Amostragem aleatória simples com 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 $\boldsymbol{y}_i = (y_{i1}, ..., y_{ip})^t$). 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 ($\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$) (esta, as vezes, tem de ser estimada para se poder fazer inferência para parâmetros de interesse como a média, total, proporção etc), com base na amostra de tamanho n, com reposição.
- Amostragem aleatória simples com reposição ($AAS_c \equiv A_1$).
- 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 $\{y_2, y_3, y_3\}$.

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 números aleatórios (link 1, link 2).
2. Repõe-se esse elemento na população.
3. Repete-se os procedimentos 1 e 2, $n-1$ vezes.
4. No R: ee
   - Função "sample".
   - Pacote "sampling".
   - Pacote "survey".

Notações/exemplo (univariado)

- População:
  - "Labels" - $\mathcal{U} = \{1, ..., N\}$.
  - Variável (não aleatória) - $\boldsymbol{y} = (y_1, ..., y_N)^t$ (valores da característica de interesse para cada elemento na população).
- Amostra
  - "Labels" - $\{1, ..., n\}$.
  - Índices dos elementos a serem selecionados: $\boldsymbol{S} = (K_1, ..., K_n)$ (variável aleatória) e $\boldsymbol{s} = (k_1, ..., k_n)$ (os respectivos valores observados, índices sorteados).
  - Variável (aleatória) - $\boldsymbol{Y} = (Y_1, ..., Y_n)^t$ (valores da característica de interesse para cada elemento sorteado).
- Exemplo: Suponha N=3, $\boldsymbol{y} = (y_1, y_2, y_3)^t$, n = 2 e que $\boldsymbol{s} = (2, 3)$. Assim, $k_1 = 2, k_2 = 3$, $Y_1 = y_2$ e $Y_2 = y_3$.

Estimação da média

- Estimador natural (replicar na amostra a fórmula do parâmetro de interesse) (sob duas formas diferentes):

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i \in \boldsymbol{S}} y_i \quad (1)$$

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{N} F_i y_i, \quad (2)$$

em que, na Equação (1) $Y_i, i = 1, .., n$ e $\boldsymbol{S} = (K_1, ..., K_n)$ são variáveis (vetores) aleatórios, enquanto que na Equação (2) F_i é uma v.a. que representa o número de vezes que o i-ésimo elemento da população apareceu na amostra.

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 (usuais) de Inferência Estatística e pode levar a inferências exatas, desde que as suposições sobre a distribuição de $\boldsymbol{Y}$ sejam válidas.
- Note que, neste caso, como há reposição $Y_i \overset{i.i.d}{\sim} D(\boldsymbol{\theta})$.
- Utilizar a distribuição de $\boldsymbol{S}$ pode ser bem complicado.

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 (link 1, link 2, link 3) (i.d. - identicamente distribuídas):
  1. $F_i \overset{i.d.}{\sim}$ binomial$(n, 1/N)$. Ou seja,

$$P(F_i = f_i) = \binom{n}{f_i} p^{f_i}(1-p)^{n-f_i} \mathbb{1}_{\{0,1,2,\ldots,n\}}(f_i)$$

  2. $(F_i, F_j) \overset{i.d.}{\sim}$ trinomial$(n, 1/N, 1/N)$ $(i \neq j)$, ou seja, os $F_i's$ são **dependentes** e:

$$\begin{aligned} P(F_i = f_i, F_j = f_j) &= \frac{n!}{f_i! f_j! (n - f_i - f_j)!} p_1^{f_i} p_2^{f_j} (1 - p_1 - p_2)^{n - f_i - f_j} \\ &\times \mathbb{1}_{\{0,1,2,\ldots,n-f_j\}}(f_i) \mathbb{1}_{\{0,1,2,\ldots,n\}}(f_j) \end{aligned}$$

- Exercício: obtenha a distribuição de $(F_{k_1}, F_{k_2}, ..., F_{k_r})^t$, em que $\{k_1, k_2, ..., k_r\}$ é um subconjunto de $\{1, 2, ..., n\}$.
- Exercício: obtenha a distribuição de $(F_1, F_2, ..., F_N)^t$.

Estimação da média

- Prova (1): para um número n de repetições, repetimos um procedimento cujo a probabilidade de sucesso (selecionar o elemento i) é constante (1/N) e contabilizamos o número de sucessos.
- Prova (2): para um número n de repetições, repetimos um procedimento cujas probabilidades de interesse (selecionar os elmentos i e j, $i \neq j$) são constantes (1/N,1/N) e contabilizamos o número de ocorrências de cada categoria.

Propriedades dos F_i

1. $\mathcal{E}_{A_1}(F_i) = \frac{n}{N}, \mathcal{V}_{A_1}(F_i) = n\frac{1}{N}\left(\frac{N-1}{N}\right) = n\frac{N-1}{N^2}$.
2. $Cov_{A_1}(F_i, F_j) = -n\frac{1}{N}\frac{1}{N} = -\frac{n}{N^2}$.
3. $\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n$ (probabilidade do i-ésimo elemento aparecer na amostra). Prova

$$\pi_i = 1 - P(F_i = 0) = 1 - \left(1 - \frac{1}{N}\right)^n.$$

4. $\pi_{ij} = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n$ (probabilidade do i-ésimo e j-ésimo elementos aparecerem na amostra). Exercício: provar esta propriedade.

As provas dos resultados 1, 2 e 3 decorrem das respectivas distribuições de F_i e $(F_i, F_j)^t$.

Estimação da média : Propriedades do estimador sob AAS_c

- Valor esperado

$$\begin{aligned}\mathcal{E}_{A_1}(\widehat{\mu}) &= \frac{1}{n}\mathcal{E}_{A_1}\left(\sum_{i=1}^{N} F_i y_i\right) = \frac{1}{n}\sum_{i=1}^{N}\mathcal{E}_{A_1}(F_i)y_i = \frac{1}{n}\sum_{i=1}^{N}\frac{n y_i}{N} \\ &= \sum_{i=1}^{N}\frac{y_i}{N} = \mu\end{aligned}$$

- Portanto $\widehat{\mu}$ é uma estimador não viciado (env)

Estimação da média : Propriedades do estimador sob AAS_c

- Variância do estimador

$$
\begin{aligned}
\mathcal{V}_{A_1}(\widehat{\mu}) &= \frac{1}{n^2}\mathcal{V}\left(\sum_{i=1}^{N} F_i y_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 \mathcal{V}_{A_1}(F_i) + \sum_{\substack{i\neq j,\\ i,j=1,2,\ldots,N}} Cov_{A_1}(F_i y_i, F_j y_j)\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 \left(\frac{n(N-1)}{N^2}\right) - \sum_{\substack{i\neq j,\\ i,j=1,2,\ldots,N}} y_i y_j \frac{n}{N^2}\right) \\
&= \frac{1}{N^2 n}(N-1)\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{\substack{i\neq j,\\ i,j=1,2,\ldots,N}} y_i y_j\right)
\end{aligned}
$$

Estimação da média : Propriedades do estimador sob AAS_c

- Cont.

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

$$
\begin{aligned}
\mathcal{V}_{A_1}(\widehat{\mu}) &= \frac{1}{N^2 n}(N-1)\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\left(N^2\mu^2 - \sum_{i=1}^{N} y_i^2\right)\right) \\
&= \frac{N-1}{(N-1)N^2 n}\left(N\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{n} y_i^2 - N^2\mu^2 + \sum_{i=1}^{N} y_i^2\right) \\
&= \frac{1}{N^2 n}\left(N\sum_{i=1}^{N} y_i^2 - N^2\mu^2\right) = \frac{1}{n}\left(\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \mu^2\right) = \frac{\sigma^2}{n}
\end{aligned}
$$

Estimação da média : Propriedades do estimador sob AAS_c

- Resumidamente, $\mathcal{E}_{A_1}(\widehat{\mu}) = \mu$ e $\mathcal{V}_{A_1}(\widehat{\mu}) = \frac{\sigma^2}{n}$. Podemos provar, sob AAS_c, que $\widehat{\mu}$ é consistente.
- A distribuição exata de $\widehat{\mu}$ (sob A_1) 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 TCL 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_c

- 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

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

ou

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

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_c

- Vamos considerar o seguinte estimador

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

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

$$\begin{aligned}
\mathcal{E}_{A_1}(\widehat{\sigma}^2) &= \frac{1}{n-1}\mathcal{E}_{A_1}\left(\sum_{i=1}^{n} Y_i^2 - n\widehat{\mu}^2\right) = \frac{1}{n-1}\left[\mathcal{E}_{A_1}\left(\sum_{i=1}^{N} y_i^2 F_i\right) - n\mathcal{E}_{A_1}(\widehat{\mu}^2)\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{N} y_i^2\mathcal{E}_{A_1}(F_i) - n\mathcal{E}_{A_1}\left(\widehat{\mu}^2\right)\right] \\
&= \frac{1}{n-1}\left[N(\sigma^2+\mu^2)\frac{n}{N} - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - n\mu^2 - \sigma^2\right] = \sigma^2
\end{aligned}$$

Estimação da média : Propriedades do estimador sob AAS_c

- A prova de sua consistência, i.e.,

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

  também será discutida mais à frente.

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

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

  por Slutsky.

Intervalo de Confiança

- Estimativa da variância:

  $\widetilde{\sigma}^2 = \frac{1}{n-1}\sum_{i\in \boldsymbol{s}}(y_i - \widetilde{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^{N} f_i\,(y_i - \widetilde{\mu})^2$.

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

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

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

- Erro da estimativa: $z_\gamma\sqrt{\frac{\widehat{\sigma}^2}{n}}$.

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 $Z_t = \frac{\widehat{\mu}-\mu_0}{\widehat{\sigma}/\sqrt{n}}$, em que $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$.
- Sob H_0, vimos que $Z_t \approx N(0,1)$, para n e N-n suficientemente grandes.
- Defina $z_t = \frac{\widetilde{\mu}-\mu_0}{\widetilde{\sigma}/\sqrt{n}}$ 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)
  - $p - valor = 2[1 - P(Z \leq |z_t||H_0)]$.

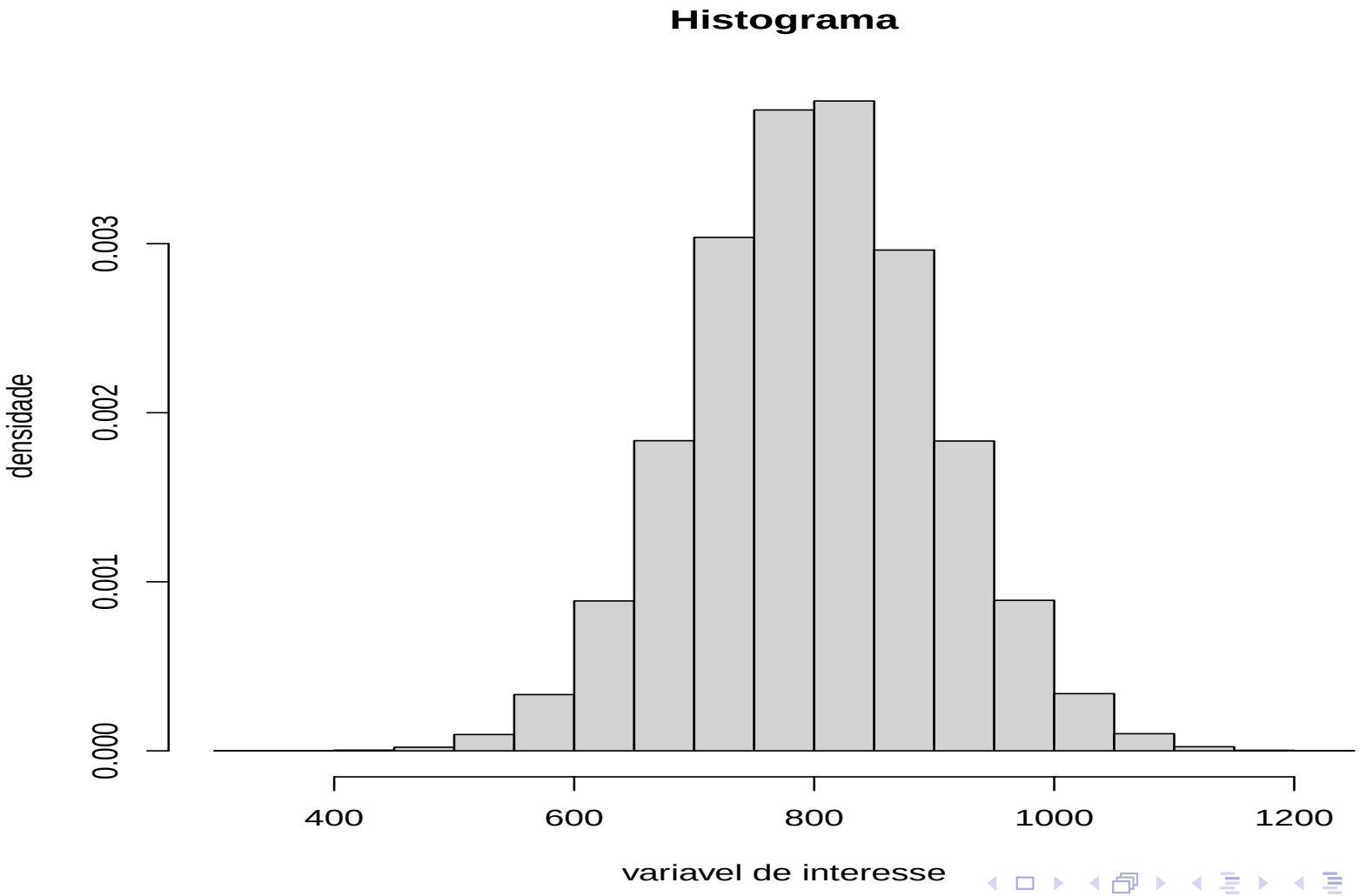
Estudos de simulação

- Distribuição assintótica do estimador para a média. Tamanho da população $N = 100.000$.
- Cinco cenários, variando em função da variável de interesse na população (X).
  - $X \sim N(800, 10.000)$
  - $X \sim \text{gama}(5; 0,00625),\ E(X) = 800, V(X) = 128.000.$
  - $X \sim t_{(7)}(800, 5.000),\ E(X) = 800,\ V(X) = 7.000.$
  - $X \sim U[400; 1.200].$
  - $X \sim 0,5N(200, 5.000) + 0,5N(600, 5.000)$

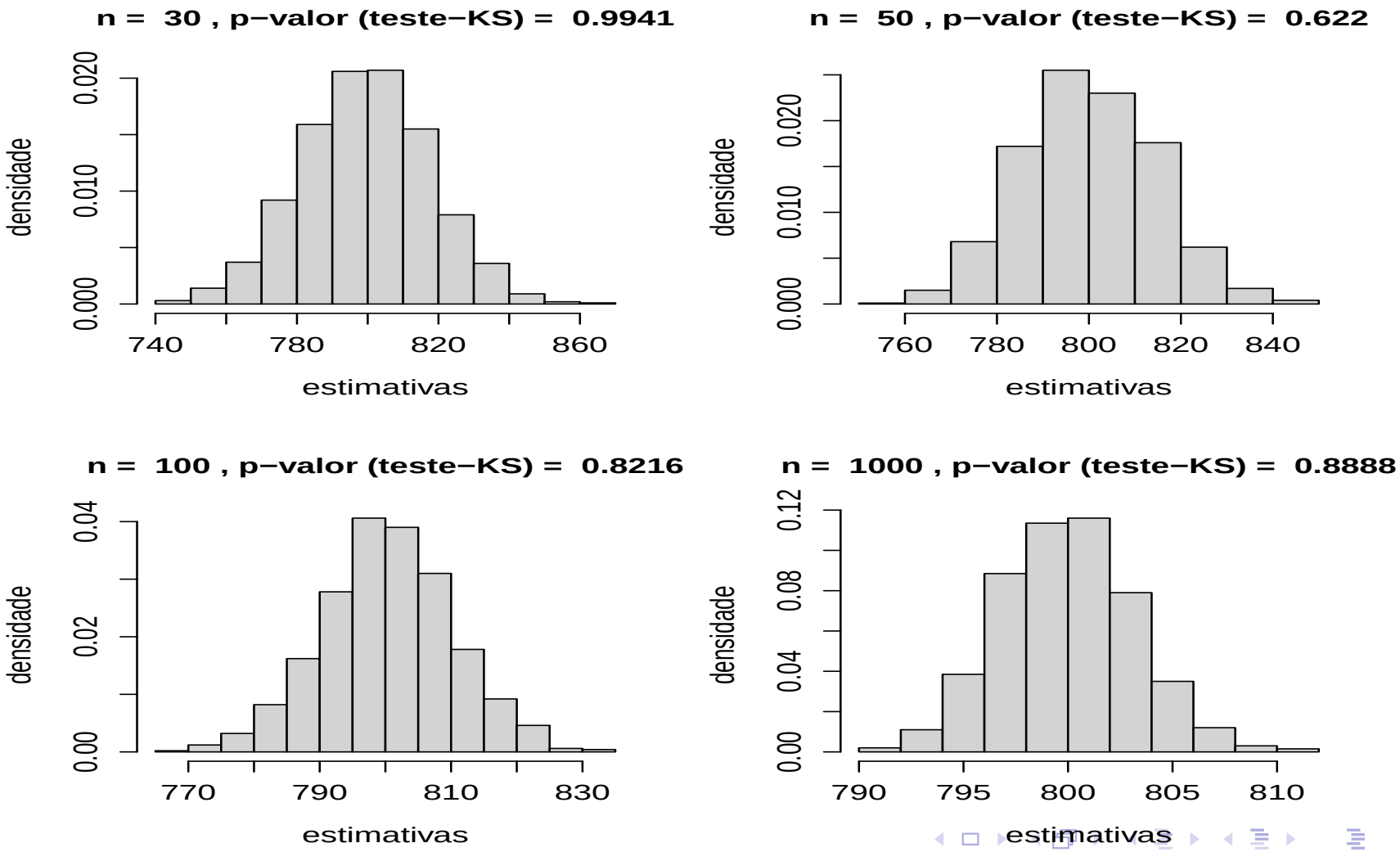
Estudos de simulação

- Quatro tamanhos amostrais $(30, 50, 100, 1.000)$, 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).

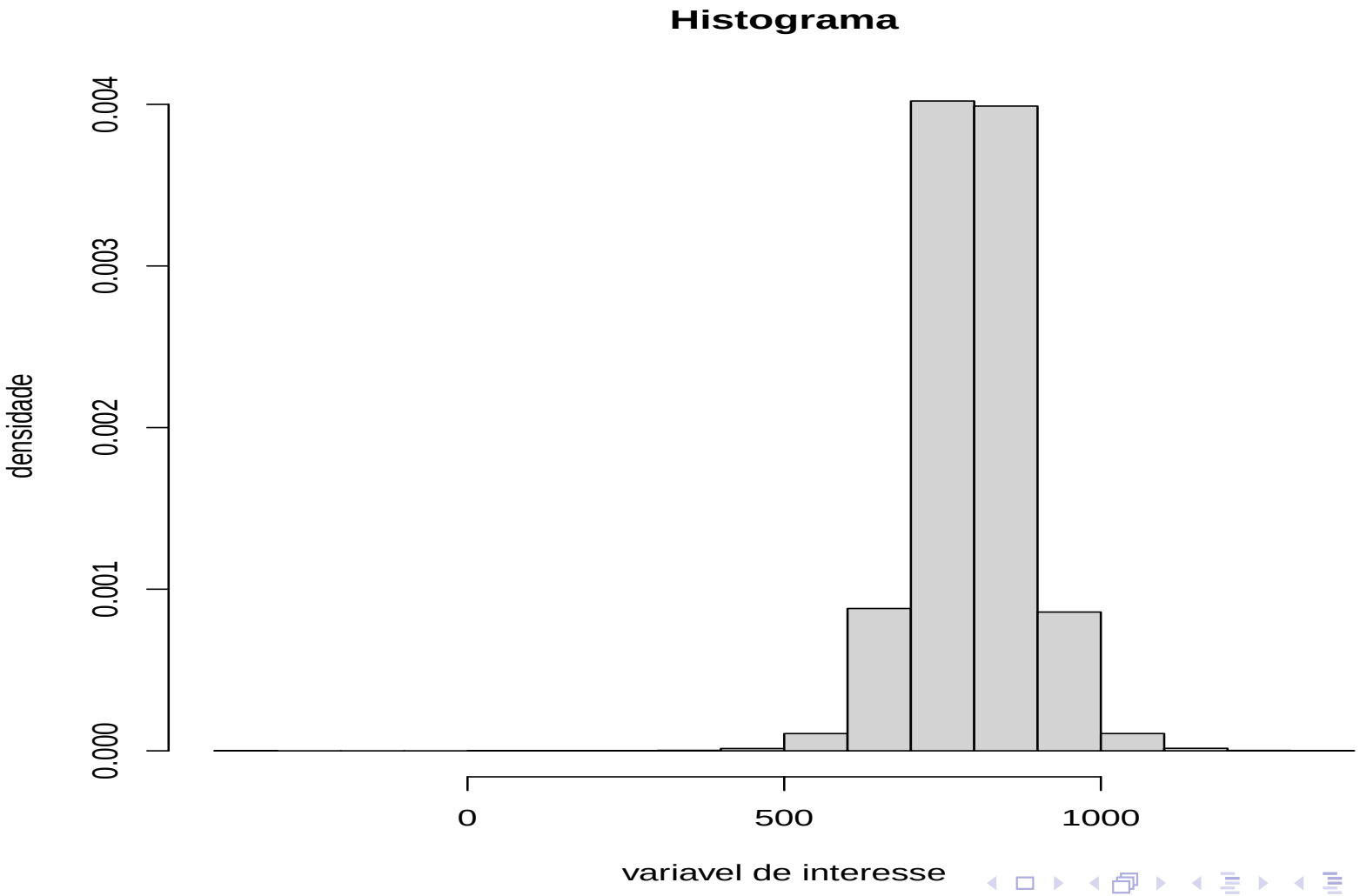
normal


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

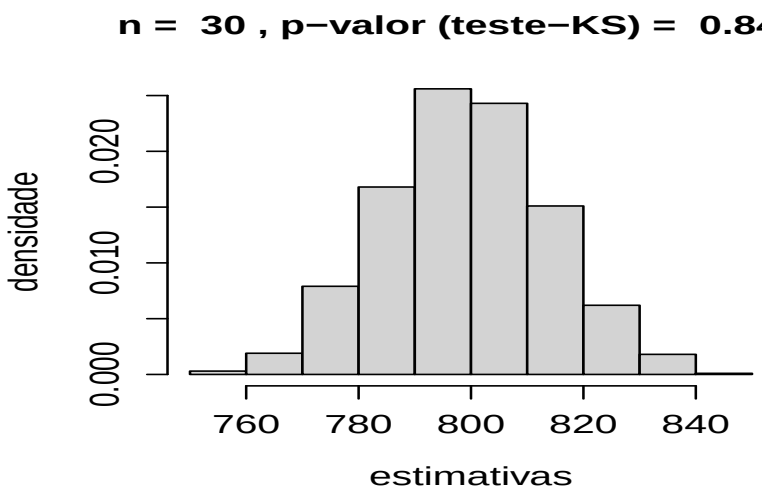
normal

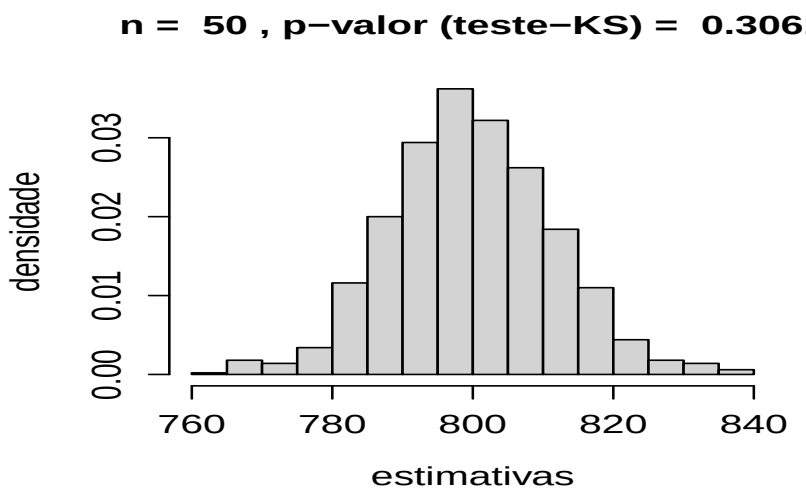

n = 30 , p−valor (teste−KS) = 0.9941
densidade
0.000
0.010
0.020
740
780
820
860
estimativas
n = 50 , p−valor (teste−KS) = 0.622
densidade
0.000
0.010
0.020
760
780
800
820
840
estimativas
n = 100 , p−valor (teste−KS) = 0.8216
densidade
0.00
0.02
0.04
770
790
810
830
estimativas
n = 1000 , p−valor (teste−KS) = 0.8888
densidade
0.00
0.04
0.08
0.12
790
795
800
805
810
estimativas

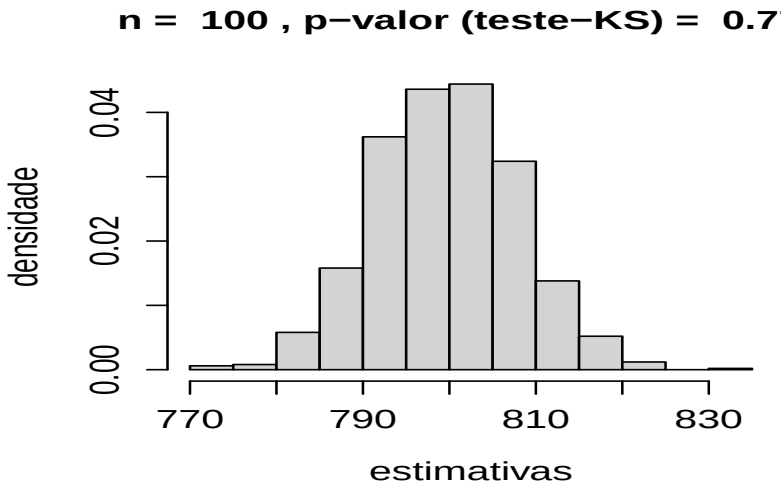
t de Student

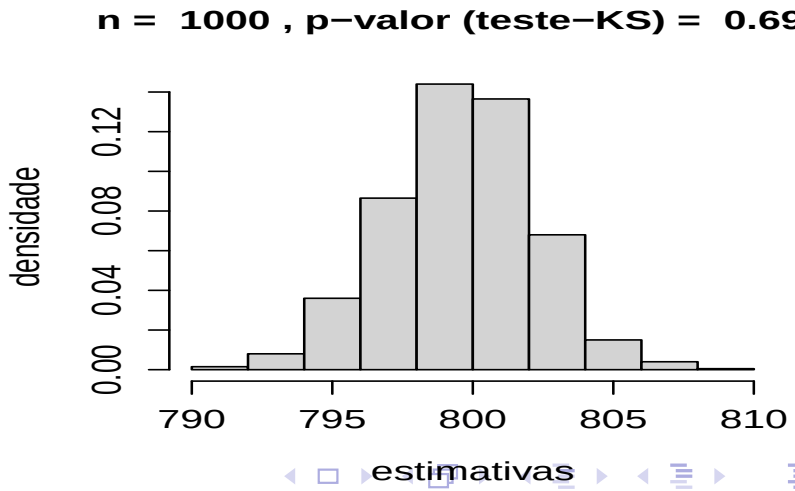

Histograma
densidade
0.000
0.001
0.002
0.003
0.004
0
500
1000
variavel de interesse

t de Student

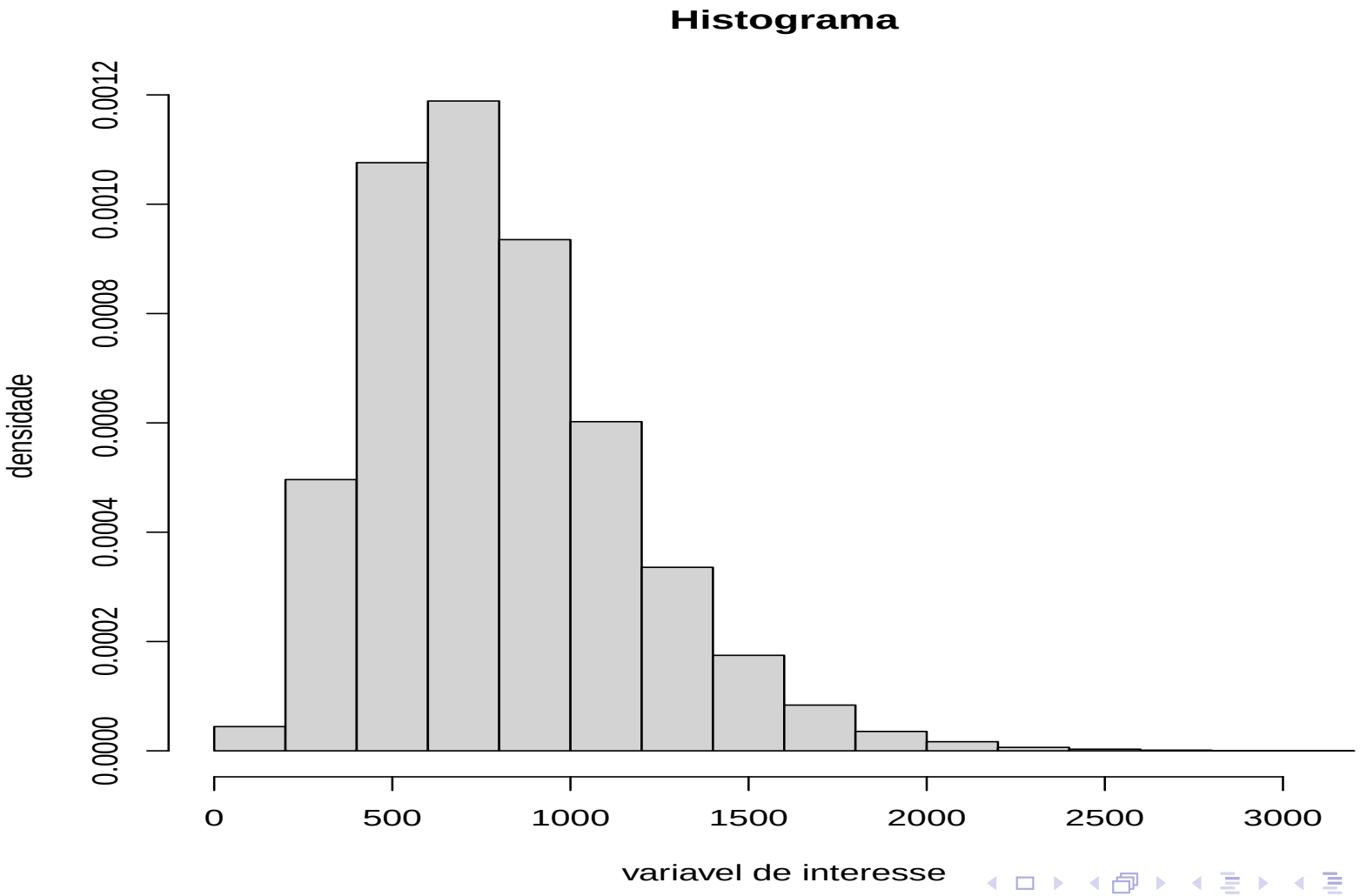

n = 30 , p−valor (teste−KS) = 0.8446
densidade
0.000
0.010
0.020
760
780
800
820
840
estimativas


n = 50 , p−valor (teste−KS) = 0.3061
densidade
0.00
0.01
0.02
0.03
760
780
800
820
840
estimativas

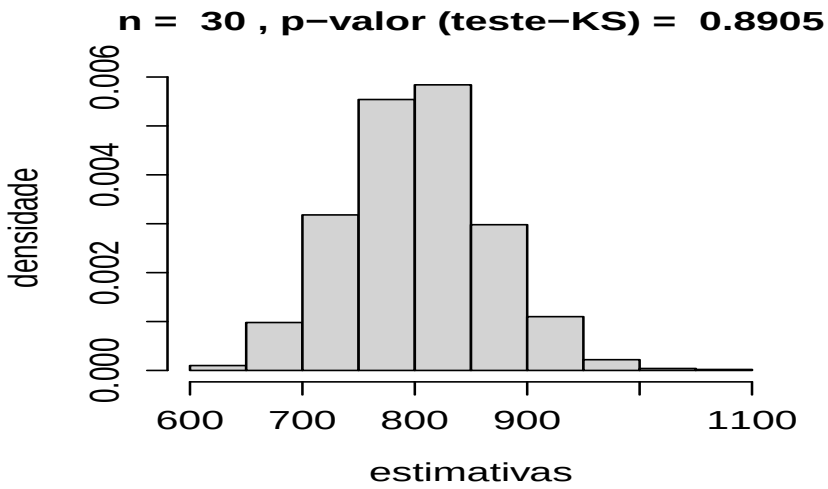

n = 100 , p−valor (teste−KS) = 0.776
densidade
0.00
0.02
0.04
770
790
810
830
estimativas

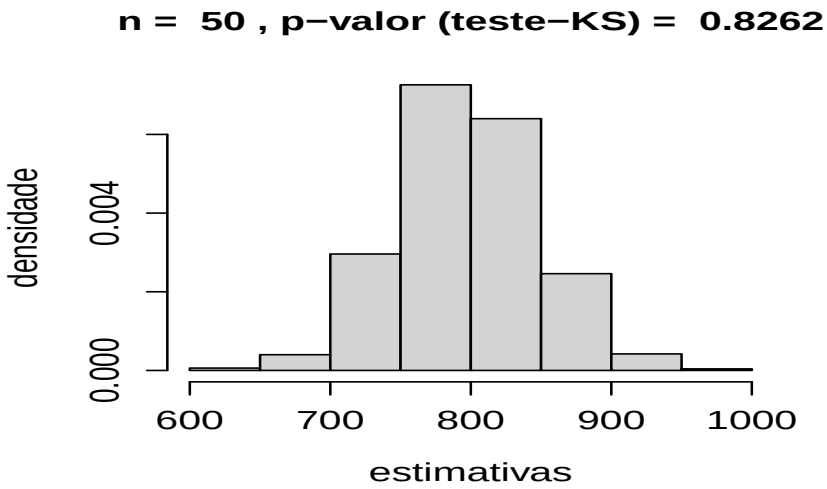

n = 1000 , p−valor (teste−KS) = 0.6938
densidade
0.00
0.04
0.08
0.12
790
795
800
805
810
estimativas

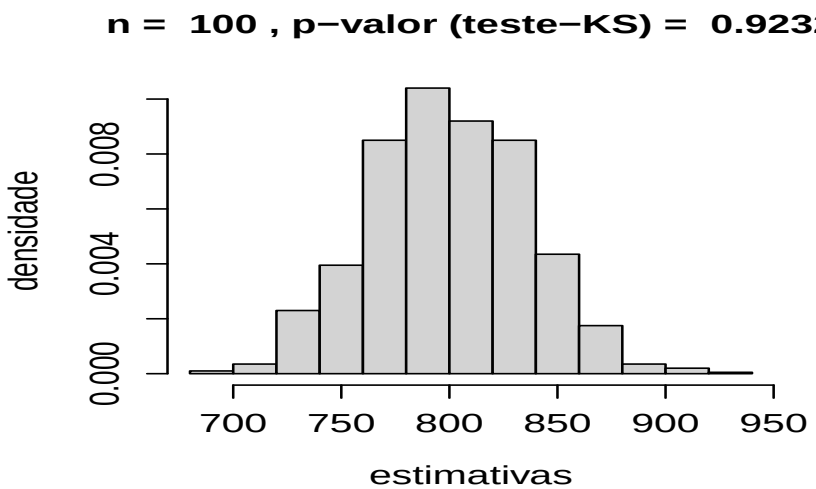
gama

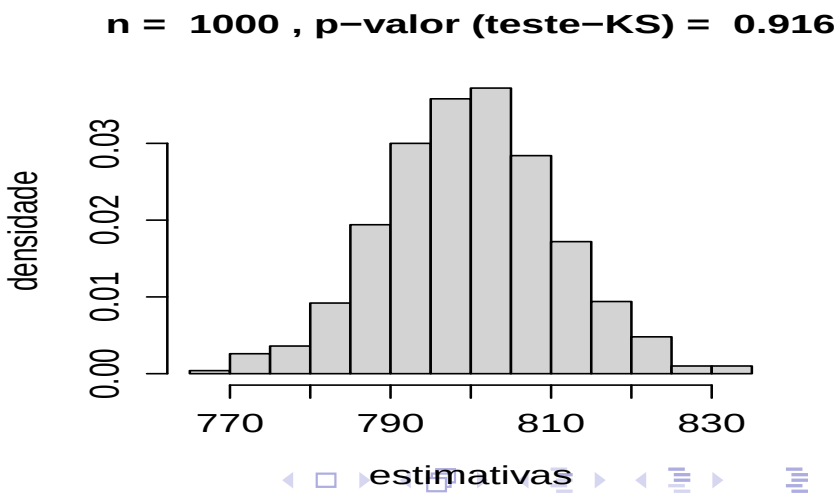

Histograma
densidade
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0
500
1000
1500
2000
2500
3000
variavel de interesse

gama

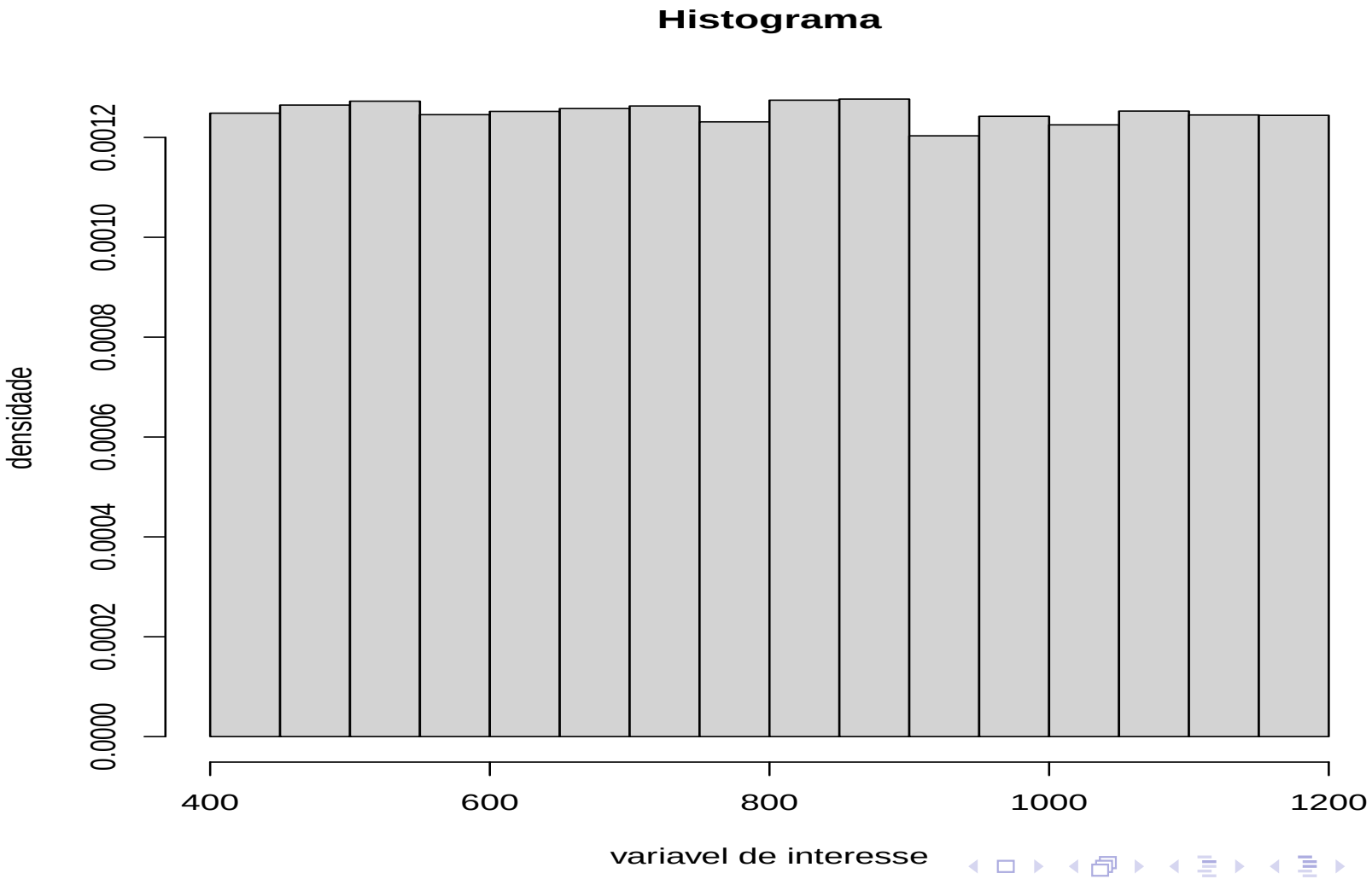

n = 30 , p−valor (teste−KS) = 0.8905
densidade
0.000
0.002
0.004
0.006
600
700
800
900
1100
estimativas


n = 50 , p−valor (teste−KS) = 0.8262
densidade
0.000
0.004
600
700
800
900
1000
estimativas

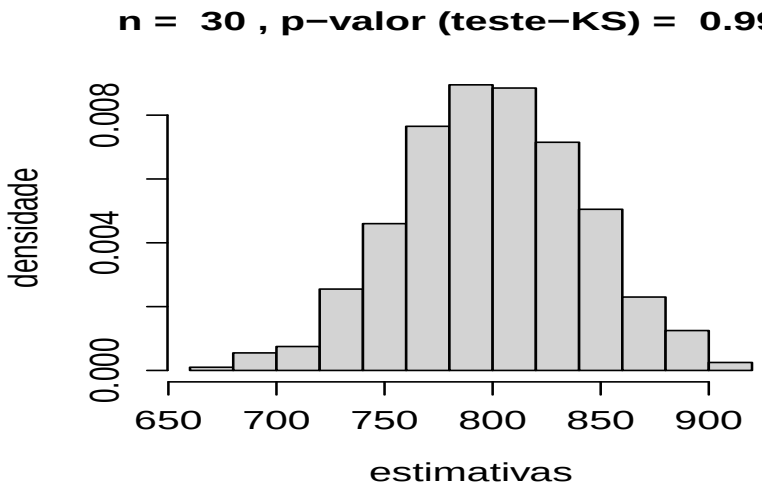

n = 100 , p−valor (teste−KS) = 0.9232
densidade
0.000
0.004
0.008
700
750
800
850
900
950
estimativas

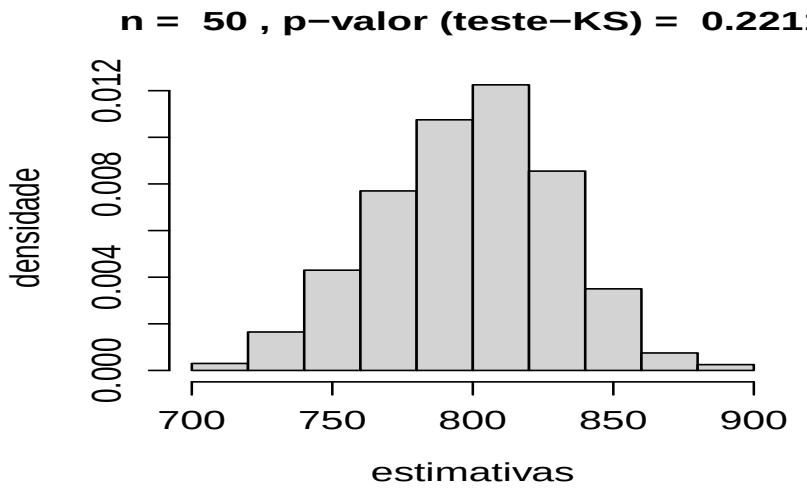

n = 1000 , p−valor (teste−KS) = 0.916
densidade
0.00
0.01
0.02
0.03
770
790
810
830
estimativas

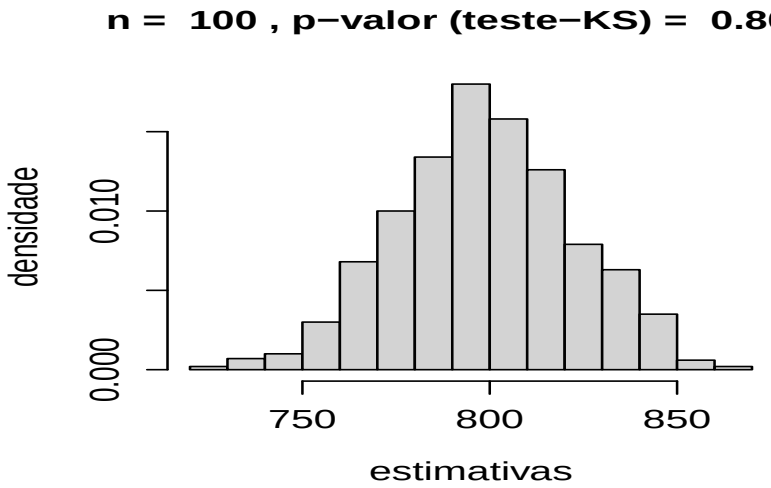
uniforme

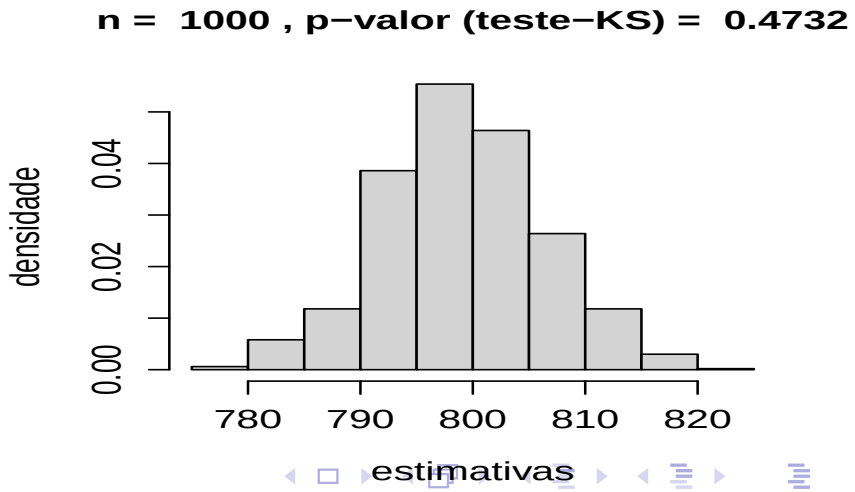

Histograma
densidade
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
400
600
800
1000
1200
variavel de interesse

uniforme

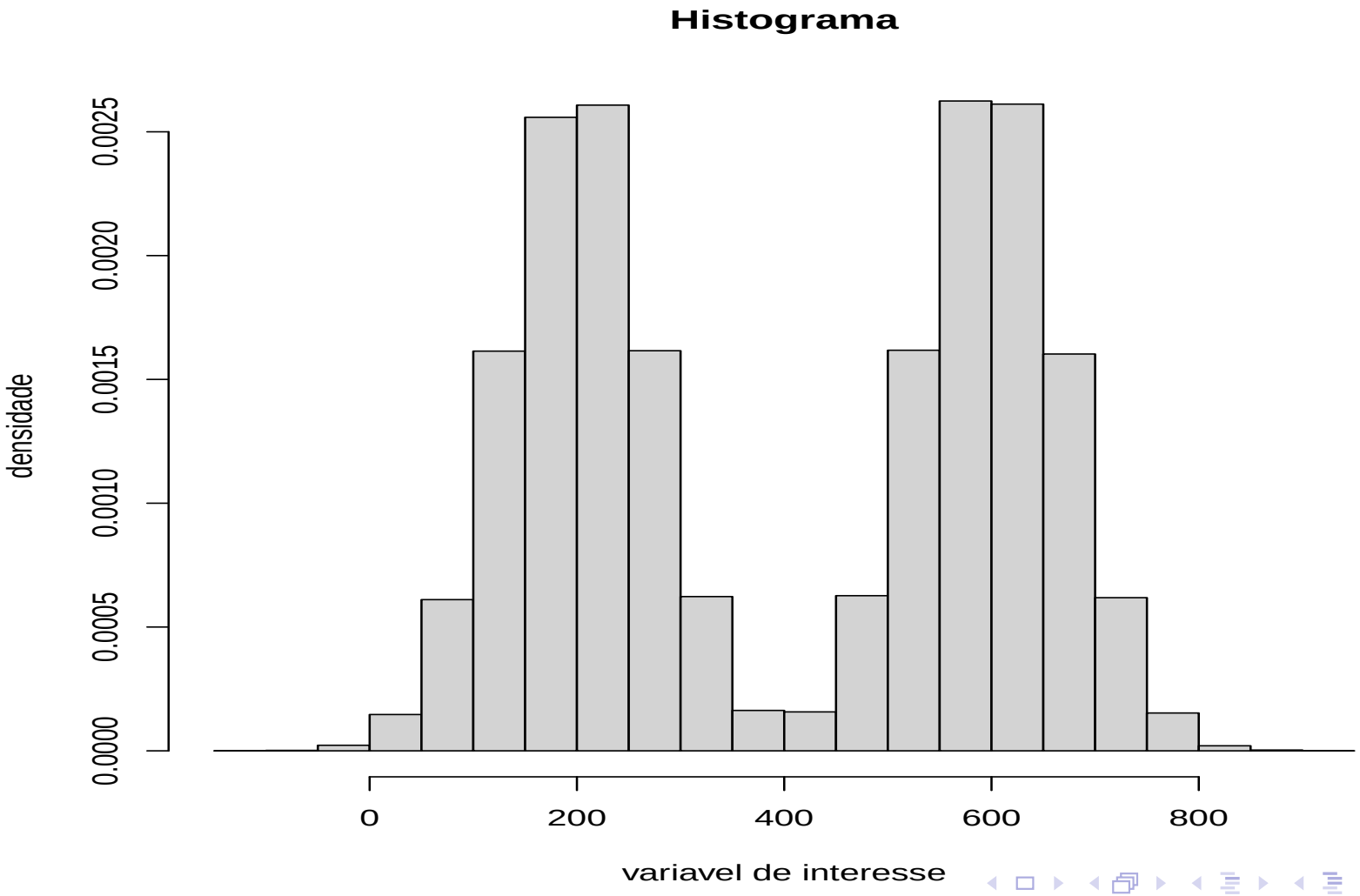

n = 30 , p−valor (teste−KS) = 0.9923
densidade
0.000
0.004
0.008
650
700
750
800
850
900
estimativas


n = 50 , p−valor (teste−KS) = 0.2211
densidade
0.000
0.004
0.008
0.012
700
750
800
850
900
estimativas

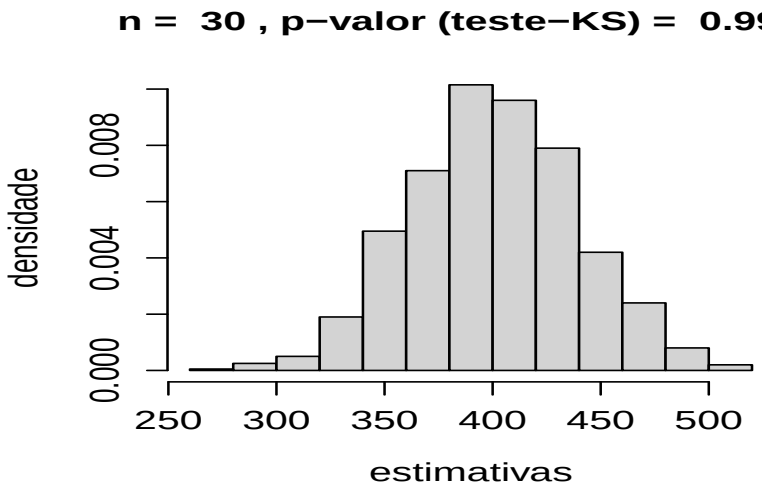

n = 100 , p−valor (teste−KS) = 0.8009
densidade
0.000
0.010
750
800
850
estimativas

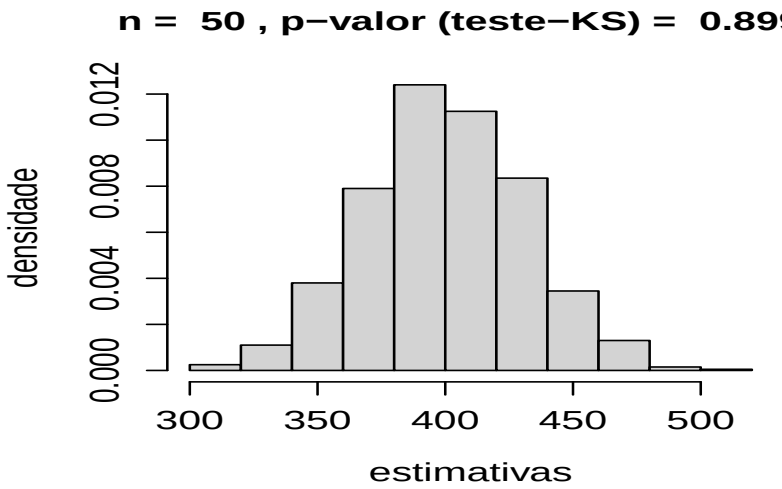

n = 1000 , p−valor (teste−KS) = 0.4732
densidade
0.00
0.02
0.04
780
790
800
810
820
estimativas

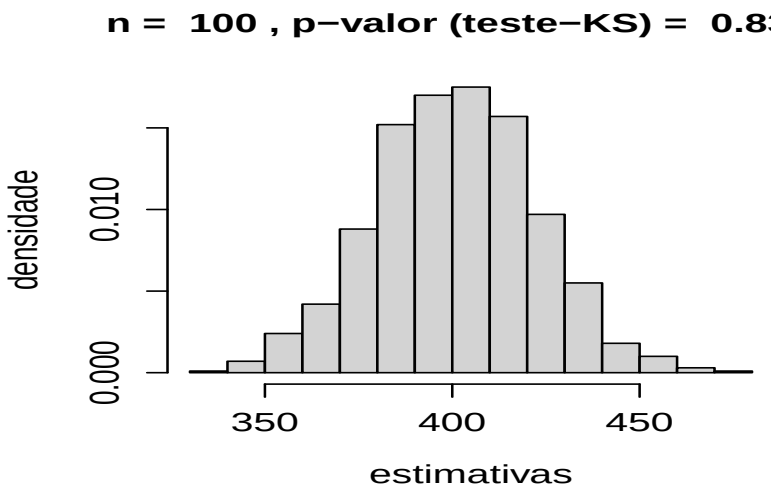
mistura de duas normais

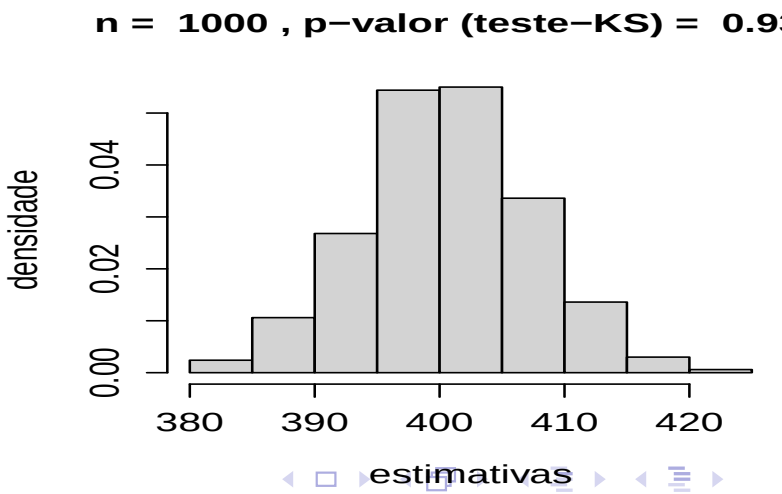


mistura de duas normais









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.
- Critérios:
  - Erro de estimativa: $z_\gamma \sqrt{\frac{\widehat{\sigma}^2}{n}}$. Fixa-se um erro de estimativa de interesse.
  - Precisão (Probabilidade do módulo da diferença): $P\left(|\widehat{\mu} - \mu| < \delta\right) > \gamma$, $\delta > 0$, $\gamma \in (0, 1)$ (fixa-se δ).

Determinação do tamanho amostral: erro da estimativa

$$\delta = z_\gamma \sqrt{\frac{\sigma^2}{n}} \rightarrow n = \frac{z_\gamma^2 \sigma^2}{\delta^2}$$

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

Determinação do tamanho amostral: precisão

$$P_{A_1}\left(|\widehat{\mu}-\mu|<\delta\right)>\gamma \quad \leftrightarrow \quad P_{A_1}\left(\left|\frac{\widehat{\mu}-\mu}{\sqrt{\sigma^2/n}}\right|<\frac{\sqrt{n}\delta}{\sigma}\right)>\gamma$$

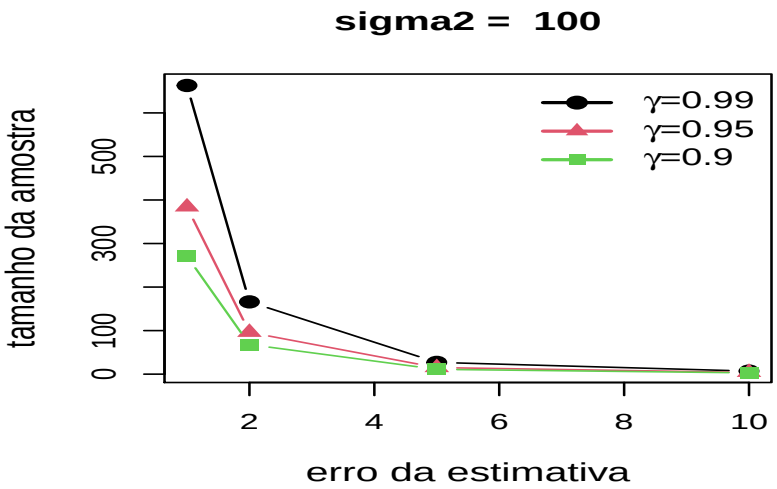
$$\leftrightarrow \quad P_{A_1}\left(|Z|<\frac{\sqrt{n}\delta}{\sigma}\right)>\gamma \leftrightarrow \frac{\sqrt{n}\delta}{\sigma}=z_\gamma$$

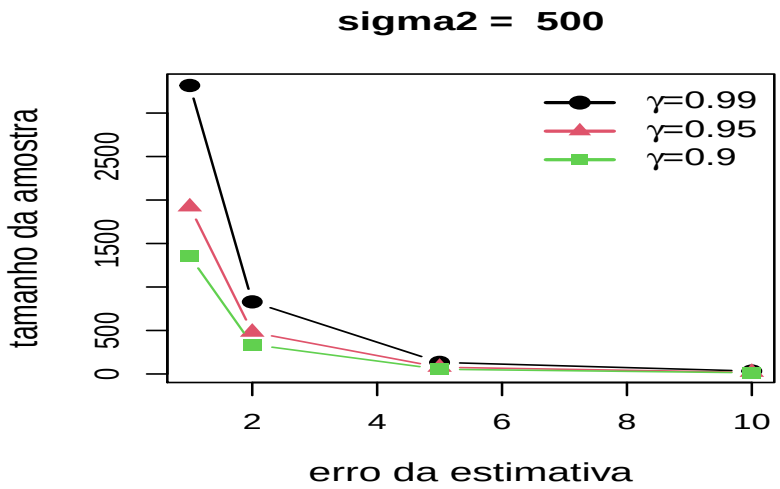
em que $Z \approx N(0,1)$. O que leva ao mesmo procedimento oriundo de se fixar o erro da estimativa.

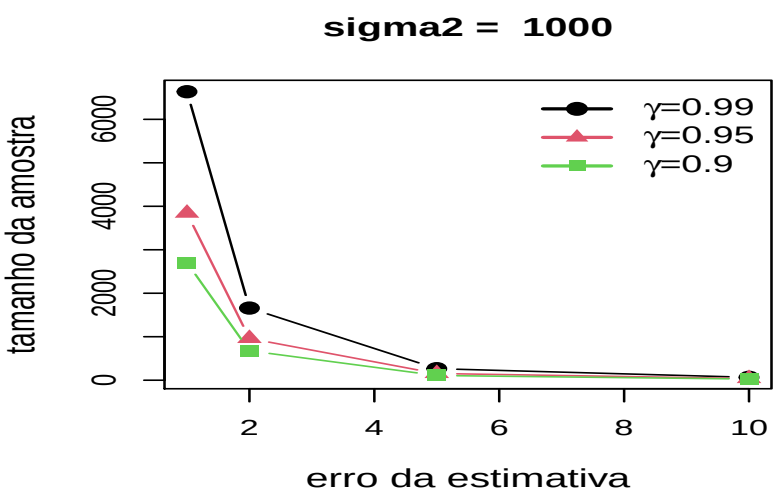
Tamanhos amostrais

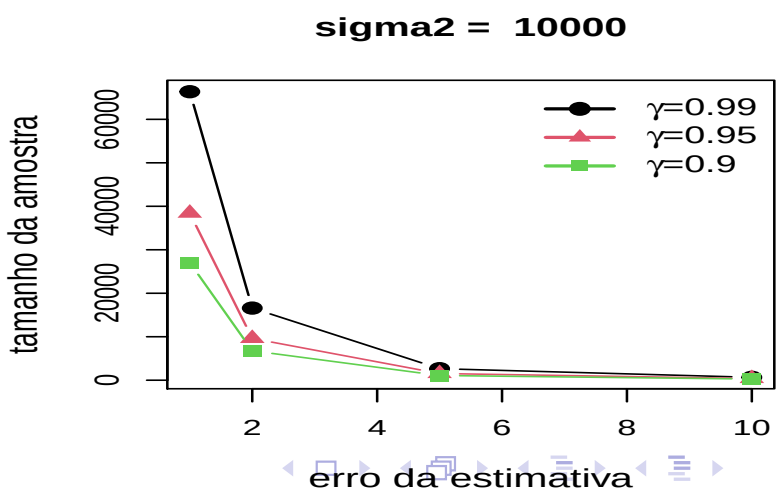
- Situações hipotéticas: cruzamento entre os níveis de diferentes fatores de interesse
  - $\delta \in \{1, 2, 5, 10\}$.
  - $\gamma \in \{0,9; 0,95; 0,99\}$.
  - $\sigma^2 \in \{100, 500, 1.000, 10.000\}$.

Tamanhos amostrais









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

- $\mathcal{E}_{A_1}(\widehat{\tau}) = \mathcal{E}_{A_1}(N\widehat{\mu}) = N\mathcal{E}(\widehat{\mu}) = N\mu = \tau$ (não viciado).
- $\mathcal{V}_{A_1}(\widehat{\tau}) = N^2\mathcal{V}_{A_1}(\widehat{\mu}) = N^2\frac{\sigma^2}{n}$ (a imprecisão associada à estimação do total é maior do que aquela associada à média).

Propriedades do estimador

- Normalidade assintótica, como

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

lembrando que N é fixo, temos que

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

Intervalo de Confiança

- Estimativa da variância:

  $\widetilde{\sigma}^2 = \frac{1}{n-1}\sum_{i\in \boldsymbol{s}}(y_i - \widetilde{\mu})^2 = \frac{1}{n-1}\sum_{=1}^{N} f_i\,(y_i - \widetilde{\mu})^2.$

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

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

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

- Erro da estimativa: $z_\gamma N\sqrt{\frac{\widehat{\sigma}^2}{n}}$.

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 $Z_t = \frac{\widehat{\tau} - \tau_0}{N\widehat{\sigma}/\sqrt{n}}$, em que $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$.
- Sob H_0, vimos que $Z_t \approx N(0,1)$, para n e N-n suficientemente grandes.
- Defina $z_t = \frac{\widetilde{\tau} - \tau_0}{N\widetilde{\sigma}/\sqrt{n}}$ 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 (Exercício).

Determinação do tamanho amostral: erro da estimativa

Analogamente ao caso da média populacional, temos que

$$\delta = z_\gamma N \sqrt{\frac{\sigma^2}{n}} \rightarrow n = \frac{z_\gamma^2 N^2 \sigma^2}{\delta^2}$$

Isto vale para qualquer um dos dois critérios: erro da estimativa e precisão.