1. (a) i. $\overline{y} = \sum_{i=1}^{9} \frac{n_{R_i}}{n} \overline{y}_{R_i} = 3.522$. Quanto à variância, somando e subtraindo $\overline{y}_{R_i}$ desenvolvendo o quadrado obtemos

$$\begin{aligned}(n-1)s_{\mathbf{Y}}^2 &= \sum_{i=1}^{100} (Y_i - \overline{y})^2 \\ &= \sum_{i=1}^{9} \sum_{j=1}^{n_{R_i}} \Big((Y_{i,j} - \overline{y}_{R_i}) + (\overline{y}_{R_i} - \overline{y}) \Big)^2 \\ &= \sum_{i=1}^{9} (n_{R_i} - 1) s_{R_i}^2 + \sum_{i=1}^{9} n_{R_i} (\overline{y}_{R_i} - \overline{y})^2,\end{aligned}$$

onde $Y_{i,j}$ se refere à $j^{\text{ésima}}$ unidade amostrada na subregião i e $s_{R_i}^2$ à variância da porção de amostra correspondente à subregião i. Substituindo os valores temos $99 \times s_{\mathbf{Y}} = 311.66 + 499.922 = 811.582$, logo $s_{\mathbf{Y}}^2 = 8.198$.

ii. O intervalo de confiança tem extremos definidos pr

$$\overline{y} \pm z_{0.025} \times \sqrt{\widehat{\text{Var}}(\overline{y})} = \overline{y} \pm 1.96 \times \sqrt{\left(\frac{1}{100} - \frac{1}{882}\right) s_{\mathbf{Y}}^2}.$$

Substituindo os valores obtemos que, com confiança 95%, $\mu_{\mathbf{Y}} \in (2.994, 4.050)$.

iii. A amplitude do intervalo com confiança 95% é aproximadamente igual a

$$2 \times 1.96 \times \sqrt{\left(\frac{1}{n} - \frac{1}{882}\right) \times 8.198}.$$

Este valor é apenas aproximado porque utilizamos o valor de variância da amostra obtida na amostragem já recolhida. A resolução desta equação dá

$$\frac{1}{n} = \frac{1}{882} + \frac{1}{8.198} \left(\frac{0.5}{2 \times 1.96}\right)^2 = 0.003118 \quad \Leftrightarrow \quad n = 320.675.$$

Logo, seria aconselhável recolher uma amostra com, pelo menos, 320 unidades. (Nota: O arredondamento simples diria 321. No entanto, estamos a trabalhar com valores aproximados para a variância e, consequentemente, para a amplitude, pelo que o valor 320 é aceitável.)

(b) i. $\overline{y}_{str} = \sum_{i=1}^{9} \frac{N_{R_i}}{N} \overline{y}_{R_i} = 3.554$. A estratificação efectuada é proporcional, o que deveria fazer com que $\overline{y}_{str} = \overline{y}$. A diferença deve-se aos arredondamentos que foram necessários para definir os tamanhos de amostragem em cada estrato.

ii. Atendendo à independência das amostragem efectuadas em cada estrato:

$$\widehat{\text{Var}}(\overline{y}_{str}) = \sum_{i=1}^{9} \left(\frac{N_{R_i}}{N}\right)^2 \left(\frac{1}{n_{R_i}} - \frac{1}{N_{R_i}}\right) s_{R_i}^2.$$

Esta expressão para a variância não se simplifica mais, utilizando as caracterizações para a estratificação proporcional, atendendo aos arredondamentos efectuados nas definições dos tamanhos de amostragem por estrato que fazem com que se não verifiquem as igualdades $\frac{N_{R_i}}{N} = \frac{n_{R_i}}{100}$. Efectuando os cálculos temos $\widehat{\mathrm{Var}}(\overline{y}_{str}) = 0.0295$.

iii. $\widehat{p}_{str} = \sum_{i=1}^{9} \frac{N_{R_i}}{N} \widehat{p}_{R_i} = 0.6107$.

Atendendo à independência entre as amostragens em cada estrato:

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{p}_{str}) &= \sum_{i=1}^{9} \left(\frac{N_{R_i}}{N}\right)^2 \widehat{\mathrm{Var}}(\widehat{p}_{R_i}) \\ &= \sum_{i=1}^{9} \left(\frac{N_{R_i}}{N}\right)^2 \left(\frac{1}{n_{R_i}} - \frac{1}{N_{R_i}}\right) \frac{n_{R_i}}{n_{R_i}-1} \widehat{p}_{R_i}(1-\widehat{p}_{R_i}) = 0.001422.\end{aligned}$$

iv. Como são conhecidas aproximações para as variâncias dentro da cada estrato podemos utilizar uma estratificação óptima: n_{R_i} proporcional a $N_{R_i} S_{R_i} \approx N_{R_i} s_{R_i}$. Como $\sum_{i=1}^{9} N_{R_i} s_{R_i} = 1411.77$, então, numa amostragem estratificada com tamanho total de amostra n,

$$n_{R_i} = \frac{N_{R_i} s_{R_i}}{1411.77} \times n.$$

Para o caso $n = 100$ temos então os seguintes tamanhos de estratos na amostra:

R_i	1	2	3	4	5	6	7	8	9
n_{R_i}	16	6	14	22	14	2	11	7	8

(c) Para seleccionar os grupos utilizamos o método da função inversa generalizada com a seguinte ordenação:

$(\mathcal{C}_1, \mathcal{C}_2)$	$(\mathcal{C}_2, \mathcal{C}_3)$	$(\mathcal{C}_2)$
0.5	0.8	1

Admitindo que o número em [0,1] escolhido é $u = 0.6157$, seleccionamos $s = (\mathcal{C}_2, \mathcal{C}_3)$.

i. Uma estimativa não enviesada para o total da população é dada por

$$\widehat{T}(p) = \frac{\widehat{T}_2(p)}{1} + \frac{T_3(p)}{0.8},$$

onde $\widehat{T}_2(p)$ é uma estimativa não enviesada para o total, relativamente a p, no conjunto $\mathcal{C}_2$, e $T_3(p) = 195$, conforme o enunciado. $\widehat{T}_2(p)$ calcula-se utilizando o estimador de Horvitz-Thompson relativamente a uma amostragem estratificada com estratos de tamanhos 71, 84, 67 e 79 na população e 8, 9 , 8, 9, na amostra. Temos então

$$\widehat{T}_2(p) = 72 \times 0.89 + 84 \times 0.29 + 67 \times 0.91 + 79 \times 0.11 = 158.1,$$

pelo que $\widehat{T}(p) = 158.1 + \frac{195}{0.8} = 314.1$.

Conforme referido, esta estimativa é não enviesada.

ii. Trata-se de uma amostragem por grupos em dois passos, com um dos grupos seleccionados a ser completamente amostrado. Logo, a variância associada ao valor para o total no conjunto $\mathcal{C}_3$ é nula. Assim,

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{T}(p)) &= \frac{T_3^2(p)}{0.8}\left(\frac{1}{0.8}-1\right)+\widehat{\mathrm{Var}}(\widehat{T}_2(p)) \\ &= \frac{195}{0.8}\left(\frac{1}{0.8}-1\right)+\left(\frac{882}{302}\right)^2\times 0.000438 = 60.9412.\end{aligned}$$

(Nota: O valor para $\widehat{(}\mathrm{Var})(\widehat{T}_2(p))$ é calculado de forma análoga ao efectuado em 1.-(a)-iii.)

2. (a) Como $(\pi_1,\ldots,\pi_5) = 2\mathbf{p}$, segue-se que $\pi_3 = 0.667$, $\pi_4 = \pi_5 = 0.222$. Segue-se então que $p_1 = 0.012$, pelo que $p_2 = 0.099$.

   (b) Uma possível função de distribuição para o problema em causa é:

S	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
P	.012	.247	.346	.445	.853	.865	.887	.899	.901	1.000

   Admitindo que $u = .2143$, seleccionamos $s = (1,3)$.

   (c) $\widehat{T} = \dfrac{Y_1}{0.445} + \dfrac{Y_3}{0.667} = \dfrac{-3}{0.445} + \dfrac{12}{0.667} = 11.2494.$

   $$\sigma^2_{\mathbf{Y}} = \frac{1}{5^2}\frac{(Y_1-Y_3)^2}{\pi_{1,3}} = \frac{(-3-12)^2}{25\times 0.235} = 38.2979.$$

   (d) Como neste plano de amostragem todas as probabilidades de inclusão de segunda ordem são não nulas, a estimativa anterior é não enviesada para a variância da população.

   (e) $\widehat{\mu}_{\mathbf{Y}} = \dfrac{1}{N}\widehat{T} = 2.2499.$

   $$\widehat{\mathrm{Var}}(\widehat{\mu}_{\mathbf{Y}}) = \frac{1}{25}\widehat{\mathrm{Var}}(\widehat{T}).$$

   $$\widehat{\mathrm{Var}}(\widehat{T}) = \widehat{\mathrm{Var}}_{\mathrm{SYG}} = \tfrac{\pi_1\pi_3-\pi_{1,3}}{\pi_{1,3}}\left(\tfrac{Y_1}{\pi_1}-\tfrac{Y_3}{\pi_3}\right)^2 = 160.903.$$

   O intervalo com confiança 95% tem extremos $\widehat{\mu}_{\mathbf{Y}} \pm 1.96\times\sqrt{160.903}$, isto é, com confiança 95%, $\mu_{\mathbf{Y}} \in (-22.6123,\ 27.112)$.

   (Nota: É possível, neste plano de amostragem, que a estimativa para a variância de $\widehat{\mu}_{\mathbf{Y}}$ seja negativa. Em rigor, isso não permite construir o intervalo de confiança. Uma resposta possível consiste em utilizar a expressão genérica para obter estimativas para a variância do estimador de Horvitz-Thompson. Caso esta também seja negativa, então não é possível obter o intervalo de confiança pretendido.)