

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, \dots, y_N (variáveis não aleatórias), em que y_i é a observação relativa ao indivíduo i (podemos também considerar observações multivariadas $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})^t$). 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 ($\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 / s^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu)^2$) (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 sem reposição ($AAS_s \equiv A_2$).
- Amostra $\{y_{k_1}, y_{k_2}, \dots, y_{k_n}\}$, em que $k_i \in \{1, 2, \dots, N\}$. Por exemplo, Se $N = 5$ e $n = 3$, podemos ter $\{y_2, y_3, y_5\}$.

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 O elemento selecionado é retirado da população.
- 3 Repete-se os procedimentos 1 e 2, $n - 1$ vezes.
- 4 No R:
 - Função “[sample](#)”.
 - Pacote “[sampling](#)”.
 - Pacote “[survey](#)”.

Notações/exemplo (univariado)

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

Estimação da média

- Estimador natural (sob duas formas diferentes):

$$\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i \in \mathbf{S}} y_i \quad (1)$$

$$\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{i=1}^N \Delta_i y_i, \quad (2)$$

em que, na Equação (1) $Y_i, i = 1, \dots, n$ e $\mathbf{S} = (K_1, \dots, K_n)$ são variáveis (vetores) aleatórios, enquanto que na Equação (2) Δ_i é uma v.a. que indica se o elemento i elemento da população apareceu na amostra, e \mathbf{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 $\mathbf{Y} = (Y_1, \dots, Y_n)^t$ (esses desenvolvimentos dependem da distribuição considerada para a amostra).
- Neste caso, $Y_i, i = 1, \dots, n$ 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 \mathbf{Y} sejam válidas.
- Note que, neste caso, como não há reposição as variáveis Y_i 's não são independentes nem identicamente distribuídas.
- Utilizar a distribuição de \mathbf{S} pode ser bem complicado.

Estimação da média

- Utilizar a forma (2) e considerarmos a distribuição das variáveis $\Delta = (\Delta_1, \dots, \Delta_N)^t$ (mais geral, ou seja, em princípio, os resultados se aplicam, independentemente da distribuição de \mathbf{Y}).
- Neste caso, $\Delta_i, i = 1, \dots, N$ são variáveis aleatórias ($y_i, i = 1, 2, \dots, N$ 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 \mathbf{Y} .

Estimação da média

- Resultados ([link 1](#), [link 2](#), [link 3](#)) (i.d. - identicamente distribuídas):

1 Note que, nesse caso, $\Delta_i \equiv F_i$.

2 $F_i \stackrel{i.d.}{\sim} \text{Bernoulli}(n/N)$. Ou seja, $P(F_i = f_i) = p^{f_i}(1-p)^{1-f_i} \mathbf{1}_{\{0,1\}}(F_i)$,
 $\mathcal{E}(F_i) = \frac{n}{N}$, $\mathcal{V}(F_i) = \frac{n}{N} \left(1 - \frac{n}{N}\right)$

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:

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

Demonstrações

Por outro lado, temos que

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

mas

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

Demonstrações

Assim

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

Estimação da média : Propriedades do estimador sob AAS_s

- Valor esperado

$$\begin{aligned}\mathcal{E}_{A_2}(\hat{\mu}) &= \frac{1}{n} \mathcal{E}_{A_2} \left(\sum_{i=1}^N F_i y_i \right) = \frac{1}{n} \sum_{i=1}^N \mathcal{E}_{A_2}(F_i) y_i = \frac{1}{n} \sum_{i=1}^N \frac{ny_i}{N} \\ &= \sum_{i=1}^N \frac{y_i}{N} = \mu\end{aligned}$$

Estimação da média : Propriedades do estimador sob AAS_s

■ Variância do estimador

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

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

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

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 $\mathcal{E}_{A_i}(\hat{\mu}) = \mu$, $i = 1, 2$.
- Por outro lado, $\mathcal{V}_{A_1}(\hat{\mu}) = \frac{\sigma^2}{n}$ e $\mathcal{V}_{A_2}(\hat{\mu}) = (1 - f)\frac{s^2}{n}$, em que $f = \frac{n}{N} \in (0, 1)$ e $\sigma^2 = \frac{N-1}{N}s^2$.
- Portanto, o efeito do planejamento (EPA), do estimador sob o plano A_2 em relação ao plano A_1 , é dado por:

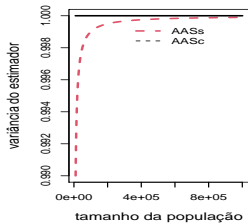
$$EPA = \frac{\mathcal{V}_{A_2}(\hat{\mu})}{\mathcal{V}_{A_1}(\hat{\mu})} = (1 - f)\frac{s^2}{\sigma^2} = (1 - f)\frac{N}{N - 1} \approx (1 - f)$$

Estimação da média : Propriedades do estimador sob AAS_s

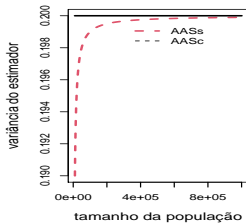
- Assim, se $N \rightarrow \infty$, então $\mathcal{V}_{A_2}(\hat{\mu}) \rightarrow \mathcal{V}_{A_1}(\hat{\mu})$.
- Consequentemente, temos que o plano AAS_s é melhor do que AAS_c , tendendo ambos a serem equivalentes, à medida que o tamanho da população (e o da amostra) tende(m) a infinito.

Comparação dos planos amostrais

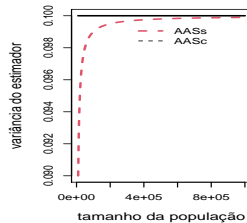
n = 100 , sigma2 = 100



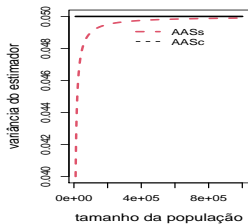
n = 500 , sigma2 = 100



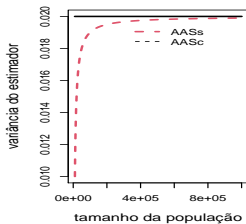
n = 1000 , sigma2 = 100



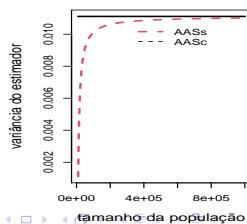
n = 2000 , sigma2 = 100



n = 5000 , sigma2 = 100

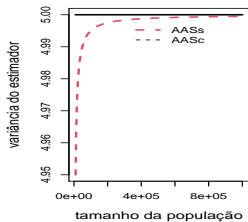


n = 9000 , sigma2 = 100

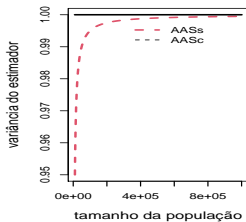


Comparação dos planos amostrais

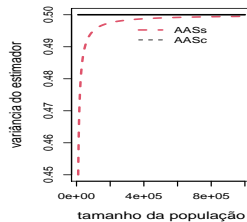
n = 100 , sigma2 = 500



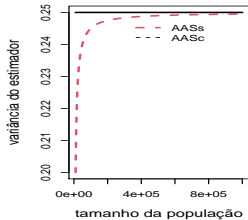
n = 500 , sigma2 = 500



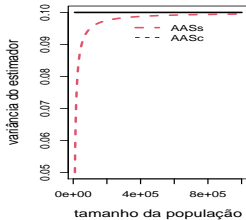
n = 1000 , sigma2 = 500



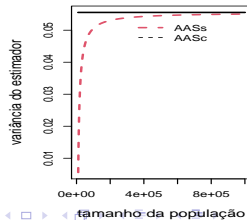
n = 2000 , sigma2 = 500



n = 5000 , sigma2 = 500

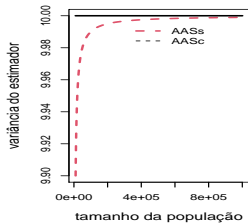


n = 9000 , sigma2 = 500

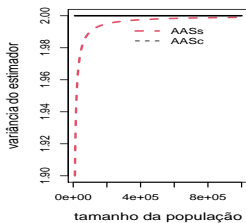


Comparação dos planos amostrais

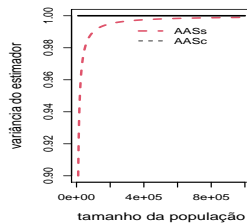
n = 100 , sigma2 = 1000



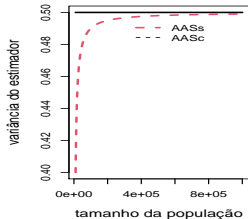
n = 500 , sigma2 = 1000



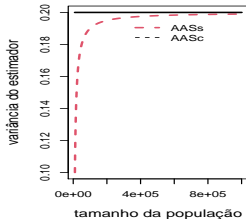
n = 1000 , sigma2 = 1000



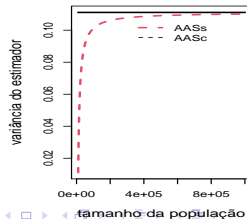
n = 2000 , sigma2 = 1000



n = 5000 , sigma2 = 1000

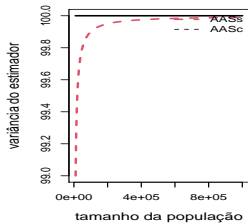


n = 9000 , sigma2 = 1000

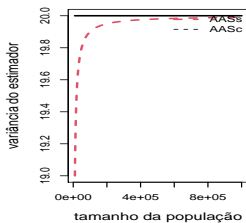


Comparação dos planos amostrais

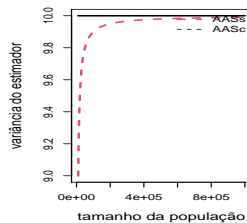
n = 100 , sigma2 = 10000



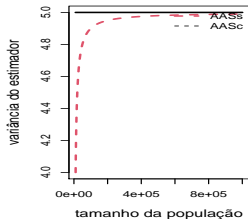
n = 500 , sigma2 = 10000



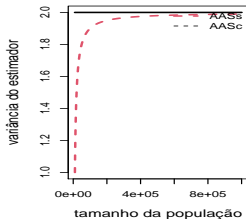
n = 1000 , sigma2 = 10000



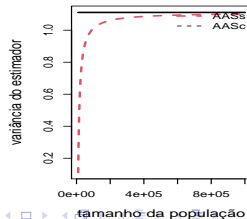
n = 2000 , sigma2 = 10000



n = 5000 , sigma2 = 10000



n = 9000 , sigma2 = 10000



Estimação da média : Propriedades do estimador sob AAS_s

- Resumidamente, $\mathcal{E}_{A_2}(\hat{\mu}) = \mu$ e $\mathcal{V}_{A_2}(\hat{\mu}) = (1 - f)\frac{s^2}{n}$. Podemos provar, sob AAS_s , que $\hat{\mu}$ 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 TCL padrão não se aplica.
- Estimativa $\tilde{\mu} = \frac{1}{n} \sum_{i \in s} y_i = \frac{1}{n} \sum_{i=1}^N f_i y_i$.

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

$$\frac{\hat{\mu} - \mu}{\sqrt{(1-f)s^2/n}} \xrightarrow[N \rightarrow \infty]{D} N(0, 1) \quad (4)$$

ou

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

- Problema: s^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

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

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

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

continua no próximo slide.

Estimação da média : Propriedades do estimador sob AAS_s

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

Estimação da média : Propriedades do estimador sob AAS_s

- A prova de sua **consistência**, i.e.,

$$\hat{s}^2 \xrightarrow[\substack{n \rightarrow \infty, \\ N-n \rightarrow \infty}]{P} s^2 \quad (5)$$

também será discutida mais à frente.

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

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

pele **Teorema de Slutsky**.

Intervalo de Confiança

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

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

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

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

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 $Z_t = \frac{\hat{\mu} - \mu_0}{\hat{s} \sqrt{(1-f)/\sqrt{n}}}$, em que $\hat{s} = \sqrt{\hat{s}^2}$.

- Sob H_0 , vimos que $Z_t \approx N(0, 1)$, para n e $N-n$ suficientemente grandes.

- Defina $z_t = \frac{\tilde{\mu} - \mu_0}{\tilde{s} \sqrt{(1-f)/\sqrt{n}}}$ 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)
 - $p - \text{valor} = P(Z \leq z_t | H_0)$

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)
 - $p - \text{valor} = P(Z \geq z_t | H_0)$

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).
 - $X \sim N(800, 10.000)$
 - $X \sim \text{gama}(5; 0, 00625)$, $E(X) = 800$, $V(X) = 128.000$.
 - $X \sim t_{(7)}(800, 5000)$, $E(X) = 800$, $V(X) = 7.000$.
 - $X \sim U[400; 1.200]$.
 - $X \sim 0.5N(200, 5.000) + 0.5N(600, 5.000)$

Estudos de simulação

- Quatro tamanhos amostrais (30, 50, 100, 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, sem reposição, 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} , $i=1, \dots, n$, (tamanho da amostra) $j=1, \dots, m$ (réplica).

$$\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \dots & v_{nm} \end{bmatrix}$$

Cont.

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

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

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

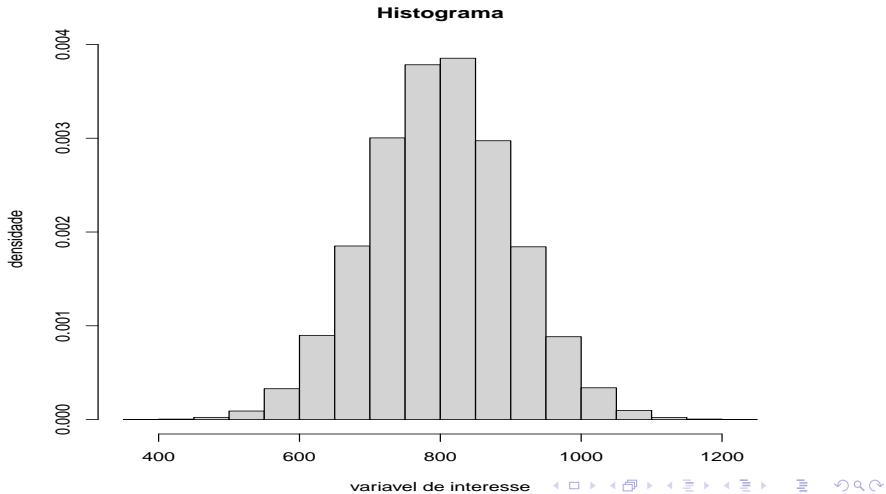
Cont.

- 5) Porém, na prática considera-se $v_{(i)l} = \frac{v_{(i)(2)} + v_{(i)(3)}}{2}$ 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, $i = 1, 2, \dots, n$.

- Além disso, consideramos como a linha de referência

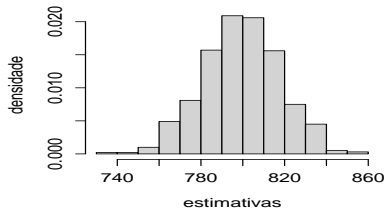
$$v_{(i)} = \frac{1}{m} \sum_{j=1}^m v_{(i)j}, i = 1, 2, \dots, n.$$

normal



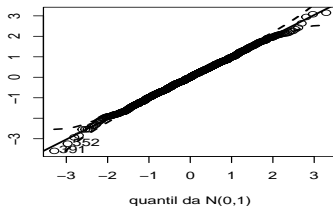
normal

n = 30 , p-valor (teste-KS) = 0.9922

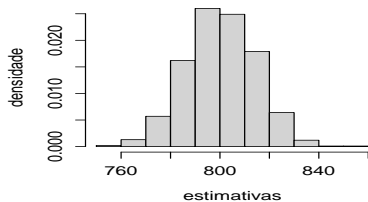


quantil da distribuição padronizada do estimador

n = 30 , p-valor (teste-KS) = 0.9922

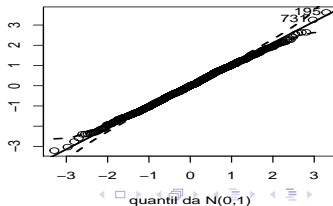


n = 50 , p-valor (teste-KS) = 0.9122



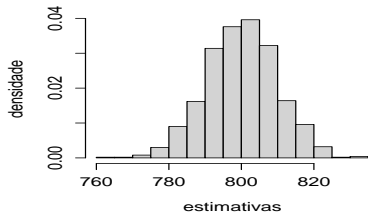
quantil da distribuição padronizada do estimador

n = 50 , p-valor (teste-KS) = 0.9122



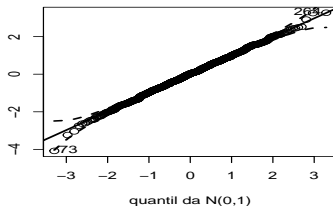
normal

n = 100 , p-valor (teste-KS) = 0.9932

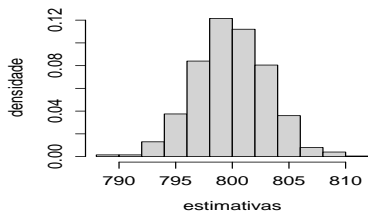


n = 100 , p-valor (teste-KS) = 0.9932

quantil da distribuição padronizada do estimador

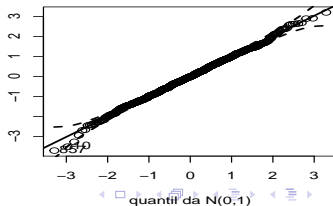


n = 1000 , p-valor (teste-KS) = 0.9863

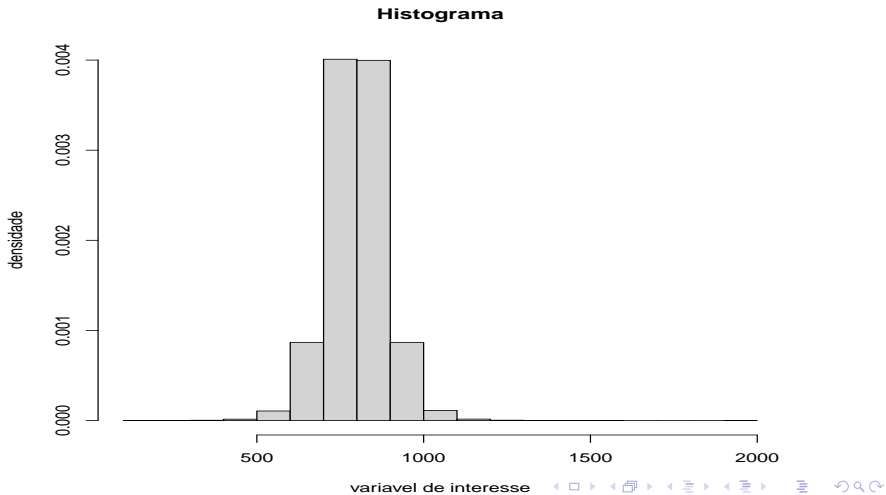


n = 1000 , p-valor (teste-KS) = 0.9863

quantil da distribuição padronizada do estimador

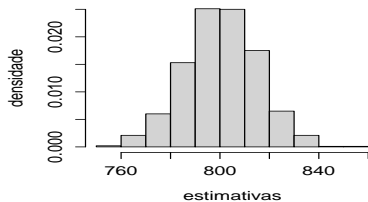


t de Student



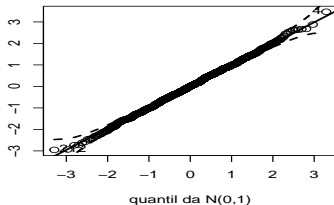
t de Student

n = 30 , p-valor (teste-KS) = 0.9304

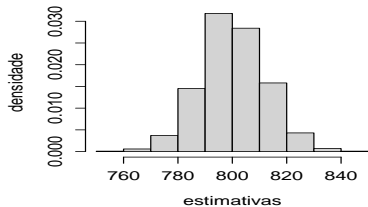


quantil da distribuição padronizada do estimador

n = 30 , p-valor (teste-KS) = 0.9304

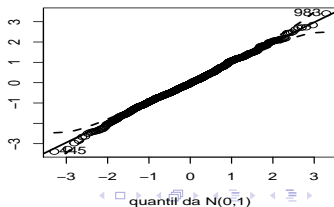


n = 50 , p-valor (teste-KS) = 0.7541



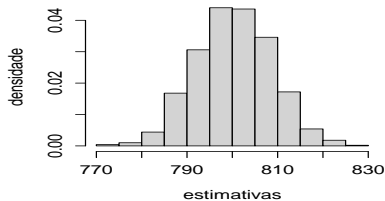
quantil da distribuição padronizada do estimador

n = 50 , p-valor (teste-KS) = 0.7541



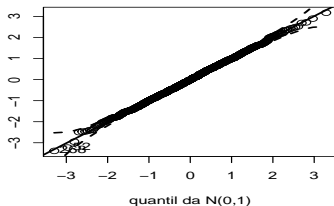
t de Student

n = 100 , p-valor (teste-KS) = 0.9826

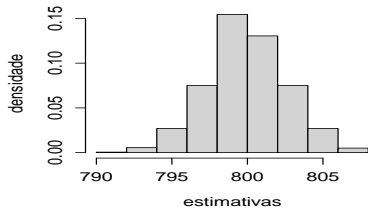


quantil da distribuição padronizada do estimador

n = 100 , p-valor (teste-KS) = 0.9826

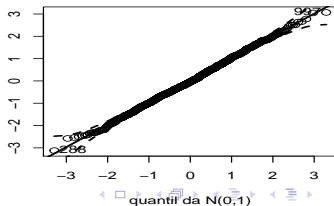


n = 1000 , p-valor (teste-KS) = 0.6335

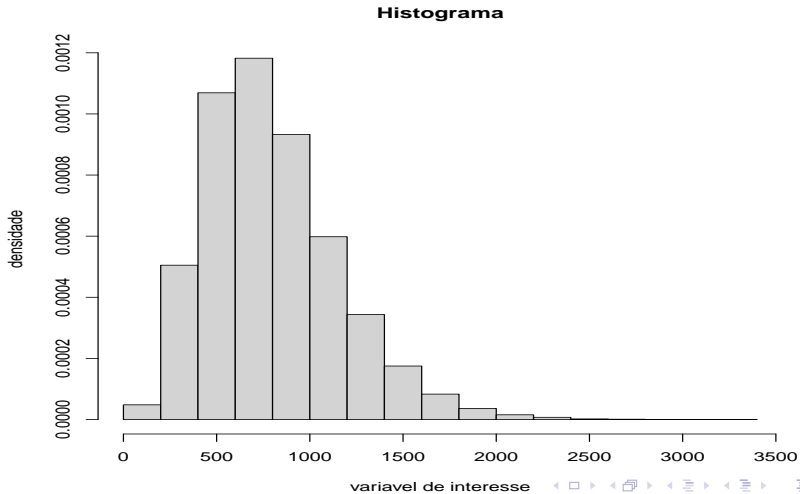


quantil da distribuição padronizada do estimador

n = 1000 , p-valor (teste-KS) = 0.6335

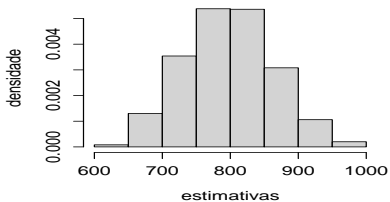


gama



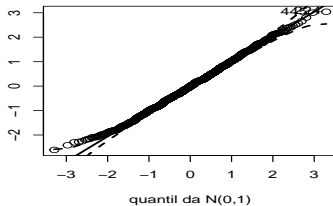
gama

n = 30 , p-valor (teste-KS) = 0.8637

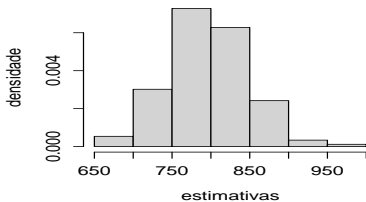


n = 30 , p-valor (teste-KS) = 0.8637

quantil da distribuição padronizada do estimador

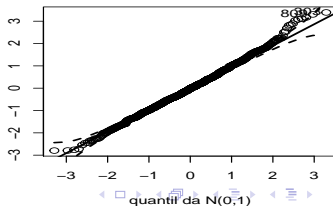


n = 50 , p-valor (teste-KS) = 0.6977



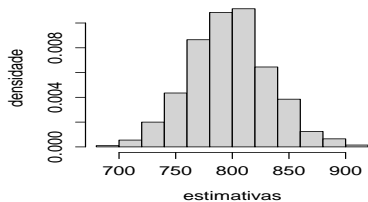
n = 50 , p-valor (teste-KS) = 0.6977

quantil da distribuição padronizada do estimador



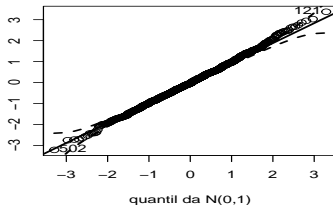
gama

n = 100 , p-valor (teste-KS) = 0.8282

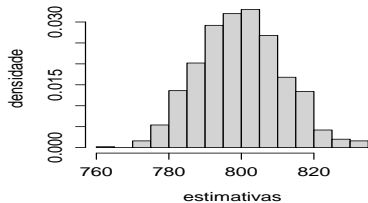


n = 100 , p-valor (teste-KS) = 0.8282

quantil da distribuição padronizada do estimador

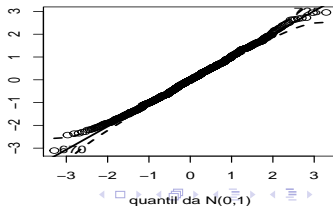


n = 1000 , p-valor (teste-KS) = 0.3915

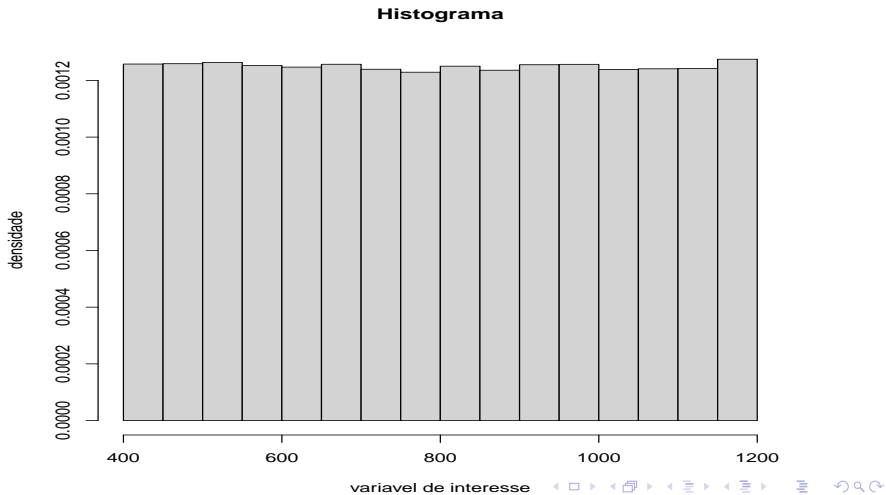


n = 1000 , p-valor (teste-KS) = 0.3915

quantil da distribuição padronizada do estimador

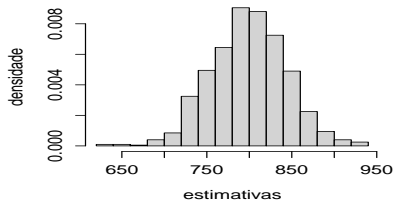


uniforme



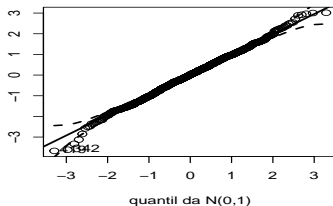
uniforme

n = 30 , p-valor (teste-KS) = 0.8445

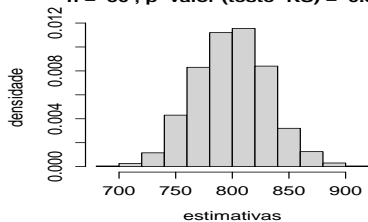


quantil da distribuição padronizada do estimador

n = 30 , p-valor (teste-KS) = 0.8445

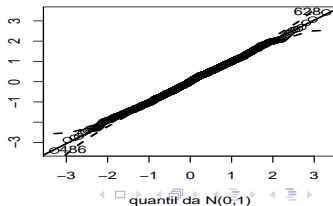


n = 50 , p-valor (teste-KS) = 0.539



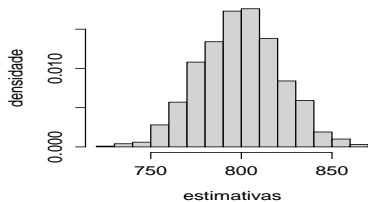
quantil da distribuição padronizada do estimador

n = 50 , p-valor (teste-KS) = 0.539



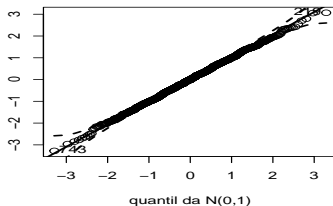
uniforme

n = 100 , p-valor (teste-KS) = 0.9864

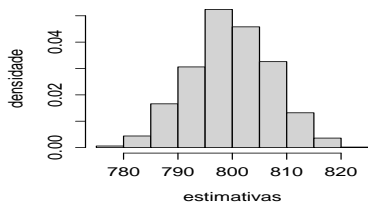


quantil da distribuição padronizada do estimador

n = 100 , p-valor (teste-KS) = 0.9864

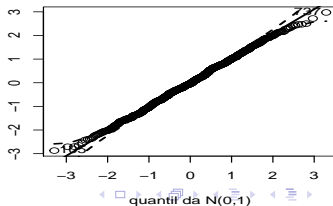


n = 1000 , p-valor (teste-KS) = 0.6529

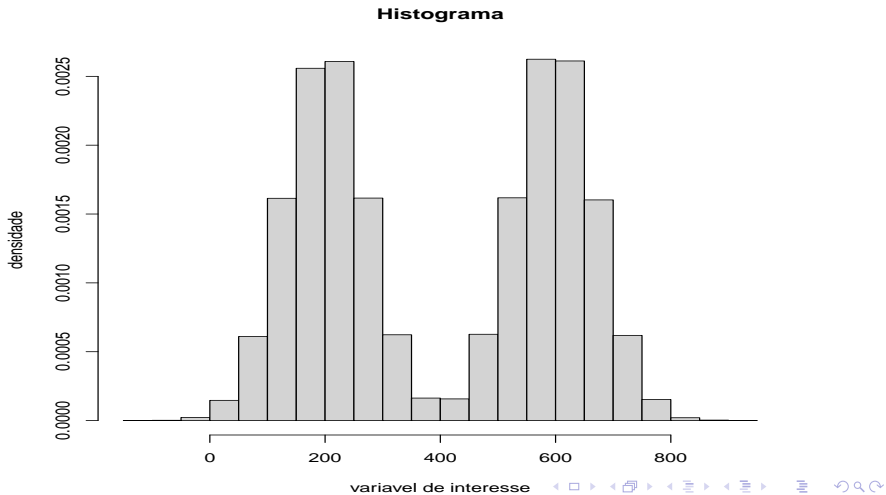


quantil da distribuição padronizada do estimador

n = 1000 , p-valor (teste-KS) = 0.6529

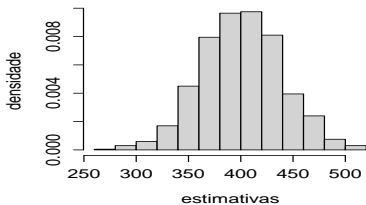


mistura de duas normais



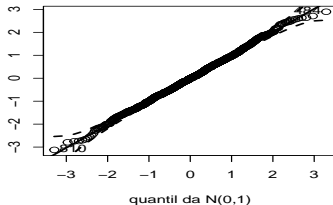
mistura de duas normais

n = 30 , p-valor (teste-KS) = 0.981

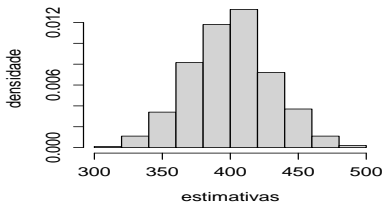


n = 30 , p-valor (teste-KS) = 0.981

quantil da distribuição padronizada do estimador

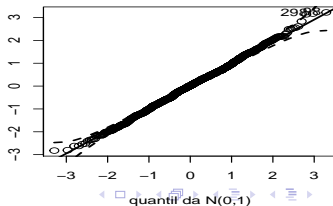


n = 50 , p-valor (teste-KS) = 0.7892



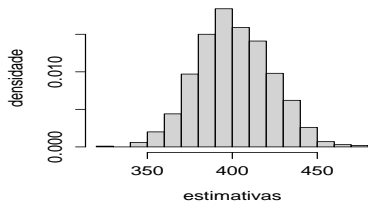
n = 50 , p-valor (teste-KS) = 0.7892

quantil da distribuição padronizada do estimador



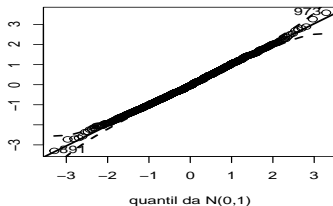
mistura de duas normais

n = 100 , p-valor (teste-KS) = 0.6017

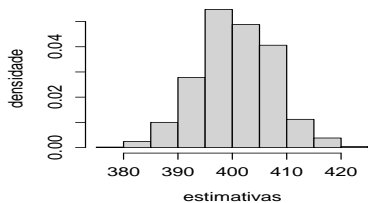


n = 100 , p-valor (teste-KS) = 0.6017

quantil da distribuição padronizada do estimador

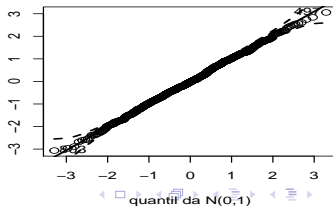


n = 1000 , p-valor (teste-KS) = 0.7449



n = 1000 , p-valor (teste-KS) = 0.7449

quantil da distribuição padronizada do estimador



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: $z_\gamma \sqrt{\frac{(1-f)s^2}{n}}$. Fixa-se um erro de estimativa de interesse.
- Precisão - Probabilidade do módulo da diferença:
 $P(|\hat{\mu} - \mu| < \delta) > \gamma, \delta > 0, \gamma \in (0, 1)$.

Determinação do tamanho amostral: erro da estimativa

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

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

$$\begin{aligned} P_{A_2} (|\hat{\mu} - \mu| < \delta) > \gamma &\Leftrightarrow P_{A_2} \left(\left| \frac{\hat{\mu} - \mu}{\sqrt{(1-f)\sigma^2/n}} \right| < \frac{\sqrt{n}\delta}{\sigma} \right) > \gamma \\ &\Leftrightarrow P_{A_2} \left(|Z| < \frac{\sqrt{n}\delta}{\sqrt{1-f}\sigma} \right) > \gamma \Leftrightarrow \frac{\sqrt{n}\delta}{\sqrt{1-f}\sigma} = z_\gamma \end{aligned}$$

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

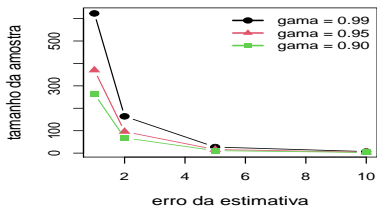
Os tamanhos amostrais, sob os planos amostrais A_1 e A_2 são dados, respectivamente, por $n_{A_1} = \frac{1}{\frac{\delta^2}{z_\gamma^2 \sigma^2}}$ e $n_{A_2} = \frac{1}{\frac{\delta^2}{s^2 z_\gamma^2} + \frac{1}{N}}$. Assim, nota-se que $n_{A_1} \geq n_{A_2}$.

Tamanhos amostrais

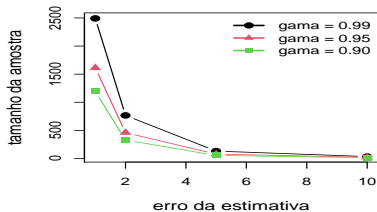
- Situações hipotéticas: cruzamento entre os níveis de diferentes fatores de interesse
 - $\delta \in \{1, 2, 5, 10\}$.
 - $\gamma \in \{0, 9; 0, 95; 0, 99\}$.
 - $s^2 \in \{100, 500, 1000, 10000\}$.

Tamanhos amostrais

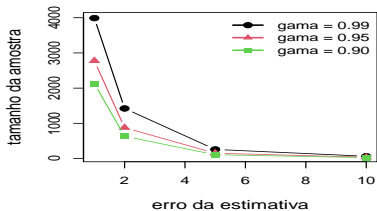
$s^2 = 100$, $N = 10000$



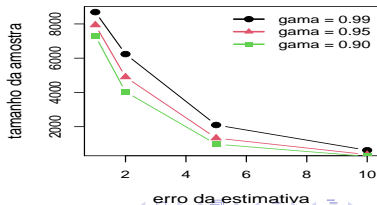
$s^2 = 500$, $N = 10000$



$s^2 = 1000$, $N = 10000$

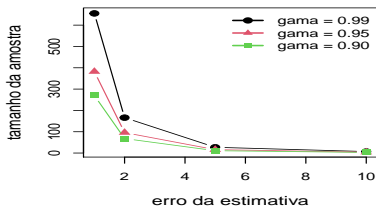


$s^2 = 10000$, $N = 10000$

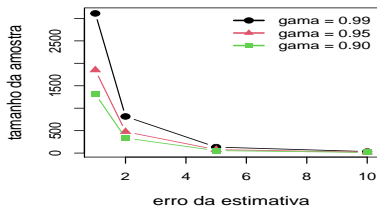


Tamanhos amostrais

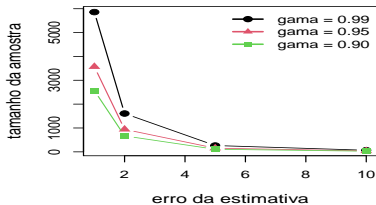
$s^2 = 100$, $N = 50000$



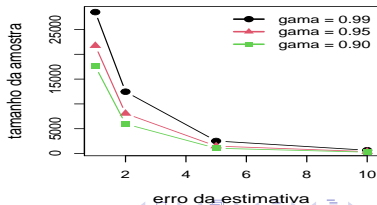
$s^2 = 500$, $N = 50000$



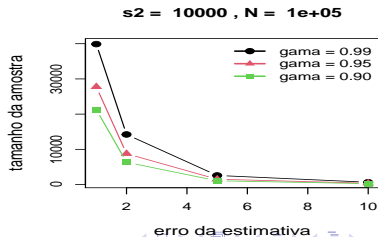
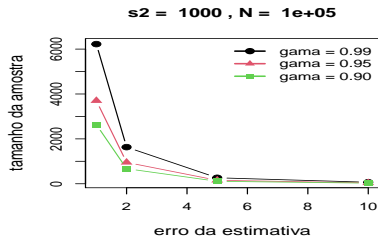
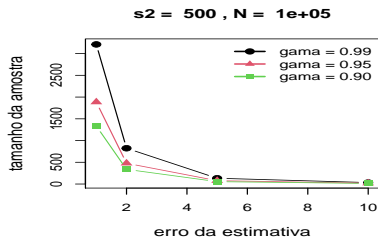
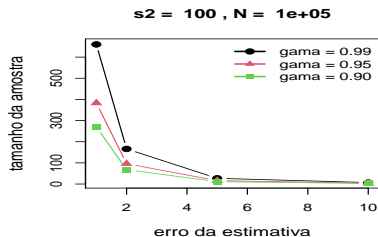
$s^2 = 1000$, $N = 50000$



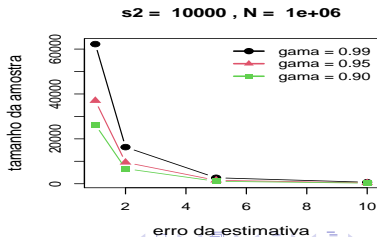
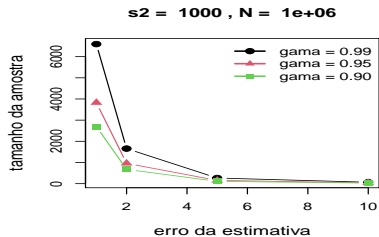
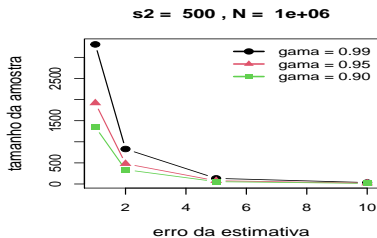
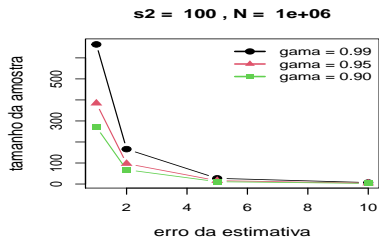
$s^2 = 10000$, $N = 50000$



Tamanhos amostrais

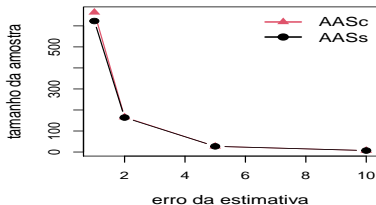


Tamanhos amostrais

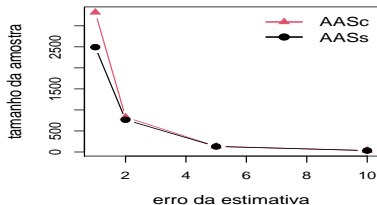


Comparação dos tamanhos amostrais ($\gamma = 0,99$)

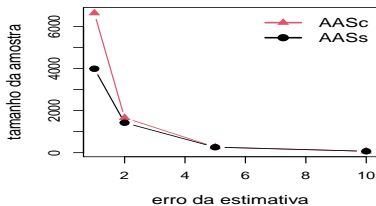
$\sigma^2 = 99.99$, $s^2 = 100$, $N = 10000$



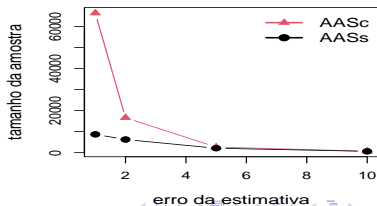
$\sigma^2 = 499.95$, $s^2 = 500$, $N = 10000$



$\sigma^2 = 999.9$, $s^2 = 1000$, $N = 10000$

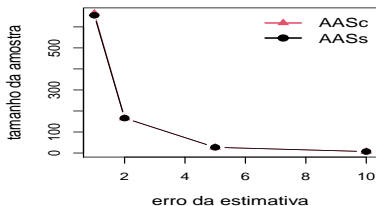


$\sigma^2 = 9999$, $s^2 = 10000$, $N = 10000$

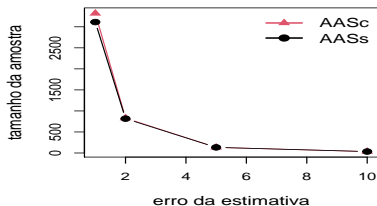


Comparação dos tamanhos amostrais ($\gamma = 0,99$)

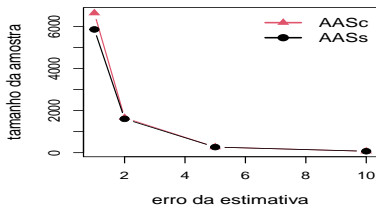
$\sigma^2 = 99.998$, $s^2 = 100$, $N = 50000$



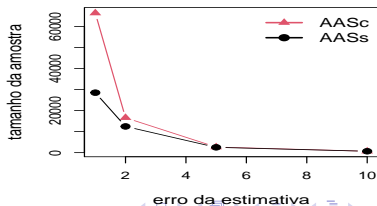
$\sigma^2 = 499.99$, $s^2 = 500$, $N = 50000$



$\sigma^2 = 999.98$, $s^2 = 1000$, $N = 50000$

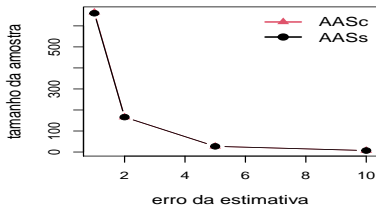


$\sigma^2 = 9999.8$, $s^2 = 10000$, $N = 50000$

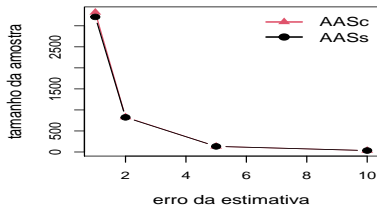


Comparação dos tamanhos amostrais ($\gamma = 0,99$)

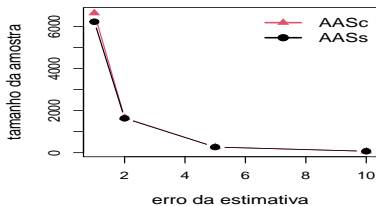
$\sigma^2 = 99.999$, $s^2 = 100$, $N = 1e+05$



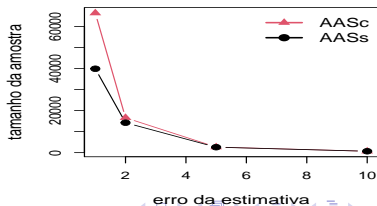
$\sigma^2 = 499.995$, $s^2 = 500$, $N = 1e+05$



$\sigma^2 = 999.99$, $s^2 = 1000$, $N = 1e+05$

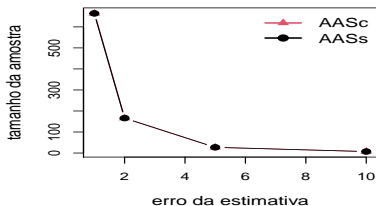


$\sigma^2 = 9999.9$, $s^2 = 10000$, $N = 1e+05$

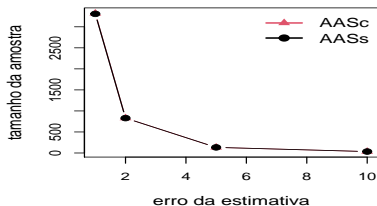


Comparação dos tamanhos amostrais ($\gamma = 0,99$)

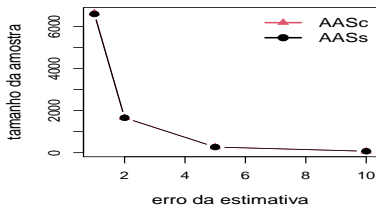
$\sigma^2 = 99.9999$, $s^2 = 100$, $N = 1e+06$



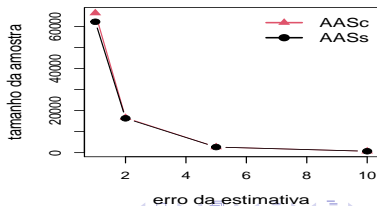
$\sigma^2 = 499.9995$, $s^2 = 500$, $N = 1e+06$



$\sigma^2 = 999.999$, $s^2 = 1000$, $N = 1e+06$



$\sigma^2 = 9999.99$, $s^2 = 10000$, $N = 1e+06$



Estimação do total populacional

- $\tau = \sum_{i=1}^N y_i = N\mu.$
- Estimador “natural”: $\hat{\tau}_u = \sum_{i=1}^n Y_i.$ Problema: se os y_i 's foram positivos, $\hat{\tau}_u$ sempre subestimar \hat{a} $\tau.$
- Alternativa $\hat{\tau} = N\hat{\mu}.$
- Estimativa $\tilde{\tau} = N\tilde{\mu}$

Propriedades do estimador

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

Propriedades do estimador

- Normalidade assintótica, como

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

lembrando que N é fixo, temos que

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

Intervalo de Confiança

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

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

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

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

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 $H_1 : \tau > \tau_0$.

3 $H_0 : \tau = \tau_0$ vs $H_1 : \tau \neq \tau_0$.

■ Estatística do teste $Z_t = \frac{\hat{\tau} - \tau_0}{N\sqrt{(1-f)\hat{s}/\sqrt{n}}}$, em que $\hat{s} = \sqrt{\hat{s}^2}$.

■ Sob H_0 , vimos que $Z_t \approx N(0, 1)$, para n e $N-n$ suficientemente grandes.

■ Defina $z_t = \frac{\tilde{\tau} - \tau_0}{N\sqrt{(1-f)\tilde{s}/\sqrt{n}}}$ 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

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

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.