Universidade Federal Fluminense
Instituto de Matemática e Estatística
Departamento de Estatística uff

Notas de Aula
04 - Amostra Aleatória Simples sem Reposição

1. Distribuição de f_i e probabilidade de inclusão

   Na AASs, sorteamos, sem reposição e igual probabilidade n elementos de $\mathcal{U}$. Assim cada elemento aparece uma única vez na amostra. Vamos considerar amostras ordenadas, para manter a coerência com o estudo da AASc.

   Distribuição de f_i

   Seja f_i o número de vezes que o i-ésimo elemento aparece na amostra, então

   $$f_i \sim Bernoulli\,(\pi_i)\,,$$

   com $E(f_i) = \pi_i$ e $Var(f_i) = \pi_i(1-\pi_i)$.

   Probabilidade de inclusão de primeira ordem

   $$\pi_i = \frac{n}{N}.$$

   Probabilidade de inclusão de segunda ordem

   $$\pi_{ij} = \frac{n}{N}\frac{(n-1)}{(N-1)}.$$

   A covariância entre f_i e f_j é dada por

   $$Cov(f_i, f_j) = -\frac{n}{N^2}\frac{N-n}{N-1}.$$

   Esses mesmos resultados valeriam em caso de amostras não ordenadas?

2. A estatística $t(\boldsymbol{s})$ e suas propriedades

   A estatística $t(\boldsymbol{s})$, total da amostra, definida por

   $$t(\boldsymbol{s}) = t = \sum_{i \in \boldsymbol{s}} Y_i = \sum_{i=1}^{N} f_i Y_i$$

   tem para o plano AASs as seguintes propriedades

   $$E(t) = n\mu \qquad \text{e} \qquad Var(t) = n(1-f)S^2,$$

   em que $f = n/N$ denominada fração amostral.

3. Estimação do total populacional τ

   Com base nas propriedades do t, podemos deduzir um estimador não viesado para o total populacional.

   O estimador

   $$\hat{\tau} = \frac{N}{n} t(\boldsymbol{s})$$

   é um estimador não viesado para o total populacional τ, pois $E(\hat{\tau}) = \tau$ e a sua variância é dada por $Var(\hat{\tau}) = N^2(1-f)S^2/n$.

4. Estimação da média populacional μ

   Da relação entre a média e o total populacionais e amostrais, resulta que

$$\overline{y} = \frac{t(\boldsymbol{s})}{n}$$

   é um estimador não viesado para a média populacional μ, pois $E(\overline{y}) = \mu$ e a sua variância é $Var(\overline{y}) = (1-f)S^2/n$.

5. Estimação da variância populacional S^2

   Dentro do plano AASs, a estatística

$$s^2 = \frac{1}{n-1}\sum_{i \in \boldsymbol{s}}(Y_i - \overline{y})^2$$

   é um estimador não viesado da variância populacional S^2.

   Deste modo, podemos fornecer estimadores não viesados para as variâncias dos estimadores $\hat{\tau}$ e $\overline{y}$, isto é,

$$\widehat{Var(\hat{\tau})} = N^2(1-f)\frac{s^2}{n} \qquad \text{e} \qquad \widehat{Var(\overline{y})} = (1-f)\frac{s^2}{n}.$$

   Exercício 1: Considere uma população com 3 elementos, cuja característica de interesse é a renda familiar. O parâmetro populacional é $\boldsymbol{D} = (12, 30, 18)$. Para esta população $\tau = 60$, $\mu = 20$ e $S^2 = 84$. Definido o plano amostral AASs, com $n = 2$, defina: (i) o espaço amostral induzido pelo plano amostral. (ii) as distribuições amostrais de $\hat{\tau}$, $\overline{y}$ e s^2. (iii) com base nas distribuições amostrais calcule $E(\hat{\tau})$, $Var(\hat{\tau})$, $E(\overline{y})$, $Var(\overline{y})$, $E(s^2)$.

6. Normalidade assintótica e intervalos de confiança

   Pelo TLC, temos as seguintes distribuições assintóticas

$$\frac{\hat{\tau} - E(\hat{\tau})}{DP(\hat{\tau})} \sim N(0,1) \qquad \text{e} \qquad \frac{\overline{y} - E(\overline{y})}{DP(\overline{y})} \sim N(0,1).$$

   Sob AASs, resulta que

$$\frac{\hat{\tau} - \tau}{N\sqrt{(1-f)S^2/n}} \sim N(0,1) \qquad \text{e} \qquad \frac{\overline{y} - \mu}{\sqrt{(1-f)S^2/n}} \sim N(0,1).$$

   A partir destas distribuições assintóticas, podemos obter intervalos de confiança com nível de confiança de aproximadamente igual a $1-\alpha$. Para o total populacional temos

$$IC(\tau, (1-\alpha)\%) = \left(\hat{\tau} - z_{\alpha/2}N\sqrt{(1-f)\frac{s^2}{n}}; \hat{\tau} + z_{\alpha/2}N\sqrt{(1-f)\frac{s^2}{n}}\right),$$

   e para a média populacional temos

$$IC(\mu, (1-\alpha)\%) = \left(\overline{y} - z_{\alpha/2}\sqrt{(1-f)\frac{s^2}{n}}; \overline{y} + z_{\alpha/2}\sqrt{(1-f)\frac{s^2}{n}}\right).$$

7. Determinação do tamanho da amostra

   Controlando o erro absoluto para o estimador da média

   Vamos determinar o tamanho da amostra para que o estimador $\overline{y}$ tenha erro máximo igual a B para um nível de confiança $1-\alpha$, isto é,

$$P(|\overline{y} - \mu| \leq B) \approx 1 - \alpha.$$

Facilmente derivamos da expressão acima que

$$n = \frac{1}{\frac{D}{S^2} + \frac{1}{N}},$$

em que $D = B^2/z^2_{\alpha/2}$.

Controlando o erro relativo para o estimador da média

Vamos determinar o tamanho da amostra para que o estimador $\overline{y}$ tenha erro máximo relativo iguala a B para um nível de confiança $1 - \alpha$, isto é,

$$P(\left|\frac{\overline{y} - \mu}{\mu}\right| \leq B) \approx 1 - \alpha.$$

Facilmente derivamos da expressão acima que

$$n = \frac{1}{\frac{D}{CV^2} + \frac{1}{N}}.$$

Como ficaria para o total?

8. Estimação de proporções

Consideremos o caso em que se tem interesse na estimação da proporção de elementos da população que possuem determinada característica. Neste caso, temos associada a cada elemento da população a variável

$$Y_i = \begin{cases} 1, & \text{se o elemento } i \text{ possui a característica} \\ 0, & \text{caso contrário.} \end{cases}$$

A proporção populacional, o parâmetro de interesse nesse caso é

$$p = \frac{1}{N}\sum_{i=1}^{N} Y_i = \mu,$$

ou seja, a média da variável de interesse é a proporção populacional. A variância populacional neste caso será

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - p)^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - p^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i - p^2 = p - p^2 = p(1-p),$$

e

$$S^2 = \frac{N}{N-1}\sigma^2 = \frac{N}{N-1}p(1-p)$$

Estimador não viciado para a proporção

Como o parâmetro de interesse é a média, sabemos que a média amostral é um estimador não viciado. Neste caso, a média amostral é a proporção amostral

$$\hat{p} = \frac{1}{n}\sum_{i \in \boldsymbol{s}} Y_i$$

e a sua variância é dada por

$$Var(\hat{p}) = (1 - f)\frac{S^2}{n} = \frac{N-n}{N-1}\frac{p(1-p)}{n}.$$

Um estimador não viciado para S^2 é

$$s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}),$$

logo um estimador para a variância de $\hat{p}$ é dado por

$$\widehat{Var(\hat{p})} = (1-f)\frac{s^2}{n} = (1-f)\frac{\hat{p}(1-\hat{p})}{n-1}.$$

Intervalo de confiança

Pelo TLC, temos que

$$\frac{\hat{p}-p}{\sqrt{\frac{N-n}{N-1}\frac{p(1-p)}{n}}} \sim N(0,1)$$

e um intervalo de confiança de nível de confiança de aproximadamente $1-\alpha$ é dado por

$$IC(p,(1-\alpha)\%) = \left(\hat{p} - z_{\alpha/2}\sqrt{(1-f)\frac{\hat{p}(1-\hat{p})}{n-1}}; \hat{p} + z_{\alpha/2}\sqrt{(1-f)\frac{\hat{p}(1-\hat{p})}{n-1}}\right),$$

como $\hat{p}(1-\hat{p}) \leq 1/4$, um intervalo de confiança conservador é dado por

$$IC(p,(1-\alpha)\%) = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{1-f}{4(n-1)}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{1-f}{4(n-1)}}\right),$$

Tamanho da amostra

Como no caso da média, podemos considerar o tamanho da amostra n, de tal forma que

$$P(|\hat{p}-p| \leq B) \approx 1-\alpha,$$

e obtemos um tamanho de amostra dado por

$$n = \frac{N}{\frac{(N-1)D}{p(1-p)+1}}$$

e se considerarmos um cenário mais conservador

$$n = \frac{N}{4(N-1)D+1},$$

em que $D = B^2/z^2_{\alpha/2}$.

9. Otimalidade de $\overline{y}$ na AASc

Considere as variáveis $y_1, \ldots, y_n$, isto é, a variável y_i assume os valores $Y_1, \ldots, Y_N$, com probabilidade $1/N$, ou seja, $P(y_i = Y_j) = 1/N$, $j = 1, \ldots, N$. No plano AASs as variáveis y_i são independentes.

Estimador linear: um estimador linear de μ é uma função $\boldsymbol{d_s}$ dada por

$$\overline{y}_{sl} = \sum_{i=1}^{n} l_i y_i,$$

em que os l_i são constantes conhecidas.

Um estimador $\overline{y}_{sl}$ é não viciado para μ se e somente se

$$\sum_{i=1}^{n} l_i = 1.$$

Com relação à AASc, na classe dos estimadores lineares não viciados, $\overline{y}$ é o de menor variância.

10. Algoritmos de seleção

    (a) Amostra Aleatória Simples com reposição

    Algoritmo convencional

    1. Sorteamos um elemento de $\mathcal{U} = \{1, \ldots, N\}$ com probabilidade $1/N$.

    2. Repetimos o procedimento n vezes

    Isso equivale a n realizações independentes de uma $Unif(0, 1) : a_q, \ldots, a_n$, de modo que

    $$\frac{i-1}{N} < a_i \leq \frac{i}{N} \rightarrow \text{elemento } i \text{ pertence a amostra.}$$

    (b) Amostra Aleatória Simples sem reposição

    Algoritmo convencional

    Sorteamos o k-ésimo elemento $k = 1, \ldots, n$ com probabilidade $\frac{1}{N-(k-1)}$ dentre os $N - (k - 1)$ restantes.

    Algoritmo de Hajek

    1. Associamos a cada elemento de $\mathcal{U} = \{1, \ldots, N\}$ um número pseudo aleatório a_i por meio de N realizações independentes de uma $Unif(0, 1)$.

    2. Ordenamos a população, agora representada pelos pares (i, a_i) de acordo com a_i, obtendo uma permutação aleatória.

    3. Uma amostra aleatória simples é obtida selecionando n elementos consecutivos quaisquer nessa permutação.

11. Comparação entre AASc e AASs

    Quando há dois planos amostrais, é importante saber qual deles é o "melhor". Como dito anteriormente, o critério mais utilizado em Amostragem é o Erro Quadrático Médio (ou a variância para estimadores não viciados). Sendo assim, trabalharemos com o Efeito do Planejamento (EPA), que compara a variância de um plano qualquer com relação a um plano que é considerado padrão. Considerando AASc e AASs, em ambos os casos, $\overline{y}$ é um estimador não viciado para μ, ou seja,

    $$EPA = \frac{Var_{AASs}(\overline{y})}{Var_{AASc}(\overline{y})} = \frac{(1-f)S^2/n}{\sigma^2/n} = \frac{N-n}{N-1}.$$

    Quando $EPA > 1$ temos que o plano no numerador é menos eficiente, quando $EPA < 1$ temos que o plano no denominador é menos eficiente.

    Quando $EPA = 1$?