Trabalho final de amostragem

João Gabriel Teixeira Ramos Moraes

Amostra Aleatória simppes

Dados: amostra aleatória simples sem reposição, encontrada na planilha "AmostraSorteada".

1. Estimar a média populacional da renda e calcula a variância dessa estimativa.
2. Estimar o tamanho médio das famílias e calcule a variância dessa estimativa.
3. Estimar o total de pessoas que moram na região e apresente a variância dessa estimativa.
4. Estimar a proporção de fumantes e apresente a variância dessa estimativa. Construir um intervalo de 98% confiança para a proporção populacional.

```
paste("media pop renda",mean(df_pop$renda))
```

```
## [1] "media pop renda 4.55341792877609"
```

```
paste("media pop tamanho",mean(df_pop$tam))
```

```
## [1] "media pop tamanho 2.74310274252968"
```

```
paste("tamanho pop", sum(df_pop$tam))
```

```
## [1] "tamanho pop 33507"
```

```
paste("proporção", sum(df_pop$fumantes/sum(df_pop$tam)))
```

```
## [1] "proporção 0.685021040379622"
```

1) Estimar a média populacional da renda e calcula a variância dessa estimativa.

```
n <- nrow(df_aas) ##tamanho da amostra
N <- nrow(df_pop) ##tamanho da população
i <- df_aas$renda ## vetor contendo renda da amostra

media_estimada <- mean(i) ## média estimada

um_por_n_menos_um_por_N <- (  ( 1/n )-( 1/N ) )
variancia_estimada<- um_por_n_menos_um_por_N*var(i) ## variância do estimador
```

Resposta item 1: O estimador utilizado para estimar o parâmetro "média populacional" $= \bar{Y} = \frac{1}{N}\sum_{i\in U} y_i$ foi o estimador "média amostral" $= \bar{y} = \frac{1}{n}\sum_{i\in s} y_i$. Este estimador foi escolhido pois é não víciado para uma amostragem aleatória simples sem reposição.

A estimativa da variância do estimador é dada por $\hat{V}_{\text{AASs}}(\bar{y}) = (\frac{1}{n} - \frac{1}{N})\hat{S}_y^2$, onde $\hat{S}_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2$

A estimativa da média populacional da renda é de 4.465 salários mínimos por família, com variância de 0.0168205.

2) Estimar o tamanho médio das famílias e calcule a variância dessa estimativa.

```
n <- nrow(df_aas) ##tamanho da amostra
N <- nrow(df_pop) ##tamanho da população
i <- df_aas$tam ## vetor com o tamanho das famílias da amostra

media_estimada <- mean(i) ## tamanho médio das famílias estimado

um_por_n_menos_um_por_N <- (  ( 1/n )-( 1/N ) )
variancia_estimada <- um_por_n_menos_um_por_N*var(i) ## variância do estimador
```

Resposta item 2: O estimador utilizado para estimar o parâmetro "média populacional" $= \bar{Y} = \frac{1}{N}\sum_{i\in U} y_i$ foi o estimador "média amostral" $= \bar{y} = \frac{1}{n}\sum_{i\in s} y_i$. Este estimador foi escolhido pois é não víciado para uma amostragem aleatória simples sem reposição.

A estimativa da variância do estimador é dada por $\hat{V}_{\text{AASs}}(\bar{y}) = (\frac{1}{n} - \frac{1}{N})\hat{S}_y^2$, onde $\hat{S}_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2$

A estimativa do tamanho médio das famílias é de 2.836 pessoas por família, com variância 0.0022258.

3) Estimar o total de pessoas que moram na região e apresente a variância dessa estimativa.

```
n <- nrow(df_aas) ##tamanho da amostra
N <- nrow(df_pop) ##tamanho da população
i <- df_aas$tam ## vetor com o tamanho das famílias da amostra

total_estimado <- mean(i)*N ## total estimado

um_por_n_menos_um_por_N <- (1/n-(1/N))
variancia_estimada <- N*N*(1-(n/N))*var(i)/n
```

Resposta do item 3: O estimador utilizado para estimar o total populacional $= Y = \sum_{i\in U} y_i$ foi o $\hat{Y} = N\bar{y}$, pois ele é um estimador não víciado para uma amostra aleatória simples sem reposição.

A estimativa da variãncia do estimador é dada por $\hat{V}_{\text{AASs}}(\hat{Y}_{\text{ASSs}}) = \frac{N^2}{(1-n/N)}S^2$

A estimativa do total de pessoas que moram na região foi de 3.4647317×10^4, com variância 3.3210908×10^5.

4) Estimar a proporção de fumantes e apresente a variância dessa estimativa. Construir um intervalo de 98% confiança para a proporção populacional.

```r
i <- df_aas$fumantes
j <- df_pop$fumantes
k <- df_aas$tam
l <- df_pop$tam

proporcao_estimada <- sum(i)/sum(k) ## proporção estimada

um_por_n_menos_um_por_N <- (  (1/n) - (1/N) )
variancia_estimada <- um_por_n_menos_um_por_N *n* proporcao_estimada * (1-proporcao_estimada) / (n-1) #

limite_inferior <- proporcao_estimada - qnorm(0.99)*sqrt(variancia_estimada)
limite_superior <- proporcao_estimada + qnorm(0.99)*sqrt(variancia_estimada)
```

Resposta do tópico 4: O estimador utilizado para estimar a proporção populacional $= p = \frac{N_A}{N}$(onde N_A representa o número de unidades populacionais que possuem o atributo de interesse, nesse caso, o número de pessoas que fumam) foi o estimador proporção amostral $= \bar{p} = \frac{n_A}{n}$,(onde n_A representa o número de unidades na amostra selecionada que possuem o atributo de interesse, nesse caso, o número de pessoas que fumam), pois ele é não víciado para uma amostra aleatória simples sem reposição.

A estimativa da variância do estimador é $\hat{V}_{\text{AASs}}(\hat{p}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n\hat{p}(1-\hat{p})}{n-1}$

A estimativa para a proporção de fumantes foi de 0.679%, com variância de 3.5390714×10^{-4}.

Um intervalo de confiança foi construído usando a distribuição assintótica do estimador $\bar{p}$, estratégia possível porque o tamanho da amostra é suficientemente grandeonde. O valor $Z_{\frac{\alpha}{2}} = -2.33$, retirado da distribuição normal padrão, é utilizado para construir o limite inferior, enquanto o valor $Z_{\frac{\alpha}{2}} = 2.33$, retirado da distribuição normal padrão, é utilizado para construir o limite superior.

Dessa forma, o intervalo de confiança obtido, com 98% de confiança, é de (0.635515046518655 , 0.723043512039904)

Amostra aleatória por conglomerado

Dados: dados populacionais, encontrados na planilha "Populacao-aTrabalhar"

```
## dados populacionais
df_pop <- read_excel("C:/Users/joaog/Downloads/Cópia de Trabalho_tipo2-BaseCriada.xlsx", sheet = "Popula
N <- nrow(df_pop) ##tamanho da população
```

1. Descrever um plano amostral via AC de um estágio e que das 7 localidades presentes, sejam sorteados 3 conglomerados da planilha Populacao-aTrabalhar. Depois, informar quais foram os 3 conglomerados sorteados, e o método de sorteio
2. Estimar a média populacional da renda e calcula a variância dessa estimativa.
3. Estimar o tamanho médio das famílias e calcule a variância dessa estimativa.
4. Estimar o total de pessoas que moram na região e apresente a variância dessa estimativa.
5. Estimar a proporção de fumantes e apresente a variância dessa estimativa.
6. Construir um intervalo de 98% confiança para a proporção populacional.

1) Descrever um plano amostral via AC de um estágio e que das 7 localidades presentes, sejam sorteados 3 conglomerados da planilha Populacao-aTrabalhar.

Resposta item 1:

1. Primeiramente, a base de dados será dividída em relação as localidades, tendo assim sete clusters(grupos) diferentes
2. Por amostragem simples sem reposição, serão selecionados 3 clusters

```
set.seed(1345678)
conglomerados <- sample(c("LocalidadeA","LocalidadeB","localidadeC","LocalidadeD","LocalidadeE","localid
```

Assim, temos, por amostragem aleatória simples, os 3 clusters selecionados, estes sendo LocalidadeA, LocalidadeE, LocalidadeD. A base de dados a partir da amostra se encontra abaixo.

```
df_ac <- df_pop %>% filter(local %in% conglomerados) ## base de dados

## lista com novos data frames para cada cluster
df_conglomerados <- list()
df_conglomerados[[1]] <- df_ac_1 <-  df_pop %>% filter(local==conglomerados[1])
df_conglomerados[[2]] <- df_ac_2 <-  df_pop %>% filter(local==conglomerados[2])
df_conglomerados[[3]] <- df_ac_3 <-  df_pop %>% filter(local==conglomerados[3])
```

2) Estimar a média populacional da renda e calcula a variância dessa estimativa.

```
m <- 3 ## número de conglometados na amostra
M <- 7 ## número de conglomerados na população
n <- nrow(df_ac) ##tamanho da amostra
N <- nrow(df_pop) ## tamanho da população
```

```
## variância do estimador de Horvitz-Thompson da média por unidade
termo <- 0
soma_t <- 0
for(i in df_conglomerados){invisible({
  termo <- (sum(i$renda) - (sum(df_ac$renda)/m))^2
  soma_t <- soma_t+termo
 })
}
var_1 <- (M*M)*((1/m)-(1/M))*soma_t/(N*N*(m-1))
## variância do estimador tipo razão da média por unidade
termo <- 0
soma_t <- 0
for (i in df_conglomerados) {
  invisible({
  termo <- (nrow(i)^2) * (mean(i$renda) - mean(df_ac$renda))^2
  soma_t <- soma_t + termo
  })
}

var_2 <- (m/n)^2 * ((1/m)-(1/M))*soma_t * (1/(m-1))

## estimador tipo razão da média
media_estimada <- mean(df_ac$renda)
variancia_real <- var_2
```

Resposta item 2: O estimador utilizado para estimar o parâmetro média populacional $= \bar{Y} = \frac{1}{N}\sum_{i \in U} y_i$ foi o estimador tipo razão para a média $= \bar{y} = \frac{1}{n}\sum_{i \in a} Y_i$.

Este estimador foi escolhido pois apresenta variância ligeiramente menor do que o esitimador de Horvitz-Thompson da média por unidade (6.2556137×10^{-4}<0.0242134).

A estimativa da média populacional da renda é de 4.551 salários mínimos por família, com variância de 6.2556137×10^{-4}.

3) Estimar o tamanho médio das famílias e calcule a variância dessa estimativa.

```
m <- 3 ## número de conglometados na amostra
M <- 7 ## número de conglomerados na população
n <- nrow(df_ac) ##tamanho da amostra
N <- nrow(df_pop) ## tamanho da população

## variância do estimador de Horvitz-Thompson da média por unidade
termo <- 0
soma_t <- 0
for(i in df_conglomerados){invisible({
  termo <- (sum(i$tam) - (sum(df_ac$tam)/m))^2
  soma_t <- soma_t+termo
 })
}
var_1 <- M*M*(1/(N*N))*((1/m)-(1/M))*soma_t
## variância do estimador tipo razão da média por unidade
termo <- 0
```

```
soma_t <- 0
for (i in df_conglomerados) {
  invisible({
  termo <- (nrow(i)^2) * (mean(i$tam) - mean(df_ac$tam))^2
  soma_t <- soma_t + termo
  })
}
var_2 <- (m/n)^2 * ((1/m)-(1/M)) * (1/(m-1))*soma_t
## estimador tipo razão da média
media_estimada <- mean(df_ac$tam)
variancia_estimada <- var_2
```

Resposta item 3: O estimador utilizado para estimar o parâmetro média populacional $= \bar{Y} = \frac{1}{N}\sum_{i\in U} y_i$ foi o estimador tipo razão para a média $= \bar{y} = \frac{1}{n}\sum_{i\in a} Y_i$.

Este estimador foi escolhido pois apresenta variância ligeiramente menor do que o esitimador de Horvitz-Thompson da média por unidade (4.2635271×10^{-4}<0.0185386).

A estimativa do tamanho médio das famílias é de 2.733 salários mínimos por família, com variância de 6.2556137×10^{-4}.

4) Estimar o total de pessoas que moram na região e apresente a variância dessa estimativa.

```
m <- 3 ## número de conglometados na amostra
M <- 7 ## número de conglomerados na população
n <- nrow(df_ac) ##tamanho da amostra
N <- nrow(df_pop) ## tamanho da população

f <- m/M
## variância do estimador natural (de Horvitz-Thompson
termo <- 0
soma_t <- 0
for(i in df_conglomerados){invisible({
  termo <- (sum(i$tam) - (sum(df_ac$tam)/m))^2
  soma_t <- soma_t+termo
 })
}

var_1 <- M*M *((1-f)/m )*(soma_t)/(m-1)

##estimador tipo razão, variância
termo <- 0
soma_t <- 0
for (i in df_conglomerados) {
  invisible({
  termo <- (nrow(i)^2) * (mean(i$tam) - mean(df_ac$tam))^2
  soma_t <- soma_t + termo
  })
}

var_2 <- M*M*((1/m)-(1/M))*(1/(m-1))*soma_t
## estimando pelo estimador tipo razão
```

```r
total_estimado <- N*mean(df_ac$tam)
variancia_real <- var_2
```

resposta item 4:O estimador utilizado para estimar o total populacional $= Y = \sum_{i \in U} y_i$ foi o estimador tipo tazão $= N\bar{y}$, pois ele é um estimador não víciado para uma amostra aleatória simples sem reposição.

A estimativa do total de pessoas que moram na região foi de 3.337798×10^4, e a estimativa da variância do estimador foi de 6.3056725×10^4.

5) Estimar a proporção de fumantes e apresente a variância dessa estimativa.

```r
m <- 3 ## número de conglometados na amostra
M <- 7 ## número de conglomerados na população
n <- nrow(df_ac) ##tamanho da amostra
N <- nrow(df_pop) ## tamanho da população

a <- m
Pc2 <- sum(df_ac$fumantes)/sum(df_ac$tam)


termo <- 0
soma_t <- 0
for (i in df_conglomerados){invisible({

  termo <- (sum(i$fumantes)-Pc2*sum(i$tam))^2
  soma_t <- soma_t+termo

  })
}
variancia_estimada <- soma_t/(a*(a-1)*((sum(df_ac$tam)/a)^2))
```

Resposta do tópico 5: O estimador utilizado para estimar a proporção populacional $= p = \frac{N_A}{N}$(onde N_A representa o número de unidades populacionais que possuem o atributo de interesse, nesse caso, o número de pessoas que fumam) foi o estimador $p_{c2} = \sum_{\alpha \in A} \frac{T_\alpha}{b_\alpha}$.

A estimativa para a proporção de fumantes foi de 0.679%, com variância de 1.8130487×10^{-5}.

6) Construir um intervalo de 98% confiança para a proporção populacional.

```r
limite_inferior <- proporcao_estimada - qnorm(0.99)*sqrt(variancia_estimada)
limite_superior <- proporcao_estimada + qnorm(0.99)*sqrt(variancia_estimada)
```

Resposta do tópico 6: Um intervalo de confiança foi construído usando a distribuição assintótica do estimador $\bar{p}$, estratégia possível porque o tamanho da amostra é suficientemente grandeonde. O valor $Z_{\frac{\alpha}{2}} = -2.33$, retirado da distribuição normal padrão, é utilizado para construir o limite inferior, enquanto o valor $Z_{\frac{\alpha}{2}} = 2.33$, retirado da distribuição normal padrão, é utilizado para construir o limite superior.

Dessa forma, o intervalo de confiança obtido, com 98% de confiança, é de (0.669373711028058 , 0.689184847530501)

Amostra estratificada

Se um plano amostral via amostragem estratificada fosse utilizada, considerando os dados populacionais, qual(ais) variável(is) seria(am) para estimar a renda?

Resposta: Ao optarmos por um plano amostral via Amostragem Estratificada (AE) para estimar a renda, a escolha das variáveis estratificadoras desempenha um papel crucial na eficiência da estimativa. A AE é particularmente útil quando a variância da população é elevada e a obtenção de uma amostra grande o suficiente para reduzir a variância do estimador a níveis satisfatórios é impraticável.

Considerando os dados populacionais, uma estratificação eficaz seria baseada na variável de ensino. Ao estratificar a população por níveis de escolaridade, podemos observar que indivíduos com graus semelhantes de educação tendem a apresentar rendas mais próximas entre si. Portanto, os estratos formados por essa variável específica têm variâncias internas inferiores à variância global da população.

Essa abordagem estratificada proporciona uma estimativa mais eficiente da renda média, uma vez que a homogeneidade dentro de cada estrato resulta em uma diminuição significativa na variância total do estimador. Dessa forma, a escolha da variável de ensino como critério estratificador é justificada pela sua capacidade de reduzir as variâncias dos estratos, tornando a estimação mais precisa e confiável.

Conclusão

Para estimar a renda média, a abordagem mais indicada seria a Amostragem Estratificada (AE), devido à sua eficácia na redução da variância. Já para o tamanho médio, e proporção de fumantes, a Amostragem por Conglomerados (AC) seria a escolha preferencial, pois apresenta menor variância. Essas metodologias são selecionadas com base na busca por estimativas mais precisas e eficientes, alinhadas às características específicas das variáveis. A combinação de AE e AASs se aproxima mais dos dados populacionais, proporcionando estimativas representativas.