Exercícios 28 a 32 – Resoluções

28.

a) *** Excluído do Programa ajustado ***

b) Os dados recolhidos são observações (57) de uma v.a. X que representa o n.º de palavras utilizadas num poema, escolhido ao acaso, no Cancioneiro da Ajuda (que inclui 310 composições poéticas, todas elas Cantigas de Amor); sendo que X representa uma contagem, então é uma v.a. discreta e, por essa razão os dados são discretos (por isso também se chama observações aos dados).
As representações adequadas para dados discretos são o **gráfico de barras** e o **diagrama de barras**.
Para construir o gráfico de barras é necessário começar por ordenar os dados.
Uma forma expedita de o fazer consiste no seguinte: uma observação rápida do conjunto dos dados (amostra) permite perceber de imediato que todos são da ordem das dezenas. Então começa-se por dividir os dados em dois algarismos – o das dezenas e o das unidades. Constrói-se uma coluna com os algarismos das dezenas – a que se pode chamar "caules" – e posteriormente os algarismos das unidades – a que se pode chamar "folhas" – de cada dado vão ser "pendurados" nos algarismos das respectivas dezenas. Por exemplo, pendura-se o algarismo 8 no algarismo 6 das dezenas, para representar 68; pendura-se o algarismo 3 no algarismo 4 das dezenas, para representar 43 e assim sucessivamente. Repetindo o processo para todos os dados, obtém-se o seguinte esquema (onde os dados foram sendo colocados por coluna da tabela onde são fornecidos):

1	2	9	6	2	2													
2	2	7	7	5	4	3	7	7	8	5	2	3	1	4	5	8	8	8	3
3	6	0	3	0	6	1	2	1	8	8								
4	3	9	9	2	5	6	4	3	2	2	9	3	7					
5	0	1	1	7														
6	8	9	5	3														
7	9	4																

É, agora, muito mais fácil ordenar as folhas de cada caule e esse processo conduz ao esquema que se segue:

1	2	2	2	6	9													
2	1	2	2	3	3	3	4	4	5	5	5	7	7	7	7	8	8	8	8
3	0	0	1	1	2	2	6	6	8	8								
4	2	2	2	3	3	3	4	5	6	7	9	9	9					
5	0	1	1	7														
6	3	5	8	9														
7	4	9																

Voltando a juntar esquerda para a direita, e de cima para baixo, caules e folhas, obtém-se o conjunto de dados já ordenados (amostra ordenada):

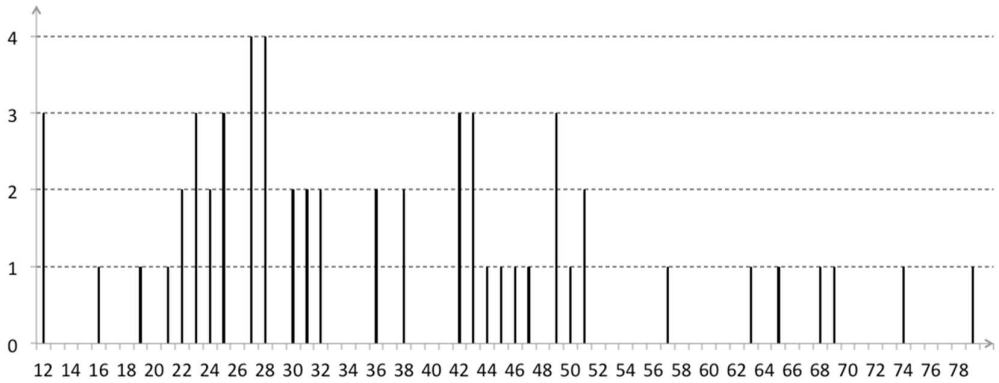
12, 12, 12, 16, 19, 21, 22, 22, 23, ..., 63, 65, 68, 69, 74, 79

A mostra ordenada contém 31 valores distintos. Designando por y_i o i-ésimo valor distinto ($y_1 = 12$, $y_2 = 16$, $y_3 = 19$, ..., $y_{30} = 74$, $y_{31} = 79$) e por n_i a correspondente frequência absoluta – o n.º de vezes que aparece na amostra ($n_1 = 3$, $n_2 = 1$, $n_3 = 1$, ..., $n_{30} = 1$, $y_{31} = 1$), pode construir-se a seguinte tabela de frequências simples:

yi	12	16	19	21	22	23	24	25	27	28	30	31	32	36	38	42	…
ni	3	1	1	1	2	3	2	3	4	4	2	2	2	2	2	3	

…	yi	43	44	45	46	47	49	50	51	57	63	65	68	69	74	79
	ni	3	1	1	1	1	3	1	2	1	1	1	1	1	1	1

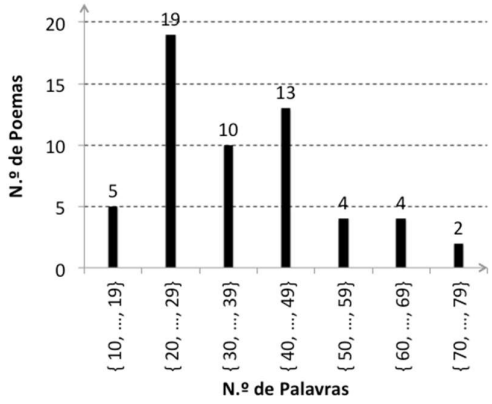
Um gráfico de barras baseado directamente na tabela anterior, tem o seguinte aspecto:



Facilmente se percebe que o gráfico anterior é muito pouco "elucidativo", já que representa muitos valores distintos e com frequências associadas muito baixas. Nesta situação é preferível agrupar os dados em "classes", juntando os poemas que, embora não tendo exactamente o mesmo n.º de palavras, são "parecidos" no que diz respeito a essa característica. Não existe uma forma de escolher esse agrupamento que seja a "melhor". A escolha depende do bom senso na procura da "excelência gráfica". De seguida, apresenta-se a tabela de frequências simples associada a uma classificação possível, classes com dimensão 10, que parece "sensata".

Classe C_i	n_i
{ 10, …, 19}	5
{ 20, …, 29}	19
{ 30, …, 39}	10
{ 40, …, 49}	13
{ 50, …, 59}	4
{ 60, …, 69}	4
{ 70, …, 79}	2
	57

Baseado na tabela anterior pode construir-se um **diagrama de barras** (já não é exactamente um gráfico porque o eixo horizontal deixou de ter abcissas para passar a ter classes, como acontece na representação de dados qualitativos) que tem o seguinte aspecto:



A análise do diagrama anterior permite estabelecer conclusões mais abrangentes e úteis como, por exemplo: os poemas mais frequentes têm entre 20 e 29 palavras; os poemas com um n.º de palavras entre 70 a 79 são os menos frequentes; a distribuição do número de palavras por poema aparenta ser enviesada, etc..

c) Média: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 2093/57 \approx 36.7$ (mais uma casa decimal do que os dados)

Variância: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] = [90949 - 2093^2/57]/56 \approx$

≈ 251.71 (mais duas casas decimais do que os dados)

Mediana: n = 57 é ímpar $\Rightarrow Q_{1/2} = x_{(58/2)} = x_{(29)} = 32$ (elemento na posição 32 da amostra ordenada)

1º Quartil: n/4 = 14.25 é não inteiro $\Rightarrow Q_{1/4} = x_{(15)} = 25$ (elemento na posição 15 da amostra ordenada)

3º Quartil: 3n/4 = 42.75 é não inteiro $\Rightarrow Q_{3/4} = x_{(43)} = 46$ (elemento na posição 43 da amostra ordenada)

Amplitude inter-quartis: $H = Q_{3/4} - Q_{1/4} = 46 - 25 = 21$

d) Para representar a caixa-com-bigodes (em inglês *boxplot* ou *box-and-whiskers*) é necessário calcular as barreiras, ou limites, de *outliers* – Limite Inferior (LI) e Limite Superior (LS) – bem como os pontos adjacentes – Adjacente Inferior (AI) e Adjacente Superior (AS) – aos mesmos.

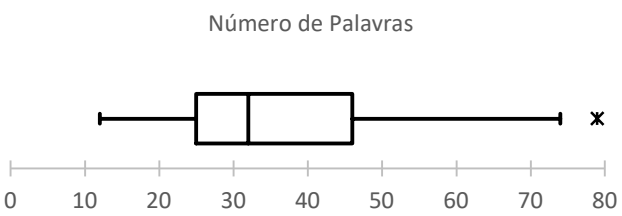
$LI = Q_{1/4} - 1.5\,H = 25 - 1.5\times21 = -6.5 \rightarrow -6.5 < x_{(1)} \Rightarrow$ Não existem *outliers* inferiores

$LS = Q_{3/4} + 1.5\,H = 46 + 1.5\times21 = 77.5 \rightarrow 77.5 < 79 \Rightarrow x_{(57)} = 79$ é um *outlier* superior

O ponto AI é o menor valor da amostra que pertence ao intervalo [LI, LS]: $AI = x_{(1)} = 12$

O ponto AS é o maior valor da amostra que pertence ao intervalo [LI, LS]: $AS = x_{(56)} = 74$

Representação dos dados em *boxplot*:



A forma da *boxplot* sugere que a distribuição do n.º de palavras num poema é assimétrica à direita (ou positiva) já que a barra que a localização da mediana está para a esquerda do centro da caixa. Esta característica já era esperada uma vez que a média é superior à mediana amostral ($\bar{x} > Q_{1/2}$) esperando-se, portanto, uma "cauda" direita da distribuição mais "pesada" (estende-se mais) do que a esquerda.

29.

a) Os dados são observações de uma v.a. contínua (tempo de vida), pelo que são contínuos.

i) *** Excluído do Programa ajustado ***

ii) Como a v.a. em estudo (população que deu origem às observações) é contínua assume um conjunto de valores distintos que é infinito e não numerável (contável). Assim, não faz sentido representar as observações através de um gráfico de barras, pois uma v.a. contínua não tem f.m.p. e sim f.d.p.. A "imagem estatística" da f.d.p. da v.a. que foi observada é a representação gráfica adequada para dados contínuos e que permite estabelecer hipóteses sobre a distribuição da população subjacente. Essa representação é o **histograma**, onde os dados são distribuídos por intervalos disjuntos e adjacentes – "classes".

É usual construir as classes por forma a que todas tenham a mesma amplitude h (a não ser que exista alguma justificação específica para que se proceda de modo distinto). Segundo este princípio, se os dados forem distribuídos por k classes, C_1, C_2, ..., C_k, então para garantir que as classes em conjunto contém todos os dados, basta que o limite inferior de C_1 seja menor ou igual do que o mínimo da amostra, $x_{(1)}$, e que o limite superior de C_k seja maior ou igual do que o máximo da amostra, $x_{(n)}$. Ou seja, pretende distribuir-se a amplitude amostral, $r = x_{(n)} - x_{(1)}$, por k classes com amplitude comum h; para tal, basta tomar h tão pequeno quanto possível e tal que

$$h \geq (x_{(n)} - x_{(1)})/k\ .$$

Resta decidir qual o n.º de classes, k, a considerar. Não existe um valor que seja o "melhor" para atribuir a k; mas uma regra comum e aceite como bastante boa numa série de situações distintas é devida a Sturges, segundo a qual se deve tomar

$$k = [\log_2 n]+1 ,$$

onde [x] representa a parte inteira de x. Excepto nos casos em que n é exactamente igual a uma potência de 2, a regra de Sturges pode ser enunciada do seguinte modo: tome-se k como o menor inteiro tal que

$$2^k \geq n .$$

Sempre que n é exactamente igual a uma potência de 2, condição anterior conduz a exactamente uma classe a menos do que a regra de Sturges. No entanto, a regra é meramente indicadora de um valor em torno do qual é adequado tomar o valor de k, pelo que, frequentemente se deve experimentar uma classe a mais ou uma classe a menos para se perceber qual a representação que parece descrever melhor a distribuição subjacente. Assim, é costume simplificar e utilizar a condição $2^k \geq n$ para a escolha inicial de k.

Vejamos como se faz, segundo o exposto, a classificação dos dados do problema presente.

Vai ser necessário ter a amostra ordenada pelo que, e embora não seja obrigatório, pode fazer-se uso do procedimento descrito no exercício anterior, tomando como caules os algarismos das unidades e como folhas os algarismos das décimas, chegando-se à seguinte representação esquemática das observações que compõem a amostra:

0	2	2	2	3	3	4	5	7
1	0	2	3	5	5	8		
2	0	3	5					
3	0	3						
4	0	5	7					
5	0	5	6	9				
6	0	0	0	5				

n = 30: menor k tal que $2^k \geq 30 \Rightarrow$ k = 5

$x_{(1)} = 0.2$, $x_{(30)} = 6.5 \Rightarrow r = x_{(30)} - x_{(1)} = 6.3$

$r/k = 6.3/5 = 1.26 \Rightarrow$ h = 1.26

As classes C_j vão ser consideradas da forma $C_j = [\ a_j, a_j+h[$, fechadas à esquerda e abertas à direita, j = 1, ..., k–1, tomando $a_1 = x_{(1)}$; a última classe será da forma $C_k = [\ a_k, a_k+h]$, fechada também à direita para incluir $x_{(n)}$ que coincide com a_k+h (o que é resultado de não se ter feito qualquer arredondamento a r/k). Deste modo, os dados vão ser distribuídos pelas classes:

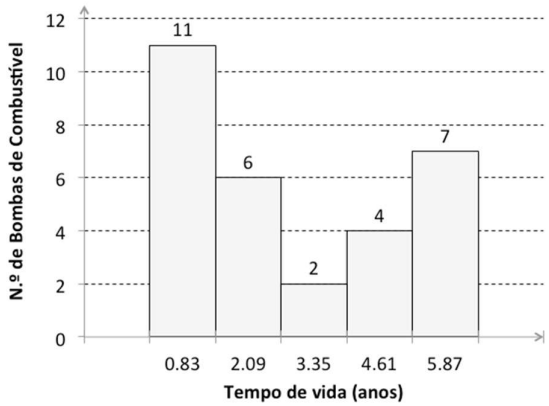
$C_1 = [\ 0.20, 1.46[$; $C_2 = [\ 1.46, 2.72[$; $C_3 = [\ 2.72, 3.98[$; $C_4 = [\ 3.98, 5.24[$ e $C_5 = [\ 5.24, 6.50]$.

A cada classe C_j associa-se uma frequência absoluta n_j que não é mais do que o n.º de observações que pertencem a C_j, obtendo-se a seguinte tabela de frequências simples para os dados classificados:

Classe C_j	m_j	n_j
[0.20, 1.46[	0.83	11
[1.46, 2.72[	2.09	6
[2.72, 3.98[	3.35	2
[3.98, 5.24[	4.61	4
[5.24, 6.50]	5.87	7
		30

Note-se que a tabela anterior inclui uma coluna onde estão os pontos médios das classes, m_j, que frequentemente são utilizados como os representantes das classe na representação gráfica do histograma.

Um histograma para os dados observados obtém-se associando a cada classe uma barra com altura igual à correspondente frequência absoluta e representando classes *versus* frequências num referencial, obtendo-se:



b) Calcule as características amostrais: média, mediana, quartis e desvio-padrão.

Média: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 83.9/30 \approx \underline{2.797}$ (pelo menos mais uma casa decimal do que os dados)

Mediana: n = 30 é par $\Rightarrow Q_{1/2} = [x_{(30/2)} + x_{(30/2+1)}]/2 = [x_{(15)} + x_{(16)}]/2 = (2.0 + 2.3)/2 = \underline{2.15}$

1º Quartil: n/4 = 7.5 é não inteiro $\Rightarrow Q_{1/4} = x_{(8)} = \underline{0.7}$

3º Quartil: 3n/4 = 22.5 é não inteiro $\Rightarrow Q_{3/4} = x_{(23)} = \underline{5.0}$

Desvio-padrão: $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]} = \sqrt{(378.51 - 83.9^2/30)/29} \approx$

$\approx \underline{2.23}$ (mais duas casas decimais do que os dados)

c) Amplitude inter-quartis: $H = Q_{3/4} - Q_{1/4} = 5.0 - 0.7 = 4.3$; $1.5\ H = 6.45$

Barreiras de *outliers*:

$LI = Q_{1/4} - 1.5\ H = 0.7 - 6.45 = -5.75 \rightarrow -5.75 < x_{(1)} \Rightarrow$ Não existem *outliers* inferiores; $AI = x_{(1)} = 0.2$

$LS = Q_{3/4} + 1.5\ H = 5.0 + 6.45 = 11.45 \rightarrow 11.45 > x_{(n)} \Rightarrow$ Não existem *outliers* superiores; $AS = x_{(30)} = 6.5$

Representação dos dados em *boxplot*:

30.

a) Médias: $\overline{x}_I = 149.7/10 = \underline{14.97}$; $\overline{x}_{II} = 155.6/10 = \underline{15.56}$

Variâncias: $s_I^2 = (2257.57 - 149.7^2/10)/9 \approx \underline{1.8401}$; $s_{II}^2 = (2432.98 - 155.6^2/10)/9 \approx \underline{1.3160}$

Desvios-padrão: $s_I = \sqrt{s_I^2} \approx \sqrt{1.8401} \approx \underline{1.36}$; $s_{II} = \sqrt{s_{II}^2} \approx \sqrt{1.316} \approx \underline{1.15}$

b) Medianas: $Q^I_{1/2} = [x_{(5)} + x_{(6)}]/2 = \underline{14.75}$; $Q^{II}_{1/2} = [x_{(5)} + x_{(6)}]/2 = \underline{16.16}$

1º.s Quartis: $Q^I_{1/4} = x_{(3)} = \underline{13.8}$; $Q^{II}_{1/4} = x_{(3)} = \underline{14.4}$

3º.s Quartis: $Q^{I}_{3/4} = x_{(8)} = 16.2$; $Q^{II}_{3/4} = x_{(8)} = 16.3$

Amplitudes inter-quartis: $H_I = Q^{I}_{3/4} - Q^{I}_{1/4} = 2.4$; $1.5\ H_I = 3.6$

$H_{II} = Q^{II}_{3/4} - Q^{II}_{1/4} = 1.9$; $1.5\ H_{II} = 2.85$

Barreiras de *outliers*:

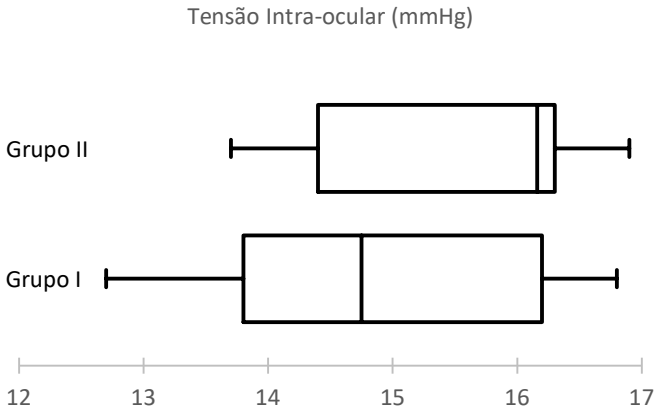
$LI^{I} = Q^{I}_{1/4} - 1.5\ H_I = 10.2 < x^{I}_{(1)} \Rightarrow$ Não existem *outliers* inferiores no Grupo I; $AI^{I} = x^{I}_{(1)} = 12.7$

$LS^{I} = Q^{I}_{3/4} + 1.5\ H_I = 19.8 > x^{I}_{(n)} \Rightarrow$ Não existem *outliers* superiores no Grupo I; $AS^{I} = x^{I}_{(10)} = 16.8$

$LI^{II} = Q^{II}_{1/4} - 1.5\ H_{II} = 11.55 < x^{II}_{(1)} \Rightarrow$ Não existem *outliers* inferiores no Grupo II; $AI^{II} = x^{II}_{(1)} = 13.7$

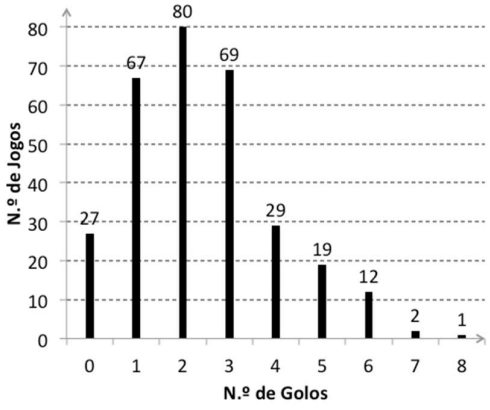
$LS^{II} = Q^{II}_{1/4} + 1.5\ H_{II} = 19.15 > x^{II}_{(n)} \Rightarrow$ Não existem *outliers* superiores no Grupo II; $AS^{II} = x^{II}_{(10)} = 16.9$

c) Representação dos dois grupos de dados em *boxplot*s paralelas:



31.

a) Os dados recolhidos são observações de uma população discreta (n.º de golos num jogo, escolhido ao acaso, na 1.ª Liga Profissional de Futebol portuguesa – é uma contagem), pelo que são discretos. A representação mais adequada é utilizando um gráfico de barras e, como os dados já são apresentados numa tabela de frequências simples, a sua construção é directa, obtendo-se:



b) Para responder à questão é necessário calcular o limite superior das barreiras de *outliers* (basta este já que os valores em questão estão na cauda direita da distribuição amostral e não poderiam, portanto, ser *outliers* inferiores).

1º Quartil: $n/4 = 306/4 = 76.5 \Rightarrow Q_{1/4} = x_{(77)} = 1$

3º Quartil: $3n/4 = 306/4 = 229.5 \Rightarrow Q_{3/4} = x_{(230)} = 3$

Amplitude inter-quartis: $H = Q_{3/4} - Q_{1/4} = 2$; $1.5\,H = 3$

Barreira superior de *outliers*:

$LS = Q_{3/4} + 1.5\,H = 6 \Rightarrow$ Todos os jogos com n.º de golos superiores a 6 são *outliers*

c) Determine a média e o desvio padrão do número de golos por jogo.

Média: $\bar{x} = 739/306 \approx \underline{2.4}$

Desvio-padrão: $s = \sqrt{(2541 - 739^2/306)/305} \approx \underline{1.6}$

32. Sabe-se que $X_i \cap \mathcal{N}(\mu, \sigma = 0.5)$, $i = 1, \ldots, n$. A média das n pesagens é dada por $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Ora, sabe-se que a soma de v.a.'s com distribuição Normal tem ainda distribuição Normal. Por outro lado, se multiplicarmos uma v.a. com distribuição Normal por uma constante não nula, o resultado é uma v.a. também com distribuição Normal, com os respectivos valor esperado e variância. Deste modo, $\bar{X}$ tem distribuição Normal. Sendo o valor esperado um operador linear, vem que:

$$\mu_{\bar{X}} = \mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathrm{E}(X_i)}_{\mu} = \frac{1}{n}n\mu = \mu,$$

ou seja, o valor esperado da média aleatória é sempre igual ao valor esperado da população, independentemente da distribuição desta.

Uma vez que as variáveis X_i, $i = 1, \ldots, n$, são independentes e a variância da soma de v.a.'s independentes é igual à soma das variâncias parcelares, as restantes propriedades da variância permitem deduzir que:

$$\sigma^2_{\bar{X}} = \mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\underbrace{\mathrm{Var}(X_i)}_{\sigma^2} = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n},$$

isto é, a variância da média aleatória é sempre igual à variância da população dividida pelo número de parcelas, independentemente da distribuição daquela, bem como:

$$\sigma_{\bar{X}} = \sqrt{\mathrm{Var}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \sigma/\sqrt{n}.$$

Em conclusão, a distribuição da média aleatória de v.a.'s i.i.d. com distribuição comum $\mathcal{N}(\mu, \sigma)$ é

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \cap \mathcal{N}(\mu_{\bar{X}} = \mu,\ \sigma_{\bar{X}} = \sigma/\sqrt{n}).$$

No caso presente, como $\sigma = 0.5$, tem-se que $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \cap \mathcal{N}(\mu_{\bar{X}} = \mu,\ \sigma_{\bar{X}} = 0.5/\sqrt{n})$ e, portanto,

$$\frac{\bar{X} - \mu}{0.5/\sqrt{n}} \cap \mathcal{N}(0, 1).$$

A = "A média das n pesagens não se afasta do valor médio do peso do corpo mais de 0.20 gramas" $\Leftrightarrow$

$\Leftrightarrow$ $A = \{|\bar{X} - \mu| \le 0.2\}$. Pretende determinar-se n: $P(A) = 0.99 \Leftrightarrow P(|\bar{X} - \mu| \le 0.2) = 0.99$.

$$P(|\bar{X} - \mu| \le 0.2) = 0.99 \Leftrightarrow P\left(\frac{|\bar{X} - \mu|}{0.5/\sqrt{n}} \le \frac{0.2}{0.5/\sqrt{n}}\right) = 0.99 \Leftrightarrow P\left(\left|\frac{\bar{X} - \mu}{0.5/\sqrt{n}}\right| \le 0.4\sqrt{n}\right) = 0.99 \Leftrightarrow$$

$$\Leftrightarrow P\left(-0.4\sqrt{n} \le \frac{\bar{X} - \mu}{0.5/\sqrt{n}} \le 0.4\sqrt{n}\right) = 0.99 \Leftrightarrow \Phi(0.4\sqrt{n}) - \Phi(-0.4\sqrt{n}) = 0.99 \Leftrightarrow$$

$$\Leftrightarrow 2\,\Phi(0.4\sqrt{n}) - 1 = 0.99 \Leftrightarrow \Phi(0.4\sqrt{n}) = 0.995 \Leftrightarrow 0.4\sqrt{n} = \Phi^{-1}(0.995) \Leftrightarrow \sqrt{n} = 2.5 \times 2.576 \Rightarrow$$

$$\Rightarrow n = (2.5 \times 2.576)^2 \Leftrightarrow n \approx 41.47 \Rightarrow n = \underline{42 \text{ pesagens}}$$