

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

Trabalho de ME430

Grupo

Victor 206493, Jordão 170844, Nicole 204186, Leticia 201357

Prof. Dr. Caio Azevedo

Campinas

2018

Questão 2 da Lista IV

1 Introdução

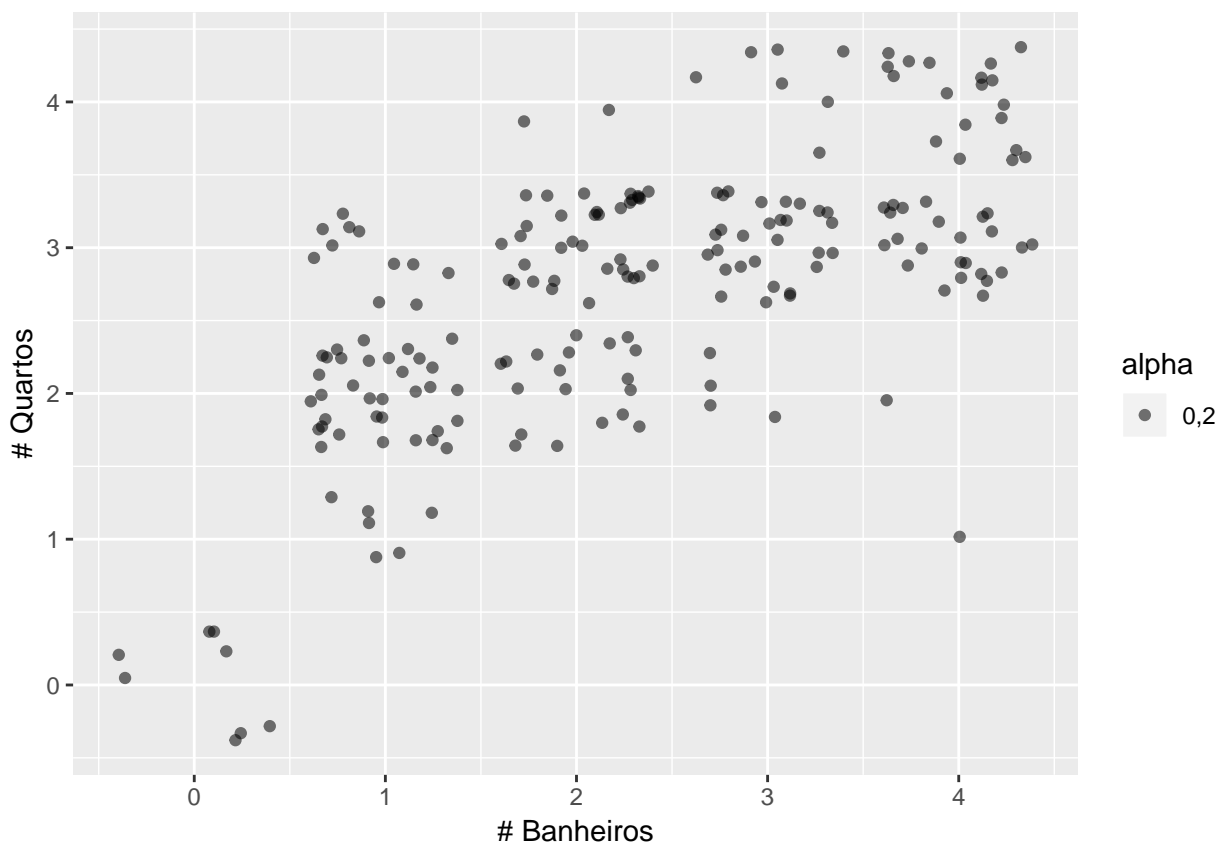
Este trabalho consiste na aplicação de técnicas aprendidas na disciplina ME430 - Técnicas de Amostragem - em um conjunto de dados da COMVEST. Esse conjunto de dados possui informações de 73498 candidatos ao vestibular 2017 e seus desempenhos nele. O objetivo é estimar i) a média da pontuação total de cada candidato, 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 nossas análises sobre os dados, tanto nas informações que conhecemos no nível populacional, quanto nas obtidas por meio de amostras. Em seguida na Seção 3 apresentamos nossas análises, estimadores pontuais das estatísticas de interesse e seus intervalos de confiança. As seções são divididas em subseções para cada estimador, o quais chamamos de, *Média*, *Proporção* e *Total*, respectivamente aos itens i), ii) e iii) da Tarefa.

2 Análise Descritiva

2.1 Média

2.2 Proporção

2.3 Total



```
## $concordant
## [1] 9762
##
## $discordant
```

```
## [1] 900
##
## $extra.x
## [1] 4625
##
## $extra.y
## [1] 2307
##
## $spare
## [1] 2506
##
## $kendall.tau
## [1] 0,6294
##
## $goodman.kruskal.gamma
## [1] 0,8312

## [1] 1,515

## [1] 0,7575
```

3 Análise Inferencial

3.1 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 $\sigma_\mu^2 = \frac{1}{N} \sum_{i=1}^N y_i$ da pontuação e a variância nos estratos $\sigma_{\mu h}^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{ih}$, 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 candidato 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 $\hat{\sigma}_\mu^2 = \hat{\sigma}_d^2 + \hat{\sigma}_e^2 = 331,8968$ (**HELP WANTED**) para a variância σ_μ^2 , onde $\hat{\sigma}_d^2 = \sum_{i=1}^H \frac{N_h}{N} \hat{\sigma}_h^2$ é a variância estimada no h-ésimo estrato e $\hat{\sigma}_e^2 = \sum_{i=1}^H \frac{N_h}{N} (\hat{\mu}_h - \mu)^2$ 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 $n \geq \frac{z_{0.95}^2}{\delta^2} \sum_{i=1}^H \frac{N_h}{N} \hat{\sigma}_h^2$ (**OU SERIA S^2 AQUI?**) para garantir que $P(|\mu - \hat{\mu}| \leq \delta) \geq 0.95$ [1], onde $z_{0.95}$ é o 95-quantil da normal padrão. Portanto, para os dados coletados, são necessárias pelo menos 388 unidades amostrais.

Por outro lado, uma amostragem aleatória simples sem reposição (AASs) requer $\left(\frac{\delta^2}{s_\mu^2} + \frac{1}{N} \right)^{-1}$ para garantir que $P(|\mu - \hat{\mu}| \leq \delta) \geq 0.95$ [1], onde $s_\mu^2 = \frac{n\sigma^2}{n-1}$. Para os dados observados, esse plano amostral precisa de 1254 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 $\hat{\sigma}_h^2$. As informações obtidas da amostra piloto e referentes à amostragem AE com AON estão resumidas na Tabela 1. Observe que $n = 391$ tem 3 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
0	2223	596,8	17
1	31859	312,0027	170
2	21828	269,1964	108
4	17588	324,1968	96

3.2 Proporção

3.3 Total

4 Conclucões

Referências

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