Amostragem aleatória simples sem reposição (parte 1)

Prof. Caio Azevedo

Estrutura geral

- Temos uma população de interesse de tamanho N e desejamos realizar inferências sobre algum parâmetro (média, total, proporção, variância) dessa população, com base em uma amostra de tamanho n.
- Com algumas adaptações, os resultados a serem vistos poderão ser utilizados mesmo se N for infinito.
- 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). Exemplos: peso, altura, intenção de voto, conhecimento em alguma área.

Estrutura geral

- Objetivo: estimar $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ (média, ou proporção se os $y_i's$ foram variáveis binárias), $\tau = \sum_{i=1}^{N} y_i = N\mu$ (total), variância(s) ($\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$, $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$) (esta(s), as vezes, tem de ser estimada(s) para se poder fazer inferência para parâmetros de interesse como a média, total, proporção etc), com base na amostra de tamanho n, sem reposição.
- Amostragem aleatória simples com reposição ($AAS_s \equiv A_2$).
- Amostra $\{y_{k_1}, y_{k_2}, ..., y_{k_n}\}$, em que $k_i \in \{1, 2, ..., N\}$. Por exemplo, Se N = 5 e n = 3, podemos ter $\{y_5, y_2, y_3\}$.

Mecanismo de sorteio da amostra

1. Dado que os elementos da população estão numerados de 1 a N, sorteia-se um elemento, segundo algum procedimento de geração de um número aleatório (função "sample", no R).
2. O elemento selecionado é retirado da população.
3. Repete-se os procedimentos 1 e 2, n-1 vezes.

Estimação da média

- Estimador natural (sob duas formas diferentes):

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{1}$$

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{N} \Delta_i y_i, \tag{2}$$

em que Δ_i é uma v.a. que indica se o elemento i elemento da população apareceu na amostra, e $\boldsymbol{s}$ representa a amostra sorteada, ou seja:

$$\Delta_i = \begin{cases} 1, \text{se o indivíduo } i \text{ foi selecionado} \\ 0, \text{caso contrário} \end{cases}$$

Estimação da média

- Utilizar a forma (1) e considerarmos a distribuição das variáveis $\boldsymbol{Y} = (Y_1, ..., Y_n)^t$ (esses desenvolvimentos dependem da distribuição considerada para a amostra). Neste caso, $Y_i, i = 1, .., n$ são variáveis aleatórias. Esta abordagem é vista nos cursos de Inferência Estatística e pode levar a inferências exatas, desde que as suposições sejam válidas. Note que, neste caso, como não há reposição não temos mais que $Y_i \overset{i.i.d}{\sim} D(\boldsymbol{\theta})$.

Estimação da média

- Utilizar a forma (2) e considerarmos a distribuição das variáveis $\boldsymbol{F} = (F_1, ..., F_N)^t$(mais geral, ou seja, em princípio, os resultados se aplicam, independentemente da distribuição de $\boldsymbol{Y}$). Neste caso, $F_i, i = 1, .., N$ são variáveis aleatórias ($y_i, i = 1, 2, ..., N$ são variáveis não aleatórias). Esta abordagem leva a inferências aproximadas ("n" e "N-n"suficientemente grandes). Uma vantagem é que ela se aplica, em princípio, independentemente da forma da distribuição de $\boldsymbol{Y}$.

Estimação da média

- Resultados:

1. Note que, nesse caso, $\Delta_i \equiv F_i$.
2. $F_i \overset{i.d.}{\sim}$ Bernoulli(n/N). Ou seja, $P(F_i = f_i) = p^{f_i}(1-p)^{1-f_i}\mathbb{1}_{\{0,1\}}(F_i)$, $\mathcal{E}(F_i) = \frac{n}{N}$, $\mathcal{V}(F_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$
3. $\text{Cov}(F_i, F_j) = -\frac{n}{N^2}\frac{N-n}{N-1}$
4. $\pi_i = \frac{n}{N}$ (probabilidade do i-ésimo elemento aparecer na amostra).

   Prova
$$\pi_i = P(F_i = 1) = \frac{n}{N}.$$
5. $\pi_{ij} = \frac{n}{N}\frac{n-1}{N-1}$ (probabilidade do i-ésimo e j-ésimo elementos aparecerem na amostra).

Demonstrações

Sob a mesma probabilidade de seleção de cada indivíduo e usando técnicas de contagem, temos que:

$$
\begin{aligned}
P(F_i = 1) &= \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} \\
&= \frac{\frac{(N-1)!}{(n-1)!}}{\frac{N(N-1)!}{n(n-1)!}} = \frac{n}{N}
\end{aligned}
$$

Demonstrações

Por outro lado, temos que

$$Cov(F_i, F_j) = \mathcal{E}(F_iF_j) - \mathcal{E}(F_i)\mathcal{E}(F_j)$$

mas

$$\begin{aligned}
\mathcal{E}(F_iF_j) &= \sum_{f_i=0}^{1}\sum_{f_j=0}^{1} f_if_jP(F_i = f_i, F_j = f_j) \\
&= P(F_i = 1, F_j = 1) = \frac{\binom{2}{2}\binom{N-2}{n-2}}{\binom{N}{n}} \\
&= \frac{\binom{2}{2}\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n(n-1)(n-2)!}{(n-2)!}\frac{(N-2)!}{N(N-1)(N-2)!} \\
&= \frac{n(n-1)}{N(N-1)}
\end{aligned}$$

Demonstrações

Assim

$$
\begin{aligned}
Cov(F_i, F_j) &= \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N} = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \\
&= \frac{Nn(n-1) - n^2(N-1)}{N^2(N-1)} = \frac{-n(N-n)}{N^2(N-1)}
\end{aligned}
$$

Estimação da média : Propriedades do estimador sob AAS_s

- Valor esperado

$$\mathcal{E}_{A_2}(\widehat{\mu}) = \frac{1}{n}\mathcal{E}_{A_2}\left(\sum_{i=1}^{N} F_i y_i\right) = \frac{1}{n}\sum_{i=1}^{N}\mathcal{E}_{A_2}(F_i)y_i = \frac{1}{n}\sum_{i=1}^{N}\frac{n y_i}{N}$$

$$= \sum_{i=1}^{N}\frac{y_i}{N} = \mu$$

Estimação da média : Propriedades do estimador sob AAS_s

- Variância do estimador

$$
\begin{aligned}
\mathcal{V}_{A_2}(\widehat{\mu}) &= \frac{1}{n^2}\mathcal{V}\left(\sum_{i=1}^{N} F_i y_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 \mathcal{V}_{A_2}(F_i) + \sum_{\substack{i\neq j,\\ i,j=1,2,\ldots,N}} Cov_{A_2}(F_i y_i, F_j y_j)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 \left[\frac{n}{N}\left(1-\frac{n}{N}\right)\right] - \sum_{\substack{i\neq j,\\ i,j=1,2,\ldots,N}} y_i y_j \frac{n(N-n)}{N^2(N-1)}\right) \\
&= \frac{1}{n^2}\left(\frac{n(N-n)}{N^2}\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{\substack{i\neq j,\\ i,j=1,2,\ldots,N}} y_i y_j\right)
\end{aligned}
$$

Estimação da média : Propriedades do estimador sob AAS_s

- Cont.

Lembrando que (Exercício) $\sum_{i \neq j} y_i y_j = -\sum_{i=1}^{N} y_i^2 + N^2\mu^2$ e $\sum_{i=1}^{N}(y_i - \mu)^2 = \sum_{i=1}^{N} y_i^2 - N\mu^2$ (e denotando $f = \frac{n}{N}$), vem que

$$
\begin{aligned}
\mathcal{V}_{A_2}(\widehat{\mu}) &= \frac{1}{n^2}\left(\frac{n(N-n)}{N^2}\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\left(N^2\mu^2 - \sum_{i=1}^{N} y_i^2\right)\right) \\
&= \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N(N-1)}\left(N\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{n} y_i^2 - N^2\mu^2 + \sum_{i=1}^{N} y_i^2\right) \\
&= \frac{1}{n}(1-f)\frac{1}{N(N-1)}\left(N\sum_{i=1}^{N} y_i^2 - N^2\mu^2\right) \\
&= \left(\frac{1-f}{n}\right)\frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - N\mu^2\right) = (1-f)\frac{s^2}{n}
\end{aligned}
$$

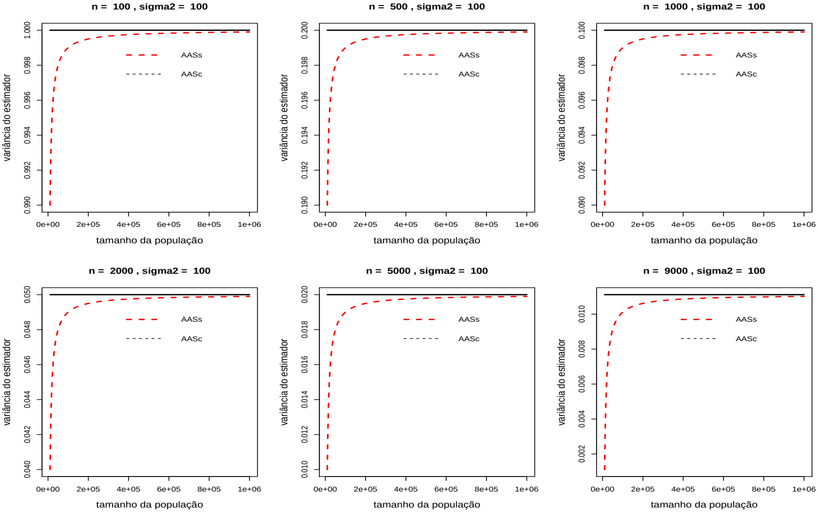
Estimação da média : Propriedades do estimador sob AAS_s

- O estimador para a média sob AAS_c ou AAS_s, é o mesmo.
- Temos que $\mathcal{E}_{A_i}(\widehat{\mu}) = \mu$, $i = 1, 2$.
- Além disso, $\mathcal{V}_{A_1}(\widehat{\mu}) = \frac{\sigma^2}{n}$ e $\mathcal{V}_{A_2}(\widehat{\mu}) = (1 - f)\frac{s^2}{n}$, em que $f \in (0, 1)$.
- Portanto, o efeito do planejamento (EPA), do estimador sob o plano A_2 em relação ao plano A_1, é dado por:

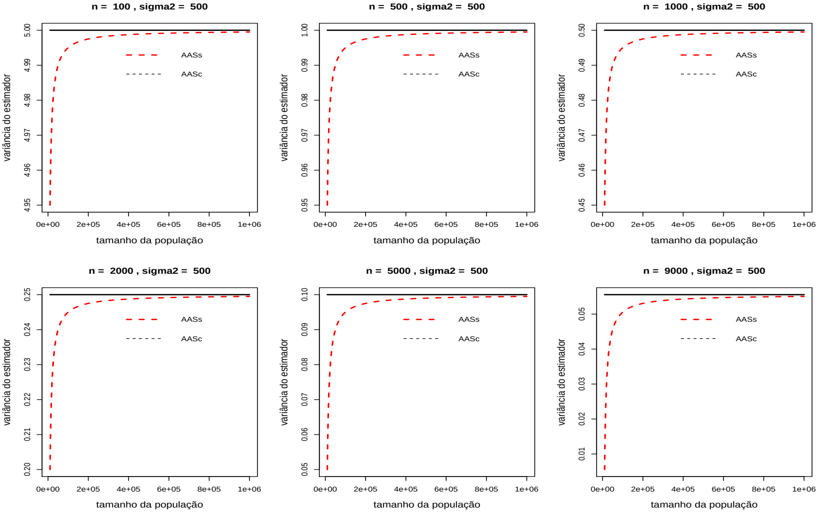
$$EPA = \frac{\mathcal{V}_{A_2}(\widehat{\mu})}{\mathcal{V}_{A_1}(\widehat{\mu})} \approx (1 - f)$$

- Ademais, quando $N \to \infty$, $\mathcal{V}_{A_2}(\widehat{\mu}) \to \mathcal{V}_{A_1}(\widehat{\mu})$.
- Consequentemente, temos que o plano AAS_s é melhor do que AAS_c, tendendo ambos a serem equivalentes, à medida que o tamanho da população tende a infinito.

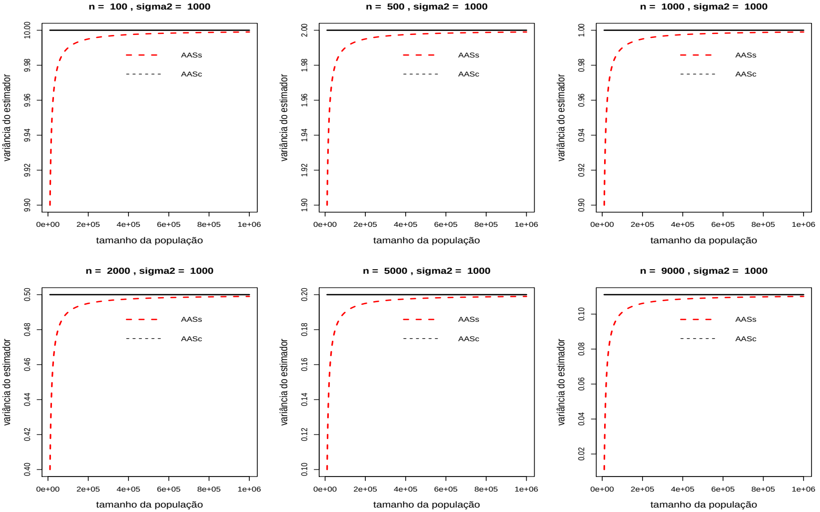
Comparação dos planos amostrais



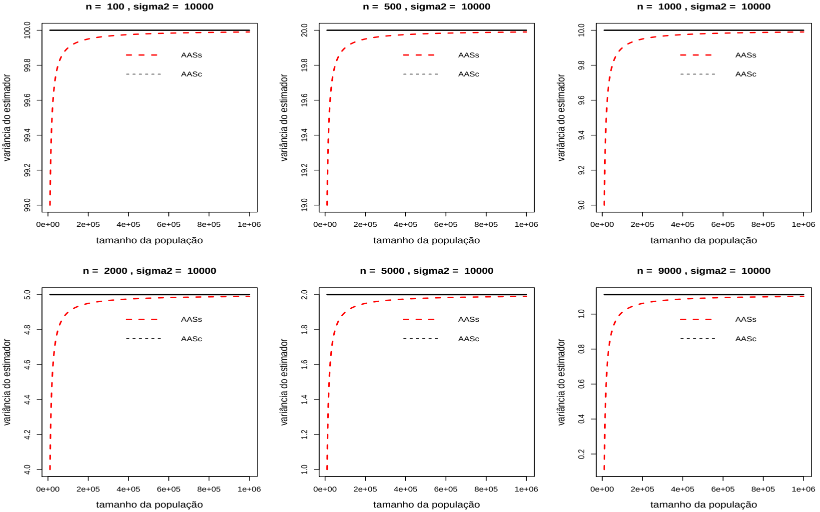
Comparação dos planos amostrais



Comparação dos planos amostrais



Comparação dos planos amostrais



Estimação da média : Propriedades do estimador sob AAS_s

- Resumidamente, $\mathcal{E}_{A_2}(\widehat{\mu}) = \mu$ e $\mathcal{V}_{A_2}(\widehat{\mu}) = (1-f)\frac{s^2}{n}$. Podemos provar, sob AAS_s, que $\widehat{\mu}$ é consistente.
- A distribuição exata é bastante complicada de ser obtida (média de uma combinação linear de um vetor aleatório com distribuição multinomial).
- Distribuição assintótica: note que em $\{F_i\}_{i\geq 1}$ os F_i's são identicamente distribuídos mas não independentes. O TLC padrão não se aplica.
- Estimativa $\widetilde{\mu} = \frac{1}{n}\sum_{i\in \boldsymbol{s}} y_i = \frac{1}{n}\sum_{i=1}^{N} f_i y_i$.

Estimação da média : Propriedades do estimador sob AAS_s

- Discutiremos, com mais detalhes (mais a frente), como se obter os resultados assintóticos mas, por enquanto, sob certas condições, entre elas, n e N-n suficientemente grandes, temos que

$$\frac{\widehat{\mu} - \mu}{\sqrt{(1-f)s^2/n}} \xrightarrow[\substack{n \to \infty, \\ N-n \to \infty}]{D} N(0,1) \tag{4}$$

ou

$$\widehat{\mu} \approx N(\mu, (1-f)s^2/n), \text{para n e N-n suficientemente grandes.}$$

Problema: σ^2, quase sempre, é desconhecido. Faz-se necessário considerar um estimador consistente (de preferência não viciado), para se poder usar o Teorema de Slutsky .

Estimação da média : Propriedades do estimador sob AAS_s

- Vamos considerar o seguinte estimador

$\widehat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \widehat{\mu}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{N} F_i\left(y_i - \widehat{\mu}\right)^2$.

Note que (lembrando que $\sum_{i=1}^{N} y_i^2 = (N-1)s^2 + N\mu^2$)

$$
\begin{aligned}
\mathcal{E}_{A_2}(\widehat{s}^2) &= \frac{1}{n-1}\mathcal{E}_{A_2}\left(\sum_{i=1}^{n} Y_i^2 - n\widehat{\mu}^2\right) \\
&= \frac{1}{n-1}\left[\mathcal{E}_{A_2}\left(\sum_{i=1}^{N} y_i^2 F_i\right) - n\mathcal{E}_{A_2}(\widehat{\mu}^2)\right]
\end{aligned}
$$

Estimação da média : Propriedades do estimador sob AAS_s

$$
\begin{aligned}
\mathcal{E}_{A_2}(\widehat{s}^2) &= \frac{1}{n-1}\left[\sum_{i=1}^{N} y_i^2 \mathcal{E}_{A_2}(F_i) - n\mathcal{E}_{A_2}\left(\widehat{\mu}^2\right)\right] \\
&= \frac{1}{n-1}\left[((N-1)s^2 + N\mu^2)\frac{n}{N} - n\left[\left(1-\frac{n}{N}\right)\frac{s^2}{n} + \mu^2\right]\right] \\
&= \frac{1}{n-1}\left[ns^2\left(1-\frac{1}{N}\right) + n\mu^2 - s^2 + s^2\frac{n}{N} - n\mu^2\right] \\
&= \frac{1}{n-1}\left(ns^2 - s^2 - \frac{ns^2}{N} + \frac{s^2 n}{N}\right) = s^2
\end{aligned}
$$

Estimação da média : Propriedades do estimador sob AAS_s

- A prova de sua consistência, i.e.,

$$\widehat{s}^2 \xrightarrow[\substack{n\to\infty,\\ N-n\to\infty}]{P} s^2 \tag{5}$$

também será discutida mais à frente.

- Portanto, dos resultados (4) e (5), temos que

$$\frac{\widehat{\mu}-\mu}{\sqrt{(1-f)\widehat{s}^2/n}} = \frac{\widehat{\mu}-\mu}{\sqrt{(1-f)s^2/n}}\frac{s}{\widehat{s}} \xrightarrow[\substack{n\to\infty,\\ N-n\to\infty}]{D} N(0,1)$$

por Slutsky.

Intervalo de Confiança

- Estimativa: $\widetilde{s}^2 = \frac{1}{n-1}\sum_{i \in \boldsymbol{s}} (y_i - \widetilde{\mu})^2$.
- Assim, um intervalo de confiança (assintótico) com coeficiente de confiança de aproximadamente γ é dado por

$$IC(\mu, \gamma) \approx \left[\widehat{\mu} - z_\gamma \sqrt{\frac{(1-f)\widehat{s}^2}{n}}; \widehat{\mu} + z_\gamma \sqrt{\frac{(1-f)\widehat{s}^2}{n}}\right]$$

em que $P(Z \leq z_\gamma) = \frac{1+\gamma}{2}$ e $Z \sim N(0,1)$.
- Erro da estimativa: $z_\gamma \sqrt{(1-f)\frac{\widehat{s}^2}{n}}$.

Testes de Hipótese

- Hipóteses usuais (μ_0 conhecido)
  1. $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$.
  2. $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$.
  3. $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.
- Estatística do teste $Z_t = \frac{\widehat{\mu}-\mu_0}{\widehat{s}\sqrt{(1-f)}/\sqrt{n}}$, em que $\widehat{s} = \sqrt{\widehat{s^2}}$.
- Sob H_0, vimos que $Z_t \approx N(0,1)$, para n e N-n suficientemente grandes.
- Defina $z_t = \frac{\widetilde{\mu}-\mu_0}{\widetilde{s}\sqrt{(1-f)}/\sqrt{n}}$ o valor calculado da estatística do teste e z_c o(s) valor(es) crítico(s).
- Defina ainda $Z \sim N(0,1)$.

Testes de Hipótese

- Procedimento para testar o conjunto de hipóteses [1]
- Valor crítico
  - $P(Z \leq z_c | H_0) = \alpha.$
  - Se $z_t \leq z_c$ rejeita-se H_0, caso contrário, não se rejeita.
- p-valor (nível descritivo)
  - $p - valor = P(Z \leq z_t | H_0)$

Testes de Hipótese

- Procedimento para testar o conjunto de hipóteses [2]
- Valor crítico
  - $P(Z \geq z_c | H_0) = \alpha$.
  - Se $z_t \geq z_c$ rejeita-se H_0, caso contrário, não se rejeita.
- p-valor (nível descritivo)
  - $p - valor = P(Z \geq z_t | H_0)$

Testes de Hipótese

- Procedimento para testar o conjunto de hipóteses [3]
- Valor crítico
  - $P(Z \leq z_c | H_0) = \frac{1+\alpha}{2}$.
  - Se $|z_t| \geq z_c$ rejeita-se H_0, caso contrário, não se rejeita.
- p-valor (nível descritivo)
  - $p - valor = 2[1 - P(Z \leq |z_t| | H_0)]$.

Estudos de simulação

- Distribuição assintótica do estimador para a média. Tamanho da população $N = 100000$.
- Cinco cenários, variando em função da variável de interesse na população (X).
  - $X \sim N(800, 10000)$
  - $X \sim \text{gama}(5; 0,00625)$, $E(X) = 800, V(X) = 128000$.
  - $X \sim t_{(7)}(800, 5000)$, $E(X) = 800$, $V(X) = 7000$.
  - $X \sim U[400; 1200]$.
  - $X \sim 0.5N(200, 5000) + 0.5N(600, 5000)$

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 $R = 1000$ réplicas (amostras selecionadas da população de interesse).

Procedimento para se gerar o gráfico de envelopes (quantil-quantil)

1) Simule n variáveis aleatórias independentes de interesse $(V_{ij} \overset{i.i.d.}{\sim} N(0,1))$. Repita este processo m vezes.
2) Ao final teremos uma matriz com valores simulados dessas variáveis aleatórias, digamos V_{ij}, i=1,...,n, (tamanho da amostra) j=1,...,m (réplica).

$$\boldsymbol{V} = \begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1m} \\ v_{21} & v_{22} & \ldots & v_{v2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \ldots & v_{nm} \end{bmatrix}$$

Cont.

3) Dentro de cada amostra, ordena-se, de modo crescente, os valores simulados, obtendo-se $v^*_{(i)j}$ (estatísticas de ordem):

$$\boldsymbol{V}^* = \begin{bmatrix} v_{(1)1} & v_{(1)2} & \cdots & v_{(1)m} \\ v_{(2)1} & v_{(2)2} & \cdots & v_{(2)m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{(n)1} & v_{(n)2} & \cdots & v_{(n)m} \end{bmatrix}$$

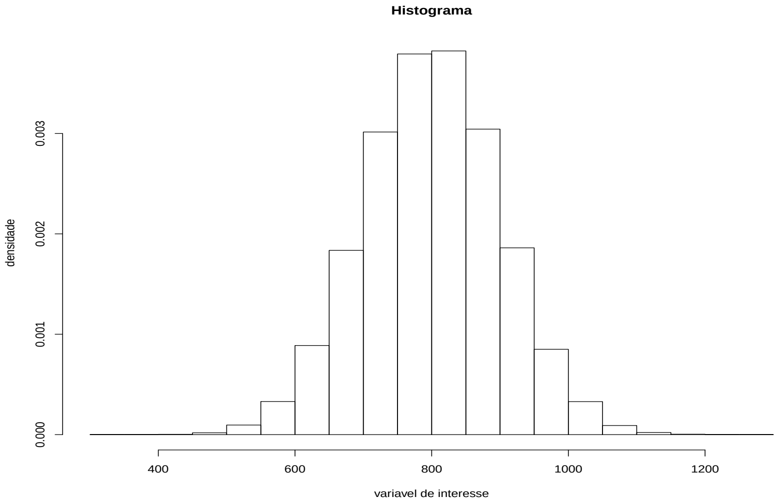
4) Pode-se obter os limites $v_{(i)I} = \min_{1 \leq j \leq m} v_{(i)j}$ e $v_{(i)S} = \max_{1 \leq j \leq m} v_{(i)j}$, $i = 1, 2, ..., n$.

Cont.

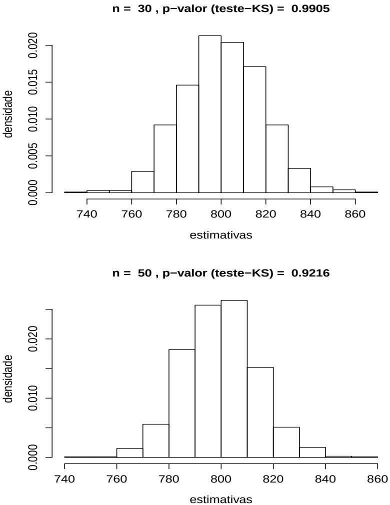
5) Porém, na prática considera-se $v_{(i)I} = \frac{v_{(i)(2)}+v_{(i)(3)}}{2}$ e $v_{(i)S} = \frac{v_{(i)(m-2)}+v_{i(m-1)}}{2}$ (para se gerar limites de confiança), em que $v_{(i)(r)}$ é a r-ésima estatística de ordem dentro de cada linha, $i = 1, 2, ..., n$.

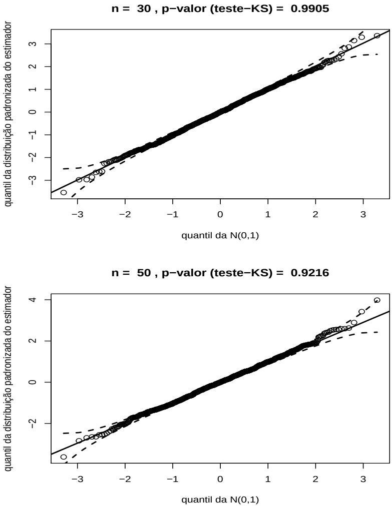
- Além disso, consideramos como a linha de referência $v_{(i)} = \frac{1}{m}\sum_{j=1}^{m} v_{(i)j}, i = 1, 2, ..., n$.

normal

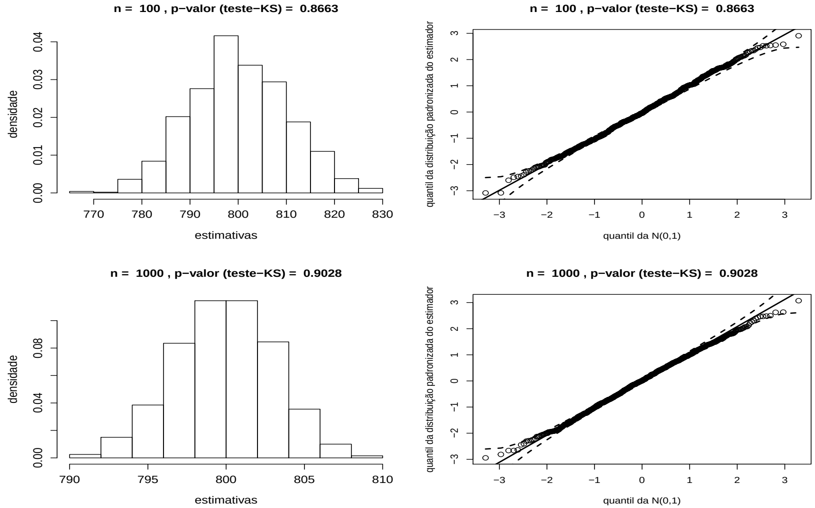

Histograma
densidade
0.000
0.001
0.002
0.003
400
600
800
1000
1200
variavel de interesse

normal

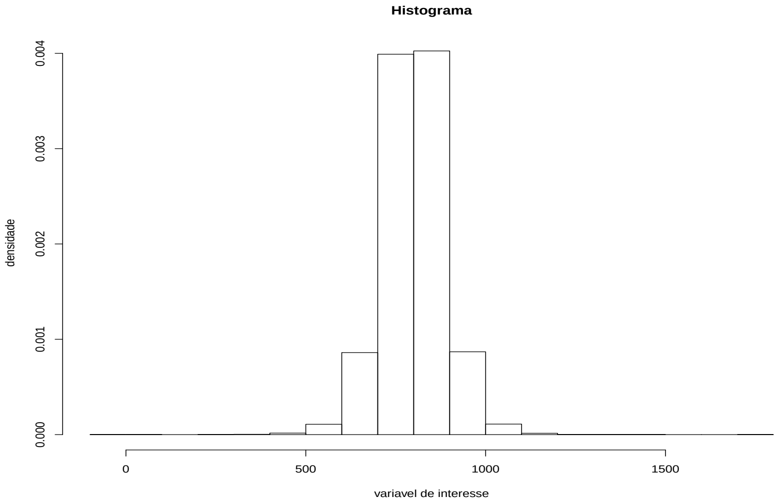

n = 30 , p−valor (teste−KS) = 0.9905
densidade
estimativas
n = 50 , p−valor (teste−KS) = 0.9216
densidade
estimativas


n = 30 , p−valor (teste−KS) = 0.9905
quantil da distribuição padronizada do estimador
quantil da N(0,1)
n = 50 , p−valor (teste−KS) = 0.9216
quantil da distribuição padronizada do estimador
quantil da N(0,1)

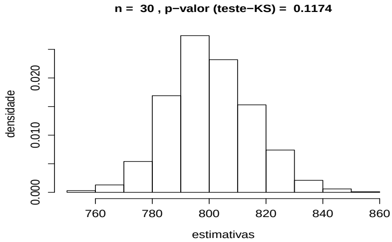
normal

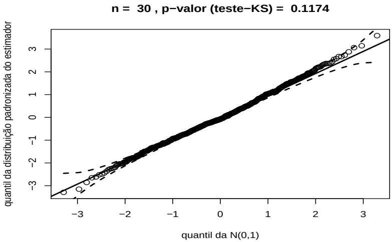

n = 100 , p−valor (teste−KS) = 0.8663
densidade
0.00
0.01
0.02
0.03
0.04
770
780
790
800
810
820
830
estimativas
n = 100 , p−valor (teste−KS) = 0.8663
quantil da distribuição padronizada do estimador
quantil da N(0,1)
−3
−2
−1
0
1
2
3
n = 1000 , p−valor (teste−KS) = 0.9028
densidade
0.00
0.04
0.08
790
795
800
805
810
estimativas
n = 1000 , p−valor (teste−KS) = 0.9028
quantil da distribuição padronizada do estimador
quantil da N(0,1)

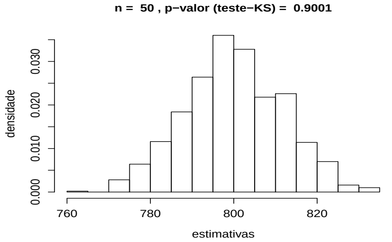
t de Student

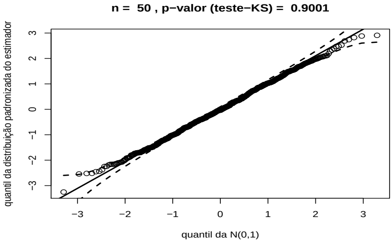

Histograma
densidade
0.000
0.001
0.002
0.003
0.004
0
500
1000
1500
variavel de interesse

t de Student

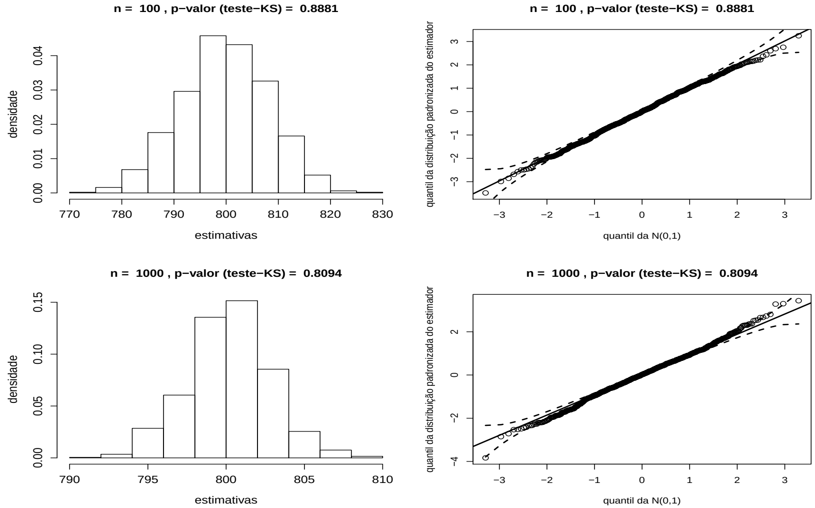

n = 30 , p−valor (teste−KS) = 0.1174
densidade
0.000
0.010
0.020
760
780
800
820
840
860
estimativas


n = 30 , p−valor (teste−KS) = 0.1174
quantil da distribuição padronizada do estimador
−3
−2
−1
0
1
2
3
quantil da N(0,1)

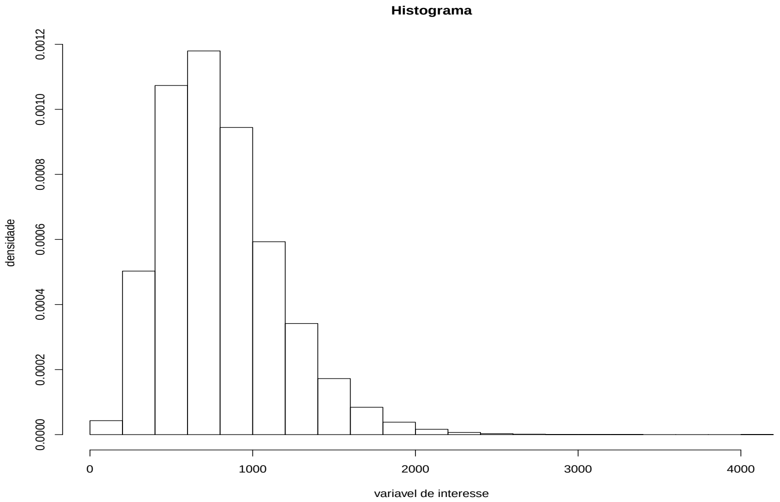

n = 50 , p−valor (teste−KS) = 0.9001
densidade
0.000
0.010
0.020
0.030
760
780
800
820
estimativas


n = 50 , p−valor (teste−KS) = 0.9001
quantil da distribuição padronizada do estimador
−3
−2
−1
0
1
2
3
quantil da N(0,1)

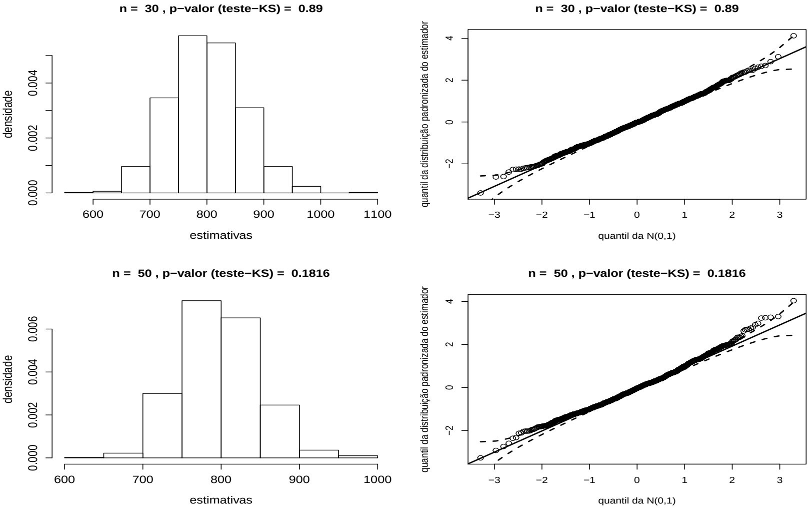
t de Student


n = 100 , p−valor (teste−KS) = 0.8881
densidade
estimativas
n = 100 , p−valor (teste−KS) = 0.8881
quantil da distribuição padronizada do estimador
quantil da N(0,1)
n = 1000 , p−valor (teste−KS) = 0.8094
densidade
estimativas
n = 1000 , p−valor (teste−KS) = 0.8094
quantil da distribuição padronizada do estimador
quantil da N(0,1)

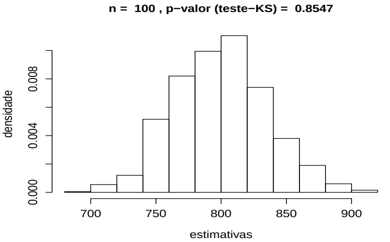
gama

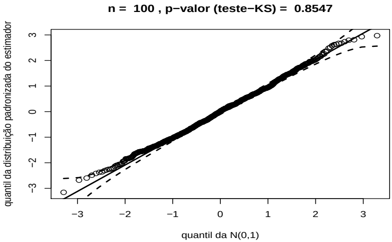

Histograma
densidade
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0
1000
2000
3000
4000
variavel de interesse

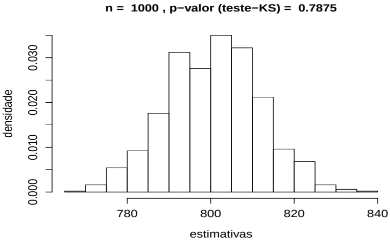
gama

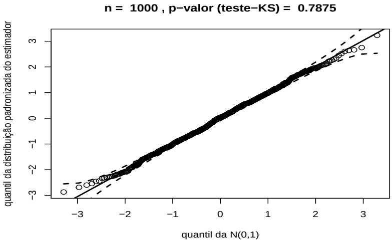

n = 30 , p−valor (teste−KS) = 0.89
densidade
0.000
0.002
0.004
600
700
800
900
1000
1100
estimativas
n = 30 , p−valor (teste−KS) = 0.89
quantil da distribuição padronizada do estimador
−2
0
2
4
−3
−2
−1
0
1
2
3
quantil da N(0,1)
n = 50 , p−valor (teste−KS) = 0.1816
densidade
0.000
0.002
0.004
0.006
600
700
800
900
1000
estimativas
n = 50 , p−valor (teste−KS) = 0.1816
quantil da distribuição padronizada do estimador
−2
0
2
4
−3
−2
−1
0
1
2
3
quantil da N(0,1)

gama

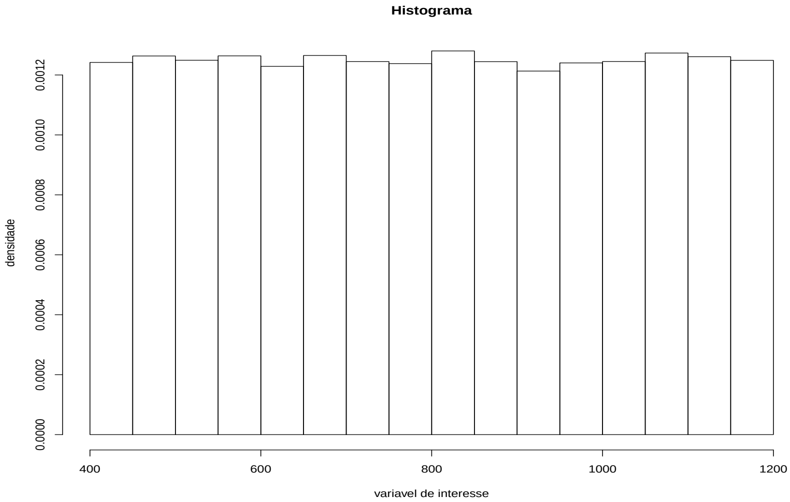

n = 100 , p−valor (teste−KS) = 0.8547
densidade
0.000
0.004
0.008
700
750
800
850
900
estimativas


n = 100 , p−valor (teste−KS) = 0.8547
quantil da distribuição padronizada do estimador
quantil da N(0,1)
−3
−2
−1
0
1
2
3

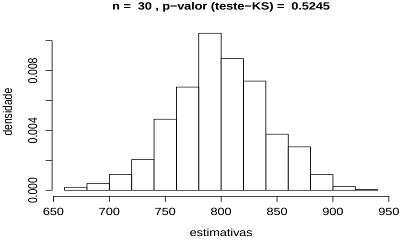

n = 1000 , p−valor (teste−KS) = 0.7875
densidade
0.000
0.010
0.020
0.030
780
800
820
840
estimativas

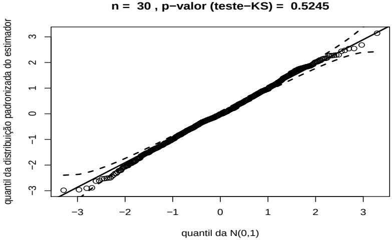

n = 1000 , p−valor (teste−KS) = 0.7875
quantil da distribuição padronizada do estimador
quantil da N(0,1)
−3
−2
−1
0
1
2
3

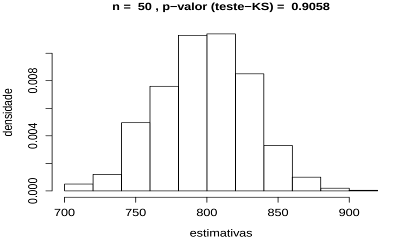
uniforme

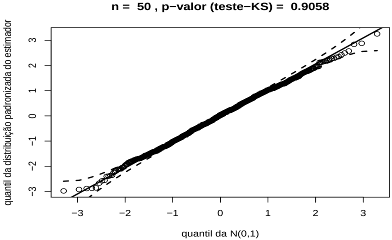

Histograma
densidade
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
400
600
800
1000
1200
variavel de interesse

uniforme

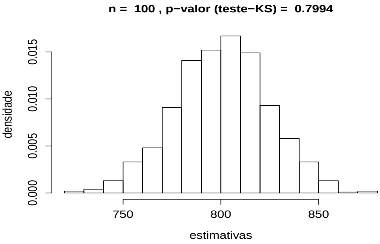

n = 30 , p−valor (teste−KS) = 0.5245
densidade
0.000
0.004
0.008
650
700
750
800
850
900
950
estimativas

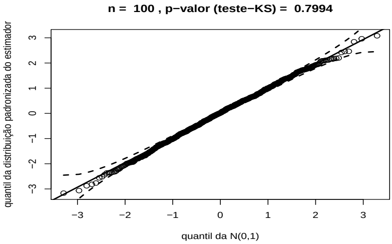

n = 30 , p−valor (teste−KS) = 0.5245
quantil da distribuição padronizada do estimador
−3
−2
−1
0
1
2
3
quantil da N(0,1)

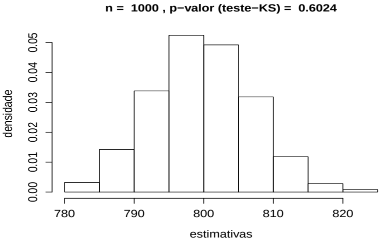

n = 50 , p−valor (teste−KS) = 0.9058
densidade
0.000
0.004
0.008
700
750
800
850
900
estimativas

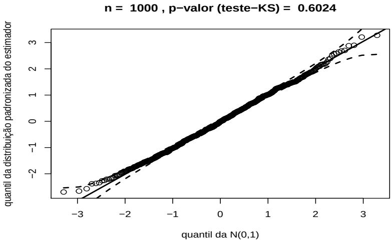

n = 50 , p−valor (teste−KS) = 0.9058
quantil da distribuição padronizada do estimador
−3
−2
−1
0
1
2
3
quantil da N(0,1)

uniforme

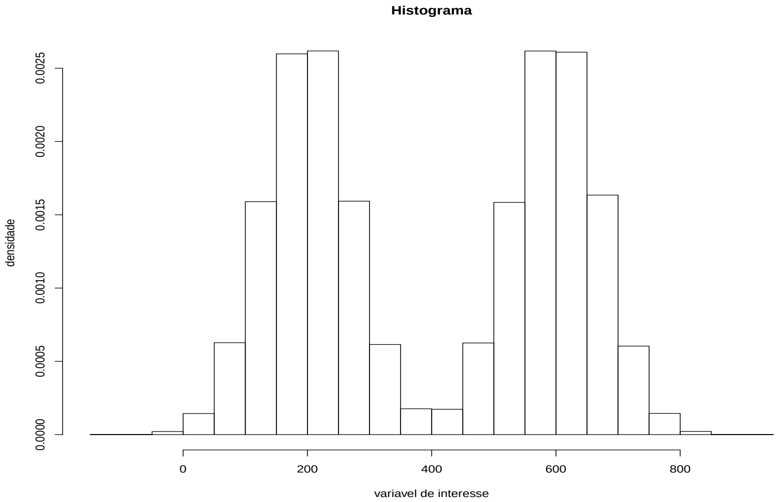

n = 100 , p−valor (teste−KS) = 0.7994
densidade
0.000
0.005
0.010
0.015
750
800
850
estimativas


n = 100 , p−valor (teste−KS) = 0.7994
quantil da distribuição padronizada do estimador
quantil da N(0,1)

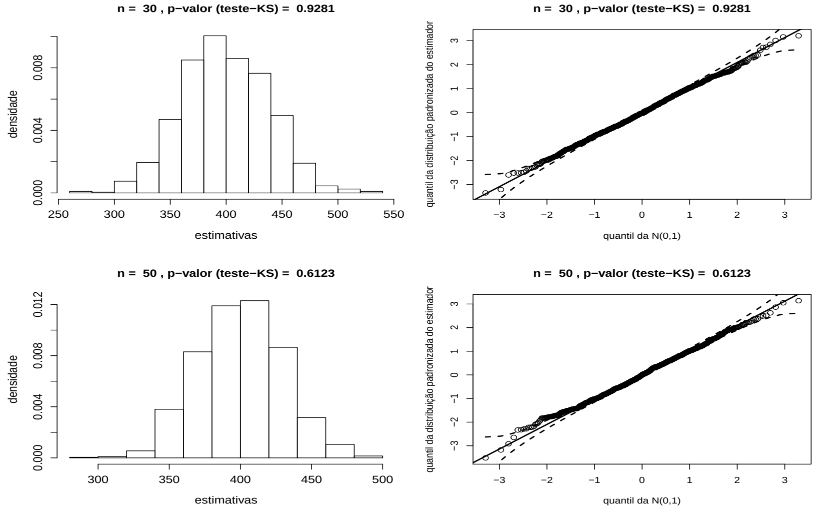

n = 1000 , p−valor (teste−KS) = 0.6024
densidade
0.00
0.01
0.02
0.03
0.04
0.05
780
790
800
810
820
estimativas


n = 1000 , p−valor (teste−KS) = 0.6024
quantil da distribuição padronizada do estimador
quantil da N(0,1)

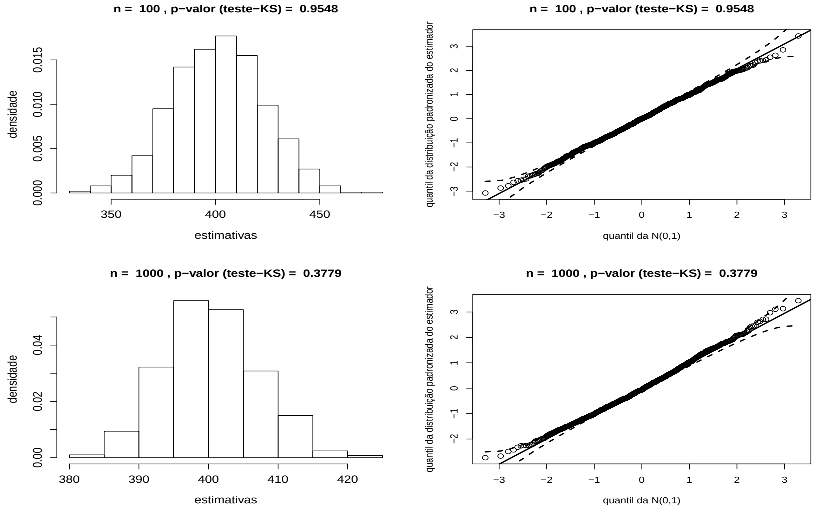
mistura de duas normais



mistura de duas normais



mistura de duas normais



Determinação do tamanho amostral

- Estabelece-se algum critério de interesse acerca da acurácia/precisão na estimativa da média populacional.
- Sob o estimador proposto, calcula-se o tamanho da amostra, com base em sua distribuição assintótica e critério estabelecido.
- Erro de estimativa: $z_\gamma \sqrt{\frac{(1-f)\widehat{s}^2}{n}}$. Fixa-se um erro de estimativa de interesse.
- Probabilidade do módulo da diferença $P\left(|\widehat{\mu} - \mu| < \delta\right) > \gamma$, $\delta > 0$, $\gamma \in (0,1)$.

Determinação do tamanho amostral: erro da estimativa

$$
\begin{aligned}
\delta &= z_{\gamma}\sqrt{\frac{(1-f)s^2}{n}} \rightarrow \left(\frac{1}{n}-\frac{1}{N}\right) = \frac{\delta^2}{z_{\gamma}^2 s^2} \rightarrow \frac{1}{n} = \frac{\delta^2}{z_{\gamma}^2 s^2} + \frac{1}{N} \\
\rightarrow & \quad \frac{1}{n} = \frac{\delta^2 N + z_{\gamma}^2 s^2}{N z_{\gamma}^2 s^2} \rightarrow n = \frac{N z_{\gamma}^2 s^2}{\delta^2 N + z_{\gamma}^2 s^2} = \frac{1}{\frac{\delta^2}{s^2 z_{\gamma}^2} + \frac{1}{N}}
\end{aligned}
$$

Em geral, o (um) valor de s^2 é obtido através de pesquisas anteriores ou de uma amostra piloto, de tamanho apropriado.

Determinação do tamanho amostral: precisão

$$
\begin{aligned}
P_{A_1}\left(|\widehat{\mu}-\mu|<\delta\right)>\gamma \quad &\leftrightarrow \quad P_{A_1}\left(\left|\frac{\widehat{\mu}-\mu}{\sqrt{(1-f)\sigma^2/n}}\right|<\frac{\sqrt{n}\delta}{\sigma}\right)>\gamma \\
&\leftrightarrow \quad P_{A_1}\left(|Z|<\frac{\sqrt{n}\delta}{\sqrt{1-f}\sigma}\right)>\gamma \leftrightarrow \frac{\sqrt{n}\delta}{\sqrt{1-f}\sigma}=z_\gamma
\end{aligned}
$$

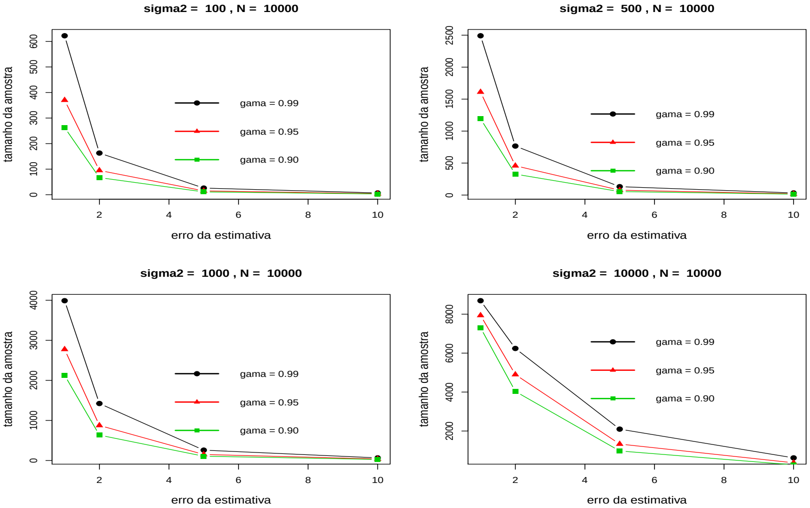
em que $Z \approx N(0,1)$. O que leva ao mesmo procedimento oriundo de se fixar o erro da estimativa.

Os tamanhos amostrais, sob os planos amostrais A_1 e A_2 são dados, respectivamente, por $n_{A_1} = \frac{1}{\frac{\delta^2}{z_\gamma^2\sigma^2}}$ e $n_{A_2} = \frac{1}{\frac{\delta^2}{s^2 z_\gamma^2}+\frac{1}{N}}$. Assim, nota-se que $n_{A_1} \geq n_{A_2}$.

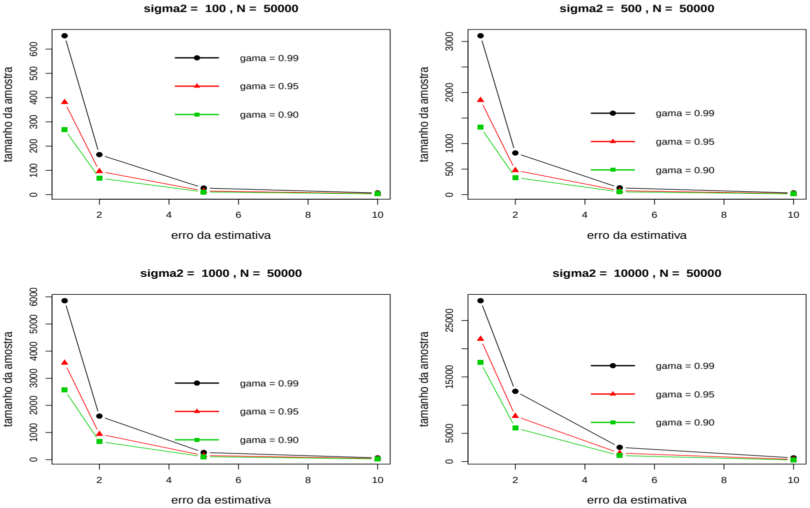
Tamanhos amostrais

- Situções hipotéticas: cruzamento entre os níveis de diferentes fatores de interesse
  - $\delta \in \{1, 2, 5, 10\}$.
  - $\gamma \in \{0,9; 0,95; 0,99\}$.
  - $\sigma^2 \in \{100, 500, 1000, 10000\}$.

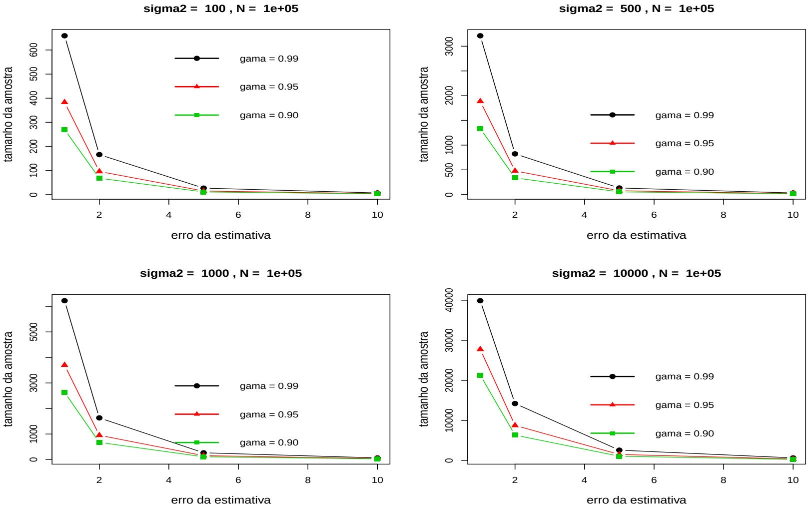
Tamanhos amostrais



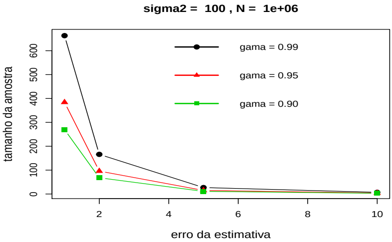
Tamanhos amostrais

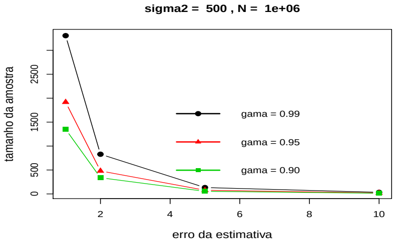


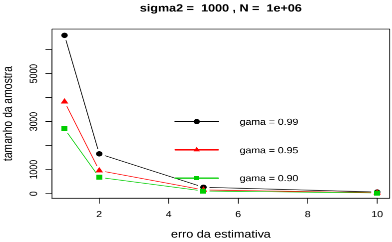
Tamanhos amostrais

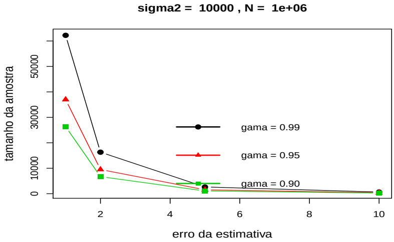


Tamanhos amostrais

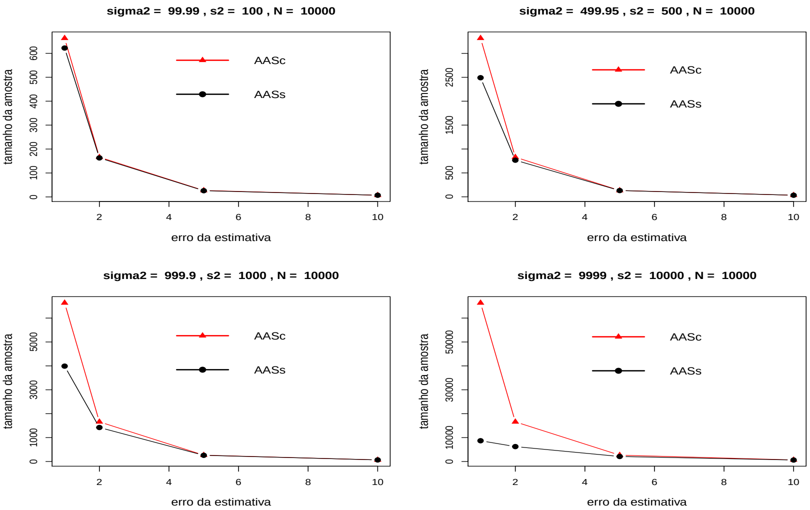




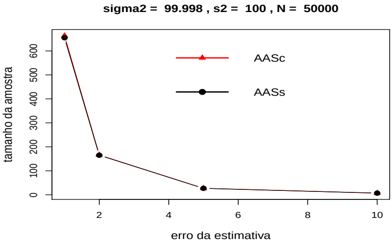


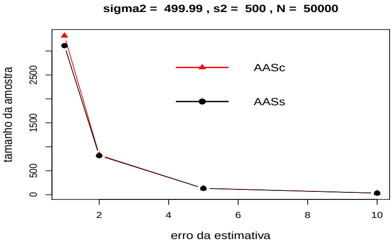


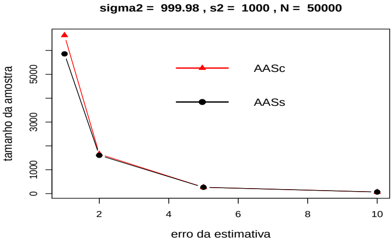
Comparação dos tamanhos amostrais ($\gamma = 0,99$)

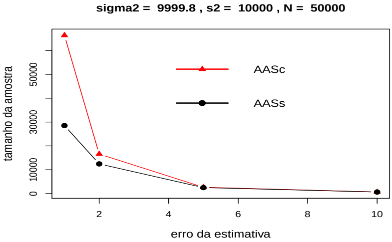


Comparação dos tamanhos amostrais ($\gamma = 0,99$)

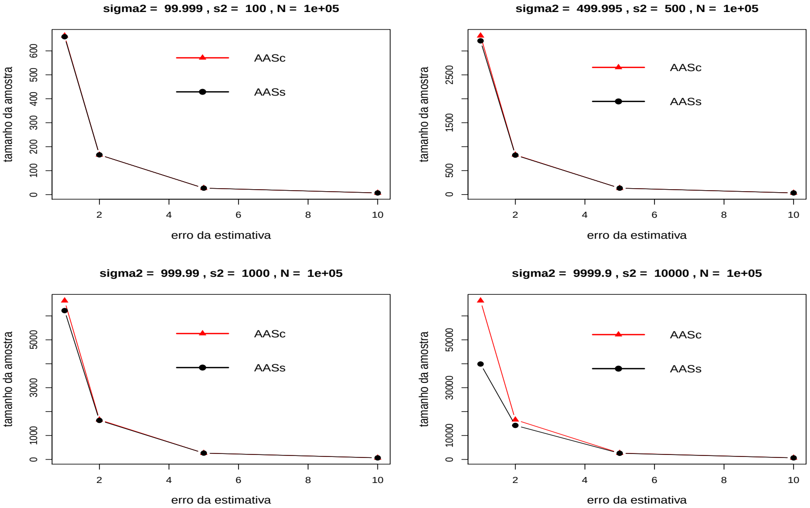




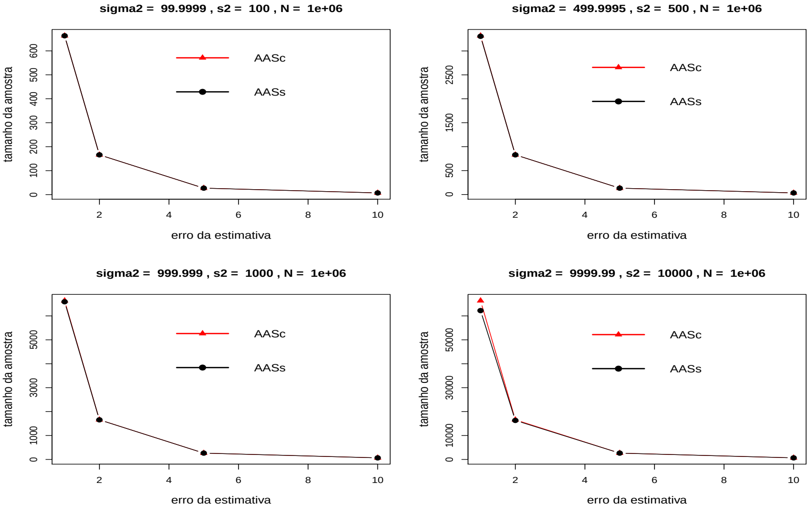




Comparação dos tamanhos amostrais ($\gamma = 0,99$)



Comparação dos tamanhos amostrais ($\gamma = 0,99$)



Estimação do total populacional

- $\tau = \sum_{i=1}^{N} y_i = N\mu$.
- Estimador "natural": $\widehat{\tau}_u = \sum_{i=1}^{n} Y_i$. Problema: se os y_i's foram positivos, $\widehat{\tau}_u$ sempre subestimará τ.
- Alternativa $\widehat{\tau} = N\widehat{\mu}$.
- Estimativa $\widetilde{\tau} = N\widetilde{\mu}$

Propriedades do estimador

- $\mathcal{E}_{A_2}(\widehat{\tau}) = \mathcal{E}_{A_2}(N\widehat{\mu}) = N\mathcal{E}(\widehat{\mu}) = N\mu = \tau$ (não viciado).
- $\mathcal{V}_{A_2}(\widehat{\tau}) = N^2\mathcal{V}_{A_2}(\widehat{\mu}) = N^2(1-f)\frac{s^2}{n}$ (a imprecisão associada à estimação do total é maior do que aquela associada à média).

Propriedades do estimador

- Normalidade assintótica, como

$$\frac{\widehat{\mu} - \mu}{\sqrt{(1-f)\widehat{s}^2/n}} \xrightarrow[\substack{n \to \infty, \\ N-n \to \infty}]{D} N(0,1),$$

lembrando que N é fixo, temos que

$$\frac{N\widehat{\mu} - N\mu}{\sqrt{N^2(1-f)\widehat{s}^2/n}} \xrightarrow[\substack{n \to \infty, \\ N-n \to \infty}]{D} N(0,1) \rightarrow \frac{\widehat{\tau} - \tau}{\sqrt{(1-f)N^2\widehat{s}^2/n}} \xrightarrow[\substack{n \to \infty, \\ N-n \to \infty}]{D} N(0,1)$$

Intervalo de Confiança

- Assim, um intervalo de confiança (assintótico) com coeficiente de confiança de aproximadamente γ é dado por

$$IC(\mu,\gamma) \approx \left[\widehat{\mu} - z_{\gamma} N\sqrt{(1-f)\frac{\widehat{s}^2}{n}}; \widehat{\mu} + z_{\gamma} N\sqrt{(1-f)\frac{\widehat{s}^2}{n}}\right]$$

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

- Erro da estimativa: $z_{\gamma} N\sqrt{(1-f)\frac{\widehat{s}^2}{n}}$.

Testes de Hipótese

- Hipóteses usuais (τ_0 conhecido)
  1. $H_0 : \tau = \tau_0$ vs $H_1 : \tau < \tau_0$.
  2. $H_0 : \tau = \tau_0$ vs $H_0 : \tau > \tau_0$.
  3. $H_0 : \tau = \tau_0$ vs $H_0 : \tau \neq \tau_0$.
- Estatística do teste $Z_t = \frac{\widehat{\tau}-\tau_0}{N\sqrt{(1-f)}\widehat{s}/\sqrt{n}}$, em que $\widehat{s} = \sqrt{\widehat{s}^2}$.
- Sob H_0, vimos que $Z_t \approx N(0,1)$, para n e N-n suficientemente grandes.
- Defina $z_t = \frac{\widetilde{\tau}-\tau_0}{N\sqrt{(1-f)}\widetilde{s}/\sqrt{n}}$ 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

$$
\begin{aligned}
\delta &= z_{\gamma}\sqrt{\frac{(1-f)s^2N^2}{n}} \rightarrow \left(\frac{1}{n}-\frac{1}{N}\right) = \frac{\delta^2}{z_{\gamma}^2 s^2 N^2} \rightarrow \frac{1}{n} = \frac{\delta^2}{z_{\gamma}^2 s^2 N^2} + \frac{1}{N} \\
&\rightarrow \frac{1}{n} = \frac{\delta^2 + z_{\gamma}^2 s^2 N}{z_{\gamma}^2 s^2 N^2} \rightarrow n = \frac{z_{\gamma}^2 s^2 N^2}{\delta^2 + z_{\gamma}^2 s^2 N} = \frac{1}{\frac{\delta^2}{N^2 s^2 z_{\gamma}^2} + \frac{1}{N}}
\end{aligned}
$$

Em geral, o (um) valor de s^2 é obtido através de pesquisas anteriores ou de uma amostra piloto, de tamanho apropriado. Isto vale para qualquer um dos dois critérios: erro da estimativa e precisão.