MAT02025 - Amostragem 1

Uso da distribuição normal, viés e seus efeitos

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2022

O uso da distribuição normal

O uso da distribuição normal

- Um estimador $\hat{\theta}$ de θ dado por um plano de amostragem é denominado **imparcial**[1] se o valor médio de $\hat{\theta}$, tomado em todas as amostras possíveis fornecidas pelo plano, for igual a θ.
- Na notação da aula anterior, esta condição pode ser escrita como:

$$\mathrm{E}\left(\hat{\theta}\right) = \sum_{j=1}^{V} \pi_j \hat{\theta}_j = \theta,$$

em que $\hat{\theta}_j$ é a estimativa dada pela j-ésima amostra.

[1] Ou **não-tendencioso**, ou **não-enviesado**. **Exercício:** no exercício da aula passada, mostre que a média amostral é um estimador não enviesado para a média populacional.

O uso da distribuição normal

- As amostras nos levantamentos são frequentemente grandes o suficiente para que as estimativas feitas a partir delas tenham **distribuição aproximadamente normal**.
- Além disso, com a amostragem probabilística, **temos fórmulas** que fornecem a **média** e a **variância das estimativas**.

O uso da distribuição normal

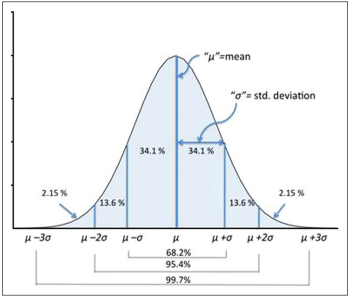
- Suponha que tenhamos obtido **uma amostra** por um procedimento conhecido por fornecer um estimador imparcial e calculado **a estimativa da amostra** $\hat{\theta}$ e seu desvio padrão[2] $\sigma_{\hat{\theta}}$.
  - **Qual a qualidade da estimativa?**
- Não podemos saber o valor exato do erro **dessa estimativa** $(\hat{\theta} - \theta)$, mas, . . .

[2] Frequentemente chamado de **erro padrão**.

O uso da distribuição normal

... a partir das **propriedades da curva normal**, as chances são

- 0,32 (cerca de 1 em 3) que o erro absoluto $|\hat{\theta} - \theta|$ excede $\sigma_{\hat{\theta}}$.
- 0,05 (1 em 20) que o erro absoluto $|\hat{\theta} - \theta|$ excede $1{,}96\sigma_{\hat{\theta}} \approx 2\sigma_{\hat{\theta}}$.
- 0,01 (1 em 100) que o erro absoluto $|\hat{\theta} - \theta|$ excede $2{,}58\sigma_{\hat{\theta}}$.



O uso da distribuição normal

- Por exemplo, se uma amostra probabilística dos registros de baterias em uso rotineiro em uma grande fábrica mostrar uma vida média $\hat{\mu} = 394$ dias, com um erro padrão $\sigma_{\hat{\mu}} = 4,6$ dias, as chances são de 99 em 100 de que a vida média da população de baterias esteja entre

$$\hat{\mu}_I = 394 - (2,58)(4,6) = 382 \text{ dias e } \hat{\mu}_S = 394 + (2,58)(4,6) = 406 \text{ dias.}$$

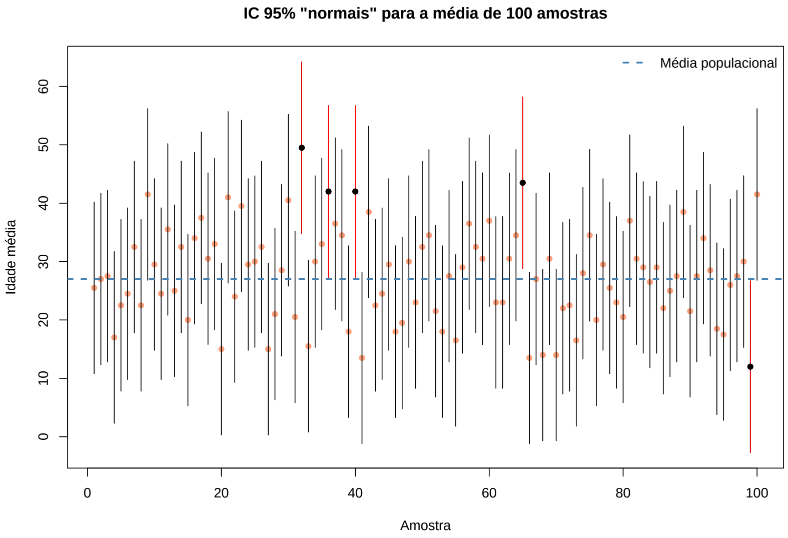
O uso da distribuição normal

- Os limites, 382 dias e 406 dias, são chamados de limites de confiança inferior e superior.
- Com **uma estimativa única de um único levantamento**, a afirmação **"μ está entre 382 e 406 dias"** **não é, necessariamente, correta**.
- O número de **"99% de confiança"** implica que **se o mesmo plano de amostragem fosse usado muitas vezes em uma população**, e **uma declaração de confiança sendo feita para cada amostra**, **cerca de 99% dessas declarações estariam corretas** e **1% erradas**.

O uso da distribuição normal

A verificação prática de que aproximadamente a proporção declarada de afirmações está correta contribui muito para educar e tranquilizar os pesquisadores (administradores e/ou tomadores de decisão) sobre a natureza da amostragem.

O uso da distribuição normal



O uso da distribuição normal

- O exemplo anterior assume que $\sigma_{\hat{\mu}}$ (ou mais genericamente, $\sigma_{\hat{\theta}}$) é conhecido.
- Na verdade, $\sigma_{\hat{\mu}}$, como $\hat{\mu}$, está sujeito a um erro de amostragem.
- Com uma **variável normalmente distribuída**, as tabelas da distribuição t de Student são usadas em vez das tabelas normais para calcular os limites de confiança para μ quando a amostra é pequena.
- A substituição da tabela normal pela tabela t faz quase nenhuma diferença se o número de graus de liberdade em $\sigma_{\hat{\mu}}$ exceder 50.

Viéses e seus efeitos

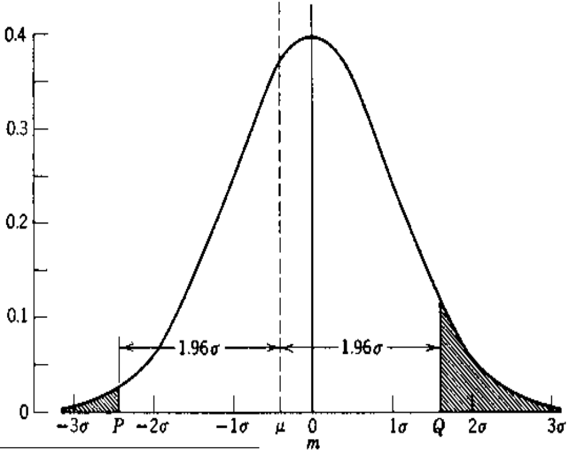
Viéses e seus efeitos

Na teoria dos levantamentos por amostragem, é necessário considerar **estimadores enviesados** por duas razões.

1. Em alguns dos problemas mais comuns, particularmente na estimativa de razões (índices), estimadores convenientes e adequados são considerados tendenciosos.
2. Mesmo com estimadores que são imparciais na amostragem probabilística, os **erros de medição** e **não resposta** podem produzir vieses nos resultados calculados nos dados.
    - Isso acontece, por exemplo, se as pessoas que se recusam a ser entrevistadas quase todas se opõem a algum gasto de fundos públicos, ao passo que as que são entrevistadas se dividem igualmente a favor e contra.

Viéses e seus efeitos

- Suponha que a estimativa $\hat{\mu}$ seja **normalmente distribuída** em torno de uma média m que é uma distância B do valor real da população μ, como mostrado na figura[3].



[3] Ou seja, o estimador é enviesado, mas normalmente distribuído.

Viéses e seus efeitos

- A quantidade de viés[4] é $B = m - \mu$.
- Suponha que não saibamos que existe qualquer viés.
  - Calculamos o desvio padrão, σ (estamos usando σ no lugar de $\sigma_{\hat{\mu}}$), da distribuição de frequência da estimativa, que será, naturalmente,o desvio padrão sobre a média da distribuição, m, não sobre a verdadeira média μ.
- Como uma afirmação sobre a precisão da estimativa, declaramos que **a probabilidade é de 0,05** de que a estimativa $\hat{\mu}$ contenha um erro de mais de 1,96σ, ou seja,

$$\Pr(|\hat{\mu} - \mu| > 1,96\sigma) \approx 0,05.$$

[4] De maneira mais geral, $B(\hat{\theta}) = \mathrm{E}\left(\hat{\theta} - \theta\right) = \mathrm{E}\left(\hat{\theta}\right) - \theta$ é o definido como o viés (vício) do estimador $\hat{\theta}$. Se $B(\hat{\theta}) = 0$, então $\hat{\theta}$ é não enviesado.

Viéses e seus efeitos

- Examinaremos como a presença de viés distorce essa probabilidade. Note que

$$\Pr(\hat{\mu} - \mu > 1,96\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu+1,96\sigma}^{\infty} e^{-(\hat{\mu}-m)^2/2\sigma^2} d\hat{\mu}.$$

- Fazendo $t = (\hat{\mu} - m)/\sigma$, então o limite inferior do intervalo de integração de t é

$$\frac{\mu - m}{\sigma} + 1,96 = 1,96 - \frac{B}{\sigma}.$$

Viéses e seus efeitos

Assim,

$$\Pr(\hat{\mu} - \mu > 1,96\sigma) = \frac{1}{\sqrt{2\pi}} \int_{1{,}96-(B/\sigma)}^{\infty} e^{-t^2/2} dt.$$

E semelhantemente, temos

$$\Pr(\hat{\mu} - \mu < -1,96\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1{,}96-(B/\sigma)} e^{-t^2/2} dt.$$

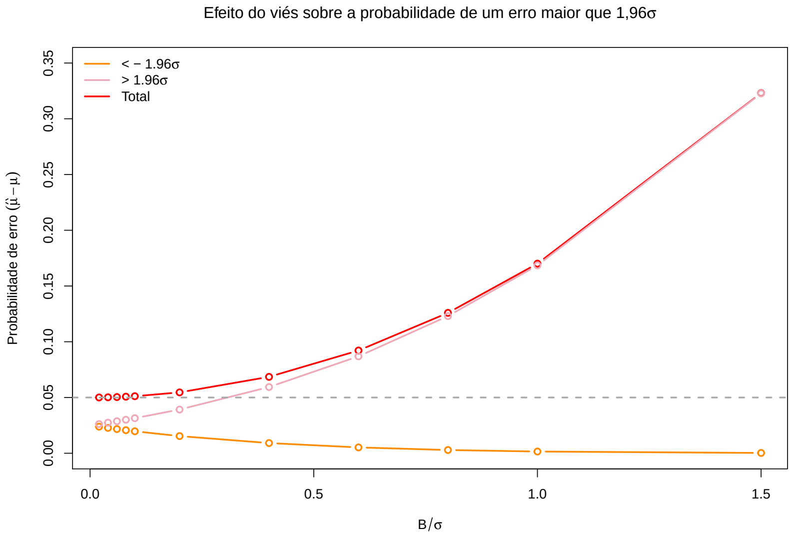
Viéses e seus efeitos

- A quantidade de perturbação depende, unicamente, da razão entre o viés e o desvio padrão.

Table 1: Efeito do viés sobre a probabilidade de um erro maior que $1,96\sigma$

B/σ	$< -1,96\sigma$	$> 1,96\sigma$	Total
0.02	0.0239	0.0262	0.0500
0.04	0.0228	0.0274	0.0502
0.06	0.0217	0.0287	0.0504
0.08	0.0207	0.0301	0.0507
0.10	0.0197	0.0314	0.0511
0.20	0.0154	0.0392	0.0546
0.40	0.0091	0.0594	0.0685
0.60	0.0052	0.0869	0.0921
0.80	0.0029	0.1230	0.1259
1.00	0.0015	0.1685	0.1701
1.50	0.0003	0.3228	0.3230

Viéses e seus efeitos



Comentários

- Para a probabilidade total de um erro maior que $1,96\sigma$, o viés tem pouca influência, desde que seu valor seja inferior a $1/10$ do desvio padrão.
  - $B = \sigma/10$, a probabilidade total é 0,0511, em vez de 0,05.
- A medida que o viés aumenta, a perturbação se torna mais grave.
  - $B = \sigma$, a probabilidade total de erro será 0,1701 $> 3\times$ 0,05.

Comentários

- As duas caudas são afetadas de forma diferente.
- Com um viés positivo, como neste exemplo, a probabilidade de uma subestimativa em mais de 1,96σ diminui rapidamente do presumido 0,025, para se tornar insignificante quando $B = \sigma$.
- A probabilidade da superestimação correspondente aumenta continuamente.

Comentários

- Como regra de trabalho, o efeito do viés sobre a precisão de uma estimativa é insignificante se o viés for menor que um décimo do desvio padrão da estimativa.
- Se tivermos um processo de estimativa enviesado, no qual $B/\sigma < 0{,}1$, em que B é o valor absoluto do viés, pode-se afirmar que o viés não é uma desvantagem considerável do processo.
- Mesmo com $B/\sigma = 0{,}2$, a perturbação na probabilidade do erro total é modesta.

Comentários

- Ao usar esses resultados, uma distinção deve ser feita entre as duas fontes de viés mencionadas no início desta seção.
- Com vieses do tipo que surgem na estimativa de razões, um limite superior para a razão B/σ pode ser encontrado matematicamente.
  - Se a amostra for grande o suficiente, podemos ter certeza de que B/σ não excederá 0,1.
- Com vieses causados por erros de medição ou não resposta, geralmente é impossível encontrar um limite superior garantido para B/σ que seja pequeno.

Para casa

- Estude a função `integrate` do `R` (material suplementar no Moodle), e avalie o efeito do viés quando B é negativo (e mesmas condições do exemplo).

Próxima aula

- Erro quadrático médio;
- Exercícios.

Por hoje é só!

Bons estudos!