Lista de exercício

Estimadores razão e regressão

Paulo Ricardo Seganfredo Campana

3 de maio de 2023

Questão 1.

Considerando os estimadores $\hat{\bar{y}}$: média amostral e $\hat{\bar{y}}_{reg}$: estimador regressão para a média populacional e sabendo-se que:

$$\mathrm{Var}\left[\hat{\bar{y}}\right] - \mathrm{Var}\left[\hat{\bar{y}}_{reg}\right] = \rho_{XY}^2 \frac{\sigma_Y^2}{n} \tag{1}$$

$$\mathrm{Var}\left[\hat{\bar{y}}_R\right] - \mathrm{Var}\left[\hat{\bar{y}}_{reg}\right] = \frac{1}{n}(\rho_{XY}\sigma_Y - B\sigma_X)^2 \tag{2}$$

Então é correto afirmar que:
$\mathrm{Var}\left[\hat{\bar{y}}\right] < \mathrm{Var}\left[\hat{\bar{y}}_R\right]$ é verdadeira se e somente se $\mathrm{Cov}[X, Y] < \frac{B}{2}\mathrm{Var}[X]$, em que B é a razão populacional entre X e Y.

subtraindo as equações (1) e (2) temos que:

$$\begin{aligned}
\mathrm{Var}\left[\hat{\bar{y}}\right] - \mathrm{Var}\left[\hat{\bar{y}}_R\right] &= \rho_{XY}^2 \frac{\sigma_Y^2}{n} - \frac{1}{n}(\rho_{XY}\sigma_Y - B\sigma_X)^2 \\
&= \frac{1}{n}\left(\rho_{XY}^2\sigma_Y^2 - (\rho_{XY}\sigma_Y - B\sigma_X)^2\right) \\
&= \frac{1}{n}\left(\rho_{XY}^2\sigma_Y^2 - \rho_{XY}^2\sigma_Y^2 + 2\rho_{XY}\sigma_Y\sigma_X B - B^2\sigma_X^2\right) \\
&= \frac{1}{n}\left(2\rho_{XY}\sigma_Y\sigma_X B - B^2\sigma_X^2\right)
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}\left[\hat{\bar{y}}\right] < \mathrm{Var}\left[\hat{\bar{y}}_R\right] &\iff \mathrm{Var}\left[\hat{\bar{y}}\right] - \mathrm{Var}\left[\hat{\bar{y}}_R\right] < 0 \\
&\iff \frac{1}{n}\left(2\rho_{XY}\sigma_Y\sigma_X B - B^2\sigma_X^2\right) < 0 \\
&\iff 2\rho_{XY}\sigma_Y\sigma_X B < B^2\sigma_X^2 \\
&\iff 2\rho_{XY}\sigma_Y < B\sigma_X \\
&\iff \rho_{XY} < \frac{R}{2}\frac{\sigma_X}{\sigma_Y} \\
&\iff \mathrm{Cov}[X,Y] < \frac{R}{2}\frac{\mathrm{Var}[X]}{\mathrm{Var}[Y]}
\end{aligned}
$$

Questão 2.

A afirmativa:

> "Os estimadores do tipo razão e regressão para a média não são viesados e são mais eficientes do que o estimador da média no método aleatório simples"

É verdadeira ou falsa? Justifique sua resposta.

É falsa, pois calculando o viés do estimador $\hat{\bar{y}}_R$ vemos que o mesmo é diferente de 0. Porém, eles são mais eficientes, pois comparando a variância dos estimadores, fica claro que o estimador regressão é o mesmo do estimador da média populacional multiplicado por um termo entre $[0, 1]$.

$$
\begin{aligned}
\mathrm{Bias}(\hat{\bar{y}}_R) &= \mathrm{E}\left[\hat{\bar{y}}_R - \bar{y}_U\right] \\
&= \mathrm{E}\left[\frac{\bar{y}}{\bar{x}}\bar{x}_U - \bar{y}_U\right] \\
&= \mathrm{E}\left[\bar{y}\frac{\bar{x}_U}{\bar{x}} - \bar{y}_U\right] \\
&= \mathrm{E}\left[\bar{y}\left(1 - \frac{\bar{x}-\bar{x}_U}{\bar{x}}\right) - \bar{y}_U\right] \\
&= \mathrm{E}\left[\bar{y} - \frac{\bar{y}}{\bar{x}}(\bar{x}-\bar{x}_U) - \bar{y}_U\right] \\
&= \mathrm{E}[\bar{y}] - \mathrm{E}\left[\hat{B}(\bar{x}-\bar{x}_U)\right] - \mathrm{E}[\bar{y}_U] \\
&= -\mathrm{E}\left[\hat{B}(\bar{x}-\bar{x}_U)\right] \\
&= -\mathrm{Cov}\left[\hat{B}, \bar{x}\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}\left[\hat{\bar{y}}_{reg}\right] &= \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}(1 - R^2) \\
&= \mathrm{Var}\left[\hat{\bar{y}}\right](1 - R^2) \\
&\leq \mathrm{Var}\left[\hat{\bar{y}}\right]
\end{aligned}
$$

$$
\text{com } R = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} \text{ (correlação)}
$$

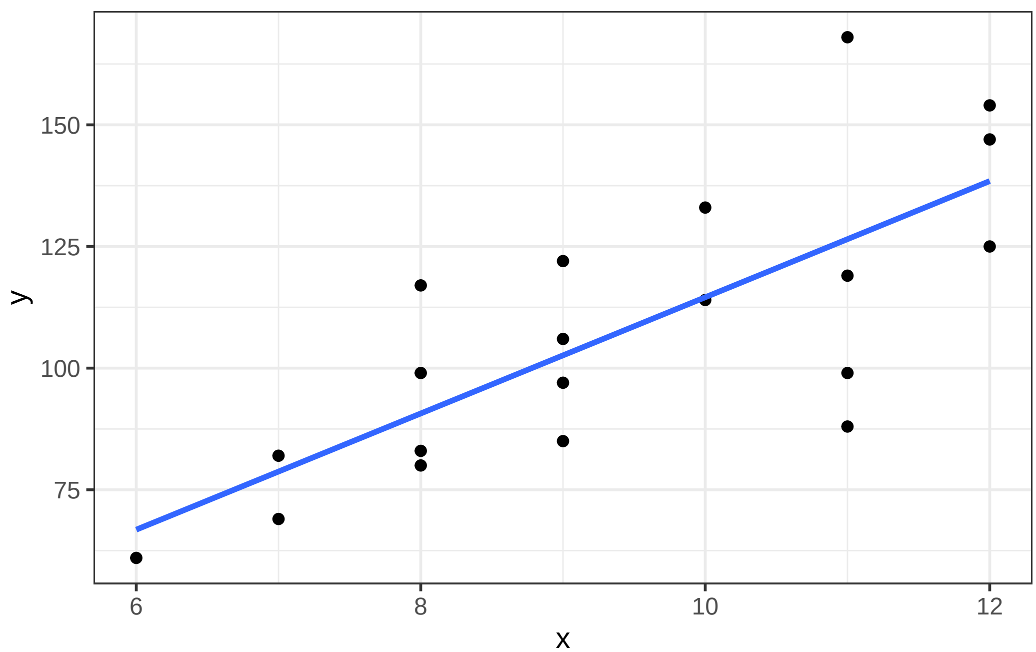
Questão 3.

Pesquisadores desejam estimar a idade média de determinada espécie de árvore em uma reserva florestal e isso não é simples pois precisam arrancar pedaços da árvore e contar os anéis de crescimento presentes neste pedaço. Sabe-se que quanto mais velha for a árvore maior será o seu diâmetro X. Então foi mais fácil medir o diâmetro de todas as 1132 árvores desta reserva e obtiveram diâmetro médio igual a 10 m. Em seguida coletaram uma amostra de 20 árvores para avaliar sua idade pelo processo complexo Y e obteve-se as seguintes informações:

x	12	11	8	9	11	8	7	10	12	11	6	8	10	12	9	9	7	11	9	8
y	125	119	83	85	99	117	69	133	154	168	61	80	114	147	122	106	82	88	97	99

a) Construa o gráfico de dispersão de X e Y e calcule o seu coeficiente de correlação.

```
q3 |> ggplot(aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```




```
cor(q3$x, q3$y)
# > [1] 0.7573225
```

```r
N <- 1132
n <- 20
μx <- 10
```

b) e c) Calcule a estimativa para a média das árvores utilizando o estimador razão e sem utilizar a variável auxiliar X, considerando que a amostra de Y foi selecionada com o plano amostral amostragem aleatória simples.

```r
B <- mean(q3$y) / mean(q3$x)
B * μx
# > [1] 114.2553
```

```r
mean(q3$y)
# > [1] 107.4
```

d) Calcule um IC a 95% de confiança para os itens b) e c).

```r
resid <- B * q3$x - q3$y
var <- (1 - n/N) *
    (μx / mean(q3$x))^2 *
    var(resid) / n
erro <- qt(0.95, df = n-1) * sqrt(var)
B * μx + c(-erro, +erro)
# > [1] 106.6151 121.8955
```

```r
var <- (1 - n/N) * var(q3$y) / n
erro <- qt(0.95, df = n-1) * sqrt(var)
mean(q3$y) + c(-erro, +erro)
# > [1]  96.41625 118.38375
```

e) Qual dos dois estimadores você escolheria como o melhor? Justifique sua resposta.

O estimador razão, pois apresenta menor variância.

f) Como você justificaria a escolha da estimativa pelo estimador razão?

Pela correlação entre as variáveis X e Y ser razoável (75%).

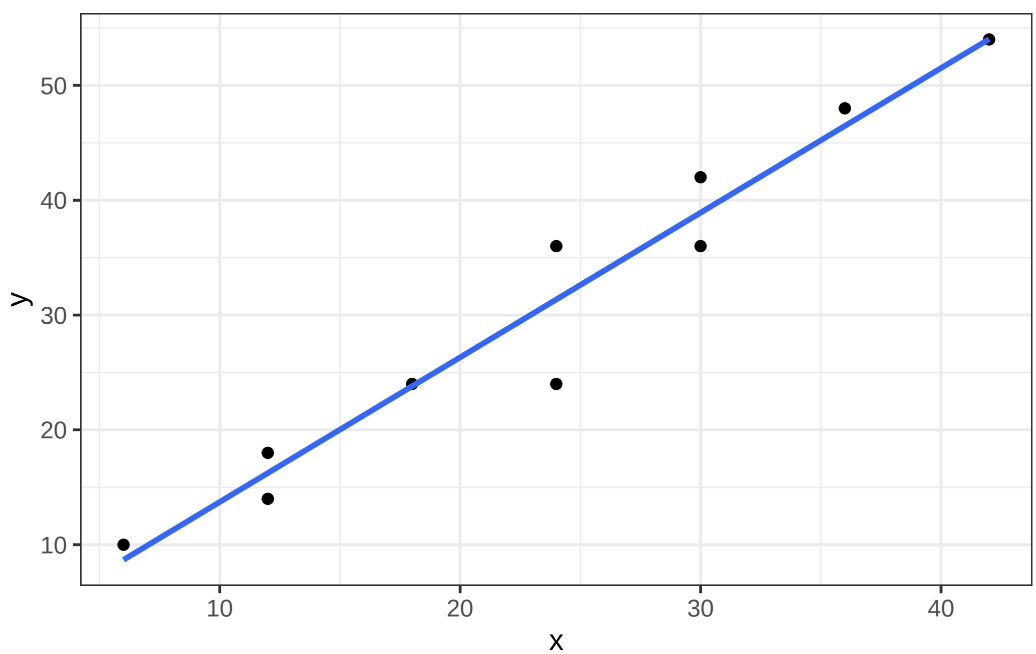
Questão 4.

Deseja-se estimar o número de árvores mortas de determinada espécie em uma reserva florestal. Dividiu-se esta reserva em áreas de 1.5 hectare e o número de árvores mortas foi avaliada por fotografia aérea X nas 200 áreas que apresentou uma contagem de 15600 árvores mortas da espécie em estudo. Em 10 das 200 áreas selecionadas o número de árvores mortas foi avaliada por contagem terrestre Y e já se sabia por fotografia aérea X. Estas informações estão na Tabela seguinte

x	12	30	24	24	18	30	12	6	36	42
y	18	42	24	36	24	36	14	10	48	54

a) Construa um gráfico de dispersão entre X e Y e calcule o coeficiente de correção entre X e Y.

```r
q4 |> ggplot(aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```




```r
cor(q4$x, q4$y)
# > [1] 0.9729456
```

```r
N <- 200
n <- 10
tx <- 15600
```

b) e c) Calcule a estimativa para o número de árvores mortas utilizando o estimador razão. e sem utilizar a variável auxiliar X, considerando que a amostra de Y foi selecionada com o plano amostral amostragem aleatória simples.

```r
B <- mean(q4$y) / mean(q4$x)
B * tx
# > [1] 20400
```

```r
mean(q4$y) * N
# > [1] 6120
```

d) e e) Calcule as variâncias estimadas e um IC a 95% de confiança para cada um dos estimadores obtidos nos itens b) e c).

```r
resid <- B * q4$x - q4$y
var <- (1 - n/N) *
    (tx / mean(q4$x))^2 *
    var(resid) / n
erro <- qt(0.95, df = n-1) * sqrt(var)
B * tx + c(-erro, +erro)
# > [1] 19091.04 21708.96
```

```r
var <- (1 - n/N) * var(q4$y) / n * N^2
erro <- qt(0.95, df = n-1) * sqrt(var)
mean(q4$y) * N + c(-erro, +erro)
# > [1] 4442.068 7797.932
```

f) Qual dos dois estimadores você escolheria como o melhor? Justifique sua resposta.

O estimador razão, pois apresenta menor variância.

g) Como você justificaria a possível escolha da estimativa pelo estimador Razão?

Pela correlação entre as variáveis X e Y ser muito boa (97%).

h) O estimador razão é o mais eficiente?

Sim, pois possui menor variância.

Questão 5.

Observe a seguinte população do exemplo 4.4 (LOHR, p. 121)

i	1	2	3	4	5	6	7	8
x	4	5	5	6	8	7	7	5
y	1	2	4	4	7	7	7	8

Utilizando a linguagem R como você determinaria a distribuição de $\hat{t}_y$ e $\hat{t}_{y_R}$ para a situação de amostras de tamanho $n = 4$.

```
N <- 8
n <- 4
µx <- mean(q5$x)
```

```
simples <- q5$i |>
    sample(size = n) |>
    replicate(n = 10000, simplify = FALSE) |>
    lapply(sort) |>
    unique() |>
    sapply(function(i) mean(q5$y[i]) * N) |>
    round(0) |>
    as_tibble() |>
    summarise(prop = n() / 70, .by = value) |>
    add_column(tipo = "simples")

razão <- q5$i |>
    sample(size = n) |>
    replicate(n = 10000, simplify = FALSE) |>
    lapply(sort) |>
    unique() |>
    sapply(function(i) mean(q5$y[i]) / mean(q5$x[i]) * µx * N) |>
    round(0) |>
    as_tibble() |>
    summarise(prop = n() / 70, .by = value) |>
    add_column(tipo = "razão")
```

```
rbind(simples, razão) |>
    ggplot(aes(x = value, y = 0)) +
    geom_segment(aes(xend = value, yend = prop)) +
    facet_grid(cols = vars(fct_rev(tipo))) +
    lims(x = c(20,60), y = c(0,0.2))
```