PROJETO DE AMOSTRAGEM

MOISÉS SALES

Resumo. O seguinte projeto visa utilizar técnicas de amostragem, em conjunto com um banco de dados, para estimar a média das variáveis IDH e Esperança de Vida no Brasil no ano de 2010.

1. Introdução

O presente relatório busca utilizar os conhecimentos e técnicas de amostragem para realizar a estimação da média das variáveis IDH e Esperança de Vida no Brasil em 2010. Será utilizado um banco de dados contendo as variáveis, nos anos de 1990 e 2000, para 5.565 municípios do Brasil. Esse banco de dados pode ser encontrado no arquivo CAD. Serão descritas as seguintes etapas: a escolha do plano amostral utilizado; a estimação do tamanho de amostra; a apresentação dos estimadores para os parâmetros de acordo com o plano amostral escolhido, e, também, um estimador centrado para a variância; a seleção da amostra e a geração das estimativas; por fim, a conclusão e o código computacional utilizado.

2. Metodologia

O plano amostral utilizado será **Amostragem Estratificada**, e cada elemento do estrato será selecionado por meio de uma **Amostra Aleatória Simples sem Reposição**. A motivação da escolha desse plano é de certa forma intuitiva, visto que temos 5 regiões distintas no banco de dados; também podemos perceber pela figura 1 que, ao dividirmos o banco de dados em 5 estratos, sendo cada um uma respectiva região, temos estratos heterogêneos entre si, e os valores em cada estrato são homogêneos, sendo este um bom indício de que a utilização desse plano amostral pode ser proveitosa. Sendo divido assim, a região norte como o estrato $H = 1$, a região Nordeste como o estrato $H = 2$, a região Centro Oeste como o estrato $H = 3$, a região Sudeste como o estrato $H = 4$ e a região Sul como o estrato $H = 5$.

Date: Setembro 2024.
Key words and phrases. Amostragem, Estimativas, Brasil.

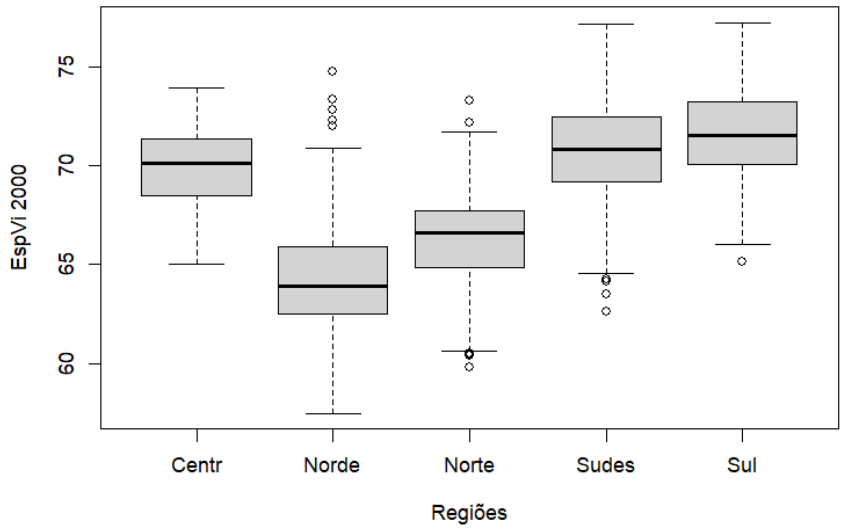

FIGURA 1. Boxplots da Esperança de Vida, por região do Brasil, no ano de 2000.

3. Estimadores e Estimativas

3.1. **Tamanho da amostra.** Por Bolfarine e Oliveira Bussab (2005) o estimador da média populacional sob uma AAS tem a seguinte forma:

$$\bar{y} = \frac{1}{n}\sum_{i\in\mathbf{s}}^{n} Y_i \tag{1}$$

com isso, podemos determinar o tamanho n da amostra da seguinte forma:

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

em que B é um erro máximo admitido. Ao desenvolver a equação, obtemos:

$$n = \frac{1}{(D/S^2) + (1/N)} \tag{2}$$

em que $D = B^2/z_\alpha^2$. No caso da amostragem estratificada, como realizaremos uma AAS em cada estrato, (2) será o tamanho da amostra de cada estrato, e o tamanho total da amostra será a soma do tamanho de cada estrato.

Primeiro é necessário estimar o valor de S^2, para isso vamos recorrer a informação do levantamento do ano 2000. Considerando S_{yH}^2 como a variância da variável idh00 para cada estrato $H = 1, ..., 5$; temos:

Região	Norte	Nordeste	Cent_Oeste	Sudeste	Sul
S^2_{yH}	0.005	0.004	0.003	0.006	0.003

TABELA 1. Variância da variável IDH, por região do Brasil, no ano 2000.

Considerando um nível de confiança de 95% temos que $z^2_{5\%} = 3.8416$, assumindo uma margem de erro máxima de 0.01, temos que $B^2 = 0.01^2 = 0.0001$. Calculando o tamanho de amostra para cada estrato temos:

- Região Norte: $N_1 = 449$; $S^2_{y1} = 0.005$; $\mathbf{n_1} = 135$;
- Região Nordeste: $N_2 = 1794$; $S^2_{y2} = 0.004$; $\mathbf{n_2} = 139$;
- Região Centro-Oeste: $N_3 = 466$; $S^2_{y3} = 0.003$; $\mathbf{n_3} = 91$;
- Região Sudeste: $N_4 = 1668$; $S^2_{y4} = 0.006$; $\mathbf{n_4} = 196$;
- Região Sul: $N_5 = 1188$; $S^2_{y5} = 0.003$; $\mathbf{n_5} = 117$;

Logo, o tamanho total de nossa amostra é $\mathbf{n}_k = \sum_{h=1}^{5} n_h = 678$.

Como queremos estimar a média de duas variáveis, precisamos também calcular qual seria o tamanho da amostra para a variável Esperança de Vida e , após isso, utilizar o maior tamanho de amostra entre as duas.

Região	Norte	Nordeste	Cent_Oeste	Sudeste	Sul
S^2_{yH}	4.639	6.577	3.284	5.223	5.037

TABELA 2. Variância da variável Esperança de Vida, por região do Brasil, no ano 2000.

Realizando os mesmos cálculos anteriores e considerando $B = 0.5$ ano, logo $B^2 = 0.25$, temos:

- Região Norte: $N_1 = 449$; $S^2_{y1} = 4.639$; $\mathbf{n_1} = 62$;
- Região Nordeste: $N_2 = 1794$; $S^2_{y2} = 6.577$; $\mathbf{n_2} = 96$;
- Região Centro-Oeste: $N_3 = 466$; $S^2_{y3} = 3.284$; $\mathbf{n_3} = 46$;
- Região Sudeste: $N_4 = 1668$; $S^2_{y4} = 5.223$; $\mathbf{n_4} = 77$;
- Região Sul: $N_5 = 1188$; $S^2_{y5} = 5.037$; $\mathbf{n_5} = 73$;

Logo, o tamanho total de nossa amostra é $\mathbf{n}_l = \sum_{h=1}^{5} n_h = 354$.

Por fim, utilizaremos o maior tamanho de amostra encontrado, $\mathbf{n} = 678$.

3.2. **Estimadores.** Como já foi visto, utilizaremos o estimador

$$\bar{y}_{AAS} = \frac{1}{n} \sum_{i \in \mathbf{s}}^{n} Y_i$$

A variância do estimador é dada por:

$$\mathrm{Var}_{AAS}(\bar{y}_{AAS}) = (1 - f)\frac{S^2}{n}$$

em que $f = n/N$.

Sabemos que a variância amostral

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

é um estimador não viesado da variância populacional S^2 para o planejamento AAS. E para um plano AAS, a estatística

$$\mathrm{var}_{AAS}(\bar{y}) = \widehat{\mathrm{Var}}_{AAS}(\bar{y}) = (1-f)\frac{s^2}{n}$$

é um estimador não viesado de $\mathrm{Var}_{AAS}(\bar{y})$.

Por Bolfarine e Oliveira Bussab (2005) vemos que, no caso de uma amostragem estratificada, o estimador

$$\bar{y}_{es} = \frac{1}{N}\sum_{h=1}^{H} N_h\bar{y}_h = \sum_{h=1}^{H} W_h\bar{y}_h \tag{3}$$

é um estimador não viesado para a média populacional $\bar{Y}$ com

$$\mathrm{Var}_{AAS}(\bar{y}_{es}) = \sum_{h=1}^{H} W_h^2\mathrm{Var}_{AAS}(\bar{y}_h). \tag{4}$$

Um estimador não viesado para (4) é dado por

$$\mathrm{var}_{AAS}(\bar{y}_{es}) = \widehat{\mathrm{Var}}_{AAS}(\bar{y}_{es}) = \sum_{h=1}^{H} W_h^2\mathrm{var}_{AAS}(\bar{y}) = \sum_{h=1}^{H} W_h^2(1-f_h)\frac{s_h^2}{n_h} \tag{5}$$

3.3. **Alocação da amostra pelos estratos.** A amostra de tamanho $\mathbf{n} = 678$ será distribuída proporcionalmente ao tamanho dos estratos, da seguinte forma (Bolfarine e Oliveira Bussab 2005)

$$n_h = nW_h = n\frac{N_h}{N}.$$

O estimador $\bar{y}_{es}$ se torna

$$\bar{y}_{es} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i\in\mathbf{s}_h} Y_{hi},$$

com variância igual a

$$\mathrm{V}_{pr} = \mathrm{Var}_{AAS}(\bar{y}_{es}) = \sum_{h=1}^{H} W_h^2\mathrm{Var}_{AAS}(\bar{y}_h) = \sum_{h=1}^{H} W_h^2(1-f)\frac{S^2}{n},$$

que é estimado por

$$\mathrm{var}(\bar{y}_{es}) = \sum_{h=1}^{H} W_h^2(1-f)\frac{s_h^2}{n_h}.$$

O tamanho de amostra de cada estrato será:

- Região Norte: 55;
- Região Nordeste: 219;
- Região Centro-Oeste: 57;
- Região Sudeste: 203;
- Região Sul: 145;

3.4. **Estimativas.** Após a seleção da amostra conforme foi descrito por este relatório, chegamos as seguintes estimativas:

3.4.1. *Estimativas da média populacional e sua variância.* As estimativas resultantes foram:

$$\bar{y}_{es}(\text{idh10}) = 0.6573$$
$$\text{var}(\bar{y}_{es})(\text{idh10}) = 2.623066e-06$$

$$\bar{y}_{es}(\text{espvi10}) = 73.05935$$
$$\text{var}(\bar{y}_{es})(\text{espvi10}) = 0.00307$$

3.4.2. *Intervalos de confiança.* Um intervalo de confiança para $\bar{Y}$ com 95% de confiança é dado por:

$$\bar{Y} \in \left(\bar{y}_{es} - z_\alpha \sqrt{\sum_{h=1}^{H} W_h^2 (1-f_h) \frac{s_h^2}{n_h}};\ \bar{y}_{es} + z_\alpha \sqrt{\sum_{h=1}^{H} W_h^2 (1-f_h) \frac{s_h^2}{n_h}} \right).$$

No nosso caso, isso resulta em

$$\bar{Y}(\text{idh10}) \in [0.65412; 0.66047];$$
$$\bar{Y}(\text{espvi10}) \in [72.95078; 73.16792].$$

Os valores das médias populacionais das variáveis, utilizando o arquivo RESP, são:

$$\bar{Y}(idh10) = 0.65916$$
$$\bar{Y}(espvi10) = 73.089$$

4. Conclusão

Podemos notar que ambas as estimativas feitas pelo processo descrito apontam uma ótima precisão, sendo notória a eficácia dos processos de amostragem. Ao invés de visitarmos e analisarmos 5565 municípios, analisamos apenas uma amostra de 678 que nos produziu valores tão precisos quanto, e, além disso, possui uma agilidade muito maior na produção de estimativas, tudo isso com um custo financeiro muito inferior.

5. Código computacional

```
library(readxl)
library(dplyr)

set.seed(1)

#importando CAD
dataset = read_excel("CAD.xlsx")

#importando RESP
dataset2 = read_excel("RESP.xlsx")

#boxplots
boxplot(espvi00 ~ regiao, data=dataset, xlab="Regiões",
                                                ylab="EspVi 2000")

regioes = c("Norte", "Norde", "Centr", "Sudes", "Sul")
N = count(dataset)
#estimar tamanho da amostra idh
tam_amostra_idh = function() {
  tamanho = vector()
  #erro maximo admitido
  D = 0.0001 / 3.8416
  for (x in regioes){
    s2_h = var(subset(dataset, regiao==x)$idh00)
    N_h = count(subset(dataset, regiao==x))
    tamanho = append(tamanho, ceiling(1 / ((D/s2_h) + (1/N_h))) )
  }
  return(tamanho)
}

#estimar tamanho da amostra espvi
tam_amostra_espvi = function() {
  tamanho = vector()
  #erro maximo admitido
  D = 0.25 / 3.8416
  for (x in regioes){
    s2_h = var(subset(dataset, regiao==x)$espvi00)
    N_h = count(subset(dataset, regiao==x))
    tamanho = append(tamanho, ceiling(1 / ((D/s2_h) + (1/N_h))) )
  }
  return(tamanho)
```

```
}

#tamanho da amostra final, usando idh ja que tam_amostra_idh >
#tam_amostra_espvi
n = 0
for(x in tam_amostra_idh()) {n = n + x}

#tamanho do estrato
tam_estrato = function(){
  tam = vector()
  for (x in regioes){
    N_h = count(subset(dataset, regiao==x))
    tam = append(tam, round(n * (N_h/N), digits=0))
  }
  return (tam)
}

#amostra por estrato
amostra_norte = subset(dataset2, regiao=="Norte")
                           %>% sample_n(tam_estrato()[[1]])

#retirei 1 elemento por conta de arredondamento, o valor
#exato era 219,56..., se não retirar a amostra tem tamanho n = 679

amostra_nordeste = subset(dataset2, regiao=="Norde")
                           %>% sample_n(tam_estrato()[[2]] - 1)

amostra_co = subset(dataset2, regiao=="Centr")
                           %>% sample_n(tam_estrato()[[3]])

amostra_sudes = subset(dataset2, regiao=="Sudes")
                           %>% sample_n(tam_estrato()[[4]])

amostra_sul = subset(dataset2, regiao=="Sul")
                           %>% sample_n(tam_estrato()[[5]])

#amostra com todas as regioes
amostra_final = rbind(amostra_norte, amostra_nordeste,
                      amostra_co, amostra_sudes, amostra_sul)

#criando os estimadores
```

```
#estimador media idh
est_media_idh = function(){
  return(amostra_final$idh10 %>% mean())
}

#estimador media espvi
est_media_espvi = function(){
  return(amostra_final$espvi10 %>% mean())
}

#est variancia idh
est_var_idh = function(){
  s2 = 0
  for (x in regioes){
    n_h = count(subset(amostra_final, regiao == x))
    N_h = count(subset(dataset, regiao==x))
    W_h = N_h / N
    f_h = n_h/N_h
    s2_h = var(subset(amostra_final, regiao==x)$idh10)
    s2 = s2 + (W_h**2 * (1-f_h) * s2_h/n_h)
    }
  return(s2)
}

#est variancia espvi
est_var_espvi = function(){
  s2 = 0
  for (x in regioes){
    n_h = count(subset(amostra_final, regiao == x))
    N_h = count(subset(dataset, regiao==x))
    W_h = N_h / N
    f_h = n_h/N_h
    s2_h = var(subset(amostra_final, regiao==x)$espvi10)
    s2 = s2 + (W_h**2 * (1-f_h) * s2_h/n_h)
  }
  return(s2)
}

#intervalo de confiança idh
int_conf_idh = function(){
  inf = est_media_idh() - 1.96 * sqrt(est_var_idh())
  sup = est_media_idh() + 1.96 * sqrt(est_var_idh())
```

```
  return(c(inf, sup))
  }

#intervalo de confiança espvi
int_conf_espvi = function(){
  inf = est_media_espvi() - 1.96 * sqrt(est_var_espvi())
  sup = est_media_espvi() + 1.96 * sqrt(est_var_espvi())
  return(c(inf, sup))
}

#valor populacional idh10
dataset2$idh10 %>% mean()

#valor populacional espvi10
dataset2$espvi10 %>% mean()
```

Referências

Bolfarine, Heleno e Wilton de Oliveira Bussab (2005). *Elementos de amostragem*. Editora Blucher.