Tema 1

Estadística descriptiva univariante

Carlos Montes – uc3m

¿Qué es la Estadística?

Es una herramienta de aprendizaje
a partir de la observación.

Nos ayuda a extraer conclusiones generalizables a partir de un conjunto de datos observados ⇨ *inducción* o *inferencia.*

DATOS (MUESTRA)
*realizaciones de una **variable***

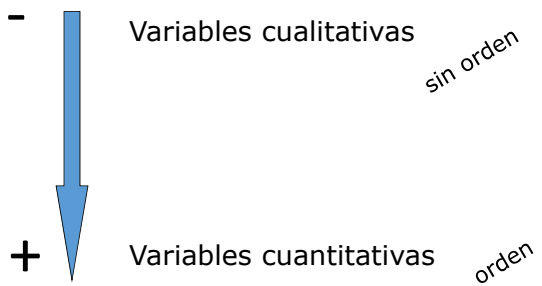
CONCLUSIONES
sobre el fenómeno que los ha originado

* Según su naturaleza, los datos pueden ser:

Datos cuantitativos.
Toman valores numéricos
❖Discretos: toman valores finitos.
❖Continuos: toman valores en un intervalo.

Datos cualitativos, categóricos o atributos.
No toman valores numéricos
Su realización concreta es una cualidad o modalidad.

La cantidad de información
aportada por ambos tipos de variables
es muy distinta:



OBJETIVO:

inferir cómo será la población
de la variable de interés
a partir
de la información limitada
que nos aporta la muestra.

A la hora de enfrentarse
a un conjunto de datos
hay que comenzar realizando
dos operaciones básicas.

ORDENAR

RESUMIR

2.1. Análisis básico. Generalidades

- *Frecuencia*
  - *absoluta* (*f*): el número de veces que aparece cada dato de la variable.
  - *total* (*n*): número total de datos de la variable (suma de frecuencias absolutas).
  - *relativa* (*fr*): cociente entre frecuencia absoluta y frecuencia total.

2.1. Análisis básico. Generalidades

- *acumulada*: supuesta la ordenación de los datos de menor a mayor, la frecuencia acumulada de x_i es la suma de frecuencias hasta el valor x_i.
  - Absoluta (*F*)
  - Relativa (*Fr*)

Tabla de distribución de frecuencias

2.1. Análisis básico. Generalidades

$$0{,}4842 + 0{,}3789 + 0{,}1263$$

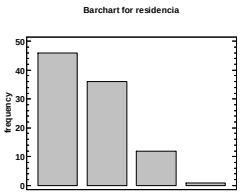
$$\frac{46}{95} \qquad \frac{94}{95}$$

$$46 + 36 + 12$$

		Relative	*Cumulative*	*Cum. Rel.*
Value	*Frequency*	*Frequency*	*Frequency*	*Frequency*
1	46	0,4842	46	0,4842
2	36	0,3789	82	0,8632
3	12	0,1263	94	0,9895
4	1	0,0105	95	1,0000

2.2. Gráficos para variables cualitativas

Diagrama de barras
Eje 1:valor o categoría de la variable.
Eje 2:altura proporcional a la frecuencia.



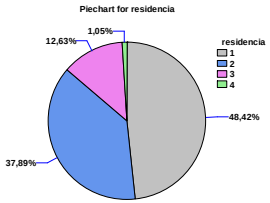
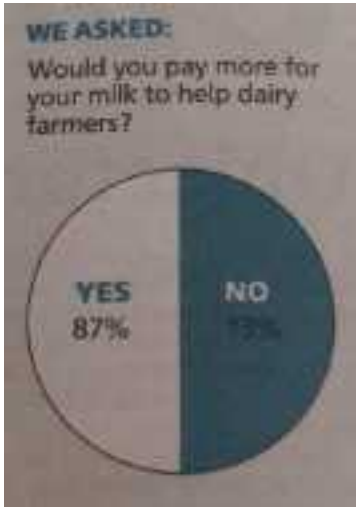


Diagrama de tarta
círculo dividido en sectores proporcionales a la frecuencia de cada valor.

2.2. Gráficos para variables cualitativas

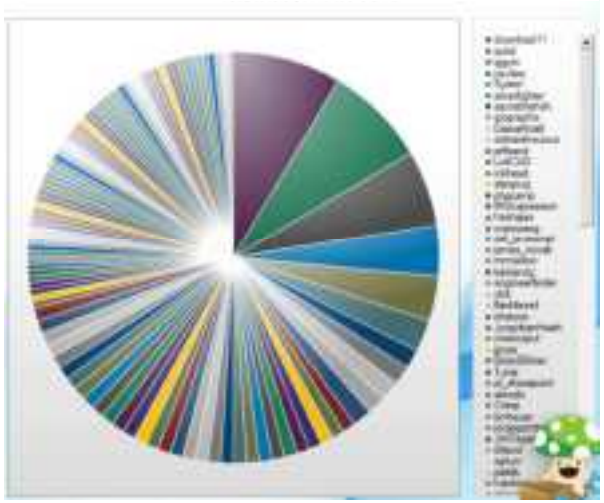
Encuesta en un periódico local



2.2. Gráficos para variables cualitativas

Los 100 usuarios de Twitter más activos

Más de 4 o 5 sectores dificultan la lectura del diagrama.



2.3. Variables cuantitativas

En variables cuantitativas el análisis de frecuencias se realiza de la misma manera que en variables cualitativas.

- ✓ Absolutas
- ✓ Relativas
- ✓ Absolutas acumuladas
- ✓ Relativas acumuladas

Muchos valores diferentes

↓

valores en clases o intervalos (generalmente de la misma longitud)

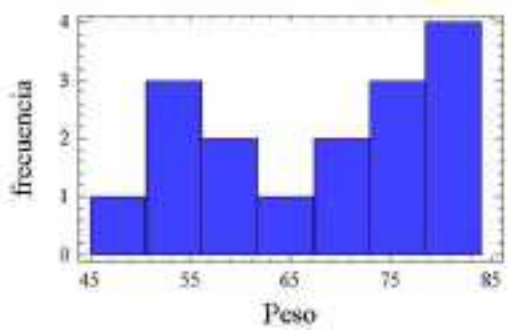
2.3. Variables cuantitativas

No confundir con el rango intercuartílico.

- *Rango o recorrido de una variable*: diferencia entre el mayor y el menor valor de ésta.
- *Amplitud de un intervalo*: diferencia entre el extremo superior e inferior del mismo.
- *Marca de clase (m_i):* punto medio de cada intervalo o clase, valor representativo de todos los datos del intervalo.

El número de clases *r* debe oscilar entre 5 y 20; a menudo se escoge el entero más próximo a $\sqrt{n}$

El ***histograma*** es una representación para variables agrupadas en intervalos.



- Abscisas: intervalo de valor de la variable.
- Ordenadas: altura proporcional a la frecuencia, de manera que las áreas de los rectángulos sean proporcionales a las frecuencias.

Muestra las tendencias generales de los datos:

• Concentraciones: más de una concentración ⇨ datos heterogéneos.

• Huecos: indicio de que los datos proceden de poblaciones diferentes.

• Valores atípicos: aquellos que se separan mucho del patrón general que siguen los datos.

• Asimetrías: tendencia de los datos cuando nos alejamos de las zonas de concentración.

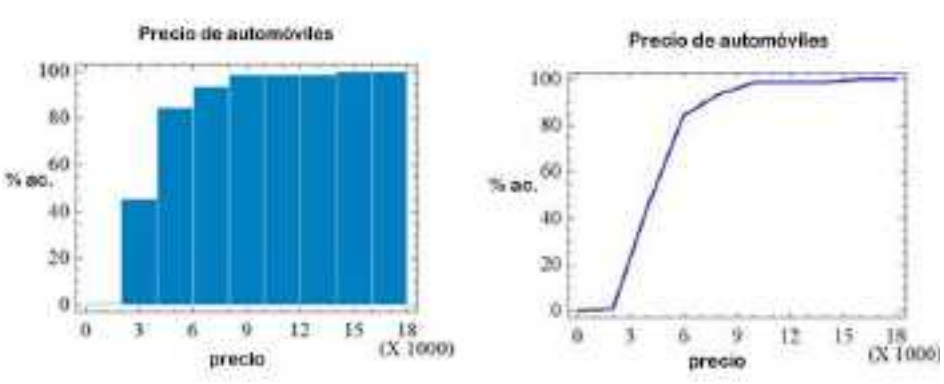
  • Cola de la distribución de los datos hacia $+\infty$, ⇨ asimetría positiva.

  • Cola de la distribución de los datos hacia $-\infty$ ⇨ asimetría negativa.

* El polígono de frecuencias es una línea poligonal que resulta al unir los puntos centrales de la parte superior del histograma.

2.4. Gráficos para variables cuantitativas

* Ambos pueden construirse a partir de las frecuencias acumuladas.



3.1. Medidas características. Generalidades

* Son aquellas que nos permiten resumir con un solo número los rasgos fundamentales de la distribución.

* Deben acompañarse de herramientas gráficas para evitar errores.

3.1. Medidas características. Generalidades

Podemos distinguir:

- Tendencia central o centralización: indican el valor medio de los datos.
- Dispersión: indican la variabilidad de los datos.
- Forma:
  - Simetría
  - Apuntamiento

3.2. Medidas de tendencia central

Media aritmética

$$\overline{x} = \frac{\sum_n x_j f(x_j)}{n}$$

$$\overline{x} = \frac{\sum_n m_j f(m_j)}{n}$$

⇨ Error de agrupamiento

Propiedades de la media aritmética $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

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

2) $y = x + k$

$$\bar{y} = \frac{\sum_{i=1}^{n}(x_i + k)}{n} = \frac{\sum x_i}{n} + \frac{\sum k}{n} = \frac{\sum x_i}{n} + \frac{nk}{n} = \bar{x} + k$$

3) $y = kx$ $\qquad \bar{y} = \frac{\sum_{i=1}^{n} kx_i}{n} = k\frac{\sum x_i}{n} = k\bar{x}$

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Stnd. skewness	-1.20518
Stnd. kurtosis	-1.70142

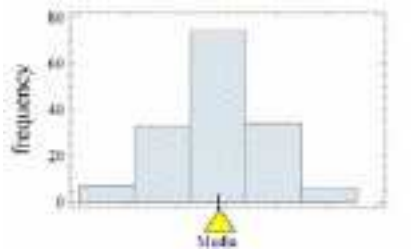
Es muy sensible a los datos atípicos.

1, 2, 4, 5, 7, 9, 11, 13 $\qquad \bar{x} = 6{,}5$

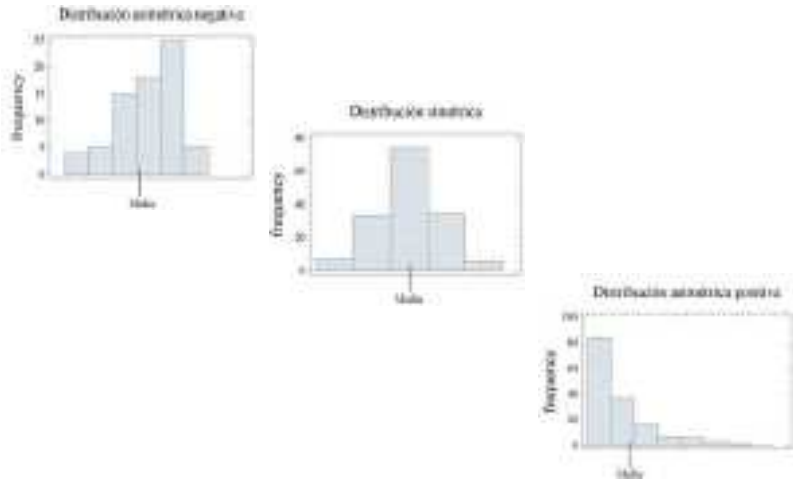
1, 2, 4, 5, 7, 9, 11, 130 $\qquad \bar{x} = 21{,}125$

Para muestras muy asimétricas o con muchos datos atípicos, la mediana es mejor medida de tendencia central.

Es el centro de gravedad de los datos.



Si la distribución es asimétrica, se desplaza respecto a la clase más frecuente, y deja de ser una buena medida de centralización.



Mediana

Valor de la muestra
que la divide en dos partes iguales.

* Para calcular la mediana se ordenan los datos de menor a mayor:

- nº impar de datos: valor central.

2, 3, 4, 5, 7, 7, 9

- nº par de datos: media aritmética de los valores centrales.

2, 3, 4, 5, 7, 7, 9, 11

$$\frac{5+7}{2} = 6$$

- Si tenemos los datos organizados en forma de tabla.

Accidentes mortales	n	f	F	N
0	7	0,039	0,039	7
1	26	0,144	0,183	33
2	33	0,182	0,365	66
3	38	0,210	0,575	104
4	29	0,160	0,735	133
5	20	0,110	0,846	153
6	15	0,083	0,929	168
7	9	0,050	0,978	177
8	2	0,011	0,989	179
9	2	0,011	1,000	181
10	0	0,000	1,000	181
>10	0	0,000	1,000	181
Total	181			

La mediana
es el primer valor
donde se alcanza la
frecuencia relativa
acumulada 0,5.

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Stnd. skewness	-1.20518
Stnd. kurtosis	-1.70142

La mediana NO es sensible a datos atípicos.

Robustez

2, 3, 4, 5, 7, 7, 9

2, 3, 4, 5, 7, 7, 87

Moda

Es el valor más frecuente de la distribución.

- Es apropiada para datos cualitativos o cuantitativos discretos.
- Pueden existir una o varias modas.
- En una muestra continua solo podemos hablar de un intervalo modal (el de mayor densidad de frecuencia)

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Stnd. skewness	-1.20518
Stnd. kurtosis	-1.70142

En variables continuas puede que no se repita ningún valor.

Pueden existir distribuciones con más de una moda.

Medidas de la separación de los datos (generalmente, respecto a la media).

medida
+ representativa

- dispersión

Varianza

$$s_x{}^2 = \frac{\sum_n (x_j - \bar{x})^2 f(x_j)}{n}$$

Propiedades de la varianza

1) Es una cantidad acotada y positiva

2) La varianza NO se ve afectada por los cambios de origen (transformaciones aditivas)

$$s_x^2 = \frac{\sum_n (x_i - \bar{x})^2}{n} \qquad y = x + k$$

$$s_y^2 = \frac{\sum_n (y_i - \bar{y})^2}{n} = \frac{\sum_n (x_i + k - \bar{x} - k)^2}{n} = \frac{\sum_n (x_i - \bar{x})^2}{n} = s_x^2$$

3) La varianza SÍ se ve afectada por los cambios de escala (transformaciones multiplicativas)

$$s_x^2 = \frac{\sum_n (x_i - \bar{x})^2}{n} \qquad y = kx$$

$$s_y^2 = \frac{\sum_n (y_i - \bar{y})^2}{n} = \frac{\sum_n (kx_i - k\bar{x})^2}{n} = \frac{k^2 \sum_n (x_i - \bar{x})^2}{n} = k^2 s_x^2$$

$$\boldsymbol{S_y^2 = k^2 \cdot S_x^2}$$

Fórmula de cálculo

$$\boldsymbol{s_x^2 = \frac{\sum x_j^2 f(x_j)}{n} - \bar{x}^2}$$

Una medida alternativa es la ***cuasivarianza***

$$\hat{s}_x^2 = \frac{\sum_n (x_j - \bar{x})^2 f(x_j)}{n - 1}$$

La mayoría de los programas estadísticos calculan la cuasivarianza en lugar de la varianza, y la llaman varianza.

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Stnd. skewness	-1,20548
Stnd. kurtosis	-1.70142

No aparece por defecto en el programa.

- La varianza mide el promedio de las desviaciones (al cuadrado) de las observaciones respecto a la media.
- Al ser un cuadrado, siempre es positiva.
- Es muy sensible a datos atípicos.
- Problema: unidades *67,68 cm²*

desviación típica

Desviación típica

Es la raíz cuadrada positiva de la varianza.

$$s_x = \sqrt{\frac{\sum_n (x_j - \bar{x})^2 \cdot f(x_j)}{n}}$$

Summary Statistics for altura	
Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Stnd. skewness	-1,20548
Stnd. kurtosis	-1.70142

Desviación típica

- Toma siempre valores no negativos.
- Ventaja: tiene las mismas unidades que la variable.

8,22 cm

- Inconveniente: raíz cuadrada. La varianza es más fácil de usar en operaciones matemáticas al evitar la raíz.

Cuasidesviación típica

$$\hat{s}_x = \sqrt{\frac{\sum_n (x_j - \bar{x})^2 f(x_j)}{n-1}}$$

- Para tamaños de muestra grande, casi no hay diferencia.

Coeficiente de variación

Es una medida de dispersión relativa.

$$CV = \frac{s}{/\bar{x}/} \bullet 100 \quad \bar{x} \neq 0$$

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

Nos permite:
1) Comparar la dispersión entre distribuciones.
2) Evaluar la representatividad de la media.

Cuantiles

Son los valores de la variable
que dividen la distribución en *c* partes iguales.

- **Cuartiles** (Q) c=4
- **Quintiles** (K) c=5
- **Percentiles** (p) c=100

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

Rango intercuartílico (RI)

Es la diferencia
entre los percentiles 75 y 25
(o entre los cuartiles 3 y 1)

3.3. Medidas de dispersión

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
Stnd. skewness	-1,20518
Stnd. kurtosis	-1,70142

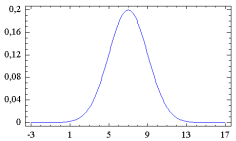
Coeficiente de asimetría de Fisher

Ronald Aylmer Fisher (1890-1962)

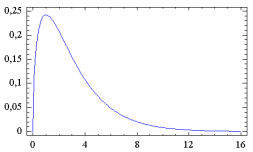
$$CA = \gamma_1 = \frac{\sum (x_i - \bar{x})^3}{ns^3}$$

3.4. Medidas de forma

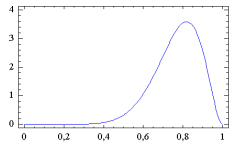
$\gamma_1 = 0$ ⇨ Distribución *simétrica*



$\gamma_1 > 0$ ⇨ Distribución *asimétrica positiva*
o asimétrica a derechas



$\gamma_1 < 0$ ⇨ Distribución *asimétrica negativa*
o asimétrica a izquierdas



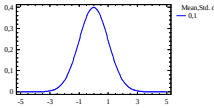
Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
Skewness	-0,302876
Stnd. skewness	-1,20518
Kurtosis	-0,855173
Stnd. kurtosis	-1,70142

Coeficiente de apuntamiento o curtosis

Indica el mayor o menor agrupamiento de los datos en torno a la media.

Como referencia se toma el apuntamiento de la distribución normal, que cumple:

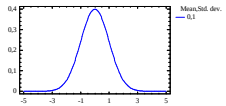


$$CA_p = \frac{\sum (x_i - \bar{x})^4}{ns^4} = 3$$

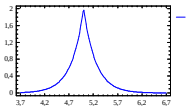
$$CA_p = \frac{\sum (x_i - \bar{x})^4}{ns^4} - 3$$

(Exceso de curtosis)

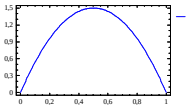
CAp=0: *mesocúrtica*



CAp>0: *leptocúrtica*



CAp<0: *platicúrtica*



Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
Skewness	-0,302876
Stnd. skewness	-1,20518
Kurtosis	-0,855173
Stnd. kurtosis	-1,70142

4. Diagrama de caja

Representación gráfica de una distribución,
construida para mostrar
sus características principales
y señalar los posibles datos atípicos.

Mínimo **Máximo** **Cuartiles**

LI= Q_1-1,5(Q_3-Q_1) **LS= Q_3+1,5(Q_3-Q_1)**

LIE= Q_1-3(Q_3-Q_1) LSE= Q_3+3(Q_3-Q_1)

4. Diagrama de caja