Gabarito Prova I

Amostragem

1. Lembremos da definição de probabilidade condicional, i.e. $P(A|B) = P(A \cap B)/P(B)$.

Também lembremos que um planejamento amostral é aleatório simples sem substituição se todas as possíveis amostras de tamanho n tem a mesma probabilidade de serem selecionadas, i.e,

$$P(S=s) = \begin{cases} \frac{1}{\binom{N}{n}}, & \text{se } \#s = n \\ 0, & \text{caso contrário} \end{cases} \qquad (*)$$

em que $\pi_k = \frac{n}{N}$ para todo $k=1,\dots N$

Agora, numa amostragem Bernoulli com $\pi_k = \pi$ $\forall k \in U$ temos que

$$P(S=s) = \pi^{n_s}(1-\pi)^{N-n_s}. \qquad (**)$$

Se consideramos que o tamanho $n_s = n$ (fixado) então

$$P(S=s \mid n_s=n) = \frac{P(S=s \text{ e } n_s=n)}{P(n_s=n)}$$

Note também que para o plano amostral Bernoulli,

$$n_s = \sum_{k \in U} I_k(s) \quad \text{em que}$$

$$I_k(s) = \begin{cases} 1 & \text{se o elemento } k \in s \\ 0 & \text{se o elemento } k \notin s \end{cases}$$

logo $I_k(s) \sim \text{Bernoulli}(\pi_k) = \text{Bernoulli}(\pi)$

$$\sum_{k \in U} I_k(s) \sim \text{Binomial}(N, \pi_k), \quad \underbrace{\text{Binomial}(N, \pi)}_{\text{hipótese}}$$

logo,

$$P(n_s = n) = \binom{N}{n} \pi^n (1-\pi)^{N-n}$$

Agora. Note que

$$P(S = s \text{ e } n_s = n) = P((S=s) \cap (n_s = n))$$

$$= \pi^n (1-\pi)^{N-n}$$

Daí,

$$P(S=s \mid n_s = n) = \frac{\cancel{\pi^n (1-\pi)^{N-n}}}{\binom{N}{n} \cancel{\pi^n (1-\pi)^{N-n}}} = \frac{1}{\binom{N}{n}}$$

que coincide com (*), logo queda verificado.

2.

$$\overbrace{9-10}^{1} \quad \overbrace{10-11}^{2} \quad \overbrace{11-12}^{3} \quad \overbrace{12-13}^{4} \quad \overbrace{13-14}^{5}$$

A primeira hora é selecionada com probabilidade de $1/5$ mas a segunda hora é extraída de forma condicional

Se a primeira hora foi 1 ou 5 a probabilidade é igual a $1/3$ pela hipótese, e probabilidade de $1/2$ no resto dos casos.

1ª hora selecionada

2ª hora	1	2	3	4	5
1	0	0	$1/15$	$1/15$	$1/15$
2	0	0	0	$1/10$	$1/10$
3	$1/10$	0	0	0	$1/10$
4	$1/10$	$1/10$	0	0	0
5	$1/15$	$1/15$	$1/15$	0	0

$\} \; P(s)$

As amostras são de tamanho 2

Note que $\sum_s P(s) = 1$

De fato, $\frac{6}{15} + \frac{6}{10} = \frac{150}{150} = 1$

K = primeira hora, l = segunda hora.

A amostra $s = (1,2) \Rightarrow P(K=1, l=2) = 0$

horas contíguas.

$s = (3,1) \Rightarrow P(K=3, l=1) = \frac{1}{5} \cdot \frac{1}{3} = \frac{1}{15}$

$s = (3,2) \Rightarrow P(K=4, l=2) = \frac{1}{5} \cdot \frac{1}{2} = \frac{1}{10}$

Agora lembremos que

$\pi_K = \sum_{s \in S} P(s)$, em que S é o conjunto de todas as amostras, i.e., $\{1,2,3,4,5\} \times \{1,2,3,4,5\}$.

a)

$\pi_1 = \sum_{\substack{s \in S \\ 1 \in s}} P(s) = 4/15 + 2/10 = \frac{8+6}{30} = \frac{12}{30}$

$\pi_2 = \sum_{\substack{s \in S \\ 2 \in s}} P(s) = 3/10 + 1/15 = \frac{11}{30}$

$\pi_3 = \sum_{\substack{s \in S \\ 3 \in s}} P(s) = 2/15 + \frac{2}{10} = \frac{10}{30}$

$\pi_4 = \sum_{\substack{s \in S \\ 4 \in s}} P(s) = \frac{3}{10} + \frac{1}{15} = \frac{11}{30}$

$\pi_5 = \sum_{\substack{s \in S \\ 5 \in s}} P(s) = \frac{4}{15} + \frac{2}{10} = \frac{12}{30}$

Note que

$$E(n_s) = \sum_{U} \pi_k = \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5$$
$$= \frac{12}{30} + \frac{11}{30} + \frac{10}{30} + \frac{11}{30} + \frac{12}{30} = \frac{56}{30}$$
$$= 1,8667 \approx 2.$$

Agora
$$\pi_{kl} = \sum_{\substack{s \in S \\ k \in s \wedge \\ l \in s}} p(s)$$

b)

$\pi_{11} = 0 \quad = 0$

$\pi_{12} = 0 \quad = 0 = \pi_{21}$

$\pi_{13} = 1/15 + \frac{1}{10} = \frac{5}{30} = \pi_{13}$

$\pi_{14} = \frac{1}{15} + 1/10 = 5/30 = \pi_{41}$

$\pi_{15} = \frac{1}{15} + 1/15 = 2/30 = \pi_{51}$

$\pi_{23} = 0$

$\pi_{24} = \frac{1}{10} + \frac{1}{10} = \frac{2}{10}$

$\pi_{25} = \frac{1}{10} + \frac{1}{15} = \frac{5}{30}$

$\pi_{34} = 0$

$\pi_{35} = \frac{1}{10} + \frac{1}{15} = \frac{5}{30}$

$\pi_{45} = 0$

$\pi_{55} = 0$

$$(\pi)_{kl} = \begin{bmatrix} 0 & 0 & 5/30 & 5/30 & 2/30 \\ 0 & 0 & 0 & 6/30 & 5/30 \\ 5/30 & 0 & 0 & 0 & 5/30 \\ 5/30 & 6/30 & 0 & 0 & 0 \\ 2/30 & 5/30 & 5/30 & 0 & 0 \end{bmatrix} = \underset{\sim}{\pi}$$

logo o plano amostral é não mensurável pois tem $\pi_{kl} = 0$.

c)

$$\underset{\sim}{\Delta} = C(I_k, I_l) = \begin{cases} \pi_{kl} - \pi_k \pi_l, & k \neq l \\ \pi_k (1 - \pi_k), & k = l \end{cases}$$

$k, l \in \{1, 2, 3, 4, 5\}$

$$\Delta_{11} = \pi_1(1-\pi_1) = \frac{12}{30}\left(1-\frac{12}{30}\right) = \frac{12}{30} \times \frac{18}{30} = \frac{216}{900}$$

$$\Delta_{12} = \pi_{12} - \pi_1\pi_2 = 0 - \frac{12}{30} \times \frac{11}{30} = -\frac{12}{30} \times \frac{11}{30} = -\frac{132}{900}$$

$$\Delta_{13} = \pi_{13} - \pi_1\pi_3 = 5/30 - 12/30 \times \frac{10}{30} = \frac{30}{900}$$

$$\Delta_{14} = \pi_{14} - \pi_1\pi_4 = 5/30 - 12/30 \times \frac{11}{30} = \frac{18}{900}$$

$$\Delta_{15} = \pi_{15} - \pi_1\pi_5 = 2/30 - \frac{12}{30} \times \frac{12}{30} = -\frac{84}{900}$$

$$\Delta_{22} = \pi_2(1-\pi_2) = \frac{11}{30}\left(1-\frac{11}{30}\right) = \frac{11}{30} \times \frac{19}{30} = \frac{209}{900}$$

$$\Delta_{23} = \pi_{23} - \pi_2\pi_3 = 0 - \frac{11}{30}\left(\frac{10}{30}\right) = -\frac{110}{900}$$

$$\Delta_{24} = \pi_{24} - \pi_2\pi_4 = \frac{2}{10} - \left(\frac{11}{30} \times \frac{11}{30}\right) = \frac{180}{900} - \frac{121}{900} = \frac{59}{900}$$

$$\Delta_{25} = \pi_{25} - \pi_2\pi_5 = \frac{5}{30} - \left(\frac{11}{30} \times \frac{12}{30}\right) = \frac{150}{900} - \frac{132}{900} = \frac{18}{900}$$

$$\Delta_{33} = \pi_3(1-\pi_3) = \frac{10}{30} \times \left(1-\frac{10}{30}\right) = \frac{10}{30} \times \left(\frac{20}{30}\right) = \frac{200}{900}$$

$$\Delta_{34} = \pi_{34} - \pi_3\pi_4 = 0 - \left(\frac{10}{30} \times \frac{11}{30}\right) = -\frac{110}{900}$$

$$\Delta_{35} = \pi_{35} - \pi_3\pi_5 = 5/30 - \left(\frac{10}{30} \times \frac{12}{30}\right) = \frac{150}{900} - \frac{120}{900} = \frac{30}{900}$$

$$\Delta_{44} = \pi_4(1-\pi_4) = \frac{11}{30}\left(1-\frac{11}{30}\right) = \frac{11}{30} \times \left(\frac{19}{30}\right) = \frac{209}{900}$$

$$\Delta_{45} = \pi_{45} - \pi_4\pi_5 = 0 - \frac{11}{30} \times \frac{12}{30} = -\frac{132}{900}$$

$$\Delta_{55} = \pi_5(1-\pi_5) = \frac{12}{30}\left(1-\frac{12}{30}\right) = \frac{12}{30} \times \frac{18}{30} = \frac{216}{900}$$

$$\underset{\sim}{\Delta} = \frac{1}{900}\begin{bmatrix} 216 & -132 & 30 & 18 & -84 \\ -132 & 209 & -110 & 59 & 18 \\ 30 & -110 & 200 & -110 & 30 \\ 18 & 59 & -100 & 209 & -132 \\ -84 & 18 & 30 & -132 & 216 \end{bmatrix}$$

3)

a) Com base nas informações da tabela pode-se obter estimações para o gasto médio da população e I.C.

Seja N = 20000 então o gasto médio estimado é

$$\hat{\bar{t}}_{\pi} = \bar{Y}_{est} = 0,3(1,2) + 0,2(2,4) + 0,5(0,6)$$

$$= \underbrace{1,14 \text{ salários mínimos}}$$ (média populacional)

O gasto total

$$\hat{t}_{\pi} = \hat{Y}_{est} = N(\bar{Y}_{est}) = 20.000\,(1,14) = 22800$$

$\underbrace{\text{salários mínimos}}$

$$Var(\bar{Y}_{est}) = \frac{(0,3)^2(0,36)}{40} + \frac{(0,2)^2(1,21)}{36} + \frac{(0,5)^2(0,04)}{44}$$

$$- \frac{(0,3)(0,36) + (0,2)(1,21) + (0,5)(0,04)}{20000}$$

$$Var(\bar{Y}_{est}) = Var(\hat{\bar{t}}_{\pi}) = \sum_{h=1}^{3} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{3} \frac{W_h S_h^2}{N}, \text{ logo}$$

$$Var(\bar{Y}_{est}) = 0,001262106$$

$d.p(\bar{y}_{est}) = \sqrt{Var(\bar{y}_{est})} = \underline{0{,}03552}$ salários mínimos

logo um intervalo de confiança de 95% para $\bar{y}$ é

$$\bar{y}_{est} \pm t_{n-1; 1-\alpha} \times d.p(\bar{y}_{est})$$

$$[1{,}14 - t_{119} \times 0{,}03552 ; 1{,}14 + t_{119} \times 0{,}03552]$$

O valor de t_{119} é 1,98 (que não é muito diferente de $Z_{1-\alpha}$ que seria 1,96),

$$1{,}14 \pm 1{,}98 \times 0{,}03552$$

$$(1{,}06967 ; 1{,}21033)$$ em salários mínimos.

Agora para o total temos que

$$Var(\hat{t}_\pi) = Var(N \bar{y}_{est}) = N^2 Var(\bar{y}_{est})$$

logo

$$d.p(\hat{t}_\pi) = N\sqrt{Var(\bar{y}_{est})} = N \times d.p(\bar{y}_{est})$$

Daí

$\hat{t}_{\pi} \pm t_{n-1,\, 1-\alpha}\, d_p(\hat{t}_{\pi})$ é um

IC de $(1-\alpha)$% para o total.

Aqui $d.p(\hat{t}_{\pi}) = 710{,}5226$ logo o I.C.

$22\,800 \pm 1{,}98\,(710{,}5226)$ que é

$(21393{,}17 \ ;\ 24206{,}83)$ em salários mínimos.

5) Se o orçamento para a coleta da informação não pode ser superior a R\$ 2,5 milhões podemos pensar da seguinte forma:

O tamanho de amostra numa amostragem estratificada aleatória simples que minimiza a variância da média (ou total) dado um orçamento fixo pode ser visto como

$$n = \frac{(C - C_0)\displaystyle\sum_{h=1}^{3}\left(\frac{W_h S_h}{\sqrt{C_h}}\right)}{\displaystyle\sum_{h=1}^{3}\left(W_h S_h \sqrt{C_h}\right)} \qquad (\ast)$$

lembrando que a função custo pode ser linear, i.e.

$$C = C_0 + \sum_{h=1}^{H} C_h n_h$$

H = número total de estratos.

C = orçamento total para a coleta da informação
C_0 = custo fixo que não depende do número de elementos a selecionar e
C_h = custo de realizar a amostragem no estrato h.

Vimos que o tamanho de amostra ótimo por estrato é proporcional a n e é dado por

$$n_h = n \frac{P_h}{\sum_{h=1}^{H} P_h}, \quad (☺) \quad \text{em que}$$

$$P_h = \frac{W_h S_h}{\sqrt{C_h}}$$ } P_h é uma variável de proporcionalidade

Utilizando (☺) temos

$$n = \frac{2500000 \left(\frac{4.000\sqrt{0{,}36}}{\sqrt{5000}} + \frac{6000\sqrt{1{,}21}}{\sqrt{3000}} + \frac{10000\sqrt{0{,}04}}{\sqrt{1000}} \right)}{M}$$

$$M = 4000\sqrt{0,36}\sqrt{5000} + 6000\sqrt{1,21}\sqrt{6000} + 10000\sqrt{0,04}\sqrt{1000}$$

$$n = \frac{217,6856 \times 25000}{594448.1}$$

$n = 915,4948$ (tamanho global)

$\underbrace{n \cong 916}$

E os tamanhos por estratos devem seguir (☺), i.e

$$\underbrace{n_1} = 915,4948 \times \frac{\left(\frac{4.000\sqrt{0,36}}{\sqrt{5000}}\right)}{217,6856} \approx 142,7422 \underbrace{\approx 143}$$

$$\underbrace{n_2} = 915,4948 \times \frac{\left(\frac{6000\sqrt{1,121}}{\sqrt{3000}}\right)}{217,6856} \approx 506,7682 \underbrace{\approx 507}$$

$$\underbrace{n_3} = 915,4948 \times \frac{\left(\frac{10000\sqrt{0,04}}{\sqrt{1000}}\right)}{217,6856} \approx 265,9843 \underbrace{\cong 266}$$

4) Questão 4 desenvolvida em R (arquivo adjunto)