ANADI
Análise de dados em Informática

Aulas TP - Estatística Descritiva

Instituto Superior de Engenharia do Porto

Ano letivo 2018/2019

- Média: é o quociente entre a soma de todos os valores observados e o número de observações.

Dados não classificados:

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

Dados classificados:

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

onde c é o número de classes, x_i representa o valor característico da classe, n_i a frequência absoluta da classe e f_i a frequência relativa da classe i, respetivamente.

Exercício

- Calcule a média da amostra $\{4, 6, 7, 8, 9, 10\}$.

```
> amostra=c(4,6,7,8,9,10)
> mean(amostra) # usar função mean ou em alternativa
[1] 7.333333
> sum(amostra)/length(amostra)
[1] 7.333333
```

- Calcule a média de uma amostra com a distribuição de frequências seguinte: (R: $\bar{x} = 7$)

Intervalo	Número de observações
[2,4[	5
[4,6[	10
[6,8[	12
[8,10[	10
[10,12[	5

```
> liminf=c(2,4,6,8,10) # limite inferior de cada classe
> limsup=c(4,6,8,10,12) # limite superior de cada classe
> xi=(liminf+limsup)/2 # representante de cada classe
> num.obs=c(5,10,12,10,5) # número de observações em cada classe
> total=sum(num.obs) # número total de observações
> fi=num.obs/total # frequência relativa de cada classe
> media=sum(fi*xi) #média
> media
[1] 7
```

- **Mediana**: divide ao meio o conjunto de valores observados.

Dados discretos ou contínuos não classificados:

Sejam $x_1 \leq x_2 \leq \cdots \leq x_n$ os valores observados. Então

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & \text{se } n \text{ é ímpar,} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{se } n \text{ é par.} \end{cases}$$

Dados contínuos classificados:

$$\tilde{x} = l_{c-1} + \frac{0.5 - F_{c-1}}{f_c}(l_c - l_{c-1})$$

onde c é a classe da mediana, f_c é a frequência relativa da classe c, F_{c-1} é a frequência relativa acumulada até à classe $c-1$, l_{c-1} e l_c são os limites inferior e superior da classe c.

Exercício

a) Determine a mediana das amostras seguintes: $\{5, 5, 7, 15, 16, 17, 24\}$; $\{18, 7, 6, 1, -6, -30\}$.

b) Determine a mediana dos dados classificados seguintes: (R: $\tilde{x} = 7.17$)

Intervalo	Número de observações
[2,4[	5
[4,6[	8
[6,8[	12
[8,10[	10
[10,12[	5

```
> # alínea a)
> dados1=c(5,5,7,15,16,17,24)
> med.dados1=median(dados1)
> med.dados1
[1] 15
> dados2=c(18,7,6,1,-6,-30)
> med.dados2=median(dados2)
> med.dados2
[1] 3.5
```

```
> # alínea b)
> liminf=c(2,4,6,8,10) # limite inferior de cada classe
> limsup=c(4,6,8,10,12) # limite superior de cada classe
> num.obs=c(5,8,12,10,5) # frequência (observações) de cada classe
> total=sum(num.obs) # número total de observações
> num.obs.rel=num.obs/total # frequência relativa
> num.obs.acum=cumsum(num.obs) # frequência acumulada
> num.obs.acum.rel=num.obs.acum/total # frequência relativa acumulada
> n_metade=total/2
> n_metade
[1] 20
> num.obs.acum
[1]  5 13 25 35 40
> # n_metade=20 está na classe 3 das observações acumuladas
> # logo c=3
> c=3
> mediana=liminf[c]+(0.5-num.obs.acum.rel[c-1])/num.obs.rel[c]*(limsup[c]-liminf[c])
> mediana
[1] 7.166667
```

- Moda: é o valor mais comum no conjunto de observações.

  Dados discretos ou contínuos não classificados: é o valor que ocorre com maior frequência num conjunto de observações.

  Dados contínuos classificados: é a classe com maior frequência (numa distribuição de frequência com intervalos de classe de igual amplitude).

- Quantil de ordem q, $0 < q < 1$: é o valor que separa os $100q\%$ valores menores da amostra dos $100(1-q)\%$ valores maiores.
  - Percentis: $p_{0.01}, \ldots, p_{0.99}$.
  - Decis: $d_{0.1}, \ldots, d_{0.9}$.
  - Quartis: $q_{0.25}, q_{0.50}, q_{0.75}$.

Exercício

Determine o quantil de ordem 0.2 da amostra $\{3, 5, 7, 15, 16, 1, 17, 24, 0, -1, 5, 2, 9\}$.

```
> dados=c(3,5,7,15,16,1,17,24,0,-1,5,2,9)
> qt.2=quantile(dados,0.2)
> qt.2
20%
1.4
> # OBS:
> qt.5=quantile(dados,0.5)
> qt.5
50%
  5
> mediana=median(dados)
> mediana
[1] 5
```

- Amplitude total: é a diferença entre o maior e o menor dos valores do conjunto de observações.

$$A = x_n - x_1$$

- Amplitude interquartil: é a diferença entre o terceiro e o primeiro quartil.

$$AIQ = x_{q_3} - x_{q_1}$$

```
> # exemplo:
> dados=c(0.9,0.8,0.3,1.1,1.2,1.3,0.7,0.5)
> # Amplitude total
> A=max(dados)-min(dados)
> A
[1] 1
> # ou alternativamente
> AA=diff(range(dados))
> AA
[1] 1
> # Amplitude interquartil
> AIQ1=quantile(dados,0.75)-quantile(dados,0.25)
> AIQ1
  75%
0.475
> # ou alternativamente
> AIQ2=IQR(dados)
> AIQ2
[1] 0.475
```

- Desvio (absoluto) de x_i em relação a x_0:

$$x_i - x_0 \qquad\qquad |x_i - x_0|$$

- Desvio absoluto médio dos valores x_i em relação a x_0:

Dados não classificados:

$$\delta_{x_0} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_0|$$

Dados classificados:

$$\delta_{x_0} = \frac{1}{n}\sum_{i=1}^{c}n_i|x_i - x_0|$$

x_i valor característico da classe

- Desvio absoluto médio: é a média dos desvios absolutos em relação à média amostral.

Dados não classificados:

$$\delta_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Dados classificados:

$$\delta_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{c}n_i|x_i - \bar{x}|$$

- Variância: indica a proximidade com que os valores estão agrupados à volta da média; um valor pequeno significa que os valores estão pouco espalhados.

Dados não classificados:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right] \end{aligned}$$

Dados classificados:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{c}n_i(x_i-\bar{x})^2 \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{c}n_i x_i^2 - n\bar{x}^2\right] \end{aligned}$$

- Desvio padrão: é a raíz quadrada positiva da variância.

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

```
> # exemplo (para dados não classificados)
> dados=c(0.9,0.8,0.3,1.1,1.2,1.3,0.7,0.5)
> var.dados=var(dados)
> var.dados
[1] 0.12
> dp1.dados=sqrt(var.dados)
> dp1.dados
[1] 0.3464102
> # ou alternativamente
> dp2.dados=sd(dados); dp2.dados
[1] 0.3464102
```

- Coeficiente de variação: é adimensional (sem unidades de medida) baseada no desvio padrão e na média, permitindo comparar a variabilidade de distribuições de frequência diferentes.

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

Nota: Não está definido quando $\bar{x} = 0$; só deve ser usado quando as observações têm todas o mesmo sinal.

```
> # exemplo:
> dados=rnorm(100,-10,2)
> dp.data=sd(dados); dp.data
[1] 2.045684
> med.dados=mean(dados); med.dados
[1] -10.16505
> CV= round(dp.data/abs(med.dados)*100,1) ; CV
[1] 20.1
```

Medidas de forma

São medidas que sintetizam a deformação ou assimetria da distribuição, ou que avaliam o *peso* das caudas da distribuição.

- Momentos amostral simples de ordem r: é a média dos valores observados elevados à potência r.

  Dados não classificados:

  $$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

  Dados classificados:

  $$m'_r = \frac{1}{n}\sum_{i=1}^{c} n_i x_i^r$$

- Momentos amostral centrado de ordem r: é a média dos desvios em relação a $\bar{x}$ elevados à potência r.

  Dados não classificados:

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

  Dados classificados:

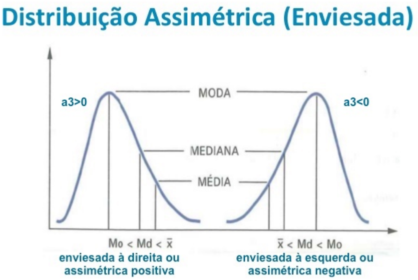
  $$m_r = \frac{1}{n}\sum_{i=1}^{c} n_i(x_i - \bar{x})^r$$

Para calcular momentos pode-se instalar um package que contém a função moments ou alternativamente usar a definição para calculá-los.

```
> install.packages("e1071") # instalar package "e1071"
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.4/e1071_1.6-8.tgz'
Content type 'application/x-gzip' length 763504 bytes (745 KB)
==================================================
downloaded 745 KB


The downloaded binary packages are in
        /var/folders/1l/k0884jy528lb86wlk5zjtmnr0000gn/T//RtmpiEkRkt/downloaded_packages

>library(e1071)
> moment(dados, order=3, center=TRUE) # momento centrado de ordem 3
[1] -3.286185

# alternativamente

> 1/length(dados)*sum((dados-mean(dados))^3)
[1] -3.286185
```

- Coeficiente de assimetria: enviesamento, deformação ou assimetria da distribuição de frequência.

$$a_3 = \frac{m_3}{s^3}$$



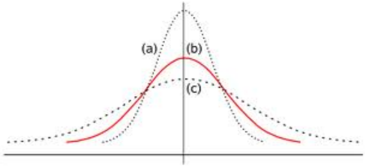
  - É adimensional.
  - $$\begin{cases} a_3 > 0 & \text{assimétrica positiva} \\ a_3 < 0 & \text{assimétrica negativa} \end{cases}$$

```
# exemplo:
>dataki=rchisq(100,0.95,df=7)
> a3=skewness(dataki); a3
[1] 1.381949
> # ou alternativamente
> aa3=moment(dataki, order=3, center=TRUE)/sd(dataki)^3
> aa3
[1] 1.381949
```

- Curtose: *peso* das caudas da distribuição de frequência.

$$a_4 = \frac{m_4}{s^4}$$

  - É adimensional.
  - $$\begin{cases} a_4 > 3 & \text{mais esguia e com cauda mais pesada (a) Leptocúrtica} \\ a_4 = 3 & \text{curva normal (b) Mesocúrtica} \\ a_4 < 3 & \text{mais achatada e com cauda menos pesada (c) Platicúrtica} \end{cases}$$




```
> kurtosis(dataki)
[1] 2.605
> # Note que a função kurtosis subtrai 3 unidades à fórmula dada...
> aa4=moment(dataki, order=4, center=TRUE)/sd(dataki)^4
> aa4
[1] 5.605
```

Organização dos dados

Dados originais (brutos)

é o conjunto de dados na sequência em que foram registados, antes de terem sido organizados ou tratados.

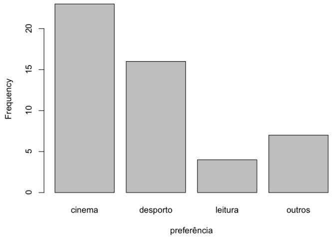
Dados Qualitativos

- Tabelas de frequências.
  - absolutas: nº de observações associadas a cada categoria;
  - relativas: quociente entre a frequência absoluta dessa categoria e o número total de observações.

Preferência	Nº pessoas	Freq. relativa
Leitura	4	8%
Cinema	23	46%
Desporto	16	32%
Outros	7	14%
Total	50	100%

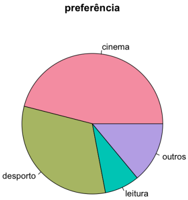
```
> amostra=c(rep('L',4),rep('C',23),
+          rep('D',16),rep('O',7))
> data=factor(amostra,c('L','C','D','O'),
+ labels=c("Leitura", "Cinema",
+              "Desporto", "Outros"))
> Tabela=table(data); Tabela
data
 Leitura    Cinema Desporto    Outros
       4        23       16         7
> Tabela.rel=prop.table(Tabela); Tabela.rel
data
 Leitura    Cinema Desporto    Outros
    0.08      0.46     0.32      0.14
```

- **Gráficos de barras:** barras separadas, de igual largura, cuja altura é proporcional à frequência (absoluta e relativa) da categoria correspondente.




```
>barplot(Tabela)
```

- **Gráficos circulares:** círculo dividido em sectores cujo ângulo ao centro (e área) é proporcional à frequência da categoria correspondente.



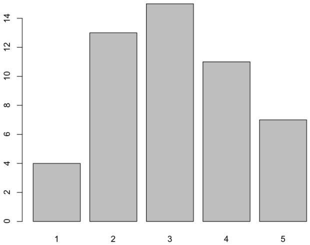

```
> pie(Tabela)
```

Dados Quantitativos Discretos

- Tabela de frequências.

agregado familiar	1	2	3	4	5	Total
número pessoas	4	13	15	11	7	50

- Gráfico de barras.



- Diagrama de caixa de bigodes (adiante).
- Gráfico de caule-e-folhas (a seguir).

Dados Quantitativos Contínuos

- Distribuição de frequências.

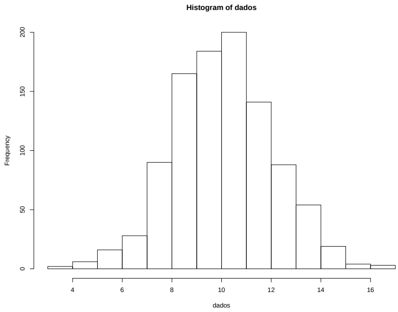
dist. etária	]15, 20]	]20, 25]	]25, 30]	]30, 35]	]35, 40]	]40, 45]	]45, 50]	]50, 55]	Total
n° pessoas	8	10	8	7	4	7	4	2	50

Classe de uma variável quantitativa contínua: é um intervalo da forma $[a, b[$ ou $]a, b]$ que representa um conjunto de valores que a variável pode tomar.
Amplitude da classe: é a distância entre o limite superior e o limite inferior da classe, i. e., amplitude $= b - a$.
Centro, marca, ponto médio ou valor característico da classe: é o ponto médio da classe, i. e., $\frac{a+b}{2}$.

Regras classificar dados quantitativos contínuos:

- Regra de Sturges: $c = int[1 + 3.3 \log_{10} n]$, c é o número de classes, n o número de observações, e $int(x)$ é a parte inteira de x;
- ou $c = int[\sqrt{n}]$;
- Amplitude das classes: $\frac{\text{maior observ. } - \text{ menor observ.}}{c}$;
- Arredondar a amplitude das classes para um número conveniente, alterando ligeiramente o valor inicial do número de classes;
- O limite inferior da primeira classe é qualquer número conveniente que seja inferior à primeira observação.

- **Histograma:** marcam-se as classes no eixo horizontal, as frequências no eixo vertical, e usam-se barras de área proporcional à frequência da classe correspondente. As barras são contíguas.



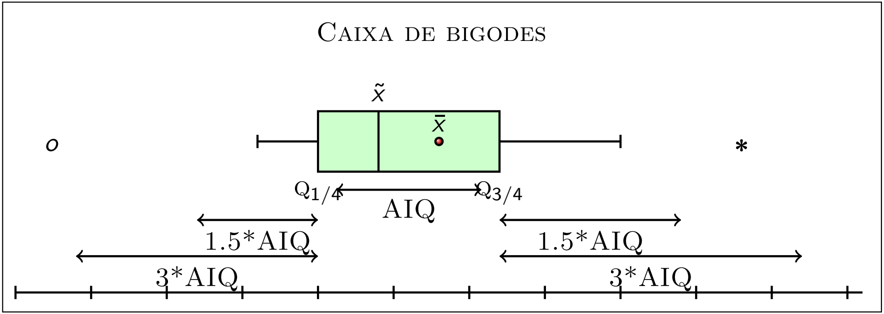

```
> dados=rnorm(1000,10,2)
> hist(dados)
```

- Usa o método de Sturges por defeito mas pode-se personalizar as classes (pesquisar breaks e cut)

- Gráfico de caule-e-folhas.

```
> stem(dados)

  The decimal point is at the |

   3 | 47
   4 | 14779
   5 | 0222455667778899
   6 | 00111236666666778888888999
   7 | 00000000111111111222222222222333333344444445555555555666666677777777+7
   8 | 00000000000111111111111111111222222222222222222333333333333333333333+78
   9 | 00000000000000000000111111111111111111111222222222222222233333333333+103
  10 | 00000000000000000000000011111111111111111222222222222222222333333333+124
  11 | 00000000000000011111111111111111111222222222222222233333333333333333+66
  12 | 00000000000000011111111111222222222222222233333344444444555555555566+12
  13 | 0000000011111111112222222333333444445556677778888899999
  14 | 122233344445566778
  15 | 00357
  16 | 256
```

Diagrama de caixa de bigodes



- Permite interpretar a localização e dispersão dos dados.
- A *caixa* tem como limites o 1º e o 3º quartis.
- Barreiras inferior e superior: $BI = q_{1/4} - 1.5\,AIQ \qquad BS = q_{3/4} + 1.5\,AIQ$
- Limites dos *bigodes*: são marcados pelo menor valor do conjunto de dados que não é menor que a BI, e pelo maior valor do conjunto de dados que não é maior que a BS, respetivamente.
- outliers: valores que são inferiores a BI ou superiores a BS.
- outliers severos: valores que são inferiores a BI* ou superiores a BS*:

  $BI^* = q_{1/4} - 3\,AIQ \qquad BS^* = q_{3/4} + 3\,AIQ$

```
> dados1=rchisq(100,.2, df=7)
> dados2=rchisq(100,5, df=7)
> classes=c("dados1","dados2")
> boxplot(dados1,dados2,names=classes,col=c(4,2))
```