Capítulo 2

Estatística Descritiva

AMG, JFO (v8 – 2017)
adaptado de: *Estatística*,
Rui Campos Guimarães,
José A. Sarsfield Cabral

Slide 2.-1

Conteúdo

Slide 2.0

Resultados de aprendizagem

- Distinguir os diferentes tipos de escalas em que dados podem ser expressos
- Classificar e sintetizar manualmente pequenos conjuntos de dados amostrais
- Calcular, manualmente ou com recurso a calculadoras, estatísticas de pequenos conjuntos de dados amostrais
- Utilizar folhas de cálculo e software estatístico para classificar e sintetizar dados amostrais (representação tabular e gráfica, cálculo de estatísticas)
- Distinguir os vários tipos de estatísticas: de localização, de dispersão, de forma e de associação
- Interpretar correctamente representações tabulares e gráficas de dados amostrais, assim como os resultados de estatísticas

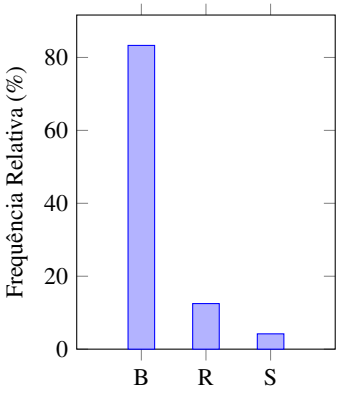
Slide 2.1

2.1 Introdução

Objectivo da Estatística Descritiva
Classificar e *sintetizar* a informação contida em conjuntos de dados (normalmente uma amostra e apresentados de forma desorganizada)

B	B	B	B	B	B	B	S	B	B	B	B	B	B	B	B	B	B
B	B	B	B	S	B	B	B	B	B	B	S	R	B	B	B	R	B	B
B	R	B	B	B	R	B	B	B	B	B	B	R	B	B	B	B	B	B
B	B	R	B	R	R	B	R	B	B	B	S	R	B	B	B	R	B	B
B	B	B	R	B	B	B	B	B	B	B	B	S	R	B	B	B	B
B	B	B	B	B	B	B	B	B	R	B	B	B	B	B	B	R	B	B

Categoria de Peças	Frequência Absoluta	Frequência Relativa
Sem defeito (B)	100	83.3 %
Recuperáveis (R)	15	12.5 %
Sucata (S)	5	4.2 %
Total	**120**	**100** %



Slide 2.2

Características importantes

Localização – Valor central ou representativo que indica aonde o conjunto de dados está centrado

Dispersão – Medida da variabilidade dos dados em relação ao seu valor central

Forma – Forma como os dados se distribuem pela respectiva gama de valores (uniforme, em forma de *sino*, de forma simétrica, . . .)

Valores extremos (*outliers*) – Valores da amostra que se encontram muito afastados da maioria dos restantes dados amostrais (podem ser resultado de erros de medida)

Tempo – Alteração das características dos dados ao longo do tempo (característica a estudar apenas em Estatística Multivariada e sem grande profundidade)

Slide 2.3

2.2 Tipos de Dados e Escalas

Classificação de dados segundo a escala em que são expressos

Dados qualitativos (ou categóricos)

Escala nominal – Dados são identificados apenas pela atribuição de um nome que designa uma classe (ou categoria) (classes: exaustivas, mutuamente exclusivas e não ordenáveis)

Escala ordinal – Idêntica à escala nominal, mas em que é possível estabelecer uma ordenação das classes, segundo um critério relevante

Dados quantitativos (ou numéricos)

Escala de intervalo – Os dados são diferenciados e ordenados por números expressos numa escala cuja origem é arbitrária (a diferença entre esses números tem significado, mas o quociente não)

Escala absoluta – Tem uma origem fixa (além da diferença, o quociente entre os dados também tem significado)

Dados quantitativos podem ser:

- *Discretos* – resultantes de contagens
- *Contínuos* – resultantes de medições

Slide 2.4

Exemplos

- Classif. de pessoas pela cor do cabelo: preto, castanho, branco, loiro, . . .
- Classificação dos consumidores de bens de primeira necessidade pelo género: feminino ou masculino

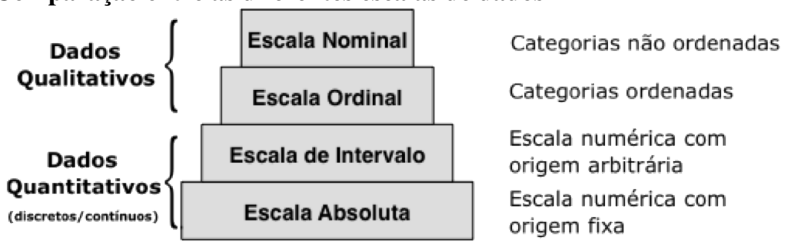
- Classificações obtidas pelos alunos num teste de Estatística: mau, medíocre, suficiente, bom ou muito bom
- Classificação dos clientes segundo o volume de encomendas que colocam: clientes A (muito importantes), B (importantes) e C (pouco importantes)

- Temperaturas registadas, em ^oC, às 8 horas de dias sucessivos
  (note-se que não faz sentido dizer que em dias consecutivos a temperatura duplicou, p.e. passou de 5^oC para 10^oC, ... porquê?)

- Peso de pessoas, expressos em kg
- Volumes de investimento, espressos em milhares de euros
- Resultados de 150 lançamentos de um dado

Slide 2.5

Comparação entre as diferentes escalas de dados

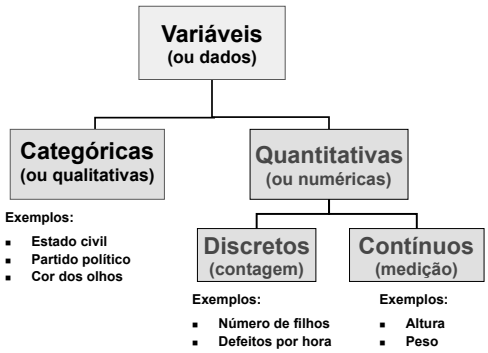


- escalas apresentam um grau crescente de informação ($\downarrow$)
- dados numa escala podem ser transformados em escalas precedentes ($\uparrow$)

Exemplo
Se se dispuser das temperaturas máximas diárias, em ^oC, registadas na cidade de Faro ao longo último ano, é possível reclassificá-las nas seguintes classes:

- dias frios ($< 18^oC$), dias amenos ($[18^oC, 25^oC]$) e dias quentes ($> 25^oC$)

Slide 2.6



Slide 2.7

2.3 Caracterização e Representação de Dados Categóricos

Amostra univariada
Quando cada um dos dados que a integram mede, numa escala qualquer, apenas um atributo

Dados categóricos (ou qualitativos) – distribuem-se por um conjunto de diferentes classes ou categorias

- representação tabular ou gráfica da distribuição de frequências (ou empírica ou dos dados amostrais)

Dados quantitativos – dados discretos ou contínuos

- representação tabular ou gráfica da distribuição de frequências
- cálculo de estatísticas (ou medidas) amostrais
- representação gráfica de estatísticas

Slide 2.8

Dados Categóricos (ou Qualitativos)

- Caracterização de dados qualitativos: *tabelas de frequências*, *diagramas de barras* e diagramas circulares
- *Objectivo*: representar a forma como os dados se distribuem por um conjunto de diferentes categorias

Frequência absoluta da categoria k (N_k) – número de dados contidos na categoria k
Número total de dados: $N = \sum_{k=1}^{K} N_k$

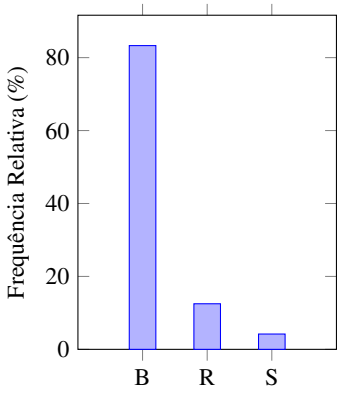
Frequência relativa da categoria k (f_k) – número de dados contidos na categoria k, expresso como uma proporção do número total de dados: $f_k = \frac{N_k}{N}$

Slide 2.9

Tabela de frequências

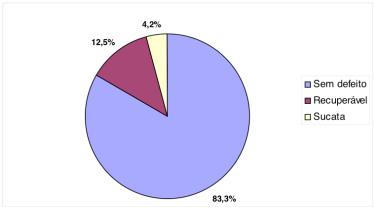
Categoria de Peças	Frequência Absoluta	Frequência Relativa
Sem defeito (B)	100	83.3 %
Recuperáveis (R)	15	12.5 %
Sucata (S)	5	4.2 %
Total	**120**	**100** %

Diagrama de barras de frequências relativas



Slide 2.10

Diagrama circular de frequências relativas



2.4 Caracterização e Representação de Dados Quantitativos

Descrição de dados quantitativos

- *Representações tabulares e gráficas*
  - tabela de frequências
  - histograma de frequências
  - polígono de frequências
  - polígono de frequências acumuladas
- *Estatísticas*
  - de localização: média, mediana, moda
  - de dispersão: amplitude, intervalo interquartis, desvio médio, desvio quadrático médio, variância, desvio-padrão
  - momentos de ordem i
- *Representação gráfica de estatísticas*
  - diagramas tipo caixa (*boxplots*)

Slide 2.11

2.4.1 Representações Tabulares e Gráficas

Representação de dados quantitativos contínuos

Peso (em gramas) do conteúdo de 100 garrafas

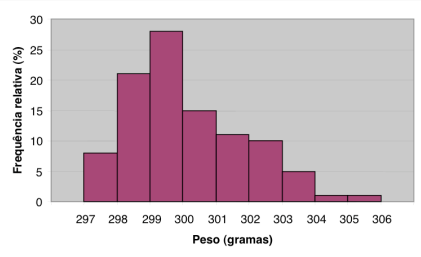
302.25	299.20	300.24	297.72	301.10	299.70	302.17	297.47
298.35	303.76	298.65	299.38	301.11	298.86	298.61	304.22
300.36	299.19	300.86	299.83	299.68	299.75	301.99	302.85
302.52	300.12	301.81	297.99	298.49	299.60	299.26	302.45
299.23	298.73	303.07	299.07	303.66	302.10	298.35	302.83
297.83	299.75	299.55	298.76	299.95	299.76	299.13	299.07
301.85	299.83	300.52	298.98	298.72	303.49	298.00	305.13
300.73	300.14	298.34	299.07	298.56	301.05	302.39	297.84
299.62	299.12	298.11	299.48	299.70	301.35	298.85	300.20
302.45	302.04	300.83	298.94	298.34	299.70	298.36	301.59
298.07	298.29	303.91	298.16	297.59	300.96	299.08	300.02
299.57	300.84	299.18	301.33	297.87	301.63	300.48	299.16
298.23	302.00	300.86	299.80				

Como representar este conjunto de dados?
Criar *classes* (também denominadas categorias ou células) e contar . . . Slide 2.12

Tabela de frequências (relativas)

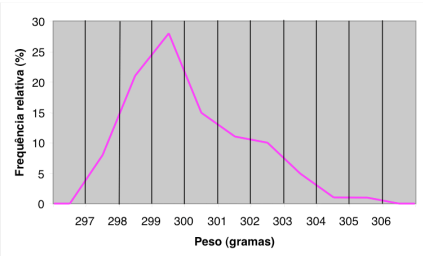
k	Limites	$f'_k(\%)$.	$fa'_k(\%)$.
1	$[297,0^+;298,0]$	8	8
2	$[298,0^+;299,0]$	21	29
3	$[299,0^+;300,0]$	28	57
4	$[300,0^+;301,0]$	15	72
5	$[301,0^+;302,0]$	11	83
6	$[302,0^+;303,0]$	10	93
7	$[303,0^+;304,0]$	5	98
8	$[304,0^+;305,0]$	1	99
9	$[305,0^+;306,0]$	1	100
Total		100	100

Histograma de freq. relativas

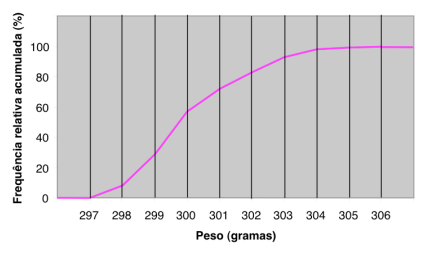


Slide 2.13

Polígono de frequências relativas



Polígono de freq. relativas acum.



Em quantas classes agrupar os dados?

Número de classes

Não existe uma regra que permita definir inequivocamente o número de classes, existem sim algumas *regras práticas* e *bom senso*

- Regra do livro recomendado:

$$\text{Número de classes } (K) \approx \sqrt{\text{Número de dados}(N)}$$

- Número de classes não deve ser muito elevado sob pena de se perderem os benefícios da *síntese*, nem muito pequeno sob pena de se perder *detalhe* na representação

$$\text{Número de classes } (K) \text{ entre 5 e 20 classes}$$

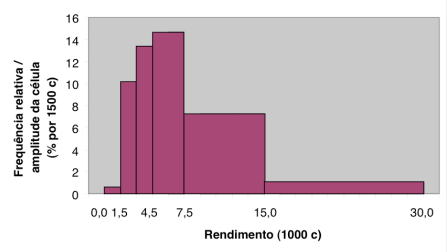
- Deve-se experimentar com diferentes números de classes e com diferentes limites para as classes (existem vantagens óbvias na utilização de números redondos)

(http://www.shodor.org/interactivate/activities/Histogram/) Slide 2.14

Classes de amplitude variável

- adequadas a distribuições de frequência com concentrações elevadas em alguns valores
- permitem homogeneizar o número de elementos em cada classe
- altura de cada barra do histograma (h) é proporcional à frequência por unidade de amplitude da classe

Rendimentos colectáveis (contos)			Frequências relativas (%)
0^+	–	1500	0,6
1500^+	–	3000	10,2
3000^+	–	4500	13,4
4500^+	–	7500	29,3
7500^+	–	15000	36,4
15000^+	–	30000	10,1
Total			100,0



Exemplo (classe 15000^+– 30000): $h = \frac{\frac{10.1}{30000-15000}}{1500} = \frac{10.1}{10} = 1.01$

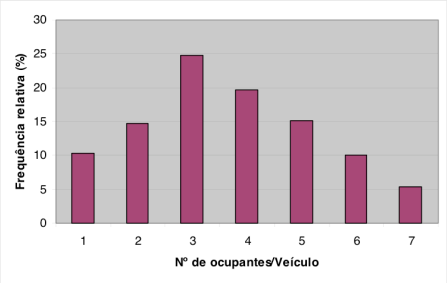
Unidade de amplitude: 1500 (= amplitude da célula de menor amplitude)

Slide 2.15

Representação de dados quantitativos discretos

- Conjunto de valores que os dados tomam é limitado (ou reduzido)
- Não é necessário definir classes (um valor $\longrightarrow$ uma classe)
- Representação por diagramas de barras

Nº de ocupantes por veículo	Frequência absoluta	Frequência relativa (%)
1	103	10,3
2	147	14,7
3	248	24,8
4	197	19,7
5	152	15,2
6	100	10,0
7	53	5,3
Total	1000	100,0



E se o conjunto de valores que os dados tomam for elevado?

$\Rightarrow$ proceder como nos dados contínuos (criar classes e usar histogramas)

Slide 2.16

2.4.2 Estatísticas

- São *medidas* calculadas com base nos dados amostrais
- Descrevem globalmente o conjunto de valores que tais dados tomam
- *Objectivo*: traduzir em números o que se apreende da observação de uma tabela de frequências ou de um histograma

- Estatísticas de *localização*
- Estatísticas de *dispersão*
- Outras estatísticas: *momentos* (ordinários e centrados), coeficientes de *assimetria* e de *achatamento*

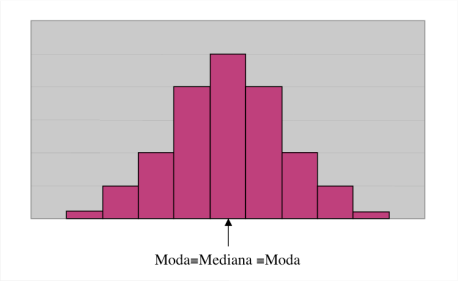
Notas

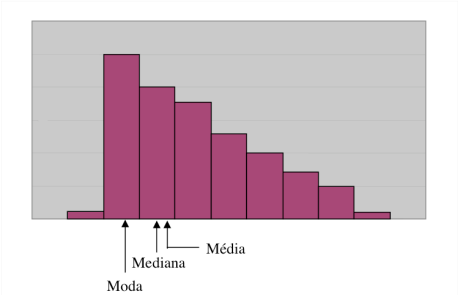
- Só se calculam estatísticas para dados quantitativos
- Dependendo do tipo de dados (discretos ou contínuos) e de como são apresentados (não agrupados, *i.e.* em bruto, ou agrupados em classes) existem fórmulas apropriadas para o cálculo de cada estatística

Slide 2.17

- Pretendem indicar onde se concentram os dados de uma distribuição de frequências ou histograma

  $\longrightarrow$ *tendência central* ou *valor central*
- Principais estatísticas de localização: *Média*, *Mediana* e *Moda*





Slide 2.18

Estatísticas de localização – Média amostral ($\bar{x}$)

Média amostral ($\bar{x}$)

- Valor *central* em relação aos dados
  (ponto de equilíbrio da distribuição de frequências ou histograma)

$\rightarrow$ média aritmética dos dados

Dados não agrupados (dados em bruto): $\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$

Dados discretos (valores x_k): $\bar{x} = \sum_{k=1}^{K} f_k \cdot x_k$

Dados contínuos agrupados (em classes): $\bar{x} = \sum_{k=1}^{K} f_k \cdot M_k$

Notas

- f_k: frequência relativa, M_k: o ponto central da classe k
- dados agrupados apenas permitem cálculo aproximado, porquê?
- outras médias: geométrica, harmónica

Slide 2.19

Propriedades da média amostral

- A média amostral toma um *valor central* em relação aos dados que constituem a amostra:

$$\sum_{n=1}^{N}(x_n - \bar{x}) = \sum_{n=1}^{N} x_n - N\bar{x} = N\bar{x} - N\bar{x} = 0$$

- O valor da constante C, que *minimiza a soma dos desvios quadráticos* relativamente a C, é a média amostral:

$$\frac{\mathrm{d}}{\mathrm{d}C}\sum_{n=1}^{N}(x_n - C)^2 = -2\sum_{n=1}^{N}(x_n - C) = 0 \Leftrightarrow C = \bar{x}$$

- Como a segunda derivada é sempre positiva, este valor minimiza de facto a soma dos desvios quadráticos:

$$\frac{\mathrm{d}^2}{\mathrm{d}C^2}\sum_{n=1}^{N}(x_n - C)^2 = \frac{\mathrm{d}^2}{\mathrm{d}C^2}\left[-2\sum_{n=1}^{N}(x_n - C)\right] = 2N > 0, \forall_C$$

Slide 2.20

Estat. de localização – Mediana amostral (*Med*)

- Valor tal que cerca de metade dos dados são maiores a esse valor e os restantes são menores
- Medida mais adequada para representar o valor central em distribuições assimétricas com valores extremos atípicos (*outliers*)

Vector de dados $(x_1, x_2, \ldots, x_N)$ ordenados por ordem crescente (ou decrescente), a *mediana amostral* (*Med*) vem

- para N ímpar: $Med = x_{(N+1)/2}$
- para N par: $Med = \frac{x_{N/2} + x_{N/2+1}}{2}$

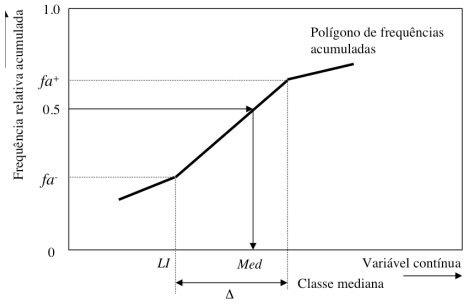
Propriedade da mediana amostral
A mediana é o valor que minimiza a soma dos desvios absolutos:

$$MIN \sum_{n=1}^{N} |x_n - C| \Rightarrow C = Med$$

Slide 2.21

Mediana para dados contínuos agrupados

$$Med = LI + \frac{0.5 - fa^-}{fa^+ - fa^-} \times \Delta = LI + \frac{0.5 - fa^-}{f} \times \Delta$$



- Classe mediana – classe onde a frequência acumulada ultrapassa 0.5
- Assume que as observações contidas na classe mediana se distribuem regularmente ao longo dessa mesma classe ⇒ *cálculo aproximado*

Slide 2.22

Estat. de localização – Moda amostral (*Mod*)

- Valor ou gama de valores nos quais a concentração de dados é máxima
- Uma amostra pode ter mais do que uma moda
  (amostras *unimodais versus amostras multimodais*)

Dados discretos – Valor dos dados mais frequente

Se houver dois ou mais valores adjacentes para os quais a frequência seja máxima, a moda será a média desses valores

Dados contínuos – Ponto central da classe modal (classe com maior frequência) ou do conjunto de classes modais, se estas forem adjacentes

ou ...

Slide 2.23

Dados contínuos (cont.) – ter em atenção as frequências (absolutas ou relativas) da classe modal (N_{Mod} ou f_{Mod}) e das classes que lhe são adjacentes (N_1 ou f_1 para a célula à esquerda e N_2 ou f_2 para a célula à direita)

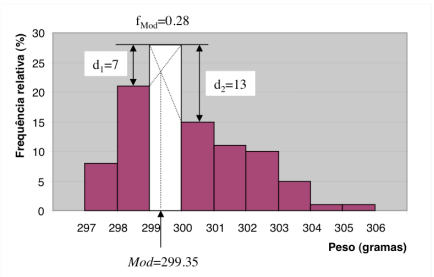
$$Mod = LI + \frac{d_1}{d_1 + d_2} \times \Delta$$

LI – limite inferior da classe modal

Δ – amplitude da classe modal

d_1: $(N_{Mod} - N_1)$ ou $(f_{Mod} - f_1)$

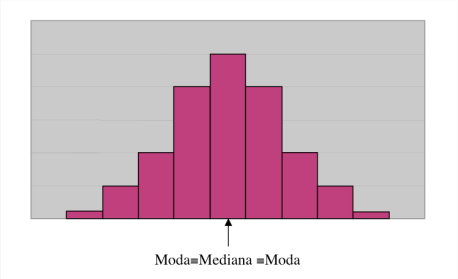
d_2: $(N_{Mod} - N_2)$ ou $(f_{Mod} - f_2)$



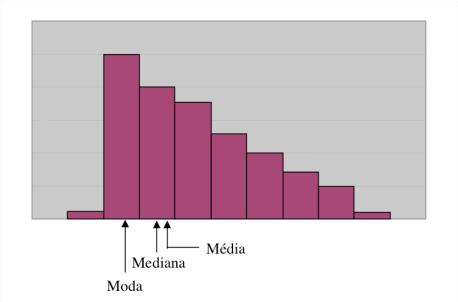
Slide 2.24

Comparação entre média, mediana e moda

Histograma simétrico



Histograma assimétrico à direita



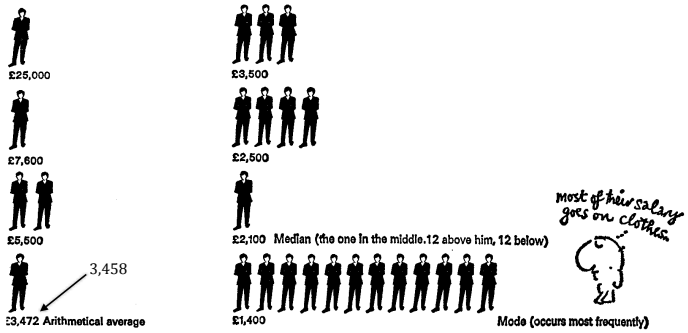
Histograma simétrico: $\bar{x} = Med = Mod$

Histograma assimétrico unimodal à direita: $\bar{x} > Med > Mod$ *

* Trata-se de uma regra prática que nem sempre se verifica (ver detalhes em `http://www.amstat.org/publications/jse/v13n2/vonhippel.html`)

Slide 2.25

Que medida de localização utilizar?



(Darrell Huff, "How to Lie with Statistics", Penguin Books, 1973)

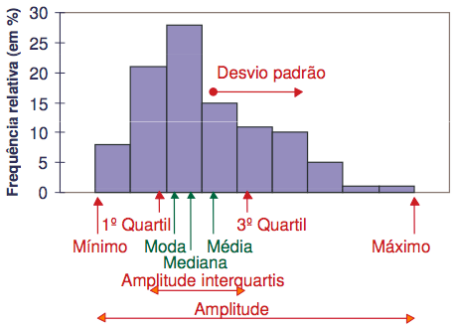
Moda: não é muito útil para dados contínuos não-agrupados (pode haver mais do que uma moda ou mesmo nenhuma)

Mediana: pouco afectada por valores extremos (utilizada como medida quando se pretende caracterizar por exemplo o rendimento da população)

Média: de longe a medida de localização mais utilizada, embora seja afectada por valores extremos (popularidade: por ser afectada por todos os valores e propriedades matemáticas atractivas para a inferência estatística)

Slide 2.26

- Pretendem caracterizar a *variabilidade* de uma distribuição de frequências ou histograma
- Principais estatísticas de dispersão: *Amplitude*, *Amplitude e intervalo interquartis*, Desvio absoluto médio, Desvio quadrático médio, *Variância*, *Desvio padrão*, *Quartis* e *Percentis*



(Nóvoa, 2008)

Slide 2.27

Estat. dispersão – Amplitude, Intervalo e Ampl. interquartis

Amplitude (A) – Diferença entre os valores máximo e mínimo dos dados

$$A = (x_n)_{MAX} - (x_n)_{MIN}$$

Desvantagem: afectada por valores extremos atípicos

Intervalo interquartis (IIQ) – Intervalo entre o 1^o e o 3^o quartil

$$IIQ = [Q1, Q3]$$

Amplitude interquartis (AIQ) – Diferença entre o 3^o e o 1^o quartil

$$AIQ = Q3 - Q1$$

Slide 2.28

Estat. de dispersão – Desvio (absoluto) médio

- Média dos desvios absolutos dos dados em relação à média

Dados não agrupados: $DAM = \frac{1}{N}\sum_{n=1}^{N} |x_n - \bar{x}|$

Dados discretos agrupados: $DAM = \sum_{k=1}^{K} f_k \cdot |x_k - \bar{x}|$

Dados contínuos agrupados: $DAM = \sum_{k=1}^{K} f_k \cdot |M_k - \bar{x}|$

Notas

- f_k: frequência relativa, M_k: o ponto central da classe k
- Dados agrupados apenas permitem cálculo aproximado, porquê?
- Porquê desvio absoluto?

Slide 2.29

Estat. de dispersão – Desvio quadrático médio

- Média dos quadrados dos desvios dos dados em relação à média
- Dá maior peso a desvios maiores (devido ao quadrado)

Dados não agrupados: $DQM = \frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{x})^2$

Dados discretos agrupados: $DQM = \sum_{k=1}^{K} f_k \cdot (x_k - \bar{x})^2$

Dados contínuos agrupados: $DQM = \sum_{k=1}^{K} f_k \cdot (M_k - \bar{x})^2$

Notas

- Preferível ao DAM (maior facilidade de manipulação matemática)
- Utilização do valor central da classe em dados agrupados leva à sobreavliação sistemática do DQM $\Rightarrow$ *Correcção de Sheppard*

Slide 2.30

Correcção de Sheppard

- Tomar o ponto médio da classe M_k, no caso de dados contínuos agrupados, é um procedimento já adoptado no cálculo da média amostral e do desvio absoluto médio.
- No cálculo da média e da mediana não se introduz qualquer erro sistemático, pois os desvios introduzidos por pontos abaixo do ponto médio M_k tendem a ser compensados por desvios de sinal contrário introduzidos por pontos acima de M_k, isto é, se nuns casos se subavalia noutros sobreavalia-se.

- No cálculo do *DQM*, ao elevar-se ao quadrado os desvios, tende-se a sobreavaliar sistematicamente o *DQM*.
- Para histogramas unimodais, com classes de igual amplitude Δ, com a classe modal relativamente afastada dos valores extremos dos dados e com classes progressivamente decrescentes nas caudas (histograma em forma de sino), este efeito pode ser minorado pela introdução de um termo negativo no cálculo do *DQM* – a correcção de Sheppard (para o desvio quadrático médio):

$$DQM = \sum_{k=1}^{K} f_k (M_k - \bar{x})^2 - \frac{\Delta^2}{12}$$

Slide 2.31

Estatísticas de dispersão – Variância amostral

- Semelhante ao *DQM* mas divide-se por $N-1$ em vez de N

$$s^2 = \frac{N}{N-1} \cdot DQM$$

Dados não agrupados: $s^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2$

Dados discretos agrupados: $s^2 = \frac{N}{N-1} \sum_{k=1}^{K} f_k (x_k - \bar{x})^2$

Dados contínuos agrupados: $s^2 = \frac{N}{N-1} \sum_{k=1}^{K} f_k (M_k - \bar{x})^2$

Porquê dividir por $N-1$?
Para fazer inferências sobre a população a partir de valores de amostras.
A justificar quando se abordar a estimação pontual

Slide 2.32

Estatísticas de dispersão – Desvio padrão amostral

- Desvio padrão amostral (s) é igual à raiz quadrada da variância amostral (s^2)

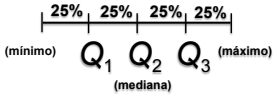
$$s = \sqrt{s^2}$$

- Preferível em relação à variância amostral por se exprimir nas mesmas unidades que os dados a partir dos quais é calculada
- Interpretável como o valor absoluto de um desvio *típico* dos dados em relação à média amostral
- Da mesma ordem de grandeza do desvio absoluto médio (DAM)

Slide 2.33

Estat. de dispersão – Quartis e Percentis

Quartis (Q_1, Q_2. Q_3)
Dividem um conjunto *ordenado* de dados em 4 segmentos com igual número de elementos



Cálculo dos quartis (livro recomendado)
Vector de dados $(x_1, x_2, \ldots, x_N)$ ordenados por ordem crescente (ou decrescente), a posição do 1^o e 3^o quartil vem

$$pos(Q1) = \frac{N+1}{4} \qquad pos(Q2) = \frac{2(N+1)}{4} \qquad pos(Q3) = \frac{3(N+1)}{4}$$

- Se $pos(Q_i)$ é um número inteiro então $Q_i = x_{pos(Q_i)}$
- Senão Q_i é igual à média entre os dois valores correspondentes às posições mais próximas $\left(Q_i = (x_{\lfloor pos(Q_i) \rfloor} + x_{\lceil pos(Q_i) \rceil})/2\right)$

Slide 2.34

Não há uma regra única para o cálculo dos quartis. Outras regras:

- arredondar posição para o valor inteiro mais próximo
- média pesada dos valores correspondentes às posições mais próximas

$\Rightarrow$ na prática só se notam diferenças significativas para conjuntos de dados de pequenas dimensões

Exemplo:
Dados ordenados ($N = 11$): $11, 12, 13, 16, 16, 17, 18, 19, 21, 21, 22$

Percentis ($P_1, P_2, \ldots, P_{99}$)

- Dividem o conjunto de dados em 100 grupos, com cerca de 1% de valores em cada grupo

$$\text{Percentil do valor } X = \frac{\text{número de valores inferiores a } X}{\text{número total de valores}} \%$$

Slide 2.35

Momentos ordinários (m'_i), ou em relação à origem, de ordem i

$$\textbf{Dados não agrupados: } m'_i = \frac{1}{N} \sum_{n=1}^{N} (x_n)^i \quad (i = 1, 2, \ldots)$$

$$\textbf{Dados discretos agrupados: } m'_i = \sum_{k=1}^{K} f_k (x_k)^i \quad (i = 1, 2, \ldots)$$

$$\textbf{Dados contínuos agrupados: } m'_i = \sum_{k=1}^{K} f_k (M_k)^i \quad (i = 1, 2, \ldots)$$

- A média amostral corresponde ao momento ordinário de primeira ordem:

$$m'_1 = \overline{x}$$

Slide 2.36

Momentos centrados (m_i), ou em relação à média, de ordem i

$$\textbf{Dados não agrupados: } m_i = \frac{1}{N} \sum_{n=1}^{N} (x_n - \overline{x})^i \quad (i = 1, 2, \ldots)$$

$$\textbf{Dados discretos agrupados: } m_i = \sum_{k=1}^{K} f_k (x_k - \overline{x})^i \quad (i = 1, 2, \ldots)$$

$$\textbf{Dados contínuos agrupados: } m_i = \sum_{k=1}^{K} f_k (M_k - \overline{x})^i \quad (i = 1, 2, \ldots)$$

- O momento centrado de 1^a ordem é sempre nulo e o momento centrado de 2^a ordem corresponde ao desvio quadrático médio:

$$m_1 = 0 \qquad \wedge \qquad m_2 = DQM$$

Slide 2.37

Relação entre momentos ordinários e momentos centrados
Qualquer momento centrado de ordem i pode ser expresso em função dos momentos ordinários de ordem não superior a i, por exemplo:

- $m_2 = m'_2 - (m'_1)^2$
- $m_3 = m'_3 - 3m'_2 m'_1 + 2(m'_1)^3$
- $m_4 = m'_4 - 4m'_3 m'_1 + 6m'_2 (m'_1)^2 - 3(m'_1)^4$

- Os momentos centrados não contêm mais informação que os momentos ordinários. A utilização de uma ou outra forma depende da facilidade do cálculo e da facilidade de interpretação de um ou outro

- Embora a definição não imponha limites à ordem dos momentos, raramente se calculam momentos de ordem superior a 4 pois a partir daí são extremamente difícies de interpretar

Slide 2.38

Momento centrado de 3ª ordem – desvio cúbico médio (m_3)

- Mede a *assimetria* com que os dados se distribuem em torno da média
- Dados assimétricos à direita darão origem a um m_3 positivo enquanto uma assimetria à esquerda originará m_3 negativo
- A correcção de Sheppard não é necessária (nem útil) por a ordem do momento ser ímpar (não há sobreavaliação sistemática)

Coeficiente de assimetria (g_1)

- Para fazer inferências sobre populações a partir de amostras deve-se substituir m_3 por k_3:

$$k_3 = \frac{N^2}{(N-1)(N-2)} m_3$$

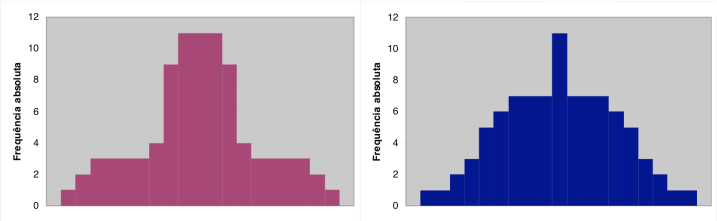
- Para N grande k_3 e m_3 praticamente não se distinguem
- Habitualmente usa-se uma medida adimensional, o *coeficiente de assimetria* (g_1), em vez de k_3 directamente:

$$g_1 = \frac{k_3}{s^3}$$

Slide 2.39

Momento centrado de 4ª ordem (m_4)

- A característica a que está associado este momento é a *kurtose* – medida de concentração dos dados em torno da média (*achatamento*)
- Exemplo: Amostras de igual dimensão, com iguais medidas de localização, de dispersão e de simetria mas diferentes kurtoses:



- A correcção de Sheppard para o quarto momento tem a seguinte expressão (nas condições anteriormente expostas):

$$CS(m_4) = \frac{\Delta^2}{2} \sum_{k=1}^{K} f_k (M_k - \bar{x})^2 - \frac{7\Delta^4}{240}$$

Slide 2.40

Coeficiente de kurtose (g_2)

- *Alternativa*: Utilizar $m_4 - 3m_2^2$ como medida de kurtose
- *Porquê?* Para a distribuição Normal $m_4 = 3 \cdot m_2^2$. Então a medida de kurtose de uma amostra pode ser lida relativamente à kurtose da população normal: superior, igual ou inferior, conforme é maior, igual ou menor do que zero.
- *Ajuste de m_4 a inferências em relação a uma população*

$$k_4 = \frac{N^2}{(N-1)(N-2)(N-3)} [(N+1)m_4 - 3(N-1)m_2^2]$$

- *Reescalamento de k_4 – o coeficiente de kurtose*

$$g_2 = \frac{k_4}{s^4} = \frac{N^2(N+1)}{(N-1)(N-2)(N-3)} \times \frac{m_4}{s^4} - 3\frac{(N-1)^2}{(N-2)(N-3)}$$

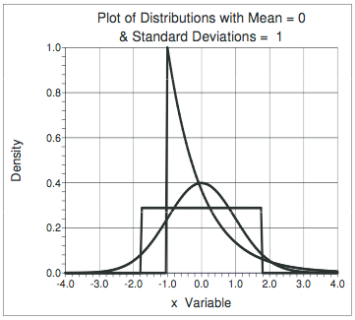
- Para valores de N elevados o coeficiente de kurtose aproxima-se de:

$$g_2 = \frac{m_4}{s^4} - 3$$

Slide 2.41

Porquê momentos de ordem superior a dois?

⇒ A média e o desvio padrão (ou a variância) não são suficientes para distinguir distribuições

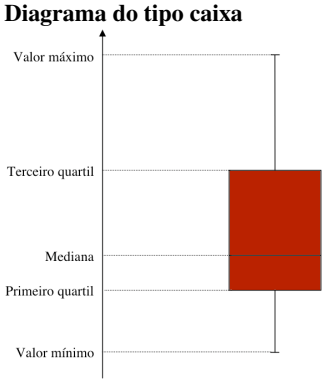


Slide 2.42

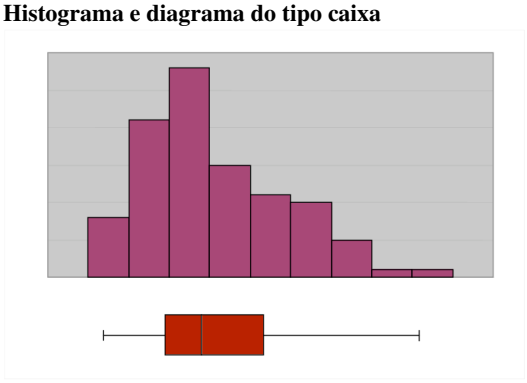
(Nóvoa, 2008)

2.4.3 Representação gráfica de estatísticas

Diagrama do tipo caixa



Histograma e diagrama do tipo caixa



- Não existe uma forma única para o diagrama do tipo caixa
- A forma apresentada é a forma básica do diagrama do tipo caixa
- Formas mais elaboradas permitem identificar potenciais valores extremos atípicos (*outliers*)

Slide 2.43

50	37	31	31	28	27	24	23	23	22	22
21	21	21	21	20	20	20	19	19	18	18
18	18	17	17	14	14	14	13	11	11	10

N = 33 Q_3 = 23.0

Max = 50.0 Med = 20.0

Min = 10.0 Q_1 = 17.0

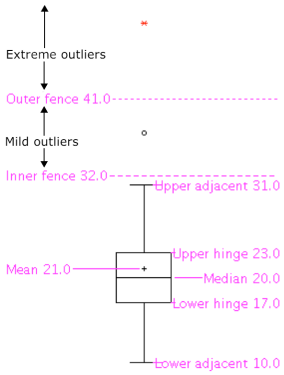
$\bar{x}$ = 21.0 AIQ = Q_3 - Q_1 = 6.0

UpperOuterFence = $Q_3 + 3 \cdot$ AIQ = 41.0

UpperInnerFence = $Q_3 + 1.5 \cdot$ AIQ = 32.0

LowerInnerFence = $Q_1 - 1.5 \cdot$ AIQ = 8.0

LowerOuterFence = $Q_1 - 3 \cdot$ AIQ = -1.0



(*onlinestatbook*)

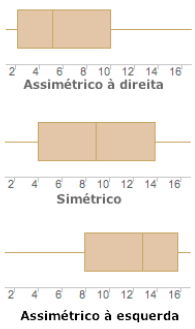
(http://onlinestatbook.com/chapter2/boxplot_demo.html)

- Valores extremos atípicos podem ser causados por erros inerentes ao processo de aquisição de dados
- Podem influenciar bastante o valor de algumas estatísticas

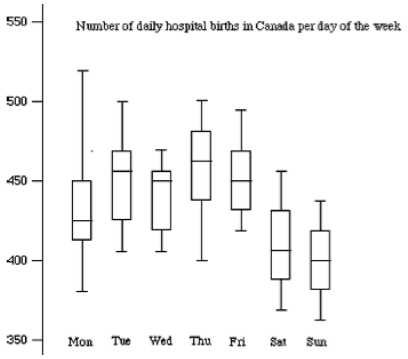
Slide 2.44

Interpretação de diagramas do tipo caixa

- a caixa central inclui 50% dos dados (os 50% centrais)
- os "bigodes" mostram a amplitude dos dados
- a simetria é indicada pela caixa central, "bigodes" e valor médio
- comparação entre diferentes grupos ⟶ construção de diagramas do tipo caixa lado a lado (ver caracterização de amostras bivariadas)



(*StatTrek*)



(*Schwarz, 2006*)

Slide 2.45

Como lidar com valores extremos (*outliers*)?
Duas alternativas:

(a) Utilizar medidas resistentes

Medidas resistentes: pouco sensíveis a valores extremos atípicos (baseadas em ordenações – quartis, mediana, *trimean*, *trimmean*, ...)
Notas:

- *trimean*: $TM = (Q1 + 2Q2 + Q3)/4$
- o *trimmean* é uma função do *Excel* que calcula a média ignorando uma percentagem dos valores mais extremos

Medidas não resistentes: são sensíveis a valores extremos atípicos (consideram sempre todos os valores – média, DQM, variância, ...)

(b) Eliminar *outliers*

1. Identificar possíveis valores extremos atípicos
2. Verificar cada um dos valores extremos atípicos (erro de medida?)
3. Corrigir ou Eliminar valores extremos atípicos confirmados
4. Calcular as estatísticas amostrais pretendidas

Slide 2.46

2.5 Caracterização e Representação de Dados Bivariados

Amostra bivariada
Amostra constituída por *pares ordenados* de dados (medem-se 2 atributos diferentes do mesmo objecto)

- Para além da análise univariada a cada um dos atributos ...
- Pretende-se também *analisar as interacções* entre os dois atributos

Dados qualitativos:

- Tabela de informação cruzada
- Diagramas de barras sobrepostas

Dados quantitativos:

- Diagrama (x, y)
- Coeficientes de correlação e de determinação

Dados mistos:

- Um dos atributos é quantitativo e o outro é qualitativo

Slide 2.47

2.5.1 Dados Qualitativos

Dados qualitativos – Tabela de inform. cruzada

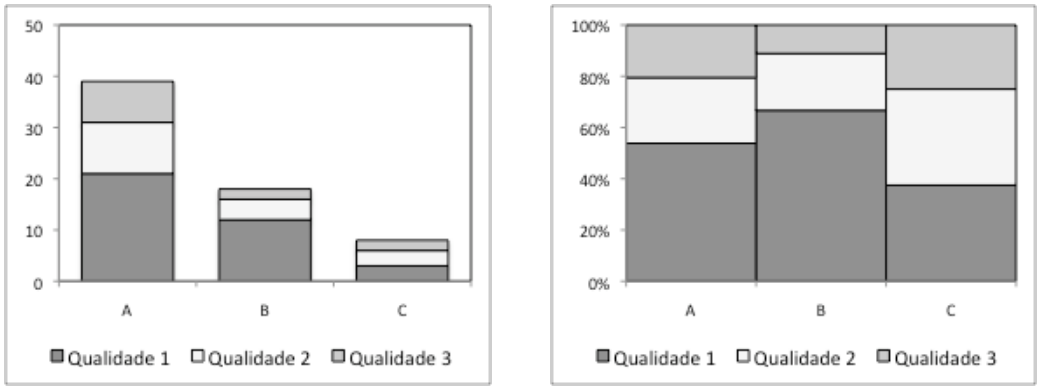
Nota: também conhecida como tabela de contingência

	Qualidade das entregas			
Distribuidor	1	2	3	Total da linha
A	21	10	8	39
	32.3%	15.4%	12.3%	60.0%
	53.8%	25.6%	20.5%	100.0%
	58.3%	58.8%	66.7%	–
B	12	4	2	18
	18.5%	6.2%	3.1%	27.7%
	66.7%	22.2%	11.1%	100.0%
	33.3%	23.5%	16.7%	–
C	3	3	2	8
	4.6%	4.6%	3.1%	12.3%
	37.5%	37.5%	25.0%	100.0%
	8.3%	17.6%	16.7%	–
Total	36	17	12	65
da coluna	55.4%	26.2%	18.5%	100.0%
	100.0%	100.0%	100.0%	–

Em cada célula figuram, de cima para baixo: n^o de observações, % do n^o total de observações, % do n^o de observ. da linha e % do n^o de observ. da coluna

Slide 2.48

Dados qualitativos – Diagr. de barras sobrepostas



Nota o diagrama da direita também é conhecido como diagrama mosaico

Slide 2.49

Dados qualitativos – Paradoxo de Simpson

Business School		
	Admit	Deny
Male	480 = 80%	120 = 20%
Female	180 = 90%	20 = 10%

Law School		
	Admit	Deny
Male	10 = 10%	90 = 90%
Female	100 = 33%	200 = 67%

- Aparentemente a taxa de admissão de mulheres é mais alta quer na Business school quer na Law school

Business and Law Schools		
	Admit	Deny
Male	490 = 70%	210 = 30%
Female	280 = 56%	220 = 44%

- Agora a taxa de admissão de mulheres é inferior à dos homens!
- O que aconteceu? Esta situação é conhecida como Paradoxo de Simpson: inversão de resultados quando se juntam grupos
- Causado pela diferença nas percentagens nas suas tabelas (não se devem combinar percentagens) e não pelo tamanho das amostras

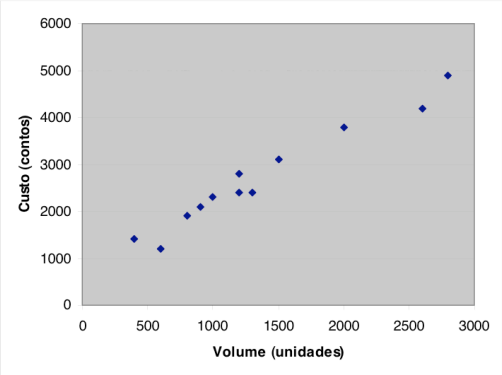
Slide 2.50

2.5.2 Dados Quantitativos

Dados quantitativos – Diagrama (x, y)

Lote	Volume de produção (unidades)	Custo de produção (contos)
1	1500	3100
2	800	1900
3	2600	4200
4	1000	2300
5	600	1200
6	2800	4900
7	1200	2800
8	900	2100
9	400	1400
10	1300	2400
11	1200	2400
12	2000	3800

Diagrama (x, y)



- Caracterização da interacção entre X e Y

Slide 2.51

Ajuste de uma relação linear → MMQ

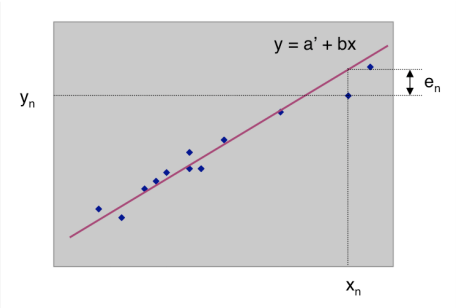
Método dos mínimos quadrados (MMQ)

- Considere-se uma recta que, no plano (x, y), tem a equação:

$$y = a' + b \cdot x$$

- Os parâmetros a' e b são fixados atendendo ao *erro* associado a cada ponto (x_n, y_n):

$$e_n = y_n - (a' + bx_n)$$



Slide 2.52

- Como medir o ajuste de uma dada recta a um conjunto de pontos

  ⇒ Utilizando os erros e_n

  Soma dos erros: cancelamento dos erros positivos com os negativos

  Soma do valor absoluto dos erros: dificuldade de manipulação matemática

  Soma dos erros quadráticos:

$$SEQ = \sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} (y_n - a' - b \cdot x_n)^2$$

- Para um dado conjunto de pontos (x_n, y_n), SEQ apenas varia com os parâmetros a' e b

Slide 2.53

- Determinar os valores de a' e b que minimizam SEQ

  ⇒ calcular as derivadas parciais e igualá-las a zero

$$\begin{cases} \frac{\mathrm{d}SEQ}{\mathrm{d}a'} = (-2) \cdot \sum_{n=1}^{N} (y_n - a' - b \cdot x_n) = 0 \\ \frac{\mathrm{d}SEQ}{\mathrm{d}b} = (-2) \cdot \sum_{n=1}^{N} x_n \cdot (y_n - a' - b \cdot x_n) = 0 \end{cases} \Leftrightarrow \begin{cases} a' = \bar{y} - b\bar{x} \\ b = \frac{s_{XY}}{s_{XX}} \end{cases}$$

onde: $s_{XY} = \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})$ $\quad s_{XX} = \sum_{n=1}^{N} (x_n - \bar{x})^2$

Nota: Utiliza-se a' por razões de consistência com notação a introduzir no capítulo sobre Regressão

Slide 2.54

Ajuste de uma relação linear → MMQ (Exemplo)

	x_n	y_n	$(x_n-\bar{x})(y_n-\bar{y})$	$(x_n-\bar{x})^2$
	1500	3100	55486.11	20069.44
	800	1900	451319.44	311736.11
	2600	4200	1852152.78	1541736.11
	1000	2300	146319.44	128402.78
	600	1200	1143819.44	575069.44
	2800	4900	3159652.78	2078402.78
	1200	2800	-14513.89	25069.44
	900	2100	278819.44	210069.44
	400	1400	1253819.44	918402.78
	1300	2400	17986.11	3402.78
	1200	2400	48819.44	25069.44
	2000	3800	700486.11	411736.11
$\sum$	16300	32500	9094166.67	6249166.67
$\sum/N$	1358.33	2708.33	s_{XY}	s_{XX}

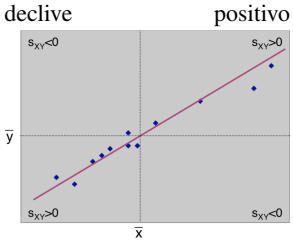
$$b=\frac{S_{XY}}{S_{XX}}=1.455 \qquad a'=\bar{y}-b\cdot\bar{x}=731.6$$

$$y=a'+b\cdot x=731.6+1.455\cdot x$$

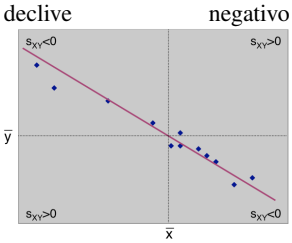
Slide 2.55

Tipo de relações

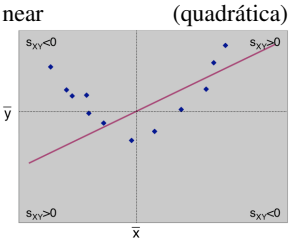
Relação linear com declive positivo



Relação linear com declive negativo



Relação não linear (quadrática)



Observações

- Recta passa sempre pelo ponto $(\bar{x},\bar{y})$, sejam quais forem os dados
- Desde que os x_n não sejam todos iguais, $s_{XX}>0$, o sinal do declive da recta (b) será definido pelo termo $s_{XY}=\sum_{n=1}^{N}(x_n-\bar{x})(y_n-\bar{y})$
- Perigoso fazer extrapolações para fora da gama de valores dos dados
- MMQ pode também ser utilizado para ajustar relações não lineares

Slide 2.56

Medidas de associação (ou qualidade de ajuste)

- *Associação* é um qualquer tipo de relacionamento entre os dados
- Vamos apenas estudar medidas de relacionamento *linear*

Medidas do grau de ajustamento da relação linear aos dados

- *Produto cruzado médio*: $\dfrac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})(y_n-\bar{y})$
- *Covariância amostral*: (para inferências acerca da população)

$$c_{XY}=\frac{1}{N-1}\sum_{n=1}^{N}(x_n-\bar{x})(y_n-\bar{y})$$

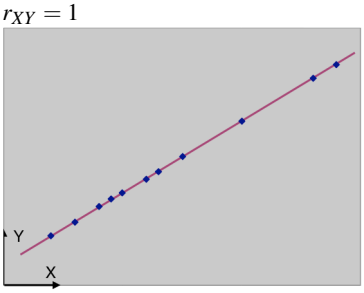
- *Coeficiente de correlação amostral*: (medida adimensional)

$$r_{XY}=\frac{c_{XY}}{s_x s_Y}=\frac{\frac{1}{N-1}\sum_{n=1}^{N}(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_n-\bar{x})^2}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(y_n-\bar{y})^2}}=\frac{s_{XY}}{\sqrt{s_{XX}s_{YY}}}$$

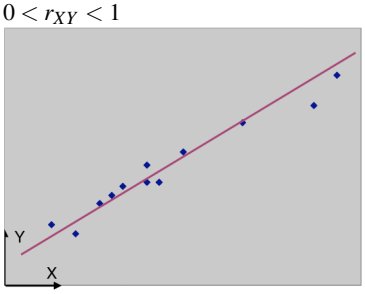
Nota: $s_X=\frac{1}{N-1}\sum_{n=1}^{N}(x_n-\bar{x})^2$ é diferente de $s_{XX}=\sum_{n=1}^{N}(x_n-\bar{x})^2$

Slide 2.57

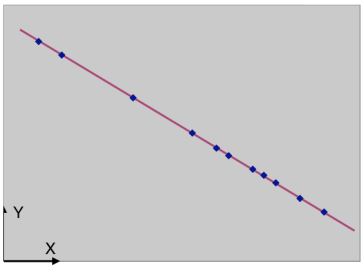
$r_{XY} = 1$



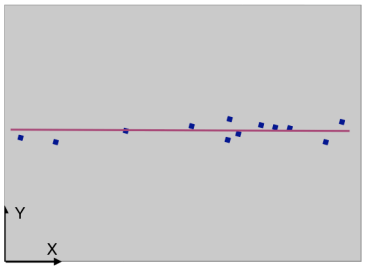
$0 < r_{XY} < 1$



$r_{XY} = -1$



$r_{XY} = 0$



Slide 2.58

- $r_{XY} = \pm 1$: relação linear perfeita entre os valores x_n e y_n
- $r_{XY} = 0$: inexistência de qualquer relação linear
- $0 < |r_{XY}| < 1$: interpretação difícil $\Rightarrow$ *Coef. de determinação* r^2_{XY}

  Notas: O coeficiente de correlação só mede relações lineares

  É sensível a valores extremos atípicos

Relação causa-efeito e Coeficiente de correlação

- Existência de relação causa-efeito $\Longrightarrow |r_{XY}| \approx 1$
- Será que $|r_{XY}| \approx 1 \Longrightarrow$ existência de relação causa-efeito ? *Não*

  1. X pode ser a causa da ocorrência de Y
  2. Y pode ser a causa da ocorrência de X
  3. Pode existir uma 3^a variável Z que seja a causa comum da ocorrência de X e Y
  4. Combinação das situações anteriores
  5. A relação entre X e Y pode ser pura coincidência

$\Longrightarrow$ necessário analisar *contexto* de cada situação e/ou realizar experiência específica para verificar relação causa-efeito

Slide 2.59

Coeficiente de determinação amostral (r^2_{XY})

- Expressando a variação em Y em variação não explicada e explicada ...

$$S_{YY} = \sum_{n=1}^{N} (y_n - \bar{y})^2 = \sum_{n=1}^{N} [(y_n - \hat{y}_n) + (\hat{y}_n - \bar{y})]^2 = \overbrace{\sum_{n=1}^{N} e_n^2}^{\text{Não expl.}} + \overbrace{b^2 \sum_{n=1}^{N} (x_n - \bar{x})^2}^{\text{Explicada}}$$

- ...e dividindo a variação explicada pela variação total ...

$$\frac{b^2 \sum_{n=1}^{N} (x_n - \bar{x})^2}{\sum_{n=1}^{N} (y_n - \bar{y})^2} = \frac{\left(\frac{s_{XY}}{s_{XX}}\right)^2 \cdot s_{XX}}{s_{YY}} = \frac{s^2_{XY}}{s_{XX} \cdot s_{YY}} = \mathbf{r^2_{XY}} = 1 - \frac{\sum_{n=1}^{N} e_n^2}{s_{YY}}$$

$\Rightarrow$ r^2_{XY}: representa a proporção da variação dos valores de y que são explicados por variações nos valores de x

- Para inferências relativamente a populações a partir de amostras utiliza-se o *coeficiente de determinação corrigido*:

$$r^2_{XY} \text{ (corrigido)} = 1 - \frac{\sum_{i=1}^{N} (e_n)^2 / N - 2}{s_{XY} / N - 1}$$

Slide 2.60

2.5.3 Dados mistos

Na caracterização do relacionamento envolvendo dados mistos é necessário distinguir duas situações:

1^a situação – a variável dependente (Y) é quantitativa e a variável independente (X) é qualitativa

⇒ Nesta situação, tipicamente pretende-se comparar a performance média entre dois (ou mais) grupos

Exemplo: comparação do número de calorias por dose entre vários grupos de fibras

2^a situação – a variável dependente (Y) é qualitativa e a variável independente (X) é quantitativa

Exemplo: relação entre a quantidade fumada (maços/dia – variável quantitativa) e se um doente desenvolve ou não cancro do pulmão (sim ou não – variável qualitativa)

⇒ Requer conceitos avançados de estatística

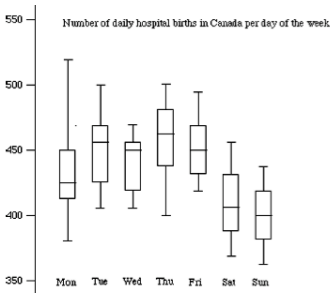
⇒ Iremos estudar apenas a 1^a situação

Slide 2.61

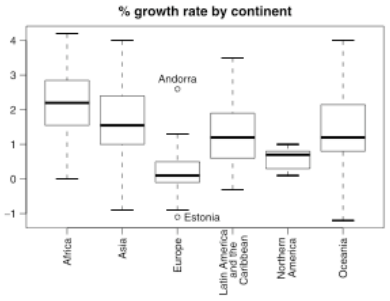
Dados mistos – Comparação entre grupos

- Atributo qualitativo caracteriza o grupo
- Cálculo de estatísticas dentro de cada grupo
- Diagramas do tipo caixa para cada grupo, colocados lado a lado e alinhados para facilitar a comparação

⇒ Procurar *diferenças de dispersão* (variação) entre os vários grupos, *alterações nas médias* de cada grupo e outliers (valores extremos atípicos)



(*Schwarz, 2006*)



(*Schwarz, 2006*)

Slide 2.62

2.6 Anexos

2.6.1 Classificação de Dados: Resumo

Summary of The Four Levels of Measurement: Appropriate Descriptive Statistics and Graphs

Level of Measurement	P roperties	Examples	Descriptive statistics	Graphs
Nominal / Categorical	Discrete Arbitrary (no order)	Dichotomous Yes / No Gender Types / Categories Colour shape	Frequencies Percentage Mode	Bar Pie
Ordinal / Rank	Ordered categories Ranks	Ranking of favourites Academic grades	Frequencies Mode Median Percentiles	Bar Pie Stem&leaf
Interval	Equal distances between values Discrete (e.g., Likert scale) Continuous (e.g., deg. F)	Discrete - Thoughts, behaviours, feelings, etc. on a Likert scale Continuous - Deg. C or F	Frequencies (if discrete) Mode (if discrete) Median Mean SD Skewness Kurtosis	Bar (if discrete) Pie (if discrete) Stem & Leaf Boxplot Histogram (if continuous)
Ratio	Continuous Meaningful 0 allows ratio statements (e.g., A is twice as large as B)	Age Weight VO_2 max Deg. Kelvin	Mean SD Skewness Kurtosis	Histogram Boxplot Stem & Leaf (may need to round leafs)

adapted from http://www.wilderdom.com/research/Summary_Levels_Measurement.html (James Neil, 2009)

Slide 2.63

Summary of Descriptive Statistics Summaries for the Four Levels of Measurement

Statistic	Nominal	Ordinal	Interval	Ratio
Mode	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Range, Min, Max	No	Yes	Yes	Yes
Mean	No	No	Yes	Yes
SD	No	No	Yes	Yes

Summary of Graphical Summaries for the Four Levels of Measurement

Graph	Nominal	Ordinal	Interval	Ratio
Bar / Pie	Yes	Yes	If discrete	No
Stem & Leaf	No	Yes	Yes	Yes
Boxplot	No	Yes	Yes	Yes
Histogram	No	No	If continuous	Yes

adapted from http://www.wilderdom.com/research/Summary_Levels_Measurement.html (James Neil, 2009)

Slide 2.64

2.6.2 Fórmulas de Cálculo Alternativas

Cálculo da média amostral ($\bar{x}$)

- Soma de uma constante C

$$y_i = x_i + C, \quad i = 1, \ldots, n \quad \rightarrow \quad \bar{y} = \bar{x} + C$$

- Multiplicação por uma constante C

$$y_i = C \cdot x_i, \quad i = 1, \ldots, n \quad \rightarrow \quad \bar{y} = C \cdot \bar{x}$$

Exemplo

The average fuel efficiencies, in miles per gallon, of cars sold in the United States in the years 1999 to 2003 were: $28.2, 28.3, 28.4, 28.5, 29.0$. Find the sample mean of this set of data.

Fazendo: $y_i = 10 \cdot x_i - 280$

x_i	28.2	28.3	28.4	28.5	29.0
y_i	2	3	4	5	10

$$\bar{y} = \frac{2+3+4+5+10}{5} = \frac{24}{5} = 4.8 \quad \rightarrow \quad \bar{x} = \frac{\bar{y}+280}{10} = \frac{4.8+280}{10} = \frac{284.8}{10} = 28.48$$

Slide 2.65

Cálculo da variância amostral (s^2)

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

- Dificuldade reside no cálculo da expressão: $\sum_{i=1}^{n} (x_i - \bar{x})^2$
- Esta expressão pode ser simplificada de forma a facilitar o cálculo manual:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2 \cdot \bar{x} \cdot \sum_{i=1}^{n} x_i + \bar{x}^2 \cdot \sum_{i=1}^{n} 1 = \sum_{i=1}^{n} x_i^2 - 2 \cdot \bar{x} \cdot n \cdot \bar{x} + \bar{x}^2 \cdot n = \left(\sum_{i=1}^{n} x_i^2 \right) - \left(n \cdot \bar{x}^2 \right)$$

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

Slide 2.66

Cálculo de s^2 pela fórmula tradicional

x_i	1	2	5	6	6
$\bar{x}$	4	4	4	4	4
$x_i - \bar{x}$	-3	-2	1	2	2
$(x_i - \bar{x})^2$	9	4	1	4	4

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

$$= \frac{9+4+1+4+4}{5-1} = \frac{22}{4} = 5.5$$

Cálculo de s^2 pela fórmula alternativa

x_i	1	2	5	6	6
x_i^2	1	4	25	36	36

$$\sum_{i=1}^{5} x_i^2 = 1+4+25+36+36 = 102$$

$$\sum_{i=1}^{N} x_i^2 - N\cdot\bar{x}^2 = 102 - 5\times 16 = 22$$

$$s^2 = \frac{1}{N-1}\cdot\left(\sum_{i=1}^{N} x_i^2 - N\cdot\bar{x}^2\right) = \frac{22}{4} = 5.5$$

Slide 2.67

Exemplo

The following data give the yearly numbers of law enforcement officers killed in the United States over 10 years:

$$164, 165, 157, 164, 152, 147, 148, 131, 147, 155$$

Find the sample variance of the number killed in these years.

To make the computation easier: $y_i = x_i - 150$

x_i	164	165	157	164	152	147	148	131	147	155
y_i	14	15	7	14	2	-3	-2	-19	-3	5
y_i^2	196	225	49	196	4	9	4	361	9	25

$$\bar{y} = \frac{1}{10}\cdot\sum_{i=1}^{5} y_i = \frac{14+15+7+14+2-3-2-19-3+5}{10} = \frac{30}{10} = 3.0$$

$$\sum_{i=1}^{5} y_i^2 = 196+225+49+196+4+9+4+361+9+25 = 1078$$

$$\sum_{i=1}^{N} y_i^2 - N\cdot\bar{y}^2 = 1078 - 10\times 3^2 = 988 \qquad s^2 = \frac{988}{9} \approx 109.78$$

Nota: $x_i - \bar{x} = x_i - \bar{x} - C + C = (x_i - C) - (\bar{x} - C) = y_i - \bar{y}$

Slide 2.68

Cálculo do coeficiente de correlação amostral (r)

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

Computational formula for r:

$$r = \frac{\sum_{i=1}^{n}(x_i\cdot y_i) - n\cdot\bar{x}\cdot\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right)\cdot\left(\sum_{i=1}^{n} y_i^2 - n\cdot\bar{y}^2\right)}}$$

Useful property — If r is the sample correlation coefficient for the data x_i, y_i with $i = 1,\ldots,n$, then for any constants a, b, c and d, it is also the sample correlation coefficient for the data $a + b\cdot x_i, c + d\cdot y_i$ with $i = 1\ldots,n$ provided that b and d have the same sign (that is, provided that $b\cdot d \geq 0$).

Slide 2.69

Exemplo

The following table gives the U.S. per capita consumption of whole milk (x) and low-fat milk (y) in three different years. Find the sample correlation coefficient of the given data.

	Per capita consumption (gallons)		
	1980	1982	1984
Whole milk (x)	17.1	14.7	12.8
Low-fat milk (y)	10.6	11.5	13.2

To make the computation easier, let us first subtract 12.8 from each of the x values and 10.6 from each of the y values.

	i		
	1	2	3
x_i'	4.3	1.9	0
y_i'	0	0.9	2.6

$$\bar{x'} = \frac{4.3+1.9+0}{3} = 2.0667$$

$$\bar{y'} = \frac{0+0.9+2.6}{3} = 1.1667$$

$$\sum_{i=1}^{3} x_i y_i = 0 + 1.9\times 0.9 + 0 = 1.71$$

$$\sum_{i=1}^{3} x_i^2 = 4.3^2 + 1.9^2 + 0 = 22.10$$

$$\sum_{i=1}^{3} y_i^2 = 0 + 0.9^2 + 2.6^2 = 7.57$$

$$r = \frac{1.71 - 3 \times 2.0667 \times 1.1667}{\sqrt{[22.10 - 3 \times 2.0667^2] \cdot [7.57 - 3 \times 1.1667^2]}} = -0.97$$

Slide 2.70

Comentários:

- Fórmula alternativa *facilita* os cálculos quando realizados com calculadoras tradicionais
- Apresenta no entanto problemas de *precisão numérica* que poderão originar resultados errados (devido à subtracção de dois valores da mesma dimensão quando estes têm muitos dígitos significativos, por serem muito grandes e/ou terem muitas casas decimais, devido à capacidade finita da representação em vírgula flutuante — ver exemplo no slide seguinte)
- Existe uma outra fórmula de cálculo baseada em iterações que é igualmente eficiente em termos computacionais e não apresenta problemas de previsão numérica (ver "Computing the standard deviation efficiently", Mark Hoemmen, `http://www.cs.berkeley.edu/~mhoemmen/cs194/Tutorials/variance.pdf`)
- Para a grande maioria das situações a fórmula alternativa apresentada atrás é adequada e a mais prática de usar

Slide 2.71

Exemplo
Calcule, usando a fórmula alternativa, a variância amostral para as seguintes situações:

10 000, 10 001, 10 002

100 000, 100 001, 100 002

1 000 000, 1 000 001, 1 000 002

10 000 000, 10 000 001, 10 000 002

100 000 000, 100 000 001, 100 000 002

...

- A correcção dos resultados depende da precisão do computador ou da calculadora usada

- Como se poderia obter o resultado correcto facilmente?

Slide 2.72