Mini curso de Probabilidade e Estatística

Por tutor Mestre Omar Barroso Khodr

Aula 7

- Lei dos Grandes Números
- Teorema do limite central (TLC)
- Proporção
- Viés
- Eficiência
- Intervalos de Confiança

Lei dos Grandes Números

- Essa lei pode nos dizer duas coisas...
- 1. A média de muitas amostras independentes têm alta probabilidade de se aproximar da média populacional.
- 2. O histograma de densidade de muitas amostras independentes têm alta probabilidade de se aproximar do gráfico da densidade da distribuição subjacente. Ou seja, ele têm alta probabilidade de se aproximar de uma curva de distribuição normal.

Lei dos Grandes Números

- Suponha que $X_1, X_2, \ldots, X_n$ são variáveis aleatórias independentes com a mesma distribuição.
- Nesse caso, dizemos que X_i é **independente e identicamente distribuída** ou seja **i.i.d**.
- Suponha a média amostral de $X_1, \ldots, X_n$:

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Note, que nesse caso $\bar{X}_n$ também é uma variável aleatória.
- Ou seja, podemos dizer que conforme n cresce, a probabilidade de $\bar{X}_n$ é perto de μ indo para 1.

Teorema do limite central (TLC)

- Esse teorema afirma que, sob certas condições, a **soma** de um grande número de variáveis aleatórias converge aproximadamente a uma distribuição normal.
- Aqui, apresentamos uma versão da TLC que se aplica a variáveis aleatórias i.i.d.
- Suponha que $\{X_1, \ldots, X_n\}$ é uma sequência i.i.d. de variáveis aleatórias com um distribuição e valor esperado μ e uma variância finita σ^2.
- Vimos que pela leis dos grandes números, a média amostral $\bar{X}_n$ converge para o valor esperado μ conforme $n \to \infty$.
- Mais precisamente, a média amostral tende a μ, escalado pelo valor $\sqrt{n}$ ele aproxima uma distribuição normal com média 0 e variância σ^2.

Distribuições amostrais da média e da proporção

- Uma distribuição amostral mostra todos os resultados possíveis que uma estatística pode obter em todas as amostras possíveis de uma população e com que frequência cada resultado acontece - e pode nos ajudar a usar amostras para fazer previsões sobre a chance de algo ocorrer.
- Muitas vezes, a amostragem é feita para estimar a **proporção** de uma população que tem uma característica específica, como a proporção de todos os itens defeituosos que saem de uma linha de montagem ou a proporção de todas as pessoas que entram em uma loja de varejo e fazem uma compra antes de sair.
- A população é reconhecida como p enquanto a amostra é $\hat{p}$.

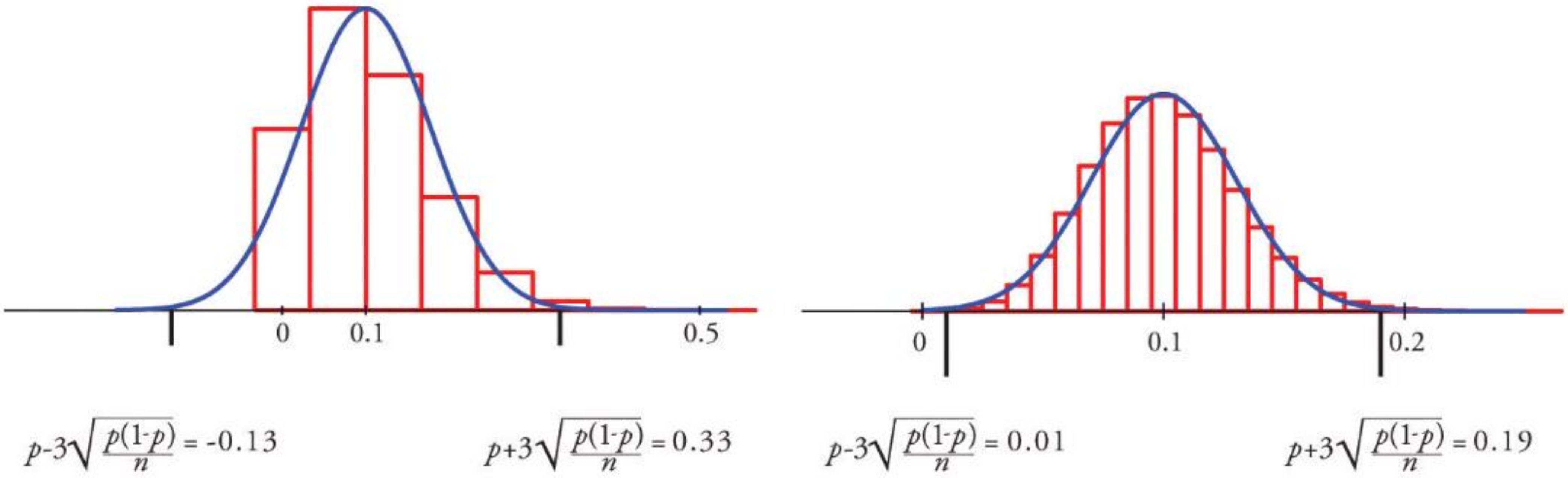
Distribuições amostrais da média e da proporção

- Exemplo:
- Supomos que 43% das pessoas que entram em uma loja fazem uma compra antes de sair. Essas 43% seriam o nosso p.
- Todavia, imagina que estamos trabalhando com um amostra coletada de 200 pessoas, Segundo nossas observações 78 dos clientes compram algum produto.
- Dessa maneira, nossa proporção é:
- $\hat{p} = \frac{78}{200} = 0.39 \equiv 39\%$
- Deste modo, a proporção da amostra é uma variável aleatória: ela varia de amostra para amostra de uma forma que não pode ser prevista com certeza.

Distribuições amostrais da média e da proporção

- Para amostras grandes, a proporção da amostra é aproximadamente distribuída normalmente, com média $\mu_{\hat{p}} = p$ e um desvio padrão, $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$.
- Se uma amostra é grande o suficiente, ela fica no intervalo [0,1].
- Ou seja, conforme uma proporção é uma variável aleatória, conforme ficamos tirando outras variáveis elas tendem a formar uma distribuição normal.

Distribuições amostrais da média e da proporção



A imagem mostra que quando o espaço amostral é pequeno nossa distribuição não segue um formato normalizado. Todavia, quando aumentamos a amostra nossa distribuição segue um padrão normal.Fonte: LibreTexts (2021)

Estimadores e Propriedades

- Um **estimador**, é uma função matemática que usa dados **amostrais** para estimar um parâmetro populacional desconhecido (e.g., média e proporção).
- Por exemplo, a média amostral ($\bar{X}$) estima a média populacional (μ).
- A proporção amostral ($\hat{p}$) estima a proporção populacional (p).
- A **estimativa** de um parâmetro é um resultado da aplicação de um procedimento estatístico aos dados para obter alguns coeficientes de um modelo. O valor calculado usando a média aritmética seria uma estimativa da média populacional;

Viés (Bias)

- Diferença entre o valor esperado do estimador e o valor verdadeiro do parâmetro. $E[\hat{\theta}] - E[\theta]$
- Não-viesado (justo): $E[\hat{\theta}] = \theta$
- Viesado: $E[\hat{\theta}] \neq \theta$

Eficiência

- Um estimador é mais eficiente se tiver menor variância (mais preciso).
- Entre dois estimadores não-viesados, o mais eficiente é o que tem menor variância.
- Por exemplo, vamos supor que temos duas bases de dados X_1 e X_2.
- Após calcular a média amostral de ambas e a variância, podemos ver que ${\hat{\sigma}^2}_{x_2} < {\hat{\sigma}^2}_{x_1}$.
- Desta maneira, o estimador X_1 é mais eficiente.

Consistência

- À medida que o tamanho da amostra (n) aumenta, o estimador converge para o valor verdadeiro do parâmetro.
- **Requer:**
- Não-viesado (ou viés tendendo a zero).
- Variância tendendo a zero quando $n \to \infty$.

Intervalos de Confiança (IC)

- O IC, é uma faixa de valores que contém o parâmetro populacional (e.g., média, proporção) com uma probabilidade conhecida (nível de confiança, e.g., 95%).
- Não é uma afirmação sobre a amostra, mas sobre onde o parâmetro verdadeiro provavelmente pode ser encontrado.
- Para compor um IC, precisamos de uma estimativa pontual (e.g., a média amostral); O erro padrão (desvio padrão da amostra).
- Nível de confiança, [e.g.] 95% → significa que, se repetirmos o estudo 100 vezes, o IC incluirá o parâmetro verdadeiro em ~95 casos.

IC fórmulas

- $\boldsymbol{\sigma\ conhecido\ (distribuição\ normal)}: IC = \mu \pm \frac{Z\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}}$
- $\boldsymbol{\sigma\ desconhecido\ (t - student)}: IC = \bar{X} \pm \frac{t\alpha}{2} \cdot \frac{S}{\sqrt{n}}$
- $\mu: Média\ pop.$
- $\bar{X}: Média\ amostral$
- $\sigma: Desvio\ Padrão$
- $s: Erro\ Padrão\ (ou\ desvio\ padrão\ da\ amostral)$

Exemplo (IC)

- Queremos saber a média de altura (μ) de atletas em um torneio de futebol (em um nível de confiança de 95%).
- Vamos supor que em **nossa amostra**, temos n = 100; $\bar{X}$ = 170 cm; e, *s* = 10 cm.
- Como a população é desconhecida utilizaremos a distribuição t:
- Vamos montar nossa operação...
- Dado que nosso n é 100, teremos que encontrar o valor t (ou Z) na tabela...
- Nesse caso, temos que descontar 1 de 100 pelos graus de liberdade (GL). Ou seja, n-1 ou 100-1= 99 (GL).

Exemplo (IC)

- $\boldsymbol{\sigma}$ ***desconhecido*** $(\boldsymbol{t} - \boldsymbol{student})$: $IC = \bar{X} \pm \frac{t\alpha}{2} \cdot \frac{s}{\sqrt{n}}$
- $\frac{t\alpha}{2}$ ~ 1,984 (de acordo com a tabela)
- Ou seja,
- $170 \pm 1{,}984 \cdot \frac{10}{\sqrt{100}} = 170 \pm 1{,}984$
- Assim, $170 + 1{,}984 = \mathbf{171{,}984}$; $170 - 1{,}984 = \mathbf{168{,}016} \rightarrow [171{,}984;\ 168{,}016]$
- Interpretação: Temos 95% de confiança de que a altura média populacional está entre 168.0 cm e 172.0 cm.

Exemplo (IC) 2

- Temos uma fazenda que se especializa em colheitas de frutas. Existem **centenas** de maçãs nas árvores, queremos escolher **aleatoriamente** 90 maçãs para colocar em embalagens para exportação. Supomos que temos essas propriedades.
- Média = 86
- S= 6,2
- Queremos saber Queremos saber a média de maças (μ) cabem em uma caixa (em um nível de 95% de confiança).

Exemplo (IC) 2

- $\boldsymbol{\sigma\ conhecido}$: $IC = \mu \pm \frac{Z\alpha}{2} \,.\, \frac{\sigma}{\sqrt{n}}$
- N = 90 (como n > 30; utilizamos a tabela Z distribuição normal (Z), mesmo que σ (desvio padrão populacional) seja desconhecido (pelo Teorema Central do Limite).
- $\bar{X} = 86$
- S = 6,2
- N = 90
- Assim, $\frac{Z\alpha}{2}$ ~ 1,960 (de acordo com a tabela)
- $86 \pm 1{,}960 \,.\, \frac{6{,}2}{\sqrt{90}} = 86 \pm 1{,}960 \,.\, 0{,}654 \ \rightarrow 86 \pm 1{,}282$
- Assim, $[84{,}72; 87{,}28]$
- Temos 95% de confiança de que a verdadeira média de maçãs por caixa (μ) na população está entre ~84,72 e ~87,28.

Student t Distribution Probabilities Table

one-tail area	0.25	0.125	0.1	0.075	0.05	0.025	0.01	0.005	0.0005
two-tail area	0.5	0.25	0.2	0.15	0.1	0.05	0.02	0.01	0.001
confidence level	0.5	0.75	0.8	0.85	0.9	0.95	0.98	0.99	0.999
d.f. 1	1.000	2.414	3.078	4.165	6.314	12.706	31.821	63.657	636.619
2	0.816	1.604	1.886	2.282	2.920	4.303	6.965	9.925	31.599
3	0.765	1.423	1.638	1.924	2.353	3.182	4.541	5.841	12.924
4	0.741	1.344	1.533	1.778	2.132	2.776	3.747	4.604	8.610
5	0.727	1.301	1.476	1.699	2.015	2.571	3.365	4.032	6.869
6	0.718	1.273	1.440	1.650	1.943	2.447	3.143	3.707	5.959
7	0.711	1.254	1.415	1.617	1.895	2.365	2.998	3.499	5.408
8	0.706	1.240	1.397	1.592	1.860	2.306	2.896	3.355	5.041
9	0.703	1.230	1.383	1.574	1.833	2.262	2.821	3.250	4.781
10	0.700	1.221	1.372	1.559	1.812	2.228	2.764	3.169	4.587
11	0.697	1.214	1.363	1.548	1.796	2.201	2.718	3.106	4.437
12	0.695	1.209	1.356	1.538	1.782	2.179	2.681	3.055	4.318
13	0.694	1.204	1.350	1.530	1.771	2.160	2.650	3.012	4.221
14	0.692	1.200	1.345	1.523	1.761	2.145	2.624	2.977	4.140
15	0.691	1.197	1.341	1.517	1.753	2.131	2.602	2.947	4.073
16	0.690	1.194	1.337	1.512	1.746	2.120	2.583	2.921	4.015
17	0.689	1.191	1.333	1.508	1.740	2.110	2.567	2.898	3.965
18	0.688	1.189	1.330	1.504	1.734	2.101	2.552	2.878	3.922
19	0.688	1.187	1.328	1.500	1.729	2.093	2.539	2.861	3.883
20	0.687	1.185	1.325	1.497	1.725	2.086	2.528	2.845	3.850
21	0.686	1.183	1.323	1.494	1.721	2.080	2.518	2.831	3.819
22	0.686	1.182	1.321	1.492	1.717	2.074	2.508	2.819	3.792
23	0.685	1.180	1.319	1.489	1.714	2.069	2.500	2.807	3.768
24	0.685	1.179	1.318	1.487	1.711	2.064	2.492	2.797	3.745
25	0.684	1.198	1.316	1.485	1.708	2.060	2.485	2.787	3.725
26	0.684	1.177	1.315	1.483	1.706	2.056	2.479	2.779	3.707
27	0.684	1.176	1.314	1.482	1.703	2.052	2.473	2.771	3.690
28	0.683	1.175	1.313	1.480	1.701	2.048	2.467	2.763	3.674
29	0.683	1.174	1.311	1.479	1.699	2.045	2.462	2.756	3.659
30	0.683	1.173	1.310	1.477	1.697	2.042	2.457	2.750	3.646
35	0.682	1.170	1.306	1.472	1.690	2.030	2.438	2.724	3.591
40	0.681	1.167	1.303	1.468	1.684	2.021	2.423	2.704	3.551
45	0.680	1.165	1.301	1.465	1.679	2.014	2.412	2.690	3.520
50	0.679	1.164	1.299	1.462	1.676	2.009	2.403	2.678	3.496
60	0.679	1.162	1.296	1.458	1.671	2.000	2.390	2.660	3.460
70	0.678	1.160	1.294	1.456	1.667	1.994	2.381	2.648	3.435
80	0.678	1.159	1.292	1.453	1.664	1.990	2.374	2.639	3.416
100	0.677	1.157	1.290	1.451	1.660	1.984	2.364	2.626	3.390
500	0.675	1.152	1.283	1.442	1.648	1.965	2.334	2.586	3.310
1000	0.675	1.151	1.282	1.441	1.646	1.962	2.330	2.581	3.300
infinity	0.674	1.150	1.282	1.440	1.645	1.960	2.326	2.576	3.291

Tabela t-student. Fonte: Crafton Hills College

Confidence Interval Critical Values, $z_{\alpha/2}$

Level of Confidence	Critical Value, $z_{\alpha/2}$
0.90 or 90%	1.645
0.95 or 95%	1.96
0.98 or 98%	2.33
0.99 or 99%	2.575

Tabela Distribuição Normal (Z). Fonte: Crafton Hills College

Bibliografia

- **BUSSAB, W. O.; MORETTIN, P. A.** Estatística Básica. Saraiva, 2017.
- **LARSON, R.; FARBER, B.** Estatística Aplicada. Pearson, 2016.
- **Pishro-Nik; H.** Introduction to Probability, Statistics, and Random Processes**.** Kappa Research, 2014.
- **TRIOLA, M. F.** Introdução à Estatística. Pearson, 2018.