

Unidade 10: Estatística

- 1 **Introducción á Estatística Descritiva**
 - 1.1 Fenómenos aleatorios e determinísticos.
 - 1.2 Concepto de poboación estatística, mostra e tamaño mostral.
- 2 **Variable estatística monodimensional.**
 - 2.1 Definición de variables estatísticas
 - 2.2 Tipos: Cualitativas e cuantitativas, discretas e continuas.
- 3 **Variables cualitativas.**
 - 3.1 Frecuencia absoluta e relativa.
 - 3.2 Gráficos.
- 4 **Variables cuantitativas**
 - 4.1 Medidas tendencia central: media, mediana e moda
 - 4.2 Medidas de dispersión: Percorrido, varianza, desviación típica e coeficiente de variación de Pearson..
 - 4.3 Variables discretas.
 - 4.4 Variables continuas.
- 5 **Variables bidimensionais**
 - 5.1 Táboas de frecuencias.
 - 5.2 Covarianza.
 - 5.3 Regresión e correlación.

Introducción

Estatística

A Estatística xurde no século XVII aplicada ó estudio de poboacións. Merecen especial mención o *Tratado sobre as anualidades da vida* de Jan de Witt, gran conselleiro de Holanda, e o estudio sobre a poboación de Londres de John Graunt.

Cando se reúnen miles ou millóns de datos referidos a un fenómeno (idade no intre de morrer, número de fillos por parella, votos a un partido político, quilómetros os que é necesario cambiar unha peza dun coche, etc.) é necesario facer comprensibles eses datos xa sexa resumíndoos nun único valor que os represente (a media, por exemplo), mediante gráficos ou táboas, ou utilizando unha serie de parámetros que os describan. Ese é o obxecto da Estatística.

Neste senso, os primeiros estudos estatísticos tiñan por obxecto a recollida dos datos máis relevantes dun estado e da súa poboación, de aí o nome de Estatística.

Estatística hoxe

Na actualidade, a Estatística é unha das partes máis utilizadas das Matemáticas e en parte segue conservando o seu carácter descritivo: Fanse estudos para valorar o comportamento dos consumidores, dos votantes, da esperanza de vida, dos efectos do tabaco, dos accidentes de tráfico, do aumento dos prezos, ... Estudos que consisten basicamente na recollida e manexo de datos.

Pero a Estatística ten tamén outro ámbito de actuación: A predicción (toda a ciencia occidental baséase na capacidade de predicción e a Estatística non é unha excepción).

Para predicir, por exemplo, cal vai se-lo comportamento dos consumidores fronte a un novo produto fanse unha serie de enquisas para recoller datos, pero iso non é suficiente. É necesario saber en que medida os datos recollidos reflicten a situación real e en que medida podemos predicir o comportamento da totalidade da poboación a partir dos recollidos de só a uns poucos individuos.

Para responder esas e outras preguntas foi necesario dotar a Estatística dunha forte base teórica, a Probabilidade, polo que hoxe en día non ten moito senso falar de Estatística como unha parte independente das Matemáticas, a Estatística está totalmente englobada na Probabilidade.

Estatística

Terminoloxía

- **Poboación:** Conxunto de individuos nos que se fai o estudo.
- **Mostra:** Parte da poboación na que se recollen os datos.
- **Variable estatística:** É a característica que se vai estudar. Utilizaremos o termo distribución para referirnos á totalidade dos valores dunha variable.
- **Frecuencia absoluta:** É o número de veces que se repite un certo valor ou conxunto de valores.
- **Frecuencia relativa:** Calcúlase dividindo a frecuencia absoluta polo número total de casos. Multiplicando por 100 transfórmase nunha porcentaxe.

As variables estatísticas podemos clasificalas en:

- **Cualitativas:** Cando esa característica non pode describirse numericamente (cor dos ollos, partido ó que se vota, etc).
- **Cuantitativas:** Cando a característica é de tipo numérico
 - ✓ **Discretas:** toman valores puntuais (número de fillos dunha familia, número de alumnos nun centro, etc.).
 - ✓ **Continuas:** toman calquera valor nun intervalo posible (altura das árbores dun bosque, cantidade de chuvia caída, temperatura, etc.).

Exemplo: Queremos estudar, mediante unha enquisa, a intención de voto nunhas eleccións xerais e relacionalo coa idade:

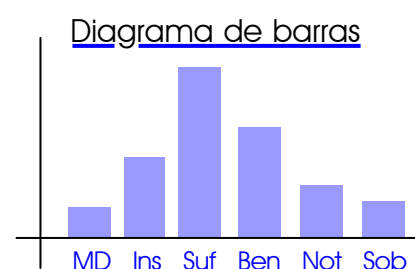
- **Poboación:** Españóis con dereito a voto.
- **Mostra:** 2000 persoas elixidas ó chou polas guías telefónicas.
- **Var. estatística:** Estudiamos dúas variables, partido ó que pensan votar (variable cualitativa) e idade (cuantitativa discreta).

Variables cualitativas

Para o manexo e representación de valores non numéricos empréganse sobre todo gráficos, táboas, porcentaxes...

Trátase de elixir os xeitos de representación axeitados para poder facerse unha imaxe da distribución deses valores.

Exemplo: As notas na avaliación de Matemáticas dos alumnos dun grupo foron: Ins, Suf, Suf, Not, MD, Suf, Ben, Sob, Ben, Ins, Ben, Suf,

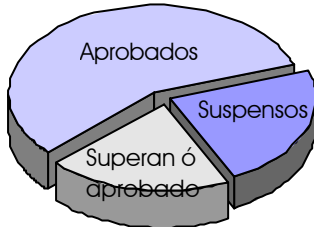


Ins, Suf, Not, Suf, Ben, Sob, Ben, Ins, Suf, Suf, Ins, Suf, Suf, Not, Not, MD, Suf, Ben, Sob, Ben, Ins, Suf, Ben.

Utiliza diferentes representacións para facilitar a valoración deses datos.

Solución: Empecemos por tabular os datos.

Diagrama de sectores



Valores da variable	MD	Ins	Suf	Ben	Not	Sob
Frecuencia absoluta	2	6	12	8	4	3
Frecuencia relativa	$\frac{2}{35}$	$\frac{6}{35}$	$\frac{12}{35}$	$\frac{8}{35}$	$\frac{4}{35}$	$\frac{3}{35}$

Un xeito de facilitala valoración dos datos e pasalos a porcentaxes:

MD	Ins	Suf	Ben	Notable	Sobre
5'7%	17'1%	35'3%	22'9%	11'4%	8'6%

Tamén pode ser útil simplificala distribución agrupando resultados:

Calificación baixa	Calificación media	Calificación alta
22'9%	57'1%	20%

Pódese utilizar diferentes tipos de gráficos para representar eses valores e facilitar a súa interpretación: Diagramas de barras, de sectores, pictogramas, etc.

Variables cuantitativas discretas.

Ademais das táboas e gráficos, ás variables cualitativas podemos asignarlle unha serie de parámetros que describan a distribución.

- **Medidas de tendencia central:** Son valores numéricos que intentan resumir toda a distribución. Os principais son a **media**, a **mediana** e a **moda**.
- **Medidas de dispersión:** Indican o espallados que están os valores da variable. Os máis utilizados son varianza, desviación típica, percorrido, período intercuartílico e coeficiente de variación.

Notación:

- Á variables por unha letra maiúscula, X ou Y,
- Os valores cunha letra minúscula cun subíndice (x_i ou y_i).
- As frecuencias absolutas por n_i
- Frecuencias relativas por f_i

1º alumno			2º alumno		
x_i	n_i	$x_i \cdot n_i$	y_i	n_i	$y_i \cdot n_i$
2'5	2	5	4'5	2	9
4	2	8	5'5	2	11
7	1	7	6	1	12
8	5	40	7	4	28
	10	60		10	60
$\bar{x} = \frac{60}{10} = 6$			$\bar{y} = \frac{54}{9} = 6$		

Exemplo: Dous alumnos obtiveron as seguintes puntuacións:

1º alumno: 8, 2'5, 4, 8, 8, 4, 7, 8, 2'5, 8.

2º alumno: 7, 7, 4'5, 5'5, 7, 4'5, 7, 5'5, 6.

a) Calcular as frecuencias de cada valor.

b) ¿Que nota se lle debe asignar a cada alumno?

Solución: b) A resposta semella evidente: **a nota media**.

$$\bar{x}_1 = \frac{2'5 + 2'5 + 4 + 4 + 7 + 8 + 8 + 8 + 8 + 8}{10} = 6$$

$$\bar{x}_2 = \frac{4'5 + 4'5 + 5'5 + 5'5 + 6 + 7 + 7 + 7 + 7}{9} = 6$$

Podemos dicir que a **media** é un valor que *resume* todo o conxunto de valores da variable estatística.

A media calcúlase máis facilmente agrupando os datos iguais, tal como aparece na táboa.

Medidas de tendencia central

As variables estatísticas abarcan multitude de datos, polo que resulta difícil facerse unha imaxe global.

Os parámetros de tendencia central son medidas que intentan resumir nun único valor tódolos valores da variable.

- **Media:** É a máis utilizada. $\bar{x} = \frac{\sum_{i=1}^n x_i n_i}{N} = \sum_{i=1}^n x_i \frac{n_i}{N} = \sum_{i=1}^n x_i fr_i$

Como pode verse pola súa fórmula, ten en conta os valores da variable e a frecuencia de cada valor.

- **Mediana:** Valor que está no centro cando se colocan os valores da variable por orden. Só ten en conta as frecuencias pero non os valores da variable.
- **Moda:** Valor que máis se repite. É moi doado de calcular pero non ten en conta ós demais valores da distribución nin as frecuencias.

Exemplo: Calcula a mediana e a moda das notas dos alumnos do exercicio anterior.

Solución: Para calcular as medianas, situamos por orden os valores das variables.

1º alumno: 2'5, 2'5, 4, 4, 7, 8, 8, 8, 8, 8.

2º alumno: 4'5, 4'5, 5'5, 5'5, 6, 7, 7, 7, 7.

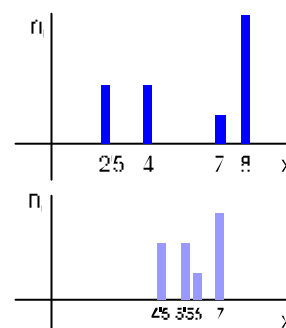
Ó ser par o número de datos do 1º alumno non hai un valor central.

Optouse, por convención, coller a media dos dous centrais: 7'5

A mediana do segundo é: 6

A diferenza entre a media e a mediana é unha medida da asimetría da distribución tal como se aprecia nos diagramas de barras.

As modas son 8 para o primeiro e 7 para o segundo.



Medidas de dispersión:

Miden o espallados que están os datos dunha distribución, o que permite comparar distintas distribucións.

Un xeito de facelo é mediante as diferenzas dos valores da variable estatística cun certo valor de referencia. O valor de referencia máis utilizado é a media.

Exemplo: A nota media dos estudantes dos exercicios anteriores é a mesma pero as notas son moi diferentes: As notas do 1ª son moi variables mentres que as do 2º son moito máis homoxéneas.

1º alumno			
x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x}) \cdot n_i$
2'5	2	-3'5	-7
4	2	-2	-4
7	1	1	1
8	5	2	10
	10		0

2º alumno			
y_i	n_i	$y_i - \bar{y}$	$(y_i - \bar{y}) \cdot n_i$
4'5	2	-1'5	-3
5'5	2	-0'5	-1
6	1	0	0
7	4	1	4
	9		0

¿Como podemos medir a variación das notas?

Solución: Podemos comparar as notas de cada un ca súa media.

Non podemos simplemente sumar as desviacións con respecto a media porque obtemos sempre 0 ó compensarse as desviacións por exceso coas desviacións por defecto.

Un xeito de evitalo é medir a dispersión coa *media do valor absoluto das desviacións*.

$$DM_X = \frac{|2'5 - 6| \cdot 2 + |4 - 6| \cdot 2 + |7 - 6| \cdot 1 + |8 - 6| \cdot 5}{10} = 2'2$$

$$DM_Y = \frac{|4'5 - 6| \cdot 2 + |5'5 - 6| \cdot 2 + |6 - 6| \cdot 1 + |7 - 6| \cdot 4}{9} = 0'89$$

Outro xeito é elevar ó cadrado e calcular a *media dos cadrados das desviacións*:

$$s_X^2 = \frac{(2'5 - 6)^2 \cdot 2 + (4 - 6)^2 \cdot 2 + (7 - 6)^2 \cdot 1 + (8 - 6)^2 \cdot 5}{10} = 5'35$$

$$s_Y^2 = \frac{(4'5 - 6)^2 \cdot 2 + (5'5 - 6)^2 \cdot 2 + (6 - 6)^2 \cdot 1 + (7 - 6)^2 \cdot 4}{9} = 1$$

Desviación media D_m : É a media dos valores absolutos das diferencias a media. Utilízase pouco:

$$D_m = \frac{\sum_i |x_i - \bar{x}| n_i}{N}$$

Varianza, s^2 : Media dos cadrados das diferencias a media.

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

Desviación típica, s : É a raíz cadrada da varianza (ó calcular a varianza elévase ó cadrado as desviacións, polo que as súas unidades non coinciden coas da variable, calcúlase a raíz cadrada para utilizar as mesmas).

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

A desviación típica é a medida de dispersión máis utilizada. Sen embargo non é axeitada para comparar variables con medias diferentes (por exemplo, non ten a mesma importancia unha desviación de 1 punto sobre 100 que 1 punto sobre 10). Nese caso é preferible o coeficiente de variación que relativiza as desviacións.

Coeficiente de variación de Pearson: É a razón entre a

desviación típica e a media: $\gamma = \frac{s}{\bar{x}}$ (desviación por unidade)

É unha cantidade sen dimensións, polo que o seu valor é independente da media.

Percorrido: Diferencia entre os valores extremos. É a máis fácil de calcular pero é pouco descritivo ó non ter en conta ó resto dos valores nin ás frecuencias.

Varianza e desviación típica

Hai outra fórmula para calcula-la varianza:

$$s^2 = \frac{\sum x_i^2 \cdot n_i}{N} - \bar{x}^2$$

Demostración:

$$s^2 = \frac{\sum (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \cdot n_i}{N} = \frac{\sum (x_i^2 n_i - 2x_i\bar{x} n_i + \bar{x}^2 n_i)}{N} =$$

$$\frac{\sum x_i^2 n_i - \sum 2x_i\bar{x} n_i + \sum \bar{x}^2 n_i}{N} = \frac{\sum x_i^2 n_i}{N} - \frac{\sum 2x_i\bar{x} n_i}{N} + \frac{\sum \bar{x}^2 n_i}{N}$$

No numerador da 2ª fracción tódolos sumandos teñen un factor $2\bar{x}$, e na 3ª tódolos sumandos teñen un factor $\bar{x}^2$, sacando factor común queda:

$$\frac{\sum x_i^2 n_i}{N} - 2\bar{x} \frac{\sum x_i n_i}{N} + \bar{x}^2 \frac{\sum n_i}{N} \quad \begin{matrix} \sum n_i = N \\ \downarrow \\ \bar{x} = \frac{\sum x_i n_i}{N} \end{matrix} = \frac{\sum x_i^2 n_i}{N} - 2\bar{x} \cdot \bar{x} + \bar{x}^2 \cdot 1 = \frac{\sum x_i^2 n_i}{N} - \bar{x}^2$$

Exemplo: As estaturas dun grupo de persoas son: 1'75, 1'59, 1'64, 1'83, 1'75, 1'66, 1'65, 1'72, 1'89, 1'56, 1'70, 1'68, 1'69, 1'91, 1'59, 1'81, 1'65, 1'68, 1'64, 1'65, 1'60, 1'62, 1'62

Calcular

- Media, mediana e moda.
- Percorrido, varianza e desviación típica.

Solución: Colocamos os datos por orden para calcular a mediana.

1'56, 1'59, 1'59, 1'60, 1'62, 1'62, 1'64, 1'64, 1'65, 1'65, 1'65, 1'66, 1'68, 1'68, 1'69, 1'70, 1'72, 1'75, 1'75, 1'81, 1'83, 1'89, 1'91.

Mediana: 1'66	Media: $\bar{x} = \frac{38'88}{23} = 1'69$	Moda: 1'65
Percorrido: 1'91-1'56=0'35	Varianza: $s^2 = \frac{65'92}{23} - 1'69^2 = 0'0085$	Desviación típica: $s = \sqrt{0'0085} = 0'092$

A estatura é unha variable continua pois entre dous valores dados pode tomar calquera valor intermedio, de aí que teñamos tantos valores diferentes.

Nestes casos, agrupar os datos iguais non aporta case ningunha vantaxe ó haber moitos diferentes. Para facilita-lo traballo podemos agrupalos por clases e asignar a cada grupo un valor intermedio:

Media: $\bar{x} = \frac{38'86}{23} = 1'69$	Varianza: $s^2 = \frac{65'84}{23} - 1'69^2 = 0'0080$	Desviación típica: $s = \sqrt{0'008} = 0'0896$
---	---	---

x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
1'56	1	1'56	2'43
1'59	2	3'18	5'05
1'60	1	1'6	2'56
1'62	2	3'24	5'25
1'64	2	3'28	5'38
1'65	3	4'95	8'17
1'66	1	1'66	2'76
1'68	2	3'36	5'64
1'69	1	1'69	2'86
1'70	1	1'7	2'89
1'72	1	1'72	2'96
1'75	2	3'5	6'13
1'81	1	1'81	3'28
1'83	1	1'83	3'35
1'89	1	1'89	3'57
1'91	1	1'91	3'65
	23	38'88	65'92

	x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[1'55,1'61)	1'58	4	6'32	9'99
[1'61,1'67)	1'64	8	13'12	21'52
[1'67,1'73)	1'70	5	8'5	14'45
[1'73,1'79)	1'76	2	3'52	6'2
[1'79,1'85)	1'82	2	3'64	6'63
[1'85,1'91]	1'88	2	3'76	7'07
		23	38'86	65'84

Variables cuantitativas continuas

Son variables estadísticas que poden tomar calquera valor nun intervalo, o que orixina unha gran cantidade de valores diferentes con frecuencias pequenas para cada un.

Para evitar ter que manexar esa gran cantidade de valores diferentes da variable, agrúpanse os datos en **clases** e asígnaselle a cada unha un número, normalmente o punto medio, que chamaremos **marca de clase**.

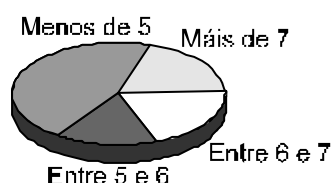
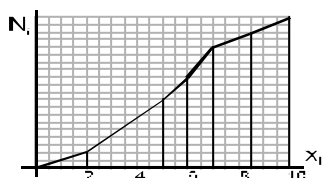
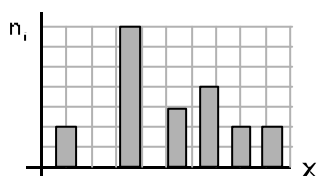
Dentro de cada clase suponse que os valores distribúense de xeito uniforme.

Dese xeito, transformamos unha variable continua noutra discreta que ten por valores as marcas de clase e como frecuencia de cada un o número de elementos da clase respectiva.

Neste proceso prodúcese unha perda de información pois os valores de cada clase non teñen porque distribuírse uniformemente (no exemplo anterior os valores no intervalo $[1'73, 1'79]$, do que tomamos como representante $1'76$, están todos por debaixo deste: $1'72$, $1'75$ e $1'75$).

A media, varianza e desviación típica dunha variable continua calcúlanse a partir das marcas de clase (polo xeral non se producirán diferencias importantes respecto ós cálculos feitos sen agrupar os valores como sucede no exemplo anterior).

	Datos agrupados	Datos sen agrupar
Media	1'689	1'69
D. típica	0'089	0'092



Gráficos.

Para facilita-la interpretación dos datos dunha variable estatística tamén se utilizan, ademais dos parámetros estatísticos, representacións gráficas. As máis usuais son:

Diagramas de barras: Gráficos feitos con barras da mesma base e coa altura proporcional as frecuencias (elixindo convenientemente as unidades no eixe vertical mesmo a altura pode coincidir coa frecuencia). Utilízanse para as variables cualitativas e para as discretas.

Polígonos de frecuencias: Poden ser de frecuencias absolutas ou relativas e tamén de frecuencias simples e acumuladas. Utilízanse fundamentalmente para variables discretas.

Diagramas de sectores: Dividindo un círculo en sectores proporcionais as frecuencias de cada valor.

Pictogramas: Debuxos alusivos ó tema de estudo e de tamaño proporcional a frecuencia de cada valor.

Histogramas: Empréganse cando os datos están agrupados en clases (variables estatísticas continuas).

É un diagrama de áreas. Nos histogramas, a **área** de cada barra é proporcional á frecuencia da clase a que corresponde.

As bases das barras son os intervalos das diferentes clases e as alturas calcúlanse simplemente dividindo a frecuencia de cada clase pola amplitude do intervalo:

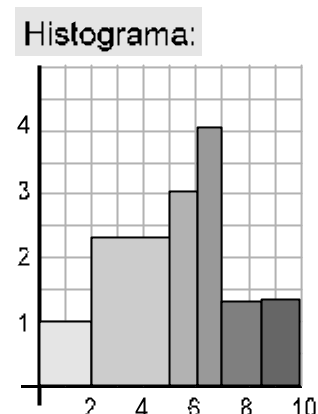
$$a_i = \frac{n_i}{x_{i+1} - x_i}$$

Comunmente tómanse intervalos coa mesma amplitude, polo que a altura da barra respectiva xa nos indica directamente a frecuencia.

A **moda** dunha variable continua é a clase máis alta no histograma e a **mediana** é o valor que divide ó histograma en dúas partes coa mesma área.

Propiedades dos histogramas:

- A área total das barras é N (1 nos histogramas de frecuencias relativas).
- A área do histograma sobre un intervalo corresponde á frecuencia (frecuencia relativa se é de frecuencias relativas) coa que a variable estatística toma valores no intervalo.
- A mediana é o punto que deixa a mesma área a cada lado.
- Defínese a moda, para datos agrupados, como a clase correspondente a barra máis alta do histograma.



Problemas coa media

Vimos que a media permite resumir nun só valor toda unha serie de datos. Desafortunadamente, en moitas situacións reais, as cousas non son tan sinxelas e a media non describe axeitadamente a situación. Nestas ocasións é convinte utilizar outras mediadas de tendencia central como a mediana.

Exemplo: No deserto do Sahara, as temperaturas durante o día acadan os 40°C, mentres que pola noite baixan ata os 0°C. Se calculámo-la media obtemos $(40+0)/2=20$ que é unha temperatura ideal, pero a ti non che gustaría vivir alí.

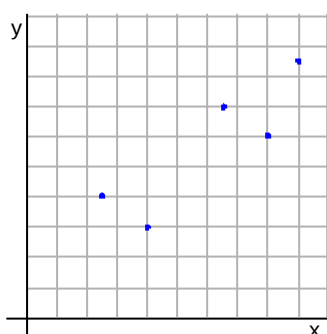
Exemplo: Nun pesqueiro traballan 4 mariñeiros, cun soldo de 50.000 pts mensuais e un patrón que cobra 1.000.000 de pts. O soldo medio é $(50.000 \cdot 4 + 1.000.000)/5 = 240.000$ pts, pero non parece que esta cantidade describa axeitadamente a situación.

Variables bidimensionais

En ocasións estudiamos sobre unha mesma poboación dúas variables diferentes (idade e preferencias políticas, estatura e peso, notas en dúas materias, consumo alcohólico e número de accidentes, etc).

Chamámoslle variable **estatística bidimensional** a distribución conxunta das dúas variables.

Representando graficamente os valores da distribución bidimensional obtemos a **nube de puntos** da distribución ou **diagrama de dispersión**.



Exemplo: As calificacións en dous exames de Matemáticas de 5 alumnos foron:

1º exame	2.5	4	6'5	8	9
2º exame	4	3	7	6	8'5

Representa a nube de puntos e calcula as medias e desviacións típicas desas variables (fíxate que a frecuencia de cada resultado é 1).

Solución: Poñemos os datos nunha táboa para facilita-la súa manipulación:

x_i	y_i	x_i^2	y_i^2
2.5	4	6'25	16
4	3	16	9
6'5	7	42'25	49
8	6	64	36
9	8'5	81	72'25
30	28'5	209'5	182'25

$$\bar{x} = \frac{\sum_{i=1}^5 x_i}{5} = \frac{30}{5} = 6 \quad \bar{y} = \frac{\sum_{i=1}^5 y_i}{5} = \frac{28'5}{5} = 5'7$$

$$s_x^2 = \frac{\sum_{i=1}^5 x_i^2}{5} - \bar{x}^2 = \frac{209'5}{5} - 6^2 = 5'9 \rightarrow s_x = \sqrt{5'9} = 2'43$$

$$s_y^2 = \frac{\sum_{i=1}^5 y_i^2}{5} - \bar{y}^2 = \frac{182'25}{5} - 5'7^2 = 3'96 \rightarrow s_y = \sqrt{3'96} = 1'99$$

Regresión e correlación

Nunha variable estatística cualitativa bidimensional, ademais do estudio de cada un dos caracteres por separado, podemos estar interesados en coñecer se existe algunha relación entre os valores dos dous caracteres (notas en dous exames, estatura e peso, consumo alcohólico e número de accidentes, etc.).

Para estudar a relación entre as variables necesitamos “medir” a variación conxunta das mesmas. Esa medida pódese obter por medio da **covarianza**:

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N}$$

Fíxate que se un individuo ten os datos altos en relación ás medias, aportará para a covarianza (+ + = +) e tamén

se os seus datos son baixos ($-- = +$), pero se os datos non se corresponden entre si perxudicará á covarianza ($+ - = -$ $- + = -$).

De xeito semellante á varianza, dispoñemos dunha segunda fórmula para o cálculo da covarianza:

$$s_{xy} = \frac{\sum_{i=1}^n x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}$$

Correlación lineal

É o estudio do grao de dependencia lineal entre os caracteres dunha variable bidimensional.

Para medir ese grao de dependencia utilízase o coeficiente de **correlación lineal**:

$$\rho = \frac{s_{xy}}{s_x \cdot s_y}$$

O coeficiente de correlación lineal sempre comprendido entre -1 e 1 .

Poden darse diferentes graos de dependencia lineal entre os caracteres dunha variable bidimensional.

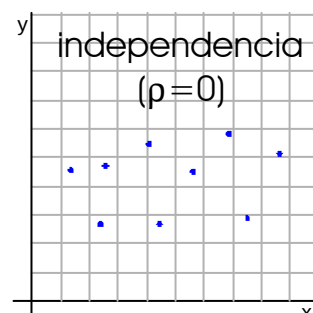
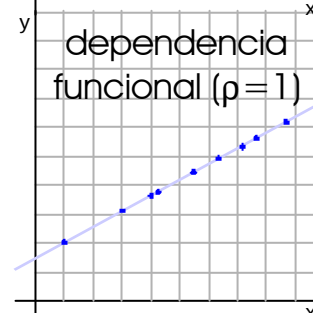
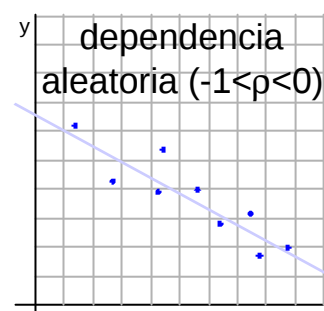
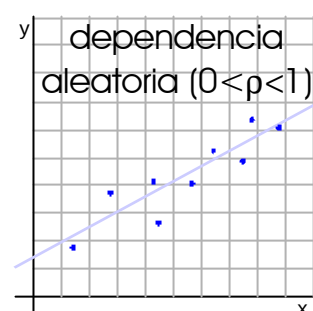
- Dependencia aleatoria** ou **probabilística**: A nube de puntos aproxímase a unha recta. O coeficiente de correlación ten un valor entre 0 e 1 (ou entre -1 e 0), canto máis próximo a 0 menor grao de dependencia.
- Dependencia funcional**: É un caso extremo cando hai unha dependencia total entre as variables.
A nube de puntos está sobre unha recta. O coeficiente de correlación é 1 (recta crecente) ou -1 (decrecente).
- Independencia**: Cando non hai relación lineal entre as variables (pode haber unha relación doutro tipo).
A nube de puntos non se asemella a unha recta. O coeficiente de correlación é 0 .

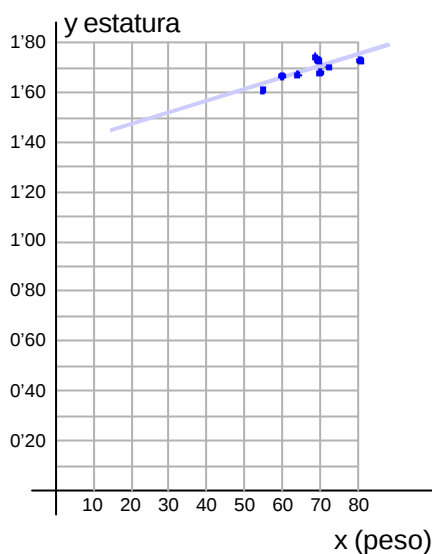
Exemplo: O peso e a estatura dun grupo de persoas son:

Peso (en kg)	55	60	64	68	70	70	72	81
Estatura (en cm)	162	166	166	175	168	174	170	173

Representar a nube de puntos e estudar o grao de dependencia lineal entre as variables peso e estatura.

Solución: O grao de dependencia ven dado polo coeficiente de correlación lineal, polo que debemos calcular as medias, desviacións típicas e covarianza das variables:





$$\bar{x} = \frac{\sum_{i=1}^8 x_i}{8} = \frac{540}{8} = 67'5 \text{ kg} \quad s_x = \sqrt{\frac{\sum_{i=1}^8 x_i^2}{8} - \bar{x}^2} = \sqrt{\frac{36890}{8} - 67'5^2} = 7'4 \text{ kg}$$

$$\bar{y} = \frac{\sum_{i=1}^8 y_i}{8} = \frac{1354}{8} = 169'25 \text{ m} \quad s_y = \sqrt{\frac{\sum_{i=1}^8 y_i^2}{8} - \bar{y}^2} = \sqrt{\frac{229310}{8} - 169'25^2} = 4'26 \text{ m}$$

$$s_{xy} = \frac{\sum_{i=1}^8 x_i \cdot y_i}{8} - \bar{x} \cdot \bar{y} = \frac{91587}{8} - 67'5 \cdot 169'25 = 24$$

O coeficiente de correlación lineal será: $\rho = \frac{24}{7'4 \cdot 4'26} = 0'76$

0'76 é un valor entre 0 e 1, o que indica un certo grao de relación lineal positiva (crecente) entre as variables. Hai tendencia a que as persoas de maior peso sexan máis altas e as de menor peso sexan máis baixas pero non é radicalmente así e son bastante frecuentes os casos en que non ocorre.

Regresión lineal

Se unha nube de puntos se poidese aproximar por medio dunha recta, esta facilitaríanos enormemente o seu estudio (aínda que perderíamos algunha exactitude).

Chámase regresión lineal á búsqueda da fórmula lineal ($y=ax+b$) que mellor se axuste a nube de puntos dunha distribución bidimensional.

Segundo cal sexa o criterio que utilizamos para medir o axuste dos puntos á recta, obteremos diferentes ecuacións para a recta de regresión (pode haber varias rectas “parecidas” que nos resulten de utilidade).

O criterio máis estendido é o chamado *de mínimos cadrados* que consiste en medir o axuste polas diferencias de ordenadas (ou abscisas) dos puntos da nube á recta.

Hai dúas rectas de mínimos cadrados (a correspondente as diferencias de ordenadas e a das diferencias de abscisas):

$$\text{Recta de Y sobre X: } y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x})$$

$$\text{Recta de X sobre Y: } x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y})$$

Exemplo: O volume dunha certa cantidade de ar a diferentes temperaturas foi:

Temperatura (en °C)	15'4	20'1	24	28	32'6	36
Volume (en cm ³)	15'5	15'7	16	16'4	16'9	17'2

- Estudia a correlación lineal entre as variables temperatura e volume.
- ¿Cal será o volume a 40°C?
- ¿A que temperatura se conxela o aire (volume practicamente 0)?

Solución:

$$\bar{x} = \frac{\sum_{i=1}^6 x_i}{6} = \frac{1561}{6} = 260'16^\circ\text{C} \quad s_x = \sqrt{\frac{\sum_{i=1}^6 x_i^2}{6} - \bar{x}^2} = \sqrt{\frac{435993}{6} - 260'16^2} = 7'056^\circ\text{C}$$

$$\bar{y} = \frac{\sum_{i=1}^6 y_i}{6} = \frac{977}{6} = 16'283\text{cm}^3 \quad s_y = \sqrt{\frac{\sum_{i=1}^6 y_i^2}{6} - \bar{y}^2} = \sqrt{\frac{159315}{6} - 16'283^2} = 0'615\text{cm}^3$$

$$s_{xy} = \frac{\sum_{i=1}^6 x_i \cdot y_i}{6} - \bar{x} \cdot \bar{y} = \frac{256761}{6} - 260'16 \cdot 16'283 = 4'316$$

O coeficiente de correlación lineal será: $\rho = \frac{4'316}{7'056 \cdot 0'615} = 0'995$

É un valor moi próximo a 1, o que indica un forte grao de relación lineal entre as variables.

Para *estimar* o volume a unha certa temperatura utilizamos a recta de regresión de Y sobre X (volume en función da temperatura):

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x}) \rightarrow y - 16'283 = \frac{4'316}{7'056^2} (x - 26'016) \rightarrow y = 0'087x + 14'028$$

Para $x=40$: $y = 0'087 \cdot 40 + 14'028 = 17'508 \text{ cm}^3$

Para *estimar* a temperatura a partir dun volume utilizamos a recta de regresión de X sobre Y:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y}) \rightarrow x - 26'016 = \frac{4'316}{0'615^2} (y - 16'283) \rightarrow x = 11'411y - 159'793$$

Para $y \approx 0$: $x = 11'411 \cdot 0 - 159'793 = -159'793^\circ\text{C}$

Estatística

Xurde no século XVII a raíz de estudos sobre a poboación humana.

O seu obxecto é reunir, manexar e interpretar grandes cantidades de datos referidos a un certo fenómeno.

- **Poboación:** Conxunto de individuos nos que se fai o estudo.
- **Mostra:** Parte da poboación na que se recollen os datos.
- **Variable estatística:** Característica que se vai estudar.
 - **Cualitativas:** Cando a característica non pode describirse numericamente.
 - **Cuantitativas:** Cando a característica é de tipo numérico.
 - ✓ **Discretas:** Valores puntuais.
 - ✓ **Continuas:** Calquera valor nun intervalo. Transfórmanse en discretas agrupando os datos en clases e asignándolle a cada clase un valor, marca de clase.
- **Frecuencia absoluta n_i :** Veces que se repite un valor ou conxunto de valores.
- **Frecuencia relativa f_i :** Relación entre a frecuencia absoluta e o número total de casos. Multiplicando por 100 transfórmase nunha porcentaxe.
- **Medidas de tendencia central:** Valores que intentan resumir toda a distribución.
 - **Media:**

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

- **Mediana:** Valor central da distribución
 - **Moda:** Valor ou valores que máis se repiten.
- **Medidas de dispersión:** Miden o espallados que están os valores da variable.

- **Desviación media D_m :** $D_m = \frac{\sum_i |x_i - \bar{x}| n_i}{N}$

- **Varianza, s^2 :** $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 n_i}{N}$ ou $s^2 = \frac{\sum_i x_i^2 \cdot n_i}{N} - \bar{x}^2$

- **Desviación típica, s :** Raíz cadrada da varianza $s = \sqrt{s^2}$

- **Coeficiente de variación de Pearson:** $\gamma = \frac{s}{\bar{x}}$

- **Percorrido:** Diferencia entre os valores extremos.

Gráficos.

- **Diagramas de barras:** Barras ca mesma base e coa altura proporcional as frecuencias.
- **Polígonos de frecuencias:** Utilízanse fundamentalmente para variables discretas.
- **Diagramas de sectores:** Sectores de ángulos proporcionais as frecuencias.

- **Pictogramas:** Debuxos de tamaño proporcional a frecuencia de cada valor.
- **Histogramas:** Só cando os datos están agrupados en clases (variables continuas).
A **área** de cada barra é proporcional á frecuencia da clase a que corresponde.
A moda e a mediana das variables continuas calcúlanse no histograma.

Variables bidimensionais

- **Diagrama de dispersión:** Representación conxunta das variables.

- **Covarianza:**
$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N} = \frac{\sum_{i=1}^n x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}$$

- **Correlación lineal:** Estudio do grao de dependencia lineal entre os caracteres dunha variable bidimensional.

- ✓ **Coeficiente de correlación lineal:** Mide o grao de correlación
$$\rho = \frac{s_{xy}}{s_x \cdot s_y}$$

- ✓ **Independencia:** Non hai relación lineal entre as variables. A nube de puntos non se asemella a unha recta. O coeficiente de correlación é 0.
- ✓ **Dependencia funcional:** Unha variable é función da outra. A nube de puntos está sobre unha recta. O coeficiente de correlación é 1 (recta crecente) ou -1 (decrecente).
- ✓ **Dependencia aleatoria ou probabilística:** Situación intermedia entre as anteriores. A nube de puntos aproxímase a unha recta. O coeficiente de correlación ten un valor entre 0 e 1 (ou entre -1 e 0). Canto máis próximo a 0 menor grao de dependencia.
- **Regresión lineal:** Búsqueda da fórmula lineal ($y=ax+b$) que mellor se axuste a nube de puntos dunha distribución bidimensional.

Rectas de regresión lineal de mínimos cadrados:

$$\text{Recta de Y sobre X: } y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x})$$

$$\text{Recta de X sobre Y: } x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y})$$

Ampliación

Relacións funcionais

Xa comentamos en moitas ocasións que a ciencia occidental basea a súa metodoloxía na elaboración de modelos matemáticos e na capacidade de predicción a partir deses modelos.

Nos fenómenos nos que interveñen dúas magnitudes o modelo matemático adoita ser unha fórmula que relaciona ámbalas dúas magnitudes.

Determinar cal é a fórmula axeitada para describir a relación é un problema complexo: Cando se efectúa un experimento é inevitable cometer pequenos erros na medidas que fan que os valores obtidos non correspondan exactamente a unha fórmula, ¿como podemos estar seguros que as diferencias entre os valores obtidos e os teóricos débense os erros experimentais e non a que a fórmula sexa diferente?

Un xeito de contestar a esa pregunta, cando a dependencia é lineal, é mediante a correlación e a regresión lineais, pero ¿que sucede cando a dependencia non é lineal?

Modelos exponenciais

As funcións do tipo $y=a \cdot b^x$ son as seguintes en importancia, despois das lineais, para a descrición de fenómenos físicos. Podemos utilizar o modelo de regresión lineal para estudar a dependencia exponencial con só tomar logaritmos:

$$y = a \cdot b^x \xrightarrow{\text{tomando logaritmos}} \ln(y) = \ln(a \cdot b^x) \rightarrow \ln(y) = \ln(a) + x \cdot \ln(b)$$

Vemos que hai unha dependencia lineal entre x e $\ln(y)$.

Exemplo: A poboación española desde 1960 ata 1996 foi:

Ano	X	1960	1970	1975	1981	1986	1991	1996
Poboación	Y	30.903.137	34.041.531	36.012.682	37.682.355	38.473.418	38.872.268	39.669.394

¿Podemos asegurar que é exponencial?

Solución: Calculamos o logaritmo neperiano de Y:

X	1960	1970	1975	1981	1986	1991	1996
$\ln(Y)$	17'246	17'343	17'399	17'445	17'465	17'476	17'496

O coeficiente de correlación lineal é 0'969, o que indica unha forte relación lineal entre os valores de X e $\ln(Y)$. A recta de regresión de $\ln(Y)$ sobre X é: $y=0'006885x+3'77958$.

A función exponencial correspondente a Y sobre X será:

$$y = e^{3'77958} \cdot (e^{0'006885})^x = 43'7979 \cdot 1'0069^x \text{ (indica un aumento do 0'69\% anual)}$$

Problema: A xulgar polos datos anteriores, ¿coidas que se pode afirmar un comportamento diferente nos períodos 1960-1975 e 1986-1996?

Modelos potenciais

Un método semellante permite abordar as dependencias de tipo potencial, $y=a \cdot x^b$ (fíxate que agora a variable x non está no expoñente senón na base)

Problema: Estudia a relación entre os seguintes datos correspondentes á caída libre dun obxecto:

Tempo (en s)	0'145	0'178	0'21	0'237	0'264	0'288
Espacio (en m)	0'1026	0'1576	0'2139	0'2771	0'3402	0'4075

Nota: As dependencias funcionais tamén poden abordarse sen partir do modelo lineal

Exercicios e actividades

- 1 Nunha enquisa pregúntaselle a 40 alumnos dun instituto cantos irmáns son na súa familia. As respostas foron as seguintes: 1, 2, 2, 3, 1, 2, 1, 1, 4, 1, 3, 1, 5, 2, 2, 1, 2, 4, 1, 2, 1, 1, 3, 2, 1, 5, 1, 4, 2, 2, 1, 4, 1, 3, 3, 1, 2, 3, 1, 2.

- a) Calcula os parámetros da distribución: media, mediana, moda, varianza, desviación típica e coeficiente de variación de Pearson.
b) Debuxa o diagrama de barras e o polígono de frecuencias acumuladas.

Estatuta	nº
< 1.54	29
[1.54,1.58)	49
[1.58,1.62)	74
[1.62,1.66)	91
[1.66,1.70)	92
[1.70,1.74)	76
[1.74,1.78)	42
≥ 1.78	30

- 2 Queremos estudar as alturas dos estudantes de C.O.U. da comarca do Salnés. Témo-los datos que aparecen na táboa:

- a) Calcula os parámetros da distribución.
b) Debuxa o histograma e o polígono de frecuencias acumuladas.

[1.62,1.66)	91
[1.66,1.70)	92
[1.70,1.74)	76
[1.74,1.78)	42

- 3 Nun taller traballan 14 operarios cun soldo de 120.000 pts. mensuais, 8 oficiais que cobran 180.000 pts. ó mes e un xefe de taller con soldo de 4.500.000 pts.

- a) Calcula as medidas de tendencia central da distribución.
b) ¿Cal deses parámetros describe máis axeitadamente a situación? ¿Por que?
c) Representa os valores nun diagrama de sectores.

Ingrsos mensuais (en miles)	n_i
50 ou menos	35
(50,100]	104
(100,150]	215
(150,175]	185
(175,200]	120
(200,250]	85
Máis de 250	40

- 4 Para estudar o poder adquisitivo das familias españolas, fíxose unha enquisa na que se obtiveron os datos da táboa.

- a) Calcula os parámetros da distribución.
b) Debuxa o histograma.

- 5 Queremos medir con precisión o tempo empregado por un obxecto en caída libre en percorrer 5 m para o que realizamos varias medicións obtendo os seguintes resultados: 0'98, 1'4, 1'03, 0'99, 0'6

- a) ¿Que tempo emprega en facer o percorrido?
b) ¿Como podemos mellorar o valor e cal é o valor corrixido?
c) Calcula a desviación típica deses valores e explica o seu significado.
d) ¿É importante que os datos sexan simétricos en relación a media? Xustifica a resposta.

- 6 Un profesor que imparte clase en dous grupos diferentes, un con 40 alumnos e outro con 25, puxo as notas que aparecen no cadro.

- a) Calcula os parámetros das dúas distribucións e compara os seus valores.
b) Debuxa os histogramas de frecuencias relativas das distribucións e compara as notas de cada grupo.

Notas	Ins	Suf	Ben	Not	Sob
Grupo A	15	8	10	5	2
Grupo B	7	6	6	4	2

- 7 Un alumno decide comprobar se hai algunha relación entre as notas nos exames dunha certa materia e o número de horas de estudo, para o que realiza unha serie de probas cos seguintes resultados:

Horas de estudio	0	4	5	7	12
Notas no exame	2	4'5	6	7	9

- a) Estudia o grao de correlación lineal entre as horas de estudo e a nota nos exames.
b) ¿Que nota podemos supoñer que obterá nun exame para o que estudou 8 horas?
c) ¿Cantas horas se supón que debe estudar para obter un 5?

- 8 A altura e o peso ó nacer de seis nenos foron os seguintes:

Peso (grms)	3200	3700	2800	3000	2800	3500
Altura (cm)	52	54	50	49	51	55

- Calcula a ecuación da recta de regresión da altura (Y) sobre o peso (X) e represéntaa no diagrama de dispersión.
- Calcula coeficiente de correlación e interprétao.
- ¿Que altura podemos supoñer que terá un neno de 3800 grms?

- 9 As notas de 10 alumnos dun grupo de 1º de bacharelato en Matemáticas e Lingua foron:

Matemáticas	2	4	4	5'5	6	6	6	7'5	8	10
Lingua	3	3	5	6	5	7	5	7'5	10	8

- Estudia o grao de correlación lineal entre as notas en ámbalas dúas materias.
- Se un alumno ten un 7 en Matemáticas ¿que nota pode esperar en Lingua?
- ¿Que podemos dicir do xeito de calificar do profesor de cada materia?
- Estudia se os alumnos teñen un comportamento similar nas dúas materias salvando as diferencias debidas o xeito de puntuar de cada profesor tipificando as notas (a nota de cada materia réstaslle a media e divídela pola desviación típica) e comparando os histogramas das novas variables obtidas ó tipificar.

Actividade: As seguintes táboas recollen as notas medias do expediente e as notas de selectividade dos alumnos dun instituto galego (son datos reais pero seguramente ti preferirás facer o estudio coas notas do teu propio centro)

- Calcula a nota media e o coeficiente de variación das notas medias do expediente e das notas nos exames de selectividade de homes e mulleres e compara e interpreta os resultados.

		Notas examen selectividade				
Homes		[2,4)	[4,5)	[5,6)	[6,7)	[7,8)
Nota Media	[5,6)	2	2	-	-	-
	[6,7)	5	7	4	-	1
	[7,8)	-	1	5	2	-
	[8,9]	-	1	-	3	-

		Notas examen selectividade				
Mulleres		[2,4)	[4,5)	[5,6)	[6,7)	[7,8)
Nota Media	[5,6)	-	2	-	-	-
	[6,7)	10	11	4	1	-
	[7,8)	3	9	5	3	-
	[8,9]	-	3	1	2	2

- Estudia a regresión e correlación das notas de selectividade (Y) en relación ás notas dos expedientes (X) tanto para homes como para mulleres.
- Calcula a nota que podes esperar en selectividade a partir da nota do teu expediente.

Problema: Un conductor quere cambiar de coche. Está indeciso entre dous modelos de prezo semellante e decidiu optar polo de menor consumo. O modelo A ten un consumo, cada 100 km, de 8 litros en estrada e 12 litros en cidade, o modelo B consume 7 litros en estrada e 14 litros en cidade. ¿Que modelo lle convén se, de promedio, o conductor fai 200 km. de estrada por cada 150 en cidade?