

ESTATÍSTICA

Michelle Hanne Soares de Andrade

michellehanne.andrade@gmail.com



Teorema do Limite Central

Teorema do Limite Central - Resumo

Distribuição de X	Distribuição de \bar{X}	Tamanho de amostra
$N(\mu; \sigma)$	$N(\mu; \sigma/\sqrt{n})$	n qualquer
Qualquer	aproximadamente $N(\mu; \sigma/\sqrt{n})$	$n > 30$ (em geral)

Exemplo Binomial

Corolário 5.3. *Seja X uma v.a. com distribuição Binomial de parâmetros n e p . Se $n \geq 30$ e p tal que $np > 5$ e $n(1-p) > 5$, então:*

$$X \stackrel{a}{\sim} N(np, np(1-p)).$$

Exemplo 5.4. *Considere-se a v.a. $X \sim \text{Bin}(100, 0.1)$. Calculemos $P(X = 10)$ Como $n = 100 \geq 30$, $np = 100 \times 0.1 = 10 > 5$ e $n(1-p) = 100 \times 0.9 = 90$,*

$$\begin{aligned} P(X = 10) &= P(X \leq 10) - P(X \leq 9) \approx \Phi\left(\frac{10-10}{3}\right) - \Phi\left(\frac{9-10}{3}\right) = \Phi(0) - \Phi(-0.33) = \\ &= 0.5 - 0.3707 = 0.1293. \end{aligned}$$

Nota: *O valor exacto é $P(X = 10) = \binom{100}{10} 0.1^{10} 0.9^{90} = 0.1319$.*

$$E(X_n) = np, \quad \sigma = \text{DP}(X) = \sqrt{np(1-p)}.$$

Exemplo - Poisson

Corolário 5.5. *Seja X uma v.a. com distribuição Poisson de parâmetro λ . Se $\lambda > 5$, então:*

$$X \stackrel{a}{\sim} N(\lambda, \lambda).$$

Exemplo 5.6. *Considere $X \sim P(230)$. Calculemos um valor aproximado de $P(X = 241)$.*

$$\begin{aligned} P(X = 241) &= P(X \leq 241) - P(X \leq 240) \approx P\left(Z \leq \frac{241-230}{\sqrt{230}}\right) - P\left(Z \leq \frac{240-230}{\sqrt{230}}\right) = \\ &= \Phi(0.73) - \Phi(0.66) = 0.7673 - 0.7454 = 0.0219 \end{aligned}$$

Nota: *O valor exacto é $P(X = 241) = e^{-230} \cdot \frac{230^{241}}{241!} = 0.0198$.*

- Valor esperado
 $\mu = E(X) = \lambda$
- Variância
 $\sigma^2 = \text{Var}(X) = \lambda,$

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!},$$

Distribuição Amostral de uma Proporção

\hat{p} é uma média de variáveis aleatórias $X_i \sim \text{Binomial}(m=1; p)$.

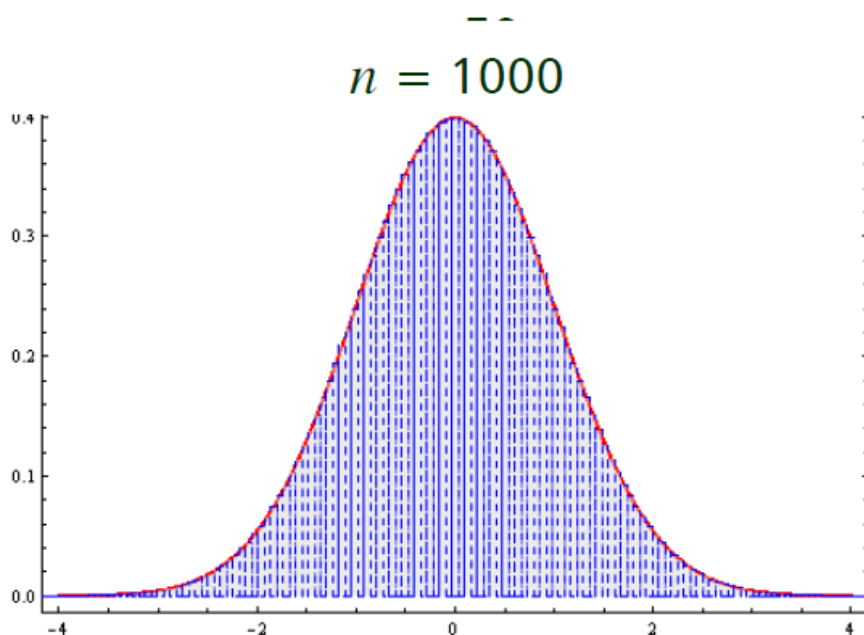
$$\text{Assim, } \mu = E[X_i] = p \text{ e } \sigma = \sqrt{\text{Var}[X_i]} = \sqrt{p(1-p)}$$

Pelo Teorema Central do Limite, a distribuição amostral de \hat{p} pode ser aproximada por uma Normal com média p e desvio-padrão $\sqrt{p(1-p)}$ quando $np \geq 5$ e $n(1-p) \geq 5$.

$$\text{Ou seja, } Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \stackrel{\text{aproximadamente}}{\sim} N(0,1)$$

Distribuição Amostral de uma Proporção

Ilustração: o que acontece com a distribuição dos valores de $Z = \frac{\hat{p} - 0.50}{\sqrt{0.50(1 - 0.50)/n}}$ quando n cresce?



Exemplo - Distribuição Amostral de uma Proporção

Um biólogo está estudando a preferência de uma espécie de aranha (espécie A) quanto ao local de confecção de sua teia em árvores: perto do tronco ou ao final dos galhos.

Em 40 teias de aranha da espécie A, ele observou que 22 delas foram tecidas perto do tronco, ou seja, $\hat{p} = 22 / 40 = 0.55$.

Para aranhas de uma espécie B, estudos mostram que a proporção das que preferem fazer teias perto do tronco é igual a 0.75.

Supondo que a proporção populacional das aranhas da espécie A que fazem teias perto do tronco também seja $p=0.75$, qual é a probabilidade de o resultado amostral ter ocorrido?

Exemplo - Distribuição Amostral de uma Proporção

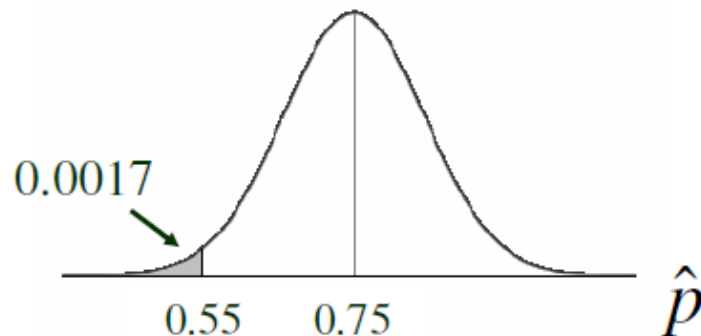
Como a proporção amostral é uma variável aleatória contínua, já sabemos que o cálculo de $P[\hat{p} = 0.55]$ não faz sentido.

Assim, vamos calcular a probabilidade de amostras com proporções ainda mais extremas do que a obtida. Ou seja,

$$P[\hat{p} < 0.55] = P\left[\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < \frac{0.55 - 0.75}{\sqrt{0.75(1-0.75)/40}}\right]$$

$$= P[Z < -2.92]$$

$$= 0.0017$$



Exemplo - Distribuição Amostral de uma Proporção

Sob a hipótese de que a proporção populacional das aranhas da espécie A que fazem teias perto do tronco também seja $p=0.75$, a amostra coletada é pouco verossímil (probabilidade de 0.0017).

Sendo assim, a hipótese de que as duas espécies têm a mesma proporção de aranhas que tecem suas teias perto do tronco deve ser revista.



Distribuição Qui-Quadrado

Distribuição Qui-Quadrado

- A distribuição qui-quadrado é obtida diretamente das variáveis aleatórias independentes normais.
- padrões. Sejam $Z_i, i = 1, 2, \dots, n$ variáveis aleatórias independentes, cada uma distribuída como normal padrão. Defina uma nova variável aleatória como a soma dos quadrados de Z_i :

$$X = \sum_{i=1}^n Z_i^2.$$

Distribuição Qui-Quadrado

- A fdp de uma distribuição qui-quadrado com diversos graus de liberdade é mostrada na próxima Figura.
- Uma variável aleatória qui-quadrada é sempre não-negativa, e que, diferentemente da distribuição normal, a distribuição qui-quadrado não é simétrica em torno de qualquer ponto.

Distribuição Qui-Quadrado

- Então, X terá o que é conhecido como **distribuição qui-quadrado** com n **graus de liberdade** (ou gl abreviadamente). Escrevemos isso como $X \sim X_n^2$. Os gl em uma distribuição qui-quadrado correspondem ao número de termos na fórmula anterior.

Distribuição Qui-Quadrado

Representação gráfica

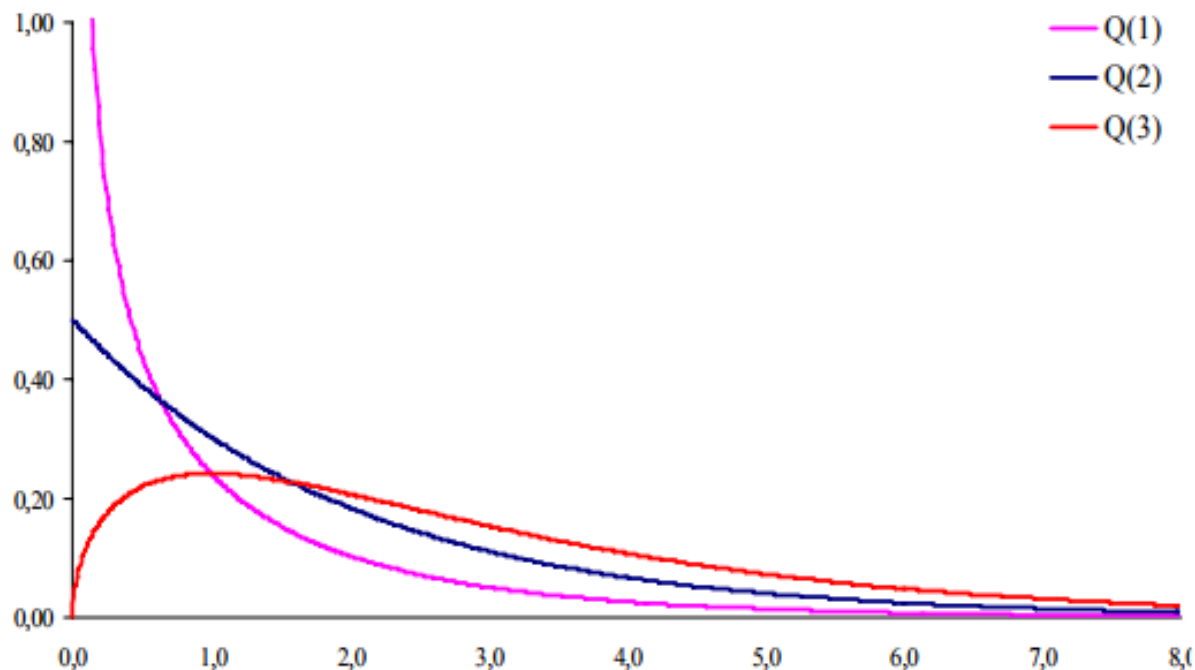


Figura 1.6 - Gráfico da distribuição χ^2 (Qui-Quadrado) para os gl de 1, 2 e 3

Distribuição Qui-Quadrado

Definição 9.1: Uma variável aleatória contínua X tem distribuição qui-quadrado com n graus de liberdade, denotada por χ_n^2 , se sua função densidade for dada por:

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x > 0, \quad n > 0$$

Sendo, $\Gamma(w) = \int_0^\infty x^{w-1} e^{-x} dx, \quad w > 0.$

IDEIA Graus de liberdade: Considere um conjunto de dados qualquer. Graus de liberdade é o número de valores deste conjunto de dados que podem variar após terem sido impostas certas restrições a todos os valores.

Distribuição Qui-Quadrado

Devido a sua importância a distribuição qui-quadrado está tabulada para diferentes valores do parâmetro n .

Assim, poderemos achar na tabela o valor χ^2_α que satisfaça $P(X \leq \chi^2_\alpha) = \alpha$ ou $P(X \geq \chi^2_\alpha) = \alpha$, dependendo da tabela.

O que é tabelado é a função inversa, em relação a área à direita ou à esquerda de cada curva. Isto é, dado um valor de área na cauda direita, a tabela retorna um valor χ^2_α tal que $P(X \geq \chi^2_\alpha) = \alpha$ e dado um valor de área na cauda esquerda a tabela retorna um valor χ^2_α tal que $P(X \leq \chi^2_\alpha) = \alpha$.

Tabela Qui-Quadrado

$$P(X^2_{(2)} \geq 0,2107) = 0,90$$

Distribuição Qui-Quadrado



L	0,990	0,950	0,900	0,500	0,100	0,050	0,025	0,010	0,005
1	0,0002	0,0039	0,0158	0,4549	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0201	0,1026	0,2107	1,3863	4,6052	5,9915	7,3778	9,2104	10,5965
3	0,1148	0,3518	0,5844	2,3660	6,2514	7,8147	9,3484	11,3449	12,8381
4	0,2971	0,7107	1,0636	3,3567	7,7794	9,4877	11,1433	13,2767	14,8602
5	0,5543	1,1455	1,6103	4,3515	9,2363	11,0705	12,8325	15,0863	16,7496
6	0,8721	1,6354	2,2041	5,3481	10,6446	12,5916	14,4494	16,8119	18,5475
7	1,2390	2,1673	2,8331	6,3458	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,6465	2,7326	3,4895	7,3441	13,3616	15,5073	17,5345	20,0902	21,9549
9	2,0879	3,3251	4,1682	8,3428	14,6837	16,9190	19,0228	21,6660	23,5893
10	2,5582	3,9403	4,8652	9,3418	15,9872	18,3070	20,4832	23,2093	25,1881
11	3,0535	4,5748	5,5778	10,3410	17,2750	19,6752	21,9200	24,7250	26,7569
12	3,5706	5,2260	6,3038	11,3403	18,5493	21,0261	23,3367	26,2170	28,2997
13	4,1069	5,8919	7,0415	12,3398	19,8119	22,3620	24,7356	27,6882	29,8193
14	4,6604	6,5706	7,7895	13,3393	21,0641	23,6848	26,1189	29,1412	31,3194
15	5,2294	7,2609	8,5468	14,3389	22,3071	24,9958	27,4884	30,5780	32,8015
16	5,8122	7,9616	9,3122	15,3385	23,5418	26,2962	28,8453	31,9999	34,2671
17	6,4077	8,6718	10,0852	16,3382	24,7690	27,5871	30,1910	33,4087	35,7184
18	7,0149	9,3904	10,8649	17,3379	25,9894	28,8693	31,5264	34,8052	37,1564
19	7,6327	10,1170	11,6509	18,3376	27,2036	30,1435	32,8523	36,1908	38,5821
20	8,2604	10,8508	12,4426	19,3374	28,4120	31,4104	34,1696	37,5663	39,9969
21	8,8972	11,5913	13,2396	20,3372	29,6151	32,6706	35,4789	38,9322	41,4009
22	9,5425	12,3380	14,0415	21,3370	30,8133	33,9245	36,7807	40,2894	42,7957
23	10,1957	13,0905	14,8480	22,3369	32,0069	35,1725	38,0756	41,6383	44,1814
24	10,8563	13,8484	15,6587	23,3367	33,1962	36,4150	39,3641	42,9798	45,5584
25	11,5240	14,6114	16,4734	24,3366	34,3816	37,6525	40,6465	44,3140	46,9280
26	12,1982	15,3792	17,2919	25,3365	35,5632	38,8851	41,9231	45,6416	48,2898
27	12,8785	16,1514	18,1139	26,3363	36,7412	40,1133	43,1945	46,9628	49,6450
28	13,5647	16,9279	18,9392	27,3362	37,9159	41,3372	44,4608	48,2782	50,9936
29	14,2564	17,7084	19,7677	28,3361	39,0875	42,5569	45,7223	49,5878	52,3355

Propriedades do Qui-Quadrado

Propriedades

$$E(X) = n$$

$$Var(X) = 2n$$

Aplicação – Teste de Hipótese

Se um dado não viciado for jogado 6 vezes, espera-se obter 1 vez cada face (1, 2, 3, 4, 5 e 6) já que a probabilidade de cair qualquer face é $1/6$.

Supondo que um dado foi jogado 186 vezes e se obteve:

Face 1	Face 2	Face 3	Face 4	Face 5	Face 6
34	29	30	32	28	33

- Qual será o valor de χ^2 ?
- Como se pode interpretar esse valor?

Aplicação – Teste de Hipótese

Como calcular

Karl Pearson propôs a seguinte fórmula para medir as possíveis discrepâncias entre proporções observadas e esperadas:

$$\chi^2 = \sum [(o - e)^2 / e]$$

em que

- o = frequência observada para cada classe,
- e = frequência esperada para aquela classe.

Note-se que $(o - e)$ = desvio (d), portanto a fórmula também pode ser escrita como

$$\chi^2 = \sum (d^2 / e)$$

Aplicação – Teste de Hipótese

Resolvendo:

As frequências esperadas em cada classe são calculadas por: $p.N$. Portanto:

$$E_{(\text{face 1})} = E_{(\text{face 2})} = E_{(\text{face 3})} = E_{(\text{face 4})} = E_{(\text{face 5})} = E_{(\text{face 6})} = p.N = 1/6 . 186 = 31$$

a. Qual será o valor de χ^2 ?

Assim, os valores parciais são somados: e chega-se ao valor de χ^2 :

observado	34	29	30	32	28	33
esperado	31	31	31	31	31	31
χ^2 parcial	0,2903	0,1290	0,0322	0,0322	0,2903	0,1290

$$\chi^2 = (0,2903 + 0,1290 + 0,0322 + 0,0322 + 0,2903 + 0,1290) = 0,903$$

b. Como se pode interpretar esse valor?

Lembrando que G.L. = número de classes -1, como há há 6 classes, G.L. = 5.

Verificando-se a tabela de χ^2 na linha em G.L. = 5 encontra-se χ^2_c igual a 11,070. Como o valor de Qui Quadrado obtido (0,903) foi *menor* que o esperado ao acaso (11,070) admite-se que o dado seja honesto.



Distribuição T - Student

T- Student

- Padronizar variável aleatória normal requer que o μ e σ sejam conhecidos. Na prática, porém, não podemos calcular $z = (x - \mu) / \sigma$ porque σ é desconhecido. Em vez disso, substituímos σ por s e calculamos a estatística t .

$$t = \frac{x - \mu}{s}$$

Distribuição Amostral da Média e da variância

- Se discrepâncias nas observações sobre a média são aleatórios e independentes, então a distribuição amostral da média tem μ e variância, σ^2/n .
- A quantidade σ^2/n é a variância da média.
- Sua raiz quadrada é chamada o ***erro padrão da média***:

$$\sigma = \frac{\sigma}{\sqrt{n}}$$

- A estimativa do **erro padrão da média** é:

$$s = \frac{s}{\sqrt{n}}$$

Distribuição t

- Normalmente, a variância da população, σ^2 não é conhecida e não podemos usar a distribuição normal como a distribuição de referência para a média da amostra. Em vez disso, substituir e usar a distribuição t .
- Se a distribuição de referência é normal e a variância da população é estimado por s^2 , a quantidade:

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

- que é conhecido como a média padronizada ou como a estatística t , terá a distribuição com $\nu = n - 1$ graus de liberdade.

Exemplo:

Exemplo para a tabela: supõe-se uma distribuição t-student com 5 graus de liberdade e área A (ou probabilidade de valores acima) de 0,1 ou 10%:

Na interseção da coluna $v = 5$ e $A = 0,1000$, o valor é $t = 1,476$. Isso significa que

$$P(t > 1,476) = 0,1000 \text{ OU } P(t \leq 1,476) = 1 - 0,1000 = 0,9.$$

Considerando a simetria da distribuição,

$$P(t < -1,476) = 0,1000 \text{ e também } P(-1,476 \leq t \leq 1,476) = 1 - 2 \times 0,1 = 0,8.$$

v / A	0,2500	0,2000	0,1500	0,1000	0,0500	0,0250	0,0100	0,0050	0,0025	0,0010	0,0005
001	1,000	1,376	1,963	3,078	6,314	12,710	31,820	63,660	127,300	318,300	636,600
002	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	14,090	22,330	31,600
003	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	7,453	10,210	12,920
004	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
005	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
006	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
007	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
008	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
009	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
010	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
011	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
012	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
013	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
014	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140

Distribuição T

- Suponhamos que Z tenha uma distribuição normal padrão e que X tenha uma distribuição qui-quadrado com n graus de liberdade. Adicionalmente, suponhamos que Z e X sejam independentes. Então, a variável aleatória terá uma **distribuição t** com n graus de liberdade.

$$T = \frac{Z}{\sqrt{X/n}}$$

Distribuição T

$T \sim tn$. A distribuição t obtém seus graus de liberdade da variável aleatória qui-quadrada no denominador da equação anterior.

A fdp da distribuição t tem uma forma semelhante à da distribuição normal padrão, exceto pelo fato de que ela é mais espalhada e, portanto, tem mais área nos extremos.

O valor esperado de uma variável aleatória com distribuição t **é zero** (no sentido exato, o valor esperado somente existirá para $n > 1$).

A **variância** será $n/(n - 2)$ para $n > 2$. Não existe variância de $n \leq 2$ devido à distribuição ser tão espalhada.

Distribuição T

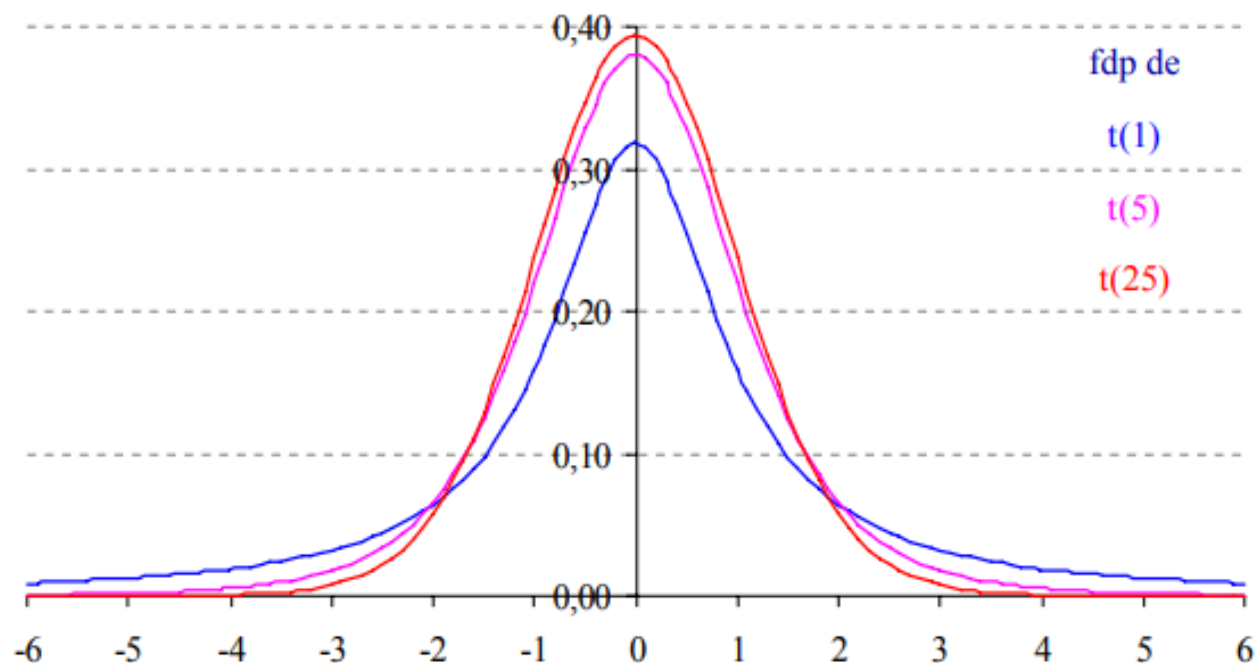


Figura 1.5 - Gráfico da distribuição t (de Student) para os gl de 1, 5 e 25

Distribuição T

A distribuição t de Student é uma das distribuições mais utilizadas na estatística, com aplicações que vão desde a modelagem estatística até testes de hipóteses.

Definição 9.2: Uma variável aleatória contínua X tem distribuição t de Student com ν graus de liberdade, denotada por t_ν , se sua função densidade for dada por:

$$f(x) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}, \quad \nu = 1, 2, 3, \dots \quad \forall x \in \mathbb{R}$$

A expressão acima é assustadora???

Boa Notícia: Não precisaremos dela para calcular probabilidades.

Mais uma vez, o parâmetro ν , chamado de graus de liberdade, está associado ao número de parcelas independentes em uma soma.

Distribuição T

Propriedades

$$E(X) = 0 \quad \text{para } \nu > 1$$

$$\text{Var}(X) = \frac{\nu}{\nu - 2}, \quad \text{para } \nu > 2$$

Distribuição T

Ao contrário da distribuição normal, não existe uma relação entre as diferentes distribuições t, assim seria necessária uma tabela para cada valor de v .

É comum que os livros didáticos apresentem tabelas da distribuição t que envolvem os valores críticos.

O motivo para isso é que a maioria das aplicações da distribuição t envolve a construção de intervalos de confiança ou de testes de hipóteses.

Nessas aplicações, nosso interesse está no valor crítico associado a um nível de significância α que, como visto no gráfico a seguir, é o valor da abscissa que deixa probabilidade (área) α acima dela.

Distribuição T

- Na tabela t, cada linha corresponde a um número diferente de graus de liberdade e cada coluna corresponde a uma área α na cauda superior. No corpo da tabela temos a abscissa t_{α} que deixa a área α acima dela

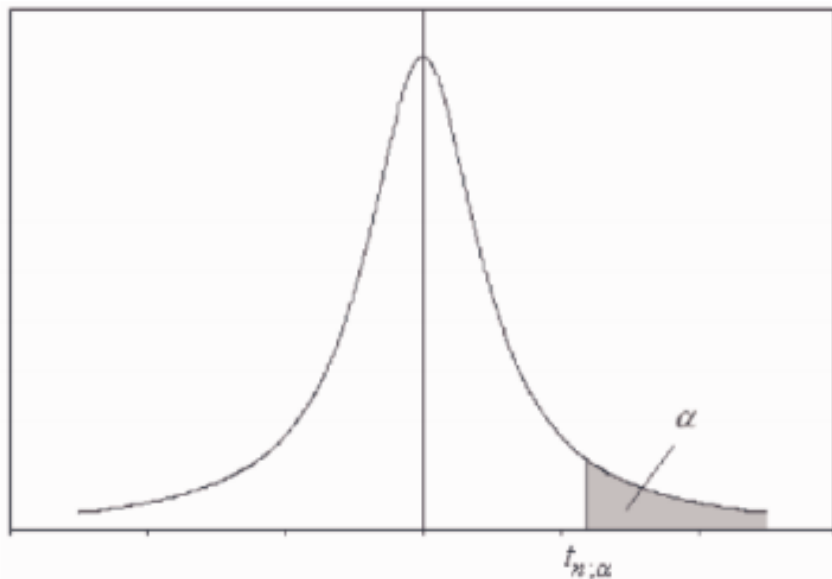


Figura : Ilustração do valor crítico $t_{n;\alpha}$ da distribuição $t(v)$.

Distribuição T

$P(t \text{ de Student} \geq \text{valor tabelado}) = \alpha \Leftrightarrow$ Valores unilaterais

	0.5000	0.2000	0.1000	0.0500	0.0400	0.0200	0.0100	0.0050	0.0010
1	1.000	3.078	6.314	12.706	15.894	31.821	63.656	127.321	636.578
2	0.816	1.886	2.920	4.303	4.849	6.965	9.925	14.089	31.600
3	0.765	1.638	2.353	3.182	3.482	5.841	8.451	11.917	24.478
4	0.741	1.533	2.132	2.776	2.998	5.294	7.709	10.847	21.448
5	0.727	1.476	2.015	2.571	2.777	5.051	7.400	10.595	20.154
6	0.718	1.440	1.943	2.447	2.612	4.876	7.171	10.350	19.648
7	0.711	1.415	1.895	2.365	2.517	4.759	6.998	10.229	19.408
8	0.706	1.397	1.860	2.306	2.449	4.681	6.896	10.133	19.241
9	0.703	1.383	1.833	2.262	2.398	4.621	6.821	10.060	19.171
10	0.700	1.372	1.812	2.228	2.359	4.574	6.764	10.000	19.107
11	0.697	1.363	1.796	2.201	2.328	4.537	6.718	9.947	19.047
12	0.695	1.356	1.781	2.177	2.299	4.500	6.681	9.899	18.990
13	0.694	1.350	1.771	2.160	2.282	4.473	6.650	9.856	18.941
14	0.692	1.345	1.761	2.145	2.264	4.449	6.624	9.817	18.896
15	0.691	1.341	1.753	2.131	2.249	4.427	6.602	9.782	18.854

$P(T_9 < -2,262) = 2,5\%$ ou $P(T_9 > 2,262) = 2,5\%$

$P(|T_9| \geq 2,262) = 5\%$