Amostragem aleatória simples com reposição (parte 2)

Prof. Caio Azevedo

Estimação da proporção populacional

- População: observações univariadas - $y_1, ..., y_N$ (variáveis não aleatórias), em que y_i é a observação relativa ao indivíduo i (podemos também considerar observações multivariadas).
- Temos que $y_i = 1$ se o indivíduo i possui a característica de interesse e 0 caso contrário.

Estimação da proporção

- Exemplos: presença de alguma doença, procedência (1 se é oriundo de determinado lugar, 0, caso contrário), inadimplência (1 se inadimplente, 0 caso contrário).
- Os procedimentos definidos anteriormente, em princípio, se mantêm (slides AAS_c parte 1, link). A principal diferença, de forma geral, reside na estrutura da variável de interesse.
- Parâmetro de interesse: $p = \frac{1}{N}\sum_{i=1}^{N} y_i$.
- Lembremos que $y_i = y_i^k$, $\forall k \in \mathbb{R}^+$.

Estimação da proporção

- Estimador "natural":

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i \in \boldsymbol{s}} y_i$$
$$= \frac{1}{n}\sum_{i=1}^{N} F_i y_i$$

- Note que, neste caso, a variância populacional $(\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - p)^2)$, toma a seguinte forma:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^2 - 2y_i p + p^2\right) = \frac{1}{N}\left(Np - Np^2\right)$$
$$= p(1-p) = pq, q = 1 - p$$

Propriedades do estimador

- Note que, essencialmente, $\widehat{p}$ é uma média amostral (de variáveis binárias), semelhante à $\widehat{\mu}$ em link.
- Portanto, as propriedades de $\widehat{p}$ são semelhantes as de $\widehat{\mu}$, por exemplo:
  - $\mathcal{E}_{A_1}(\widehat{p}) = \mathcal{E}_{A_1}(\widehat{\mu}) = \frac{1}{n}\sum_{i=1}^{N} y_i \mathcal{E}(F_i) = \frac{1}{n}\sum_{i=1}^{N} y_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = p.$
  - $\mathcal{V}_{A_1}(\widehat{p}) = \mathcal{V}_{A_1}(\widehat{\mu}) = \frac{\sigma^2}{n} = \frac{pq}{n}.$
- Estimativa: $\widetilde{p} = \frac{1}{n}\sum_{i \in \boldsymbol{s}} y_i = \frac{1}{n}\sum_{i=1}^{N} f_i y_i$

Propriedades do estimador

- Vimos também que um estimador não viciado para a variância populacional (σ^2) é dado por

$$
\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{p})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{N} F_i\,(y_i - \widehat{p})^2
\end{aligned}
$$

- Note, no entanto, que neste caso

$$
\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i^2 - 2Y_i\widehat{p} + \widehat{p}^2\right) = \frac{1}{n-1}\left(n\widehat{p} - n\widehat{p}^2\right) \\
&= \frac{n}{n-1}\widehat{p}\widehat{q}, \widehat{q} = 1 - \widehat{p}
\end{aligned}
$$

Propriedades do estimador

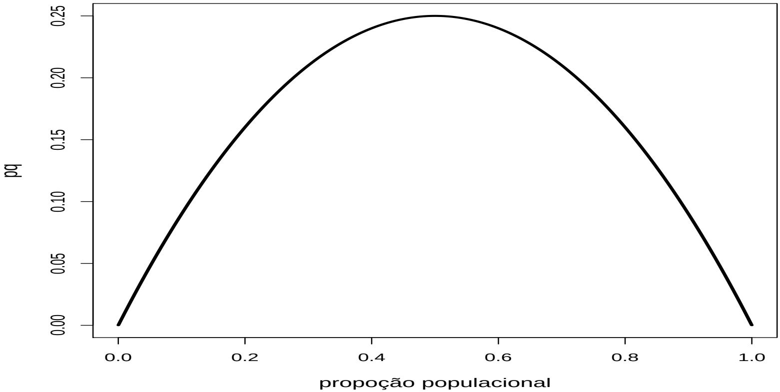
- Consequentemente, um estimador não viciado para a variância do estimador (σ^2/n) é dado por:

$$\frac{\widehat{\sigma}^2}{n} = \frac{n\widehat{p}\widehat{q}}{n(n-1)} = \frac{\widehat{p}\widehat{q}}{n-1}$$

- Analogamente ao caso da média, temos que

$$\frac{\widehat{p}-p}{\sqrt{pq/n}} \xrightarrow[\substack{n\to\infty,\\ N-n\to\infty}]{D} N(0,1) \; ; \; \frac{\widehat{p}-p}{\sqrt{\widehat{p}\widehat{q}/(n-1)}} \xrightarrow[\substack{n\to\infty,\\ N-n\to\infty}]{D} N(0,1)$$

Variância do estimador pq



Intervalo de Confiança

- Assim, dois intervalos de confiança (assintóticos) com coeficiente de confiança de aproximadamente γ, são dados por

$$IC(p,\gamma) \approx \left[\widehat{p} - z_{\gamma}\sqrt{\frac{\widehat{p}\widehat{q}}{n-1}}; \widehat{p} + z_{\gamma}\sqrt{\frac{\widehat{p}\widehat{q}}{n-1}}\right] \quad (1)$$

$$IC(p,\gamma) \approx \left[\widehat{p} - z_{\gamma}\sqrt{\frac{1}{4(n-1)}}; \widehat{p} + z_{\gamma}\sqrt{\frac{1}{4(n-1)}}\right] \quad (2)$$

em que $P\left(Z \leq z_{\gamma}\right) = \frac{1+\gamma}{2}$ e $Z \sim N(0,1)$.

- Erro da estimativa: $z_{\gamma}\sqrt{\frac{\widehat{p}\widehat{q}}{n-1}}$ ou $z_{\gamma}\sqrt{\frac{1}{4(n-1)}}$.
- O comprimento do intervalo (2) (variância (pq) máxima) sempre será maior (ou igual) ao comprimento do intervalo (1). Exercício: provar que o máximo de $g(p) = p(1-p)$ é obtido para p $= 1/2$.

Testes de Hipótese

- Hipóteses usuais (p_0 conhecido, $q_0 = 1 - p_0$)
  1. $H_0 : p = p_0$ vs $H_1 : p < p_0$.
  2. $H_0 : p = p_0$ vs $H_0 : p > p_0$.
  3. $H_0 : p = p_0$ vs $H_0 : p \neq p_0$.
- Estatística do teste $Z_t = \frac{\widehat{p} - p_0}{\sqrt{p_0 q_0 / n}}$.
- Sob H_0, vimos que $Z_t \approx N(0,1)$, para n e N-n suficientemente grandes.
- Defina $z_t = \frac{\widetilde{p} - p_0}{\sqrt{p_0 q_0 / n}}$ o valor calculado da estatística do teste e z_c o(s) valor(es) crítico(s).
- Defina ainda $Z \sim N(0,1)$. Os procedimentos são análogos ao caso da média, com as devidas adaptações.

Determinação do tamanho amostral: erro da estimativa

Analogamente ao caso da média populacional, temos que

$$\delta = z_{\gamma}\sqrt{\frac{pq}{n}} \rightarrow n = \frac{z_{\gamma}^{2}pq}{\delta^{2}} \qquad (3)$$

Podemos usar estimativas de p obtidas em pesquisas anteriores, sob uma amostra piloto ou, considerar o pior caso, em termos da variabilidade dos dados. Nessee último caso, temos que:

$$n = \frac{z_{\gamma}^{2}}{4\delta^{2}} \qquad (4)$$

Isto vale para qualquer um dos dois critérios: erro da estimativa e precisão. Note que o tamanho da amostra fornecido por (4) será maior ou igual àquele fornecido por (3).

Estudos de simulação

- Distribuição assintótica do estimador para a proporção. Tamanho da população $N = 100.000$.
- Vários cenários, variando em função do valor verdadeiro da proporção populacional p.
- $p = (0,01; 0,05; 0,10; 0,25; 0,35; 0,50; 0,65; 0,75; 0,9; 0,95; 0,99)^T$.
- A distribuição, (em princípio), da variável de interesse é Bernoulli(p).

Estudos de simulação

- Quatro tamanhos amostrais $(30, 50, 100, 1000)$, em termos percentuais, com relação ao tamanho da população (0,03%,0,05%,0,1%,1%).
- Estudar a distribuição amostral (empírica) com base em R $=$ 1.000 réplicas (amostras selecionadas da população de interesse).

p=0,01

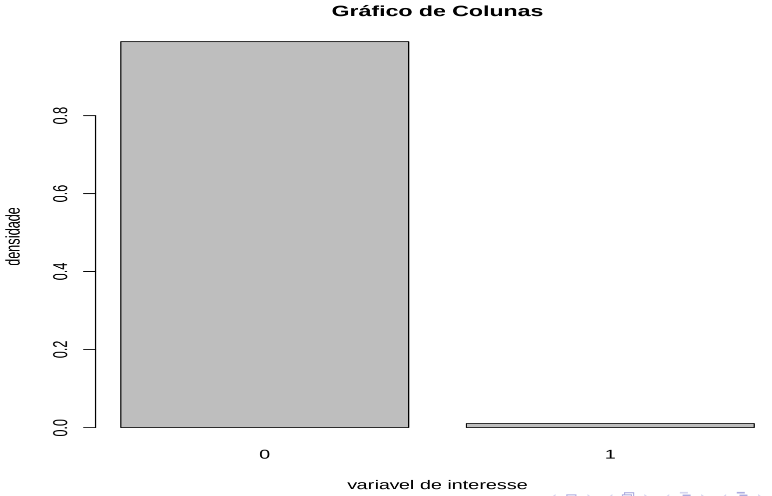
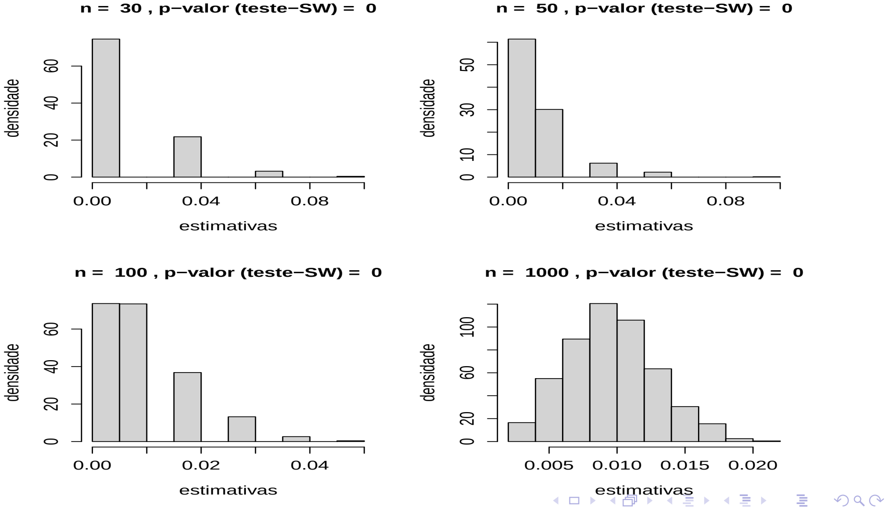


Gráfico de Colunas
densidade
0.0
0.2
0.4
0.6
0.8
0
1
variavel de interesse

p=0,01



n = 30 , p−valor (teste−SW) = 0
densidade
0
20
40
60
0.00
0.04
0.08
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
0
10
30
50
0.00
0.04
0.08
estimativas
n = 100 , p−valor (teste−SW) = 0
densidade
0
20
40
60
0.00
0.02
0.04
estimativas
n = 1000 , p−valor (teste−SW) = 0
densidade
0
20
60
100
0.005
0.010
0.015
0.020
estimativas

p=0,05

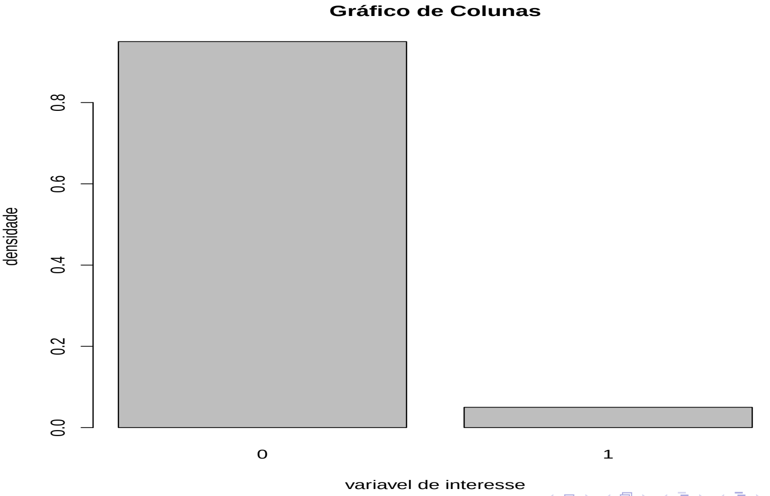
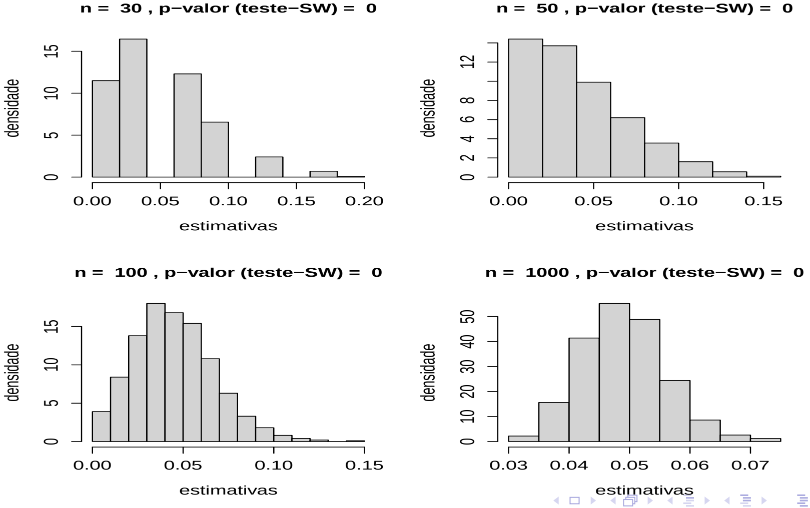


Gráfico de Colunas
densidade
0.0
0.2
0.4
0.6
0.8
0
1
variavel de interesse

p=0,05


n = 30 , p−valor (teste−SW) = 0
densidade
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
estimativas
n = 100 , p−valor (teste−SW) = 0
densidade
estimativas
n = 1000 , p−valor (teste−SW) = 0
densidade
estimativas

p=0,10

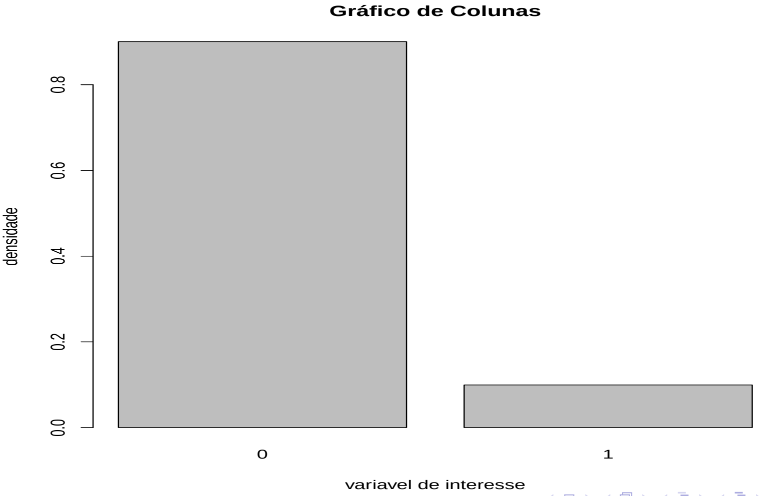
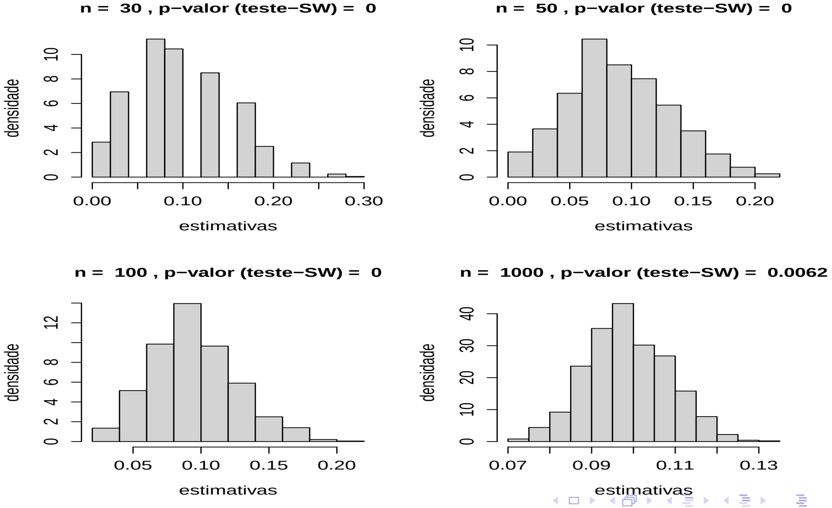


Gráfico de Colunas
densidade
0.0
0.2
0.4
0.6
0.8
0
1
variavel de interesse

p=0,10


n = 30 , p−valor (teste−SW) = 0
densidade
0
2
4
6
8
10
0.00
0.10
0.20
0.30
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
0
2
4
6
8
10
0.00
0.05
0.10
0.15
0.20
estimativas
n = 100 , p−valor (teste−SW) = 0
densidade
0
2
4
6
8
12
0.05
0.10
0.15
0.20
estimativas
n = 1000 , p−valor (teste−SW) = 0.0062
densidade
0
10
20
30
40
0.07
0.09
0.11
0.13
estimativas

p=0,25

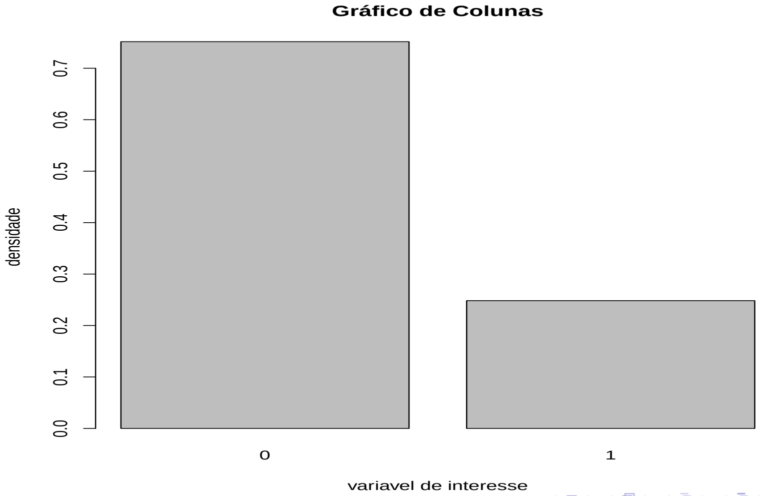
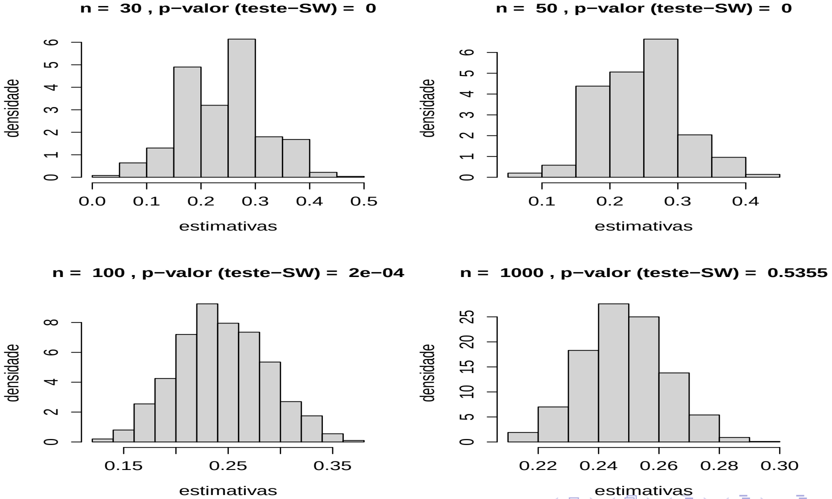


Gráfico de Colunas
densidade
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
variavel de interesse

p=0,25



n = 30 , p−valor (teste−SW) = 0
n = 50 , p−valor (teste−SW) = 0
n = 100 , p−valor (teste−SW) = 2e−04
n = 1000 , p−valor (teste−SW) = 0.5355
densidade
estimativas

p=0,35

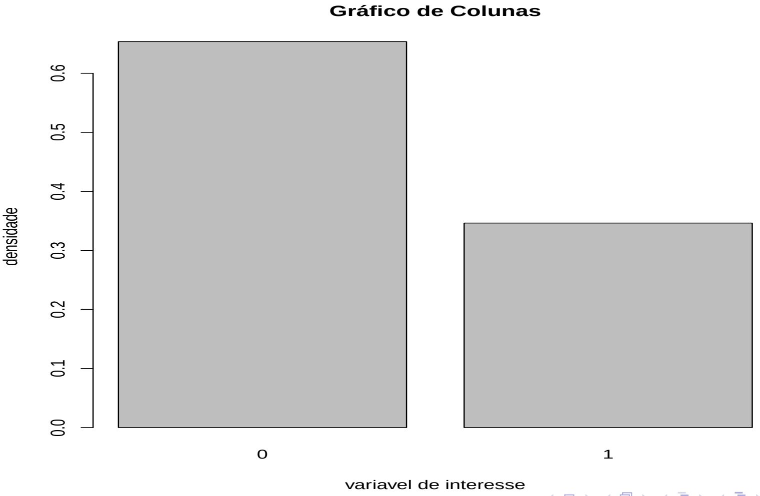
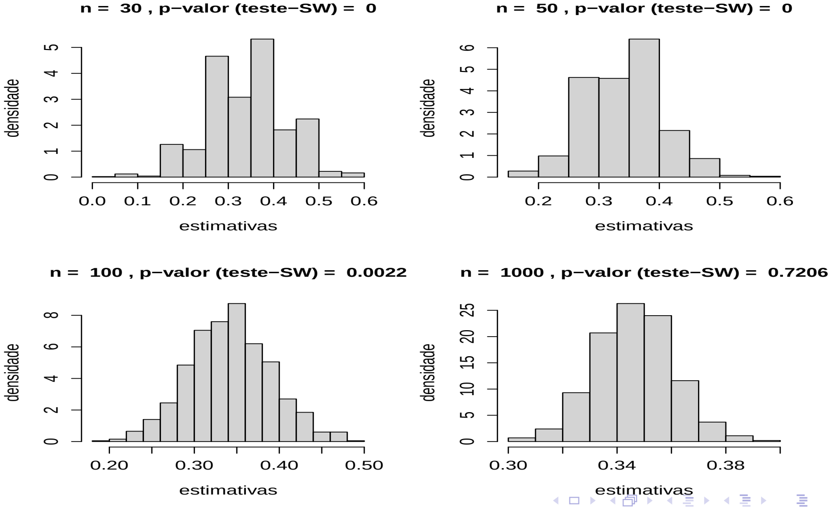


Gráfico de Colunas
densidade
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
variavel de interesse

p=0,35



n = 30 , p−valor (teste−SW) = 0
densidade
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
estimativas
n = 100 , p−valor (teste−SW) = 0.0022
densidade
estimativas
n = 1000 , p−valor (teste−SW) = 0.7206
densidade
estimativas

p=0,50

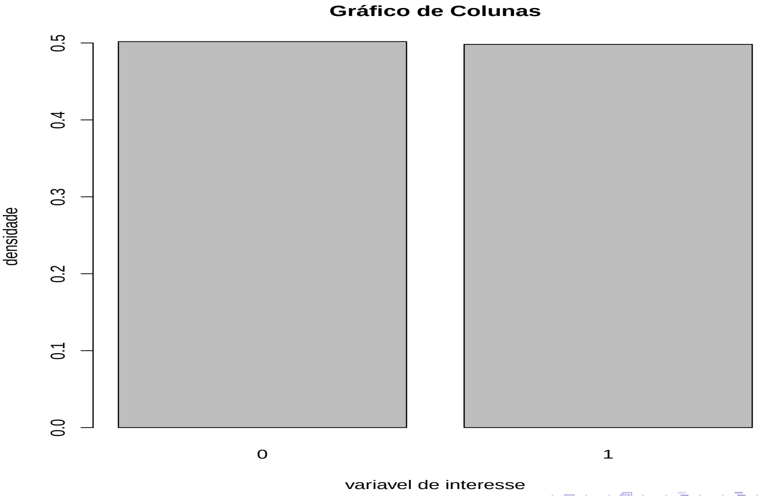
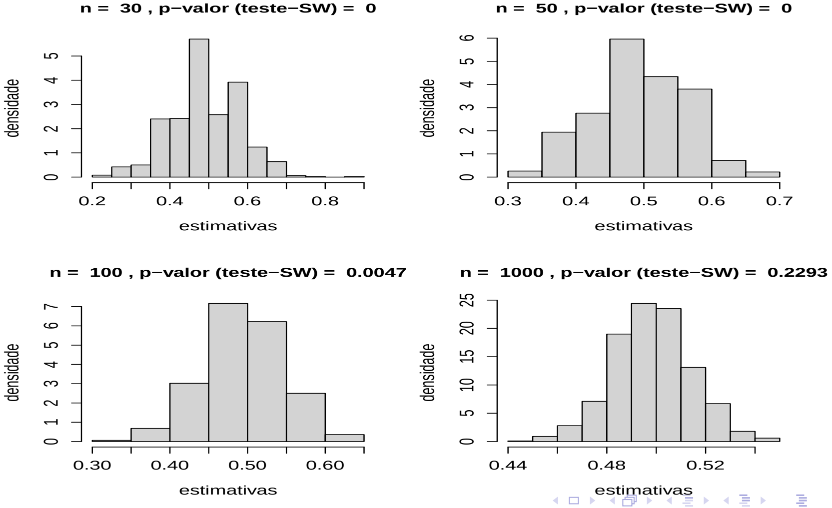


Gráfico de Colunas
densidade
0.0
0.1
0.2
0.3
0.4
0.5
0
1
variavel de interesse

p=0,50



n = 30 , p−valor (teste−SW) = 0
densidade
0 1 2 3 4 5
0.2 0.4 0.6 0.8
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
0 1 2 3 4 5 6
0.3 0.4 0.5 0.6 0.7
estimativas
n = 100 , p−valor (teste−SW) = 0.0047
densidade
0 1 2 3 4 5 6 7
0.30 0.40 0.50 0.60
estimativas
n = 1000 , p−valor (teste−SW) = 0.2293
densidade
0 5 10 15 20 25
0.44 0.48 0.52
estimativas

p=0,65

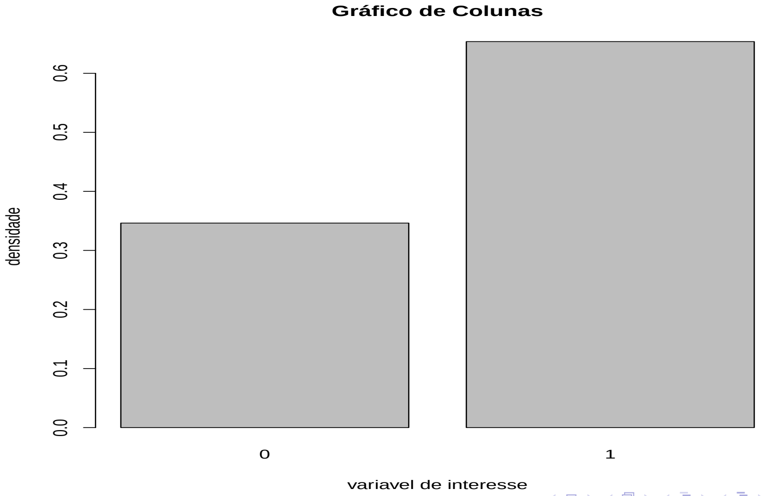
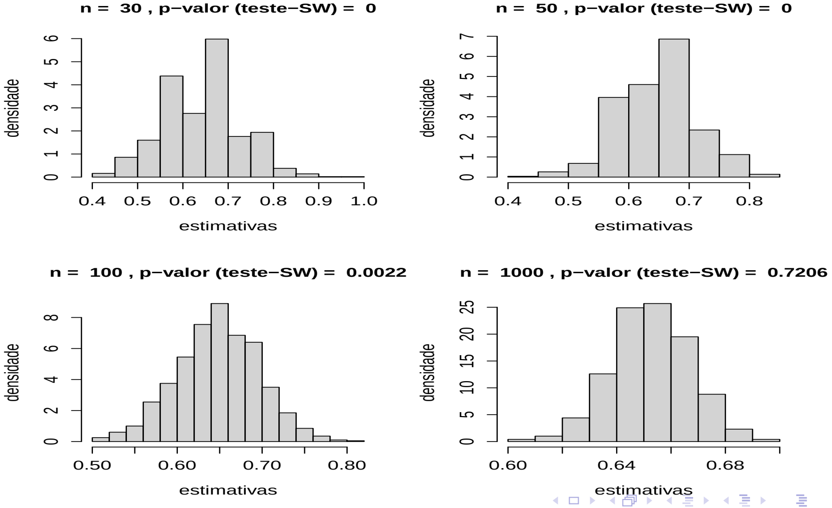


Gráfico de Colunas
densidade
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
variavel de interesse

p=0,65



n = 30 , p−valor (teste−SW) = 0
densidade
0 1 2 3 4 5 6
0.4 0.5 0.6 0.7 0.8 0.9 1.0
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
0 1 2 3 4 5 6 7
0.4 0.5 0.6 0.7 0.8
estimativas
n = 100 , p−valor (teste−SW) = 0.0022
densidade
0 2 4 6 8
0.50 0.60 0.70 0.80
estimativas
n = 1000 , p−valor (teste−SW) = 0.7206
densidade
0 5 10 15 20 25
0.60 0.64 0.68
estimativas

p=0,75

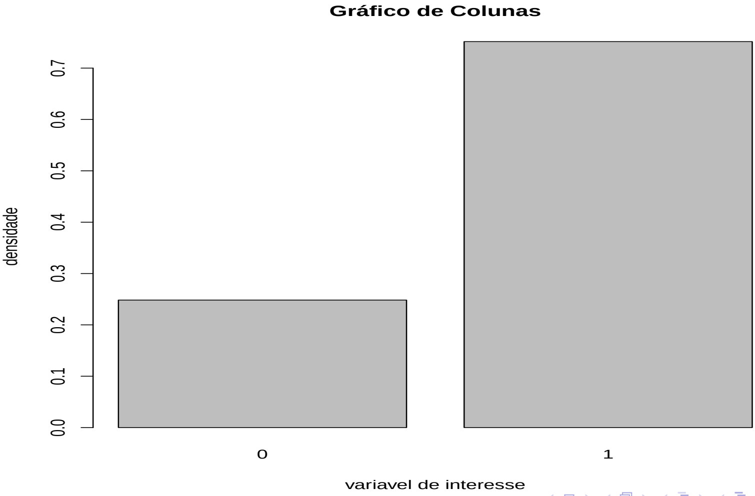
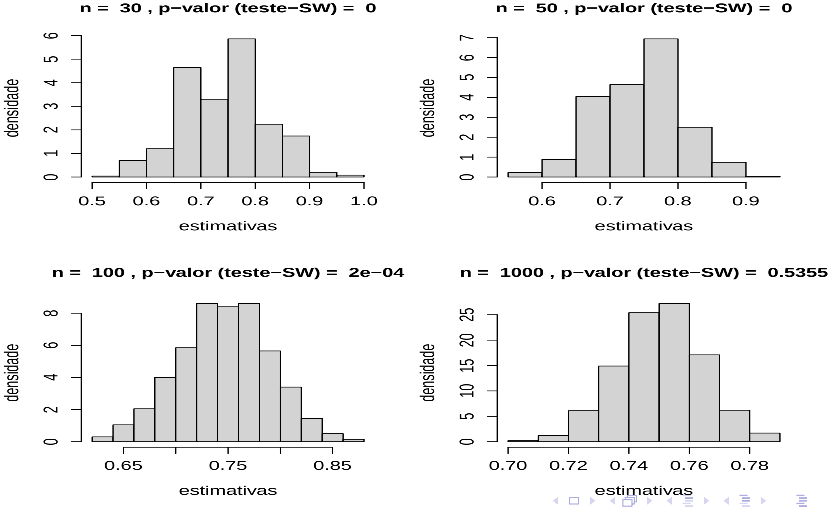


Gráfico de Colunas
densidade
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
variavel de interesse

p=0,75


n = 30 , p−valor (teste−SW) = 0
densidade
0 1 2 3 4 5 6
0.5 0.6 0.7 0.8 0.9 1.0
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
0 1 2 3 4 5 6 7
0.6 0.7 0.8 0.9
estimativas
n = 100 , p−valor (teste−SW) = 2e−04
densidade
0 2 4 6 8
0.65 0.75 0.85
estimativas
n = 1000 , p−valor (teste−SW) = 0.5355
densidade
0 5 10 15 20 25
0.70 0.72 0.74 0.76 0.78
estimativas

p=0,90

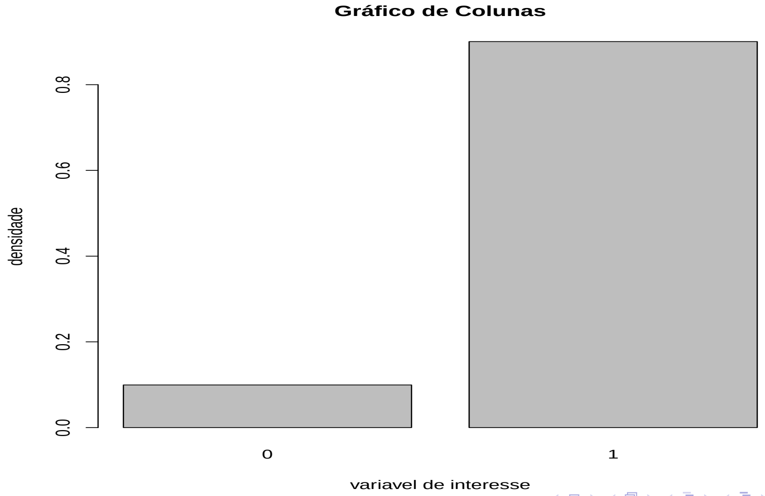
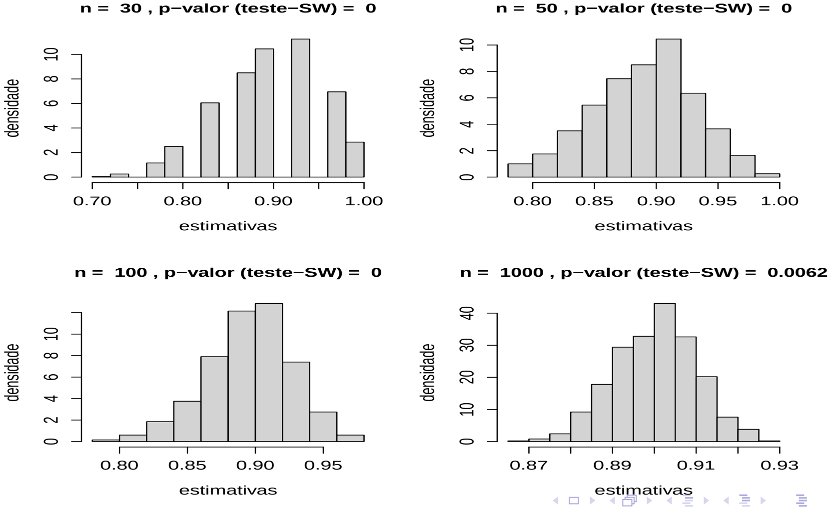


Gráfico de Colunas
densidade
0.0
0.2
0.4
0.6
0.8
0
1
variavel de interesse

p=0,90


n = 30 , p−valor (teste−SW) = 0
densidade
0
2
4
6
8
10
0.70
0.80
0.90
1.00
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
0
2
4
6
8
10
0.80
0.85
0.90
0.95
1.00
estimativas
n = 100 , p−valor (teste−SW) = 0
densidade
0
2
4
6
8
10
0.80
0.85
0.90
0.95
estimativas
n = 1000 , p−valor (teste−SW) = 0.0062
densidade
0
10
20
30
40
0.87
0.89
0.91
0.93
estimativas

p=0,95

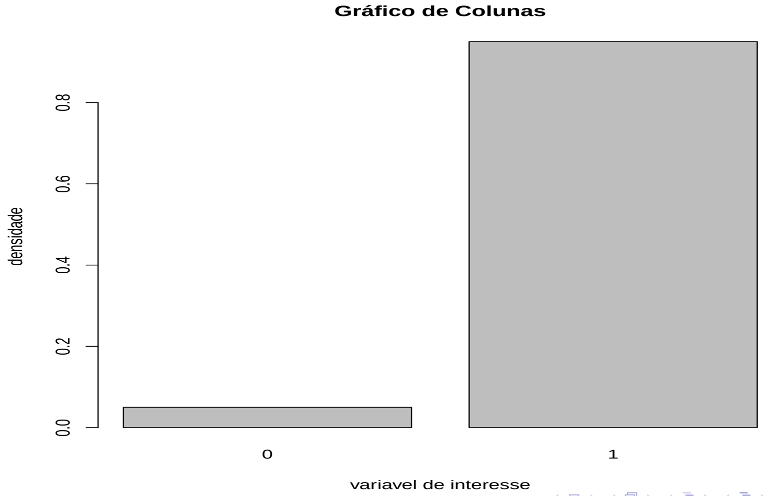
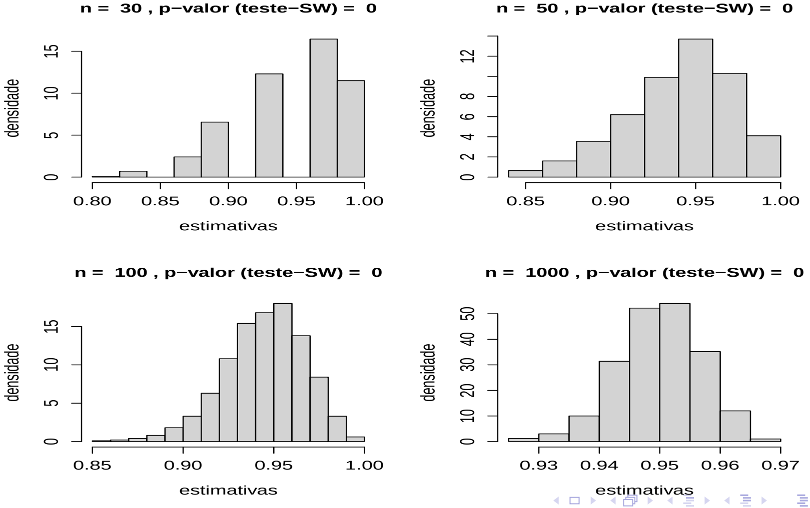


Gráfico de Colunas
densidade
0.0
0.2
0.4
0.6
0.8
0
1
variavel de interesse

p=0,95


n = 30 , p−valor (teste−SW) = 0
n = 50 , p−valor (teste−SW) = 0
n = 100 , p−valor (teste−SW) = 0
n = 1000 , p−valor (teste−SW) = 0
densidade
estimativas
0 5 10 15
0.80 0.85 0.90 0.95 1.00
0 2 4 6 8 12
0.85 0.90 0.95 1.00
0 5 10 15
0.85 0.90 0.95 1.00
0 10 20 30 40 50
0.93 0.94 0.95 0.96 0.97

p=0,99

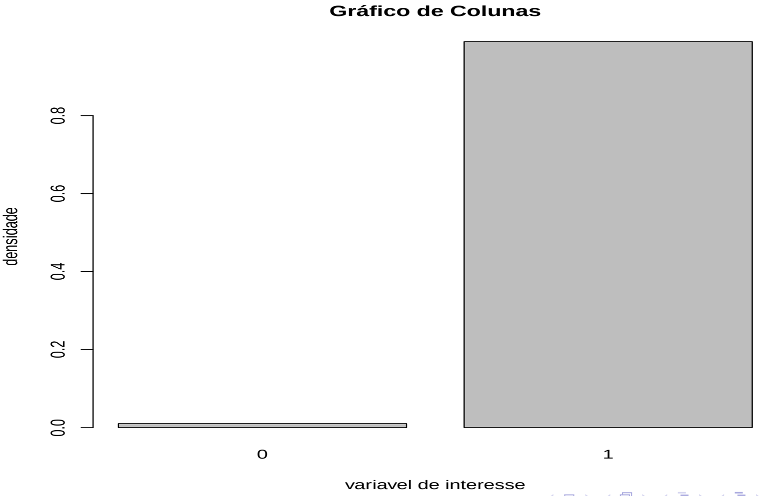
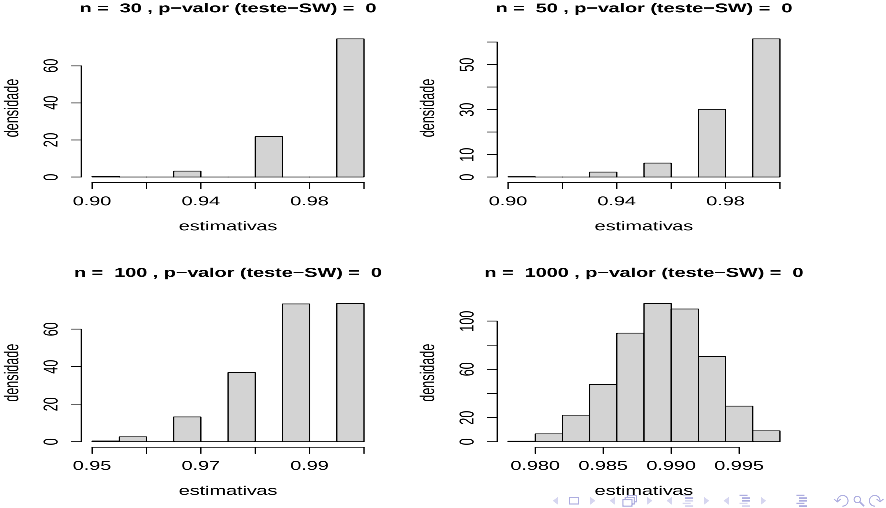


Gráfico de Colunas
densidade
0.0
0.2
0.4
0.6
0.8
0
1
variavel de interesse

p=0,99


n = 30 , p−valor (teste−SW) = 0
densidade
0
20
40
60
0.90
0.94
0.98
estimativas
n = 50 , p−valor (teste−SW) = 0
densidade
0
10
30
50
0.90
0.94
0.98
estimativas
n = 100 , p−valor (teste−SW) = 0
densidade
0
20
40
60
0.95
0.97
0.99
estimativas
n = 1000 , p−valor (teste−SW) = 0
densidade
0
20
60
100
0.980
0.985
0.990
0.995
estimativas

Otimalidade dos estimadores

- Vamos nos concentrar na média amostral e na classe de estimadores não viciados que sejam combinações lineares das variáveis aleatórias $(Y_1, ..., Y_n)'$.
- Os resultados para os outros parâmetros são análogos.
- A forma geral do estimador em questão é dada por

$$\widehat{\mu}_{sc} = \sum_{i=1}^{n} c_i Y_i$$

- Note que, sob $AAS_c(A_1)$ temos que $Y_i \overset{i.i.d.}{\sim} D(\mu, \sigma^2)$, em que $\mathcal{E}(Y_i) = \mu$ e $\mathcal{V}(Y_i) = \sigma^2$. Além disso, $c_i, i = 1, 2, ..., n$ são não aleatórios.

Otimalidade dos estimadores

- Note que $\mathcal{E}(\widehat{\mu}_{sc}) = \sum_{i=1}^{n} c_i \mathcal{E}(Y_i) = \mu \sum_{i=1}^{n} c_i$ e $\mathcal{V}(\widehat{\mu}_{sc}) = \sum_{i=1}^{n} c_i^2 \mathcal{V}(Y_i) = \sum_{i=1}^{n} \sigma^2 = \sigma^2 \sum_{i=1}^{n} c_i^2$.
- Exercício: provar que $\widehat{\mu}_{sc}$ é um estimador não viciado se e somente se

$$\sum_{i=1}^{n} c_i = 1 \tag{5}$$

- Defina $\overline{c} = \frac{1}{n}\sum_{i=1}^{n} c_i$. Pode-se provar que (usando (5)).

$$\begin{aligned}\mathcal{V}(\widehat{\mu}_{sc}) &= \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \left\{\sum_{i=1}^{n} (c_i - \overline{c})^2 + n\overline{c}^2\right\} \\ &= \sigma^2 \left\{\sum_{i=1}^{n} \left(c_i - \frac{1}{n}\right)^2 + \frac{1}{n}\right\}\end{aligned}$$

- A expressão acima atinge seu mínimo quando $c_i = \frac{1}{n}, i = 1, 2, ..., n$. Portanto, segue-se que o estimador $\widehat{\mu}$ (e consequentemente, $\widehat{\tau}$ e $\widehat{p}$) são ótimos, sob as condições postas (estimadores não viciados que são combinações lineares das observações, sob AAS_c).
- Exercício: como poderia ser demonstrado para AAS_s?
- Resolvas os exercícios dos Capítulo 1 do livro "Elementos de Amostragem" e do Capítulo 2 do livro "Amostragem: Teoria e Prática usando R".