Amostragem

Luiz Fernando, Lucas de Mattos, Gabriel Tracina

19 de novembro de 2019

```
library(dplyr)
library(purrr)
library(tidyr)
library(magrittr)

munic = read.csv2("dados_municipioV.csv")
renda = read.csv2("dados_renda.csv")

df = inner_join(munic, renda) # Base de dados
```

Questão 1

Deseja-se realizar uma pesquisa a fim de fazer um levantamento para estimar a porcentagem de indivíduos que possuem a renda superior a R$ 1500,00. A população considerada para tal pesquisa foram os indivíduos moradores do município V. Para tal foi implementado um modelo de amostra probabilística do tipo amostra aleatória simples sem reposição, que para ser efetuado é necessário um cadastro prévio da população observada, e para os moradores do município V já se tem um cadastro contendo informações sobre o sexo, escolaridade do pai e a localidade do município onde o morador reside. Para estimar a proporção de renda maior que mil e quinhentos reais, primeiramente com o cadastro em mãos e com uma amostragem aleatória sistemática definida como, ir de três em três casas coletando informações sobre a renda do morador ou moradores, será de responsabilidade do entrevistador realizar a pesquisa nas casas, preenchendo um formulário de acordo com as respostas do morador. Caso não houver ninguém em casa, é instruído ao entrevistador voltar no dia seguinte na mesma casa. E se mesmo assim não houver sucesso é orientado a efetuar a pesquisa na casa ao lado. Com a listagem da renda feita, o segundo passo é definir o tamanho da amostra, que será representada por "n", e tem sua fórmula observada abaixo:

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

Onde N é o tamanho da população, $p = \hat{p}$ que é a proporção estimada e $D = \frac{B^2}{z^2_{\frac{\alpha}{2}}}$, onde B^2 é o erro máximo relativo para um nível de confiança de $1 - \alpha$ e $z\alpha/2^2$ é o quantil relacionado a probabilidade $\frac{\alpha}{2}$. Após determinado o tamanho da amostra, será realizado um sorteio aleatório para definir os n selecionados para a entrevista, para assim estimar a proporção de moradores com renda maior que R$1500,00.

Para estimar a variância do estimador, ou seja, estimar a variância de $\hat{p}$ será usado a fórmula abaixo:

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

Onde $\hat{p}$ é o estimador da proporção e $f = \frac{n}{N}$.

Assim, vamos calcular n e $\hat{p}$.

```
# Cálculo de 'D'
set.seed(2019); D = ((0.05)^2)/(qnorm(0.975))^2
```

```r
# Amostra piloto
set.seed(2019); amost_piloto = sample_n(df, 50)

amost_phat = amost_piloto %>% filter(renda>=1500)

# Cálculo do 'p chapéu'
phat = nrow(amost_phat)/nrow(amost_piloto)
phat
```

```
## [1] 0.32
```

```r
df_modificada = anti_join(df, amost_piloto, by = "codigo")

N = nrow(df_modificada)

# Cálculo de 'n'
n = (N/( ( (N-1)*D/phat*(1-phat) ) + 1 )) %>% ceiling() # ceiling() arredonda para cima
n
```

```
## [1] 632
```

Para realizar o processo de amostragem utilizaremos o algoritmo de Hajek:

```r
set.seed(2019); hajek = runif(nrow(df_modificada), 0, 1)

df_hajek = cbind(df_modificada, hajek) %>% arrange(hajek) %>% .[1:n, ]
df_hajek %>% head()
```

```
##   codigo localidade escolaridade_pai sexo   renda        hajek
## 1   2059          B            Medio    M  997.70 0.0001178489
## 2   5578          C      Fundamental    M 1506.93 0.0001771618
## 3   4860          C            Medio    M 1499.94 0.0003254064
## 4   2052          B         Superior    F 1011.38 0.0005220606
## 5   4245          B       Analfabeto    F  997.04 0.0006059173
## 6   3395          B            Medio    F 1000.08 0.0006086985
```

```r
sumYi_amostra = nrow(df_hajek %>% filter(renda>=1500))
sumYi_amostra
```

```
## [1] 207
```

```r
phat = sumYi_amostra/n
phat
```

```
## [1] 0.3275316
```

Assim, temos que a variância do estimador é dada por:

```r
varhat = (1-n/N)*( phat*(1-phat)/(n-1) )
varhat
```

```
## [1] 0.0003044901
```

Questão 2

Estimativa pontual

Com base na questão anterior, temos que nossa estimativa pontual é o $\hat{p}$, obtido pela fórmula:

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

Assim,

```
phat = nrow(df_hajek %>% filter(renda>=1500))/n
phat
```

```
## [1] 0.3275316
```

Estimativa intervalar

O intervalo de confiança de nível de confiança de aproximadamente 95% é dado por:

$$IC(p, 95\%) = (\hat{p} - z_{\frac{\alpha}{2}}\sqrt{(1-f)\frac{\hat{p}(1-\hat{p})}{n-1}}; \hat{p} + z_{\frac{\alpha}{2}}\sqrt{(1-f)\frac{\hat{p}(1-\hat{p})}{n-1}})$$

Assim,

```
set.seed(2019); zphat = qnorm(0.975)*sqrt( (1-(n/N) )*( (phat*(1-phat))/(n-1) ) )

ic = paste("IC(p, 0.95%):", "[", phat-zphat %>% round(4), ";", phat+zphat %>% round(4), "]")
ic
```

```
## [1] "IC(p, 0.95%): [ 0.29333164556962 ; 0.36173164556962 ]"
```

Questão 3

Para realizar uma pesquisa a fim de estimar a proporção de indivíduos com renda maior que R$1500,00 foi considerado uma população, os moradores do município V, e para realizar a estimação será implementada a técnica de amostragem estratificada com alocação proporcional sem reposição, onde os estratos serão diferenciados pela localidade do município onde o morador reside. Para utilizar este modelo é necessário que a população tenha um cadastro prévio com informações sobre os moradores. Satisfazendo esta condição, tendo um cadastro dos moradores do município V, com informações sobre a escolaridade do pai, sexo e localidade onde o morador reside dentro do município. Com o cadastro em mãos, o método de coleta de informações sobre a renda será por uma amostragem aleatória sistemática, onde o entrevistador a partir de uma casa inicial é instruído a ir de três em três casas solicitando ao morador ou moradores que respondam o formulário sobre a renda. Caso na casa onde será feita a entrevista não houver ninguém no momento, é instruído ao entrevistador que volte no dia seguinte nesta mesma casa para tentar mais uma vez. Caso mesmo assim não houver ninguém em casa, é orientado que prossiga a pesquisa na casa seguinte. Dado esta listagem de informações feita, será calculado o tamanho da amostra em cada estrato, que será definido por n_h e tem sua fórmula observada abaixo:

$$n_h = nW_h$$

Onde n é o tamanho da amostra e N_h é o tamanho da população no estrato h e N é o tamanho total da população.

```
n_piloto = 5

set.seed(2019); amostra_piloto = df %>%
  group_by(localidade, escolaridade_pai, sexo) %>%
  group_split() %>%
```

```
  map(~ .x %>% sample_n(n_piloto)) %>%
  bind_rows()

amostra_piloto
```

```
## # A tibble: 120 x 5
##    codigo localidade escolaridade_pai sexo  renda
##     <int> <fct>      <fct>            <fct> <dbl>
##  1   1028 A          Analfabeto       F     1991.
##  2   1064 A          Analfabeto       F     2067.
##  3   1547 A          Analfabeto       F     2033.
##  4   1833 A          Analfabeto       F     2009.
##  5   1684 A          Analfabeto       F     2060.
##  6   1991 A          Analfabeto       M     2026.
##  7   1805 A          Analfabeto       M     2019.
##  8   1536 A          Analfabeto       M     2005.
##  9   1960 A          Analfabeto       M     1958.
## 10   1056 A          Analfabeto       M     2043.
## # ... with 110 more rows
```

Localidade

```
var_loc =
  amostra_piloto %>%
  nest(-localidade) %>%
  mutate(vars = map(
    data, ~ .x %$% var(renda)
  )) %$% vars %>% unlist() %>% sum()
```

Escolaridade do pai

```
var_escol =
  amostra_piloto %>%
  nest(-escolaridade_pai) %>%
  mutate(vars = map(
    data, ~ .x %$% var(renda)
  )) %$% vars %>% unlist() %>% sum()
```

Sexo

```
var_sexo =
  amostra_piloto %>%
  nest(-sexo) %>%
  mutate(vars = map(
    data, ~ .x %$% var(renda)
  )) %$% vars %>% unlist() %>% sum()
```

Menor variância

```
df_vars = data.frame("Localidade" = var_loc,
                     "Escolaridade do pai" = var_escol,
                     "Sexo" = var_sexo)
```

```
df_vars
```

```
##   Localidade Escolaridade.do.pai   Sexo
## 1   2272.532            682936.1 335801
```

```
which.min(df_vars)
```

```
## Localidade
##          1
```

Assim, temos que a variável *localidade* possui a menor variância em relação a renda. Logo, utilizaremos a variável *localidade* para estratificar a população.

Agora calcularemos o valor de \hat{p_h} baseado na amostra piloto e de n, utilizando a fórmula

$$n = \frac{\sum_{h=1}^{H} W_h S_h^2}{D + \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2}$$

```
Y1 = amostra_piloto %>% filter(localidade=="A" & renda>=1500) %>% nrow()
Y2 = amostra_piloto %>% filter(localidade=="B" & renda>=1500) %>% nrow()
Y3 = amostra_piloto %>% filter(localidade=="C" & renda>=1500) %>% nrow()

data.frame(Y1, Y2, Y3)
```

```
##   Y1 Y2 Y3
## 1 40  0 14
```

```
N1_amostra = amostra_piloto %>% filter(localidade=="A") %>% nrow()
N2_amostra = amostra_piloto %>% filter(localidade=="B") %>% nrow()
N3_amostra = amostra_piloto %>% filter(localidade=="C") %>% nrow()

data.frame(N1_amostra, N2_amostra, N3_amostra)
```

```
##   N1_amostra N2_amostra N3_amostra
## 1         40         40         40
```

```
N_amostra = N1_amostra+N2_amostra+N3_amostra
N_amostra
```

```
## [1] 120
```

```
phat1 = Y1/N1_amostra
phat2 = Y2/N2_amostra
phat3 = Y3/N3_amostra

data.frame(phat1, phat2, phat3)
```

```
##   phat1 phat2 phat3
## 1     1     0  0.35
```

```
W1 = (N1_amostra/N_amostra)
W2 = (N2_amostra/N_amostra)
W3 = (N3_amostra/N_amostra)

data.frame(W1, W2, W3, "soma" = sum(W1, W2, W3))
```

```
##          W1        W2        W3 soma
## 1 0.3333333 0.3333333 0.3333333    1
```

```r
S1 = ( N1_amostra/(N1_amostra-1) )*( phat1*(1-phat1) )
S2 = ( N2_amostra/(N2_amostra-1) )*( phat2*(1-phat2) )
S3 = ( N3_amostra/(N3_amostra-1) )*( phat3*(1-phat3) )

data.frame(S1, S2, S3)
```

```
##   S1 S2        S3
## 1  0  0 0.2333333
```

```r
numerador = (W1*S1)+(W2*S2)+(W3*S3)

D = ((0.03)^2)/(qnorm(0.975))^2

denominador = D + ((1/N)*numerador)

n = (numerador/denominador) %>% ceiling()
n
```

```
## [1] 312
```

Encontrando os n_h.

```r
df_modificada_h = anti_join(df, amostra_piloto, by = "codigo")

N1 = df_modificada_h %>% filter(localidade=="A") %>% nrow()
N2 = df_modificada_h %>% filter(localidade=="B") %>% nrow()
N3 = df_modificada_h %>% filter(localidade=="C") %>% nrow()

N = N1+N2+N3

W1 = (N1/N)
W2 = (N2/N)
W3 = (N3/N)

n1 = ( n*W1 )
n2 = ( n*W2 )
n3 = ( n*W3 )

data.frame(n1, n2, n3)
```

```
##         n1       n2       n3
## 1 61.37705 157.2787 93.34426
```

```r
# Assim,

n1 = ( n*W1 ) %>% floor()
n2 = ( n*W2 ) %>% round()
n3 = ( n*W3 ) %>% round()

data.frame(n1, n2, n3)
```

```
##   n1  n2 n3
## 1 61 157 93
```

```r
sum(n1, n2, n3)
```

```
## [1] 311
```

Assim, $\hat{p}_{es} = \sum_{h=1}^{H} W_h \hat{p}_h$, onde $\hat{p}_h = \frac{1}{n_h} \sum_{i \in S} Y_{i_h}$

```
set.seed(2019); sumY1_amostra = df_modificada_h %>% sample_n(n1) %>%
  filter(localidade=="A" & renda>=1500) %>% nrow()
set.seed(2019); sumY2_amostra = df_modificada_h %>% sample_n(n2) %>%
  filter(localidade=="B" & renda>=1500) %>% nrow()
set.seed(2019); sumY3_amostra = df_modificada_h %>% sample_n(n3) %>%
  filter(localidade=="C" & renda>=1500) %>% nrow()

data.frame(sumY1_amostra, sumY2_amostra, sumY3_amostra)
```

```
##   sumY1_amostra sumY2_amostra sumY3_amostra
## 1             9             0            14
```

```
phat_h1 = sumY1_amostra/n1
phat_h2 = sumY2_amostra/n2
phat_h3 = sumY3_amostra/n3

data.frame(phat_h1, phat_h2, phat_h3)
```

```
##    phat_h1 phat_h2   phat_h3
## 1 0.147541       0 0.1505376
```

```
phat_es = (W1*phat_h1) + (W2*phat_h2) + (W3*phat_h3)
phat_es
```

```
## [1] 0.07406235
```

Calculado o tamanho da amostra em cada estrato é possível calcular a estimativa da variância do estimador, onde o estimador é a proporção de indivíduos com renda maior que R$1500,00. A fórmula pode ser observada abaixo:

$$\widehat{Var(\hat{p}_{es})} = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{\hat{p_h}(1 - \hat{p_h})}{n_h - 1}$$

Calculando a estimativa para a variância do estimador usado neste contexto:

```
parcela_1 = (W1^2)*( 1-(n1/N1) )*( phat_h1*(1-phat_h1)/(n1-1) )
parcela_2 = (W2^2)*( 1-(n2/N2) )*( phat_h2*(1-phat_h2)/(n2-1) )
parcela_3 = (W3^2)*( 1-(n3/N3) )*( phat_h3*(1-phat_h3)/(n3-1) )

varhat_phat_es = parcela_1 + parcela_2 + parcela_3
varhat_phat_es
```

```
## [1] 0.0001924557
```

Questão 4

Estimativa pontual

Com base na questão anterior, temos que nossa estimativa pontual é o $\hat{p}_{es}$, obtido pela fórmula:

$$\hat{p}_{es} = \sum_{h=1}^{H} W_h \hat{p}_h$$

Assim,

```
phat_es = (W1*phat_h1) + (W2*phat_h2) + (W3*phat_h3)
phat_es
```

```
## [1] 0.07406235
```

Estimativa intervalar

O intervalo de confiança de nível de confiança de aproximadamente 95% é dado por:

$$IC(p, 95\%) = (\hat{p}_{es} - z_{\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{H} W_h^2(1-f_h)\frac{\hat{p_h}(1-\hat{p_h})}{n_h - 1}}; \hat{p}_{es} + z_{\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{H} W_h^2(1-f_h)\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}})$$

Assim, utilizando resultados de questões anteriores (`parcela_1`, `parcela_2` e `parcela_3`):

```
set.seed(2019); z_alfa2 = qnorm(0.975)

zphat_es = z_alfa2*sqrt(parcela_1 + parcela_2 + parcela_3)

ic = paste(
  "IC(p, 0.95%):", "[", phat_es-zphat_es %>% round(4), ";", phat_es+zphat_es %>% round(4), "]"
)
ic
```

```
## [1] "IC(p, 0.95%): [ 0.0468623546104209 ; 0.101262354610421 ]"
```

Questão 5

Para verificarmos qual plano amostral é o mais eficiente na execução da pesquisa, vamos aplicar o método do Efeito do Planejamento. Dessa maneira, a partir das variâncias observadas em cada um dos planos vamos tomar a decisão de qual plano amostral vamos adotar.

A fórmula do Efeito do Planejamento é dada por:

$$EPA = \frac{V_{AE_{pr}}(\hat{p})}{V_{AAS_S}(\hat{p})} = \frac{\widehat{Var(\hat{p}_{es})}}{\widehat{Var(\hat{p})}}$$

.

Veja que se $EPA > 1$, o plano no numerador é o menos eficiente, no entanto se $EPA < 1$, o plano no denominador é o menos eficiente e caso $EPA = 1$ ambos os planos são de igual maneira eficientes.

```
EPA = varhat_phat_es/varhat
EPA
```

```
## [1] 0.6320591
```

Logo, podemos concluir, como esperado, que o plano AAS_s é menos eficiente em relação ao plano AE_{pr}.