Teste de amostragem estratificada

Paulo R. S. Campana

19 de abril de 2023

```
library(tidyverse)
```

Questão 1.

Apresente os cálculos de como podem ser encontrados na amostra de cinco estratos a alocação para os estratos de forma proporcional e a alocação ótima de Neyman. Seus resultados podem divergir dos apresentados abaixo.

```
q1 <- tibble(
    stratum = c("BA", "EA", "PA", "RA", "XF"),
    Nh      = c(2371, 1442, 1710,  659, 1346),
    Sh      = c(42.4, 15.4, 40.7, 43.7, 54.2)
)
```

```
q1 |> mutate(
    nh_prop   = (40 * Nh      / sum(Nh     )) |> ceiling(),
    nh_neyman = (40 * Nh * Sh / sum(Nh * Sh)) |> ceiling()
)
```

stratum	Nh	Sh	nh_prop	nh_neyman
BA	2371	42.4	13	14
EA	1442	15.4	8	4
PA	1710	40.7	10	10
RA	659	43.7	4	4
XF	1346	54.2	8	10

Questão 2.

Considera a população com $N = 6$ domicílios listada com os respectivos valores de variáveis de interesse na Tabela

```
q2 <- tibble(
    domicílio     = c(  1,    2,    3,   4,   5,    6),
    renda         = c(800, 4200, 1600, 500, 900, 2000),
    moradores     = c(  2,    4,    2,   2,   4,    1),
    trabalhadores = c(  2,    3,    1,   1,   2,    1)
)
```

1. Para cada variável de interesse (Renda, Moradores, Trabalhadores), calcule os seguintes parâmetros populacionais: total, média e variância.

```
q2 |> summarise(across(
    c(renda, moradores, trabalhadores),
    list(total = sum, média = mean, variância = var)
)) |>
round(3) |>
pivot_longer(
    everything(),
    names_to = c("variável", "parâmetro"),
    names_sep = "_",
    values_to = "valor"
)
```

variável	parâmetro	valor
renda	total	10000.000
renda	média	1666.667
renda	variância	1846666.667
moradores	total	15.000
moradores	média	2.500
moradores	variância	1.500
trabalhadores	total	10.000
trabalhadores	média	1.667
trabalhadores	variância	0.667

2. Liste o conjunto S de todas as amostras possíveis de tamanho 2 da população, considerando apenas amostras de unidades distintas.

```
S <- sample(x = 1:6, size = 2) |>
    sort() |>
    replicate(n = 1000, simplify = FALSE) |>
    unique() |>
    as_tibble(.name_repair = "minimal")
```

2	1	1	1	1	4	1	3	2	5	2	3	3	4	2
6	5	4	3	6	6	2	5	5	6	4	4	6	5	3

3. Supondo que todas as amostras listadas no conjunto S são equiprováveis (plano A), calcule:

a. As probabilidades de inclusão das unidades.

```
SA <- S |>
    t() |>
    as_tibble() |>
    arrange(V1, V2) |>
    mutate(renda = map2_dbl(V1, V2,
        function(V1, V2) q2$renda[q2$domicílio == V1] + q2$renda[q2$domicílio == V2]
    ))
SA |> reframe(
    domicílio     = unique(c(V1, V2)),
    frequência    =  table(c(V1, V2)),
    probabilidade = frequência / sum(frequência)
)
```

domicílio	frequência	probabilidade
1	5	0.1666667
2	5	0.1666667
3	5	0.1666667
4	5	0.1666667
5	5	0.1666667
6	5	0.1666667

b. As probabilidades de inclusão dos pares de unidade.

```
SA |> mutate(probabilidade = 1 / n())
```

V1	V2	renda	probabilidade
1	2	5000	0.0666667
1	3	2400	0.0666667
1	4	1300	0.0666667
1	5	1700	0.0666667
1	6	2800	0.0666667
2	3	5800	0.0666667
2	4	4700	0.0666667
2	5	5100	0.0666667
2	6	6200	0.0666667
3	4	2100	0.0666667
3	5	2500	0.0666667
3	6	3600	0.0666667
4	5	1400	0.0666667
4	6	2500	0.0666667
5	6	2900	0.0666667

c. Os valores possíveis para o estimador Horvitz-Thompson do total populacional para a variável renda.

```
HTA <- SA |>
    mutate(
        probabilidade = 1 / n(),
        estimador = renda / probabilidade * (3 / 15)
    )
```

V1	V2	renda	probabilidade	estimador
1	2	5000	0.0666667	15000
1	3	2400	0.0666667	7200
1	4	1300	0.0666667	3900
1	5	1700	0.0666667	5100
1	6	2800	0.0666667	8400
2	3	5800	0.0666667	17400
2	4	4700	0.0666667	14100
2	5	5100	0.0666667	15300
2	6	6200	0.0666667	18600
3	4	2100	0.0666667	6300
3	5	2500	0.0666667	7500
3	6	3600	0.0666667	10800
4	5	1400	0.0666667	4200
4	6	2500	0.0666667	7500
5	6	2900	0.0666667	8700

d. O valor esperado e a variância para o estimador Horvitz-Thompson do total populacional para a variável renda.

```
HTA |> summarise(across(
    estimador,
    list(média = mean, variância = var)
))
```

estimador_média	estimador_variância
10000	23742857

4. Considere agora que o conjunto S é formado pelas amostras (1,2), (2,3), (2,4), (2,5), (2,6), tendo cada uma delas probabilidade 1/5 de ser a amostra selecionada (plano B), Repita os cálculos do item 3. para o novo plano amostral.

```
SB <- tibble(
        V1 = c(1,2,2,2,2),
        V2 = c(2,3,4,5,6)
    ) |>
    mutate(renda = map2_dbl(V1, V2,
        function(V1, V2) q2$renda[q2$domicílio == V1] + q2$renda[q2$domicílio == V2]
    ))
```

```
HTB <- SB |>
    mutate(
        probabilidade = 1 / n(),
        estimador = renda / probabilidade * (3 / 5)
    )
```

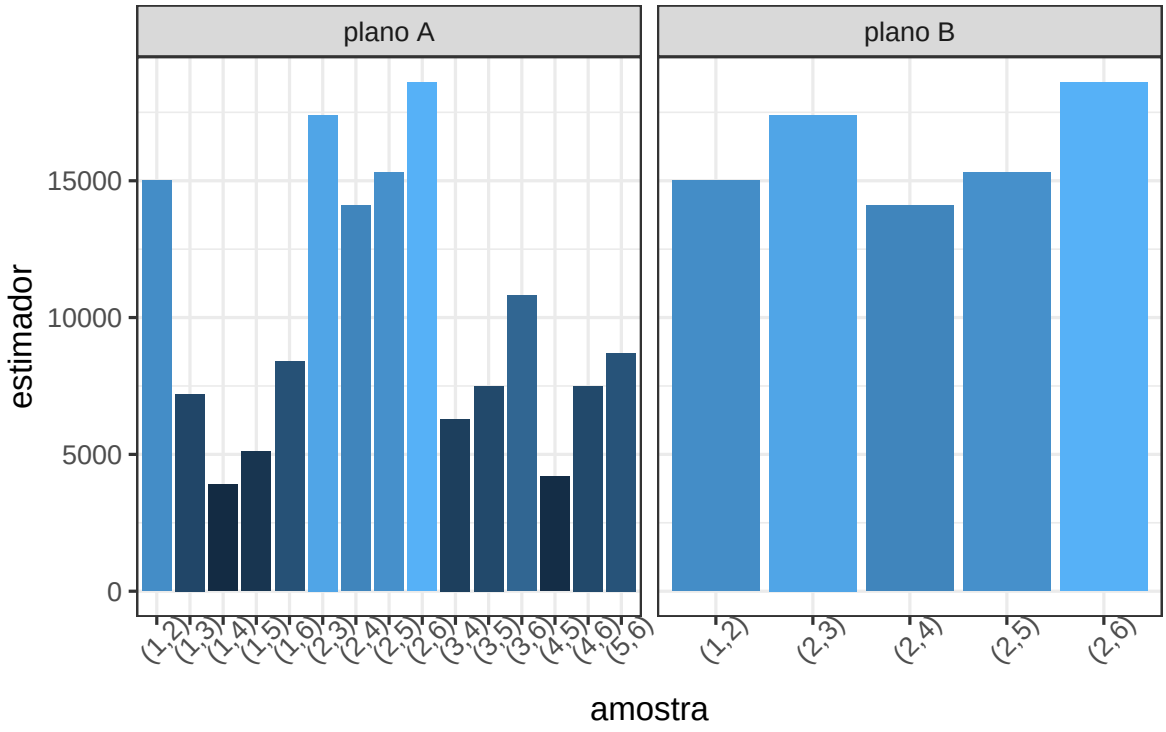
V1	V2	renda	probabilidade	estimador
1	2	5000	0.2	15000
2	3	5800	0.2	17400
2	4	4700	0.2	14100
2	5	5100	0.2	15300
2	6	6200	0.2	18600

```
HTB |> summarise(across(
    estimador,
    list(média = mean, variância = var)
))
```

estimador_média	estimador_variância
16080	3447000

5. Faça gráficos dos valores possíveis do estimador de total sob os dois planos amostrais para comparar as respectivas distribuições.

```
rbind(
    HTA |> mutate(plano = "plano A"),
    HTB |> mutate(plano = "plano B")
) |>
    mutate(amostra = paste0("(", V1, ",", V2, ")")) |>
    ggplot(aes(x = amostra, y = estimador, fill = estimador)) +
    geom_col() +
    facet_grid(cols = vars(plano), scales = "free") +
    theme_bw() +
    theme(legend.position = "none", axis.text.x = element_text(angle = 45))
```



6. Use os resultados obtidos em 3. e 4. para comparar os dois planos amostrais e indique qual deles seria preferível usar, caso fosse necessário amostrar duas unidades distintas da população ($n = 2$) para estimar o total da renda. Justifique.

É preferível usar o plano amostral A pois leva em consideração todas as possíveis amostras de tamanho 2 da população de tamanho 6, enquanto o segundo plano selecionou apenas 5 dessas amostras, que acabaram sendo as amostras com maior total de renda, que aumentou consideravelmente o estimador da média de 10000 para 16080.