lista01_resolucao

Heitor Carvalho Pinheiro - 11833351

2023-04-05

Lista 01 - SME0807 Elementos de Amostragem

Exercícios Capítulo 01

1.1.1

Imagine que queremos amostrar usuários do Spotify para melhor entender o perfil musical dos clientes.

a. A unidade de pesquisa é cada usuário do aplicativo. Será esse o objeto a ser observado.
b. A população corresponde a todos os usuários do aplicativo.
c. O instrumento de coleta de dados seria através de técnicas de *web scraping* e requisições HTTPs, através das quais as informações de uso do aplicativo seriam coletadas e posteriormente armazenadas em um banco de dados.
d. A unidade respondente seria o portador da conta.
e. O sistema de referência, muito provavelmente, seria o banco de dados da empresa.
f. A unidade amostral mais provável, nesse caso, é a própria unidade de pesquisa.
g. As unidades amostrais alternativas poderiam variar de acordo com o plano amostral escolhido. Poderíamos estratificar os usuários por Idade ou Sexo, ou ainda criar conglomerados por regiões ou países.

Como fixar o tamanho da amostra?

Ex 1.1.4

- População alvo: todos os alunos de pós-graduação
- Objetivo: estimar a proporção dos favoráveis à mudança. Espera-se que seja da ordem de 5%

Ex 1.1.10

Objetivo: Estimar o consumo médio de água por domicílio em uma cidade.

a. Unidade Domiciliar como UPA (Unidade Primária de Amostragem)

- Vantagens: maior exatidão e granularidade nos resultados.
- Desvantagens: será necessário uma grande amostra para se obter resultados representativos, alto custo.

b. Blocos de domicílios

- Vantagens: possibilidade de amostrar um número maior de domicílios. Diminuição de custos. Maior variedade amostral.

- Desvantagens: será mecessário que se garanta que os blocos sejam homogêneos e representativos da população. Será necessário o sorteio de uma Unidade Secundária de Amostragem (USA) o que além de tornar o plano mais complexo, também aumentará os custos da pesquisa.

c. Quarteirões

- Vantagens: possibilidade de mais provável de ser capaz de mapear, de modo representativo, os domicílios da cidade. Diminuição do tamanho amostral comparado aos outros. Maior capacidade de estratificação da amostra.
- Desvantagens: necessidade de múltiplos estágios para a seleção da unidade elementar, aumentando a complexidade e, inevitalvelmente, o custo. Maior custo no que tange ao treinamento de pessoal para a realização de uma pesquisa tão ampla.

Ex 1.1.13

a. A entrevista pessoal, nesse caso, é inviável na prática. A entrevista por telefone, seria viável porém custosa, especialmente de tempo. Correios nem existem mais como forma de marketing. Por fim, a Internet, sem dúvida, seria a melhor escolha, dado que seria possível utilizar métodos como o CASI (Computer Assisted Self-Interviewing) em que os expectadores do programa, mediante algum benefício poderiam responder a alguma pesquisa através de um formulário.

b. De novo, a entrevista pessoal é sempre mais custosa, do mesmo modo que o telefone. Correios, poderiam ser uma opção caso a Internet não existisse e esse editor poderia mandar cartas pedindo a resposta dos entrevistados. A internet, novamente, é a melhor opção e, dado que o editor conhece o seu público seria possível elaborar perguntas bem específicas para corroborar sua crença e levantar uma amostra representativa.

c. Nesse caso, a entrevista pessoal ou o telefonema poderia ser aplicados, porém a unidade amostral seriam postos de vacinação de animais dentro dos limites geográficos estabelecidos. Seria uma alternativa também, buscar informações na internet sobre dados históricos, se disponíveis.

Ex 2.

Os méritos são semelhantes aos do exercício 1.1.10. No que tange ao sistema de referência em cada caso, para as famílias individuais o uso do Censo de Amostras de Família do IBGE, me parece uma boa escolha. Para as unidades habitacionais o censo do IBGE por domicílios, também me parece uma escolha razoável para o SR e, por fim, para os quarteirões como unidade de amostragem, um mapa com a relação de todos os quarteirões da cidade, ou um banco de dados com essas informações também seria a melhor escolha de SR, na minha opinião.

Ex 4.

Exercícios Capítulo 02

Ex 2.1

Simulando uma população $X \sim \mathcal{N}(50, 16)$

Total τ *Média populacional* μ *Variância Populacional* S^2

```
#simulando a populacao
set.seed(42)
X <- rnorm(100, 50, 4)
```

```
total <- length(X)
media <- mean(X)
variancia <- var(X)
```

```
data.02 <- tibble(total = total,
                  media = media,
                      var = variancia)

knitr::kable(data.02)
```

total	media	var
100	50.13006	17.35079

Ex 2.2

Lembrando que na table a seguir:

X corresponde ao numero de apartamentos no condominio Y corresponde ao numero de apartamentos alugados

```
#lendo os dados

df <- read.csv("../../dados/CAP_02_tab_2_8.csv")

head(df)
```

```
##   indice  Y   X
## 1      1 19  23
## 2      2 17  18
## 3      3 25  33
## 4      4 84  89
## 5      5 91 114
## 6      6 48  66
```

```
Y.zeros <- sum(df$Y == 0)
X.zeros <- sum(df$X == 0)

Y.zeros
```

```
## [1] 19
```

```
X.zeros
```

```
## [1] 0
```

Existem 19 condominios com valores nulos

Estimando os seguintes valores

a. μ_Y, τ_Y, S_Y

```
media.Y <- mean(df$Y)
total.Y <- sum(df$Y)
var.Y <- var(df$Y)

statistics.Y <- tibble(media = media.Y,
                       total = total.Y,
                       variancia = var.Y)

knitr::kable(statistics.Y)
```

media	total	variancia
18.6	3348	409.4369

b. μ_X, τ_X, S_X

```
media.X <- mean(df$X)
total.X <- sum(df$X)
var.X <- var(df$X)

statistics.X <- tibble(media = media.X,
                       total = total.X,
                       variancia = var.X)

knitr::kable(statistics.X)
```

media	total	variancia
27.37778	4928	609.4096

```
statistics.Final <- bind_rows(statistics.X, statistics.Y)
rownames(statistics.Final) <- c("X","Y")

head(statistics.Final)
```

```
## # A tibble: 2 x 3
##   media total variancia
##   <dbl> <int>     <dbl>
## 1  27.4  4928      609.
## 2  18.6  3348      409.
```

```
row.names(statistics.Final)
```

```
## [1] "X" "Y"
```

c. A proporção P de condomínios com mais de 20 apartamentos alugados

```
proportion.P <- sum(df$Y > 20) / length(df$X)
proportion.P
```

```
## [1] 0.3166667
```

Aproximadamente, 32% dos condomínios possuem mais de 20 apartamentos alugados.

Criando a coluna binário W_i

```
df$W <- ifelse(df$Y > 20, 1, 0)
head(df)
```

```
##   indice  Y   X W
## 1      1 19  23 0
## 2      2 17  18 0
## 3      3 25  33 1
## 4      4 84  89 1
## 5      5 91 114 1
## 6      6 48  66 1
```

Variância de W

```
var(df$W)
```

```
## [1] 0.2175978
```

Portanto, a variância da variável W é cerca de 0.217.

Exercício 2.6

```
#criando os dados

df2 <- tibble(
  N = seq(1,6),
  D = c(2,6,10,8,10,12)
)

df2
```

```
## # A tibble: 6 x 2
##       N     D
##   <int> <dbl>
## 1     1     2
## 2     2     6
## 3     3    10
## 4     4     8
## 5     5    10
## 6     6    12
```

Realizando uma amostragem simples sem reposição, de acordo com o plano amostral A

Os elementos da população $\mathcal{U}$ é o conjunto $\{2, 6, 8, 10, 12\}$

A frequência dos elementos da amostra

```
df2$F <- as.numeric(table(df2$D))[match(df2$D, sort(unique(df2$D)))]
```

```
df2
```

```
## # A tibble: 6 x 3
##       N     D     F
##   <int> <dbl> <dbl>
## 1     1     2     1
## 2     2     6     1
## 3     3    10     2
## 4     4     8     1
## 5     5    10     2
## 6     6    12     1
```

a. Calcule $Var_A(f_i)$ e $Cov_A(f_i, f_j)$ para algum i e j que você escolher.

Bom, aqui temos que calcular a basicamente a Variância e a Covariância com que cada elemento i ou i, j aparece na amostra acima.

Lembrando que possuímos uma AAS_s.

Vamos considerar f_2. Ou seja, estamos considerando a probabilidade de o valor 2 pertencer à amostra. Para isso, precisamos fazer a distribuição de f_2

O total de amostras possíveis em um plano AAS_s, no nosso caso, é um arranjo de 6 elementos tomados 2 a 2.

Logo $A_{5,2} = 6!/4! = 30$

Temos um plano amostral em que todos as amostras tem igual probabilidade de serem selecionadas, logo, $p(s) = 1/30$.

Logo a distribuição amostral de f_2 é a seguinte:

```
dist_f2 <- tibble(
  f_2 = c(0,1),
  p_h = c(22/30, 8/30)
)

knitr::kable(dist_f2)
```

f_2	p_h
0	0.7333333
1	0.2666667

Logo, o valor esperado $E(f_2) = 0.26$.

Assim, podemos calcular a variância de f_2, do seguinte modo:

$$Var(f_2) = (1 - 8/30)^2 \cdot 8/30 + (0 - 8/30)^2 \cdot 22/30 \approx 0.316$$

E a $Cov(f_2, f_6)$ seria:

$$Cov(f_2, f_6) = \sum (f_2 - E(f_2)) \cdot (f_6 - E(f_6)) \cdot P(s) = (1-8/30)\cdot(1-8/30)\cdot 8/30 + (-8/30)\cdot(-8/30)\cdot(22/30) \approx 0.195$$

Dúvida

Por fim, teoricamente o plano deveria ser simétrico. Porém, como f_{10} pode assumir os valores $0, 1$ e 2. O $E(f_{10})$ seria diferente, o que quebraria a hipotése de que um plano AAS_S deveria ser simétrico.

b. Encontre a distribuição de $t(s)$

Exercício 2.10