Técnicas de Investigación Social

Grado en Relacións Laborais e Recursos Humanos

Curso 2019-2020

Tema 0. Nocións básicas en Estatística

Alejandro Saavedra Nieves

Introdución

A solución de problemas da vida real responden a razoamentos de tipo "indutivo". Polo xeral, preténdese estender a un todo as conclusións obtidas nunha parte. Así, fanse continuamente afirmacións sobre un grupo de individuos habendo observado na realidade só unha parte (a veces pequena) deles. O método científico axeitado para validar tales xeralizacións é o método estadístico.

- Un psicólogo poderá caracterizar os seus pacientes, diagnosticar síndromes, recomendar tratamentos...
- Un sociólogo será capaz de orientar ós políticos, interpretar estados sociais, adiantarse ás crises...
- Un médico coñecerá os riscos de determinados medicamentos, das particularidades de certos pacientes en relación con algúns fármacos...
- Un economista axudará á empresa a previr problemas, a identificar riscos, a deseñar campañas atendendo ós perfís dos clientes...

Introdución

A repetición dos experimentos en condicións idénticas dá lugar a resultados distintos, debido a factores ou causas como poden ser os erros pola manipulación do experimentador ou polo aparato de medida; pero además temos a "variabilidade" dos individuos obxecto de estudo: dos seres vivos nunca son iguais, nin un individuo é igual a sí mesmo en diferentes etapas da súa vida.

Temos entonces fenómenos que son esencialmente impredecibles nos seus resultados, e as afirmacións acerca deles só poden facerse en termos de probabilidade ou posibilidade *(experimentos aleatorios)*. O modo de obter resultados científicos válidos a partir de datos que son fundamentalmente impredecibles é a través das técnicas estatísticas, pois son capaces de ter en conta a variabilidade aludida.

Introdución

En toda investigación experimental poden distinguirse tres etapas:

1. Deseño
2. Recopilación de datos
3. Análise dos resultados e obtención das conclusións

* Nas tres etapas, a Estatística é fundamental.

Ejemplo: O servizo municipal de transportes, co propósito de mellorar o seu funcionamiento, desexa coñecer o tempo de espera dos seus usuarios antes de subirse a un autobús. ¿Cómo debe proceder?

Conceptos básicos

Poboación: Conxunto de obxectos, persoas, entidades da máis diversa índole, que constitúen o obxectivo do noso estudo. É o universo de individuos ó cal se refire o estudo que se pretende realizar.

Variable: Rasgo ou característica dos elementos da poboación que se pretende analizar.

Mostra: Subconxunto da poboación onde os valores da variable que se pretende analizar son coñecidos. No noso contexto, no procedimento de extracción vai intervir o azar. Polo tanto, a mostra consistirá nun conxunto de realizacións dun experimento aleatorio.

Tamaño mostral: Número de individuos que compoñen a mostra. Representámolo por n.

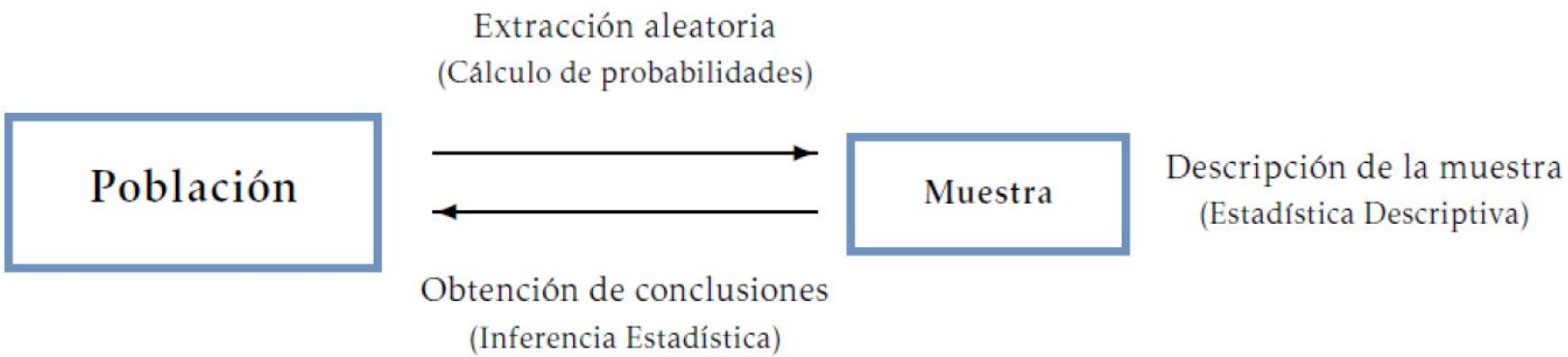
Estatística

Clasificamos as tareas vinculadas á Estatística en tres grandes disciplinas:

Estatística Descritiva. Ocúpase de recoller, clasificar e resumir a información contida na mostra.

Cálculo de Probabilidades. É unha parte da matemática teórica que estuda as leis que rexen os mecanismos aleatorios.

Inferencia Estatística. Pretende extraer conclusións para a poboación a partir do resultado observado na mostra.



A Inferencia Estatística ten un obxectivo máis ambicioso que o da mera descrición da mostra (Estatística Descritiva). Dado que a mostra se obtén mediante procedementos aleatorios, o Cálculo de Probabilidades é unha ferramenta esencial da Inferencia Estatística.

Tipos de Variables

Variables cualitativas: Non aparecen en forma numérica, senón como categorías ou atributos.

- o sexo
- cor de ollos
- nivel de estudos
- deporte favorito
- ...

Variables cuantitativas: Toman valores numéricos porque son frecuentemente o resultado dunha medición.

- idade (m) dunha persoa
- o peso (kg.) dunha persoa
- número de fillos
- número de empregados en empresas
- ...

Tipos de Variables. Variables cualitativas

A variables cualitativas, también chamadas *atributos ou variables categóricas* poden clasificarse á súa vez en:

- Cualitativas nominales: Miden características que non toman valores numéricos (sen orde).
  - o sexo (home ou muller)
  - creencias relixiosas
  - cor de ollos
  - ...
- Cualitativas ordinais: Os seus posibles valores admiten unha relación de orde.
  - máximo curso no que se está matriculado
  - categoría hoteleira
  - hábitos de consumo de tabaco
  - ...

É moi común asignar códigos numéricos ás categorías dos datos cualitativos. Esto non os converte en datos cuantitativos: esos códigos numéricos son meros símbolos representando ás categorías.

Tipos de Variables. Variables cuantitativas

Clasifícanse á súa vez en:

- Cuantitativas discretas: Toman un número discreto de valores (no conxunto de números naturales). Os seus posibles valores están separados entre sí.
  - número de multas nun ano
  - número de fillos
  - número de pasaxeiros en voos nacionais
  -
- Cuantitativas continuas: Toman valores numéricos dentro dun intervalo real.
  - o peso
  - a idade
  - nivel de glucosa en sangue
  - salario bruto anual
  - ...

Exemplo

O servicio médico dunha empresa recibe a visita de oito dos seus empregados con dor lumbar durante unha semana. Tódolos datos quedan recollidos na seguinte táboa. Clasifica as variables recollidas (sexo, peso, estatura, temperatura, número de visitas previas ó servicio e dor).

Sexo	Peso (kg.)	Estatura (m.)	Temperatura (°C)	Visitas	Dor
M	63	1.74	38	0	Leve
M	58	1.63	36.5	2	Intenso
H	84	1.86	37.2	0	Intenso
M	47	1.53	38.3	0	Moderado
M	70	1.75	37.1	1	Intenso
M	57	1.68	36.8	0	Leve
H	87	1.82	38.4	1	Leve
M	55	1.46	36.6	1	Intenso

Descrición de variables cualitativas e cuantitativas discretas

Supoñamos que os n valores que pode tomar unha variable X son: $x_1, x_2, \ldots, x_m$.

Frecuencia absoluta: Denótase por n_i e representa o número de veces que ocorre o resultado x_i.

Frecuencia relativa: Denótase por f_i e representa a proporción de datos en cada unha das clases,

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

Frecuencia absoluta acumulada. É o número de veces que se observou o resultado x_i ou valores anteriores. Denotámola por

$$N_i = n_1 + n_2 + n_3 + \cdots + n_i = \sum_{x_j \leq x_i} n_j$$

Frecuencia relativa acumulada. É a frecuencia absoluta acumulada dividida polo tamaño mostral. Denotámola por

$$F_i = \frac{N_i}{n} = f_1 + f_2 + f_3 + \cdots + f_i = \sum_{x_j \leq x_i} f_j$$

As frecuencias pódense escribir ordeadamente mediante unha **táboa de frecuencias**, que adopta esta forma:

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	n_m	f_m	N_m	F_m

Propiedades:

Frecuencias absolutas	$0 \leq n_i \leq n$	$\sum_{i=1}^{m} n_i = n$
Frecuencias relativas	$0 \leq f_i \leq 1$	$\sum_{i=1}^{m} f_i = 1$
Frecuencias absolutas acumuladas	$0 \leq N_i \leq n$	$N_m = n$
Frecuencias relativas acumuladas	$0 \leq F_i \leq 1$	$F_m = 1$

Non calculamos as frecuencias acumuladas se a variable é cualitativa nominal.

- **Variables cualitativas**:
  - Diagrama de barras.
    Consiste en erguer sobre cada valor ou modalidade da variable unha barra (segmento de recta ou rectángulo) de altura igual ou proporcional á correspondente frecuencia absoluta (n_i) ou relativa (f_i).
  - Polígono de frecuencias.
    Óbtense unindo mediante segmentos de recta os puntos (x_i, n_i) o (x_i, f_i) para todo $i = 1, \ldots, k$.
  - Gráfico de sectores.
    Divídise un círculo en sectores circulares, un por cada valor ou modalidade da variable, de forma que o ángulo de cada sector sexa $\alpha_i = 360 \times f_i$ grados.

Exemplo. Variable cualitativa nominal. Procedencia

Nunha mostra de 50 turistas de Vigo estúdase a súa "procedencia" e obtense a información detallada a continuación:

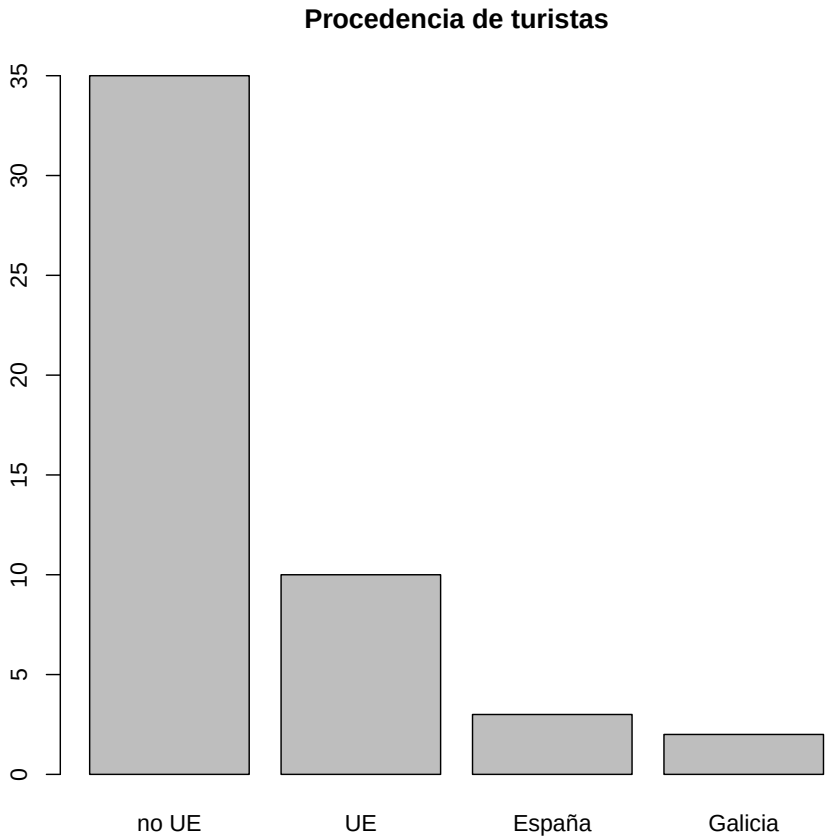
- 35 persoas residen fóra da UE,
- 10 persoas residen na UE (non en España),
- 3 persoas son españolas e
- 2 persoas son galegas.

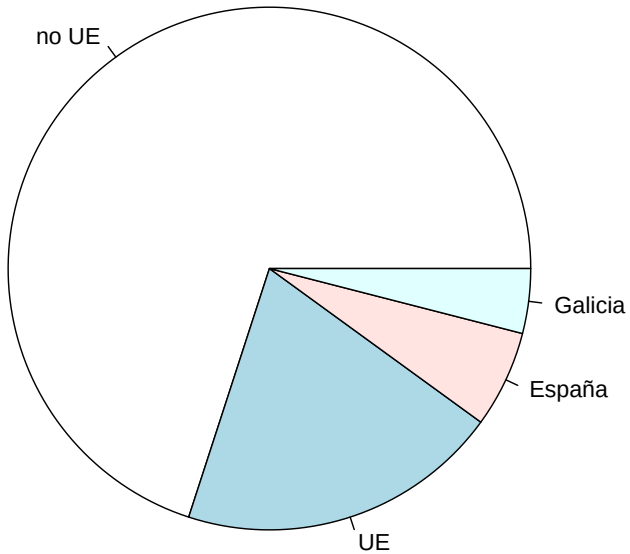
Tamaño mostral: $n = 50$

Procedencia (x_i)	n_i	f_i
Fóra da UE	35	0'7
UE (non España)	10	0'2
España	3	0'06
Galicia	2	0'04

Nótese que non calculamos as frecuencias acumuladas xa que a variable Procedencia é nominal.

Representacións gráficas. Variable cualitativa nominal. Procedencia



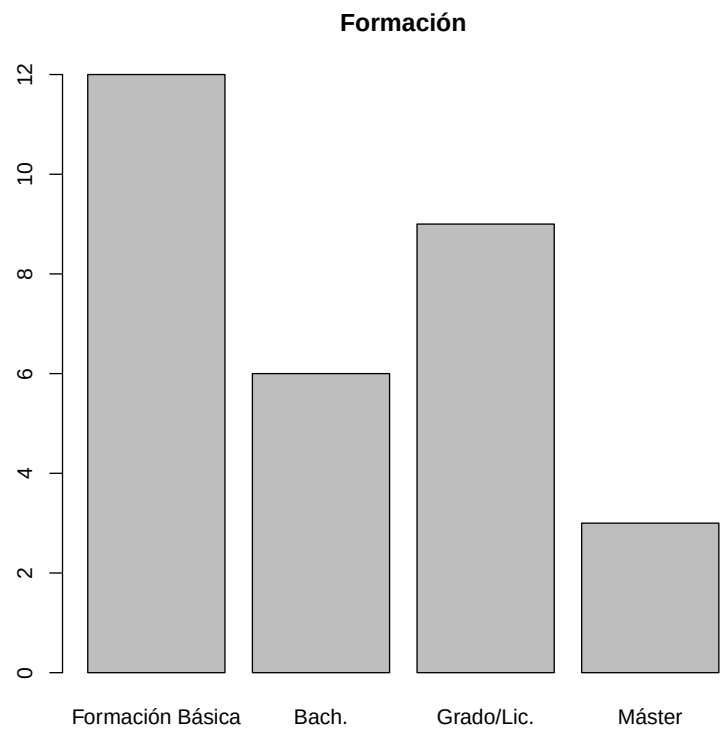


Exemplo. Variable cualitativa ordinal. Nivel máximo de Estudos

Nunha mostra de 30 candidatos a unha oferta de emprego analízase o "máximo nivel de estudos", e obtivéronse os seguintes resultados:

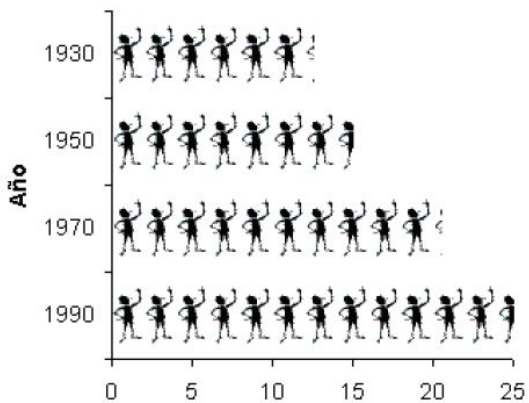
- 12 persoas teñen formación básica,
- 6 persoas acabaron bacharelato,
- 9 persoas estudiaron unha carreira e
- 3 persoas realizaron un máster.

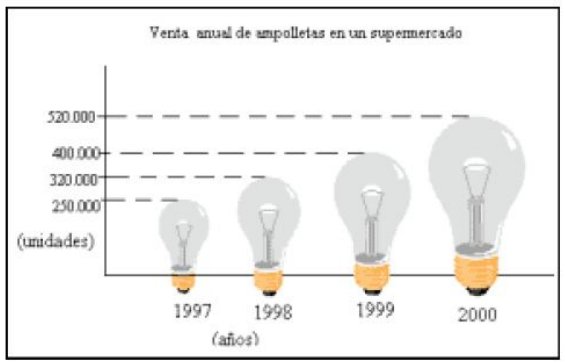
Nivel máximo de estudos	n_i	f_i	N_i	F_i
Formación Básica	12	0'4	12	0'4
Bacharelato	6	0'2	18	0'6
Grado/Licenciatura	9	0'3	27	0'9
Máster	3	0'1	30	1



Representacións gráficas. Variables cuantitativas

- **Variables cuantitativas discretas**: diagrama de barras ou tamén o de sectores (cando os valores que toma X son poucos)
- **Variables cuantitativas continuas agrupadas**: histograma, diagrama de tallos e follas.
- Otros:
  - Pictograma. Substitúese a típica barra por un debuxo relacionado coa variable que se representa.





Nº de bombillas vendidas en un supermercado

  - Cartograma. Represéntase a variable sobre un mapa.

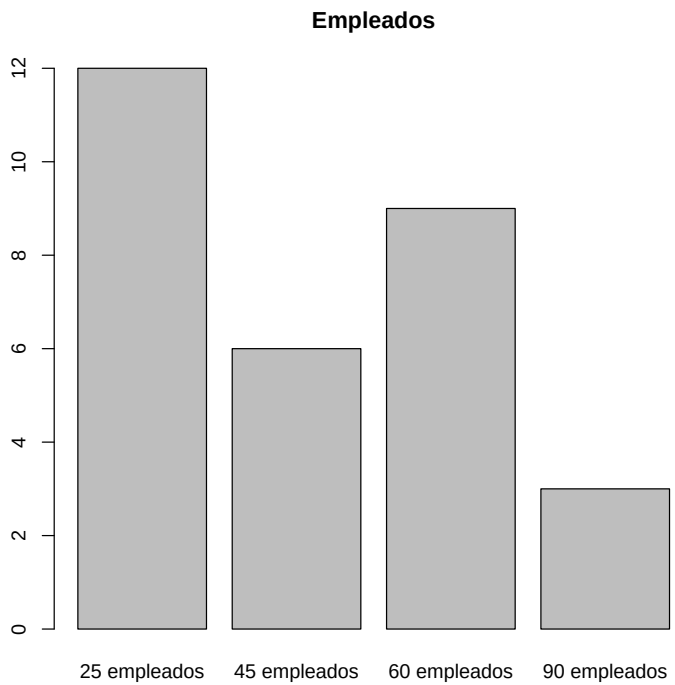
Exemplo. Variable cuantitativa discreta. Número de empregados

Consideremos unha mostra de 80 empresas, nas que observamos o número de empregados. Os datos que obtiveron son:

- 8 empresas con 25 empregados,
- 20 empresas con 45 empregados,
- 28 empresas con 60 empregados e
- 24 empresas con 90 empregados.

Tamaño mostral: $n = 80$

x_i	n_i	f_i	N_i	F_i
25	8	0.1	8	0.1
45	20	0.25	28	0.35
60	28	0.35	56	0.7
90	24	0.3	80	1



Construción dun histograma

- Para construir as frecuencias é habitual agrupar os valores que pode tomar a variable en intervalos. Deste modo contamos o número de veces que a variable cae en cada intervalo
- A cada un destos intervalos chamámoslles **intervalo de clase (Clase)**: $(L_{i-1}, L_i]$:
  - O seu punto medio é a **marca de clase**: $c_i = \dfrac{L_{i-1} + L_i}{2}$,
  - a **amplitude** do intervalo é $a_i = L_i - L_{i-1}$ y
  - a **densidade de datos** do intervalo: $d_i = \dfrac{n_i}{a_i}$.
- Polo tanto, para a definición das frecuencias e a construción da táboa de frecuencias substituiremos os valores x_i polos intervalos de clase e as marcas de clase.

Descrición de variables cuantitativas continuas

Algunhas consideracións a ter en conta:

- *Número de intervalos a considerar:*
  - Cantos menos intevalos tomemos, menos información se recolle.
  - Cantos máis intervalos tomemos, máis difícil é manexar as frecuencias.

  Acostúmase tomar como número de intervalos o entero máis próximo a $\sqrt{n}$.

- *Amplitude de cada intervalo:* O máis común, salvo xustificación na súa contra, é tomar tódolos intervalos de igual lonxitude.

- *Posición dos intervalos:* Os intervalos deben situarse alí donde se atopan as observacións e de forma contigua.

Exemplo. Variable cuantitativa continua

Considérase unha mostra de 10 personas, e pregúntaselles a idade (en anos) na que firmaron o seu primero contrato indefinido. As respostas foron as seguintes:

$$52,\ 47,\ 51,\ 28,\ 64,\ 31,\ 22,\ 53,\ 29,\ 23$$

¿Cómo resumimos a información contida nos datos da variable IDADE?

- Mostra ordeada: $22,\ 23,\ 28,\ 29,\ 31,\ 47,\ 51,\ 52,\ 53,\ 64$.
- Percorrido$= 64 - 22 = 42$.
- Número de intervalos$\simeq \sqrt{10} \simeq 3'162 \simeq 3$.
- Como $42/3 = 14$, tomaremos 15 como amplitude de cada intervalo.

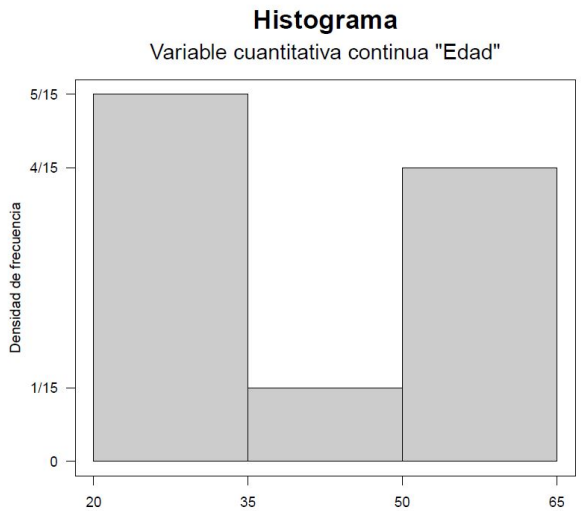
Intervalo de clase $(L_{i-1}, L_i]$	Marca de clase c_i	n_i	f_i	N_i	F_i	Densidade de frecuencia $d_i = n_i/(L_i - L_{i-1})$
$(20, 35]$	$27'5$	5	$0'5$	5	$0'5$	$5/15$
$(35, 50]$	$42'5$	1	$0'1$	6	$0'6$	$1/15$
$(50, 65]$	$57'5$	4	$0'4$	10	1	$4/15$

Representacións gráficas. Variable cuantitativa continua. Idade

A distribución de frecuencias dunha variable continua represéntase mediante o chamado **histograma.**
A área (e non a altura) dos rectángulos debe ser proporcional á frecuencia. O eixo de ordenadas no reflexa a frecuencia, senón que a altura de cada rectángulo representa a **densidade de frecuencia** sobre ese intervalo:

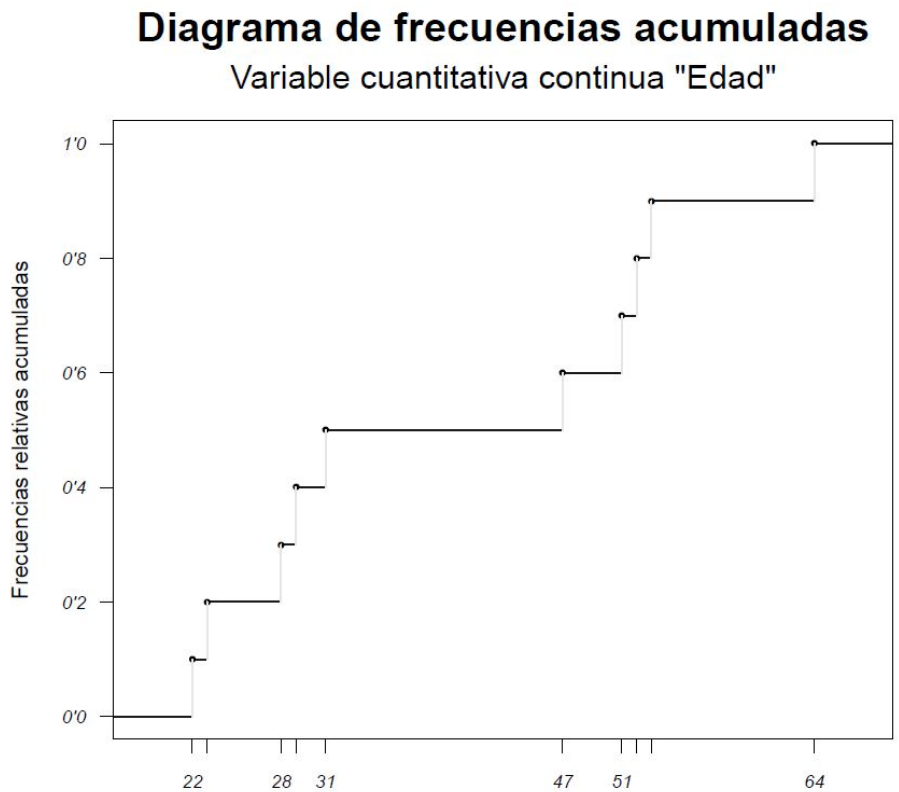
$$\text{Densidad de frecuencia} = \frac{\text{frecuencia absoluta}}{\text{amplitud}}$$

Só cando tódolos intervalos teñan a mesma amplitude, será equivalente representar a frecuencia ou a densidade de frecuencia.

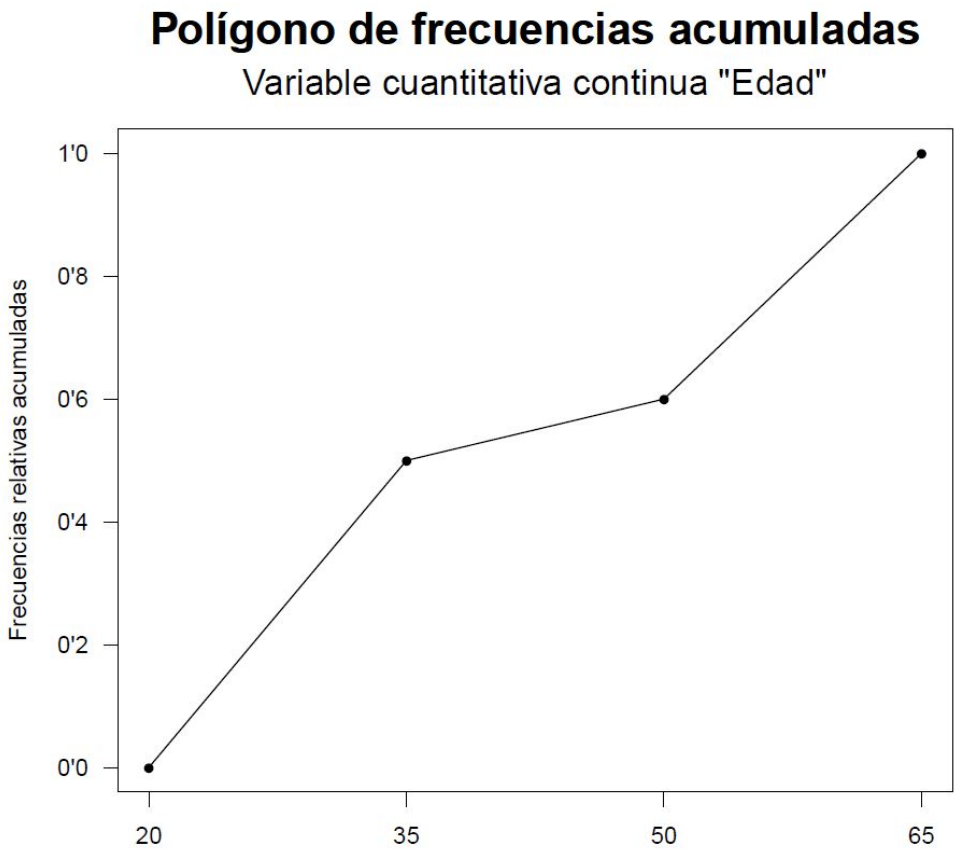


Representacións gráficas. Variable cuantitativa continua. Idade

Se se dispón de tódolos datos, representaremos as frecuencias acumuladas mediante o **diagrama de frecuencias acumuladas** ou **diagrama escalonado**, sabendo que todos os saltos son de amplitude $\frac{1}{n}$. Os saltos prodúcense nos datos da mostra, de modo que nas zonas onde hai máis datos a escaleira medra máis rapidamente.



En caso de agrupación en intervalos, as frecuencias acumuladas represéntanse mediante o **polígono de frecuencias acumuladas**. Como non se coñece o lugar exacto no que se atopa cada individuo da mostra, repártese a frecuencia de cada intervalo de maneira uniforme dentro do intervalo, o cal resulta en segmentos con pendente igual á densidad de frecuencia en cada intervalo.



Exercicio

O servicio médico dunha empresa recibe a visita de oito dos seus empregados con dor lumbar durante unha semana. Tódolos datos quedan recollidos na seguinte táboa. Clasifica as variables recollidas (sexo, peso, estatura, temperatura, número de visitas previas ó servicio e dor).

Sexo	Peso (kg.)	Estatura (m.)	Temperatura (°C)	Visitas	Dor
M	63	1.74	38	0	Leve
M	58	1.63	36.5	2	Intenso
H	84	1.86	37.2	0	Intenso
M	47	1.53	38.3	0	Moderado
M	70	1.75	37.1	1	Intenso
M	57	1.68	36.8	0	Leve
H	87	1.82	38.4	1	Leve
M	55	1.46	36.6	1	Intenso

Resume la información contenida en los datos de las diferentes variables.

Medidas características: Medidas de posición, de dispersión e de forma

Por **medida** entendemos un número que se calcula sobre a mostra, efectuando operacións cos valores que toma a variable, e que reflite certa cualidade da mesma.. Por este motivo, no que resta tratamos só con variables cuantitativas.

- Medidas de posición: son medidas que indican a posición que ocupa a mostra
- Medidas de dispersión: describen a variabilidade ds datos da mostra respecto á posición central
- Medidas de forma: miden o grado de simetría e apuntamento nos datos.

Medidas de posición. Media aritmética

Defínese a media aritmética (ou simplemente media) como:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{o bien} \quad \bar{x} = \sum_{i=1}^{k} (c_i \text{ ó } x_i) f_i$$

onde a primeira expresión emprégase cando se dispón de tódolos datos (sen agrupar), mentres que a segunda expresión aplícase a datos agrupados, empregando as frecuencias de cada valor diferente.

No caso dunha variable continua, temos dúas opcións: ou calculamos a media con tódolos datos, que denotamos por x_i (sumámolos e dividimos polo tamaño mostral), ou usamos a táboa de frecuencias considerando as marcas de clase (c_i en lugar de x_i) e as frecuencias en cada clase.

Propiedades

1. A media mídese nas mesmas unidades que os datos orixinais.

2. É o centro de gravidade dos datos:
mín $x_i \leq \bar{x} \leq max\ x_i$
$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min_{a\in\mathbb{R}} \sum_{i=1}^{n}(x_i - a)^2$

3. Se $y_i = a + bx_i$ entón $\bar{y} = a + b\bar{x}$, esto é, se se multiplican por b as observacións de X (por exemplo, ó cambiar de unidades) e se trasladan sumando unha constante a, entón a media de X cambia as súas unidades e trasládase na mesma constante.

4. Se a distribución de frecuencias é simétrica respecto a un valor M, entón $\bar{x} = M$.

Exemplo. Medidas de posición. Media. Número de horas extra

A dirección dunha empresa avalía o número de horas extra dos seus 20 empregados, detallados na seguinte táboa..

N.º de horas extra

x_i	n_i	f_i	N_i	F_i	$x_i \cdot f_i$
0	5	$0'25$	5	$0'25$	0
1	8	$0'40$	13	$0'65$	$0'40$
2	4	$0'20$	17	$0'85$	$0'40$
3	2	$0'10$	19	$0'95$	$0'30$
4	1	$0'05$	20	1	$0'20$
SUMAS	20	1			$1'3$

Polo tanto, a media é:

$$\bar{x} = \sum_{i=1}^{k} x_i f_i = 1'3 \text{ horas extra.}$$

Exemplo. Medidas de posición. Media. Idade

- Xa que disponemos de tódolos datos, calcularemos a media da variable Idade con todos eles:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(52 + 47 + 51 + \cdots + 53 + 29 + 23) = \frac{400}{10} = 40 \text{ anos}$$

- A continuación, calculamos a media da variable Idade a partir dos datos xa agrupados en intervalos de clase:

Intervalo de clase $(L_{i-1}, L_i]$	Marca de clase c_i	n_i	f_i	N_i	F_i	$c_i f_i$
$(20, 35]$	$27'5$	5	$0'5$	5	$0'5$	$13'75$
$(35, 50]$	$42'5$	1	$0'1$	6	$0'6$	$4'25$
$(50, 65]$	$57'5$	4	$0'4$	10	1	23
	SUMAS	20	1			41

Neste caso:

$$\bar{x} = \sum_{i=1}^{k} c_i f_i = 41 \text{ anos}$$

Se queremos dar a idade en meses:

$y_i = 12 * x_i$ y $\bar{y} = 12 * \bar{x} = 12 * 40 = 480$ meses.

Medidas de posición. Mediana

Ordeados os datos de menor a mayor, a mediana é o valor máis pequeno da variable que deixa á súa esquerda, como mínimo, a metade dos valores de dita variable.
Se a variable está agrupada en intervalos de clase, buscamos sobre a táboa de frecuencias o primeiro intervalo cunha frecuencia relativa acumulada que é maior ou igual que $\frac{1}{2}$ (a clase mediana) e dentro dela pódese obter a mediana por interpolación lineal, ó supoñer que os datos se distribuén de maneira uniforme dentro do intervalo.

Número de horas extra

N.º de horas extra

x_i	n_i	f_i	N_i	F_i
0	5	0′25	5	0′25
1	8	0′40	13	0′65
2	4	0′20	17	0′85
3	2	0′10	19	0′95
4	1	0′05	20	1
SUMAS	20	1		

Polo tanto, a mediana é 1 hora extra.

Medidas de posición. Moda

- Denótase por *Mo* e o intervalo con maior frecuencia será a **clase modal.**
- É o valor da variable que se presenta con maior frecuencia.
  - Datos non agrupados: valor da variable de maior frecuencia absoluta ou relativa.
  - Datos agrupados: búscase o intervalo modal (i), o de maior densidade de datos. Entón,

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} a_i$$

    - Se $i = 1$, $d_{i-1} = 0$. Polo tanto, $Mo = L_i$.
    - Se $i = \text{n}^\text{o}$ de clases, $d_{i+1} = 0$. Polo tanto, $Mo = L_{i-1}$.
- Pode ocorrer que exista unha única moda, e falamos de distribución de frecuencias **unimodal.** Se hai máis dunha moda, diremos que a distribución é **multimodal.**

Exemplo. Medidas de posición. Moda

Número de horas extra

N.º de horas extra

x_i	n_i	f_i	N_i	F_i
0	5	0′25	5	0′25
1	8	0′40	13	0′65
2	4	0′20	17	0′85
3	2	0′10	19	0′95
4	1	0′05	20	1
SUMAS	20	1		

Idade

Intervalo de clase	Marca de clase					
$(L_{i-1}, L_i]$	c_i	n_i	f_i	N_i	F_i	$c_i f_i$
(20, 35]	27′5	5	0′5	5	0′5	13′75
(35, 50]	42′5	1	0′1	6	0′6	4′25
(50, 65]	57′5	4	0′4	10	1	23
	SUMAS	20	1			

Medidas de posición. Cuantís

- Vimos que a mediana divide os datos en dúas partes iguais. Pero tamén ten interés estudar outros parámetros, os cuantís, que dividen os datos da distribución en partes iguais, é dicir, en intervalos que comprenden o mesmo número de valores.
- Sea $p \in (0, 1)$. Unha vez ordeados os datos de menor a mayor, defínese o **cuantil p**, como o valor máis pequeno da variable que deixa a súa esquerda np observacións. Isto é, a frecuencia relativa acumulada ata o **cuantil p** é maior ou igual que **p**. Nótese que a mediana é o **cuantil 0'5**. Os cuantís, o igual que a mediana, só se poderán calcular con variables que admitan un orden.
- Algúns ordenacións dos cuantís teñen nomes específicos. Así o **cuartís** son os cuantís de orde (0.25, 0.5, 0.75) e represéntanse por Q_1, Q_2, Q_3. Os cuartís dividen a distribución en catro partes. Os **decís** son os cuantís de orde (0.1, 0.2,..., 0.9). Os **percentís** son os cuantís de orde $j/100$ onde j=1,2,...,99.

Se a variable é discreta, ou se é continua e dispoñemos de tódolos datos, empezamos ordeando a mostra.
O **cuantil p** é o menor dato da mostra (primero da mostra ordeada) que ten por frecuencia relativa acumulada máis ou igual que **p**.

Para **datos non agrupados** búscase a primeira frecuencia acumulada tal que $N_i \geq pn$:

1. Se $N_i > pn$, entón $x_p = x_i$.
2. Se $N_i = pn$, entón $x_p = \dfrac{x_i + x_{i+1}}{2}$.

Se a variable é continua e está **agrupada** en intervalos de clase, buscamos o primeiro intervalo que teña unha frecuencia relativa acumulada maior ou igual que **p**, correspóndese co valor

$$x_p = L_{i-1} + \frac{pn - N_{i-1}}{n_i} a_i = L_{i-1} + \frac{p - F_{i-1}}{f_i} a_i$$

Medidas de dispersión

As medidas de dispersión utilízanse para describir a variabilidade ou esparcimento dos datos da mostra respecto á posición central. A continuación describimos as máis importantes:

- Percorrido ou rango: $R = \text{máx}_i\, x_i - \text{mín}_i\, x_i$
- Percorrido intercuartílico: Diferencia entre el cuartil terceiro e primeiro
- Varianza e desviación típica
- Cuasivarianza e cuasidesviación típica
- Coeficiente de variación

Medidas de dispersión. Percorrido e Percorrido intercuartílico

- Percorrido ou rango: $R = \text{máx}_i\, x_i - \text{mín}_i\, x_i$
  No exemplo da xornada laboral,
  Percorrido $= 6'81 - 6'14 = 0'67$
- Percorrido intercuartílico: Diferencia entre o cuartil terceiro e primeiro
  No exemplo da xornada laboral,
  Percorrido intercuartílico $= 6'67 - 6'46 = 0'21$

*levan asociadas as unidades de medida.

Diagrama de caixa (boxplot)

O diagrama de caixa é unha representación gráfica que se utiliza con variables continuas. Permite describir a dispersión e a simetría da distribución de datos. O diagrama de caixa está formado por:

- unha caixa delimitada polos cuartiles Q1 y Q3, que no seu interior ten representada unha liña horizontal á altura da mediana. Nótese que dentro da caixa atoparase a metade das observacións. Se a mediana non se atopa no centro da caixa, a distribución non é simétrica.
- unña liña vertical desde o terceiro cuartil ata o valor maior da mostra que non sexa un valor atípico,
- unña liña vertical desde o primeiro cuartil ata o valor menor da mostra que non sexa un valor atípico,
- círculos que representan os valores atípicos da mostra.

Diagrama de caixa (boxplot)

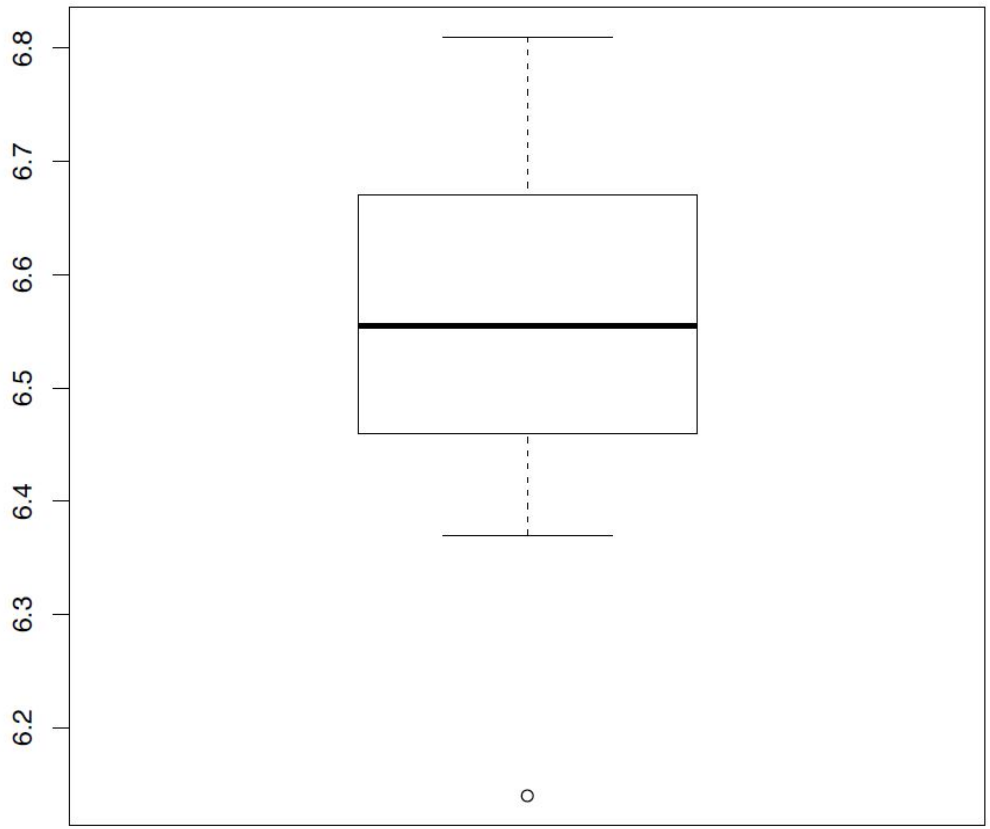
- Os segmentos horizontales inferior e superior (ou bigotes) e que delimitan as liñas verticais discontinuas, acadan as últimas observacións da mostra que non son atípicas. Polo tanto, o extremo inferior será a menor observación maior ou igual que $Q1 - 1,5 \cdot RIC$ e o extremo superior será a mayor observación menor ou igual que $Q3 + 1,5 \cdot RIC$.

Nota Considérase que un dato x é **atípico** se está en algunha destas dúas circunstancias:

$$x < Q1 - 1'5 \cdot RIC \quad \text{o} \quad x > Q3 + 1'5 \cdot RIC$$

sendo RIC = rango intercuartílico = $Q3 - Q1$.

Ejemplo. Diagrama de caja (boxplot). Xornada laboral



Medidas de dispersión. Varianza

A media emprégase como medida de posición. Entón, parece razoable tomar como medida de dispersión algún criterio de discrepancia dos puntos respecto á media.
Recorda que a simple diferencia dos puntos á media, ó ponderala, dá cero. Polo tanto, elevamos esas diferencias ó cadrado para que non se cancelen os sumandos positivos cos negativos. O resultado é a **varianza**:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\,; \qquad S^2 = \sum_{i}^{k}(x_i \text{ ó } c_i - \bar{x})^2\, f_i$$

Propiedades

1.- $S^2_{a+X} = S^2_X$. A varianza non se ve afectada por cambios de localización.

2.- $S^2_{b\cdot X} = b^2 \cdot S^2_X$. A varianza mídese no cadrado da escala da variable.

Medidas de dispersión. Desviación típica

A propiedade 2 dá pé a calcular a raíz cadrada da varianza, obtendo así unha medida de dispersión que se expresa nas mesmas unidades da variable. Esta medida é a **desviación típica**, ou *desviación estándar*, que en coherencia denotamos por S.

Exemplo. Medidas de dispersión. N° de horas extra

Calculemos a varianza e a desviación típica da variable n.° de horas extra:

N.° de horas extra	n_i	f_i	N_i	F_i	$x_i f_i$	$(x_i - \bar{x})^2 f_i$
0	5	0,25	5	0,25	0	0,4225
1	8	0,40	13	0,65	0,40	0,0360
2	4	0,20	17	0,85	0,40	0,0980
3	2	0,10	19	0,95	0,30	0,2890
4	1	0,05	20	1	0,20	0,3645
SUMAS	20	1			1,3	1,21

Para o cálculo da varianza e desviación típica necesitamos previamente a media da variable ($\bar{x} = 1{,}3$).
Polo tanto, a varianza da variable n.° de horas extra é:

$$S^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 f_i = \sum_{i=1}^{k} (x_i - 1{,}3)^2 f_i = 1{,}21$$

E a desviación típica:
$S = \sqrt{1{,}21} = 1{,}1$ (aproximadamente, unha hora extra)

Exemplo. Medidas de dispersión. Idade

Calculemos a varianza e a desviación típica da variable Idade a partir da táboa de frecuencias onde os datos foron agrupados en intervalos de clase:

Intervalo de clase $(L_{i-1}, L_i]$	Marca de clase c_i	n_i	f_i	N_i	F_i	$c_i f_i$	$(c_i - \bar{x})^2 f_i$
$(20, 35]$	27,5	5	0,5	5	0,5	13,75	91,125
$(35, 50]$	42,5	1	0,1	6	0,6	4,25	0,225
$(50, 65]$	57,5	4	0,4	10	1	23	108,900
	SUMAS	20	1			41	200,25

A varianza e a desviación típica son, respectivamente
$S^2 = \sum_{i=1}^{k}(c_i - \bar{x})^2 f_i = \sum_{i=1}^{k}(c_i - 41)^2 f_i = 200{,}25$ y $S = \sqrt{200{,}25} \simeq 14{,}15$ anos.

- Para calcular a desviación típica con tódolos datos:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 40 \text{ anos}$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10}\left[(52-40)^2 + (47-40)^2 + \cdots + (23-40)^2\right] = 201{,}8$$

$$S = \sqrt{201{,}8} \simeq 14{,}21 \text{ anos.}$$

Medidas de dispersión. Cuasivarianza e cuasidesviación típica

Se dividimos a varianza por $(n-1)$ en lugar de por n, obtemos a **cuasivarianza**:

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

Coñecendo a varianza pódese calcular a cuasivarianza, e viceversa, pois $S_c^2 = n \cdot S^2/(n-1)$. Ademáis, ambas medidas exprésanse nas unidades da variable ó cadrado.
A **cuasidesviación típica** é simplemente a raíz cadrada da cuasivarianza, e polo tanto denotámola por S_c.

Nº horas extra
$S_c^2 = \frac{n}{n-1}S^2 = \frac{20}{19} \cdot 1{,}21 \simeq 1{,}27$
$S_c = \sqrt{1{,}27} \simeq 1{,}13$ (aproximadamente unha hora extra)

Idade
$S_c^2 = \frac{n}{n-1}S^2 = \frac{10}{9} \cdot 201{,}8 \simeq 224{,}22$
$S_c = \sqrt{224{,}22} \simeq 14{,}97$ años

Medidas de dispersión. Coeficiente de variación

Hai situacións nas que temos que comparar poblacións nas que

- as unidades de medida son distintas

Exemplo:

Peso de formigas en gramos: ($s = 2{,}41$ gramos)

8.180881	10.503650	8.210198	13.096271	9.259044
15.540982	7.854185	12.010111	8.725924	11.712810

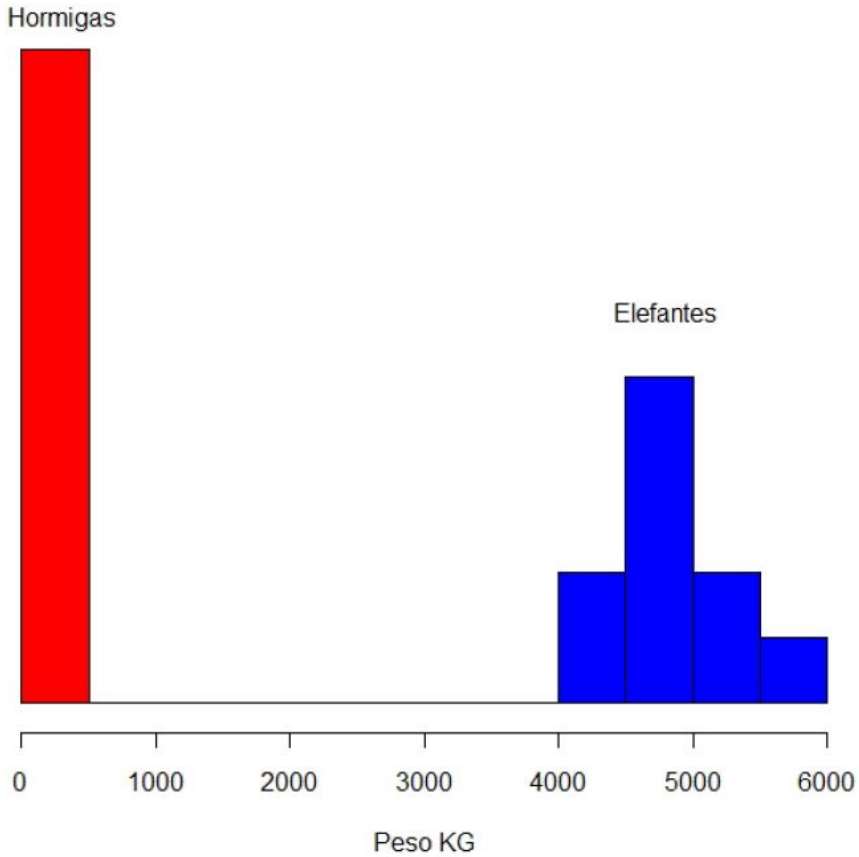
Peso de elefantes en kg: ($s = 320{,}0495$ kilos)

5100.636	4987.702	5035.441	5321.591	5502.833
4737.402	4537.105	4731.434	4742.981	4444.282

Medidas de dispersión. Coeficiente de variación

Hai situacións nas que temos que comparar poblacións nas que

- ou que aínda coa mesma unidade de medida difiren nas súas magnitudes.



Para estos casos precisamos unha medida da dispersión na que no inflúan as unidades, sería conveniente ter unha medida adimensional.

Medidas de dispersión. Coeficiente de variación

Si queremos unha medida de dispersión que non dependa da escala e que, polo tanto, permita unha comparación das dispersiones relativas de varias mostras, existen varias propostas, pero quedámonos co **coeficiente de variación**:

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

Nº de horas extra

$$CV = \frac{1'1}{1'3} \simeq 0'846 \ (84'6\,\%)$$

Idade

$$CV = \frac{14'21}{40} \simeq 0'356 \ (35'6\,\%)$$

Distribución de frecuencias bidimensional

En adiante, trabállase con n pares de observacións da variable (X, Y) que poden presentarse:

- Individualmente ou en extensión: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- En táboas de frecuencias de dobre entrada, representando unha distribución bidimensional de frecuencias

Exemplo Unha mostra de 500 vivendas en Galicia clasificouse segundo a súa superficie en m^2 (X) e o número de dormitorios (Y) resultando a seguinte distribución conxunta de frecuencias:

X/Y	1	2	3	4
0-60	9	23	6	0
60-90	7	65	120	22
90-120	2	15	87	50
120-200	0	5	29	60

Distribución de frecuencias bidimensional

- A variable X toma r valores diferentes: $x_1, x_2, \ldots, x_r$.
- A variable Y toma c valores diferentes: $y_1, y_2, \ldots, y_c$.
- A frecuencia absoluta conxunta do valor x_i de X co valor y_j de Y é n_{ij}, para todo $i = 1, \ldots, r$ y para todo $j = 1, \ldots, c$.

Propiedades

- O número total de observacións é n:

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

- A frecuencia relativa conxunta para todo $i = 1, \ldots, r$ e para todo $j = 1, \ldots, c$ é

$$f_{ij} = \frac{n_{ij}}{n}$$

- A suma de tódalas frecuencias relativas é igual a 1, é dicir,

$$\sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij} = 1.$$

Táboa de dobre entrada

A distribución bidimensional de frecuencias adoita presentarse nunha táboa de dobre entrada coñecida como:

- Táboa de correlación se as variables son cuantitativas.
- Táboa de continxencia se algunha das variables é cualitativa

Notación da táboa

X/Y	y_1	y_2	$\dots$	y_c	
x_1	n_{11}	n_{12}	$\dots$	n_{1c}	
x_2	n_{21}	n_{22}	$\dots$	n_{2c}	
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
x_r	n_{r1}	n_{r2}	$\dots$	n_{rc}	
					n

Táboa de dobre entrada: exemplo

Exemplo Unha mostra de 500 vivendas en Galicia clasificouse segundo a súa superficie en m^2 (X) e o número de dormitorios (Y):

X/Y	1	2	3	4
0-60	9	23	6	0
60-90	7	65	120	22
90-120	2	15	87	50
120-200	0	5	29	60

a) ¿Cal será a distribución en frecuencias relativas?

X/Y	1	2	3	4
0-60	0.018	0.046	0.012	0
60-90	0.014	0.130	0.240	0.044
90-120	0.004	0.030	0.174	0.100
120-200	0	0.010	0.058	0.120

Táboa de dobre entrada: exemplo

Exemplo Unha mostra de 500 vivendas en Galicia clasificouse segundo a súa superficie en m^2 (X) e o número de dormitorios (Y):

X/Y	1	2	3	4
0-60	9	23	6	0
60-90	7	65	120	22
90-120	2	15	87	50
120-200	0	5	29	60

b) ¿Que porcentaxe de vivendas ten máis de 120 m^2 e 3 dormitorios?

$$\frac{29}{500} = 0{,}058 \longrightarrow 5{,}8\,\%$$

c) Entre as vivendas con superficie entre 60 y 90 m^2, ¿que porcentaxe ten máis de 2 dormitorios?

$$\frac{120+22}{7+65+120+22} = \frac{142}{214} = 0{,}6636 \longrightarrow 66{,}36\,\%$$

Táboa de dobre entrada: exemplo

Exemplo Unha mostra de 500 vivendas en Galicia clasificouse segundo a súa superficie en m^2 (X) e o número de dormitorios (Y):

X/Y	1	2	3	4
0-60	9	23	6	0
60-90	7	65	120	22
90-120	2	15	87	50
120-200	0	5	29	60

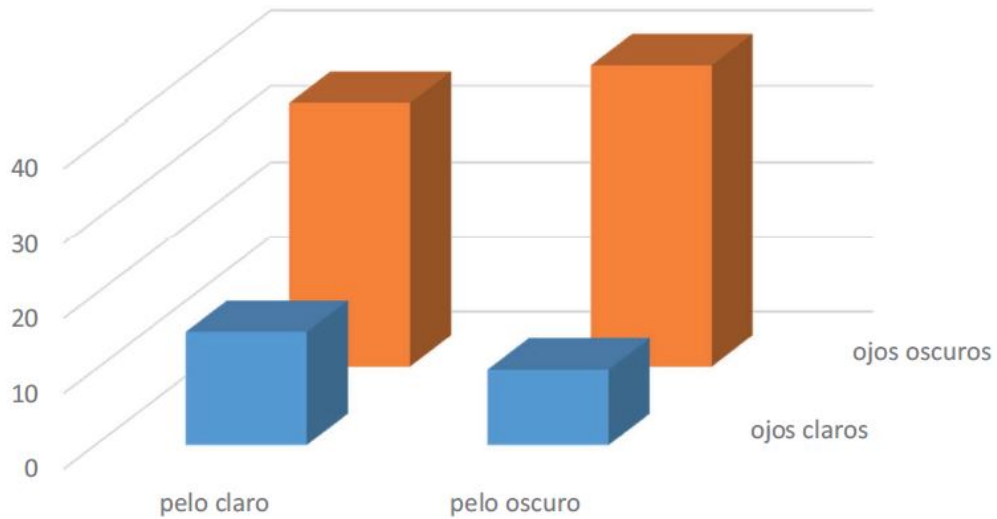
d) Entre as vivendas de menos de 3 dormitorios, ¿qué porcentaxe ten unha superficie superior a 90 m^2?

$$\frac{2+15+5}{9+23+7+65+2+15+5} = \frac{22}{126} = 0{,}1746 \longrightarrow 17{,}46\,\%$$

e) Distribución de frecuencias da variable X:

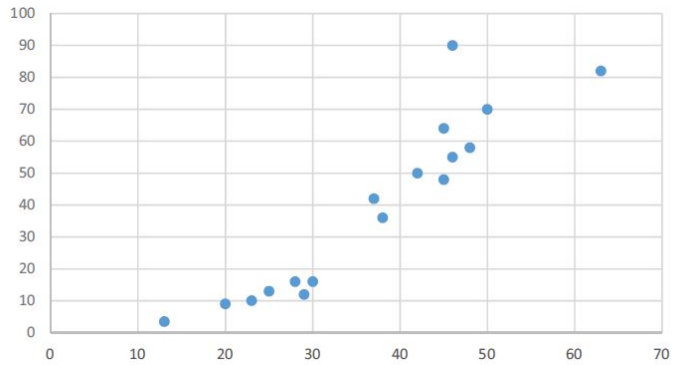
$L_{i-1} - L_i$	$n_{i\cdot}$	$f_{i\cdot}$
0-60	38	0.076
60-90	214	0.428
90-120	154	0.308
120-200	94	0.188
	500	1

- **Diagrama de barras**: Para variables cualitativas ou cuantitativas sen agrupar.

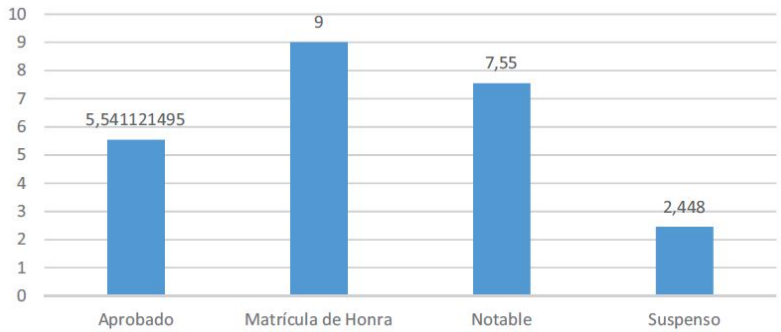


Representacións gráficas

- **Diagrama de dispersión**: Para variables cuantitativas sen agrupar.



- **Gráficos de resumo**: Unha variable é explicada en función da outra.



Distribucións marxinais

A partir da táboa de dobre entrada podemos obter as distribucións de X ou de Y.

X/Y	y_1	y_2	$\dots$	y_c	
x_1	n_{11}	n_{12}	$\dots$	n_{1c}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\dots$	n_{2c}	$n_{2\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
x_r	n_{r1}	n_{r2}	$\dots$	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot c}$	n

- A frecuencia marxinal de x_i, $i = 1, \dots, r$:

$$n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij}$$

(correspóndese con sumar por filas a táboa)

- A frecuencia marxinal de y_j, $j = 1, \dots, c$:

$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{rj} = \sum_{i=1}^{r} n_{ij}$$

(correspóndese con sumar por columnas a táboa)

Distribucións marxinais

Chamaremos distribucións marxinais ás distribucións unidimensionais de frecuencias das variables X e Y; respectivamente:

$$(x_i; n_{i\cdot}),\ i = 1, 2, \ldots, r \text{ y } (y_j; n_{\cdot j}),\ j = 1, 2, \ldots, c.$$

- A frecuencia relativa marxinal de X,

$$f_{i\cdot} = \frac{n_{i\cdot}}{n}$$

- A frecuencia relativa marxinal de Y,

$$f_{\cdot j} = \frac{n_{\cdot j}}{n}$$

Táboa de dobre entrada: exemplo

a) Superficie media e mediana das vivendas e varianza da superficie (X):

$(L_{i-1}, L_i]$	$n_{i\cdot}$	c_i	$c_i n_{i\cdot}$	$N_{i\cdot}$	$c_i^2 n_{i\cdot}$
0-60	38	30	1140	38	34200
60-90	214	75	16050	252	1203750
90-120	154	105	16710	406	1697850
120-200	94	160	15040	500	2406400
	500		48400		5342200

Media:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} c_i n_{i\cdot} = \frac{48400}{500} = 96{,}8\ m^2$$

Mediana:

$$\frac{n}{2} = 250 \rightarrow Me_x \in (60, 90]$$

$$Me_X = L_{i-1} + \frac{n/2 - N_{i-1}}{n_i} \times a_i = 60 + \frac{250 - 38}{214} \times 30 = 89{,}72\ m^2$$

Varianza:

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{r} c_i^2 n_{i\cdot} - \bar{x}^2 = \frac{5342200}{500} - 96{,}8^2 = 1314{,}16(m^2)^2.$$

Táboa de dobre entrada: exemplo

b) Número medio e máis frecuente de dormitorios nunha vivenda e varianza do número de dormitorios (Y):

y_j	$n_{\cdot j}$	$y_j n_{\cdot j}$	$y_j^2 n_{\cdot j}$
1	18	18	18
2	108	216	432
3	242	726	2178
4	132	528	2112
	500	1488	4740

Media:

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{c} y_j n_{\cdot j} = \frac{1488}{500} = 2{,}976 \ m^2$$

Moda: $Mo_Y = 3$

Varianza:

$$S_Y^2 = \frac{1}{n}\sum_{j=1}^{c} y_j^2 n_{\cdot j} - \bar{y}^2 = \frac{4740}{500} - 2{,}976^2 = 0{,}6234$$

c) ¿Cal das dúas variables presenta maior dispersión?

$$CV_X = \frac{S_X}{\bar{x}} = \frac{36{,}251}{96{,}8} = 0{,}3745 \text{ y } CV_Y = \frac{S_Y}{\bar{y}} = \frac{0{,}7896}{2{,}976} = 0{,}2653$$

Exemplo. Distribución da superficie para as vivendas de 2 dormitorios:

$X\|Y=2$	$n_{i\|Y=2}$	$f_{i\|Y=2}=\frac{n_{i\|Y=2}}{108}$
0-60	23	0.213
60-90	65	0.602
90-120	15	0.139
120-200	5	0.0462
	108	1

Exemplo. Distribución da superficie para as vivendas de 3 dormitorios:

$X\|Y=3$	$n_{i\|Y=3}$	$f_{i\|Y=3}=\frac{n_{i\|Y=3}}{242}$
0-60	6	0.025
60-90	120	0.496
90-120	87	0.360
120-200	29	0.120
	242	1

Distribucións condicionadas

Distribución de X condicionada ó valor y_j da variable Y $(X|Y = y_j)$

- Represéntase por $(x_i; n_{i|Y=y_j} = n_{ij})$, con $i = 1, 2, \ldots, r$.
- Número total de observacións: $\sum_{i=1}^{r} = n_{\cdot j}$
- Frecuencias relativas: $f_{i|Y=y_j} = \frac{n_{i|Y=y_j}}{n_{\cdot j}} = \frac{n_{ij}}{n_{\cdot j}}$

Distribución de Y condicionada ó valor x_i da variable Y $(Y|X = x_i)$

- Represéntase por $(y_j; n_{j|X=x_i} = n_{ij})$, con $i = 1, 2, \ldots, c$.
- Número total de observacións: $\sum_{j=1}^{c} = n_{i\cdot}$
- Frecuencias relativas: $f_{i|Y=y_j} = \frac{n_{j|X=x_i}}{n_{i\cdot}} = \frac{n_{ij}}{n_{i\cdot}}$

Distribucións condicionadas

- Poden definirse tamén distribucións condicionadas a un conxunto ou intervalo de valores; por exemplo, a distribución de $X|Y \leq y_j$ ou a de $Y|X > x_i$.
- Nas distribucións condicionadas o estudo redúcese á parte da táboa determinada pola condición.
- A distribucións condicionadas son distribucións unidimensionales.

Exemplo Número medio de dormitorios nas vivendas de máis de 90 m^2, isto é, trátase da distribución de $Y|X > 90$.

y_j	$n_{j\|X>90}$	$y_j n_{j\|X>90}$
1	2	2
2	20	40
3	116	348
4	110	440
	248	830

Media:

$\bar{y}_{|X>90} = \frac{1}{248}\sum_{j=1}^{c} y_j n_{j|X>90} = \frac{830}{248} = 3{,}347$, maior que a media global (2.976).

Independencia Estatística

Dúas variables X, Y son independentes estatísticamente se o comportamento dunha delas non se ve afectado polos valores que toma a outra, é dicir:

$$f_{i|Y=y_j} = f_{i\cdot} \text{ para calquera par de valores } (x_i, y_j)$$

$$e$$

$$f_{j|X=x_i} = f_{\cdot j} \text{ para calquera par de valores } (x_i, y_j).$$

Equivalentemente, X e Y son independentes se e só se

$$f_{ij} = f_{i\cdot} f_{\cdot j} \text{ y } n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \text{ para todo } i, j$$

Independencia Estadística: ejemplo

X/Y	y_1	y_2	y_3	$n_{i\cdot}$	$f_{i\cdot}$
x_1	1	3	5	9	9/54=1/6
x_2	2	6	10	18	18/54=1/3
x_3	3	9	15	27	27/54=1/2
$n_{\cdot j}$	6	18	30	$n = 54$	

Temos que ver que $f_{i|Y=y_j} = f_{i\cdot}$ con $i, j = 1, 2, 3$.

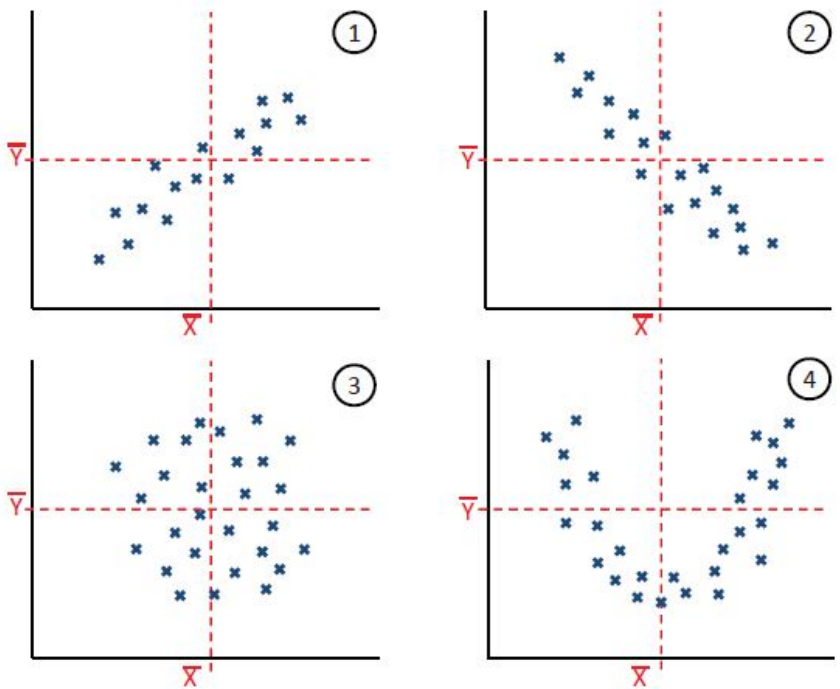
X	$n_{i\|Y=y_1}$	$f_{i\|Y=y_1}$	$n_{i\|Y=y_2}$	$f_{i\|Y=y_2}$	$n_{i\|Y=y_3}$	$f_{i\|Y=y_3}$
x_1	1	1/6	3	3/18=1/6	5	5/30=1/6
x_2	2	2/6=1/3	6	6/18=1/3	10	10/30=1/3
x_3	3	3/6=1/2	9	9/18=1/2	15	15/30=1/2
	$n_{\cdot 1} = 6$		$n_{\cdot 2} = 18$		$n_{\cdot 3} = 30$	

Covarianza: asociación entre variables cuantitativas

No caso de que as variables non sexan independentes, vamos a estudar cómo medir a posible relación **lineal** entre ambas.

Covarianza: $S_{XY} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}(x_i - \overline{x})(y_j - \overline{y})n_{ij}$, para datos tabulados

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_j - \overline{y}), \text{ para } n \text{ pares de datos}$$



1. Relación directa: $S_{XY} > 0$
2. Relación inversa: $S_{XY} < 0$
3. Independentes: $S_{XY} \approx 0$
4. Sen relación lineal: $S_{XY} \approx 0$

Coeficiente de correlación: asociación entre variables cuantitativas

¿Cómo sabemos se a relación lineal existente entre as variables é intensa ou non?

Usaremos o **coeficiente de correlación de Pearson**:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

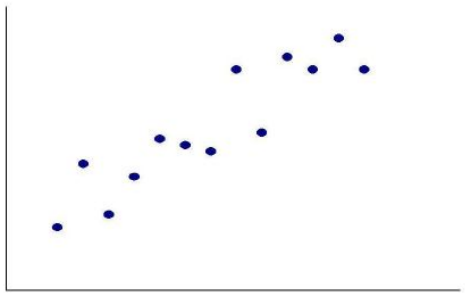
Propiedades

1. É unha medida adimensional.
2. $-1 \leq r_{XY} \leq 1$.
3. $r_{aX+b,cY+d} = r_{XY}$.
4. Cando $r_{XY} = 0$ dise que as variables están **incorreladas**: non existe relación lineal entre elas.

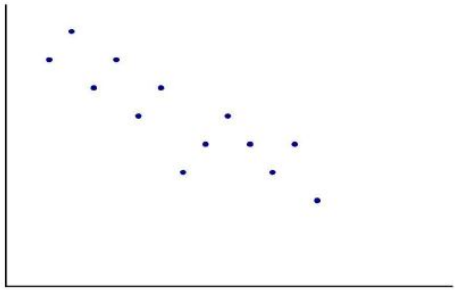
Máis comentarios

- O seu signo coincide co da covarianza.
- Se X e Y son independentes, $r_{XY} = 0$. O recíproco non é certo en xeral.
- Se $r_{XY} \neq 0$ existe asociación lineal, máis forte canto máis se acerque o coeficiente a 1 o a -1.

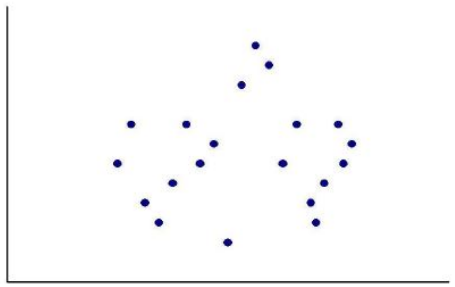
Coeficiente de correlación



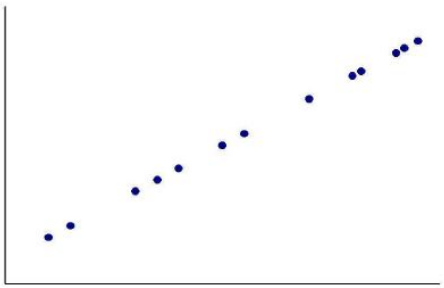
$r_{XY} > 0$



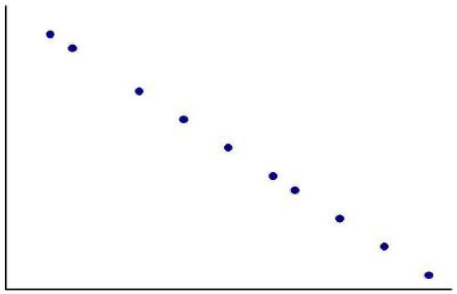
$r_{XY} < 0$



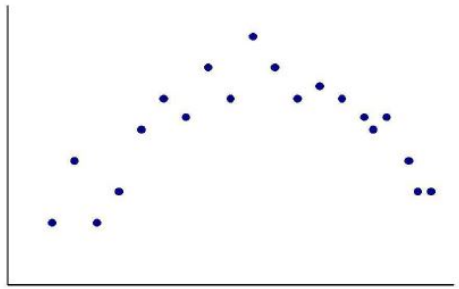
r_{XY} cercano a 0



$r_{XY} = 1$



$r_{XY} = -1$



r_{XY} cercano a 0

Coeficiente χ^2 de Pearson

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}\, n_{\cdot j}}{n}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i\cdot}\, n_{\cdot j}} - 1\right)$$

- Se X e Y son independentes, entón $\chi^2 = 0$.
- Se $\chi^2 \neq 0$, existe asociación entre as variables, de maior intensidade canto máis alto sexa o seu valor.

Coeficiente V de Crámer

- Defínese como $V = \sqrt{\frac{\chi^2}{n(\text{mín}\{r,c\}-1)}}$.
- $0 \leq V \leq 1$.
- Maior grado de dependencia ou asociación canto máis próximo a 1.

Asociación entre variables cualitativas: exemplo

Da poboación adulta con idade comprendida entre 25 y 65 anos seleccionouse unha mostra de 500 personas clasificándoas segundo a frecuencia de asistencia ó cine e o nivel educativo (educación superior ou non):

	Superior	No superior	$n_{i\cdot}$
Cada semana	8	2	10
Cada mes	35	18	53
Alguna vez al año	77	83	160
Nunca	35	242	277
$n_{\cdot j}$	155	345	500

Vamos a analizar a asociación utilizando o coeficiente V de Crámer, para o cal necesitamos calcular o coeficiente χ^2 de Pearson.

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i\cdot}\,n_{\cdot}} - 1\right) = 500\left(\frac{8^2}{10\times 155} + \frac{2^2}{10\times 345} + \cdots + \frac{242^2}{277\times 345} - 1\right) = 107{,}255$$

$$V = \sqrt{\frac{\chi^2}{n(\text{mín}\{r,c\}-1)}} = \sqrt{\frac{107{,}255}{500\times 1}} = 0{,}463$$

Recta de regresión

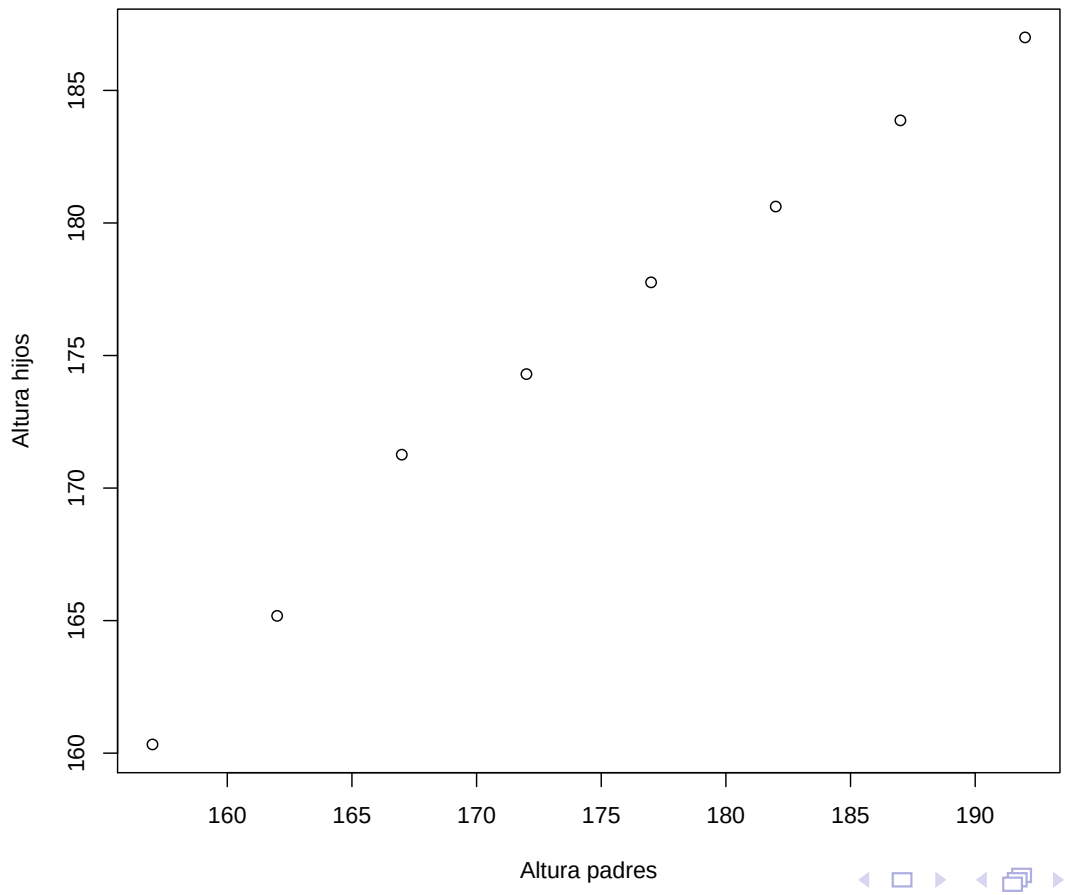
Tal e como vimos ata o momento, representaremos por (X, Y) a variable bidimensional estudada, onde X e Y son as variables unidimensionais correspondentes ás primeiras e segundas características, respectivamente, medidas para cada individuo.

- ¿Existe relación entre a altura e o peso? ¿de qué tipo é esa relación?
- ¿Cómo se relaciona a cantidade de cartos invertidos por un laboratorio para anunciar un novo fármaco coas cifras de ventas durante o primeiro mes?
- ¿Está relacionada a altura dun pai coa do seu fillo? ¿cómo?
- ¿Está relacionado o Volume Expiratorio Forzado coa estatura?

Recta de regresión: altura dos fillos vs. altura dos pais

A seguinte táboa detalla os datos (por pares) da altura de 8 pais (X) e a altura dos seus fillos (Y).

	1	2	3	4	5	6	7	8
Altura dos pais (X)	157	162	167	172	177	182	187	192
Altura dos fillos (Y)	160.33	165.18	171.26	174.30	177.76	180.62	183.87	187.00



Recta de regresión: introdución

A recta de regresión determina a relación lineal entre dúas variables continuas. Esta recta describe cómo varía a media dunha variable en función dos valores da outra.

O coeficiente de correlación lineal é unha medida resumo da asociación lineal entre dúas variables continuas.

Unha maneira alternativa é indicar a ecuación da recta que describe a situación dos puntos.

En Ciencias Sociais, establécense relacións entre variables en promedio.

Recta de regresión: introdución

Exemplo Existe relación entre a renta e o gasto das familias se, por exemplo, ó aumentar a renta das familias diminúe a proporción de gasto. Esto non implica que tódalas familias de maior renta gasten menos que as de menor renta.

En promedio, as familias de renta máis alta destinarán menos á alimentación.

- Relación positiva: se ó aumentar unha variable, aumenta en promedio a outra variable.
- Relación negativa: se ó aumentar unha variable, diminúe en promedio a outra variable.

Parece natural medir a dependencia entre dúas variables describindo cómo varía a variable dependiente en función da variable independente.

- Por exemplo, cómo varía a media da variable dependente condicionada ós valores da independente.

Exemplo. Volume Expiratorio Forzado e estatura

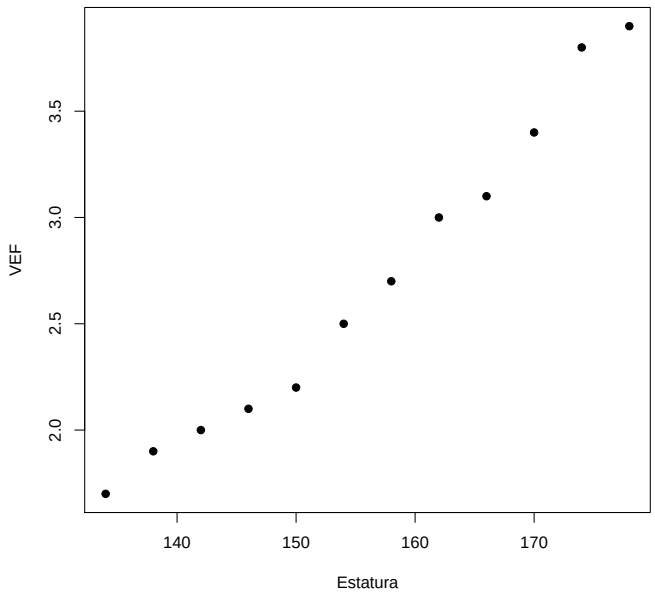
- O Volume Expiratorio Forzado (VEF) é unha medida da capacidade pulmonar.
- Crese que o VEF está relacionado coa estatura.
- Interésanos estudar a variable bidimensional (X, Y):
  - X é a estatura de nenos de 10 a 15 anos de idade.
  - Y é o VEF.
- A continuación móstrase a estatura (en cm.) e o VEF (en l.) de 12 ninos nese rango de idade:

Estatura	134	138	142	146	150	154	158	162	166	170	174	178
VEF	1.7	1.9	2.0	2.1	2.2	2.5	2.7	3.0	3.1	3.4	3.8	3.9

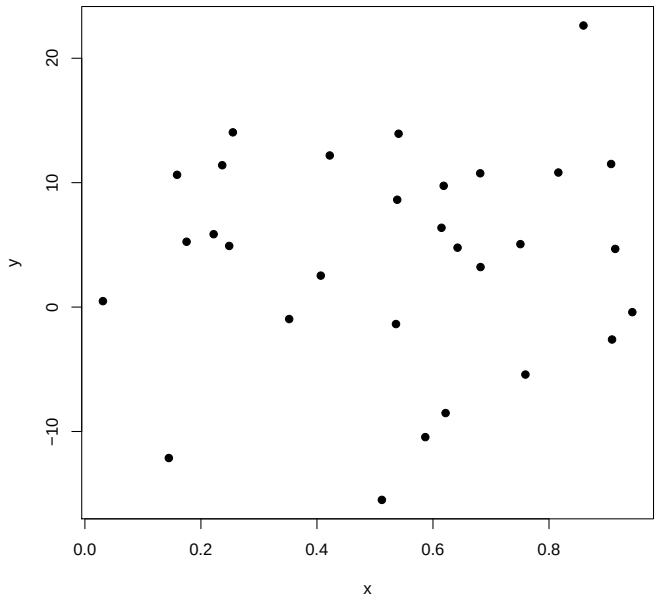
O diagrama de dispersión

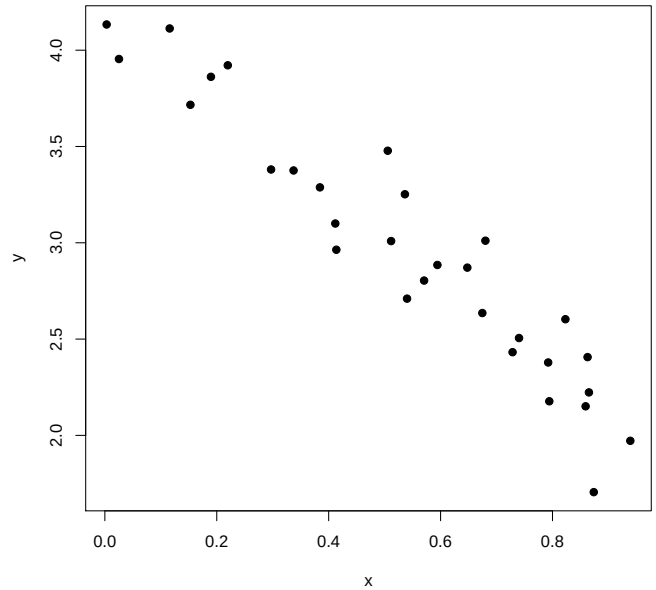
- A representación gráfica máis útil de dúas variables continuas é o **diagrama de dispersión**.
- Consiste en representar nun eixo de coordenadas os pares de observacións (x_i, y_i).
- A nube así debuxada reflite a posible relación entre as variables.
- A maior relación entre as variables máis estreita e alongada será la nube.

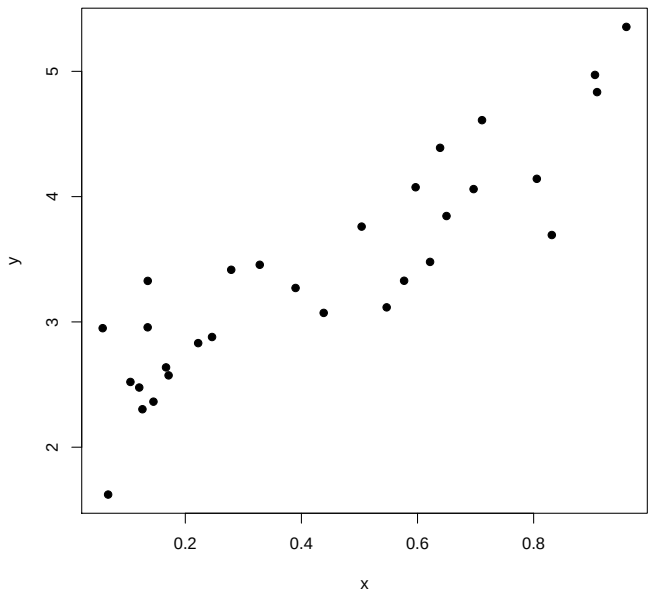
Estatura	134	138	142	146	150	154	158	162	166	170	174	178
VEF	1.7	1.9	2.0	2.1	2.2	2.5	2.7	3.0	3.1	3.4	3.8	3.9

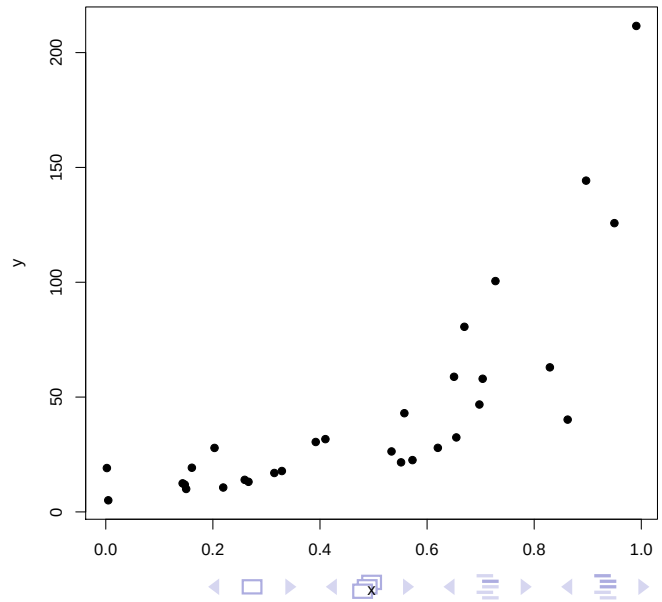


Algúns diagramas de dispersión









Coeficiente de correlación lineal

- A covarianza cambia se modificamos as unidades de medida das variables.
- Esto é un inconveniente porque non nos permite comparar a relación entre distintos pares de variables medidas en diferentes unidades.
- A solución é utilizar o **coeficiente de correlación lineal**

Coeficiente de correlación lineal entre X e Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

Coeficiente de correlación lineal

Coeficiente de correlación lineal entre X e Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

- A correlación lineal toma valores entre -1 e 1 e serve para investigar a relación lineal entre as variables.
- Se toma valores cercanos a -1 diremos que hai unha relación inversa entre X e Y.
- Se toma valores cercanos a $+1$ diremos que hai unha relación directa entre X e Y.
- Se toma valores cercanos a cero diremos que non existe relación lineal entre X e Y.

Exemplo. Volume Expiratorio Forzado e estatura: correlación

Para os datos do exemplo sobre o VEF e a estatura óbtense que:

- A desviación típica da estatura é $S_x = 13{,}808$ centímetros.
- A desviación típica do VEF é $S_y = 0{,}717$ litros.
- O coeficiente de correlación lineal entre X e Y será

$$r_{XY} = \frac{9{,}783}{13{,}808 \cdot 0{,}717} = 0{,}988$$

- A correlación é próxima a 1 e polo tanto a relación entre ambas variables é directa.

Modelo de regresión lineal

- O tipo de relación máis sinxela que se establece entre un par de variables é a **relación lineal** $Y = \beta_0 + \beta_1 X$
- Sen embargo, este modelo supón que unha vez determinados os valores dos parámetros β_0 e β_1 é posible predicir exactamente a resposta Y dado cualquer valor da variable de entrada X.
- Na práctica tal precisión case nunca é acadable, de modo que o máximo que se pode esperar é que a ecuación anterior sexa válida suxeita a un erro aleatorio, é dicir, a relación entre a **variable dependiente** (Y) e a **variable regresora** (X) articúlase mediante unha **recta de regresión**.

Modelo de regresión lineal simple

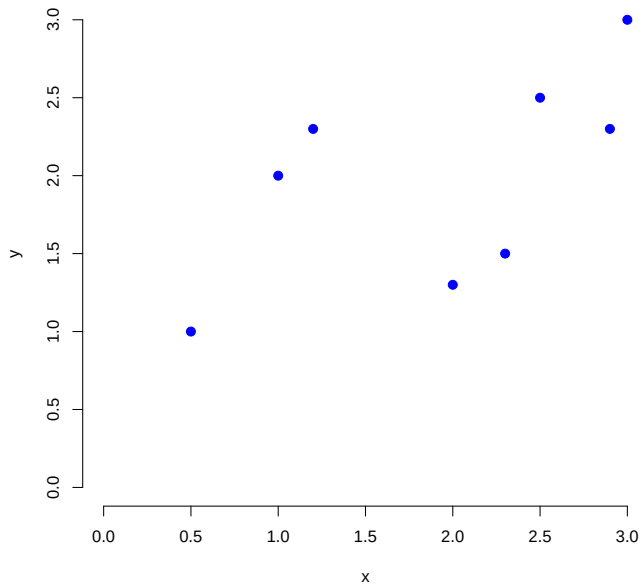
$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Modelo de regresión lineal

Modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

- Dada unha mostra $(x_1,y_1),\ldots,(x_n,y_n)$ da variable bidimensional (X, Y), ¿Cal é a recta que mellor axusta os datos?

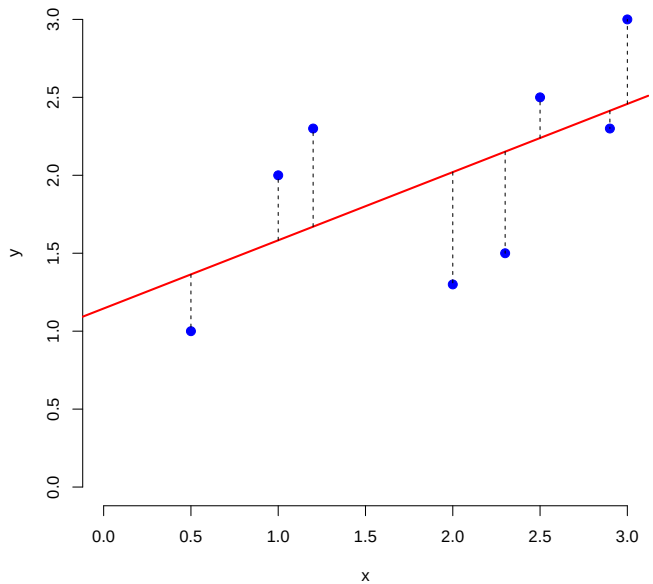


- O obxectivo é determinar os valores dos parámetros descoñecidos β_0 y β_1 (mediante estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$) de maneira que a recta definida axuste da mellor forma posible os datos.

O método de mínimos cadrados

Coeficientes estimados polo método de mínimos cadrados

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Recta de regresión de Y sobre X

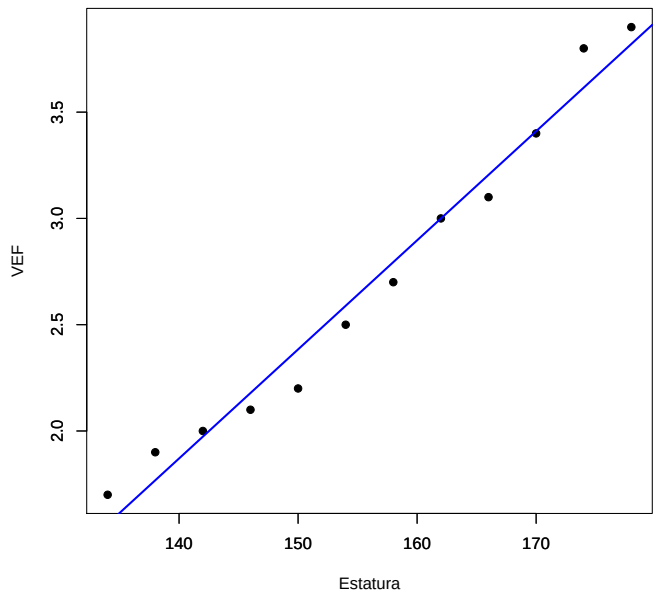
$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Exemplo. Volume Expiratorio Forzado e estatura: recta de regresión

Para os datos do exemplo sobre o VEF e a estatura obtense que:

- $\hat{\beta}_1 = \frac{9{,}783}{13{,}808^2} = 0{,}0513$
- $\hat{\beta}_0 = 2{,}691 - 156 \cdot 0{,}0513 = -5{,}312$
- A recta de regresión será entón

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = -5{,}312 + 0{,}0513x$$



Valores observados e previstos: erro na predición

- Obtida a recta de regresión de Y sobre X, podemos predicir as observacións de Y a partir das de X. Isto é,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

  para todo $i = 1, \ldots, n$.

- Dados os valores de X inicialmente considerados, os valores que se predín de Y ($\hat{y}$) non coinciden cos valores inicialmente observados.

- Polo tanto, existe un erro de predición:

$$\text{residuo} = \text{valor observado - valor de la recta} = y_i - \hat{y}_i$$

Descomposición da variabilidade

- A variabilidad de toda a mostra denomínase **variabilidade total (VT)**.

$$\text{VT} = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

- A variabilidade total descomponse en dous sumandos:
  - A variabilidad explicada (VE).

$$\text{VE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

  - A variabilidad non explicada (VNE) pola regresión.

$$\text{VNE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Descomposición da variabilidade

$$\text{VT} = \text{VE} + \text{VNE}.$$

Coeficiente de determinación

- O **coeficiente de determinación** (R^2) defínese como a proporción de variabilidade da variable dependente que é explicada pola regresión

Coeficiente de determinación

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{VNE}}{\text{VT}}.$$

- No modelo de regresión lineal simple, o coeficiente de determinación coincide co cadrado do coeficiente de correlación.

$$R^2 = r_{XY}^2$$

Exemplo. Volume Expiratorio Forzado e estatura: coeficiente de determinación

Para os datos do exemplo sobre o VEF e a estatura obtense que:

- $R^2 = 0{,}9881^2 = 0{,}976$
- Co modelo de regresión lineal simple, a variable X é capaz de explicar o 97,6 % da variación de Y.