_j + \frac{\frac{N}{4} - N_{j-1}}{n_j}\left(a_{j+1} - a_j\right) \Rightarrow \qquad Q_{1/4} = 7.405 + \frac{5.25 - 0}{7}(0.7)$$

$$Q_{1/4} = 7.93$$

Segundo cuartil ?

$j = 2 \qquad a_2 = 8.105 \qquad a_3 = 8.805$

$$Q_{3/4} = a_j + \frac{3 \times \frac{N}{4} - N_{j-1}}{n_j}\left(a_{j+1} - a_j\right) \Rightarrow \qquad Q_{3/4} = 8.105 + \frac{15.75 - 7}{9}(0.7) \Rightarrow Q_{3/4} = 8.79$$

2 Medidas de dispersión

Medidas de dispersión Indican la variabilidad de los datos en torno a su valor promedio, es decir si se encuentran muy o poco esparcidos en torno a su centro.

2.1 Recorrido

También llamado rango, se define como la diferencia entre el valor máximo y mínimo que toma la variable estadística.

Con el fin de eliminar la excesiva influencia de los valores extremos en el recorrido, se define

Recorrido intercuartílico: Diferencia entre el tercer y primer cuartil

$$R_I = Q_{3/4} - Q_{1/4}$$

Este recorrido nos dará entonces el rango que ocupan el 50% central de los datos

En ocasiones se utiliza el recorrido semiintercuartílico, o mitad del recorrido intercuartílico

2. 2 Desviación media

También llamada con mas precisión desviación media respecto a la media aritmética

Se define como la media aritmética de las diferencias absolutas entre los valores de la variable y la media aritmética de la muestra.

Se N el tamaño de la muestra y k los distintos valores x_i de la variable con frecuencias absolutas n_i, la expresión de la desviación media será

$$D_{\bar{x}} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| n_i}{N}.$$

en el caso de que la variable no tome valores repetidos, ni esté agrupada en intervalos, la expresión anterior se simplifica a

$$D_{\bar{x}} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}|}{N}.$$

Si no se tomaran valores absolutos, unas desviaciones se anularían con otras, alcanzando finalmente la desviación media un valor de 0, debido a la propiedad de la media aritmética

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

Desviación media respecto a la mediana $D_{M_e} = \frac{\sum_{i=1}^{k} |x_i - M_e| n_i}{N}.$

La medida de variabilidad más común empleada en estadística es la varianza, que es una función de las desviaciones (o distancias) de las mediciones muestrales desde la media

2.3 Varianza y desviación típica

Se define entonces la **varianza** de una muestra con datos repetidos como

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

Evidentemente la varianza no tiene las mismas unidades que los datos de la muestra. Para conseguir las mismas unidades se define la **desviación típica** (algunas veces llamada desviación estándar) como la raíz cuadrada de la varianza

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

Aun cuando está estrechamente relacionada con la varianza, la desviación estándar se puede usar para dar una imagen **más o menos precisa de la variación de datos para un solo conjunto de mediciones**.

$$s^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i - \frac{1}{N}\left(\sum_{i=1}^{k} x_i n_i\right)^2}{N-1}$$

$$s^2 = \frac{\sum_{i=1}^{k} c_i^2 n_i - \frac{1}{N}\left(\sum_{i=1}^{k} c_i n_i\right)^2}{N-1}$$

Muchas distribuciones de datos de la vida real tienen forma de montículo. Los datos que poseen distribuciones en forma de montículo tienen características definidas de variación, como se expresa en el enunciado siguiente.

s^2 es “casi” el promedio del cuadrado de las desviaciones de los valores observados desde su media.

Significado

La varianza es de valor al comparar la variación relativa de dos conjuntos de mediciones, pero **proporciona información acerca de la variación en un solo conjunto únicamente cuando se interpreta en términos de la desviación estándar**.

Regla empírica

Para una distribución de mediciones que sea aproximadamente normal (forma de campana),se deduce que el intervalo con puntos extremos

$\mu \pm \sigma$	contiene aproximadamente 68% de las mediciones.
$\mu \pm 2\sigma$	contiene aproximadamente 95% de las mediciones.
$\mu \pm 3\sigma$	contiene casi todas las mediciones.

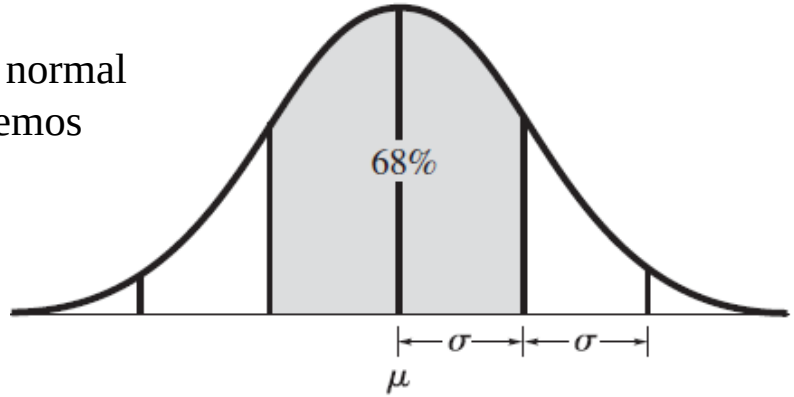


Fig. Curva normal

La regla se aplica a distribuciones que no son exactamente normales pero sí en forma de montículo

El conocimiento de la media y la desviación estándar nos da una imagen más o menos buena de la distribución de las frecuencias.

En el caso de que los datos no se repitan, estas definiciones se simplifican a

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2}{N-1} \qquad s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}.$$

✓La desviación típica, será siempre positiva y sólo tendrá un valor nulo cuando todas las observaciones coincidan con el valor de la media.
✓La desviación típica no es una medida robusta de la dispersión.
✓ El hecho de que se calcule evaluando los cuadrados de las desviaciones hace que sea muy sensible a observaciones extremas, bastante más que la desviación media (dado que aparece un cuadrado).
✓La desviación típica no es una buena medida de dispersión cuando se tiene algún dato muy alejado de la media.

Propiedades

La desviación típica, siempre es positiva y sólo tendrá un valor nulo cuando todas las observaciones coincidan con el valor de la media

Ejercicio. Supongamos que el numero de hijos de una muestra de 20 familias es la siguiente

4	1	1	3	1	2	5	1	2	3
2	2	3	2	1	4	2	3	2	1

x_i	n_i	$x_i \times n_i$	$x_i^2 \times n_i$
1	6	6	**6**
2	7	14	**28**
3	4	12	**36**
4	2	8	**32**
5	1	5	**25**

$$s^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i - \frac{1}{N}\left(\sum_{i=1}^{k} x_i n_i\right)^2}{N-1}$$

$$s^2 = \frac{\sum_{i=1}^{5} x_i^2 n_i - \frac{1}{20}\left(\sum_{i=1}^{5} x_i n_i\right)^2}{N-1}$$

$$\sum_{i=1}^{5} x_i^2 n_i = 6 + 28 + 36 + 32 + 25 = 127$$

$$\frac{1}{20}\left(\sum_{i=1}^{5} x_i n_i\right)^2 = \frac{(6+14+12+8+5)^2}{20} = \frac{45^2}{20} = 101.25$$

$$s^2 = \frac{127 - 101.25}{20-1} = 1.355 \Rightarrow \boxed{s = \sqrt{1.355} = 1.16}$$

2. 4 Coeficientes de variación

Un problema que plantean las medidas de dispersión vistas anteriormente es que vienen expresadas en las unidades en que se ha medido la variable. Para solucionar esto, se definen unas medidas de dispersión relativas, independientes de la unidades usadas. *Estas dispersiones relativas van a permitir además comparar la dispersión entre diferentes muestras (con unidades diferentes).*

2.4. 1Coeficiente de variación de Pearson

Definido como el cociente entre la desviación típica y la media aritmética

$$CV = \frac{s}{|\bar{x}|}$$

Cuanto mayor sea CV**, mayor dispersión tendrán los datos**. Normalmente se expresa en porcentaje

En general, el coeficiente de variación es un estadístico útil para comparar la variabilidad de variables que tienen desviaciones estándar distintas y medias distintas.

2.4.2 Coeficiente de variación media

Es similar al coeficiente de variación de Pearson, pero empleando una desviación media en vez de la media aritmética. Se tienen entonces dos coeficientes de variación media dependiendo de que se calcule respecto a la desviación media respecto a la media aritmética o respecto a la mediana

$$CVM_{\bar{x}} = \frac{D_{\bar{x}}}{|\bar{x}|} \qquad CVM_{M_e} = \frac{D_{Me}}{|M_e|}.$$

Ejercicio. Supongamos que el numero de hijos de una muestra de 20 familias es la siguiente

4	1	1	3	1	2	5	1	2	3
2	2	3	2	1	4	2	3	2	1

$s = 1.16$

$\bar{x} = 2.25$

$$CV = \frac{s}{|\bar{x}|} = \frac{1.16}{2.25} \qquad CV = 0.516 \Rightarrow \qquad CV = 52\%$$

Sueldo inicial de profesores

Profesor	1	2	3	4	5	6	7	8	9	10	11	12
Sueldo x_i	3450	3550	3650	3480	3355	3310	3490	3730	3540	3925	3520	3480
$(x_i - \bar{x})^2$	8100	100	12100	3600	34225	52900	2500	36100	0	148225	400	3600

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{n} \qquad \bar{x} = 3540$$

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

$$s = \sqrt{s^2} = \sqrt{\frac{301850}{11}} \Rightarrow s = 165.65$$

$$CV = \frac{s}{|\bar{x}|} = \frac{165.65}{3540} \Rightarrow CV = 0.046$$

$$CV = 0.046 \times 100 = 4.67\%$$

3 Momentos

La media aritmética y la varianza, son en realidad casos particulares de una definición más general

Si tenemos una muestra de la variable estadística, x, la cual toma los valores $x_1, x_2, \ldots, x_k$ con frecuencias absolutas $n_1, n_2, \ldots, n_k$, se define el momento de orden r respecto al parámetro c como

$$M_r(c) = \frac{\sum_{i=1}^{k} (x_i - c)^r n_i}{N}$$

3.1 Momentos respecto al origen ($c = 0$)

Se define el momento de orden r respecto al origen como

$$a_r = \frac{\sum_{i=1}^{k} x_i^r n_i}{N}$$

Los momentos respecto al origen suministran entonces medidas de tendencia central

Los primeros momentos respecto al origen son

$$a_0 = \frac{\sum_{i=1}^{k} n_i}{N} = 1 \qquad a_1 = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \bar{x} \qquad a_1 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} = \bar{x}_Q{}^2$$

La media aritmética es el momento de primer orden respecto al origen.

Momentos respecto a la media

Sustituyendo c por la media aritmética. Se define los momento de orden r respecto a la media

$$m_r = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^r n_i}{N},$$

donde los primeros momentos son entonces

$$m_0 = \frac{\sum_{i=1}^{k} n_i}{N} = 1. \qquad m_1 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})\, n_i}{N} = 0. \qquad m_2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i}{N} = \frac{N-1}{N} s^2.$$

El momento de orden 1 se anula por la propiedad de la media aritmética expresada anteriormente. Puede observarse que el momento de orden 2 respecto a la media es, aproximadamente, la varianza

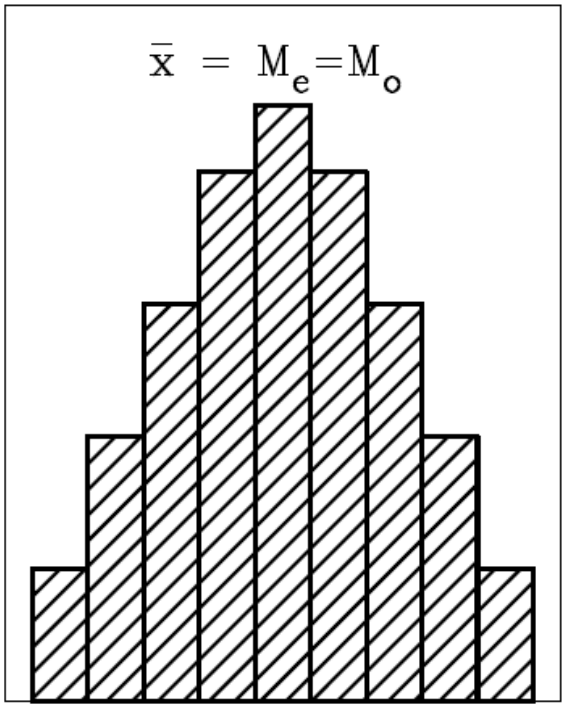
4 Asimetría y Curtosis

La descripción estadística de una muestra de datos no concluye con el cálculo de su tendencia central y su dispersión. Para dar una descripción completa es necesario estudiar también el grado de simetría de los datos respecto a su medida central y la concentración de los datos alrededor de dicho valor.

Se dice que una distribución de medidas es simétrica cuando valores de la variable equidistantes, a uno y otro lado, del valor central tienen la misma frecuencia. Es decir, en este caso tendremos simetría en el histograma (o en el diagrama de barras) alrededor de una vertical trazada por el punto central

Una distribución perfectamente simétrica. Media aritmética, mediana y moda coinciden

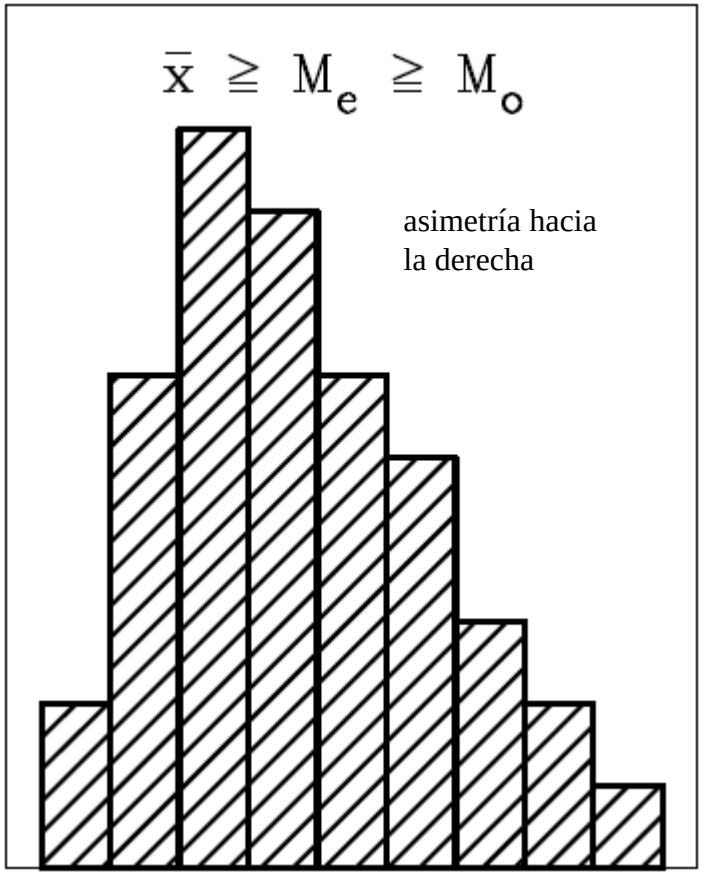
$$\bar{x} = M_e = M_o$$



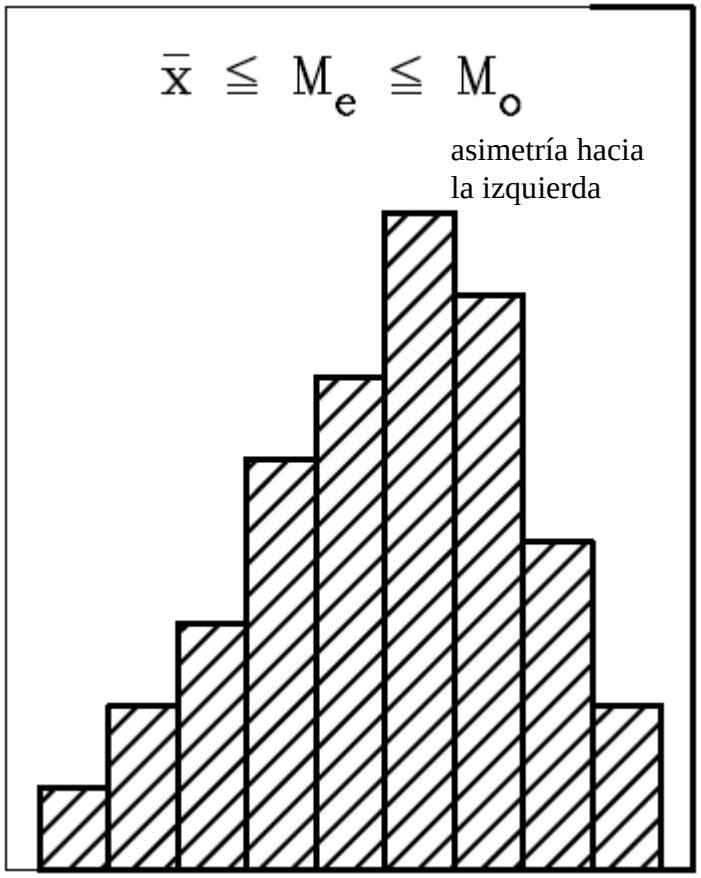
En el caso de no tener simetría, diremos que tenemos asimetría a la derecha (o positiva) o a la izquierda (o negativa) dependiendo de que el histograma muestre una cola de medidas hacia valores altos o bajos de la variable respectivamente. También se puede decir que la distribución está sesgada a la derecha (sesgo positivo) o a la izquierda (sesgo negativo)

una distribución asimétrica La media, mediana y moda no coinciden

Asimetría positiva $\bar{x} \geq M_e \geq M_o$



Asimetría negativa $\bar{x} \leq M_e \leq M_o$



4.1 Coeficientes de Asimetría

Con el fin de cuantificar el grado de asimetría de una distribución se pueden definir los coeficientes de asimetría. Aunque no son los únicos, existen dos coeficientes principales:

4.1.1 Coeficiente de asimetría de Fisher

Se define como el cociente entre el momento de orden 3 respecto a la media y el cubo de la desviación típica

$$g_1 = \frac{m_3}{s^3} \qquad m_3 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^3 n_i}{N}.$$

Distribución simétrica

Las desviaciones respecto a la media se anularán (puesto que en $m_3$el exponente es impar se sumarán números positivos y negativos)

$$g_1 = 0$$

Distribución Asimétrica

g_1 Tendrá valores positivos para una asimetría positiva (a la derecha)

g_1 Tendrá valores negativos cuando la asimetría sea negativa (izquierda)

La división por el cubo de la desviación típica se hace para que el coeficiente sea adimensional y, por lo tanto, comparable entre diferentes muestras.

4.1.2 Coeficiente de asimetría de Pearson

Este coeficiente, también adimensional, se define como

$$A_P = \frac{\bar{x} - M_o}{s}.$$

Distribución simétrica

$A_P = 0$ media y moda coinciden

Distribución Asimétrica

$A_P > 0$ La distribución sesgada esté hacia la derecha

$A_P < 0$ La distribución sesgada esté hacia la izquierda

Ejercicio. Supongamos que el numero de hijos de una muestra de 20 familias es la siguiente

4	1	1	3	1	2	5	1	2	3
2	2	3	2	1	4	2	3	2	1

$s = 1.16$

$\bar{x} = 2.25$

x_i		$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
1	6	-1.25	-11.718
2	7	-0.25	-0.109
3	4	0.75	1.687
4	2	1.75	10.718
5	1	2.75	20.796
Total	20		21.373

Coeficiente de asimetría de Fisher

$$m_3 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^3 n_i}{N}, \quad \sum_{i=1}^{k} (x_i - \bar{x})^3 n_i = 21.373$$

$$m_3 = \frac{21.373}{20} = 1.068 \qquad s^3 = 1.561$$

$$g_1 = \frac{m_3}{s^3} = \frac{1.068}{1.561} \Rightarrow g_1 = 0.68$$ **Positiva**

Coeficiente de asimetría de Pearson

$$A_P = \frac{\bar{x} - M_o}{s}.$$

$$A_P = \frac{2.25 - 2}{1.16}. \Rightarrow A_P = 0.215$$ **Positiva**

4.2 Curtosis

Además de la simetría, otra característica importante de la forma en que se distribuyen los datos de la muestra es cómo es el agrupamiento en torno al valor central.

➢Los datos se pueden distribuir de forma que tengamos un gran apuntamiento (o pico en el histograma) alrededor del valor central, en cuyo caso diremos que tenemos una **distribución leptocúrtica.**

➢o en el extremo contrario, el histograma puede ser muy aplanado, lo que corresponde a una **distribución platicúrtica**

➢el caso intermedio, diremos que la **distribución es mesocúrtica** y el agrupamiento corresponderá al de una distribución llamada normal, o en ***forma de campana de Gauss***

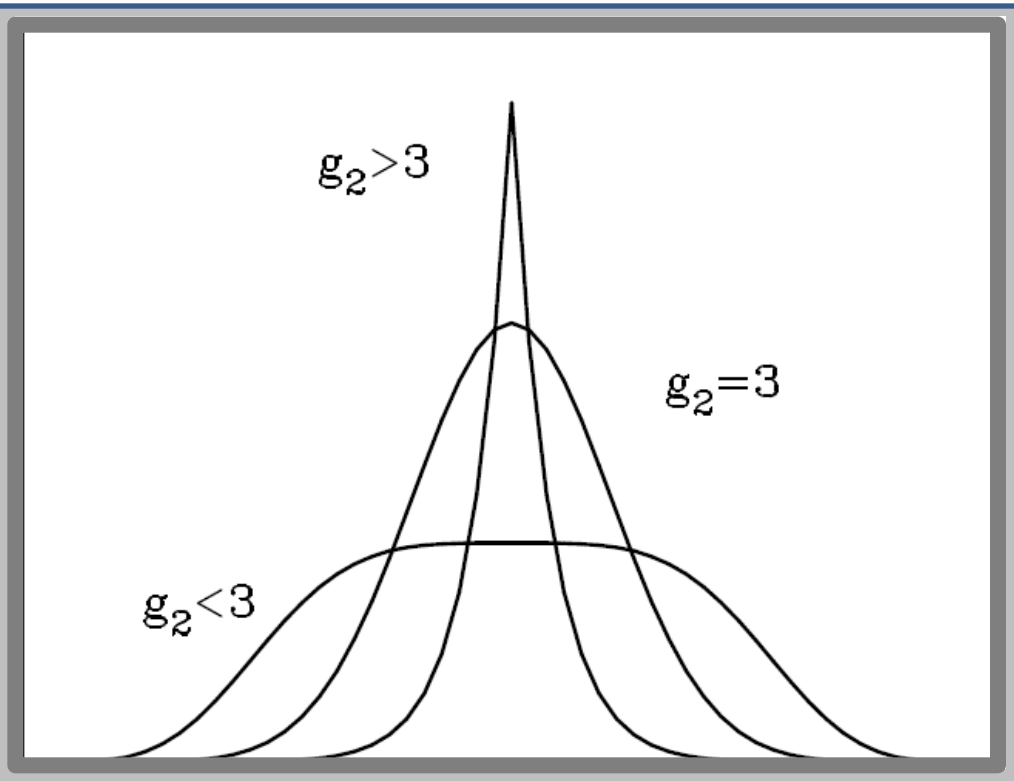
Esta característica del agrupamiento de los datos se denomina Curtosis

4.2.1 Coeficiente de Curtosis

Se define como el cociente entre el momento de cuarto orden respecto a la media y la cuarta potencia de la desviación típica

$$g_2 = \frac{m_4}{s^4}. \qquad m_4 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^4 n_i}{N}.$$

Este coeficiente adimensional alcanza valores mayores cuanto mas puntiaguda es la distribución



Distribuciones con diferente grado de apuntamiento: leptocúrtica ($g_2 > 3$), mesocúrtica ($g_2 = 3$) y platicúrtica ($g_2 < 3$).