Tamanho de amostra

Prof. Paulo Justiniano Ribeiro Jr

Departamento de Estatística
Universidade Federal do Paraná

Por que dimensionar amostras?

- Dimensionar esforço, economizar recursos.
- Planejar pesquisas de opinião pública.
- Controle de qualidade.
- Estudos demográficos.
- Inspeções de qualidade de água.
- Volume de madeira em florestas cultivadas.
- Biodiversidade.
- Estoque de peixes.
- Testes de medicamentos.

Figura 1. Foto de fauxels no Pexels.

Exemplos básicos

- Exemplos (simples) discutidos aqui
    1. Tamanho da amostra para estimar a média.
    2. Tamanho da amostra para estimar a proporção.
    3. Tamanho da amostra para estimar a variância.
- Outros contextos
- Teoria e prática

Figura 2. Foto de Karolina Grabowska no Pexels.

Tamanho da amostra para a média

Tamanho da amostra para estimar a média

Nosso objetivo é coletar dados para estimar a **média populacional** μ.

A questão é:

Quantos elementos (itens, objetos, pessoas, ...) devemos amostrar?

A resposta pode ser:

Uma quantidade que permita obter estimativas com uma incerteza aceitável.

Amplitude do intervalo de confiança para a média

O intervalo de confiança para média é

$$\text{IC}(\mu): \left(\overline{y} - z_{\alpha/2} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) ; \overline{y} + z_{\alpha/2} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)\right).$$

Logo, a **amplitude** do intervalo dada pela diferença entre o limite superior e inferior é

$$A_{\text{IC}(\mu)} = 2 \cdot z_{\alpha/2} \cdot (\sigma/\sqrt{n}).$$

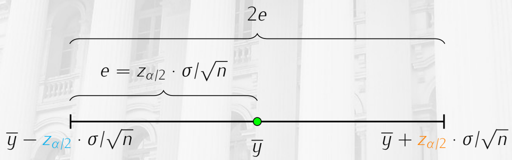


Figura 3. Margem de erro do intervalo de confiança para a média.

Componentes do intervalo de confiança para a média

A amplitude do intervalo de confiança depende de três componentes:

$$A_{\text{IC}(\mu)} = 2 \cdot z_{\alpha/2} \cdot (\sigma/\sqrt{n})$$

1. Coeficiente de confiança $1 - \alpha$, que determina o valor crítico $z_{\alpha/2}$.
2. Desvio-padrão populacional σ.
3. Tamanho da amostra n.

Efeitos na amplitude do intervalo de confiança para a média

$$\uparrow 1-\alpha \Rightarrow \uparrow z_{\alpha/2} \Rightarrow \uparrow \mathrm{A}_{\mathrm{IC}(\mu)}$$

$$\uparrow \sigma \Rightarrow \uparrow \mathrm{A}_{\mathrm{IC}(\mu)}$$

$$\mathrm{A}_{\mathrm{IC}(\mu)} = 2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\uparrow n \Rightarrow \downarrow \mathrm{A}_{\mathrm{IC}(\mu)}$$

1. $z_{\alpha/2}$: cada vez que aumentamos a confiança $1-\alpha$, o valor de $z_{\alpha/2}$ fica maior e, consequentemente, a amplitude do intervalo aumenta.
2. σ: um grande desvio padrão indica a possibilidade de um considerável distanciamento dos valores amostrais em relação à média populacional.
3. n: quanto maior for o tamanho da amostra, maior será a quantidade de informação disponível. Com isso, valores maiores de n produzem intervalos mais informativos (estreitos).

Invertendo a equação da margem de erro

A partir da equação do **erro máximo provável**,

$$e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

e fixando e, podemos obter n a partir da seguinte equação

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{e} \right)^2.$$

Tamanho da amostra para estimar a média

Note que, em

$\uparrow 1-\alpha \Rightarrow \uparrow z_{\alpha/2} \Rightarrow \uparrow n$

$\uparrow \sigma \Rightarrow \uparrow n$

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{e}\right)^2$$

$\uparrow e \Rightarrow \downarrow n$

- O tamanho da amostra (n) depende do
  - **nível de confiança** ($1-\alpha$) desejado (expresso pelo valor crítico $z_{\alpha/2}$).
  - **desvio-padrão** (σ) (embora veremos que não é estritamente necessário).
  - **erro máximo admitido** (e).
- Como o tamanho da amostra precisa ser um número inteiro, usamos o número inteiro logo acima, denotado por $\lceil n \rceil$.

Exercício: cálculo de tamanho de amostra para a média

Considere uma característica $Y \sim \mathrm{N}(\mu, \sigma^2 = 36)$.

1. Calcule o tamanho da amostra, para que com 95% de probabilidade, a média amostral não difira da média populacional por mais de
   a) 0.5 unidades e 2 unidades.
   b) Qual o impacto do erro máximo assumido para o tamanho da amostra?

Solução do item 1

- Temos $\sigma = 6$, $e = 0.5$ e $z_{0.025} = 1.96$. Assim,

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{e}\right)^2 = \left(\frac{1.96 \cdot 6}{0.5}\right)^2 \approx 554.$$

- Temos $\sigma = 6$, $e = 2$ e $z_{0.025} = 1.96$. Assim,

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{e}\right)^2 = \left(\frac{1.96 \cdot 6}{2}\right)^2 \approx 35.$$

- Quanto menor o erro admitido, maior o tamanho da amostra.

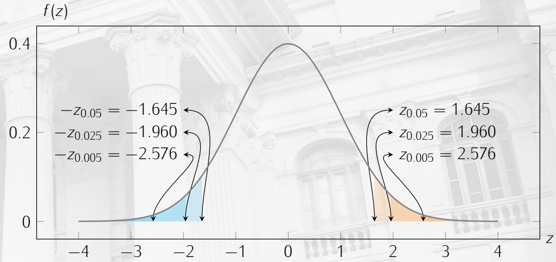


Figura 4. Quantis da distribuição Normal Padrão.

Exercício: cálculo de tamanho de amostra para a média

Considere uma característica $Y \sim \mathrm{N}(\mu, \sigma^2 = 36)$.

1. Calcule o tamanho da amostra, para que com 95% de probabilidade, a média amostral não difira da média populacional por mais de
   a) 0.5 unidades e 2 unidades.
   b) Qual o impacto do erro máximo assumido para o tamanho da amostra?
2. **Calcule o tamanho da amostra, para que a diferença da média amostral para a média populacional (em valor absoluto) seja menor ou igual a 2 unidades, com níveis de confiança de**
   a) **90% e 99%.**
   b) **Compare as estimativas do item anterior e analise o impacto do nível de confiança para a determinação do tamanho amostral.**

Solução do item 2

- Temos $\sigma = 6$, $e = 2$ e $z_{0.05} = 1.645$. Assim,

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{e}\right)^2 = \left(\frac{1.645 \cdot 6}{2}\right)^2 \approx 25.$$

- Temos $\sigma = 6$, $e = 2$ e $z_{0.005} = 2.576$. Assim,

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{e}\right)^2 = \left(\frac{2.576 \cdot 6}{2}\right)^2 \approx 60.$$

- Quanto maior o nível de confiança, maior o tamanho da amostra.

Quando a variância é desconhecida

Se σ for desconhecido?

1. Estime o valor de σ com base em algum estudo feito anteriormente.
2. Faça uma amostra piloto e estime o desvio-padrão amostral s, e use-o como uma aproximação para o desvio-padrão populacional σ.
3. Use a **regra empírica da amplitude** para dados com distribuição (aproximadamente) Normal.

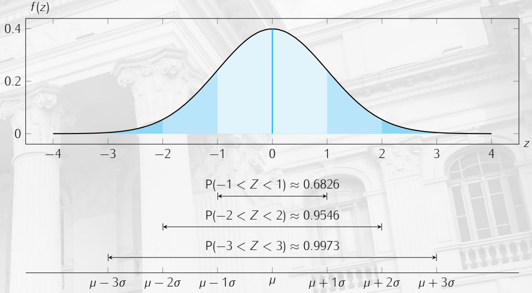


Figura 5. Áreas simétricas na distribuição Normal para a regra empírica.

A regra empírica da amplitude

Definem-se como **valores usuais** aqueles que são típicos (não extremos).

Como sabemos que em uma distribuição (aproximadamente) Normal praticamente 95% dos valores encontram-se a 2 desvios-padrões acima e abaixo da média, temos que

$$
\begin{aligned}
4\sigma &= (\mu + 2\sigma) - (\mu - 2\sigma) \\
4\sigma &= Y_{(n)} - Y_{(1)} \\
\tilde{\sigma} &= \frac{Y_{(n)} - Y_{(1)}}{4}
\end{aligned}
$$

pode ser utilizado como um estimador para σ. $Y_{(n)}$ é maior valor da amostra e $Y_{(1)}$ é o menor.

Exercício: gasto com educação dos filhos

Um cientista social deseja estimar o gasto mensal com educação dos filhos nas famílias de uma cidade. Quantas famílias devem ser selecionadas para termos 90% de confiança que a média amostral esteja a menos de R$ 30.00 da média populacional? Apura-se que gastos estão entre R$ 800.00 e R$ 1200.00. Assume-se que isto ocorre 95% das vezes.

Figura 6. Foto de August de Richelieu no Pexels.

Solução

Para confiança de $1 - \alpha$ de 0.90,
temos que $z_{\alpha/2} = 1.645$, $e = 30$ e o desvio padrão pode ser aproximado por

$$\tilde{\sigma} = (1200 - 800)/4 = 100.$$

Usando a equação apresentada, temos

$$n = \left(\frac{1.645 \cdot 100}{30} \right)^2 \approx 31.$$

Tamanho da amostra para proporção

Tamanho da amostra para estimar a proporção

Seguindo o mesmo raciocínio do tamanho de amostra para a média, a partir da equação do erro máximo provável para a distribuição amostral (aproximada pela Normal) da proporção,

$$e = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}},$$

podemos isolar n e chegar à seguinte equação

$$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 \cdot p(1-p).$$

Quando não conhecemos p, usamos $\hat{p}$ (estimativa otimista) ou $p = 0.5$ (estimativa conservadora) como valores para p.

Exemplo: proporção de troncos defeituosos

Um engenheiro florestal deseja estimar a verdadeira proporção de troncos defeituosos (impróprios para a marcenaria por terem rachaduras ou nós), com um erro máximo de 3% e nível de confiança de 99%. Calcule o tamanho da amostra necessário para se estimar esta proporção se:

1. O engenheiro tem uma estimativa de que, em uma amostra anterior, aproximadamente 10% dos troncos eram defeituosos.
2. O fabricante não tem nenhuma informação prévia sobre a proporção de troncos defeituosos.

Figura 7. Foto de Pexels.

Solução

1. Temos $p = 0.1$, $e = 0.03$ e $z_{\alpha/2} = z_{0.005} = 2.576$. Assim,

$$n = \left(\frac{2.576}{0.03}\right)^2 0.1 \cdot (1 - 0.1) \approx 664.$$

2. Temos $p = 0.5$, $e = 0.03$ e $z_{\alpha/2} = z_{0.005} = 2.576$. Assim,

$$n = \left(\frac{2.576}{0.03}\right)^2 0.5 \cdot (1 - 0.5) \approx 1844.$$

Tamanho da amostra para a variância

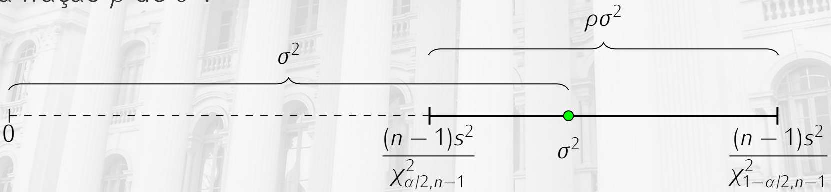
Tamanho da amostra para a variância

- O intervalo de confiança para σ^2 é

$$\text{IC}_{1-\alpha}(\sigma^2) = \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \right),$$

  em que $\chi^2_{\alpha/2,n-1}$ e $\chi^2_{1-\alpha/2,n-1}$ são os valores da cauda direita e esquerda da distribuição χ^2 com $n-1$ graus de liberdade.

- Queremos determinar o tamanho da amostra n para que a amplitude do intervalo seja uma fração ρ de σ^2.



Tamanho da amostra para a variância

- Note que neste caso queremos um intervalo de amplitude tal que satisfaça

$$A_{IC\sigma^2} = \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} - \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \leq \rho s^2.$$

- Simplificando os termos, precisamos encontrar n tal que

$$(n-1)\left(\frac{1}{\chi^2_{\alpha/2,n-1}} - \frac{1}{\chi^2_{1-\alpha/2,n-1}}\right) \leq \rho.$$

- Neste caso, a margem de erro é definida como um percentual de σ^2.
- Após simplificação, o valor particular de s^2 não entra diretamente no cálculo.
- Neste caso não é possível expressar n por uma fórmula. Cálculo feito por algum algoritmo.

Curvas de tamanho amostra

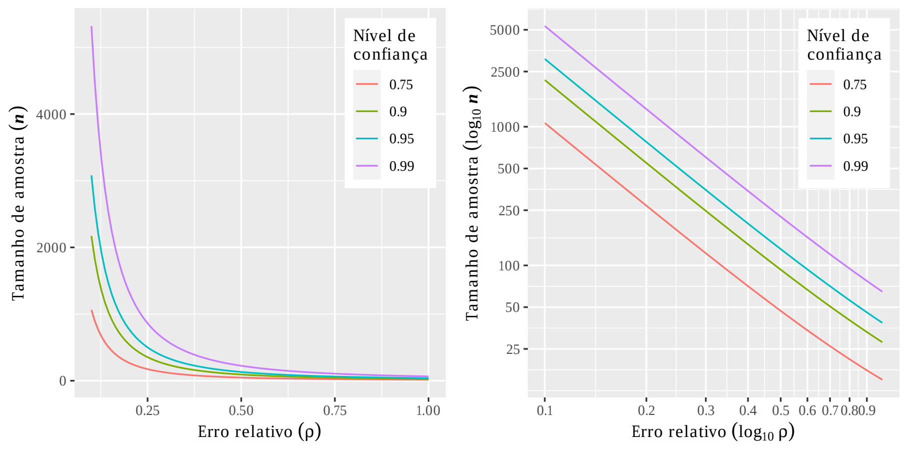


Figura 8. Curvas para determinar o tamanho da amostra para estimação da variância.

Considerações práticas sobre tamanho de amostra

- Para média e proporção, é simples determinar o tamanho de amostra.
- Para outros parâmetros populacionais, pode não ser de fácil obtenção.
- Em algumas situações é possível empregar simulação computacional para determinar tamanho de amostra.
- Para esquemas complexos de amostragem ou delineamentos experimentais, todas as características do plano amostral/experimental devem ser consideradas.
- Quase sempre os tamanhos de amostra determinados superam a capacidade logística/operacional disponível para a sua execução.