Instituto Superior de Agronomia–Secção de Matemática

Amostragem e Análise Ambiental
2020/2021
Exercícios do módulo I – Amostragem

1. Uma população é formada de 35 árvores de uma determinada espécie, pertencentes a um parque ecológico, que possuem os seguintes "diâmetros à altura do peito" (DAP), em cm :

25, 20, 35, 21, 22, 22, 24, 25, 30, 38, 24, 20, 21, 25, 20, 15, 25, 23, 20, 24, 28, 24, 24, 22, 28, 26, 23, 19, 22, 27, 25, 23, 28, 27, 42.

Suponha que lhe é pedido para extrair uma amostra aleatória simples de dimensão 10 com o objectivo de estimar o DAP médio de todas as árvores. No Anexo tem uma exemplificação de procedimentos e alguns cálculos associados. Responda às seguintes questões usando resultados do Anexo.

a) Como extrairia uma amostra aleatória simples de tamanho 10? No Anexo há 2 procedimentos, qual o que escolheria e porquê?

b) Estime o DAP médio das árvores daquele parque. Compare com o valor médio da população, determinando um valor da medida designada por *Erro Relativo de Estimação Percentual* assim definido:
$ER = |\overline{X} - \mu|/\mu \times 100\%$.

c) Complete o *output* do Anexo indicando os valores ou expressões para as letras **A**, **B**, **C** e **D**.

Anexo

```
> x<-c(25, 20, 35, 21, 22, 22, 24, 25, 30, 38, 24, 20,
+ 21, 25, 20, 15, 25, 23, 20, 24, 28, 24, 24, 22, 28, 26,
+ 23, 19, 22, 27, 25, 23, 28, 27, 42)

> sum(x)
[1] 867

>sum(x^2)
[1] 22415

> set.seed(35)
> amostra1<-sample(x,10,rep=TRUE)
```

```
> amostra1
 [1] 27 25 24 27 23 21 25 28 42 27

> set.seed(35)
> amostra2<-sample(x,10)
> amostra2
 [1] 27 25 20 25 15 35 25 25 23 26


> sum(amostra1)
[1] 269
> mean(amostra1)
[1]A

> var(amostra1)
[1] 32.76667
> sd(amostra1)
[1] B

#vamos pedir para calcular o coeficiente de variação da amostra 1
>CV<- C
[1] D

> mean(amostra2)
[1] 24.6

> var(amostra2)
[1] 25.82222
```

2. Considere os seguintes dados referentes ao peso seco, x (em g), de 30 plantas de uma dada espécie, recolhidos ao acaso numa dada região, com o objectivo de se obter informação sobre características do solo da região.

12.27	9.71	6.76	6.43	7.5	9.88	9.54	13.62	9.33	7.66
5.86	9.89	6.19	7.41	5.48	8.64	9.6	7.73	6.74	10.83
11.68	8.04	7.23	7.35	8.94	12.05	8.23	7.42	11.2	10.12

$$\sum_{i=1}^{30} x_i = 263.33 \text{ g}; \qquad \sum_{i=1}^{30} x_i^2 = 2435.799 \text{ g}^2$$

a) Diga qual o estimador do peso seco médio das plantas daquela espécie que devemos usar e indique duas propriedades.

b) Com base nos dados recolhidos indique uma estimativa para o verdadeiro peso médio seco das plantas daquela espécie na região em estudo e uma estimativa para a variância do estimador considerado na alínea anterior.

c) Obtenha um intervalo de confiança a 95% para o peso médio seco das plantas daquela espécie naquela região.

d) Considerando este estudo como um estudo piloto, indique a dimensão de amostra a recolher de modo a responder à alínea anterior com um erro inferior a 1g.

3. Para estimar o número de árvores existentes numa floresta, dividiu-se a área ocupada pela floresta em 800 parcelas de igual dimensão. Escolheram-se ao acaso 20 dessas parcelas, tendo-se contado em cada um deles o número de árvores existentes da espécie preponderante (x_i), bem como o número de árvores de todas as outras espécies diferentes (y_i). Os valores registados em cada parcela encontram-se no quadro abaixo:

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
61	12	23	13	38	10	52	3
38	17	31	9	29	27	44	11
75	14	46	22	31	14	74	12
43	8	63	25	52	11	38	10
37	15	72	12	27	8	42	9

a) Considere apenas a espécie preponderante. Estime:

i) o número médio de árvores por parcela;

ii) a variância do número de árvores por parcela;

iii) o total de árvores daquela espécie na floresta.

b) Estime o número total de árvores da floresta e a variância do estimador que utilizou.

c) Construa um intervalo de confiança a 99% para o número total de árvores na floresta.

d) Considerando este plano de amostragem como um estudo piloto, em quantas parcelas deveria ser feita a contagem das árvores para estimar, com 95% de confiança o número médio de árvores por parcela da espécie preponderante, com um erro inferior a 3 árvores?

e) (*) No estudo da manutenção da espécie preponderante é importante a avaliação da razão entre o número de árvores dessa espécie e o número de árvores das outras espécies.

i) Indique uma estimativa daquela razão.

ii) Indique uma estimativa da variância do estimador usado.

(*) - esta alínea será para fazer mais tarde

4. Resolver o exercício anterior com apoio do R.

5. Considere que tem um campo rectangular de 5 km × 4 km. Pretende-se dividir em parcelas de 100 m × 100 m e dessas seleccionar ao acaso 20 parcelas para recolher valores de características do solo. Elabore um programa em R que lhe permita fazer aquela escolha.

6. Um Engenheiro do Ambiente pretende estimar a proporção p de árvores de uma dada espécie que apresentam uma certa moléstia numa região.

a) Ele deseja que a probabilidade de que a sua estimativa não se desvie do verdadeiro valor de p por mais que 0,02 com uma confiança de 95%. Que dimensão de amostra deve recolher?

b) Suponha que aquele Engenheiro decide realizar um estudo piloto que lhe possa dar mais alguma informação para estimar o tamanho da amostra. Escolhidas 50 árvores ao acaso, encontrou 4 com a referida moléstia:

i) Determine uma estimativa de p e um intervalo a 95% de confiança para p.

ii) Usando estes resultados como um estudo piloto, o que poderia dizer sobre o tamanho da amostra a recolher para estimar p com o mesmo erro considerado em a)?

7. Os dados que se apresentam abaixo referem-se aos valores observados de pH em amostras recolhidas em dois tipos de solos existentes numa dada região.

Solo 1 (x_1)

9.0	8.2	7.8	8.0	6.7	12.3	7.7	8.3	7.8	8.7	9.4	10.1	9.5	6.4	9.2	6.6

$$\sum_{j=1}^{16} x_{1j} = 135.7; \qquad \sum_{j=1}^{16} x_{1j}^2 = 1183.75$$

Solo 2 (x_2)

3.6	5.3	7.0	6.3	9.2	5.0	4.0	4.1	3.7	9.1	2.4	5.3	6.4	4.3	3.9	6.1
3.8	3.0	6.5	3.2	4.3	7.3	3.0	6.3	3.2	7.2	6.0	6.2	6.9	5.1	4.7	6.1
2.8	8.4	2.9	4.1	3.9	6.3	10.0	5.2								

$$\sum_{j=1}^{40} x_{2j} = 212.1; \qquad \sum_{j=1}^{40} x_{2j}^2 = 1265.81$$

a) Calcule e compare a estimativa do valor médio de pH e respectiva precisão nos seguintes casos:

   i) Para cada tipo de solo separadamente.

   ii) Considerando os 56 valores observados como se se tratasse de dados provenientes de uma amostra aleatória simples.

   iii) Considerando os dados provenientes de uma amostra aleatória estratificada em que se considerou uma afectação proporcional na recolha das amostras.

b) Construa um intervalo de confiança a 95% para o valor médio de pH considerando a situação a) iii).

c) Supondo fixo o custo de amostragem de cada observação individual, qual a dimensão óptima de amostra a recolher em cada Solo, no caso de se pretender uma amostra de dimensão total 60.

8. Suponha que numa dada região existem 588 explorações agrícolas para as quais se pretende estudar o investimento feito em maquinaria agrícola e equipamento no último ano. O registo completo dos dados referentes a todas as explorações (registados em unidades monetárias adequadas) encontra-se no quadro abaixo, dividido em 3 conjuntos consoante a classificação das explorações:

Explorações Pequenas

17	38	9	7	11	14	17	10	31	24	22	21	9	41	19
9	13	26	36	18	8	11	23	19	16	14	14	17	20	20
9	18	6	19	52	14	5	27	14	14	28	17	9	11	12
25	19	28	15	18	24	23	27	24	20	21	27	21	34	26
21	9	29	22	10	18	45	24	16	95	40	42	11	17	17
13	14	23	17	27	18	34	18	16	17	20	23	18	42	22
18	23	16	26	11	37	23	32	24	16	24	34	37	31	29
15	41	38	21	34	23	24	27	34	5	34	29	22	26	30
26	27	39	30	31	28	39	28	34	28	24	44	22	23	40
16	5	19	36	36	17	21	43	21	19	14	14	31	27	39
30	41	28	19	32	18	19	33	27	28	26	23	32	36	21
24	32	19	18	31	25	26	21	18	36	29	47	26	31	26
32	27	43	45	45	25	17	30	27	28	16	44	20	15	31
21	42	27	32	33	21	35	44	24	26	38	57	54	24	37
21	33	19	20	32										

Explorações Médias

37	30	41	17	38	29	32	21	39	41	28	33	35	24	36
28	20	23	27	34	33	36	25	28	39	36	22	25	54	53
36	14	22	32	21	35	35	39	32	40	24	48	41	30	42
20	38	23	17	38	16	23	28	32	18	60	28	47	61	25
22	25	48	53	35	25	23	44	18	56	42	55	39	24	38
42	27	30	34	43	29	35	43	62	25	15	66	34	25	11
45	28	40	32	38	33	48	46	54	45	35	31	30	42	22
23	46	14	42	33	31	75	50	44	33	41	32	45	44	51
39	35	22	44	35	24	29	23	32	30	35	50	28	21	21
12	30	28	60	35	49	33	22	58	25	23	39	40	44	41
14	37	32	22	27	23	37	59	50	46	40	47	41	38	48
40	32	31	22	24	25	33	54	36	52	39	61	46	36	16
37	38	51	25	35	49	9	46	35	53	43	59	41	52	51
47	72	46	29	25	42	42	43	46	43	29	58	47	85	52
48	23	39	40	43	52	36	35	27	56	47	39	51	48	48
23	24	39	30	59	35	39	32	51	18	27	38	36	41	11
42	42	65	27	34	72	49	39	44	57	64	51	53	55	63
39	31	48												

Explorações Grandes

53	63	44	66	40	42	48	44	27	56	37	39	37	40	66
49	39	54	30	68	36	42	28	29	41	57	30	39	28	80
79	61	81	53	57	54	29	94	77	52	61	49	52	67	36
35	57	63	32	48	57	50	62	51	52	59	55	22	18	84
57	86	50	54	96	45	28	59	64	42	41	77	76	83	36
42	39	72	84	34	55	51	66	96	63	88	87	63	91	117
107	48	56	71	54	64	45	61	59	68	50	74	100	144	80
64	101	105	77	85	60	63	66	36	95					

a) Obtenha uma amostra aleatória simples de dimensão 25 desta população, explicando o procedimento usado para a obtenção da amostra. Registe-a.

b) A partir dos dados obtidos determine uma estimativa do investimento médio por exploração, uma estimativa do investimento total, assim como um intervalo de confiança a 95% para cada um daqueles parâmetros.

c) Considerando os valores obtidos em b) como um estudo piloto determine a dimensão da amostra que se deveria considerar por forma a estimar o investimento médio com uma tolerância de 3 unidades monetárias com um risco inferior a 5%.

d) Suponha agora que considera os três estratos em que a população é dividida, pequenas, médias e grandes empresas. Usando de novo números aleatórios obtenha uma amostra de dimensão 25 admitindo que decide fazer uma afectação proporcional.

e) Com os dados obtidos na alínea anterior calcule o investimento médio por exploração, assim como um intervalo de confiança a 95% para aquele parâmetro. Compare com os resultados obtidos na alínea b) e comente.

f) Usando as estimativas da variância em cada estrato obtidas na alínea anterior determine as dimensões óptimas a recolher em cada estrato para um erro de 3 unidades monetárias com um risco inferior a 5%.

9. Uma amostra aleatória simples de 20 apartamentos foi retirada de uma zona da cidade contendo 12400 apartamentos. O número de pessoas por apartamento na amostra foi o seguinte:

```
5  6  3  3  2  3  4  4  3  3  4  3  3  1  2  4  3  4  2  4
```

a) Estime o número total de pessoas vivendo naquela zona e determine o intervalo de confiança a 90% para esse parâmetro.

b) i) Estime a percentagem de apartamentos com mais do que 3 pessoas. Indique a precisão da sua estimativa.

ii) Indique uma estimativa do total de apartamentos nas condições referidas.

c) Admita que, para cada um dos apartamentos usados na amostra, se regista o número de automóveis existentes. Suponha que se obtiveram os seguintes valores:

```
2  3  1  0  1  2  0  3  1  0  1  0  2  1  0  2  1  0  1  3
```

Uma medida de interesse para o estudo do parque automóvel é dada pelo número de habitantes por automóvel.

i) Indique uma estimativa daquela medida.

ii) Sabendo que o número de automóveis existentes na zona em estudo é de 11000, estime o número médio de habitantes por apartamento.

10. Realizou-se um inquérito com o objectivo de obter informações sobre o número de herdades que produzem trigo, bem como a área total de trigo, numa dada região. As herdades foram estratificadas de acordo com a respectiva dimensão. No seguinte quadro encontra-se a caracterização dos estratos, o número de herdades em cada estrato, o número de herdades seleccionadas em cada estrato, o número de herdades observadas possuindo trigo e a área de trigo observada

Classes (em ha)	N_i	Nº de herdades na amostra	Nº de herdades com trigo	Área com trigo (em ha)
1-5	435	22	0	0
6-20	519	26	1	7
21-50	357	18	5	35
51-150	519	26	21	386
151-300	400	20	16	710
300-	266	13	11	873

a) Considere os dados referentes apenas ao estrato relativo às maiores herdades:
   i) Indique uma estimativa da área total de trigo naquele estrato. Calcule o I.C. a 95% que lhe está associado.
   ii) Qual a estimativa da área média de trigo por herdade com trigo.
   iii) Calcule uma estimativa da verdadeira proporção de herdades com trigo naquele estrato.
   iv) Investigue a dimensão da amostra que aconselharia recolher para responder à alínea anterior com um erro inferior a 1%.

b) Indique uma estimativa da área total de trigo naquela região. Como calcularia a precisão da sua estimativa?

c) Calcule uma estimativa da verdadeira proporção de herdades com trigo. Construa um intervalo de confiança a 95% para essa proporção.

d) Indique uma estimativa da área com trigo relativamente às herdades com trigo, na classe com herdades maiores.

e) No caso da alínea anterior suponha que conhece o total de herdades com trigo no último estrato e que é de 180. Usando esta informação indique uma estimativa da área total de trigo na referida classe.

f) Como procederia para comparar as estimativas obtidas nas alíneas b) e e). Quando é que a estimativa definida na alínea e) é preferível à definida na alínea b).

g) Usando os dados do quadro como uma amostragem piloto determine a dimensão da amostra a recolher em cada estrato que permite uma variância mínima na estimação da proporção de herdades com trigo. Suponha fixos os custos de amostragem.

11. Um investigador pretende estudar certas características de uma população de amêijoas. Para isso decide recolher uma amostra numa dada região que, atendendo às características morfológicas, se considera dividida em quatro estratos, que designaremos por A, B, C e D. No seguinte quadro encontram-se a dimensão N_i de cada estrato, a dimensão de amostra n_i recolhida em cada estrato, assim como a média e a variância do número de amêijoas observadas em cada estrato.

Estrato	N_i	n_i	$\overline{x}_i$	$s_i'^{\,2}$
A	5703.9	14	0.50	0.068
B	1270.0	16	1.25	0.042
C	1286.4	13	4.00	1.146
D	5063.9	15	1.80	0.794

As unidades de amostragem consideradas foram parcelas de 5 m^2 de área.

a) Indique uma estimativa do total de amêijoas naquela região. Qual a precisão da sua estimativa?

b) Indique um I.C. a 95% para o n.º médio de amêijoas em cada parcela de 5 m^2.

c) Considere os dados do quadro resultantes de uma amostragem piloto. Qual seria a dimensão total de amostra a recolher (e a dimensão em cada estrato) se pretendesse com 95% de confiança ter uma estimativa para o valor médio definido na alínea anterior com uma variância de 0.002.

d) Admita que, após recolhidas as amostras em cada estrato se regista o n.º de amêijoas impróprias para consumo, tendo-se encontrado 1, 2, 5 e 2 respectivamente nos estratos A,B,C e D.

   i) Indique uma estimativa da proporção de amêijoas impróprias para consumo naquela região.

   ii) Indique um intervalo de confiança para a verdadeira proporção de amêijoas impróprias no estrato C.

12. No seguinte quadro apresentam-se os dados de um estudo realizado em 1000 doentes seleccionados de entre 10000 que constam nos processos de um pequeno hospital. Encontrou-se 30 que revelaram valores positivos de existência de uma dada bactéria. Os dados foram recolhidos categorizados por classe etária (apresentam-se valores que possibilitam cálculos simples).

Classe de idade	nºde doentes por classe	dimensão da amostra	n. de pessoas do sexo feminino	n. de resultados positivos
A–Menos de 25 anos	1000	100	20	12
B–De 25 a 40 anos	4000	500	200	8
C–Acima de 40 anos	5000	400	350	10
Total	10000	1000	570	30

a) Considere os dados referentes apenas à **classe A**.

   i) Estime o número total de doentes que revelariam valores positivos daquela bactéria e o intervalo de confiança a 95% para esse parâmetro.

ii) Estime a proporção, p, de doentes com valores positivos e indique a precisão dessa estimativa.

iii) Suponha que, com base em estudos anteriores, se pode admitir que a proporção referida na alínea anterior verifica $p \in [0.01, 0.15]$. Qual a dimensão mínima da amostra que assegura uma precisão da estimativa da proporção de resultados positivos com erro inferior a 0.05?

iv) Indique uma estimativa do número total de doentes so sexo feminino que, naquela faixa etária acorrem ao hospital. Determine um intervalo de confiança para o parâtero em estudo.

b) Considere agora todos valores dados no quadro acima.

i) Responda de novo à questão colocada na alínea a)i).

ii) Usando os dados recolhidos como uma amostragem piloto, qual a dimensão óptima da amostra a recolher em cada classe para estimar a proporção de doentes revelando resultados positivos, ignorando custos e de modo a obter a mesma dimensão final, $n = 1000$.

13. Considere a seguinte amostra, retirada aleatoriamente de uma população para a qual se pretende estimar σ.

16 43 13 7 12 14 6 25 0 54

Obtenha estimativas e intervalos de confiança para σ usando procedimentos clássicos e a metodologia *bootstrap*.

Nota: Esta amostra é dada em Wonnacott and Wonnacott (1990), pág. 280, e os autores dizem que foi obtida por simulação de uma distribuição com $\sigma = 10$.

14. Considere que dispõe da seguinte amostra e respectiva média e variância:

$$\mathbf{x} = \{10, 9, 10, 11, 16, 15, 8, 6, 18, 17, 11, 12\} \qquad \overline{x} = 11.92 \qquad s_x^2 = 14.26515$$

a) Determine uma estimativa do coeficiente de variação da variável em estudo.

b) Explique como procederia para obter uma estimativa *bootstrap* do coeficiente de variação e para construir um intervalo de confiança *bootstrap* para aquele parâmetro.

15. O coeficiente de variação (CV) é uma medida de dispersão que descreve a quantidade de variabilidade relativa à média, numa característica quantitativa positiva. Como o coeficiente de variação não possui unidades, ele pode ser usado para comparar a dispersão dos dados em conjuntos com unidades diferentes.

Dada uma amostra de observações $(x_1, x_2, \ldots, x_n)$ o coeficiente de variação é definido como $CV = s_x/\overline{x} \times 100\%$

Considere a amostra relativa a 30 valores de pH registados no solo de uma dada região.

```
pH<-c(8.20,8.15,8.25,8.17,8.31,8.30,8.25,8.20,8.15,8.20,
8.23,8.10,8.25,8.50,8.20,8.15,8.20,8.10,8.20,8.21,8.22,
8.20,8.21,8.25,8.25,8.20,8.22,8.20,8.15,8.17)
```

Obtenha uma estimativa *bootstrap* para o CV do pH do solo daquela região e um intervalo de confiança a 95%.

16. Os dados `patch` no `package (bootstrap)` contêm medidas de uma certa hormona na corrente sanguínea de oito indivíduos depois de usarem um medicamento. O parâmetro de interesse é

$$\theta = \frac{E[novo] - E[antigo]}{E[antigo] - E[placebo]}$$

Se $\theta \leq 0.20$ isso indica bioequivalência dos antigo e novo medicamentos. A estatística de interesse é a **razão**

$$\widehat{\theta} := \frac{\overline{Y}}{\overline{Z}}$$

com $Y := novo - antigo$ e $Z := antigo - placebo$.

a) Calcule a estimativa *Jackknife* do desvio padrão da **razão** da bioequivalência, $\widehat{\theta}$.

b) Obtenha um I.C. *boostrap* para aquela **razão**.