MAT02025 - Amostragem 1

AAS: validade da aproximação normal ou o TCL para população finita

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

Teorema Central do Limite para população finita

TCL para população finita

- Quando as observações individuais $Y_1, Y_2, \ldots, Y_n$ **não são normalmente distribuídas**, os níveis de confiança aproximados dos intervalos de confiança usuais dependem da **distribuição normal aproximada da média amostral** $\bar{y}$.
- Se $Y_1, Y_2, \ldots, Y_n$ são uma sequência de variáveis aleatórias **independentes** e **identicamente distribuídas** com média e variância finitas, a distribuição de

$$\frac{\bar{y} - \overline{Y}}{\sqrt{\mathrm{Var}\left(\bar{y}\right)}}$$

aproxima-se de uma **distribuição normal padrão** à medida que n **aumenta**, pelo **teorema central do limite (TCL)**.

- O resultado também é válido se a variância for substituída por um estimador razoável da variância.

TCL para população finita

- Quando uma população finita é amostrada usando **amostragem aleatória simples com reposição**, as n observações **são** de fato **independentes** e **identicamente distribuídas**, de modo que o **TCL** usual se aplica.
- No entanto, com a **amostragem aleatória simples sem reposição**, as observações da amostra **não são independentes**.
  - Selecionar uma unidade com um grande valor de Y no primeiro sorteio, por exemplo, remove essa unidade da lista de seleção, e portanto, reduz a probabilidade de obter um valor alto de Y nos sorteios subsequentes.

TCL para população finita

- Uma versão especial do **TCL** se aplica à **amostragem aleatória simples sem reposição** de uma **população finita**[1][2][3].

[1] Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Pub. Math. Inst. Hungarian Acad. Sci.*, 5, 361-374.

[2] Erdos, P., and Renyi, A. (1959). On the central limit theorem for samples from a finite population. *Pub. Math. Inst. Hungarian Acad. Sci.*, 4, 49-57.

[3] Madow, W. G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Ann. Math. Stat.*, 19, 535-545.

TCL para população finita

- Para amostrar sem reposição de uma população finita, é necessário pensar em uma **sequência de populações**, com o tamanho da população N se tornando grande junto com o tamanho da amostra n.
- Para a população com um determinado tamanho N na sequência, seja $\overline{Y}_N$ a média da população e $\overline{y}_N$ a média amostral de uma amostra aleatória simples selecionada dessa população.

TCL para população finita

- De acordo com o **teorema central do limite para população finita**, a distribuição de

$$\frac{\overline{y}_N - \overline{Y}_N}{\sqrt{\mathrm{Var}\left(\overline{y}_N\right)}}$$

aproxima-se da distribuição normal padrão à medida que n e $N - n$ se tornam grandes.

TCL para população finita

- O resultado também é válido para $\text{Var}\left(\overline{y}_N\right)$ substituída pela variância estimada $\widehat{\text{Var}}\left(\overline{y}_N\right)$ da média amostral de uma amostra aleatória simples de tamanho n de uma população de tamanho N.
- Uma condição técnica do teorema requer que, na progressão das populações hipotéticas de tamanho crescente, a contribuição (proporcional) de qualquer unidade na variância populacional não seja muito grande.

TCL para população finita

Comentário

- Por este resultado, e as demais propriedades dos estimadores, estudados nas aulas anteriores, para o esquema de amostragem aleatória simples, é possível contruir estimativas intervalares do tipo

$$\overline{y} \pm \frac{zs}{\sqrt{n}}\sqrt{1-f}.$$

- É importante salientar que este é um intervalo de confiança aproximado, pois a distribuição do estimador é aproximada.
  - A qualidade da aproximação, como visto no TCL para populações finitas, depende do tamanho da amostra n e da relação $N - n$.

Alguns detalhes

- O **TCL** para população finita requer o conceito de uma sequência de populações $U_1, U_2, \ldots$.
  - A N-ésima população na sequência tem N unidades e valores $Y_{1N}, Y_{2N}, \ldots, Y_{NN}$.
- O tamanho da amostra aleatória simples selecionada da N-ésima população, n_N, também depende de N, e a média (amostral) dessa amostra é $\overline{y}_N = \sum_{i=1}^{n_N} Y_{iN}/n_N$.

Alguns detalhes

- Para qualquer constante $\epsilon > 0$, denote o conjunto de unidades com valores de Y mais distantes da média na N-ésima população por

$$A_N = \left\{ i : |Y_{iN} - \overline{Y}_N| > \epsilon \sqrt{n_N(1 - f_N)S_N^2} \right\},$$

em que $\overline{Y}_N$, f_N e S_N^2 são, respectivamente, a média populacional, a fração de amostragem e a variância populacional na N-ésima população da sequência de populações.

Alguns detalhes

- Se a **condição de Lindeberg-Hájek**

$$\lim_{N\to\infty} \frac{\sum_{A_N} (Y_{iN} - \overline{Y}_N)^2}{(N-1)S_N^2} = 0,$$

é satisfeita, então

$$\frac{\overline{y}_N - \overline{Y}_N}{\sqrt{\text{Var}\left(\overline{y}_N\right)}} \xrightarrow{\mathcal{D}} N(0,1),$$

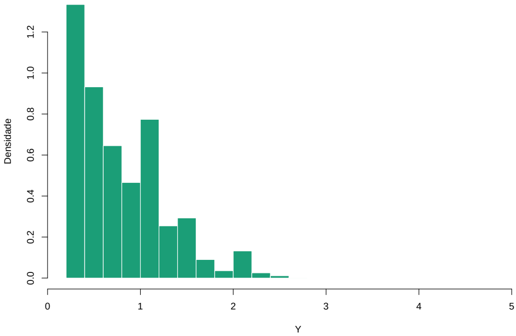
conforme $n_N \to \infty$ e $(N - n_N) \to \infty$ quando $N \to \infty$.

Algumas avaliações empíricas

Exemplo 1

- A população[4] de 53940 diamantes (objeto `diamonds` no pacote `ggplot2` do `R`) pode ser usada para ilustrar o **TCL para população finita**.
- Distribuição do peso em quilates (Y) dos diamantes na própria população.

Exemplo 1

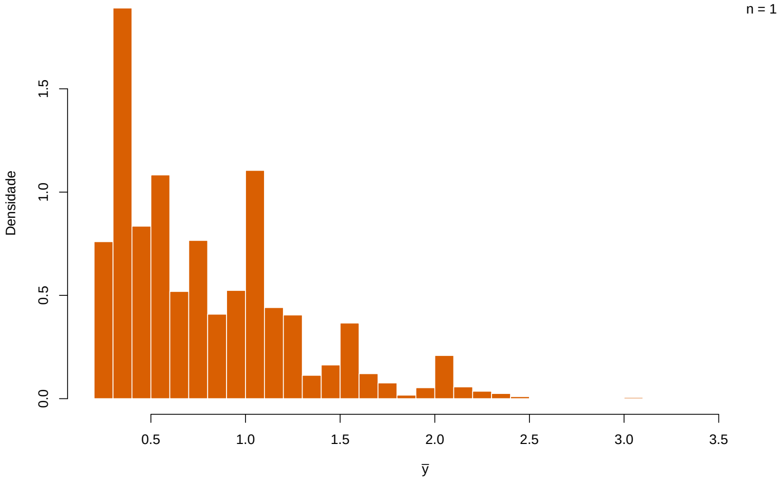


[4] Para fins ilustrativos, vamos supor que este conjunto de dados forma uma população de diamantes.

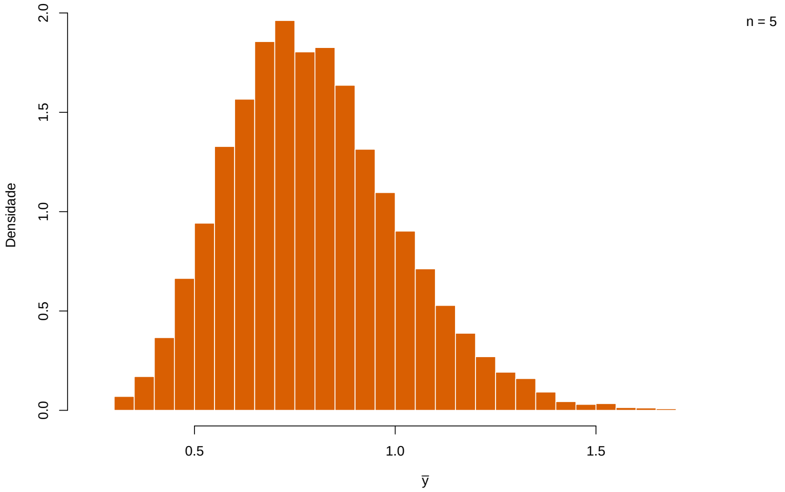
Exemplo 1

- Observe que a variável de interesse na população **não tem distribuição normal**, tendo uma forma assimétrica e acidentada.
- Em seguida, vamos apresentar distribuições da média amostral, obtidas pela repetição da seleção da amostra, utilizando amostragem amostragem aleatória simples sem reposição, nesta população de tamanho $N = 53940$, com tamanhos de amostra $n \in \{1, 5, 20, 50, 100, 1000\}$.

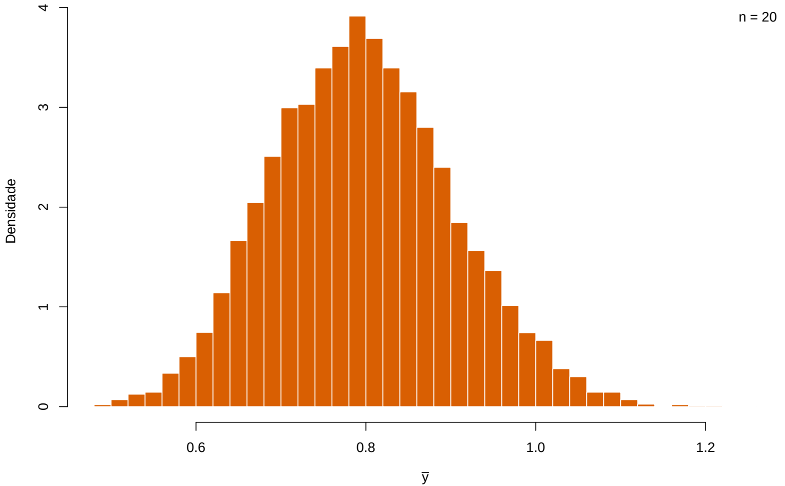
Exemplo 1



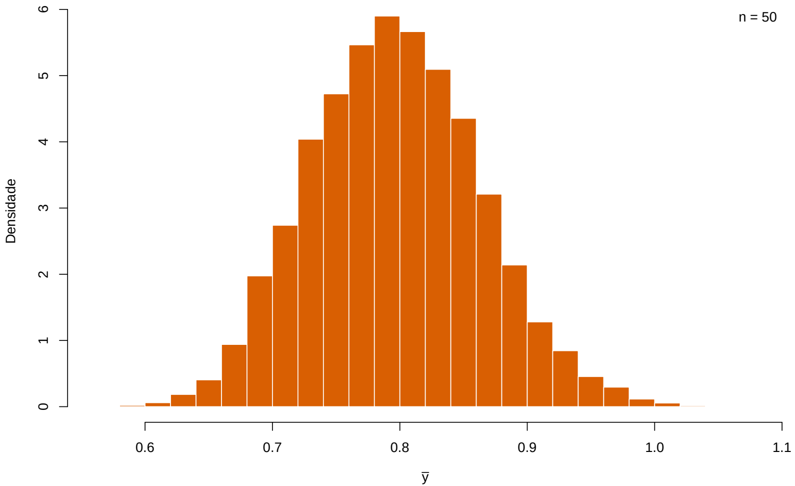
Exemplo 1



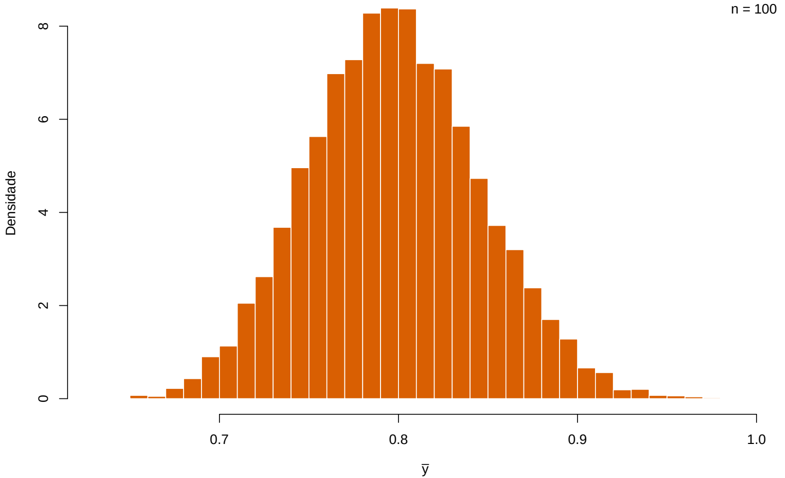
Exemplo 1



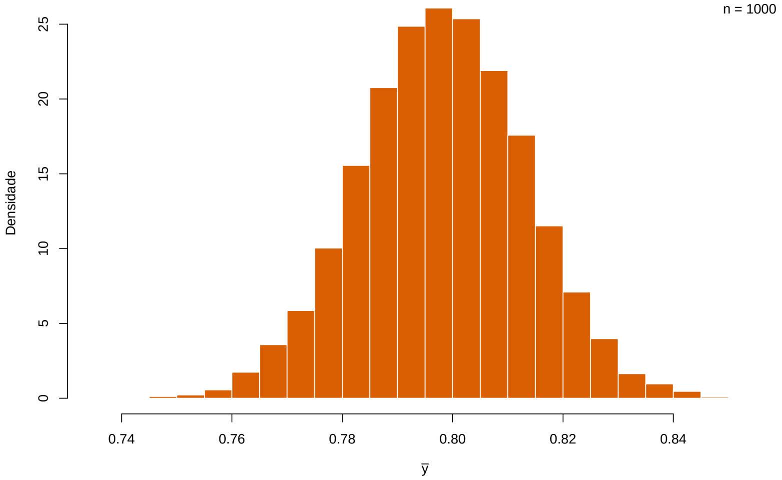
Exemplo 1



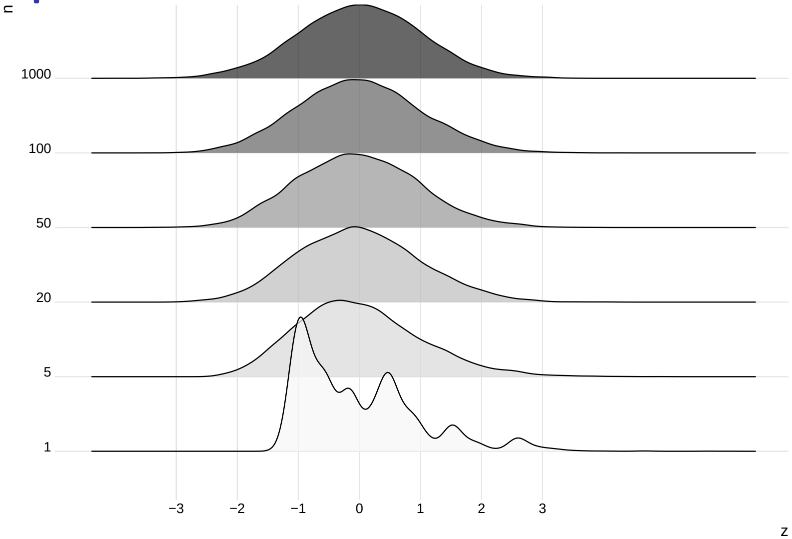
Exemplo 1



Exemplo 1



Exemplo 1



Processo centralizado.

Exemplo 1

- Os histogramas dos slides anteriores mostram que conforme o tamanho da amostra n aumenta, a distribuição de $\overline{y}$ torna-se cada vez mais próxima da normal, desde que $N - n$ também seja suficientemente grande.

Exemplo 2

10.5 Ilustração numérica

Nesta seção, vamos ilustrar o comportamento da aproximação normal para a distribuição da média amostral $\overline{y}$. Conforme visto na Seção 10.1 com relação à AASs, a distribuição de $\sqrt{n}(\bar{y}-\mu)/\sqrt{(1-f)s^2}$ é aproximadamente $N(0,1)$. Portanto, a probabilidade de cobertura do intervalo de confiança para a média populacional μ,

$$\left(\overline{y} - z_\alpha\sqrt{(1-f)\frac{s^2}{n}}, \overline{y} + z_\alpha\sqrt{(1-f)\frac{s^2}{n}}\right), \tag{10.9}$$

deve ser próxima de $\gamma = 1-\alpha$ em grandes amostras. Para $\gamma = 0,95$ ($z_\alpha = 1,96$) devemos ter cobertura próxima de 95%, ou seja, para cada 100 intervalos construídos, aproximadamente 95% devem conter o verdadeiro valor da média populacional μ. Para demonstrar este fato empiricamente, simulamos populações de tamanho

Exemplo 2

$N = 1.000$ a partir das distribuições normal, t-Student (4 graus de liberdade), gama e Gumbel com média 400 e desvio padrão 150. Para cada população, foram retiradas 100.000 amostras, segundo a AASs, de tamanhos n =10, 20, 30, 40, 50, 100 e 200. Para cada amostra retirada foi calculado o intervalo (10.9) e verificado se contém ou não a média populacional μ para cada uma das distribuições. Estas probabilidades de coberturas estimadas (empíricas) estão apresentadas na Tabela 10.1. Pode-se notar claramente que mesmo para n pequenos as probabilidades de cobertura estimadas estão relativamente próximas das correspondentes probabilidades teóricas de cobertura e que a medida que n cresce, elas vão ficando mais próximas ainda.

Exemplo 2

Tabela 10.1: Probabilidades de coberturas estimadas (em porcentagem)

γ	n	normal	t_4	gama	Gumbel
90%	10	86,5	86,6	85,9	85,7
	20	88,4	88,1	88,0	87,8
	30	88,9	89,0	88,7	88,5
	40	89,2	89,2	89,0	88,9
	50	89,2	89,1	89,1	89,0
	100	89,7	89,5	89,7	89,5
	200	89,9	89,6	89,6	89,9
95%	10	91,8	92,2	91,1	90,7
	20	93,5	93,6	93,1	92,7
	30	94,0	94,0	93,8	93,4
	40	94,4	94,3	94,0	93,8
	50	94,4	94,4	94,2	94,1
	100	94,8	94,7	94,6	94,5
	200	95,0	94,9	94,9	94,8
99%	10	97,1	97,3	96,2	96,1
	20	98,1	98,3	97,7	97,5
	30	98,5	98,5	98,2	98,1
	40	98,5	98,7	98,4	98,3
	50	98,7	98,7	98,5	98,4
	100	98,9	98,9	98,7	98,7
	200	98,9	98,9	98,9	98,8
médias populacionais (μ)		403,9	384,9	393,4	398,8
desvios padrões pop. (S)		148,1	145,9	144,8	146,3

Para casa

- Revisar os tópicos discutidos nesta aula.
- Ler o capítulo 10 do Livro "Elementos de amostragem"[5] (disponível no Sabi+).

[5] Bolfarine, H. e Bussab, W. O. **Elementos de amostragem, Blucher, 2005.**

Próximas aulas

- Atividade de Avaliação II.
- Estimativa de proporções e percentagens.

Por hoje é só!

Bons estudos!