Introdução à Probabilidade e Estatística
Estatística descritiva

Ano letivo: 2017/2018

Docente: Manuela Oliveira

Universidade de Évora

26 de Setembro de 2017

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

- População ou universo:
  - Conjunto de todos os elementos que têm uma caraterística de interesse em comum (ex: todas as árvores de uma dada espécie);
- Unidades estatísticas:
  - são os elementos da população (ex: as árvores);
- Variável:
  - caraterística de interesse (X: altura das árvores de uma espécie; x: altura observada de uma árvore);
- Amostra:
  - subconjunto da população, efetivamente observado.

Estatística descritiva a uma dimensão

- Aos valores das caraterísticas de interesse, observadas nos elementos da amostra, costuma chamar-se **dados**
- Os dados podem ser de natureza:
  - **quantitativa:**
    - **discreta:** contagens (nº de paras em cada pereira), nº de machos por ninhada de coelhos;
    - **contínua:** peso, comprimento, altura, tempo.
  - **qualitativa:**
    - **nominal:** sexo de um indivíduo, categoria taxonómica de uma espécie;
    - **ordinal:** avaliação numa escala de A (ótima) a E (péssima), da qualidade do almoço numa cantina.

Estatística descritiva a uma dimensão

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 de uma **tabela de frequências***

Descrição dos dados por tabelas

Tabela de frequências:

- Caso de 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

X_i - nº de sementes germinadas;
n_i - frequência absoluta;
$f_i = \frac{n_i}{n}$ - frequência relativa
F_i - frequência relativa acumulada.

Descrição dos dados por tabelas

Exemplo 2:

Um dos principais indicadores da poluição atmosférica nas grandes cidades é a concentração de ozono na atmosfera. Num dado Verão, registaram-se 78 valores dessa concentração (em $\mu g/m^3$), 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****.*

Descrição dos dados por tabelas

*Para **dados de natureza contínua**, como é este caso (ou quando temos dados de natureza discreta com um elevado número de valores distintos), elabora-se a **tabela de frequências**, do seguinte modo:*

- *Determina-se **max(x_i)** e **min(x_i)**, em que max(x_i)-min(x_i) é a **amplitude total**;*
- *Escolhe-se um número de subintervalos (**classes**);*
- *Para cada classe calcula-se a **frequência absoluta**, n_i e a **frequência relativa**, f_i.*

Exemplo de uma regra para a escolha do número de classes:

- ***Regra de Sturges:** toma-se como número de classes, **o inteiro m mais próximo de** $1+\log_2(n) = 1+\frac{\log_{10}(n)}{\log_{10}(2)}$*

Descrição dos dados por tabelas

Exemplo 3:

$min(x_i) = 1,1$ $max(x_i) = 11,7$

Pela Regra de Sturges, $m \approx 7,285$ *(considere-se* $m = 7$*)*

Amplitude das classes $h = 1,51$ *(considere-se* $h = 1,5$*)*

A tabela de frequências possível para este caso (com 8 classes) é:

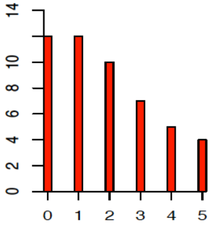
c_i	x'_i	n_i	f_i	F_i
]1,0; 2,5]	1,75	10	0,128	0,128
]2,5; 4,0]	3,25	16	0,205	0,333
]4,0; 5,5]	4,75	18	0,231	0,564
]5,5; 7,0]	6,25	26	0,333	0,897
]7,0; 8,5]	7,75	5	0,064	0,962
]8,5; 10,0]	9,25	2	0,026	0,987
]10,0; 11,5]	10,75	10	0,00	0,987
]11,5; 13,0]	12,25	1	0,013	1

x'_i *é o ponto médio da classe* c_i

Métodos gráficos

Os métodos gráficos mais usados para representar um conjuntos de dados são:

- ***Diagrama de barras:*** *para dados de natureza discreta, com um número pequeno de valores distintos;*
- ***Histograma:*** *para dados de natureza contínua, ou quando o número de valores distintos é muito elevado;*



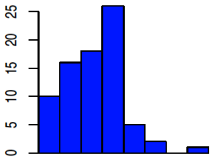


Diagrama de barras (lado esquerdo) e histograma (lado direito), das frequências absolutas.

Indicadores numéricos

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 caraterizar um conjunto de dados é constituído pelos **indicadores numéricos**, também chamados **indicadores amostrais**.*

Falaremos nas:

- *medidas de localização;*
- *medidas de dispersão.*

As medidas de localização que iremos estudar são:

- *média;*
- *mediana;*
- *quantis;*
- *moda.*

Média

Considere-se $x_1, x_2, ..., x_n$, uma amostra de n observações.

Definição: *chama-se* ***média aritmética***, ***média empírica*** *ou simplesmente* ***média***, *e representa-se por $\bar{x}$.*

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

Propriedades da média

- *Sejam $x_1, x_2, ..., x_n$ observações cuja média é $\bar{x}$ e considerando $y_i = a + bx_i$, $i = 1, ..., n$, em que $a, b \in \mathbb{R}$.*
  As observações transformadas $y_1, y_2; ..., y_n$ têm média $\bar{y} = a + b\bar{x}$.
- *Se $x_1, ..., x_n$ são n observações de média $\bar{x}$ e $y_i, ..., y_n$ são m observações de média $\bar{y}$, a média das $n + m$ observações é dada por $\frac{m\bar{x}+m\bar{y}}{n+m}$.*

Mediana e moda

Definição: *A **mediana** é o valor que divide a amostra ordenada em duas partes iguais (isto é, cada parte com o mesmo número de observações).*

*Dada a amostra $x_i, ..., x_n$, seja $x_{(1)} \leq ... \leq x_{(n)}$, a amostra ordenada. A **mediana** é dada por:*

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

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, sendo um valor da classe que tem maior frequência (ver medidas para dados agrupados).*

Quantis empíricos

*Se considerarmos a amostra ordenada, dividida em quatro partes, cada uma com o mesmo número de observações, os pontos de divisão entre as partes designam-se **quantis empíricos**, ou apenas **quartis**, e costumam representar-se por Q_1, Q_2 e Q_3. $Q_2 \equiv \bar{x}$.*

Generalização do conceito de quantil

***Definição:** chama-se **quantil de ordem** θ, $(0 \leq \theta \leq 1)$, o valor de Q^*_θ, tal que há uma proporção θ de observações inferiores ou iguais a Q^*_θ e 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 o maior inteiro contido em $n\theta$.

Nota: $Q^*_{0.25} \equiv Q_1$, $Q^*_{0.50} \equiv Q_2$ e $Q^*_{0.75} \equiv Q_3$

Medidas de localização - dados agrupados

Dados agrupados *em* $c(c < n)$ *classes (ou grupos). Sejam* $x_1', x_2', ..., x_c'$ *pontos médios de cada classe (ou valores de cada grupo);* $n_1, n_2, ..., n_c$*, as frequências absolutas de cada classe (ou grupo).*

$$\textbf{\textit{Média agrupada}} = \bar{x} \approx \frac{n_1 x_1' + n_2' x_2' + \ldots + n_c' x_c'}{n} = \frac{\sum_{i=1}^{c} n_i x_i'}{n}$$

Moda amostral *para* ***dados agrupados****:*

1º determina-se a classe amostral (classe com maior frequência);

2º de várias fórmulas que existem, vamos considerar a seguinte:

$$mo \approx x_k^{min} + (x_k^{max} - x_k^{min}) \frac{f_{k+1}}{f_{k-1} + f_{k+1}}$$

sendo k *a classe modal;* f_{k-1} *a frequência relativa da classe anterior e posterior à classe modal, respetivamente,* x_k^{min} *e* x_k^{max} *limites inferior e superior da classe* k*, respetivamente.*

Medidas de localização - dados agrupados

Quantil de ordem θ:

Identifica-se a primeira classe cuja frequência relativa acumulada seja superior ou igual a θ, seja k essa classe e F_k a frequÊncia relativa acumulada correspondente.

Uma das fórmulas usadas para determinas o quantil de ordem θ é:

$$Q_\theta^* \approx x_k^{min} + (x_k^{max} - x_k^{min}) \frac{\theta - F_{k-1}}{f_k}$$

onde F_{k-1} é a frequência relativa acumulada da classe anterior à classe k.

*Nota: A **mediana** para dados agrupados obtém-se considerando a fórmula acima, com $\theta = 0,5$.*

Indicadores de dispersão

- ***Amplitude total:*** $A_{tot} = max(x_i) - min(x_i)$
- ***Amplitude inter-quartil:*** $AIQ = Q_3 - Q_1$
- ***Variância:*** $s_x^2 = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}^2)}{n-1}$
- ***Desvio padrão:*** $s_x = S = \sqrt{s^2}$

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

Uma medida de dispersão relativa (as que se acabaram de indicar são medidas de dispersão absolutas) é o ***coeficiente de variação****, que só se calcula quando as observações têm todas o mesmo sinal. Permite a comparação entre distribuições e defini-se como:*

$$CV = \frac{s}{\bar{x}} x100\%$$

Variância e desvio padrão

Propriedades:

- $s_x^2 \geq 0$
- *Sejam* $x_1, ..., x_n$*, observações com variância* s_x^2*, considere-se* $y_i = a + bx$*,* $i = 1, ..., n$ *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* $s_y = |b| s_x$*.*

Dados agrupados em c classes, a variância calcula-se:

$$s_x^2 = s^2 = \frac{\sum_{i=1}^{c} n_i x_i'^2}{n} - \bar{x}^2$$

Caixa de bigodes

Um modo gráfico que permite facilmente interpretar a localização e a dispersão de um conjunto de dados, efetuando e, simultâneo a sua síntese, é o ***diagrama de extremos e quartis****.*

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

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

Definição:

Um valor de x_i é um candidato a ***outlier*** *se $x < B_I$ ou $x_i > B_S$*

sendo B_I ***barreira interior*** *e B_S* ***barreira superior****, definidas como:*

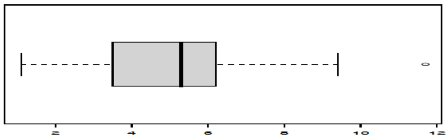
$$B_I = Q_1 - 1,5(Q_3 - Q_1) \qquad B_S = Q_3 + 1,5(Q_3 - Q_1)$$

Caixa de bigodes

*Como desenhar uma **caixa de bigodes**?*

- *Marcar o **valor adjacente inferior** (é o menor valor do conjunto de 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 **outliers**.*

***Exemplo:** caixa de bigodes referente aos dados do Exemplo 2:*



Caixa de bigodes paralelas

Quando se pretende comparar várias amostras, o recurso a caixas de bigodes paralelas é uma ferramenta muito útil, permitindo de forma fácil, obter uma primeira interpretação e comparação dos conjuntos de dados.

Exemplo: *As seguintes caixas de bigodes referem-se a um conjunto de dados* `Insect Sprays`*, disponíveis no package* `datasets` *do programa R. São contagens de insetos em unidades agrícolas experimentais, às quais foram aplicados seis tipos de inseticidas.*

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

Caixa de bigodes paralelas