Amostragem

Luiz Fernando, Lucas de Mattos, Gabriel Tracina

14 de novebro de 2019

```
library(dplyr)

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

df = inner_join(munic, renda)
```

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. 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, que para esta pesquisa são os moradores do município V, onde já se tem um cadastro contendo informações suficientes para a efetuar a amostragem simples. Dado esta condição satisfeita, o primeiro passo é definir o tamanho da amostra, que será representada por "n", que tem sua fórmula representada 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 $\hat{p}$.

Utilizaremos em nossos cálculos um erro (B) de 0.05.

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

set.seed(2019); amost_piloto = sample_n(df, 50)
amost_phat = amost_piloto %>% filter(renda>=1500)

phat = nrow(amost_phat)/nrow(amost_piloto)

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

N = nrow(df_modificada)
```

```r
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))

phat = sumYi_amostra/n

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,

```r
phat = (nrow(df_hajek %>% filter(renda>=1500))/n) %>% round(2)
phat
```

```
## [1] 0.33
```

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,

```r
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%): [ 0.2957 ; 0.3643 ]"
```

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, que 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, já se tem um cadastro dos moradores do município V, com isso é possível prosseguir com a amostragem estratificada. Dentro de cada estrato é definido um tamanho de amostra que será definido por n_h, que tem sua fórmula 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.

```r
h = df %>% select(localidade) %>% distinct() %>% nrow()

h1 = df %>% filter(localidade=="A")
h2 = df %>% filter(localidade=="B")
h3 = df %>% filter(localidade=="C")

set.seed(2019); piloto_A = sample_n(h1, 10)
set.seed(2019); piloto_B = sample_n(h2, 10)
set.seed(2019); piloto_C = sample_n(h3, 10)

phat1 = (piloto_A %>% filter(renda>=1500) %>% nrow())/10
phat2 = (piloto_B %>% filter(renda>=1500) %>% nrow())/10
phat3 = (piloto_C %>% filter(renda>=1500) %>% nrow())/10
```

Descartando a amostra piloto do nosso banco de dados:

```r
h1_modificada = anti_join(h1, piloto_A, by = "codigo")
h2_modificada = anti_join(h2, piloto_B, by = "codigo")
h3_modificada = anti_join(h3, piloto_C, by = "codigo")

N1 = nrow(h1_modificada)
N2 = nrow(h2_modificada)
N3 = nrow(h3_modificada)

W1 = (N1/N) %>% round(1)
W2 = (N2/N) %>% round(1)
W3 = (N3/N) %>% round(1)

sum(W1, W2, W3)
```

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

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

```
phat_es = (W1*phat1) + (W2*phat2) + (W3*phat3)
phat_es
```

```
## [1] 0.29
```

Encontrando os n_h. Utilizaremos o valor de n já calculado em questão anterior:

```
n1 = n*W1
n2 = n*W2
n3 = n*W3

sum(n1, n2, n3)
```

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

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:

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

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

```
parcela_1 = (W1^2)*( 1-(n1/N1) )*( phat1*(1-phat1)/(n1-1) )
parcela_2 = (W2^2)*( 1-(n2/N2) )*( phat2*(1-phat2)/(n2-1) )
parcela_3 = (W3^2)*( 1-(n3/N3) )*( phat3*(1-phat3)/(n3-1) )

varhat_phat_es = parcela_1 + parcela_2 + parcela_3
varhat_phat_es %>% round(6)
```

```
## [1] 0.000087
```

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*phat1) + (W2*phat2) + (W3*phat3)
phat_es
```

```
## [1] 0.29
```

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%): [ 0.2717 ; 0.3083 ]"
```