Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística - ME430

Trabalho de ME430

Grupo

Bragantini, J. RA170844
Nogueira, N. RA204186
Betini, L. RA201357
Cunha, V. RA206493

Prof. Dr. Caio Azevedo

Campinas
2018

Questão 2 da Lista IV

1 Introdução

Este trabalho consiste na aplicação de técnicas aprendidas na disciplina ME430 - Técnicas de Amostragem - em um conjunto de dados da COMVEST. Esse conjunto de dados possui informações de 73498 candidatos ao vestibular 2017 e seus desempenhos nele. O objetivo é estimar i) a média da pontuação total de cada cadiadato, ii) a proporção de candidatos que cursaram todo o ensino médio em escola pública e iii) o total de quartos nas casas de todos os candidatos. A Seção 2 contém análises descritivas a nível populacional e obtidas por meio de amostras pilotos. A Seção 3 apresenta análises inferenciais, estimadores pontuais dos parâmetros de interesse e seus intervalos de confiança. Essas duas seções são dividas em três subseções cada, uma para cada parâmetro: *Média*, *Proporção* e *Total*, que se referem, respectivamente, aos parâmetros i), ii) e iii).

2 Análise Descritiva

Tratamos as análises realizas abaixos como indepentes visto que o âmbito de cada análise são diferentes.

2.1 Média

Inicialmente, para estimar μ, a pontuação média de todos os candidatos na 1ª fase, foi coletada uma amostra piloto de tamanho 200 sob uma amostragem estratificada (AE) com $H = 4$ estratos, especificados a seguir, a fim de determinar o tamanho amostral e o plano amostral mais adequado. Para isso, foi estimado a variância $\sigma^2_\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ da pontuação e a variância nos estratos $\sigma^2_{\mu h} = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{ih}$, onde $N = 73498$ é o número de candidatos, N_h é o tamanho populacional do h-ésimo estrato, y_i é a pontuação do i-ésimo candiadato e y_{ih} é a pontuação do i-ésimo candidato no h-ésimo estrato. A alocação dos estratos foi feita segundo alocação proporcional (AP). Nesse tipo de alocação, o tamanho amostral do h-ésimo estrato é $n_h = n\frac{N_h}{N}$.

Os estratos escolhidos foram as respostas agrupadas da Questão 14 no questionário que cada candidato deveria responder. Essa questão é como segue: "Somando a sua renda com a renda das pessoas que moram com você, quanto é, aproximadamente, a renda familiar mensal? O valor do salário mínimo (SM) é de R$ 724,00". As respostas agrupadas determinam os seguintes estratos: 1 - dados faltantes, 2 - até 5 SM, 3 - entre 5 e 10 SM e 4 - mais que 10 SM.

A amostra piloto resultou em uma estimativa de $\hat{\sigma^2_\mu} = \hat{\sigma^2_d} + \hat{\sigma^2_e} = 397,214$ **(HELP WANTED)** para a variância σ^2_μ, onde $\hat{\sigma^2_d} = \sum_{i=1}^{H} \frac{N_h}{N}\hat{\sigma^2_h}$ é a variância estimada no h-ésimo estrato e $\hat{\sigma^2_e} = \sum_{i=1}^{H} \frac{N_h}{N}(\hat{\mu}_h - \mu)^2$ é a variância estimada das médias dos estratos. De posse dessa informação, para realizar AE com AP e erro de estimativa $\delta = 1$, é preciso um tamanho de amostra n de pelo menos $n \geq \frac{z^2_{0,95}}{\delta^2}\sum_{i=1}^{H} \frac{N_h}{N}\hat{\sigma^2}$ **(OU SERIA S^2 AQUI?)** para garantir que $P(|\mu - \hat{\mu}| \leq \delta) \geq 0,95$ [1], onde $z_{0,95}$ é o 0,95-quantil da normal padrão. Portanto, para os dados coletados, são necessárias pelo menos 870 unidades amostrais.

Por outro lado, uma amostragem aleatória simples sem reposição (AASs) requer $\left(\frac{\delta^2}{\hat{s^2} z_{0,95}} + \frac{1}{N}\right)^{-1}$ unidades amostrais para garantir que $P(|\mu - \hat{\mu}| \leq \delta) \geq 0,95$ [1], onde $s^2_\mu = \frac{n\sigma^2}{n-1}$. Para os dados observados, esse plano amostral precisa de 1060 unidades amostrais, número maior que no caso AE com AP. Portanto, é mais vantajoso realizar AE com AP.

É possível tornar a amostragem ainda mais robusta, utilizando a alocação ótima de Neyman (AON), que minimiza a variância da estimativa $\hat{\mu} = \sum$ para μ quando o custo de amostragem é homogêneo entre os estratos. Usando AON, temos que $n_h = n\frac{N_h\sigma_h}{\sum_{i=1}^{H} N_h\sigma_h}$ [1]. Como σ_h não são valores conhecidos, foi usado $\hat{\sigma^2_h}$. As informações obtidas da amostra piloto e referentes à amostragem AE com AON estão

resumidas na Tabela 1. Observe que $n = 871$ tem 1 unidades a mais sob AON, pelo fato de ter sido pego o menor inteiro maior que a expressão que determina n_h.

Tabela 1: Informações de cada estrato h: N_h - número de cadidatos no estrato, $\hat{\sigma_h^2}$ - variância no estrato estimada na amostra piloto, n_h - tamanho amostral do estrato segundo AON.

h	N_h	$\hat{\sigma_h^2}$	n_h
1	2223	562,5532	35
2	31859	378,2201	409
3	21828	283,4026	243
4	17588	251,2857	184

2.2 Proporção

A nível populacional, observa-se que existe uma correlação de 0,6207 entre o setor (público ou privado) da escola de um candidato em um nível (Ensino Fundamental 1, Ensino Fundalmental 2 e Ensino Médio) ao nível subsequente. De encontro a essa informação, também se observa que alunos que estudaram o Ensino Fundamental 1 completamente em instituições públicas frequentaram o Ensino Fundamental 2 público numa proporção maior dos restantes dos indivíduos **APRESENTAR DADOS QUE CORROBOREM AFIRMAÇÂO**.

Assim dividiremos nossa população em dois estratos: 1 - candidatos que cursaram o Ensino Fundamental 1 por completo em escolas públicas e 2 - candidatos restantes.

Sob essa estratificação, selecionamos uma amostra piloto de tamanho 201, onde cada estrato foi amostrado com um tamanho proporcional a sua população e verificamos se o comportamento de se manter no mesmo setor do ensino quando se passa do Ensino Fundamental 2 para o Ensino Médio é similar ao do observado na Subseção 2.2.

Na Tabela 3 apresentamos as estatísticas da nossa amostra piloto. Utilizando essas informações, vemos que a variância dentro dos estratos, $\hat{s_d^2} = 0,0897$ é consideravelmente menor que a estimativa da variância da amostra, $\hat{s^2} = \hat{s_d^2} + \hat{s_d^2} = 0,2057$.

Com as estatísticas obtidas da amostra piloto estabelecemos nossa margem de erro desejada, $\delta = 0,01$, e nosso intervalo de confiança desejado, $\gamma = 0,95$, assim conforme a Equação 4 apresentada no Apêndice, obtemos o tamanho amostral $n = 2349$.

Tabela 2: Informações de cada estrato h para amostra piloto: $\hat{p}_h$ - proporção de canditatos estudaram o ensino medio completo em escolas públicas, N_h - número de cadidatos no estrato, $\hat{s_h^2}$ - variância no estrato estimada na amostra piloto, n_h - tamanho amostral do estrato.

h	N_h	$\hat{p}_h$	$\hat{s_h^2}$
1	19617	0,8519	0,1286
2	53881	0,0816	0,0755

2.3 Total

No banco de dados há informações sobre eletrônicos domésticos, cômodos das casas e da estrutura familiar dos inscritos no vestibular da COMVEST. Para estimar o total de quartos de uma casa, foi verificado se existe alguma ligação entre a quantidade de banheiro, número de pessoas dependentes da mesma renda familiar e quantidade de televisões na casa com o total de quartos da casa. Essas variáveis foram selecionados supondo-se que normalmente uma casa com muitos quartos possui mais banheiros, mais dependentes e a possibilidade de haverem televisões em cômodos diferentes da casa.

3 Análise Inferencial

3.1 Média

3.2 Proporção

Uma vez definido o tamanho da amostra definimos o tamanho de cada estrato utilizando a Alocação Ótima de Neyman, Equação 4 vide Apêndice, dado isso obtemos nossa amostra, **... continuar blah blah blah**

Tabela 3: Informações de cada estrato h: $\hat{p}_h$ - proporção de canditatos estudaram o ensino medio completo em escolas públicas, N_h - número de cadidatos no estrato, $\hat{s_h^2}$ - variância no estrato estimada na amostra piloto, n_h - tamanho amostral do estrato.

h	N_h	$\hat{p}_h$	$\hat{Var}(\hat{p}_h)$	$\hat{s_h^2}$	n_h
1	19617	0,8349	0,002258	0,138	757
2	53881	0,1042	0,0004337	0,0934	1593

3.3 Total

Tabela 4: Blah blah balh

N	n	$\hat{\mu}$	$\hat{Var}(\hat{\mu})$	$\hat{s^2}$
463	463	2,6878	0	0,7587

4 Conclusões

Referências

[1] Heleno Bolfarine. *Elementos de amostragem*. Blucher, 2005.

Apêndice

Descrição das equacoes blabh blah

Equação para tamanho amostral dado uma amostra estratifica sem reposição:

$$n = \left\lceil \frac{1}{\frac{\delta^2}{z_\gamma \sum_{h=1}^{H} W_h \hat{s_h^2}} + \frac{1}{N}} \right\rceil$$

Adicionar comentários sobre esta formula (aloc otima de neyney)

$$n_h = \left\lceil n \frac{N_h \hat{s}_h}{\sum_{h=1}^{H} N_h \hat{s}_h} \right\rceil$$