Questões:

1. Os seguintes dados referem-se ao peso das bagagens individuais numa amostra de 90 passageiros que embarcaram no aeroporto de Lisboa num dado voo.

peso da bagagem (kg)	$[0,5[$	$[5,10[$	$[10,15[$	$[15,20[$	$[20,25[$	$[25,30]$
número de bagagens	5	5	A	30	B	10

[0.5] (a) Indique e classifique a variável em estudo.

1. Os dados referem-se ao peso das bagagens individuais numa amostra de dimensão $n = 90$.

   (a) Variável em estudo: peso da bagagem, em kg
   Classificação da variável em estudo: Variável Quantitativa Contínua

[1.0] (b) Sabendo que 20% das bagagens desse voo tem peso inferior a 15 kg, calcule os valores de A e B.

(b) número de bagagens com peso inferior a 15 kg $= 5 + 5 + A = 10 + A$
a proporção de bagagens com peso inferior a 15 kg $= \frac{10+A}{90}$
logo

$$\frac{10 + A}{90} = 0.20 \Leftrightarrow A = 8$$

Como a amostra tem dimensão $n = 90$, tem-se

$$5 + 5 + A + 30 + B + 10 = 90 \underset{A=8}{\Leftrightarrow} 5 + 5 + 8 + 30 + B + 10 = 90 \Leftrightarrow B = 32$$

[1.0] (c) Considere $A = B = 20$, construa a tabela de frequências completa.

(c) $A = B = 20$.
Tabela de frequências:

i	Peso da bagagem (em kg) Classe - c_i	Freq. Absoluta n_i	Freq. Relativa f_i	Freq. Abs. Acumulada N_i	Freq. Rel. Acumulada F_i
1	$[0,5[$	5	$\frac{5}{90} = 0.056$	5	0.056
2	$[5,10[$	5	$\frac{5}{90} = 0.056$	$5 + 5 = 10$	$0.056 + 0.056 = 0.112$
3	$[10,15[$	20	$\frac{20}{90} = 0.222$	$10 + 20 = 30$	$0.112 + 0.222 = 0.334$
4	$[15,20[$	30	$\frac{30}{90} = 0.333$	$30 + 30 = 60$	$0.334 + 0.333 = 0.667$
5	$[20,25[$	20	$\frac{20}{90} = 0.222$	$60 + 20 = 80$	$0.667 + 0.222 = 0.889$
6	$[25,30]$	10	$\frac{10}{90} = 0.111$	$80 + 10 = 90$	$0.889 + 0.111 = 1$
		$n = 90$	1		

2. O RMS Titanic foi um navio britânico construído em Belfast, na Irlanda do Norte, que teve a sua viagem inaugural (e única) em 10 de Abril de 1912. No caminho entre a cidade inglesa de Southampton e a cidade de Nova York colidiu com um icebergue e naufragou nas águas geladas do Atlântico norte às 23h40 do dia 15 de abril de 1902. Estima-se que o navio levava 2224 pessoas a bordo, entre passageiros e tripulação, e mais de 1500 pessoas morreram em decorrência do naufrágio. No ficheiro titanic0.txt (Moodle) tem informação sobre alguns dos passageiros do Titanic. Nesse ficheiro encontram-se os seguintes campos:

- PassengerId = número de identificação do passageiro
- Survived = se o passageiro sobreviveu (0 = Não; 1 = Sim)
- Pclass = Classe do camarote do passageiro (1 = 1ª classe; 2 = 2ª classe; 3 = 3ª classe)
- Sex = género (female = feminino, male = masculino)
- Age: idade do passageiro (em anos)
- Fare = Preço da passagem (em libras)

[1.5] (a) Caso seja possível, identifique a População e a Amostra indicando as suas dimensões, a unidade estatística, as variáveis estatísticas e os dados estatísticos, classifcando-os.

2. (a) População: todos os passageiros do Titanic

Dimensão da População: Não se sabe, estima-se que eram 2224 pessoas a bordo, entre passageiros e tripulação

Amostra: os passageiros no ficheiro titanic0.txt

Dimensão da Amostra: $n = 712$ passageiros

Unidade estatística: passageiros

Variável estatística: Survived

Dados estatísticos: Não, Sim

Classificação: Qualitativa nominal

Variável estatística: Pclass

Dados estatísticos: 1ª classe, 2ª classe, 3ª classe

Classificação: Qualitativa ordinal

Variável estatística: Sex

Dados estatísticos: feminino, masculino

Classificação: Qualitativa nominal

Variável estatística: Age

Dados estatísticos: qualquer número maior que zero

Classificação: Quantitativa contínua

Variável estatística: Fare

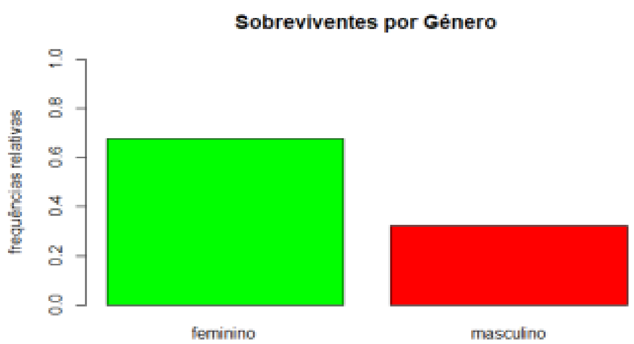
Dados estatísticos: qualquer número maior que zero

Classificação: Quantitativa contínua

[1.5] (b) Qual o género que sobreviveu mais e com que percentagem? Represente essa informação graficamente.

(b) O género feminino foi o que sobreviveu mais, 67.71% dos sobreviventes era do género feminino.

Representação gráfica: gráfico de barras (o gráfico podia estar só no script)

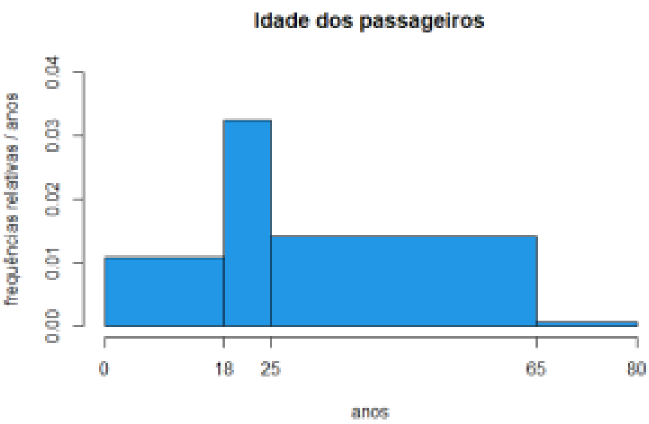


[2.0] (c) Para fazer uma análise descritiva dos dados pretende-se classificar os passageiros de acordo com a sua idade. Os passageiros com idade máxima de 18 anos são classificados de "crianças", os passageiros com idade entre os 18 e 25 anos (inclusive) são classificados de "jovens", a classificação de "adultos"refere-se aos passageiros com idade entre os 25 anos e os 65 anos (inclusive), os restantes são classificados de "idosos". Construa a tabela de frequências completa usando a classificação definida para a variável idade e represente-a graficamente.

(c) Tabela de frequências:

i	Classificação passageiros x_i	Idade (em anos) Classe - c_i	Freq. Absoluta n_i	Freq. Relativa f_i	Freq. Abs. Acumulada N_i	Freq. Rel. Acumulada F_i
1	crianças	$]0, 18]$	139	0.195	139	0.195
2	jovens	$]18, 25]$	162	0.228	301	0.423
3	adultos	$]25, 65]$	403	0.566	704	0.989
4	idosos	$]65, 80]$	8	0.011	712	1
			$n = 712$	1		

Representação gráfica: histograma (o gráfico podia estar só no script)



[1.5] (d) Faça uma análise descritiva dos dados referentes ao preço da passagem calculando as medidas de localização central e dispersão adequadas e, caso considere adequado, construa o correspondente diagrama de extremos e quartis indicando se existem (e quantos) "outliers" moderados e severos.

(d) Medidas de localização central e dispersão:

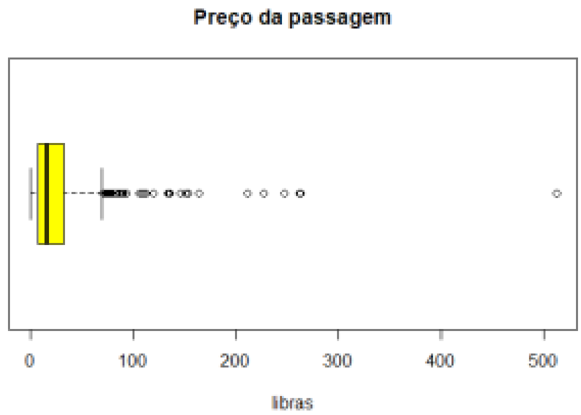
Medidas de localização central

- moda = 13 libras
- média = 34.567 libras
- mediana = 15.645 libras

Medidas de dispersão

- amplitude total = 512.329 libras
- amplitude interquartil = 24.95 libras
- variância = 2802.5 libras^2
- desvio padrão = 52.939 libras
- coeficiente de variação = 153.147%

Diagrama de extremos e quartis (com a indicação de "outliers" a partir dos moderados) (o gráfico podia estar só no script)



Só há dados considerados "outliers" no extremo superior dos dados. Há 95 dados considerados "outliers", 49 são considerados "outliers" moderados e 46 são considerados "ouliers" severos.

3. Um posto de gasolina tem uma loja de conveniência que tem um pequeno bar onde se vende café. O número de cafés vendidos diariamente nesse bar é uma variável aleatória discreta X com a seguinte função de probabilidade:

x	50	100	150	200
$f(x)$	0.2	a	b	0.1

[1.5] (a) Determine os valores a e b sabendo que, em média, são vendidos diariamente 115 cafés.

3. Seja $X-$ número de cafés vendidos diariamente nesse bar, uma variável aleatória discreta.

(a) Como, em média, são vendidos diariamente 115 cafés, então

$$E[X] = 115 \Leftrightarrow 50 \times 0.2 + 100 \times a + 150 \times b + 200 \times 0.1 = 115 \Leftrightarrow 100a + 150b = 85$$

Por outro lado, como $f(x)$ é função de probabilidade, sabe-se que

$$\sum_x f(x) = 1 \Leftrightarrow 0.2 + a + b + 0.1 = 1 \Leftrightarrow a + b = 0.7 \Leftrightarrow b = 0.7 - a$$

Voltando à primeira expressão

$$100a + 150 \times (0.7 - a) = 85 \Leftrightarrow 100a + 105 - 150a = 85 \Leftrightarrow a = 0.4$$

Desta forma,

$$b = 0.7 - 0.4 \Leftrightarrow b = 0.3$$

[1.5] (b) Considere $a = 0.3$ e $b = 0.4$. Qual a probabilidade do número de cafés vendidos num dado dia ser superior a 150 sabendo que já foram vendidos pelo menos 100 cafés?

(b) Considerando $a = 0.3$ e $b = 0.4$ tem-se

x	50	100	150	200
$f(x)$	0.2	0.3	0.4	0.1

Assim:

$$P(X > 150 | X \geq 100) = \frac{P(X > 150 \wedge X \geq 100)}{P(X \geq 100)} = \frac{P(X > 150)}{P(X \geq 100)} =$$

$$= \frac{f(200)}{f(100) + f(150) + f(200)} = \frac{0.1}{0.4 + 0.3 + 0.1} = \frac{0.1}{0.8} = 0.125$$

[2.0] (c) O posto de gasolina é abastecido uma vez por semana e as vendas semanais de gasolina, em milhares de litros, é uma variável aleatória contínua Y com função densidade de probabilidade dada por:

$$f(y) = \begin{cases} y-1 & , \quad 1 \leq y < 2 \\ 3-y & , \quad 2 \leq y < 3 \\ 0 & , \quad \text{caso contrário} \end{cases} .$$

Calcule a variância do lucro semanal sabendo que o lucro semanal é dado por $2Y-1$ e que, em média, semanalmente são vendidos 2000 litros de gasolina.

(c) Seja $Y-$ vendas semanais de gasolina, em milhares de litros, uma variável aleatória contínua

Pretende-se a variância do lucro semanal, ou seja, $V[2Y-1]$. Utilizando as propriedades da variância, vem que:

$$V[2Y-1] = 2^2 \times V[Y] = 4 \times \left(E\left[Y^2\right] - E^2\left[Y\right]\right)$$

Como $E\left[Y\right] = 2$, falta calcular $E\left[Y^2\right]$. Assim

$$E\left[Y^2\right] = \int_{-\infty}^{+\infty} y^2 f\left(y\right) dy =$$

$$= \int_{-\infty}^{1} y^2 \times 0 dy + \int_{1}^{2} y^2 \times \left(y-1\right) dy + \int_{2}^{3} y^2 \times \left(3-y\right) dy + \int_{3}^{+\infty} y^2 \times 0 dy =$$

$$\underset{(*)}{=} 0 + \int_{1}^{2} \left(y^3 - y^2\right) dy + \int_{2}^{3} \left(3y^2 - y^3\right) dy + 0 = \left[\frac{y^4}{4} - \frac{y^3}{3}\right]_1^2 + \left[\frac{3y^3}{3} - \frac{y^4}{4}\right]_2^3 =$$

$$= \left(\frac{2^4}{4} - \frac{2^3}{3}\right) - \left(\frac{1^4}{4} - \frac{1^3}{3}\right) + \left(3^3 - \frac{3^4}{4}\right) - \left(2^3 - \frac{2^4}{4}\right) = \frac{25}{6} = 4.1667$$

$(*)$ os seguintes cálculos podem estar apenas no script

Desta forma,

$$V[2Y-1] = 4 \times \left(E\left[Y^2\right] - E^2\left[Y\right]\right) = 4 \times \left(\frac{25}{6} - 2^2\right) = \frac{2}{3}$$

4. De 1000 declarações de IRS, sabe-se que 100 apresentam erros. Um fiscal das finanças selecionou 20 declarações, ao acaso, para analisar. Sabe-se que o número de declarações analisadas por hora tem distribuição de Poisson com variância 3.

[1.5] (a) Qual a probabilidade do fiscal vir a encontrar mais do que uma declaração com erro?

4. De um total de 1000 declarações de IRS, das quais se sabe que 100 apresentam erros, foram selecionadas aleatoriamente 20 declarações. Sabe-se que o número de declarações analisadas por hora tem distribuição de Poisson com variância 3.

(a) Seja X a variável aleatória discreta:

$X \rightarrow$ número de declarações analisadas com erro, em 20. $X \sim B\left(20, 0.1\right)$ pois

$n = 20$ declarações selecionadas

$p = P\left(\text{Sucesso}\right) = P\left(\text{declaração conter erros}\right) = \frac{100}{1000} = 0.1$

então

$$P\left(X > 1\right) = 1 - P\left(X \leq 1\right) = 1 - F\left(1\right) = 1 - 0.3917 = 0.6083$$

[1.5] (b) Sabendo que os funcionários das repartições de finanças trabalham 6 horas por dia, será razoável admitir que o funcionário conseguirá, num dia de trabalho, analisar pelo menos um quarto das declarações selecionadas?

(b) Seja Y a variável aleatória discreta:

$Y \rightarrow$ número de declarações analisadas em 1 hora. $Y \sim P(3)$ pois

$Y \sim P(\lambda) \qquad \text{com} \qquad V(Y) = \lambda = 3$

então

$Y' \rightarrow$ número de declarações analisadas em 6 horas. $Y' \sim P(18)$ pois

$$\begin{array}{lcr} 1 \text{ hora} & \mapsto & \lambda = 3 \\ 6 \text{ horas} & \mapsto & \lambda = 3 \times 6 = 18 \end{array}$$

Pretende-se

$$P\left(Y' \geq \frac{1}{4} \times 20\right) = P(Y' \geq 5) = 1 - P(Y' < 5) \underset{\text{v.a. discreta}}{=} 1 - P(Y' \leq 4) = 1 - F(4) =$$
$$= 1 - 0.0001 = 0.9999$$

(c) O tempo que um contribuinte demora a preencher a declaração de IRS é uma variável aleatória que se admite ter distribuição normal com média 25.6 minutos e desvio padrão igual a 1.6 minutos.

[1.5] i. Qual a probabilidade do tempo que um contribuinte demora a preencher a declaração de IRS estar entre 24.256 e 27.472 minutos?

[1.5] ii. Qual o tempo máximo que 1.79% dos contribuintes demoram a preencher a declaração de IRS?

(c) Seja W a variável aleatória contínua:

$W \rightarrow$ tempo que demora a preencher a declaração de IRS, em minutos.

$W \sim N(25.6, 1.6)$ pois $E[W] = \mu = 25.6$ minutos e $\sqrt{V[W]} = \sigma = 1.6$ minutos.

i. $P(24.256 < W < 27.472) \underset{\text{v.a. contínua}}{=} F(27.472) - F(24.256) = 0.8790 - 0.2005 = 0.6785$

ii. Pretende-se determinar m tal que:

$$P(W \leq m) = 0.0179 \Leftrightarrow F(m) = 0.0179 \Leftrightarrow m = 22.24 \text{ minutos}$$