

ESTATÍSTICA 2022/2023

Maria João Martins

(adaptado dos slides de 2021/22 da Prof^a Manuela Neves)

Slides de apoio às Aulas

Objetivos da Estatística Descritiva:

- condensar, sob a forma de tabelas, os dados observados;
- fazer a representação gráfica;
- calcular indicadores de localização e de dispersão.

Conceitos básicos em Estatística (definição e um exemplo):

- **população ou universo** → conjunto de todos os elementos que têm uma característica de interesse em comum (ex: todas as árvores de uma dada espécie)
- **unidades estatísticas** → são os elementos da população (ex: as árvores)
- **variável** → característica de interesse (ex: altura de árvores de uma espécie).
- **amostra** → subconjunto da população, efectivamente observado.

Ao(s) valor(es) da(s) característica(s) de interesse observadas nos elementos da amostra costuma chamar-se **dado(s)**.

Os **dados** podem ser de natureza:

- **quantitativa** → **discreta** (contagens: nº de peras em cada pereira, nº de machos por ninhada de coelhos) ou
→ **contínua** (medições: peso, comprimento, altura, tempo)
- **qualitativa** → **nominal** (cor dos olhos de um indivíduo, categoria taxonómica de uma espécie) ou
→ **ordinal** (avaliação numa escala de A (ótima) a E (péssima) da qualidade do almoço numa cantina)

Exemplo 1.

Num estudo para analisar a taxa de germinação de um certo tipo de cereal foram semeadas cinco sementes em cada um de 50 vasos iguais com o mesmo tipo de solo.

O número de sementes germinadas em cada vaso está registado a seguir:

1	0	1	2	1	3	2	0	0	1	4	0	2	1	0
2	4	1	2	0	3	5	3	0	2	1	3	3	0	4
0	2	5	3	0	2	5	1	1	0	4	4	1	2	1
0	5	1	2	3										

Neste caso os **dados são de natureza discreta, com um número pequeno de valores distintos.**

Dados deste tipo podem ser condensados numa tabela da forma

Dados de natureza discreta, com um número pequeno de valores distintos

x_i	n_i	f_i	F_i
0	12	0.24	0.24
1	12	0.24	0.48
2	10	0.20	0.68
3	7	0.14	0.82
4	5	0.10	0.92
5	4	0.08	1

n é a **dimensão da amostra** (n° de vasos), $n = 50$

$x_i \rightarrow n^\circ$ de sementes germinadas;

$n_i \rightarrow$ frequência absoluta, $\sum_i n_i = n$;

$f_i = \frac{n_i}{n} \rightarrow$ frequência relativa;

$F_i \rightarrow$ frequência relativa acumulada

Exemplo 2.

Um dos principais indicadores da poluição atmosférica nas grandes cidades é a concentração de ozono na atmosfera. Num dado Verão registou-se 78 valores dessa concentração (em $\mu\text{g} / \text{m}^3$), numa dada cidade:

3.5	6.2	3.0	3.1	5.1	6.0	7.6	7.4	3.7	2.8	3.4	3.5
1.4	5.7	1.7	4.4	6.2	4.4	3.8	5.5	4.4	2.5	11.7	4.1
6.8	9.4	1.1	6.6	3.1	4.7	4.5	5.8	4.7	3.7	6.6	6.7
2.4	6.8	7.5	5.4	5.8	5.6	4.2	5.9	3.0	3.3	4.1	3.9
6.8	6.6	5.8	5.6	4.7	6.0	5.4	1.6	6.0	9.4	6.6	6.1
5.5	2.5	3.4	5.3	5.7	5.8	6.5	1.4	1.4	5.3	3.7	8.1
2.0	6.2	5.6	4.0	7.6	4.7						

Agora estamos em presença de dados de **natureza contínua**

ED univariada | tabelas de frequências

Para dados de natureza contínua (ou dados de natureza discreta com um elevado número de valores distintos) elabora-se a **tabela de frequências** procedendo assim:

- Determina-se $\max(x_i)$ e $\min(x_i)$,
 $A_T = \max(x_i) - \min(x_i)$ é a **amplitude total**
- Escolhe-se um número m de subintervalos, as **classes**
- Define-se a amplitude cada classe. Em tabelas com classes de igual amplitude, será $h \approx A_T/m$
- Definem-se as classes
- Para cada classe calcula-se a **frequência absoluta**, n_i , e a **frequência relativa**, f_i .

O número m de classes pode ser escolhido através da

Regra de Sturges: m é o **o inteiro mais próximo de**

$$1 + (\log_2 n) = 1 + \frac{\log n}{\log 2}.$$

ED univariada | tabelas de frequências

Para o **Exemplo 2**: $\min(x_i) = 1.1$, $\max(x_i) = 11.7$, $A_T = 10.6$.

- pela regra de Sturges $m \approx 7.285 \rightarrow m = 7$
- a amplitude de cada classe é $h = 1.51 \rightarrow h = 1.5$
- a primeira classe começa em 1.0 (tem que conter $\min(x_i)$) e a última termina em 13.0 (tem que conter $\max(x_i)$)

Nota: com estas escolhas, será necessário considerar 8 classes

A **tabela de frequências** associada a estas escolhas é:

c_i	n_i	f_i	F_i
]1.0, 2.5]	10	0.128	0.128
]2.5, 4.0]	16	0.205	0.333
]4.0, 5.5]	18	0.231	0.564
]5.5, 7.0]	26	0.333	0.897
]7.0, 8.5]	5	0.064	0.962
]8.5, 10.0]	2	0.026	0.987
]10.0, 11.5]	0	0.00	0.987
]11.5, 13.0]	1	0.013	1

ED univariada | representação gráfica

- **Diagrama de barras** → para dados de natureza discreta com um número pequeno de valores distintos
- **Histograma** → para dados de natureza contínua, ou de natureza discreta com muitos valores distintos.

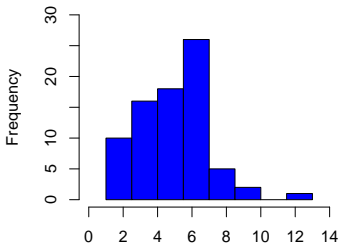
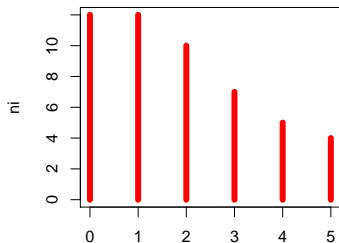



Diagrama de barras (exemplo 1) e **histograma** (exemplo 2) das frequências absolutas

Nota:

No caso de o agrupamento dos dados ser realizados com **classes de amplitude variável**, a **área** (e não a altura) do retângulo de cada classe deve ser **proporcional à frequência** dessa classe. Assim, a classe i é representada no histograma por um retângulo de largura h_i e altura f_i/h_i (também poderia ser n_i/h_i , mas há vantagem em ter as áreas dos retângulos a somar 1). O software  faz, por omissão, histogramas em que a

- altura é proporcional à frequência absoluta no caso de amplitudes constantes e
- altura é proporcional a frequência relativa / amplitude no caso de amplitudes variáveis.

As **tabelas e gráficos** constituem um primeiro conjunto de ferramentas usadas pela Estatística Descritiva para resumir e descrever um conjunto de dados

Outro conjunto de ferramentas que permite caracterizar um conjunto de dados é constituído pelos **indicadores numéricos** ou **indicadores amostrais**.

- **Indicadores de localização:** **média, mediana, quantis, moda**
- **Indicadores de dispersão:** **amplitude total, amplitude inter-quartis, variância, desvio padrão, coeficiente de variação.**

Somatórios | notação e propriedades

Seja $\{x_k\}_{k \in \mathbb{N}} = x_1, x_2, \dots$ uma sequência de números reais,

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n \quad \sum_{k=1}^n x_k = nx_k \quad \sum_{i=m}^n x_i = x_m + x_{m+1} + \dots + x_n.$$

Propriedades:

- $\sum_{i=m}^n (x_i \pm y_i) = \sum_{i=m}^n x_i \pm \sum_{i=m}^n y_i$
- $\sum_{i=m}^n \alpha x_i = \alpha \sum_{i=m}^n x_i, \alpha \in \mathbb{R}$

Exemplos:

- $\sum_{n=1}^3 (2n - 1) = (2 \times 1 - 1) + (2 \times 2 - 1) + (2 \times 3 - 1) = 1 + 3 + 5 = 9$
- $\sum_{n=1}^3 (2n - 1) = \sum_{n=1}^3 2n - \sum_{n=1}^3 1 = 2 \sum_{n=1}^3 n - 3 \times 1 = 2 \times (1 + 2 + 3) - 3 = 9$

Somatórios úteis em Estatística

- $\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n$
- $\sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$
- $\left(\sum_{i=1}^n x_i \right)^2 = (x_1 + x_2 + \cdots + x_n)^2$
- $\sum_{i=1}^n x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$
- $\left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n y_i \right) = (x_1 + x_2 + \cdots + x_n) \cdot (y_1 + y_2 + \cdots + y_n)$

Resolver o Exercício 1.9a) do Caderno de Exercícios para praticar.

ED univariada | média

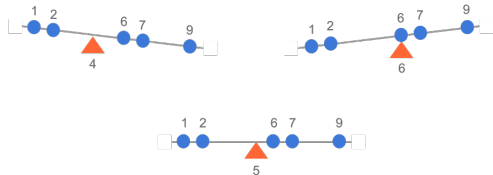
Considere-se (x_1, x_2, \dots, x_n) , uma amostra de n observações de x .

Definição

Chama-se **média aritmética**, **média empírica** ou simplesmente **média** e representa-se por \bar{x} a

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

\bar{x} é o centro de gravidade, ou ponto de equilíbrio, dos dados:



(ver <http://www.gastonsanchez.com/matrix4sl/>)

Propriedades da média

- A soma dos desvios de todas as observações relativamente à

média é zero:
$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

- Seja (x_1, x_2, \dots, x_n) uma amostra cuja média é \bar{x} e considere-se $y_i = a + bx_i$, $i = 1, \dots, n$ e $a, b \in \mathbb{R}$.

A amostra constituída pelas observações transformadas (y_1, y_2, \dots, y_n) tem média $\bar{y} = a + b\bar{x}$.

- Seja $(x_1^{(1)}, \dots, x_n^{(1)})$ uma primeira amostra de dimensão n , de média $\bar{x}^{(1)}$, e $(x_1^{(2)}, \dots, x_m^{(2)})$ uma segunda amostra de dimensão m da mesma variável, de média $\bar{x}^{(2)}$. A média das $n + m$

observações pode calcular-se como:
$$\frac{n \bar{x}^{(1)} + m \bar{x}^{(2)}}{n + m}.$$

Definição

A **mediana** é um valor que divide a amostra ordenada em duas partes com igual número de observações.

Dada a amostra (x_1, \dots, x_n) , seja $x_{(1)} \leq \dots \leq x_{(n)}$ a amostra **ordenada**. A **mediana** define-se por:

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{se } n \text{ ímpar} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{se } n \text{ par} \end{cases}$$

Se considerarmos a amostra ordenada dividida em quatro partes, cada uma com o mesmo número de observações, os pontos da divisão chamam-se **quantis empíricos** ou apenas **quantis** e costumam representar-se por Q_1 , Q_2 e Q_3 .

É claro que $Q_2 \equiv \tilde{x}$.

Definição – Generalização do conceito de quartil

Chama-se **quantil de ordem θ** , ($0 \leq \theta \leq 1$), o valor Q_{θ}^* tal que há pelo menos uma proporção θ de observações inferiores ou iguais a Q_{θ}^* e pelo menos uma proporção $(1 - \theta)$ de observações maiores ou iguais a esse valor.

Uma fórmula de cálculo pode ser

$$Q_{\theta}^* = \begin{cases} \frac{X_{(n\theta)} + X_{(n\theta+1)}}{2} & \text{se } n\theta \text{ inteiro} \\ X_{([n\theta]+1)} & \text{se } n\theta \text{ não inteiro} \end{cases}$$

onde $[n\theta]$ designa a parte inteira de $n\theta$.

Nota: $Q_{0.25}^* \equiv Q_1$; $Q_{0.5}^* \equiv Q_2 \equiv \tilde{x}$ e $Q_{0.75}^* \equiv Q_3$

Definição

A **moda, mo** , é a observação mais frequente (se existir).

Caso discreto → é a observação que tem maior frequência.

Caso contínuo → só faz sentido definir-se sobre dados agrupados →
é **um valor da classe que tem maior frequência**

Amplitude total

$$A_T = \max(x_i) - \min(x_i)$$

Amplitude inter-quartis

$$AIQ = Q_3 - Q_1$$

Variância

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

Outra fórmula de cálculo da variância: $s^2 = \frac{n \sum x_i^2 - (\sum x_i)^2}{n(n-1)}$

Desvio padrão

$$s_x = s = \sqrt{\text{variância}}$$

Propriedades

- $s_x^2 \geq 0$
- Sejam x_1, \dots, x_n , observações com variância s_x^2 . Considere-se $y_i = a + bx_i$, $i = 1, \dots, n$ e $a, b \in \mathbb{R}$. As observações transformadas têm como variância $s_y^2 = b^2 s_x^2$.
Para o **desvio padrão** tem-se $s_y = |b|s_x$.

ED univariada | coeficiente de variação

A variância e o desvio padrão medem dispersão relativamente à média. São medidas de dispersão absolutas, cujas unidades são o quadrado da unidade dos dados para a variância, ou a unidade dos dados para o desvio padrão.

Uma medida de dispersão relativa é o

Coeficiente de variação

$$CV = \frac{s}{\bar{x}} \times 100\%$$

O *CV* só se calcula quando as observações têm todas o mesmo sinal. É uma grandeza adimensional e permite comparar as dispersões de conjuntos de dados com diferentes unidades.

Nota:

Quando apenas se dispõe dos **dados agrupados em classes** (e não da totalidade dos dados) é possível obter valores aproximados de alguns indicadores de localização e dispersão. Ver por exemplo Slides 17, 18, 19 e 21 em https://fenix.isa.ulisboa.pt/downloadFile/563022967881513/slides_estat_desc_21_22.pdf.

Um modo gráfico que permite facilmente visualizar a localização e a dispersão de um conjunto de dados, efectuando em simultâneo a sua síntese → **o diagrama de extremos e quartis**.

Se nesse gráfico identificarmos as observações que se afastam do padrão geral dos dados (candidatos a *outliers*) é hábito designá-lo por **caixa de bigodes** (*boxplot*).

Existem vários critérios para classificar uma observação como **um outlier**.

Definição

Um valor x_i é um candidato a **outlier** se

$$x_i < B_I \quad \text{ou} \quad x_i > B_S$$

sendo B_I **barreira inferior** e B_S **barreira superior** definidas como:

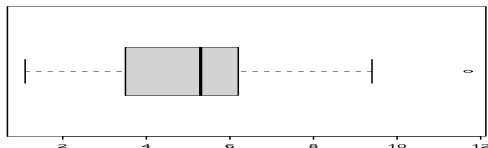
$$B_I = Q_1 - 1.5(Q_3 - Q_1) \quad B_S = Q_3 + 1.5(Q_3 - Q_1)$$

ED univariada | caixa de bigodes

Para desenhar **uma caixa de bigodes**


- Marcar **o valor adjacente inferior** → é o **menor** valor do conjunto dos dados (podendo ser o *mínimo*) maior ou igual à barreira inferior;
- Marcar **o valor adjacente superior** → é o **maior** valor do conjunto dos dados (podendo ser o *máximo*) menor ou igual à barreira superior;
- Marcar **a mediana, primeiro e terceiro quartis** (que vão permitir desenhar uma “caixa”) e marcar os candidatos a *outlier*.

Caixa de bigodes referente os dados do **Exemplo 2**.



Quando se pretende comparar várias amostras, o recurso a caixas de bigodes paralelas é uma ferramenta muito útil.

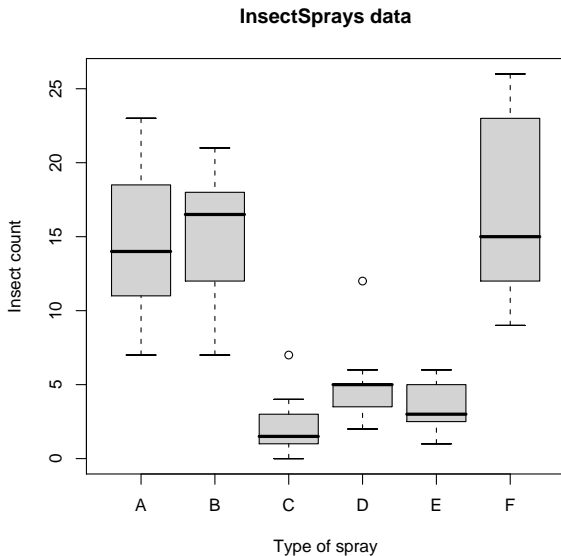
Exemplo 3.

As seguintes caixas de bigodes referem-se a um conjunto de dados `InsectSprays` disponíveis no *package* `datasets` do . São contagens de insectos em 6 unidades agrícolas experimentais, às quais foram aplicados diferentes tipos de insecticidas.

Fonte: Beall, G., (1942) The Transformation of data from entomological field experiments, *Biometrika*, 29, 243-262.

(stat.ethz.ch/R-manual/R-devel/library/datasets/html/InsectSprays.html)

ED univariada | caixas de bigodes paralelas



Estatística descritiva bivariada

Nas aulas anteriores, em cada unidade estatística, estudámos **uma única característica**. Muitas vezes, porém, interessa averiguar a existência de alguma **relação entre duas características** (variáveis) e descrever essa relação.

Exemplos: peso e altura de recém nascidos em Portugal; comprimento e largura das folhas de uma espécie vegetal.

Não são relações determinísticas que interessam à Estatística, mas é o comportamento em média (**relação estatística**) das duas características.

Se duas variáveis estão ligadas por uma **relação estatística** diz-se haver **correlação** entre elas.

Correlação **positiva** se as duas características variam no mesmo sentido e **negativa** caso contrário.

Sejam $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ observações efectuadas em n unidades estatísticas.

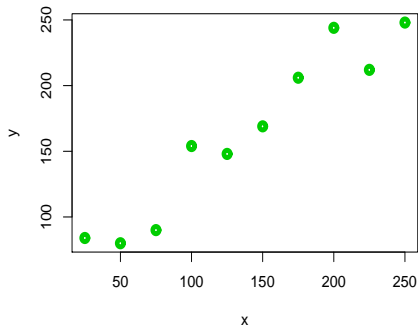
Num **diagrama de dispersão** ou **nuvem de pontos**, cada observação i é representada por 1 ponto num sistema de dois eixos cartesianos. A abcissa desse ponto é x_i e a ordenada é y_i .

Exemplo 4.

Pretende-se estudar o efeito da aplicação de diferentes quantidades de um dado fertilizante (x) na produção de relva (y). A relva é semeada uniformemente numa dada área na qual são marcados ao acaso 10 talhões de 1 m^2 , a cada um dos quais é aplicada uma certa quantidade de fertilizante. A relva é depois cortada, seca e pesada. Os dados obtidos e a **nuvem de pontos** correspondente são:

ED bivariada | nuvem de pontos

x (g/m ²)	y (g/m ²)
25	84
50	80
75	90
100	154
125	148
150	169
175	206
200	244
225	212
250	248



Médias marginais de x e y , respetivamente, são

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$(\bar{x}, \bar{y}) \rightarrow$ centro de gravidade da nuvem de pontos.

Dispersões marginais de x e y , respetivamente

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

Mas... há uma medida que dá **informação sobre a relação entre as duas variáveis**.

Definição

Dadas as variáveis x e y , chama-se **covariância de x e y** a

$$\mathit{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

Uma fórmula de cálculo alternativa é:

$$\mathit{cov}(x, y) = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n(n - 1)}$$

(demonstrar como exercício).

Propriedades da covariância

1. Seja (x_i, y_i) um conjunto de n observações e considere-se:
 $x'_i = a + bx_i$ e $y'_i = c + dy_i$, com $i = 1, \dots, n$ e $a, b, c, d \in \mathbb{R}$.
Então $cov(x', y') = bd cov(x, y)$.
2. $|cov(x, y)| \leq s_x s_y$. A igualdade só se verifica quando todos os pontos observados se encontram sobre uma reta.

Nota

Importância da covariância:

$cov(x, y) > 0 \rightarrow$ há correlação positiva

$cov(x, y) < 0 \rightarrow$ há correlação negativa.

Desvantagem da covariância:

é fortemente afectada por mudanças de escala nas observações (ver propriedade 1.)

Definição

O **coeficiente de correlação** é definido como

$$r = r_{x,y} = \frac{\text{COV}(x, y)}{s_x s_y} \quad \text{com } s_x \neq 0 \text{ e } s_y \neq 0$$

Propriedades

1. r tem sempre o mesmo sinal da covariância;
2. $-1 \leq r \leq 1$ (se $|r_{xy}| = 1$ todos os valores observados se encontram sobre uma reta);
3. Se (x, y) têm coeficiente de correlação r_{xy} , $x'_i = a + bx_i$ e $y'_i = c + dy_i$ (com $bd \neq 0$), tem-se:
 $r_{x'y'} = r_{xy}$ (se $bd > 0$) e $r_{x'y'} = -r_{xy}$ (se $bd < 0$).

Nota:

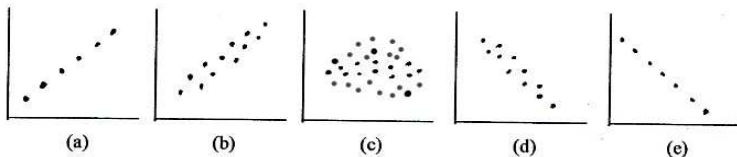
O coeficiente de correlação é uma grandeza **adimensional**.

Não é afectado, em valor absoluto, por transformações afins.

ED bivariada | coeficiente de correlação

- (a) $r = 1$ todos os pontos observados se encontram sobre uma reta de declive positivo.
- (b) $r \simeq 1$ todos os pontos observados se encontram próximos de uma reta de declive positivo.
- (c) $r \simeq 0$ a nuvem apresenta um aspecto arredondado ou alongado segundo um dos eixos.
- (d) $r \simeq -1$ todos os pontos observados se encontram próximos de uma reta de declive negativo.
- (e) $r = -1$ todos os pontos observados se encontram sobre uma reta de declive negativo.

Nota: O coeficiente de correlação mede *a nitidez da ligação* existente entre duas variáveis, quando essa ligação *é linear ou aproximadamente linear*



Regressão linear simples

Se $|r| \simeq 1$ e a nuvem de pontos sugere a existência de uma relação linear entre os valores observados, faz sentido determinar a equação de uma reta que possa traduzir bem a relação observada, i.e., pretende-se determinar

$y = b_0 + b_1x$ → **reta de regressão**, que permita:

- descrever a relação entre y (variável resposta ou dependente) e x (variável explicativa, regressora ou independente);
- prever um valor de y para um dado valor de x .

Mas ... a equação $y = b_0 + b_1x$ não é verificada para todos os pares (x_i, y_i) (note-se que só o seria se $|cov(x, y)| = s_x s_y$, ou seja se $|r| = 1$).

Regressão linear simples

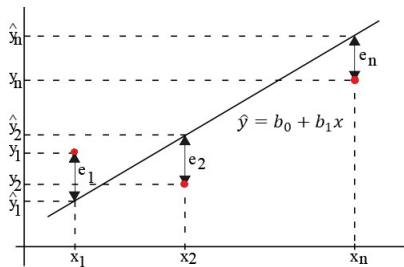
Para cada par (x_i, y_i) tem-se $y_i = b_0 + b_1 x_i + e_i$

$\hat{y}_i = b_0 + b_1 x_i$ é o **valor estimado** de y para x_i , pela reta de regressão.

Então pode-se escrever $y_i = \hat{y}_i + e_i$

$e_i = y_i - \hat{y}_i$ é o **resíduo** da observação i .

Obter a reta \iff determinar b_0 (ordenada na origem) e b_1 (declive).



(adaptado de <https://rce.casadasciencias.org/rceapp/art/2019/045/>)

Regressão linear simples

No **método dos mínimos quadrados** \rightarrow b_0 e b_1 são determinados de modo a

Minimizar a soma dos quadrados dos resíduos ou seja, minimizar

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = Q(b_0, b_1)$$

Pretende-se então determinar os minimizantes de uma função de duas variáveis. As condições de estacionaridade são:

$$\begin{cases} \frac{\partial Q}{\partial b_0} = 0 \\ \frac{\partial Q}{\partial b_1} = 0 \end{cases} \Leftrightarrow \begin{cases} 2 \sum (y_i - b_0 - b_1 x_i) = 0 \\ 2 \sum x_i (y_i - b_0 - b_1 x_i) = 0 \end{cases}$$

A estas equações chama-se **equações normais**

Regressão linear simples

Algumas conclusões podem ser tiradas destas equações:

- $\sum(y_i - b_0 - b_1 x_i) = 0 \Rightarrow \sum(y_i - \hat{y}_i) = \sum e_i = 0$ a soma dos resíduos é nula.
- $\sum(y_i - \hat{y}_i) = 0 \Rightarrow \bar{y} = \bar{\hat{y}}$ a média dos valores observados é igual à média dos valores estimados.
- a reta de regressão passa no ponto (\bar{x}, \bar{y}) .

Regressão linear simples

- **Solução do sistema**

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{s_x^2} = r \frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

A b_1 chama-se **coeficiente de regressão de y sobre x** .

Observações:

- b_1 tem o mesmo sinal que $\text{cov}(x, y)$ e r .

- Dado x_i e sendo $x'_i = x_i + 1$ tem-se

$$\hat{y}_i = b_0 + b_1 x_i \quad \hat{y}'_i = b_0 + b_1 (x_i + 1).$$

$b_1 = \hat{y}'_i - \hat{y}_i, \rightarrow b_1$ representa a **variação esperada para y quando x aumenta uma unidade.**

Um dos objectivos da reta de regressão é o de **predizer** o valor de uma variável conhecido o valor assumido pela outra **mas** é necessário avaliar o **grau de precisão** atingido pelas estimativas.

O método dos mínimos quadrados permite uma importante decomposição de $\sum (y_i - \bar{y})^2$.

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

cujas parcelas se costumam representar por

$$SQ_T = SQ_{RE} + SQ_R$$

Soma dos Quadrados Totais =

Soma dos Quadrados devidos aos REsíduos +

Soma dos Quadrados devidos à Regressão.

Definição

O coeficiente de determinação da reta de regressão, definido por

$R^2 = \frac{SQ_R}{SQ_T} = \frac{s_{\hat{y}}^2}{s_y^2}$ é uma medida da **precisão da reta de regressão**.

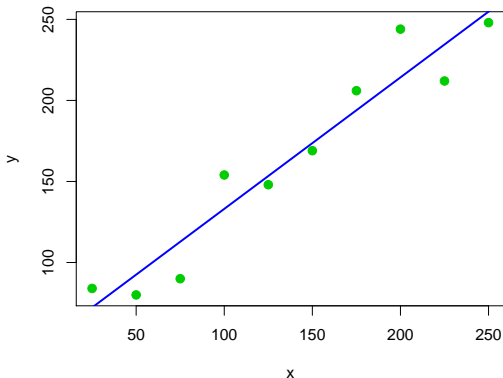
R^2 é a **fração de variabilidade dos valores observados da variável resposta que é explicada pela regressão**

No contexto da **regressão linear simples**:

$$R^2 = \frac{SQ_R}{SQ_T} = \frac{b_1^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} = \frac{b_1^2 s_x^2}{s_y^2} = \frac{r^2 s_y^2}{s_x^2} \frac{s_x^2}{s_y^2} = r^2,$$

ou seja, o coeficiente de determinação da reta é o quadrado do coeficiente de correlação (linear) entre as duas variáveis.

Exemplo 4



$$\bar{x} = 137.5 \text{ g/m}^2$$

$$\bar{y} = 163.5 \text{ g/m}^2$$

$$s_x^2 = 5729.167 \text{ (g/m}^2\text{)}^2$$

$$s_y^2 = 4092.722 \text{ (g/m}^2\text{)}^2$$

$$\text{cov}_{xy} = 4648.611 \text{ (g/m}^2\text{)}^2$$

A nuvem de pontos mostra a existência de uma tendência linear de fundo entre a produção de relva e a quantidade de fertilizante. O coeficiente de correlação é $r = 0.96$, indicando que essa tendência linear é forte. A reta de regressão ajustada (representada no gráfico sobre a nuvem de pontos) tem a equação $y = 51.933 + 0.8114x$. De acordo com este modelo, por cada g/m^2 a mais na quantidade de fertilizante, a produção de relva aumenta, em média, $0.8114 \text{ g}/\text{m}^2$. A reta explica 92.16% da variabilidade dos valores observados para a produção de relva; trata-se de uma reta com precisão elevada.

(Obter os valores assinalados a azul como exercício).