

Estadística Descriptiva e Introducción a la Probabilidad

Juan Antonio Maldonado Jurado

Departamento de Estadística e Investigación Operativa
Universidad de Granada



UNIVERSIDAD
DE GRANADA

Doble Grado en Ingeniería Informática y Matemáticas

Tema 1. Introducción a la Estadística. Estadística descriptiva unidimensional

¿Qué es la Estadística?

Estadística Descriptiva: conceptos básicos

Distribución de frecuencias de una variable estadística unidimensional

Tablas estadísticas

Representaciones gráficas

Características unidimensionales

Medidas de posición

Medidas de dispersión

 Medidas de dispersión absolutas

 Medidas de dispersión relativas

Momentos

Medidas de forma

 Medidas de asimetría

 Medidas de apuntamiento o curtosis

“La Estadística es la ciencia que estudia cómo debe emplearse la información y cómo dar una guía de acción en situaciones prácticas que envuelven incertidumbre” V. Barnett (1973)

La Estadística no debe entenderse simplemente como un conjunto de valores numéricos, ya que hoy por hoy está constituida como una Ciencia que no sólo facilita métodos precisos para obtener información numérica y describirla (lo que se denomina la *Estadística Descriptiva*), sino que además proporciona métodos para analizar esta información desde distintos puntos de vista y obtener conclusiones de la misma (*Estadística Inferencial*), ayudando así al proceso de toma de decisiones y a la investigación en cualquier otra ciencia.

En términos generales, podemos decir que *el objetivo de la Estadística es crear (obtener) conocimiento sobre algún tópico, situación o fenómeno real, recogiendo información sobre el mismo, analizando, sintetizando e interpretando (dando sentido) dicha información.*

Documento de trabajo

Estadística Descriptiva. Conceptos básicos

- ▶ **Fenómenos determinísticos:** aquellos que dan lugar al mismo resultado siempre que se realicen bajo idénticas condiciones.
- ▶ **Fenómenos aleatorios:** se caracterizan porque sus resultados pueden variar, incluso si el experimento se realiza bajo idénticas condiciones iniciales.
- ▶ **Población, colectivo o universo:** conjunto de unidades o elementos con alguna/s característica/s en común, sobre el que se desea obtener cierta información.
- ▶ **Muestra:** subconjunto de la población elegido en términos de representatividad.
- ▶ **Carácter o característica estadística:** toda propiedad que se desea estudiar en la población, que debe poder observarse sobre todos y cada uno de los individuos que la componen.
 - ▶ **Modalidad:** formas posibles en las que se puede manifestar el carácter), y cada individuo o unidad de la población debe presentar una y sólo una de ellas.
 - ▶ **Carácter cualitativo (atributo):** aquél cuyas modalidades no son medibles o cuantificables por números.
 - ▶ **Carácter cuantitativo:** aquél cuyas modalidades son numéricamente medibles o cuantificables.

resultado de una medida, visto los datos en mi saco

Estadística Descriptiva. Conceptos básicos

- ▶ **Escalas de medida:** La realización de cualquier estudio estadístico requiere, como primer paso, la identificación precisa de las modalidades del carácter bajo estudio y la asignación de símbolos o números a las distintas modalidades; esto es lo que se denomina *medición* del carácter.

Si denotamos por X al carácter, y A y B son dos individuos cuyas medidas de X son x_A y x_B , se distinguen cuatro tipos de escala:

- ▶ **Escala nominal:** Sólo se puede decir que $x_A = x_B$, o bien que $x_A \neq x_B$.
- ▶ **Escala ordinal:** No sólo se puede decir $x_A = x_B$ ó $x_A \neq x_B$, sino, en este caso, que $x_A < x_B$ ó $x_A > x_B$.
- ▶ **Escala de intervalo:**

En este caso, se podría decir que $x_A = x_B$, $x_A \neq x_B$, $x_A < x_B$, $x_A > x_B$, y, además, que A es $x_A - x_B$ unidades diferente (superior o inferior) que B .

- ▶ **Escala de razón:**

Se puede decir ya que A es x_A/x_B veces superior a B .

→ *ejemplos reflejados en el aterisco.*

	<u>X</u>
	Sexo
x_A	A Mujer
x_B	B Hombre

$x_A = x_B$ ó $x_A \neq x_B \rightarrow$ Escala nominal

X

Nivel de estudios

x_A	A	Primario	\rightarrow Escala de orden que no inventó
x_B	B	Universitario	$x_A < x_B$
		/	

Describo lo que hay, no todos los posibles,
a veces que sean variables abiertas

X

Tallo de maíz

$$48 - 36 = 12$$

x_A A 48

Diferencia de 6 Tallos

x_B B 36

Pero hay que saber más

Escala de intervalos A es $x_A - x_B$ veces def a B

Habilidades o votos

X

Notas

y_A	A	80	A es $\frac{x_A}{x_B}$ veces B \rightarrow Razón.
x_B	B	60	

\bar{X} , n , $x_1, x_2, x_3, \dots, x_k$ frecuencia absoluta
 $n_1, n_2, n_3, \dots, n_k$ número de elementos
cuantas veces los contado (elementos seco)
frecuencia relativa

$$\sum_{i=1}^k n_i = n$$

$$f_i = \frac{n_i}{n}$$

$$\sum_{i=1}^k f_i = \frac{\sum_{i=1}^k n_i}{n} = \approx 1$$

x_i se puede poner con intervalos

Si pudieran ordenar x_i

- $N_i = \sum_{j=1}^i n_j$, & seva hasta ese

$$- F_i = \frac{N_i}{n} \equiv F_i = \sum_{j=1}^i f_j$$

Estadística Descriptiva. Conceptos básicos

- ▶ **Variable:** En general, una variable es un símbolo que representa a distintos valores numéricos. Cuando estos valores son el resultado de mediciones u observaciones estadísticas, hablaremos de *variable estadística*. Así, un *carácter cuantitativo* irá representado por una *variable estadística*, y sus diversas modalidades serán los valores que toma dicha variable (hay que indicar que, por abuso del lenguaje, los caracteres cualitativos son también referidos como *variables cualitativas*). Existen diferentes clasificaciones. Entre otras:
 - ▶ **variables discretas:** diremos que una variable estadística es discreta si el paso de un valor de la variable al siguiente representa un salto (el conjunto de números reales que soporta la variable está formado sólo por puntos aislados)
 - ▶ **variables continuas:** diremos que una variable estadística es continua si *a priori* puede tomar cualquier valor entre dos valores dados; es decir, puede tomar todos los valores comprendidos en un intervalo de la recta real.
 - ▶ Por otra parte, según el número de caracteres cuantitativos que se estudia en cada unidad observada, las variables serán de distinta dimensión y hablaremos de variables *unidimensionales*, *bidimensionales*, *tridimensionales*, etc.

Distribución de frecuencias de una variable estadística unidimensional

Supongamos una población estadística de n elementos o individuos, en la que se desea estudiar una variable (o atributo) que presenta los valores (o modalidades) x_1, x_2, \dots, x_k . Se denomina:

- ▶ *Frecuencia absoluta del valor o modalidad x_i* al número total de individuos en la población que presenta dicho valor (modalidad). Se suele notar por n_i , $i = 1, \dots, k$.
- ▶ *Frecuencia relativa del valor o modalidad x_i* a la proporción del número de individuos que presenta dicho valor (modalidad), $f_i = \frac{n_i}{n}$, $i = 1, \dots, k$.

$$\sum_{i=1}^k n_i = n; \quad \sum_{i=1}^k f_i = 1.$$

Si las modalidades se pueden ordenar, supuesto $x_1 < x_2 < \dots < x_k$, se denomina:

- ▶ *Frecuencia absoluta acumulada del valor o modalidad x_i* al número de individuos que presentan un valor (modalidad) menor o igual que x_i , $N_i = \sum_{j=1}^i n_j$, $i = 1, \dots, k$.
- ▶ *Frecuencia relativa acumulada del valor o modalidad x_i* a la proporción de individuos que presentan un valor (modalidad) menor o igual que x_i .

$$F_i = \frac{N_i}{n} = \sum_{j=1}^i f_j, \quad i = 1, \dots, k.$$

Distribución de freq. de una v. estadística

En una población de Tamaño "n" SE HA OBSERVADO una variable estadística X que HA PRESENTADO k modalidades distintas x_1, x_2, \dots, x_k ; cada una de ellas con freq. absoluta $n_1, n_2, \dots, n_i, \dots, n_k$ con distribución de frecuencias $\{x_i; n_i\}_{i=1, \dots, k}$

Ya tengo los datos !!!

$$n, Y = a\bar{x} + b; \quad y_i = ax_i + b; n_i \quad \{i=1, \dots, k\}$$

$$\bar{y} = a\bar{x} + b \quad \bar{x} = \frac{\bar{y} - b}{a}$$

Distribución de frecuencias de una variable estadística unidimensional

Se denomina **distribución de frecuencias** de una variable (o atributo) al conjunto formado por cada uno de los valores (modalidades) junto con sus frecuencias. Según el tipo de frecuencias consideradas hablaremos de:

- ▶ Distribución de frecuencias absolutas: $\{(x_i, n_i); i = 1, \dots, k\}$
- ▶ Distribución de frecuencias relativas: $\{(x_i, f_i); i = 1, \dots, k\}$
- ▶ Distribución de frecuencias absolutas acumuladas: $\{(x_i, N_i); i = 1, \dots, k\}$
- ▶ Distribución de frecuencias relativas acumuladas: $\{(x_i, F_i); i = 1, \dots, k\}$.

14/02/20

Documento de trabajo

Tablas estadísticas

Usualmente todas las frecuencias se suelen presentar en la misma tabla estadística que, según el tipo de carácter o variable considerado, tiene una de las siguientes formas:

Variables discretas y atributos

Modalidades	Frec. Abs.	Frec. Rel.	Frec. Abs. Acum.	Frec. Rel. Acum.
x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	f_2	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_i	f_i	$N_i = n_1 + n_2 + \dots + n_i$	$F_i = f_1 + f_2 + \dots + f_i$
\vdots	\vdots	\vdots	\vdots	\vdots
x_k	n_k	f_k	$N_k = n_1 + n_2 + \dots + n_k = n$	$F_k = f_1 + f_2 + \dots + f_k = 1$

Si lo tienen a una amplitud $\rightarrow d_i = \frac{u_i}{a_i} \rightarrow$ *densidad de frecuencia.*

Intervalos	Marcas	Amplitud	Frec. Abs.	Frec. Rel.	Frec. Acum.
$(e_0, e_1]$	$c_1 = \frac{e_0 + e_1}{2}$	$a_1 = e_1 - e_0$	n_1	f_1	$N_1 = n_1$
$(e_1, e_2]$	$c_2 = \frac{e_1 + e_2}{2}$	$a_2 = e_2 - e_1$	n_2	f_2	$N_2 = n_1 + n_2$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$(e_{i-1}, e_i]$	c_i	a_i	n_i	f_i	$N_i = n_1 + n_2 + \dots + n_i$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$(e_{k-1}, e_k]$	c_k	a_k	n_k	f_k	$N_k = n_1 + n_2 + \dots + n_k = n$

Documento de trabajo

Representaciones gráficas

Según la naturaleza de carácter estudiado, se utilizan distintos tipos de representaciones; las más frecuentes son:

► **Atributos:**

- ▶ Diagrama de sectores
- ▶ Diagrama de rectángulos (o de barras)
- ▶ Pictograma

*{ Son importantes
son los*

► **Variables discretas:**

- ▶ Diagrama de barras
- ▶ Curva acumulativa o de distribución

► **Variables continuas:**

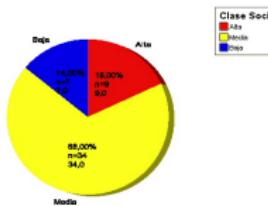
- ▶ Histograma
- ▶ Polígonal de frecuencias
- ▶ Curva acumulativa o de distribución

Documento de trabajo

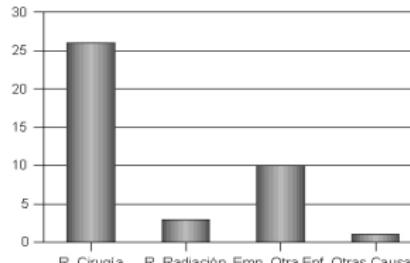
Representaciones gráficas

Atributos:

- ▶ **Diagrama de sectores:** Es un círculo dividido en tantos sectores circulares como modalidades tenga el carácter, siendo el área de cada uno proporcional a la frecuencia absoluta o relativa de la modalidad.



- ▶ **Diagrama de rectángulos o barras:** Consiste en varios rectángulos (uno por modalidad) de base constante y alturas proporcionales a las frecuencias (absolutas o relativas) de cada modalidad.

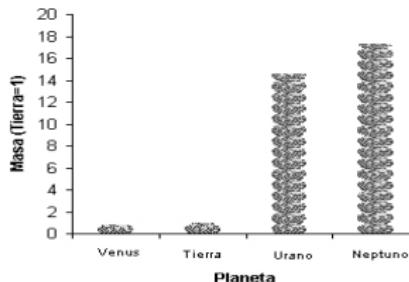
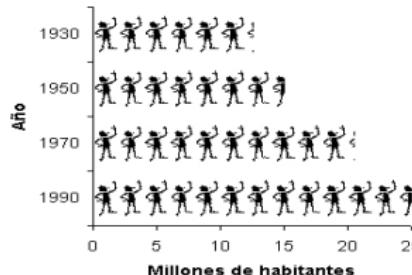


Documento de trabajo

Representaciones gráficas

Atributos:

- ▶ **Pictograma:** Se dibujan figuras, normalmente alusivas al carácter que se está estudiando, bien una para cada modalidad con tamaño proporcional a su frecuencia, o bien repitiendo la figura tantas veces como requieran las frecuencias.



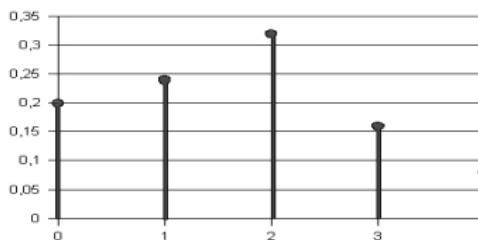
Documento de trabajo

Representaciones gráficas

Variables discretas:

- ▶ **Diagrama de barras:** Similar al de atributos: en un sistema de ejes cartesianos se representa en el eje de abcisas los valores de la variable, y se trazan barras verticales con longitudes proporcionales a sus frecuencias (absolutas o relativas).

x_i	n_i	$f_i = n_i/100$
0	20	0,2
1	24	0,24
2	32	0,32
3	16	0,16
4	8	0,08



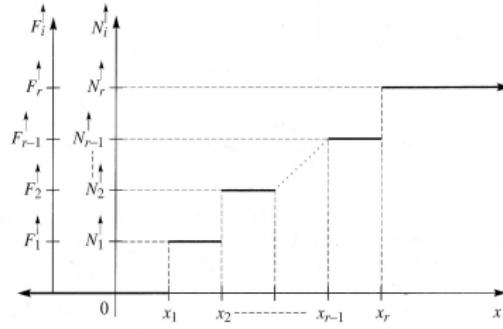
Documento de trabajo

Representaciones gráficas

Variables discretas:

- ▶ **Curva acumulativa o de distribución:** Es la representación de la denominada *función acumulativa, de repartición o de distribución*, que es una función definida para cada número real, x , como la proporción de datos menores o iguales que x . Así, si $x_1 < x_2 < \dots < x_k$ son los valores de la variable ordenados,

$$F(x) = \begin{cases} 0 & \forall x < x_1 \\ \frac{\sum_{j=1}^i n_j}{n} = \sum_{j=1}^i f_j = \frac{N_i}{n} = F_i & \forall x / x_i \leq x < x_{i+1} \\ 1 & \forall x \geq x_k \end{cases}$$



Documento de trabajo

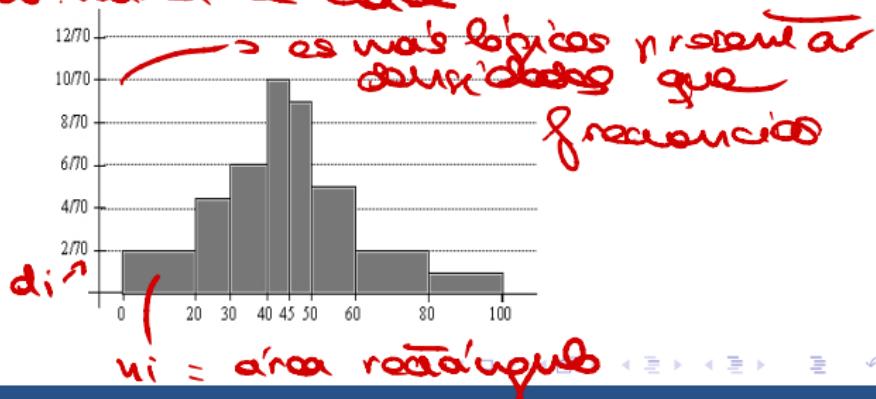
Representaciones gráficas

Variables continuas:

- ▶ **Histograma:** Está formado por rectángulos yuxtapuestos, cuyas bases son las diferentes clases o intervalos de definición de la variable, y cuyas alturas son las frecuencias medias $h_i = f_i/a_i$ por unidad de amplitud (o $h_i = n_i/a_i$), también denominadas densidades de frecuencia.

I_i	n_i	f_i	h_i	Altura $f_i = 5h_i$
(0, 20]	8	8/70	4/700	2/70
(20, 30]	9	9/70	9/700	4,5/70
(30, 40]	12	12/70	12/700	6/70
(40, 45]	10	10/70	20/700	10/70
(45, 50]	9	9/70	1,8/700	9/70
(50, 60]	10	10/70	10/700	5/70
(60, 80]	8	8/70	4/700	2/70
(80, 100]	4	4/70	2/700	1/70

Otro intervalo marca de clase

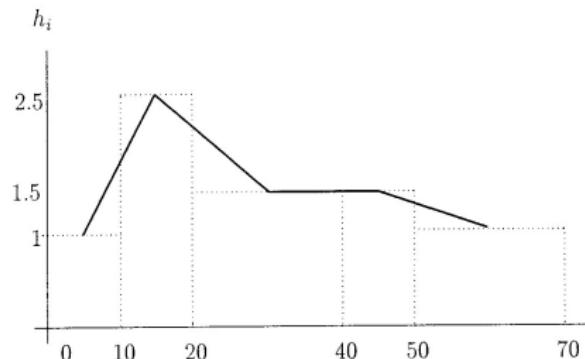


Documento de trabajo

Representaciones gráficas

Variables continuas:

- ▶ **Polygonal de frecuencias:** Es la poligonal que resulta de unir los puntos correspondientes a los techos de las marcas de clase de los intervalos en el histograma.



Documento de trabajo

Representaciones gráficas

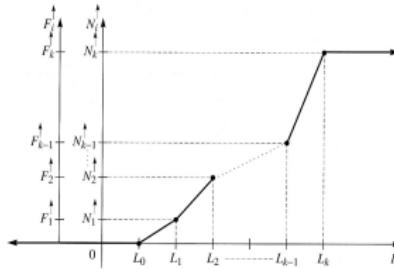
Variables continuas:

► Curva acumulativa o de distribución:

Análogamente al caso de las variables discretas, la función de distribución $F(x)$ es la proporción (o total) de individuos en la población cuyo valor de la variable es inferior o igual a x . Esta función, en este caso, se conoce únicamente para los valores de x que son extremos superiores de cada intervalo:

$$F(e_i) = \sum_{j=1}^i f_j$$

es monótona no decreciente, con $F(-\infty) = 0$ y $F(+\infty) = 1$ (o n , si trabajamos con frecuencias absolutas).



Documento de trabajo

Características unidimensionales

Datos → Tabla → Gráfico → Estadístico
↓
Medida estadística

Entendemos por **medidas, características estadísticas o estadísticos** resúmenes cuantitativos de los datos que reflejan la información de los mismos, haciendo más fácil su interpretación y facilitando, además, la comparación entre distintos conjuntos de datos.

Propiedades deseables (G.U. Yule (1857-1951)): *No importa el conjunto, va el dato*

1. Deben definirse de manera objetiva, de forma que dos personas diferentes deben dar iguales resultados.
2. Deben usar todas las observaciones y no algunas solamente.
3. Deben tener un significado concreto, para que sean rápida y fácilmente interpretables.
4. Deben ser sencillas de calcular.
5. Deben prestarse fácilmente al cálculo algebraico.
6. Deben ser poco sensibles a fluctuaciones muestrales, de forma que si cambiaseen valores extremos de los datos, no cambiaseen en gran medida las características del conjunto.

→ Ejemplo media y mediana.

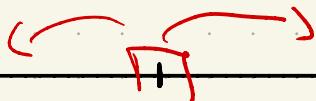
Tipos:

- **Medidas de posición:** permiten situar o localizar una distribución en la recta real. Las más importantes son las de centralización o tendencia central, también denominadas **promedios** (que proporcionan un valor central representativo, alrededor del cual se agrupan los datos) y **los cuantiles** (que proporcionan valores representativos de partes de la distribución). *o estadísticas relativas*
- **Medidas de dispersión:** miden el grado de **esparcimiento** de los datos de una distribución.
- **Medidas de forma:** caracterizan de manera precisa la **forma** de una distribución sin tener que llevar a cabo su representación gráfica.

Posición

Tendencia central	Media (cada variable se tiene una)
	Mediana
	Moda

Tendencia no central



Dispersión: como se parecen los valores a ese central que los representa.

- Absoluta: valores y unidades de medida
- Relativa: % uds. de medida (cociente)

Forma

Asimetría
Curtosis

Características unidimensionales

Medidas de posición

Media aritmética

Definición: La media aritmética de una variable (o de su distribución de frecuencias) es la suma de todos los valores de la variable dividida por el número total de observaciones.

- Si consideramos una variable estadística discreta en una población de tamaño n , con distribución de frecuencias $\{(x_i, n_i(f_i)); i = 1, \dots, k\}$, la media aritmética es

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

ordenados de menor a mayor

~~esta no se usa~~

- Si la variable es continua y los datos están agrupados en *intervalos de clase*, la distribución de frecuencias es del tipo $\{(I_i, n_i(f_i)); i = 1, \dots, k\}$, y la media aritmética se calcula suponiendo que todos los datos de cada intervalo I_i son idénticos al centro o *marca de clase*, c_i :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i c_i = \sum_{i=1}^k f_i c_i.$$

Documento de trabajo

Características unidimensionales
Medidas de posición

Media aritmética

Propiedades:

- La media aritmética está acotada por los valores extremos de la variable; esto es:
 $x_1 \leq \bar{x} \leq x_k$. $x_1 < \bar{x} < x_k$ No puede valer una sola
- La media aritmética de las desviaciones de los datos respecto de la media aritmética es igual a cero:
nuevos datos

$$\sum_{i=1}^k f_i(x_i - \bar{x}) = 0. = \sum_i f_i x_i - \bar{x} \sum_i f_i$$

- Si se somete una variable X a una transformación lineal afín, la media aritmética de la nueva variable es la imagen de la media de X por la misma transformación:

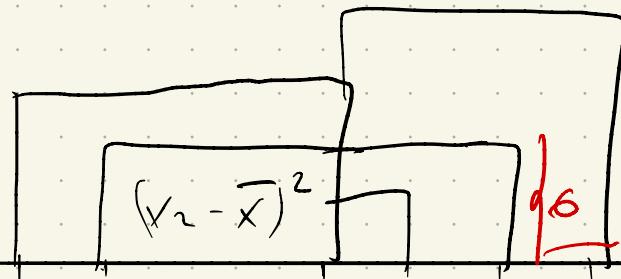
$$Y = aX + b \implies \bar{y} = a\bar{x} + b.$$

$$\bar{y} = \sum f_i y_i = \sum f_i (ax_i + b) = a \sum f_i x_i + b$$

- La media aritmética de los cuadrados de las desviaciones respecto a la media aritmética es mínima:

$$\sum_{i=1}^k f_i(x_i - \bar{x})^2 < \sum_{i=1}^k f_i(x_i - a)^2, \quad \forall a \neq \bar{x}.$$

varianza

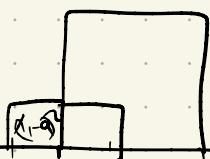


~ Son áreas

desviación típica

$$x_1 \ x_2 \ \dots \ \bar{x} \ x_8 \ x_9 \ x_{10}$$

Usa siempre la aritmética,
no su media real



$$x_1 \ x_2 \ \dots$$

El de la media (el área total) es mínimo, es óptimo si
a vale \bar{x}

$$\text{Mínimo } Q(a) = \sum_{i=1}^n f_i (x_i - a)^2$$

$$Q'(a) = \cancel{\sum} f_i (x_i - a)$$

$$\bar{x} - a = 0$$

$$\bar{x} = a \sim \text{Llega la siguiente derivada}$$

Características unidimensionales

Medidas de posición

Media geométrica

Se usa cuando se desea promediar datos de una variable que tiene efectos multiplicativos acumulativos en la evolución de una determinada característica con un valor inicial fijo.

Definición: Es la raíz n -ésima del producto de los n valores (o marcas de clase) de la distribución:

$$G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}} = \sqrt[n]{\prod_{i=1}^k x_i^{n_i}}$$

Nota: El logaritmo de la media geométrica es la media aritmética de los logaritmos de los valores de la variable:

$$G = e^{\bar{b}_g G}$$

$$\log G = \log \sqrt[n]{\prod_{i=1}^k x_i^{n_i}} = \frac{1}{n} \sum_{i=1}^k n_i \log x_i = \sum_{i=1}^k f_i \log x_i.$$

↳ Esto es más fácil de calcular,
cumple rule.

Média geométrica

$C = \text{capital}$

$$1^{\text{º}} \text{ año } i_1 \quad C_1 = C + i_1 C = C(1+i_1)$$

$$2^{\text{º}} \text{ año } i_2 \quad C_2 = C_1 + C_1 i_2 = C(1+i_1)(1+i_2)$$

⋮

⋮

$$n^{\text{º}} \text{ año } i_n \quad C_n = C_{n-1} + C_{n-1} i_n = \dots = C(1+i_1)(1+i_2) \dots (1+i_n)$$

↓

interés

¿ C_{real} es el interés medio?

Si queremos todos los i_x por i el último capital debe ser igual

$$\sqrt[n]{(1+i_1)(1+i_2) \dots (1+i_n)} = \sqrt[n]{(1+i)}^n$$

$$\sqrt[n]{(1+i_1)(1+i_2) \dots (1+i_n)} = 1+i$$

$$i = \sqrt[n]{(1+i_1)(1+i_2) \dots (1+i_n)} - 1$$

Média geométrica

Características unidimensionales

Medidas de posición

Media armónica

Se usa para promediar datos de magnitudes que son cocientes de dos magnitudes; esto es, magnitudes relativas (su unidad de medida es referida a una unidad de otra variable). Por ejemplo, para promediar velocidades, rendimientos o productividades, etc.

Se define como la inversa de la media aritmética de los valores inversos de la variable:

$$H = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_k}{x_k}} = \frac{n}{\sum_{i=1}^k \frac{n_i}{x_i}}.$$

Media cuadrática

Es la menos utilizada y, fundamentalmente, su uso se reduce al cálculo de promedios sobre superficies. Se define como la raíz cuadrada de la media aritmética de los cuadrados de los valores de la variable:

$$Q = \sqrt{\sum_{i=1}^k f_i x_i^2}.$$

Con esta haré
áreas no relacionadas
con la Desviación Típica

Se puede decir que la disponibilidad de los medios
nos da información, pero no tiene sentido hacer por ejemplo
la aritmética en el uso capital, solo con información.

Características unidimensionales

Medidas de posición

Mediana

La mediana de una distribución es un valor que divide a los individuos de la población en dos efectivos iguales, supuestos ordenados por valor creciente del carácter; esto es, si el conjunto de observaciones se ordena de menor a mayor, la mediana Me es un número que divide esta ordenación en dos partes con el mismo número de datos en cada una.

- ▶ *Variables discretas:* Se busca el primer valor de la variable cuya frecuencia absoluta acumulada sea mayor o igual que $n/2$ (o, equivalentemente, cuya frecuencia relativa acumulada sea mayor o igual que $1/2$):

$$x_i / N_{i-1} < \frac{n}{2} \leq N_i \text{ ó } F_{i-1} < \frac{1}{2} \leq F_i.$$

Pueden presentarse dos situaciones:

- ▶ $N_i > n/2$ (ó $F_i > 1/2$) $\Rightarrow Me = x_i$.
- ▶ $N_i = n/2$ (ó $F_i = 1/2$). En este caso, cualquier número en el intervalo $[x_i, x_{i+1}]$ tiene la misma frecuencia acumulada que x_i y, a efectos prácticos o de comparación entre distintas distribuciones, se conviene en tomar como mediana la media aritmética de ambos.

Documento de trabajo

Características unidimensionales Medidas de posición

Mediana

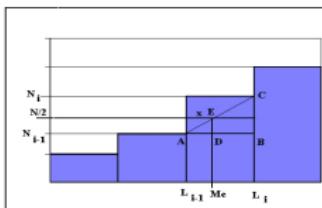
Cuartil orden 2, estudiamos una mitad.

- *Variables continuas:* Se busca

$$I_i = (e_{i-1}, e_i] / N_{i-1} < \frac{n}{2} \leq N_i \text{ ó } F_{i-1} < \frac{1}{2} \leq F_i.$$

- $N_i = n/2$ (ó $F_i = 1/2$) $\Rightarrow Me = e_i$.
- $N_i > n/2$ (ó $F_i > 1/2$). En este caso, la mediana está dentro de I_i , que se denomina el *intervalo mediano*, y para determinarla, se interpola linealmente:

$$Me = e_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i}(e_i - e_{i-1}) = e_{i-1} + \frac{\frac{1}{2} - F_{i-1}}{f_i}(e_i - e_{i-1}).$$



Propiedad: La desviación absoluta media respecto a la mediana es mínima.

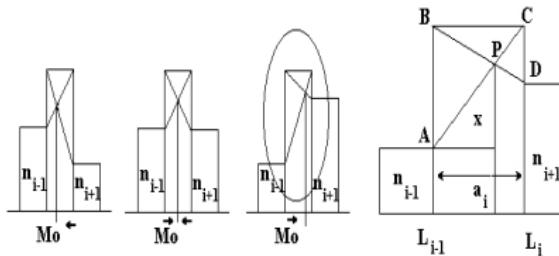
Características unidimensionales

Medidas de posición

Moda

La moda de una distribución es el valor de mayor frecuencia (absoluta o relativa), el que más se repite.

- ▶ **Variables discretas:** La moda es el valor x_i tal que $n_i \geq n_j$ (o, equivalentemente, $f_i \geq f_j$) para cualquier $j = 1, \dots, k$. → **moda valor de las modas**
- ▶ **Variabes continuas:** La moda está en el denominado *intervalo modal*, el de mayor densidad de frecuencia $h_i = f_i/a_i$ o, equivalentemente, el de mayor altura en el histograma.



$$\frac{M_o - e_{i-1}}{h_i - h_{i-1}} = \frac{e_i - M_o}{h_i - h_{i+1}} = \frac{a_i - (M_o - e_{i-1})}{h_i - h_{i+1}} :$$

$$M_o = e_{i-1} + \frac{h_i - h_{i-1}}{2h_i - h_{i-1} - h_{i+1}}(e_i - e_{i-1}).$$

Características unidimensionales

Medidas de posición

Percentiles

El *percentil de orden r* ($r = 1, \dots, 100$): es un valor, P_r , que divide al conjunto ordenado de datos en dos partes, tales que el $r\%$ del total son inferiores o iguales a P_r .

- ▶ *Variables discretas:* Se busca:

$$x_i / N_{i-1} < \frac{nr}{100} \leq N_i \text{ ó } F_{i-1} < \frac{r}{100} \leq F_i.$$

- ▶ $N_i > nr/100$ (ó $F_i > r/100$) $\Rightarrow P_r = x_i$.
- ▶ $N_i = nr/100$ (ó $F_i = r/100$). En este caso, todo número del intervalo $[x_i, x_{i+1})$ es percentil de orden r y, a efectos prácticos o de comparación entre distintas distribuciones, se conviene en tomar como P_r el punto medio de dicho intervalo.

- ▶ *Variables continuas:* Se busca:

$$I_i = (e_{i-1}, e_i] / N_{i-1} < \frac{nr}{100} \leq N_i \text{ ó } F_{i-1} < \frac{r}{100} \leq F_i.$$

- ▶ $N_i = nr/100$ (ó $F_i = r/100$) $\Rightarrow P_r = e_i$.
- ▶ $N_i > nr/100$ (ó $F_i > r/100$). En este caso, se usa la siguiente fórmula de interpolación lineal, obtenida de forma similar a la de la mediana:

$$P_r = e_{i-1} + \frac{\frac{nr}{100} - N_{i-1}}{n_i}(e_i - e_{i-1}). = e_{i-1} + \frac{\frac{r}{100} - F_{i-1}}{f_i}(e_i - e_{i-1})$$

Destacamos entre los percentiles a los *cuartiles*, Q_1, Q_2, Q_3 , que equivalen a los percentiles de orden 25, 50 y 75, respectivamente ($Q_2 = P_{50}$ es la mediana) y los *deciles*, percentiles de orden 10, 20, ..., 90.

26/02/20

Documento de trabajo

Características unidimensionales Medidas de dispersión

- ▶ Medidas de dispersión absolutas:
 - ▶ Recorrido o rango.
 - ▶ Recorrido intercuartílico.
 - ▶ Desviación absoluta media respecto a la media aritmética.
 - ▶ Desviación absoluta media respecto a la mediana.
 - ▶ Varianza.
 - ▶ Desviación típica.
- ▶ Medidas de dispersión relativas:
 - ▶ Coeficiente de apertura.
 - ▶ Recorrido relativo.
 - ▶ Recorrido semi-intercuartílico.
 - ▶ Coeficiente de variación.
 - ▶ Índice de dispersión respecto a la mediana.

Documento de trabajo

Características unidimensionales Medidas de dispersión

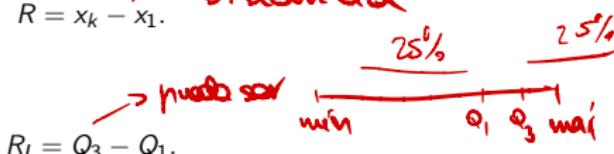
Medidas de dispersión absolutas:

- Recorrido o rango:

Bajo el supuesto de que los valores de la variable estén ordenados en sentido creciente,

$$R = x_k - x_1.$$

→ **Dispersión**



- Recorrido intercuartílico:

Indica la longitud del intervalo en el que está incluido el 50 % central de los datos.

- Desviación absoluta media respecto a \bar{x}

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

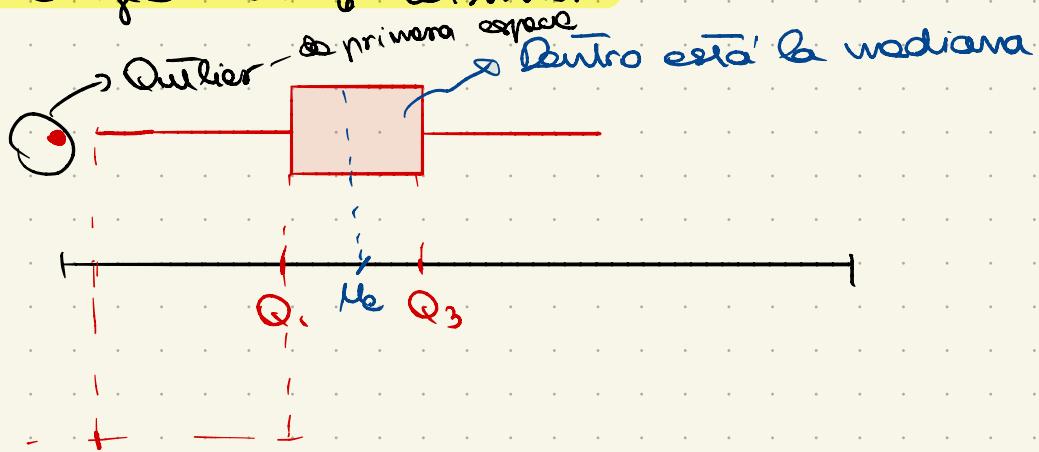
Quedía

- Desviación absoluta media respecto a M_e

$$D_{M_e} = \frac{\sum_{i=1}^k |x_i - M_e| n_i}{n}.$$

Is la desviación media que se suele usar para representar la dispersión del promedio.

Grafico Box & Whisker



Patilla inferior:

$$\min \left\{ \min(x_i), Q_1 - 1,5RI \right\}$$

Patilla superior:

$$\min \left\{ \max(x_i), Q_3 + 1,5RI \right\}$$

Sirve para observar datos atípicos.

Cuanto más pequeño sea RI, más concentrados se tiene al medio.

Documento de trabajo

Características unidimensionales
Medidas de dispersión

Medidas de dispersión absolutas:

- Varianza:

$$\text{Var}(X) = \sigma^2 = \frac{\sum_{i=1}^k n_i(x_i - \bar{x})^2}{n}$$

Desgranar y sacar
 → área media de
 todos los cuadrados ...

Propiedades de la varianza

- La varianza nunca puede ser negativa ($\sigma^2 \geq 0$).
- La varianza es la media cuadrática de dispersión óptima; esto es,

$$\frac{\sum_{i=1}^k n_i(x_i - \bar{x})^2}{n} < \frac{\sum_{i=1}^k n_i(x_i - b)^2}{n}, \forall b \neq \bar{x}.$$

- (Teorema de König). Para cualquier número real a y cualquier variable estadística X , se verifica:

$$\sum_{i=1}^k f_i(x_i - a)^2 = \sum_{i=1}^k f_i(x_i - \bar{x})^2 + (a - \bar{x})^2.$$

s con la media aritmética
 es mínima

$$\text{Así, tomando } a = 0, \text{ se tiene: } \sigma^2 = \sum_{i=1}^k f_i(x_i - \bar{x})^2 = \sum_{i=1}^k f_i x_i^2 - \bar{x}^2.$$

→ demostrar

- La varianza está acotada superior e inferiormente en cada distribución de frecuencias.

- Si se someten los datos a una transformación afín $y_i = ax_i + b$, $i = 1, \dots, k$, la varianza de los datos transformados es $\sigma_y^2 = a^2 \sigma_x^2$.

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n (ax_i + b - \bar{y})^2 = a^2 \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = a^2 \sigma_x^2$$

21/02/20

Documento de trabajo

Características unidimensionales Medidas de dispersión

Medidas de dispersión absolutas:

- **Desviación típica:** Se define como la raíz cuadrada positiva de la varianza

$$\sigma = +\sqrt{\sigma^2}.$$

Propiedades de la desviación típica

- Es no negativa ($\sigma \geq 0$).
- Es una medida de dispersión óptima ($\sigma < \sqrt{\sum_{i=1}^k f_i(x_i - b)^2}, \forall b \neq \bar{x}$).
- Está acotada superior e inferiormente en cada distribución
- No se ve afectada por cambios de origen.
- Sí se ve afectada por cambios de escala: $Y = X/a \Rightarrow \sigma_y = \frac{1}{|a|} \sigma_x$.
- $D_{Me} < D_{\bar{x}} < \sigma$.

Documento de trabajo

Características unidimensionales
Medidas de dispersión

Medidas de dispersión relativas:

- Coeficiente de apertura: Se define como el cociente entre los dos valores extremos de una distribución. Supuestos ordenados crecientemente los valores, su expresión matemática es

$$C_A = \frac{x_k}{x_1}$$

Si x_1 es 0 no se hace

- Recorrido relativo: Se define como el cociente entre el recorrido y la media aritmética

$$R_R = \frac{R}{\bar{x}} = \frac{x_k - x_1}{\bar{x}}.$$

→ no se usa mucho.

- Recorrido semi-intercuartílico: Se define como el cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartil

$$R_{SI} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

→ cuánto mayor sea mayor variabilidad del 50% de la desv.

- Coeficiente de variación de Pearson: Se define como la relación por cociente entre la desviación típica y la media

$$C.V.(X) = \frac{\sigma_x}{|\bar{x}|}.$$

Variabilidad en Término a la media.

- Índice de dispersión respecto a la mediana: Se define como el cociente entre la desviación absoluta media respecto a la mediana, y la mediana

$$V_{Me} = \frac{D_{Me}}{Me}.$$

$$Y = ax + b$$

$$\frac{\partial Y}{\partial x} = \frac{a\cancel{6x}}{a\cancel{x} + b} \neq \frac{6x}{x}$$

raiz positiva

↓
Divido por \cancel{x}

$$\frac{|a| \frac{\partial x}{x}}{a \frac{x}{x} + b \frac{x}{x}} = \frac{|a| CV(x)}{a + \frac{b}{x}}$$

Características unidimensionales

Momentos

Definición: Sea r un número entero y positivo. Se llama **momento de orden r** respecto al valor "a" a la cantidad

$${}_a m_r = \sum_{i=1}^k f_i (x_i - a)^r = \frac{1}{n} \sum_{i=1}^k n_i (x_i - a)^r.$$

Según los valores de "a" se definen dos clases de momentos:

- ▶ **Momentos no centrales**, o momentos respecto del origen: en este caso, el valor $a = 0$, y se denotan por

$$m_r = \sum_{i=1}^k f_i x_i^r.$$

- ▶ **Momentos centrales**, o momentos respecto de la media aritmética: en este caso, el valor $a = \bar{x}$, y se denotan por

$$\mu_r = \sum_{i=1}^k f_i (x_i - \bar{x})^r.$$

Momentos

$a \in \mathbb{R}$ $(x_i - a)^r \rightarrow$ es jugar con esta diferencia

$$a \in \mathbb{R}$$

$$r \in \mathbb{Z}^+$$

Si tenemos la media

$$am_r = \frac{1}{n} \sum_{i=1}^n n_i (x_i - a)^r$$

$$\text{no centrados } m_r = \frac{1}{n} \sum_{i=1}^n n_i x_i^r$$

$$a=0$$

$$m_0 = 1$$

$$m_1 = \bar{x}$$

Centrados ~~áreas de los~~ ~~cuadrados~~ cuadros ~~que~~ que ~~la~~ es
~~la dist a x~~.

$$\mu_r = \frac{1}{n} \sum_{i=1}^n n_i (x_i - \bar{x})^r$$

$$\mu_0 = 1$$

$\mu_1 = 0 \rightsquigarrow$ La media es la dif a la media
es nula

$$\mu_2 = \sigma_x^2 = m_2 - m_1^2 \quad ***$$

Para qué? Tendremos nuestras definiciones los momentos
son fundamentales para ver como se ajustan con el problema.

Características unidimensionales

Momentos

Relación entre momentos:

- ▶ Momentos centrales en función de los no centrales:

$$\begin{aligned}\mu_r &= \sum_{i=1}^k f_i (x_i - m_1)^r = \sum_{i=1}^k f_i \sum_{t=0}^r (-1)^t \binom{r}{t} m_1^t x_i^{r-t} = \\ &\quad \sum_{t=0}^r (-1)^t \binom{r}{t} m_1^t \sum_{i=1}^k f_i x_i^{r-t} = \sum_{t=0}^r (-1)^t \binom{r}{t} m_1^t m_{r-t}\end{aligned}$$

- ▶ Momentos no centrales en función de los centrales y de m_1 :

$$\begin{aligned}m_r &= \sum_{i=1}^k f_i x_i^r = \sum_{i=1}^k f_i [(x_i - m_1) + m_1]^r = \sum_{i=1}^k f_i \sum_{t=0}^r \binom{r}{t} m_1^t (x_i - m_1)^{r-t} = \\ &= \sum_{t=0}^r \binom{r}{t} m_1^t \sum_{i=1}^k f_i (x_i - m_1)^{r-t} = \sum_{t=0}^r \binom{r}{t} m_1^t \mu_{r-t}\end{aligned}$$

$$\mu_2 = m_2 - m_1^2 \quad \text{---> demostrar}$$

$$\mu_3 = m_3 - 3 m_2 m_1 + 2 m_1^3$$

$$\mu_4 = m_4 - 4 m_3 m_1 + 6 m_2^2 m_2 - 3 m_1^4$$

$$m_2 = \mu_2 + m_1^2$$

$$m_3 = \mu_3 + 3 \mu_2 m_1 + m_1^3$$

$$m_4 = \mu_4 + 4 \mu_3 m_1 + 6 \mu_2 m_1^2 + m_1^4$$

04/03/20

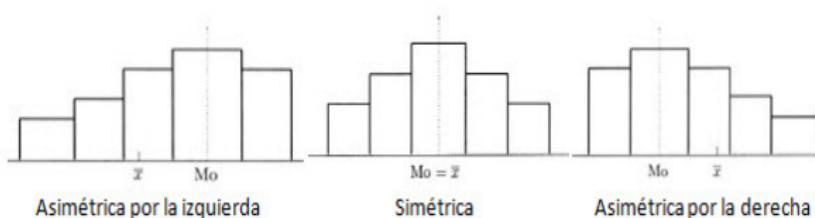
Características unidimensionales

Medidas de forma

→ Visualizar la distribución sin representarla

Medidas de asimetría:

Dada una variable estadística X , se entiende por asimetría de X a la falta de simetría respecto del eje vertical $x = \bar{x}$. Diremos, pues, que una distribución es simétrica si la perpendicular que pasa por la media aritmética divide al *diagrama diferencial* (histograma, en el caso continuo; o diagrama de barras, en el discreto) en dos partes iguales. Esto equivale a decir que a ambos lados de ese eje, y equidistantes de él, hay pares de valores/intervalos con la misma frecuencia. De lo contrario, diremos que es asimétrica.



Características unidimensionales

Medidas de forma

Medidas de asimetría:

► *Coeficiente de asimetría de Fisher:*

$$\gamma_1(X) = \frac{\mu_3}{\sigma_X^3} = \frac{1}{n} \sum_{i=1}^n n_i \left(\frac{x_i - \bar{x}}{\sigma_X} \right)^3$$

- Si $\gamma_1(X) > 0$ la distribución es asimétrica por la derecha o positiva.
- Si $\gamma_1(X) < 0$ la distribución es asimétrica por la izquierda o negativa.
- Si la distribución es simétrica, entonces $\gamma_1(X) = 0$.

► *Coeficientes de asimetría de Pearson:*

$$A_p = \frac{\bar{x} - Mo}{\sigma_X} \quad A_p^* = \frac{3(\bar{x} - Me)}{\sigma_X}$$

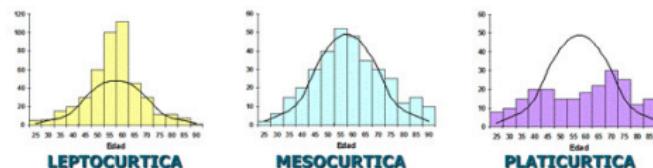
con la misma interpretación que $\gamma_1(X)$

Características unidimensionales

Medidas de forma

Medidas de apuntamiento o curtosis:

Miden la menor o mayor concentración central de frecuencias de una distribución respecto a la que presenta una distribución *Normal* de su misma media y su misma desviación típica. Se pueden dar tres casos, denominados como indica la figura:



Coefficiente de curtosis de Fisher:

$$\gamma_2(X) = \frac{\mu_4}{\sigma^4} - 3,$$

de tal forma que diremos que una distribución es: **platicúrtica** si $\gamma_2(X) < 0$; **mesocúrtica** si $\gamma_2(X) = 0$ y **leptocúrtica** si $\gamma_2(X) > 0$.

Coefficiente de curtosis de Kelley:

$$K = \frac{1}{2} \frac{Q_3 - Q_1}{D_9 - D_1} - 0,263$$

(0.263 es el valor para una distribución Normal); la interpretación es la misma que la del coeficiente de curtosis de Fisher.