Trabalho de Amostragem

Prática da amostragem aleatória simples, sistemática e estratificada em banco de dados

Paulo Ricardo Seganfredo Campana

20 de março de 2023

Objetivos

Desejamos obter valores para os estimadores da média e total, além da variância dos estimadores e intervalo de confiança sobre as variáveis Tempo de internação e uma das 8 doenças do banco de dados. Tal análise será aplicada em amostras de tamanho 30 para amostragem aleatória simples, sistemática e por fim, estratificada por unidade do PSF.

Leitura e organização do banco de dados

Fiquei responsável pela variável Alergia, que será gerada aleatóriamente de uma distribuição Bernoulli com $p = 0.3$, segundo a ASBAI (Associação Brasileira de Alergia e Imunologia).

```
library(dplyr)
set.seed(8100)

TRAB07 <- readxl::read_excel("TRAB07.xlsx")
Banco <- TRAB07 |>
    select(Paciente, Unidade, `Tempo de internação` = Tempoint, Alergias) |>
    mutate(
        Unidade = case_when(
            Unidade == 1 ~ "Mangabeira",
            Unidade == 2 ~ "Bancários",
            Unidade == 3 ~ "Valentina",
        ),
        Alergias = rbinom(n(), size = 1, prob = 0.3) |> as.logical()
    )
```

```r
head(Banco, n = 30)
```

Tabela 1: Primeiras 30 observações do banco de dados após organização

Paciente	Unidade	Tempo de internação	Alergias
1	Bancários	29	TRUE
2	Mangabeira	20	FALSE
3	Valentina	16	TRUE
4	Mangabeira	28	TRUE
5	Mangabeira	14	FALSE
6	Valentina	5	TRUE
7	Bancários	14	TRUE
8	Bancários	15	FALSE
9	Mangabeira	26	FALSE
10	Bancários	21	FALSE
11	Mangabeira	12	FALSE
12	Valentina	13	FALSE
13	Bancários	12	FALSE
14	Bancários	11	FALSE
15	Mangabeira	20	FALSE
16	Mangabeira	6	TRUE
17	Bancários	17	TRUE
18	Mangabeira	13	TRUE
19	Valentina	8	FALSE
20	Bancários	8	FALSE
21	Bancários	18	TRUE
22	Mangabeira	10	FALSE
23	Mangabeira	22	FALSE
24	Bancários	1	TRUE
25	Valentina	22	FALSE
26	Bancários	1	FALSE
27	Bancários	9	FALSE
28	Bancários	24	FALSE
29	Bancários	16	FALSE
30	Valentina	27	FALSE

Implementação dos estimadores, variância e IC

Faremos uso dos seguintes estimadores não viesados para média, total e proporção, suas variâncias, e o intervalo de confiança da distribuição t com nível de significância $\alpha = 0.1$:

$$\bar{y} = \frac{1}{n}\sum_{i\in S} y_i \qquad \hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

$$\hat{t} = N\frac{1}{n}\sum_{i\in S} y_i \qquad \hat{V}(\hat{t}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}N^2 \qquad \mathrm{IC}(\bar{y}) = \left[\bar{y} \mp t_{\alpha/2,\, n-1}\sqrt{\hat{V}(\bar{y})}\right]$$

$$\hat{p} = \frac{1}{n}\sum_{i\in S} p_i \qquad \hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n-1}$$

```
Estimador <- function(Amostra, N, tipo) {
    case_when(
        tipo == "média"      ~ mean(Amostra),
        tipo == "total"      ~ mean(Amostra) * N,
        tipo == "proporção" ~ mean(Amostra),
    )
}

VariânciaEstimador <- function(Amostra, N, tipo) {
    n <- length(Amostra)
    s2 <- var(Amostra)
    p <- Estimador(Amostra, N, tipo)
    case_when(
        tipo == "média"      ~ (1 - n/N) * s2 / n,
        tipo == "total"      ~ (1 - n/N) * s2 / n * N^2,
        tipo == "proporção" ~ (1 - n/N) * p * (1-p) / (n-1),
    )
}

IntervaloConfiança <- function(Amostra, N, tipo, alpha = 0.1) {
    n <- length(Amostra)
    t <- qt(1 - alpha/2, df = n-1)
    SE <- sqrt(VariânciaEstimador(Amostra, N, tipo))
    Estimador(Amostra, N, tipo) + c(-t * SE, t * SE)
}
```

Pipeline de análise

Para cada amostra retirada da população, aplicarei a função `Análise`, que utiliza os estimadores implementados acima para descrever a amostra.

```
Análise <- function(Amostra, N) {
    Amostra |>
        reframe(
            Variável = rep(c("Alergias", "Tempo de internação"), each = 2),
            Estimador = c("Proporção", "Total", "Média", "Total"),
            Estimativa = c(
                Estimador(Alergias, N, tipo = "proporção"),
                Estimador(Alergias, N, tipo = "total"),
                Estimador(`Tempo de internação`, N, tipo = "média"),
                Estimador(`Tempo de internação`, N, tipo = "total")
            ),
            `Variância do Estimador` = c(
                VariânciaEstimador(Alergias, N, tipo = "proporção"),
                VariânciaEstimador(Alergias, N, tipo = "total"),
                VariânciaEstimador(`Tempo de internação`, N, tipo = "média"),
                VariânciaEstimador(`Tempo de internação`, N, tipo = "total")
            ),
            `Intervalo de confiança` = list(
                IntervaloConfiança(Alergias, N, tipo = "proporção"),
                IntervaloConfiança(Alergias, N, tipo = "total"),
                IntervaloConfiança(`Tempo de internação`, N, tipo = "média"),
                IntervaloConfiança(`Tempo de internação`, N, tipo = "total")
            )
        ) |>
        mutate(across(
            Estimativa:`Intervalo de confiança`,
            function(x) lapply(x, function(x) round(x, digits = 3))
        )) |>
        mutate(across(
            Estimativa:`Variância do Estimador`,
            as.numeric
        ))
}
```

Análise para amostra aleatória simples

Na amostragem aleatória simples, cada membro da população possui mesma probabilidade de pertencer a amostra, é usado geração de números aleatórios para sortear n diferentes observações de um total de N.

```
AASimples <- slice_sample(Banco, n = 30)
ANSimples <- Análise(AASimples, N = nrow(Banco))
ANSimples
```

Tabela 2: Estimadores para amostra aleatória simples

Variável	Estimador	Estimativa	Variância do Estimador	Intervalo de confiança
Alergias	Proporção	0.233	0.005	0.110, 0.356
Alergias	Total	46.667	209.732	22.060, 71.274
Tempo de internação	Média	17.833	1.972	15.447, 20.219
Tempo de internação	Total	3566.667	78870.881	3089.485, 4043.849

Comparação com o banco de dados

Por usamos toda a população para fazer as estimativas, a variância dos estimadores é 0 pelo termo de correção $(1 - \frac{n}{N})$ portanto estas estimativas são pontuais.

```
Análise(Banco, N = nrow(Banco))
```

Tabela 3: Estimadores do banco de dados

Variável	Estimador	Estimativa	Variância do Estimador	Intervalo de confiança
Alergias	Proporção	0.250	0	0.25, 0.25
Alergias	Total	50.000	0	50, 50
Tempo de internação	Média	16.005	0	16.005, 16.005
Tempo de internação	Total	3201.000	0	3201, 3201

Análise para amostra aleatória sistemática

Para amostragem aleatória sistemática, criei uma função que calcula $k = \lfloor N/n \rfloor$ e R um número aleatório entre 1 e k, esses valores serão usados para obter os indíces da forma $R, R+k, R+2k, R+3k, \cdots$. Devido ao nosso tamanho da população 200 não ser divisível pelo tamanho desejado da amostra 30 e por arrendondar k para o inteiro anterior, essa amostragem sistemática resulta em uma amostra de tamanho 33 ou 34 dependendo do valor de R.

```
AmostragemSistemática <- function(População, n){
    N <- nrow(População)
    k <- floor(N / n)
    R <- sample(k, size = 1)
    amostra <- (1:N + R) %% k == 0
    População[amostra, ]
}
```

```
AASistemática <- AmostragemSistemática(Banco, 30)
ANSistemática <- Análise(AASistemática, N = nrow(Banco))
ANSistemática
```

Tabela 4: Estimadores para amostra aleatória sistemática

Variável	Estimador	Estimativa	Variância do Estimador	Intervalo de confiança
Alergias	Proporção	0.206	0.004	0.097, 0.314
Alergias	Total	41.176	164.486	19.472, 62.881
Tempo de internação	Média	16.618	1.537	14.519, 18.716
Tempo de internação	Total	3323.529	61488.927	2903.875, 3743.183

Análise para amostra aleatória estratificada por Unidade do PSF

Já para amostragem aleatória estratificada, usaremos como estratos as Unidades do PSF que tem como valores distintos os bairros Bancários, Mangabeira e Valentina, é utilizado alocação de Neyman para o tamanho de cada estrato na amostra, que leva em consideração o desvio padrão populacional de cada estrato como peso além do tamanho do estrato.

```
AmostragemEstratificada <- function(População, n, VarEstrato, VarDesvio) {
    VarEstrato <- pull(População, {{ VarEstrato }})
    VarDesvio <- pull(População, {{ VarDesvio }})
    Nh <- table(VarEstrato)
    Sh <- tapply(VarDesvio, VarEstrato, sd)
    tamanhos <- ceiling(
        n * Nh * Sh / sum(Nh * Sh)
    )
    unique(VarEstrato) |>
        lapply(function(x) {
            População |>
                filter({{ VarEstrato }} == x) |>
                slice_sample(n = tamanhos[x])
        })
}
```

```r
AAEstratificada <- AmostragemEstratificada(Banco, 30, Unidade, `Tempo de internação`)

AAEstratificada |>
    lapply(function(x) {
        Análise(x, N = nrow(Banco)) |>
            mutate(Unidade = unique(x$Unidade), .before = Variável)
    }) |>
    Reduce(f = rbind)
```

Tabela 5: Estimadores para amostra aleatória estratificada por unidade do PSF

Unidade	Variável	Estimador	Estimativa	Variância do Estimador	Intervalo de confiança
Bancários	Alergias	Proporção	0.429	0.018	0.194, 0.663
Bancários	Alergias	Total	85.714	700.785	38.834, 132.595
Bancários	Tempo de internação	Média	13.214	1.514	11.035, 15.394
Bancários	Tempo de internação	Total	2642.857	60574.097	2206.998, 3078.716
Mangabeira	Alergias	Proporção	0.250	0.026	-0.054, 0.554
Mangabeira	Alergias	Total	50.000	1028.571	-10.762, 110.762
Mangabeira	Tempo de internação	Média	19.750	7.329	14.621, 24.879
Mangabeira	Tempo de internação	Total	3950.000	293142.857	2924.225, 4975.775
Valentina	Alergias	Proporção	0.111	0.012	-0.091, 0.313
Valentina	Alergias	Total	22.222	471.605	-18.161, 62.605
Valentina	Tempo de internação	Média	16.111	6.299	11.444, 20.778
Valentina	Tempo de internação	Total	3222.222	251954.938	2288.820, 4155.624

Esta amostra estratificada possui 14 observações da unidade Bancários, 8 de Mangabeira e 9 de Valentina, devido ao pequeno tamanho da amostra, a variância dos estimadores é grande e o intervalo de confiança pode ficar negativo.

Análise para amostra aleatória estratificada

E finalmente, analisando a amostra estratificada como um todo.

```
AAEstratificada <- Reduce(x = AAEstratificada, f = rbind)
ANEstratificada <- Análise(AAEstratificada, N = nrow(Banco))
ANEstratificada
```

Tabela 6: Estimadores para amostra aleatória estratificada

Variável	Estimador	Estimativa	Variância do Estimador	Intervalo de confiança
Alergias	Proporção	0.290	0.006	0.161, 0.420
Alergias	Total	58.065	232.133	32.205, 83.924
Tempo de internação	Média	15.742	1.288	13.815, 17.668
Tempo de internação	Total	3148.387	51533.569	2763.092, 3533.682

Comparação entre amostragems

Vale notar a diferença entre os tipos de amostragem através da variância dos estimadores:

```
tibble(
    Variável = rep(c("Alergias", "Tempo de internação"), each = 2),
    Estimador = c("Proporção", "Total", "Média", "Total"),
          Simples = pull(      ANSimples, `Variância do Estimador`),
      Sistemática = pull(  ANSistemática, `Variância do Estimador`),
    Estratificada = pull(ANEstratificada, `Variância do Estimador`)
)
```

Tabela 7: Comparação das variâncias dos estimadores entre os tipos de amostragem

Variável	Estimador	Simples	Sistemática	Estratificada
Alergias	Proporção	0.005	0.004	0.006
Alergias	Total	209.732	164.486	232.133
Tempo de internação	Média	1.972	1.537	1.288
Tempo de internação	Total	78870.881	61488.927	51533.569