MAT02025 - Amostragem 1

AAS: estimativa do erro padrão e intervalos de confiança

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, 2021

Estimativa do erro padrão de uma amostra

Estimativa do erro padrão

As fórmulas que nos dão os **erros padrões** das estimativas da **média** $(\overline{y} = (1/n)\sum_{i=1}^{n} Y_i)$ e **total** $(\hat{Y}_T = N\overline{y})$ populacionais são usadas principalmente para três finalidades:

1. comparar a precisão obtida por amostragem aleatória simples com a dada por outros métodos de amostragem;
2. para estimar o tamanho da amostra necessária em um levantamento que está sendo planejado;
3. para estimar a precisão realmente alcançada em um levantamento que foi concluído.

Estimativa do erro padrão

- As fórmulas envolvem S^2, a **variância da população**.
- Na prática, isso não será conhecido, mas pode ser estimado a partir dos dados da amostra.
  - Se vamos estimar S^2, não deveríamos fazer a mesma avaliação[1] que fizemos para $\overline{y}$ e $\hat{Y}_T$? *(Sim ou sim?)*

Estimativa do erro padrão

Teorema

Em uma amostra aleatórica simples,

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{y})^2}{n-1}$$

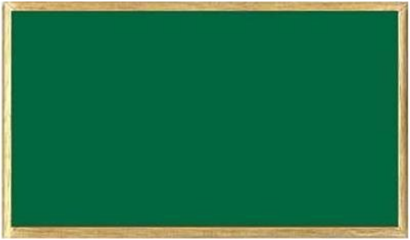
é uma estimativa (estimador) **não enviesada** de

$$S^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}.$$

Estimativa do erro padrão

Demonstração. Podemos escrever:

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}[(Y_i - \overline{Y}) - (\overline{y} - \overline{Y})]^2 \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - n(\overline{y} - \overline{Y})^2\right].
\end{aligned}
$$



Estimativa do erro padrão

Agora, calculamos a média de todas as amostras aleatórias simples de tamanho n:

$$
\begin{aligned}
\mathrm{E}\left(s^2\right) &= \mathrm{E}\left\{\frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i-\overline{Y})^2 - n(\overline{y}-\overline{Y})^2\right]\right\} \\
&= \frac{1}{n-1}\left\{\mathrm{E}\left[\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right] - n\mathrm{E}\left[(\overline{y}-\overline{Y})^2\right]\right\} \\
&= \frac{1}{n-1}\left[\frac{n}{N}\sum_{i=1}^{N}(Y_i-\overline{Y})^2 - n\mathrm{Var}\left(\overline{y}\right)\right] \\
&= \frac{1}{n-1}\left[\frac{n(N-1)}{N}S^2 - n\frac{S^2}{n}\frac{(N-n)}{N}\right] \\
&= \frac{S^2}{(n-1)N}\left[n(N-1)-(N-n)\right] = S^2.
\end{aligned}
$$

Estimativa do erro padrão

Corolário

As estimativas (estimadores) imparciais das variâncias de $\overline{y}$ e $\hat{Y}_T = N\overline{y}$ são:

$$\widehat{\mathrm{Var}}\,(\overline{y}) = s^2_{\overline{y}} = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{s^2}{n}(1-f),$$

$$\widehat{\mathrm{Var}}\,(\hat{Y}_T) = s^2_{\hat{Y}_T} = \frac{N^2 s^2}{n}\left(\frac{N-n}{N}\right) = \frac{N^2 s^2}{n}(1-f).$$

Para os erros padrões, teremos

$$s_{\overline{y}} = \frac{s}{\sqrt{n}}\sqrt{1-f}, \quad s_{\hat{Y}_T} = \frac{Ns}{\sqrt{n}}\sqrt{1-f}$$

- Essas estimativas apresentam um pequeno viés, mas, para a maioria das aplicações, o viés é desprezível.

Estimativa do erro padrão

Observação

- É preciso observar os símbolos empregados para a variância real e estimada dos estimadores. Assim, para $\overline{y}$, podemos escrever:
  - Variância real: $\text{Var}\left(\overline{y}\right) = \sigma^2_{\overline{y}}$
  - Variância estimada: $\widehat{\text{Var}}\left(\overline{y}\right) = s^2_{\overline{y}}$

[1] Ou seja, avaliar a imparcialidade deste estimador.

Intervalos de confiança

Intervalos de confiança

- É geralmente assumido que as estimativas $\overline{y}$ e $\hat{Y}_T$ são normalmente distribuídas em torno dos correspondentes valores populacionais.
- As razões para esta suposição e suas limitações são consideradas em aulas futuras.

Intervalos de confiança

- Se a suposição for válida, os **limites de confiança inferior** e **superior** para a média e o total da população são os seguintes:

Para o valor médio

$$\widehat{\overline{Y}}_I = \overline{y} - \frac{zs}{\sqrt{n}}\sqrt{1-f}, \quad \widehat{\overline{Y}}_S = \overline{y} + \frac{zs}{\sqrt{n}}\sqrt{1-f};$$

Para o valor total

$$\hat{Y}_{T_I} = N\overline{y} - \frac{zNs}{\sqrt{n}}\sqrt{1-f}, \quad \hat{Y}_{T_S} = N\overline{y} + \frac{zNs}{\sqrt{n}}\sqrt{1-f};$$

Intervalos de confiança

- O símbolo z é o valor do desvio normal, correspondente à desejada probabilidade de confiança. Seus valores mais comuns são:

Probabilidade de confiança (%)	50	80	90	95	99
Valor de z	0,67	1,28	1,64	1,96	2,58

Intervalos de confiança

Observação

- Se o tamanho da amostra é inferior a 60, os valores percentuais podem ser obtidos da **tabela t de Student** com $(n-1)$ **graus de liberdade**, sendo estes os graus de liberdade na variância estimada, s^2.
    - A distribuição t só se verifica, exatamente, quando as observações Y_i são elas próprias normalmente distribuídas e N é infinito.
    - Os desvios moderados da normalidade não a afetam grandemente.
- Para pequenas amostras, com distribuições muito assimétricas, são necessários processos especiais.

Exemplo

Exemplo

- As assinaturas de uma petição foram coletadas em 676 folhas.
- Cada folha tinha espaço suficiente para 42 assinaturas, mas em muitas folhas um número menor de assinaturas foi coletado.
- O número de assinaturas por folha foi contado em uma amostra aleatória de 50 folhas (cerca de 7% da amostra), com os resultados apresentados na tabela a seguir.

Y_i	n_i	Y_i	n_i
42	23	14	1
41	4	11	1
36	1	10	1
32	1	9	1
29	1	7	1
27	2	6	3
23	1	5	2
19	1	4	1
16	2	3	1
15	2	-	-

Exemplo

- O objetivo do levantamento amostral é estimar o **número total de assinaturas** para a petição e os limites de confiança de 80%.
- A **unidade de amostragem é a folha** e as **observações** Y_i **são os números de assinaturas por folha**.
- Observe que a distribuição original parece estar longe da normal, a maior frequência estando na extremidade superior.
- No entanto, há razões para acreditar, por experiência, que as médias das amostras de 50 unidades são aproximadamente normalmente distribuídas.

Exemplo

- Assim, temos

$$n = \sum n_i = 50, \quad \sum n_i Y_i = 1471, \quad \sum n_i Y_i^2 = 54497.$$

Portanto, o total estimado do número de assinaturas é

$$\hat{Y}_T = N\overline{y} = \frac{(676)(1471)}{50} = 19887,92 \approx 19888.$$

Para a variância da amostra, s^2, temos

$$s^2 = \frac{1}{n-1}\left[\sum n_i(Y_i - \overline{y})^2\right] = \frac{1}{n-1}\left[\sum n_i Y_i^2 - \frac{(\sum n_i Y_i)^2}{\sum n_i}\right]$$
$$= \frac{1}{49}\left[54497 - \frac{(1471)^2}{50}\right] = 228,9833 \approx 229.$$

Exemplo

Os limites de confiança de 80% são

$$19888 \pm \frac{zNs}{\sqrt{n}}\sqrt{1-f} = 19888 \pm \frac{(1,28)(676)(15,13)}{\sqrt{50}}\sqrt{1-0,0740}.$$

Isso fornece 18107 e 21669 para os limites de confiança de 80%[2].

- Uma contagem completa mostrou, neste caso, que havia 21045 assinaturas.

[2] Uma estimativa mais precisa (sem considerar arrendondamentos ao longo dos cálculos) é dada por 18104 e 21672.

Para casa

- Refaça as demonstrações da aula de hoje.
- Utilize o **Teorema** apresentado na aula para demonstrar o **Corolário**.
- Com os dados do exemplo da aula:
  - construa um intervalo de 95% de confiança para o número total de assinaturas do requerimento;
  - construa um intervalo de 90% de confiança para o número médio de assinaturas por folha do requerimento;
  - interprete os seus resultados.

Próxima aula

- O pacote `survey`.

Por hoje é só!

Bons estudos!