ESTADÍSTICA DESCRIPTIVA

TRATA DE DESCRIBIR CONJUNTOS DE DATOS RESUMIENDO LA INFORMACIÓN QUE ESTOS PROPORCIONAN, UTILIZANDO:

- TABLAS DE FRECUENCIAS
- GRÁFICAS
- **MEDIDAS NUMÉRICAS REPRESENTATIVAS (POSICIÓN, DISPERSIÓN Y FORMA)**

1

CONCEPTOS FUNDAMENTALES

- POBLACIÓN: conjunto de elementos o individuos de los que interesa estudiar alguna característica.
- MUESTRA: subconjunto de elementos de una población.

RAZONES PARA ESTUDIAR UNA MUESTRA

- Coste
- Tiempo
- Personal cualificado
- Procesos destructivos

2

•Llamamos **CARÁCTER** a la cualidad objeto de nuestro estudio.

Los caracteres pueden ser:

- **Cuantitativos:** la característica toma valores numéricos (número de peticiones a un servidor, tiempo entre peticiones consecutivas,etc)
- **Cualitativos**:la característica no toma valores numéricos (sexo, color de pelo, etc)

Los caracteres cuantitativos se llaman **VARIABLES ESTADÍSTICAS.**

Los caracteres cualitativos se llaman **VARIABLES CUALITATIVAS.**

SEGÚN EL TIPO DE VALORES QUE PUEDEN TOMAR las variables estadísticas pueden ser de dos tipos **:**

- Discretas: **si SÓLO PUEDE TOMAR UN NÚMERO finitos o infinito numerable DE VALORES DISTINTOS.**
- Continuas: si **PUEDE TOMAR** cualquier valor de uno o varios **intervalos**.

3

EJEMPLOS

• POBLACIÓN:	**ESTUDIANTES DE LA EUI**
•MUESTRA:	**ALUMNOS DE ESTADÍSTICA DEL SM22**
•VARIABLE ESTADÍSTICA:	**EDAD, PESO, NÚMERO DE…**
•VARIABLE CUALITATIVA:	**SEXO, CARA/CRUZ, SENSACIÓN**
•VARIABLE DISCRETA:	**EDAD, NÚMERO DE…**
•VARIABLE CONTINUA:	**PESO, ESTATURA**

4

1.2 DISTRIBUCIÓN DE FRECUENCIAS

Sea una muestra de tamaño *n*; supongamos que X toma como valores distintos $x_1, x_2, ..., x_k$.

•FRECUENCIA ABSOLUTA DE x_i: Es el número, n_i , de veces que se repite x_i.

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

•FRECUENCIA RELATIVA DE x_i: es el cociente entre la frecuencia absoluta y *n*.

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

•FRECUENCIA ABSOLUTA(RELATIVA) ACUMULADA DE x_i. Si llamamos $x^*_1, x^*_2, ..., x^*_k$ a los valores ordenados de menor a mayor:

$N_i = \sum_{j=1}^{i} n_j$ Frecuencia absoluta acumulada de x^*_i

$F_i = \frac{N_i}{n}$ Frecuencia relativa acumulada de x^*_i

5

•Si el número de valores distintos que toma la variable es grande (mayor que 20), se **agrupan los datos en intervalos** para construir la tabla de frecuencias.

VARIABLES NO AGRUPADAS: SI TOMA **MENOS DE 20** VALORES **DISTINTOS**.

VARIABLES AGRUPADAS: SI TOMA **MÁS DE 20** VALORES **DISTINTOS**.

Ejemplos:

VARIABLES NO AGRUPADAS: **EDAD, Nº ASIGANTURAS,…**

VARIABLES AGRUPADAS: **PESO, ESTATURA, SENSACIÓN,…**

6

VARIABLES AGRUPADAS EN INTERVALOS

A estos intervalos se les llama intervalos de clase.

Al punto medio de cada clase se le denomina marca de clase.

El número de intervalos de clase lo determina la persona que está realizando el estudio, aunque una posibilidad razonable es tomar el entero más próximo a $1+3.3\log_{10}(n)$.

CRITERIO ESENCIAL: SENTIDO COMÚN Y QUE LA LECTURA SEA FÁCIL Y SIGNIFICATIVA

EJEMPLO: PESO (Statg)

7

VARIABLES AGRUPADAS EN INTERVALOS

EJEMPLO: PESO . Por defecto, Statgraphics ofrece esta tabla:

Frequency Tabulation for Peso

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		49,0		0	0,0000	0	0,0000
1	49,0	57,5714	53,2857	6	0,1538	6	0,1538
2	57,5714	66,1429	61,8571	12	0,3077	18	0,4615
3	66,1429	74,7143	70,4286	11	0,2821	29	0,7436
4	74,7143	83,2857	79,0	9	0,2308	38	0,9744
5	83,2857	91,8571	87,5714	0	0,0000	38	0,9744
6	91,8571	100,429	96,1429	1	0,0256	39	1,0000
7	100,429	109,0	104,714	0	0,0000	39	1,0000
above	109,0			0	0,0000	39	1,0000

Cambiamos el **número de clases** y los **valores extremos**, intentando que los intervalos y las marcas de clase sean **fáciles de identificar**. Esta nueva tabla nos da un información más "digerible":

Frequency Tabulation for Peso

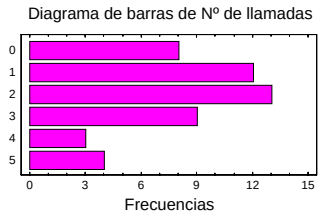
Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		45,0		0	0,0000	0	0,0000
1	45,0	55,0	50,0	3	0,0769	3	0,0769
2	55,0	65,0	60,0	15	0,3846	18	0,4615
3	65,0	75,0	70,0	13	0,3333	31	0,7949
4	75,0	85,0	80,0	7	0,1795	38	0,9744
5	85,0	95,0	90,0	0	0,0000	38	0,9744
6	95,0	105,0	100,0	1	0,0256	39	1,0000
above	105,0			0	0,0000	39	1,0000

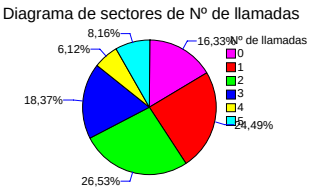
8

1.3 MÉTODOS GRÁFICOS

•VARIABLES SIN AGRUPAR:

- ♦Diagrama de barras
- ♦Diagrama de sectores





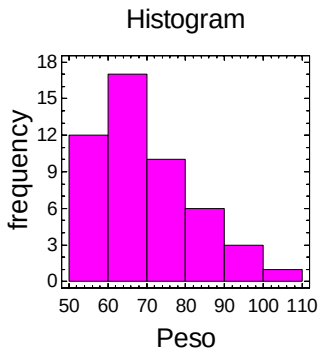
9

1.3 MÉTODOS GRÁFICOS

•VARIABLES AGRUPADAS:

- ♦Histograma

Cada clase se representa mediante un rectángulo cuyo **área es proporcional a su frecuencia** (absoluta o relativa)



10

1.3 MÉTODOS GRÁFICOS

•OTRAS REPRESENTACIONES

Diagrama de Tallo y Hojas (sintetiza información sobre los datos y sus frecuencias y da una buena imagen gráfica). Ejemplo: PESO.

```
Stem-and-Leaf Display for Peso: unit = 1,0   1|2 represents 12,0

      3      5|234
      7      5|6778
     16      6|000022233
     19      6|557
   (10)      7|0001122344
     10      7|556788
      4      8|002

            HI|100,0
```

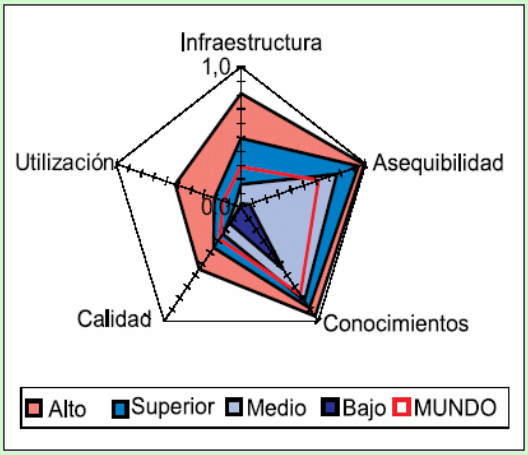
Representa que hay 4 personas con 60kg, 3 personas con 62kg y 2 personas con 63kg

11

1.3 MÉTODOS GRÁFICOS

•OTRAS REPRESENTACIONES

Indicadores del Índice de Acceso Digital según el nivel de ingresos:



12

MEDIDAS NUMÉRICAS REPRESENTATIVAS

MEDIDAS DE TENDENCIA CENTRAL

- Moda
- Media
- Mediana
- Cuantiles: cuartiles, deciles y percentiles

MEDIDAS DE DISPERSIÓN

- Rango o recorrido
- Recorrido intercuartílico
- Varianza y desviación típica
- Desviación media
- Coeficientes de variación

MEDIDAS DE FORMA

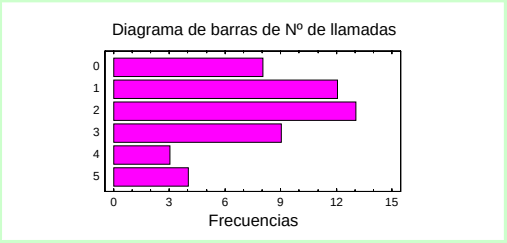
- Coeficientes de asimetría

13

1.4 MEDIDAS DE TENDENCIA CENTRAL

MODA, Mo: Es el dato que más se repite. Puede haber más de una moda.

Por ejemplo, con los datos muestrales: 2,2,3,3,4,4,4,5,6,6,6,7,7,8,8 se tienen dos modas: 4 y 6.



Hacer 2 llamadas al día es lo más frecuente. Moda=2

14

1.4 MEDIDAS DE TENDENCIA CENTRAL

MEDIA ARITMÉTICA: si todo se repartiera de forma homogénea, ¿cuánto tendrá cada uno?

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

Con los datos 2,2,3,3,4,4,4,5,6,6,6,7,7,8,8 , se tiene:

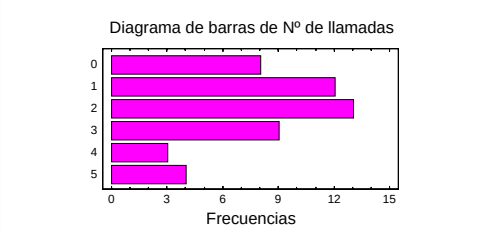
$$\bar{x} = \frac{2+2+3+3+4+4+4+5+6+6+6+7+7+8+8}{15} = 5$$

$$\bar{x} = \frac{2\cdot 2+3\cdot 2+4\cdot 3+5+6\cdot 3+7\cdot 2+8\cdot 2}{15} = 5$$

15

1.4 MEDIDAS DE TENDENCIA CENTRAL

MEDIA ARITMÉTICA:



$$\bar{x} = \frac{0\cdot 8+1\cdot 12+2\cdot 13+3\cdot 9+4\cdot 3+5\cdot 4}{49} = 1.98$$

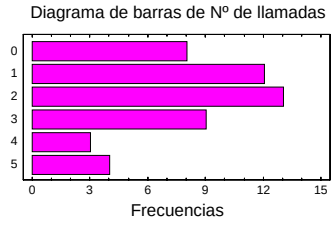
16

MEDIANA Me: Es el valor que **están en el medio** si ordenamos en magnitud los datos (el 50% es menor que él y el 50% mayor).

2 2 3 3 4 4 4 **5** 6 6 6 7 7 8 8

¿Mediana del nº de llamadas?

Observando las **frecuencias relativas acumuladas**, se ve que el 1 llega hasta el 40%, y el 2 va del 40% al 67% (incluye al 50%). Por tanto, **Mediana=2**



Frequency Table for Nº de llamadas

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0	8	0,1633	8	0,1633
2	1	12	0,2449	20	0,4082
3	2	13	0,2653	33	0,6735
4	3	9	0,1837	42	0,8571
5	4	3	0,0612	45	0,9184
6	5	4	0,0816	49	1,0000

17

CUANTIL DE ORDEN α, Cα: Es un valor tal que, ordenados en magnitud los datos, el 100 α % es menor que él y el resto mayor.

Los más utilizados son:

- los **cuartiles** Q_1 (α=0.25), Q_3 (α=0.75), Q_2=Mediana (α=0.5)
- los **deciles** D_1,...,D_9 (α= 0.1,... , 0.9)
- los **percentiles** P_1,...,P_{99} (α= 0.01,...,0.99)

Cálculo de cuantiles: (mediana, cuartiles y percentiles)

- A partir de las **frecuencias relativas acumuladas**:

Frequency Table for Nº de llamadas

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0	8	0,1633	8	0,1633
2	1	12	0,2449	20	0,4082
3	2	13	0,2653	33	0,6735
4	3	9	0,1837	42	0,8571
5	4	3	0,0612	45	0,9184
6	5	4	0,0816	49	1,0000

Hallar el tercer cuartil y el percentil 40 de la variable nº de llamadas.

18

Cα: Es un valor tal que, ordenados en magnitud los datos, el 100 α % es menor que él y el resto mayor.

Frequency Table for Nº de llamadas

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0	8	0,1633	8	0,1633
2	1	12	0,2449	20	0,4082
3	2	13	0,2653	33	0,6735
4	3	9	0,1837	42	0,8571
5	4	3	0,0612	45	0,9184
6	5	4	0,0816	49	1,0000

Tercer cuartil (α=0.75): buscamos en la última columna el valor 0.75, que se encuentra justo en el 3. Por tanto Q_3=3 y se puede **interpretar**:

- *el 75% de los estudiantes hace 3 llamadas o menos*
- *el 25% de los estudiantes hace 3 llamadas o más*
- *el porcentaje de estudiantes que hace más de 3 llamadas no llega al 25%*

3

19

Cα: Es un valor tal que, ordenados en magnitud los datos, el 100 α % es menor que él y el resto mayor.

Frequency Table for Nº de llamadas

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0	8	0,1633	8	0,1633
2	1	12	0,2449	20	0,4082
3	2	13	0,2653	33	0,6735
4	3	9	0,1837	42	0,8571
5	4	3	0,0612	45	0,9184
6	5	4	0,0816	49	1,0000

Percentil cuarenta (α=0.40): buscamos en la última columna el valor 0.40, que se encuentra justo en el 1. Por tanto P_{40}=1 y se puede **interpretar**:

- *el 40% de los estudiantes hace 1 llamada o menos*
- *el 60% de los estudiantes hace 1 llamada o más*

1

20

ROBUSTEZ DE LA MEDIANA

Consideremos los datos del ejemplo anterior:

2 2 3 3 4 4 4 5 6 6 6 7 7 8 8

Si añadimos un nuevo dato $x_{16} = 34$ y calculamos de nuevo la media y la mediana, obtenemos:

- Nueva media: $\overline{x} = 6.8$
- Nueva mediana: Me = 5.5

La media cambia más que la mediana.

¿Qué ocurriría si en el ejemplo del nº de llamadas introducimos los datos de dos estudiantes más que han hecho 0 llamadas?

21

COMPARACIÓN MEDIA-MEDIANA

- La media contiene más información porque usa los valores de todos los datos.
- La mediana es más robusta frente a los cambios en los datos.
- La media es más sencilla de calcular y se presta mejor a los cálculos algebraicos.
- Deben calcularse ambas pues proporcionan información complementaria.

22

1.5 MEDIDAS DE DISPERSIÓN

Las medidas de centralización proporcionan una información incompleta del conjunto de datos.

Ejemplo: sean X e Y las notas de dos grupos de cuarenta alumnos, con distribuciones de frecuencias:

x_i	n_i
0	20
10	20

y_i	n_i
4.5	3
5	34
5.5	3

Para ambas variables la media es 5, pero en el segundo caso 5 es un valor más representativo de los datos que en el primero.

Las medidas de dispersión nos permiten valorar si el valor de la medida de tendencia central es , o no es , representativo.

23

MEDIDAS DE DISPERSIÓN

MEDIDAS DE **AMPLITUD**:

- RANGO O RECORRIDO: R = Max - Min
- RECORRIDO INTERCUARTÍLICO: RQ = Q3 - Q1

Ejemplo: sean X e Y las notas de dos grupos de cuarenta alumnos, con distribuciones de frecuencias:

x_i	n_i
0	20
10	20

y_i	n_i
4.5	3
5	34
5.5	3

$R_X = 10-0=10$; $R_Y = 5.5 - 4.5=1$

$RQ_X = 10-0=10$; $RQ_Y = 5 - 5=0$

Con estas medidas sí se detectan las diferencias.

24

MEDIDAS DE DISPERSIÓN

MEDIDAS DE **DISTANCIA A LOS VALORES CENTRALES**:

Distancia a la media:

• VARIANZA: $V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ó $V = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 \cdot n_j$

• DESVIACIÓN TÍPICA: $Dt = \sqrt{V}$

La desviación típica es la **distancia media al valor medio de la variable**. Está en las mismas unidades que la propia variable (no así a varianza).

En el **ejemplo** anterior:

$$V_X = \frac{1}{40}\left((0-5)^2 \cdot 20 + (10-5)^2 \cdot 20\right) = 25 \Rightarrow Dt_X = 5$$

$$V_Y = \frac{1}{40}\left((4.5-5)^2 \cdot 3 + (5-5)^2 \cdot 34 + (5.5-5)^2 \cdot 3\right) = 0{,}0375 \Rightarrow Dt_Y = 0.19$$

25

MEDIDAS DE DISPERSIÓN

MEDIDAS DE **DISTANCIA A LOS VALORES CENTRALES**:

Estos resultados se interpretan considerando que la **distancia media** de la variable *X* **a su valor medio (5)** es de 5. La **distancia media** de la variable *Y* **a su valor medio (5)** es de 0.19.

(Se interpreta el valor de la desviación típica, pues está en las mismas unidades que la propia variable, no así a varianza.)

26

MEDIDAS DE DISPERSIÓN

MEDIDAS DE **DISTANCIA A LOS VALORES CENTRALES**:

Distancia a la media:

Por razones que se verán más adelante, los programas estadísticos calculan los valores de:

- CUASIVARIANZA: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- CUASIDESVIACIÓN TÍPICA: $s = \sqrt{s^2}$

27

MEDIDAS DE DISPERSIÓN

MEDIDAS DE **DISTANCIA A LOS VALORES CENTRALES**:

Distancia a la mediana:

- DESVIACIÓN MEDIA: $Dm = \frac{1}{n}\sum_{i=1}^{n}|x_i - Me|$ *ó* $Dm = \frac{1}{n}\sum_{j=1}^{k}|x_j - Me|\cdot n_j$

En el ejemplo de las variables X e Y que representan unas notas:

$$DM_X = \frac{1}{40}\left(|0-5|\cdot 20 + |10-5|\cdot 20\right) = 5$$

$$DM_Y = \frac{1}{40}\left(|4.5-5|\cdot 3 + |5-5|\cdot 34 + |5.5-5|\cdot 3\right) = 0{,}075$$

28

MEDIDAS DE DISPERSIÓN

MEDIDAS DE **DISTANCIA RELATIVA A LOS VALORES CENTRALES**. Sirven para **comparar la dispersión entre dos variables** que toman valores de magnitudes diferentes (no indica la misma dispersión una Dt=1 si el peso está medido en gramos que si está medido en Kg)

• COEFICIENTE DE VARIACIÓN DE PEARSON (dispersión **respecto de la media**):

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

• COEFICIENTE DE VARIACIÓN MEDIA (dispersión **respecto de la mediana**):

$$CVm = \frac{Dm}{|Me|}$$

29

MEDIDAS DE DISPERSIÓN

Se quiere **comparar qué variable tiene mayor dispersión**, si la Edad o la Estatura:

Como datos, tenemos:

$$\bar{X}_{EDAD} = 21.8 \; ; \; Dt_{EDAD} = 2.35$$
$$\bar{X}_{ESTATURA} = 174 \; ; \; Dt_{ESTATURA} = 7.36$$

La desviación es mayor en ESTATURA, sin embargo en términos relativos se tiene que:

$$CV_{EDAD} = \frac{2.35}{21.8} = 0.1077; \; CV_{ESTATURA} = \frac{7.36}{174} = 0.0423$$

La desviación de EDAD representa el 10.77% del valor de su media y, en cambio, la desviación de ESTATURA representa sólo el 4.23% del valor de su media. Proporcionalmente, **la variable EDAD tiene una mayor dispersión que la variable ESTATURA.**

30

PROPIEDADES DE MEDIA Y VARIANZA

Sean $x_1,\ldots, x_k$ los valores diferentes de una variable estadística X y $n_1,\ldots, n_k$ sus frecuencias correspondientes, entonces,

$$V_X = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \bar{x}^2 \quad \text{siendo} \quad n = \sum_{i=1}^{k} n_i$$

Ejemplo: sean X e Y las notas de dos grupos de cuarenta alumnos, con distribuciones de frecuencias:

x_i	n_i
0	20
10	20

y_i	n_i
4.5	3
5	34
5.5	3

$$V_X = \frac{1}{40}\left(0^2\cdot 20 + 10^2\cdot 20\right) - 5^2 = 25$$

$$V_Y = \frac{1}{40}\left(4.5^2\cdot 3 + 5^2\cdot 34 + 5.5^2\cdot 3\right) - 5^2 = 0{,}0375$$

31

PROPIEDADES DE MEDIA Y VARIANZA

Si se define una nueva variable $Y = aX+b$, entonces

A) $\bar{y} = a\bar{x} + b$

B) $V_Y = a^2 V_X$

32

DESIGUALDAD DE CHEBYCHEV

En el intervalo $(\overline{X} - k \cdot Dt, \overline{X} + k \cdot Dt)$ se encuentran, como mínimo el $100\left(1 - \frac{1}{k^2}\right)\%$ de los datos.

Por ejemplo, con la variable ESTATURA, media de 174 y desviación típica de 7.36, consideramos ***k*=2**:

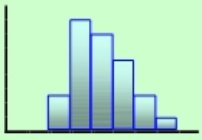
$$(\overline{X} - 2 \cdot Dt, \overline{X} + 2 \cdot Dt) = (159.28, 188.72) \; ; \; 100\left(1 - \frac{1}{4}\right)\% = 75\,\%$$

Se tiene que al menos el 75% de los estudiantes tienen una estatura entre 159.28 y 188.72 cm.

Para ***k*=3**, se tiene que al menos el 89% de los estudiantes tienen una estatura entre 152 y 196 cm.

33

1.6 MEDIDAS DE ASIMETRÍA

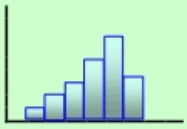


Simetría a la derecha

Skewness > **0**

Simétrica

Skewness ~ **0**



Simetría a la izquierda

Skewness < **0**

¿Cómo cuantificarla?

• Como los valores extremos influyen más en la media que en la mediana, una forma es observar la diferencia entre media y mediana, teniendo en cuenta su signo. **COEFICIENTE DE ASIMETRÍA DE PEARSON**:

$$CAP = \frac{3(\overline{x} - Me)}{Dt}$$

34

1.6 MEDIDAS DE ASIMETRÍA

Otra forma, es cuantificar las diferencias con la media (valor central), pero teniendo en cuenta el signo de esas diferencias. **COEFICIENTE DE ASIMETRÍA DE FISHER:**

$$CAF = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{(Dt)^3}$$

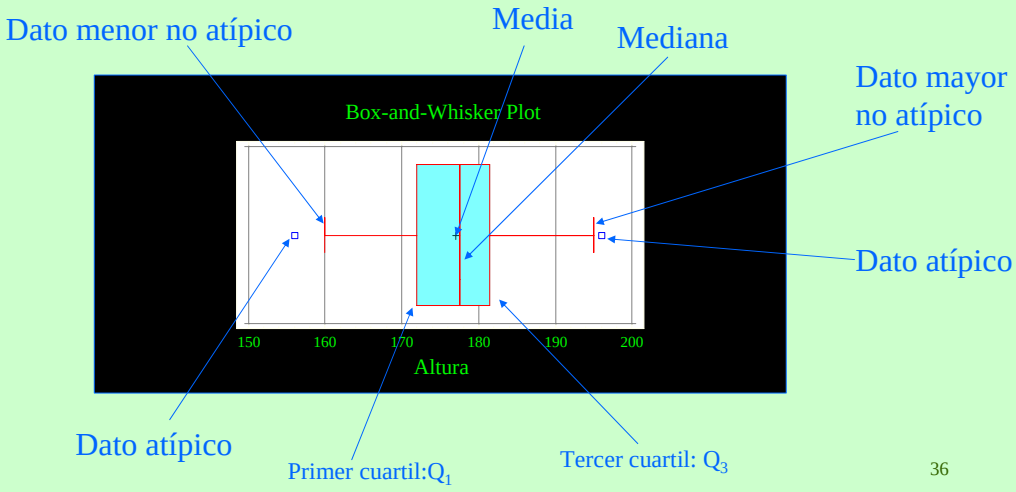
Para ambos coeficientes, si:

- CAF>0 o CAP>0, la distribución es asimétrica a la derecha.
- CAF=0 o CAP=0, la distribución es simétrica.
- CAF<0 o CAP<0, la distribución es asimétrica a la izquierda.

35

DIAGRAMA DE CAJA (BOX-PLOT)

Es una representación gráfica que sintetiza la información sobre los valores centrales, la dispersión y la simetría de una variable.



36

DIAGRAMA DE CAJA (BOX-PLOT)

Se construye del siguiente modo:

•Con los datos ordenados se obtienen los tres cuartiles

•Se dibuja un rectángulo cuyos extremos son Q1 y Q3 y se indica la posición de la mediana mediante una línea.

•Se calculan los límites de admisión (los valores que queden fuera se consideran **atípicos**)

$$LI = Q_1 - 1'5(Q_3 - Q_1)$$

$$LS = Q_3 + 1'5(Q_3 - Q_1)$$

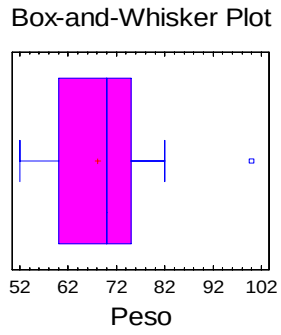
•Se dibuja una línea desde cada extremo del rectángulo hasta el valor más alejado no atípico.

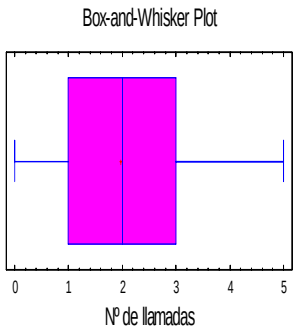
•Se marcan todos los datos considerados como atípicos.

37

DIAGRAMA DE CAJA (BOX-PLOT)

Ejemplos:





38

1.5. REGRESIÓN

Regresión (en la presentación de la asignatura)

OBJETIVO: Estudiar la posible **relación funcional** entre dos o más variables.

¿tiene **relación** el nº de asignaturas aprobadas con el de matriculadas?, ¿y con el tiempo que se tarda en llegar a la EUI?, ¿o con el nº de películas que se han visto en el curso?

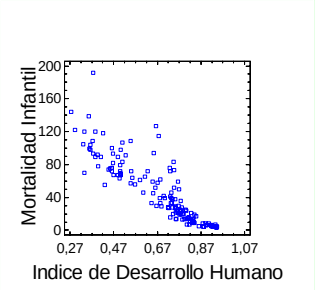
¿se puede conocer el peso de alguien a partir de su estatura?

¿tiene relación el nº de llamadas con la duración de las llamadas?

Fuente: Gonick,L. "La Estadística en cómic"

Distintos diagramas de dispersión

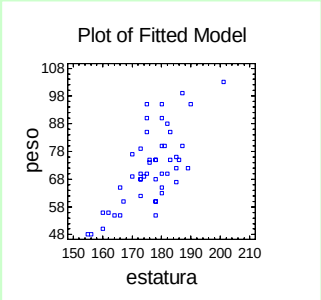
Para intuir qué tipo de relación puede haber entre dos variables, un buen instrumento es el **diagrama de dispersión**:



A **mayor** Indice de Desarrollo Humano, **menor** tasa de mortalidad infantil.

(Relación **inversa**)

Forma de recta con pendiente negativa (relación **lineal negativa**)



A **mayor** Estatura, **mayor** Peso.

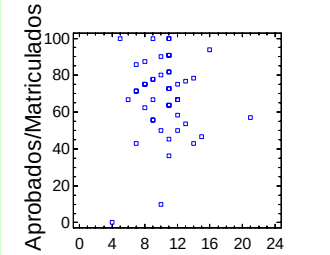
(Relación **directa**)

Forma de recta con pendiente positiva (relación **lineal positiva**)

40

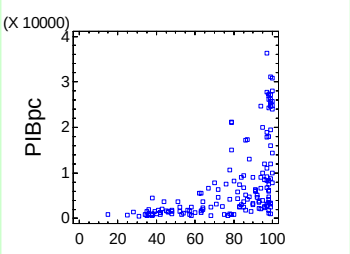
Distintos diagramas de dispersión

Para intuir qué tipo de relación puede haber entre dos variables, un buen instrumento es el **diagrama de dispersión**:



No parece que haya relación entre el número de asignaturas matriculadas con el porcentaje de asignaturas aprobadas.

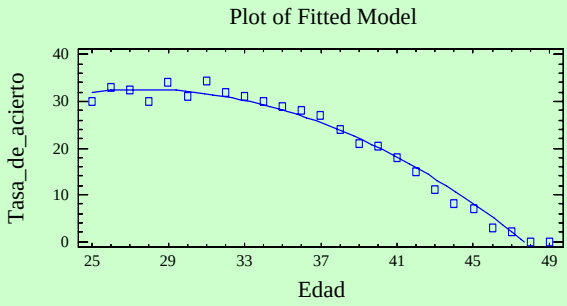
(Variables **independientes**)



A **mayor** nivel de alfabetización, **mayor** PIB (relación **directa**). Pero el modelo funcional no parece ser una recta (relación **no lineal**).

41

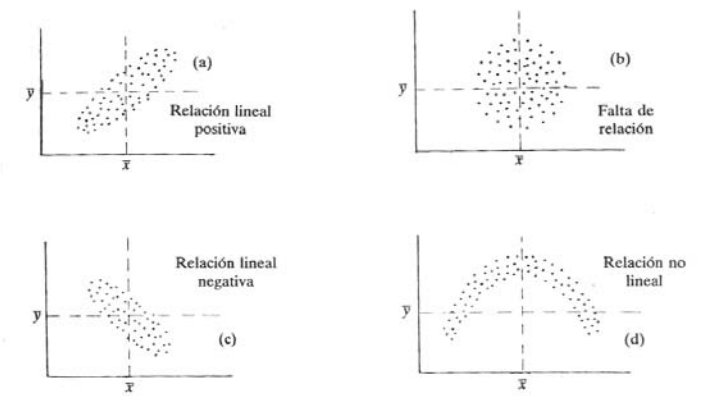
ESTUDIO DE REGRESIÓN: INFLUENCIA DE LA EDAD DE LA MUJER EN LA TASA DE ACIERTO EN LA REPRODUCCIÓN ASISTIDA



A **mayor** edad, **menor** tasa de acierto (relación **inversa**). Pero el modelo funcional que más se ajusta no es una recta (relación **no lineal**).

42

Distintos diagramas de dispersión



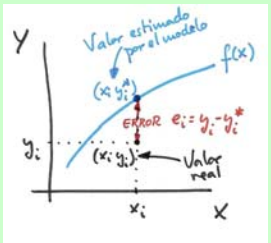
43

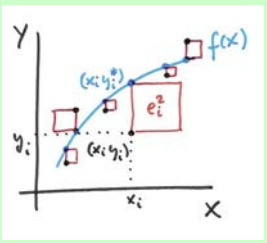
¿Cómo sabemos cuál es el modelo que mejor explica la relación entre dos variables?

CASO GENERAL: $Y = f(X)$

Cuando aproximamos el valor de y_i por el de $y_i^* = f(x_i)$, estamos cometiendo un **error:** $e_i = y_i - f(x_i) = y_i - y_i^*$ (se suele denominar **residuo**).

La idea es encontrar la función $f(x)$ que hace que la suma de los cuadrados de estos errores sea mínima (**mínimos cuadrados**).





La media de los cuadrados de dichos errores se denomina **VARIANZA RESIDUAL:**

$$V_R = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$$

44

¿Cómo se obtiene esa función *f(x)*?

- A la vista del diagrama de dispersión se propone un modelo para *f*:
  - Lineal: $f(X)= aX + b$
  - Polinómico: $f(X)= aX^2 + bX + c$
  - Exponencial: $f(X)= a\, e^{bX}$
  - …
- Se hallan los valores de *a* y *b* que hacen mínima la varianza residual.

Para el caso lineal, $f(X)= aX + b$, se buscan *a* y *b* que hagan mínimo:

$$V_R = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (ax_i + b)\right)^2$$

y se obtiene:

$$a = \frac{\sum_{i=1}^{n} y_i\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \;;\; b = \bar{y} - a\bar{x}$$

NOTA: En este curso utilizaremos Statgraphics para obtener las expresiones de los modelos de regresión.

¿Cómo sabemos si esa función *f(x)* es un *"buen modelo"*?

Cuanto menor sea su varianza residual, mejor será el modelo.

Pero para poder comparar, se ha de tener una medida *adimensional* (que no dependa de las unidades en las que estemos trabajando) y que tenga en cuenta la propia variabilidad de Y (V_Y).

Esta medida es el **COEFICIENTE DE DETERMINACIÓN**: $R^2 = 1 - \frac{V_R}{V_Y}$

Propiedades:

$$0 \le \boldsymbol{R}^2 \le 1$$

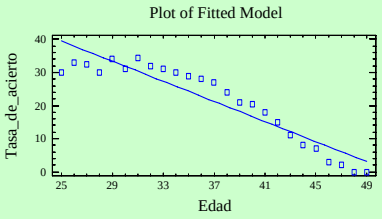
Se interpreta como el **porcentaje de la varianza de Y que es explicado por el modelo propuesto.**

El modelo será más adecuado cuanto más cercano esté de 1 (o del 100%) y será peor cuanto más cerca esté de 0.

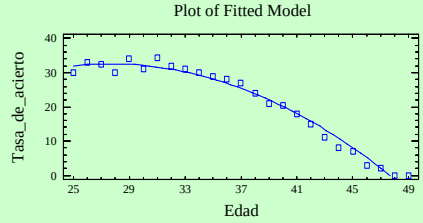
(Se considera un "buen modelo" a partir de 0.5 o 50%)

46

Ejemplo: INFLUENCIA DE LA EDAD DE LA MUJER EN LA TASA DE ACIERTO EN LA REPRODUCCIÓN ASISTIDA



Con el **modelo lineal**, se obtiene un valor de R^2= 0.8756. Es decir, el 87.56% de la variabilidad de la *tasa de acierto en la reproducción asistida* se puede explicar por la relación lineal con la variable *Edad*. (Es un buen modelo)



Con el **modelo parabólico**, se obtiene un valor de R^2= 0.9789. Es decir, el 97.89% de la variabilidad de la *tasa de acierto en la reproducción asistida* se puede explicar por la relación cuadrática con la variable *Edad*.

Es un buen modelo y, además, explica mejor el comportamiento de la tasa de acierto que el modelo lineal.

47

Regresión lineal simple

Es el modelo más sencillo, ***f(X)= aX + b***, y tiene algunas propiedades particulares que vamos a estudiar.

Para medir la relación lineal entre dos variables, se utiliza en estadística descriptiva la **COVARIANZA**:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

Cov(X,Y) > 0, indica una relación lineal **directa**,

Cov(X,Y) < 0, indica una relación lineal **inversa**.

Cov(X,Y) = 0, indica que **no hay relación lineal** entre las variables. (Puede haber otro tipo de relación)

48

Regresión lineal simple

A partir de la covarianza se tiene una expresión más simple de la recta de regresión:

$$Y - \bar{y} = \frac{Cov(X,Y)}{V(X)}(X - \bar{x})$$

Observar que la recta siempre pasa por el punto $(\bar{x}, \bar{y})$.

Se define el **coeficiente de correlación lineal**: $r = \frac{Cov(X,Y)}{dt(X)\,dt(Y)}$

Se verifica que: $-1 \le r \le 1$ y $|r| = \sqrt{R^2}\,; R^2 = r^2$

Interpretación de los parámetros del modelo

Nos interesa sobre todo:

- ✓ El tipo de relación lineal: directa o inversa
- ✓ La "bondad" del ajuste: ¿en qué grado el modelo explica el comportamiento de la variable Y?

49

Regresión lineal simple Interpretación

Tipo de relación lineal:

Observamos el coeficiente de correlación r, Cov(X,Y) o la pendiente de la recta de regresión

- r = 0 ó Cov(X,Y) = 0 ó pendiente = 0: no hay relación lineal.
- r > 0 ó Cov(X,Y) > 0 ó pendiente > 0: relación lineal directa.
- r < 0 ó Cov(X,Y) < 0 ó pendiente < 0: relación lineal inversa.

Bondad del ajuste:

Será mejor cuanto más cerca esté de 1 el valor de $|r|$ y R^2

Será peor cuanto más cerca esté de 0 el valor de $|r|$ y R^2

50

Regresión lineal simple

Observación: correlación lineal $\neq$ causalidad

-Si se aumenta de peso no se tiene porqué aumentar de estatura.

-Hay una correlación positiva entre esperanza de vida y porcentaje de población urbana, ¿es más sano vivir en la ciudad que en el campo?

-Hay una correlación negativa entre nivel de estudios universitarios de la mujer y porcentaje de mujeres casadas, ¿si estudias no te casas?

- Correlación negativa entre tasa de mortalidad y tasa de divorcios, ¿cuánto más nos divorciemos más vivimos?

51