

Obxectivos

Nesta quincena aprenderás a:

- Distinguir os conceptos de poboación e mostra.
- Diferenciar os tres tipos de variables estatísticas.
- Facer recontos e gráficos.
- Calcular e interpretar as medidas estatísticas de centralización máis importantes.
- Calcular as principais medidas de dispersión.
- Entender a importancia da elección da mostra para que sexa representativa.
- Utilizar e representar variables bidimensionais.
- Calcular o centro de gravidade, a covarianza, o coeficiente de correlación e a recta de regresión nunha distribución bidimensional.

1. Estadística descriptiva	páx. 4
Poboación e mostra	
Variables estatísticas	
Gráficos variables cualitativas	
Gráficos variables cuantitativas discretas	
Gráficos variables cuantitativas continuas	
2. Medidas de centralización	páx. 7
Media, moda e mediana	
Evolución da media	
Evolución da mediana	
Media e mediana comparadas	
Medidas de posición	
3. Medidas de dispersión	páx. 10
Desviación típica e percorrido	
Cálculo das medidas de dispersión	
A media e a desviación típica	
4. Representatividade das mostras ...	páx. 12
Mostraxe estratificada	
Mostraxe aleatoria. Nesgo	
5. Estadística bidimensional	pág. 14
Distribucións bidimensionais	
Correlación lineal	
Recta de regresión	

Exercicios para practicar

Para saber máis

Resumo

Auto-avaliación

Antes de empezar

Lembra

O curso pasado xa estudaches estatística, e en numerosas ocasións fixeches estatística aínda que non te deras conta. Vexamos algúns exemplos.

Nota media

Ao longo dun curso escolar terás moitas ocasións onde calcular este valor. Se unha nota depende de dous exames e nun tes un 4, intentarás sacar polo menos un 6 na outra.

Ao final do instituto, as medias do bacharelato e da proba selectividade. Comparacións coa media local ou nacional. As medias de corte para determinadas carreiras

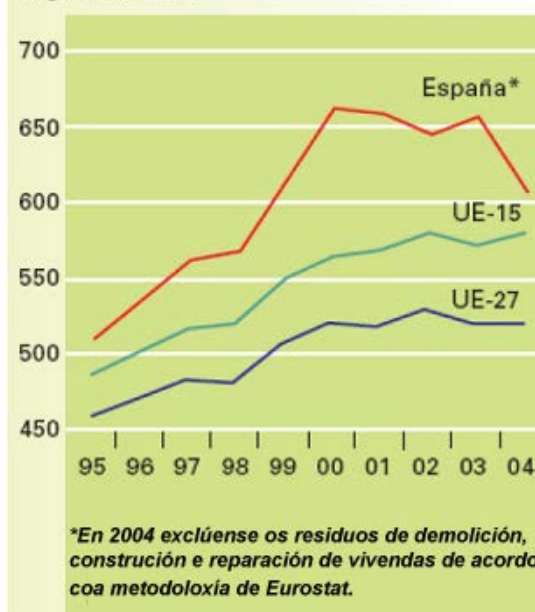
Fútbol

O xogador que máis goles marcou, o porteiro que menos encaixou. A clasificación da liga. A mellor metade de liga. Os postos de competicións europeas, os de descenso, nº de veces internacional, nº de fases finais, minutos xogados, tiros a porta, faltas.

Consumo medio de auga dos fogares. 2004 (litros/hab./día)



Residuos urbanos (kg/hab./ano)



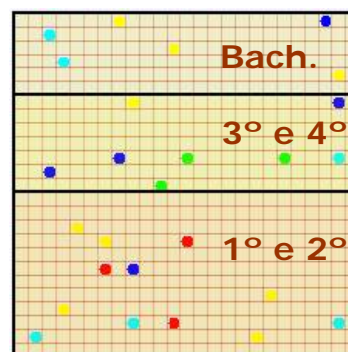
1. Estatística descritiva

Poboación e mostra.

Poboación é o conxunto de individuos, con algunha característica común, sobre o que se fai un estudo estatístico.

A **mostra** é un subconxunto da poboación, seleccionada de modo que poña de manifesto as características da mesma, de aí que a propiedade máis importante das mostras é a súa representatividade.

O proceso seguido na extracción da mostra chámase **mostraxe**



Se cada cadriño representa a cada un dos alumnos dun instituto ficticio e se lles pregunta sobre a súa cor favorita, o total dos cadros é a poboación, 625 alumnos, e os 26 enquisados constitúen a mostra.

Variables estatísticas

A característica a estudar nunha poboación é a **variable estatística**.

As variables estatísticas poden ser esencialmente de dous tipos **cuantitativas e cualitativas**.

As variables cualitativas son as que non aparecen en forma numérica senón como unha categoría ou atributo.

As variables cuantitativas son as que se poden expresar numericamente, e á súa vez poden ser:

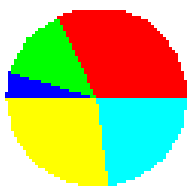
- ✓ Cuantitativas discretas, se só poden tomar un número finito de valores.
- ✓ Cuantitativas continuas cando poden tomar calquera valor dun intervalo.

- A cor dos ollos, o queixo preferido, o continente onde vives, son **variables estatísticas cualitativas**.
- O nº de ordenadores na casa, ou de televisores e o nº de habitantes por vivenda, por exemplo, son variables estatísticas **cuantitativas discretas**.
- O peso, a altura, a velocidade, a densidade, a presión, son **variables estatísticas cuantitativas continuas**.

Os datos:

- xi fi
- 7
- 3
- 1
- 6
- 5

Total 22



Teñen este diagrama de sectores

Gráficos en variables cualitativas.

O **diagrama de sectores** é o máis indicado para este tipo de información. A porcentaxe de datos de cada valor nunca mostra correspondécese coa mesma porcentaxe de sector dun círculo. Así por exemplo, se os datos son A, A, A, A, A, B, B, B, C e C. As frecuencias son (A,5), (B,3) e (C,2), as porcentaxes serán (A,50%), (B,30%) e (C,20%) os que corresponde a un gráfico de sectores con (A, 180°), (B,108°) e (C, 72°).

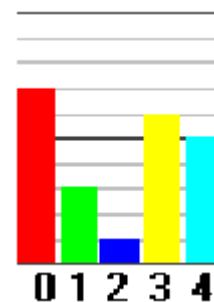
$$\frac{\text{frecuencia}}{\text{n}^\circ \text{ total de datos}} = \frac{\text{graos do sector}}{360}$$

Gráficos en variables discretas.

Diagrama de barras. Abondará con que observes un exemplo.

Aos datos,

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

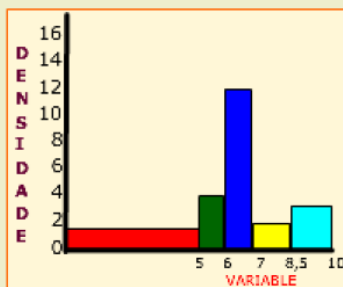
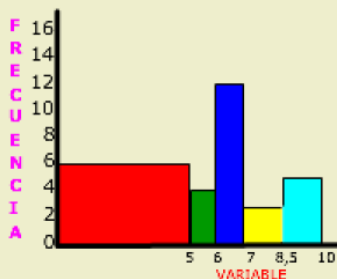


correspóndelles o gráfico da dereita.

Intervalos	Reconto	fr.	Dens.
[0 5)		6	1,2
[5 6)		4	4
[6 7)		12	12
[7 8,5)		3	2
[8,5 10)		5	3,3

RECONTO DAS NOTAS EN 30 EXÁMENES

No diagrama de frecuencias o área maior corresponde á columna vermella que non é a de máis frecuencia



$$\text{Densidade} = \frac{\text{Frecuencia}}{\text{Lonxitude do intervalo}}$$

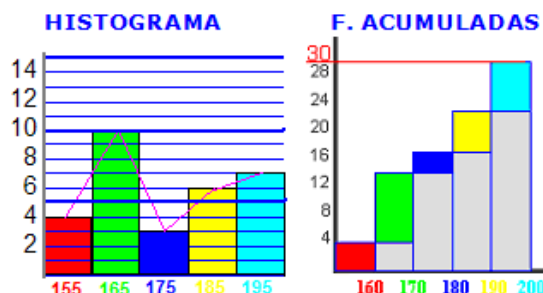
As áreas das barras-densidade resultan proporcionais ás frecuencias no intervalo.

Gráficos en variables continuas.

Histograma. Os datos represéntanse por rectángulos cuxa base é a amplitude do intervalo representado e coa altura que nos indica a frecuencia absoluta, se todos os intervalos son da mesma amplitude. Se non é o caso, as alturas calcúlanse de xeito que as áreas sexan proporcionais ás frecuencias absolutas. Á esquerda tes un exemplo feito.

Polígono de frecuencias. Uniremos os centros da parte superior de todos os rectángulos para obtelo. Tamén se adoita debuxar o histograma das **frecuencias acumuladas**, en cada dato acumúlase a frecuencia dos datos anteriores.

- [150, 160]→4
- [160, 170]→10
- [170, 180]→3
- [180, 190]→6
- [190, 200]→7



EXERCICIOS resoltos

3. Clasifica os seguintes exemplos de variables estatísticas: Lonxitude dun camiión, carga máxima, nº de rodas, nº de eixes, tipo de camiión, marcas de pneumáticos, tipo de tapicería, nº de portas, altura máxima.

Cualitativas: Tipo de camiión, marcas de pneumáticos, tipo tapicería

C. discretas: Nº de rodas, nº de eixes, nº de portas

C. continuas: Lonxitude dun camiión, Carga máxima e altura máxima.

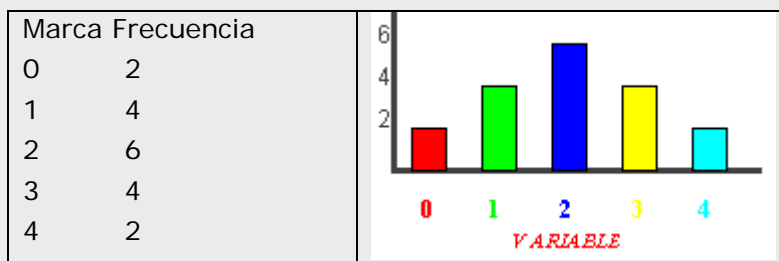
4. Calcula os graos que corresponden a cada valor en un gráfico de sectores feito a partir dos datos: R, R, V, V, V, V, V, V, A, A e A

Facemos o recuento $R \rightarrow 2$, $V \rightarrow 5$ e $A \rightarrow 3$ E calculamos

$$\frac{2}{10} = \frac{\text{Grados R}}{360}, \quad \frac{5}{10} = \frac{\text{Grados V}}{360} \quad \text{e} \quad \frac{3}{10} = \frac{\text{Grados A}}{360} \quad \text{e obtemos}$$

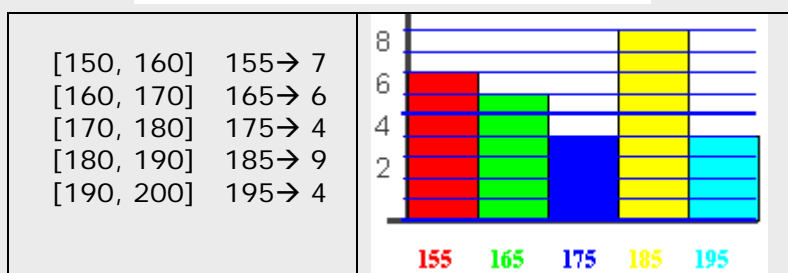
Graos R = 72, Graos V = 180 e Graos A = 108

5. Agrupa os datos seguintes e fai un diagrama de barras adecuado. Datos = { 0 1 0 0 2 3 4 1 2 2 1 2 2 3 4 3 2 1 3 }



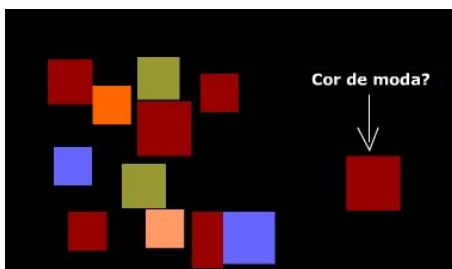
6. Clasifica os datos en intervalos e debuxa un histograma adecuado.

180 197 154 181 189 162 152 162 167 190
 189 160 166 197 187 194 152 181 173 154
 177 184 186 174 177 159 158 189 160 150



1ª AVALIACIÓN // Cor de moda?

1ª EVALUACIÓN	
5	
6	
4	NOTA
1	MEDIA
9	5,5
7	
6	
6	



Por exemplo, se temos as observacións 6,7,8,6,7,6,8,6,9 e agrupamos os datos vemos claramente que o valor 6 aparece máis que ningún outro. Neste caso a **moda** é 6.

xi	fr
6	4
7	2
8	2
9	1

Se ordenamos os datos, e dado que o nº de datos é impar xusto o 7 queda no centro.

6 6 6 6 7 7 8 8 1

Se os datos fosen 6,7,8,6,7,6,8,6,5 unha vez ordenados, e como hai unha cantidade par de datos, dous deles ocuparían o centro:

5 6 6 6 6 7 7 8 8 1

e a mediana será $(6+7)/2 = 6.5$

2. Medidas de centralización

Media, mediana e moda.

Un conxunto N de observacións, N números, pode que por si só non nos diga nada. En cambio, se ademais nos din que están situados arredor de un ou varios valores centrais xa temos unha referencia que sintetiza a información.

Media. A suma dos N números dividida entre N. Por exemplo, para 3, 4 e 5, $(3+4+5)/3 = 12/3 = 4$; para 1, 1, 4, 8, 8 e 8, $(1 \cdot 2 + 4 + 8 \cdot 3)/6 = 5$.

$$\text{Media} = \frac{X_1 \cdot f_1 + X_2 \cdot f_2 + \dots + X_n \cdot f_n}{f_1 + f_2 + \dots + f_n}$$

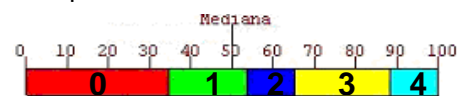
Moda. Se unha observación se repite máis que calquera outra, será considerada a moda deses datos. Por exemplo, se temos as observacións 6,7,8,6,7,6,8,6,9 e agrupamos os datos 6→4, 7→2, 8→2 e 9→1 vemos claramente que o valor 6 aparece máis que ningún outro. Neste caso a moda é 6.

No caso de variable continua, consideraremos por moda a marca do intervalo de maior frecuencia, cando isto ocorra. Tamén pode ocorrer que haxa dúas modas ou que non haxa ningunha que destaque.

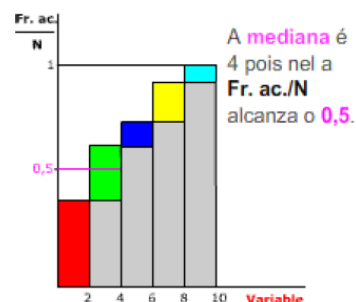
Mediana. O número tal que a metade das observacións son maiores ca el e a outra metade menores.

En xeral, para poucos datos o mellor é proceder segundo o exemplo da esquerda, segundo sexa unha cantidade par ou impar.

Para cantidades maiores, haberá que agrupar os datos primeiro nunha táboa. E determinar segmentos de lonxitude proporcional á súa frecuencia, dispoñelos de forma lineal e marcar o centro como mostra o seguinte exemplo.



Neste outro gráfico vemos indicada a mediana nun diagrama de Frecuencias relativas acumuladas:

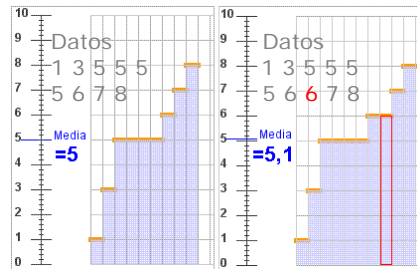
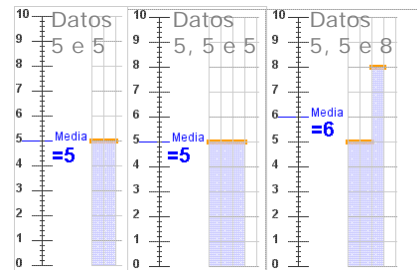


Media. Evolución desta ao engadir e/ou cambiar un dato.

1 Para os datos 5 e 5 a media é 5. Se engadimos un 5 mantense en 5. Se engadimos un 8 a media pasa a ser 6. (Figura dereita).

2 Se temos 9 datos con media 5, necesitamos engadir un 6 para que a media pase a ser 5,1. Se temos 19 datos con media 5, necesitamos un dato de valor 7 para que a media suba a 5,1. (Figura dereita).

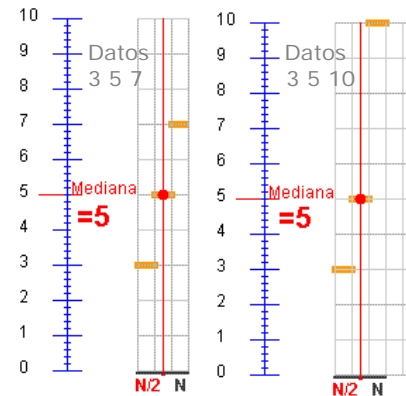
3 Para un conxunto de datos con media 5, se engadimos outro con media 5, por exemplo 6 e 4, o novo conxunto conserva a media.



Mediana. Evolución desta ao engadir e/ou cambiar un dato.

1 A mediana, para os datos 2, 3 e 4 é $Me=3$. Se cambiamos o 4 por 5 ou por 6 ou por calquera outro valor maior segue a ser $Me = 3$.

2 En cambio, se engadimos outro dato e temos 2, 3, 4 e 4, por exemplo, a $Me = 3,5$. E se agora engadimos un quinto valor, un 4 ou un 5 ou un 6 ou calquera outro maior que 4, a mediana en 2,3, 4, 4 e ?? pasa a ser 4. Dá igual o valor ?? é 5, 10 ou 25.

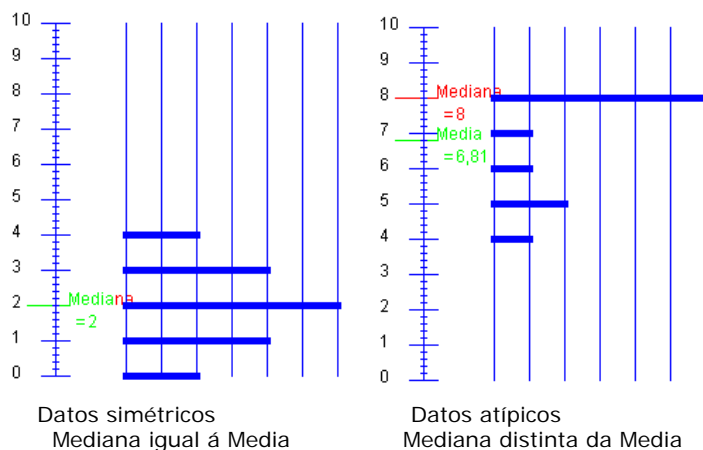


Para ver a mediana trázase unha vertical desde o eixe horizontal en $N/2$

Media e mediana comparadas

Para os datos 4 e 6 a media e a mediana coinciden en 5. Engadir un 8 ou un 11 dá o mesmo para a mediana, que pasa a ser en ambos os casos 6. Non obstante a media cun 8 pasa a ser 6 e cun 11 pasa a ser 7. Os valores 8 e 11 considéranse observacións atípicas, están distanciados do resto de valores, tiran da media e non afectan á mediana.

Se os datos estivesen repartidos simetricamente respecto a un valor, ese valor sería á vez a media e a mediana. En cambio, se os valores a un lado da mediana están máis afastados dela que os do outro lado, a media desprázase cara a eses valores afastados que tiran dela. Hai unha asimetría.



Datos simétricos
Mediana igual á Media

Datos atípicos
Mediana distinta da Media

Por exemplo, se temos as observacións

1. 20, 24 e 28.

$Me = 24$

2. E para 20, 24, 28 e 30

$Me = (24+28)/2 = 26$

3. Para 20, 24, 28 e 100

$Me = (24+28)/2 = 26$

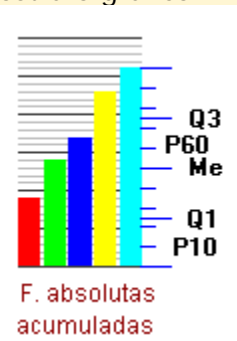
En cambio a media non se comporta da mesma forma para os mesmos datos

1 $\bar{X} = 24$

2 $\bar{X} = 25,5$

3 $\bar{X} = 43$

Tamén podemos facer un diagrama de frecuencias acumuladas e dividir en partes iguais como mostra o gráfico.

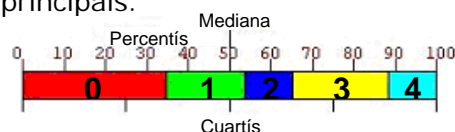


Medidas de posición: cuartís e percentís

Dado un conxunto de datos numéricos ademais da mediana podemos considerar outras medidas de posición

- Se nos fixamos no primeiro valor que supera o 25% ou o 75%, estamos a falar o **primeiro e terceiro cuartil**, Q_1 e Q_3 .
- Para outros valores como o 10%, ou o 80% falamos de **percentís**, P_{10} e P_{80} .

Exemplo. Para a variable de valores 0, 1, 2, 3, 4, e frecuencias $0 \rightarrow 9$, $1 \rightarrow 5$, $2 \rightarrow 3$, $3 \rightarrow 6$, $4 \rightarrow 3$, debuxamos barras de lonxitude proporcional ás frecuencias e dividimos o total en partes iguais: en dúas partes para a mediana, catro para os cuartís e 10 para os percentís principais.



EXERCICIOS resoltos

7. Calcula a media en cada caso:

- a) 4, 6, 8
b) 4, 6, 8, 6
c) 100, 120, 180, 200

- Solucións: a) $(4+6+8)/3 = 6$
b) $(4+6+8+6)/4 = 6$
c) $(100+120+180+200)/4 = 150$

8. Calcula a media en cada caso:

a

Marca	Fr
10	2
20	4
30	3
40	2

b

Marca	Fr
100	2
200	4
300	3
400	2

a) $\bar{X} = \frac{10 \cdot 2 + 20 \cdot 4 + 30 \cdot 3 + 40 \cdot 2}{11} = 24,54$

b) $\bar{X} = \frac{100 \cdot 2 + 200 \cdot 4 + 300 \cdot 3 + 400 \cdot 2}{11} = 245,45$

9. Determina a moda e a mediana

- a) 5,6,6 c) 1,2,3,4,2
b) 1,1,2,3 d) 3,2,3,2,2,2

- Solucións: a) $Me=6$, $Mo=6$ c) $Me=2$ $Mo=2$
b) $Me=1,5$ $Mo=1$ d) $Me=2$ $Mo=2$

10. Calcula a moda e a mediana en cada caso:

a

Marca	Fr
10	2
20	4
30	3
40	2

b

Marca	Fr
100	2
200	3
300	4
400	1

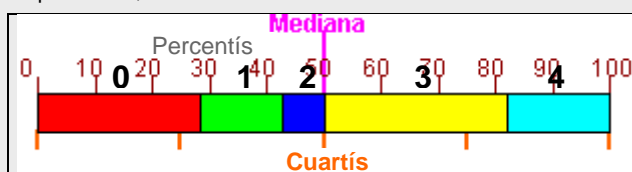
Solucións:

- a) $Me=20$ $Mo=4$
b) $Me=250$ $Mo=300$

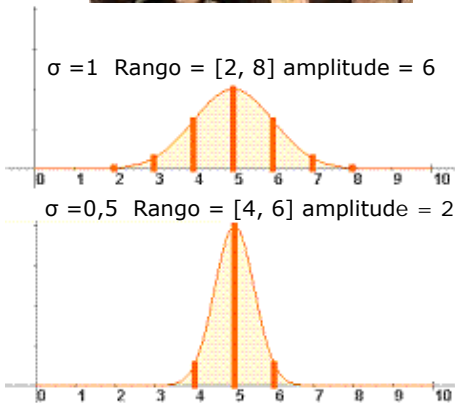
11. Calcula a mediana, cuartís primeiro e 3º, e o percentil 30 60 e 90 de os datos.

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

Facemos o reconto: $0 \rightarrow 8$, $1 \rightarrow 4$, $2 \rightarrow 2$, $3 \rightarrow 9$ e $4 \rightarrow 5$ e barras de lonxitude proporcional á frecuencia para cada valor. Ademais partimos a lonxitude total da barra en 2, 4 e 10 anacos para obter a mediana, cuartís e percentís, tal e como mostra a imaxe.



Vemos que a mediana está entre o azul e o amarelo, $(3+2)/2 = 2,5$, Q_1 entre vermello e verde, $Q_1=0$, $Q_3=3$, $p_{30}=1$, $P_{60}=3$ e $P_{90}=4$



En ambos os gráficos a media, mediana e moda valen 5

Na práctica adóitase usar a fórmula reducida para o cálculo da desviación típica.

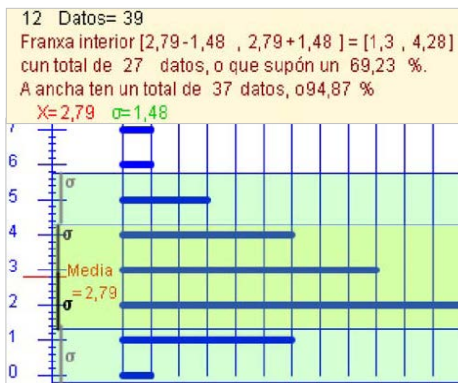
$$\sigma = \sqrt{\frac{\sum f_i X_i^2}{n} - \bar{X}^2}$$

Así, para

Marca	Fr
4	3
5	3
6	2

Tense que a media $\bar{X} = 4,85$ e

$$\sigma = \sqrt{\frac{3 \cdot 4^2 + 3 \cdot 5^2 + 2 \cdot 6^2}{8} - 4,85^2}$$



3. Medidas de dispersión.

Varianza, Desviación típica e rango

"A estatística é unha ciencia segundo a cal, se eu como un polo e ti non comes ningún, comemos como media medio polo cada un".

A estatística indicará que todos comen o mesmo cando as medidas de dispersión sexan todas nulas.

Rango. O intervalo definido polo menor e o maior dato. Tamén se chama rango a diferenza entre o maior e o menor de os datos.

Varianza. A media aritmética dos cadrados das diferenzas dos datos coa media.

$$\sigma^2 = \frac{\sum f_i (X_i - \bar{X})^2}{n} \text{ que equivale a } \sigma^2 = \frac{\sum f_i (X_i)^2}{n} - (\bar{X})^2$$

Desviación típica. A raíz cadrada positiva da varianza.

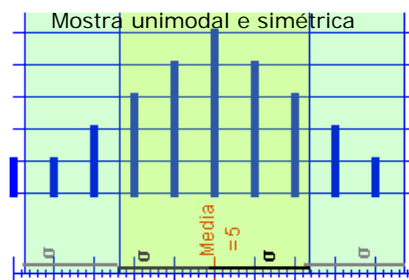
$$\sigma = \sqrt{\frac{\sum f_i (X_i - \bar{X})^2}{n}} \text{ ou } \sigma = \sqrt{\frac{\sum f_i X_i^2}{n} - \bar{X}^2}$$

Medir a dispersión

Ese é o obxectivo destas medidas. Por exemplo, os datos $A = \{20, 20\}$, $B = \{15, 20, 20, 25\}$ teñen a mesma media, moda e mediana. En todos os casos igual a 20. Non obstante, podes comprobar que en ningunha das tres medidas de dispersión definidas arriba coinciden.

Media e desviación típica.

Para mostras unimodais (unha soa moda) e case simétricas, arredor da media podemos considerar un intervalo que conteña a maioría dos datos. Por exemplo, para unha mostra con media 100 e desviación típica 10, a maior parte dos datos estarán entre 90 e 110, aproximadamente o 68%; entre 80 e 120 estará o 95% aproximadamente. E case todos entre 70 e 130. Hai unha forma de distribución de datos chamada **normal** que cumpre co anterior, e dunha maneira ou outra, de todas as poboacións grandes pódense extraer datos que se axustan a ela. En cursos superiores verás a importancia destas distribucións.



EXERCICIOS resoltos

12. Calcula a media e a desviación típica en

- a) 200, 250
- b) 175, 275
- c) 250, 250

$$a) \bar{X} = \frac{250 + 200}{2} = 225 \quad \sigma = \sqrt{\frac{(250 - 225)^2 + (200 - 225)^2}{2}} = \sqrt{\frac{25^2 + 25^2}{2}} = 25$$

$$b) \bar{X} = \frac{175 + 275}{2} = 225 \quad \sigma = \sqrt{\frac{(175 - 225)^2 + (275 - 225)^2}{2}} = \sqrt{\frac{50^2 + 50^2}{2}} = 50$$

$$c) \bar{X} = \frac{250 + 250}{2} = 250 \quad \sigma = \sqrt{\frac{(250 - 250)^2 + (250 - 250)^2}{2}} = \sqrt{\frac{0^2 + 0^2}{2}} = 25$$

13. Calcula a media e a desviación típica en:

- a) 7, 5, 3, 2, 4, 5
- b) 20, 25, 20, 22, 21

$$a) \bar{X} = \frac{7 + 5 + 3 + 2 + 4 + 5}{6} = \frac{26}{6} = 4,33$$

$$\sigma = \sqrt{\frac{7^2 + 5^2 + 3^2 + 2^2 + 4^2 + 5^2}{6} - 4,33^2} = \sqrt{\frac{128}{6} - 18,75} = 1,59$$

$$b) \bar{X} = \frac{20 + 25 + 20 + 22 + 21}{5} = \frac{108}{5} = 21,6$$

$$\sigma = \sqrt{\frac{20^2 + 25^2 + 20^2 + 22^2 + 21^2}{5} - 21,6^2} = \sqrt{\frac{2350}{5} - 466,56} = 1,85$$

(Nota.- Observa a fórmula utilizada para a desviación)

14. Organiza os datos seguintes en intervalos de 10 cm desde 150 a 200. Amplía a táboa con dúas columnas, unha para o produto das marcas coas frecuencias e outra para o produto das frecuencias cos cadrados das diferenzas coa media. Calcula a media e a desviación típica.

174	158	150	185	186	178	166	185	199
183	175	173	175	164	173	178	179	164
176	159	190	173	189	163	156	169	

	xi	fi	xi·fi	fi·(xi-X) ²
[150,160)	155	5	775	1733,65
[160,170)	165	5	825	371,58
[170,180)	175	10	1750	19,02
[180,190)	185	7	1295	906,42
[190,200)	195	2	390	914,14
Total		29	5035	3944,82

Cos datos da táboa é máis fácil, e tense:

$$\bar{X} = \frac{5035}{29} = 173,62 \quad \sigma = \sqrt{\frac{3944,82}{29}} = 11,66$$

4. Representatividade

Representatividade. Mostraxe estratificada.

REPRESENTATIVIDADE. Unha mostra é representativa da poboación cando nela podemos encontrar as mesmas proporcións das características de estudo que no conxunto da poboación. O proceso de elixir unha mostra, a qué individuos eliximos como representantes da poboación, é o punto importante e disto vai depender que o estudo sexa útil ou non (representativo ou non).

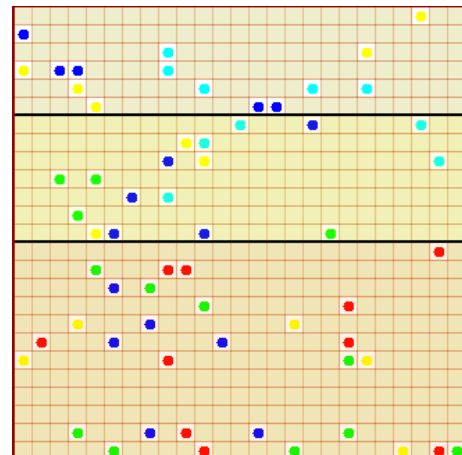
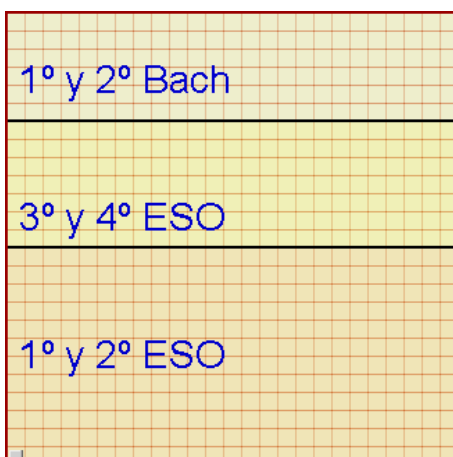
Elixir ben a mostra non é sinónimo de representatividade, pero elixila mal case é sinónimo de non representatividade.

Por exemplo, se queremos estudar o poder adquisitivo dunha poboación, e só eliximos individuos dunha determinada zona, ou principalmente dunha determinada zona, a mostra con toda seguridade non será representativa. A mostra tense que elixir tomando mostraxas de individuos proporcionais á poboación de cada zona. Se hai tres zonas con 12.000, 18.000 e 20.000 habitantes, a mostra deberá ter un 24% da primeira zona, 36% da segunda e 40% da última.

Este tipo de **mostraxe**, escollendo un reparto proporcional aos estratos, chámase **estratificado**.

Exemplo

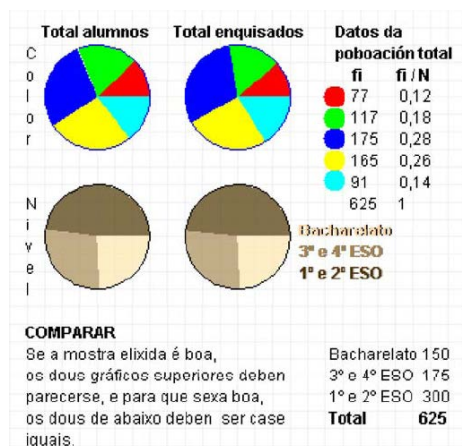
Na imaxe tes 625 cadros que representan aos alumnos dun instituto ficticio



fi	fi / N	DATOS DA
9	0,14	● BACHARELATO
15	0,24	● 3º e 4º ESO
11	0,17	● 1º ciclo ESO
17	0,27	● 2º ciclo ESO
10	0,16	● 1º e 2º ESO
62	1	MOSTRA

MOSTRAXE
 Ten en conta os alumnos que hai en cada nivel:
 1º e 2º Bacharelato 150 alumnos
 3º e 4º ESO 175 alumnos
 1º e 2º ESO 300 alumnos

Bacharelato	15
2º ciclo ESO	17
1º ciclo ESO	30
Total	62
Porcentaxe	9,92 %



Á dereita vemos a mostra estratificada que se elixiu e o resultado da enquisa. Os últimos diagramas de sectores comparan a realidade cos resultados da enquisa.

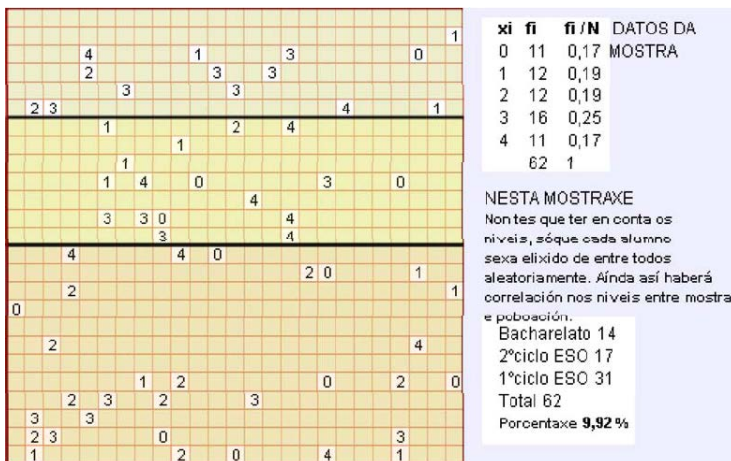
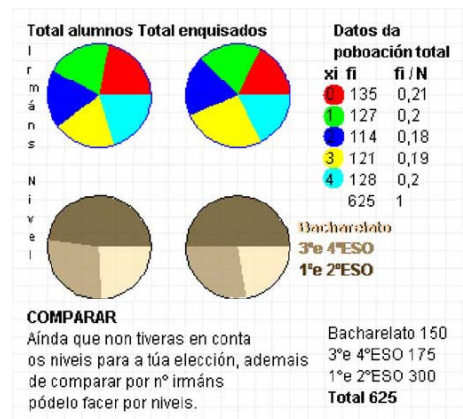
Nesgo. Mostraxe aleatoria

Nesgo. Dise que a mostra está nesgada cando hai diferenza entre os datos da mostra e os datos de toda a poboación.

Exemplo: Chamadas telefónicas voluntarias. Estas enquisas teñen varias fontes de nesgo. Hai familias que non teñen teléfono, o custo da chamada non todo o mundo está disposto a asumilo. Pero sobre todo, o factor de resposta voluntaria, os enquisados auto-seleccionáanse. Adoitan contestar aqueles cunha forte opinión negativa sobre o tema. O enfado animaos a participar.

Mostraxe aleatoria total. A diferenza do estratificado, que garda as proporcións, esta forma de elixir a mostra considera a toda a poboación e elixe individuos aleatoriamente. Considérase unha boa forma de proceder.

No seguinte exemplo escolleuse con ordenador unha mostra aleatoria total entre os 625 alumnos dun instituto, esta mostraxe pode saír estratificada ou non, no exemplo non saíu moi ben estratificada.



EXERCICIOS resoltos

15. Unha grande empresa ten traballadores en catro áreas. Operarios, Representantes, administración e dirección. As condicións de traballo son bastantes diferentes en cada área, polo que o grao de satisfacción non é igual en cada unha delas. Para descubri-lo, se hai 1000, 500, 300 e 200 traballadores nas áreas de operarios, representantes, administrativos e directivos, cantos hai que seleccionar de cada área para unha mostra de tamaño?
- 200
 - 100
 - 300
- a) Dun total de 2000 empregados, as porcentaxes para operarios, repartidores, administrativos e directivos son do 50%, 25%, 15% e 10%. O cal fai que a mostra tome 100 operarios, 50 repartidores, 30 administrativos e 20 directivos.
- 50, 25, 15 e 10.
 - 150, 75, 45 e 30

5. Estatística bidimensional

Distribucións bidimensionais

Unha distribución bidimensional é aquela na que interveñen dúas variables X e Y que poden ou non estar relacionadas.

Podemos representar conxuntamente as dúas variables nun **diagrama de dispersión** ou **nube de puntos**, simplemente asociando un punto do plano a cada par (x_i, y_i) .

Cando hai moitos datos e os pares de valores se repiten, acudimos a unha **táboa de continxencia**, como a da dereita. Neste caso a representación gráfica faise mediante un gráfico tridimensional chamado prismograma, ou máis polo miúdo poñendo puntos de tamaño proporcional á frecuencia.

Correlación lineal

O obxectivo dun estudo bidimensional é observar se hai algún tipo de relación entre as dúas variables. Esta relación, que chamaremos **correlación**, pódese apreciar vendo se a nube de puntos se achega ou non á gráfica dunha función, no noso caso a unha recta, por iso falaremos de correlación lineal.

Canto máis se achegue a nube de puntos a unha recta máis forte será a correlación lineal, ademais será positiva ou directa se a recta é crecente (se crece X crece Y) e negativa ou inversa en caso contrario (se crece X decrece Y ou viceversa).

Para cuantificar esta relación empregaremos un parámetro **r**, o **coeficiente de correlación lineal**, que toma valores entre -1 e 1. Canto máis se achega a valer 1 ou -1 máis forte será a correlación.

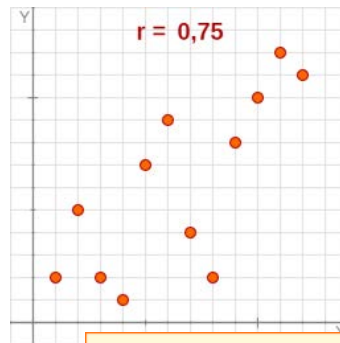
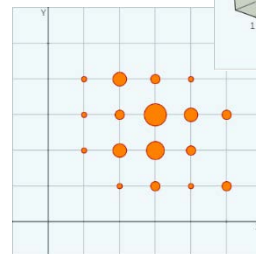
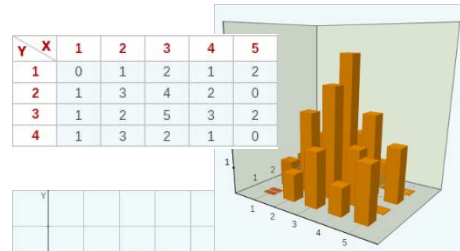
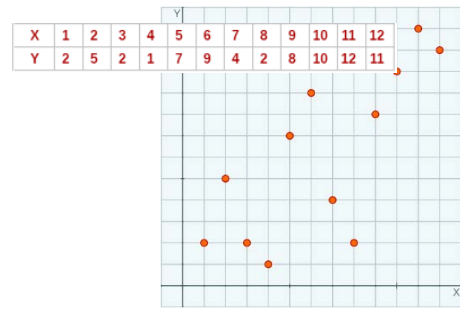
- Se **r=1** ou **r=-1**, hai dependencia funcional, os puntos están sobre unha recta.
- Se **0,5 < r < 1** consideraremos que a correlación é forte e directa ou forte e inversa se **-1 < r < -0,5**.
- Se **r=0** ou moi próximo a 0, non hai correlación lineal entre as dúas variables.

Dispoñemos os datos en columnas e calculamos a **media** e a **desviación típica** das dúas distribucións X e Y. Para calcular r temos que calcular antes outro parámetro, a **covarianza**, definida como:

$$\sigma_{XY} = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} = \frac{\sum x_i \cdot y_i}{n} - \bar{x} \cdot \bar{y}$$

E o **coeficiente de correlación lineal**:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \quad (-1 \leq r \leq 1)$$



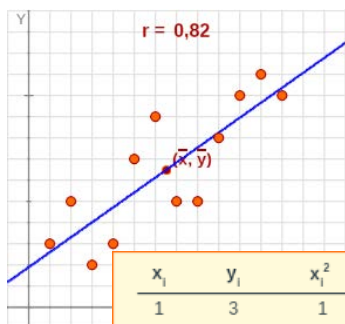
x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	2	1	4	2
2	5	4	25	10
3	2	9	4	6
4	1	16	1	4
5	7	25	49	35
6	9	36	81	54
7	4	49	16	28
8	2	64	4	16
9	8	81	64	72
10	10	100	100	100
11	12	121	144	132
12	11	144	121	132
78	73	650	613	591

$$\bar{x} = \frac{78}{12} = 6,5 \quad \sigma_x = \sqrt{\frac{650}{12} - 6,5^2} = 3,45$$

$$\bar{y} = \frac{73}{12} = 6,08 \quad \sigma_y = \sqrt{\frac{613}{12} - 6,08^2} = 3,75$$

$$\sigma_{xy} = \frac{591}{12} - 6,5 \cdot 6,08 = 9,71$$

$$r = \frac{9,71}{3,45 \cdot 3,75} = 0,75$$



x_i	y_i	x_i^2	$x_i \cdot y_i$
1	3	1	3
2	5	4	10
3	2	9	6
4	3	16	12
5	7	25	35
6	9	36	54
7	5	49	35
8	5	64	40
9	8	81	72
10	10	100	100
11	11	121	121
12	10	144	120
78	78	650	608

$\bar{x} = 6,5$ $\bar{y} = 6,5$ $\sigma_x = 3,45$ $\sigma_{xy} = 8,42$

Recta de regresión de Y sobre X

$$y = 6,5 + \frac{8,42}{3,45^2} (x - 6,5)$$

$$y = 0,7x + 1,9$$

Rectas de regresión

Cando se aprecia un certo grao de correlación entre as dúas variables dunha distribución bidimensional, búscase a recta que mellor se axusta á nube de puntos.

A recta de regresión de Y sobre X pasa polo punto (\bar{x}, \bar{y}) , centro de gravidade da nube de puntos. A súa ecuación é:

$$y = \bar{y} + \frac{\sigma_{XY}}{\sigma_X^2} (x - \bar{x})$$

A pendente $\frac{\sigma_{XY}}{\sigma_X^2}$ é o **coeficiente de regresión**.

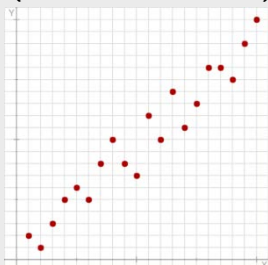
Esta recta de regresión serve para estimar o valor que tomará a variable Y para un determinado valor de X. O valor desta estimación será tanto máis fiable canto:

- Máis se achegue o coeficiente de correlación a 1 ou a -1.
- O valor quede dentro do rango de valores de X e máis cerca esté do centro de gravidade.

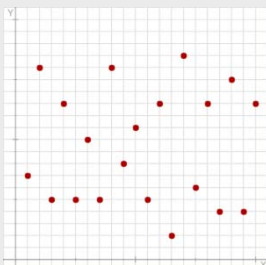
Se se quere estimar valores de X a partir dos de Y hai que utilizar outra recta análoga, a recta de regresión de X sobre Y.

EXERCICIOS resoltos

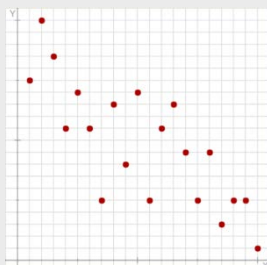
1. En vista da nube de puntos indica se cres que a correlación é moi forte, forte (directa ou inversa), débil ou moi débil.



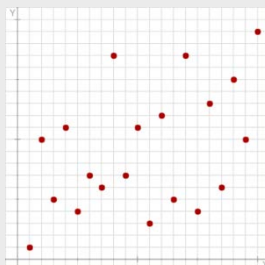
Moi forte e directa



Moi débil

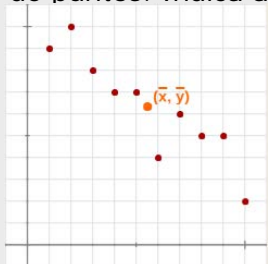


Forte e inversa



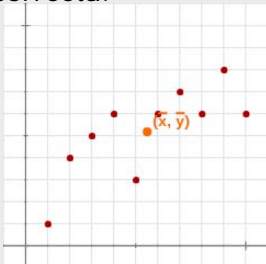
Débil

2. Unha das catro ecuacións corresponde á recta de regresión de Y sobre X da nube de puntos. Indica a correcta.



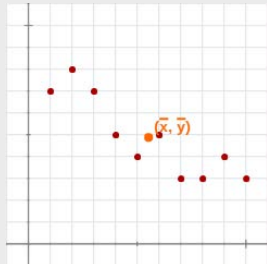
- $y = 0,7x + 2,4$
- $y = -0,7x + 10,2$
- $y = 0,7x + 10,2$
- $y = -0,7x + 2,4$

Sol: (b)



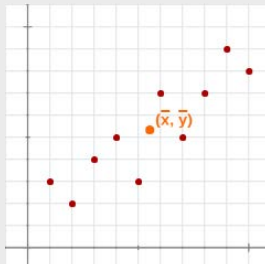
- $y = -0,5x + 2,5$
- $y = 0,5x + 8$
- $y = -0,5x + 8$
- $y = 0,5x + 2,5$

Sol: (d)



- $y = -0,5x + 2,2$
- $y = 0,5x + 2,2$
- $y = -0,5x + 7,7$
- $y = 0,5x + 7,7$

Sol: (c)



- $y = -0,7x + 1,4$
- $y = 0,7x + 9,2$
- $y = -0,7x + 9,2$
- $y = 0,7x + 1,4$

Sol: (d)

Alguns dos exercicios propostos a continuación están elaborados a partir desta publicación de INE. Podes ver artigos similares en

<http://www.ine.é/prodyser/pubfolletos.htm>

4/2007

cifras

INÉ

Boletín Informativo del Instituto Nacional de Estadística



Encuesta de Empleo del Tiempo

Qué hacemos y durante cuánto tiempo



El Instituto Nacional de Estadística (INE) presenta en esta publicación algunos de los principales resultados de la **Encuesta de Empleo del Tiempo**, primera y única encuesta de ámbito nacional sobre la utilización del tiempo. Se realizó en España entre los años 2002 y 2003 de manera armonizada con las de otros países europeos, siguiendo las recomendaciones de la Oficina Estadística de la Unión Europea (Eurostat). Entre los años 1998 y 2004 otros países de la Unión llevaron a cabo investigaciones similares.

La encuesta facilita información, entre otras cosas, del **porcentaje de personas que realizan una determinada actividad en el transcurso del día y la duración media diaria dedicada a esa actividad por dichas personas**. Esta información primaria nos permite analizar con rigor la dimensión del trabajo no remunerado realizado por los hogares, la distribución de las responsabilidades familiares en el hogar, la participación de la población en actividades culturales y de ocio, etc. Por otra parte, la información recogida también permite comparar **datos nacionales de uso del tiempo en relación con los demás países europeos que han realizado la encuesta**.

Como principales resultados, cabe destacar el dato de que **las tareas domésticas y el cuidado de niños y ancianos son tareas eminentemente femeninas, ya que el 93% de las mujeres las realizan, frente al 70% de los varones**. En el contexto europeo, es de señalar la **primera posición de España en tiempo dedicado a caminar y pasear; pero también el último lugar por lo que se refiere a tiempo dedicado a la lectura**.

El INE quiere aprovechar esta ocasión para expresar su **agradecimiento a los cerca de 24.000 hogares de la muestra**, y pone a su disposición los resultados obtenidos.

Distribución del tiempo por actividades



Actividad	Porcentaje
Cuidados personales	47,4%
Hogar y familia	12,4%
Trabajo	11,0%
Medios de comunicación	9,5%
Vida social y diversión	8,2%
Trayectos y tiempo no especificado	4,9%
Deportes y actividades al aire libre	3,3%
Estudios	3,0%
Aficiones y juegos	1,4%
Trabajo voluntario y reuniones	0,9%

NOTA: Los informantes de 10 y más años han anotado las actividades realizadas en un día concreto (de lunes a domingo) elegido al azar. El tiempo así estimado se refiere a un "día promedio" obtenido al concentrar todas las actividades de todos los informantes en un solo día. Los datos que aquí se presentan se refieren a toda la población investigada, salvo que se indique expresamente lo contrario.



Más información en:

www.ine.es

DEPÓSITO LEGAL: M-1.2047-2001 ISSN: 1578-2227 N.º P.º: 685-07-005-1

Fuentes estadísticas utilizadas:
Procedentes del INE: Encuesta de Empleo del Tiempo. La información internacional procede de Eurostat.



Para practicar

1. Clasifica as seguintes variables: a) Peso, b) densidade, c) nº de plantas dos edificios, d) Tipo de fachada dos edificios, e) nº de ventás, f) metros de fachada, g) nº de habitantes por edificio, h) tipo de porta principal.

2. Escribe tres variables cualitativas que teñan que ver con embarcacións.

3. Escribe tres variables cuantitativas discretas que teñan que ver con avións.

4. Escribe tres variables cuantitativas continuas que teñan que ver con trens.

5. Se as frecuencias para R, V, A e T son $R \rightarrow 3$, $V \rightarrow 2$, $A \rightarrow 4$ e $T \rightarrow 1$. Cantos graos lle corresponde a cada letra nun gráfico de sectores?

6. Fai unha táboa e un gráfico de sectores dos datos: R R A R A R V N V R N

7. Fai unha táboa e un gráfico de barras cos datos:
3 3 4 5 4 5 3 2 1 2 3 4 5 4 5 4 3 3 4 4

8. Agrupa os datos seguintes en intervalos e fai un histograma.

195 194 194 182 168 179 191 154 177 189
184 187 155 167 177 187 161 171 190 162
190 152 166 180 156 186 184 167 184 162

9. Calcula a media en cada caso:

- a) 4, 6, 8
- b) 4, 6, 8, 6
- c) 100, 120, 180, 200

10. Calcula a media en cada caso:

Marca	Fr
1	3
2	5
3	3
4	2

Marca	Fr
1000	3
2000	5
3000	3
4000	2

11. Determina a moda e a mediana

- a) 50,60,60
- b) 12,12,22,32
- c) 10,20,30,40,20
- d) 35,25,35,25,25,25

12. Calcula a moda e a mediana en cada caso:

Marca	Fr
100	5
200	4
300	6
400	3

Marca	Fr
100	2
200	7
300	9
400	2

13. Cal ou cales dos datos seguintes se pode considerar unha observación atípica en cada unha das dúas series?

- a) 4 5 6 5 7 8 4 5 8 7 5 12 6 7 6 5 4
- b) 8 9 1 9 8 9 7 9 6 7 8

14. Calcula a mediana, o primeiro e o terceiro cuartil e o percentil 90 de:

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

15. Calcula a mediana, primeiro e terceiro cuartil e o percentil 20 de

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

16. Calcula a media e a desviación típica en cada un dos seguintes casos:

- a) 100 e 100
- b) 99 e 101
- c) 110 e 90
- d) 120 e 80

17. Completa a táboa cos datos:

190 151 193 187 158 175 165 158 184 172
197 161 157 157 183 180 150 161 182 169
162 177 160 155 188 157 189 167 186 157

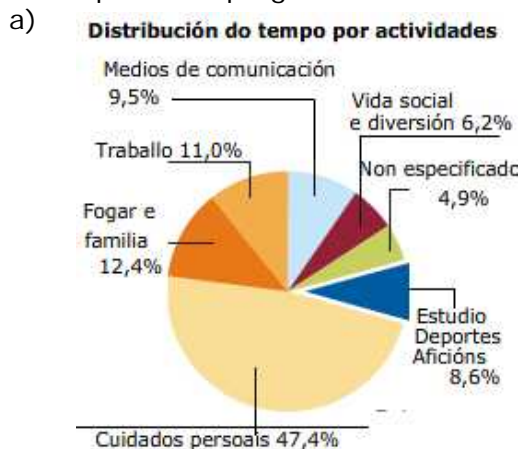
Intervalo	Marca	Frec.		
	x_i	f_i	$f_i \cdot x_i$	$f_i(x-x_i)^2$
[150,160)	155			
[160,170)	165			
[170,180)	175			
[180,190)	185			
[190,200)	195			

18. Determina a media e a desviación típica, dos datos da táboa anterior.

19. Determina os intervalos $(\bar{X} - \sigma, \bar{X} + \sigma)$ e $(\bar{X} - 2\sigma, \bar{X} + 2\sigma)$ e o número de elementos que hai en cada un.

Marca	Fr
0	5
1	4
2	7
3	3
4	2

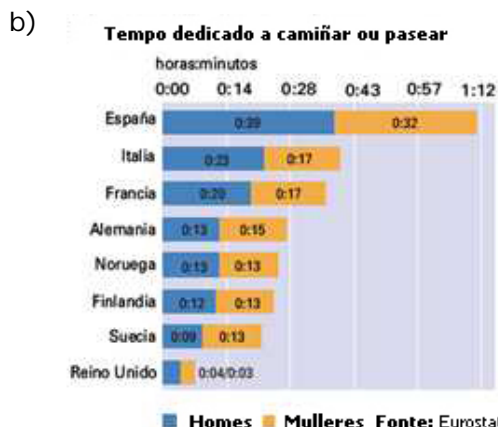
20. Observa os seguintes gráficos e responde ás preguntas de cada un



a1. Cal é a variable estudada? e a frecuencia?

a2. A que grupo de actividades dedicamos máis tempo os españois?

a3. Calcula canto tempo dedicamos ao fogar e a familia cantos graos ocupa este sector no diagrama?



b1. En que países pasean máis as mulleres que os homes?

b2. Calcula o tempo medio que se dedica en cada país a pasear.

b3. Que país está no percentil 50?



c1. Cres que durmir se contou como actividade de coidado persoal?

c2. Ás 15:00 hai un máximo local na gráfica a que se debe?

c3. Á hora da comida o 38% das persoas dedícase ao coidado persoal. Significa isto que un 62% das persoas non come?



d1. Cales son as comunidades nas que se dedica menos tempo á vida social e á diversión?

d2. Canto tempo dedican á diversión ou á vida social a maior parte das comunidades?

d3. Cal é o tempo medio que se dedica en España a esta actividade?

21. As cualificacións de 8 alumnos en Lingua e Inglés foron:

Lingua (X)	4	4	4	5	7	8	9	9
Inglés (Y)	3	5	6	5	8	9	10	9

Debuxa a nube de puntos e calcula o coeficiente de correlación lineal.

22. As horas semanais que adican 10 persoas a facer deporte e ver TV son:

Dep (X)	1	3	5	6	7	8	9	10	11	12
TV (Y)	14	14	13	10	8	9	4	8	5	5

Debuxa a nube de puntos e calcula o coeficiente de correlación lineal.

23. Dunha distribución bidimensional conocemos $\bar{x} = 8, \bar{y} = 7, \sigma_x = 1,5, \sigma_y = 2,7$ e $\sigma_{XY} = 3,28$. Calcula o coeficiente de correlación lineal, a recta de regresión de Y sobre X e o valor estimado de y para $x = 8$.

24. Dunha distribución bidimensional conocemos $\bar{x} = 8, \bar{y} = 5, \sigma_x = 1,9, \sigma_y = 2,5$ e $r = 0,83$. Calcula a recta de regresión de Y sobre X e o valor estimado de y para $x = 10$. És fiable esta predición?.



Para saber máis

A profesión de enfermería.

Florence Nightingale (1820-1910), coñecida por ser a fundadora da profesión de enfermería. Durante a guerra de Crimea decatouse de que a causa principal das mortes de feridos en combate era a falta de medidas sanitarias. Ao aplicalas, a taxa de mortalidade pasou dun 42,7% a un 2,2%. Grazas a un uso eficaz dos datos conseguiu modificar o sistema de atención sanitaria á súa volta a Gran Bretaña. Cambiou o sistema de rexistro de datos e foi unha das primeiras persoas en utilizar os gráficos estadísticos para representar os datos dunha forma sinxela de forma que ata os parlamentarios e xenerais puidesen entender.

Para Florence, os datos non eran algo abstracto, eran unha forma de poder salvar vidas humanas.

O pai da estatística.

Sir Ronald A. Fisher (1890-1962) está considerado o pai da estatística. Os escritos de Fisher axudaron a organizar a estatística como campo de estudo preciso cuxos métodos se aplican a problemas prácticos de moitas disciplinas. Como case todos os pioneiros na estatística, os seus traballos naceron da necesidade de resolver problemas prácticos.

Inferencia estatística

A estatística desenvolvida neste tema é o que se coñece como estatística descritiva, nela recóllese información e fanse cálculos que describen como están repartidos. Poñamos o caso que unha mostra elixida ao azar dáunos unha media. A verdadeira media está próxima á da mostra? Se considero un intervalo arredor da media da mostra, con qué probabilidade estará ou non nel a verdadeira? Destas preguntas e outras encárgase a inferencia estatística.

Principais campos de aplicación da estatística



A estatística aplícase en moitos campos como en **Industria e empresas**. Para o control de calidade na produción en cadea, para a análise de mercados, para o estudo de prezo de venda ao público dos artigos fabricados, en xestión financeira,...

Na parte dereita cítanse algunhas outras das súas aplicacións.

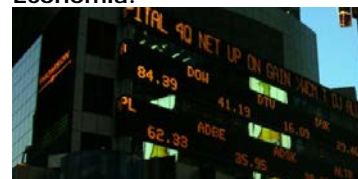
Alguns campos de aplicación da estatística

Administración pública



A través das Delegacións territoriais e provinciais, recóllese datos para analízalos e sometelos a procesos estadísticos. Desta forma coñécense datos referidos a nacementos, defuncións, matrimonios, prezos, salarios, traballo, ensino, sanidade,... Todos estes datos adóitao publicar o INE.

Economía.



Neste campo é imprescindible, sobre todo en macro-magnitudes.

Psicología.



A maior parte dos traballos científicos en psicología experimental teñen como principal ferramenta de traballo a estatística.

Medicina.



En calquera estudo experimental destas áreas existe unha materia específica chamada Bioestatística para cubrir estes estudos experimentais. En Xenética e antropometría encontramos dous dos campos de maior aplicación.



Lembra o máis importante

Poboación. Alumnos dun instituto ficticio.

Mostra. Alumnos enquisados

Variables estadísticas: Cualitativa, cor preferida; Cuantitativa discreta, nº de irmáns e cuantitativa continua, altura.

Consideremos as dúas mostras seguintes:

Nº de irmáns: 4 3 2 3 1 2 0 2 0 1 2 3 1 2 4 0 1 1 4 1 1 4 0 4 2 0 4 1

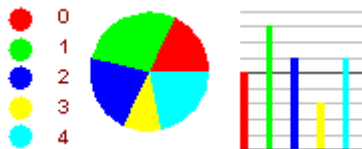
Altura: 182 172 157 194 150 166 163 196 167 199 172 185 172 168 173 160 162 173 161 192 156 164 173 180 193 172

Reconto de datos:

x_i	f	Intervalo	x_i	f_i
0	5	[150,160)	155	3
1	8	[160,170)	165	8
2	6	[170,180)	175	7
3	3	[180,190)	185	3
4	6	[190,200)	195	5
	28	Total		26

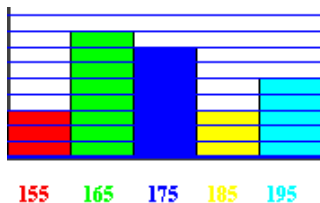
Gráficos de **sectores** e **barras**

Nº de irmáns



Altura.

Histograma



Media e moda e desviación típica

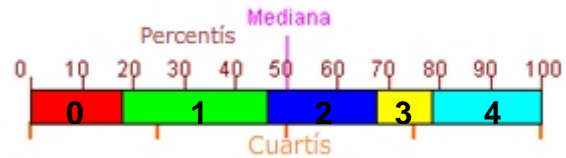
x_i	f_i	$x_i \cdot f_i$	$f_i \cdot (x_i - \bar{x})^2$
0	5	0	0
1	8	8	6,37
2	6	12	0,06
3	3	9	3,67
4	6	24	26,64
Total	28	53	54,67

Media: $\bar{x} = \frac{53}{28} = 1,89$

Moda = 1

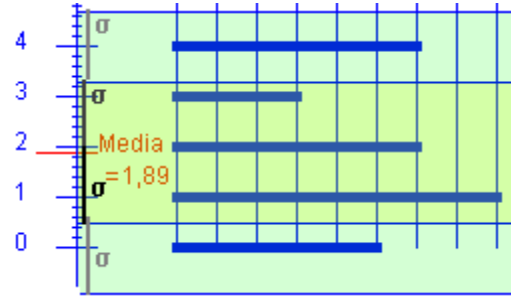
Desviación típica: $\sigma = \sqrt{\frac{54,67}{28}} = 1,39$

Cuartil, mediana, percentil



Me=2, Q1=1, Q3=3,
P20=1, P60=2, P90=4
Percorrido. De 0 a 4, de amplitude 4

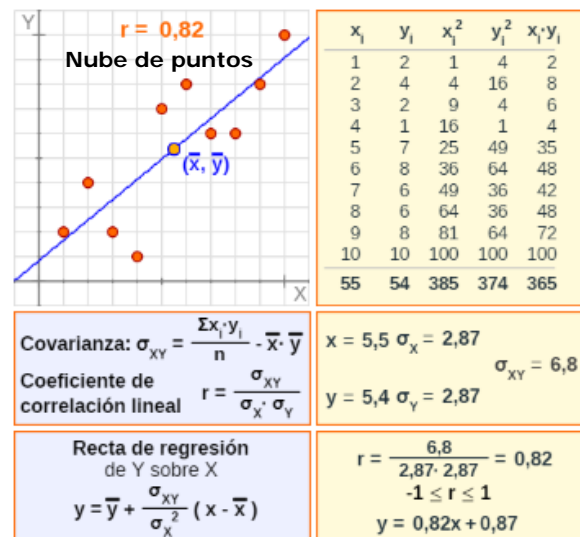
Media e desviación No noso exemplo, 17 de 28 datos non se afastan da media máis da desviación típica, son o 60,7%, e o 100% non se afastan da media mais de dúas veces a desviación.



Representatividade

Unha mostra é representativa da poboación cando nela podemos atopar as mesmas proporcións das características de estudo que no conxunto da poboación.

Distribucións bidimensionais



Autoavaliación



1. Cantos graos corresponden ao valor de frecuencia 3?

X_i	F_i
0	1
1	2
2	3
3	8
4	2

2. A mediana da distribución anterior é?

3. Cal é a moda?

X_i	F_i
3	1
4	3
5	4
6	2

4. Cal é a porcentaxe da mostra que corresponde ás dúas primeiras marcas ?

X_i	F_i
0	1
1	2
2	3
3	8
4	2

5. Cal é o percentil máis pequeno que deixa por debaixo os valores menores a 3?

X_i	F_i
0	1
1	2
2	3
3	8
4	2

6. Cal é a media?

X_i	F_i
155	2
165	2
175	8
185	1
195	5

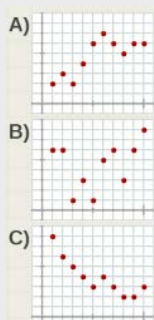
7. Cal é a desviación típica dos datos anteriores?

8. Asocia cada nube de puntos (A, B, C) co seu coeficiente de correlación:

- 1) 0.32 2) 0.79 3) -0,88

9. Na distribución da esquerda calcula a covarianza.

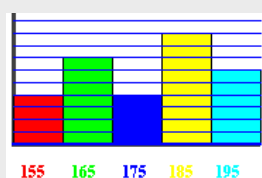
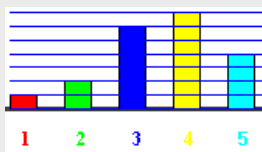
10. O centro de gravidade dunha distribución bidimensional é $(4,5, 3,75)$ e a pendente da recta de regresión de Y sobre X é 0,57. Estima o valor de y para $x = 7$



x_i	y_i	$x_i \cdot y_i$
1	1	1
2	2	4
3	3	9
4	5	20
5	4	20
6	5	30
7	5	35
8	5	40

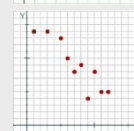
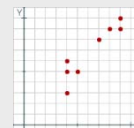
Soluciones dos exercicios para practicar

- Cualitativas: d) h)
Cuantitativas discretas c) e) g)
C. continuas: a) b) f)
- Propulsión, Carga, Tipo de travesía
- Nº de pasaxeiros, nº rodas, nº ventás
- Velocidade máxima, carga máxima, potencia.
- $R \rightarrow 108^\circ$, $V \rightarrow 72^\circ$, $A \rightarrow 144^\circ$, $T \rightarrow 36^\circ$
- $R \rightarrow 5$,
 $A \rightarrow 3$,
 $V \rightarrow 2$,
 $N \rightarrow 2$
- $1 \rightarrow 1$, $2 \rightarrow 2$, $3 \rightarrow 6$,
 $4 \rightarrow 7$, $5 \rightarrow 4$
- Intervalo x_i f_i
[150,160) 155 4
[160,170) 165 7
[170,180) 175 4
[180,190) 185 9
[190,200) 195 6
- a) 6 b) 6 c) 150
- a) 2.3 b) 2307
- a) $Mo=60$, $Me=60$ b) $Mo=12$, $Me=17$
c) $Mo=20$, $Me=20$
d) $Mo=25$ $Me=25$
- a) $Mo=300$, $Me=250$ b) $Mo=300$,
 $Me=300$
- a) 12 b) 1
- $Me=2$, $Q1=2$, $Q3=3$, $P90=4$
- $Me=3$, $Q1=1$, $Q3=4$ e $P20=1$
- A media é 100 nos 4, e a desviación 0, 1, 10 e 20.



Intervalo	Marca	Frec.	$f_i \cdot x_i$	$f_i (x-x_i)^2$
[150,160)	155	9	1395	2401
[160,170)	165	7	1155	280,77
[170,180)	175	3	525	40,33
[180,190)	185	8	1480	1494,22
[190,200)	195	3	585	1680,33
		30	5140	5896,66

- $\bar{x} = 171,3$ $\sigma \approx 14,02$
- En (0.42, 2.9) hai 11,
e en (-0.88, 4.14) todos
- a1) variable: actividades. Fr: porcentaxe de tempo diario que se dedica a cada actividade
a2) coidados persoais
a3) 2 h 58 m 34 s 44,64grados
b1) Alemaña, Suecia e Finlandia
b2) E35,5 I20, F18,5 A14 N13 F12,5 S11 R3,5 en minutos
b3) Francia
c1) Si. c2) Comida e Sesta
c3) Non, o pico ocupa dúas horas e algúns comen en media hora
d1) País Vasco, Cataluña e Madrid
d2) entre 1:30 e 1:40 horas: minutos
d3) 1:29
- $r = 0,93$
- $r = -0,91$
- $r = 0,81$
 $y = 1,46x - 1,66$
 $x = 8, y' = 7$
- $y = 1,09x - 3,73$
 $x = 10, y' = 7,18$ Bastante fiable



Soluciones AUTOAVALIACIÓN

- | | |
|---------------|----------------|
| 1. Sol 67,5° | 6. Sol 177,78 |
| 2. Sol 3 | 7. Sol 12,83 |
| 3. Sol 5 | 8. 1-B 2-A 3-C |
| 4. Sol 18,75% | 9. Sol 3 |
| 5. Sol 37 | 10. Sol 5,18 |