Tópicos de uma resolução do exame de Amostragem e Análise Ambiental (2019/2020)
2ª Chamada — Módulo I

Nota: Vou apresentar a resolução da versão do exame cuja 1ª pergunta tratava de "investimentos na criação de infraestruturas de saneamento básico".

1. Estamos perante uma amostragem estratificada; os estratos são: (1) contém as freguesias de **tipo urbano** (1) e (2) contém as freguesias de **tipo rural**.

Comecemos por indicar/calcular os **dados** de que dispomos e os necessários para responder:

$N_1 = 210; \quad N_2 = 130; \quad N = 340$

$n_1 = 30 \quad n_2 = 15 \quad n = 45 \qquad f_1 = n_1/N_1 = 0.1428571 \quad f_2 = n_2/N_2 = 0.1153846$

$\sum_{j=1}^{30} x_{1j} = 312; \qquad \sum_{j=1}^{30} x_{1j}^2 = 4670; \qquad \sum_{j=1}^{15} x_{2j} = 92; \qquad \sum_{j=1}^{15} x_{2j}^2 = 736$

x_{ij} representa o investimento (em unidades monetárias adequadas) na $j-$ésima freguesia do tipo i, observado nas amostras recolhidas.

Seja y_i a variável que conta o número de freguesias (note que é uma soma de 0 e 1) que tinham empresas que escoavam produtos com impacte ambiental negativo no estrato i, $\quad y_1 = 12 \quad y_2 = 5$

$\overline{x}_1 = 312/30 = 10.4; \quad \overline{x}_2 = 92/15 = 6.133$ (nas unidades adequadas)

$$s_1'^2 = \frac{30\ \sum_{j=1}^{30} x_{1j}^2 - (\sum_{j=1}^{30} x_{1j})^2}{30 \times 29} = 49.14483 \qquad s_2'^2 = 12.26667$$

a) Pretende-se $x_T^* = \sum_{i=1}^{2} N_i \overline{x}_i = 2981.333$ (nas unidades adequadas)

$\widehat{Var[X_T^*]} = \sum_{i=1}^{2} N_i^2 (1 - f_i) \dfrac{s_i'^2}{n_i} = 74148.26$

Intervalo de confiança a 95% para X_T

$\left] x_T^* - z_{0.025}\sqrt{\widehat{Var[X_T^*]}}\ ;\ x_T^* + z_{0.025}\sqrt{\widehat{Var[X_T^*]}} \right[\ =]2447.622\,;\,3515.045[$

b) $\hat{p}_i = \overline{y_i}$, então $\qquad \hat{p}_1 = 12/30 \qquad \hat{p}_2 = 5/15$

$\widehat{p_{st}} = \sum_{i=1}^{2} \dfrac{N_i \hat{p}_i}{N} = 0.3745098$

$\widehat{\widehat{Var[P_{st}]}} = \sum_{i=1}^{2} \dfrac{N_i^2}{N^2} \dfrac{\hat{p}_i \hat{q}_i}{n_i - 1} \dfrac{N_i - n_i}{N_i - 1} = 0.0047877$

Um IC para a verdadeira proporção, P, é

$\left] \widehat{p_{st}} - z_{0.025}\sqrt{\widehat{\widehat{Var[P_{st}]}}}\ ;\ \widehat{p_{st}} + z_{0.025}\sqrt{\widehat{\widehat{Var[P_{st}]}}} \right[\ =]0.2388902\ ;\ 0.5101294[$

c) Seja T o total de freguesias possuindo empresas nas condições referidas acima

$\widehat{T} = N \times \widehat{p_{st}} = 127.33$

A precisão é dada pela variância

$\widehat{Var[\widehat{T}]} = \sum_{i=1}^{2} N_i^2 \dfrac{\hat{p}_i \hat{q}_i}{n_i - 1} \dfrac{N_i - n_i}{N_i - 1} = 553.4655$

ou pelo desvio padrão: $\sqrt{\widehat{Var[\widehat{T}]}} = 23.52585$

d) Pretende-se a dimensão de amostra a recolher em cada estrato, para estimar o valor médio do investimento, para ter $n = 45$

i) afectação proporcional

$$\frac{n_i}{N_i} = \frac{n}{N} \Longrightarrow n_i = \frac{n}{N} \times N_i \Longrightarrow n_1 = 27.79; \quad n_2 = 17.20$$

Portanto $n_1 = 28; \quad n_2 = 17$

ii) afectação óptima

$$n_i = n\frac{N_i\ s_i^{'}}{\sum_{i=1}^{2} N_i\ s_i^{'}} \Longrightarrow n_1 = 34.3701; \quad n_2 = 10.6299$$

Portanto $n_1 = 34; \quad n_2 = 11$

e) Pretende-se fazer uma amostragem estratificada com uma tolerância d com uma confiança de 95%. Cálculo de n? – é baseado na semi-amplitude do IC. Pretende-se que esta semi-amplitude (a 95%) seja:

$$z_{0.025}\sqrt{Var[\overline{X}_{st}]} \leq d \Longrightarrow z_{0.025}\sqrt{\sum_{i=1}^{n} \frac{N_i^2}{N^2}(1-f_i)\frac{\sigma_i^{'2}}{n_i}} \leq d$$

Dado que se pretende uma afectação proporcional, i.e. , $\frac{n_i}{N_i} = \frac{n}{N} \Longrightarrow \frac{N_i}{N} = \frac{n_i}{n}$ e considerando $f_i \simeq 0$, podemos substituir acima vindo

$$z_{0.025}\sqrt{\sum_{i=1}^{n} \frac{n_i}{n} W_i \frac{\sigma_i^{'2}}{n_i}} \leq d \Longrightarrow z_{0.025}\sqrt{\sum_{i=1}^{n} \frac{1}{n} W_i \sigma_i^{'2}} \leq d \Longrightarrow (z_{0.025})^2 \sum_{i=1}^{n} \frac{1}{n} W_i \sigma_i^{'2} \leq d^2$$

Logo $n \geq \frac{(z_{0.025})^2 \sum_{i=1}^{n} W_i \sigma_i^{'2}}{d^2}$, muitas vezes substitui-se $z_{0.025}$ por 2, vindo

$$n \geq \frac{4 \sum_{i=1}^{n} W_i \sigma_i^{'2}}{d^2}$$

2. Dados : $n = 35$ agregados

x_1 – nº de pessoas no agregado familiar; x_2 – rendimento familiar semanal e x_3 – despesa semanal em alimentação (estas duas últimas variáveis expressas em unidades monetárias).

$$\sum_{i=1}^{35} x_{1i} = 123; \qquad \sum_{i=1}^{35} x_{1i}^2 = 533; \qquad \sum_{i=1}^{35} x_{2i} = 2394; \qquad \sum_{i=1}^{35} x_{2i}^2 = 177254$$

$$\sum_{i=1}^{35} x_{3i} = 907.2; \qquad \sum_{i=1}^{35} x_{3i}^2 = 28224; \qquad \sum_{i=1}^{35} x_{1i}\ x_{3i} = 3595.5; \qquad \sum_{i=1}^{35} x_{2i}\ x_{3i} = 66678$$

a) i) $\overline{x_3} = 907.2/35 = 25.92$ unidades monetárias/agregado.

ii) Trata-se de uma razão $r_1^* = \frac{\sum_{i=1}^{35} x_{3i}}{\sum_{i=1}^{35} x_{1i}} = \frac{907.2}{123} = 7.375$ unidades monetárias/pessoa.

iii) $r_2^* = \frac{\sum_{i=1}^{35} x_{3i}}{\sum_{i=1}^{35} x_{2i}} = \frac{907.2}{2394} = 0.3789$, portanto a percentagem do rendimento gasta semanalmente em alimentação é 37.89% .

b) Como são razões tem-se, para estimativa da variância do estimador em a)ii)

$$\widehat{Var[R_1^*]} = \frac{(1-f)}{n\overline{x_1}^2}\left[\frac{\sum_{i=1}^{35} x_{3i}^2 - 2r_1^*\sum_{i=1}^{35} x_{1i}x_{3i} + r_1^{*2}\sum_{i=1}^{35} x_{1i}^2}{n-1}\right] = 0.2844839$$

Para estimativa da variância do estimador em a)iii) tem-se

$$\widehat{Var[R_2^*]} = \frac{(1-f)}{n\overline{x_2}^2}\left[\frac{\sum_{i=1}^{35} x_{3i}^2 - 2r_2^*\sum_{i=1}^{35} x_{2i}x_{3i} + r_2^{*2}\sum_{i=1}^{35} x_{2i}^2}{n-1}\right] = ??$$

Vamos admitir $f \approx 0$ e então (faltam as contas!)

c) IC a 95% para a despesa semanal média em alimentação por agregado familiar é dados por

$$]\overline{x_3} - z_{0.025}\sqrt{\widehat{Var[\overline{x_3}]}} \quad ; \quad \overline{x_3} - z_{0.025}\sqrt{\widehat{Var[\overline{x_3}]}}$$

$\widehat{Var[x_3]} = (35 * \sum_{i=1}^{35} x_{3i}^2 - (\sum_{i=1}^{35} x_{3i})^2)/(35 * 34)) = 138.5111$

$\widehat{Var[\overline{x_3}]} = \widehat{Var[x_3]}/(35)) = 3.957$

$$IC =]22.0209\,;\, 29.8191[$$

d) Total de agregados familiares N=50000, permite uma estimativa do total de pessoas
$X_1^{*T} = N\overline{x_1} = 50000 \times 123/35$
Como $r_1^* = \dfrac{\overline{x_3}}{\overline{x_1}}$ então uma etimativa da despesa semanal total dessas pessoas, usando um estimador da razão, é

$X_3^{*T} = N\overline{x_3} = Nr_1^*\overline{x_1} = 1296000$ unidades monetárias

3. a) $n = 48 \quad Q_{0.25}^* = ?$

$n \times 0.25 = 12$, que é inteiro, logo $Q_{0.25}^* = \dfrac{x_{(12)} + x_{(13)}}{2} = 8.075$

b) Ver abaixo os comandos e respectiva execução

```
> pH<-c(7.7, 7.8,7.85, 7.9, 7.9, 7.95, 7.95, 8 ,8 ,8, 8, 8.05,
+ 8.1, 8.1, 8.1, 8.1, 8.1, 8.2, 8.2, 8.2, 8.25, 8.25, 8.3, 8.3,
+ 8.3, 8.3, 8.35, 8.35, 8.35, 8.35, 8.35, 8.4, 8.4, 8.4, 8.4, 8.4,
+ 8.4, 8.4, 8.4, 8.4, 8.4, 8.4, 8.4, 8.4, 8.4, 8.45, 8.45, 8.45)
> n<-length(pH);n
[1] 48
> amostra<-pH
> (quantile(amostra,prob=0.25,type=2))
  25%
8.075   #Notem que é a resposta à alínea a)

> pseudo_val<-vector()
> valor<-vector()
> for(i in 1:n)
+      {
+ valor[i]<-quantile(amostra[-i],prob=0.25,type=2)    ###equivale aos Tn-1,(-i)
+      pseudo_val[i]<-n*quantile(amostra,prob=0.25,type=2)-(n-1)*valor[i]
+  }
>
>
> #Usando o estimador de jackknife TJn que é a média dos pseudo-valores
> TJn<-mean(pseudo_val);TJn
```

```
[1] 8.6625    ## Não é uma estimativa muito boa!!!
>
>
> #IC para o 1ž quartil (aproximado)
> se <- sqrt(mean((pseudo_val - TJn)^2)/(n-1))
>
> IC<-c(TJn-qt(0.975,n-1)*se,TJn+qt(0.975,n-1)*se);IC
[1] 8.363899 8.961101
```

Nota: Alternativamente odiamos usar directamente o package `bootstrap` do R, que intergra tab o Jackknife e que dá o mesmo resultado Só é necessário definir o estimador em causa.

```
> ############
> ####usando o R para fazer o Jacknife
> #########
>
> library(bootstrap)
> x<-pH
> theta<-function(x){quantile(x,prob=0.25,type=2)}
> n<-length(x)
> results<-jackknife(x,theta);results
$jack.se
[1] 0.1484293

$jack.bias
    25%
-0.5875

$jack.values
 [1] 8.10 8.10 8.10 8.10 8.10 8.10 8.10 8.10 8.10 8.10 8.10 8.10 8.05
[14] 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05
[27] 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05
[40] 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05 8.05

$call
jackknife(x = x, theta = theta)

>
> TJn<-n*theta(x)-(n-1)*mean(results$jack.values);TJn
   25%
8.6625
> c(TJn-qt(0.975,n-1)*sd(pseudo_val)/sqrt(n), TJn+qt(0.975,n-1)*sd(pseudo_val)/sqrt(n))
     25%      25%
8.363899 8.961101
>
```