



**UNIVERSIDADE
ESTADUAL DE LONDRINA**

UNIVERSIDADE ESTADUAL DE LONDRINA
CCE - Centro de Ciências Exatas
DSTA - Departamento de Estatística (sala 11)
Prof. M.e Eng.^o Felinto Junior Da Costa
fjcosta@uel.br

Londrina, 05 de dezembro de 2024.

Índice

7

1	Introdução histórica daquilo que veio a se chamar estatística	9
1.1	Filosofia da ciência (teoria do conhecimento, epistemologia)	9
1.2	Diferentes usos relacionados ao termo, primeiros levantamentos, estudos e publicações (o passado distante)	12
1.3	Visualização de dados & Estudos e primeiras publicações	25
1.4	Pesquisadores cuja contribuição foi fundamental na área	28
1.5	Revista Biometrika	29
1.6	Eugenio	30
1.7	Estatística e <i>machine learning</i> : uma livre tradução deste link	32
2	Introdução conceitual essencial	37
2.1	Estatística descritiva	37
2.2	Estatística inferencial	38
2.3	Produção de conhecimento	38
2.4	População (universo) & amostra	41
2.5	Parâmetros e estatísticas	41
2.6	Tipos de variáveis	42
2.7	Indexação de dados (<i>i</i>)	43
2.8	Noções básicas sobre somatórios (Σ)	43
2.9	Análise combinatória (métodos de enumeração)	48
2.10	Fatoriais	59
2.11	Conectivos lógicos	60
2.12	Leis de De Morgan	60
2.13	Noções básicas para o uso de calculadora (Cassio fx-82MS)	61
2.14	Instalação do software R em conjunto com a interface gráfica RStudio	64

3 Introdução à estatística descritiva	67
3.1 Análise exploratória	67
3.2 Dados brutos, em rol, diagrama de ramos & folhas e de dispersão unidimensional	69
3.3 Sínteses numéricas descritivas	71
3.4 Medidas de forma (assimetria & curtose)	92
3.5 Apresentação tabular de dados	95
3.6 Apresentação gráfica de dados	109
4 Introdução ao cálculo de probabilidades	125
4.1 Introdução histórica	125
4.2 Conceitos essenciais	127
4.3 Probabilidade da união de eventos	150
4.4 Probabilidade de eventos condicionados	156
4.5 Dependência e independência de eventos	165
4.6 Probabilidade de eventos independentes (regra da cadeia)	172
4.7 Teorema de Bayes	175
4.8 Teoremas da Teoria das probabilidades	189
5 Introdução a variáveis aleatórias	195
5.1 Função massa de probabilidade (<i>Probability Mass Function - PMF</i>)	197
5.2 Função de densidade de probabilidade (<i>Probability Density Function - PDF</i>)	202
5.3 Esperança e variância de uma variável aleatória discreta	207
5.4 Esperança e variância de uma variável aleatória contínua	209
6 Introdução a modelos teóricos de probabilidade	211
6.1 Modelos teóricos discretos	211
6.2 Modelos teóricos do tempo de espera	226
6.3 Modelos teóricos contínuos	240
6.4 Tabelas	275
7 Introdução ao planejamento de pesquisas	281
7.1 Planejamento de pesquisas	283
7.2 Tipos de pesquisas	284
7.3 Principais etapas de uma pesquisa:	286
7.4 População	287
7.5 Censo	287
7.6 Amostra	287
7.7 Planejamento do levantamento amostral	288
7.8 Elaboração dos questionários	289
7.9 Técnicas de amostragem	290

7.10 Amostragem probabilística	290
7.11 Amostragem não probabilística	307
7.12 Dimensionamento de amostras	308
8 Introdução às estatísticas epidemiológicas	317
8.1 Tipos de estudos epidemiológicos	317
8.2 Estudos transversais	319
8.3 Estudos longitudinais	320
8.4 Terminologia	322
8.5 Medidas de risco, morte, associação e correlação	325
8.6 Sobrevida	329
8.7 Medidas de associação em estudos de coorte	330
8.8 <i>Odds ratio</i> (Razão das chances) em estudos de casos e controles	335
8.9 Correlação linear de Pearson	337
8.10 Intervalos de confiança	342
9 Introdução à distribuição das médias e diferenças entre médias amostrais e seus intervalos de confiança	347
9.1 Distribuições amostrais	347
9.2 Intervalos de confiança	352
9.3 Distribuição das médias amostrais e seus intervalos de confiança	357
9.4 Distribuição das diferenças de médias amostrais independentes e seus intervalos de confiança	394
9.5 Distribuição das diferenças de médias amostrais dependentes e seus intervalos de confiança	410
10 Introdução à distribuição das proporções amostrais e seus intervalos de confiança	413
10.1 Conceito elementar de uma proporção	413
10.2 Distribuição das proporções amostrais	414
10.3 Pobabilidades associadas à observação de uma proporção amostral \hat{p}	422
10.4 A aleatoriedade das proporções amostrais e o tamanho amostral	423
10.5 Intervalos de confiança para proporções amostrais	427
11 Introdução a testes de hipóteses	435
11.1 Filosofia da ciência	435
11.2 História	438
11.3 Conceitos	442
11.4 Natureza dos erros	444
11.5 Recomendações gerais	450
11.6 Efeito do limite central	450
11.7 Estruturas das hipóteses	452
11.8 Teste de uma média amostral	460

11.9 Teste de médias amostrais independentes de duas populações Normais	483
11.10 Teste de uma proporção amostral	516
11.11 Testes não paramétricos	526
11.12 Fluxograma auxiliar para escolha da estatística do teste de hipóteses	552
11.13 Tabelas	555
12 Introdução à Correlação Linear de Pearson e Regressão Linear Simples	561
12.1 Contexto histórico	561
12.2 Conceitos	562
12.3 Diagrama de dispersão	563
12.4 Coeficiente de correlação linear de Pearson	563
12.5 Teste de hipóteses para a correlação linear na população	566
12.6 Regressão linear simples	567
12.7 Modelo de regressão linear sob erros Normais	573
12.8 Teste de significância (global) do modelo	581
12.9 Teste de hipóteses para o coef. angular β	582
12.10 Teste de hipóteses para o coef. angular α	583
12.11 Coeficiente de determinação R^2	584
12.12 Intervalos de confiança	587
12.13 (SIMULADOR 2 COM t)	591
12.14 Verificações gráficas (visuais) das premissas do MMQO	591
12.15 Verificações adicionais	592
13 Orientações Gerais	597
13.1 Informações administrativas	597
13.2 Programas de atividade acadêmica	600

Módulo 1

Introdução histórica daquilo que veio a se chamar estatística

- do latim: *statisticum collegium* (colegiado dos assuntos do Estado);
- do alemão: *statistik* (Gottfried Achenwall, 1719-1772);
- no inglês: *statistics* (Enciclopédia Britânica, 1797).

De acordo com a revista do *Instituto Internacional de Estatística* cinco homens são chamados de fundadores dos primórdios da estatística:

- Hermann Conring;
- Gottfried Achenwall;
- Johan Peter Süßmilch;
- John Graunt, e
- William Petty.

1.1 Filosofia da ciência (teoria do conhecimento, epistemologia)

Estritamente falando, todo o conhecimento fora da matemática, da lógica demonstrativa (um ramo da mesma) e da taxonomia, encontra-se fundamentado em hipóteses (naturalmente há inúmeros

tipos de hipóteses, mas as que estamos a nos referir são altamente confiáveis, como as expressas em certas leis gerais da física e da química como, por exemplo, a Lei de Hooke as Leis de Kepler dentre tantas outras).

O *raciocínio lógico demonstrativo* permeia as ciências até onde a matemática lhe suporta; todavia, em si (assim como também a matemática), é incapaz de gerar novos conhecimentos sobre o mundo que nos rodeia.

O *método lógico demonstrativo* é próprio para objetos que existem apenas *idealmente*, que são construídos inteiramente pelo nosso pensamento.

O *método hipotético experimental* é próprio das ciências naturais (física, química, biologia, etc.), que observam seus objetos e realizam experimentos.

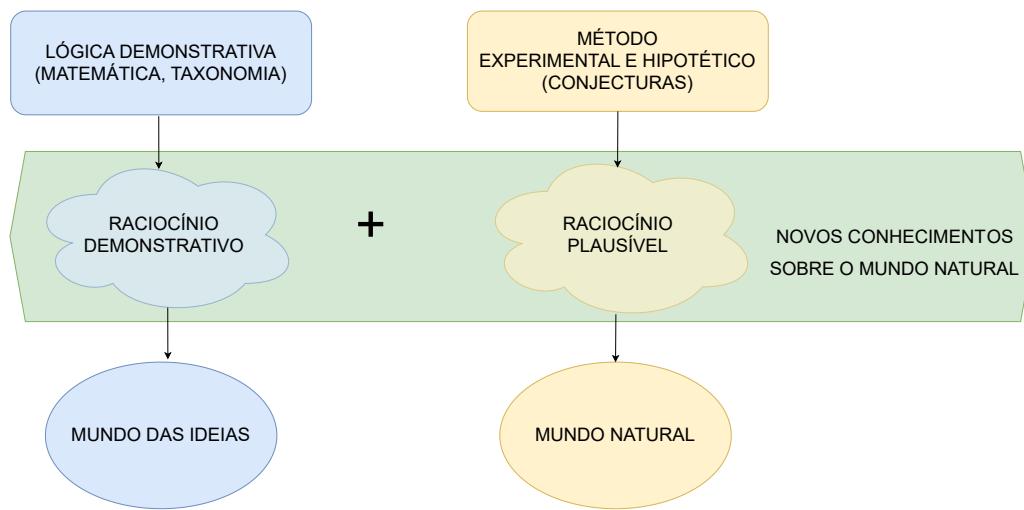


Figure 1.1: Método demonstrativo e Método experimental hipotético (George Polya, 1954)

Hipotético porque os cientistas partem de hipóteses sobre os objetos que guiam os experimentos e a avaliação dos resultados e *experimental* porque se baseia em observações e em experimentos, tanto para formular quanto para verificar as teorias.

O método hipotético experimental pode ser indutivo (fatos → lei geral) ou dedutivo (lei geral → fatos).

Isso é observado em qualquer que seja a área do conhecimento:

- ciências biológicas;
- ciências exatas;
- ciências agrárias;
- ciências humanas;
- ciência sociais e outras.

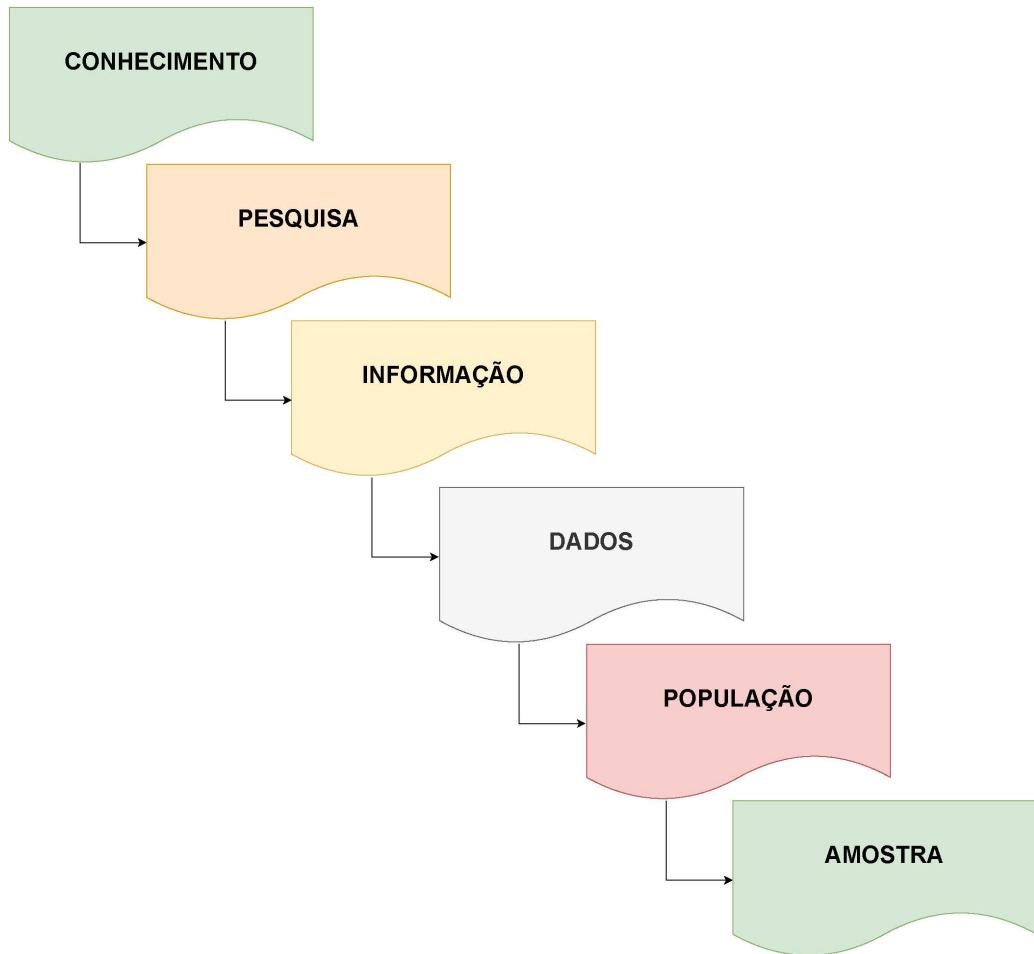


Figure 1.2: Representação esquemática do fluxo de infomações da amostra à produção de conhecimento

Assim, na investigação científica é imposto ao pesquisador formular perguntas que deverão ser apropriadamente respondidas.

- comparar esses resultados a outros valores; ou,
- comparar resultados obtidos pela aplicação de diferentes métodos/ou produtos (valores centrais, variabilidade, proporções) observados em diferentes amostras.

Uma hipótese é uma conjectura racional feita após um grande número de observações e experimentos; é uma tese que precisa ser confirmada ou verificada por meio de novas observações e experimentos.

Uma hipótese estatística é uma suposição feita sobre uma determinada característica de interesse de uma população sob estudo (um parâmetro) que subsiste (perdura, sobrevive, permanece incontestável) até que alguma informação sobre essa população seja estatisticamente significativa para contradizê-la.

“A ciência não consegue provar coisa alguma. Ela pode apenas refutar as coisas’’ (Karl Popper)

Uma teoria científica é, portanto, transitória. Uma conjectura temporariamente sustentada que um dia poderá ser refutada e substituída por outra. Conclusões baseadas em raciocínios plausíveis são provisórias, ao contrário daquelas produzidas por raciocínios lógico demonstrativos.

1.2 Diferentes usos relacionados ao termo, primeiros levantamentos, estudos e publicações (o passado distante)

O *Domesday Book* (link) foi encomendado em dezembro de 1085 por Guilherme, o Conquistador (*King William I*), que invadiu a Inglaterra em 1066.

O primeiro esboço foi concluído em agosto de 1086 e continha registros de 13.418 assentamentos nos condados ingleses ao sul dos rios Ribble e Tees (a fronteira com a Escócia) com informações sobre terras, proprietários, uso da terra, empregados e animais cujo propósito básico era fundamentar a taxação (Figura 1.4).

1.2. DIFERENTES USOS RELACIONADOS AO TERMO, PRIMEIROS LEVANTAMENTOS, ESTUDOS E PUBLICAÇÕES



Figure 1.3: Método experimental hipotético



Figure 1.4: Domesday Book

1.2. DIFERENTES USOS RELACIONADOS AO TERMO, PRIMEIROS LEVANTAMENTOS, ESTUDOS E PUBLICAÇÕES

O dramaturgo inglês William Shakespeare usou a palavra **statists** (estadistas e, portanto, num sentido não relacionado com números ou matemática) no diálogo da Cena II de Hamlet (link).

“Hamlet: Cercado assim por tantas vilanias, mesmo antes de eu poder dizer o prólogo, representava o cérebro. Sentei-me e escrevi com capricho nova carta. Já pensei, como os nossos estadistas, que é feio escrever bem, tendo insistido, até, em desaprendê-lo; mas, nessa hora muito bom me foi isso. Quererias saber qual o conteúdo da mensagem? [...]”

Um ponto de partida para a compreensão a ligação entre *estado* e *estatística* é Hermann Conring (1606-1681), professor de filosofia, medicina e política da Universidade de Helmstadt (atual Alemanha), criou um curso de Ciência política em 1660 para funcionários do estado, que descrevia e examinava as questões fundamentais do Estado. Nele a palavra *estatística* (parece ter) passado a ser considerada como uma disciplina autônoma que tinha por objetivo a descrição das coisas do Estado (Figura 1.5).

Microscopium Statisticum: quo status imperii Romano-Germanici cum primis extraordinariis, ad vivum reprezentatur (Statistical Microscope: An Analysis of the State, in which the State of the Germanic Roman Empire is vividly represented, above all extraordinary) é um livro cujo título normalmente chama mais atenção do que seu conteúdo: o uso de *statisticum* – que significa *do estado* – é reconhecido como um dos primeiros passos em direção ao uso da palavra *estatística* como hoje empregamos (Figura 1.6). Foi publicado sob o pseudônimo de Helenus Politanus em 1672.

Johan Peter Süßmilch é mais conhecido por seu notável trabalho de 1741 sobre população, conectando o direito natural e a aritmética política. O trabalho de Süßmilch foi amplamente referenciado por Robert Malthus (1798). Ele reconheceu não apenas a “proporção geométrica” da fertilidade humana, mas também o efeito de padrões de vida mais elevados sobre o problema populacional (Figura 1.7).

Com um sentido não relacionado com números ou matemática, a palavra **estatística** parece ter sido também proposta no século XVII, pelo historiador e professor alemão (à época Transilvânia) Martin Schmeitzel (1679-1747) da Universidade de Jena e, posteriormente adotada por seu aluno, (igualmente) historiador e jurista Gottfried Achenwall (1719-1772) em 1749, em *Abriß der neuen Staatswissenschaft der vornehmen Europäischen Reiche und Republiken* (Esboço da nova ciência política dos nobres impérios europeus e repúblicas, Figura 1.8).



Figure 1.5: Hermann Conring (1606-1681)



Figure 1.6: Microscopium Statisticum (ed. de 1672)



Figure 1.7: The Divine order in the circumstances of the human sex (1741)



Figure 1.8: Abriß der neuen Staatswissenschaft der vornehmen Europäischen Reiche und Republiken (1749)

Muitos anos depois, William Hooper usou a palavra **estatística** em sua tradução de *The Elements of Universal Erudition* (Elementos da Erudição Universal) escrita por Jacob Friedrich Freiherr von Bielfeld (1717-1770). Nesse livro, a *estatística* foi definida como a ciência que nos ensina o arranjo político de todos os estados modernos do mundo conhecido (novamente num sentido não associado a números ou matemática, Figura 1.9).



Figure 1.9: The Elements of Universal Erudition (1771)

Na Inglaterra a palavra estatística estava mais associada ao estudo de dados numéricos como modo de se obter *insights* sobre questões sociais e demográficas no país. Dois pioneiros nessa linha foram William Petty e John Graunt. Em 1687 o economista e filósofo inglês William Petty (1623-1687) publicou *Several Essays on Political Arithmetic* (Vários ensaios sobre aritmética política), sugerindo ao governo inglês a criação de um departamento para registro de *estatísticas* vitais (Figura 1.10).

O negociante inglês John Graunt (1620-1674) substituiu a crença pela evidência em *Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality* (Observações naturais e políticas feitas sobre as notas de mortalidade).

Nesse trabalho, realizado com dados coletados das paróquias de Londres entre 1604 e 1660, Graunt tirou as seguintes conclusões: que havia maior nascimento de crianças do sexo masculino, mas havia distribuição



Figure 1.10: Several Essays in Political Arithmetick (ed. de 1699)

aproximadamente igual de ambos os sexos na população geral; alta mortalidade nos primeiros anos de vida; maior mortalidade nas zonas urbanas em relação às zonas rurais (Figura 1.11).

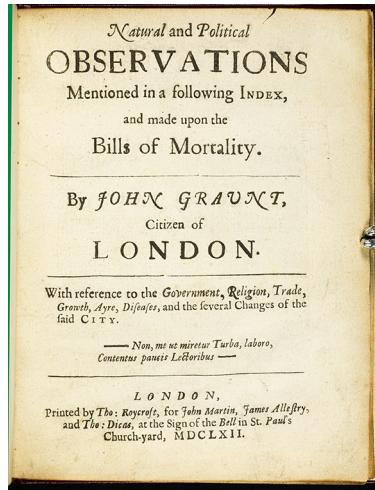


Figure 1.11: Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality (ed. de 1662)

O matemático e astrônomo inglês Edmond Halley (1656-1742) construiu em 1693, baseado em dados coletados na cidade (à época) alemã de Bresláu, uma *Life Table* (Tábua de sobrevivência), um estudo que analisa as probabilidades de sobrevivência e morte em relação à idade (Figura 1.12).

Age.	Per- sons.	Age.	Per- sons.										
1	1000	8	680	15	628	22	585	29	539	36	481	7	5547
2	855	9	670	16	622	23	579	30	531	37	472	14	4584
3	798	10	661	17	616	24	573	31	523	38	453	21	4270
4	760	11	653	18	610	25	567	32	515	39	454	28	3964
5	732	12	646	19	604	26	560	33	507	40	445	35	3604
6	710	13	640	20	598	27	553	34	499	41	436	42	3178
7	692	14	634	21	592	28	546	35	490	42	427	49	2709
												56	2194
Age.	Per- sons.	63	1694										
Curt.		Curt.		Curt.		Curt.		Curt.		Curt.		70	1204
43	417	50	346	57	272	64	202	71	131	78	58	77	692
44	407	51	335	58	262	65	192	72	120	79	49	84	253
45	397	52	324	59	252	65	182	73	109	80	41	100	107
46	387	53	313	60	242	67	172	74	98	81	34		
47	377	54	302	61	232	68	162	75	88	82	28		34000
48	367	55	292	62	222	69	152	76	78	83	23		
49	357	56	282	63	212	70	142	77	68	84	20		Sum Total.

Figure 1.12: Halley's life table (1693)

1.2. DIFERENTES USOS RELACIONADOS AO TERMO, PRIMEIROS LEVANTAMENTOS, ESTUDOS E PUBLICAÇÕES

O jurista e político escocês John Sinclair propôs que se realizasse uma detalhada pesquisa em 938 paróquias para elucidar a história natural e política de seu país (*Statistics Accounts*). Essa pesquisa fazia parte de um projeto muito mais ousado: *The Pyramid of Statistical Enquiry* (A Pirâmide da Pesquisa Estatística, Figura 1.13).



Figure 1.13: The Pyramid of Statistical Enquiry (1814)

Outro trabalho com vertente demográfica foi o de Sébastien Le Prestre, Marquês de Vauban (1633– 1707), intitulado *Méthode générale et facile pour faire le dénombrement des peuples*, detakhado num livro de 1686 (Figura 1.14).

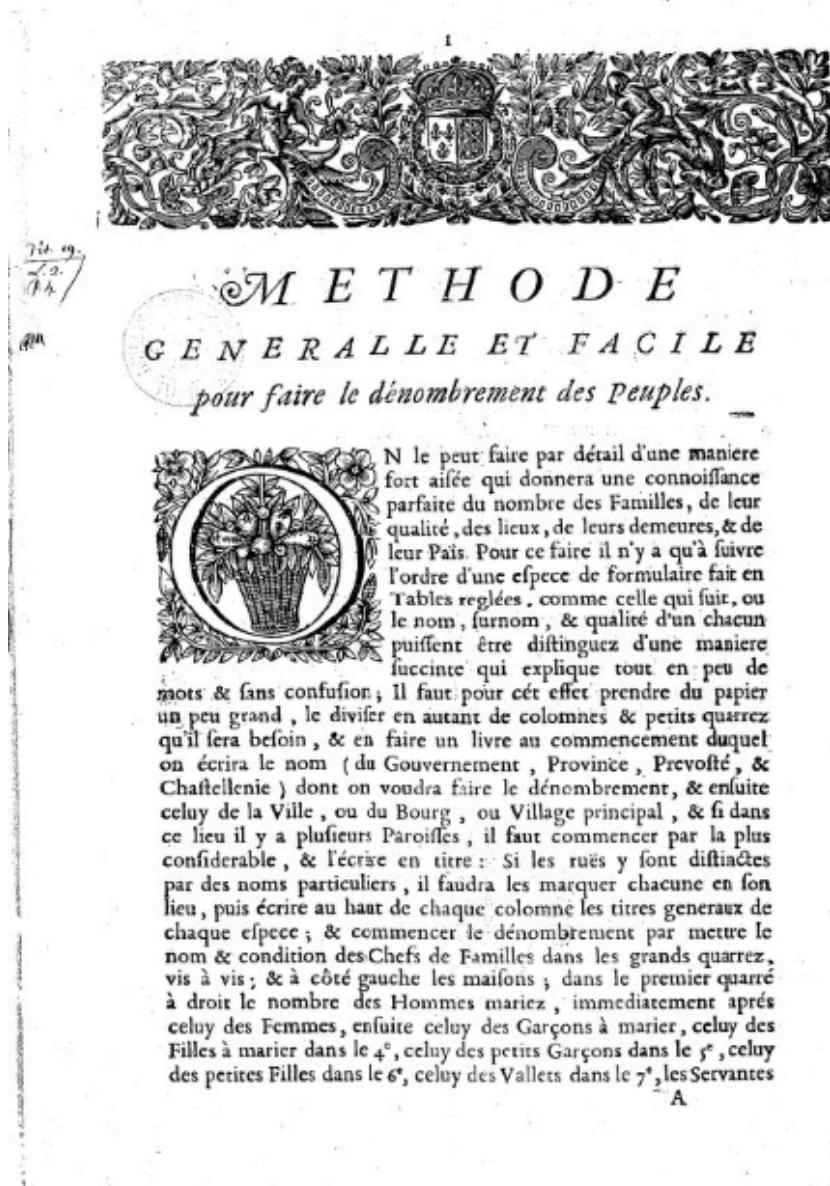


Figure 1.14: Méthode générale et facile pour faire le dénombrement des peuples (1686)

O médico inglês (considerado por alguns como o “pai” da epidemiologia moderna) John Snow (1813-1858) estudou a dispersão espacial dos casos de cólera em Londres e concluiu que sua causa residia na contaminação da água consumida (poço localizado na *Broad Street*, no distrito do *Soho*): *Report to the Cholera Outbreak in the Parish of St. James, Westminster during the Autumn of 1854* (Relatório sobre o surto de cólera na paróquia de St. James, Westminster durante o outono de 1854, Figura 1.15).

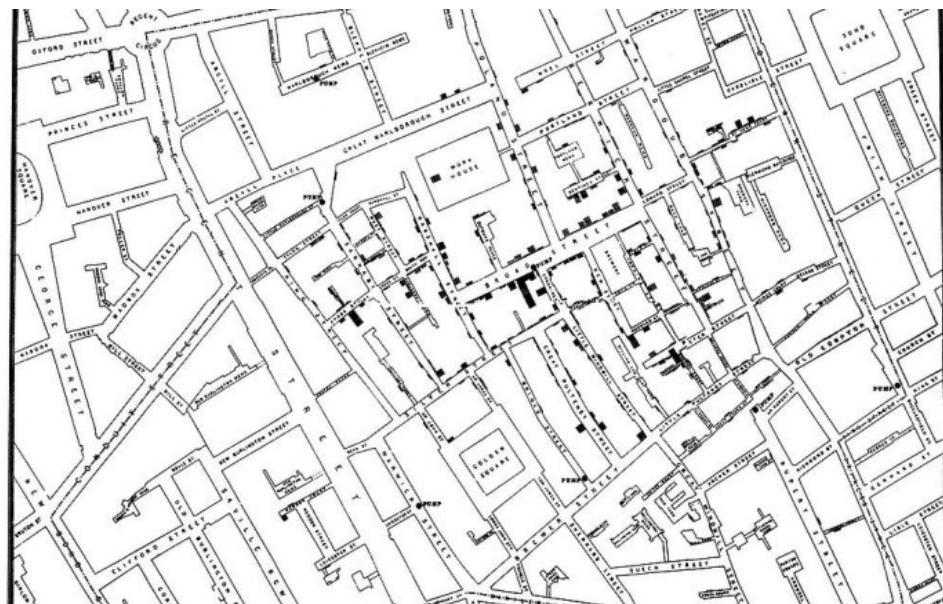


Figure 1.15: Mapa dos casos de cólera (1854)

1.3 Visualização de dados & Estudos e primeiras publicações

O teólogo e filósofo inglês Joseph Priestley (1733-1804) introduziu como inovação os primeiros gráficos com linha temporal, em que barras individuais eram usadas para visualizar o tempo de vida de uma pessoa e o todo pode ser usado para comparar a expectativa de vida de várias pessoas (Figura 1.16).

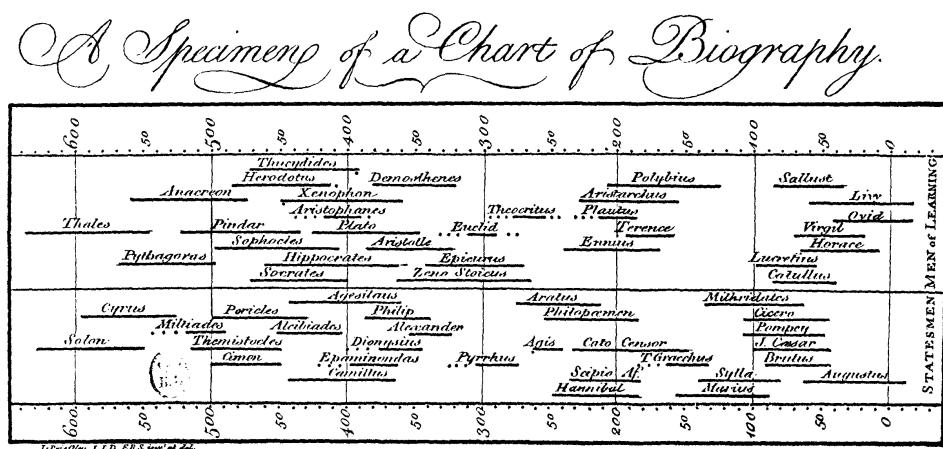


Figure 1.16: Expectativa de vida de diversas pessoas (1765)

O engenheiro e economista escocês William Playfair (1759-1823) é considerado comumente como fundador dos métodos gráficos para apresentação de estatísticas. Playfair concebeu vários tipos de diagramas para visualização de dados:

- em 1786, o gráfico de barras (Figura 1.17); e,
- em 1801, o gráfico de setores (Figura 1.18).



Figure 1.17: Commercial and Political Atlas (Atlas Comercial e Político de 1786): cada barra representa as exportações e importações da Escócia para 17 países em 1781



Figure 1.18: Statistical Breviary (Breviário Estatístico de 1801): proporção da extensão do Império Turco em diferentes regiões do mundo: África, Europa e Ásia, antes de 1789

A enfermeira inglesa Florence Nightingale (1820-1910) conduziu um trabalho pioneiro ao chegar no hospital militar britânico na Turquia em 1856, estabelecendo uma ordem e um método muito necessários aos registros médicos estatísticos e que indicaram serem as precárias práticas sanitárias o culpado da alta mortalidade (link) , Figuras 1.19 e 1.20.



Figure 1.19: Esse diagrama (coxcomb) feito durante a Guerra da Crimeia foi dividido igualmente em 12 setores, representando os meses do ano, com a área sombreada do setor de cada mês proporcional à taxa de mortalidade naquele mês. Seu sombreamento com código de cores indicava a causa da morte em cada área do diagrama



Figure 1.20: Gráfico de barras de Florence Nightingale mostrando as diferenças de mortalidade entre soldados britânicos e a população masculina inglesa geral (civis)

1.4 Pesquisadores cuja contribuição foi fundamental na área

Uma breve biografia de cada um dos pesquisadores a seguir relacionados pode ser obtida em: ([link](#)).

- Niccolò Fontana Tartaglia (Veneza à época, hoje Itália: 1499-1557)
- Girolamo Cardano (Pávia à época, hoje Itália: 1501-1576)
- Galileu Galilei (Florencia à época, hoje Itália: 1564-1642)
- Pierre de Fermat (França: 1607-1665)
- Blaise Pascal (França: 1623-1662)
- Jakob Bernoulli (Suíça: 1655-1705)
- Abraham de Moivre (França: 1667-1754)
- Thomas Bayes (Inglaterra: 1702-1761)
- Pierre-Simon Laplace (França: 1749-1827)
- Johann Carl Friedrich Gauss (Alemanha: 1777-1856)
- Lambert Adolphe Jacques Quêtelet (França à época, hoje Bélgica: 1796-1874)
- Pafnuti Lvovitch Chebyshev (Rússia: 1821-1894)
- Francis Galton (Inglaterra: 1822-1911)
- Wilhelm Lexis (Alemanha: 1837-1914)
- Thorvald Nicolai Thiele (Dinamarca: 1838-1910)

- Friedrich Robert Helmert (Saxônia: 1843-1917)
- Francis Ysidro Edgeworth (Inglaterra: 1845-1926)
- James Douglas Hamilton Dickson (Escócia: 1849-1931)
- Andrei Andreyevich Markov (Rússia: 1856-1922)
- Aleksandr Mikhailovich Lyapunov (Rússia: 1857-1918)
- Walter Frank Raphael Weldon (Inglaterra: 1860-1906)
- Karl Pearson (Inglaterra: 1857-1936)
- William Seally Gosset (Inglaterra: 1876-1937)
- Ronald Aylmer Fisher (Inglaterra: 1890-1962)
- Andrei Nikolaevich Kolmogorov (Rússia: 1903-1987)

1.5 Revista Biometrika

Karl Pearson (1857-1936) é amplamente considerado o fundador da disciplina moderna de **estatística**, e também é famoso como um filósofo da ciência, como escritor sobre o darwinismo social e como um dos principais impulsionadores para instalar a eugenia como a ciência social chave.

“Pretende-se que a *Biometrika* sirva como um meio não apenas de coletar ou publicar, sob um título, dados biológicos de um tipo não coletados sistematicamente ou publicados em outro lugar em qualquer outro periódico, mas também de disseminar um conhecimento de tal teoria estatística para o seu tratamento científico[...]”

Em outubro de 1901 foi fundada a *Biometrika, the Journal for the Statistical Study of Biological Problems* (*Biometrika*, o Jornal para o Estudo Estatístico de Problemas Biológicos) com o propósito de promover a análise estatística de fenômenos biológicos, isto é, a matematização da biologia.

Os fundadores da *Biometrika* foram Sir Francis Galton (primo de Charles Darwin), Walter Frank Raphael Weldon e Karl Pearson. A maior parte do trabalho foi feita por Pearson e Weldon, este último focando na edição do conteúdo (ou seja, o aspecto biológico) e o primeiro nos detalhes, incluindo correções de prova. Galton e o eugenista americano Charles Davenport atuaram, respectivamente, como consultor e editor.

Alguns dos tópicos abordados na revista incluem criminologia, botânica, zoologia, epidemiologia e outros aspectos da saúde humana. Na década de 1930, o caráter da *Biometrika* mudou, e “representou a vanguarda internacional da pesquisa em métodos estatísticos e sua aplicação na ciência e tecnologia”, ao invés de focar a hereditariedade.

Sir Francis Galton, que serviu como editor da primeira edição (1901), escreveu a Introdução, que incluiu uma declaração de propósito para a revista (link).

1.6 Eugenia

Em 16 de maio de 1883 *Sir Francis Galton* cunhou o termo *eugenia*, posteriormente descrevendo-o como “o estudo das agências sob controle social que podem melhorar ou reparar as qualidades raciais das gerações futuras, seja fisicamente ou mentalmente”.

Galton detalha o conceito em seu livro *Inquiries into Human Faculty and its Development*, e recomenda que indivíduos de famílias altamente classificadas em seu sistema de mérito sejam encorajados a se casar cedo e receber incentivos para ter filhos. Ele também condenou os casamentos tardios dentro desse mesmo grupo como “disgênicos” ou desvantajosos para a espécie humana.

A palavra *eugenia* foi extraída da palavra grega *eu*, que significa bem, e *genos*, que significa prole. Juntos, significaria *bem-nascido*.

Este livro caiu em domínio público e pode ser lido na íntegra online. A caracterização original de eugenica de Galton pode ser encontrada na página 17 desta edição de domínio público (Parte 1 do pdf):

“uma breve palavra para expressar a ciência de melhorar o rebanho, que não está de modo algum confinado a questões de acasalamento criterioso, mas que, especialmente no caso do homem, toma conhecimento de todas as influências que tendem, mesmo que em grau remoto, a dar ao raças ou linhagens de sangue mais adequadas uma melhor chance de prevalecer rapidamente sobre os menos adequados do que teriam de outra forma [...]”(Galton, 1883, p.17)

Mais recentemente, alguns grupos sociais viram no trabalho e opiniões de Fisher endossos ao colonialismo, à supremacia branca e à eugenica (como hoje interpretada).

Outros grupos, todavia, afirmam que Fisher não era racista nem eugenista, embora ele achasse que havia diferenças comportamentais e de inteligência entre os grupos humanos.

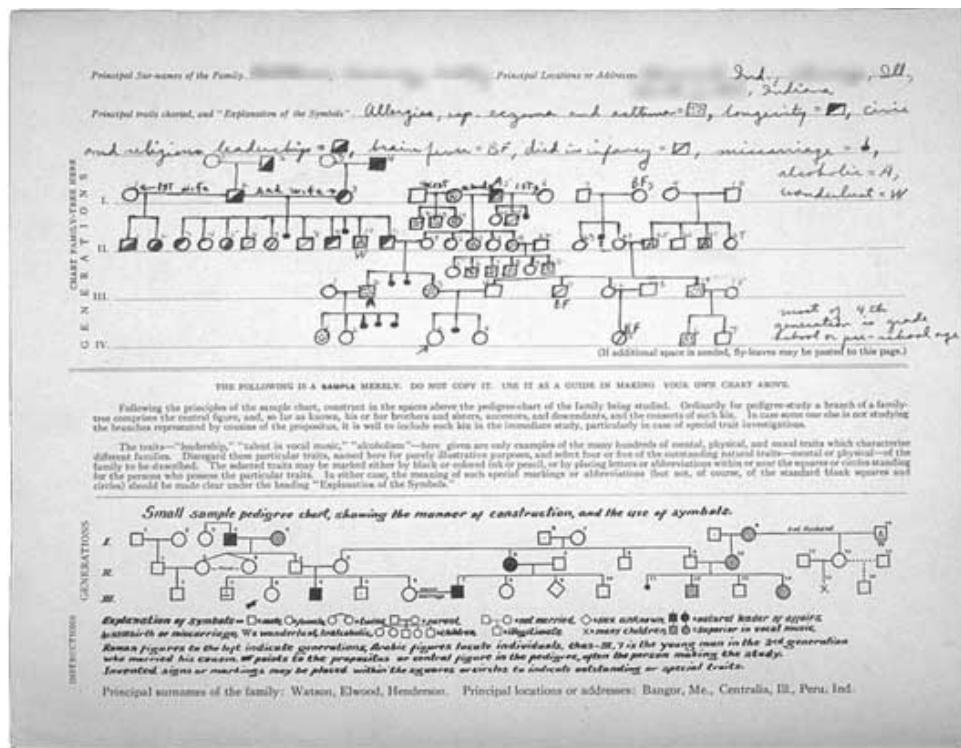


Figure 1.21: Gráfico de linhagens para alergias

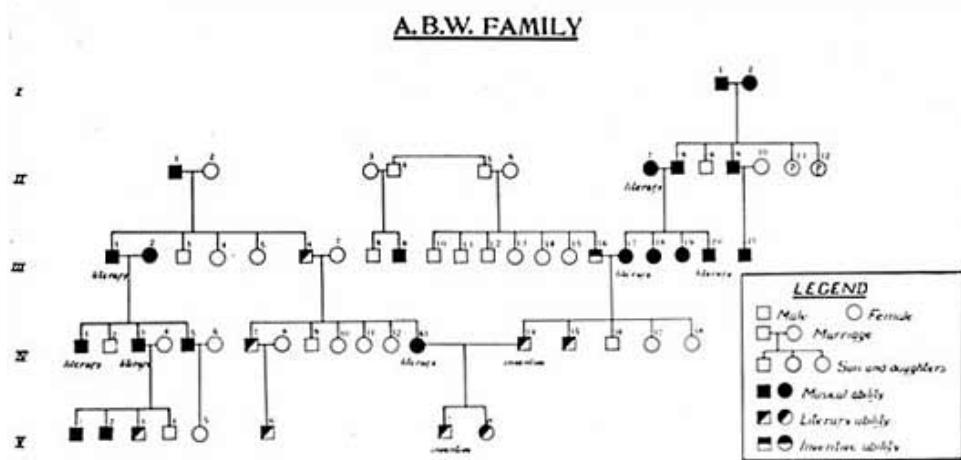


Figure 1.22: Gráfico de linhagens para aptidão musical



Figure 1.23: Linhas “normais” e “degeneradas” da família Kallikak (New Jersey)

1.7 Estatística e *machine learning* : uma livre tradução deste link

Para se raciocinar rigorosamente sob incerteza, precisamos invocar a linguagem da probabilidade (Zhang et al. 2020). Qualquer modelo que não forneça a quantificação da incerteza associada ao seu resultado provavelmente produzirá uma imagem incompleta e potencialmente enganosa.

Embora este seja um consenso irrevogável na estatística, um equívoco comum, embora muito persistente, é que os algoritmos de *machine learning* geralmente carecem de formas adequadas de quantificar a incerteza.

Apesar do fato dos dois termos existirem em paralelo e serem indistintamente utilizados, a percepção de que algoritmos de *machine learning* e a estatística implicam um conjunto de técnicas não sobrepostas permanece viva, tanto entre profissionais como acadêmicos.



Figure 1.24: Lei da Inegridade Racia (Virginia, EUA, 1924)



Figure 1.25: Licença para casamento

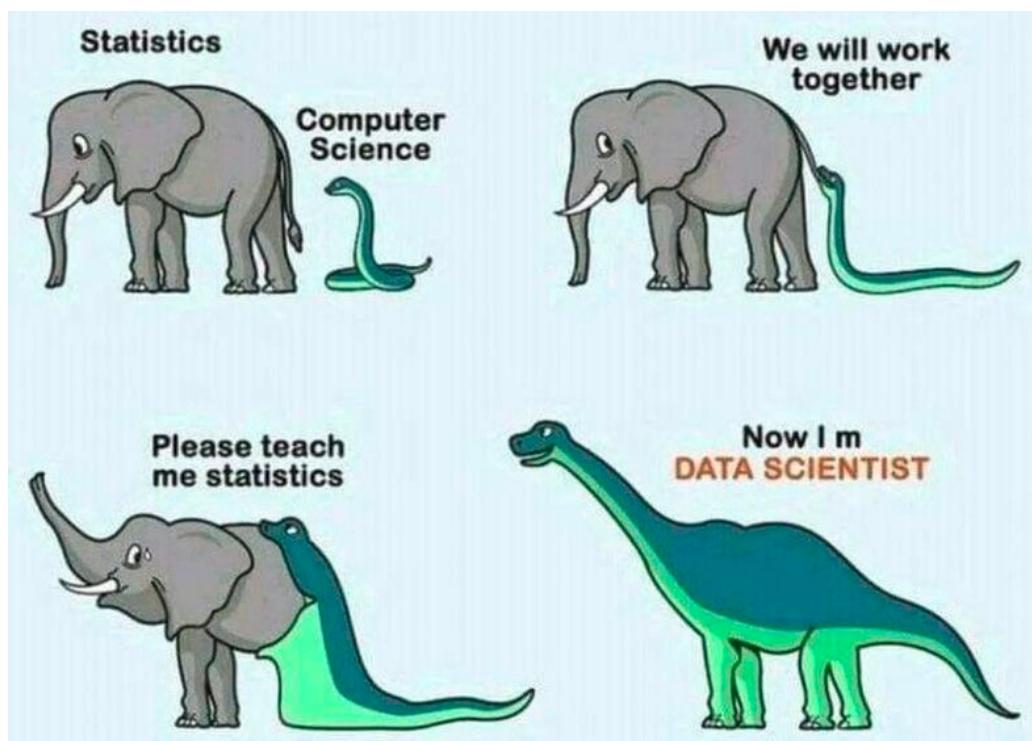


Figure 1.26: Autor: estatístico anônimo

Isso é vividamente retratado pela declaração provocativa (e potencialmente irônica) de Brian D. Ripley de que “o aprendizado de máquina é estatística menos qualquer verificação de modelos e suposições” que ele fez durante a “useR! 2004”, conferência de Viena que serviu para ilustrar a diferença entre aprendizado de máquina e estatística.

Na verdade, a relação entre estatística e algoritmos de *machine learning* é artificialmente complicada por tais afirmações e, na melhor das hipóteses, isto é lamentável, pois implica numa distinção profunda e qualitativa entre as duas disciplinas (Januschowski et al. 2020). O artigo de Leo Breiman (2001) é uma exceção notável, pois propõe diferenciar os dois com base na cultura científica, e não apenas nos métodos.

Embora as abordagens discutidas em Breiman (2001) constituam uma divisão admissível do espaço de análise e modelação de dados, os avanços mais recentes tornaram gradualmente esta distinção menos clara.

Na verdade, a tendência atual de investigação tanto em estatística com algoritmos de *machine learning* gravita no sentido de aproximar ambas as disciplinas. Numa era de necessidade crescente de que os resultados dos modelos de previsão sejam transformados em conhecimentos explicáveis e confiáveis, este é um desenvolvimento extremamente promissor e encorajador, uma vez que ambas as disciplinas têm muito a aprender uma com a outra. Junto com Januschowski et al. (2020) , argumentamos que é mais construtivo procurar um terreno comum do que introduzir fronteiras artificiais.

Módulo 2

Introdução conceitual essencial

“Estatística é a ciência de coletar, organizar, apresentar, analisar e interpretar dados[...]” (Ronald A. Fisher)

De modo geral, a estatística pode ser dividida em três grandes áreas:

- descritiva;
- probabilidade; e,
- inferencial.

2.1 Estatística descritiva

Nos primeiros trabalhos estatísticos, os dados coletados eram inicialmente apresentados na forma de tabelas e gráficos.

A **estatística descritiva** se ocupa de tudo o que seja relacionado a dados: coleta, processamento, descrição (seja na forma tabular ou gráfica) e sínteses numéricas (de locação, de dispersão, de repartição) sem inferir coisa alguma além da informação trazida pelos dados. Vem experimentando crescente uso em todas as áreas científicas e desenvolvimento:

- crescente uso de uma abordagem quantitativa em todas as ciências;
- disponibilidade de recursos computacionais;
- quantidade de dados coletados.

A palavra **estatística** pode assumir diferentes significados:

- no singular: **estatística**

- refere-se à ciência que comprehende métodos que são usados na coleta, análise, interpretação e apresentação de dados quantitativos ou qualitativos (numéricos ou não); e,
- denota uma medida ou fórmula específica (tais como uma média, um intervalo de valores, uma taxa de crescimento, um índice).
- no plural: **estatísticas**
 - refere-se a dados coletados de maneira sistemática com um propósito específico definido em qualquer campo de estudo (nesse sentido, as *estatísticas* também podem ser consideradas como agregados de fatos expressos em forma numérica).

2.2 Estatística inferencial

A **estatística inferencial** tem o objetivo de estabelecer níveis de confiança da tomada de decisão de associar uma estimativa amostral a um parâmetro populacional. Divide-se em estimação e testes de significância.

“Dedução e indução são procedimentos racionais que nos levam do já conhecido ao ainda não conhecido; isto é, permitem que adquiramos conhecimentos novos graças a conhecimentos já adquiridos.[...]"

Dedução.

Na dedução parte-se de uma verdade já conhecida para demonstrar que ela se aplica a todos os casos particulares iguais. Vai do geral ao particular.

Indução.

Na indução parte-se de alguns casos particulares iguais ou semelhantes para se estipular uma **lei geral**. Vai do particular ao geral.

Na dedução, dado **X**, infiro (concluo) **a, b, c, d**.

Na indução, dados **a, b, c, d**, infiro (concluo) **X**.

Exemplo: testes de aceleração (0-60 mph) feitos com 6 carros importados em 1999 resultaram nas seguintes medidas: 12,9 s; 16,50 s; 11,30 s; 15,20 s; 18,20 s e 17,70 s. Um estudo descritivo poderia afirmar que:

- metade dos dados coletados acelera de 0-60 mph em menos de 16,00 s; e
- a aceleração média de 0-60 mph é de 15,30 s.

Mas, a partir dessa amostra concluir que a aceleração média de **todos** os carros importados em 1999 seja de 15,30 s; ou, que **metade** dos carros importados em 1999 acelerem de 0-60 mph em menos de 16,00 s são afirmações que pertencem à **inferência estatística**.

2.3 Produção de conhecimento

“A ciência não consegue provar coisa alguma. Ela pode apenas refutar as coisas [...]” (Karl Popper)

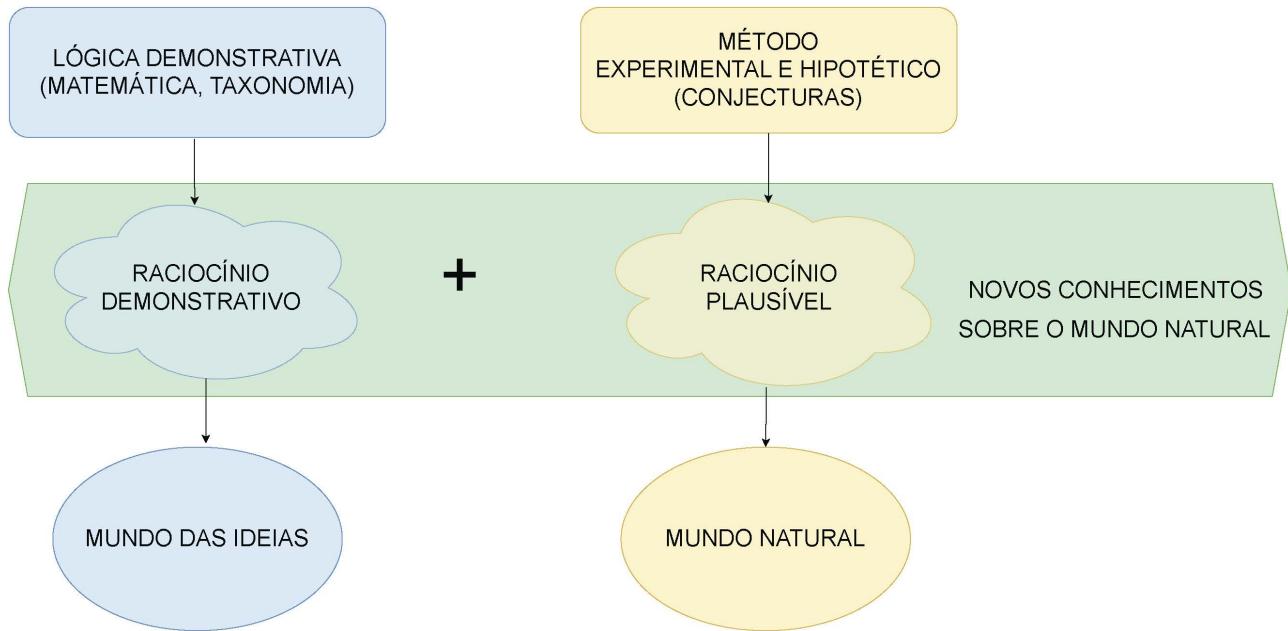


Figure 2.1: Método demonstrativo e Método experimental hipotético (George Polya, 1954)

Na expansão de qualquer área do conhecimento propomos hipóteses que serão avaliadas mediante a coleta de dados que, depois de analisados, revelarão informações que, eventualmente, nos conduzirão ao afastamento da hipótese original e à proposição de outras, num processo contínuo.

Uma investigação científica deve envolver, em linhas gerais:

- observação dos fatos;
- descrição das características essenciais, segundo o que se obteve através da observação;
- explicação dessas características descritivas;
- previsão; e,
- decisão pertinente à investigação.

O planejamento de uma pesquisa deve envolver, em linhas gerais:

- definição do *universo*: é necessário delimitar claramente, no tempo e espaço, o âmbito do inquérito, definindo, em termos precisos, o *universo* a ser trabalhado;
- exame das informações disponíveis: deve-se reunir todo o material existente: mapas, artigos, livros, relatórios relativos a levantamentos semelhantes;
- tipos de levantamentos: completo ou amostral;
- prazo;

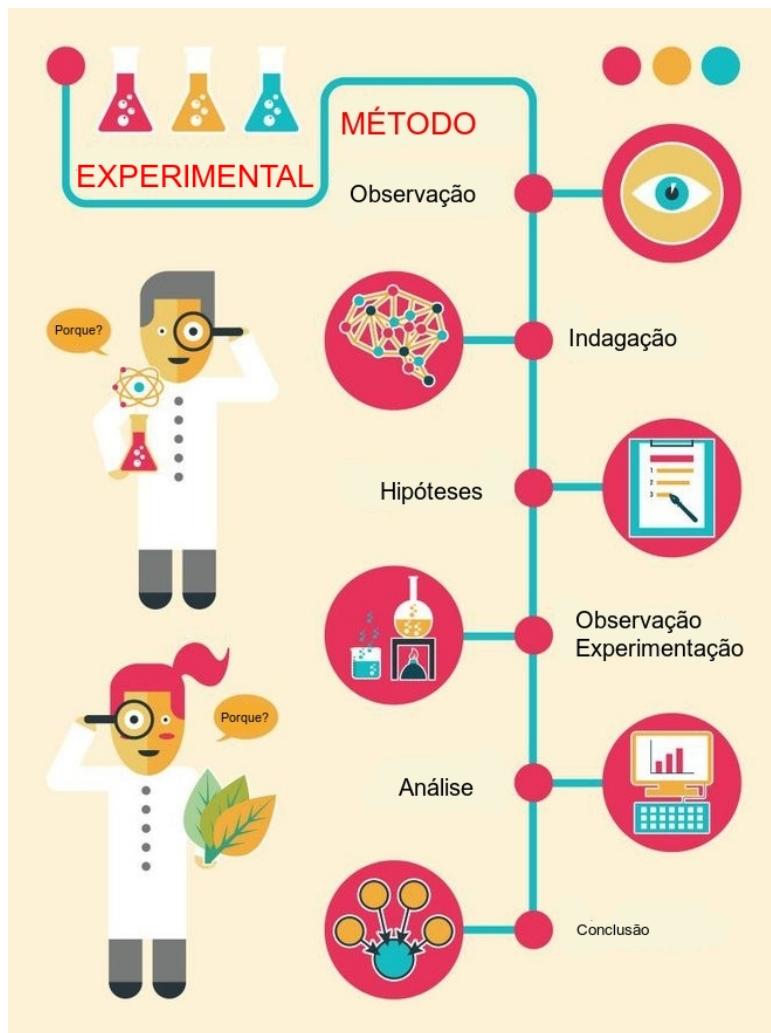


Figure 2.2: Método experimental hipotético

- custo;
- precisão.

2.4 População (universo) & amostra



Figure 2.3: Universo e amostra

Quase que, invariavelmente, em todo ramo de conhecimento, o pesquisador esbarra em uma série de limitações das mais variadas ordens (econômica, técnica, ética, geográfica, temporal,...) que impossibilitam o estudo dos dados e informações associados a todos os casos existentes (**população ou universo**).

Por essa razão, através de um procedimento estatístico denominado de amostragem, estuda-se uma população (universo) a partir de uma amostra. Amostra é, portanto, um subconjunto finito e representativo da população (universo), extraído de modo sistemático (planejado).

2.5 Parâmetros e estatísticas

É comum a adoção de letras gregas para as características descritivas que se referirem à população (universo) e letras do alfabeto latino para aquelas relativas à amostra extraída:

Característica estudada	Notação populacional	Notação amostral
Número de elementos	N	n
Média	μ ("mi")	\bar{x}
Variância	σ^2 ("sigma")	s^2
Desvio padrão	σ ("sigma")	s

Característica estudada	Notação populacional	Notação amostral
Proporção	Π (“pi”)	p ou \hat{p}

$A\alpha$ Alpha	$B\beta$ Beta	$\Gamma\gamma$ Gamma	$\Delta\delta$ Delta	$E\varepsilon$ Epsilon	$Z\zeta$ Zeta	$H\eta$ Eta	$\Theta\theta$ Theta
$I\iota$ Iota	$K\kappa$ Kappa	$\Lambda\lambda$ Lambda	$M\mu$ Mu	$N\nu$ Nu	$\Xi\xi$ Xi	$O\o$ Omicron	$\Pi\pi$ Pi
$P\rho$ Rho	$\Sigma\sigma\varsigma$ Sigma	$T\tau$ Tau	$Y\upsilon$ Upsilon	$\Phi\phi$ Phi	$X\chi$ Chi	$\Psi\psi$ Psi	$\Omega\omega$ Omega

Figure 2.4: Alfabeto grego

2.6 Tipos de variáveis

Variáveis quantitativas

- contínuas: são os dados com maior potencial de produzir informação significativa dentre todos: comprimentos, áreas, pesos, densidades; e,
- discretas: são dados com um pouco menos de informação que os de natureza contínua mas possuem mais informação que dados qualitativos: número de andares de um prédio, de degraus de uma escada, número de filhos de um casal.

Variáveis qualitativas

- ordinais: apresentam um pouco mais de informação que os dados qualitativos puramente nominais na medida que suas classes podem ser interpretadas como possuindo um ordenamento inerente: padrão construtivo (baixo, médio, alto), classe econômica de rendimento (baixa, média, alta), nível de escolaridade (fundamental, médio e superior); e,
- nominais: são dados a menor quantidade de informação: sexo, cor, códigos postais de cidades;

Codificação de variáveis qualitativas

- binárias: pela associação de valores numéricos: 0 ou 1 a uma variável qualitativa nominal que se apresente com apenas dois aspectos: sim ou não, ausência ou presença. Pela composição de mais variáveis binárias pode-se codificar variáveis que possuam um número maior de classes; e,
- *proxy*: pela associação de valores numéricos contínuos que guardam “correlação” com as classes da variável qualitativa nominal.

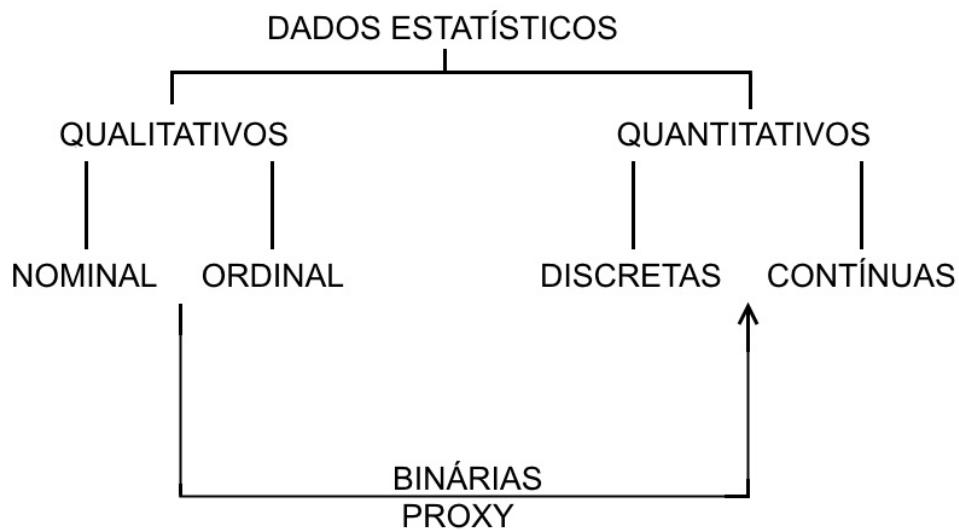


Figure 2.5: Tipos e codificações de variáveis

2.7 Indexação de dados (i)

Muitas operações matemáticas são representadas trazendo os valores dos dados indicados de modo genérico por letras (gregas ou romanas) e índices como, por exemplo, x_i . Tal notação está a indicar que, se dispuséssemos os dados em uma linha virtual (às vezes necessitando que estejam ordenados, como para a determinação de uma separatriz), cada um de seus valores estaria a ocupar uma *posição* indicada pelo índice i :

2.8 Noções básicas sobre somatórios (Σ)

Somatório é um operador matemático utilizado para simplificar expressões que envolvam soma de mais de um elemento.



Figure 2.6: Entendendo a indexação de dados

Digamos, por exemplo, que estamos interessados saber o total de comissões a pagar em um determinado setor de uma empresa.

Admita que esse setor tenha 6 funcionários: Pedro, Guilherme, Lucas, Maria, Fernanda e Roberto e que suas comissões sejam R\$ 3000; R\$ 3300; R\$ 3900; R\$ 2950; R\$ 3150 e R\$ 3450.

A representação da soma das comissões pode ser expressa de vários modos como, por exemplo, nesse extensa frase:

O total de comissões a pagar em um determinado setor de uma empresa é a Renda do Pedro mais a Renda do Guilherme mais a Renda do Lucas mais a Renda da Maria mais a Renda da Fernanda mais Renda do Roberto.

Atribuindo os valores para cada uma das rendas:

O total de comissões a pagar em um determinado setor de uma empresa é: : R\$ 3000 + R\$ 3300 + R\$ 3900 + R\$ 2950 + R\$ 3150 + R\$ 3450.

Chamando-se “O total de comissões a pagar em um determinado setor de uma empresa é” de X , teremos:

$$X = R\$ 3000 + R\$ 3300 + R\$ 3900 + R\$ 2950 + R\$ 3150 + R\$ 3450.$$

Para simplificar a representação dessa operação, vamos enumerar os funcionários: Pedro (1), Guilherme (2), Lucas (3), Maria (4), Fernanda (5) e Roberto (6). Além disso, vamos chamar a comissão a ser paga pela letra X .

Para diferenciar a fração da comissão X a ser paga a cada um dos funcionários podemos por um índice na letra X para indicar a quem estamos nos referindo. Assim X_1 seria a comissão do Pedro, X_2 a do Guilherme, X_3 a do Lucas, X_4 a da Maria, X_5 a da Fernanda e X_6 a do Roberto.

Com essa notação podemos representar matematicamente o total das comissões a pagar em um determinado setor de uma empresa por:

$$X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$$

Cada um desses fatores pode ser generalizado como um X_i , a comissão de um i -ésimo funcionário qualquer. Sabendo que o setor tem apenas 6 funcionários (Pedro, Guilherme, Lucas, Maria, Fernanda e Roberto) então esse i irá variar de 1 a 6 (Pedro:1, Guilherme: 2, Lucas: 3, Maria: 4, Fernanda: 5 e Roberto: 6).

Com todas essas considerações podemos representar a soma das comissões utilizando a notação matemática do somatório.

A letra grega maiúscula Σ (“sigma”) é habitualmente adotada na matemática para representar o somatório de uma quantidade de fatores. Assim, nosso exemplo da soma de 6 fatores (comissões) pode ser representada matematicamente por:

$$\sum_{i=1}^6 X_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$$

Observe que abaixo da letra Σ (“sigma”) vemos $i = 1$ indicando que o índice dos fatores (X) a serem somados (a i -ésima comissão) irá se iniciar pela comissão do primeiro funcionário, quando então $i = 1$.

Acima da letra Σ (“sigma”) vemos o número 6 indicando que o índice dos fatores (X) a serem somados irá se dar até o valor da comissão do sexto funcionário, quando então $i=6$.

Generalizando-se para uma soma de n fatores X :

$$\sum_{i=1}^n X_i.$$

A representação matemática do somatório pode ser inserida junto a qualquer outra operação como, por exemplo, podemos, depois de realizar a soma, dividi-la por um valor n qualquer

$$\frac{\sum_{i=1}^n X_i}{n}$$

ou elevá-la ao quadrado:

$$\left(\sum_{i=1}^n X_i \right)^2$$

Atenção para a diferença entre essas duas operações:

$$\left(\sum_{i=1}^n X_i \right)^2$$

e

$$\sum_{i=1}^n X_i^2$$

A primeira indica que devemos realizar a soma dos fatores e só então elevar esse resultado ao quadrado. A segunda indica que devemos realizar a soma dos quadrados de cada um dos fatores.

```
library(formattable)
comissoes=c(3000, 3300, 3900, 2950, 3150, 3450)
```

```
#Somatório das comissões
currency(sum(comissoes),
  symbol = "R$",
  digits = 2L,
  format = "f",
  big.mark= ".",
  decimal.mark= ",",
  sep= " ")
```

```
## [1] R$ 19.750,00
```

```
#Somatório das comissões dividido pelo número de comissões
currency(sum(comissoes)/length(comissoes),
  symbol = "R$",
  digits = 2L,
  format = "f",
  big.mark= ".",
  decimal.mark= ",",
  sep= " ")
```

```
## [1] R$ 3.291,67
```

```
#Quadrado do somatório das comissões
currency(sum(comissoes)^2,
  symbol = "R$",
  digits = 2L,
  format = "f",
  big.mark= ".",
  decimal.mark= ",",
  sep= " ")
```

```
## [1] R$ 390.062.500,00
```

```
#Somatório dos quadrados das comissões
currency(sum(comissoes^2),
  symbol = "R$",
  digits = 2L,
  format = "f",
  big.mark= ".",
  decimal.mark= ",",
  sep= " ")
```

```
## [1] R$ 65.627.500,00
```

2.9 Análise combinatória (métodos de enumeração)

A análise combinatória (ou métodos de enumeração) é um conjunto de técnicas para agrupamento de objetos conforme regras definidas e obtenção, através de cálculos, do número de agrupamentos possíveis.

2.9.1 Princípio básico da contagem (regra da multiplicação)

Suponha a realização de dois experimentos. Se o experimento E pode gerar qualquer um de n resultados possíveis ($E_1, E_2, E_3, \dots, E_n$) e se, para cada um dos resultados do experimento, houver m resultados possíveis para o experimento F ($F_1, F_2, F_3, \dots, F_m$), então os dois experimentos possuem conjuntamente $n \cdot m$ diferentes resultados possíveis.

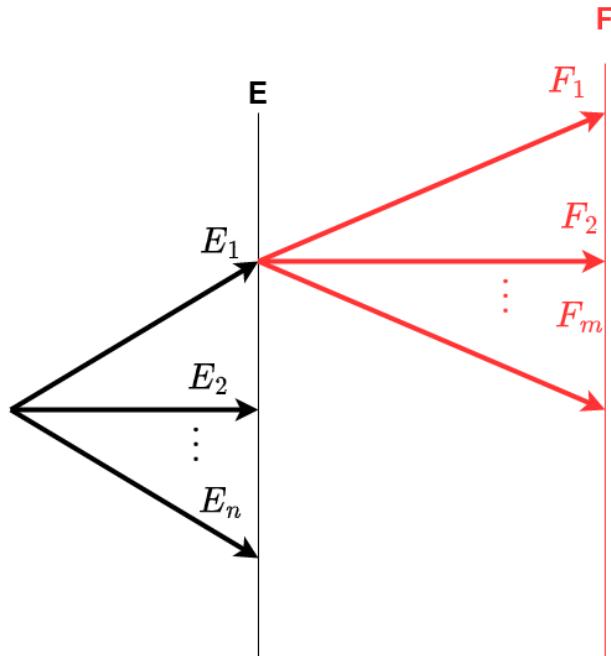


Figure 2.7: Regra da multiplicação

Esse princípio recebe o nome de *Princípio multiplicativo*, e é aplicado nos casos em que os eventos são interligados pelo conectivo **e**, característico de decisões sucessivas: ocorrem os dois.

Se um homem tem 2 camisas e 4 gravatas, então ele tem $2 \times 4 = 8$ formas de combinar uma camisa com uma gravata.

Um diagrama como ilustrado na Figura 2.8 (denominado **diagrama de árvore** em virtude de sua aparência) geralmente é usado para explicar o princípio acima

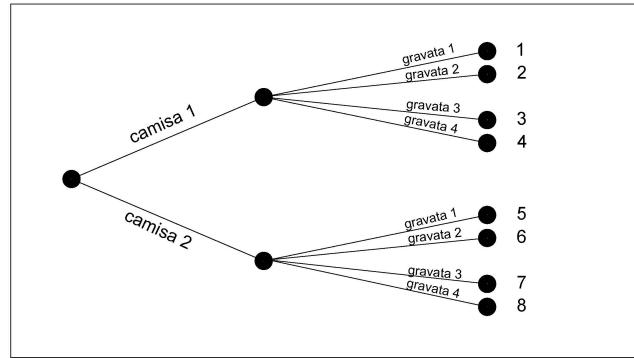


Figure 2.8: Diagrama de árvore

Ao lançarmos uma moeda três vezes (assumindo-se que K: cara e C: coroa) haverá $2 \times 2 \times 2 = 8$ possibilidades distintas.

O **diagrama de árvore** associado será (cf. Figura 2.9):

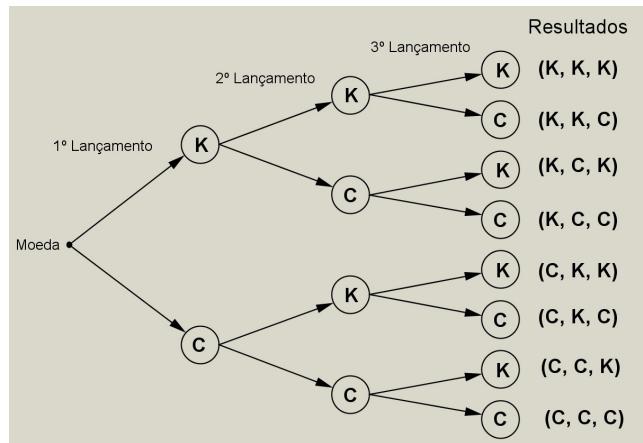


Figure 2.9: Diagrama de árvore

2.9.2 Regra da adição

Suponha agora os mesmos experimentos E e F que geram n e m resultados possíveis ($E_1, E_2, E_3, \dots, E_n$ e $F_1, F_2, F_3, \dots, F_m$), mas que esses experimentos não estejam mais alinhados sequencialmente: ocorre o evento E ou o evento F . Então o número de maneiras pelas quais o evento E ou o evento F poderão se manifestar será de $n + m$ maneiras diferentes:

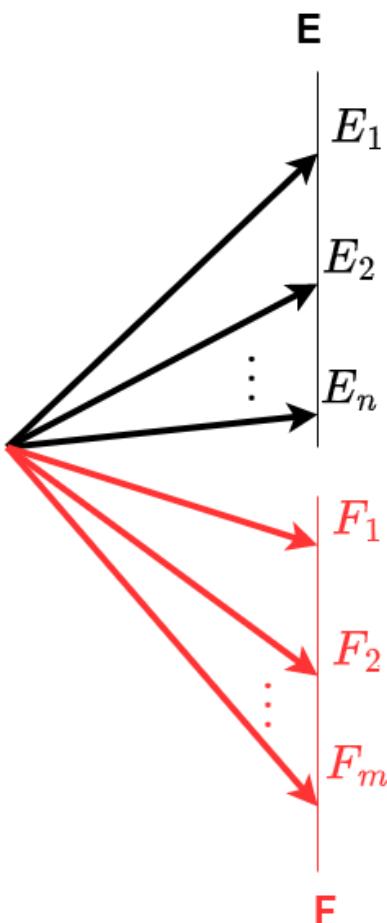


Figure 2.10: Regra da adição

Esse princípio recebe o nome de *Princípio aditivo*, e é aplicado nos casos em que os eventos são interligados pelo conectivo **ou**, característico de eventos mutuamente exclusivos: ocorre um ou outro.

Uma cantina de um colégio possui três tipos de sucos e dois tipos de refrigerantes. Um aluno pode adquirir apenas 1 suco ou 1 refrigerante. Quantas possibilidades de escolha ele tem?

Seja E_1 definido como escolher um tipo de suco ($n_1 = 3$) e E_2 definido como escolher 1 tipo de refrigerante ($n_2 = 2$). Então o número total de possíveis escolhas será dado aplicando-se o princípio aditivo:

$$n_1 + n_2 = 5$$

2.9.3 Permutações (ordenação de elementos)

Suponha n objetos diferentes. Permutar os n objetos equivaleria a colocá-los dentro de uma caixa com n compartimentos em **alguma ordenação**.

O primeiro compartimento pode ser ocupado por qualquer um dos n objetos, o segundo por $n - 1$ e o último por apenas 1 objeto.

Assim, pelo princípio básico vemos que essa caixa poderá ser carregada de:

$$n.(n - 1).(n - 2).1 = n! \text{ maneiras diferentes.}$$

Exemplo: quantos diferentes arranjos ordenados das letras a, b e c são possíveis?

Pela enumeração direta vemos que são 6, ou seja, ‘abc’, ‘acb’, ‘bac’, ‘bca’, ‘cab’ e ‘cba’, resultado de $3! = 6$.

Exemplo: quantas diferentes ordens de rebatedores são possíveis em um time de beisebol formado por 9 jogadores?

$9! = 362.880$ ordenamentos possíveis para os rebatedores.

Exemplo: uma turma de teoria da probabilidade é formada por 6 estudantes do sexo masculino e 4 do sexo feminino. Aplica-se uma prova e os estudantes são classificados de acordo com o seu desempenho. Suponha que nenhum dos estudantes tenha tirado a mesma nota. (a) Quantas diferentes classificações são possíveis? (b) Se os estudantes do sexo masculino forem classificados apenas entre si e também os do sexo feminino apenas entre si, quantas diferentes classificações são possíveis?

- a) Como cada classificação corresponde a um arranjo particular das 10 pessoas, a resposta é $10! = 3.628.800$.
- (b) Como há 6! possíveis classificações dos homens entre si e 4! classificações possíveis das mulheres entre si, segue do princípio básico que há $(6!)(4!) = (720)(24) = 17.280$ classificações possíveis neste caso.

2.9.4 Arranjos sem repetição

Considere um conjunto de n objetos diferentes. Suponha que desejamos selecionar uma quantidade p desses objetos, onde $p < n$, e dispor os p objetos escolhidos em uma ordem específica.

Quando selecionamos e ordenamos um subconjunto de objetos de um conjunto maior, estamos formando arranjos ordenados. Nesse caso, o número total de arranjos possíveis de p elementos retirados de um conjunto de n objetos distintos (**sem repetir nenhum elemento**) é dado por:

$$P_{(n,p)} = \frac{n!}{(n-p)!}$$

A fórmula para $P_{(n,p)}$ considera tanto a seleção de p elementos quanto a ordenação deles, de modo que qualquer alteração na ordem dos elementos resulta em um novo arranjo.

Dessa forma, arranjos com os mesmos objetos em ordens distintas são considerados distintos.

Exemplo: quantos agrupamentos distintos, formados por 3 letras cada, podem ser formados com as 7 letras: A, B, C, D, E, F, G, considerando que não é permitido repetir nenhuma letra e que a ordem dos elementos importa?

$$\begin{aligned}
 n &= 7 \\
 p &= 3 \\
 P_{(n,p)} &= \frac{7!}{(7-3)!} \\
 &= \frac{7!}{4!} = \\
 &= \frac{7 \times 6 \times 5 \times 4!}{4!} \\
 &= 7 \times 6 \times 5 = 210
 \end{aligned}$$

$ABC, ACB, BAC, BCA, CAB, CBA, \dots$

2.9.5 Arranjos com repetição

Considere um conjunto de n objetos diferentes. Suponha que desejamos selecionar uma quantidade p desses objetos, onde $p < n$, e dispor os p objetos escolhidos em uma ordem específica.

Ao permitirmos a repetição dos objetos no agrupamento, cada posição pode ser ocupada por qualquer um dos n objetos, independentemente das escolhas anteriores.

Dessa forma, o número de arranjos com repetição de p elementos selecionados de um conjunto de n objetos distintos é dado por:

$$P_{(n,p)} = n^p$$

Essa fórmula ocorre porque, ao preencher cada uma das p posições com qualquer um dos n objetos, multiplicamos n possibilidades para cada posição, resultando em:

$$n \times n \times \dots \times n = n^p$$

Exemplo: Quantos agrupamentos diferentes (onde a ordem dos elementos é razão para distinção: **permutações**) formados por **3 letras cada** podem ser formados com as **7 letras**: A, B, C, D, E, F, G **com repetição**?

$$\begin{aligned} n &= 7 \\ p &= 3 \\ P_{(n,p)} &= n^p \\ &= 7^3 = 343 \end{aligned}$$

Primeira posição: temos 7 opções (uma das 7 letras). Segunda posição: também temos 7 opções, pois a repetição é permitida. Terceira posição: novamente, temos 7 opções.

AAA, AAB, AAC, ...

2.9.6 Combinações sem repetição

Em uma *permutação*, a *ordem* dos objetos em cada agrupamento é essencial: qualquer alteração na ordem cria um **agrupamento distinto**. Por exemplo, o agrupamento *abc* é diferente de *bca* numa permutação, pois a ordem importa.

Em muitos problemas, entretanto, estamos interessados somente na seleção dos objetos, **sem considerar a ordem** em que eles aparecem.

Nesse caso, os agrupamentos onde os mesmos elementos aparecem em *ordens diferentes* são considerados **equivalentes**. Tais seleções são chamadas de combinações. Por exemplo, *abc* e *bca* representam uma mesma combinação, pois contêm os **mesmos** elementos *independentemente da ordem*.

O conceito de uma combinação refere-se a um conjunto de n objetos distintos, dos quais escolhemos p objetos ($p < n$) **sem repetição**. Assim:

- a *ordem* não importa: então *abc* é igual a *bca* (os agrupamentos que contêm os mesmos objetos em qualquer ordem **são considerados iguais**);
- a *repetição* não permitida: cada objeto só aparece uma vez em cada agrupamento.

O número total de combinações possíveis de p objetos selecionados de um conjunto de n objetos distintos é dado por:

$$C_{(n,p)} = \frac{n!}{p! \times (n-p)!}$$

em que o **numerador** ($n!$) representa o número total de maneiras de permutar todos os n objetos, o **denominador** ($p! \times (n-p)!$) remove as redundâncias causadas pela permutação dos p objetos escolhidos ($p!$) e das maneiras de ordenar os $(n-p)$ objetos restantes ($(n-p)!$).

A fórmula resulta no número de agrupamentos únicos onde **apenas a composição do conjunto importa**, e a ordem dos elementos dentro de cada conjunto não afeta a contagem.

Exemplo: Qual é número de formas nas quais 3 cartas podem ser escolhidas ou selecionadas de um total de 8 cartas diferentes?

$$n = 8$$

$$p = 3$$

$$\begin{aligned} C_{(n,p)} &= \frac{n!}{p! \times (n-p)!} \\ C_{(8,3)} &= \frac{8!}{3!(8-3)!} \\ &= \frac{8!}{3! \times 5!} \\ &= \frac{8 \times 7 \times 6 \times 5!}{3! \times 5!} \\ &= \frac{8 \times 7 \times 6}{3!} = 56 \end{aligned}$$

O número total de combinações com repetição, de p objetos selecionados de n (também chamado de combinações de n elementos tomados p a cada vez com repetição) é representado por:

$$C_{(n+p-1,p)} = \frac{(n+p-1)!}{p! \times (n-1)!}$$

Exemplo: um comitê de três pessoas deve ser formado a partir de um grupo de 20 pessoas. Quantos comitês diferentes são possíveis?

$$n = 20$$

$$p = 3$$

$$\begin{aligned} C_{(n,p)} &= \frac{n!}{p! \times (n-p)!} \\ C_{(20,3)} &= \frac{20!}{3! \times (20-3)!} \\ &= \frac{20 \times 19 \times 18 \times 17!}{3! \times 17!} \\ &= \frac{20 \times 19 \times 18}{3 \times 2 \times 1} \\ &= \frac{6840}{6} = 1140 \end{aligned}$$

Exemplo: de um grupo de cinco mulheres e sete homens, quantos comitês diferentes formados por duas mulheres e três homens podem ser formados? E se dois dos homens estiverem brigados e se recusarem a trabalhar juntos?

Para resolver esse problema, vamos dividi-lo em duas partes:

1. calcular o número de comitês possíveis formados por duas mulheres e três homens (combinações $C_{(5,2)}$ e $C_{(7,3)}$)
2. Calcular o número de comitês possíveis se dois dos homens estiverem brigados e se recusarem a trabalhar juntos.

Parte 1: Comitês sem restrição

Escolha das Mulheres: Temos 5 mulheres e precisamos escolher 2. O número de maneiras de fazer isso é uma combinação, dada por:

$$C_{(5,2)} = \frac{5!}{2! \times (5-2)!} = \frac{5 \times 4}{2 \times 1} = 10$$

Escolha dos Homens: Temos 7 homens e precisamos escolher 3. O número de maneiras de fazer isso também é uma combinação, dada por:

$$C_{(7,3)} = \frac{7!}{3! \times (7-3)!} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35$$

O **número total de comitês** é o produto das duas combinações (princípio básico):

$$\text{Total} = C_{(5,2)} \times C_{(7,3)} = 10 \times 35 = 350$$

Portanto, sem restrições, há **350 comitês diferentes** que podem ser formados.

Parte 2: Comitês com restrição (Dois Homens Brigados)

Para calcular o número de comitês possíveis com essa restrição, usamos o princípio da exclusão: (total de comitês sem restrição - total de comitês com restrição). Começamos contando quantos comitês incluem os dois homens brigados: A e B .

Se ambos estão no comitê então precisamos escolher apenas mais 1 homem entre os 5 homens restantes. O número de maneiras de fazer isso é:

$$C_{(5,1)} = 5$$

Então, do total de **35 possíveis grupos de 3 homens formados sem restrição** apenas 5 são grupos onde A e B estão presentes, resultando então em $35 - 5 = 30$ grupos onde A e B estão ausentes.

Compondo com o número de possíveis grupos de mulheres (princípio básico) teremos:

$$\text{Total} = C_{(5,2)} \times 30 = 10 \times 30 = 300$$

Portanto, com essa restrição, há **300 comitês diferentes** que podem ser formados.

2.9.7 Combinações com Repetição

Em muitos problemas de contagem, estamos interessados em **selecionar** um subconjunto de objetos de um conjunto maior, sem nos preocupar com a ordem dos objetos. Isso é chamado de **combinação**. No entanto, ao contrário das combinações convencionais, em alguns casos é permitido que os mesmos objetos sejam escolhidos mais de uma vez. Essas são conhecidas como **combinações com repetição**.

Nas **combinações com repetição**, selecionamos p objetos a partir de um conjunto de n objetos distintos, **permitindo** que cada objeto seja escolhido mais de uma vez. Nesse contexto: a **ordem dos objetos não importa**, ou seja, uma seleção de A, B, A é considerada igual a A, A, B e a **repetição é permitida**, então podemos escolher o mesmo objeto mais de uma vez.

Combinações com repetição são úteis em contextos onde é necessário selecionar subconjuntos com elementos repetidos, como: - escolha de itens com reposição (como selecionar moedas em valores específicos). - problemas de contagem em álgebra combinatória. - distribuição de recursos em cenários onde itens podem ser alocados mais de uma vez.

O número de combinações **com repetição** para selecionar p objetos de um conjunto de n objetos distintos é dado por:

$$C_{(n+p-1,p)} = \frac{(n+p-1)!}{p! \times (n-1)!}$$

Essa fórmula reflete o fato de que, ao permitir repetições, cada seleção possível pode ser visualizada como uma combinação com um conjunto “expandido” de opções, onde as escolhas de um mesmo objeto múltiplas vezes são válidas e contadas.

Exemplo: quantas maneiras existem de selecionar 3 frutas de um conjunto com 5 tipos diferentes (maçã, banana, laranja, uva e pera), onde repetições são permitidas?

$$\begin{aligned}
n &= 5 \\
p &= 3 \\
C_{(n+p-1,p)} &= \frac{(n+p-1)!}{p! \times (n-1)!} \\
C_{(5+3-1,3)} &= \frac{(5+3-1)!}{3! \times (5-1)!} = \frac{7!}{3! \times 4!} \\
&= \frac{7 \times 6 \times 5 \times 4!}{3! \times 4!} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = \frac{210}{6} = 35
\end{aligned}$$

Exemplo: supondo que você queira comprar um sorvete com 4 bolas em uma sorveteria que possui 3 sabores disponíveis: chocolate, baunilha e morango. De quantos modos diferentes você pode fazer esta compra? (Note que nesta combinação é possível repetir a ordem de dois ou mais sabores, assim tratando de uma combinação com repetição).

$$\begin{aligned}
n &= 3 \\
p &= 4 \\
C_{(n+p-1,p)} &= \frac{(n+p-1)!}{p! \times (n-1)!} \\
C_{(n+p-1,p)} &= \frac{(3+4-1)!}{4!(3-1)!} = 15
\end{aligned}$$

2.10 Fatoriais

$$n! = n.(n-1).(n-2) \dots (n-(n-1)).(n-n)$$

$$1! = 1$$

$$0! = 1$$

$$\begin{aligned}
P_{(n,n)} &= \frac{n!}{(n-n)!} = \frac{n!}{0!} = n! \\
C_{(n,0)} &= \frac{n!}{0! \times (n-0)!} = \frac{n!}{1 \times (n)!} = 1 \\
C_{(n,1)} &= \frac{n!}{1!(n-1)!} \\
&= \frac{n!}{(n-1)!} \\
&= \frac{n \times (n-1)!}{(n-1)!} = n
\end{aligned}$$

2.11 Conectivos lógicos

Muitos dos problemas ligados à probabilidade de ocorrência de eventos são propostos com o auxílio de conectivos lógicos:

- **Proposição:** a afirmação de que algo é verdadeiro. Após analisarmos qualquer proposição, podemos defini-la como verdadeira ou falsa como, por exemplo: “o céu é azul”;
- **Negação:** negação do valor lógico de uma proposição. A negação de uma proposição verdadeira é falsa. A negação de uma proposição falsa é verdadeira. Os símbolos da negação são o til \neg ou \neg ;
- **Conjunção:** proposição composta com a utilização do conectivo “e” como, por exemplo: “o céu é azul e as nuvens são brancas”. Os símbolos usuais para uma conjunção são: \cap ou a letra “V” invertida; e,
- **Disjunção:** proposição composta com a utilização do conectivo “ou” como, por exemplo, “o céu é azul ou os pássaros são pretos”. Os símbolos usuais para uma disjunção são: \cup ou a letra V .

2.12 Leis de De Morgan

Augustus de Morgan foi um matemático e lógico indiano.



Figure 2.11: Augustus De Morgan (1806 - 1871)

Primeira Lei de De Morgan:

Negar duas proposições ligadas com “e” (\cap); ou seja, uma **conjunção**, é o mesmo que negar duas proposições e ligá-las com “ou”’ (\cup); ou seja, transformá-las em uma disjunção). Considerando as proposições “p” e “q” teremos:

- $\sim(p \cap q) = (p) \cup (q)$; ou,
- $(p \cap q)^c = (p^c) \cup (q^c)$.

Segunda Lei de De Morgan:

Negar duas proposições ligadas por “ou”’ (\cup); ou seja, uma **disjunção**, é o mesmo que negar as duas proposições e ligá-las com “e” (ou seja, transformá-las em uma conjunção). Considerando as proposições “p” e “q” teremos:

- $\sim(p \cup q) = (p) \cap (q)$; ou,
- $(p \cup q)^c = (p^c) \cap (q^c)$.

2.13 Noções básicas para o uso de calculadora (Cassio fx-82MS)

Em estatística trabalha-se muito com a análise de um ou mais conjuntos de dados, sendo comum a realização de diversas operações matemáticas com esses dados. Muitas dessas operações envolvem somatórios, por exemplo, e para simplificar essas operações o uso da calculadora se torna essencial.

Neste curso recomenda-se o uso de uma calculadora científica. Existem diversas calculadoras que cumprem as funções necessárias nesse curso. Para padronizar as aulas, alguns professores sugerem a calculadora científica de código: FX82MS, que é a calculadora que cujo funcionamento será exibido a seguir, passo a passo. A seguir serão descritas algumas das funções básicas mais importantes no uso desta calculadora.

Primeiro vamos deixar a calculadora no modo de regressão linear. Esse modo permite que a calculadora funcione normalmente para as operações comuns (soma, subtração, multiplicação e divisão), e ainda libera todas as funções importantes nesse curso. Sempre que o aluno for utilizar a calculadora, ele deve se certificar que ela esteja no modo de regressão linear, da seguinte forma:

PASSO 1:

- 1. ON
- 2. MODE
- 3. Aperte 3 para escolher REG
- 4. Aperte 1 para escolher LIN

Repare que no topo do visor da calculadora apareceu o símbolo **REG**, que indica que a calculadora está em modo de regressão. Desde que esteja no modo de regressão, podemos passar para o passo seguinte.

O nosso objetivo aqui é inserir o conjunto de dados na calculadora para então realizarmos as operações necessárias. Mas antes de inserir os dados, temos que garantir que a calculadora esteja **vazia** para o novo conjunto de dados. Ou seja, devemos limpar a calculadora:

PASSO 2:

- 1. SHIFT
- 2. MODE
- 3. Aperte 1 para escolher Scl (*Stat Clear*)
- 4. Aperte = para limpar a calculadora

Entrada de dados.

Agora que a calculadora está em modo de regressão e está limpa, podemos inserir o conjunto de dados. Para ilustrar esta função, vamos inserir o seguinte conjunto de dados: $X = 5, 3, 6, 2$.

Para inserir cada um desses elementos você deve digitar o número e em seguida o botão M+.

A sequência fica assim: 5 M+ 3 M+ 6 M+ 2 M+.

A cada vez que você insere uma observação, a calculadora atualiza o número de observações inseridas. No final, nesse caso, aparece **n=4** porque inserimos 4 observações.

Funções envolvendo somatórios.

Observe na calculadora os botões **shift** e **alpha**. Geralmente estes botões aparecem nas cores amarela e vermelha, respectivamente. Observe ainda que alguns botões da calculadora possuem termos nessas cores. Para selecionar as funções em **amarelo**, antes devemos ligar o modo **shift**. Enquanto que para selecionar as funções em **vermelho** deve-se ligar o modo **alpha**.

Por exemplo, para abrir a função **S-SUM** que está em **amarelo** no botão 1, faz-se: SHIFT 1. A função **S-SUM** é a que contém todos os somatórios importantes. Ao abrir esta função aparecem três opções da seguinte forma:

$$\Sigma(x) \Sigma(x^2) n$$

Aperta-se 1 = para ter o somatório de x ; 2 = para ter o somatório de x^2 ou 3 = para saber o número n de observações inseridas.

Funções para obter a média e o desvio padrão.

A função **S-VAR** fornece a média e o desvio padrão dos dados. Essas são medidas importantes, que serão utilizadas durante todo o curso. Para abrir esta função faz-se: SHIFT 2.

$$\bar{x} \sigma_x S_x$$

A opção 1 retorna a média dos dados, a opção 2 retorna o desvio padrão populacional e a opção 3 o desvio padrão amostral.

Como inserir dois conjuntos de dados.

Quando se deseja estudar dois conjuntos de dados, de mesmo tamanho, pode-se inseri-los de forma simultânea na calculadora. Para ilustrar vamos inserir os seguintes conjuntos de dados: $X = 2, 7, 4, 3, 2$ e $Y = 1, 2, 3, 6, 5$. **Antes de inserir os dados, lembre-se de limpar a calculadora.**

Em seguida vamos inserir os dados de 2 em 2: o primeiro de X com o primeiro de Y e assim por diante. Repare que ao lado do botão M+ tem um botão com uma vírgula. Esta vírgula é utilizada para separar as observações de X das de Y . A sequência fica assim:

- 2,1 M+
- 7,2 M+
- 4,3 M+
- 3,6 M+
- 2,5 M+

Se você usar a função **S-SUM**, na tela vai aparecer os somatórios apenas de X, que foi pela ordem, o primeiro a ser inserido. Na calculadora tem um botão grande e style="color:gray;">S-SUM, com 4 setas. Depois de selecionar a função **amarelo** aperte a seta para frente que aparecerão os somatórios para Y . O mesmo acontece para a função **S-VAR**.

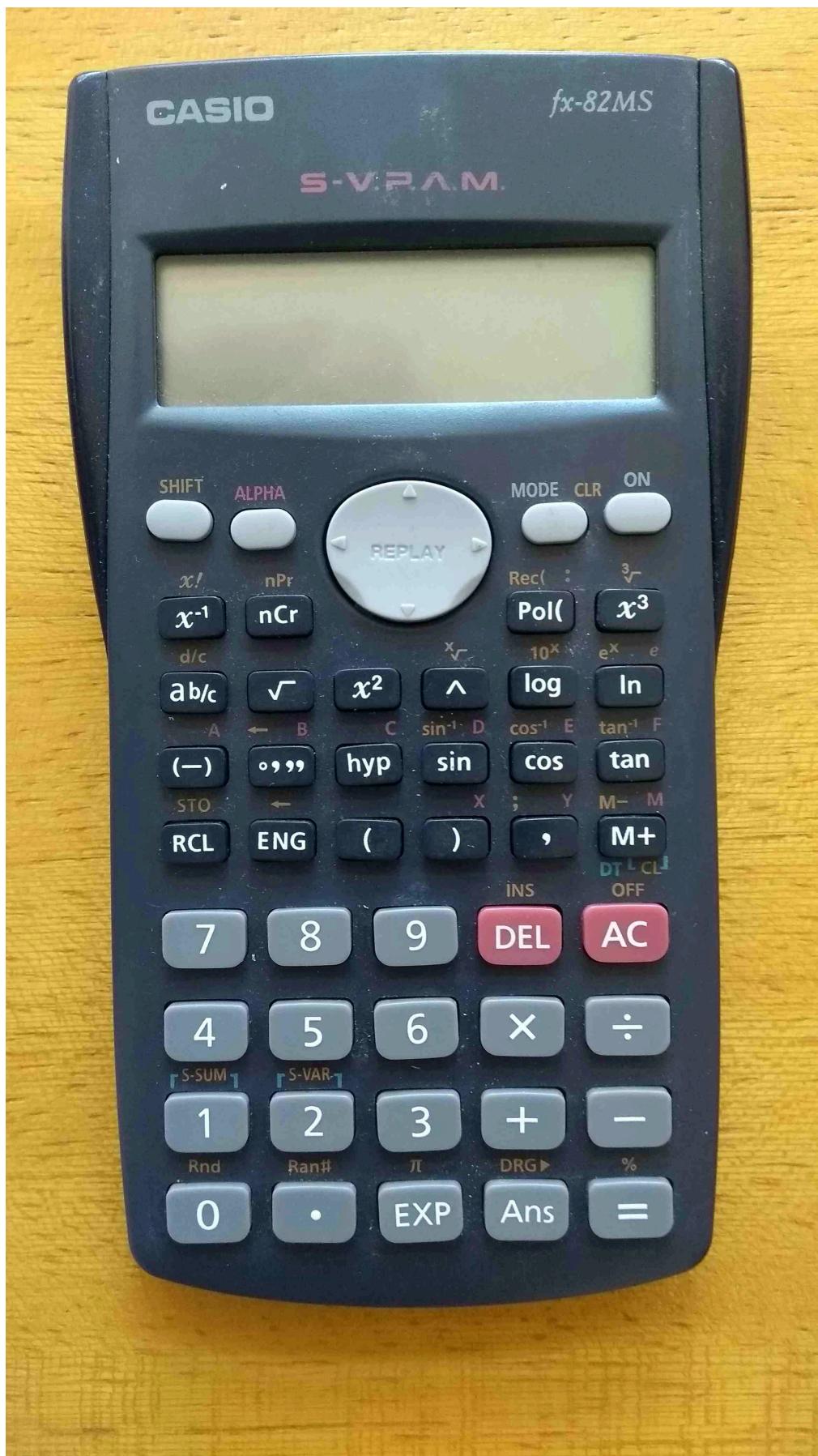


Figure 2.12: Calculadora Cassio

2.14 Instalação do software R em conjunto com a interface gráfica RStudio

“A pergunta não é se o R faz; mas sim, como ele faz [...] (anônimo)”

R é uma linguagem e ambiente para computação estatística e gráficos. É um projeto GNU que é semelhante à linguagem e ambiente S que foi desenvolvido nos Laboratórios Bell (anteriormente AT&T, agora *Lucent Technologies*) por John Chambers e colegas. R pode ser considerado como uma implementação diferente de S. Existem algumas diferenças importantes, mas muito código escrito para S roda inalterado sob R.

R fornece uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, *clustering*, ...) e gráficas, e é altamente extensível. A linguagem S costuma ser o veículo escolhido para pesquisa em metodologia estatística, e R fornece uma rota de código aberto para participação nessa atividade.

Um dos pontos fortes do R é a facilidade com que gráficos de qualidade de publicação bem projetados podem ser produzidos, incluindo símbolos matemáticos e fórmulas quando necessário. Grande cuidado foi tomado sobre os padrões para as escolhas de design menores em gráficos, mas o usuário mantém o controle total.

R está disponível como Software Livre sob os termos da Licença Pública Geral GNU da *Free Software Foundation* em forma de código-fonte. Ele compila e roda em uma ampla variedade de plataformas UNIX e sistemas similares (incluindo FreeBSD e Linux), Windows e MacOS.

R é um conjunto integrado de recursos de software para manipulação de dados, cálculo e exibição gráfica. Inclui:

- uma instalação eficaz de manipulação e armazenamento de dados,
- um conjunto de operadores para cálculos em arrays, em particular matrizes, uma coleção grande, coerente e integrada de ferramentas intermediárias para análise de dados,
- facilidades gráficas para análise de dados e exibição na tela ou em cópia impressa, e uma linguagem de programação bem desenvolvida, simples e eficaz que inclui condicionais, loops, funções recursivas definidas pelo usuário e recursos de entrada e saída.

O termo “ambiente” destina-se a caracterizá-lo como um sistema totalmente planejado e coerente, em vez de um acréscimo incremental de ferramentas muito específicas e inflexíveis, como é frequentemente o caso de outros softwares de análise de dados.

R, como S, é projetado em torno de uma verdadeira linguagem de computador e permite aos usuários adicionar funcionalidades adicionais definindo novas funções. Grande parte do sistema é escrito no dialeto R de S, o que torna mais fácil para os usuários seguirem as escolhas algorítmicas feitas. Para tarefas de computação intensiva, os códigos C, C++ e Fortran podem ser vinculados e chamados em tempo de execução. Usuários avançados podem escrever código C para manipular objetos R diretamente.

Muitos usuários pensam no R como um sistema estatístico. Preferimos pensar nisso como um ambiente no qual as técnicas estatísticas são implementadas. R pode ser estendido (facilmente) via packages . Existem cerca de oito pacotes fornecidos com a distribuição R e muitos mais estão disponíveis através da família CRAN de sites da Internet, cobrindo uma ampla gama de estatísticas modernas.

R tem seu próprio formato de documentação semelhante ao LaTeX, que é usado para fornecer documentação abrangente, tanto on-line em vários formatos quanto em cópia impressa.

A página principal pode ser acessa em: The R Project for Statistical Computing e as informações acima foram traduzidas de Fonte das informações.

2.14.1 RStudio

RStudio é um ambiente de desenvolvimento integrado (IDE) para R e Python. Ele inclui um console, editor de realce de sintaxe que oferece suporte à execução direta de código e ferramentas para plotagem, histórico, depuração e gerenciamento de espaço de trabalho. O RStudio está disponível em código aberto e edições comerciais e é executado na área de trabalho (Windows, Mac e Linux). A página principal pode ser acessada em: RStudio.

Há inúmeros tutoiais para a instalação do *R* e o *RStudio* (uma IDE: *Integrated development environment* para poder utilizar o software de um modo mais amigável), dentre os quais: Tutorial de instalação (UFPr).

2.14.2 Pacotes

Os pacotes na linguagem de programação R são um conjunto de funções R , código compilado e dados de amostra. Estes são armazenados em um diretório chamado “biblioteca” dentro do ambiente R. Por padrão, o R instala um grupo de pacotes durante a instalação. Assim que iniciarmos o console R, apenas os pacotes padrão estarão disponíveis por padrão. Outros pacotes que já estão instalados precisam ser carregados explicitamente para serem utilizados pelo programa R que os usará.

Uma lista de todos os pacotes disponibilizados para os mais variados problemas de análise estatística pode ser vista em Lista de pacotes.

Módulo 3

Introdução à estatística descritiva

Sobre o estudo da estatística por áreas nas quais, aparentemente, não se vislumbra sua utilidade trazemos o prefácio da tradução do livro de Jack Levin (Estatística aplicada às ciências humanas) por Sérgio Francisco Costa, ao dizer que o livro:

“[...] destina-se a um público muito específico: estudantes de Ciências Humanas, refúgio errôneo dos que fogem das equações e dos cálculos, pois que, embora humanas - e talvez por isso mesmo - não podemos prescindir das tão odiadas quantificações [...]”

3.1 Análise exploratória

A análise exploratória de dados (*EDA: Exploratory Data Analysis*, originalmente desenvolvida pelo matemático e estatístico norte-americano John Tukey na década de 1970) é usada para se investigar conjuntos de dados e resumir suas principais características, muitas vezes usando métodos de visualização de dados por gráficos e apresentação de tabelas.

Habitualmente uma *EDA* envolve:

- verificar quais são os tipos de variáveis presentes nos dados;
- sintetizar os valores assumidos por cada uma das variáveis;
- verificar os padrões de cada variável e eventuais associações entre duas ou mais delas; e,
- apresentação de tabelas e gráficos expositivos variados.



Figure 3.1: John Tukey (1915-2000)

3.2 Dados brutos, em rol, diagrama de ramos & folhas e de dispersão unidimensional

Consideremos os dados obtidos da medição das alturas em metros de 60 estudantes de uma determinada classe de um certo curso aqui na UEL:

```
alturas=c(1.63,1.67,1.47,1.64,1.66,1.73,2.00,1.62,1.65,1.56,1.65,1.85,1.73,
        1.78,1.82,1.68,1.67,1.83,1.72,1.71,1.73,1.67,1.66,1.95,1.76,1.73,
        1.77,1.68,1.65,1.64,1.66,1.68,1.61,1.73,1.72,1.83,1.69,1.84,1.66,
        1.78,1.54,1.74,1.56,1.66,1.56,1.62,1.55,1.86,1.44,1.67,1.76,1.79,
        1.75,1.41,1.65,1.58,1.93,1.57,1.71,1.58,0.1,3.68,0,NA)
alturas

## [1] 1.63 1.67 1.47 1.64 1.66 1.73 2.00 1.62 1.65 1.56 1.65 1.85 1.73 1.78 1.82
## [16] 1.68 1.67 1.83 1.72 1.71 1.73 1.67 1.66 1.95 1.76 1.73 1.77 1.68 1.65 1.64
## [31] 1.66 1.68 1.61 1.73 1.72 1.83 1.69 1.84 1.66 1.78 1.54 1.74 1.56 1.66 1.56
## [46] 1.62 1.55 1.86 1.44 1.67 1.76 1.79 1.75 1.41 1.65 1.58 1.93 1.57 1.71 1.58
## [61] 0.10 3.68 0.00    NA
```

Garbage in, garbage out. Não são raras as vezes nas quais o relatório com os dados coletados em uma pesquisa apresentam uma série de erros. Não estamos a nos referir aqui aos **erros amostrais** mas sim aos erros experimentais (não amostrais), aqueles decorrentes de dados coletados incorretamente, tais como aqueles resultantes de omissões na transcrição das informações, da leitura de instrumentos descalibrados ou de informações simplesmente não coletadas.

Denomina-se pré-processamento essa etapa de *limpeza* do conjunto de dados na qual busca-se corrigir de modo extremamente criterioso esses problemas e, para tanto, um profundo conhecimento do objeto que está sendo pesquisado é necessário de modo a não serem liminarmente eliminados dados simplesmente por destoarem da alguma tendência (para essas situações há ferramentas estatísticas apropriadas).

O conjunto original de dados (*dataset*) refere-se a alturas de pessoas (estudantes) e assim, trata-se de uma variável quantitativa e contínua e como tal será analisada. As omissões de informação “NA” (*not available*) e as medidas transcritas com erros grosseiros (0 m; 0,10 m; 3,68 m) serão removidas.

Assim, o *dataset* será composto pelos dados abaixo:

```
alturas=c(1.63,1.67,1.47,1.64,1.66,1.73,2.00,1.62,1.65,1.56,1.65,1.85,1.73,
        1.78,1.82,1.68,1.67,1.83,1.72,1.71,1.73,1.67,1.66,1.95,1.76,1.73,
        1.77,1.68,1.65,1.64,1.66,1.68,1.61,1.73,1.72,1.83,1.69,1.84,1.66,
        1.78,1.54,1.74,1.56,1.66,1.56,1.62,1.55,1.86,1.44,1.67,1.76,1.79,
        1.75,1.41,1.65,1.58,1.93,1.57,1.71,1.58)
alturas
```

```
## [1] 1.63 1.67 1.47 1.64 1.66 1.73 2.00 1.62 1.65 1.56 1.65 1.85 1.73 1.78 1.82
## [16] 1.68 1.67 1.83 1.72 1.71 1.73 1.67 1.66 1.95 1.76 1.73 1.77 1.68 1.65 1.64
## [31] 1.66 1.68 1.61 1.73 1.72 1.83 1.69 1.84 1.66 1.78 1.54 1.74 1.56 1.66 1.56
## [46] 1.62 1.55 1.86 1.44 1.67 1.76 1.79 1.75 1.41 1.65 1.58 1.93 1.57 1.71 1.58
```

Esse conjunto de dados certamente contém diversas informações acerca da altura dessas pessoas; todavia, da maneira como estão expostos, a visualização dessas informações fica bastante difícil. Esse modo de apresentação é chamado de dados *brutos*.

Com um pequeno refinamento, como pela simples ordenação desses dados (são medidas numéricas contínuas), algumas informações começam a se destacar:

```
sort(alturas)
```

```
## [1] 1.41 1.44 1.47 1.54 1.55 1.56 1.56 1.56 1.57 1.58 1.58 1.61 1.62 1.62 1.63
## [16] 1.64 1.64 1.65 1.65 1.65 1.65 1.66 1.66 1.66 1.66 1.66 1.67 1.67 1.67 1.67
## [31] 1.68 1.68 1.68 1.69 1.71 1.71 1.72 1.72 1.73 1.73 1.73 1.73 1.73 1.74 1.75
## [46] 1.76 1.76 1.77 1.78 1.78 1.79 1.82 1.83 1.83 1.84 1.85 1.86 1.93 1.95 2.00
```

A interpretabilidade das informações trazidas por esses dados começa a ficar mais fácil como, por exemplo, as alturas:

- mínima; e,
- máxima dos estudantes.

A uma listagem de valores ordenada (de modo crescente ou decrescente) dá-se o nome de *rol*.

Outra forma de apresentação desses dados é por um *Diagrama de Ramos e Folhas*, uma apresentação híbrida pois ao mesmo tempo que espelha a quantidade de medidas observadas para cada altura, mantém as informações da listagem.

```
stem(alturas)
```

```
## The decimal point is 1 digit(s) to the left of the |
## 14 | 147
## 15 | 45666788
## 16 | 1223445555666677778889
## 17 | 1122333345667889
## 18 | 233456
## 19 | 35
## 20 | 0
```

À esquerda do traço vertical (os ramos) são apresentadas frações das medidas das alturas (no caso, decímetros) e à direita (as folhas) são apresentadas os complementos dessas medidas (os centímetros) de tal modo que cada um dos dados da amostral original possa ter sua medida resgatada fazendo-se a leitura dos valores à esquerda com cada um deles à direita.

Essa apresentação também oferece uma apreciação visual a respeito de como os valores se distribuem.

Um *Gráfico de dispersão unidimensional (stripchart)* expressa visualmente duas informações: a localização de cada uma das medidas e a dispersão dos dados.

```
stripchart(alturas, method = "stack", offset=1,
          pch=20, at=0.5,
          main="Gráfico de dispersão unidimensional",
          col="blue", cex=1,
          xlab="Alturas dos estudantes (m)",
          ylab="Quantidades observadas (un)")
```

3.3 Sínteses numéricas descritivas

Além da apresentação elementar de algumas informações relacionadas aos dados brutos da amostra, tais como os valores *mínimo* e *máximo* observados, a estatística descritiva possui muitas outras ferramentas para *condensar* a informação contida nos dados.

São chamadas de *sínteses numéricas*, medidas que condensam variados aspectos relacionados aos valores dos dados. As principais *sínteses numéricas* são:

Gráfico de dispersão unidimensional

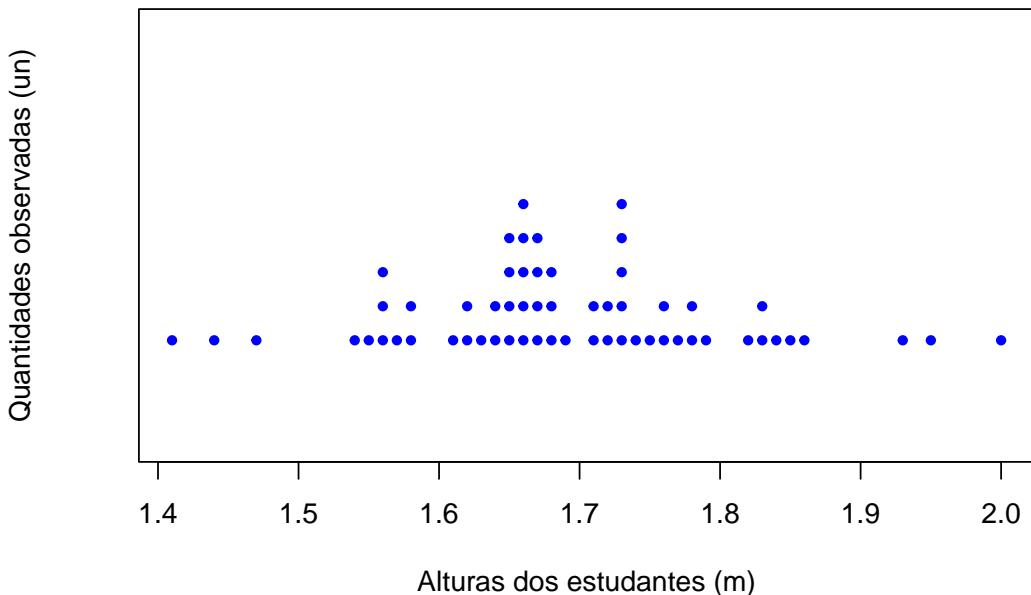


Figure 3.2: Gráfico de dispersão unidimensional (stripchart)

- de tendência central (posição): média (simples ou aritmética, geométrica, harmônica, anarmônica, quadrática, biquadrática), moda e mediana;
- de dispersão (variabilidade): absolutas (amplitude total, variância e desvio padrão) ou relativas (coeficiente de variação, unidades padronizadas); e,
- de subdivisão (separatrizes, quantis): mediana (50%), quartis (25%, 50%, 75%), decis (10%, ..., 90%) e percentis (1%....99%).

Uma medida de posição ou dispersão é dita **resistente** quando forem pouco afetadas pela alteração de uma pequena porção dos dados. A mediana é uma medida resistente, já a média e a variância não são.

3.3.1 Medidas de tendência central (posição)

3.3.1.1 Média

Sejam x_1, x_2, \dots, x_n os n valores assumidos pela variável X (dados brutos). A *média aritmética simples* será dada por:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Algumas propriedades da média aritmética:

- somando-se (ou subtraindo-se) cada um dos elementos do conjunto de dados por uma constante arbitrária qualquer k , a média aritmética ficará adicionada (ou subtraída) dessa essa constante k

```
alturas_ad=alturas+0.05

par(mfrow=c(1,2))

stripchart(alturas,method = "stack", at=0.5,
main="",pch = 20,
col="blue", cex=1, xlab="Alturas originais dos estudantes (m)",
ylab="Quantidades observadas (un)")
abline(v=mean(alturas), col="red")
text(mean(alturas)-0.2, 1, "Média=1,69 m", col = "red", srt=90)

stripchart(alturas_ad,method = "stack", at=0.5,
main="",pch = 20,
col="blue", cex=1, xlab="Alt. dos estudantes (m) adic. de 5cm",
ylab="Quantidades observadas (un)")
abline(v=mean(alturas_ad), col="red")
text(mean(alturas_ad)-0.2, 1, "Média=1,74 m", col = "red", srt=90)
```

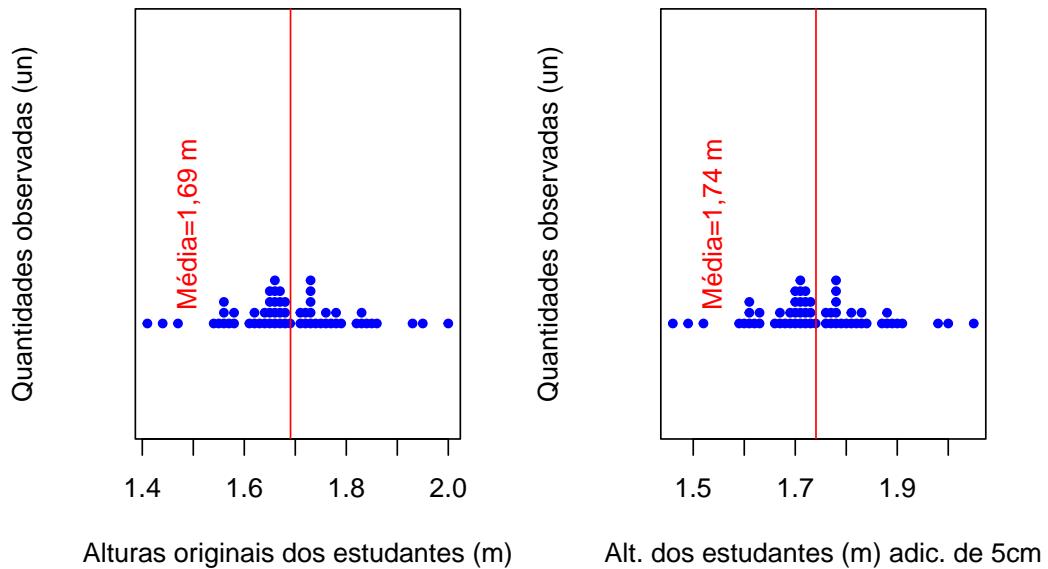


Figure 3.3: Mudanças na média pela adição (subtração) de uma constante $k = 0.05$

- multiplicando-se (ou dividindo-se) cada um dos elementos do conjunto de dados por uma constante arbitrária k , a média aritmética ficará multiplicada (ou dividida) por essa constante k

```

alturas_mult=alturas*1.2

par(mfrow=c(1,2))

stripchart(alturas,method = "stack", at=0.5,
main="",pch = 20,
col="blue", xlab="Alturas originais dos estudantes (m)",
ylab="Quantidades observadas (un)")
abline(v=mean(alturas), col="red")
text(mean(alturas)-0.1, 1, "Média=1,69 m", col = "red", srt=90)

stripchart(alturas_mult,method = "stack", at=0.5,
main="",pch = 20,
col="blue", xlab="Alt. dos estudantes (m) mult. por 1,2",
ylab="Quantidades observadas (un)")
abline(v=mean(alturas_mult), col="red")
text(mean(alturas_mult)-0.1, 1, "Média= 2,02 m", col = "red", srt=90)

```

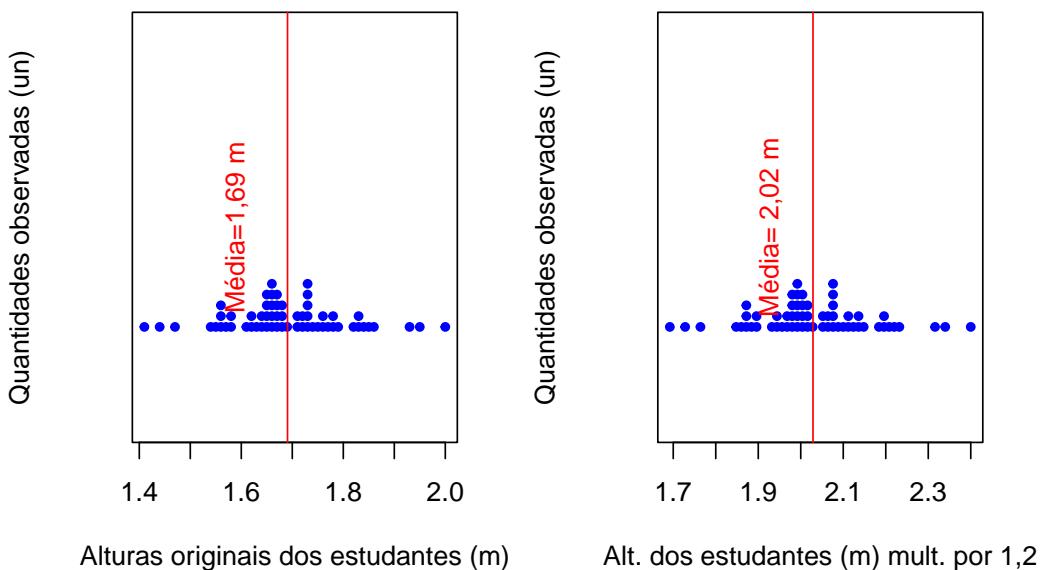


Figure 3.4: Mudanças na média pela multiplicação (divisão) de uma constante $k = 1.2$

- a soma dos desvios observados entre cada um dos valores assumidos pela variável X e sua média \bar{x} é nula;
- a soma dos quadrados dos desvios é mínima;
- em uma distribuição de frequências, a soma dos produtos dos desvios entre a média o valor médio de cada uma das classes, pelas respectivas frequências é nula; e,
- multiplicando-se (ou dividindo-se) todas as frequências de uma distribuição por uma constante arbitrária, a média aritmética não se altera.

Usando os dados das medidas das alturas dos 60 estudantes teremos o seguinte valor para a **média**:

```
round(mean(alturas),2)
```

```
## [1] 1.69
```

3.3.1.2 Moda

Moda é o valor que ocorre com maior frequência na amostra. Uma amostra pode se apresentar como:

- unimodal;
- bimodal;
- plurimodal; ou,
- amodal.

```
tab_alturas=table(alturas)
```

```
tab_alturas
```

```
## alturas
## 1.41 1.44 1.47 1.54 1.55 1.56 1.57 1.58 1.61 1.62 1.63 1.64 1.65 1.66 1.67 1.68
##   1     1     1     1     1     3     1     2     1     2     1     2     4     5     4     3
## 1.69 1.71 1.72 1.73 1.74 1.75 1.76 1.77 1.78 1.79 1.82 1.83 1.84 1.85 1.86 1.93
##   1     2     2     5     1     1     2     1     2     1     1     2     1     1     1     1
## 1.95 2
##   1     1
```

```
barplot(tab_alturas,
        main="Valores observados da alturas dos estudantes",
        xlab="Altura (cm)",
        ylab="Quantidade observada (un)",
        ylim=c(0,6),
        col="blue",
        las=0,
        hor="FALSE")
```

Usando os dados das medidas das alturas dos 60 estudantes teremos os seguintes valores para a **moda**:

Valores observados da alturas dos estudantes

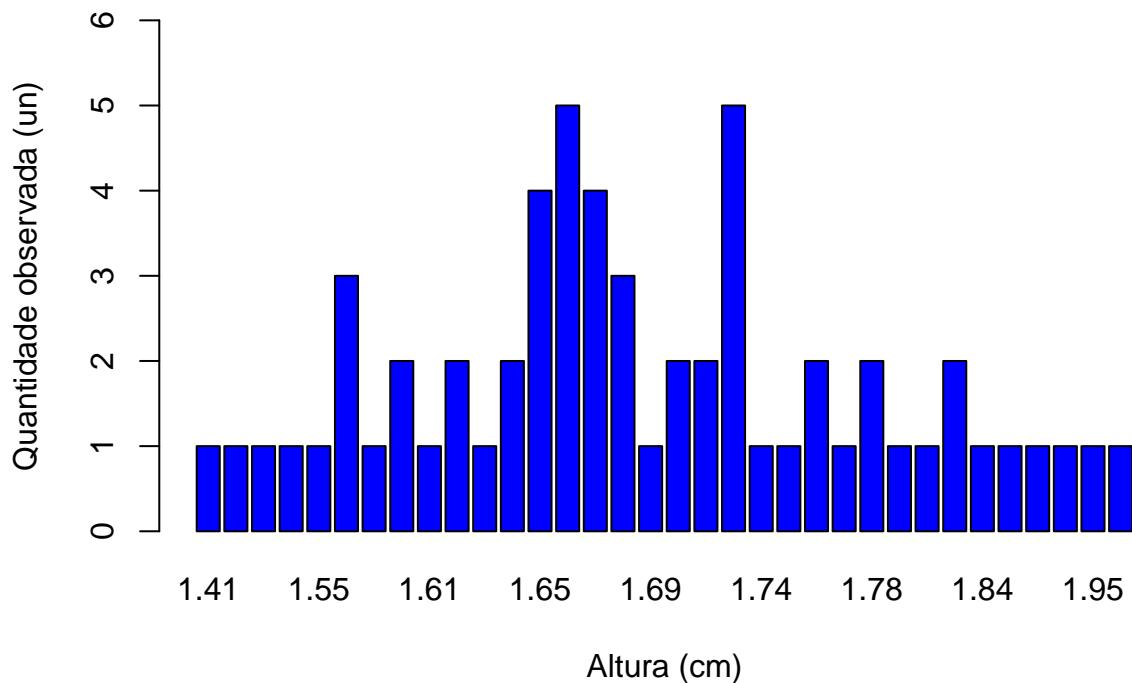


Figure 3.5: Bimodal: 1,66 m e 1,73 m

```
# função em R para extrair a moda:
```

```
Modes <- function(x) {
  ux <- unique(x)
  tab <- tabulate(match(x, ux))
  ux[tab == max(tab)]
}

Modes(alturas)
```

```
## [1] 1.66 1.73
```

3.3.1.3 Mediana

Mediana é uma medida quantitativa tal que divide a amostra ordenada dos dados em duas partes com *igual quantidade de dados* tais que na primeira delas as observações possuem valores menores que sua medida e na outra parte as observações possuem valores superiores a ela.

Por essa razão, a mediana é uma separatriz (de subdivisão) de 50%, equivalente ao 2º quartil, ao 5º decil e ao 50º percentil. Para sua estimação necessitamos saber qual a **posição** que ela ocupa no rol de dados e assim,

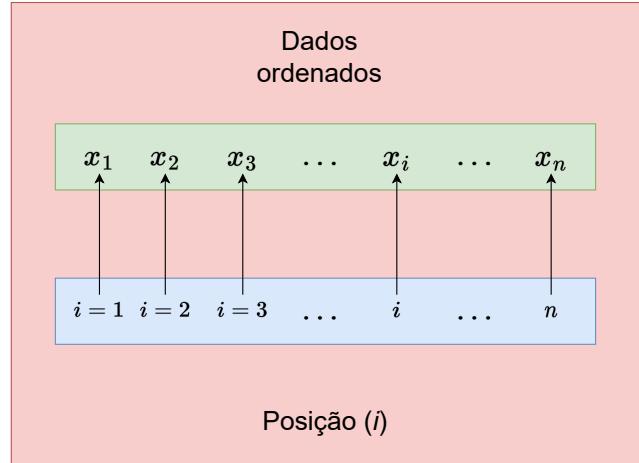


Figure 3.6: Entendendo a indexação de dados

duas situações podem ocorrer:

1- se a amostra possui um número **ímpar** (n) de elementos: a medida da mediana igual ao valor do $i - simo$ elemento da **amostra ordenada** (a medida da mediana será um valor, de fato, observado) tal que:

$$Md = x_i$$

com:

- $i = \frac{n+1}{2}$ (n é o número de observações);

2- se a amostra possui um número **par** (n) de elementos: a medida da mediana será a **média aritmética** dos valores dos elementos nas posições imediatamente anterior (i_{ant}) e posterior (i_{post}) à sua posição central virtual (a medida de mediana não será, portanto, um valor observado):

$$Md = mdia(x_{i_{ant}}; x_{i_{post}})$$

com:

- $i_{ant} = \frac{n}{2}$ e $i_{post} = \frac{n}{2} + 1$ (n é o número de observações).

Sendo uma **separatriz**, sua posição L pode ser também calculada pela expressão mais geral (para qualquer percentil) que logo mais será apresentada.

Usando os dados das medidas das alturas dos 60 estudantes teremos o seguinte valor para a **mediana**:

```
sort(alturas)
```

```
## [1] 1.41 1.44 1.47 1.54 1.55 1.56 1.56 1.56 1.57 1.58 1.58 1.61 1.62 1.62 1.63
## [16] 1.64 1.64 1.65 1.65 1.65 1.66 1.66 1.66 1.66 1.66 1.67 1.67 1.67 1.67
## [31] 1.68 1.68 1.68 1.69 1.71 1.71 1.72 1.72 1.73 1.73 1.73 1.73 1.74 1.75
## [46] 1.76 1.76 1.77 1.78 1.78 1.79 1.82 1.83 1.83 1.84 1.85 1.86 1.93 1.95 2.00
```

```
median(alturas)
```

```
## [1] 1.675
```

3.3.1.4 Diferentes posições da média, moda e mediana

Essas três medidas podem se apresentar com valores em posições alternadas quando as comparamos:

- quando a moda=mediana=média temos uma distribuição de frequências razoavelmente **simétrica**;
- quando a moda \leq mediana \leq média (há uma quantidade maior de dados com grandes valores, arrastando a média para a direita, para cima) temos uma distribuição de frequências **positivamente assimétrica**, ; e,
- quando a moda \geq mediana \geq média (há uma quantidade maior de dados com pequenos valores, arrastando a média para a esquerda, para baixo) temos uma distribuição de frequências **negativamente assimétrica**.

```
barplot(tab_alturas,
        main="Valores observados da alturas dos estudantes",
        xlab="Altura (cm)",
        ylab="Quantidade observada (un)",
        ylim=c(0,6),
        col="blue",
```

```

    las=0,
    hor="FALSE")
abline(v=mean(19.9, 21.1), col="red")
text( mean(19.9, 21.1)-0.5, 5, "Média=1,69 m", col = "red", srt=90)
abline(v=median(18.7 , 19.9), col="darkgreen")
text(median(18.7 , 19.9)-0.5, 5, "Mediana=1,675 m", col = "darkgreen", srt=90)
abline(v=c(16.3, 23.5), col="darkgrey")
text(c(16.3-0.5, 23.5-0.5), 5, c("Moda=1,66","Moda=1,73"), col = "darkgray", srt=90)

```

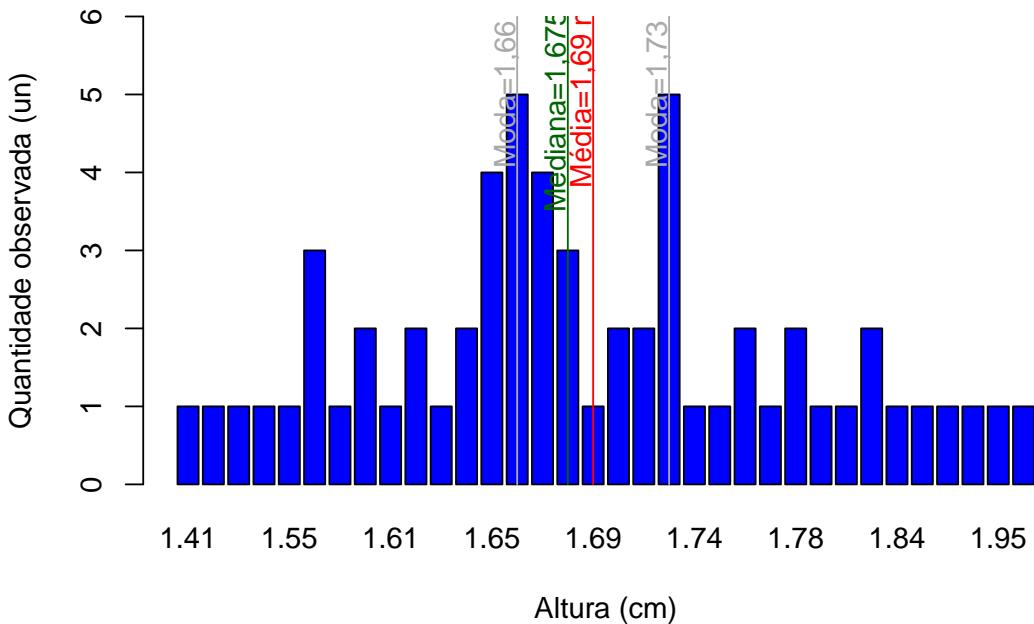
Valores observados da alturas dos estudantes

Figure 3.7: Valores observados das alturas dos estudantes e as posições da média, moda e mediana

3.3.2 Medidas de dispersão (variabilidade)

O conhecimento de uma medida de tendência central nos provê uma informação útil mas incompleta. As medidas de dispersão nos ajudam a ter uma perspectiva melhor dos dados.

- amplitude total dos dados;
- desvio padrão (variância): é considerada a mais útil das medidas de dispersão;
- coeficiente de variação; e,
- unidades padronizadas.

Comparação entre medidas de posição

	Média	Mediana	Moda
Definição	$\bar{x} = \frac{\sum x}{n}$	Valor do meio	Valor mais freqüente
Existência	Sempre existe	Sempre existe	Pode não existir, pode haver mais de uma
Leva em conta todos os valores	Sim	Não	Não
Afetada por valores discrepantes	Sim	Não	Não
Vantagens	Usada em muitos métodos estatísticos	Menos sensível a valores discrepantes	Apropriada para dados qualitativos

Figure 3.8: Quadro comparativo entre as medidas de tendência central (posição)

Diferentes tipos quanto à dimensão (unidade):

- **medidas absolutas** são aquelas expressas na mesma unidade de medida da variável do fenômeno estudado ($m; kg; \frac{R\$}{ms}; ...$);
- **medidas relativas** são adimensionais e assim podem ser usadas para se comparar a variabilidade de dois ou mais conjuntos de dados, mesmo quando as variáveis se refiram a diferentes fenômenos ou que sejam expressas, originalmente, em diferentes unidades.

3.3.2.1 Amplitude total dos dados

A amplitude total dos dados é a simples diferença entre o **maior** e o **menor** dos valores observados:

$$A = x_{max} - x_{min}$$

3.3.2.2 Estimação da variância (e desvio padrão)

Sejam x_1, x_2, \dots, x_n os n valores assumidos pela variável X . Dá-se o nome de desvios a contar da média as diferenças entre cada uma das observações e a média: $x_i - \bar{x}$ com $i = 1, 2, \dots, n$.

Não é possível considerar a possibilidade de se adotar o valor médio desses desvios pois uma das propriedades da média é que a soma dos desvios em torno de si é nula.

$$\bar{d} = \frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$$

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

constitui-se numa restrição linear dos desvios porque qualquer $n - 1$ deles completamente determina o outro. Tampouco se considera a possibilidade de se adotar o valor médio desses desvios em módulo, pelas dificuldades teóricas em problemas de inferência.

$$\bar{d} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Uma alternativa é adotar o valor médio do **quadrado** desses desvios.

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

ou,

$$S^2 = \frac{1}{(n - 1)} \times \left[\sum_{i=1}^n (x_i^2) - \frac{(\sum_{i=1}^n x_i)^2}{n} \right]$$

Diz-se que a variância amostral (variância *ajustada*) possui $(n - 1)$ graus de liberdade, denotado pela letra grega ν . A perda de *um* grau de liberdade deve-se à necessidade de se substituir a média populacional desconhecida (μ) por sua estimativa amostral (\bar{x}), deduzida a partir dos dados coletados.

Pode-se demonstrar que em razão dessa restrição a melhor estimativa para a variância populacional é obtida dividindo-se a soma dos quadrados dos desvios por $(n - 1)$. Assim S^2 será um estimador não tendencioso para a variância amostral ao ser dividido por $(n - 1)$.

```
IC.Na = function (N, n, mu, sigma) {
  dados=data.frame()
  plot(0, 0,
    type="n",
    xlim=c(sigma-0.1*sigma,sigma+0.1*sigma),
    ylim=c(0,N),
    bty="l",
    xlab="Desvio padrão",
    ylab="Amostras extraídas",
    main=paste0("Flutuação dos valores dos desvios padrão \nobtidos em ", N," amostras de
      tamanho ",n),
```

```

sub=paste0("A população de origem tem uma distribuição ~ N (\u03bc:",mu," ;
           \u03c3:", sigma,")")
abline(v=sigma, col='darkgreen', lwd=2, lty=2)
for (i in 1:N) {
  x = rnorm(n, mu, sigma)
  media = mean(x)
  sd = sqrt(sum((x-mean(x))^2)/(n-1))
  sd_vies = sqrt(sum((x-mean(x))^2)/(n))
  temp=cbind(mu, media, sd, sd_vies)
  dados=rbind(dados, temp)
  plotx = c(sd)
  ploty = c(i,i)
  if ( sd < sigma) points(sd, i, col="blue",cex=1)+text(y=i+3,x=sd, labels=round(sd,3),
    ↪ cex=1, col='blue')
  else
    points(sd, i, col="blue", cex=1)+text(y=i+3,x=sd, labels=round(sd,3), cex=1, col='blue')
  ↪
  plotx = c(sd_vies)
  ploty = c(i,i)
  if ( sd_vies < sigma) points(sd_vies, i, col="red",cex=1)+text(y=i+3,x=sd_vies,
    ↪ labels=round(sd_vies,3), cex=1, col='red')
  else
    points(sd_vies, i, col="red", cex=1)+text(y=i+3,x=sd_vies, labels=round(sd_vies,3),
      ↪ cex=1, col='red')
}
abline(v=mean(dados$sd), col='blue', lwd=2, lty=2)
abline(v=mean(dados$sd_vies), col='red', lwd=2, lty=2)
}

```

IC.Na(N=100, n=15, mu=170, sigma=7)

Uma medida de dispersão que apresenta a mesma unidade que a das observações originais é o **desvio-padrão**, definido como a raiz quadrada positiva da variância.

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Tanto a variância quanto o desvio padrão indicam, em média, qual será o erro (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (média).

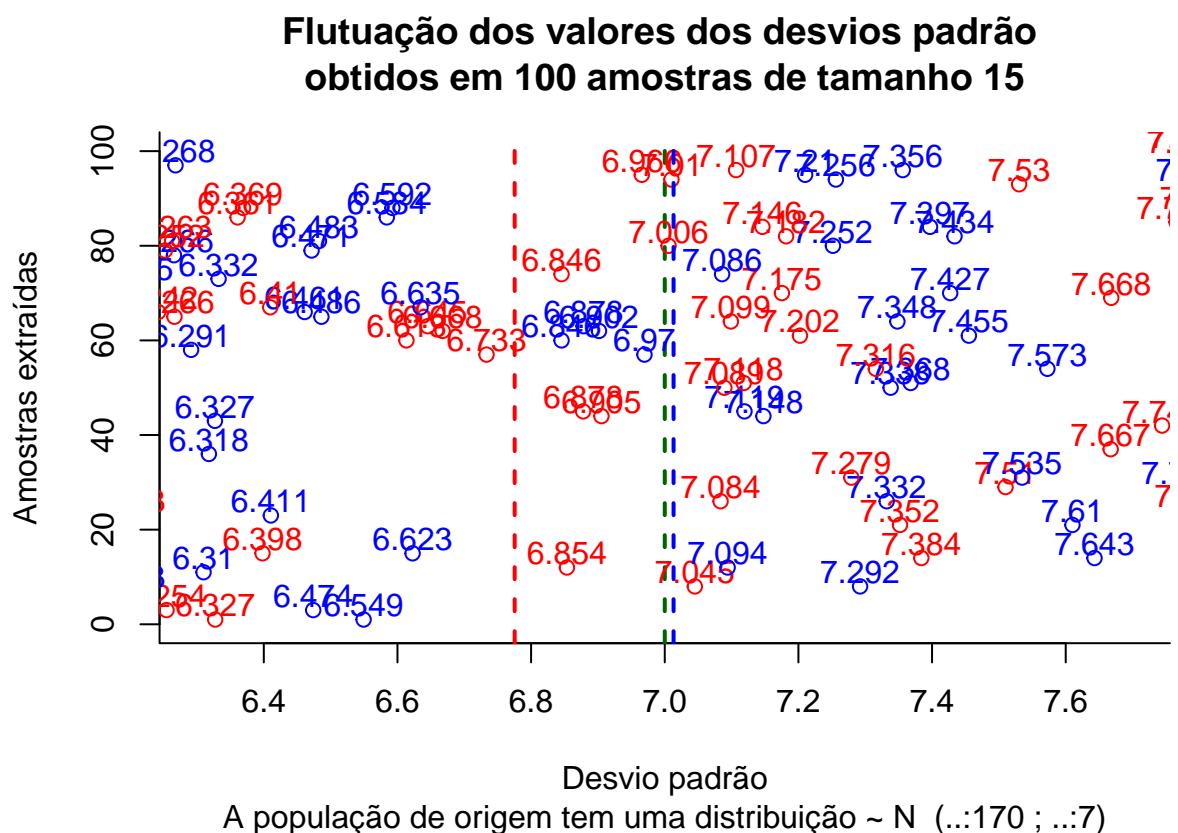


Figure 3.9: Flutuação dos valores do desvio padrão obtidos pelo estimador não viésado (em azul) e pelo estimador viésado (em vermelho) para diversas amostras extraídas de uma mesma população distribuição $\sim N(\mu; \sigma)$ (em verde o desvio padrão populacional, em azul a média dos desvios padrão amostrais correta e em vermelho a estimada de modo viésado)

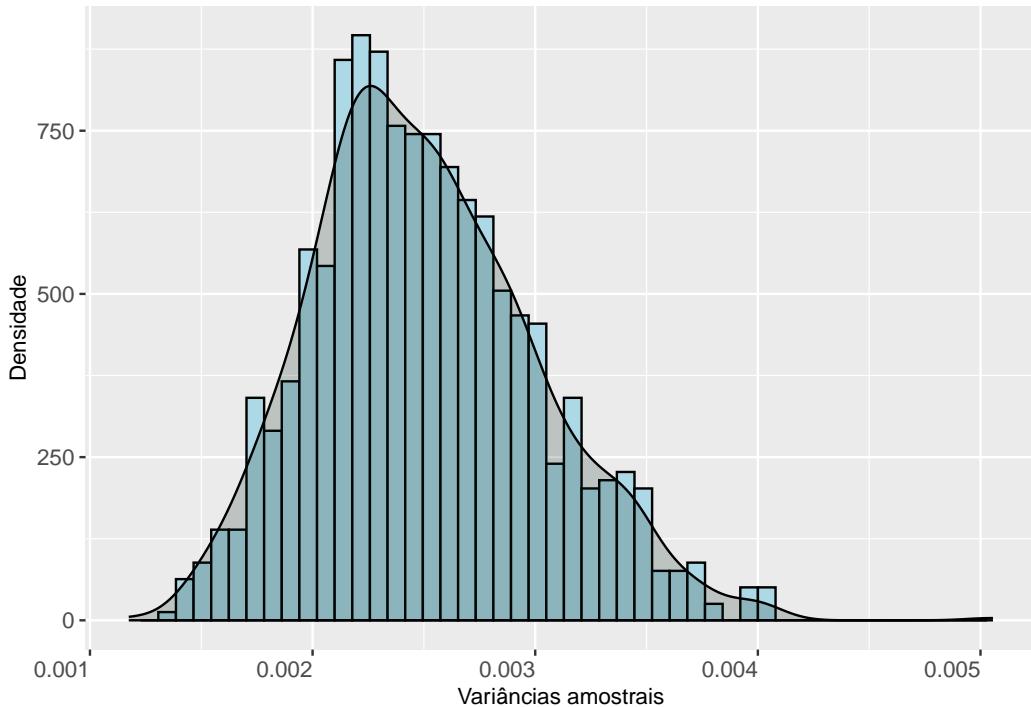


Figure 3.10: A distribuição das variâncias amostrais segue uma curva aproximada pela distribuição Qui-quadrado com $(n-1)$ graus de liberdade

Usando os dados das medidas das alturas dos 60 estudantes teremos o seguinte valor para a **variância** (com unidade igual a m^2) e o **desvio padrão** (com unidade igual a m):

```
# Variância
var(alturas)
```

```
## [1] 0.0130809
```

```
# Desvio padrão
sd(alturas)
```

```
## [1] 0.1143718
```

Propriedades da variância:

- somando-se (ou subtraindo-se) cada um dos elementos do conjunto de dados por uma constante arbitrária, a variância (e o desvio padrão) não se altera; e,
- multiplicando-se (ou dividindo-se) cada um dos elementos do conjunto de dados por uma constante arbitrária, a variância ficará multiplicada (ou dividida) pelo quadrado dessa constante. O desvio padrão fica multiplicado (ou dividido) por essa constante

```

# Adicionando-se uma constante k=0.05
alturas_ad=alturas+0.05

# Variância não se altera
var_ad= var(alturas_ad)
var_ad

## [1] 0.0130809

# Multiplicando-se uma constante k=1.2
alturas_mult=alturas*1.2

# Variância fica multiplicada (dividida) pelo quadrado dessa constante)
var(alturas_mult)

## [1] 0.0188365

all.equal(var(alturas_mult), var(alturas)*(1.2^2))

## [1] TRUE

```

3.3.2.3 Coeficiente de variação.

O coeficiente de variação (uma medida adimensional) é dado pela razão do desvio padrão pela média:

$$CV(\%) = 100 \cdot \left(\frac{s}{\bar{x}} \right)$$

Table 3.1: Classificação da variabilidade a partir da medida do Coeficiente de variação

Classificação	Medida do Coeficiente de variação (CV %)
Baixo	$CV \leq 10\%$
Médio	$10\% \leq CV \leq 20\%$
Alto	$20\% \leq CV \leq 30\%$
Muito alto	$CV \geq 30\%$

3.3.2.4 Padronização (*z-scores*)

À conversão do valor assumido por uma variável em unidades de desvio padrão acima (ou abaixo) do valor médio de sua distribuição é dado o nome de *padronização*. Essa métrica permite comparações com outras, procedentes de outros fenômenos.

Para padronizar (achar o seu *z-score* Z) o valor de uma variável procede-se segundo a fórmula:

$$Z = \frac{x_i - \bar{x}}{s}$$

O valor Z expressa quantos desvios esse dado está acima (ou abaixo) da média da distribuição.

Pelo *Teorema de Tchebichev* pode-se estimar a probabilidade mínima dos dados situados a certa distância de k desvios da média dessa distribuição:

$$P(|X - \mu| \geq k\sigma) \leq 1 - \frac{1}{k^2}$$

Assim, se $k = 2$ **ao menos** 75% das observações devem estar entre a média e dois desvios padrões acima ou abaixo da média.

```
med=round(mean(alturas),2)
desv= round(sd(alturas),2)
```

No exemplo das alturas dos estudantes temos a média de 1.69 m e um desvio padrão de 0.11 m. Assim, **ao menos** 75% das alturas deverão estar entre 1.47 m e 1.91 m.

```
sort(alturas)
```

```
## [1] 1.41 1.44 1.47 1.54 1.55 1.56 1.56 1.56 1.57 1.58 1.58 1.61 1.62 1.62 1.63
## [16] 1.64 1.64 1.65 1.65 1.65 1.66 1.66 1.66 1.66 1.66 1.67 1.67 1.67 1.67
## [31] 1.68 1.68 1.68 1.69 1.71 1.71 1.72 1.72 1.73 1.73 1.73 1.73 1.74 1.75
## [46] 1.76 1.76 1.77 1.78 1.78 1.79 1.82 1.83 1.83 1.84 1.85 1.86 1.93 1.95 2.00
```

```
# Duas observações menores que 1,47m e três maiores que 1,91m.  
# Assim, 54 observações dentro do intervalo, equivalendo a 91,66% do total.
```

3.3.3 Medidas de subdivisão (separatrizes)

Separatrizes (quantis) são valores que delimitam uma proporção de observações existentes de um conjunto de dados previamente ordenados menores que ele.

De modo geral, um *quantil* de ordem p (ou também p – *quantil*, indicado por q_p) é uma medida onde p é uma proporção qualquer (limitada no intervalo $0 < p < 1$), tal que $100p\%$ das observações sejam menores que seu valor q_p .

Desse modo, o valor q_p de uma variável aleatória X remete à medida da probabilidade:

$$P(X = x | x \leq q_p) = p$$

Os quantis mais informativos (e que por essa razão são usados para um importante gráfico que mais adiante será exposto em detalhes - *Boxplot*) são: \

- 1º Quartil ($q_{0,25}$): 25% dos dados possuem valores abaixo desse valor e 75% estão acima;
- 2º Quartil ou mediana ($q_{0,50}$): 50% dos dados possuem valores abaixo desse valor e 50% estão acima; e,
- 3º Quartil ($q_{0,75}$): 75% dos dados possuem valores abaixo desse valor e 25% estão acima.

Há muitos modos de se estabelecer os quantis descritos na literatura. O próprio R apresenta 9 modos diferentes:

```
quantile(alturas, type=1)
```

```
##   0%  25%  50%  75% 100%
## 1.41 1.63 1.67 1.75 2.00
```

```
quantile(alturas, type=2)
```

```
##    0%   25%   50%   75% 100%
## 1.410 1.635 1.675 1.755 2.000
```

```
quantile(alturas, type=3)
```

```
##    0%   25%   50%   75% 100%
## 1.41 1.63 1.67 1.75 2.00
```

```
quantile(alturas, type=4)
```

```
##    0%   25%   50%   75% 100%
## 1.41 1.63 1.67 1.75 2.00
```

```
quantile(alturas, type=5)
```

```
##    0%   25%   50%   75% 100%
## 1.410 1.635 1.675 1.755 2.000
```

```
quantile(alturas, type=6)
```

```
##    0%   25%   50%   75% 100%
## 1.4100 1.6325 1.6750 1.7575 2.0000
```

```
quantile(alturas, type=7)
```

```
##    0%   25%   50%   75% 100%
## 1.4100 1.6375 1.6750 1.7525 2.0000
```

```
quantile(alturas, type=8)
```

```
##    0%   25%   50%   75% 100%
## 1.410000 1.634167 1.675000 1.755833 2.000000
```

```
quantile(alturas, type=9)
```

```
##    0%   25%   50%   75% 100%
## 1.410000 1.634375 1.675000 1.755625 2.000000
```

Para grandes conjuntos de dados a diferença entre os quantis determinados sob esses diferentes modos será desprezível.

De modo geral, para se calcular a posição L de um quantil qualquer de ordem p em um rol de dados há algumas regras empíricas:

$$L_p = \frac{p}{100} \times (n) L_p = \frac{p}{100} \times (n + 1) L_p = [\frac{p}{100} \times (n - 1)] + 1$$

Onde:

- p é a **ordem** do quantil em % (50% no caso mediana, por exemplo);
- n é o número de dados do rol; e,
- L é a **posição** do valor referente ao quantil desejado.

Os quartis calculados a partir das posições determinadas por essas regras *aproximadamente* subdividem o conjunto de dados em 25%, 50% e 75%.

Assim, para a determinação dos quartis pela primeira regra o valor de p seria:

- para o *primeiro quartil* (Q_1): $L_{q_{0,25}} = \frac{25}{100} \times (n)$;
- para o *segundo quartil* (a mediana ou Q_2): $L_{q_{0,50}} = \frac{50}{100} \times (n)$; ou,
- para o *terceiro quartil* (Q_3): $L_{q_{0,75}} = \frac{75}{100} \times (n)$.

Novamente podemos nos deparar com **duas situações possíveis** para o valor calculado para a posição L qualquer que seja a regra:

- se o valor calculado da **posição L** for um **inteiro**, nessa posição encontraremos o valor referente ao quantil desejado;
- se o valor calculado da **posição L** for **fracionário**, o valor desse quantil será determinado pela média entre os dois valores dos dados que estão nas **posições** imediatamente anterior e imediatamente posterior à posição **L** calculada.

Juntamente com as observações mínima (x_i) e máxima (x_n), o 1º, 2º e 3º Quartis são importantes para se ter uma boa idéia da assimetria da distribuição dos dados.

Para uma distribuição simétrica (ou aproximadamente simétrica) deveremos observar (Distribuição Gaussiana):

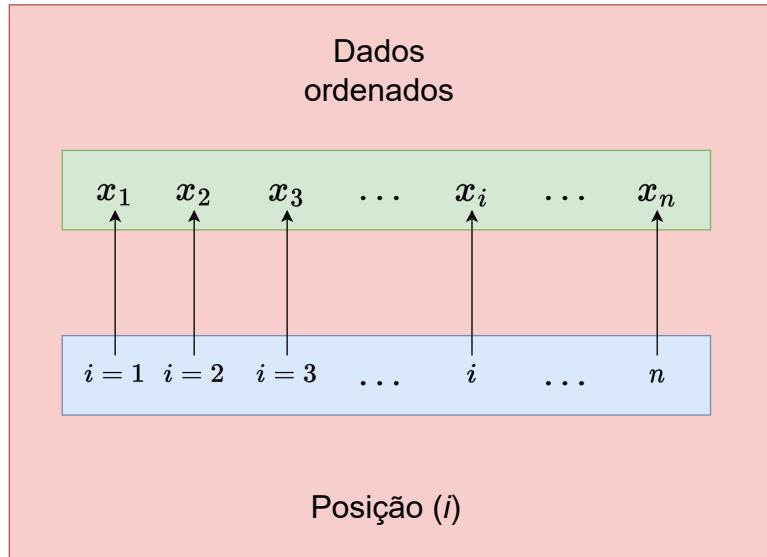


Figure 3.11: Entendendo a indexação de dados

- a dispersão inferior: $q_2 - x_1 \approx x_n - q_2$ à dispersão superior ;
- $q_2 - q_1 \approx q_3 - q_2$; e,
- $q_1 - x_1 \approx x_n - q_3$.

Para nosso conjunto de dados, segundo a regra empírica apresentada teremos as seguintes posições para determinação dos valores dos quartis:

- para o *primeiro quartil*:

$$\begin{aligned}
 L_{Q_1} &= \frac{p}{100} \times (n) \\
 &= \frac{25}{100} \times (60) \\
 &= 0,25 * 60 \\
 &= 15
 \end{aligned}$$

- para o *segundo quartil*:

$$\begin{aligned}
 L_{Q_2} &= \frac{p}{100} \times (n) \\
 &= \frac{50}{100} \times (60) \\
 &= 0,5 * 60 \\
 &= 30
 \end{aligned}$$

- para o *terceiro quartil*:

$$\begin{aligned}
 L_{Q_3} &= \frac{p}{100} \times (n) \\
 &= \frac{75}{100} \times (60) \\
 &= 0,75 * 60 \\
 &= 45
 \end{aligned}$$

E os quartis serão:

$-Q_1=1,63$
 $-Q_2=1,67$
 $-Q_3=1,75$

3.4 Medidas de forma (assimetria & curtose)

Quando analisamos o histograma (a representação gráfica da distribuição das frequências dos valores agrupados em classes) de uma determinada variável, não é muito comum que ele se mostre simétrico tal como seria se os dados fossem distribuídos de modo exatamente Normal.

Ao observarmos que a cauda se mostra mais alongada para a direita (indicativo da existência de uma quantidade maior de dados com grandes valores, arrastando a média para a direita: moda < mediana < média) diz-se que a *distribuição é assimétrica à direita*. Na situação oposta (moda > mediana > média) diz-se que ela é *assimétrica à esquerda*.

```

a=rbeta(10000,5,2)
c=rbeta(10000,5,5)
b=rbeta(10000,2,5)

par(mfrow=c(1,3))
hist(a,
      xlab="Valores", col = 'lightblue',
      ylab="Frequência",
      main="Assimetria à esq.")
hist(c,
      xlab="Valores", col = 'lightblue',
      ylab="Frequência",
      main="Assimetria à dir")
hist(b,
      xlab="Valores", col = 'lightblue',
      ylab="Frequência",
      main="Assimetria neutra")

```

```
main="Relativa simetria")
hist(b,
  xlab="Valores", col = 'lightblue',
  ylab="Frequência",
  main="Assimetria à dir.")
```

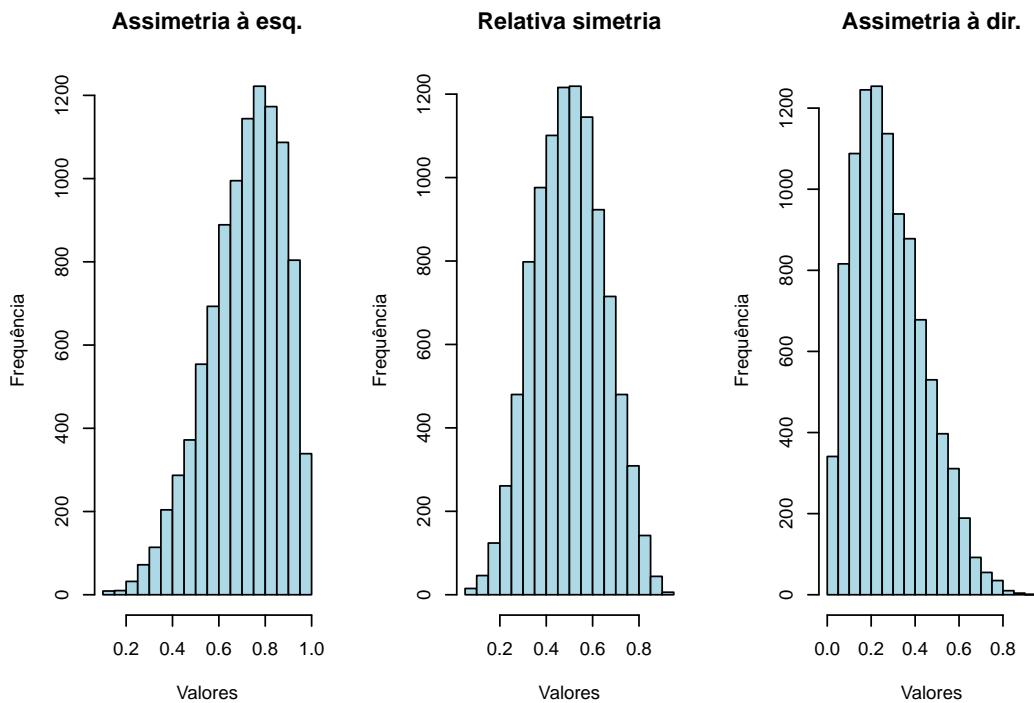


Figure 3.12: Diferentes formas na distribuição dos dados

De modo assemelhado, o histograma pode denotar uma forma mais *plana* ou menos *aguda*, onde um *cume* mostra-se mais destacado.

Nesse aspecto da forma, uma variável com distribuição Gaussiana apresentaria uma curva a que denominamos *mesocúrtica*. Distribuições com um aspecto mais plano são denominadas de *platicúrticas* e as com um cume agudo são denominadas *leptocúrticas*.

A curtose é uma medida da agudeza da distribuição dos dados em relação à distribuição Gaussiana.

Essas possíveis variações na forma de uma distribuição podem ser numericamente quantificadas através dos *coeficientes de assimetria e curtose*.

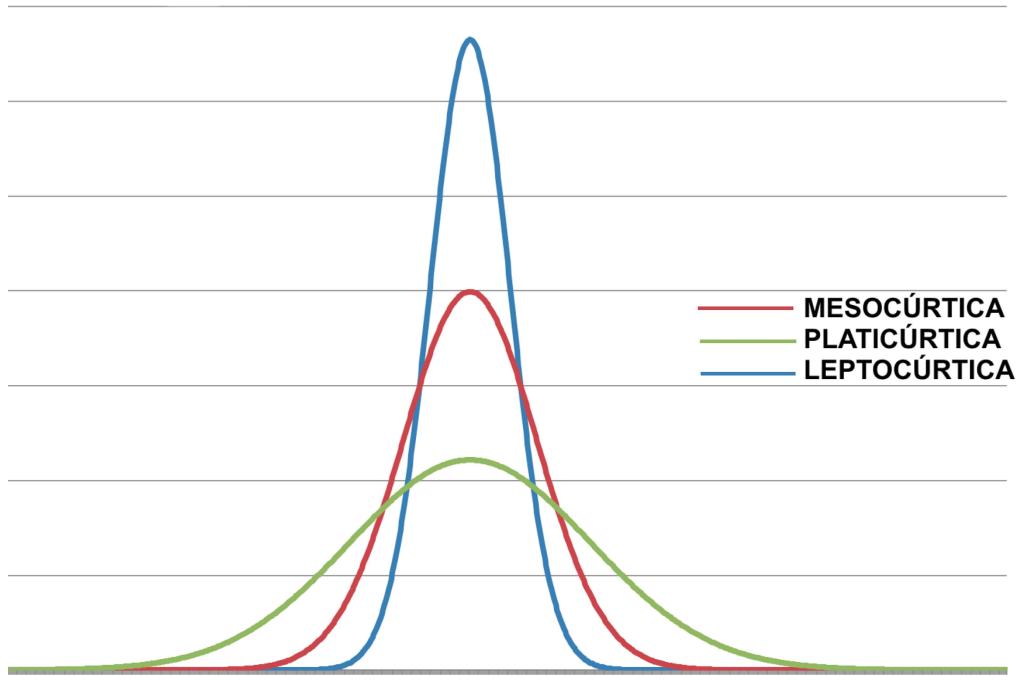


Figure 3.13: Diferentes aspectos de uma distribuição quanto à sua inclinação

Uma das medidas do coeficiente de assimetria é através do *primeiro ou segundo coeficientes de Pearson*, dados pelas seguintes relações:

- Primeiro coeficiente de assimetria de Pearson: $AS = \frac{\bar{x} - M_o}{s}$
- Segundo coeficiente de assimetria de Pearson: $AS = \frac{3(\bar{x} - M_d)}{s}$

Onde:

- \bar{x} é a média;
- M_o é a moda;
- S é o desvio padrão; e,
- M_d é a mediana.

A *assimetria* é classificada do modo seguinte:

- $-1 \leq AS \leq 1$: distribuição simétrica;

- $AS < -1$: distribuição com assimetria negativa; e,
- $AS > 1$: distribuição com assimetria positiva.

Uma das medidas do coeficiente de curtose é através da seguinte relação entre *quartis* e *percentis*:

$$K = \frac{Q_3 - Q_1}{2 \times (P_{90} - P_{10})}$$

Onde:

- $Q_3 = 3^{\circ}$ quartil;
- $Q_1 = 1^{\circ}$ quartil;
- $P_{90} = 90^{\circ}$ percentil; e,
- $P_{10} = 10^{\circ}$ percentil.

O *coeficiente de curtose* é classificado do modo seguinte:

- $k = 0,263$: distribuição mesocúrtica;
- $k < 0,263$: distribuição leptocúrtica; e,
- $k > 0,263$: distribuição platicúrtica.

3.5 Apresentação tabular de dados

As sínteses numéricas expostas condensam ao máximo a informação trazida pelos dados na forma de estatísticas associadas à:

- posição: média, moda, mediana;
- dispersão: amplitude total dos dados, variância (esvio padrão), coeficiente de variação;
- separatrizes (repartição): como por exemplo os quartis (Q_1 ; Q_2 /mediana e Q_3).

A correta exposição dos dados na forma de tabelas e gráficos auxilia o entendimento de muitas outras características relacionadas aos dados trabalhados por parte do leitor com grande riqueza visual.

Ao se lidar com grandes conjuntos de dados a visualização da informação contida nos dados fica comprometida se eles forem simplesmente apresentados como uma listagem, mesmo que depurados de eventuais inconsistências e ordenados como a lista abaixo:

```
sort(alturas)
```

```
## [1] 1.41 1.44 1.47 1.54 1.55 1.56 1.56 1.56 1.57 1.58 1.58 1.61 1.62 1.62 1.63
## [16] 1.64 1.64 1.65 1.65 1.65 1.66 1.66 1.66 1.66 1.66 1.67 1.67 1.67 1.67 1.67
## [31] 1.68 1.68 1.68 1.69 1.71 1.71 1.72 1.72 1.73 1.73 1.73 1.73 1.73 1.74 1.75
## [46] 1.76 1.76 1.77 1.78 1.78 1.79 1.82 1.83 1.83 1.84 1.85 1.86 1.93 1.95 2.00
```

Um dos modos de se lidar com isso é condensando a informação dos dados brutos em tabelas.

Uma tabela é uma forma não discursiva de apresentar informações nas quais o dado numérico se destaca como informação central. Uma tabela se diferencia de um quadro por este ter todos os seus campos delimitados por linhas e conter apenas informações de natureza qualitativa.

Uma tabela deve conter algumas **informações essenciais**, fora daquela estritamente relacionada aos dados, para que a compreensão do leitor acerca dos dados expostos seja a mais imediata possível:

- título que explique o que a tabela contém, local, data;
- cabeçalho nas colunas e linhas com a explicação, ainda que resumida, a que se referem as quantidades expostas no corpo;
- corpo formado pelos dados referentes às variáveis;
- fonte dos dados;
- uniformidade no número de casas decimais apresentadas no corpo;
- todas as casas devem apresentar valores ou símbolos que expliquem a ausência da informação (NI, NE, ou 0-zero).

Trabalhos de natureza acadêmica ou científica deveriam obrigatoriamente seguir, quando publicados no Brasil, a norma vigente publicada pela ABNT: Associação Brasileira de Normas Técnicas e algumas publicações do IBGE: Instituto Brasileiro de Geografia e Estatística (como em link).

Observa-se frequentemente, todavia, que as publicações seguem normas particulares das instituições de ensino (para trabalhos de conclusão de curso, monografias, dissertações e teses) ou das editoras (artigos), muitas vezes mescladas com recomendações da ABNT. Na Universidade Estadual de Londrina o portal da biblioteca possui uma ligação para a seção “Normas para trabalhos” (link).

3.5.1 Apresentação tabular de dados qualitativos

3.5.1.1 Dados qualitativos em entrada única

Para alguns tipos de dados, a apresentação tabular é bastante imediata.

Admita que tenha sido realizada uma pesquisa junto a um terminal de desembarque internacional em algum aeroporto sobre o continente de procedência do passageiro, num determinado período de um certo dia, tendo sido anotados os seguintes valores: AM, AM, A, A, A, AM, EU, EU, EU, EU, AM, AS, AS, AS, OC, AS, EU, AM, onde os continentes anotados são assim identificados: americano (AM); africano (A), europeu (EU); asiático (AS) e da oceania (OC). Uma tabela para a apresentação dos resultados poderia ser:

Table 3.2: Desembarques no terminal internacional A em Cumbica (SP, Brasil-10/10/2021: 8 h 00min às 12 h 00 min)

Continente de procedência	Desembarques
América	5
África	3
Europa	5
Ásia	4
Oceania	1
Total	18

Fonte: Próprio autor

Outro exemplo de apresentação tabular onde são apresentadas as proporções relativas observadas de cada nível da variável estudada (“tipo de família”, com quatro níveis diferentes), de um levantamento amostral feito pela Agência do Censo dos Estados Unidos em 2005.

Table 3.3: Estrutura domiciliar dos Estados Unidos

Estrutura domiciliar	Número (milhões)	Freq. rel.	Freq. rel. (%)
Casal com filhos	24,1	0,22	22
Casal sem filhos	31,1	0,28	28
Solteiro, sem parceiro	19,1	0,17	17
Morando sozinho	30,1	0,27	27
Outros domicílios	6,7	0,06	6
Total	111,1	1,00	100%

Fonte: Próprio autor

3.5.1.2 Dados qualitativos em entrada dupla

Outros tipos de dados são provenientes de pesquisas que têm por base respostas de natureza binária como, por exemplo:

- sim ou não;
- gosto ou não gosto;
- voto em “A” ou voto em “B”; ou,
- concordo ou não concordo.

Como resultado final, são obtidas contagens que expressam as frequências absolutas observadas para cada uma das variáveis (ou seus níveis) como na apresentação tabular de dados qualitativos por *Tabelas de Contingência*.

As *tabelas de contingência* são usadas para associar duas ou mais variáveis qualitativas (ou seus níveis) às contagens das respostas obtidas, na forma das frequências absoluta e relativa observadas em cada uma dessas variáveis (ou seus níveis).

O uso desse tipo de tabela é comum quando se pretende investigar se as variáveis estudadas têm alguma associação por meio de testes não paramétricos. Esse tipo de apresentação facilita a extração de informações relacionadas às probabilidades marginais ou condicionadas de cada uma variáveis ou seus níveis.

Admita agora que a pesquisa anterior junto ao terminal de desembarque internacional tenha também apontado o sexo do passageiro em seu desembarque. Uma tabela de dupla entrada com aqueles dados assumiria a forma:

Table 3.4: Desembarques no terminal internacional A em Cumbica (SP, Brasil - 10/10/2021: 8 h 00min às 12 h 00 min)

Desembarques no terminal internacional A em Cumbica (SP, Brasil)	Sexo do passageiro		Total
	M	F	
América	3	2	5
África	3	0	3
Europa	1	4	5
Ásia	2	2	4
Oceania	0	1	1
Total	9	9	18

Fonte: Próprio autor

Table 3.5: Incidência de baixo peso ao nascer em recém-nascidos de Pelotas, RS, segundo o hábito tabágico da mãe durante a gravidez (1982)

Classificação da mãe	Baixo peso ao nascer		Total
	Sim	Não	
Fumante	275	2.144	2.419
Não fumante	311	4.496	4.807
Total	586	6.640	7.226

Fonte: Próprio autor

Um outro exemplo, usando dados da incidência de baixo peso ao nascer em recém-nascidos de Pelotas (RS) segundo o hábito tabágico da mãe durante a gravidez (1982):

Ou ainda neste outro estudo que analisa a inclinação partidária de dois tipos de núcleos familiares em relação à presença de filhos:

Table 3.6: Inclinação partidária (frequências absolutas)

Estrutura domiciliar	Democrata	Republicano	Totais
Casal com filho(s)	762	468	1230
Casal sem filhos	484	477	961
Totais	1246	945	2191

Fonte: Próprio autor

A partir das contagens obtidas na pesquisa (as frequências absolutas), uma tabela com as frequências relativas pode ser construída, passando a apresentar as proporções relativas de cada categoria em relação aos níveis pesquisados:

Table 3.7: Inclinação partidária (frequências relativas)

Estrutura domiciliar	Democrata (%)	Republicano (%)	Totais (%)
Casal com filho(s)	34,78	21,36	56,14
Casal sem filhos	22,09	21,77	43,86
Totais (%)	56,87	43,13	100

Fonte: Próprio autor

3.5.2 Apresentação tabular de dados quantitativos

Todavia, para grandes quantidades de observações de dados quantitativos, a apresentação na forma de tabelas deve ser precedida do agrupamento dos valores observados em classes. O procedimento estatístico de agrupar os dados em *classes* ou *categorias* envolve construir uma *tabela de distribuição de frequências*.

Uma *tabela de distribuição de frequências* associa cada *classe* (intervalo) de valores da variável estudada ao número de ocorrências observadas. Como *regra prática*, a repartição dos dados brutos em classes deve sempre observar para que não haja um número excessivo de classes (diminuição da finalidade de resumir os dados, criação de classes sem nenhuma observação) nem tampouco poucas (que não possibilitem a visualização da distribuição e promovam perda da informação original).

A construção de uma *distribuição de frequências* consiste essencialmente em:

- escolher as *classes* ou *intervalos* (dados quantitativos) ou *categorias* (dados qualitativos);
- separar ou enquadrar os dados nessas *classes* ou *categorias*; e,
- contar o número de dados de cada *classe* ou *categoria*.

A literatura propõe vários modos para se determinar o número k de classes:

Crítério	Tamanho da amostra (n)	Fórmula
Raiz quadrada	$n \leq 25$	$k=5$
Raiz quadrada	$25 \leq n \leq 220$	$k=\sqrt{n}$
Herbert Sturges	$25 \leq n \leq 220$	$2^k > n$
Giuseppe Milone	$135 \leq 572237$	$k=1+3,22\log(n)^{(1)}$
	$20 \leq 36315$	$k=-1+2\ln(n)^{(2)}$

- ⁽¹⁾: logarítmico na base 10; e
- ⁽²⁾: logarítmico na base e .

Ao se escolher um número (k) de classes deve-se **ponderar** para que:

- os intervalos das classes tenham, geralmente, a mesma amplitude (raramente se necessita dispor de classes com amplitudes diferentes);
- os intervalos, a faixa de variação que vai do limite inferior da **primeira classe** ao limite superior da **última classe***, devem conter todos os valores possíveis da variável;
- cada valor observado deve pertencer **apenas a uma classe**;
- nenhuma classe deverá estar vazia (sem observação alguma);
- não adotar um número muito elevado de classes de modo que cada classe possua poucas observações (ou mesmo nenhuma); e,
- não adotar um número muito reduzido de classes de modo a esconder a variabilidade dos dados ao se reunir todas as observações em poucas faixas de valores;
- alguns autores recomendam um número mínimo de 5 classes e um máximo de 15;
- podemos considerar a amplitude de cada classe com **uma casa decimal a mais que os dados** de modo a facilitar a incorporação do último valor (mais elevado) na última classe.

Em nosso exemplo das alturas dos estudantes, a determinação do número de classes pelo critério da *raiz quadrada* ($n=60$) sugere 8 classes (outros critérios: pelo menor inteiro tq. $2^k > n; k = 6$, pelo critério de Sturges $k = 6, 86 \sim 7$, de Giuseppe Milone $k = 8, 18 \sim 9$).

$$\begin{aligned} k &= \sqrt{n} \\ &= 7,74 \end{aligned}$$

Arredondar para **mais**: $k = 8$.

A *amplitude total* (C) dos valores observados é a simples diferença entre o *valor máximo* (2,00 m) e o *valor mínimo* (1,41 m):

$$\begin{aligned} C &= 2,00 - 1,41 \\ &= 0,59m \end{aligned}$$

A amplitude de cada uma das classes (c) será dada pelo quociente da *amplitude total* (C) pelo *número de classes* (k).

$$\begin{aligned} c &= \frac{C}{k} \\ &= \frac{0,59}{8} \\ &= 0,07375m \end{aligned}$$

Arredondar para **mais**: $c = 0,075m$.

As classes são então assim construídas:

- Limite inferior da 1^a classe (LI_1): valor mínimo observado; e,
- Limite superior da 1^a classe (LS_1): $LI_1 + c$.

e assim sucessivamente até a última classe.

Símbolos gráficos para intervalos:

- Os símbolos abaixo indicam que o valor situado à sua esquerda **está incluído** no intervalo e o da direita **não está**:

$\vdash \bullet - \circ$

- Os símbolos abaixo indicam que o valor situado à sua esquerda **não está** incluído no intervalo e o da direita **está incluído***:

$\neg \circ - \bullet$

As tabelas que serão apresentadas a seguir estão sem os requisitos essenciais expostos anteriormente uma vez que o propósito é explicar a construção e cálculo dos valores de suas células.

Com $c = 0,075m$ as 8 classes ficam assim estabelecidas, tendo-se como ponto de partida o valor mínimo observado: 1,41 m - 1,485 m; 1,485 m - 1,56 m; 1,56 m - 1,635 m; 1,635 m - 1,71 m; 1,71 m - 1,785 m; 1,785 m - 1,86 m; 1,86 m - 1,935m; 1,935 - 2,01 m.

{1,41 ; 1,44 ; 1,47 ; 1,54 ; 1,55 ; 1,56 ; 1,56 ; 1,56; 1,57 ; 1,58 ; 1,58 ; 1,61 ; 1,62 ; 1,62 ; 1,63; 1,64 ;1,64 ; 1,65 ; 1,65 ; 1,65 ; 1,66 ; 1,66 ; 1,66 ; 1,66 ; 1,66 ; 1,67 ; 1,67 ; 1,67 ; 1,67 ; 1,68 ; 1,68 ;1,68 ; 1,68 ; 1,69 ; 1,71 ; 1,71 ; 1,72 ; 1,72 ; 1,73 ; 1,73 ; 1,73 ; 1,73 ; 1,74 ; 1,74 ; 1,75 ; 1,75 ; 1,76 ; 1,76 ; 1,76 ; 1,77 ; 1,78 ; 1,78; 1,79 1,82 ; 1,83 ; 1,83 ; 1,84 ; 1,85 ; 1,86 ; 1,93; 1,95 ; 2,00}

A tabela de distribuição de frequências com 8 classes, cada uma com amplitude 0,075 m, assume a forma:

Classe	Frequência absoluta (f_i)
1,41 m - 1,485 m	3
1,485 m - 1,56 m	2
1,56 m - 1,635 m	10
1,635 m - 1,71 m	19
1,71 m - 1,785 m	16
1,785 m - 1,86 m	6
1,86 m - 1,935 m	2
1,935m - 2,01 m	2
Total	60

Alternativamente, caso adotássemos como ponto de partida (um pouco abaixo do valor mínimo observado) o valor de 1,40 m e como amplitude de classe 0,08 m, uma tabela alternativa de distribuição de frequências teria como classes : 1,40 m - 1,48 m; 1,48 m - 1,56 m; 1,56 m - 1,64 m; 1,64 m - 1,72 m; 1,72 m - 1,80 m; 1,80 m - 1,88 m; 1,88 m - 2,06 m e, para facilitar a contagem das observações pertencentes a cada uma das classes ordenamos os dados:

$\{ 1,41 ; 1,44 ; 1,47 ; 1,54 ; 1,55 ; 1,56 ; 1,56 ; 1,56 ; 1,57 ; 1,58 ; 1,58 ; 1,61 ; 1,62 ; 1,62 ; 1,63 ; 1,64 ; 1,64 ; 1,65 ; 1,65 ; 1,65 ; 1,66 ; 1,66 ; 1,66 ; 1,66 ; 1,66 ; 1,67 ; 1,67 ; 1,67 ; 1,67 ; 1,68 ; 1,68 ; 1,68 ; 1,69 ; 1,71 ; 1,71 ; 1,72 ; 1,72 ; 1,73 ; 1,73 ; 1,73 ; 1,73 ; 1,74 ; 1,75 ; 1,76 ; 1,76 ; 1,77 ; 1,78 ; 1,78 ; 1,79 ; 1,82 ; 1,83 ; 1,83 ; 1,84 ; 1,85 ; 1,86 ; 1,93 ; 1,95 ; 2,00; \}$

A tabela de distribuição de frequências com 7 classes, cada uma com amplitude 0,08 m, assume a forma:

Classe	Frequência absoluta (f_i)
1,40 m \leftarrow 1,48 m	3
1,48 m \leftarrow 1,56 m	2
1,56 m \leftarrow 1,64 m	10
1,64 m \leftarrow 1,72 m	21
1,72 m \leftarrow 1,80 m	15
1,80 m \leftarrow 1,88 m	6
1,88 m \leftarrow 2,06 m	3
Total	60

Também podemos cogitar adotar alternativamente um intervalo de classe $c = 0,10$ m, com a primeira classe começando (um pouco abaixo do valor mínimo observado) na altura de 1,40 m; todavia, a última classe não iria contemplar o valor máximo observado (2,00 m) e necessitáfamos abrir mais uma classe apenas para incluí-lo.

Mas começando-se no valor mínimo obseravado (1,41 m) estariamos assegurando que o limite superior da última classe incluiria o valor máximo observado (2,00 m). Assim, essas seriam as classes sob uma amplitude de 0,10 m: 1,41 m - 1,51 m; 1,51 m - 1,61 m; 1,61 m - 1,71 m; 1,71 m - 1,81 m; 1,81 m - 1,91 m; 1,91 m - 2,01 m. O total de 6 classes (1,41 m a 2,01 m) cobre toda faixa de variação dos valores dos dados (de 1,41 m a 2,00 m) e é de rápida assimilação pelo leitor.

Ordenando-se os dados para facilitar a contagem das observações pertencentes a cada uma das classes:

$\{ 1,41 ; 1,44 ; 1,47 ; 1,54 ; 1,55 ; 1,56 ; 1,56 ; 1,56 ; 1,57 ; 1,58 ; 1,58 ; 1,61 ; 1,62 ; 1,62 ; 1,63 ; 1,64 ; 1,64 ; 1,65 ; 1,65 ; 1,65 ; 1,66 ; 1,66 ; 1,66 ; 1,66 ; 1,66 ; 1,67 ; 1,67 ; 1,67 ; 1,67 ; 1,68 ; 1,68 ; 1,68 ; 1,69 ; 1,71 ; 1,71 ; 1,72 ; 1,72 ; 1,73 ; 1,73 ; 1,73 ; 1,73 ; 1,74 ; 1,75 ; 1,76 ; 1,76 ; 1,77 ; 1,78 ; 1,78 ; 1,79 ; 1,82 ; 1,83 ; 1,83 ; 1,84 ; 1,85 ; 1,86 ; 1,93 ; 1,95 ; 2,00 \}$

A tabela de distribuição de frequências com 6 classes, cada uma com amplitude 0,10 m, assume a forma:

Classe	Frequência absoluta (f_i)
1,41 m \leftarrow 1,51 m	3
1,51 m \leftarrow 1,61 m	8
1,61 m \leftarrow 1,71 m	23
1,71 m \leftarrow 1,81 m	17
1,81 m \leftarrow 1,91 m	6
1,91 m \leftarrow 2,01 m	3
Total	60

Tabelas de distribuição de frequências mais completas podem montadas agregando muitas informações adicionais em novas colunas, mediante simples operações aritméticas.

Essas informações servem para tornar a visualização mais imediata e muitas delas são obtidas com operações matemáticas elementares:

- Classe i : é a simples identificação de cada classe;
- Amplitude (Δ_i) da classe i : a diferença entre o valor do limite superior e o do inferior de cada classe;
- Intervalo de valores da classe i (onde seu limite inferior **está contido** e o limite superior **não está contido**);
- Valor médio (\bar{x}_i) de cada classe i : o valor de seu **limite inferior** mais a metade da amplitude da classe;
- Frequência absoluta (f_i) da classe i : o número de observações contidas no intervalo da classe considerada;
- Frequência relativa ($fr_i = \frac{f_i}{n}$) da classe i (ou frequência relativa percentual, se assim apresentada): o quociente do número de observações n_i contidas no intervalo da classe f_i , pelo número total de observações (n);
- Frequência acumulada (fac_i) da classe i (ou frequência acumulada percentual, se assim apresentada): o número de observações com medidas contidas na classe i e nas anteriores a ela;
- Densidade absoluta ($\delta_i = \frac{f_i}{\Delta_i}$): o quociente do número de observações da classe (f_i) pela sua amplitude (Δ_i);
- Densidade relativa $\delta_{fr_i} = \frac{fr_i}{\Delta_i}$: o quociente da frequência relativa (fr_i) pela amplitude (Δ_i) da classe.

Vejo como exemplo as tabelas abaixo:

Classe	Int. de valores	Alt. média	Freq. abs.	Freq. rel.	Freq. rel. (%)	Freq. acumulada	Freq. acum. (%)
		(\bar{x}_i)	(f_i)	(fr_i)	($fr_i\%$)	(fac_i)	($fac_i\%$)
1	1,41 \leftarrow 1,51	1,46	3	0,05	5	3	5,00
2	1,51 \leftarrow 1,61	1,56	8	0,13	13,33	11	18,33
3	1,61 \leftarrow 1,71	1,66	23	0,38	38,33	34	56,66
4	1,71 \leftarrow 1,81	1,76	17	0,28	28,34	51	85,00
5	1,81 \leftarrow 1,91	1,86	6	0,10	10	57	95,00
6	1,91 \leftarrow 2,01	1,96	3	0,05	5	60	100,00

Classe	Int. de valores	Alt. média	Freq. abs.	Freq. rel.	Freq. rel. (%)	Freq. acumulada	Freq. acum. (%)
Totais	-		60	1,00	100,00	-	-

Classe	Int. de valores	Freq. abs.	Amplitude	Dens. abs	Freq. rel.	Dens. rel.
1	1,41 ⊢ 1,51	(f_i) 3	(Δ_i) 0,10	(δ_i) 30	(fr_i) 0,05	(δ_{fr_i}) 0,5
2	1,51 ⊢ 1,61	8	0,10	80	0,13	1,33
3	1,61 ⊢ 1,71	23	0,10	230	0,39	3,83
4	1,71 ⊢ 1,81	17	0,10	170	0,28	2,83
5	1,81 ⊢ 1,91	6	0,10	60	0,10	1
6	1,91 ⊢ 2,01	3	0,10	30	0,05	0,5
Totais	-	60	-	-	1,00	-

3.5.3 Média

Nas tabelas de *distribuições de frequências* os resultados estão agrupados em *intervalos de classes* (i). Por essa razão, os dados perdem sua identidade individual e passam a se representados pelo valor médio de cada intervalo (\bar{x}_i).

A média será então dada pelo produto deste valor médio de cada intervalo (\bar{x}_i) pela frequência absoluta que ele apresentou (n_i), dividido pela quantidade de dados (n).

Sejam n_1, n_2, \dots, n_n as frequências apresentadas para cada intervalo i dos valores assumidos pela variável X para o total n de observações. Assim a *média aritmética simples* para dados agrupados será dada por:

$$\bar{x} = \frac{\sum_{i=1}^k f_i \cdot \bar{x}_i}{n}$$

onde:

- \bar{x}_i : o valor médio do intervalo da classe i ;
- f_i : a frequência absoluta da classe i ;
- k é o número de classes da tabela de distribuição de frequências;
- n é o número de dados da tabela (eventualmente, os dados podem se referir a toda a população sob estudo)

3.5.4 Moda

Moda para dados apresentados na forma de uma distribuição de frequências:

$$Mo = l_{inf} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \times \Delta_i$$

Primeiramente identificamos a(s) classe(s) modal(is), que é (são) a(s) classe(s) com maior(es) frequência(s) absoluta f_i . Os demais elementos da expressão da moda são:

- l_{inf} : limite inferior da classe modal;
- Δ_1 frequência absoluta da **classe modal** menos a frequência absoluta da **classe anterior** à classe modal;
- Δ_2 frequência absoluta da **classe modal** menos a frequência absoluta da **classe posterior** à classe modal; e,
- Δ_i é o intervalo de cada classe.

3.5.5 Mediana

Mediana para dados apresentados na forma de uma **distribuição de frequências**:

$$Md = l_{inf} + \left[\frac{\frac{\sum_i^k f_i}{2} - f_{ac_{(md-1)}}}{f_{md}} \right] \times \Delta_i$$

Primeiramente identificamos a classe mediana, que é a classe que contém o elemento de posição $\frac{n}{2}$ (basta observar a coluna da frequência absoluta acumulada: f_{ac_i} , percorrendo-a até a classe i que tenha valor $> \frac{n}{2}$). Os demais elementos da expressão da mediana são:

- l_{inf} : limite inferior da **classe mediana**;
- $f_{ac_{(i-1)}}$: é a frequência absoluta acumulada até a **classe anterior à classe mediana**;
- f_{md} : é a frequência absoluta da **classe mediana**; e,

- Δ_i : é o intervalo de cada classe.

3.5.6 Variância

Variância para dados agrupados:

$$S^2 = \frac{1}{n-1} \times \left[\sum_{i=1}^k (\bar{x}_i)^2 \cdot f_i - \frac{\left(\sum_{i=1}^k \bar{x}_i \cdot f_i \right)^2}{n} \right]$$

em que:

- \bar{x}_i : o valor médio do intervalo da classe i ;
- f_i : a frequência absoluta da classe i ;
- k é o número de classes da tabela de distribuição de frequências;
- n é o número de dados da tabela (eventualmente, os dados podem se referir a toda a população sob estudo)

3.5.7 Quartis

Quartis para dados agrupados:

$$Q_i = l_{inf_{Q_i}} + \Delta_i \frac{L_{Q_i} - f_{ac_{Q_{i-1}}}}{f_{Q_i}}$$

em que:

- n é o número de dados;
- Q_i é o quartil desejado: $i = 1, 2, 3$;
- L_{Q_i} é posição do quartil desejado tal que:

- $L_{Q_1} = 0.25n$
- $L_{Q_2} = 0.5n$
- $L_{Q_3} = 0.75n$

- *classe quartílica* é a classe onde a posição do quartil desejado (L_{Q_i}) se localiza;
- $l_{inf_{Q_i}}$ é o limite inferior da *classe quartílica*;
- $f_{ac_{Q_{i-1}}}$ é a *frequência acumulada* da classe imediatamente anterior à classe quartílica;
- f_{Q_i} é a *frequência absoluta* de classe quartílica;
- Δ_i é a amplitude de cada classe (frequentemente igual para todas).

3.6 Apresentação gráfica de dados

Uma apresentação na forma gráfica torna ainda mais fácil a visualização das informações contidas nos dados. Há uma gama enorme de gráficos para a representação de dados a depender de sua natureza (qualitativa ou quantitativa).

3.6.1 Gráficos para uma variável qualitativa

- ranking: barras;
- parte em relação ao todo: setores;

3.6.1.1 Colunas

A partir das tabelas mostradas na seção 3.5.1.1 Dados qualitativos em entrada única poderíamos elaborar a apresentação gráfica na forma de *Gráficos de colunas*:

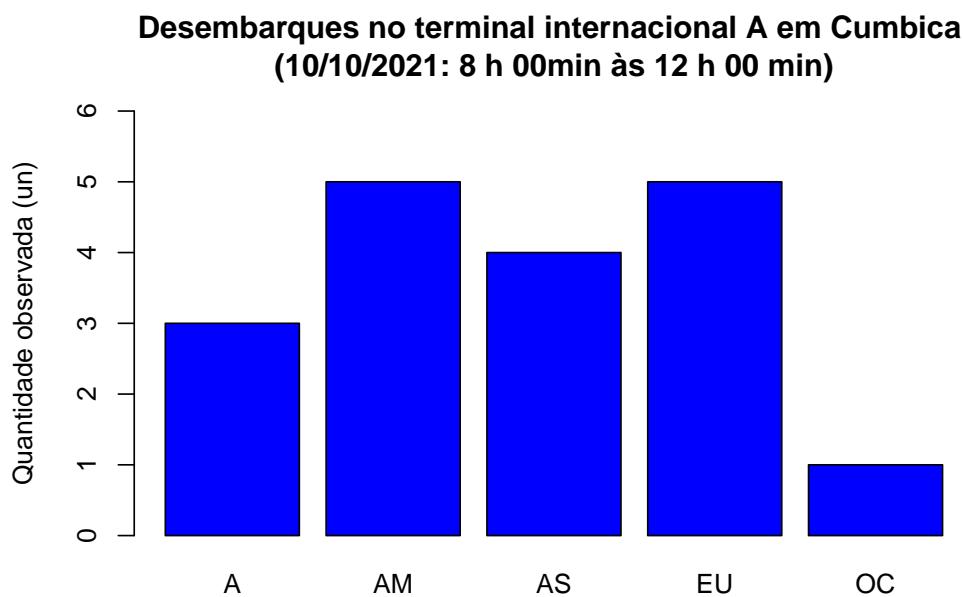
```
desembarque=c('AM', 'AM', 'A', 'A', 'A', 'AM', 'EU', 'EU', 'EU', 'AM', 'AS', 'AS', 'AS', 'OC', 'AS', 'EU', 'AM')
tab_desembarque=table(desembarque)

barplot(tab_desembarque,
```

```

main="Desembarques no terminal internacional A em Cumbica \n(10/10/2021: 8 h 00min
    ↵ às 12 h 00 min)",
sub= "Continente de procedência: América: AM; África: A; Europa: EU; Ásia: AS;
    ↵ Oceania: OC \nfonte: próprio autor",
xlab="",
ylab="Quantidade observada (un)",
ylim=c(0,6),
col="blue",
las=0,
hor="FALSE")

```



Continente de procedência: América: AM; África: A; Europa: EU; Ásia: AS; Oceania: C
fonte: próprio autor

Figure 3.14: Gráfico de barras dos dados observados no terminal de desembarque internacional do aeroporto

```

library(ggplot2)
dados=data.frame(tipo=c("Casal com filhos",
                        "Casal sem filhos",
                        "Solteiro, s/parceiro",
                        "Morando sozinho",
                        "Outros domicílios"),
                 quant=c(24.1, 31.1,
                        19.1, 30.1,
                        6.7))

ggplot(dados, aes(x=tipo, y=quant, color=tipo)) +
  geom_bar(stat="identity", position=position_dodge())+
  ggtitle("Estrutura domiciliar dos Estados Unidos, 2005")+
  theme(legend.position="bottom")+
  geom_text(aes(label=quant), vjust=1.6, color="white", position = position_dodge(0.9),
            size=3.5)+

```

```
scale_fill_brewer(palette="Paired")+
theme_minimal()+
xlab("") +
ylab("Frequência absoluta observada (milhões)")+
labs(colour = "Tipos de domicílios")
```

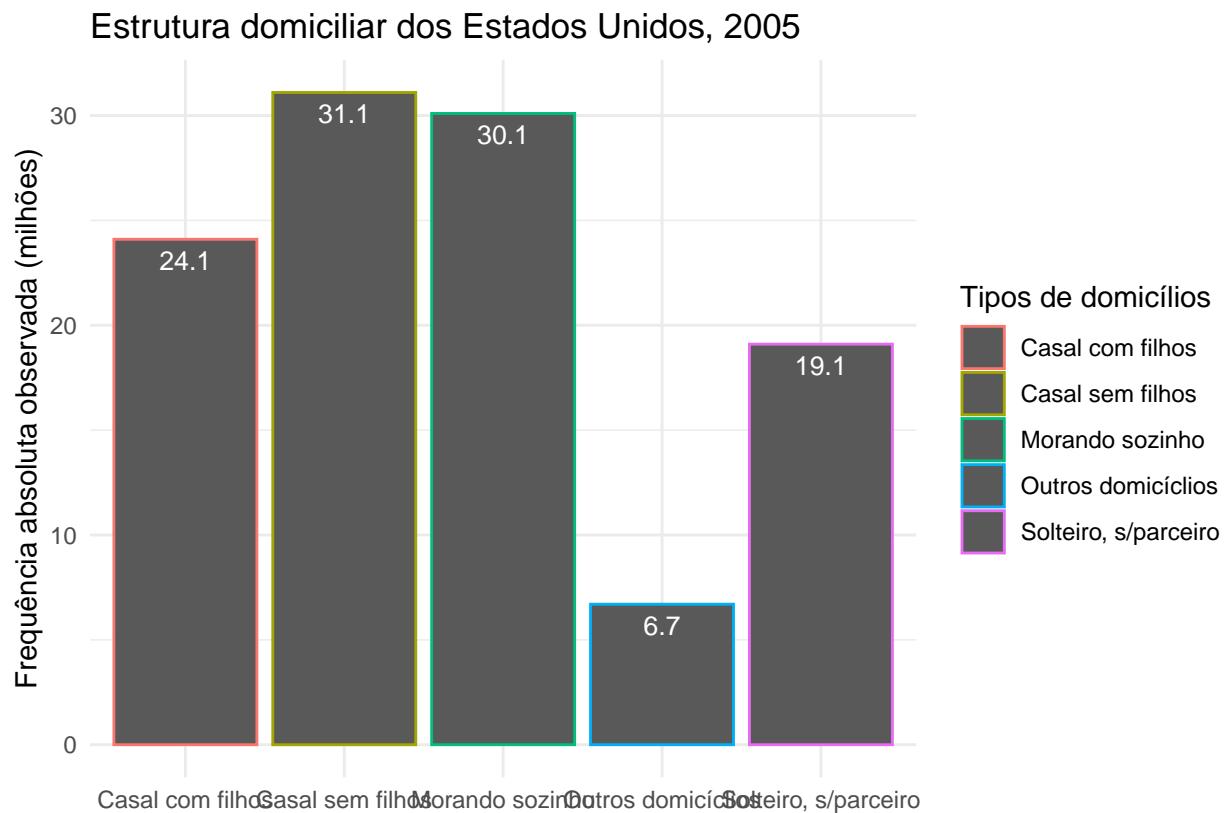


Figure 3.15: Gráfico de barras da estrutura domiciliar dos Estados Unidos

3.6.1.2 Setores

Em um *Gráfico de setores* a representação das quantidades está associada a uma fração do comprimento de um círculo. Para sua confecção considera-se a proporção da quantidade observada específica da quantidade total de dados, expressa na forma de fração do ângulo de um setor circular em relação ao ângulo interno total de um círculo (360°).

```
library(scales)
##
## Attaching package: 'scales'
```

```

## The following objects are masked from 'package:formattable':
##
##     comma, percent, scientific

library(ggplot2)

desembarques_classes=data.frame(
  group = c("América","África","Europa","Ásia","Oceania"),
  value = c(5,3,5,4,1))

blank_theme=theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    plot.title=element_text(size=14, face="bold")
  )

ggplot(desembarques_classes, aes(x="", y=value, fill=group)) +
  blank_theme +
  scale_fill_brewer("Blues")+
  labs(title="Desembarques no terminal internacional A em Cumbica",
       subtitle="(10/10/2021: 8 h 00min às 12 h 00 min)",
       caption = "Fonte: próprio autor") +
  theme(axis.text.x=element_blank()) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  geom_text(aes(y = value/2 + c(0, cumsum(value)[-length(value)])),
            label = percent(value/18 )), size=5)+
  guides(fill = guide_legend(title = "Legenda",
                             label.position = "right",
                             title.position = "top", title.vjust = 1))

```

```

library(ggplot2)
library(scales)

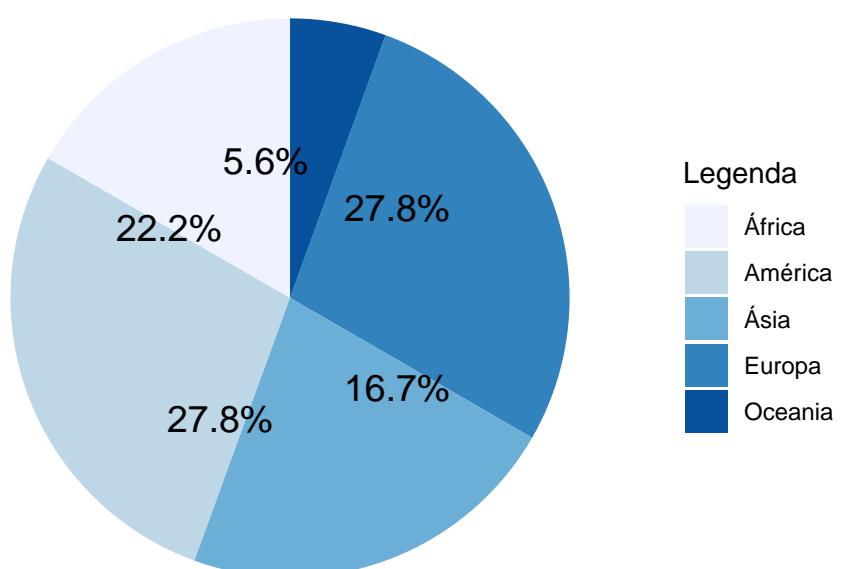
blank_theme=theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    plot.title=element_text(size=14, face="bold")
  )

bp=ggplot(dados, aes(x="", y=quant, fill=tipo))+#
  geom_bar(width = 1, stat = "identity")
pie=bp + coord_polar("y", start=0)

```

Desembarques no terminal internacional A em Cumbica

(10/10/2021: 8 h 00min às 12 h 00 min)



Fonte: próprio autor

Figure 3.16: Gráfico de setores dos desembarques observados no terminal de desembarque internacional do aeroporto

```

pie +
  scale_fill_brewer("Blues") +
  blank_theme +
  theme(axis.text.x=element_blank()) +
  geom_text(aes(x = 1.2,label = quant), position = position_stack(vjust = 0.5)) +
  ggtitle("Estrutura domiciliar dos Estados Unidos, 2005") +
  theme(legend.position = "right", legend.justification = "center", legend.direction =
    "vertical",
        legend.spacing.x = unit(0.5, 'cm'), legend.spacing.y = unit(0.5, 'cm')) +
  guides(fill = guide_legend(title = "Tipos de domicílios",
                             label.position = "right",
                             title.position = "top", title.vjust = 1))

```

Estrutura domiciliar dos Estados Unidos, 2005

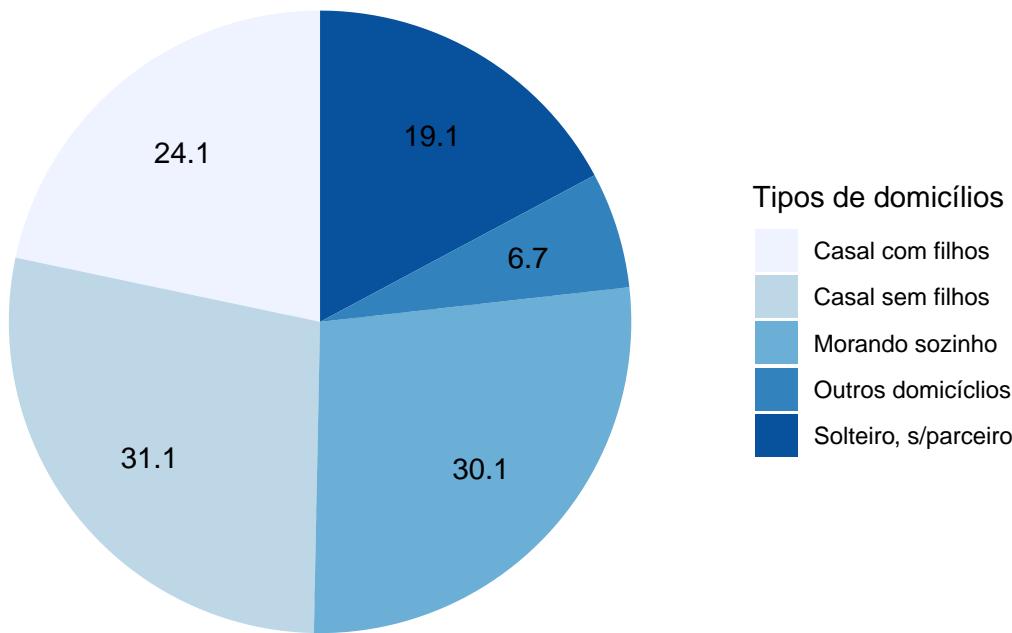


Figure 3.17: Gráfico de setores da estrutura domiciliar dos Estados Unidos

3.6.1.3 Colunas para dados em uma tabela de dupla entrada

```

library(ggplot2) # Carrega a biblioteca ggplot2

# Dados fornecidos
casal_com_filho_democratas <- 3478
casal_com_filho_republicano <- 2136
casal_sem_filho_democratas <- 2209

```

```

casal_sem_filho_republicano <- 2177

# Criar um dataframe com os dados
dados <- data.frame(
  Categoria = c("Com Filhos", "Com Filhos", "Sem Filhos", "Sem Filhos"),
  Partido = c("Democratas", "Republicanos", "Democratas", "Republicanos"),
  Contagem = c(casal_com_filho_democratas, casal_com_filho_republicano,
              casal_sem_filho_democratas, casal_sem_filho_republicano)
)

# Criar o gráfico de barras empilhadas
ggplot(dados, aes(x = Categoria, y = Contagem, fill = Partido)) +
  geom_bar(stat = "identity") +
  labs(title = "Contagem de Votos por Categoria e Partido (Censo dos EUA,2005)",
       x = "Categoria",
       y = "Contagem") +
  scale_fill_manual(values = c("Democratas" = "lightgreen", "Republicanos" = "lightblue")) +
  theme_minimal()

```

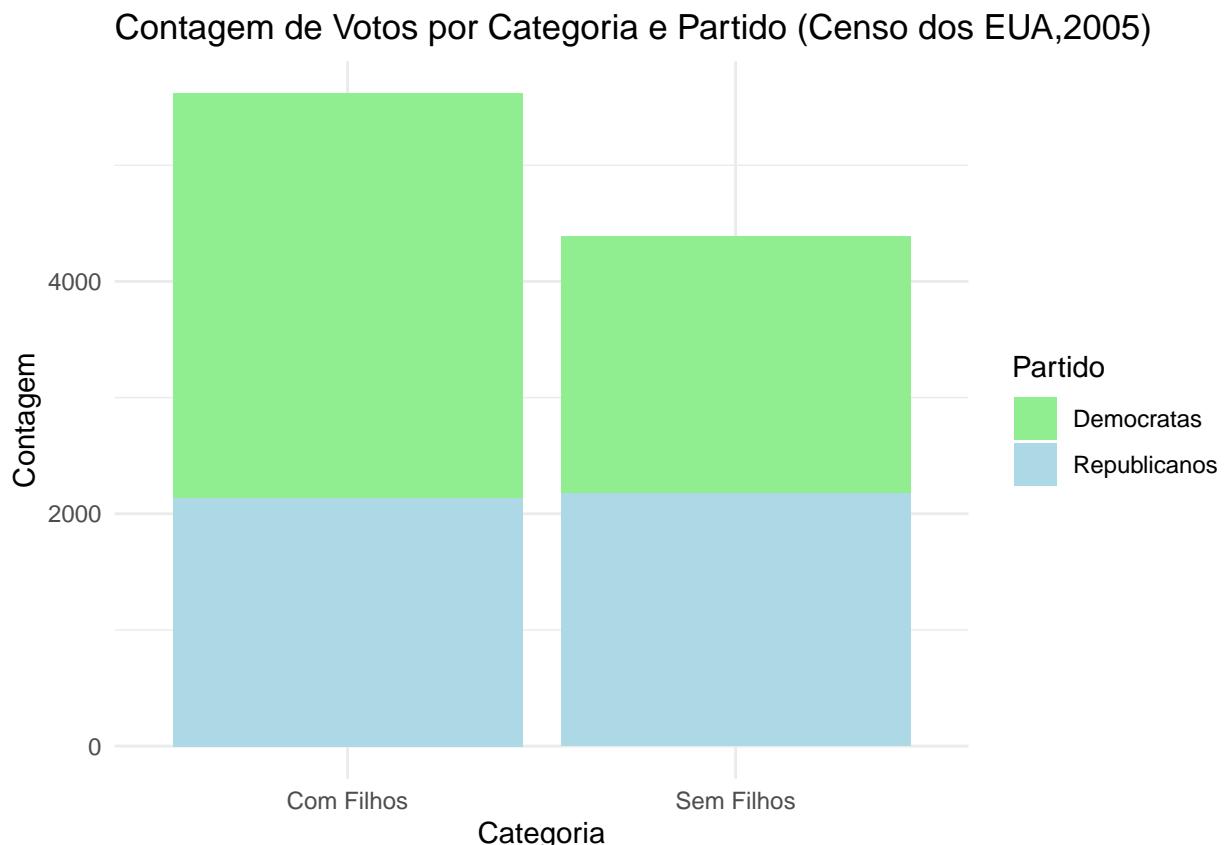


Figure 3.18: Gráfico de barras da estrutura familiar em relação à inclinação partidária nos Estados Unidos

```

library(ggplot2)                      # Carrega a biblioteca ggplot2

# Dados fornecidos

```

```

fumantes_filho_bp = 275
fumantes_filho_pn = 2144
n_fumantes_filho_bp = 311
n_fumantes_filho_pn = 6640

# Criar um dataframe com os dados
dados <- data.frame(
  Risco = c("Fumante", "Fumante", "Não fumante", "Não fumante"),
  Peso = c("Baixo peso", "Peso normal", "Baixo peso", "Peso normal"),
  Contagem = c(fumantes_filho_bp, fumantes_filho_pn,
               n_fumantes_filho_bp, n_fumantes_filho_pn)
)

# Criar o gráfico de barras empilhadas
ggplot(dados, aes(x = Risco, y = Contagem, fill = Peso)) +
  geom_bar(stat = "identity") +
  labs(title = "Peso de recém nascidos em Pelotas (RS, 1982)",
       x = "Exposição ao risco",
       y = "Contagem") +
  scale_fill_manual(values = c("Baixo peso" = "gray", "Peso normal" = "lightgreen")) +
  theme_minimal()

```

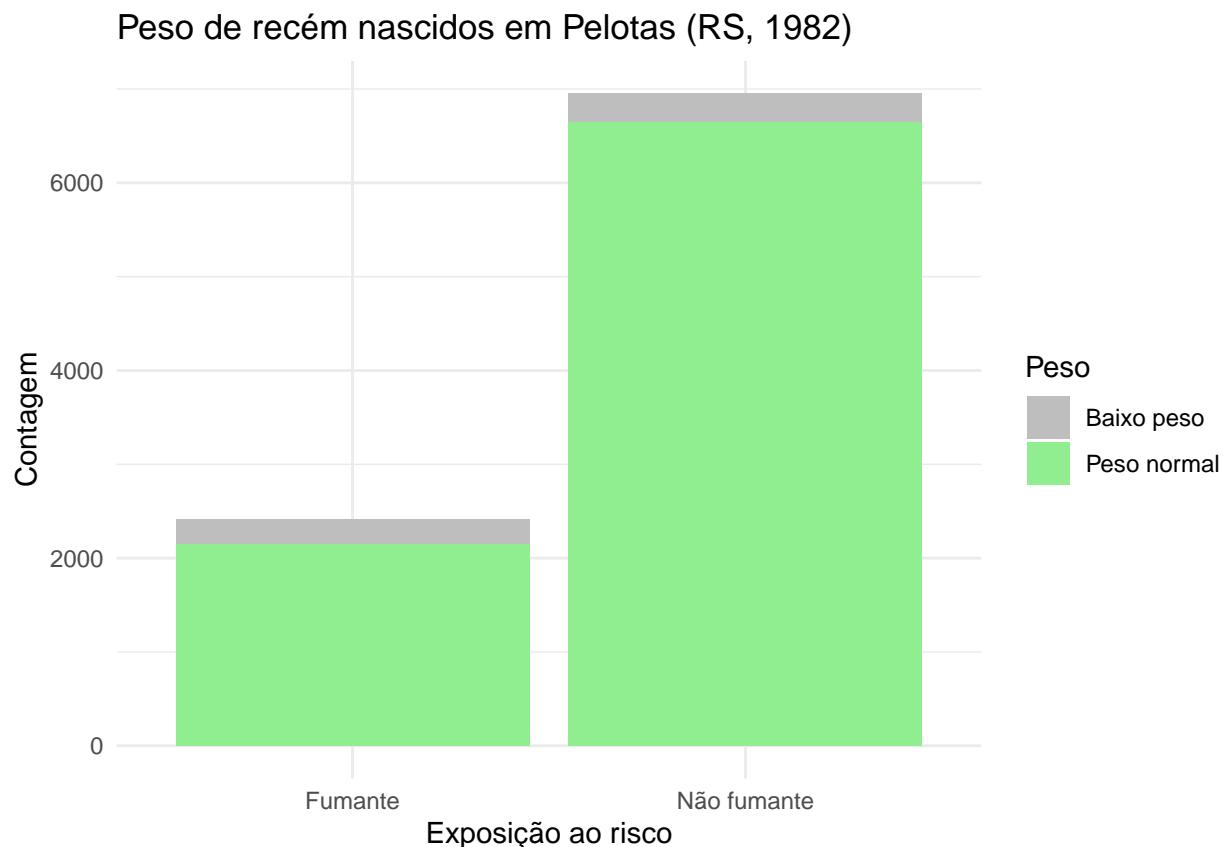


Figure 3.19: Gráfico de barras da exposição ao fator de risco e o efeito

3.6.2 Gráficos para uma variável quantitativa

- ranking: barras;
- parte em relação ao todo: setores;
- dispersão *unidimensional*;
- distribuição: histograma e o *box plot*.

3.6.2.1 Barras

Se modificarmos o diagrama de ramos e folhas dos comprimentos e quantidades observadas, representando cada uma das alturas medidas por um *retângulo* cujas alturas sejam proporcionais à quantidade contada de cada uma dessas alturas teremos um *Gráfico de barras*.

```
tab_alturas=table(alturas)

barplot(tab_alturas,
        main="Valores observados da alturas dos estudantes",
        xlab="Altura (cm)",
        ylab="Quantidade observada (un)",
        ylim=c(0,6),
        col="blue",
        las=0,
        hor="FALSE")
```

3.6.2.2 Histograma

Para dados quantitativos, o agrupamento dos valores brutos observados em classes (cada uma com um valor mínimo e máximo fixado) permite a geração de um *Histograma*, um tipo diferente de *Gráfico de barras* onde cada coluna está unida às colunas imediatamente adjacentes (indicando a continuidade de valores das medidas) e sua altura expressa a quantidade de observações contidas nessa classe.

Para as classes estabelecidas na seção anterior o histograma das alturas dos estudantes terá esse aspecto:

Valores observados da alturas dos estudantes

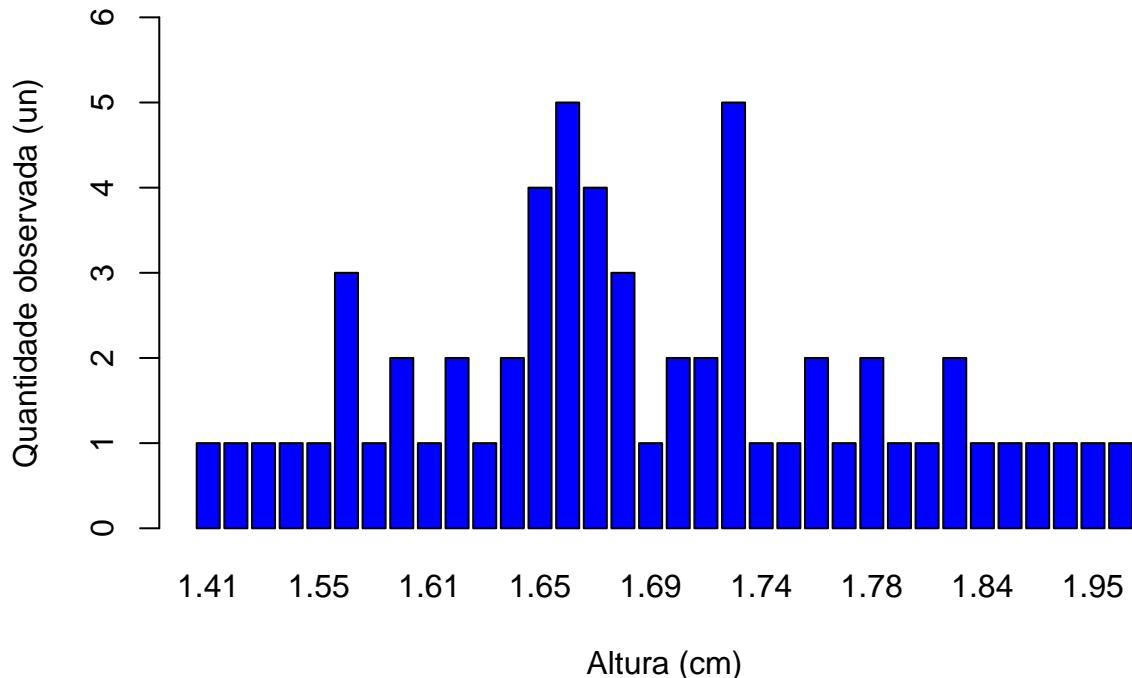


Figure 3.20: Gráfico de barras dos dados brutos: uma barra para cada observação e sua altura expressando o número de observações com esse valor

```

h1=hist(alturas, breaks=seq(1.41 , 2.01 , 0.1), include.lowest = TRUE, right = FALSE, main=
  "Histograma das alturas dos estudantes", col="blue",
xlab="Classes de alturas (m)", ylab="Frequência absoluta observada (un)" , cex=0.7,
  ylim=c(0,30))
text(h1$mid,h1$counts,labels=h1$counts, adj=c(0.5, -0.5))
abline(v=mean(alturas), col="red")
text(mean(alturas)-0.01, 28, "Média=1,69 m", col = "red", srt=90)
abline(v=median(alturas), col="darkgreen")
text(median(alturas)-0.01, 27.2, "Mediana=1,675 m", col = "darkgreen", srt=90)
abline(v=Modes(alturas), col="darkgrey")
text(Modes(alturas)+c(-0.01, -0.01), 27, c("Moda=1,66","Moda=1,73"), col = "darkgray",
  srt=90)
  
```

Um *histograma* é a representação gráfica de uma *tabela de distribuição de frequências* em colunas (retângulos).

A base de cada retângulo representa o intervalo de cada classe e a altura, a quantidade ou a *frequência absoluta* com que aquele valor da classe ocorre no conjunto de dados.

O termo *histograma* foi cunhado por Karl Pearson (c. 1891) e vem da composição em grego de *istos* (mastro) com *gramma* (escrita), convertida em inglês para *historical diagram: histogram*.

Histograma das alturas dos estudantes

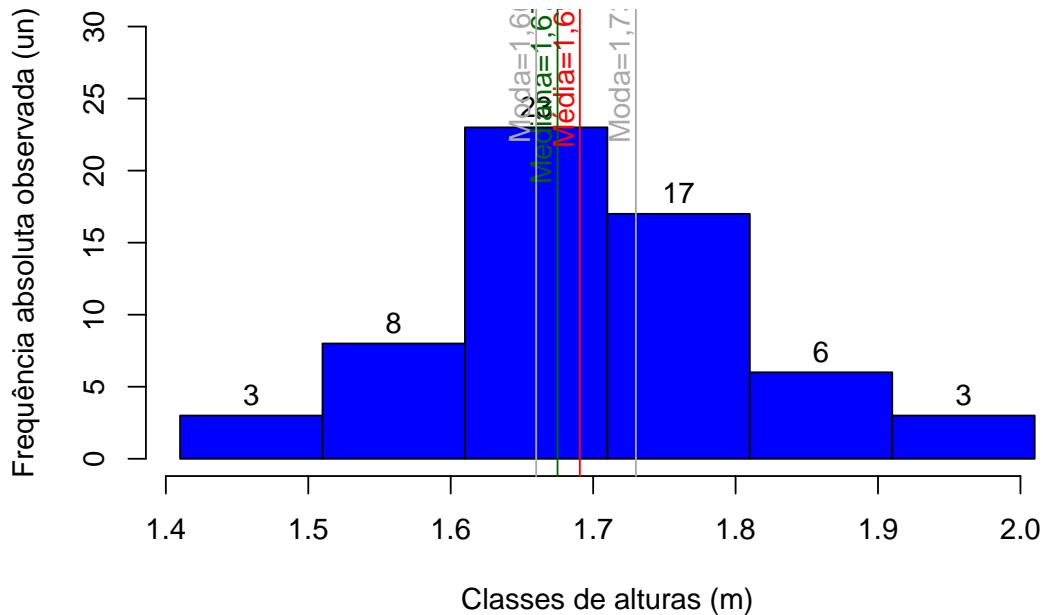


Figure 3.21: Histograma das alturas dos estudantes com as posições da média, moda e mediana

Como elemento gráfico, seu uso é anterior à sua denominação (maiores detalhes em: [\(link\)](#)).

Num *histograma de densidade*, a altura de cada retângulo representa uma *densidade* relacionada à *frequência relativa* no intervalo de cada classe.

```
h2=hist(alturas,breaks=seq(1.41 , 2.01 , 0.10), include.lowest = TRUE, right = FALSE, main=
  "Histograma das densidades das alturas dos estudantes", col="blue",
  xlab="Classes de alturas (m)", ylab="Densidade da freq. relativa", prob="TRUE", ylim=c(0,5))
text(h2$mid,h2$density,labels=round(h2$density, 5), adj=c(0.5, -0.5), cex=0.7)
lines(density(alturas), col="red")
lines(density(alturas, adjust=2), col="orange")
```

Como a área de cada retângulo é igual à proporção (fr_i) da classe (i) a soma de todas essas áreas será igual a 1:

$$(0.10*0.5)+(0.10*1.333)+(0.10*3.833)+(0.10*2.833)+(0.10*1)+(0.10*0.50)$$

```
## [1] 0.9999
```

Histograma das densidades das alturas dos estudantes

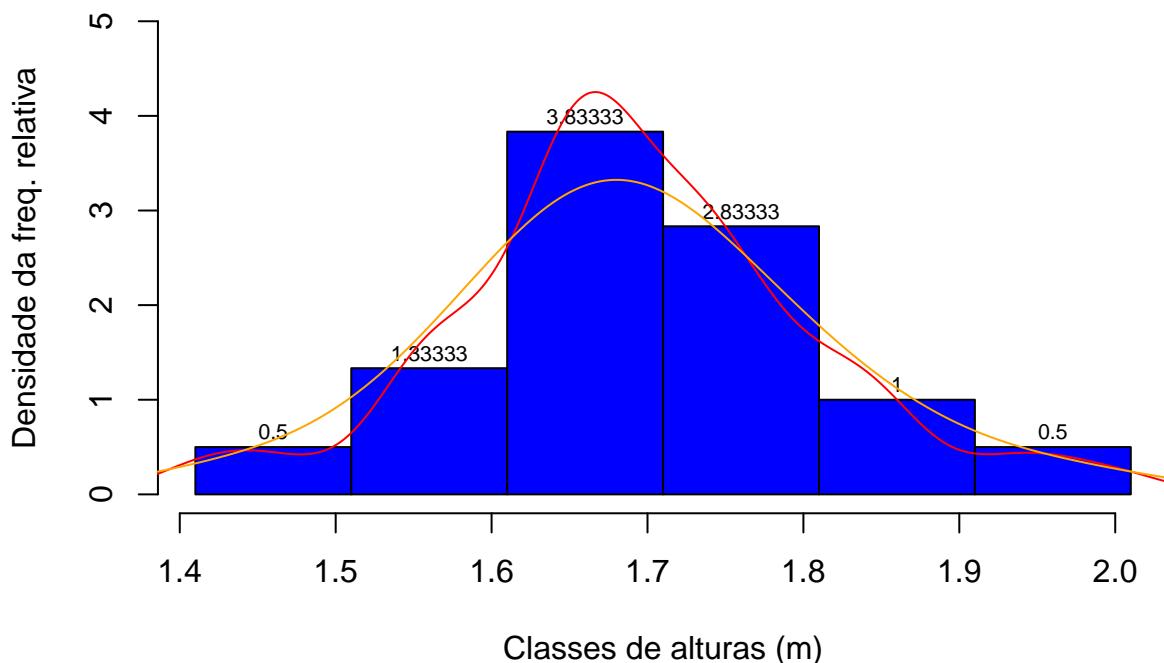


Figure 3.22: A linha vermelha é uma aproximação da Função de Densidade da frequência relativa de observação (a linha preta é a curva da função densidade de uma distribuição Normal com média e variâncias dadas pelos dados

Uma aproximação para a **área sob a curva da Função de Densidade** pode ser soma das áreas de um dos retângulo com:

- Base = Δ_i ; e,\
- Altura = $\frac{f_{r_i}}{\Delta_i}$.

A **área da curva da Função de Densidade delimitada por dois valores quaisquer** é uma analogia para a probabilidade de que um determinado valor de altura de um estudante (amostrado aleatoriamente dentre todos os 60 estudantes) esteja contida nesse intervalo.

Equivale dizer que, amostrando-se aleatoriamente um estudante dentre todos os 60 alunos, a probabilidade de que a altura desse estudante esteja contida entre os valores mínimo e máximo da amostra é, **naturalmente**, igual a 1 (100%)

3.6.2.3 Setores

Em um *Gráfico de setores* a representação das quantidades está associada a uma fração do comprimento de um círculo. Para sua confecção considera-se a proporção da quantidade observada específica da quantidade total de dados, expressa na forma de fração do ângulo de um setor circular em relação ao ângulo interno total de um círculo (360°).

```
library(scales)
library(ggplot2)

alturas_classes=data.frame(
  group = c("1,41-1,51",
            "1,51-1,61",
            "1,61-1,71",
            "1,71-1,81",
            "1,81-1,91",
            "1,91-2,01"),
  value = c(3,8,23,17,6,3))

blank_theme=theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    plot.title=element_text(size=14, face="bold")
  )

ggplot(alturas_classes, aes(x="", y=value, fill=group)) +
  blank_theme +
  scale_fill_brewer("Blues") +
  ggtitle("Alturas dos estudantes") +
  theme(axis.text.x=element_blank()) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  geom_text(aes(y = value/2 + c(0, cumsum(value)[-length(value)]),
                label = percent(value/60 )), size=5) +
  guides(fill = guide_legend("Classes de valores (m)",
                             label.position = "right",
                             title.position = "top", title.vjust = 1))
```

3.6.2.4 Box-plot (gráfico de caixas)

Alturas dos estudantes

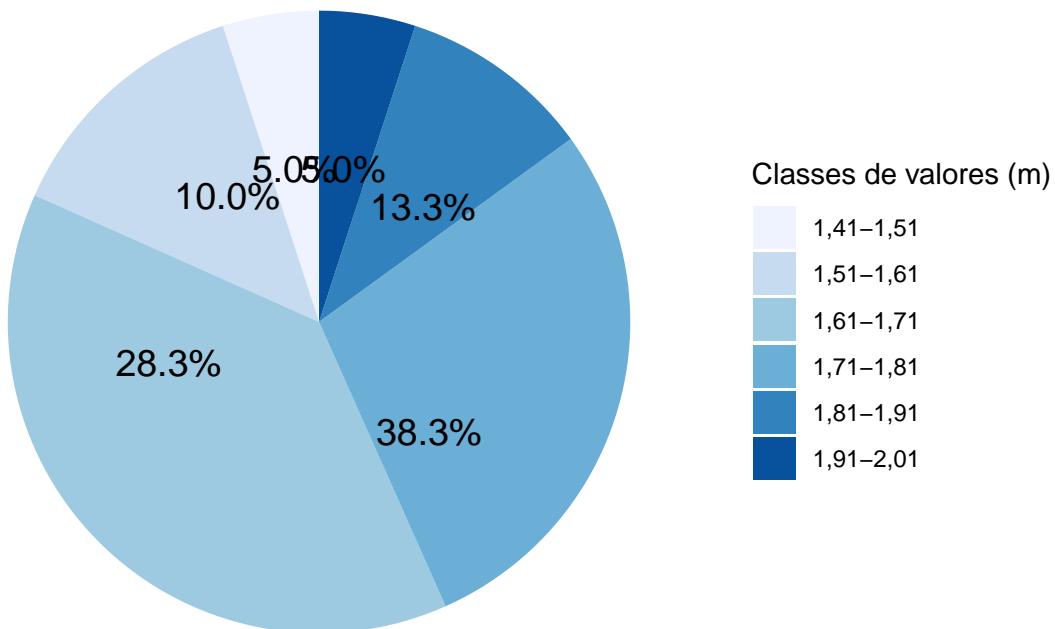


Figure 3.23: Gráfico de setores das alturas dos estudantes

O gráfico **Box-plot** (*box and whisker plot*): esse gráfico apresenta de modo conjunto, informações sobre a posição, dispersão, assimetria e dados discrepantes do conjunto analisado:

- a mediana (Q_2);
- os valores mínimo: x_1 e máximo: x_n (dados ordenados);
- o 1º e 3º quartis;
- a dispersão (intervalo interquartílico: $d_q = (Q_3 - Q_1)$);
- os limites superior: $LS = Q_3 + 1,50.d_q$, e inferior: $LI = Q_1 - 1,50.d_q$ (*bigodes*);
- os valores mínimo e máximo observados (caso não existam valores superiores aos limites LI e LS); ou
- as observações mais extremas, situadas fora dos limites LI e LS (que **podem ou não ser outliers**, dados atípicos).

```
min=min(alturas)
q1=1.635
q2=1.675
med=mean(alturas)
q3=1.755
max=max(alturas)
iq=q3-q1
ls=q3+1.5*iq
li=q1-1.5*iq
head(sort(alturas,TRUE)) #2.00 1.95 >>1.93<< 1.86 1.85 1.84
```

```

## [1] 2.00 1.95 1.93 1.86 1.85 1.84

tail(sort(alturas,TRUE)) # 1.56 1.55 1.54 1.47 1.44 >>1.41<<

## [1] 1.56 1.55 1.54 1.47 1.44 1.41

boxplot(alturas,
         main="Boxplot do conjunto de dados de alturas",
         ylim=c(1.2, 2.1))

lines( y=c(1.47, 1.47), x=c(0.6,1), col="blue")
text(x=0.60, y=1.47-0.05, "Delimitador inferior do bigode=1,47", col = "blue", srt=0)

lines( y=c(1.93,1.93), x=c(0.6,1), col="blue")
text(x=0.60, y=1.93+0.05, "Delimitador superior do bigode=1,93", col = "blue", srt=0)

lines(y=c(med, med), x=c(1,1.4), col="blue")
text(x=1.4 , y= med+0.05 , "Média=1,6907", col = "blue", srt=0)

lines(y=c(q1, q1), x=c(1, 1.4), col="blue")
text(x=1.4 , y=q1 -0.05, "Primeiro quartil: Q1=1,635", col = "blue", srt=0)

lines(y=c(q2, q2), x=c(0.6,1), col="blue")
text(x=0.60 , y= q2 - 0.05, "Mediana: Q2=1,675", col = "blue", srt=0)

lines(y=c(q3, q3), x=c(1, 1.4), col="blue")
text(x= 1.4 , y=q3 + 0.05, "Terceiro quartil: Q3=1,755", col = "blue", srt=0)

lines(y=c(li,li) , x=c(1.01,1.4) , col="red", lty=2)
text(x=1.2, y=q1-1.5*iq-0.05 , "Limite inferior teórico: LI=1,455) ", col = "red", srt=0)

lines(y=c(ls,ls) , x=c(1.01,1.4) , col="red", lty=2)
text(x=1.2, y=q3+1.5*iq +0.05 , "Limite superior teórico: LS=1,935", col = "red", srt=0)

points (y=1.47, x=1 , col="green", cex=1, lwd=5)
text(x=1, y=1.47-0.05 , "Última observação dentro do LI: h=1,47 ", col = "green", srt=0)

points (y=1.93, x=1 , col="green", cex=1, lwd=5)
text(x=1, y=1.93+0.05 , "Última observação dentro do LS: h=1,93 ", col = "green", srt=0)

```

Boxplot do conjunto de dados de alturas

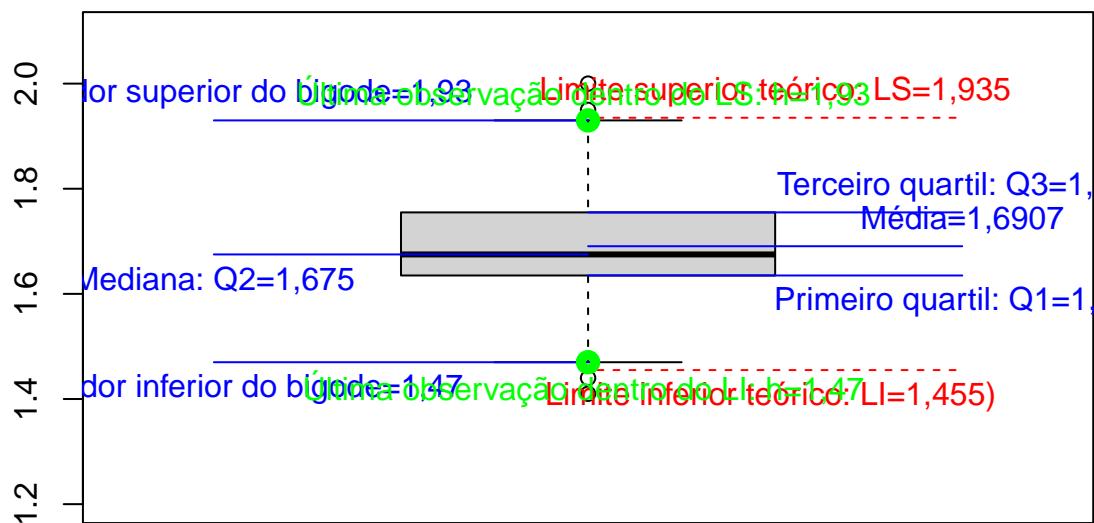


Figure 3.24: Box-plot de um rol de valores com Distribuição Normal (média 20 e variância 5

Módulo 4

Introdução ao cálculo de probabilidades

Seria bom começar o capítulo sobre teoria das probabilidades, dando uma definição concisa, simples e intuitiva, todavia formalmente rigorosa. Infelizmente, isto não será possível.

Se por um lado, uma definição rigorosa de probabilidade requer um aparato matemático sofisticado e é bem pouco intuitiva; por outro lado as definições simples e frequentemente encontradas são tautológicas como:

Probabilidade é um *número* que quantifica, uma *medida da informação* disponível sobre a *possibilidade* de ocorrência de um determinado *evento* quando ainda não se sabe se ele ocorrerá ou não.

Essa definição é “circular” (*definiendum = definien*) uma vez que se vale de um sinônimo de probabilidade (possibilidade) para se auto definir.

Todavia ela nos introduz **dois conceitos** que iremos usar como ponto de partida:

1. probabilidade refere-se a *experimentos de resultado incerto (aleatórios)*;
2. que probabilidade é uma *quantidade*.

4.1 Introdução histórica

Os estudos de probabilidade surgiram no século XVII, motivados por questões práticas relacionadas a jogos de azar e decisões econômicas.

Uma das situações que estimulou essas discussões foi o problema apresentado pelo **Cavaleiro de Méré** (Chevalier de Méré), que envolvia jogos de azar com dados. Ele levantou duas questões principais sobre a probabilidade de certos resultados ao lançar dados, que acabaram influenciando o desenvolvimento da teoria probabilística.

A primeira questão envolvia o lançamento de um dado seis vezes, onde Méré acreditava que havia uma alta chance de obter pelo menos um “6”. Sua intuição estava correta: a probabilidade de não obter um “6” em seis lançamentos consecutivos é $(\frac{5}{6})^6$, aproximadamente 33%, o que significa que a chance de obter pelo menos um “6” é de cerca de 67%.

O segundo problema que Méré trouxe era mais intrigante e envolvia o lançamento de dois dados 24 vezes. Ele acreditava que deveria obter pelo menos um duplo “6”, mas errou em sua previsão. A probabilidade de não obter um duplo “6” em 24 lançamentos consecutivos é $(\frac{35}{36})^{24}$, aproximadamente 51%, ou seja, a chance de obter um duplo “6” é apenas cerca de 49%, e não tão alta quanto ele esperava ao observar os resultados do jogo.

Essa discrepância entre intuição e realidade levou Méré a buscar ajuda com Pascal, e a subsequente troca de ideias com Fermat. Foi nessa correspondência entre **Blaise Pascal** e **Pierre de Fermat** em 1654, na qual discutiam problemas de divisão de apostas em jogos interrompidos, que se estabeleceu a base para o conceito de probabilidade esperada.

A formalização desses estudos avançou no século XVIII com a publicação da obra *Ars Conjectandi* (1713) de **Jacob Bernoulli**, que introduziu a *lei dos grandes números*. Essa lei estabelece que, com um número crescente de experimentos, a frequência observada de um evento tende a se aproximar de sua probabilidade verdadeira, fornecendo assim uma base teórica sólida para a análise de fenômenos aleatórios.

Outro avanço significativo veio com **Abraham de Moivre**, que em *The Doctrine of Chances* (1718) aplicou a teoria da probabilidade ao estudo de distribuições estatísticas, introduzindo a curva normal para modelar variáveis aleatórias. Ele também formalizou o conceito de **esperança matemática**, essencial para a análise de risco e a tomada de decisões em situações de incerteza.

No século XIX, **Pierre-Simon Laplace** sistematizou a teoria da probabilidade em sua obra *Théorie Analytique des Probabilités* (1812), onde ele introduziu a *regra de Bayes*, expandindo a aplicação da probabilidade para áreas como astronomia e ciências sociais. Sua abordagem permitiu que a probabilidade fosse utilizada para fazer inferências sobre eventos desconhecidos com base em informações prévias.

A evolução da teoria da probabilidade levou, no século XX, à sua formalização por meio da *teoria da medida*. Esta teoria, desenvolvida por matemáticos como **Andrey Kolmogorov** na década de 1930, deu à probabilidade

um arcabouço rigoroso dentro da matemática, utilizando conceitos de medida para definir a probabilidade como uma função que mapeia eventos (subconjuntos de um espaço amostral) para valores numéricos entre 0 e 1. Esse formalismo ficou conhecido como *modelo axiomático da probabilidade*.

Esses axiomas são a base para o desenvolvimento de modelos probabilísticos consistentes e robustos, que hoje são amplamente utilizados em áreas como finanças, física, estatística e inteligência artificial.



Figure 4.1: Astralagus (um dos ossos que compõem o calcanhar, usado no Egito antigo como um dado rudimentar)

4.2 Conceitos essenciais

4.2.1 Experimentos determinísticos e experimentos aleatórios

Aleatório provém do latim: *aleatorium*: fato cujo desfecho depende de um acontecimento futuro e incerto, resultado da sorte ou acaso, accidental.

Probabilidade deriva do latim: *probabilitas*: qualidade do que se pode comprovar, de *probabilis*: o que pode passar por um teste, provável e de *probare*: provar, testar, examinar.

Ao contrário de um **experimento determinístico**, cujo resultado pode ser previamente determinado como :

- como a reação de dois átomos de hidrogênio com um átomo de oxigênio: $2H_2 + O_2 \rightarrow 2H_2O$ e
- a distância S percorrida no vácuo sob velocidade constante V e sem atrito num intervalo de tempo t :

$$S = V \times t$$

o conceito de experimento aleatório é o que estabelece que seu resultado **não pode ser previsto com certeza** como em:

- o lançamento de um dado. O resultado pode ser qualquer número inteiro de 1 a 6 e
- a medição da altura de uma pessoa selecionada aleatoriamente.

Os resultados observados **apresentam variações** mesmo quando esses experimentos são repetidos indefinidamente e sob as mesmas condições; todavia, é possível estabelecer um conjunto cujos elementos compõem todos os possíveis resultados:

- qualquer número inteiro de 1 a 6: e o conjunto de possíveis resultados é finito e
- qualquer valor em um intervalo contínuo por exemplo, entre 1,50 m e 2,00 m (com infinitas possibilidades dentro desse intervalo).

4.2.2 O espaço amostral

A primeira coisa que fazemos quando começamos a pensar sobre a probabilidade de ocorrência de um certo resultado em um *experimento aleatório* é tentar listar todos os resultados com *possibilidade de ocorrência*.

Esses resultados formam um conjunto a que denominamos de *espaço amostral* que, usualmente, é representado pela letra grega maiúscula Ω .

Para que Ω seja considerado o *espaço amostral* desse experimento aleatório ele precisa apresentar duas propriedades:

1. *apenas um* de seus elementos ocorre cada vez que realizamos o *experimento aleatório*; e,
2. *pelo menos um* dos possíveis resultados ocorre sempre que realizarmos o *experimento aleatório*.

Essas condições indicam que os elementos de Ω são *mutuamente exclusivos* e *exaustivos*.

4.2.2.1 Espaços aleatórios discretos

- são finitos ou contáveis (infinito numerável)
- pode-se atribuir uma probabilidade para cada resultado

Exemplo 1:

- Experimento aleatório: lançar um dado e contar o número de pontos na face que ficar exposta para cima
- Espaço amostral finito: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Exemplo 2:

- Experimento aleatório: lançar dois dados e contar o número de pontos nas faces que ficarem expostas para cima
- Espaço amostral finito: $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

Exemplo 3:

- Experimento aleatório: lançar uma moeda e contar o número de lançamentos necessários até se obter uma “cara”
- Espaço amostral infinito contável: $\Omega = \{1, 2, 3, 4, 5, \dots, k, \dots\}$

Exemplo 4:

- Experimento aleatório: lançar um dado até se obter um “6”
- Espaço amostral infinito contável: $\Omega = \{1, 2, 3, 4, 5, \dots, k, \dots\}$

Um espaço amostral consiste então da *enumeração* (finita ou infinita contável) de todos os *possíveis resultados* de serem obtidos em um experimento aleatório.

Cada um dos possíveis resultados de um experimento aleatório é chamado de um *elemento* desse espaço amostral. Assim, para o espaço amostral Ω , seus elementos serão representados por letras gregas minúsculas ω_n

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots, \omega_n, \dots\}$$

4.2.2.2 “Espaços” aleatórios contínuos

- não são contáveis: os resultados possíveis formam um intervalo contínuo de valores
- probabilidade de um resultado específico é zero: como existem infinitos resultados possíveis, a probabilidade de um valor específico é 0
- probabilidades são atribuídas a intervalos: a probabilidade de um evento é associada à extensão do intervalo (comprimento, área, volume, etc.) que o evento ocupa.

Exemplo 1:

- Experimento aleatório: a altura de uma pessoa aleatoriamente sorteada
- Intervalo amostral: $\Omega = [1,5; 2,0]$ m

Exemplo 2:

- Experimento aleatório: o peso de uma pessoa aleatoriamente sorteada
- Intervalo amostral: $\Omega = [10; 100]$ kg

Exemplo 3:

- Experimento aleatório: o teor de um minério por quilo de uma amostra de solo extraída de um local aleatório
- Intervalo amostral: $\Omega = [0,001; 0,01]$ gramas

4.2.2.3 Espaços amostrais equiprováveis e não equiprováveis

Se *todos* os elementos que compõem um espaço amostral finito de um experimento aleatório possuem a *mesma* probabilidade de ocorrência é dito que o *espaço amostral* desse *experimento aleatório* é *equiprovável* (com a mesma probabilidade para todos os seus elementos).

Exemplo 1

- Experimento aleatório: lançar um dado e contar o número de pontos na face que ficar exposta para cima
- Espaço amostral finito: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Probabilidades: $P(1) = \frac{1}{6}, P(2) = \frac{1}{6}, P(3) = \frac{1}{6}, P(4) = \frac{1}{6}, P(5) = \frac{1}{6}, P(6) = \frac{1}{6}$

> Exemplo 2

- Experimento aleatório: lançar dois dados e contar o número de pontos nas faces que ficarem expostas para cima
- Espaço amostral finito: $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
- Probabilidades: $P(2) = \frac{1}{36}, P(3) = \frac{2}{36}, P(4) = \frac{3}{36}, P(5) = \frac{4}{36}, P(6) = \frac{5}{36}, P(7) = \frac{6}{36}, P(8) = \frac{5}{36}, P(9) = \frac{4}{36}, P(10) = \frac{3}{36}, P(11) = \frac{2}{36}, P(12) = \frac{1}{36}$

Cada um dos elementos que compõem o espaço amostral (a soma dos valores numéricos das faces no lançamento de um dado por duas vezes) poderá resultar de diferentes combinações de valores.

A Tabela 4.1 apresenta todas as combinações possíveis de serem obtidas, bem como as proporções em relação ao total para cada elemento do espaço amostral.

Table 4.1: Quadro dos possíveis resultados de um experimento aleatório: somas dos valores numéricos das faces no lançamento de um dado por duas vezes

Soma	Possíveis combinações de resultados nos lançamentos	Frequência n_i	Proporção f_i
2	(primeiro,segundo) (1,1)	1	$\frac{1}{36}$
3	(1,2); (2,1)	2	$\frac{2}{36}$
4	(1,3); (2,2); (3,1)	3	$\frac{3}{36}$
5	(1,4); (2,3); (3,2); (4,1)	4	$\frac{4}{36}$
6	(1,5); (2,4); (3,3); (4,2); (5,1)	5	$\frac{5}{36}$

Soma	Possíveis combinações de resultados nos lançamentos	Frequência n_i	Proporção f_i
7	(1,6); (2,5); (3,4); (4,3); (5,2); (6,1)	6	$\frac{6}{36}$
8	(2,6); (3,5); (4,4); (5,3); (6,2)	5	$\frac{5}{36}$
9	(3,6); (4,5); (5,4); (6,3)	4	$\frac{4}{36}$
10	(4,6); (5,5); (6,4)	3	$\frac{3}{36}$
11	(5,6); (6, 5)	2	$\frac{2}{36}$
12	(6,6)	1	$\frac{1}{36}$
Totais		36	1

As probabilidades de ocorrência de cada um os elementos desse espaço amostral são diferentes e, por essa razão é dito que o *espaço amostral* desse *experimento aleatório* tem elementos *não equiprováveis*.

4.2.3 Evento

Define-se como *evento de interesse* um *subconjunto finito* do espaço amostral, composto por *um ou mais* de seus elementos que *satisfazem* o enunciado estabelecido no experimento aleatório proposto.

A expressão *evento de interesse* (também chamado de *sucesso*) refere-se, no contexto do cálculo de probabilidades, à ocorrência de um resultado *desejado* durante a realização de um experimento aleatório.

Frequentemente, eventos de interesse são representados por letras maiúsculas do alfabeto romano e podem ser acompanhados de uma notação explicativa, como $E(\dots)$.

Por exemplo, considere um experimento aleatório que consiste em lançar um dado uma única vez. Um possível evento de interesse pode ser: E(obtenção do número 2) ($E(2)$) e, nesse contexto pode ser a obtenção do número 2 como resultado.

Podemos ter variados tipos de *eventos de interesse* como:

1. *simples* ou *compostos*;
2. *certos* ou *impossíveis*;
3. *dependentes* ou *independentes* ;
4. *mutuamente exclusivos* ;
5. *complementares*;

4.2.3.1 Diagramas de Venn para representar o espaço amostral e eventos de interesse

Em muitos dos problemas de probabilidade, o *evento de interesse* pode se definido como *associações* de *dois ou mais* eventos formados, por sua vez, por um ou mais elementos do espaço amostral do experimento aleatório. Uniões, interseções e complementos são algumas dessas associações que, doravante, serão muito utilizados.

Por essa razão, a representação do espaço amostral e esses eventos por meio de Diagramas de Venn pode ajudar a compreensão de um problema de cálculo probabilístico.

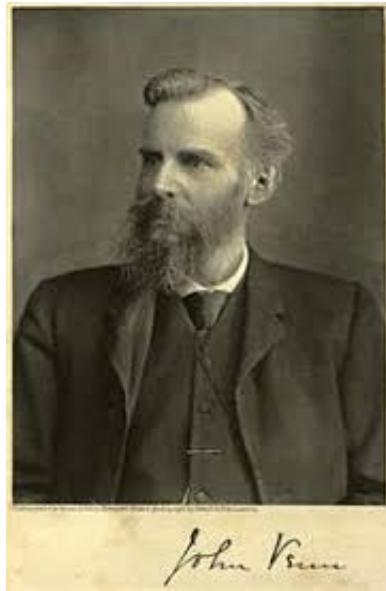
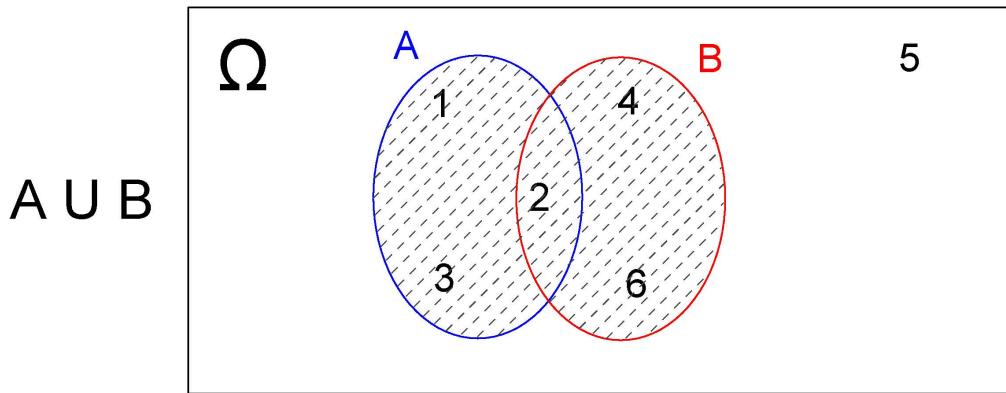


Figure 4.2: John Venn, 1834–1923

4.2.3.1.1 União $A \cup B$ Sejam A e B dois *eventos de interesse* definidos sobre o *espaço amostral* $\Omega = \{1, 2, 3, 4, 5, 6\}$ (lançamento de um dado) tais que $A = \{1, 2, 3\}$ e $B = \{2, 4, 6\}$.

Um *evento de interesse* E expresso como a *união* desses dois outros, representado por $E = (A \cup B)$, será o subconjunto do espaço amostral Ω que contém os elementos que pertençam a **A , ou a B ou a ambos**.

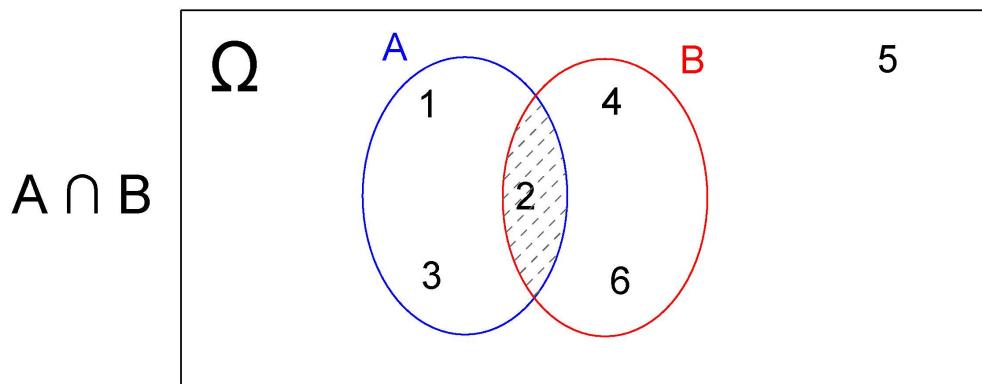
Desse modo, $E = A \cup B = \{1, 2, 3, 4, 6\}$ e o Diagrama de Venn correspondente será:

Figure 4.3: União: $A \cup B$

Na realização desse *experimento aleatório* (lançar um dado) o *evento de interesse* E ocorrerá quando qualquer um dos resultados for um elemento pertencente a A , ou a B ou a *ambos*.

4.2.3.1.2 Interseção $A \cap B$ Um *evento de interesse* E definido como a *interseção* dos eventos A e B anteriormente definidos, representado por $E = (A \cap B)$, será o subconjunto do espaço amostral Ω que contém todos os elementos que pertençam a **ambos os eventos A e B simultaneamente**.

Desse modo, $E = (A \cap B) = \{2\}$ e o Diagrama de Venn correspondente será:

Figure 4.4: Interseção: $A \cap B$

Na realização desse *experimento aleatório* (lançar um dado) o *evento de interesse E* ocorrerá apenas quando o resultado for um elemento simultaneamente pertencente a A e B .

Quando o evento de interesse é definido pela interseção de dois outros, todavia essa interseção é vazia, representa-se E como

$$E(A \cap B) = \emptyset$$

4.2.3.1.3 Complemento A^c Um *evento de interesse* pode também ser definido como o *complemento* de outros como, por exemplo, de A , sendo representado por $E = (A^c)$ (ou $E = (\bar{A})$).

Desse modo, $E = (A^c) = \{4, 5, 6\}$ e o Diagrama de Venn correspondente será:

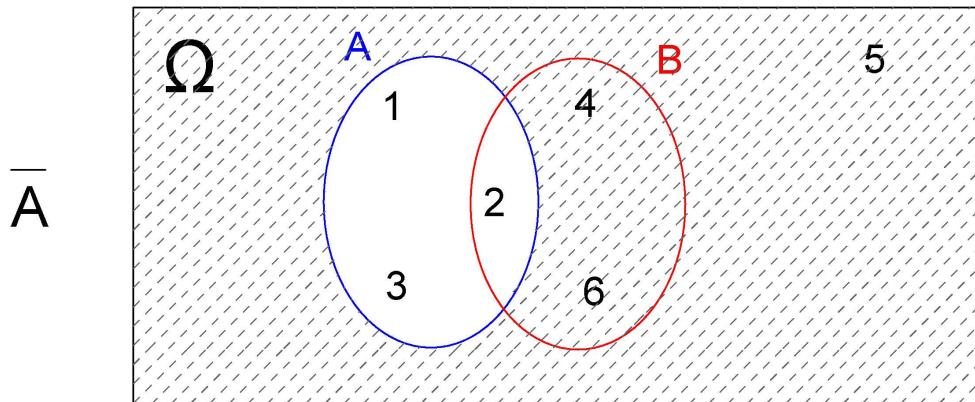


Figure 4.5: Complementar A^c

De modo análogo, para $E = (B^c) = \{1, 3, 5\}$ e o Diagrama de Venn correspondente será :

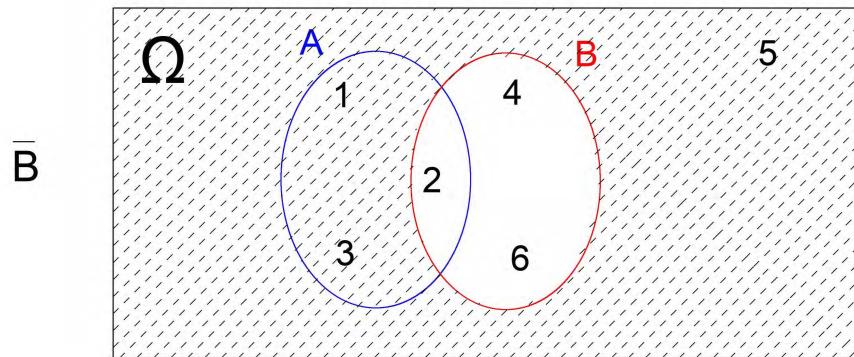
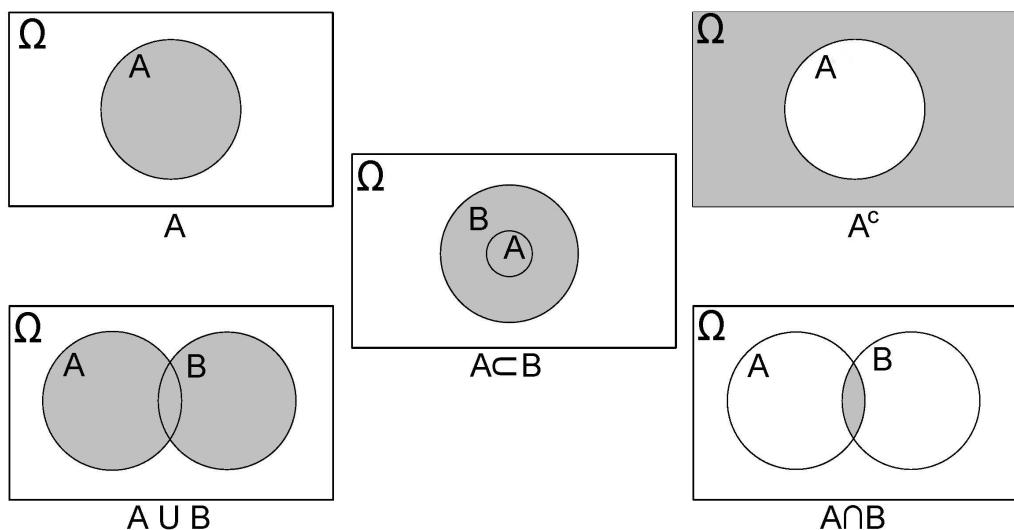


Figure 4.6: Complementar de B



DIAGRAMAS DE VENN

Figure 4.7: Diagramas de Venn

4.2.3.2 Eventos simples e eventos compostos

O evento de interesse ($E(2)$) definido no experimento aleatório anterior (obter o número 2) é formado por apenas um elemento do espaço amostral. Eventos formados por apenas um elemento do espaço amostral são denominados de *evento simples*.

$$\Omega = \{1; 2; 3; 4; 5; 6\} E(2) = \{2\}$$

Admita agora o mesmo *experimento aleatório* todavia definindo como *evento de interesse* (obter-se um número par). Um *evento de interesse* assim definido é um evento composto uma vez que é formado por mais de um elemento do espaço amostral:

$$\Omega = \{1; 2; 3; 4; 5; 6\} E(\text{par}) = \{2; 4; 6\}$$

Outro exemplo, a partir de um *experimento aleatório* que consiste em se lançar uma moeda *duas* vezes, cujo *espaço amostral* é representado por um conjunto composto por *quatro* elementos

$$\Omega = \{(\text{Cara}, \text{Coroa}), (\text{Coroa}, \text{Cara}), (\text{Cara}, \text{Cara}), (\text{Coroa}, \text{Coroa})\}$$

Se definirmos como *evento de interesse* na realização desse experimento aleatório obter-se $E = \{(\text{Cara}, \text{Cara})\}$, o evento E será um *evento simples* pois é formado por apenas *um* elemento do espaço amostral.

Se, por outro lado, definimos como *sucesso* obter-se $E_1 = \{(\text{Cara}, \text{Coroa}) \text{ ou } (\text{Coroa}, \text{Cara})\}$, o evento E_1 será um *evento composto* pois é formado por *dois* elementos do espaço amostral.

Se codificarmos *Cara=1* e *Coroa=0*, podemos representar num plano XY o espaço amostral Ω desse experimento aleatório e o *evento de sucesso* E_1

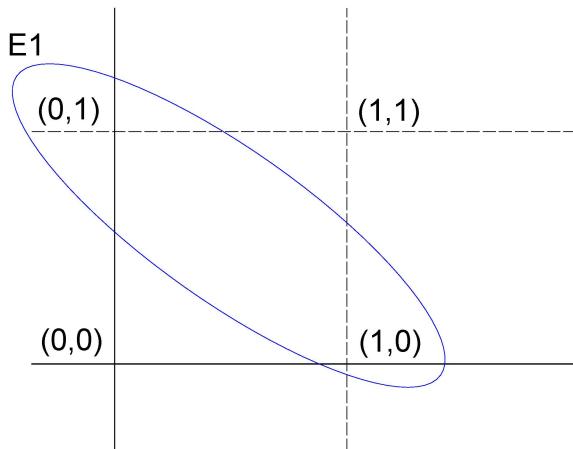


Figure 4.8: Representação gráfica do espaço amostral do experimento aleatório e do evento de interesse definido

4.2.3.3 Eventos certos e eventos impossíveis

Um evento de interesse G , definido sobre o espaço amostral Ω , em que $G = \Omega$, expressa que qualquer elemento de Ω satisfaz o evento G , ou seja, qualquer um dos possíveis resultados do experimento aleatório corresponde ao evento.

Um evento de interesse assim definido ocorrerá com certeza, razão pela qual tais eventos são denominados eventos certos.

Por outro lado, se definirmos um evento de interesse I que não contém resultados pertencentes a Ω — o espaço amostral, ou seja, todos os resultados possíveis — como, por exemplo, obter o número 7 no lançamento de um dado de seis faces, esse evento será impossível de ocorrer.

Eventos assim definidos são chamados de eventos impossíveis.

4.2.3.4 Eventos independentes

Dois eventos são considerados *independentes* quando a probabilidade de ocorrência de um evento de interesse em um determinado experimento aleatório não é influenciada pelo *resultado prévio* de outro evento.

Em outras palavras, a ocorrência de um evento não altera a probabilidade do outro. Caso contrário, esses eventos são classificados como dependentes ou condicionados.

Este conceito será explorado em maior detalhe em seções posteriores.

4.2.3.5 Eventos mutuamente exclusivos

Dois eventos que *nunca* poderão ocorrer simultaneamente são ditos *mutuamente exclusivos*. No experimento do lançamento da moeda por uma vez, nunca observaremos, simultaneamente, dois eventos como $E = \{(Cara)\}$ e $F = \{(Coroa)\}$.

Um evento assim definido teria sua interseção vazia

$$G = (E \cap F) = \emptyset$$

e, por essa razão, sua probabilidade será $P(G) = P(E \cap F) = 0$.

4.2.3.6 Eventos complementares

Definido um *evento de interesse* qualquer pode-se observar apenas dois resultados:

1. *ocorrer*;
2. *não ocorrer* o sucesso.

Ou seja, um ou outro deverá forçosamente ocorrer.

Chama-se de *evento complementar* (E^c ou \bar{E}) a um evento (E) e sua probabilidade de sucesso será:

$$P(E^c) = 1 - P(E)$$

Se a probabilidade de sucesso de que ele ocorra for $P(E) = p$ e a de que ele não ocorra for $P(E^c) = q$ vê-se que a soma dessas quantidades deverá ser $p + q = 1$, novamente antecipando um dos postulados do conceito axiomático de probabilidade.

4.2.4 Probabilidade

4.2.4.1 Conceito clássico ou *a priori*

Sob uma visão intuitiva, a probabilidade como uma medida da informação que temos sobre a possibilidade de ocorrência de um evento aleatório, pode ser definida como a medida numérica expressa em termos relativos (percentuais), obtida pela razão (proporção) entre o número de eventos favoráveis (sucessos) pelo número total de eventos prováveis no experimento (espaço amostral).

Esse conceito de probabilidade é denominado *clássico* ou *a priori*, baseado em um conhecimento prévio ou uma crença subjetiva sobre a probabilidade de um evento ocorrer.

Por exemplo, um jogador de cartas pode ter uma crença a priori de que a probabilidade de uma carta ser um ás é de 1 em 13, independentemente do número de baralhos no jogo

A distribuição de frequências é um instrumento importante para a análise da variabilidade de experimentos aleatórios e, em particular, as frequências relativas são estimativas das probabilidades.

$$P(E) = \frac{\text{número de resultados de interesse (sucessos)}}{\text{número total de resultados possíveis no espaço amostral}}$$

Com o estabelecimento de suposições adequadas, um modelo teórico de probabilidade pode ser empregado sem a realização *a priori* do experimento aleatório, reproduzindo de modo razoável a distribuição das frequências quando o experimento é realizado.

Consideremos o exemplo do experimento que consiste em se lançar um dado e observar o valor numérico de sua face. As suposições que deveriam ser estabelecidas *a priori* são:

- só pode ocorrer uma das seis faces; e,
- o dado utilizado não possui viés algum (não favorece face alguma).

Como todos os N resultados do espaço amostral apresentam uma **mesma probabilidade** de ocorrência, então a proporção teórica de ocorrência de qualquer um desse resultados poderá ser apresentado na forma vista na forma vista na Tabela 4.2.

$$P(E) = \frac{1}{N}$$

Table 4.2: Distribuição das proporções teóricas do um experimento aleatório: lançamento de um dado

Face	1	2	3	4	5	6	Total
Proporção teórica	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Sendo equiprováveis todos os elementos do espaço amostral, todos terão a mesma probabilidade de ocorrência que será:

$$\begin{aligned} P(E) &= \frac{1}{N} \\ &= \frac{1}{6} \\ &= \frac{1}{6} \end{aligned}$$

Por essa razão sabe-se, *a priori* a probabilidade de ocorrência de qualquer evento ao se realizar esse tipo de experimento aleatório uma única vez.

4.2.4.2 Conceito frequentista ou *a posteriori*

Todavia, se realizarmos o experimento aleatório anterior apenas algumas, tal regularidade poderá não ser comprovada: as frequências observadas (as quantidades obtidas para cada um dos valores numéricos das faces) apresentarão uma **grande irregularidade** diferindo das frequências teóricas definidas.

Observa-se que os resultados das frequências observadas irá se estabilizar, aproximando-se das frequências teóricas, à medida que se repete esse experimento um número suficientemente grande de vezes.

A definição frequencial (*a posteriori*):

1- refere-se à probabilidade empírica observada *a posteriori*; 2- tem por objetivo estabelecer um modelo adequado à interpretação de alguns tipos de experimentos aleatórios; e, 3- é a base para se formular um modelo teórico de distribuição de probabilidades como os que serão abordados mais adiante.

Ao se repetir o experimento aleatório um grande número de vezes (n tendendo a infinitas vezes), a quantidade de vezes que um determinado resultado foi verificado dividida por o número de repetições realizadas (n) irá se aproximar de sua proporção teórica. É o que se denomina como *regularidade estatística dos resultados* por essa propriedade não mais se necessita que os eventos sejam *equiprováveis*.

Formalmente conhecida como Lei Fraca dos Grandes Números (um dos pilares da teoria da probabilidade, foi formalizada pelo matemático suíço Jakob Bernoulli em 1713) e estabelece uma convergência para a probabilidade: à medida que o número de ensaios independentes de um experimento aleatório aumenta, a frequência absoluta dos resultados observados tende a se aproximar da probabilidade teórica

$$P(E) = \lim_{n \rightarrow \infty} \frac{F(E)}{n}$$

onde:

- $P(E)$ é a probabilidade de ocorrência do evento E ;
- $F(E)$ é a frequência observada do evento E (o número de vezes que ele ocorre em n repetições); e,
- n é o número de repetições do experimento.

Jakob Bernoulli in 1713

4.2.4.2.1 Simulações

As simulações desempenham um papel fundamental no entendimento prático dos conceitos probabilísticos, permitindo a reprodução de experimentos aleatórios em larga escala.

Por meio de simulações, podemos verificar empiricamente a convergência das frequências observadas para as frequências teóricas discutidas nos conceitos anteriores.

Elas fornecem uma ferramenta poderosa para ilustrar a regularidade estatística dos resultados, especialmente em situações em que realizar o experimento real seria impraticável ou custoso.

Ao simular o lançamento de um dado, por exemplo, é possível observar como a frequência relativa de qualquer face começa a se aproximar da probabilidade teórica ($P(\cdot) = \frac{1}{6}$) à medida que aumentamos o número de repetições.

```
# Função para lançar o dado n vezes e calcular a frequência de uma face específica

lancar_dado <- function(n, face_escolhida) {
  # Definindo as faces do dado
  faces <- 1:6

  # Realizando n lançamentos
  lancamentos <- sample(faces, n, replace = TRUE)

  # Calculando a frequência observada da face escolhida
  frequencia <- sum(lancamentos == face_escolhida) / n * 100

  # Exibindo a frequência em percentual
  cat("A frequência observada da face", face_escolhida, "foi de", frequencia, "%\n")
}

lancar_dado(10, 3)

## A frequência observada da face 3 foi de 10 %

lancar_dado(10000, 3)

## A frequência observada da face 3 foi de 16.73 %
```

Ao simular o lançamento de um dado, por exemplo, é possível observar como as frequências relativas de todas as faces começam a se aproximar das probabilidades teóricas ($P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$) à medida que aumentamos o número de repetições.

```
library(ggplot2)

lanca_dado <- function(numero_de_lancamentos) {
  # Gere os lançamentos do dado
  lancamentos <- sample(1:6, numero_de_lancamentos, replace = TRUE)

  # Crie um data frame com os resultados
  dados <- data.frame(Face = lancamentos)
```

```

# Contagem das ocorrências de cada face
contagem <- table(dados$Face)

# Crie um gráfico de barras com o número de lançamentos no título
grafico <- ggplot(data = data.frame(Face = names(contagem), Contagem =
  as.numeric(contagem)),
  aes(x = Face, y = Contagem)) +
  geom_bar(stat = "identity") +
  labs(x = "Face do Dado", y = "Contagem") +
  ggtitle(paste("Lançamento de um Dado por:", numero_de_lancamentos, "vezes")) +
  theme_minimal()

# Exiba o gráfico
print(grafico)
}

```

```
lanca_dado(10)
```

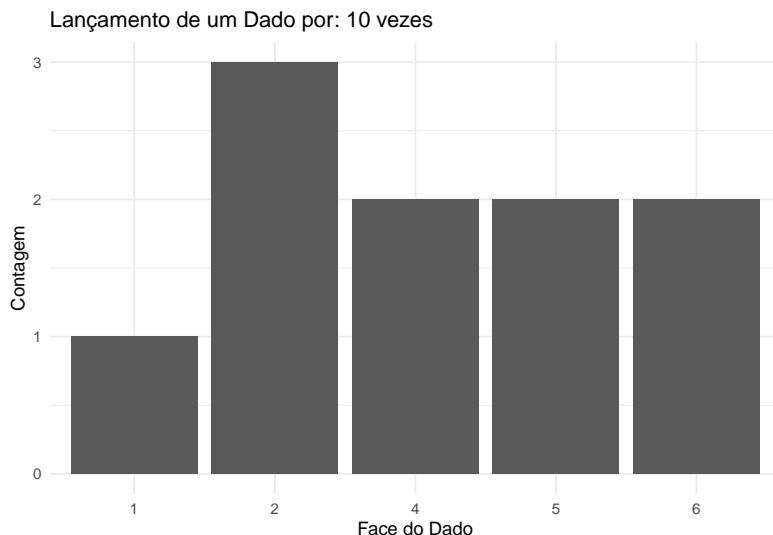


Figure 4.9: Histograma das frequências observadas em 10 lançamentos de um dado justo, evidenciando a variabilidade significativa nas frequências relativas, mesmo que todos os resultados sejam igualmente prováveis.

```
lanca_dado(10000)
```

Ao simular o lançamento de dois dados, por exemplo, é possível observar como as frequências relativas de todas as possíveis somas das faces começam a se aproximar das probabilidades teóricas:

$$P(2) = P(12) = \frac{1}{36}, P(3) = P(11) = \frac{2}{36}, P(4) = P(10) = \frac{3}{36}, P(5) = P(9) = \frac{4}{36}, P(6) = P(8) = \frac{5}{36}, P(7) = \frac{6}{36}$$

à medida que aumentamos o número de repetições.

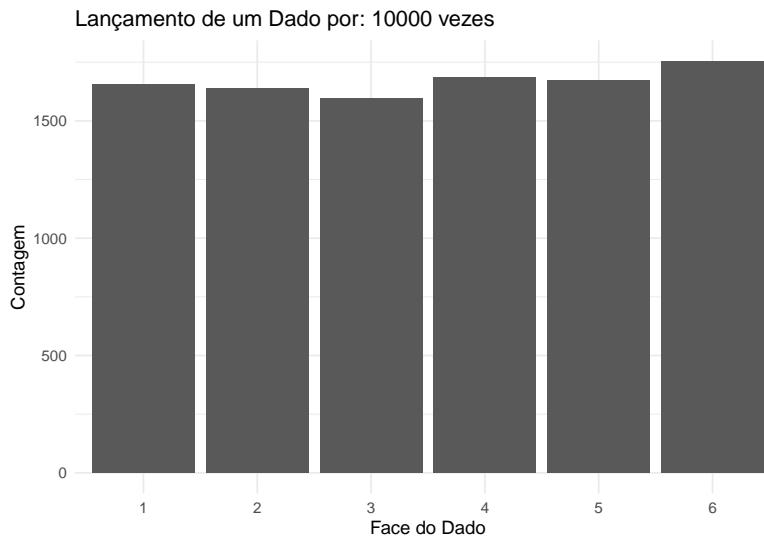


Figure 4.10: Histograma das frequências observadas em 10.000 lançamentos de um dado justo, ilustrando a convergência assintótica das frequências relativas de cada resultado para sua probabilidade teórica, considerando que todos os resultados são igualmente prováveis.

```

library(ggplot2)

lanca_dois_dados <- function(numero_de_lancamentos) {
  # Gere os lançamentos dos dois dados
  dado1 <- sample(1:6, numero_de_lancamentos, replace = TRUE)
  dado2 <- sample(1:6, numero_de_lancamentos, replace = TRUE)

  # Calcule a soma dos dois dados
  somas <- dado1 + dado2

  # Crie um data frame com os resultados
  dados <- data.frame(Soma = somas)

  # Contagem das ocorrências de cada soma
  contagem <- table(dados$Soma)

  # Crie um data frame com a proporção de cada soma
  dados_grafico <- data.frame(
    Soma = as.numeric(names(contagem)),
    Contagem = as.numeric(contagem),
    Proporcao = as.numeric(contagem) / numero_de_lancamentos
  )

  # Probabilidades teóricas de cada soma
  prob_teoricas <- c(1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36)
  somas_teoricas <- 2:12
  prob_teoricas_formatadas <- paste0("(", round(prob_teoricas * 100, 2), "%)")

  # Adicione as probabilidades teóricas ao eixo x como rótulos
  labels_eixo_x <- paste(somas_teoricas, prob_teoricas_formatadas)

  # Crie o gráfico de barras com as frequências observadas e as probabilidades teóricas no
  # eixo x
}
```

```

grafico_somas <- ggplot(data = dados_grafico, aes(x = Soma, y = Proporcao)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = scales::percent(Proporcao, accuracy = 0.1)), vjust = -0.5) + # Exibe as proporções acima das barras
  scale_x_continuous(breaks = somas_teoricas, labels = labels_eixo_x) + # Define os rótulos com soma e probabilidade
  scale_y_continuous(labels = scales::percent) + # Formata o eixo y como porcentagem
  labs(x = "Soma dos Dados (Probabilidade Teórica)", y = "Proporção das Observações (%)")
  +
  ggttitle(paste("Lançamento de Dois Dados por:", numero_de_lancamentos, "vezes")) +
  theme_minimal()

# Exiba o gráfico
print(grafico_somas)
}

lanca_dois_dados(10)

```

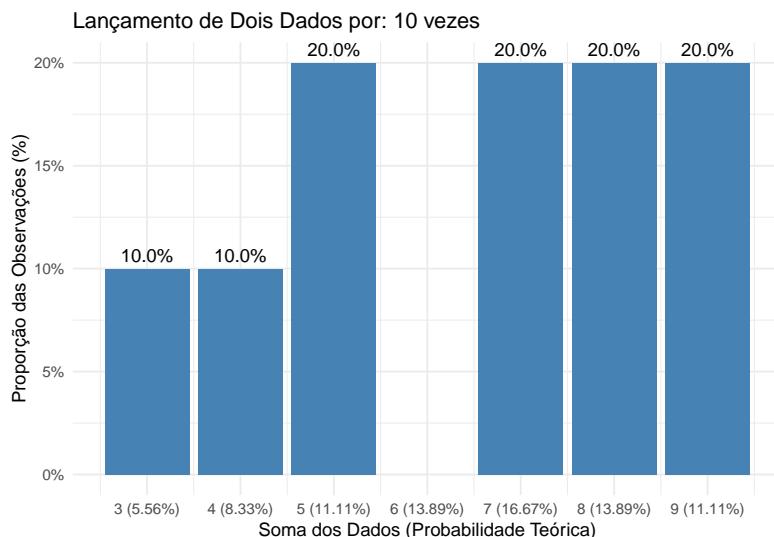


Figure 4.11: Histograma das frequências observadas em 10 lançamentos de dois dados justos, evidenciando a diferença significativa das frequências relativas observadas em relação às probabilidades teóricas de cada resultado possível.

```
lanca_dois_dados(10000)
```

4.2.4.3 Conceito axiomático

Esta abordagem é baseada em um conjunto de axiomas matemáticos que definem as propriedades básicas de probabilidades. A probabilidade é definida como uma função de conjuntos que atribui a cada conjunto de eventos

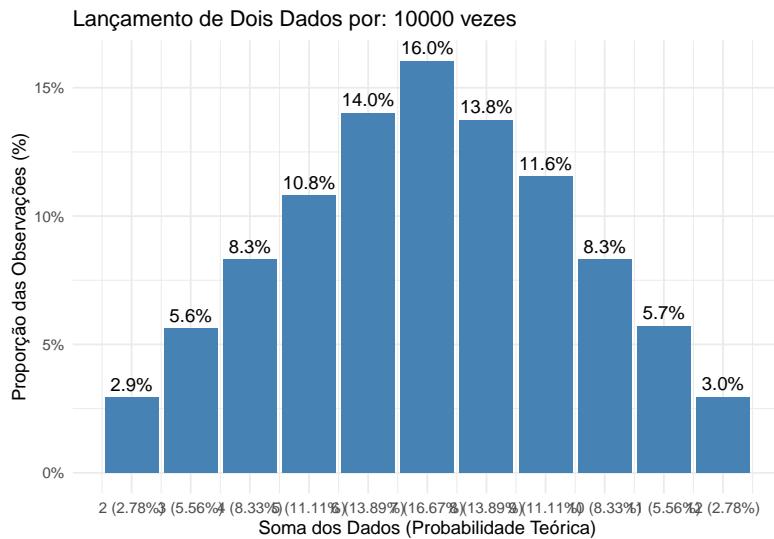


Figure 4.12: Histograma das frequências observadas em 10.000 lançamentos de dois dados justos ilustrando a convergência assintótica das frequências relativas observadas para as probabilidades teóricas de cada resultado possível.

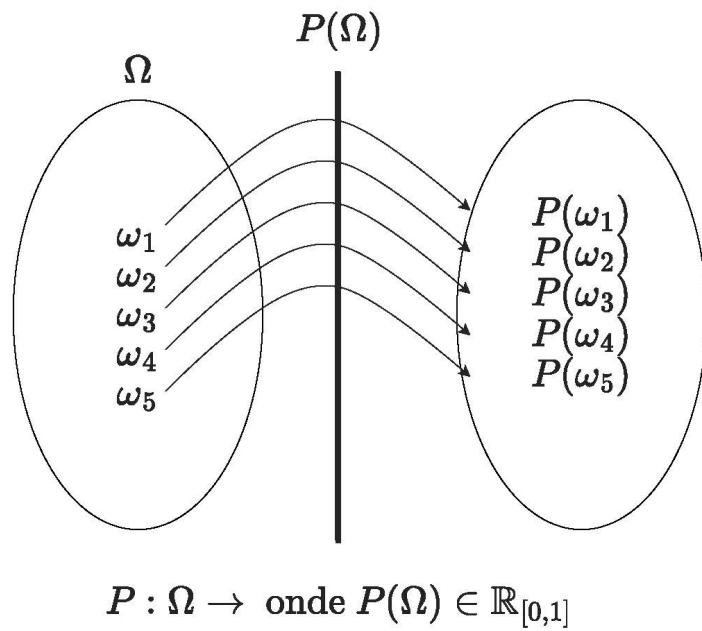
um número entre 0 e 1, satisfazendo os axiomas matemáticos de probabilidade. Essa abordagem permite que as probabilidades sejam definidas formalmente e usadas para cálculos matemáticos.

Um *axioma* é uma premissa considerada necessariamente evidente e verdadeira, fundamento de uma demonstração, porém ela mesma indemonstrável, originada, segundo a tradição racionalista, de princípios inatos da consciência ou, segundo os empiristas, de generalizações da observação empírica.

Admita P uma função que opera sobre o espaço Ω ; isto é, uma função que associa uma quantidade $P(\Omega)$ a cada elemento $\omega \in \Omega$.

Essa função P será uma **função de probabilidade** se, e somente se, satisfizer a **três axiomas** (postulados: conceitos iniciais necessários à construção ou aceitação de uma teoria) estabelecidos por Andrey Kolmogorov (1933).

Kolmogoroff afirmou que uma *Teoria das probabilidades* poderia ser desenvolvida a partir de *axiomas*, da mesma forma que a geometria e a álgebra, e a considerou como caso especial da *Teoria da medida e integração* desenvolvida por Lebesgue, Borel e Fréchet. Ele estabeleceu como postulados as propriedades comuns das noções de probabilidade clássica e frequentista que, desta forma, viraram casos particulares da definição axiomática.



$$P : \Omega \rightarrow \text{ onde } P(\Omega) \in \mathbb{R}_{[0,1]}$$

Figure 4.13: Representação gráfica da função $P(\Omega)$

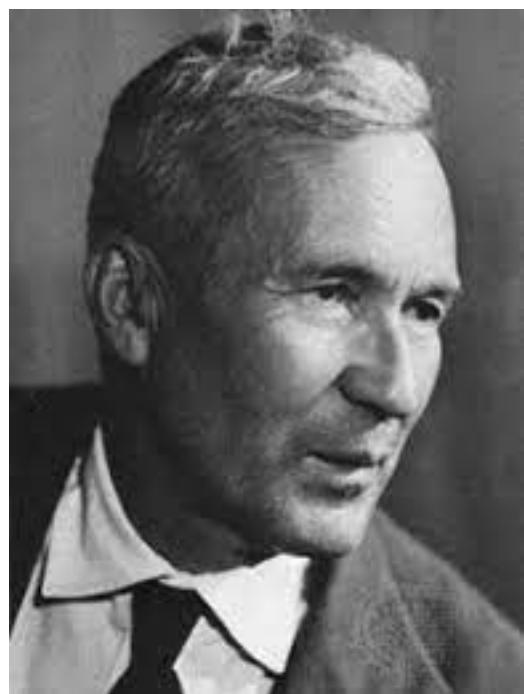


Figure 4.14: Andrey Nikolaevich Kolmogorov (1903-1987)

4.2.4.3.1 Postulado do intervalo A probabilidade de qualquer E é **um número real entre 0 e 1** (pode-se entender isso como uma convenção, onde então se estabelece a medida da probabilidade é um número positivo e que qualquer evento pode ter probabilidade de, no máximo, 1). Esse postulado está plenamente de acordo com a interpretação frequentista de probabilidade.

$$P(E) \geq 0 \text{ (não negatividade e,) mais especificamente, } 0 \leq P(E) \leq 1$$

4.2.4.3.2 Postulado da certeza (normalização) O segundo postulado refere-se à probabilidade do **evento certo** ser igual a 1. No que diz respeito à interpretação frequentista, uma probabilidade de 1 implica que o evento em questão ocorrerá 100% do tempo ou, em outras palavras, **que é certo que ele ocorra** (como, p. exemplo, um experimento aleatório de se lançar dois dados e somar o valor de suas faces o evento certo poderia ser definido como observar um valor menor que 13 ou maior que 2)

$$P(\Omega) = 1$$

4.2.4.3.3 Postulado da aditividade para eventos mutuamente exclusivos (aditividade)

$$P\left(\bigcup_{n=1}^{\infty} \omega_n\right) = \sum_{n=1}^{\infty} P(\omega_n)$$

para qualquer sequência de eventos **mutuamente exclusivos** $\{\omega_1, \omega_2, \omega_3, \dots, \omega_n, \dots\}$ (isto é, tal que $\omega_i \cap \omega_j = \emptyset$ se $i \neq j$)

Tomando o terceiro postulado no caso mais simples, isto é, para **dois** eventos mutuamente exclusivos ω_1 e ω_2 , pode ser facilmente visto que é satisfeita pela interpretação frequentista.

Se um evento ocorrer, digamos, 28% das vezes, outro evento ocorrerá 39%, e os dois eventos não podem ocorrer ao mesmo tempo (ou seja, são mutuamente exclusivos), então um ou outro evento ocorrerão em $28 + 39 = 67\%$ das vezes. Assim, o terceiro postulado é satisfeito, e o mesmo tipo de argumento se aplica quando há mais de dois eventos mutuamente exclusivos.

Recapitulando

- 1- foi definido o conceito de **experimento aleatório** como sendo aquele cujos resultados não podem ser determinados com certeza antes de sua realização;
- 2- foi definido o conceito de **espaço amostral** de um experimento aleatório como sendo o conjunto de **todos os possíveis resultados** que ele pode apresentar;
- 3- foi definido que um **evento de interesse** é um subconjunto do espaço amostral no qual estamos particularmente interessados;
- 4- foi definida uma **função** que tem como domínio o espaço amostral e associa uma quantidade (entre **0** e **1**) a **cada elemento** do espaço amostral; e, por fim,
- 5- estabelecemos que se essa função atende a **três postulados** então ela será uma **medida da probabilidade** de ocorrência de cada evento do espaço amostral em questão.

Assim, quando uma função P associa uma quantidade $P(\Omega)$ a um evento ω e $P(\Omega)$ atende aos três axiomas anteriormente estabelecidos, diz-se que ela é a **função de probabilidade** de Ω .

4.3 Probabilidade da união de eventos

Considerem o espaço amostral de um experimento que consiste no lançamento de um dado honesto: $S = \{1, 2, 3, 4, 5, 6\}$ e admitam alguns eventos constituídos sobre esse espaço amostral, abaixo descritos:

$$E_1 = \{par\}, E_2 = \{mpar\}, E_3 = \{1, 2, 3\}, E_4 = \{4, 5, 6\}, E_5 = \{\geq 4\}, E_6 = \{\leq 5\}.$$

A partir desses eventos podemos propor novos eventos de interesse a partir de *uniões* (conectivo \cup) de dois (ou mais) dos eventos originais como, por exemplo,

$$H_a = E_1 \cup E_2, H_b = E_1 \cup E_4, H_c = E_2 \cup E_3, H_d = E_2 \cup E_5, H_e = E_4 \cup E_6, H_f = E_3 \cup E_5.$$

No experimento aleatório estabelecido:

- lembrando que a *união* de dois conjuntos é o conjunto formado pelos elementos que estão em um, no outro ou em ambos, e
- pensando em probabilidade como a razão do “número de resultados favoráveis” pelo “número de resultados possíveis” (*conceito a priori*)

podemos facilmente verificar que as probabilidades de ocorrência desses eventos são:

$$P(H_a) = 1, P(H_b) = P(H_c) = P(H_d) = P(H_e) = \frac{1}{3}, P(H_f) = 0$$

Considerem agora a Tabela 4.3, de dupla entrada, na qual vemos a distribuição dos alunos de uma escola conforme seu sexo e o curso:

Table 4.3: Distribuição da quantidade de alunos segundo seu sexo e curso escolhido

Curso	Sexo		
	Masculino (M)	Feminino (F)	Total
Matemática pura (M)	70	40	110
Matemática aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Total	115	85	200

Essa tabela nos possibilita calcular a probabilidade de ocorrência de diversos eventos de interesse que desejemos estabelecer.

Exemplo: seja o experimento aleatório de se escolher, aleatoriamente, um estudante qualquer desses quatro cursos. Assim, se definimos nosso evento de interesse M como sendo **M:sexo masculino**, a probabilidade de sucesso (que o indivíduo sorteado aleatoriamente seja do sexo masculino) será:

$$P(M) = \frac{115}{200}$$

Exemplo: se nosso evento de interesse A como sendo **A : curso de matemática aplicada**, a probabilidade de sucesso (que o indivíduo sorteado aleatoriamente seja do curso de matemática aplicada) será:

$$P(A) = \frac{30}{200}$$

A partir dos eventos de interesse anteriormente estabelecidos, podemos definir outros eventos na forma de uniões (\cup) e interseções (\cap):

- uma união entre os dois eventos de interesse anteriores A e M é representada por $A \cup M$ (alternativamente lê-se também **ou**) e representa um evento onde **pelo menos** um dos dois eventos básicos pode ocorrer: **ou A , ou M ou ambos**; e,
- uma interseção dos dois eventos de interesse anteriores A e M é representada por $A \cap M$ (alternativamente lê-se também **e**) e representa um evento onde **os dois eventos** básicos devem ocorrer: A e M .

Exemplo: se definimos nosso evento de interesse ($P(A \cap M)$) como sendo **sexo masculino e cursando matemática aplicada**. Facilmente podemos visualizar na Tabela 4.3 que apenas 15 alunos do curso do evento de interesse (matemática aplicada) são do sexo do segundo evento de interesse (masculino), em relação a todo espaço amostral e assim:

$$P(A \cap M) = \frac{15}{200}$$

Exemplo: consideremos agora o evento de interesse ($P(A \cup M)$) como sendo **sexo masculino ou cursando matemática aplicada**.

Na Tabela 4.3 temos as duas probabilidades **marginais**:

1. $P(A) = \frac{30}{200}$ (curso: matemática aplicada); e, 2- $P(M) = \frac{115}{200}$ (sexo masc).

Poderíamos intuir equivocadamente que:

$$P(A \cup M) = P(A) + P(M) = \frac{30}{200} + \frac{115}{200} = \frac{145}{200}$$

Tal raciocínio é errado pois iria considerar por **duas vezes** os alunos do **sexo masculino**. Uma fração da quantidade global (115) de alunos do **sexo masculino** já considera aqueles que estão matriculados no curso de **matemática aplicada** (15). É preciso **subtrair** da soma das probabilidades marginais essa **parcela em comum** que é a interseção dos dois eventos básicos.

A resposta correta será:

$$P(A \cup M) = P(A) + P(M) - P(A \cap M) = \frac{30}{200} + \frac{115}{200} - \frac{15}{200} = \frac{130}{200}$$

Portanto, para quaisquer eventos de interesse A e B , podemos estabelecer uma **regra geral da probabilidade da união de dois eventos quaisquer** como:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Se A e B forem **mutuamente exclusivos**, a interseção entre eles será vazia ($A \cap B = \emptyset$) e, assim, essa probabilidade é zero. Nessa situação, a probabilidade de $P(A \cup B)$ fica reduzida a uma **regra particular para a adição de probabilidades de eventos mutuamente exclusivos**:

$$P(A \cup B) = P(A) + P(B)$$

Relembrando o que se denomina como *regularidade estatística dos resultados*:

$$P(E) = \lim_{n \rightarrow \infty} \frac{F(E)}{n}$$

```
# Função para simular uma população maior mantendo proporções
simPop <- function(tamanho_populacao) {
  # Proporções conforme a tabela
  prop <- data.frame(
    Curso = c("Matemática pura", "Matemática aplicada", "Estatística", "Computação"),
    Masculino = c(70, 15, 10, 20),
    Feminino = c(40, 15, 20, 10),
    Total = c(110, 30, 30, 30)
  )
}
```

```

)
# Calculando as proporções relativas
prop$propM <- prop$Masculino / prop$Total
prop$propF <- prop$Feminino / prop$Total

# Função para gerar amostra de acordo com as proporções
gerar_amostra <- function(curso, propM, propF, total, tamanho_populacao) {
  n_curso <- round((total / sum(prop$Total)) * tamanho_populacao)
  sexo <- sample(c("M", "F"), n_curso, replace = TRUE, prob = c(propM, propF))
  data.frame(Curso = rep(curso, n_curso), Sexo = sexo)
}

# Gerando a população para cada curso
populacao <- do.call(rbind, lapply(1:nrow(prop), function(i) {
  gerar_amostra(prop$Curso[i], prop$propM[i], prop$propF[i],
                 prop$Total[i], tamanho_populacao)
}))

return(populacao)
}

```

```

# Gerando uma população de 10.000 indivíduos
popSim <- simPop(100000)
table(popSim$Curso, popSim$Sexo)

```

```

##
##          F      M
## Computação     4984 10016
## Estatística    9923  5077
## Matemática aplicada 7527  7473
## Matemática pura 20010 34990

```

```

# Selecionar uma amostra de 200 com reposição (simular a tabela)
ordem1=c("Matemática pura", "Matemática aplicada", "Estatística", "Computação")
ordem2=c("M", "F")
amostPop=popSim[sample(1:nrow(popSim), 200, replace = TRUE), ]
tab=table(factor(amostPop$Curso, levels = ordem1), factor(amostPop$Sexo, levels = ordem2))
tab=addmargins(tab)
colnames(tab)=c("M", "F", "Total")
rownames(tab)=c("Matemática pura (M)", "Matemática aplicada (A)", "Estatística
  (E)", "Computação (C)", "Total")
tab

```

```

##
##          M      F Total
## Matemática pura (M)    76   42   118

```

```

##  Matemática aplicada (A) 11 11 22
##  Estatística (E)         12 15 27
##  Computação (C)         22 11 33
##  Total                  121 79 200

# Calcular a probabilidade de ser do sexo "Masculino" na amostra
pMasc<- mean(amostrPop$Sexo == "M")
pMasc

## [1] 0.605

# Calcular a probabilidade de cursar "Matemática aplicada" na amostra
pMatAp <- mean(amostrPop$Curso == "Matemática aplicada")
pMatAp

## [1] 0.11

# Calcular a probabilidade de cursar "Matemática aplicada" -E- ser do sexo "Masculino" na
# amostra
pMatAp_and_Masc <- mean(amostrPop$Curso == "Matemática aplicada" & amostrPop$Sexo == "M")
pMatAp_and_Masc

## [1] 0.055

# Calcular a probabilidade de cursar "Matemática aplicada" -OU- ser do sexo "Masculino" na
# amostra
pMatAp_or_Masc <- mean(amostrPop$Curso == "Matemática aplicada" | amostrPop$Sexo == "M")
pMatAp_or_Masc

## [1] 0.66

```

Exemplo: Uma população é composta por 20 pessoas que consomem o produto **A**, 30 pessoas que consomem o produto **B** e 50 pessoas que consomem o produto **C**. Um pesquisador de mercado seleciona aleatoriamente uma pessoa desta população. **Sabendo que uma pessoa não consome mais de um produto ao mesmo tempo**, qual a probabilidade de ter sido selecionada uma pessoa que consome os produtos **A ou C**?

Solução:

Definindo os eventos de interesse e as probabilidades associadas:

- 1- E_A = consumidor do produto A: $P(E_A = \frac{20}{100})$;
 2- E_B = consumidor do produto B: $P(E_B = \frac{30}{100})$; e,
 3- E_C = consumidor do produto C: $P(E_C = \frac{50}{100})$.

Pela regra geral da probabilidade da união de dois eventos quaisquer sabemos que:

$$P(E_A \cup E_C) = P(E_A) + P(E_C) - P(E_A \cap E_C)$$

Como foi estabelecido no enunciado que uma pessoa **não** consome mais de um produto ao mesmo tempo (esses eventos são, portanto, **mutuamente exclusivos**: $E_A \cap E_C = \emptyset$) a probabilidade pedida será:

$$\begin{aligned} P(E_A \cup E_C) &= P(E_A) + P(E_C) - P(E_A \cap E_C) \\ &= \frac{20}{100} + \frac{50}{100} - 0 \\ &= \frac{70}{100} \\ &= 0,70 \end{aligned}$$

4.4 Probabilidade de eventos condicionados

Admita dois eventos definidos sobre o experimento aleatório de se sortear uma carta de um baralho:

- $A : \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K\}_{vermelho}$: cartas vermelhas, e
- $B : \{J, Q, K\}_{qualquer}$: cartas de figuras.

A probabilidade de se sortear aleatoriamente uma carta vermelha é $P(A) = 26/52 = 0,5$ e a probabilidade de escolher uma carta de figura é $P(B) = 12/52 = 0,23$.

Se nosso interesse é agora determinar a probabilidade de um evento definido como uma **vermelha e de figura ao mesmo tempo** ou seja, uma carta que está simultaneamente nos conjuntos A e B , estamos então interessados na probabilidade da **interseção** (conectivo \cap) desses dois eventos: $P(A \cap B)$.

A interseção acaba impondo restrições no espaço amostral inicial (todas as 52 cartas do baralho):

- olhando-se por um prisma vemos que o espaço amostral agora é reduzido para A (apenas as 26 cartas vermelhas) e, **dentro desse novo espaço amostral**, estamos interessados nos elementos B (apenas cartas de figura);
- do mesmo modo se olharmos por outro prisma, quando o espaço amostral agora é reduzido para B (apenas as 12 cartas de figuras) e, **dentro desse novo espaço amostral**, estamos interessados nos elementos A (apenas cartas vermelhas);

A probabilidade de um evento A condicionada à ocorrência prévia de um outro evento B , pode ser entendida como a fração de A dentro de B , ou seja:

$$P(\text{ocorreu } A, \text{ ocorrer } B) = \frac{P(A \cap B)}{P(A)} P(\text{foi sorteada uma carta vermelha, sortear-se uma figura}) = \frac{6}{12} = 0,50$$

$$P(\text{ocorreu } B, \text{ ocorrer } A) = \frac{P(B \cap A)}{P(B)} P(\text{foi sorteada uma figura, sortear-se uma carta vermelha}) = \frac{6}{12} = 0,50$$

Reescrevendo-se

$$P(A \cap B) = P(\text{ocorreu } A, \text{ ocorrer } B) \times P(A)P(B \cap A) = P(\text{ocorreu } B, \text{ ocorrer } A) \times P(B)P(A \cap B) = P(B \cap A)$$

Dois eventos A e B definidos sobre um experimento aleatório qualquer são ditos **condicionados** quando a ocorrência prévia de um deles impõe **uma restrição** no espaço amostral do segundo.

A **probabilidade** de um evento qualquer A **condicionada** a um segundo evento B é representada como $P(A|B)$.

A *barra vertical* pode ser “lida” adotando-se termos correlatos que facilitam o entendimento da relação existente, tais como :

- probabilidade de A **posto que** ocorreu B ;

- probabilidade de A admitindo-se que ocorreu B ;
- probabilidade de A considerando-se que ocorreu B ,

A **regra geral da probabilidade de dois eventos condicionados** estabelece que:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

sendo $P(B) > 0$ e $P(A) > 0$ nas expressões acima.

Exemplo: Consideremos a Tabela 4.3 que apresenta informações cruzadas do sexo dos alunos e seus respectivos cursos. Vamos definir os eventos **Fem:sexo feminino** e **Est: cursar estatística**. Como calcular a probabilidade condicionada de nosso evento de interesse $P(\text{Fem}|\text{Est})$ (a probabilidade de um aluno aleatoriamente escolhido ser do sexo **feminino, dado que ele cursa estatística**)?

$$P(\text{Fem}|\text{Est}) = \frac{P(\text{Fem} \cap \text{Est})}{P(\text{Est})}$$

$$= \frac{20}{30} = \frac{2}{3}$$

Esse cálculo é facilmente entendido observando-se as celulas da distribuição de frequências na Tabela 4.3.

Exemplo: Considerem a Tabela 4.4 que relaciona a ida à praia de uma certa pessoa às condições climáticas do dia.

Table 4.4: Condicionamento de passeios à praia em relação às condições climáticas observadas

Dia	1	2	3	4	5	6	7	8	9	10
Foi à praia?	N	S	N	S	S	S	N	N	S	S
Fez sol?	N	S	N	S	N	S	S	N	S	S

Baseado nos dados coletados responda:

- 1- Qual a probabilidade dessa pessoa ir à praia?
- 2- Sabendo-se que fez Sol, qual a probabilidade dessa pessoa ir à praia?
- 3- Os eventos **ir à praia** e **fazer Sol** são independentes ou condicionados?

Da Tabela 4.4 extraímos as seguintes probabilidades:

$$\begin{aligned}
 P(IP) &= \frac{6}{10} = 0,60 \\
 P(FS) &= \frac{6}{10} = 0,60 \\
 P(IP \cap FS) &= \frac{5}{10} \\
 &= 0,50
 \end{aligned}$$

A partir delas podemos calcular a seguinte probabilidade condicionada:

$$\begin{aligned}
 P(IP|FS) &= \frac{P(IP \cap FS)}{P(FS)} \\
 &= \frac{5}{6} \\
 &= 0,83
 \end{aligned}$$

A probabilidade dessa pessoa ir à praia ($P(IP)$) é 0,60; mas quando faz Sol a probabilidade ($P(IP|FS)$) dela aumenta para 0,83.

Assim, os eventos IP e FS são condicionados: essa pessoa vai à praia 60% dos dias analisados; mas, quando faz sol, ela vai em 83% dos dias (a presença de Sol altera a probabilidade dela ir à praia).

Exemplo: Em uma cidade existem 15.000 usuários de telefonia, dos quais 10.000 possuem telefones fixos, 8.000 telefones móveis e 3.000 telefones fixos e móveis. Seja o experimento aleatório de uma operadora de telefone móvel selecionar uma pessoa dessa cidade para oferecer uma promoção do tipo “Fale Grátis de seu Móvel para seu Fixo”.

Responda:

- 1- Sorteando-se aleatoriamente um cliente dessa operadora, se soubermos antecipadamente que ele tem telefone móvel, qual a probabilidade de esse cliente tenha telefone fixo também?
- 2- Sabendo-se que ele tem telefone fixo, qual a probabilidade de ele tenha telefone móvel também?

O espaço amostral de todos esses possíveis eventos pode ser ilustrado pelo diagrama de Venn abaixo:

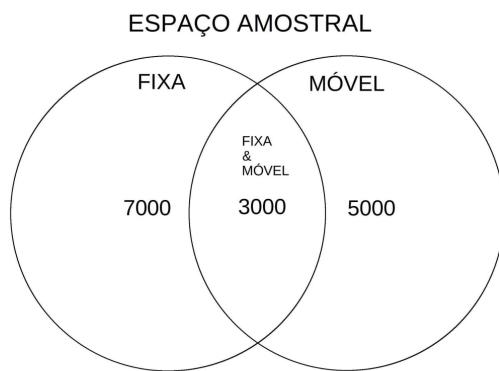


Figure 4.15: Diagrama de Venn do espaço amostral

Do diagrama apresentado na Figura 4.15 podemos extrair imediatamente as probabilidades pedidas:

- $P(F|M)$ (probabilidade de ter uma linha fixa sabendo que possui um telefone móvel); e,
- $P(M|F)$ (probabilidade de ter uma linha móvel sabendo que possui um telefone fixo):

$$\begin{aligned}
 P(F|M) &= \frac{n(MF)}{n(M)} \\
 &= \frac{3000}{8000} \\
 &= 0,375
 \end{aligned}$$

e

$$\begin{aligned}
 P(M|F) &= \frac{n(MF)}{n(F)} \\
 &= \frac{3000}{10000} \\
 &= 0,300
 \end{aligned}$$

Mas também podemos calcular as probabilidades do modo como explicado no começo desta sessão. Definindo-se os eventos F : **telefone fixo** e M : **telefone móvel**, a primeira pergunta pede $P(F|M)$: probabilidade de ter um telefone fixo sabendo que ele tem um telefone móvel:

$$\begin{aligned}
 P(F|M) &= \frac{P(F \cap M)}{P(M)} \\
 &= \frac{\frac{3000}{15000}}{\frac{8000}{15000}} \\
 &= 0,375.
 \end{aligned}$$

A segunda pede $P(M|F)$: probabilidade de ter um telefone móvel sabendo que ele tem um telefone fixo:

$$\begin{aligned}
 P(M|F) &= \frac{P(M \cap F)}{P(F)} \\
 &= \frac{\frac{3000}{15000}}{\frac{10000}{15000}} \\
 &= 0,300
 \end{aligned}$$

Exemplo: Considere a Tabela 4.5 onde são expostos os resultados de uma pesquisa relacionada ao gosto pela prática de tênis entre alunos e alunas. Definindo-se os eventos A : “**gostar de tênis**” e B : “**ser do sexo feminino**”, calcule as probabilidade pedidas ao se sortear, aleatoriamente, uma das pessoas pesquisadas.

- 1- Qual a probabilidade de que goste de tênis ($P(T)$)?
- 2- Qual probabilidade de que não goste de tênis ($P(T^c)$)?
- 3- Qual a probabilidade de que seja do sexo feminino **ou** goste de tênis: ($P(F \cup T)$)?
- 4- Sabendo-se que foi sorteada uma aluna, qual a probabilidade de que goste de tênis ($P(T|F)$)?
- 5- Verifique se os eventos T : “**gostar de tênis**” e F : “**ser do sexo feminino**” são condicionados ou independentes ($P(T \cap F) = P(T) \times P(F)$)

Table 4.5: Distribuição da quantidade de alunos segundo seu sexo e a preferência por tênis

Curso	Sexo			Total
	Masculino (M)	Feminino (F)		
Gostam de tênis (T)	400	200		600
Não gostam de tênis (NT)	50	50		100
Total	450	250		700

Exemplo: Imagine que um jogador está treinando cobranças de pênaltis. Historicamente a probabilidade de acertar uma cobrança, supondo que acertou a anterior é de 60%. Mas, se ele tiver errado a anterior a probabilidade de acertar cai para 30%. Construa a distribuição de probabilidades do número de acertos em 3 tentativas de cobrança.

A seguir vemos a cadeia de eventos necessária para que cada contagem de gols se verifique:

$$\begin{aligned}
 0 \text{ GOL} &= [E_1 \cap E_2 \cap E_3] \\
 1 \text{ GOL} &= \{[A_1 \cap E_2 \cap E_3] \cup [E_1 \cap A_2 \cap E_3] \cup [E_1 \cap E_2 \cap A_3]\} \\
 2 \text{ GOLS} &= \{[A_1 \cap A_2 \cap E_3] \cup [E_1 \cap A_2 \cap A_3] \cup [A_1 \cap E_2 \cap A_3]\} \\
 3 \text{ GOLS} &= [A_1 \cap A_2 \cap A_3]
 \end{aligned}$$

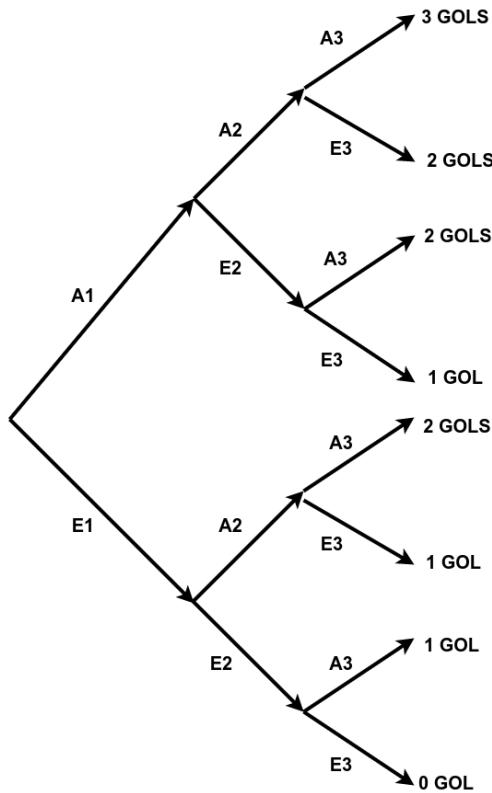


Figure 4.16: Diagrama em árvore das três repetições dependentes de um pênalti

A probabilidade associada a cada contagem de gols será:

$$P(0 \text{ GOL}) = P[E_1 \cap E_2 \cap E_3]$$

$$P(1 \text{ GOL}) = P\{[A_1 \cap E_2 \cap E_3] \cup [E_1 \cap A_2 \cap E_3] \cup [E_1 \cap E_2 \cap A_3]\}$$

$$P(2 \text{ GOLS}) = P\{[A_1 \cap A_2 \cap E_3] \cup [E_1 \cap A_2 \cap A_3] \cup [A_1 \cap E_2 \cap A_3]\}$$

$$P(3 \text{ GOLS}) = P[A_1 \cap A_2 \cap A_3]$$

A partir do enunciado, podemos deduzir as probabilidades de cada um dos eventos:

$$P(A_1) = 0,50$$

$$P(E_1) = 0,50$$

$$P(A_{i+1}|A_i) = 0,60 \text{ logo, pelo complementar, } P(E_{i+1}|A_i) = 0,40$$

$$P(A_{i+1}|E_i) = 0,30 \text{ logo, pelo complementar } P(E_{i+1}|E_i) = 0,70$$

$$\begin{aligned}
 P(0 \text{ GOL}) &= P[E_1 \cap E_2 \cap E_3] \\
 P(0 \text{ GOL}) &= P[E_1] \times P[E_2|E_1] \times P[E_3|E_2] \\
 P(0 \text{ GOL}) &= 0,50 \times 0,70 \times 0,70 \\
 P(0 \text{ GOL}) &= 0,245
 \end{aligned}$$

$$\begin{aligned}
 P(3 \text{ GOLS}) &= P[A_1 \cap A_2 \cap A_3] \\
 P(3 \text{ GOLS}) &= P[A_1] \times P[A_2|A_1] \times P[A_3|A_2] \\
 P(3 \text{ GOLS}) &= 0,50 \times 0,60 \times 0,60 \\
 P(3 \text{ GOLS}) &= 0,18
 \end{aligned}$$

$$\begin{aligned}
 P(1 \text{ GOL}) &= P\{[A_1 \cap E_2 \cap E_3]\} + P\{[E_1 \cap A_2 \cap E_3]\} + P\{[E_1 \cap E_2 \cap A_3]\} \\
 P(1 \text{ GOL}) &= P[A_1] \times P[E_2|A_1] \times P[E_3|E_2] + \\
 &\quad P[E_1] \times P[A_2|E_1] \times P[E_3|A_2] + \\
 &\quad P[E_1] \times P[E_2|E_1] \times P[A_3|E_2] \\
 P(1 \text{ GOL}) &= 0,50 \times 0,40 \times 0,70 + \\
 &\quad 0,50 \times 0,30 \times 0,40 + \\
 &\quad 0,50 \times 0,70 \times 0,30 \\
 P(1 \text{ GOL}) &= 0,305
 \end{aligned}$$

$$\begin{aligned}
 P(2 \text{ GOLS}) &= P\{[A_1 \cap A_2 \cap E_3]\} + P\{[E_1 \cap A_2 \cap A_3]\} + P\{[A_1 \cap E_2 \cap A_3]\} \\
 P(2 \text{ GOLS}) &= P[A_1] \times P[A_2|A_1] \times P[E_3|A_2] + \\
 &\quad P[E_1] \times P[A_2|E_1] \times P[A_3|A_2] + \\
 &\quad P[A_1] \times P[E_2|A_1] \times P[A_3|E_2] \\
 P(2 \text{ GOLS}) &= 0,50 \times 0,60 \times 0,40 + \\
 &\quad 0,50 \times 0,30 \times 0,60 + \\
 &\quad 0,50 \times 0,40 \times 0,30 \\
 P(2 \text{ GOLS}) &= 0,27
 \end{aligned}$$

4.5 Dependência e independência de eventos

Pela regra geral da probabilidade de dois eventos condicionados:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ P(B|A) &= \frac{P(B \cap A)}{P(A)} \end{aligned}$$

Como a probabilidade de interseção não se altera ($P(A \cap B) = P(B \cap A)$), podemos reescrever essas duas expressões:

$$\begin{aligned} P(A \cap B) &= P(A|B) \times P(B) \\ P(A \cap B) &= P(B|A) \times P(A) \end{aligned}$$

com $P(B) > 0$ e $P(A) > 0$ nas expressões acima.

Se os eventos A e B são guardam nenhuma relação de condicionamento eles são chamadas de **eventos independentes**. Equivale dizer que $P(A|B) = P(A)$ (ou $P(B|A) = P(B)$), a probabilidade de A não se altera pela prévia ocorrência de B (ou a de B pelo de A).

Portanto, **dois eventos são denominados independentes se, e somente se:**

$$P(A \cap B) = P(A) \times P(B)$$

Independência e correlação: se duas variáveis aleatórias são **independentes** não há associação de natureza alguma entre elas, **inclusive a linear**, um caso particular de correlação. Todavia uma **correlação linear nula** não implica em **independência** posto existirem várias outras formas outras de relacionamento (quadrática, cúbica, ...).

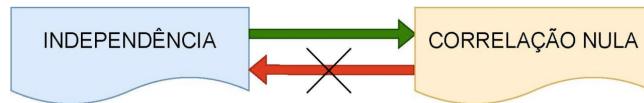


Figure 4.17: Independência implica em ausência de qualquer tipo de associação (a recíproca não se aplica)

4.5.1 Demonstração clássica de independência

Uma bolsa contém 5 bolas **vermelhas** e 5 **azuis**. Nós removemos uma bola aleatória da bolsa, registramos sua cor **e a colocamos de volta na sacola**. Em seguida, removemos outra bola aleatória da bolsa e registramos sua cor.

- Qual é a probabilidade de a primeira bola ser **vermelha** ?
- Qual é a probabilidade de a segunda bola ser **azul**?
- Qual é a probabilidade de a primeira bola ser **vermelha** e a segunda bola **azul**?
- A primeira bola retirada foi uma bola **vermelha** e a segunda bola **azul**; esses eventos foram *independentes* ?

Solução:

Probabilidade em se retirar uma bola **vermelha** em primeiro lugar:

Há 10 bolas das quais 5 são **vermelhas** . A probabilidade de se retirar uma bola **vermelha** será:

$$P(1^{\text{a}} \text{vermelha}) = \frac{5}{10} = \frac{1}{2}$$

Probabilidade em se retirar uma bola **azul** em segundo lugar:

O enunciado do experimento assegura que após a retirada da primeira bola ela é **devolvida** ao sacola; por essa razão, ao se retirar a segunda bola, há novamente 10 bolas no total, das quais 5 são **azuis**. A probabilidade de se retirar uma bola **azul** será:

$$P(2^a azul) = \frac{5}{10} = \frac{1}{2}$$

Probabilidade da primeira bola retirada ser **vermelha** e a segunda ser **azul**:

Ao se retirar duas bolas do sacola há quatro possíveis combinações de resultados. Nós podemos obter:

- 1- uma **vermelha** e depois outra **vermelha**;
- 2- uma **vermelha** e depois uma **azul**;
- 3- uma **azul** e depois uma **vermelha**; ou,
- 4- uma **azul** e depois outra **azul**;

Queremos saber a probabilidade do segundo resultado após termos obtido uma bola **vermelha** na primeira seleção.

Como existem 5 bolas **vermelhas** e 10 bolas no total, existem $\frac{5}{10}$ possibilidades de obter uma bola **vermelha** primeiro.

Agora nós colocamos a primeira bola de volta, então há novamente 5 bolas **vermelhas** e 5 bolas **azuis** na sacola.

Portanto, há $\frac{5}{10}$ possibilidades de obter uma segunda bola **azul** se a primeira bola for **vermelha**.

Isso significa que existem: $\frac{5}{10} \times \frac{5}{10} = \frac{25}{100}$ possibilidades de se obter uma bola **vermelha** em primeiro lugar e uma bola **azul** em segundo.

Então, a probabilidade associada será de $\frac{1}{4}$.

A primeira bola retirada foi uma bola vermelha e a segunda bola azul. Esses dois eventos são independentes?

Esses eventos serão *independentes se, e somente se*:

$$P(A \cap B) = P(A) \times P(B)$$

$$\begin{aligned} P(1^{\text{a}} \text{vermelha}) &= \frac{5}{10} = \frac{1}{2} \\ P(2^{\text{a}} \text{azul}) &= \frac{5}{10} = \frac{1}{2} \\ P(1^{\text{a}} \text{vermelha}, 2^{\text{a}} \text{azul}) &= \frac{25}{100} = \frac{1}{4} \end{aligned}$$

Como $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2}$, os eventos são independentes.

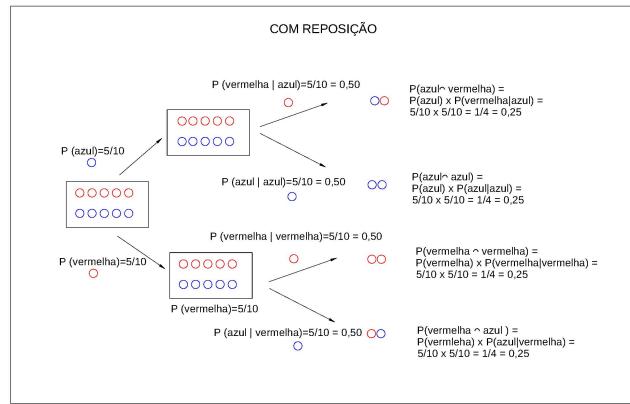


Figure 4.18: Ilustração do experimento aleatório sob a condição de reposição

4.5.2 Demonstração clássica de dependência

E se, ao retirarmos a primeira bola, não a devolvêssemos ao sacola?

Admitamos agora que o enunciado de nosso problema passou a ser:

Uma bolsa contém 5 bolas **vermelhas** e 5 **azuis**. Nós removemos uma bola aleatória da bolsa, registramos sua cor e **não a colocamos de volta na sacola**. Em seguida, removemos outra bola aleatória da bolsa e registramos sua cor.

- 1- Qual é a probabilidade de a primeira bola ser **vermelha** ?
- 2- Qual é a probabilidade de a segunda bola ser **azul**?
- 3- Qual é a probabilidade de a primeira bola ser **vermelha** e a segunda bola **azul**?
- 4- A primeira bola retirada foi uma bola **vermelha** e a segunda bola **azul**; esses eventos foram *independentes* ?

Solução:

1^a Etapa: analisar todos os possíveis resultados

Probabilidade da primeira bola retirada ser **vermelha** e a segunda ser **azul**:

Ao se retirar duas bolas do sacola há quatro possíveis combinações de resultados. Nós podemos obter:

- uma **vermelha** e depois outra **vermelha**;
- uma **vermelha** e depois uma **azul**;
- uma **azul** e depois uma **vermelha** ; ou,
- uma **azul** e depois outra **azul**.

Queremos saber a probabilidade do segundo resultado após termos obtido uma bola **vermelha** na primeira seleção.

Como existem 5 bolas **vermelhas** e 10 bolas no total, existem $\frac{5}{10}$ maneiras de obter uma bola **vermelha** primeiro.

Entretanto, nessa nova situação, nós não colocamos a primeira bola de volta, então haverá apenas 4 bolas **vermelhas** e 5 bolas **azuis** na sacola.

- Haverá $\frac{4}{9}$ maneiras de obter uma segunda bola **vermelha** se a primeira bola for **vermelha**. Isso significa que existem: $\frac{5}{10} \times \frac{4}{9} = \frac{20}{90}$ maneiras de se obter uma bola **vermelha** em primeiro lugar e uma bola **vermelha** em segundo. Então, a probabilidade associada será de $\frac{2}{9}$;

- Haverá $\frac{5}{9}$ maneiras de obter uma segunda bola azul se a primeira bola for vermelha . Isso significa que existem: $\frac{5}{10} \times \frac{5}{9} = \frac{25}{90}$ maneiras de se obter uma bola vermelha em primeiro lugar e uma bola azul em segundo. Então, a probabilidade associada será de $\frac{5}{18}$;
- Haverá $\frac{5}{9}$ maneiras de obter uma segunda bola vermelha se a primeira bola for azul. Isso significa que existem: $\frac{5}{10} \times \frac{5}{9} = \frac{25}{90}$ maneiras de se obter uma bola azul em primeiro lugar e uma bola vermelha em segundo. Então, a probabilidade associada será de $\frac{5}{18}$.
- Haverá $\frac{4}{9}$ maneiras de obter uma segunda bola azul se a primeira bola for azul. Isso significa que existem: $\frac{5}{10} \times \frac{4}{9} = \frac{20}{90}$ maneiras de se obter uma bola azul em primeiro lugar e uma bola azul em segundo. Então, a probabilidade associada será de $\frac{2}{9}$;

Resumo das probabilidades calculadas:

- 1 -uma vermelha e depois outra vermelha : $\frac{2}{9}$;
- 2- uma vermelha e depois uma azul: $\frac{5}{18}$;
- 3- uma azul e depois uma vermelha : $\frac{5}{18}$; e,
- 4- uma azul e depois outra azul: $\frac{2}{9}$.

2^a Etapa: analisar a possibilidade de se obter uma bola vermelha na primeira extração:

- uma vermelha e depois outra vermelha : $\frac{2}{9}$;
- uma vermelha e depois uma azul: $\frac{5}{18}$.

A probabilidade total de se obter uma bola vermelha na primeira extração será:

$$P(1^{\text{a}}\text{vermelha}) = \frac{2}{9} + \frac{5}{18} = \frac{1}{2}$$

3^a Etapa: analisar a possibilidade de se obter uma bola azul na segunda extração:

- uma **vermelha** e depois uma **azul**: $\frac{5}{18}$;
- uma **azul** e depois outra **azul**: $\frac{2}{9}$.

A probabilidade total de se obter uma bola **azul** na segunda extração será:

$$P(2^a \text{azul}) = \frac{5}{18} + \frac{2}{9} = \frac{1}{2}$$

4^a Etapa: analisar a possibilidade de se obter uma bola **vermelha** e em seguida **azul**:

- uma **vermelha** e depois outra **azul**: $\frac{5}{18}$;

5^a Etapa: Esses dois eventos são independentes?

Esses eventos serão *independentes se, e somente se*:

$$P(A \cap B) = P(A) \times P(B)$$

$$\begin{aligned} P(1^a \text{vermelha}) &= \frac{2}{9} + \frac{5}{18} = \frac{1}{2} \\ P(2^a \text{azul}) &= \frac{5}{18} + \frac{2}{9} = \frac{1}{2} \\ P(1^a \text{vermelha}, 2^a \text{azul}) &= \frac{5}{18} \end{aligned}$$

Como $\frac{5}{18} \neq \frac{1}{2} \times \frac{1}{2}$, os eventos **não são independentes**.

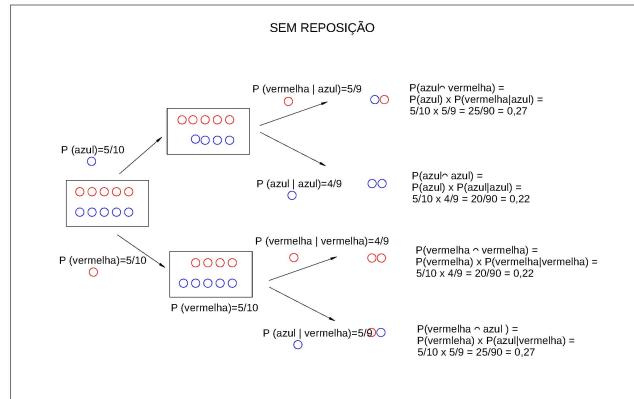


Figure 4.19: Ilustração do experimento aleatório sob a condição de não reposição

4.6 Probabilidade de eventos independentes (regra da cadeia)

Se E_1, E_2, \dots, E_n são eventos independentes entre si, então:

$$P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_1) \times P(E_2) \dots \times P(E_n)$$

Para que isso se verifique, a independência entre cada um e todos os eventos deve se verificar. Numa situação de três eventos, por exemplo, teríamos que observar:

$$P(E_1 \cap E_2) = P(E_1) \times P(E_2)$$

$$P(E_1 \cap E_3) = P(E_1) \times P(E_3)$$

$$P(E_2 \cap E_3) = P(E_2) \times P(E_3)$$

$$P(E_1 \cap E_2 \cap E_3) = P(E_1) \times P(E_2) \times P(E_3)$$

Exemplo: considere o experimento aleatório de se lançar dois dados e obter o valor **1** no primeiro deles e **5** no segundo (defina os eventos E_1 = sair face 1 e E_5 = sair face 5).

Solução:

Quando lançamos dois dados o resultado obtido em um deles (o valor numérico da face) **não condiciona ou altera** o resultado obtido no outro: os resultados são **são independentes**. Desse modo, sendo $P(E_1) = \frac{1}{6}$ e $P(E_5) = \frac{1}{6}$:

$$\begin{aligned} P(E_1 \cap E_5) &= \frac{1}{6} \times \frac{1}{6} \\ &= \frac{1}{36}. \end{aligned}$$

Exemplo: Uma empresa que compra produtos de dois fabricantes diferentes (**Fabricante 1** e **Fabricante 2**) adquiriu 168 unidades do primeiro e 84 do segundo. Sabendo que 8 unidades fabricadas pelo primeiro fornecedor não atenderam às especificações e apenas 4 do segundo, verifique se o fato de uma amostra ter atendido às especificações independe de ter sido produzida pelo **Fabricante 1**.

Solução:

Para a primeira verificação pedida defina os eventos $Fab1$: **ter sido produzida pelo Fabricante 1**, $Aprov$: **ter atendido às especificações** e $Fab2$: **ter sido produzida pelo Fabricante 2**. Na sequência podemos calcular as seguintes probabilidades:

$$\begin{aligned} P(Fab1) &= \frac{168}{252} \\ &= 0,6666 \\ P(Aprov) &= \frac{240}{252} \\ &= 0,9523 \\ P(Fab1 \cap Aprov) &= \frac{160}{252} \\ &= 0,6349 \end{aligned}$$

Se o fato de uma amostra ter sido aprovada **independe** de ter sido produzida pelo Fabricante 1 **então** $P(Aprov|Fab1) = P(Aprov)$:

$$\begin{aligned} P(Aprov|Fab1) &= \frac{P(Aprov \cap Fab1)}{P(Fab1)} \\ &= \frac{0,6349}{0,6666} \\ &= 0,9523. \end{aligned}$$

Como $P(Aprov|Fab1) = P(Aprov)$, verifica-se que o fato de uma amostra aleatoriamente sorteada entre as peças do fabricante 1 não condiciona sua aprovação.

Exemplo: A probabilidade de um consumidor (C_1) ficar satisfeito com o desempenho de certa marca de produto é de 25%. A probabilidade de um outro consumidor (C_2) ficar satisfeito com a mesma marca é de 40%. Admitamos que os dois consumidores irão consumir o produto num mesmo momento e de **forma independente (incomunicáveis)**. Qual a probabilidade de **os dois** consumidores ficarem satisfeitos simultaneamente?

Solução:

As probabilidades individuais dos consumidores 1 e 2 ficarem satisfeitos com o desempenho da marca do produto são:

$$\begin{aligned} P(C_1) &= 0,25 \\ P(C_2) &= 0,40 \end{aligned}$$

A probabilidade de **ambos** ficarem satisfeitos, dado que o enunciado afirma que esses eventos são **independente** será:

$$\begin{aligned} P(C_1 \cap C_2) &= 0,25 \times 0,40 \\ &= 0,10. \end{aligned}$$

4.7 Teorema de Bayes



Figure 4.20: Thomas Bayes (1702 - 1761)

Admita o espaço amostral de um experimento baseado no sorteio aleatório de um estudante de uma escola, com dois possíveis resultados quanto ao sexo:

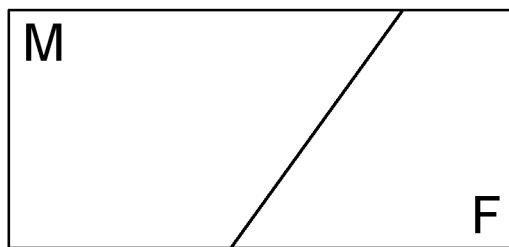


Figure 4.21: Espaço amostral

Considere agora um evento definido nesse espaço amostral como sendo “ter um carro”:

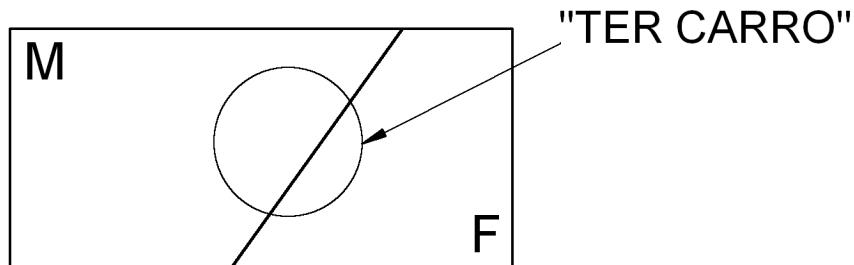


Figure 4.22: Espaço amostral

As interseções desse evento com os elementos do espaço amostral são:

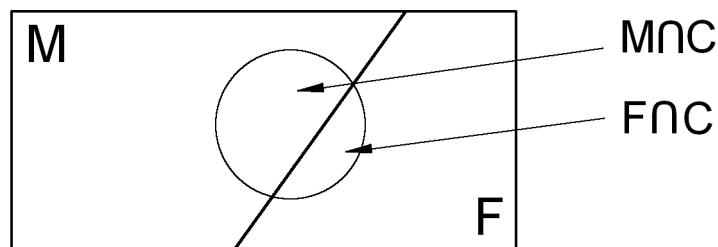


Figure 4.23: Espaço amostral

Pela **regra da probabilidade condicionada** temos que

$$P(C|F) = \frac{P(C \cap F)}{P(F)} P(C \cap F) = P(C|F)P(F)$$

e, de modo equivalente,

$$P(C|M) = \frac{P(C \cap M)}{P(M)} P(C \cap M) = P(C|M)P(M)$$

A probabilidade de se ter um carro é dada pela regra da união de eventos que, nesse caso são disjuntos e assim:

$$P(C) = P(C \cap M) \cup P(C \cap F)P(C) = P(C|M)P(M) + P(C|F)P(F)$$

Sorteado aleatoriamente um estudante da escola verificou-se possuir um carro. Qual a probabilidade de que seja do sexo feminino ($P(F|C)$)?

$$P(F|C) = \frac{P(F \cap C)}{P(C)}P(F \cap C) = P(F|C)P(C)$$

Pela igualdade $P(C \cap F) = P(F \cap C)$:

$$P(C \cap F) = P(C|F)P(F)P(F \cap C) = P(F|C)P(C)$$

substituindo-se na expressão acima chega-se a:

$$\begin{aligned} P(C \cap F) &= P(F \cap C) \\ P(C|F).P(F) &= P(F|C).P(C) \\ P(F|C) &= \frac{P(C|F)P(F)}{P(C)} \end{aligned}$$

uma **relação** entre duas probabilidades inversamente condicionadas conhecida como **Teorema de Bayes**.

Admita então serem dados:

- “M”: ser do sexo masculino: $P(M) = 0,65$;

- “F”: ser do sexo feminino: $P(F) = 0,35$.
- “C”: possuir um carro:
 - $P(C|M) = 0,30$
 - $P(C|F) = 0,18$.

A probabilidade de se ter carro ($P(C)$) resulta de união de dois únicos e possíveis eventos condicionados ao sexo e disjuntos. Assim:

$$\begin{aligned} P(C) &= P(C \cap M) \cup P(C \cap F) \\ P(C) &= [P(M).P(C|M)] \cup [P(F).P(C|F)] \\ P(C) &= [0,65.0,30] + [0,35.0,18] \\ P(C) &= 0,258 \end{aligned}$$

e podemos calcular $P(F|C)$:

$$\begin{aligned} P(F|C) &= \frac{P(F).P(C|F)}{P(C)} \\ P(F|C) &= \frac{0,35.0,18}{0,258} \\ P(F|C) &= 0,2442 \end{aligned}$$

A probabilidade de que um estudante aleatoriamente sorteado nessa escola e sabendo-se **a priori** que possui um carro ser do **sexo feminino** é de 24,42%.

Para um espaço amostral mais amplo, de modo geral consideremos, inicialmente o diagrama da Figura 4.24 onde Ω é o espaço amostral de um experimento aleatório qualquer:

Admita que E_1 , E_2 , E_3 e E_4 formem a partição do espaço amostral Ω (seus elementos são **mutuamente exclusivos**) como exposto na Figura 4.25



Figure 4.24: Espaço amostral

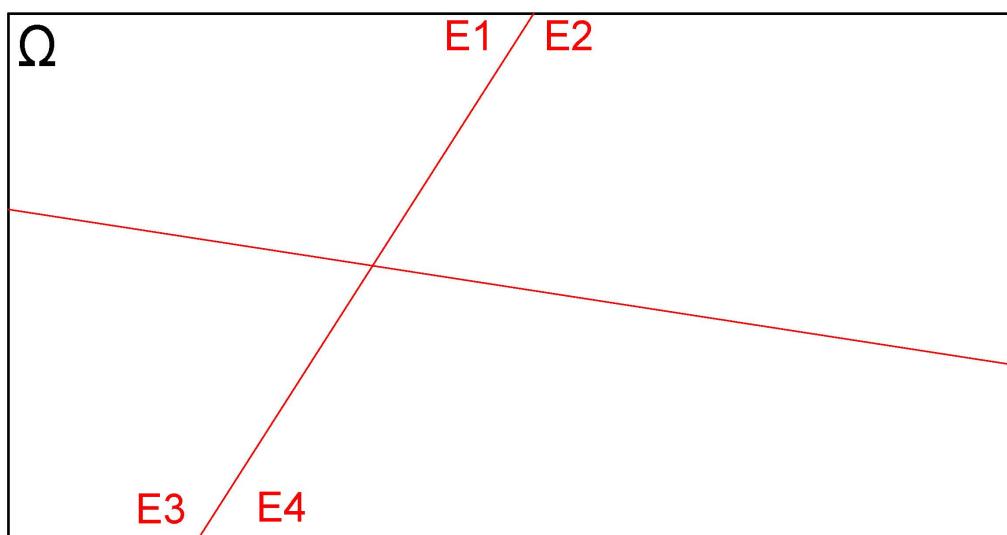


Figure 4.25: Espaço amostral e suas partições

E seja B um evento qualquer em Ω como ilustrado na Figura 4.26

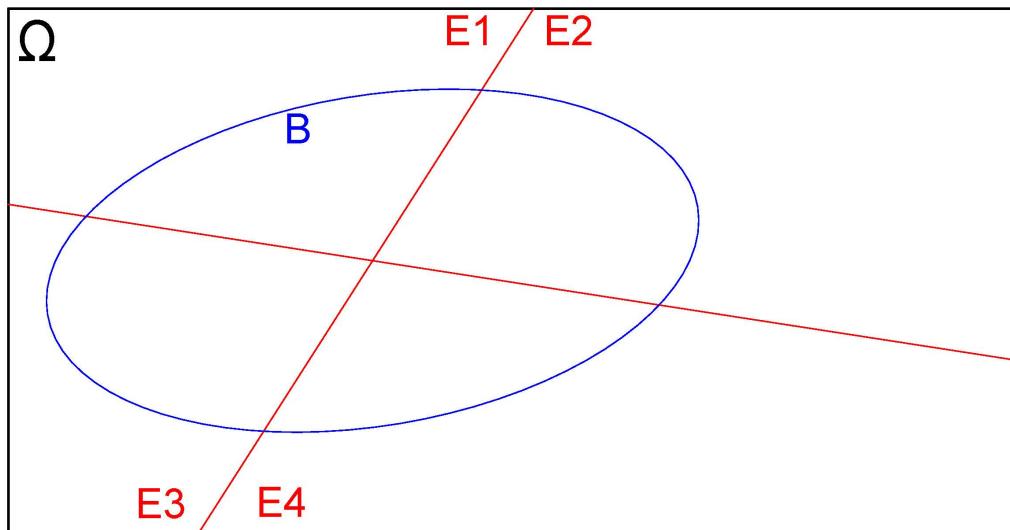


Figure 4.26: Evento definido sobre o espaço amostral

Delimitemos as interseções do evento B com as partições E_1, E_2, E_3 e E_4 do espaço amostral Ω , como ilustrado na Figura 4.27

Isso pode ser estendido, em uma forma geral, para $i = 1, \dots, n$ partições como ilustrado na Figura 4.28

Na representação esquemática da Figura 4.28 podemos identificar:

- 1- $E_1, E_2, \dots, E_i, \dots, E_n$ constituem-se em partições do espaço amostral Ω ;
- 2- Todas as partições são mutuamente exclusivas: $E_i \cap E_j = \emptyset, \forall i \neq j$ (a interseção de quaisquer partições é vazia);
- 3- Sendo vazias as interseções entre quaisquer partições, o espaço amostral Ω será a simples união de todas elas: $\Omega = E_1 \cup E_2 \cup E_3 \cup E_4 \cup \dots \cup E_i \dots \cup E_n$; e,
- 4- B é um evento qualquer definido sobre as partições de Ω

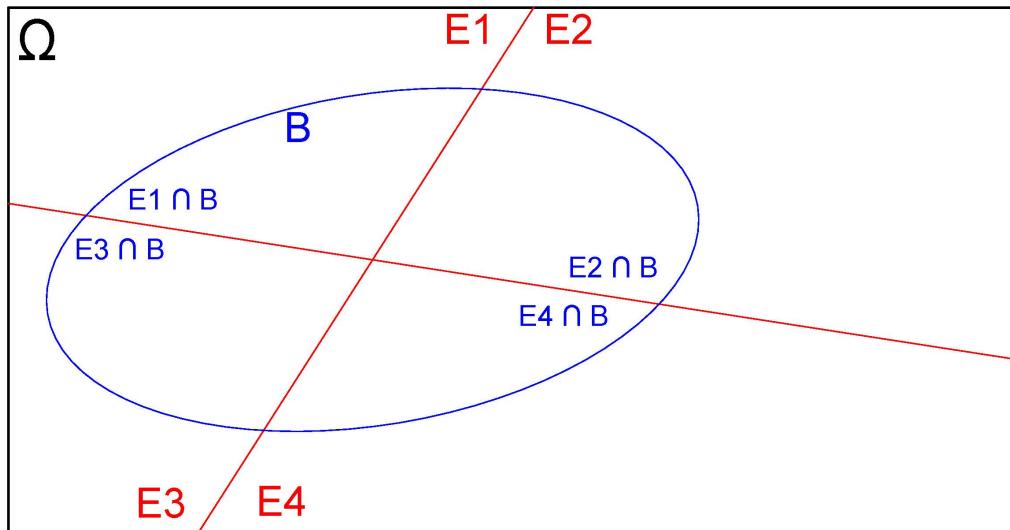
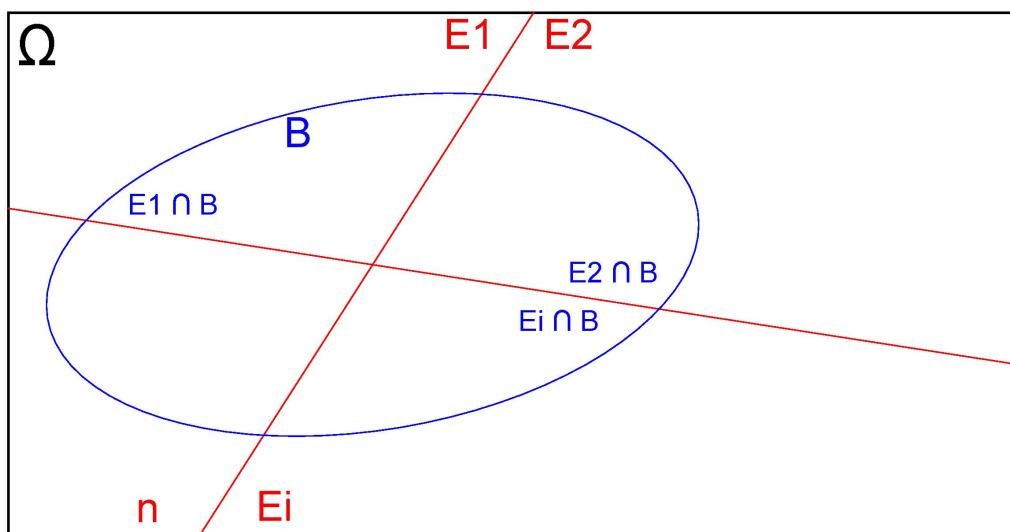


Figure 4.27: Interseções das partições do espaço amostral com o evento B

Figure 4.28: Interseções das n partições do espaço amostral com o evento B

São conhecidas as probabilidades de ocorrência de cada um dos elementos do espaço amostral Ω :

$$P(E_1); P(E_2); P(E_3); \dots; P(E_i); \dots; P(E_n)$$

e também as probabilidades do evento B condicionadas a cada elemento do espaço amostral:

$$P(B|E_1); P(B|E_2); \dots; P(B|E_i); \dots; P(B|E_n)$$

A *probabilidade de ocorrência do evento B* é dada pela soma das probabilidades de cada uma de suas interseções com os elementos do espaço amostral Ω , uma vez que essas interseções são disjuntas entre si:

$$\begin{aligned} P(B) &= P(E_1 \cap B) \cup P(E_2 \cap B) \cup \dots \cup P(E_i \cap B) \cup \dots \cup P(E_n \cap B) \\ P(B) &= \sum_{i=1}^n P(E_i \cap B) \end{aligned}$$

Pela *Regra do produto de eventos condicionados*, a *probabilidade de ocorrência do evento B posto* ter ocorrido um evento E_i é:

$$\begin{aligned} P(B|E_i) &= \frac{P(E_i \cap B)}{P(E_i)} \\ P(E_i \cap B) &= P(E_i) \times P(B|E_i) \end{aligned}$$

com $P(E) > 0$

Aplicando-se na expressão anteriormente desenvolvida da *probabilidade de ocorrência do evento B* teremos:

$$\begin{aligned} P(B) &= P(E_1 \cap B) \cup P(E_2 \cap B) \cup \dots \cup P(E_i \cap B) \cup \dots \cup P(E_n \cap B) \\ P(B) &= P(E_1) \times P(B|E_1) + P(E_2) \times P(B|E_2) + \\ &\quad \dots + P(E_i) \times P(B|E_i) + \\ &\quad \dots + P(E_n) \times P(B|E_n) \end{aligned}$$

Portanto a **probabilidade total** do evento B em Ω é dada pelo somatório:

$$P(B) = \sum_{i=1}^n [P(E_i) \cdot P(B|E_i)]$$

Pela **Regra do produto de eventos condicionados** a probabilidade de ocorrência de um evento E_i posto ter ocorrido o evento B é:

$$\begin{aligned} P(E_i|B) &= \frac{P(E_i \cap B)}{P(B)} \\ P(E_i \cap B) &= P(B) \times P(E_i|B) \\ P(B) &= \frac{P(E_i \cap B)}{P(E_i|B)} \end{aligned}$$

com $P(B) > 0$

Pela **igualdade** dos dois modos de se expressar a probabilidade total do evento B desenvolvidos:

$$P(B) = \frac{P(E_i \cap B)}{P(E_i|B)}$$

e

$$P(B) = \sum_{i=1}^n [P(E_i) \cdot P(B|E_i)]$$

tem-se

$$\frac{P(E_i \cap B)}{P(E_i|B)} = \sum_{i=1}^n [P(E_i) \cdot P(B|E_i)]$$

Rearranjando-se em termos da expressão anterior para exprimir a probabilidade de ocorrência de um evento E_i posto ter ocorrido o evento B chegamos a:

$$P(E_i|B) = \frac{P(E_i \cap B)}{\sum_{i=1}^n [P(E_i) \cdot P(B|E_i)]}$$

Sendo

$$P(E_i \cap B) = P(B) \times P(E_i|B)$$

a expressão anterior pode ser reescrita como:

$$P(E_i|B) = \frac{P(E_i) \times P(B|E_i)}{\sum_{i=1}^n [P(E_i) \times P(B|E_i)]}$$

uma forma mais geral do **Teorema de Bayes**.

O Teorema de Bayes é também chamado de Teorema da probabilidade a *posteriori* ao permitir que se calcule $P(E_i|B)$ em termos da ocorrência $P(B|E_i)$

É, de certo modo, uma conjugação do *teorema na probabilidade total* e da *regra do produto* de probabilidades.

O denominador:

$$P(B) = \sum_{i=1}^n [P(E_i) \times P(B|E_i)]$$

é a denominada **probabilidade marginal** de ocorrência do evento B no espaço amostral Ω composto por n elementos (partições).

Na expressão do Teorema de Bayes:

- $P(E_k|B)$ é a denominada probabilidade *a posteriori* do evento E_k condicionada pela ocorrência anterior do evento B ;
- $P(E_k)$ é a denominada probabilidade *a priori* do evento E_k ;
- $P(B|E_k)$ é a denominada probabilidade *a posteriori* do evento B condicionada pela ocorrência anterior do evento E_k ;
- $P(E_i)$ é a denominada probabilidade *a priori* de cada evento E_i ;
- $P(B|E_i)$ é a denominada probabilidade *a posteriori* do evento B condicionada pela ocorrência anterior de cada evento E_i .

Exemplo: Constatou-se que o aumento nas vendas de um certo produto comercializado por uma empresa num mês **pode ocorrer somente** por uma das quatro causas **mutuamente exclusivas** a seguir:

- 1- ação de marketing;
- 2- propaganda;
- 3- flutuações na economia do país; ou,
- 4- efeitos sazonais.

A probabilidade de haver uma ação da empresa no mês focada para o *marketing* é de 40%; e para propaganda é de 30%; as probabilidades de ocorrerem flutuações na economia do país é de 20% e de efeitos sazonais é de 10%. Uma pesquisa mostrou que a probabilidade de haver um aumento nas vendas do produto devido a uma ação de *marketing* é de 7%; devido à publicidade, de 7,5%, por flutuações na economia do país, de 3% e por sazonalidade de 2%.

Em um determinado mês a empresa observou um considerável incremento nas vendas. Qual seria sua causa mais provável? Qual a probabilidade de incremento das vendas em um certo mês?

Inicialmente definimos um experimento aleatório como sendo “qual fato ocorreu no mês”.

Não sabemos qual fato ocorreu, mas sabemos que as possibilidades são apenas 4 (*marketing*, propaganda, flutuações na economia ou efeitos sazonais).

Podemos então conceber que esses fatos são elementos do espaço amostral do experimento aleatório: pois são eventos exaustivos e exclusivos: não pode ocorrer mais de um ao mesmo tempo e ao menos um ocorrerá.

Assim esse espaço amostral é composto pelos seguintes “elementos” e suas probabilidades são tiradas do enunciado:

- E_1 o elemento “Ação de marketing” $\therefore \rightarrow P(E_1) = 0,40$;
- E_2 o elemento “Ação de propaganda” $\therefore \rightarrow P(E_2) = 0,30$;
- E_3 o elemento “Flutuações na economia” $\therefore \rightarrow P(E_3) = 0,20$; ou,
- E_4 o elemento “Sazonalidade” $\therefore \rightarrow P(E_4) = 0,10$.

Chamemos de B ao evento “ocorrer um incremento nas vendas”, um evento construído sobre os elementos do espaço amostral e que apresenta diferentes probabilidades a depender de qual elemento do espaço amostral ocorreu (a probabilidade de B está *condicionada* aos elementos do espaço amostral). Da leitura do enunciado extraímos as probabilidades de ocorrência de cada um dos eventos influenciadores:

As probabilidades condicionadas de ocorrer um incremento das vendas (B) pela ocorrência anterior de cada um dos elementos do espaço amostral (**posto ter ocorrido o evento E_i**) também são tiradas do enunciado:

- $P(B|E_1) = 0,07$;
- $P(B|E_2) = 0,075$;
- $P(B|E_3) = 0,03$; e,
- $P(B|E_4) = 0,02$.

Para responder à indagação do problema (“Qual a causa mais provável?”) podemos invertê-la e reformulá-la:

“Qual a probabilidade de ter ocorrido cada um dos quatro eventos (E_1, E_2, E_3, E_4) **posto** (dado) ter ocorrido um incremento nas vendas?

Calculemos para cada um deles usando o Teorema de Bayes:

$$P(E_i|B) = \frac{P(E_i) \times P(B|E_i)}{\sum_{i=1}^n [P(E_i) \times P(B|E_i)]}$$

Probabilidade da empresa ter realizado uma ação de *marketing*, **posto** ter ocorrido um incremento nas vendas de seu produto:

$$\begin{aligned} P(E_1|B) &= \frac{P(E_1) \times P(B|E_1)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]} \\ P(E_1|B) &= \frac{0,40 \times 0,07}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)} \\ P(E_1|B) &= 0,4786 \end{aligned}$$

Probabilidade da empresa ter realizado propaganda, **posto** ter ocorrido um incremento nas vendas de seu produto:

$$\begin{aligned} P(E_2|B) &= \frac{P(E_2) \times P(B|E_2)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]} \\ P(E_2|B) &= \frac{0,30 \times 0,075}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)} \\ P(E_2|B) &= 0,3846 \end{aligned}$$

Probabilidade da empresa ter ocorrido flutuações na economia, **posto** ter ocorrido um incremento nas vendas de seu produto:

$$\begin{aligned} P(E_3|B) &= \frac{P(E_3) \times P(B|E_3)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]} \\ P(E_3|B) &= \frac{0,20 \times 0,03}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)} \\ P(E_3|B) &= 0,1026 \end{aligned}$$

Probabilidade da empresa ter ocorrido efeitos sazonais, **posto** ter ocorrido um incremento nas vendas de seu produto:

$$P(E_4|B) = \frac{P(E_4) \times P(B|E_4)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]}$$

$$P(E_4|B) = \frac{0,10 \times 0,02}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)}$$

$$P(E_4|B) = 0,03419$$

Respostas:

1- Os cálculos indicam que o evento mais provável pelo incremento das vendas observado naquele mês foi o de uma **ação de marketing**;

2- A probabilidade de incremento das vendas em um determinado mês como resultado dos quatro possíveis eventos indicados é o **próprio denominador do Teorema de Bayes**: 0,0585.

Exemplo: Considere 5 urnas, cada uma delas contendo 6 bolas. Duas dessas urnas (urnas tipo C_1) possuem 3 bolas brancas em seu interior. Duas outras (urnas tipo C_2) possuem 2 bolas brancas em seu interior e a última (urna tipo C_3) possui 6 bolas brancas em seu interior (cf. Figura 4.29).

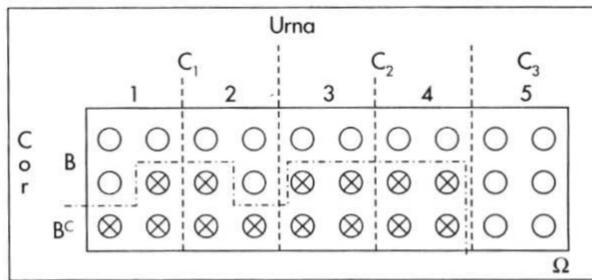


Figure 4.29: Cinco urnas cada uma com 6 bolas em cores de diferentes quantidades da cor branca

Escolhida aleatoriamente uma urna retira-se uma bola. Qual a probabilidade da urna escolhida ter sido a urna C_3 sabendo-se que a bola retirada foi branca?

Desejamos determinar $P(C_3|Branca)$

Da leitura do enunciado extraímos as seguintes informações:

$$\begin{aligned}P(C_1) &= \frac{2}{5} \\P(C_2) &= \frac{2}{5} \\P(C_3) &= \frac{1}{5} \\P(Branca|C_1) &= \frac{1}{2} \\P(Branca|C_2) &= \frac{1}{3} \\P(Branca|C_3) &= 1\end{aligned}$$

$$\begin{aligned}P(C_3|Branca) &= \frac{P(C_3) \times P(Branca|C_3)}{\sum_{i=1}^3 [P(C_i) \times P(Branca|C_i)]} \\P(C_3|Branca) &= \frac{0,20 \times 1,00}{(0,40 \times 0,50) + (0,40 \times 0,33) + (0,20 \times 1,00)} \\P(C_3|Branca) &= 0,375\end{aligned}$$

4.8 Teoremas da Teoria das probabilidades

4.8.1 Teorema 01

Se E é um evento num espaço discreto Ω , então $P(E)$ é igual à soma das probabilidades de ocorrência de todos os elementos do espaço amostral que satisfazem ao evento de interesse E .

Sejam E_1, E_2, E_3, \dots a sequência finita ou infinita de eventos que satisfazem ao evento de interesse E . Assim, $E = E_1 \cup E_2 \cup E_3 \dots$. Como E_1, E_2, E_3, \dots são eventos **mutuamente exclusivos**, pelo terceiro postulado das probabilidades teremos:

$$P(E) = P(E_1) + P(E_2) + P(E_3) + \dots$$

Exemplo: Lançamento de uma moeda duas vezes

Espaço amostral dos possíveis eventos (resultados): $\Omega = \{(cara, cara), (cara, coroa), (coroa, cara), (coroa, coroa)\}$

- Evento de interesse E : obter ao menos uma *cara*
- Eventos que satisfazem: $E_1 = \{(cara, cara)\}; E_2 = \{(cara, coroa)\}; E_3 = \{(coroa, cara)\}$

A probabilidade de E ($P(E)$) será a soma das probabilidades dos eventos que o satisfazem:

$$P(E) = P(E_1) + P(E_2) + P(E_3) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

4.8.2 Teorema 02

Se um experimento aleatório pode ter N resultados possíveis e equiprováveis e um evento E pode ter n resultados que o satisfazem, então $P(E) = \frac{n}{N}$.

Sejam $E_1, E_2, E_3, \dots, E_N$ os resultados do espaço amostral Ω , cada um deles equiprovável ($P(E_i) = \frac{1}{N}$). Se E é a união de n desses eventos **mutuamente exclusivos**, pelo terceiro postulado das probabilidades teremos:

$$\begin{aligned} P(E) &= P(E_1) + P(E_2) + P(E_3) + \dots + P(E_n) \\ P(E) &= \frac{1}{N} + \frac{1}{N} + \frac{1}{N} + \dots + \frac{1}{N} \\ P(E) &= \frac{n}{N} \end{aligned}$$

4.8.3 Teorema 03

Se E e E^c são eventos complementares no espaço amostra Ω então $P(E^c) = 1 - P(E)$.

Sendo os eventos E e E^c **mutuamente exclusivos** e também sendo $E \cup E^c = \Omega$, considerando-se que $P(\Omega) = 1$, pelos segundo e terceiro postulados tem-se:

$$\begin{aligned} P(\Omega) &= 1 \\ 1 &= P(E \cup E^c) \\ 1 &= P(E) + P(E^c) \end{aligned}$$

4.8.4 Teorema 04

$$P(\emptyset) = 0$$

Sendo Ω e \emptyset são **mutuamente exclusivos** e, como de acordo com a definição de um espaço vazio $\Omega \cup \emptyset = \Omega$, pelo terceiro postulado tem-se:

$$\begin{aligned} P(\Omega) &= P(\Omega \cup \emptyset) \\ P(\Omega) &= P(\Omega) + P(\emptyset) \\ P(\Omega) - P(\Omega) &= P(\emptyset) \\ P(\emptyset) &= 0 \end{aligned}$$

4.8.5 Teorema 05

Se A e B são eventos em um mesmo espaço amostral Ω e $A \subset B$ então $P(A) \leq P(B)$.

Se $A \subset B$ então pode-se escrever: $B = A \cup (A^c \cap B)$ (verifica-se pelo correspondente diagrama de Venn).

Como A e $A^c \cap B$ são **mutuamente exclusivos**, pelo terceiro postulado tem-se:

$$\begin{aligned} P(B) &= P(A) + P(A^c \cap B) \\ P(A) &= P(B) - P(A^c \cap B) \end{aligned}$$

4.8.6 Teorema 06

A probabilidade de qualquer evento E em Ω está compreendida entre $0 \leq P(E) \leq 1$.

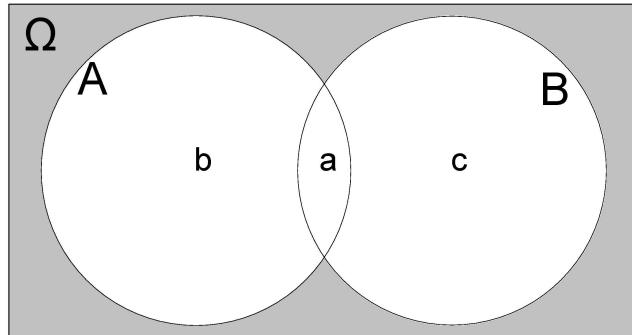
Estando $\emptyset \subset E \subset \Omega$ e considerando-se o Teorema 5 tem-se:

$$P(\emptyset) \leq P(E) \leq P(\Omega) \quad 0 \leq P(E) \leq 1$$

4.8.7 Teorema 07

Para dois eventos quaisquer em Ω , A e B tem-se que: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Sejam as seguintes probabilidades para esses eventos **mutuamente exclusivos**:



- $P(A \cap B) = a$;
- $P(A \cap B^c) = b$; e,
- $P(A^c \cap B) = c$.

$$\begin{aligned} P(A \cup B) &= a + b + c \\ P(A \cup B) &= (a + b) + (c + d) - a \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

4.8.8 Teorema 08

Para três eventos quaisquer em Ω , A , B e C tem-se que:

$$\begin{aligned} P(A \cup B \cup C) &= \\ &= P(A) + P(B) + P(C) - \\ &\quad P(A \cap B) - P(A \cap C) - P(B \cap C) + \\ &\quad P(A \cap B \cap C) \end{aligned}$$

Escrevendo-se $A \cup B \cup C$ como $A \cup (B \cup C)$ e usando o Teorema 7 duas vezes (uma para $P[A \cup (B \cup C)]$ e a outra para $P(B \cup C)$) tem-se:

$$\begin{aligned} P(A \cup B \cup C) &= P[A \cup (B \cup C)] \\ P(A \cup B \cup C) &= P(A) + P(B \cup C) - P[A \cap (B \cup C)] \\ P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(B \cap C) - P[A \cap (B \cup C)] \end{aligned}$$

Pela lei distributiva tem-se:

$$\begin{aligned} P[A \cap (B \cup C)] &= P[(A \cap B) \cup (A \cap C)] \\ P[A \cap (B \cup C)] &= P(A \cap B) + P(A \cap C) - P[(A \cap B) \cap (A \cap C)] \\ P[A \cap (B \cup C)] &= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \end{aligned}$$

Chega-se a :

$$\begin{aligned} P(A \cup B \cup C) &= \\ &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + \\ &\quad P(A \cap B \cap C) \end{aligned}$$

Módulo 5

Introdução a variáveis aleatórias

Ao realizar um experimento aleatório frequentemente estamos interessados mais em alguma função do resultado do que no próprio resultado em si:

Ao lançar dois dados, podemos estar interessados na soma “7”, sem nos importar se isso foi decorrente de (1,6),(2,5),(3,4),(4,3),(5,2) ou (6,1).

De forma semelhante, ao lançar três vezes uma moeda, podemos estar interessados no número total de “2 caras” (KK) que ocorre, sem nos preocuparmos se isso decorreu da sequência (K,K,C),(K,C,K) ou (C,K,K).

Da mesma forma, ao planejar uma família de três filhos, podemos estar interessados em ter exatamente 2 filhos do sexo masculino (MM), sem nos importar se isso resultou de (F,M,M), (M,F,M) ou (M,M,F).

Essas quantidades de interesse, ou mais formalmente, essas funções de valor real definidas no espaço amostral (o conjunto de todos os possíveis resultados do experimento), são conhecidas como variáveis aleatórias.

Em outras palavras, uma função X (geralmente representada por uma letra maiúscula) que associa cada elemento ω pertencente ao espaço amostral Ω um número real é denominada, de forma mais precisa, como uma variável aleatória ou função estocástica.

$$X(\Omega) \rightarrow \mathcal{R}_X, \text{ estando } \mathcal{R}_X \subseteq \mathcal{R}$$

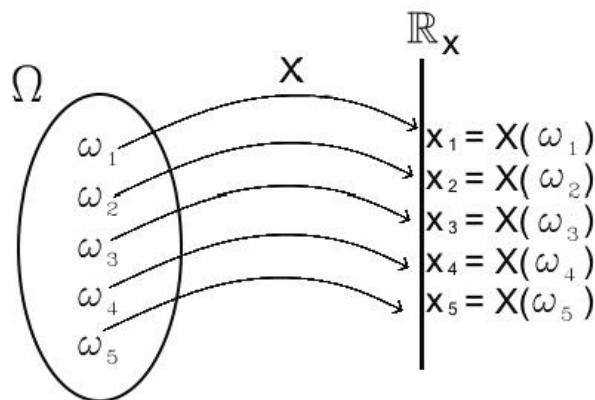


Figure 5.1: Variável aleatória

O *domínio* dessa função (X) é o conjunto de todos os possíveis valores numéricos de interesse do experimento aleatório e seu *contradomínio* está em \mathbb{R} .

Exemplo 1: considere o espaço amostral do experimento aleatório relacionado ao sexo do bebê em três gestações bem sucedidas: $\Omega = \{\omega_1 : (FFF), \omega_2 : (FFM), \omega_3 : (FMF), \omega_4 : (MFF), \omega_5 : (FMM), \omega_6 : (MFM), \omega_7 : (MMF), \omega_8 : (MMM)\}$. Se estivermos interessados no número de nascimentos do sexo masculino, podemos definir a função X para associar cada elemento ω_i em Ω a um valor $x_i \in \mathbb{R}$ que apresentará os seguintes resultados: $X(FFF) = 0; X(FFM) = 1; X(FMF) = 1; X(MFF) = 1; X(FMM) = 2; X(MFM) = 2; X(MMF) = 2; X(MMM) = 3$ $\mathcal{R}_X = \{x_1 = 0; x_2 = 1; x_3 = 2; x_4 = 3\} \subseteq \mathcal{R}$

Exemplo 2: considere o espaço amostral do experimento aleatório relacionado à sobrevivência de paciente ao final de 1 dia em uma UTI com quatro leitos: $\Omega = \{(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 1, 0), (0, 0, 1, 1), (0, 1, 0, 0), (0, 1, 0, 1), (0, 1, 1, 0), (0, 1, 1, 1), (1, 0, 0, 0), (1, 0, 0, 1), (1, 0, 1, 0), (1, 0, 1, 1), (1, 1, 0, 0), (1, 1, 0, 1), (1, 1, 1, 0), (1, 1, 1, 1)\}$. Cada elemento do espaço amostral é uma sequência de quatro valores binários (x_1, x_2, x_3, x_4) , onde: $x_i = 0$ indica que o paciente no leito i sobreviveu e $x_i = 1$ caso contrário. Se estivermos interessados no número de falecimentos podemos definir a função X para associar cada elemento ω_i em Ω a um valor $x_i \in \mathbb{R}$ que apresentará os seguintes resultados: $X(0000) = 0; X(0001) = 1; X(0010) = 1; \dots; X(1111) = 4$ $\mathcal{R}_X = \{x_1 = 0; x_2 = 1; x_3 = 2; x_4 = 3; x_5 = 4\} \subseteq \mathcal{R}$

Existem dois tipos de variáveis aleatórias: **discretas** e **contínuas**:

Os valores possíveis de uma variável aleatória discreta pertencem a um conjunto finito ou infinito enumerável, como $\{0, 1, 2, \dots\}$. Exemplos incluem: o número de acidentes em uma semana; o número de partículas emitidas por uma fonte radioativa em um intervalo de tempo; ou o número de casos de uma doença em um mês.

Os valores possíveis de uma variável aleatória contínua pertencem a um intervalo contínuo de números reais, como $[a, b]$, $[0, \infty)$ ou $-\infty, \infty$). Exemplos incluem: o peso ou a altura de um grupo de pessoas; o tempo de vida de uma lâmpada; o tempo de reação a um estímulo; a concentração de álcool em um certo volume de sangue, a temperatura mínima no inverno Antártico.

A função aleatória X não irá, em geral, simplesmente relacionar os possíveis resultados de interesse do experimento aleatório a números reais. Mas sim computar algumas informações úteis como a probabilidade de sua verificação.

5.1 Função massa de probabilidade (*Probability Mass Function - PMF*)

Considere X uma variável aleatória discreta com o contradomínio $x_1, x_2, x_3, \dots, x_n$ Uma *função (de distribuição) de probabilidade* $f(x)$ é assim denominada se, aplicada a cada um dos possíveis valores x_i da variável aleatória X , resultar em sua probabilidade de ocorrência. Assim, para $x = x_i$, $f(x_i) = P(X = x_i) = p(x_i)$.

Para que essa *função* $f(x)$ possa ser considerada uma **função de distribuição de probabilidade**, ela precisa necessariamente atender às seguintes condições:

Postulado do intervalo:

$$0 \leq f(x_i) \leq 1$$

para qualquer $x_i \in \mathcal{R}_X$, onde \mathcal{R}_X é o contradomínio de X .

Postulado do evento certo:

$$\sum_{i=1}^n f(x_i) = 1.$$

Equivale afirmar que a probabilidade de ocorrência de um dos valores que a variável aleatória pode assumir está sempre compreendida no intervalo $0 \leq P(X = x_i) \leq 1$. E mais, que a soma das probabilidades de todos os possíveis valores de X será 1.

Ponto amostral	(cara,cara)	(cara,coroa)	(coroa,cara)	(coroa,coroa)
X	2	1	1	0

Observação: somente se a soma acima for finita há a possibilidade de se ter probabilidades $p(x_i)$ iguais para todos x_i

Exemplo: Suponha que uma moeda seja lançada duas vezes e que X seja a variável aleatória que represente o número de *caras* verificado. Defina o espaço amostral, associe para cada evento possível o valor da variável aleatória e definda uma função discreta de probabilidade correspondente.

O espaço amostral desse experimento é $S = \{(cara,cara), (cara,coroa), (coroa,cara), (coroa,coroa)\}$ e a tabela abaixo relaciona o número de **caras** (o valor da variável aleatória X) associado a cada evento possível desse experimento:

As probabilidades de ocorrência de cada um desses eventos é:

$$\begin{aligned} P(cara, cara) &= \frac{1}{4} \\ P(cara, coroa) &= \frac{1}{4} \\ P(coroa, cara) &= \frac{1}{4} \\ P(coroa, coroa) &= \frac{1}{4} \end{aligned}$$

Para definir uma *função discreta de distribuição de probabilidade* deveremos associar a cada valor que a variável aleatória X assume sua correspondente *probabilidade de ocorrência*.

$$\begin{aligned} P(X = 0) &= P(coroa, coroa) = \frac{1}{4} \\ P(X = 1) &= P[(cara, coroa) \cup (coroa, cara)] \\ &= P(cara, coroa) + P(coroa, cara) \\ &= \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{2} \\ P(X = 2) &= P(cara, cara) = \frac{1}{4} \end{aligned}$$

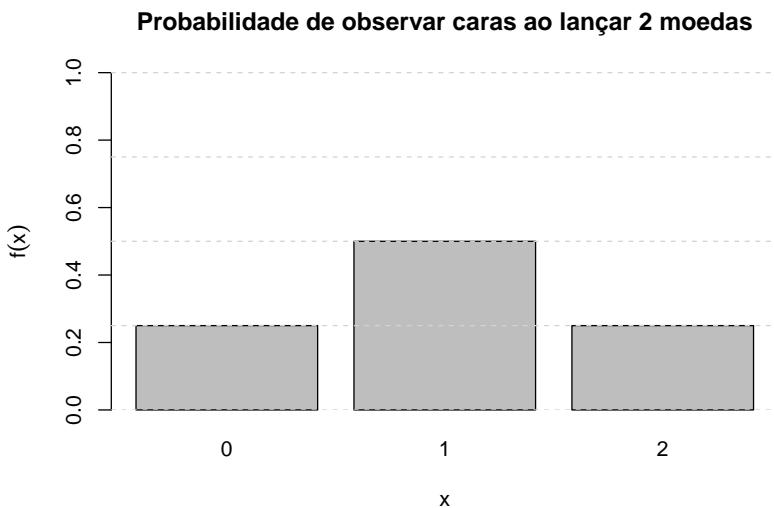
Table 5.1: *

Função discreta de probabilidades da variável aleatória X

x_k	0	1
$P(X = x_k) = f(x_k)$	1/4	1/2

```
# Valores possíveis para o número de caras
x <- c(0, 1, 2)
# Probabilidades associadas (calculadas com o modelo binomial)
p <- c(0.25, 0.5, 0.25)
# Criando o gráfico de barras
barplot(
  p,
  names.arg = x,
  xlab = "x",
  ylab = expression(f(x)),
  col = "gray",
  ylim = c(0, 1),
  main = "Probabilidade de observar caras ao lançar 2 moedas"
)

# Adicionando linhas no eixo y para reforçar a interpretação
abline(h = seq(0, 1, by = 0.25), col = "lightgray", lty = 2)
```



Uma função de distribuição cumulativa $F(x)$ para uma variável aleatória X exprime a probabilidade de que a variável aleatória X assuma um valor menor ou igual a determinado x , sendo definida por:

$$F(x) = P(X \leq x)$$

Propriedades:

- 1- $0 \leq F(x) \leq 1$ (os valores da função estão no intervalo de probabilidade);
 2- $F(x)$ é não decrescente: $F(x) \leq F(y)$ se $x \leq y$ (a probabilidade cumulativa nunca diminui); - $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$ (a probabilidade cumulativa no limite inferior é zero); 4- $F(+\infty) = \lim_{x \rightarrow \infty} F(x) = 1$ (a probabilidade cumulativa no limite superior é um).

Relação com a Função de Probabilidade:

Para uma variável aleatória discreta X , a **função massa de probabilidade** $f(x)$ pode ser derivada da **função de distribuição cumulativa** $F(x)$. Especificamente, para todo x em $(-\infty, \infty)$:

$$f(x) = P(X = x) = F(x) - F(x^-),$$

onde $F(x^-)$ é o valor de $F(x)$ imediatamente **antes** de x , considerando a natureza discreta da variável.

Além disso, a definição cumulativa da função $F(x)$ para valores discretos x pode ser escrita como:

$$F(x) = P(X \leq x) = \sum_{u \leq n} f(u),$$

em que u são os valores possíveis da variável aleatória X . Equivale dizer que é a soma sobre todos os valores u assumidos por X para os quais $u \leq x$.

Se X é discreta e assume um número finito de valores x_1, x_2, \dots, x_n , então sua função de probabilidade cumulativa $F(x)$ será dada por:

$$F(x) = \begin{cases} 0, & \text{se } -\infty < x < x_1, \\ f(x_1), & \text{se } x_1 \leq x < x_2, \\ f(x_1) + f(x_2), & \text{se } x_2 \leq x < x_3, \\ \vdots & \\ f(x_1) + f(x_2) + \dots + f(x_n), & \text{se } x_n \leq x < \infty. \end{cases}$$

Essa expressão mostra que $F(x)$ é obtida pela soma cumulativa das probabilidades associadas aos valores discretos de X , até um ponto x .

x_k	0	1	2
$P(X = x_k) = f(x_k)$	1/4	1/2	1/4

Exemplo: Suponha que uma moeda seja lançada duas vezes e que X seja a variável aleatória que represente o número de **caras** verificado. Especifique sua função de probabilidade cumulativa dessa variável aleatória e apresente seu gráfico.

Sua função de probabilidade cumulativa é dada por:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4} & 0 \leq x < 1 \\ \frac{3}{4} & 1 \leq x < 2 \\ 1 & 2 \leq x \end{cases} \quad (5.1)$$

O gráfico de sua função de probabilidade cumulativa é:

```
# Valores possíveis para o número de caras
x <- c(0, 1, 2)

# Probabilidades associadas (calculadas com o modelo binomial)
p <- c(0.25, 0.5, 0.25)

# Probabilidade acumulada
p_cumulative <- cumsum(p)

# Criando o gráfico de probabilidade acumulada com linhas verticais
plot(
  x,
  p_cumulative,
  type = "s", # Tipo "steps" para probabilidade acumulada
  xlab = "x",
  ylab = expression(F(x)),
  ylim = c(0, 1),
  xlim = c(0, 3), # Expandido para incluir margens
  main = "Função de distribuição acumulada (CDF)",
  col = "blue",
  lwd = 2,
  axes = FALSE # Desabilita os eixos padrão para customização
)

# Adicionando linhas verticais para conectar os valores
segments(
  x0 = x[-1], # Exclui o primeiro valor (X=0)
  y0 = c(0, p_cumulative[-length(p_cumulative)])[-1], # Exclui o valor inicial (F(0))
  x1 = x[-1],
```

```

y1 = p_cumulative[-1], # Exclui o valor inicial ( $F(0)$ )
col = "red",
lty = 2 # Linhas tracejadas
)

# Adicionando bolinhas abertas nos limites inferiores (apenas  $X > 0$ )
points(x[-1], c(0, p_cumulative[-length(p_cumulative)])[-1], pch = 1, col = "blue", cex =
  1.5)

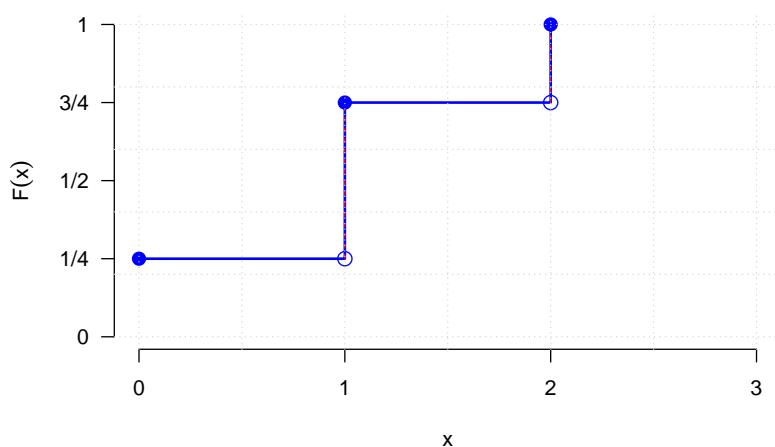
# Adicionando bolinhas fechadas no limite superior de cada degrau
points(x, p_cumulative, pch = 16, col = "blue", cex = 1.5)

# Customizando o eixo X
axis(1, at = 0:3, labels = 0:3)

# Customizando o eixo Y com frações
axis(2, at = c(0, 0.25, 0.5, 0.75, 1), labels = c("0", "1/4", "1/2", "3/4", "1"), las = 1)

# Adicionando grades para facilitar a interpretação
grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted")

```

Função de distribuição acumulada (CDF)

5.2 Função de densidade de probabilidade (*Probability Density Function - PDF*)

Considerem os espaços amostrais a seguir $(\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_5)$ representativos de 4 experimentos aleatórios e admitam também que todos os eventos possíveis são equiprováveis.

Interpretem o último deles como um espaço amostral formado por ∞ pontos amostrais.

Os eventos que compõem os quatro primeiros espaços amostrais são variável aleatória discretas.

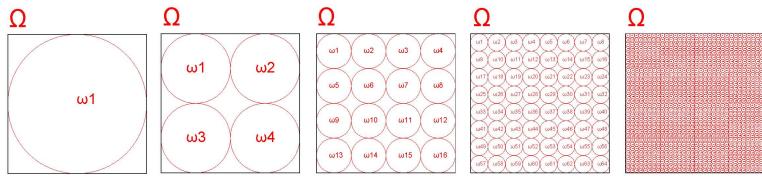


Figure 5.2: Diferentes espaços amostrais de um experimento aleatório (por razões gráficas desprezem o espaço fora dos círculos)

Discretas pois permitem a contagem dos possíveis valores (finitos ou infinitos contáveis) aleatórios que o experimento pode assumir. Mas no quinto espaço amostral temos incontáveis possibilidades.

Um *espaço amostral* com essa característica é representativo de uma *variável aleatória contínua*.

Sendo todos os eventos representados nos espaços amostrais **equiprováveis**, comparemos as probabilidades associadas a cada um desses possíveis resultados.

Em Ω_1 , $P(\omega_1) = 1$

Em Ω_2 , $P(\omega_1) = P(\omega_2) = P(\omega_3) = P(\omega_4) = 0,50$

Em Ω_3 , $P(\omega_1) = P(\omega_2) = \dots = P(\omega_{16}) = 0,0625$

Em Ω_4 , $P(\omega_1) = P(\omega_2) = \dots = P(\omega_{64}) = 0,015625$

Em Ω_5 , $P(\omega_n) \rightarrow 0$, à medida que o número de eventos $n \rightarrow \infty$

A probabilidade individual de qualquer evento do quinto espaço amostral ocorrer $\rightarrow 0$.

Por essa razão, no caso de variáveis aleatórias contínuas, não faz sentido falar em uma *probabilidade pontual exata*, associada a um resultado específico. Isso ocorre porque, para qualquer valor particular, a probabilidade é sempre igual a zero.

Em vez disso, considera-se a probabilidade associada a um intervalo de valores. A função de densidade de probabilidade contínua de uma variável aleatória X apresenta as seguintes propriedades fundamentais:

1- $f(x) \geq 0$, para todo $x \in (-\infty, \infty)$; 2- A integral da função sobre todo o domínio é igual a 1:

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Se X é uma variável aleatória contínua, a probabilidade de X assumir qualquer valor exato é $P(X = x) = 0$. No entanto, a **probabilidade intervalar** de X estar entre dois valores distintos, a e b , é dada por:

$$P(a < X < b) = \int_a^b f(x) dx.$$

Graficamente, a interpretação de uma função densidade de probabilidade contínua é representada pela **área sob a curva** da função $f(x)$, delimitada pelos valores de interesse a e b . Essa área corresponde à probabilidade de X estar no intervalo (a, b) .

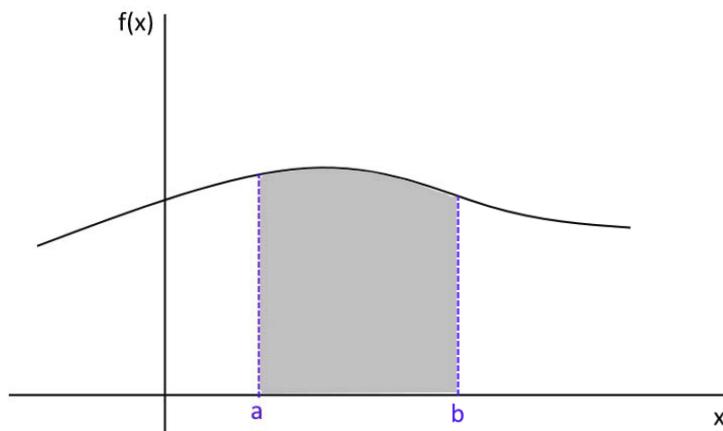


Figure 5.3: A área sob a curva de uma função de probabilidade de uma variável contínua entre dois valores quaisquer é a probabilidade de se observar valores entre esses dois pontos

Como $f(x) \geq 0$, a curva da função densidade de probabilidade estará acima do eixo x e a totalidade da área será igual a 1, conforme

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

A **função de distribuição cumulativa**, definida como $F(x) = P(X \leq x)$, também terá a forma de uma curva, crescente, que aumenta continuamente de 0 para 1. Essa característica reflete o fato de que, ao longo do eixo x , a probabilidade cumulativa acumula todos os valores de $f(x)$ até x , de modo que:

1. $F(x)$ é **não decrescente** ($F(x_1) \leq F(x_2)$ para $x_1 \leq x_2$);
2. $F(x)$ atinge valores-limite:

$$\lim_{x \rightarrow -\infty} F(x) = 0 \quad \text{e} \quad \lim_{x \rightarrow \infty} F(x) = 1.$$

Assim, $F(x)$ descreve graficamente a probabilidade acumulada até qualquer valor x , reforçando a relação entre a densidade de probabilidade e a probabilidade cumulativa.

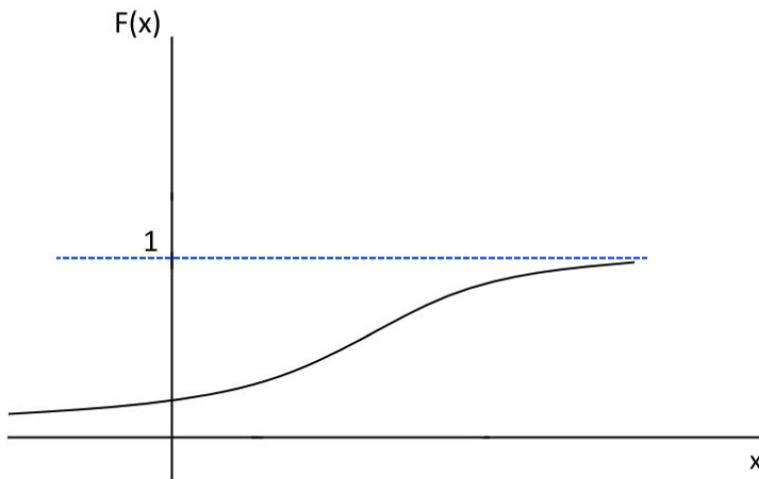


Figure 5.4: Função de probabilidade cumulativa

Exemplo: Seja a seguinte função e verifique se a função $f(x)$ pode ser a *função de densidade de probabilidade* da variável aleatória contínua X e determine qual a probabilidade associada a valores compreendidos no intervalo $0 \leq X \leq \frac{1}{2}$.

$$f(x) = \begin{cases} 2x & \text{para } 0 \leq x \leq 1 \\ 0 & \text{fora desse intervalo} \end{cases} \quad (5.2)$$

A resolução deste exemplo será feita de um modo *geométrico*.

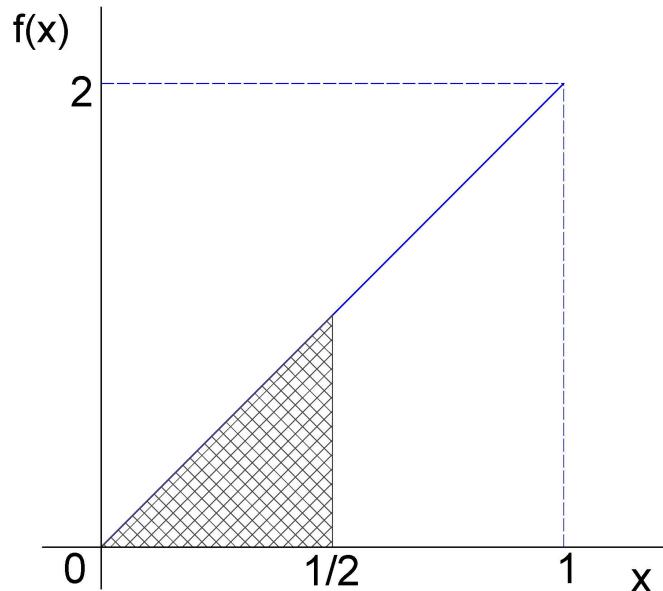


Figure 5.5: A probabilidade de se observar valores entre 0 e $1/2$ é igual à área sob a função densidade de probabilidade entre esses dois valores

- (a) Verificações para se aceitar a função como uma função de densidade de probabilidade para a variável aleatória X :

$$f(x) \geq 0$$

e,

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Resp.: Atende às duas condições (não assume valores menores que zero e a área sob a reta dessa função é unitária)

- (b) Cálculo da probabilidade para o intervalo $0 \leq X \leq \frac{1}{2}$ a partir da área do triângulo hachurado ($\frac{\text{base} \times \text{altura}}{2}$):

$$P(0 \leq X \leq \frac{1}{2}) = \frac{1}{2} \times (\frac{1}{2} \times 1) = \frac{1}{4}$$

5.3 Esperança e variância de uma variável aleatória discreta

Coletando-se dados podemos analisá-los, por exemplo, em termos de sua distribuição, pelas estatísticas da média e variância.

De maneira análoga procedemos com variáveis aleatórias (discretas ou contínuas) onde dispomos das *probabilidades* de ocorrência associadas a cada um dos valores (discretos ou infinitos numeráveis) que ela pode assumir.

A *esperança matemática* (valor esperado ou expectância) de uma variável aleatória discreta é dada pela *somatória do produto* de cada um dos valores que ela pode assumir pela probabilidade associada a cada um desses valores.

Seja X uma variável aleatória discreta que pode assumir os valores x_1, x_2, \dots, x_n ; e sejam P_1, P_2, \dots, P_n as respectivas probabilidades associadas às suas ocorrências.

A esperança da variável X , denotada por $E(X)$ será:

$$E(X) = \sum_{i=1}^n x_i \cdot P_i$$

Com n sendo o número de possíveis resultados que a variável X pode assumir.

A expressão anterior é semelhante àquela usada para se calcular a média para frequências de dados sendo que agora, no lugar de se utilizar a frequência relativa a cada dado observado, temos as probabilidades dadas por um modelo teórico pressuposto.

Algumas propriedades envolvendo a esperança:

- 1- Se c é uma constante qualquer, então: $E(c) = c$ ($c \in \mathbb{R}$);
- 2- Se c é uma constante qualquer, então: $E(cX) = c \cdot E(X)$ ($c \in \mathbb{R}$);
- 3- Se c é uma constante qualquer, então: $E(X \pm c) = E(X) \pm c$ ($c \in \mathbb{R}$);
- 4- Se X e Y são duas variáveis aleatórias quaisquer, então: $E(X + / - Y) = E(X) + / - E(Y)$;
- 5- Se X e Y são duas variáveis aleatórias independentes quaisquer, então: $E(X \cdot Y) = E(X) \cdot E(Y)$.

A variância de uma variável aleatória qualquer X , denotada por $Var(X)$, será dada por:

$$Var(X) = \sum_{i=1}^n [x_i - E(X)]^2 \cdot P_i$$

Algumas propriedades envolvendo a variância:

- 1- Se c é uma constante qualquer, então: $Var(c) = 0$ ($c \in \mathbb{R}$);
- 2- Se c é uma constante qualquer, então: $Var(cX) = c^2 \cdot Var(X)$ ($c \in \mathbb{R}$);
- 3- Se X e Y são duas variáveis aleatórias **independentes** quaisquer, então: $Var(X \pm Y) = Var(X) + Var(Y)$;
- 4- Se X e Y são duas variáveis aleatórias **quaisquer**, então: $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$ (também).

A covariância ($Cov(X, Y)$) entre duas variáveis aleatórias quaisquer X e Y é dada por:

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

Exemplo: Seja X uma variável aleatória discreta que indica o *número de pontos observados na face superior de um dado* quando ele é lançado. Calcule a esperança e a variância dessa variável aleatória.

Table 5.2: *

Função discreta de distribuição de probabilidades de X

x_i	$P(X = x_i)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

$$E(X) = \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = 3,50$$

$$\begin{aligned}
Var(X) &= (1 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + (2 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + \\
&\quad (3 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + (4 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + (5 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + \\
&\quad (6 - 3,50)^2 \cdot \left(\frac{1}{6}\right) \\
&= 2,90
\end{aligned}$$

Exemplo: Uma empresa de caminhões de aluguel possui uma frota composta de 4 veículos. O aluguel é cobrado por diária de uso de um caminhão e a função de distribuição de probabilidade de locações diárias está a seguir especificada. Calcule a esperança e a variância de locação diária dessa empresa.

Table 5.3: *

Função discreta de distribuição de probabilidade de locações diárias

x_i	$P(X = x_i)$
0	0,10
1	0,20
2	0,30
3	0,30
4	0,10

$$E(X) = (0,0,10) + (1,0,20) + 2,0,30 + (3,0,30) + (4,0,10) = 2,10 \text{ (caminhões por dia)}$$

$$\begin{aligned}
Var(X) &= (0 - 2,10)^2 \cdot 0,10 + (1 - 2,10)^2 \cdot 0,20 + (2 - 2,10)^2 \cdot 0,30 + \\
&\quad (3 - 2,10)^2 \cdot 0,30 + (4 - 2,10)^2 \cdot 0,10 \\
&= 1,29^1
\end{aligned}$$

¹: (caminhões por dia)²

5.4 Esperança e variância de uma variável aleatória contínua

A esperança e a variância de uma variável aleatória contínua são dadas, respectivamente, por:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$$Var(X) = E[(X - E(X))^2] = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x) dx$$

Módulo 6

Introdução a modelos teóricos de probabilidade

Existem variáveis aleatórias discretas ou contínuas, que apresentam certas características ou padrões de comportamento. Para essas variáveis, com base nesses comportamentos típicos, foram estruturados modelos teóricos de distribuições de probabilidade (variáveis discretas) e de densidade de probabilidade (variáveis contínuas) e derivadas as expressões de suas esperanças e variâncias.

6.1 Modelos teóricos discretos

6.1.1 Uniforme

Variável aleatória X , assumindo valores x_1, x_2, \dots, x_k tem distribuição uniforme se, e somente se,

$$P(X = x_i) = \frac{1}{k}.$$

para todo $i = 1, 2, \dots, k$. Para uma variável com distribuição uniforme:

- Esperança: $E(X) = \frac{1}{k} \sum_{i=1}^k x_i$
- Variância: $VAR(X) = \frac{1}{k} \left[\sum_{i=1}^k x_i^2 \times \frac{(\sum_{i=1}^k x_i)^2}{k} \right]$

6.1.2 Bernoulli

Variável aleatória com distribuição *Bernoulli* é uma variável definida por um experimento probabilístico em que os resultados possíveis se resumem a apenas dois: **sucesso** ou **fracasso** (ocorrência ou não).

Caracterização de uma variável aleatória X com distribuição de Bernoulli: $X \sim Ber(p)$

Para uma variável de Bernoulli:

x_i	Evento	$P(X = x_i)$
1	Sucesso	p
0	Fracasso	q=1-p
Σ	-	1

- Esperança: $E(X) = p$
- Variância: $VAR(X) = p(1 - p)$

Exemplo: Seja X uma variável aleatória resultante do lançamento de um dado uma única vez e cujo sucesso está definido como **obter a face com 5 pontos**. Calcule a probabilidade de sucesso e fracasso, assim como sua variância.

x_i (face 5 no lançamento de um dado)	Evento	$P(X = x_i)$
1	Sucesso	$p=1/6$
0	Fracasso	$q=5/6$
Σ	-	1

- Esperança: $E(X) = \frac{1}{6}$
- Variância: $Var(X) = \frac{5}{36}$

Admita agora X uma variável aleatória resultante de realização de n tentativas (repetições) de Bernoulli e definindo x como sendo o número de sucessos verificados nessas n tentativas. Desse modo, proporção de sucessos observada após n repetições é expressa como $\frac{x}{n}$.

Se p é a probabilidade de sucesso a cada repetição e se ϵ é um número qualquer positivo, tem-se:

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{x}{n} - p\right| \geq \epsilon\right) = 0$$

A Lei dos grandes números para infinitas repetições de Bernoulli afirma que, após um **grande número de repetições (n)**, a proporção de sucessos observada ($\frac{x}{n}$) **irá se aproximar** da probabilidade teórica da variável aleatória de Bernoulli p .

6.1.3 Binomial

Variável aleatória com distribuição Binomial é uma variável resultante da repetição de um **experimento modelado por uma variável de Bernoulli** (isto é, a cada repetição apenas dois resultados podem ocorrer: sucesso ou fracasso).

Para que X seja uma variável aleatória com distribuição Binomial: $X \sim b(n, p)$ é necessário que:

- o experimento deve ser realizado um número n finito de vezes;
- cada repetição deve ser independente das demais;
- cada repetição é, em essência, um ensaio de Bernoulli onde só pode haver dois resultados: sucesso ou fracasso;
- a probabilidade de sucesso p em cada repetição é **sempre a mesma**; e, consequentemente,
- a probabilidade de fracasso $q = 1 - p$ em cada repetição é **também a mesma**.

Considerem o diagrama de árvore ilustrado na Figura 6.1 que representa, esquematicamente, 3 repetições independentes de um evento modelado por uma variável de Bernoulli, com probabilidade individual de sucesso $P(X = 1) = p$ e, de fracasso, $P(X = 0) = 1 - p = q$.

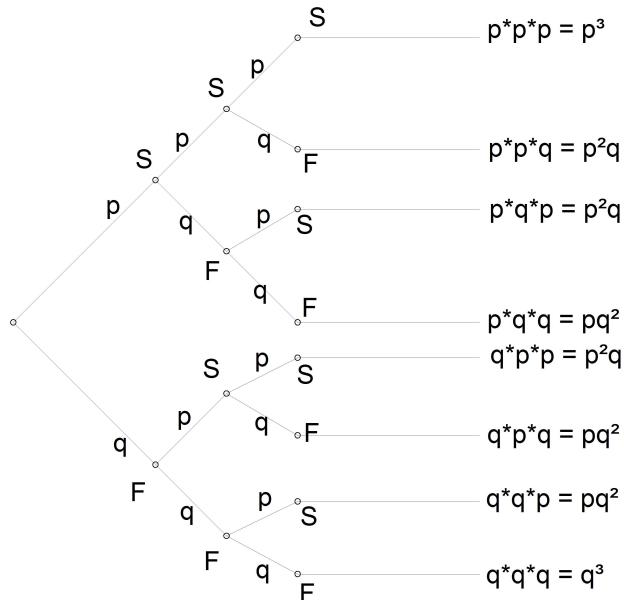


Figure 6.1: Três repetições independentes de um experimento aleatório modelado por uma variável de Bernoulli

Se p é a probabilidade de se verificar sucesso em qualquer uma das n repetições de Bernoulli realizadas no experimento aleatório então uma variável aleatória Binomial X definida sobre esse experimento apresentará k sucessos após n repetições independentes e terá a seguinte função de probabilidade:

Table 6.1: *

Função discreta de probabilidade da variável $X \sim b(n, p)$ com $n = 3$ (repetições)

Número de sucessos	Probabilidade	Probabilidade se $p = 0,50$
0	q^3	$\frac{1}{8}$
1	$3pq^2$	$\frac{3}{8}$
2	$3p^2q$	$\frac{3}{8}$
3	p^3	$\frac{1}{8}$

$$\begin{aligned} f(k) &= P(X = k) \\ f(k) &= C_k^n \cdot p^k \cdot q^{(n-k)} \\ f(k) &= \frac{n!}{k!(n-k)!} \cdot p^k \cdot q^{(n-k)} \end{aligned}$$

Sendo a probabilidade p de sucesso, igual em todas as repetições, então:

- Esperança: $E(X) = \sum_{i=1}^n x_i \cdot P(X = x_i) = n.p$
- Variância: $V(X) = E(X^2) - [E(X)]^2 = n.p.q$

Exemplo: Numa prova com 6 questões, a probabilidade de que um aluno acerte cada uma delas é de 0,30. Admitindo que a resolução dessas 6 questões é feita de modo independente, qual a probabilidade desse aluno acertar 4 questões?

- 1- cada questão apresenta apenas duas possibilidades: **acertar ou errar**; assim, esse experimento aleatório pode seguir o modelo teórico de Bernoulli tendo o evento de sucesso definido como: **a chance de acertar uma prova**, com probabilidade de ocorrência $p = 0,30$;
- 2- ao se repetir esse experimento $n = 6$ (pois este é o número de questões a serem resolvidas) o experimento passa seguir o modelo teórico Binomial pois nos foi assegurada a independência entre cada repetição bem como a constância da probabilidade p .

A probabilidade de se acertar $k = 4$ questões em $n = 6$ repetições independentes tendo cada uma uma probabilidade de sucesso $p = 0,30$ será então:

$$\begin{aligned}
 P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\
 P(X = 4) &= 15.0, 30^4 \cdot 0, 70^{(6-4)} \\
 &= 0,0595
 \end{aligned}$$

Conclusão: a probabilidade de um aluno acertar 4 questões das 6 resolvidas, considerando a probabilidade associada ao acerto de cada questão, é de 0,0595.

Exemplo: Ainda utilizando a construção teórica desse experimento, admitamos que nosso interesse reside em obter as seguintes probabilidades a ele associadas: 1- probabilidade do aluno não acertar nenhuma questão;
 2- probabilidade do aluno acertar todas as questões;
 3- probabilidade do aluno acertar no mínimo 2 questões; e a
 4- probabilidade do aluno acertar no máximo 2 questões.

A resposta aos dois primeiros itens é imediata pela simples aplicação dos dados ao modelo, pois o número de sucessos desejado é $k = 0$ no primeiro e $k = 6$ no segundo (e $p = 0,30$ para todos). Assim:

$$\begin{aligned}
 P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\
 P(X = 0) &= 1.0, 30^0 \cdot 0, 70^{(6-0)} \\
 &= 0,1176
 \end{aligned}$$

$$\begin{aligned}
 P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\
 P(X = 6) &= 1.0, 30^6 \cdot 0, 70^{(6-6)} \\
 &= 0,000729
 \end{aligned}$$

A resposta aos dois últimos itens irá demandar o uso da **regra da adição de probabilidades** e, como cada evento é disjunto dos demais, essa regra recai sobre a simples adição das probabilidades envolvidas.

Ao perguntar qual a probabilidade do aluno acertar no **mínimo** 2 questões ($P(X \geq 2)$) equivale a se perguntar qual a probabilidade do aluno acertar 2 **OU** 3 **OU** 4 **OU** 5 **OU** 6 questões. Assim, temos como elementos desses eventos de sucesso 2, 3, 4, 5, 6. Assim a solução passará pelo cálculo das probabilidades individuais para **cada** um desses eventos de sucesso que serão simplesmente somadas pois, a ocorrência de cada um desses eventos de sucesso é disjunta dos demais (se ocorrer 2 não ocorre simultaneamente 3).

$$\begin{aligned} P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\ P(X = 2) &= 15.0, 30^2 \cdot 0, 70^{(6-2)} \\ &= 0,3241 \end{aligned}$$

$$\begin{aligned} P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\ P(X = 3) &= 20.0, 30^3 \cdot 0, 70^{(6-3)} \\ &= 0,1852 \end{aligned}$$

$$\begin{aligned} P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\ P(X = 4) &= 15.0, 30^4 \cdot 0, 70^{(6-4)} \\ &= 0,0595 \end{aligned}$$

$$\begin{aligned} P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\ P(X = 5) &= 6.0, 30^5 \cdot 0, 70^{(6-5)} \\ &= 0,01020 \end{aligned}$$

$$\begin{aligned} P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\ P(X = 6) &= 1.0, 30^6 \cdot 0, 70^{(6-6)} \\ &= 0,000729 \end{aligned}$$

Assim, $P(X \geq 2) = 0,3241 + 0,1852 + 0,0595 + 0,01020 + 0,00079 = 0,5797$

```
# Defina o número de repetições (n) e o número de sucessos (k)

n=6 # Número de repetições
k=c(0,1,2,3,4,5,6) # Número de sucessos varia de nenhum (0) até dez (10) acertar as dez
# → questões
p=0.30 # A probabilidade em cada repetição de Bernoulli

# Probabilidade de k sucessos em n repetições (utilizando a função 'dbinom')

probabilidade <- dbinom(k, n, prob = p) # Neste exemplo, 0.5 é a probabilidade de sucesso

# Crie uma tabela com duas colunas (Número de Sucessos e Probabilidade)

tabela <- data.frame(Número_de_Sucessos = k, Probabilidade = probabilidade)

# Exiba a tabela

print(tabela)
```

```
##   Número_de_Sucessos Probabilidade
## 1                  0      0.117649
## 2                  1      0.302526
## 3                  2      0.324135
## 4                  3      0.185220
## 5                  4      0.059535
## 6                  5      0.010206
## 7                  6      0.000729
```

Exemplo: Uma pessoa trabalha em 3 empregos onde desenvolve atividades iguais, sendo remunerada também igualmente nos três lugares. A probabilidade de que o pagamento saia até o 2º dia útil nos três empregos é de 0,85. Qual a probabilidade de apenas um salário sair até o 2º dia útil?

- 1- a probabilidade de ocorrência do pagamento até o 2º dia útil em cada emprego pode ser modelada por uma variável aleatória de Bernoulli pois apresenta apenas duas possibilidades: ocorrer ou não, cuja probabilidade de sucesso nos foi dada: $p = 0,85$;
- 2- os três empregos podem ser considerados como repetições desse experimento básico;
- 3- esse experimento final pode ter as probabilidades modeladas por uma variável aleatória Binomial com evento de sucesso definido como **chance de se receber apenas um pagamento até o 2º dia útil** ($k = 1$) pois consiste na repetição de ($n = 3$) experimentos de Bernoulli independentes e com probabilidade individual constante ($p = 0,85$).

A probabilidade de se receber o pagamento até o 2º dia útil **em apenas um emprego** será dada por:

$$\begin{aligned} P(X = k) &= C_k^n \cdot p^k \cdot q^{n-k} \\ P(X = 1) &= 3.0,85^1 \cdot 0,15^2 \\ &= 0,0574 \end{aligned}$$

Conclusão: a probabilidade desse trabalhador receber **apenas um salário** até o 2º dia útil do mês é de 0,0574.

```
# Defina o número de repetições (n) e o número de sucessos (k)

n=3 # Número de repetições
k=c(0,1,2,3) # Número de sucessos varia de nenhum (0) até três (3) receber até o segundo
# dia nos três empregos
p=0.85 # A probabilidade em cada repetição de Bernoulli

# Probabilidade de k sucessos em n repetições (utilizando a função 'dbinom')

probabilidade = dbinom(k, n, prob = p) # Neste exemplo, 0,856 é a probabilidade de sucesso

# Crie uma tabela com duas colunas (Número de Sucessos e Probabilidade)

tabela = data.frame(Número_de_Sucessos = k, Probabilidade = probabilidade)

# Exiba a tabela

print(tabela)

##   Número_de_Sucessos Probabilidade
## 1                  0      0.003375
## 2                  1      0.057375
## 3                  2      0.325125
## 4                  3      0.614125
```

6.1.4 Poisson

A distribuição de *Poisson* (assim chamada em homenagem a Siméon Denis Poisson que a descobriu no início do século XIX) é largamente empregada quando se deseja **contar o número de eventos raros** cuja probabilidade média seja dada em termos de um **intervalo de tempo, determinada extensão, área ou volume** (uma taxa).

Uma variável aleatória discreta X com Distribuição de *Poisson* é aquela que pode assumir **infinitos valores numeráveis** ($k = 0, 1, 2, .s, \infty$). Sua representação é: $X \sim Pois(\lambda)$ e sua função de probabilidade para esses valores é:

$$\begin{aligned} f(k) &= P(X = k) \\ &= \frac{\lambda^k \cdot e^{-\lambda}}{k!} \end{aligned}$$

Com $\epsilon = 2,718$ (número irracional de Euler).

A esperança e a variância de uma variável aleatória discreta com Distribuição de *Poisson* são dados pelo seu parâmetro λ que expressa o número médio de eventos ocorrendo no **intervalo de tempo**, ou em uma **determinada extensão, área ou volume**:

- Esperança: $E(X) = \lambda$;
- Variância: $Var(X) = \lambda$

Exemplo: Uma central telefônica recebe em média 5 chamadas por minuto. Supondo que a Distribuição de Poisson seja adequada a esse contexto, obter as probabilidade de que essa central não receba chamadas num intervalo de 1 e que receba no máximo duas chamadas em 4 minutos.

Dados do problema:

- 1- λ = é o parâmetro da distribuição de Poisson (a esperança, a média); assim temos $\lambda = 5$ chamadas por **minuto** (é importante atentar para qual é a unidade associada ao valor do λ);
- 2- **não receber** chamada alguma equivale a um $k = 0$;
- 3- na sequência, ao se perguntar sobre a probabilidade de se receber **no máximo** duas chamadas em **4 minutos** equivale a não receber chamada alguma **ou** uma chamada **ou** duas chamadas (soma das probabilidades de eventos mutuamente excludentes);
- 4- **mas** é necessário reestimar o valor de λ pois agora o intervalo de tempo é de **4 minutos** e o valor que nos foi dado é para **1 minuto** (o que é feito mediante uma simples regra de três: 5 chamadas em **um minuto** passam a ser 20 chamadas em **quatro minutos**)

Probabilidade de **não receber chamada alguma**:

$$\begin{aligned} P(X = k) &= \frac{\lambda^k \cdot e^{-\lambda}}{k!} \\ P(X = 0) &= \frac{5^0 \cdot e^{-5}}{0!} \\ P(X = 0) &= \frac{1.0,00673}{1} \\ &= 0,00673 \end{aligned}$$

Probabilidade de receber no **máximo 2** chamadas em 4 minutos ($\lambda = 20$ chamadas por 4 minutos):

$$P(X = 0) = \frac{20^0 \cdot e^{-20}}{0!} = 2,061154e - 09$$

$$P(X = 1) = \frac{20^1 \cdot e^{-20}}{1!} = 4,122307e - 08$$

$$P(X = 2) = \frac{20^2 \cdot e^{-20}}{2!} = 4,122307e - 07$$

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 4,554699e - 7$$

```
# Defina o número de repetições (n) e o número de sucessos (k)

k= 0      # Número de sucessos varia de nenhum (0) até infinito
l= 5 # A probabilidade de sucesso na proporção dada (área, comprimento, tempo...) 5 chamadas
      # por minuto

k1= c(0,1,2)
l1= 20 # 5 chamadas em 1 minuto >> 20 chamadas em 20 minutos

# Probabilidade de k sucessos em n repetições (utilizando a função 'dbinom')

probabilidade = dpois(k, lambda = l) # Neste exemplo, l=5 chamadas por minuto
probabilidade1 = dpois(k1, lambda = l1) # Neste exemplo, l=20 chamadas por minuto

# Crie uma tabela com duas colunas (Número de Sucessos e Probabilidade)

tabela = data.frame(Número_de_Sucessos = k, Probabilidade = probabilidade)
tabela1 = data.frame(Número_de_Sucessos = k1, Probabilidade = probabilidade1)

# Exiba a tabela

print(tabela)

##   Número_de_Sucessos Probabilidade
## 1                      0     0.006737947

print(tabela1)

##   Número_de_Sucessos Probabilidade
```

```
## 1          0  2.061154e-09
## 2          1  4.122307e-08
## 3          2  4.122307e-07
```

Exemplo: Um posto de bombeiros recebe em média 3 chamadas por dia. Admitindo que as probabilidades associadas ao recebimento de diferentes números de chamadas podem ser modeladas por uma variável aleatória de *Poisson* qual seria a probabilidade desse posto receber 4 chamadas em 2 dias?

A unidade da esperança dessa variável de *Poisson* (λ) de chamadas nos foi dada **por dia** ao passo que a probabilidade pedida está associada a um período de **dois dias**, exigindo que a esperança λ seja convertida para essa nova unidade (uma simples regra de três: 3 chamadas por dia, então para 2 dias, 6 chamadas). Assim, a probabilidade pedida será:

$$\begin{aligned} P(X = k) &= \frac{\lambda^k \cdot e^{-\lambda}}{k!} \\ P(X = 4) &= \frac{6^4 \cdot e^{-6}}{4!} \\ &= 0,1338 \end{aligned}$$

A figura abaixo ilustra a distribuição acumulada das probabilidades de alguns sucessos para o exemplo em estudo.

```
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr    1.0.2     v tidyr    1.3.1
## -- Conflicts ----- tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()   masks scales::discard()
## x dplyr::filter()    masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
prob=c(0.00248, 0.01448, 0.044643, 0.08929, 0.1338, 0.16072, 0.16072, 0.137762, 0.256105)
k=c("k=0", "k=1", "k=2", "k=3", "k=4", "k=5", "k=6", "k=7", "soma(k=8,k=9,...,inf)")
legend_title="Sucessos"
nchamadas=data.frame(sucesso = k, proporcao= prob)
```

```
nchamadas %>%
  mutate(va_poisson = "Probabilidades segundo o modelo teórico de Poisson")
ggplot(nchamadas, aes(x = va_poisson, y = proporcao, fill = forcats::fct_rev(sucesso))) +
  geom_col( width = 0.2) +
  geom_text(aes(label = proporcao),size=3,
            position = position_stack(vjust = 0.5) ) +
  theme(legend.position = "right") +
  ylab("Probabilidade acumulada") +
  xlab(NULL) +
  scale_fill_discrete(name="Número de sucessos",
                      labels=rev(c("k=0", "k=1", "k=2", "k=3", "k=4",
                                 "k=5", "k=6", "k=7", "soma(k=8,k=9,...,inf)")))
```

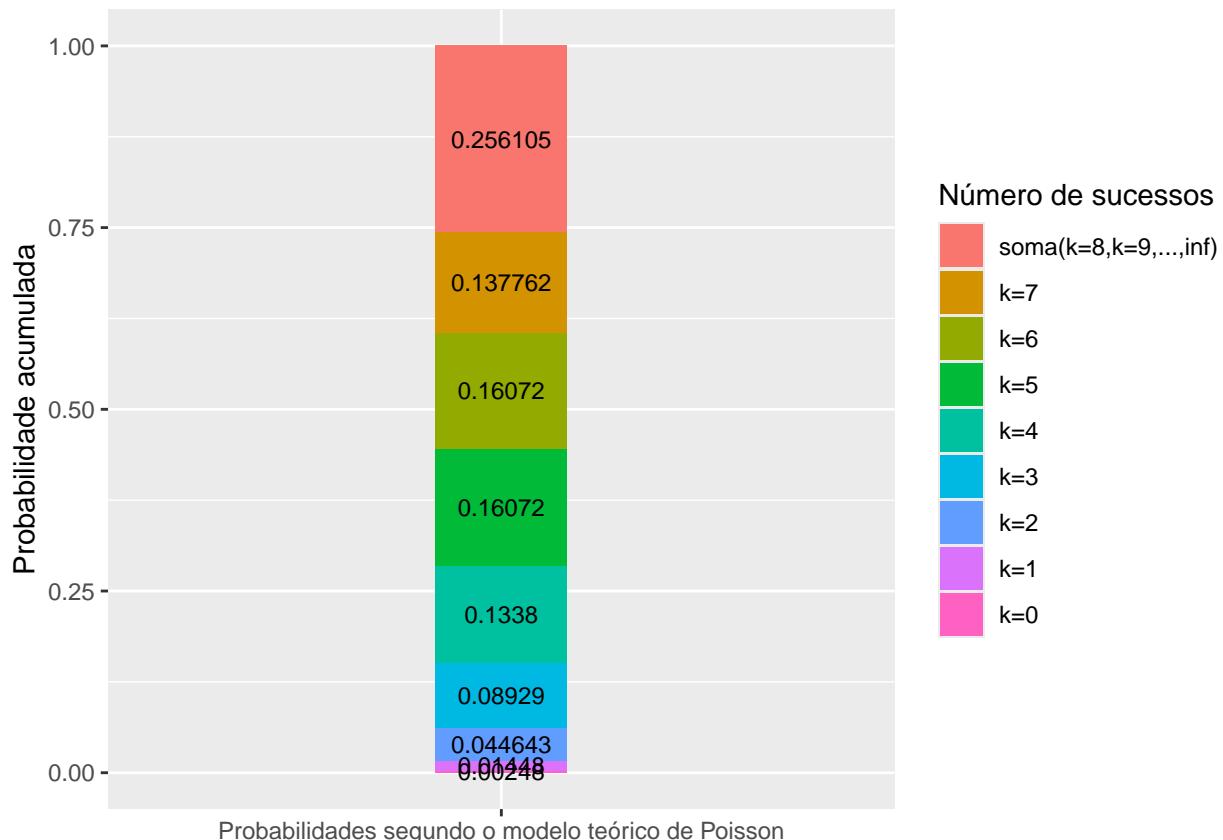


Figure 6.2: Gráfico ilustrativo das probabilidades acumuladas

Exemplo: Por um posto de pedágio passam, em média, 5 carros por minuto. Qual a probabilidade de passarem exatamente 3 carros em 1 minuto?

$$\begin{aligned} P(X = k) &= \frac{\lambda^k \cdot e^{-\lambda}}{k!} \\ P(X = 3) &= \frac{5^3 \cdot e^{-5}}{3!} \\ &= 0,1404 \end{aligned}$$

Uma variável aleatória discreta de *Poisson* modela muito bem eventos raros; ou seja, aqueles que não acontecem com grande frequência para qualquer intervalo considerado (tempo, extensão, área, volume). Trata-se de uma caso de variável Binomial no qual $n \rightarrow \infty$ e p é pequeno ($n \geq 50$ e $n.p \leq (5)$). Nesse cenário pode-se demonstrar que:

$$\lim_{n \rightarrow \infty} P(X) = C_k^n \cdot p^k \cdot q^{n-k}$$

é igual a:

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Tal aproximação era, tempos atrás (antes da era computacional), bastante útil pois, para um n muito grande o cálculo fatorial era trabalhoso! Nesse contexto pode-se modelar o experimento acima, de modo bem aproximado, por uma variável aleatória de Poisson com $\lambda = n.p$:

$$f(k) = P(X = k) = \frac{n.p^k \cdot e^{-n.p}}{k!}$$

6.1.5 Multinomial

A distribuição multinomial é uma generalização da distribuição binomial - a qual admite apenas dois resultados: sucesso e fracasso - para as situações de mais de dois valores. Assim como a distribuição binomial, a distribuição multinomial é uma função de distribuição para processos discretos nos quais prevalecem probabilidades fixas para cada valor gerado de modo independente uns dos outros.

Admita que X seja uma variável aleatória com distribuição multinomial: $X_k \sim \text{multinomial}(n, p)$ que envolve um processo aleatório que possui um conjunto de k possíveis resultados, cada um com sua probabilidade p_k definida:

$$X_k = \begin{cases} X_1 \text{ com probabilidade } p_1 \\ X_2 \text{ com probabilidade } p_2 \\ \vdots \\ X_k \text{ com probabilidade } p_k \end{cases}$$

- X_1, X_2, \dots, X_k são os k possíveis resultados assumidos pela variável aleatória multinomial X_k ;
- p é o vetor de probabilidades p_1, p_2, \dots, p_k associadas à ocorrência de cada um dos possíveis resultados da variável aleatória multinomial X_k ;
- n é o número finito de vezes que o experimento é realizado;
- n_1, n_2, \dots, n_k é o número de sucessos observados em cada um dos possíveis resultados que a variável aleatória multinomial X_k pode assumir, de tal modo que $n_1 + n_2 + \dots + n_k = n$;
- as probabilidades de sucesso p_k em cada uma das repetições são sempre as mesmas: independência de resultados entre as repetições.

A função de distribuição de probabilidades é dada por:

$$f(X = (n_1, n_2, \dots, n_k)) = P(X_1 = n_1; X_2 = n_2, \dots, X_k = n_k)$$

$$P(X_1 = n_1; X_2 = n_2, \dots, X_k = n_k) = \frac{n!}{n_1! n_2! \dots n_k!} \cdot p_1^{n_1} \cdot p_2^{n_2} \dots p_k^{n_k}$$

A esperança e a variância de uma variável aleatória discreta com Distribuição de multinomial são dadas por:

- O vetor esperança: $E(X_k) = n.p_k = \{n.p_1; n.p_2; \dots; n.p_k\};$
- O vetor variância: $Var(X_k) = n.p_k.(1 - p_k) = \{n.p_1.(1 - p_1); n.p_2.(1 - p_2); \dots; n.p_k.(1 - p_k)\}.$

Os valores do vetor da esperança numericamente calculados poderão ser arredondados se o objetivo incluir sua aplicação prática em algum contexto; caso contrário, permanecem como calculados foram.

Exemplo: Em uma partida de xadrez dois jogadores têm a probabilidade de vencer, perder ou empatar. A probabilidade do jogador “A” vencer é 0,40, do jogador “B” vencer é 0,35 e do jogo terminar empatado é de 0,25. Calcule a probabilidade de em 12 partidas, observar-se que o jogador “A” ganhou 7 partidas, o jogador “B” venceu 2 partidas e as 3 partidas restantes terminaram em empate.

Uma variável aleatória X pode ser definida para modelar a probabilidade dos diferentes resultados possíveis nesse experimento aleatório tal que $X_3 \sim multinomial(n, p)$ tal que:

$$X_3 = \begin{cases} X_1: \text{"A" ganha probabilidade } p_1 = 0,40 \\ X_2: \text{"B" ganha probabilidade } p_2 = 0,35 \\ X_3: \text{empate entre "A" e "B" probabilidade } p_3 = 0,25 \end{cases}$$

$$\begin{aligned} P(X_1 = n_1; X_2 = n_2; X_3 = n_3) &= \frac{n!}{n_1!n_2!n_3!} \cdot p_1^{n_1} \cdot p_2^{n_2} \cdot p_3^{n_3} \\ P(X_1 = 7; X_2 = 2; X_3 = 3) &= \frac{12!}{7!2!3!} \cdot 0,40^7 \cdot 0,35^2 \cdot 0,25^3 \\ P(X_1 = 7; X_2 = 2; X_3 = 3) &= 0,02483712 \end{aligned}$$

```
# Defina as probabilidades de sucesso de cada possível resultado da variável aleatória:
→ p_(k). Admita que existam apenas tês:
```

```
p1= 0.4
p2= 0.35
p3= 0.25
```

```
# O número de repetições 'n' é automaticamente definido pelo arranjo de sucessos que se
→ deseja estimar a probabilidade
x=c(7,2,3) # Nesse caso deseja-se 7 sucessos do resultado X_1 (jogador 1), 2 do X_2 (jogador
→ 2) e 3 do X_3 (jogador 3)
```

```
# Probabilidade desse vetor de sucessos (utilizando a função 'dmultinom')
```

```
probabilidade = dmultinom(x=c(7,2,3), prob = c(0.4,0.35,0.25))
print(probabilidade)
```

```
## [1] 0.02483712
```

6.2 Modelos teóricos do tempo de espera

As distribuições do tempo de espera são outra importante classe de problemas associados com a quantidade de tempo que leva para a ocorrência de um evento específico de interesse. Dentro dessa classe de problemas se enquadram duas distribuições bastante conhecidas, são elas: geométrica e Geométrica negativa.

6.2.1 Geométrica

A distribuição geométrica é uma distribuição de probabilidade discreta que modela o número de tentativas independentes necessárias para obter o primeiro sucesso em um processo de Bernoulli, onde cada tentativa tem duas possibilidades: sucesso ou fracasso e a probabilidade de sucesso é constante e denotada por p sob as seguintes condicionantes:

- 1- cada experimento é um ensaio de Bernoulli (só poderá haver dois resultados possíveis: sucesso ou fracasso);
- 2- cada repetição deve ter seu resultado independente do resultado das demais;
- 3- a probabilidade de sucesso (p) é constante para todas as repetições;
- 4- consequentemente, a probabilidade de fracasso ($q = 1 - p$) também o é; e,
- 5- o experimento é repetido segue até que se verifique o primeiro sucesso.

Considere o experimento aleatório de se lançar uma moeda **não honesta**, com probabilidade p de ocorrência de *Cara* e $(1 - p)$ de ocorrência de *Coroa*. Se definimos nosso evento de sucesso como sendo obter *Cara* no lançamento, quantos lançamentos serão necessários para se verificar a ocorrência de sucesso?

Admita uma sequência de n lançamentos: $\{Coroa, Coroa, \dots, Coroa, Cara\}$ onde no $n - \text{simo}$ lançamento verificou-se o sucesso. Assim sendo, podemos definir $j = (n - 1)$ como o número de tentativas **anteriores** fracassadas.

Uma variável aleatória X com Distribuição Geométrica, com parâmetro p ($0 \leq p \leq 1$), é aquela que pode assumir **infinitos valores numeráveis** ($j = 0, 1, 2, \dots, \infty$) para a quantidade j de tentativas que **precedem o primeiro sucesso**, que será observado na tentativa seguinte ($j + 1$).

Sua representação é $X \sim geo(p)$ e sua função de probabilidade é:

$$\begin{aligned} f(X = x; p) &= P(X = j) = p \cdot (1 - p)^j \\ &= P(X = j) = p \cdot q^j \end{aligned}$$

O Modelo geométrico pode ser escrito sob uma “forma alternativa”: o **número de tentativas n até se observar o primeiro sucesso**, agora com $x = n = 1, 2, \dots$.

$$\begin{aligned} f(X = x; p) &= P(X = n) = p \cdot (1 - p)^{(n-1)} \\ &= P(X = n) = p \cdot q^{(n-1)} \end{aligned}$$

A esperança e a variância de uma variável aleatória discreta com Distribuição geométrica ($X \sim geo(p)$) são:

- Esperança: $E(X) = \frac{1}{p}$
- Variância: $Var(X) = \frac{(1-p)}{p^2} = \frac{q}{p^2}$.

A distribuição geométrica é frequentemente usada em situações em que estamos interessados em calcular quantas tentativas independentes são necessárias até que um evento específico ocorra. Por exemplo, pode ser usada para modelar o número de lançamentos de uma moeda justa até que a primeira cara apareça, ou o número de tentativas até que um cliente faça sua primeira compra em um site de comércio eletrônico.

Lembrando que o modelo binomial é aplicado sobre a contagem de número de sucessos k em n tentativas de Bernoulli; ou seja, o número de tentativas n é **fixo** e o número de sucessos k é **aleatório**.

Já o modelo geométrico estima o número de tentativas j até se observar o primeiro sucesso; isto é, o número de sucessos k é **fixo** e o número de tentativas j é **aleatório**.

Uma variável aleatória geométrica é definida como o número de tentativas até que o primeiro sucesso fosse encontrado e, como essas tentativas são independentes entre si; ie., a probabilidade p não se altera em razão de terem sido realizadas tentativas anteriores, a contagem do número de tentativas até o próximo sucesso pode ser começada em qualquer tentativa sem alterar a distribuição de probabilidades da variável aleatória. A consequência de usar um modelo geométrico é que o sistema presumivelmente não será desgastado, a probabilidade permanece constante.

Nesse sentido à distribuição geométrica é dita **faltar qualquer memória**.

Exemplo: A probabilidade de que um *bit* transmitido através de um canal digital seja recebido **com erro** é de 0,1. Considere que as transmissões sejam eventos independentes e o erro relativamente raro. Uma variável aleatória discreta pode ser definida como $X \sim Geo(p)$. Qual a probabilidade de que **o primeiro erro** na transmissão de um *bit* ocorra na **quinta** transmissão?

Uma variável aleatória discreta com Distribuição geométrica pode ser definida para modelar a probabilidade desse experimento aleatório como $X \sim geo(p)$, onde p é a probabilidade individual de sucesso (no nosso caso, que o *bit* seja transmitido com erro).

Dados do problema:

1- a probabilidade de ocorrência de um sucesso (aqui bem entendido como sendo a transmissão de um *bit* com erro) é $p = 0,1$; e,

2- a probabilidade pedida é a de se observar a ocorrência do primeiro sucesso com 5 repetições (bem entendido aqui que o número de tentativas **sem se observar sucesso** será $j = 4$ e, em $j + 1 = 5$ teremos sucesso).

$$\begin{aligned}f(X = x; p) &= P(X = j) = (1 - p)^j \cdot p \\P(X = 4) &= (1 - 0,1)^4 \cdot 0,1 \\P(X = 4) &= 0,0656\end{aligned}$$

A probabilidade de que na **quinta transmissão** de um *bit* ocorra um erro é de 6,56%.

```
p = 0.10
n = 4 # Uma vez que na 5 repetição ocorrerá o 'sucesso' (o bit será transmitido com erro)
dgeom(x = n, prob = p)
```

```
## [1] 0.06561
```

Exemplo: Uma linha de produção está sendo analisada para fins de controle da qualidade das peças produzidas. Tendo em vista o alto padrão requerido, a produção é interrompida para regulagem **toda vez que uma peça defeituosa é observada**. Se 0,01 é a probabilidade da peça ser defeituosa, determine a probabilidade de ocorrer uma peça defeituosa entre a 4^a e 6^a peças produzidas.

Uma variável aleatória discreta com Distribuição geométrica pode ser definida para modelar esse experimento aleatório como $X \sim Geo(p)$ onde p é a probabilidade individual de sucesso (no caso, a produção de uma peça defeituosa). Pede-se a probabilidade de que essa ocorrência se verifique **OU** na quarta **OU** na quinta **OU** na setxa peça produzida.

Dados do problema:

- a probabilidade de ocorrência de um sucesso (aqui bem entendido como sendo a produção de uma peça defeituosa) é $p = 0,01$; e,

- a probabilidade pedida é a de se observar a ocorrência da produção da primeira peça defeituosa com 4, 5 **OU** 6 repetições.

Assim sendo o número de tentativas **sem se ter nenhuma peça produzida com defeito** é de $3 \leq j \leq 5$ porque assim, em $j + 1$, teremos sucesso na quarta, quinta ou sexta peça produzidas.

Considerando-se que os eventos são disjuntos (ocorrerá na quarta, na quinta ou na sexta), probabilidade pedida será:

$$P(X = j)_{3 \leq j \leq 5} = P(X = 3) + P(X = 4) + P(X = 5)$$

A probabilidade de verificarnos sucesso na 4^a peça produzida (peça produzida com defeito) será:

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 3) &= (1 - 0,01)^3 \cdot 0,01 \\ P(X = 3) &= 0,00970299 \end{aligned}$$

A probabilidade de verificarnos sucesso na 5^a peça produzida (peça produzida com defeito) será:

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 4) &= (1 - 0,01)^4 \cdot 0,01 \\ P(X = 4) &= 0,00960596 \end{aligned}$$

A probabilidade de verificarnos sucesso na 6^a peça produzida (peça produzida com defeito) será:

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^k \cdot p \\ P(X = 5) &= (1 - 0,01)^5 \cdot 0,01 \\ P(X = 5) &= 0,0095099 \end{aligned}$$

A probabilidade de termos uma peça **produzida com defeito** na quarta **OU** na quinta **OU** na sexta das peças produzidas será:

$$\begin{aligned}P(3 \leq j \leq 5) &= P(X = 3) + P(X = 4) + P(X = 5) \\P(3 \leq j \leq 5) &= 0.00970299 + 0.00960596 + 0.00950990 \\P(3 \leq j \leq 5) &= 0.02881885\end{aligned}$$

A probabilidade de termos uma **peça defeituosa** na quarta **OU** na quinta **OU** na sexta das peças produzidas é de 2,9116%.

```
p = 0.01
n = c(3,4,5) # Uma vez que na 4, 5 e 6 repetições ocorrerão os 'sucessos' (a produção de uma
  ↵ peça defeituosa)
dgeom(x = n, prob = p)
```

```
## [1] 0.00970299 0.00960596 0.00950990
```

Exemplo 9 A probabilidade de um alinhamento ótico bem sucedido na montagem de produto de armazenamento de dados é de 0,80. Assuma que as tentativas são independentes e responda:

- 1- Qual é a probabilidade de que o primeiro alinhamento bem sucedido requeira exatamente quatro tentativas?
- 2- Qual é a probabilidade de que o primeiro alinhamento bem sucedido requeira no máximo quatro tentativas?
- 3- Qual é a probabilidade de que o primeiro alinhamento bem sucedido requeira ao menos quatro tentativas?

Uma variável aleatória discreta com Distribuição geométrica pode ser definida para modelar esse experimento aleatório como $X \sim Geo(p)$ onde p é a probabilidade individual de sucesso .

Dados do problema:

- a probabilidade de ocorrência de um sucesso (alinhamento ótico bem sucedido na montagem de produto de armazenamento de dados) é $p = 0,80$;
- o item (1) pede a probabilidade de verificar o primeiro sucesso com exatamente **quatro repetições**; assim, o número de tentativas **sem se observar sucesso** é $j = 3$ (em $j + 1 = 4$ verifica-se sucesso);
- o item (2) pede a probabilidade de se verificar o primeiro sucesso com **no máximo** quatro repetições; assim, o número de tentativas **sem se observar sucesso** é de $0 \leq j \leq 3$ (em $j + 1$ teremos sucesso: no primeiro **OU** no segundo **OU** no terceiro **OU** no quarto alinhamentos realizados); e,

- o item (3) pede a probabilidade de se observar o primeiro sucesso com **no mínimo quatro** repetições; assim, o número de tentativas **sem se observar sucesso** é de $\$3 \leq j \leq \infty\$$ (em $j + 1$ teremos sucesso: no quarto **OU*** **no quinto** **OU** sexto** alinhamentos realizados).

Para o item (1) a probabilidade de termos a ocorrência de um sucesso (ou seja, um alinhamento ótico bem sucedido) na 4^a montagem será:

$$\begin{aligned}f(X = x; p) &= P(X = j) \\P(X = j) &= (1 - p)^j \cdot p \\P(X = 3) &= (1 - 0,80)^3 \cdot 0,80 \\P(X = 3) &= 0,0064\end{aligned}$$

```
p = 0.80
n = c(3) # Uma vez que na 4 repetição ocorrerá o 'sucesso' (um alinhamento ótico bem
           ↵ sucedido)
dgeom(x = n, prob = p)
```

```
## [1] 0.0064
```

Para o item (2) considerando-se que as repetições são independentes, a probabilidade pedida será:

$$P(X = j)_{0 \leq j \leq 3} = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$\begin{aligned}f(X = x; p) &= P(X = j) \\P(X = j) &= (1 - p)^j \cdot p \\P(X = 0) &= (1 - 0,80)^0 \cdot 0,80 \\P(X = 0) &= 0,80\end{aligned}$$

$$\begin{aligned}f(X = x; p) &= P(X = j) \\P(X = j) &= (1 - p)^j \cdot p \\P(X = 1) &= (1 - 0,80)^1 \cdot 0,80 \\P(X = 1) &= 0,16\end{aligned}$$

$$\begin{aligned}
 f(X = x; p) &= P(X = j) \\
 P(X = j) &= (1 - p)^j \cdot p \\
 P(X = 2) &= (1 - 0,80)^2 \cdot 0,80 \\
 P(X = 2) &= 0,032
 \end{aligned}$$

$$\begin{aligned}
 f(X = x; p) &= P(X = j) \\
 P(X = j) &= (1 - p)^j \cdot p \\
 P(X = 3) &= (1 - 0,80)^3 \cdot 0,80 \\
 P(X = 3) &= 0,0064
 \end{aligned}$$

A probabilidade pedida é de:

$$\begin{aligned}
 P(X = j)_{0 \leq j \leq 3} &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\
 P(X = j)_{0 \leq j \leq 3} &= 0,9984
 \end{aligned}$$

```

p = 0.80
n = c(0,1,2,3) # Uma vez que na 1, 2, 3, e 4 repetições ocorrerão os 'sucessos'
                  # (alinhamentos óticos bem sucedidos)
dgeom(x = n, prob = p)

```

```
## [1] 0.8000 0.1600 0.0320 0.0064
```

Para o item (3) considerando-se que os eventos pedidos são disjuntos a probabilidade pedida deverá ser calculada a partir do complemento da probabilidade total menos os eventos que não são de interesse:

$$P(X = j)_{3 \leq j \leq \infty} = 1 - P(X = 0) + P(X = 1) + P(X = 2)$$

$$\begin{aligned}
 f(X = x; p) &= P(X = j) \\
 P(X = j) &= (1 - p)^j \cdot p \\
 P(X = 0) &= (1 - 0,80)^0 \cdot 0,80 \\
 P(X = 0) &= 0,80
 \end{aligned}$$

$$\begin{aligned}
 f(X = x; p) &= P(X = j) \\
 P(X = j) &= (1 - p)^j \cdot p \\
 P(X = 1) &= (1 - 0,80)^1 \cdot 0,80 \\
 P(X = 1) &= 0,16
 \end{aligned}$$

$$\begin{aligned}
 f(X = x; p) &= P(X = j) \\
 P(X = j) &= (1 - p)^j \cdot p \\
 P(X = 2) &= (1 - 0,80)^2 \cdot 0,80 \\
 P(X = 2) &= 0,032
 \end{aligned}$$

A probabilidade é de:

$$\begin{aligned}
 P(X = j)_{3 \leq j \leq \infty} &= 1 - P(X = 0) + P(X = 1) + P(X = 2) \\
 P(X = j)_{3 \leq j \leq \infty} &= 1 - (0,80 + 0,16 + 0,032) \\
 P(X = j)_{3 \leq j \leq \infty} &= 0,008
 \end{aligned}$$

6.2.2 Binomial Negativa

A distribuição Binomial Negativa (também conhecida como de Distribuição de Pascal em homenagem ao matemático francês Blaise Pascal) pode ser considerada como uma generalização da variável Geométrica, na qual agora considera-se a situação em que se modelam as probabilidades de se verificar mais de um evento de sucesso em um certo número de repetições.

Ao se realizar repetidos experimentos de Bernoulli, uma variável aleatória Binomial Negativa modela as probabilidades de serem observados k sucessos em n repetições.

Um experimento que apresenta uma distribuição Binomial Negativa satisfaz aos seguintes pressupostos:

- 1- cada repetição é um ensaio de Bernoulli (só poderá haver dois resultados possíveis: sucesso ou fracasso);
- 2- cada repetição não altera a probabilidade original (há independência entre as repetições);
- 3- portanto, a probabilidade de sucesso (p) em cada repetição é constante;
- 4- e, consequentemente, a probabilidade de fracasso ($q = 1 - p$) em cada repetição também é constante; e,
- 5- o experimento aleatório prossegue até que sejam verificados k sucessos (na última repetição terá sido observado o k -ésimo sucesso).

A notação de uma variável aleatória Binomial Negativa é $X \sim bn(p, k)$, onde o parâmetro p ($0 \leq p \leq 1$) indica a probabilidade individual de sucesso a cada repetição de Bernoulli e k o número total de sucessos desejado e estabelecido *a priori*.

Sua função discreta de probabilidade, que calcula a probabilidade de se observar o número k sucessos estabelecido *a priori* em n de ensaios de Bernoulli realizados, é a seguinte:

$$\begin{aligned} f(X = n; (p, k)) &= P(X = n; (p, k)) = \binom{n-1}{k-1} \cdot p^k \cdot q^{(n-k)} \\ &= \frac{(n-1)!}{(k-1)! \cdot (n-k)!} \cdot p^k \cdot q^{(n-k)} \end{aligned}$$

Como são necessários no mínimo k tentativas para se obter k sucessos, os valores de $n = k, k+1, k+2, \dots$.

A esperança e a variância de uma variável aleatória discreta com Distribuição Binomial Negativa são:

- Esperança: $E(X) = \frac{k}{p}$;
- Variância: $Var(X) = \frac{k \cdot (1-p)}{p^2} = \frac{q \cdot k}{p^2}$.

Uma variável aleatória Binomial é uma contagem de número de sucessos k em n tentativas de Bernoulli; ou seja, o número de tentativas n é predeterminado (fixo) e o número de sucessos k é a variação aleatória. Em n tentativas a probabilidade de se observar k sucessos é calculada pela sua função de distribuição discreta de probabilidades.

Uma variável aleatória Binomial Negativa é uma contagem do número de tentativas n de Bernoulli em k sucessos; ou seja, o número de sucessos k é predeterminado (fixo) e o número de tentativas é a variação aleatória. Em n tentativas a probabilidade de se observar k sucessos é calculada pela sua função de distribuição discreta de probabilidades.

Exemplo 10: A probabilidade com que um *bit* transmitido através de um canal digital de transmissão seja recebido com erro é de 0,1 e que as transmissões sejam eventos independentes. Qual a probabilidade de que nas dez primeiras transmissões ocorram quatro erros?

Uma variável aleatória discreta com Distribuição Binomial Negativa pode ser definida para modelar esse experimento aleatório, tal que $X \sim bn(p, k)$ onde p é a probabilidade individual de sucesso e k o número de sucessos estabelecido *a priori*.

Dados do problema:

- 1- a probabilidade de ocorrência de um sucesso (aqui bem entendido como sendo a recepção errada de um *bit* transmitido) é $p = 0,1$; e,
- 2- o número de sucessos (aqui bem entendido como sendo a recepção errada de um *bit* transmitido) está definido em $k = 4$.

Pede-se a probabilidade de se observar quatro sucessos ($k = 4$) em dez ($n = 10$) transmissões (repetições).

A probabilidade de se observar $k = 4$ sucessos ao se realizar $n = 10$ tentativas de Bernoulli é dada pela função discreta de probabilidade da variável aleatória Binomial Negativa:

$$\begin{aligned}
 f(X = n; (p, k)) &= P(X = 10; (p = 0,1; k = 4)) = \binom{n-1}{k-1} \cdot p^k \cdot q^{n-k} \\
 &= \frac{(n-1)!}{(k-1)! \cdot (n-k)!} \cdot p^k \cdot q^{n-k} \\
 &= \frac{(10-1)!}{(4-1)! \cdot (10-4)!} \cdot 0,1^4 \cdot 0,9^{10-4} \\
 &= \frac{362880}{6 \cdot 720} \cdot 0.0001 \cdot 0.531441 \\
 &= 0,004464104
 \end{aligned}$$

A probabilidade de se observar 4 sucessos em 10 tentativas é de 0,4464104%.

```

k = 4
n = 10
p = 0.10
x = n-k # Entrar com n-k na função pois ela espera o número de falhas
dnbinom(x = x , size = k, prob = p)

```

```
## [1] 0.004464104
```

Exemplo 11: Bob é um jogador de basquete de uma escola. Ele é um lançador de arremessos livres e sua probabilidade de acertar é igual a 70%. Durante uma partida qualquer, qual a probabilidade de que Bob acerte seu **terceiro** arremesso livre na sua **quinta** tentativa?

Uma variável aleatória discreta com Distribuição Binomial Negativa pode ser definida para modelar esse experimento aleatório tal que $X \sim bn(p, k)$ onde p é a probabilidade individual de sucesso e k o número de sucessos estabelecido *a priori* sob probabilidade constante e igual a p a cada repetição (p, k são os parâmetros do modelo).

Dados do problema:

- 1- a probabilidade de ocorrência de um sucesso é $p = 0,70$, e
- 2- o número de sucessos =3\$.

Pede-se a probabilidade de se observar três sucessos ($k = 3$) em cinco ($n = 5$) arremessos(repetições).

A probabilidade de se obter $k = 3$ sucessos ao se realizar $n = 5$ tentativas é dada pela função discreta de probabilidade da variável Binomial Negativa:

$$\begin{aligned}
f(X = n; (p; r)) &= P(X = 5; (p = 0,7; k = 3)) = \binom{n-1}{k-1} \cdot p^k \cdot q^{n-k} \\
&= \frac{(n-1)!}{(k-1)! \cdot (n-k)!} \cdot p^k \cdot q^{n-k} \\
&= \frac{(5-1)!}{(3-1)! \cdot (5-3)!} \cdot 0,7^3 \cdot 0,3^{5-3} \\
&= \frac{24}{2 \cdot 2} \cdot 0,343 \cdot 0,09 \\
&= 0,18522
\end{aligned}$$

A probabilidade de Bob acertar 3 arremessos em 5 tentativas é de 18,522%.

```
k = 3
n = 5
p = 0.70
x = n-k # Entrar com n-k na função pois ela espera o número de falhas
dnbinom(x = x , size = k, prob = p)
```

```
## [1] 0.18522
```

Exemplo 12: Lançamos repetidas vezes uma moeda. Seja X o número de caras até que consigamos sete coroas. Qual é a probabilidade de que o número de caras seja igual a cinco até que consigamos as sete coroas?

Uma variável aleatória discreta com Distribuição Binomial Negativa pode ser definida para modelar esse experimento aleatório tal que $X \sim bn(p, k)$ onde p é a probabilidade individual de sucesso e k o número de sucessos estabelecido *a priori* sob probabilidade constante e igual a p a cada repetição (p, k são os parâmetros do modelo).

Dados do problema:

- a probabilidade de ocorrência de um sucesso é $p = 0,5$, e,
- o número de sucessos é $k = 7$.

Pede-se a probabilidade de se observar sete sucessos (sete Caras) em doze tentativas (sete Caras e cinco Coroas).

A probabilidade de se obter $k = 7$ sucessos ao se realizar $n = 12$ tentativas é dada pela função discreta de probabilidade da variável Binomial Negativa:

$$\begin{aligned}
 f(X = n; p; r) &= P(X = 12; p = 0.5; k = 7) = \binom{n-1}{k-1} \cdot p^k \cdot q^{n-k} \\
 &= \frac{(n-1)!}{(r-1)! \cdot (n-k)!} \cdot p^k \cdot q^{n-k} \\
 &= \frac{(12-1)!}{(7-1)! \cdot (12-7)!} \cdot 0,5^7 \cdot 0,5^{12-7} \\
 &= \frac{39916800}{720 \cdot 120} \cdot 0.0078125 \cdot 0.03125 \\
 &= 462 \cdot 0.0078125 \cdot 0.03125 \\
 &= 0.112793
 \end{aligned}$$

A probabilidade de se obter 7 sucessos em 12 tentativas é de 11,28%.

```

k = 7
n = 12
p = 0.50
x = n-k # Entrar com n-k na função pois ela espera o número de falhas
dnbinom(x = x , size = k, prob = p)

```

```
## [1] 0.112793
```

Exemplo 13: Considere o tempo para recarregar o flash de uma câmera de celular. Assuma que a probabilidade de que uma câmera instalada no celular durante sua montagem passe no teste seja de 0.80 e que cada câmera é montada de modo que a probabilidade não se altere (independência). Determine as seguintes probabilidades:

- 1- de que a segunda falha ocorra na décima câmera testada;
- 2- de que a segunda falha ocorra no teste de quatro ou menos câmeras; e,
- 3- o valor esperado do número de câmeras testadas para obter a terceira falha.

Uma variável aleatória discreta com Distribuição Binomial Negativa pode ser definida para modelar esse experimento aleatório tal que $X \sim bn(p, k)$ onde p é a probabilidade individual de sucesso e k o número de sucessos estabelecido *a priori* sob probabilidade constante e igual a p a cada repetição (p, k são os parâmetros do modelo).

Dados do problema:

- se a probabilidade de que a câmera montada no celular passe no teste é 0,80 a probabilidade de não passar é de $(1 - 0,80) = 0,20$;

- fica bem entendido aqui que o **sucesso** é a câmera montada no celular **não passar** no teste, logo $p = 0,20$;
- no item (1) pede-se a probabilidade de se observar um número de sucessos fixado *a priori* $k = 2$ em $n = 10$ câmeras testadas (repetições de Bernoulli);
- no item (2) pede-se a probabilidade de se observar um número de sucessos também fixado *a priori* em $k = 2$ mas agora no intervalo de $n \leq 4$ câmeras testadas (repetições de Bernoulli); e,
- o valor esperado para o número de câmeras testadas (repetições de Bernoulli) ($n = ?$) para que se observem $k = 3$ sucessos.

A probabilidade de se observar $k = 2$ sucessos ao se realizar $n = 10$ tentativas de Bernoulli é dada pela função discreta de probabilidade da variável Binomial Negativa:

/

$$\begin{aligned}
f(X = n; (p, k)) &= P(X = 10; (p = 0.2; k = 2)) = \binom{n-1}{k-1} \cdot p^k \cdot q^{n-k} \\
&= \frac{(n-1)!}{(k-1)! \cdot (n-k)!} \cdot p^k \cdot q^{n-k} \\
&= \frac{(10-1)!}{(2-1)! \cdot (10-2)!} \cdot 0,2^2 \cdot 0,8^{10-2} \\
&= \frac{362880}{1 \cdot 40320} \cdot 0,04 \cdot 0,1677722 \\
&= 9 \cdot 0,04 \cdot 0,1677722 \\
&= 0,06039799
\end{aligned}$$

A probabilidade de se observar $k = 2$ sucessos em $n = 10$ tentativas de Bernoulli é de 6,039%.

As probabilidades de se observar $k = 2$ sucessos ao serem realizadas $n \leq 4$ tentativas é expressa por \$P(X=2) P(X=3) P(X=4)=P(X=2)+P(X=3)+P(X=4))\$, dadas pela função discreta de probabilidade da variável Binomial Negativa aplicada a:

$$\begin{aligned}
f(X = n; (p, k)) &= P(X = 2; (p = 0.2; k = 2)) = \binom{n-1}{k-1} \cdot p^k \cdot q^{n-k} \\
&= \frac{(2-1)!}{(2-1)! \cdot (2-2)!} \cdot 0,2^2 \cdot 0,8^{2-2} \\
P(X = 2) &= 0,04
\end{aligned}$$

$$\begin{aligned}
 f(X = n; (p, k)) &= P(X = 3; (p = 0.2; k = 2)) = \binom{n-1}{k-1} \cdot p^k \cdot q^{n-k} \\
 &= \frac{(3-1)!}{(2-1)! \cdot (3-2)!} \cdot 0,2^2 \cdot 0,8^{3-2} \\
 P(X = 3) &= 0,064
 \end{aligned}$$

$$\begin{aligned}
 f(X = n; (p, k)) &= P(X = 4; (p = 0.2; k = 2)) = \binom{n-1}{k-1} \cdot p^k \cdot q^{n-k} \\
 &= \frac{(4-1)!}{(2-1)! \cdot (4-2)!} \cdot 0,2^2 \cdot 0,8^{4-2} \\
 P(X = 4) &= 0,0768
 \end{aligned}$$

A probabilidade de se obter $r = 2$ sucessos em $n \leq 4$ tentativas é de $(0,04 + 0,064 + 0,0768)$ 18,08%.

O valor esperado (esperança) do número de câmeras testadas para que se observem $k = 3$ sucessos é dado

$$\begin{aligned}
 E(X) &= \frac{k}{p} \\
 E(X) &= \frac{3}{0,2} \\
 E(X) &= 15
 \end{aligned}$$

O valor esperado (esperança) do número n de câmeras testadas para que se observem $r = 3$ sucessos é 15

6.3 Modelos teóricos contínuos

Experimentos aleatórios nos quais os possíveis resultados assumem valores resultantes de processos de mensuração tais como, por exemplo, rendas, pesos, velocidades, tempos, comprimentos, pertencentes aos números Reais, podem ser adequadamente modelados por variáveis aleatórias contínuas.

Para estes uma função densidade de probabilidade é definida de modo a retornar a probabilidade de ocorrência associada a um intervalo de valores, posto a probabilidade exata de ocorrência de um valor aleatório contínuo tender a zero ($P(X = x) \rightarrow 0$).

A função $f(x)$ é uma função densidade de probabilidade para a variável aleatória contínua X se atende às seguintes condições relacionadas aos axiomas da probabilidade:

- $f(x) \geq 0$ para todo $x \in (-\infty, \infty)$;
- a área definida por $f(x)$ é igual a 1 (área sob $f(x)$ e acima do eixo x).

Para tornar o conceito mais compreensível admita a função densidade de probabilidade (fdp) a seguir e sua representação gráfica na Figura 6.3

$$f(X = x) = \begin{cases} 2x & \text{para } 0 \leq x \leq 1 \\ 0, & \text{para qualquer outro } x \end{cases}$$

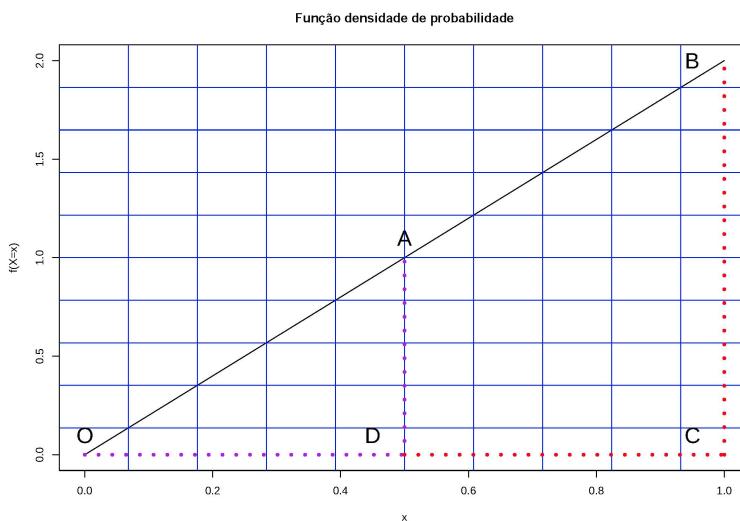


Figure 6.3: A área definida por (ODA) equivale à probabilidade de $f(X = x)$ no intervalo $0 \leq x \leq 0,50$ é notadamente menor que a área definida por (ABCD) equivalente à probabilidade de $f(X = x)$ no intervalo $0,5 \leq x \leq 1$. Tendo os intervalos $[0;0,50]$ e $[0,50; 1,00]$ igual amplitude, depreende-se que uma fdp é uma função indicadora da concentração massa (probabilidade) nos possíveis valores de X

6.3.1 Uniforme

A Distribuição Uniforme é uma das distribuições contínuas mais simples de toda a Estatística. Ela se caracteriza por ter uma função densidade contínua em um intervalo fechado $[a, b]$. Ou seja, a probabilidade de ocorrência de um certo valor é sempre a mesma.

Embora as aplicações desta distribuição não sejam tão abundantes quanto as demais distribuições que discutiremos mais adiante, utilizaremos a Distribuição Uniforme para introduzirmos as funções contínuas e darmos uma noção de como se utiliza a função densidade para determinarmos probabilidades, esperanças e variâncias.

Uma variável aleatória X tem Distribuição Uniforme no intervalo $[a, b]$, com notação $X \sim U(a, b)$, se sua função densidade de probabilidade for dada por:

$$f(X = x) = \begin{cases} \frac{1}{b-a}, & \text{para } a \leq x \leq b \\ 0, & \text{para qualquer outro } x \end{cases}$$

A esperança e a variância de uma variável aleatória contínua com Distribuição Uniforme são:

- Esperança: $E(X) = \frac{(a+b)}{2}$; e,
- Variância: $Var(X) = \frac{(b-a)^2}{12}$.

A probabilidade para um intervalo $[c, d]$ tal que $a \leq c < d \leq b$ será dada por:

$$\int_c^d \frac{1}{(b-a)} dx \frac{1}{(b-a)} \int_c^d 1 dx \frac{1}{(b-a)} \Big|_c^d \frac{(d-c)}{(b-a)}$$

Exemplo 14: Verifique se as funções a seguir atendem os pressupostos necessários para ser uma função densidade de probabilidade (assuma que toda $f(x) = 0$ para valores fora dos intervalos especificados):

1- $f(x) = 3x$ para $0 \leq x \leq 1$;

2- $f(x) = \frac{x^2}{2}$ para $x \geq 0$;

3- $f(x) = \frac{(x-3)}{2}$ para $3 \leq x \leq 5$;

4- $f(x) = 2$ para $0 \leq x \leq 2$;

5-

$$f(X = x) = \begin{cases} \frac{(2+x)}{4}, & \text{para } -2 \leq x \leq 0 \\ \frac{(2-x)}{4}, & \text{para } 0 \leq x \leq 2 \end{cases}$$

6- $f(x) = -\pi$ para $-\pi < x < 0$

Os gráficos das funções densidade de probabilidade são:

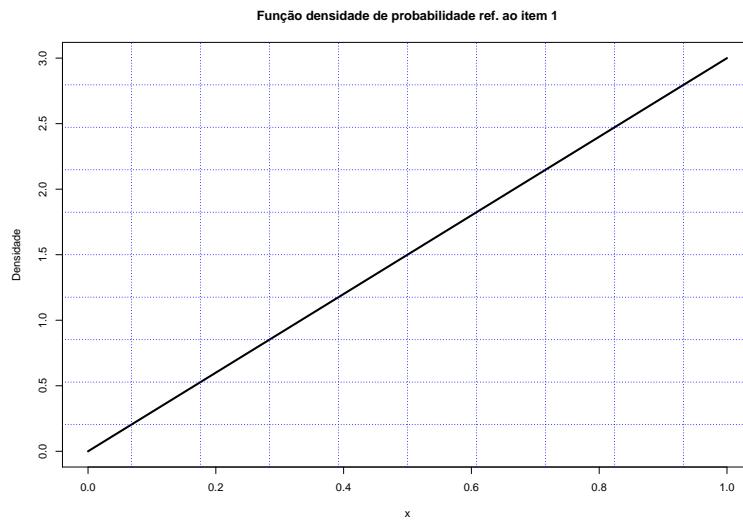


Figure 6.4: A área definida por $f(x)$ no intervalo $0 \leq x \leq 1$ é maior que 1. Por essa razão não pode ser uma fdp

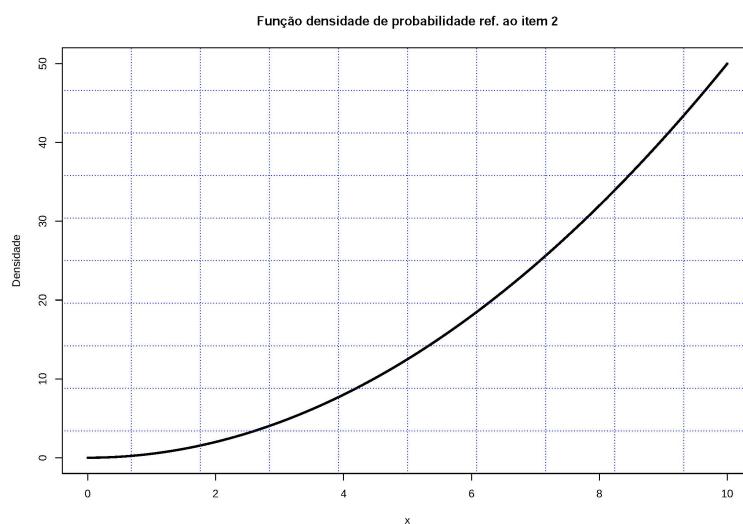


Figure 6.5: A área definida por $f(x)$ no intervalo $x \geq 0$ é maior que 1. Por essa razão não pode ser uma fdp

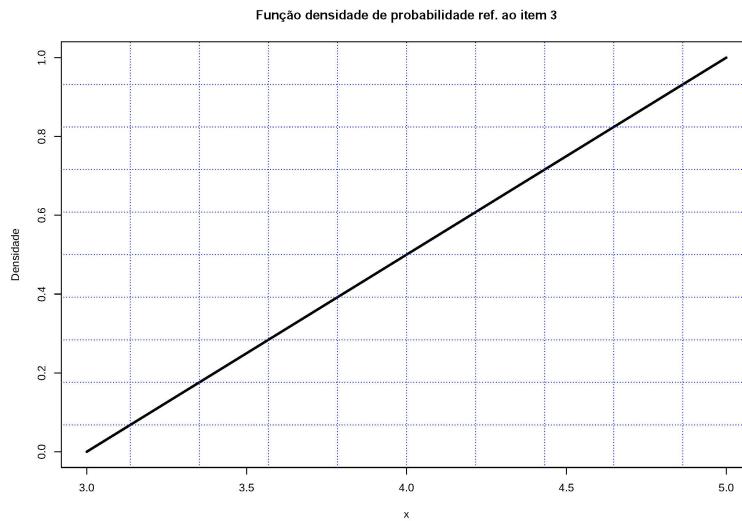


Figure 6.6: Os valores assumidos por $f(x)$ são ≥ 0 e a área definida por $f(x)$ o intervalo $3 \leq x \leq 5$ é igual a 1. Por essa razão pode ser uma fdp

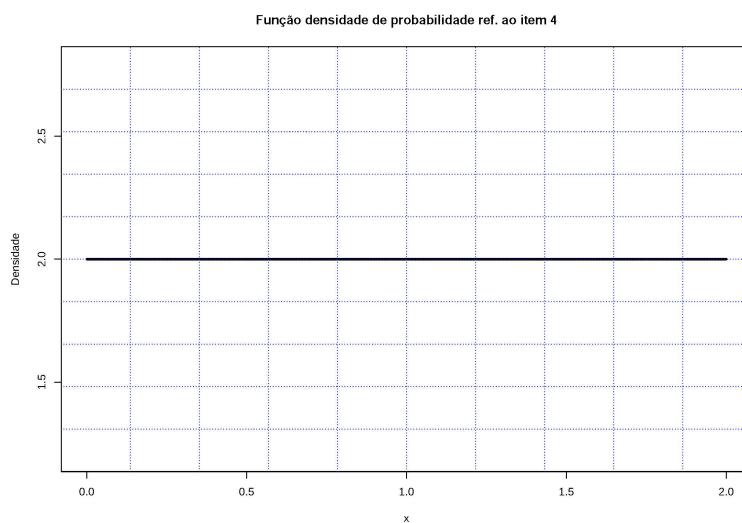


Figure 6.7: A área definida por $f(x)$ no intervalo $0 \leq x \leq 2$ é maior que 1. Por essa razão não pode ser uma fdp

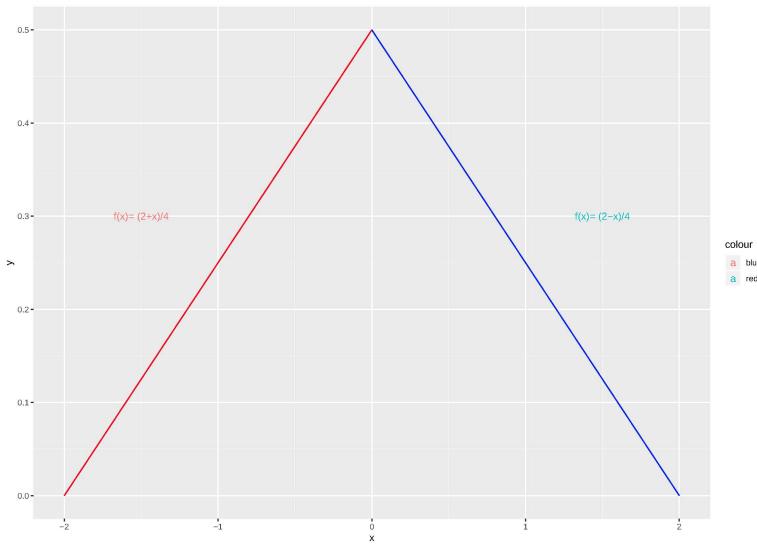


Figure 6.8: Os valores assumidos por $f(x)$ são ≥ 0 e a área definida por $f(x)$ nos intervalos $-2 \leq x \leq 0$ e $0 \leq x \leq 2$ é igual a 1. Pode ser uma fdp

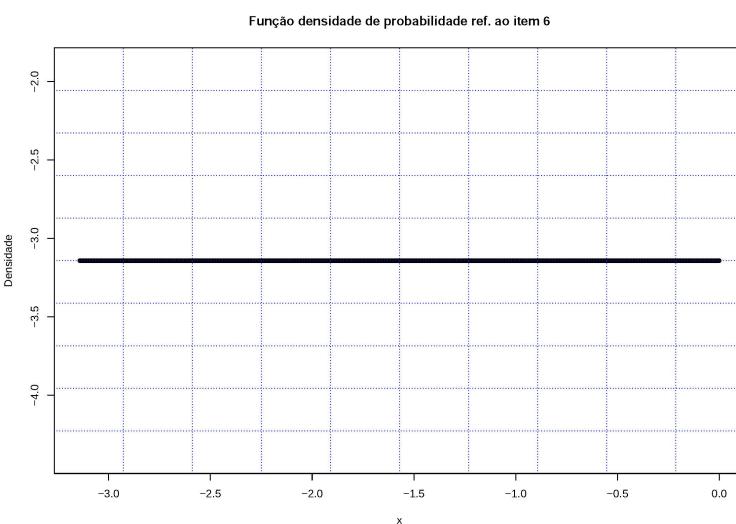


Figure 6.9: Os valores assumidos por $f(x)$ são < 0 . Por essa razão não pode ser uma fdp.

Exemplo: A dureza X de uma peça de aço pode ser entendida como sendo uma variável aleatória contínua uniforme no intervalo $(50, 70)$ da escala Rockwel. Calcule a esperança e a variância dessa variável aleatória e a probabilidade de que uma peça tenha dureza entre 55 e 60?

Definindo a variável aleatória contínua $X : X \sim U(50, 70)$:

$$f(X = x) = \begin{cases} \frac{1}{70-50} = \frac{1}{20}, & \text{para } 50 \leq x \leq 70 \\ 0, & \text{para qualquer outro } x \end{cases}$$

Sua esperança e a variância são:

- Esperança: $E(X) = \mu = \frac{(70+50)}{2} = 60$; e,
- Variância: $Var(X) = \frac{(70-50)^2}{12} = 33,33$.

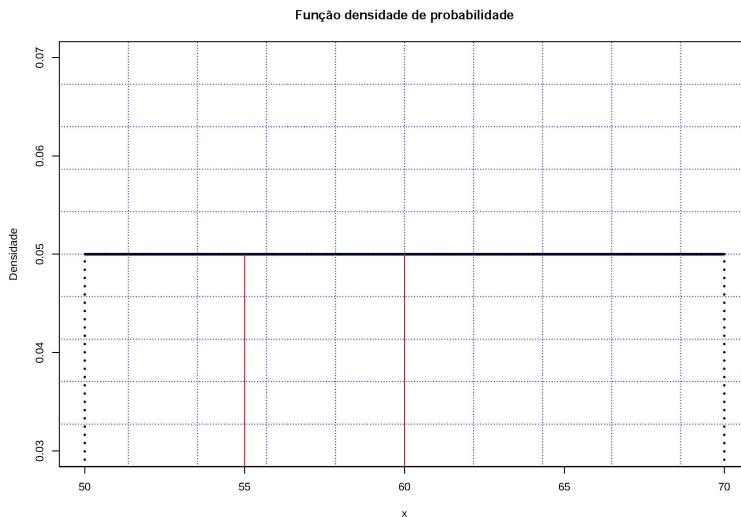


Figure 6.10: Os valores assumidos por $f(x)$ são ≥ 0 e a área definida por $f(x)$ no intervalo $50 \leq x \leq 70$ é igual a 1. Por essa razão pode ser uma fdp. A probabilidade pedida equivale à área $P(60 \leq x \leq 55) = (60 - 55) \cdot 0,05 = 0,25$.

6.3.2 Exponencial

A Distribuição Exponencial é largamente utilizada nas áreas de engenharia, física, computação e biologia para modelar variáveis tais como vida útil de equipamentos, tempos entre falhas (*TBF*), tempos de sobrevivência de espécies, intervalos de solicitação de recursos por exemplo.

Esta é uma distribuição que se caracteriza por ter uma função de taxa de falha constante, a única com esta propriedade e por essa razão tem sido usada extensivamente como um modelo para o tempo de vida de certos produtos e materiais.

Uma variável aleatória contínua X que assume valores não negativos segue o modelo teórico Exponencial com parâmetro (taxa) λ : $X \sim Exp(\lambda)$. Sua função densidade de probabilidade é dada por:

$$f(X = x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & \text{para } x \geq 0 \\ 0, & \text{para } x < 0 \end{cases}$$

Alternativamente com parâmetro (escala): $\alpha = \frac{1}{\lambda}$ e sua densidade de probabilidade é dada por:

$$f(X = x) = \begin{cases} \frac{1}{\alpha} \cdot e^{-\frac{1}{\alpha} \cdot x}, & \text{para } x \geq 0 \\ 0, & \text{para } x < 0 \end{cases}$$

Para se calcular probabilidades de uma Distribuição Exponencial torna-se necessária a resolução da integral associada, posto que a análise simplificada de figuras geométricas não mais é possível.

De modo geral temos:

$$\begin{aligned} P(a < X < b) &= \int_a^b \lambda \cdot e^{-\lambda \cdot x} dx \\ P(a < X < b) &= -e^{-\lambda \cdot x} \Big|_a^b \\ P(a < X < b) &= e^{-\lambda \cdot a} - e^{-\lambda \cdot b} \end{aligned}$$

Sua esperança e a variância são:

- Esperança: $E(X) = \mu = \frac{1}{\lambda} = \alpha$; e,
- Variância: $Var(X) = \frac{1}{\lambda^2} = \alpha^2$.

Exemplo: Uma indústria fabrica lâmpadas especiais que ficam em operação continuamente. A empresa oferece a seus clientes a garantia de reposição, caso a lâmpada dure menos de 50 horas. A vida útil dessas lâmpadas pode ser modelada adequadamente através da distribuição Exponencial com parâmetro $\lambda = \frac{1}{8000}$. Determine a probabilidade de uma lâmpada necessitar ser trocada pela indústria em razão da garantia oferecida ao cliente.

Definindo a variável aleatória contínua T como sendo a vida útil da lâmpada: $T \sim Exp(\frac{1}{8000})$ e sua função densidade de probabilidade:

$$f(T = t) = \begin{cases} \frac{1}{8000} \cdot e^{-\frac{1}{8000} \cdot t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

A probabilidade de que uma lâmpada tenha uma vida útil menor que 50 horas será dada pela integral da fdp no intervalo [0;50]:

$$\begin{aligned} P(0 < T < 50) &= \int_0^{50} \lambda \cdot e^{-\lambda \cdot x} dx \\ P(0 < T < 50) &= -e^{-\lambda \cdot x} \Big|_0^{50} \\ P(0 < T < 50) &= e^{-\frac{1}{8000} \cdot 0} - e^{-\frac{1}{8000} \cdot 50} \\ P(0 < T < 50) &= 1 - 0,939413063 \\ &= 0,006 \end{aligned}$$

A probabilidade de que uma lâmpada fabricada por essa empresa tenha uma vida útil menor que 50 h é de 0,006 (proporção de 0,60%), naturalmente muito pequena considerando que a duração média das lâmpadas é de $\mu = \frac{1}{\lambda} = \frac{1}{\frac{1}{8000}} = 8000$ h, aproximadamente 333 dias (esperança da variável).

```

# Biblioteca necessária
library(ggplot2)

# Parâmetro lambda (inverso da esperança)
lambda <- 1/8000 # horas

# A esperança é de 8000 horas ~ 1 ano

# Faixa de valores para mostrar a curvatura suave
x_values <- seq(0, 50000, length.out = 50)

# Função densidade de probabilidade exponencial:  $f(x) = \lambda * \exp(-\lambda * x)$ 
# para lambda maior que zero  $> X \sim \text{Exp}(\lambda)$ 

density_values <- dexp(x_values, rate = lambda)

# Pontos
plot_data <- data.frame(x = x_values, density = density_values)

# Gráfico
plot <- ggplot(plot_data, aes(x, density)) +
  geom_line(color = "blue", size = 1) +
  theme_minimal() +
  labs(title = "Função densidade de probabilidade exponencial",
       x = "Variável aleatória: vida útil das lâmpadas (h)",
       y = "Densidade")

```

```

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

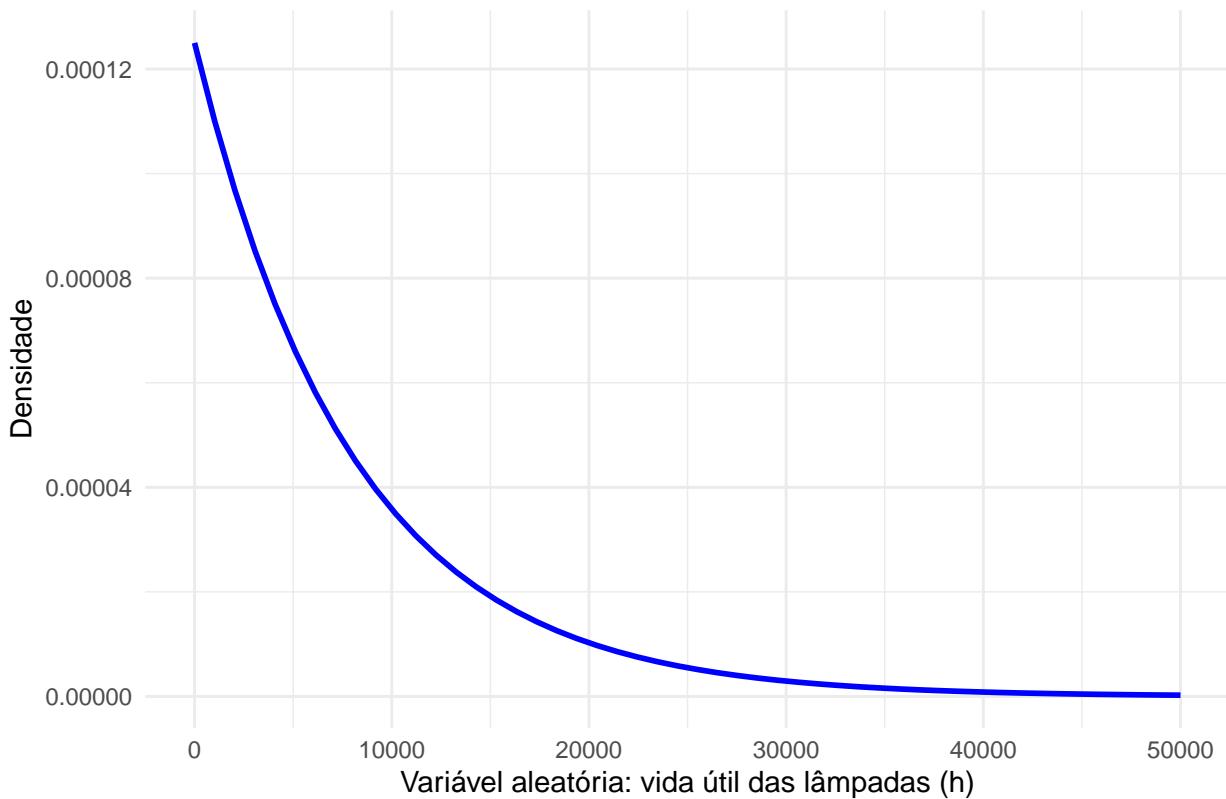
```

```

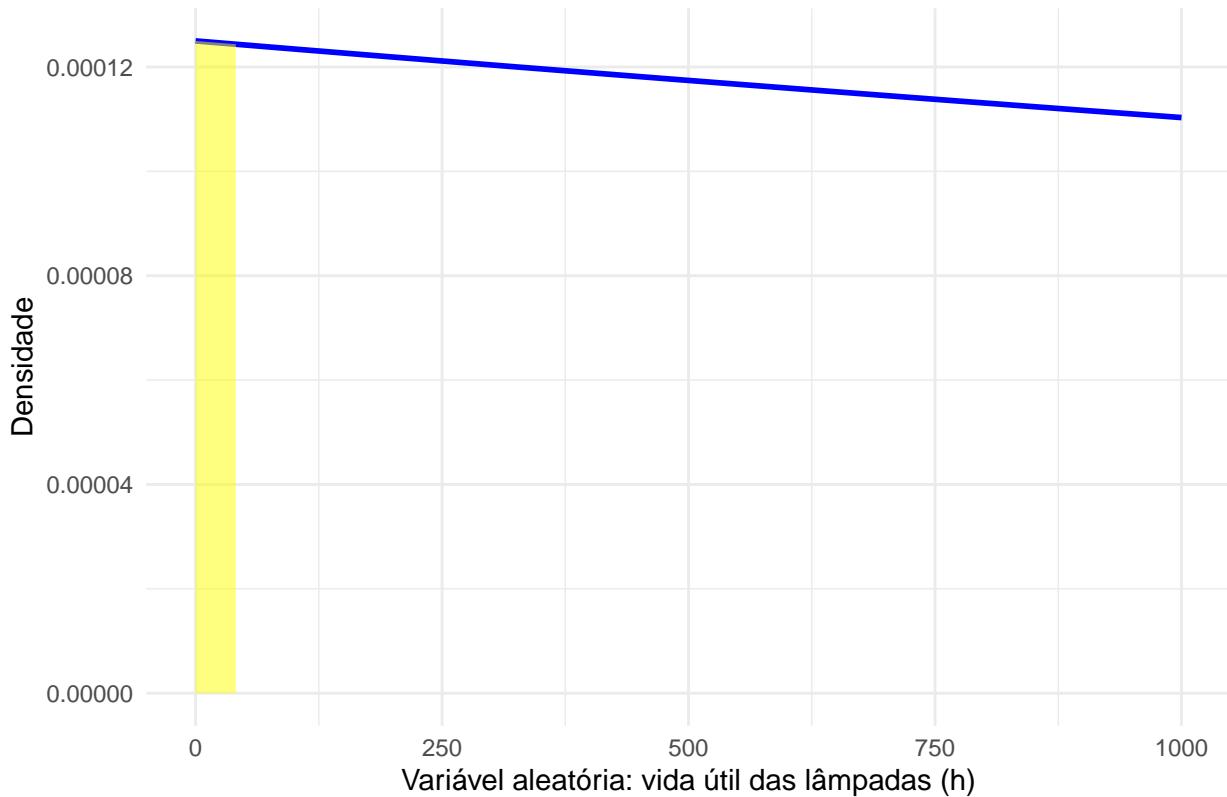
# Plote para mostrar a curvatura suave
plot

```

Função densidade de probabilidade exponencial



Função densidade de probabilidade exponencial



```
# Valores do intervalo
# 0.006230509 para [0,50]
# 0.9937695 para [50 , inf]. 1 - Prob[0,50]

valores <- c(0, 50) #a integrar
lambda <- 1/8000 # horas

# Integração numérica
probability <- integrate(function(x) dexp(x, rate = lambda), lower = valores[1], upper =
  valores[2])$value

# Valor
cat("Probabilidade de se observar valores entre ", valores[1], "e", valores[2], "é :",
  probability, "\n")

## Probabilidade de se observar valores entre 0 e 50 é : 0.006230509
```

Exemplo: O intervalo de tempo (minutos) entre as emissões de uma fonte radioativa é uma variável aleatória contínua que pode ser modelada pela Distribuição Exponencial com parâmetro $\lambda = 0,20$. Calcule a probabilidade de haver uma emissão em um intervalo de tempo inferior a 2 minutos.

Definindo a variável aleatória contínua T como sendo o intervalo de tempo entre as emissões radioativas dessa fonte: $T \sim Exp(0, 20)$ e sua função densidade de probabilidade:

$$f(T = t) = \begin{cases} 0,20 \cdot e^{-0,20 \cdot t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

A probabilidade de uma emissão em um intervalo de tempo inferior a 2 minutos será dada pela integral da fdp no intervalo $[0;2]$:

$$\begin{aligned} P(0 < T < 2) &= \int_0^2 \lambda \cdot e^{-\lambda \cdot x} dx \\ P(0 < T < 2) &= -e^{-\lambda \cdot x} \Big|_0^2 \\ P(0 < T < 2) &= e^{-0,20 \cdot 0} - e^{-0,20 \cdot 2} \\ P(0 < T < 2) &= 1 - 0,6703 \\ &= 0,3296 \end{aligned}$$

A probabilidade de uma emissão em um intervalo de tempo inferior a 2 min é de 0,3296, naturalmente considerável uma vez que o intervalo médio entre as emissões radioativas é de $\mu = \frac{1}{\lambda} = \frac{1}{0,20} = 5$ min (esperança da variável).

```
valores <- c(0, 2) # a integrar
lambda <- 0.2 # lambda já foi dado

# Integração numérica
probability <- integrate(function(x) dexp(x, rate = lambda), lower = valores[1], upper =
  valores[2])$value

# Valor
cat("Probabilidade de se observar valores entre ", valores[1], "e", valores[2], "é :",
  probability, "\n")

## Probabilidade de se observar valores entre 0 e 2 é : 0.32968
```

Exemplo: Certo tipo de fusível elétrico tem duração de vida (horas) que segue uma Distribuição Exponencial com tempo médio de vida de 100 horas. Cada peça tem um custo de R\$ 10,00 e, se durar menos de 200 horas, existe um custo adicional de R\$ 8,00. Pede-se: - a probabilidade de fusível durar mais de 150 horas; e,
- o custo esperado.

Se a vida útil média (μ) desse fusível é de 100 horas, então o valor do parâmetro dessa distribuição será $\frac{1}{100}$ (pois $\mu = \frac{1}{\lambda}$) e a variável aleatória contínua T será definida como sendo a vida útil do fusível: $T \sim Exp(\frac{1}{100})$, com sua função densidade de probabilidade:

$$f(T = t) = \begin{cases} \frac{1}{100} \cdot e^{-\frac{1}{100} \cdot t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

O primeiro item pede a probabilidade de um fusível durar mais de 150 horas poderá ser dada por 1 menos o valor da integral da fdp no intervalo [0;150]:

$$\begin{aligned} P(T > 150) &= 1 - P(0 < T < 150) \\ &= 1 - \int_0^{150} \lambda \cdot e^{-\lambda \cdot x} dx \\ &= 1 - e^{-\lambda \cdot x} \Big|_0^{150} \\ &= 1 - (e^{-0,01 \cdot 0} - e^{-0,01 \cdot 150}) \\ &= 1 - (1 - 0,22313) \\ &= 0,22313 \end{aligned}$$

A probabilidade de um fusível ter uma vida útil maior que 150 horas é de 0,22313.

```
valores <- c(0, 150) # integrar e subtrair de 1
lambda <- 1/100 # foi dada a vida útil média

# Integração numérica
probability <- 1 - integrate(function(x) dexp(x, rate = lambda), lower = valores[1], upper =
  valores[2])$value

# Valor
rotulos <- c(150, 'infinito') # integrar e subtrair de 1

cat("Probabilidade de se observar valores entre ", rotulos[1], "e", rotulos[2], "é :",
  probability, "\n")

## Probabilidade de se observar valores entre 150 e infinito é : 0.2231302
```

O custo unitário de um fusível é de R\$ 10,00 com um custo adicional de R\$ 8,00 se sua vida for inferior a 200 horas. Assim o custo esperado de um fusível será dada produto dos custos pelas respectivas probabilidades associadas:

$$C = \begin{cases} R\$10,00 & \text{se } t > 200 \\ R\$18,00 & \text{se } t < 200 \end{cases}$$

A probabilidade de um fusível durar mais de 200 horas poderá ser dada por 1 menos o valor da integral da fdp no intervalo [0;200]:

$$\begin{aligned} P(T > 200) &= 1 - P(0 < T < 200) \\ &= 1 - \int_0^{200} \lambda \cdot e^{-\lambda \cdot x} dx \\ &= 1 - e^{-\lambda \cdot x} \Big|_0^{200} \\ &= 1 - (e^{-0,01 \cdot 0} - e^{-0,01 \cdot 200}) \\ &= 1 - (1 - 0,1353) \\ &= 0,1353 \end{aligned}$$

A probabilidade de um fusível ter uma vida útil maior que 200 horas é de 0,1353.

```
valores <- c(0, 200) # integrar e subtrair de 1
lambda <- 1/100 # foi dada a vida útil média

# Integração numérica
probability <- 1 - integrate(function(x) dexp(x, rate = lambda), lower = valores[1], upper =
  valores[2])$value

# Valor
rotulos <- c(200, 'infinito') # integrar e subtrair de 1

cat("Probabilidade de se observar valores entre ", rotulos[1], "e", rotulos[2], "é :",
  probability, "\n")

## Probabilidade de se observar valores entre 200 e infinito é : 0.1353353
```

A probabilidade de um fusível durar menos de 200 horas será dada por 1 menos o valor calculado anteriormente:

$$P(0 < T < 200) = 1 - 0,1353 = 0,8647$$

A probabilidade de um fusível ter uma vida útil menor que 200 horas é de 0,8647.

```

valores <- c(0, 200) #a integrar
lambda <- 1/100 # foi dada a vida útil média

# Integração numérica
probability <- integrate(function(x) dexp(x, rate = lambda), lower = valores[1], upper =
  ↪ valores[2])$value

# Valor
cat("Probabilidade de se observar valores entre ", valores[1], "e", valores[2], "é :",
  ↪ probability, "\n")

## Probabilidade de se observar valores entre 0 e 200 é : 0.8646647

```

O custo esperado é de: $10,00 \times 0,1353 + 18,00 \times 0,8647 = R\$16,92$

6.3.3 Normal

A distribuição Normal (Gaussiana) é uma das mais importantes distribuições de probabilidades por possibilitar a adequada modelagem de fenômenos de diversas áreas: física, biologia, psicologia, ciências sociais e econômicas.

A história da curva Gaussiana está relacionada à formulação da Teoria da Probabilidade nos séculos XVIII e XIX, que contou com contribuições de muitos matemáticos dentre os quais podemos citar Abrahan De Moivre, Pierre Simon Laplace, Adrien-Marie Legendre, Francis Galton e Johann Carl Friedrich Gauss.

Esses matemáticos constataram que as variações entre repetidas medidas da mesma grandeza física apresentavam um grau surpreendente de regularidade. Com a repetição de medidas em um numero razoável observou-se que distribuição das variações poderia ser satisfatoriamente aproximada por uma curva contínua.

Em 1920 Karl Pearson relembra ter usado a expressão *curva normal* como uma substituição de *natureza diplomática* para evitar uma questão internacional sobre precedência que poderia surgir no uso comum à época da denominação “Curva de Laplace-Gauss”, dois grandes matemáticos e astrônomos. Todavia, reconheceu também que a nova denominação poderia levar pessoas a incorrer no erro de supor que todas as demais distribuições seriam anormais.

Uma variável aleatória contínua X que assuma valores x ($-\infty < x < \infty$) com média μ e variância σ^2 distribuídos segundo uma Curva Gaussiana é denotada por $X \sim N(\mu, \sigma^2)$, e sua função densidade de probabilidade é dada por:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A função de probabilidade cumulativa, a probabilidade de que a variável aleatória X apresente um valor menor ou igual a x é dada por:

$$F(x) = P(X \leq x) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(v-\mu)^2}{2\sigma^2}} dv$$

Sejam as seguintes variáveis aleatórias contínuas com Distribuição Normal:

- $X \sim N(\mu_X, \sigma_X^2)$, tal que $E(X) = \mu_X$ e $Var(X) = \sigma_X^2$; e
- $Y \sim N(\mu_Y, \sigma_Y^2)$, tal que $E(Y) = \mu_Y$ e $Var(Y) = \sigma_Y^2$.

Uma variável aleatória definida como uma soma de variáveis Normais $W = X \pm Y$ terá:

- $E(W) = \mu_X \pm \mu_Y$; e,
- $Var(W) = \sigma_X^2 + \sigma_Y^2$.

Para qualquer variável aleatória contínua com Distribuição Normal, chama-se de *padronização* à mudança da escala original dos dados para unidades padronizadas: *scores z*.

Uma variável padronizada segue possuindo Distribuição Normal, sendo denotada por $Z \sim N(0, 1)$, indicando que a média é 0 e o desvio-padrão é 1. Para a padronização de uma variável original X segue:

$$Z = \frac{X - \mu}{\sigma}$$

A função densidade de probabilidade de uma variável aleatória contínua padronizada é dada por:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$f(z) = 0,3989e^{-5z^2}$$

E a função de probabilidade cumulativa (a probabilidade de que a variável aleatória padronizada Z apresente um valor menor ou igual a z) é dada por:

$$F(z) = P(Z \leq z)$$

$$P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{u^2}{2}} du$$

A área sob a curva padronizada (probabilidade cumulativa entre dois valores z) é obtida em tabelas, dispensando a resolução numérica da integral acima (posto não possuir solução analítica).

Essas tabelas apresentam no **cruzamento** de suas **linhas** e **colunas**, a área sob a curva Normal padronizada equivalente à probabilidade associada a um **determinado intervalo* como, por exemplo:

A tabela Z possibilita:

- 1- encontrar a probabilidade (área) partindo de *score z*; e
- 2- encontrar o *score z*.

z	Segunda casa decimal de z									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706

Figure 6.11: Tabela Z mostrando a probabilidade ao intervalo $[0 ; 1,64]$ (quadro superior à esquerda explica onde a área se encontra)

Modo 1: admita que você padronizou um certo valor e obteve o *score z* igual a 1,64. Na coluna vertical à esquerda você deverá encontrar qual é a linha que apresenta a **unidade** e a **primeira casa decimal** desse valor: 1,6. Nas outras **dez** colunas verticais você deverá buscar aquela que apresenta a **segunda casa decimal** desse valor: 4. No cruzamento dessas duas colunas você irá fazer a leitura do número que lá dentro se encontra. Agora veja o desenho orientativo que há no canto superior à direita (cada tabela pode variar um pouco). Ele expõe graficamente uma área hachurada e na cor laranja entre o **zero** e um valor **z**. É exatamente o valor dessa área que você acabou de encontrar (a área sob a curva da fdp no intervalo $[0 ; 1,64]$).

Modo 2: admita que você precisa determinar qual é o valor do score z para uma probabilidade (área) no intervalo $[0 ; z] = 0,4495$. Nessa situação, simplesmente faça o caminho reverso. Encontre que célula apresenta esse valor de 0,4495 e faça a leitura da **unidade** e a **primeira casa decimal** do valor do score z na coluna lateral à esquerda (1,6) e de sua **segunda casa decimal** na linha que identifica as outras dez colunas (4).

A fdp da distribuição Normal apresenta uma **curva simétrica** centrada em sua média μ . A fdp da distribuição Normal padronizada também é simétrica e centra em sua média que agora tem valor 0.

A **totalidade da área** sob essas fdp (ou seja, o intervalo $-\infty < z < \infty$) possui área igual a 1. Cada metade, consequentemente, terá área igual a 0,50.

Por esse motivo as tabelas Z mostram apenas a **metade** da curva da fdp e muitos exercícios irão demandar que você some a área (0,50) do restante da curva da fdp, subtraia ou faça outras operações aritméticas simples para resolvê-los.

```

library(ggplot2)
options("digits"=4)
prob_desejada=0.95
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)
d_0=dnorm(0, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-4, 0),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores score (z)", breaks = z_desejado) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(0, z_desejado),
            colour="red")+
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c( z_desejado, 4),
            colour="black")+
  labs(title=
    "Curva da função densidade da distribuição Normal padronizada",
    subtitle = "P(-inf; 0)=0,50 (cinza) \nP(0 ; 1,645)=0,4495 (vermelho) \nP(1,645 ;
    \n inf)=0,0505 (cinza)")+
  geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada),
               color="blue", lty=2, lwd=0.3)+
  geom_segment(aes(x = 0, y = 0, xend = 0, yend = d_0), color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=-1, y=0.2, label="Probabilidade (área) =0,50 ", angle=0, vjust=0,
           hjust=0, color="blue",size=3)+
  annotate(geom="text", x=0.1, y=0.1, label="Probabilidade (área) =0,4495", angle=0,
           vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=2, y=0.05, label="Probabilidade (área) =0,0505", angle=0, vjust=0,
           hjust=0, color="blue",size=3)+
  theme_bw()

```

Curva da função densidade da distribuição Normal padronizada

$$\begin{aligned} P(-\infty; 0) &= 0,50 \text{ (cinza)} \\ P(0 ; 1,645) &= 0,4495 \text{ (vermelho)} \\ P(1,645 ; \infty) &= 0,0505 \text{ (cinza)} \end{aligned}$$

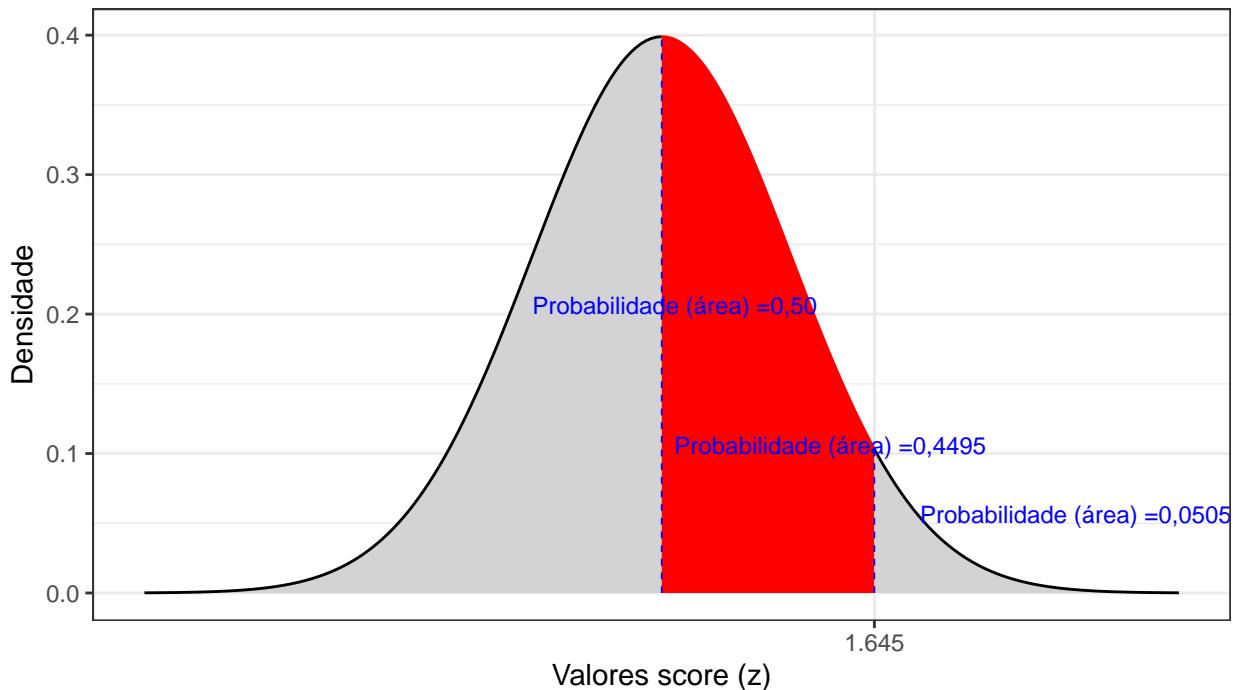


Figure 6.12: Curva da fdp da Distribuição Normal padronizada mostrando as áreas delimitadas pelo score z arbitrado (1,64)

Exemplo: Admita que o índice pluviométrico de uma cidade siga uma distribuição normal, com média de 101,60 mm/ano e desvio padrão de 12,70 mm/ano. Quais seriam as probabilidades dessa cidade ter menos de 83,82 mm/ano e mais de 96,52 mm/ano de precipitação no próximo ano?

A probabilidade de ocorrência de uma **precipitação inferior** a 83,82mm/ano equivale (graficamente) à área situada no intervalo $[-\infty; 83,82]$ na curva da fdp da distribuição Normal com média 101,60mm/ano e desvio padrão de 12,70mm/ano:

$$P(X \leq 83,82) \equiv \text{rea}[-\infty; 83,82]$$

A probabilidade de ocorrência de uma **precipitação superior** a 96,52 mm/ano equivale (graficamente) à área situada no intervalo $[96,52; +\infty]$ na curva da fdp distribuição Normal com média 101,60mm/ano e desvio padrão de 12,70mm/ano

$$P(X \geq 96,52) \equiv rea[96,52; +\infty]$$

Padronizando esses valores será possível estabelecer os valores das precipitações associadas às probabilidades pedidas em termos de scores z que podem ser obtidas em tabelas Z.

Considerando-se que a média é de 101,60mm/ano e o desvio padrão é de 12,70mm/ano, para a primeira precipitação (83,82mm/ano) teremos:

$$X_1 = 83,82$$

$$Z_n = \frac{X_n - \mu}{\sigma}$$

$$z_1 = -1,40$$

E a probailidade pedida equivale (graficamente) à área situada no intervalo $[-\infty; -1,40]$ na curva da fdp distribuição Normal padronizada:

$$P(X \leq 83,82) = P(Z \leq -1,40) \equiv rea[-\infty; -1,40]$$

Portanto, uma precipitação de 83,82mm/ano localiza-se a -1,40 desvios padrão à esquerda da média da curva Normal padronizada ($\mu = 0$).

Em uma tabela da Distribuição Normal Padronizada temos a probabilidade associada ao intervalo $P(0 < Z < z)$ tabelada para vários valores de z . No caso, veremos que para um valor $P(0 < z < 1,40) = 0,4192$ (lembre-se: a curva é simétrica por essa razão as tableas resumem-se a mostrar um dos lados).

Sendo a curva simétrica, a área total (probabilidade) sob a fdp é igual a 1: 0,50 à esquerda e 0,50 à direita. Assim, a área hachurada em vermelho na Figura 6.13 é a probabilidade pedida:

$$P(X \leq 83,82) = 0,50 - 0,4192$$

$$P(X \leq 83,82) = 0,0808$$

```
# Integração numérica no R
fx <- function(x){(1/(12.7*sqrt(2*pi))) * exp( -(1/2)*((x - 101.6)/(12.7))^2)}
p1=integrate(fx, 83.82, 101.6)
1-(p1$value + 0.5)
```

```
## [1] 0.08076
```

```
# Ou usando a função no R
pnorm(83.82, 101.6, 12.7)
```

```
## [1] 0.08076
```

```
library(ggplot2)
options("digits"=4)
prob_desejada=0.0808
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)
d_0=dnorm(0, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado),
            colour="red") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores score (z)", breaks = z_desejado) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado, 0),
            colour="black")+
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, 4),
            colour="black")+
  labs(title=
    "Curva da função densidade da distribuição Normal padronizada",
    subtitle = "P(-inf; -1,40)=0,0808 (vermelho) \nP(-1,40 ; 0 )=0,4192 (cinza) \nP(0 ;
    \n inf)=0,50 (cinza)")+
  geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada),
    color="blue", lty=2, lwd=0.3)+
  geom_segment(aes(x = 0, y = 0, xend = 0, yend = d_0), color="blue", lty=2, lwd=0.3)+
```

theme_bw()

Curva da função densidade da distribuição Normal padronizada

$$P(-\infty; -1,40) = 0,0808 \text{ (vermelho)}$$

$$P(-1,40 ; 0) = 0,4192 \text{ (cinza)}$$

$$P(0 ; \infty) = 0,50 \text{ (cinza)}$$

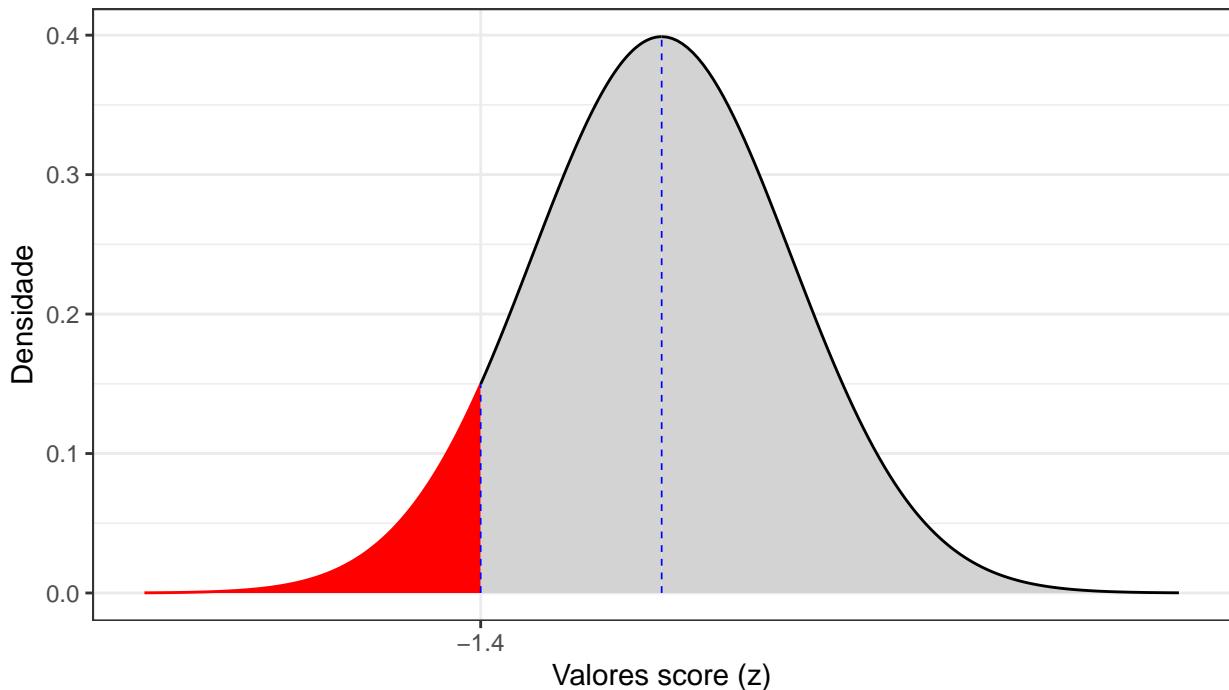


Figure 6.13: Curva da fdp da Distribuição Normal padronizada mostrando as áreas delimitadas pelo score z calculado (-1,40)

De modo análogo para a segunda questão 96,52 mm/ano) teremos:

$$X_2 = 96,52$$

$$Z_n = \frac{X_n - \mu}{\sigma}$$

$$z_2 = -0,40$$

E a probabilidade pedida equivale (graficamente) à área situada no intervalo [-0,40 ; ∞] na curva da fdp distribuição Normal padronizada:

$$P(X \geq 96,52) = P(Z \geq -0,40) \equiv rea[-\infty; -1,40]$$

Portanto, uma precipitação de 96,52 mm/ano localiza-se a -0,40 desvios padrão à esquerda da média da curva Normal padronizada ($\mu = 0$).

Em uma tabela da Distribuição Normal Padronizada temos a probabilidade associada ao intervalo $P(0 < Z < z)$ tabelada para vários valores de z . No caso, veremos que para um valor $P(0 < z < 0,40) = 0,1554$ (lembre-se: a curva é simétrica por essa razão as tabelas resumem-se a mostrar um dos lados).

Sendo a curva simétrica, a área total (probabilidade) sob a fdp é igual a 1: 0,50 à esquerda e 0,50 à direita. Assim, a área hachurada em vermelho na Figura 6.14 é a probabilidade pedida:

$$P(X \geq 96,52) = 0,50 + 0,4192 = 0,6554$$

```
# Integração numérica no R
fx <- function(x){(1/(12.7*sqrt(2*pi))) * exp( -(1/2)*((x - 101.6)/(12.7))^2)}
```

```
p1=integrate(fx, 101.6, 96.52)
```

```
1-(p1$value + 0.5)
```

```
## [1] 0.6554
```

```
# Ou usando a função no R
pnorm(96.52, 101.6, 12.7, lower.tail = FALSE) # para calcular à direita
```

```
## [1] 0.6554
```

```
library(ggplot2)
options("digits"=4)
prob_desejada=0.3446
z_desejado=round(qnorm(prob_desejada),3)
d_desejada=dnorm(z_desejado, 0, 1)
d_0=dnorm(0, 0, 1)
```

```
ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-4, z_desejado),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores score (z)", breaks = z_desejado) +
  geom_area(stat = "function",
            fun = dnorm,
```

```

fill = "red",
xlim = c(z_desejado, 0),
colour="red")+
geom_area(stat = "function",
  fun = dnorm,
  fill = "red",
  xlim = c(0, 4),
  colour="red")+
labs(title=
  "Curva da função densidade da distribuição Normal padronizada",
  subtitle = "P(-inf; -0,40)=0,3446 (cinza) \nP(-0,40 ; 0)=0,1554 (vermelho) \nP(0 ;
  \u2192 inf)=0,50 (vermelho)")+
geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada),
  color="blue", lty=2, lwd=0.3)+
geom_segment(aes(x = 0, y = 0, xend = 0, yend = d_0), color="blue", lty=2, lwd=0.3)+theme_bw()

```

Curva da função densidade da distribuição Normal padronizada

$P(-\infty; -0,40)=0,3446$ (cinza)
 $P(-0,40 ; 0)=0,1554$ (vermelho)
 $P(0 ; \infty)=0,50$ (vermelho)

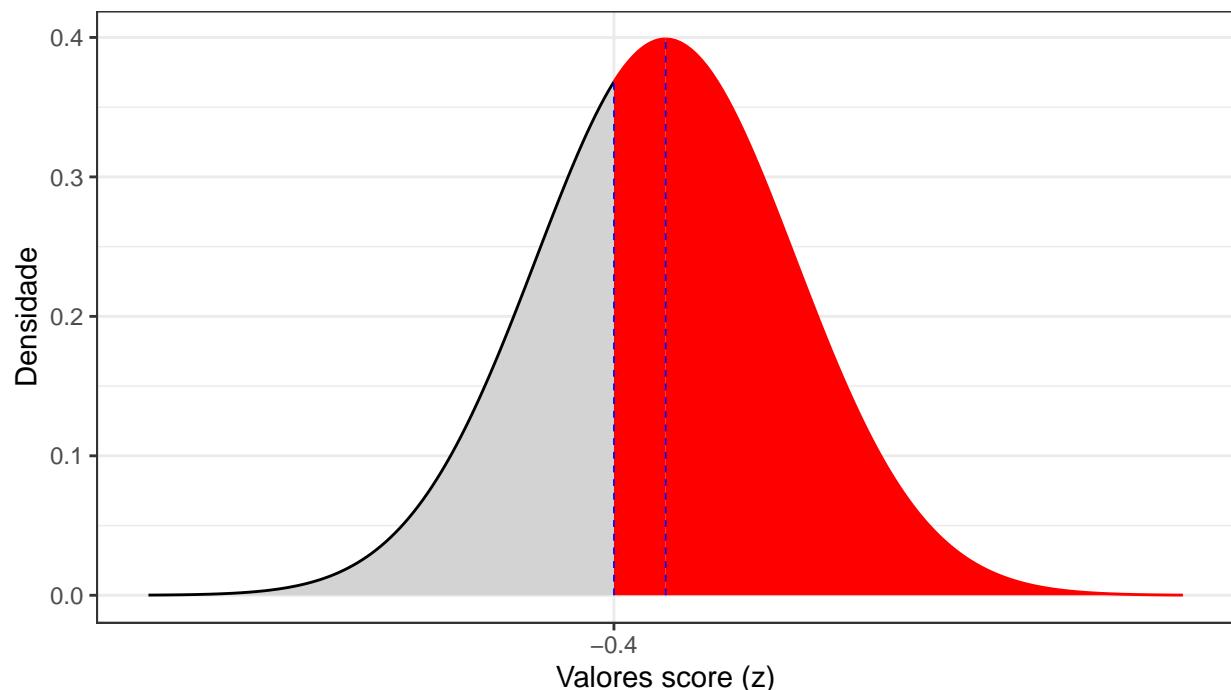


Figure 6.14: Curva da fdp da Distribuição Normal padronizada mostrando as áreas delimitadas pelo score z calculado (-0,40)

6.3.4 Student “t”

Se uma variável aleatória T contínua com ν graus de liberdade segue a *Distribuição t de Student*, sua função densidade de probabilidade é dada por:

$$f(t) = \frac{-\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{t^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}$$

com $\Gamma(n) = (n!)$

Uma variável aleatória contínua com essa distribuição possui:

- $E(T) = \mu = 0$; e,
- $Var(T) = \sigma^2 = \frac{\nu}{(\nu-2)}$, para $\nu > 2$

Admitamos que a partir de uma amostra aleatória composta por n valores retirados de uma população Normal com variância conhecida σ^2 deseje-se estimar a média μ .

Para grandes amostras ($n \geq 30$) a distribuição amostral de \bar{X} é aproximadamente Normal, com média μ e variância $\frac{\sigma^2}{n}$. Isso torna possível estabelecer a seguinte estatística padronizada anteriormente vista:

$$Z \sim \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Entretanto, para amostras de tamanho reduzido e variância desconhecida, a adoção do desvio padrão amostral S na estatística anterior conduz a uma outra distribuição.

Essa nova distribuição ainda é simétrica e com média $\mu = 0$; todavia não mais seria a Normal padronizada pois seu denominador $\frac{S}{\sqrt{n}}$ é uma variável aleatória (S é uma variável aleatória pois depende da amostra extrída ao passo o denominador anterior era uma constante: σ).

Essa família de distribuições (cuja forma tende à de uma distribuição Normam padronizada quando $n \rightarrow \infty$, $t_n \rightarrow N(0, 1)$) foi estabelecida pelo químico e estatístico inglês William Sealy Gosset.

$$T \sim \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Para se trabalhar com essa distribuição é preciso saber qual sua forma específica e isso é informado por uma estatística denominada **graus de liberdade**: ν .

Toda estatística de teste que dependa de uma variável aleatória possui graus de liberdade (ν). O número de informações independentes (ou livres) da amostra dá o número de graus de liberdade da Distribuição t de Student.

Na situação acima o propósito é estimar a média populacional μ através da média amostral \bar{X} ; todavia, tivemos também que estimar sua variância σ^2 através de S^2 , de tal modo que o número de graus de liberdade será $\nu = n - 1$: o tamanho da amostra menos 1.

A área sob a curva da fdp de uma distribuição de Student (probabilidade cumulativa entre dois valores t) é também obtida em tabelas.

Essas tabelas apresentam no **cruzamento** de suas **linhas** e **colunas**, o valor “t” para várias áreas (probabilidades) associadas:

- ao intervalo fechado: $[-t ; +t]$ (Figura 6.16);
- o intervalo aberto à esquerda: $[-\infty ; t]$ (Figura 6.17); e,
- o intervalo aberto à direita: $[t, \infty]$ (Figura 6.18).

Nas linhas horizontais lê-se os graus de liberdade ν e nas colunas as áreas (probabilidades).

A tabela t possibilita:

- 1- encontrar a probabilidade (área) partindo de um valor “t”; e
- 2- encontrar um valor “t” para determinada probabilidade

A fdp da distribuição de Student apresenta também uma **curva simétrica** centrada em sua média $\mu = 0$.

Distribuição t de Student												
gl/q	Área contida nas duas caudas laterais (bicaudal) da distribuição t de Student											
	0,990	0,980	0,975	0,950	0,900	0,800	0,200	0,100	0,050	0,025	0,020	0,010
	0,995	0,990	0,9875	0,975	0,950	0,900	0,100	0,050	0,025	0,0125	0,010	0,005
1	0,0157	0,0314	0,0393	0,0787	0,1584	0,3249	3,0777	6,3138	12,7062	25,4517	31,8205	63,6567
2	0,0141	0,0283	0,0354	0,0708	0,1421	0,2887	1,8856	2,9200	4,3027	6,2053	6,9646	9,9248
3	0,0136	0,0272	0,0340	0,0681	0,1366	0,2767	1,6377	2,5534	3,1824	4,1765	4,5407	5,8409
4	0,0133	0,0267	0,0333	0,0667	0,1338	0,2707	1,5332	2,1318	2,7764	3,4954	3,7469	4,6041
5	0,0132	0,0263	0,0329	0,0659	0,1322	0,2672	1,4759	2,0150	2,5706	3,1634	3,3649	4,0321
6	0,0131	0,0261	0,0327	0,0654	0,1311	0,2648	1,4398	1,9432	2,4469	2,9687	3,1427	3,7074
7	0,0130	0,0260	0,0325	0,0650	0,1303	0,2632	1,4149	1,8946	2,3646	2,8412	2,9980	3,4995
8	0,0129	0,0259	0,0323	0,0647	0,1297	0,2619	1,3968	1,8595	2,3060	2,7515	2,8965	3,3554
9	0,0129	0,0258	0,0322	0,0645	0,1293	0,2610	1,3830	1,8331	2,2222	2,6850	2,8214	3,2498
10	0,0129	0,0257	0,0321	0,0643	0,1289	0,2602	1,3722	1,8125	2,2281	2,6338	2,7638	3,1693
11	0,0128	0,0256	0,0321	0,0642	0,1286	0,2596	1,3634	1,7959	2,2010	2,5931	2,7181	3,1058
12	0,0128	0,0256	0,0320	0,0640	0,1283	0,2590	1,3562	1,7823	2,1788	2,5600	2,6810	3,0545
13	0,0128	0,0256	0,0319	0,0639	0,1281	0,2586	1,3502	1,7709	2,1604	2,5326	2,6503	3,0123

Figure 6.15: Tabela t mostrando duas áreas (probabilidades) para um grau de liberdade igual a 10. No intervalo fechado $[-0,1289 ; 0,1289]$ a probabilidade é de 0,90 e para os intervalos abertos à direita: $[0,1289 ; \infty]$ e à esquerda: $[-\infty ; 0,1289]$ é de 0,95.

A **totalidade da área** sob essa fdp (ou seja, o intervalo $-\infty < t < \infty$) possui área igual a 1. Cada metade, consequentemente, terá área igual a 0,50.

Muitos exercícios irão demandar que você some a área (0,50) do restante da curva da fdp, subtraia ou faça outras operações aritméticas simples para resolvê-los.

```
library(ggplot2)

alfa=0.05

prob_desejada1=alfa/2
df=10
t_desejado1=round(qt(prob_desejada1,df ),4)
d_desejada1=dt(t_desejado1,df )

prob_desejada2=1-alfa/2
df=10
t_desejado2=round(qt(prob_desejada2, df),4)
d_desejada2=dt(t_desejado2,df )

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(-4, t_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
```

```

    args=list(df),
    fill = "lightgrey",
    xlim = c(t_desejado1,0),
    colour="black") +
geom_area(stat = "function",
          fun = dt,
          args=list(df),
          fill = "lightgrey",
          xlim = c(0, t_desejado2),
          colour="black") +
geom_area(stat = "function",
          fun = dt,
          args=list(df),
          fill = "red",
          xlim = c(t_desejado2,4),
          colour="black") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores de t", breaks = c(t_desejado1, t_desejado2)) +
labs(title= "Curva da função densidade \nDistribuição t (df=10)",
     subtitle = "P(-2,228 ; 2,228)=0,90 (cinza) \nP(-inf ; -2,228)=P(2,086; inf)=0,05
                  (vermelho)")+
geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1),
             color="blue", lty=2, lwd=0.3)+
geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2),
             color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=-0.1, y=0.2, label="Probabilidade (área) =0,90 \n(gl=10)",
         angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=-3.5, y=0.1, label="Probabilidade (área) =0,05 \n(gl=10)",
         angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=2.5, y=0.1, label="Probabilidade (área) =0,05 \n(gl=10)", angle=0,
         vjust=0, hjust=0, color="blue",size=3)+

theme_bw()

```

```

alfa=0.025
prob_desejada=alfa
df=10
t_desejado=round(qt(prob_desejada,df ),4)
d_desejada=dt(t_desejado,df )

```

```

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(-4, t_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(t_desejado,0),

```

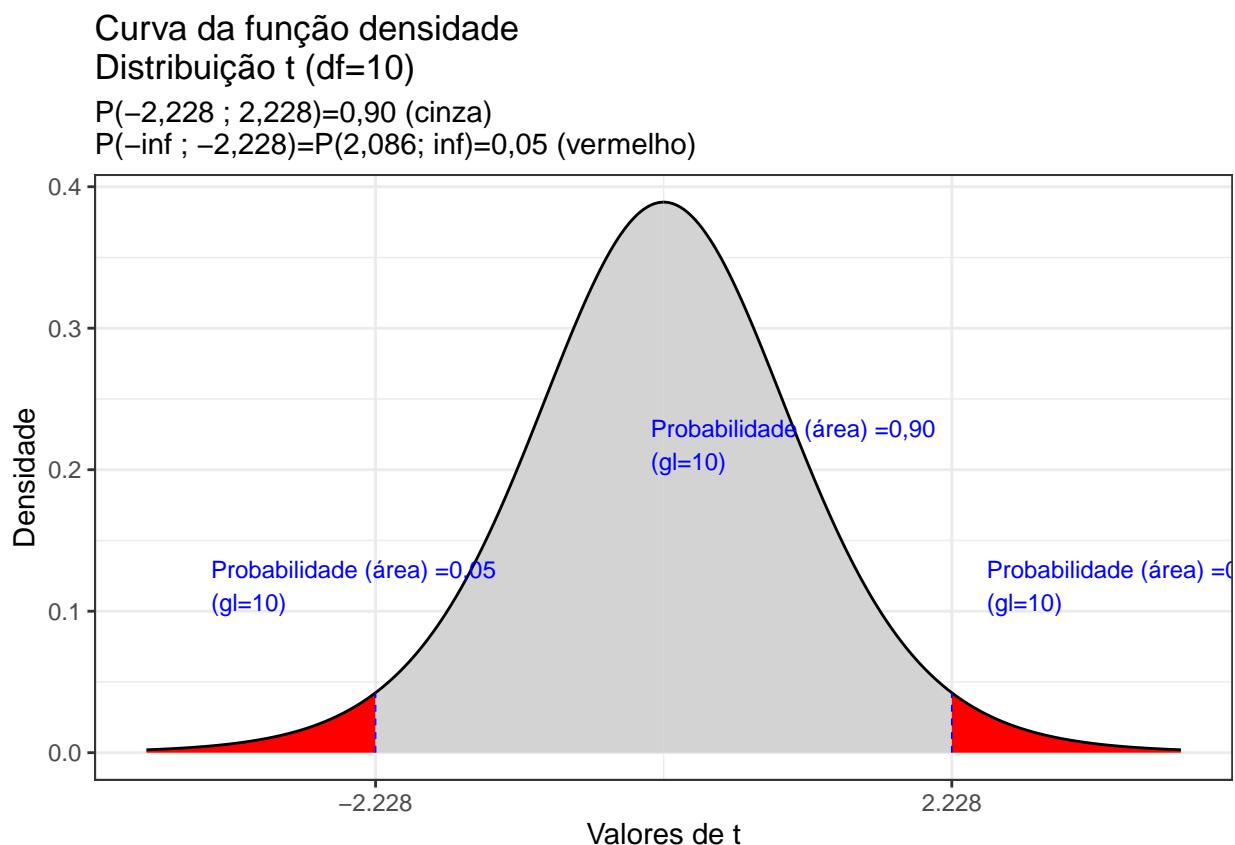


Figure 6.16: Curva da fdp da Distribuição Student para 10 graus de liberdade, mostrando as áreas delimitadas pelos valores $+/-t$ ($+/-2,28$)

```

        colour="black") +
geom_area(stat = "function",
  fun = dt,
  args=list(df),
  fill = "lightgrey",
  xlim = c(0, 4),
  colour="black")+
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores de t", breaks = c(t_desejado)) +
labs(title= "Curva da função densidade \nDistribuição t (df=10)",
     subtitle = "P(-inf ; -2,228)=0,025 (vermelho) \nP(-2,228 ; +inf)= 0,975 (cinza)")+
geom_segment(aes(x = t_desejado, y = 0, xend = t_desejado, yend = d_desejada),
             color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=-0.1, y=0.2, label="Probabilidade (área) =0,975 \n(gl=10)",
         angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=-3.5, y=0.1, label="Probabilidade (área) =0,025 \n(gl=10)",
         angle=0, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Curva da função densidade Distribuição t (df=10)

$P(-\infty ; -2,228)=0,025$ (vermelho)
 $P(-2,228 ; +\infty)= 0,975$ (cinza)

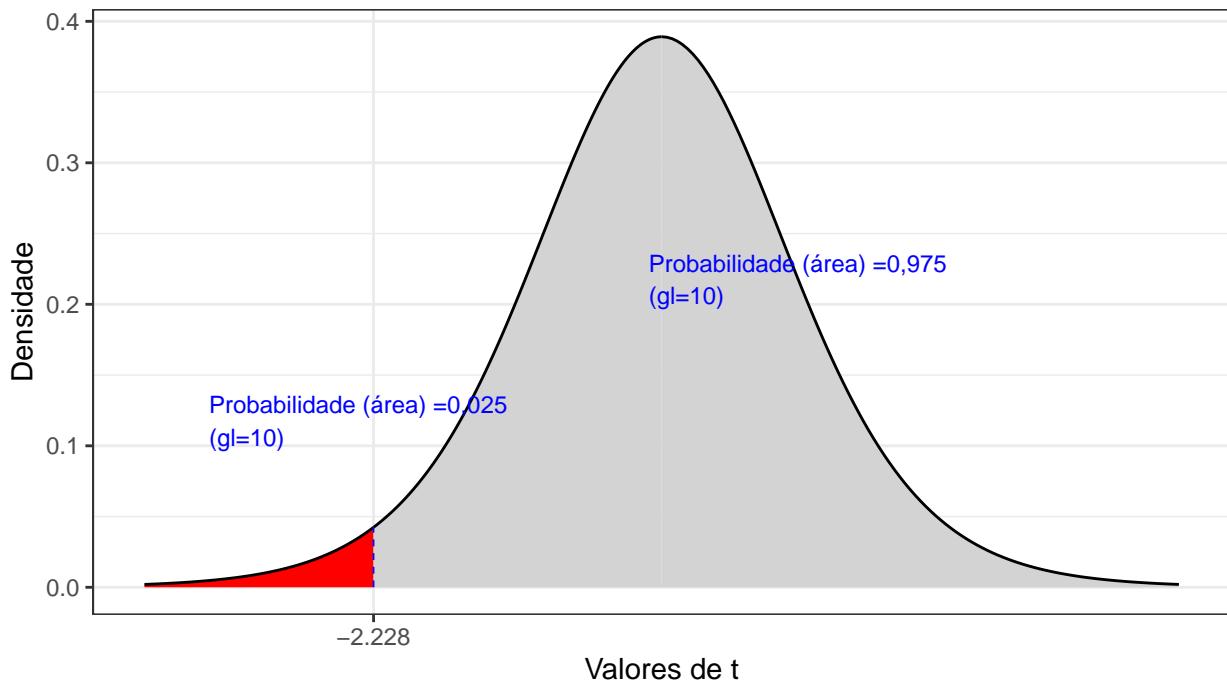


Figure 6.17: Curva da fdp da Distribuição Student para 10 graus de liberdade, mostrando as áreas delimitadas pelo valor $-t$ (-2,28)

```

alfa=0.025
prob_desejada=1-alfa

```

```

df=10
t_desejado=round(qt(prob_desejada,df ),4)
d_desejada=dt(t_desejado,df)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(-4, 0),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(0, t_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(t_desejado, 4),
            colour="black")+
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de t", breaks = c(t_desejado)) +
  labs(title= "Curva da função densidade \nDistribuição t (df=10)",
       subtitle = "P(-inf ; 2,228)=0,975 (vermelho) \nP(2,228 ; +inf)= 0,025 (cinza)")+
  geom_segment(aes(x = t_desejado, y = 0, xend = t_desejado, yend = d_desejada),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=0, y=0.2, label="Probabilidade (área) =0,975 \n(gl=10)", angle=0,
           vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=2.5, y=0.1, label="Probabilidade (área) =0,025 \n(gl=10)",
           angle=0, vjust=0, hjust=0, color="blue",size=3) +
  theme_bw()

```

6.3.5 Qui-Quadrado

Considerem X_1, X_2, \dots, X_ν como ν variáveis aleatórias contínuas independentes e normalmente distribuídas com média zero e variância 1. Definamos também uma variável aleatória resultante da soma dos quadrados das variáveis anteriormente especificadas:

$$\chi^2 = X_1^2 + X_2^2 + \dots + X_\nu^2$$

A variável aleatória χ^2 possui seguinte fdp para $x > 0$ (para $x \leq 0, f(x) = 0$), com ν graus de liberdade:

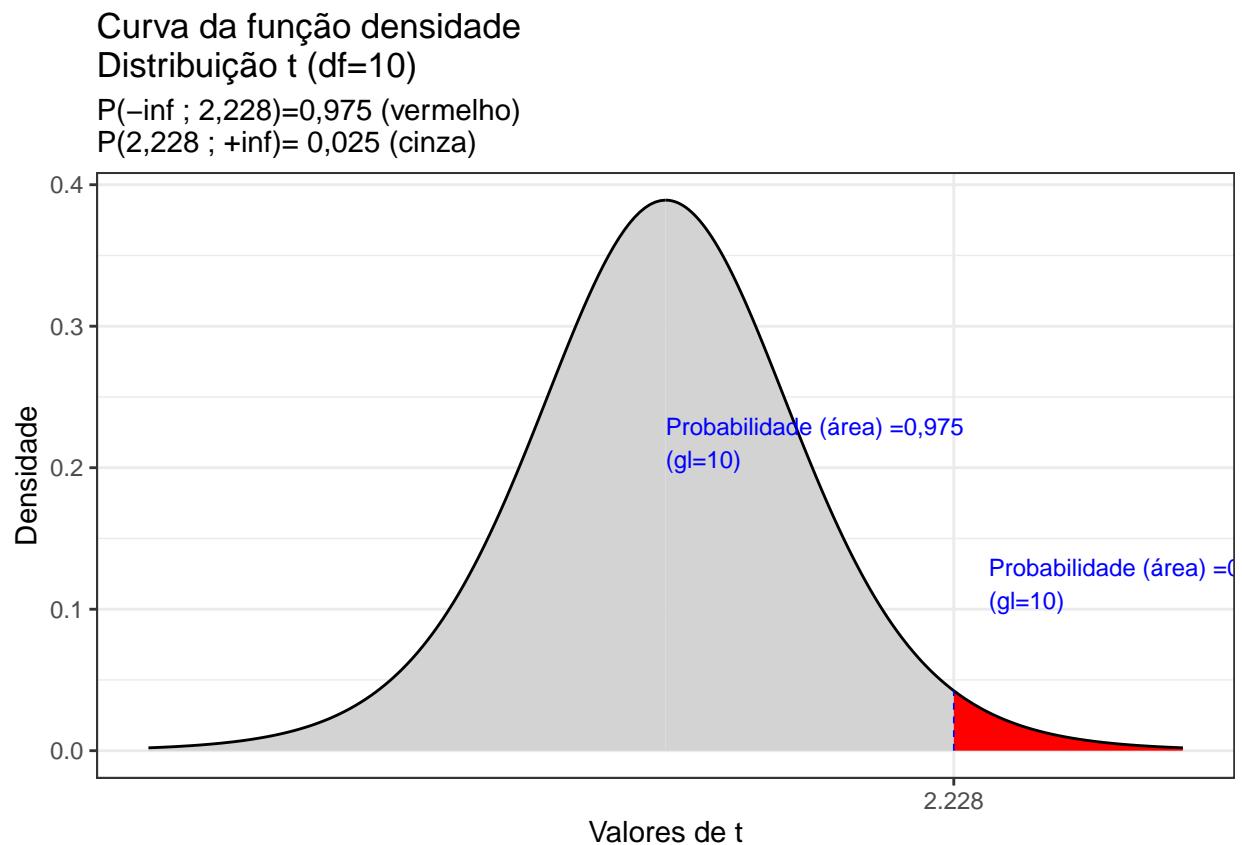


Figure 6.18: Curva da fdp da Distribuição Student para 10 graus de liberdade, mostrando as áreas delimitadas pelo valor $-t$ (-2,28)

$$f(x) = \frac{1}{2^{\frac{\nu}{2}} \cdot \Gamma\left(\frac{\nu}{2}\right)} \cdot x^{\left(\frac{\nu}{2}\right)^{-1} \cdot \epsilon^{-\frac{\nu}{2}}}$$

A função de probabilidade cumulativa é dada por:

$$P(\chi^2 \leq x) = \frac{1}{2^{\frac{\nu}{2}} \cdot \Gamma\left(\frac{\nu}{2}\right)} \int_{-\infty}^x u^{\left(\frac{\nu}{2}\right)^{-1} \cdot \epsilon^{-\frac{\nu}{2}}} du$$

Algumas propriedades da distribuição Qui-quadrado:

- Pelo Teorema Central do Limite esta família de distribuições tende a uma distribuição Normal quando o número de graus de liberdade tende ao infinito ($\nu \rightarrow \infty (\chi^2 \rightarrow N(0, 1))$);
- Se uma variável é definida como a soma de duas variáveis independentes com Distribuição Qui-quadrado com ν_1 e ν_2 graus de liberdade, essa variável também seguirá a Distribuição Qui-quadrado com $\nu_1 + \nu_2$ graus de liberdade;
- É assimétrica e definida para $x > 0$.

6.3.6 Fisher-Snedecor “F”

Uma variável aleatória contínua definida como $X \sim F(\nu_1, \nu_2)$ segue a Distribuição Fisher-Snedecor com parâmetros ν_1 e ν_2 , números inteiros positivos conhecidos como graus de liberdade do numerador e do denominador, respectivamente.

A Distribuição de Fisher-Snedecor é também conhecida como a Distribuição da razão de variâncias.

Uma variável aleatória X que segue uma Distribuição de Fisher-Snedecor com ν_1 e ν_2 graus de liberdade tem sua pdf dada por:

$$f(x) = \frac{\Gamma((\nu_1 + \nu_2)/2)(\nu_1/\nu_2)^{\nu_1/2}x^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)[(\nu_1/\nu_2)x + 1]^{(\nu_1+\nu_2)/2}} \quad x > 0,$$

com $\nu_1 = 1, 2, \dots$ e $\nu_2 = 1, 2, \dots$

6.4 Tabelas

Tabela - Normal Padrão de 0 a z

$P(0 \leq Z \leq z)$

z	Segunda casa decimal de Z									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000
4,0	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Professor Guru
professorguru.com.br

Figure 6.19: Tabela de valores “z” da Distribuição Normal padronizada

Av. Vereador José Diniz, 2804 - Campo Belo - São Paulo/SP - Brasil - CEP 04604-005

Atenção: O local é restrito à realização das aulas presenciais. Informações somente pelos telefones ou e-mail.

 (11) 3499-2828
 (11) 99828-2824
<http://AulasdeMatemática.com.br>
 Atendimento de Seg à Sáb das 10 às 23hs

Thiago Rodrigo Carneiro

 Lic. Matemática - UFRJ
 Bach. Estatística - USP

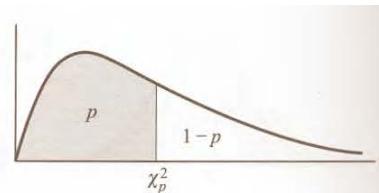
gl/q	Área contida nas duas caudas laterais (bicaudal) da distribuição t de Student												
	0,990	0,980	0,975	0,950	0,900	0,800	0,200	0,100	0,050	0,025	0,0125	0,020	0,010
	0,995	0,990	0,9875	0,975	0,950	0,900	0,100	0,050	0,025	0,0125	0,010	0,005	
1	0,0157	0,0314	0,0393	0,0787	0,1584	0,3249	3,0777	6,3138	12,7062	25,4517	31,8205	63,6567	
2	0,0141	0,0283	0,0354	0,0708	0,1421	0,2887	1,8856	2,9200	4,3027	6,2053	6,9646	9,9248	
3	0,0136	0,0272	0,0340	0,0681	0,1366	0,2767	1,6377	2,3534	3,1824	4,1765	4,5407	5,8409	
4	0,0133	0,0267	0,0333	0,0667	0,1338	0,2707	1,5332	2,1318	2,7764	3,4954	3,7469	4,6041	
5	0,0132	0,0263	0,0329	0,0659	0,1322	0,2672	1,4759	2,0150	2,5706	3,1634	3,3649	4,0321	
6	0,0131	0,0261	0,0327	0,0654	0,1311	0,2648	1,4398	1,9432	2,4469	2,9687	3,1427	3,7074	
7	0,0130	0,0260	0,0325	0,0650	0,1303	0,2632	1,4149	1,8946	2,3646	2,8412	2,9980	3,4995	
8	0,0129	0,0259	0,0323	0,0647	0,1297	0,2619	1,3968	1,8595	2,3060	2,7515	2,8965	3,3554	
9	0,0129	0,0258	0,0322	0,0645	0,1293	0,2610	1,3830	1,8331	2,2622	2,6850	2,8214	3,2498	
10	0,0129	0,0257	0,0321	0,0643	0,1289	0,2602	1,3722	1,8125	2,2281	2,6338	2,7638	3,1693	
11	0,0128	0,0256	0,0321	0,0642	0,1286	0,2596	1,3634	1,7959	2,2010	2,5931	2,7181	3,1058	
12	0,0128	0,0256	0,0320	0,0640	0,1283	0,2590	1,3562	1,7823	2,1788	2,5600	2,6810	3,0545	
13	0,0128	0,0256	0,0319	0,0639	0,1281	0,2586	1,3502	1,7709	2,1604	2,5326	2,6503	3,0123	
14	0,0128	0,0255	0,0319	0,0638	0,1280	0,2582	1,3450	1,7613	2,1448	2,5096	2,6245	2,9768	
15	0,0127	0,0255	0,0319	0,0638	0,1278	0,2579	1,3406	1,7531	2,1314	2,4899	2,6025	2,9467	
16	0,0127	0,0255	0,0318	0,0637	0,1277	0,2576	1,3368	1,7459	2,1199	2,4729	2,5835	2,9208	
17	0,0127	0,0254	0,0318	0,0636	0,1276	0,2573	1,3334	1,7396	2,1098	2,4581	2,5669	2,8982	
18	0,0127	0,0254	0,0318	0,0636	0,1274	0,2571	1,3304	1,7341	2,1009	2,4450	2,5524	2,8784	
19	0,0127	0,0254	0,0318	0,0635	0,1274	0,2569	1,3277	1,7291	2,0930	2,4334	2,5395	2,8609	
20	0,0127	0,0254	0,0317	0,0635	0,1273	0,2567	1,3253	1,7247	2,0860	2,4231	2,5280	2,8453	
21	0,0127	0,0254	0,0317	0,0635	0,1272	0,2566	1,3232	1,7207	2,0796	2,4138	2,5176	2,8314	
22	0,0127	0,0254	0,0317	0,0634	0,1271	0,2564	1,3212	1,7171	2,0739	2,4055	2,5083	2,8188	
23	0,0127	0,0253	0,0317	0,0634	0,1271	0,2563	1,3195	1,7139	2,0687	2,3979	2,4999	2,8073	
24	0,0127	0,0253	0,0317	0,0634	0,1270	0,2562	1,3178	1,7109	2,0639	2,3909	2,4922	2,7969	
25	0,0127	0,0253	0,0317	0,0633	0,1269	0,2561	1,3163	1,7081	2,0595	2,3846	2,4851	2,7874	
26	0,0127	0,0253	0,0316	0,0633	0,1269	0,2560	1,3150	1,7056	2,0555	2,3788	2,4786	2,7787	
27	0,0127	0,0253	0,0316	0,0633	0,1268	0,2559	1,3137	1,7033	2,0518	2,3734	2,4727	2,7707	
28	0,0126	0,0253	0,0316	0,0633	0,1268	0,2558	1,3125	1,7011	2,0484	2,3685	2,4671	2,7633	
29	0,0126	0,0253	0,0316	0,0633	0,1268	0,2557	1,3114	1,6991	2,0452	2,3638	2,4620	2,7564	
30	0,0126	0,0253	0,0316	0,0632	0,1267	0,2556	1,3104	1,6973	2,0423	2,3596	2,4573	2,7500	
31	0,0126	0,0253	0,0316	0,0632	0,1267	0,2555	1,3095	1,6955	2,0395	2,3556	2,4528	2,7440	
32	0,0126	0,0253	0,0316	0,0632	0,1267	0,2555	1,3086	1,6939	2,0369	2,3518	2,4487	2,7385	
33	0,0126	0,0253	0,0316	0,0632	0,1266	0,2554	1,3077	1,6924	2,0345	2,3483	2,4448	2,7333	
34	0,0126	0,0253	0,0316	0,0632	0,1266	0,2553	1,3070	1,6909	2,0322	2,3451	2,4411	2,7284	
35	0,0126	0,0252	0,0316	0,0632	0,1266	0,2553	1,3062	1,6896	2,0301	2,3420	2,4377	2,7238	
36	0,0126	0,0252	0,0316	0,0631	0,1266	0,2552	1,3055	1,6883	2,0281	2,3391	2,4345	2,7195	
37	0,0126	0,0252	0,0316	0,0631	0,1265	0,2552	1,3049	1,6871	2,0262	2,3363	2,4314	2,7154	
38	0,0126	0,0252	0,0315	0,0631	0,1265	0,2551	1,3042	1,6860	2,0244	2,3337	2,4286	2,7116	
39	0,0126	0,0252	0,0315	0,0631	0,1265	0,2551	1,3036	1,6849	2,0227	2,3313	2,4258	2,7079	
40	0,0126	0,0252	0,0315	0,0631	0,1265	0,2550	1,3031	1,6839	2,0211	2,3289	2,4233	2,7045	
45	0,0126	0,0252	0,0315	0,0631	0,1264	0,2549	1,3006	1,6794	2,0141	2,3189	2,4121	2,6896	
48	0,0126	0,0252	0,0315	0,0630	0,1263	0,2548	1,2994	1,6772	2,0106	2,3139	2,4066	2,6822	
50	0,0126	0,0252	0,0315	0,0630	0,1263	0,2547	1,2987	1,6759	2,0086	2,3109	2,4033	2,6778	
55	0,0126	0,0252	0,0315	0,0630	0,1262	0,2546	1,2971	1,6730	2,0040	2,3044	2,3961	2,6682	
60	0,0126	0,0252	0,0315	0,0630	0,1262	0,2545	1,2958	1,6706	2,0003	2,2990	2,3901	2,6603	
63	0,0126	0,0252	0,0315	0,0630	0,1262	0,2544	1,2951	1,6694	1,9983	2,2962	2,3870	2,6561	
70	0,0126	0,0252	0,0315	0,0629	0,1261	0,2543	1,2938	1,6669	1,9944	2,2906	2,3808	2,6479	
75	0,0126	0,0252	0,0314	0,0629	0,1261	0,2542	1,2929	1,6654	1,9921	2,2873	2,3771	2,6430	
80	0,0126	0,0251	0,0314	0,0629	0,1261	0,2542	1,2922	1,6641	1,9901	2,2844	2,3739	2,6387	
85	0,0126	0,0251	0,0314	0,0629	0,1260	0,2541	1,2916	1,6630	1,9883	2,2818	2,3710	2,6349	
90	0,0126	0,0251	0,0314	0,0629	0,1260	0,2541	1,2910	1,6620	1,9867	2,2795	2,3685	2,6316	
95	0,0126	0,0251	0,0314	0,0629	0,1260	0,2541	1,2905	1,6611	1,9853	2,2775	2,3662	2,6286	
99	0,0126	0,0251	0,0314	0,0629	0,1260	0,2540	1,2902	1,6604	1,9842	2,2760	2,3646	2,6264	
100	0,0126	0,0251	0,0314	0,0629	0,1260	0,2540	1,2901	1,6602	1,9840	2,2757	2,3642	2,6259	
120	0,0126	0,0251	0,0314	0,0628	0,1259	0,2539	1,2886	1,6577	1,9799	2,2699	2,3578	2,6174	
100000	0,0125	0,0251	0,0313	0,0627	0,1257	0,2533	1,2816	1,6449	1,9600	2,2414	2,3264	2,5759	

As linhas indicam o número de graus de liberdade (gl) da distribuição t de Student e as colunas indicam a soma das áreas contidas nas caudas (bicaudal). Por exemplo, a linha com 16 gl e coluna 0,10 cujo valor tabelado é 1,746 indica que o valor 1,746 deixa 10% de probabilidade nas duas caudas quando há 16 gl. Ou seja, dada a probabilidade bicaudal eu descubro o valor t correspondente.

Fonte: Microsoft Excel 2007, fórmula INVT.

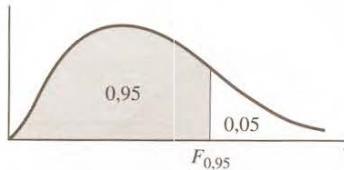
Figure 6.20: Tabela de valores "t" da Distribuição de Student

Percentis χ_p^2 da Distribuição Qui-Quadrado com v Graus de Liberdade



v	$\chi^2_{0,005}$	$\chi^2_{0,01}$	$\chi^2_{0,025}$	$\chi^2_{0,05}$	$\chi^2_{0,10}$	$\chi^2_{0,25}$	$\chi^2_{0,50}$	$\chi^2_{0,75}$	$\chi^2_{0,90}$	$\chi^2_{0,95}$	$\chi^2_{0,975}$	$\chi^2_{0,99}$	$\chi^2_{0,995}$	$\chi^2_{0,999}$
1	0,0000	0,0002	0,0010	0,0039	0,0158	0,102	0,455	1,32	2,71	3,84	5,02	6,63	7,88	10,8
2	0,0100	0,0201	0,0506	0,103	0,211	0,575	1,39	2,77	4,61	5,99	7,38	9,21	10,6	13,8
3	0,0717	0,115	0,216	0,352	0,584	1,21	2,37	4,11	6,25	7,81	9,35	11,3	12,8	16,3
4	0,207	0,297	0,484	0,711	1,06	1,92	3,36	5,39	7,78	9,49	11,1	13,3	14,9	18,5
5	0,412	0,554	0,831	1,15	1,61	2,67	4,35	6,63	9,24	11,1	12,8	15,1	16,7	20,5
6	0,676	0,872	1,24	1,64	2,20	3,45	5,35	7,84	10,6	12,6	14,4	16,8	18,5	22,5
7	0,989	1,24	1,69	2,17	2,83	4,25	6,35	9,04	12,0	14,1	16,0	18,5	20,3	24,3
8	1,34	1,65	2,18	2,73	3,49	5,07	7,34	10,2	13,4	15,5	17,5	20,1	22,0	26,1
9	1,73	2,09	2,70	3,33	4,17	5,90	8,34	11,4	14,7	16,9	19,0	21,7	23,6	27,9
10	2,16	2,56	3,25	3,94	4,87	6,74	9,34	12,5	16,0	18,3	20,5	23,2	25,2	29,6
11	2,60	3,05	3,82	4,57	5,58	7,58	10,3	13,7	17,3	19,7	21,9	24,7	26,8	31,3
12	3,07	3,57	4,40	5,23	6,30	8,44	11,3	14,8	18,5	21,0	23,3	26,2	28,3	32,9
13	3,57	4,11	5,01	5,89	7,04	9,30	12,3	16,0	19,8	22,4	24,7	27,7	29,8	34,5
14	4,07	4,66	5,63	6,57	7,79	10,2	13,3	17,1	21,1	23,7	26,1	29,1	31,3	36,1
15	4,60	5,23	6,26	7,26	8,55	11,0	14,3	18,2	22,3	25,0	27,5	30,6	32,8	37,7
16	5,14	5,81	6,91	7,96	9,31	11,9	15,3	19,4	23,5	26,3	28,8	32,0	34,3	39,3
17	5,70	6,41	7,56	8,67	10,1	12,8	16,3	20,5	24,8	27,6	30,2	33,4	35,7	40,8
18	6,26	7,01	8,23	9,39	10,9	13,7	17,3	21,6	26,0	28,9	31,5	34,8	37,2	42,3
19	6,84	7,63	8,91	10,1	11,7	14,6	18,3	22,7	27,2	30,1	32,9	36,2	38,6	43,8
20	7,43	8,26	9,59	10,9	12,4	15,5	19,3	23,8	28,4	31,4	34,2	37,6	40,0	45,3
21	8,03	8,90	10,3	11,6	13,2	16,3	20,3	24,9	29,6	32,7	35,5	38,9	41,4	46,8
22	8,64	9,54	11,0	12,3	14,0	17,2	21,3	26,0	30,8	33,9	36,8	40,3	42,8	48,3
23	9,26	10,2	11,7	13,1	14,8	18,1	22,3	27,1	32,0	35,2	38,1	41,6	44,2	49,7
24	9,89	10,9	12,4	13,8	15,7	19,0	23,3	28,2	33,2	36,4	39,4	43,0	45,6	51,2
25	10,5	11,5	13,1	14,6	16,5	19,9	24,3	29,3	34,4	37,7	40,6	44,3	46,9	52,6
26	11,2	12,2	13,8	15,4	17,3	20,8	25,3	30,4	35,6	38,9	41,9	45,6	48,3	54,1
27	11,8	12,9	14,6	16,2	18,1	21,7	26,3	31,5	36,7	40,1	43,2	47,0	49,6	55,5
28	12,5	13,6	15,3	16,9	18,9	22,7	27,3	32,6	37,9	41,3	44,5	48,3	51,0	56,9
29	13,1	14,3	16,0	17,7	19,8	23,6	28,3	33,7	39,1	42,6	45,7	49,6	52,3	58,3
30	13,8	15,0	16,8	18,5	20,6	24,5	29,3	34,8	40,3	43,8	47,0	50,9	53,7	59,7
40	20,7	22,2	24,4	26,5	29,1	33,7	39,3	45,6	51,8	55,8	59,3	63,7	66,8	73,4
50	28,0	29,7	32,4	34,8	37,7	42,9	49,3	56,3	63,2	67,5	71,4	76,2	79,5	86,7
60	35,5	37,5	40,5	43,2	46,5	52,3	59,3	67,0	74,4	79,1	83,3	88,4	92,0	99,6
70	43,3	45,4	48,8	51,7	55,3	61,7	69,3	77,6	85,5	90,5	95,0	100	104	112
80	51,2	53,5	57,2	60,4	64,3	71,1	79,3	88,1	96,6	102	107	112	116	125
90	59,2	61,8	65,6	69,1	73,3	80,6	89,3	98,6	108	113	118	124	128	137
100	67,3	70,1	74,2	77,9	82,4	90,1	99,3	109	118	124	130	136	140	149

Figure 6.21: Tabela de valores “x” da Distribuição Qui-quadrado



Valores do 95º Percentil (nível 0,05), $F_{0,95}$, para a Distribuição F

com graus de liberdade v_1 no numerador e v_2 no denominador.

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Figure 6.22: Tabela de valores “x” da Distribuição F específicos do percentil 95 (há outras tabelas, para outros percentis)

Módulo 7

Introdução ao planejamento de pesquisas

O estudo de uma realidade ainda não compreendida impõe ao pesquisador a formulação de hipóteses sobre suas possíveis causas, qualquer que seja a área do conhecimento:

- ciências biológicas;
- ciências exatas;
- ciências agrárias;
- ciências humanas;
- ciência sociais e outras.

Uma hipótese é uma conjectura racional feita após um grande número de observações e experimentos; é uma tese que precisa ser confirmada ou verificada por meio de novas observações e experimentos.

Uma teoria científica é transitória. Uma conjectura temporariamente sustentada que um dia poderá ser refutada e substituída por outra.

Conclusões baseadas em raciocínios plausíveis são provisórias, ao contrário daquelas produzidas por raciocínios demonstrativos. Considere as hipóteses a seguir:

Exemplo: Crianças socialmente isoladas assistem mais televisão do que crianças bem integradas a seus grupos?



Figure 7.1: Representação esquemática do fluxo de informações da amostra à produção de conhecimento

Exemplo: Famílias constituídas por um só dos genitores (pai ou mãe ausentes) geram mais delinquentes?

Exemplo: Diferentes tipos de uso do solo urbano influenciam na taxa de ocorrência de crimes?

Só após ter-se bem definido pelo pesquisador o que seria uma **criança socialmente isolada** e uma **criança bem integrada a um grupo**; assim como o que seria **família, genitor ausente** e até mesmo o que é **um delinquente**, o que é um **crime** e quais são os **usos do solo urbano** é que se pode avançar com o planejamento da pesquisa até a sua execução (entrevistas com crianças que responderiam o número de horas que passam defronte à televisão por dia ou um levantamento comparativo que permita verificar se há alguma correlação entre o comportamento social e o ambiente familiar de origem).

É necessário ao pesquisador testar suas hipóteses com informações trazidas da realidade estudada mesmo que, aparentemente, pareçam verdadeiras porque, caso contrário, seu julgamento seria conduzido baseado em ideias **pré-concebidas** por experiências pessoais anteriores, muitas vezes tendenciosas, resultando em conclusões científicamente nulas.

7.1 Planejamento de pesquisas

Alguns consideram o artigo publicado em 1895 pelo estatístico norueguês Anders Nicolai Kiaer (*Observations et expériences concernant les dénombremens représentatifs*) como o nascimento oficial da pesquisa por amostragem, apesar de existirem registros anteriores da realização de pesquisas por Laplace, Lavoisier e outros ([link](#)).

Pesquisa é uma investigação sistemática para se obter informações precisas que permitam descrever, explicar o fenômeno que se deseja estudar.

Pesquisas são baseadas em raciocínio lógico e envolve métodos indutivos e dedutivos.

Requerem uma análise aprofundada de todos os dados coletados para que não haja anomalias associadas a eles.

Uma pesquisa cria um caminho para gerar novas perguntas: os dados existentes ajudam a criar mais oportunidades de pesquisa.

Uma pesquisa tem natureza analítica: utiliza todos os dados disponíveis para que não haja ambiguidade na inferência.

A precisão é um dos aspectos mais importantes da pesquisa: as informações obtidas devem ser o mais precisas e verdadeiras possível: precisão nos instrumentos utilizados, nas calibrações de instrumentos ou ferramentas, treinamento de operadores.

7.2 Tipos de pesquisas

Table 7.1: Quadro de tipos de pesquisas conforme sua classificação

Classificação	Tipos de pesquisas
Finalidade	básica (fundamental) aplicada (tecnológica)
Abordagem	qualitativa quantitativa (descritiva ou analítica)
Objetivos	exploratória explicativa
Tempo	transversal longitudinal
Natureza	observacional experimental
Obtenção dos dados	observacional experimental por amostragem

7.2.1 Quanto à finalidade

- na pesquisa básica os dados coletados para aprimorar o conhecimento; a principal motivação é a expansão do conhecimento; é uma pesquisa não comercial que não tem como propósito imediato a criação ou invenção de nada; e,
- uma pesquisa aplicada se concentra na análise e solução de problemas existentes na vida real; refere-se ao estudo que ajuda a resolver problemas práticos usando métodos científicos.

7.2.2 Quanto à forma de abordagem

Os tipos de métodos de pesquisa podem ser amplamente divididos em duas categorias quantitativas e qualitativas:

- a pesquisa quantitativa descreve, infere e resolve problemas usando números; a ênfase é colocada na coleta de dados numéricos, no resumo desses dados e na realização de inferências a partir dos dados;
- a pesquisa qualitativa é baseada em palavras, sentimentos, opiniões, sons e outros elementos não numéricos e não quantificáveis.

7.2.3 Quanto aos objetivos

- uma pesquisa exploratória é conduzida para explorar um grupo de perguntas; as respostas e análises podem não oferecer uma conclusão final para o problema analisado; tem como objetivo lidar com novas problemáticas que não foram exploradas antes;
- uma pesquisa explicativa é conduzida para entender o impacto de certas alterações em procedimentos padrão já estabelecidos; a realização de experimentos é a forma mais popular de pesquisa casual

7.2.4 Quanto ao desenvolvimento no tempo

- em uma pesquisa transversal a análise está fixada em um momento específico no tempo;
- uma pesquisa longitudinal desenrola-se em um período de tempo determinado

7.2.5 Quanto à natureza

- em uma pesquisa observacional o pesquisador atua de modo passivo;
- uma pesquisa experimental o pesquisador é ativo ao promover processos de modo deliberado;
- em uma pesquisa amostral o pesquisador define uma população que apresenta a característica de interesse do estudo.

7.2.6 Quanto à forma de obtenção dos dados

- nos levantamento de dados em uma pesquisa observacional o pesquisador atua meramente como expectador de fenômenos ou fatos, sem, no entanto, realizar qualquer intervenção que possa interferir no curso natural e/ou no desfecho dos mesmos, embora possa, neste meio tempo, realizar medições, análises e outros procedimentos para coleta de dados;
- em pesquisas experimentais o delineamento do experimento estabelece o modo como as variáveis em estudo serão aplicadas ao objeto com o propósito de se obter uma informação (resposta) sobre sua influência para validação ou não de uma hipótese previamente estabelecida;

- levantamentos amostrais são aqueles nos quais os dados são extraídos de um subconjunto tecnicamente extraído de uma população bem definida por meio de procedimentos controlados pelo pesquisador e que podem ser subdivididos em probabilísticos (casuais ou aleatórios) e não probabilísticos (intencionalmente dirigidos).

7.3 Principais etapas de uma pesquisa:

- Definição precisa do objetivo;
- Planejamento;
- Execução;
- Analise dos dados obtidos;
- Resultados; e,
- Conclusões.

7.3.1 Objetivo

Ao se iniciar qualquer pesquisa deve-se ter bem muito bem definido o problema a ser pesquisado, reduzido a uma *hipótese testável*.

Os objetivos de uma pesquisa devem ser elaborados de forma bastante clara (já que as demais etapas da pesquisa tomam como base esses objetivos) e, invariavelmente, envolve uma extensa revisão da literatura existente sobre o assunto.

Exemplo: (objetivo geral) estabelecer o perfil dos estudantes universitários de Londrina para se (objetivos específicos) conhecer a renda média familiar e cidade de origem. Hipótese: a renda média familiar dos estudantes com origem diversa de Londrina é menor que do que os da própria cidade.

Uma vez que o objetivo geral está estabelecido e as hipóteses a serem testadas foram formuladas deve-se definir a população alvo cujos elementos contém a informação desejada considerando as definições estabelecidas para o problema.

- todas as universidades de Londrina (ou apenas as universidades públicas ou particulares);
- todos os cursos (ou algum em particular) ...

7.4 População

Denomina-se por população ao universo de todos os elementos que apresentam a característica (informação) sob estudo (o termo aqui é utilizado em sentido estritamente técnico, nada relacionado ao número de habitantes de um determinado local).

- os pesos dos estudantes de uma determinada escola (população: todos os alunos);
- os salários pagos por uma empresa (população: todos os funcionários legalmente existentes);
- a proporção de indivíduos favoráveis a determinado projeto em uma cidade (população: todos os habitantes dessa cidade);
- a durabilidade das peças sob produção em uma certa fábrica (população: todas as peças produzidas por essa fábrica);
- o número de horas passadas defronte à televisão por crianças até 10 anos de idade no Brasil (população: todas as crianças do Brasil com até 10 anos).

7.5 Censo

Denomina-se por censo à investigação de todos os elementos da população definida, o que resulta em apuração exata da informação requerida na pesquisa.

Todavia, muitos objetos de pesquisa impõem um grau de dificuldade e custo financeiro muito elevados para a execução de um censo o que acaba por tornarem não muito frequentes e, usualmente são realizados apenas pelo estado para dar suporte ao planejamento nacional ou local.

7.6 Amostra

A coleta de dados em toda a população é inviável (ou até mesmo impossível) por diversas razões como, por exemplo:

- tempo e/ou recursos financeiros limitados;
- grande dispersão geográfica da população impondo complicações de ordem logística;
- ensaios destrutivos (corpos de prova) para geração de informações;
- inexistência *a priori* de dados, demandando a realização de experimentos para a sua geração.

Denomina-se por amostra a qualquer subconjunto da população, extraído mediante procedimentos tecnicamente prescritos.

Se a característica em estudo em uma população fosse homogênea em todos os seus elementos, qualquer tamanho de amostra seria suficiente (na realidade, bastaria um elemento dessa população para estudar a característica em toda ela).

Considerando que existe variabilidade da característica nos elementos da população o pesquisador deve usar procedimentos estatísticos para a realização da amostragem e assegurar que tal variabilidade se reflita igualmente na amostra.

Quando a população é grande o estudo de uma fração (amostra) mostra-se mais vantajoso pelas seguintes razões:

- redução de custos;
- redução de prazos: problemas relacionados à data de referência e a imprecisões introduzidas ao se fixar uma data pretérita (dificuldade em se recordar); e,
- maior precisão nas informações: menos entrevistadores (mas com alto nível de treinamento) e procedimentos de acompanhamento mais rigorosos.

Todavia há situações nas quais a extração de uma amostra não recomendada como:

- população pequena
- a característica de interesse é de fácil mensuração na população;
- necessidade de elevada precisão na estimativa.

7.7 Planejamento do levantamento amostral

O planejamento do levantamento amostral deve considerar:

- população objeto: identificar a população total de interesse sobre a qual desejamos obter informações;
- característica populacional: delimitar o aspecto da população que interessa ao estudo;
- unidade amostral: definida de acordo com o interesse do estudo é onde a informação de interesse está; pode ser uma peça, um indivíduo, uma família, uma fazenda, um corpo de prova, etc;

- erro amostral: diferença entre um resultado obtido pela análise da informação trazida por uma amostral específica e o verdadeiro valor da informação na população;
- tamanho da amostra: decorrência do item anterior e também das probabilidades de cometimento de erros do tipo I e II estabelecidas *a priori* (testes de hipóteses)

7.8 Elaboração dos questionários

Um questionário deve ser previamente elaborado de modo a manter o foco na obtenção de dados necessários à pesquisa:

- facilitação da comunicação: a linguagem deve ser a mesma adotada pelo público-alvo; e a redação precisa ortograficamente;
- perguntas ambíguas ou não relacionadas à hipótese a ser testada devem ser evitadas, bem como o uso de termos ou simples palavras que possam induzir o respondente a uma opção;
- respostas possíveis: oferecer todas as possíveis alternativas de resposta para que o respondente possa encontrar sua melhor opção e não desistir da pesquisa;

7.8.1 Tipos de perguntas:

- pergunta desqualificatória: funciona como um filtro para evitar que respondentes que não integrem o público-alvo respondam à pesquisa;
- pergunta de resposta única: modelo de pergunta mais comum;
- pergunta de seleção múltipla: o respondente pode selecionar todas as opções que desejar dentre as alternativas oferecidas;
- pergunta em escala: formato de pergunta onde o respondente escolhe em uma escala de pontos pré-determinada (0 a 5; 0 a 10; 1 a 5, entre outros) e permite uma segunda análise a perguntas com apenas duas opções (*concordo totalmente* ou *discordo totalmente*, por exemplo).

Algumas vantagens de pesquisas virtuais:

- impessoalidade: a ausência do entrevistador induz o respondente a uma resposta sincera;
- conveniência: o respondente pode participar da pesquisa em horário mais flexível;
- abrangência: permite alcançar mais facilmente um maior número de pessoas;
- menor custo envolvido; e,
- facilidade de tabulação: as respostas apresentadas pelo respondente podem ser automaticamente tabuladas e apresentadas na forma de gráficos.

7.8.2 Execução do levantamento amostral

Encaminhamento dos questionários (ou disponibilização em meios virtuais); realização das entrevistas, do experimento ou ainda da observação.

7.8.3 Análise exploratória dos dados

Obtenção de sínteses numéricas, apresentação na forma de tabelas e gráficos de variados formatos das respostas obtidas nos questionários.

7.8.4 Resultados e conclusões

Apresentação dos resultados coerentes com os objetivos estipulados e a conclusão acerca da hipótese inicialmente proposta (rejeição ou não rejeição da hipótese nula contraposta àquela formulada).

7.9 Técnicas de amostragem

O modo de se obter uma amostra é tão importante, e existem tantos modos de fazê-lo, que esses procedimentos constituem especialidades dentro da Estatística.

Todavia os que são mais frequentemente empregados estão representados na Figura ??:

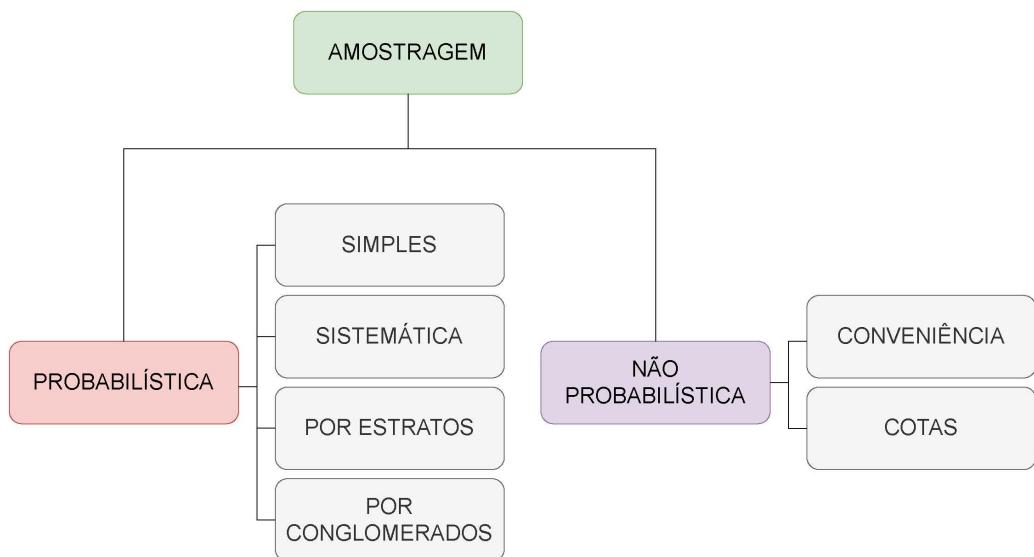


Figure 7.2: Principais procedimentos para se extrair uma amostra

7.10 Amostragem probabilística

Uma amostragem de natureza probabilística é aquela que reúne todas as técnicas pelas quais se deixa completamente ao acaso a escolha dos elementos da população a serem incluídos na amostra. A aleatorização visa

assegurar que a informação extraída da amostra possa ser generalizada na população de origem. A cada extração a probabilidade de um elemento ser incluído é igual para todos (embora ela se altere em razão de ser uma extração sem reposição).

7.10.1 Amostragem aleatória simples (AAS)

Consiste na seleção de n elementos amostrais de tal modo que cada um deles tenha a mesma probabilidade de pertencer à amostra que os demais.



Figure 7.3: Amostra aleatória simples AAS

Duas situações distintas:

- com reposição do elemento amostral escolhido: o mesmo elemento da população pode ser amostrado mais de uma vez (a probabilidade de seleção não se altera); ou,
- sem reposição: cada elemento da população é amostrado uma única vez (a probabilidade de seleção se altera)

Amostragem aleatória simples sem reposição. Admita uma população ($N = 5$) composta pelos elementos: $\{a, b, c, d, e\}$ (podem ser as rendas anuais de cinco pessoas, os pesos de cinco vacas ou cinco modelos diferentes de aviões) da qual se deseja extrair uma amostra de tamanho $n = 3$.

Haverá 10 amostras possíveis de serem extraídas com tamanho 3 ($n = 3$): $\{abc, abd, abe, acd, ace, ade, bcd, bce, bde, cde\}$ pois:

$$C_{(N,n)} = \frac{N!}{n! \times (N-n)!} = 10$$

Amostragem aleatória simples com reposição. Considere agora a mesma população anterior ($N = 5$) e o mesmo tamanho da amostra ($n = 3$). Se a amostragem for feita com reposição teremos então $N^n = 125$ amostras possíveis de serem extraídas: {aaa, aab, aac, aad, aae, aba, abb, abc, abd, abe,}

```
# Dados
conjunto=c("a", "b", "c", "d", "e")

# As 10 combinações possíveis tomando-se 3 elementos:
library(combinat)

## 
## Attaching package: 'combinat'

## The following object is masked from 'package:utils':
##       combn

#combn(conjunto, 3) (remova o # para executar)

# As 125 permutações possíveis tomando-se 3 elementos:
# permn(conjunto) (remova o # para executar)

# Extração de uma amostra (sem reposição) composta por 3 elementos do conjunto:
amostra_sr=sample(conjunto, 3, replace=FALSE)
amostra_sr

## [1] "c" "d" "e"

# Extração de uma amostra (com reposição) composta por 3 elementos do conjunto:
amostra_cr=sample(conjunto, 3, replace=TRUE)
amostra_cr

## [1] "e" "a" "d"
```

Do ponto de vista da quantidade de informação contida na amostra, a amostragem sem reposição é mais adequada.

Todavia a amostragem com reposição conduz a um tratamento teórico mais simples, pois ele implica que tenhamos independência entre as unidades selecionadas (não há alteração na probabilidade de seleção).

Para populações muito grandes a reposição ou não é irrelevante.

Uma vez determinadas as possíveis amostras, segue-se o problema de como elas serão efetivamente extraídas na prática numa amostragem aleatória simples.

Numa situação simples como a que acabamos de conceber poderíamos escrever cada uma das 10 (ou 125) possíveis amostras em um pedaço de papel e colocá-los em uma urna para serem sorteados.

Ou então enumerar os elementos da lista de possibilidades atribuindo um número a cada um e, em seguida, usar uma tabela de números aleatórios (ou um programa computacional para sua geração) para a escolha dos elementos que integrarão a amostra.

Uma AAS raramente é realizada na prática pois é necessário dispor de uma listagem bem definida *a priori*.

Assim, sob circunstâncias reais, um planejamento amostral pode ser definido de modo a assegurar que uma amostra mais informativa, mais barata e rápida possa ser extraída, principalmente quando a amostragem aleatória simples mostrar-se impraticável.

Em estudos de larga escala muitas vezes requerem uma abordagem mista.

A amostragem mista tem vantagens a nível prático, quando se conhecem algumas informações da população; assim sendo define-se uma característica dos elementos a incluir na amostra, deixando-se os restantes fatores ao acaso.

Neste tipo de amostragem salientam-se os seguintes métodos:

- 1- sistemática;
- 2- estratificada; e,
- 3- por conglomerado.

7.10.2 Amostragem aleatória sistemática

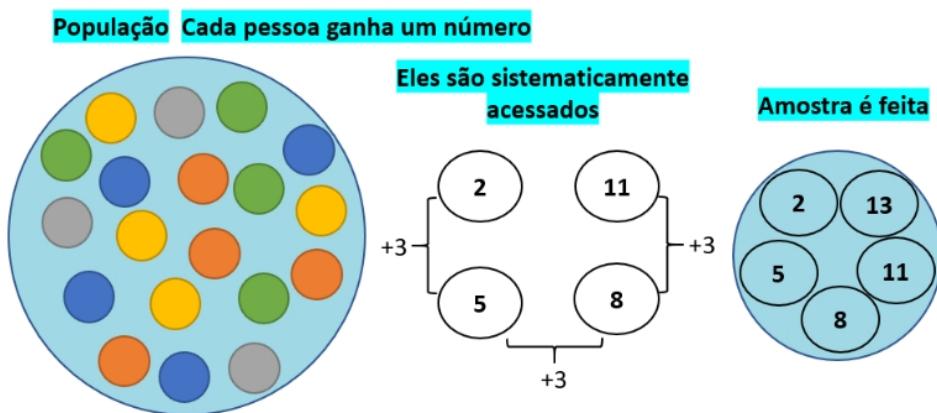


Figure 7.4: Amostra sistemática

Quando os elementos da população estão dispostos sob alguma maneira organizada e aleatória (linha de produção, listagens, ...) a extração de elementos pode ser realizada pela estipulação de um ponto de partida aleatório (o primeiro elemento a ser tomado como integrante da amostra) e de um passo (intervalo), de modo que a seleção dos demais elementos será feita a cada k elementos da listagem.

Roteiro:

- se N é o tamanho da população a ser amostrada;
- e n o tamanho da amostra que se deseja;

calcula-se o passo (intervalo) a ser adotado para a extração dos demais elementos amostrais. O primeiro elemento a ser coletado será aleatoriamente escolhido dentre os k primeiros.

$$S = \frac{N}{n}$$

Sorteia-se o ponto de partida (um dos S números do primeiro intervalo) e depois, a cada S elementos da população, retira-se um para fazer parte da amostra, até completar o valor den.

Algumas situações possíveis de se encontrar:

- se S for fracionário pode-se aumentar n até tornar S um inteiro;
- reduzir N em 1 unidade;
- se N for um número primo, excluem-se por sorteio alguns elementos da população para tornar S inteiro.

Exemplo: considerem uma população composta por pelos seguintes elementos $P=\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ ($N=10$) da qual desejamos extrair uma amostra de tamanho 3 ($n=3$).

O passo S (o intervalo de extração de cada elemento) será igual a $S = \frac{N}{n} = \frac{10}{3} = 3,33$ (fracionário). Aumentando-se para $n = 4$ resultará também em um S fracionário (2,5). Com $n=5$, $S = 2$. O primeiro elemento a integrar a amostra será será aleatoriamente escolhido dentre os 5 (S) primeiros. Assim, as duas possíveis amostras serão:

$$\begin{aligned}A1 &= 1, 3, 5, 7, 9; e, \\A2 &= 2, 4, 6, 8, 10.\end{aligned}$$

Avaliar, alternativamente, excluir aleatoriamente 1 elemento da população ($N = 9$). Mantendo-se $n = 3$ teremos $S = 3$.

$$\begin{aligned}A1 &= 1, 4, 7; \\A2 &= 2, 5, 8; e, \\A3 &= 3, 7, 9.\end{aligned}$$

Exemplo: uma operadora telefônica pretende saber a opinião de seus assinantes comerciais sobre seus serviços na cidade de Florianópolis. Supondo que há 25.037 assinantes comerciais e a amostra precisa ter no mínimo 800 elementos, mostre como seria organizada uma amostragem sistemática para selecionar os respondentes sabendo que a operadora dispõe de uma lista ordenada alfabeticamente com todos os seus assinantes.

Calculando o passo (S):

$$\begin{aligned} S &= \frac{N}{n} \\ &= \frac{25037}{800} \\ &= 31,29 \end{aligned}$$

Aumentar n não irá resolver o problema ($N = 25037$ é um número **primo**). Arredondar S para cima irá extrapolar o tamanho da população ($32 \times 800 = 25600 > 25037$).

Podemos arredondar S para baixo ($31 \times 800 = 24800$) para baixo e excluir **aleatoriamente** 237 elementos da população (é uma população relativamente grande e isso não acarretará problema algum).

Assim nossa amostra será composta por 800 elementos (n) de uma população de (reduzida a) 24800 elementos. Sorteamos **aleatoriamente** o primeiro elemento dentre os 31 primeiros da listagem. Os demais, a cada 31 **elementos**.

Na amostragem sistemática deve-se avaliar o **risco** de periodicidades sistemáticas:

- se lista de elementos estiver organizada com base em alguma informação da população (escolaridade, renda, ...) que possa induzir a algum tipo de viés;
- se em um processo produtivo for sabidamente reconhecido que falhas podem se tornar mais frequentes a cada certo número de unidades produzidas (máquinas descalibradas).

7.10.3 Amostragem aleatória estratificada

Quando se pode identificar na população a presença de **grupos distintos** (estratos) a amostragem estratificada se dá pela realização de amostragens aleatórias simples dentre os elementos de **cada um desses grupos**.

Um **estrato** é uma subdivisão da população onde se observa a existência de uma razoável **homogeneidade interna** da informação desejada. Desse modo, é essencial para que a amostra final tenha qualidade, que **entre os estratos** estabelecidos exista **heterogeneidade** e assim, cada indivíduo pertença a apenas um estrato.

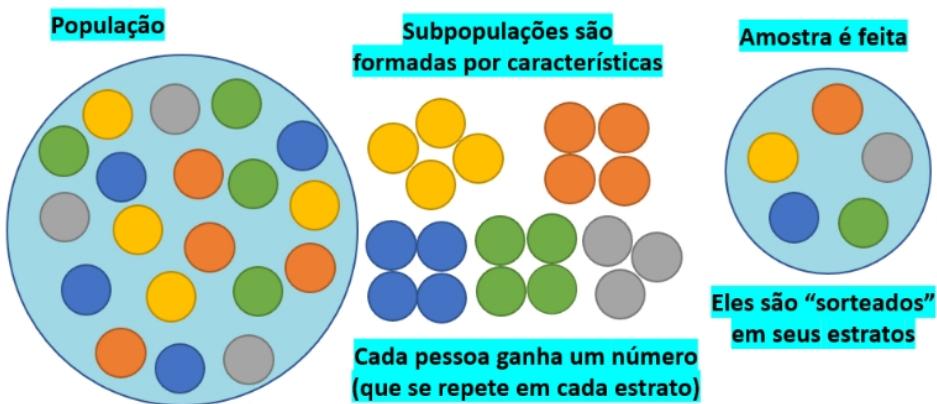


Figure 7.5: Amostra estratificada

Há dois modos possíveis de se realizar uma amostragem estratificada:

- não proporcional; e,
- proporcional.

Em uma amostragem estratificada **não proporcional** o total de elementos extraídos de cada estrato é igual à razão do tamanho da amostra pelo número de estratos (de cada estrato serão escolhidos aleatoriamente um **mesmo número** de elementos).

Esse modo de extração de elementos implica considerar **igual representatividade** de cada estrato na população, **independentemente** de quantos elementos ele abrigue (estratos menores teriam um mesmo peso que estrato maiores).

Já na amostragem estratificada **proporcional** a amostra extraída de cada um dos estratos **segue algum critério de ponderação** do peso ou variabilidade de cada estrato da população.

Na alocação proporcional ao tamanho dos estratos a proporção relativa de cada uma das k amostras extraídas (n_k) em relação ao tamanho de cada um dos k estratos (N_k) é a mesma (garantindo que estratos maiores tenham mais elementos dentro da amostra final e que estratos menores tenham menos presença nela):

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots = \frac{n_k}{N_k}$$

Onde:

- N é o tamanho da população;
- n o tamanho da amostra que se deseja extraír da população;
- N_i é o tamanho do $i - \text{simo}$ estrato da população, tal que $N = N_1 + N_2 + \dots + N_k$;
- n_i o tamanho da $i - \text{sima}$ amostra a ser extraída do $i - \text{simo}$ estrato, tal que $n = n_1 + n_2 + \dots + n_k$.

O tamanho da $i - \text{sima}$ amostra a ser extraída de um $i - \text{simo}$ estrato será determinada em razão do tamanho da amostra que se deseja extraír (n), o tamanho da população (N) e o tamanho do $i - \text{simo}$ estrato (N_i) tal que:

$$n_i = \frac{N_i}{N} \cdot n$$

para $i=1,2,\dots,k$ estratos.

Exemplo: considerem uma comunidade universitária composta 8000 indivíduos ($N=8000$) sendo 800 professores ($N_1 = 800$), 1200 funcionários ($N_2 = 1200$) e 6000 estudantes ($N_3 = 6000$), da qual se estipulou extraír uma amostra de tamanho igual a 900 elementos ($n = 900$) para fins de uma pesquisa sobre o estilo de liderança preferido, que se considera ser diferente para cada grupo componente da comunidade acadêmica.

Numa amostragem estratificada **não proporcional** os elementos são extraídos em igual quantidade de cada um dos estratos:

- 300 professores;
- 300 funcionários; e,
- 300 alunos.

Numa amostragem estratificada uniforme todas os elementos são extraídos em quantidade de modo independente do peso proporcional dos estratos na população. Esse tipo de amostragem apresenta resultados **menos precisos** mas, em contrapartida, estudar características de cada camada de forma mais eficiente.

Numa amostragem estratificada **proporcional** os elementos são extraídos de cada um dos estratos considerando-se seus diferentes tamanhos (suas proporções em relação à população total):

- o estrato dos professores possui $N_p = 800$ elementos;
- o estrato dos funcionários possui $N_f = 1200$ elementos; e,
- o estrato dos estudantes possui $N_e = 6000$ elementos.

Para uma amostra com um total de $n = 900$ elementos seguem-se as quantidades a serem extraídas aleatoriamente de cada um dos três estratos:

- $n_p = \frac{N_p}{N} \cdot n = \frac{800}{8000} \cdot 900 = 90$ professores;
- $n_f = \frac{N_f}{N} \cdot n = \frac{1200}{8000} \cdot 900 = 135$ funcionários;
- $n_e = \frac{N_e}{N} \cdot n = \frac{6000}{8000} \cdot 900 = 675$ alunos;

Partindo-se desses tamanhos amostrais determinados (90 professores, 135 funcionários e 675 alunos) pode-se recorrer à extração sistemática usando-se a listagem dessas três categorias:

- $S_p = \frac{N_p}{n_p}$ em que S_p é o passo a ser seguido na extração de n_p professores (90) da “população” de N_p professores (800);
- $S_f = \frac{N_f}{n_f}$ em que S_f é o passo a ser seguido na extração de n_f funcionários (135) da “população” de N_f funcionários (1200); e,
- $S_e = \frac{N_e}{n_e}$ em que S_e é o passo a ser seguido na extração de n_e alunos (675) da “população” de N_e alunos (6000).

Muitos ajustes devem ser feitos pois, de modo frequente, os resultados dos passos obtidos na prática resultam em números fracionários. Todavia, devemos ter sempre procurar não reduzir o tamanho amostral e ter em mente que o tamanho da população não pode ser aumentado.

Para os professores: $S_p = \frac{800}{90} = 8,88$. Se tomamos $S_p = 9$ (um professor a cada nove da lista) veremos que para extrair 90 professores a população teria de ser de 810 (a população é de 800). Uma das opções seria usar $S_p = 8$ e se extrair 100 professores (uma amostra um pouco maior). Outra possibilidade seria ainda usar $S_p = 8$ remover aleatoriamente 80 professores da população e então tomar 90 professores (pois com $S_p = 8$, $8 \cdot 90 = 720$).

Para os funcionários: $S_f = \frac{1200}{135} = 8,88$ (o mesmo porque essas amostras foram estabelecidas de modo proporcional). Se tomamos $S_f = 9$ (um funcionário a cada nove da lista) veremos que para extrair 135 funcionários a população teria de ser de 1215 (a população é de 1200). Uma das opções seria usar $S_f = 8$ e se extrair 150 funcionários (uma amostra um pouco maior). Outra possibilidade seria ainda usar $S_f = 8$ remover aleatoriamente 120 funcionários da população e então tomar 135 funcionários (pois com $S_f = 8$, $8 \cdot 135 = 1080$).

Do mesmo modo para os alunos: $S_e = \frac{6000}{675} = 8,88$ (o mesmo porque essas amostras foram estabelecidas de modo proporcional). Se tomamos $S_e = 9$ (um aluno a cada nove da lista) veremos que para extrair 675 alunos a população teria de ser de 6075 (a população é de 6000). Uma das opções seria usar $S_e = 8$ e se extrair 750 alunos (uma amostra um pouco maior). Outra possibilidade seria ainda usar $S_e = 8$ remover aleatoriamente 600 funcionários da população e então tomar 600 alunos (pois com $S_e = 8$, $8 \cdot 135 = 1080$).

Ao final poderíamos extrair então 100 professores, 150 funcionários e 750 alunos; ou, pela segunda possibilidade, extrair 90 professores, 135 funcionários e 600 alunos (eliminando-se aleatoriamente elementos das populações antes de se sistematizar a extração, como antes explicado).

A proporção de elementos extraídos de cada um dos estratos é constante entre os extratos, assegurando uma extração proporcional:

$$\frac{100}{800} = \frac{150}{1200} = \frac{750}{6000} = 0,125 \quad \frac{90}{720} = \frac{135}{1080} = \frac{675}{5400} = 0,125$$

Pode-se **otimizar** uma amostragem estratificada proporcional considerando também sua variabilidade interna. O tamanho de cada uma das amostras (n_1, n_2, \dots, n_k) dos diferentes estratos são proporcionais aos **tamanhos** dos estratos (N_1, N_2, \dots, N_k) e **também** segundo algum critério adicional (otimização), como a variabilidade interna de cada estrato ($\sigma_1, \sigma_2, \dots, \sigma_k$) de modo a se manter iguais as razões:

$$\frac{n_1}{N_1 \cdot \sigma_1} = \frac{n_2}{N_2 \cdot \sigma_2} = \dots = \frac{n_k}{N_k \cdot \sigma_k}$$

Onde:

- N é o tamanho da população;

- n o tamanho da amostra que se deseja extrair da população;
- N_i é o tamanho do $i - \text{simo}$ estrato da população, tal que $N = N_1 + N_2 + \dots + N_k$;
- n_i o tamanho da $i - \text{sima}$ amostra a ser extraída do $i - \text{simo}$ estrato, tal que $n = n_1 + n_2 + \dots + n_k$; e,
- σ_i é o desvio padrão do $i - \text{simo}$ estrato.

O tamanho da $i - \text{sima}$ amostra a ser extraída de um $i - \text{simo}$ estrato será determinada em razão do tamanho da amostra que se deseja extrair (n), o tamanho da população (N), do tamanho e variabilidade do $i - \text{simo}$ estrato (N_i e σ_i) tal que:

$$n_i = \frac{n \cdot N_i \cdot \sigma_i}{N_1 \cdot \sigma_1 + N_2 \cdot \sigma_2 + \dots + N_k \cdot \sigma_k}$$

para $i=1,2,\dots, k$ estratos.

Exemplo: considere estudar a opinião de estudantes de uma universidade com relação à legalização do aborto. A equipe possui dados descritivos relacionados ao sexo, orientação religiosa e rendimento médio familiar de toda a comunidade acadêmica. Na revisão bibliográfica identifica-se que algumas das variáveis que habitualmente implicam em opiniões diferentes (escolaridade e idade) já não mais precisam ser consideradas; todavia, outras ainda devem ser consideradas. Assim, um plano de estratificação de vários níveis pode ser estabelecido partindo-se da premissa de homogeneidade de opinião interna em cada um deles: sexo, orientação religiosa e rendimento familiar.

Considerando uma amostra de $n = 1.000$ estudantes e as seguintes medidas descritivas disponibilizadas pela universidade e relacionadas à sua população de estudantes:

- sexo: 35% masculino e 65% feminino;
- orientação religiosa: 60% católica; 20% evangélica; 10% sem; 5% espírita e 5% outras; e,
- rendimento médio mensal familiar: 35% até R\$ 4.000,00, 65% acima de R\$ 4.000,00.

podemos estabelecer várias camadas estratificadas proporcionalmente, tal como ilustrado na Figura 7.6.

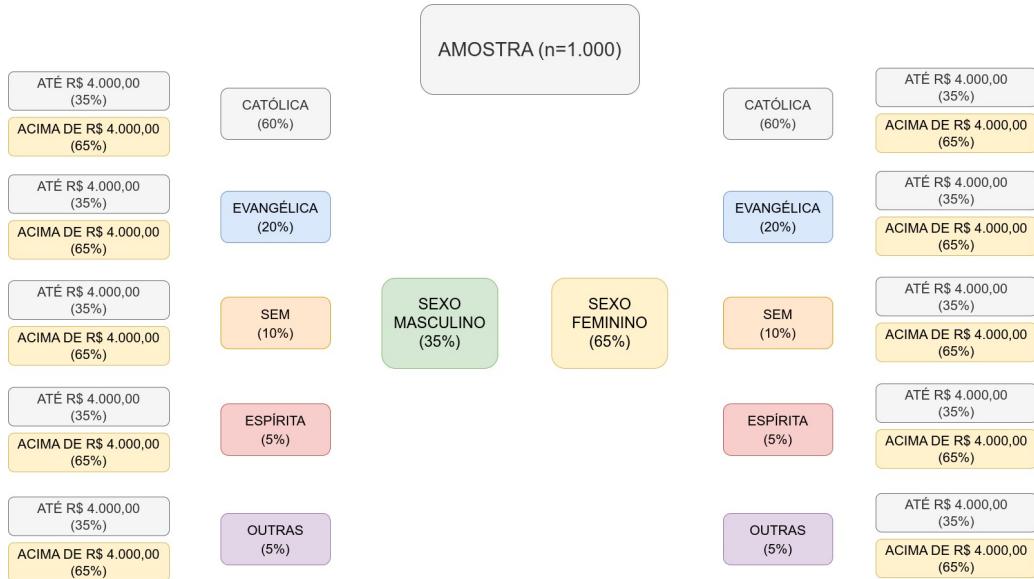


Figure 7.6: Plano de estratificação proporcional

7.10.4 Amostragem aleatória por conglomerados



Figure 7.7: Amostragem por conglomerados

Muitas vezes a **dispersão espacial** de uma população a ser investigada torna impeditiva uma amostragem aleatória simples.

Um modo de contornar essa dificuldade é dividir a área total onde se assenta a população de interesse em várias *áreas geográficas menores* e sem sobreposição, tais como cidades, regionais de cidades, bairros, quarteirões de um bairro, Essa subdivisão pode também ser realizada valendo-se de critérios organizacionais como, por exemplo, universidades, escolas, grau escolar, departamentos de uma empresa,

As subpopulações que se localizam nessas áreas menores passam a ser denominadas de conglomerados e são como que representações **em escala reduzida** da população total.

A **heterogeneidade** presente na população original passa a estar representada dentro de um conglomerado. Ou seja, é essencial para a qualidade final da amostra extraída desse modo, que os elementos dentro de cada conglomerado sejam tão **diversos** quanto a diversidade que se observa nos elementos da população total (a ideia de representação em escala reduzida).

Em uma amostragem de **apenas 1 estágio**, após serem aleatoriamente sorteados um certo número de conglomerados, todos os elementos internos desses conglomerados são estudados.

Todavia, considerando que os elementos de um conglomerado natural dentro de uma população são habitualmente mais homogêneos do que os elementos da população total (os moradores de um bairro são mais semelhantes entre si do que todos os moradores do município), **pode não ser** necessário um grande número de elementos para se representar adequadamente um conglomerado natural.

Uma diretriz científica num processo de amostragem por conglomerados é **maximizar o número de conglomerados** e **diminuir** o número de elementos aleatoriamente escolhidos **dentro** de cada um deles.

Recomenda-se observar as diferenças de tamanho existentes entre cada conglomerado, de modo a equilibrar a probabilidade. A probabilidade de seleção de um elemento num desenho de amostragem com probabilidade proporcional ao tamanho:

- na primeira etapa é dada a cada conglomerado uma oportunidade de seleção **proporcional** ao seu tamanho; e,
- na segunda etapa um **mesmo número** de elementos é escolhido dentro de cada conglomerado selecionado.

Esses procedimentos igualam as probabilidades últimas de seleção de todos os elementos da população pois:

- conglomerados com mais elementos têm maior probabilidade de serem selecionados; e,
- elementos em conglomerados maiores têm menor chance de seleção do que elementos em conglomerados menores.

Exemplo: a população universitária de Londrina (estimada em 25.000 estudantes) pode ser entendida como distribuída em vários **conglomerados organizacionais** como, por exemplo: UEL; UNIFIL; PUC; INESUL; UTFPr; Arthur Thomas; CESUMAR; Pitágoras; Positivo;

Se desejamos realizar uma pesquisa entre os estudantes universitários de Londrina (na qual sabe-se que não fará diferença se a instituição é pública ou privada) podemos sortear aleatoriamente alguns desses conglomerados.

Entretanto, lembrando que todos os elementos de um conglomerado devem ser entrevistados, pode ser que o número de estudantes em cada conglomerado escolhido ainda seja por demais elevado.

Nesse caso, um **segundo estágio** (como, por exemplo, utilizar a subdivisão administrativa que as universidades habitualmente adotam ao se subdividir em diversos centros de estudos como conglomerados dentro dela) pode ser proposto.

Assim como na estratificação, a proposição de conglomerados deve sempre considerar as variáveis condicionantes relacionadas com o objeto de estudo para que as informações de todas as unidades amostrais finais a serem entrevistadas possa ser usada seguramente para se inferir sobre a informação na população sob estudo.

Exemplo: a Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE coleta informações demográficas e socioeconômicas sobre a população brasileira. Sinteticamente, utiliza amostragem por conglomerados em três estágios:

- primeiro estágio: amostras de municípios (conglomerados) para cada uma das regiões geográficas do Brasil (Norte, Nordeste, Centro-Oeste, Sudeste e Sul);
- segundo estágio: setores censitários sorteados (subdivisão estabelecida pelo IBGE dentro de um município) em cada município (conglomerado sorteado);
- terceiro estágio: domicílios sorteados aleatoriamente em cada setor censitário.

Exemplo: considere estudar a opinião de estudantes universitários de toda uma cidade com relação à legalização do aborto. A equipe possui dados descritivos relacionados ao sexo, orientação religiosa e rendimento médio familiar de toda a comunidade acadêmica. Na revisão bibliográfica identifica-se que algumas das variáveis que habitualmente implicam em opiniões diferentes (escolaridade e idade) já não mais precisam ser consideradas; todavia, outras ainda devem ser consideradas. Assim, um plano de estratificação de vários níveis pode ser estabelecido partindo-se da premissa de homogeneidade de opinião interna em cada um deles: sexo, orientação religiosa e rendimento familiar.

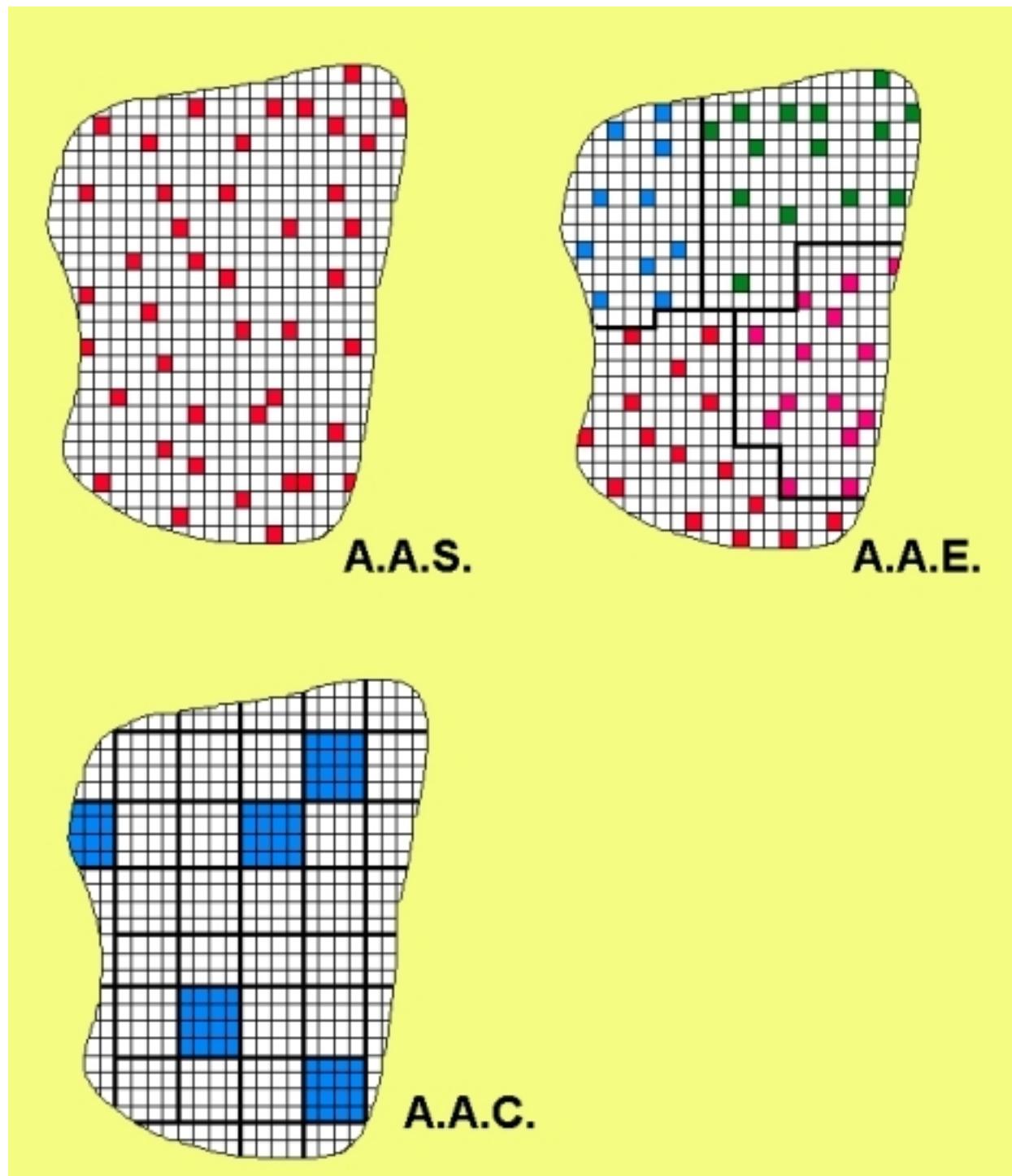


Figure 7.8: Ilustração comparativa dos principais modos de extração de amostras

Nesse caso, podemos considerar cada universidade como um conglomerado. Numa primeira etapa promovemos um sorteio e, na sequência, uma estratificação da amostra total em termos da população estudantil de cada conglomerado. A partir desse ponto, em cada universidade, estratificações suplementares são feitas para se considerar proporcionalmente as diferentes opiniões (sexo, orientação religiosa, renda).

Ao final, após vários estágios, uma amostra não probabilística pode ser extraída de cada grupo individualizado anteriormente, tal como a ilustrado na Figura 7.9.

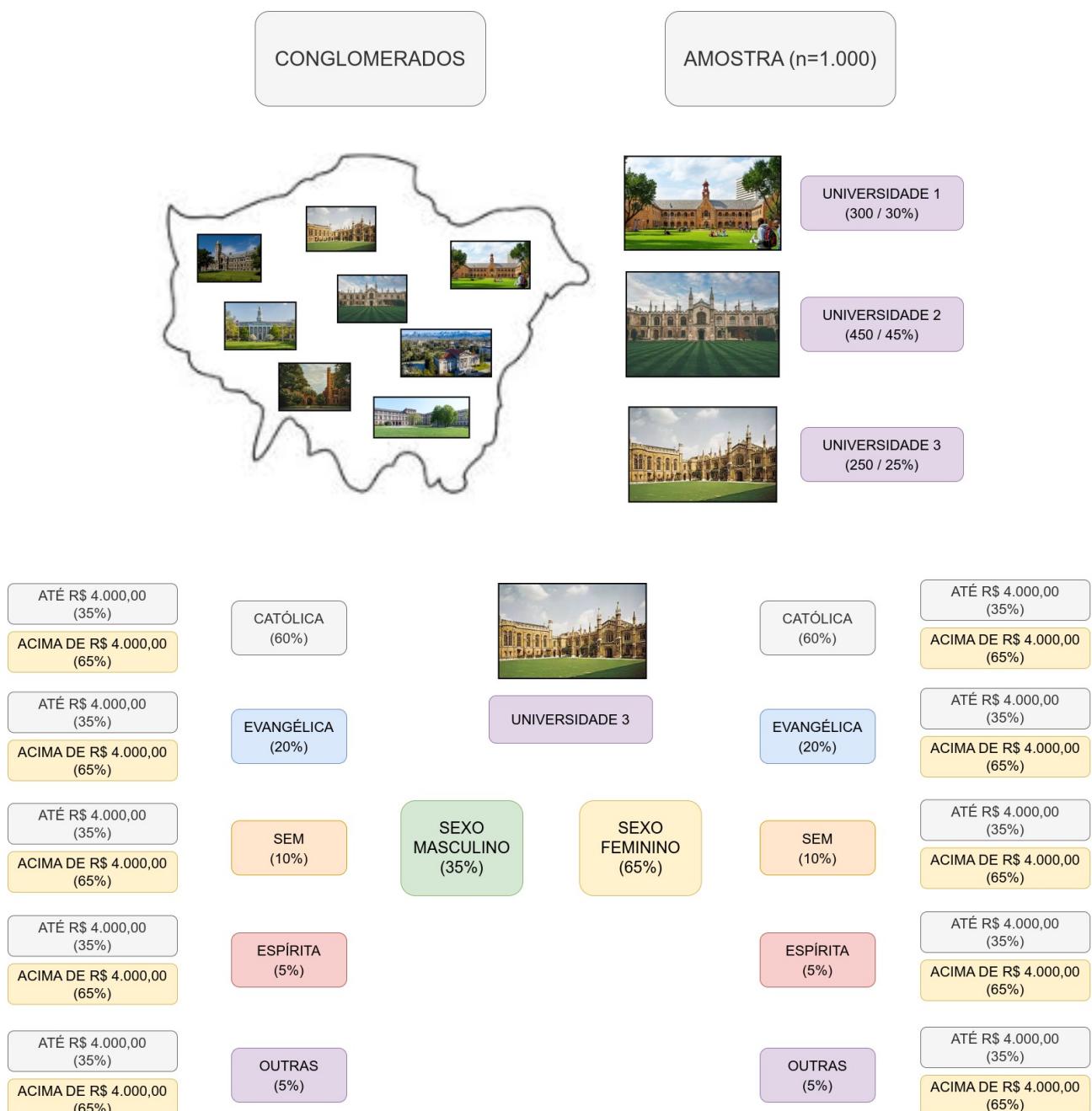


Figure 7.9: Planejamento da extração da amostra

7.11 Amostragem não probabilística

Não obstante os métodos de amostragem probabilísticos serem adequados à generalização da informação colhida, há diferentes situações para as quais podemos nos decidir por métodos probabilísticos como, por exemplo, para tornar a pesquisa menos custosa financeiramente ou ainda porque talvez não seja necessário ter um elevado rigor e precisão nas estimativas que se deseja obter.

Amostragens não probabilísticas são aquelas em que a amostra é extraída de modo *dirigido* (*intencional*, não aleatório) pelo pesquisador em decorrência da natureza de seu estudo, não sendo consideradas a probabilidade de seleção de seus elementos.

7.11.1 Amostragem por conveniência

Esta técnica é muito comum e consiste em se selecionar uma amostra da população imediatamente acessível (prontamente disponível). Considerem, por exemplo, pesquisar a opinião de estudantes universitários em Londrina sobre determinado assunto.

Poderíamos considerar cada universidade localizada em Londrina como um conglomerado e, dentro delas, realizar uma amostragem aleatória de todos os seus estudantes (ou parte, se realizarmos o delineamento em mais de um estágio).

Por conveniência podemos simplesmente decidir ir a um encontro de estudantes universitários que se realiza na cidade e perguntar a alguns deles que se declarem estudar em Londrina qual sua opinião sobre aquele assunto.

As limitações desse tipo de amostragem são óbvias posto poder haver no grupo de entrevistados diferentes segmentos sociais, econômicos, políticos, filosóficos, religiosos dentre muitos outros fatores de diferenciação, que podem ser fundamentais face às opiniões que se deseja colher sobre o assunto inquerido, resultando em graves distorções.

Esse tipo de amostragem, embora não aleatória, é bastante utilizada na área de *marketing* na qual geralmente as amostras são obtidas em locais com aglomerações, como teatros, cinemas, mercados, Neste caso, é importante o senso crítico do pesquisador para evitar vieses, por exemplo, não selecionar sempre pessoas de mesmo sexo, de mesma faixa etária,

7.11.2 Amostragem por cotas

A amostragem por cotas assemelha-se com a amostragem estratificada proporcional; mas, ao contrário da amostragem estratificada, a seleção final (no estrato) não precisa ser aleatória. A população é vista de forma segregada (estratificada), dividida em diversos subgrupos como sexo, idade, raça, local de residência, ocupação,

Para compensar a falta de aleatoriedade na seleção, costuma-se dividir a população num grande número de subgrupos e seleciona-se (não aleatoriamente) uma quantidade de elementos em cada subgrupo, proporcional ao seu tamanho.

Numa pesquisa socioeconômica, a população pode ser dividida por localidade, por nível de instrução, por faixas de renda, ...

7.12 Dimensionamento de amostras

7.12.1 Erros

Há de distinguir dois tipos de erros associados a levantamentos amostrais:

- erros amostrais, as diferenças entre o resultado obtido em uma amostra específica (uma estatística) e seu verdadeiro valor na população (o parâmetro);
- erros não amostrais (experimentais), decorrentes de dados amostrais coletados incorretamente, inconsistente, fruto de erros nas transcrições, delineamentos fracamente estabelecidos (resultando em amostras tendenciosas), leituras instrumentais imprecisas (resultantes da perda da calibração dos instrumentos ou operação por técnicos com diferentes habilidades).

Os erros amostrais ocorrem porque as amostras são aleatórias: se de um grupo de 100 números extraímos uma amostra aleatória de 10 deles a média amostral calculada teria um valor diferente a cada diferente amostra extraída (essa flutuação é assunto da teoria da distribuição das médias e proporções amostrais). Já os erros não amostrais devem ser minimizados ou melhor não existir.

A determinação do tamanho de uma amostra (n_0) é função do *erro amostral* tolerável e do *nível de significância* α estabelecido *a priori* pelo pesquisador que se relaciona ao *nível de confiança* pretendido por $(1 - \alpha)$.

Assim, se o **nível de significância** máximo admissível para o assunto pesquisado é $\alpha = 0,05$, o **nível de confiança** será $(1 - \alpha) = 0,95$ (uma vez que: $\alpha + (1 - \alpha) = 1$).

Table 7.2: Valores críticos de z_c correspondentes a alguns níveis de significância (confiança)

Níveis de confiança $(1 - \alpha)$	0,80	0,90	0,95	0,99	0,999
Níveis de significância α	0,20	0,10	0,05	0,01	0,001
z_c	1,28	1,64	1,96	2,57	3,29

Todavia, como mais adiante se verá, há situações nas quais o valor crítico referente ao nível de confiança estabelecido e que será empregado no dimensionamento da amostra será obtido de uma outra distribuição (*t* de *Student*).

7.12.2 Determinação do tamanho de uma amostra para estimação da média populacional

Determinação do tamanho n_0 de uma amostra para estimação da média considerando-se uma **população infinita** ($N \geq 20.n_0$) e seguindo uma distribuição Normal:

$$n_0 = \frac{z_c^2 \cdot \sigma^2}{\varepsilon^2}$$

em que:

- n_0 : é o tamanho amostral;
- z_c : valor crítico tabelado da distribuição Normal usado para o nível de significância desejado (por exemplo, para $\alpha=5\%$, $z_c = 1,96$);
- σ desvio padrão populacional obtido em estudos prévios; e,
- ε : é o erro amostral, a máxima diferença entre μ e \bar{x} que se espera observar sob um nível de confiança de $(1 - \alpha)$.

Exemplo: Qual o tamanho de amostra necessária para se estimar o peso médio de cervos em uma dada população sob estudo, admitida **infinita**. Sabe-se de estudos anteriores que o desvio padrão σ do peso para animais dessa idade é de 30 kg. Utilize um erro ε de 10 kg na estimativa e um nível confiança $(1 - \alpha)$ de 95%.

$$\begin{aligned} n_0 &= \frac{Z^2 \cdot \sigma^2}{\varepsilon^2} \\ n_0 &= \frac{1,96^2 \cdot 30^2}{10^2} \\ n_0 &\sim 35 \end{aligned}$$

Se a população **não pode ser considerada infinita**, ou seja, se $N < 20.n_0$, então aplica-se uma correção sobre o valor inicialmente calculado para a (n_0) obtendo-se um novo tamanho (n):

$$n = \frac{N \cdot n_0}{N + n_0}$$

No exemplo anterior, caso a população sob estudo fosse composta por apenas 200 animais ($N < 20.n_0$) o tamanho da amostra seria:

$$\begin{aligned} n &= \frac{N \cdot n_0}{N + n_0} \\ n &= \frac{200 \cdot 35}{200 + 35} \\ n &\sim 30 \end{aligned}$$

O conhecimento prévio do **desvio padrão populacional** (σ) para utilizar as expressões acima é quase que uma exceção. Na maioria dos estudos ele é desconhecido e a única informação disponível acerca da variabilidade é o **desvio padrão amostral** S .

Nesse cenário, a variável Norma padronizada Z é substituída por uma outra, que segue a distribuição “t” de *Student* e, para se obter seu valor crítico t_c para um determinado nível de confiança desejado necessitamos ter uma informação adicional: os *graus de liberdade* (gl ou df), que são iguais ao tamanho da amostra **menos 1** ($gl = n_0 - 1$). Observa-se que para $n \rightarrow \infty$, os valores críticos de $z_c = t_c$ para um mesmo nível de significância.

Ocorre porém que, não tendo ainda sido retirada a amostra, não dispomos do valor de s . Se não conhecemos nem ao menos um limite superior para σ , a única solução será colher uma amostra piloto de n_0 elementos para, com base nela obtermos uma estimativa de s e estimarmos o tamanho amostral pela expressão:

$$n = \frac{t_c^2 \cdot s^2}{\varepsilon^2}$$

com s calculado sobre a amostra piloto de n_0 elementos e com t_c obtido em uma tabela de valores da distribuição “t”.

Essas tabelas de valores consideram nas suas **colunas** variados níveis de significância α e, nas suas **linhas** uma informação chamada de **graus de liberdade (gl)**

Graus de liberdade não mais são, no contexto estudado, que o tamanho da amostra piloto n_0 menos 1 ($gl = n_0 - 1$). Assim, se a amostra piloto for de 5 elementos, $gl=4$ e será nessa linha, junto à coluna do nível de significância desejado, que o valor “t” será encontrado.

p ▶	90%	80%	70%	60%	50%	40%	30%	20%	10%	8%	6%	5%	4%	2%	1%	0,2%	0,1%
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	7,916	10,579	12,706	15,895	31,821	63,657	318,309	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	3,320	3,896	4,303	4,849	6,965	9,925	22,327	31,599
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	2,605	2,951	3,182	3,482	4,541	5,841	10,215	12,924
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,333	2,601	2,776	2,999	3,747	4,604	7,173	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,191	2,422	2,571	2,757	3,365	4,032	5,893	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,104	2,313	2,447	2,612	3,143	3,707	5,208	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,046	2,241	2,365	2,517	2,998	3,499	4,785	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,004	2,189	2,306	2,449	2,896	3,355	4,501	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	1,973	2,150	2,262	2,398	2,821	3,250	4,297	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	1,948	2,120	2,228	2,359	2,764	3,169	4,144	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	1,928	2,096	2,201	2,328	2,718	3,106	4,025	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	1,912	2,076	2,179	2,303	2,681	3,055	3,930	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	1,899	2,060	2,160	2,282	2,650	3,012	3,852	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	1,887	2,046	2,145	2,264	2,624	2,977	3,787	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	1,878	2,034	2,131	2,249	2,602	2,947	3,733	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	1,869	2,024	2,120	2,235	2,583	2,921	3,686	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	1,862	2,015	2,110	2,224	2,567	2,896	3,646	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	1,855	2,007	2,101	2,214	2,552	2,878	3,610	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	1,850	2,000	2,093	2,205	2,539	2,861	3,579	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	1,844	1,994	2,086	2,197	2,528	2,845	3,552	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	1,840	1,988	2,080	2,189	2,518	2,831	3,527	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	1,835	1,983	2,074	2,183	2,508	2,819	3,505	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	1,832	1,978	2,069	2,177	2,500	2,807	3,485	3,768
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	1,828	1,974	2,064	2,172	2,492	2,797	3,467	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	1,825	1,970	2,060	2,167	2,485	2,787	3,450	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	1,822	1,967	2,056	2,162	2,479	2,779	3,435	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	1,819	1,963	2,052	2,158	2,473	2,771	3,421	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	1,817	1,960	2,048	2,154	2,467	2,763	3,408	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	1,814	1,957	2,045	2,150	2,462	2,756	3,396	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	1,812	1,955	2,042	2,147	2,457	2,750	3,385	3,646
31	0,127	0,256	0,389	0,530	0,682	0,853	1,054	1,309	1,696	1,810	1,952	2,040	2,144	2,453	2,744	3,375	3,633
32	0,127	0,255	0,389	0,530	0,682	0,853	1,054	1,309	1,694	1,808	1,950	2,037	2,141	2,449	2,738	3,365	3,622
33	0,127	0,255	0,389	0,530	0,682	0,853	1,053	1,308	1,692	1,806	1,948	2,035	2,138	2,445	2,733	3,356	3,611
34	0,127	0,255	0,389	0,529	0,682	0,852	1,052	1,307	1,691	1,805	1,946	2,032	2,136	2,441	2,728	3,348	3,601
35	0,127	0,255	0,388	0,529	0,682	0,852	1,052	1,306	1,690	1,803	1,944	2,030	2,133	2,438	2,724	3,340	3,591
36	0,127	0,255	0,388	0,529	0,681	0,852	1,052	1,306	1,688	1,802	1,942	2,028	2,131	2,434	2,719	3,333	3,582
37	0,127	0,255	0,388	0,529	0,681	0,851	1,051	1,305	1,687	1,800	1,940	2,026	2,129	2,431	2,715	3,326	3,574
38	0,127	0,255	0,388	0,529	0,681	0,851	1,051	1,304	1,686	1,799	1,939	2,024	2,127	2,429	2,712	3,319	3,566
39	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,304	1,685	1,798	1,937	2,023	2,125	2,426	2,708	3,313	3,558
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	1,796	1,936	2,021	2,123	2,423	2,704	3,307	3,551
45	0,126	0,255	0,388	0,528	0,680	0,850	1,049	1,301	1,679	1,791	1,929	2,014	2,115	2,412	2,690	3,281	3,520
50	0,126	0,255	0,388	0,528	0,679	0,849	1,047	1,299	1,676	1,787	1,924	2,009	2,109	2,403	2,678	3,261	3,496
55	0,126	0,255	0,387	0,527	0,679	0,848	1,046	1,297	1,673	1,784	1,920	2,004	2,104	2,396	2,668	3,245	3,476
60	0,126	0,254	0,387	0,527	0,679	0,848	1,045	1,296	1,671	1,781	1,917	2,000	2,099	2,390	2,660	3,232	3,460
70	0,126	0,254	0,387	0,527	0,678	0,847	1,044	1,294	1,667	1,776	1,912	1,994	2,093	2,381	2,648	3,211	3,435
80	0,126	0,254	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,773	1,908	1,990	2,088	2,374	2,639	3,195	3,416
90	0,126	0,254	0,387	0,526	0,677	0,846	1,042	1,291	1,662	1,771	1,905	1,987	2,084	2,368	2,632	3,183	3,402
100	0,126	0,254	0,386	0,526	0,677	0,845	1,042	1,290	1,660	1,769	1,902	1,984	2,081	2,364	2,626	3,174	3,390
110	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,659	1,767	1,900	1,982	2,078	2,361	2,621	3,166	3,381
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,766	1,899	1,980	2,076	2,358	2,617	3,160	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,751	1,881	1,960	2,054	2,326	2,576	3,090	3,291

Figure 7.10: Tabela t de Student: cada linha refere-se a um gl e cada coluna a um nível de significância (no cruzamento tem-se o valor crítico de t sob essas condições)

Observe que à medida que o tamanho da amostra cresce, o valor crítico t_c se aproxima do valor crítico c para um mesmo nível de significância. Por exemplo, para um $\alpha = 5\%$ uma amostra de 121 ($df=121-1=120$) elementos possui um valor crítico $t_c = 1,96$ (na distribuição de Student) e um valor crítico $z_c = 1,96$ (distribuição Normal padrão).

Se $n \leq n_0$, a amostra piloto já terá sido suficiente para ser usada na análise. Caso contrário, deve-se retirar mais elementos da população, recalcular o tamanho da amostra n até se observe essa desigualdade.

7.12.2.1 Margem de erro em uma estimativa amostral da média

Reescrevendo-se a expressão para a determinação do tamanho amostral podemos exprimir o erro ε associado à estimativa obtida de uma amostra de tamanho n : \hat{p} da média populacional

$$\varepsilon = z_c \cdot \sqrt{\frac{\sigma^2}{n}}$$

em que ε é uma quantidade para **mais e para menos** da estimativa obtida de uma amostra de tamanho n em relação a μ sob o nível de confiança $1 - \alpha$ que determina z_c .

A expressão anterior considera que a variância populacional σ^2 é conhecida. Caso não se tenha informação alguma sobre seu valor, seguem-se as mesmas considerações relacionadas ao tamanho n da amostra:

- se $n \geq 30$, adotar a variância amostral S^2 como aproximação de σ^2 ;
- se $n < 30$, adotar a variância amostral S^2 como aproximação de σ^2 usando-se o valor crítico t_c da distribuição de *Student* (com gl/df iguais ao tamanho da amostra menos 1)

7.12.3 Determinação do tamanho de uma amostra para estimação da proporção populacional

A determinação do tamanho de uma amostra para estimação da proporção populacional considerando-se uma **população infinita** ($N \geq 20.n_0$):

$$n_0 = \frac{z_c^2 \cdot \pi \cdot (1 - \pi)}{\epsilon^2}$$

em que:

- n_0 é o tamanho da amostra;
- z_c é valor crítico tabelado da distribuição Normal para o nível de significância desejado (por exemplo, para $\alpha=5\%$, $z_c=1,96$);

- π é a proporção populacional;
- ε : é o erro amostral, a máxima diferença entre π e p que se espera observar sob um nível de confiança de $(1 - \alpha)$.

Quando não se dispõe de nenhuma informação *a priori* sobre a proporção populacional (π) a adoção do máximo valor possível ao produto: $\pi.(1 - \pi) = \frac{1}{4}$ assegura que o tamanho da amostra obtido será suficiente para a estimativa qualquer que seja a proporção populacional π .

Isso equivale a considerar:

$$n_0 = \frac{z_c^2}{\varepsilon^2} \cdot \frac{1}{4}$$

De modo análogo, se a população **não pode ser considerada infinita** ($N < 20n_0$) aplica-se uma correção sobre o valor calculado do tamanho da amostra (n_0) chegando-se a um novo tamanho (n):

$$n = \frac{N \cdot n_0}{N + n_0}$$

Exemplo: Qual o tamanho da amostra (n_0) suficiente para estimarmos a proporção da área com solo contaminado que necessita de certo tratamento de descontaminação, com precisão (ε) de 0,02 e um nível de confiança ($1 - \alpha$) de 95%, sabendo que essa proporção seguramente não é superior a 0,2?

$$\begin{aligned} n_0 &= \frac{z_c^2 \cdot \pi \cdot (1 - \pi)}{\varepsilon^2} \\ n_0 &= \frac{1,96^2 \cdot 0,20 \cdot 0,80}{0,02^2} \\ n_0 &\sim 1.537 \end{aligned}$$

Considerando-se uma estimativa conservadora para $\pi.(1 - \pi)$ pelo máximo valor possível desse produto ($\frac{1}{4}$) teremos:

$$\begin{aligned} n_0 &= \frac{z_c^2}{\varepsilon^2} \cdot \frac{1}{4} \\ n_0 &= \frac{1,96^2}{0,02^2} \cdot \frac{1}{4} \\ n_0 &= 2.401 \end{aligned}$$

7.12.3.1 Margem de erro em uma estimativa amostral da proporção

Reescrevendo-se a expressão para a determinação do tamanho amostral para a situação na qual não temos nenhuma informação sobre a proporção populacional (π), podemos exprimir o erro ε associado à estimativa da proporção (p) obtida de uma amostra de tamanho n da proporção populacional (π) sob o nível de confiança $(1 - \alpha)$ pelo critério mais conservador $(\pi \cdot (1 - \pi) = \frac{1}{4})$

$$\begin{aligned} \varepsilon &= z_c \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} \\ \varepsilon &= z_c \cdot \sqrt{\frac{\frac{1}{4}}{n}} \end{aligned}$$

em que ε é uma quantidade para **mais e para menos** da estimativa p obtida de uma amostra de tamanho n em relação a π sob o nível de confiança $1 - \alpha$ que determina z_c .

Exemplo: Uma pesquisa recente mostra o apoio dos eleitores a uma posição de liberação das restrições sobre a pesquisa de células estaminais embrionárias e permitir o uso médico do princípio ativo da *cannabis sativa*. A pesquisa realizada para o *The Detroit News* descobriram que 50% dos prováveis eleitores de Michigan apoiam a proposta de células-tronco, 32% são contra e 18% indecisos. A pesquisa telefônica ouviu 602 prováveis eleitores de Michigan. Qual a margem de erro a um nível de significância de 95% para os eleitores **a favor** da liberação das pesquisas? (link: Elgin C. College)

$$\begin{aligned} \varepsilon &= z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \\ &= 1,96 \cdot \sqrt{\frac{0,50 \cdot (1 - 0,50)}{602}} \\ &= 0,04 \end{aligned}$$

A margem de erro é de 4 pontos percentuais para *cima ou para baixo* (46%; 54%) na proporção de eleitores em relação à proporção populacional π a favor da liberação das pesquisas, sob um nível de confiança de 95%

Table 7.3: Independent Samples T-Test

cline6-7	t	df	p	95% CI for Cohen		
				Cohen	Lower	Upper
engagement	2.365	38	0.023	0.748	0.101	1.385

Módulo 8

Introdução às estatísticas epidemiológicas

8.1 Tipos de estudos epidemiológicos

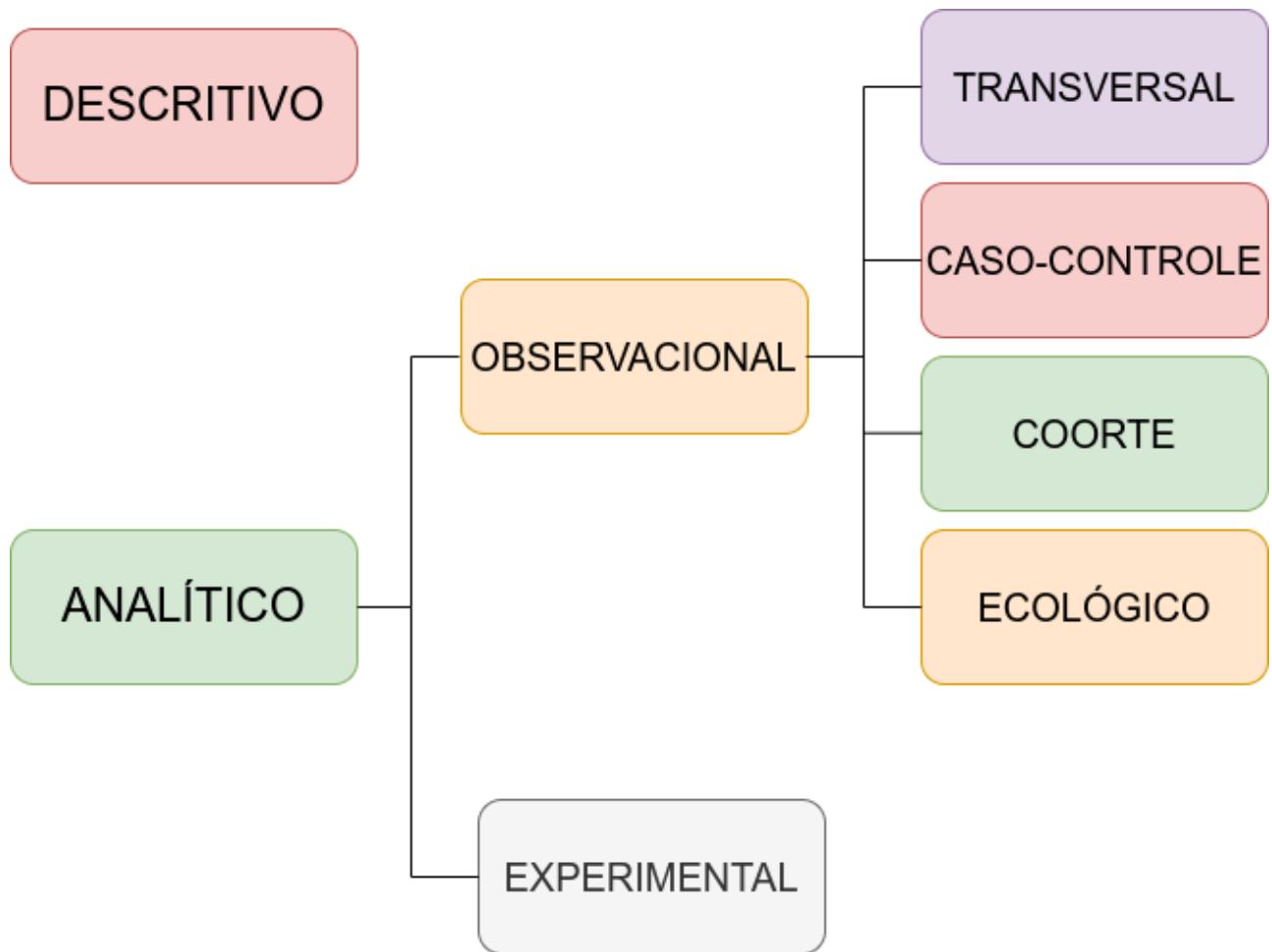


Figure 8.1: Estudos epidemiológicos

Quanto ao objetivo

- Exploratórias (descritivas)
- Explicativas (analíticas)

Quanto ao tempo

- Transversais
- Longitudinais

Quanto à natureza

- Observacional
- Experimental

Estudos descritivos têm como foco principal a caracterização de uma população ou fenômeno em um momento específico, sem tentar estabelecer relações de causa e efeito. Esses estudos fornecem uma visão geral de uma situação ou contexto, frequentemente sendo usados para medir a prevalência de doenças ou comportamentos.

Um estudo descritivo que relata a prevalência de diabetes em uma determinada região, baseado em dados coletados em um hospital.

Estudos analíticos observacionais buscam investigar possíveis associações entre variáveis, sem que o pesquisador manipule as condições em estudo. Estes estudos podem ser transversais, quando os dados são coletados em um único momento, ou longitudinais, quando há um acompanhamento ao longo do tempo.

Um estudo observacional que investiga a relação entre o consumo de alimentos ricos em gordura e doenças cardíacas, acompanhando os participantes ao longo do tempo, sem alterar seus comportamentos.

Estudos analíticos experimentais envolvem a manipulação ativa de variáveis pelo pesquisador, em um ambiente controlado, para avaliar seus efeitos sobre outras variáveis. Esses estudos, geralmente longitudinais, são considerados a melhor abordagem para investigar causalidade, pois permitem o controle de *fatores de confusão* e a comparação direta entre grupos. U

Um ensaio clínico em que um grupo de pacientes recebe um novo medicamento e outro grupo recebe um placebo para avaliar a eficácia do medicamento.

8.2 Estudos transversais

Um estudo transversal é um tipo de investigação observacional em que os dados são coletados em um único ponto no tempo ou durante um período curto. Esse tipo de estudo fornece uma “fotografia” instantânea das condições, comportamentos, ou características de uma população ou grupo em um determinado momento.

Embora útil para fornecer uma visão geral da saúde da população, esse tipo de estudo não permite inferir relações causais entre fatores de risco e desfechos, sendo mais adequado para levantar hipóteses.

Um estudo transversal pode ser usado para determinar a prevalência de hipertensão em uma população: um grupo de pessoas é examinado em um momento específico e o número de pessoas com hipertensão é registrado. Esse estudo dá uma visão geral da saúde da população naquele momento, mas não pode determinar se a hipertensão é causada por fatores observados no estudo.

8.2.1 Estudos de casos e controles

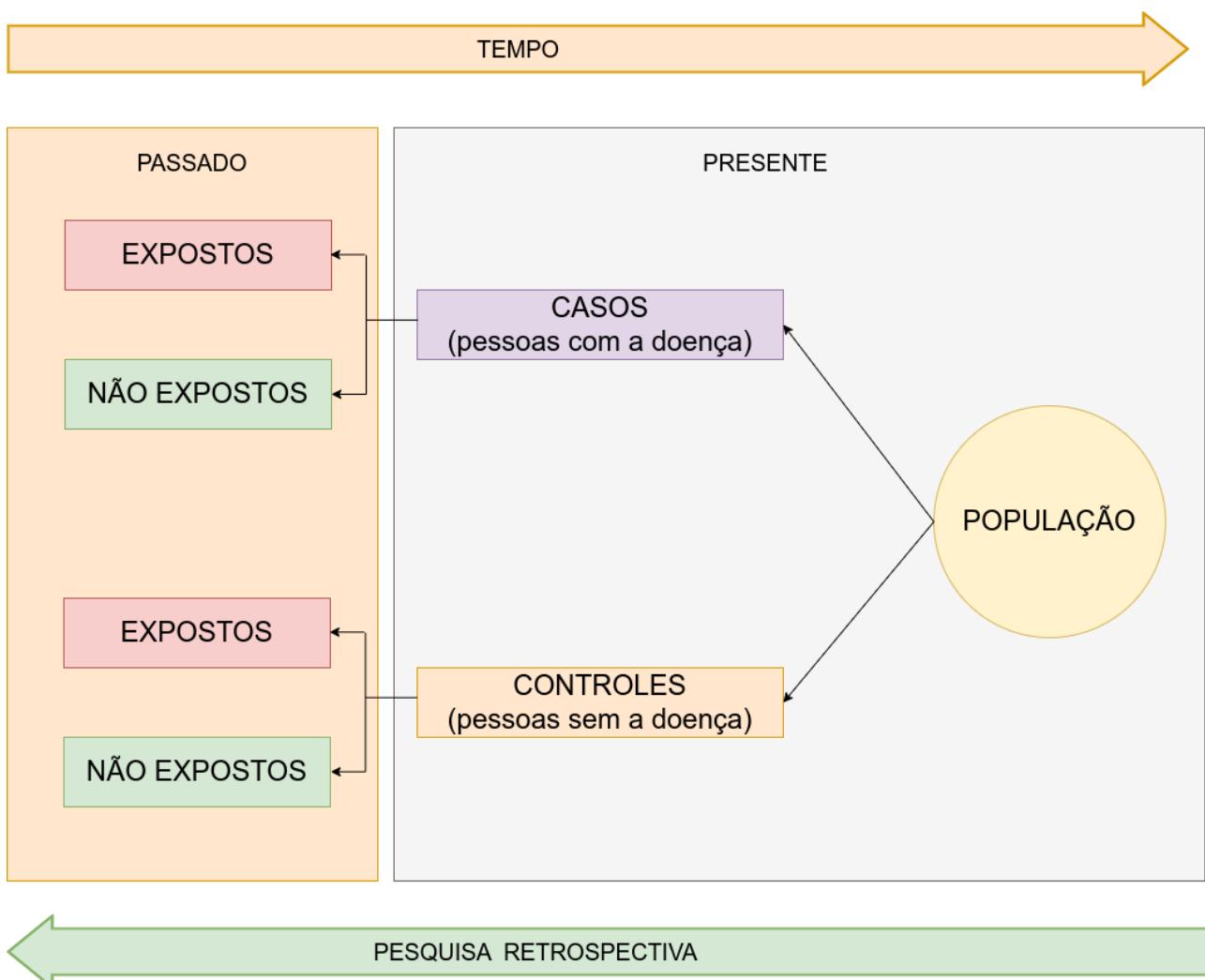


Figure 8.2: Estudo de casos e controles (retrospectivo)

Um estudo de caso-controle, geralmente retrospectivo, começa com a identificação de um grupo de casos (indivíduos com uma doença) e um grupo de controles (indivíduos sem a doença).

Por meio da anamnese o profissional de saúde ajuda o paciente a se recordar (pesquisa retrospectiva) de situações no passado que possam, de algum modo, configurar a exposição ao fator de risco pesquisado. As prevalências de exposição a um determinado fator são então medidas nos dois grupos e comparadas.

Se a prevalência da exposição for maior nos casos do que nos controles, esta exposição pode então ser um fator de risco para a doença. Se a prevalência for menor entre os casos, então esta exposição pode ser um fator de proteção para a doença.

Este método é particularmente útil para investigar doenças raras ou condições com longos períodos de incubação.

Algumas vantagens:

- são estudos relativamente baratos
- podem investigar vários possíveis fatores de risco e são úteis para doenças raras

Algumas desvantagens:

- são muito vulneráveis a vícios de seleção e observação
- não são adequados para investigar exposições raras
- não podem obter estimativas da incidência de doença

8.3 Estudos longitudinais

8.3.1 Estudos de coorte

Estudos de coorte são estudos observacionais em que um grupo de indivíduos expostos e outro grupo de não expostos a uma causa potencial de doença são acompanhados ao longo do tempo.

A incidência da doença é então comparada entre os dois grupos.

Para conduzir este tipo de estudo, é fundamental que uma hipótese clara seja formulada previamente ao início da investigação.

Considerando que esses estudos tendem a ser muito caros, eles geralmente são implementados somente depois que a hipótese foi explorada com outros desenhos de estudo mais econômicos.

Algumas vantagens:

- a exposição é medida antes do início da doença;
- as exposições raras podem ser estudadas selecionando grupos de indivíduos apropriados;
- a incidência da doença pode ser medida nos grupos de expostos e não expostos.

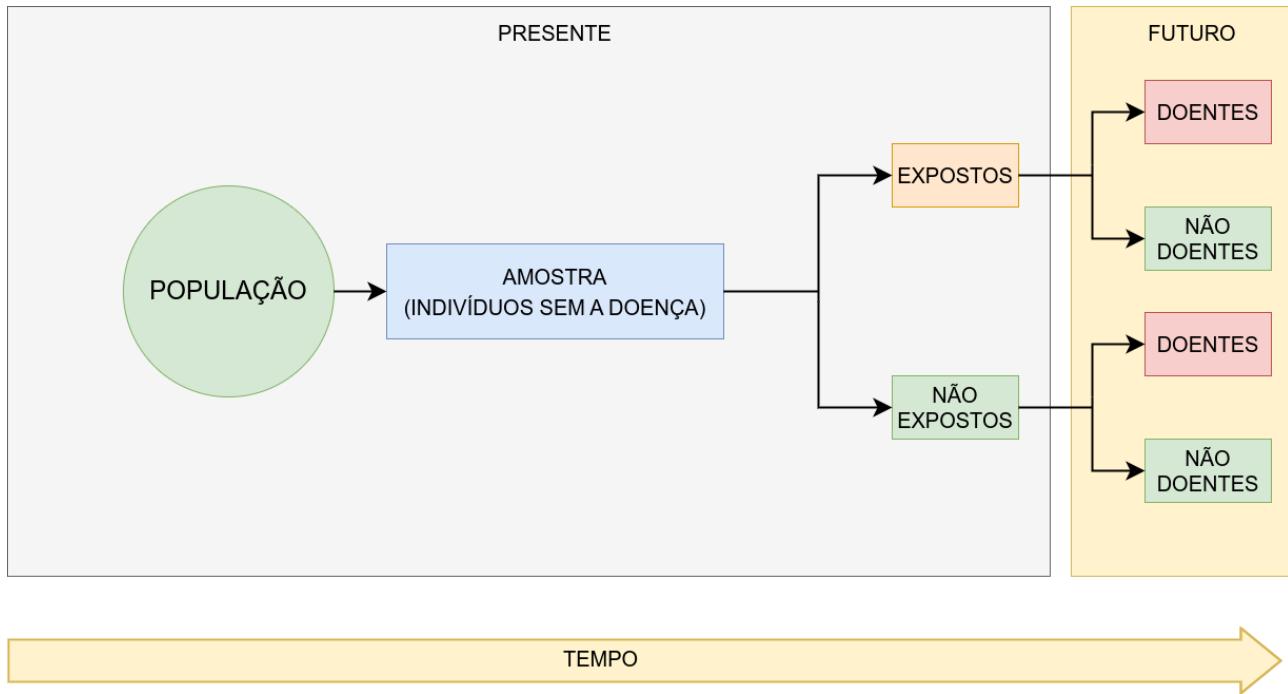


Figure 8.3: Estudos de coorte (prospectivos)

Algumas desvantagens:

- o estudo pode ser extenso e caro, especialmente se o período necessário para observar o efeito for prolongado;
- mudanças na condição de exposição e nos critérios diagnósticos podem ocorrer durante o período do estudo e isto pode afetar a classificação dos indivíduos em expostos e não expostos e em doentes e não doentes;
- a perda de indivíduos durante o seguimento pode introduzir sérios vieses no estudo.

8.3.2 Estudos clínicos aleatorizados

Em estudos epidemiológicos experimentais, também conhecidos como estudos de intervenção ou ensaios clínicos, os indivíduos participantes são alocados a diferentes grupos de acordo com a presença ou não de exposição. No entanto, nesses estudos é o pesquisador quem define quais os indivíduos que receberão a exposição e esta exposição é uma medida preventiva ou terapêutica.

Este tipo de estudo tem como principal vantagem a possibilidade de garantir a validade dos resultados.

Os estudos experimentais são classificados em dois grandes grupos: as intervenções terapêuticas e as intervenções preventivas.

As intervenções terapêuticas incluem pacientes que apresentam uma condição de saúde específica e o objetivo é avaliar a capacidade de determinada intervenção produzir a recuperação, reduzir sintomas, prevenir recrudescimento ou diminuir o risco de uma evolução desfavorável. Para este tipo de estudo, a unidade de amostragem e análise é o indivíduo.

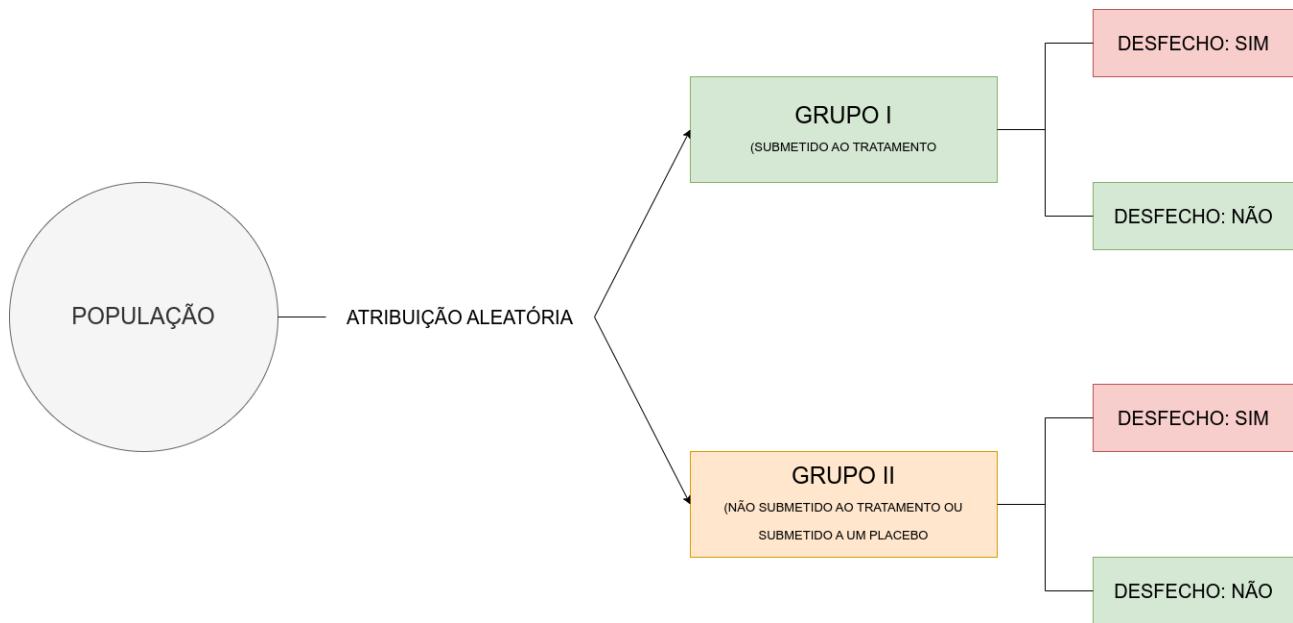


Figure 8.4: Estudos clínicos aleatorizados

As intervenções preventivas envolvem pessoas sadias e o objetivo é avaliar a capacidade de uma intervenção em prevenir a ocorrência de um evento indesejado. As unidades de amostragem nesses casos podem ser tanto os indivíduos como comunidades.

Considerando que os participantes são deliberadamente selecionados pelo pesquisador para receber ou não uma intervenção, os estudos epidemiológicos experimentais envolvem questões éticas importantes e estão sujeitos a regulação legal.

8.4 Terminologia

- Epidemiologia

A epidemiologia é uma ciéncia médica que se concentra na distribuição e nos determinantes (fatores de risco) da frequêcia das doenças na população (desfechos), examinando seus padrões em busca de determinar por que alguns grupos ou certos indivíduos desenvolvem uma doença ao passo que outros não.

- Estudos epidemiológicos

Estudos epidemiológicos são experimentos científicos realizados com o propósito mais comum de se desejar saber se determinadas características pessoais, hábitos ou aspectos do ambiente onde uma pessoa vive estão associados com certa doença, manifestações de uma doença ou outro evento de interesse do pesquisador.

- Desfecho (“sucesso”)

Desfecho é o termo usado para designar a ocorrência do evento de interesse em uma pesquisa. O desfecho pode ser o surgimento de uma doença, de um determinado sintoma, o óbito ou qualquer outro evento relacionado ao processo de saúde-doença. Uma dificuldade inerente está em quantificar a intensidade do desfecho.

- Fator de risco (fator sob estudo)

Fator de risco é a denominação usada em Epidemiologia para designar uma variável que se supõe estar associada ao desfecho. Refere-se portanto a um aspecto de hábitos pessoais ou a uma exposição ambiental, que pode estar associada a uma maior probabilidade de ocorrência de uma doença. Uma dificuldade inerente reside em como quantificar a exposição.

- Risco

Por risco entende-se a “a probabilidade de um membro de uma população definida desenvolver uma dada doença (ou condição) em um período de tempo”. Perceba que nesta definição é possível observar três elementos: base populacional, doença (ou condição) e tempo.

- População em risco

Um fator importante no cálculo das medidas da frequência de uma doença é a estimativa correta do número de pessoas em estudo. Idealmente, esses números devem incluir apenas pessoas potencialmente suscetíveis às doenças (ou condições) em estudo. Por exemplo: homens não devem ser incluídos no cálculo da frequência de câncer do colo do útero e, vice-e-versa para câncer de próstata. Uma vez que os fatores de risco geralmente podem ser modificados, intervir para alterá-los em uma direção favorável pode reduzir probabilidade de ocorrência da doença. O resultado dessas intervenções pode ser estatisticamente verificado em variados tipos de ensaios ou medidas repetidas usando-se os mesmos métodos e definições.

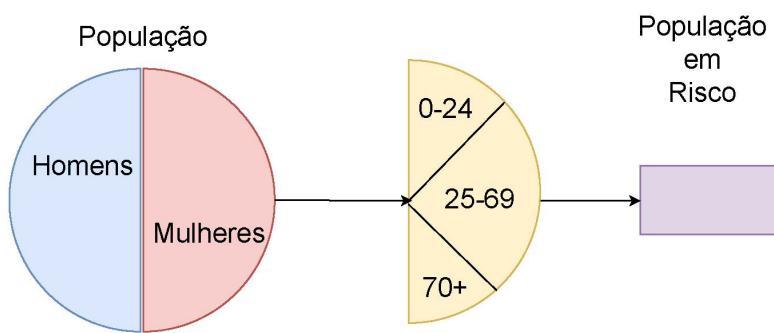


Figure 8.5: Adaptação: Basic Epidemiology: R. Bonita, R. Beaglehole, T Kjellström, 2006 (p. 17)

- Confundimento

A palavra “confundir” vem do latim *confundere* e significa misturar (fundir junto). O confundimento é outra importante questão em estudos epidemiológicos. Em um estudo da associação entre a exposição a uma causa (fator de risco) e a ocorrência de uma doença, o confundimento pode ocorrer quando existe outra exposição na população e está associada tanto à doença quanto ao fator de risco em estudo. O confundimento pode ter uma influência muito importante, podendo até alterar a direção aparente de uma associação. Uma variável que aparece como fator de proteção pode, após o controle de confundimento, ser considerada um fator de risco. Ou então o confundimento pode criar a aparência de uma relação causa-efeito que, na verdade, não existe. O confundimento ocorre quando os efeitos de duas exposições (fatores de risco) **não foram separados** e a análise conclui que o efeito é devido a um fator e não a outro. O confundimento surge porque a distribuição não aleatória de fatores de risco na fonte também ocorre na população de estudo, fornecendo estimativas enganosas de efeito. Nesse sentido, pode parecer um viés, mas na verdade não resulta de um erro sistemático no projeto de pesquisa.

Um exemplo de confundimento pode ser a explicação para a relação demonstrada entre beber café e o risco de doenças cardíacas coronarianas, pois sabe-se que o consumo de café está associado com o uso de tabaco: as pessoas que bebem café são mais propensas a fumar do que as pessoas que não bebem café.

Também é sabido que o tabagismo é uma causa de doença cardíaca coronariana. É, portanto, possível que a relação entre o consumo de café e doenças cardíacas seja meramente reflete a associação causal conhecida do uso de tabaco e doenças cardíacas. Nesta situação, fumar causa confundimento na aparente relação entre o consumo de café e doença cardíaca coronariana porque o tabagismo está correlacionado com beber café e é um fator de risco mesmo para quem não bebe café.

Para se contornar esse tipo de problema deve-se, na etapa de delineamento do experimento, estabelecer os fatores envolvidos e, na realização da pesquisa observar a:

- casualização: as amostras devem ser de tal modo constituídas que variáveis e confundimento nelas existam, potencialmente, em igual proporção (como, por exemplo, fumantes e não fumantes);
 - restrição: se estamos estudando a relação do café com doenças coronarianas, admitir apenas não fumantes.
-
- Vícios de seleção e de observação

Vícios de seleção ocorrem quando os casos e controles são escolhidos de maneira que não representem corretamente a população. Vícios de observação ocorrem quando há erros na forma como a exposição ou os desfechos são medidos.

8.5 Medidas de risco, morte, associação e correlação

- Incidência (I);
- Prevalência (P);
- Incidência cumulativa (risco - IC);
- Fatalidade dos casos (FC);
- Taxa de mortalidade (TM);
- Diferença de risco (risco atribuível - RA);
- Razão de risco (risco relativo - RR);
- Risco atribuível proporcional (fração etiológica - FE);

- *Odds ratio* (razão de chances - OR); e,
- Correlação linear de Pearson.

A morbidade é um dos importantes indicadores de saúde. É um termo genérico usado para designar o conjunto de casos de uma dada doença ou a soma de agravos à saúde que atingem um grupo de indivíduos.

Medir morbidade nem sempre é uma tarefa fácil, pois são muitas as limitações que contribuem para essa dificuldade, como a subnotificação.

Para fazer essas mensurações, utilizam-se principalmente as medidas de incidência e prevalência.

8.5.1 Incidência

Incidência representa a **proporção** de número de **novos casos** de uma determinada doença em um **intervalo de tempo** em uma população exposta ao risco. É, por conseguinte, uma medida dinâmica pois pode sofrer alteração em razão do tempo no qual o estudo foi realizado.

Para um indivíduo pertencente à população exposta, indica a probabilidade de desenvolver a doença (risco).

Observe como calcular a incidência:

$$I = \frac{\text{Número de novos casos de uma doença durante um determinado período de tempo}}{\text{Tamanho da população exposta ao risco nesse determinado período de tempo}} \text{ apresentação usualmente na forma:}$$

Exemplo: para se determinar a incidência de meningite no Maranhão no ano de 2014, será necessário saber o número de casos de meningite que ocorreram naquele período de tempo entre os residentes do Maranhão e o número de habitantes do estado no mesmo período de tempo (todos os possíveis expostos à doença):

$$I = \frac{177 \text{ novos casos notificados de meningite no Maranhão em 2014}}{2.648.532 (\text{população do Maranhão em 2014})} (\times 10^5) I = \frac{6,68}{100.000}$$

Os dados sobre prevalência e incidência tornam-se muito mais úteis se convertidos em taxas!

Como você pode notar, os **casos novos**, ou incidentes, são aqueles que **não existiam no início** do período de observação (tempo analisado), mas que vieram a ocorrer no decorrer desse período.

As taxas de incidência tendem a variar conforme o número de episódios da doença analisada, o número de pessoas que tiveram um episódio de uma doença, tempo para diagnosticá-la e a duração da investigação.

8.5.2 Prevalência

Prevalência representa a proporção de indivíduos de uma população que é acometida por uma determinada doença (ou agravo) em um determinado **momento**. É considerada uma medida **estática**.

Ela engloba tanto os casos existentes, quanto os novos que ocorreram no período.

Indica a probabilidade de ter a doença.

Observe como calcular a prevalência:

$$P = \frac{\text{Número de casos existentes de doença em um determinado momento no tempo}}{\text{Tamanho da população em risco nesse mesmo momento no tempo}} \text{apresentação usualmente na forma: } P(%)$$

Exemplo: se em uma determinada comunidade mensurou-se 89 casos de indivíduos portadores de hipertensão em um determinado momento. Sabendo-se que a população (todos estão potencialmente expostos) dessa comunidade é de 3.500 a prevalência será:

$$P = \frac{89 \text{ casos de hipertensão na comunidade no dia 01/01/2014}}{3.500 \text{ indivíduos como população em risco na comunidade em 01/01/2014}} (\times 10^2) P = \frac{2,54}{100}$$

Os dados sobre prevalência e incidência tornam-se muito mais úteis se convertidos em taxas!

8.5.3 Relação entre prevalência e incidência

A prevalência depende tanto da incidência quanto da duração da doença. Se os casos de incidentes não forem resolvidos e continuarem ao longo do tempo eles se tornarão casos prevalentes. Nesse sentido:

$$P = \text{Incidência} \times \text{Duração média da doença}$$

8.5.4 Quadro comparativo entre medidas de incidência e de prevalência

Table 8.1: Quadro comparativo entre medidas de incidência e de prevalência

	Incidência	Prevalência
Numerador	Número de novos casos de doença durante um determinado período de tempo	Número de casos existentes de doença em um determinado momento no tempo
Denominador	Tamanho da população em risco	Tamanho da população em risco
Foco	Se o evento é um caso novo Tempo de início da doença	Presença ou ausência de uma doença O período de tempo é arbitrário Um “instantâneo” no tempo
Uso	Expressa o risco de adoecer A principal medida de doenças ou condições agudas, mas também usado para doenças crônicas Mais útil para estudos de causalidade	Estima a probabilidade da população estar doente no período de tempo estudado Útil no estudo da carga de doenças crônicas e implicações para os serviços de saúde

8.5.5 Incidência cumulativa - IC (Risco)

Incidência Cumulativa (ou risco) é uma medida da ocorrência de uma doença.

Ao contrário da Incidência, no denominador temos agora o número de pessoas na população exposta **sem a doença** no começo do período do estudo:

$$IC = \frac{\text{Núm. de novos casos de uma doença durante um determ. período de tempo}}{\text{Tam. da pop. em risco (exposta) livre (sem) da doença no começo de um determ. período de tempo}} \text{apresentação usu...}$$

8.5.6 Quadro comparativo entre medidas de risco e prevalência

Table 8.2: Quadro comparativo entre medidas de risco e prevalência

Característica	Risco	Prevalência
O que é medido	Probabilidade da doença	Percentagem da população com a doença
Unidade	adimensional	adimensional
Momento do diagnóstico da doença:	Casos novos (recém diagnosticados)	Existentes
Sinônimos	Incidência cumulativa	-

8.5.7 Fatalidade dos Casos (FC)

Fatalidade dos casos é uma medida da severidade da doença, definida como a proporção de casos com desfecho em óbito pelo total de acometidos (portadores da condição) em um determinado período de tempo.

$$FC(\%) = \frac{\text{Número de mortes de casos diagnosticados da doença durante um determinado período de tempo}}{\text{Número de casos diagnosticados nesse período de tempo}} \text{ apresentação us}$$

8.6 Sobrevida

Uma vez que a TM representa a proporção de pessoas afetadas por uma doença e que faleceram em decorrência dela, a sobrevida S pode ser considerada como seu complemento:

$$S = 1 - TM$$

8.6.1 Taxas de mortalidade (TM)

A principal desvantagem da Taxa bruta de mortalidade é que ela não leva em conta o fato de que a chance de morrer varia de acordo com idade, sexo, etnia e incontáveis outros fatores (sociais, econômicos, ...).

Geralmente não é apropriado usá-la para comparar diferentes períodos de tempo ou áreas geográficas. Por exemplo, padrões de morte em núcleos urbanos recentemente constituídos e formados predominantemente por famílias jovens provavelmente serão muito diferentes das estâncias balneares escolhidas frequentemente por aposentados.

A Taxa bruta de mortalidade para todas as mortes ou uma causa específica de morte é calculado da seguinte forma:

$$TM(\%) = \frac{\text{Número de mortes durante um determinado período de tempo}}{\text{Número de pessoas sob risco de morte nesse período de tempo}} \text{ apresentação usualmente na forma: } TM(\times 10^n)$$

8.6.2 Taxas mais específicas

- taxa de mortalidade infantil;
- taxa de mortalidade maternal;
- taxa de mortalidade entre adultos; ou,
- taxas de mortalidade ajustadas por faixa etária.

Quantificar a ocorrência de doenças ou alterações nos estados de saúde é o primeiro passo de um estudo epidemiológico.

8.7 Medidas de associação em estudos de coorte

Uma tabela é uma forma de representação retangular que permite mostrar clara e resumidamente os dados correspondentes a uma ou mais variáveis, visualizar o comportamento dos dados e facilitar o entendimento das informações. Uma tabela de dupla entrada permite extrair facilmente as proporções **individuais, marginais e associadas** relativas a duas variáveis (tabelas com mais variáveis são possíveis de serem construídas).

Especificamente para estudos epidemiológicos, admita que as variáveis envolvidas se refiram a contagens relacionadas à ocorrência de uma doença em dois grupos de pessoas sob diferentes exposições. O grupo não exposto ao fator de risco é frequentemente usado como referência.

- (a) o grupo de pessoas expostas a um determinado fator de risco;
- (b) o grupo de pessoas não expostas.

Table 8.3: Casos classificados em relação ao desfecho a partir da exposição ao fator de risco

Fator de risco	Desfecho observado (doença)		Total
	Ausente		
Presente	(a)	(b)	(e)
Exposto	(c)	(d)	(f)
Não exposto			
Total	(a) + (c)	(b) + (d)	(e) + (f)

Exemplo: Incidência de baixo peso ao nascer em recém-nascidos de Pelotas (RS) segundo o hábito tabágico da mãe durante a gravidez (1982)

$$I_e = \frac{(a)}{(e)} \times 100 = \frac{275}{2.419} \times 100 = 11,37\%$$

Interpretação: 11,37% das crianças analisadas e que têm mães tabagistas nasceram com baixo peso.

8.7.1 Incidência observada de nascimentos com baixo peso entre mães não expostas ao risco (não fumantes): I_0

Table 8.4: Incidência de baixo peso ao nascer em recém-nascidos de Pelotas, RS, segundo o hábito tabágico da mãe durante a gravidez (1982)

Classificação da mãe	Baixo peso ao nascer		Total
	Sim	Não	
Fumante	275 (a)	2.144 (b)	2.419 (e)
Não fumante	311 (c)	4.496 (d)	4.807 (f)
Total	586	6.640	7.226

$$I_0 : \frac{(c)}{(f)} \times 100 = \frac{311}{4.807} \times 100 = 6,47\%$$

Interpretação: 6,47% das crianças analisadas e que têm mães não tabagistas nasceram com baixo peso.

8.7.2 Prevalência de nascimentos com baixo peso na população estudada

$$\frac{(a) + (c)}{(e) + (f)} \times 100 = \frac{586}{7.226} \times 100 = 8,11\%$$

Interpretação: 8,11% das crianças avaliadas nasceram com baixo peso.

8.7.3 Diferença de risco (Risco atribuível - RA)

A diferença de risco (também chamada de excesso de risco ou risco atribuível) é a diferença nas taxas de ocorrência entre os grupos expostos e não expostos da população. Essa medida quantifica o excesso absoluto de risco associado a uma dada exposição. É uma medida útil do problema de saúde pública causado pela exposição ao fator de risco.

Analisando-se as incidências na Tabela vemos que a diferença de risco de nascimento de bebês com baixo peso entre mães fumantes e não fumantes é:

$$\begin{aligned} RA &= \frac{(a)}{(e)} - \frac{(c)}{(f)} \\ &= \frac{275}{2.419} - \frac{311}{4.807} \\ &= 0,11368334 - 0,064697316 \\ &= 4,9\% \end{aligned}$$

Interpretação: 4,9% do risco de nascer com baixo peso pode ser atribuído ao fato de terem mães tabagistas.

8.7.4 Razão de risco (Risco relativo - RR)

A razão de risco (também chamada de risco relativo) é o quociente entre as taxas de ocorrência entre os grupos expostos e não expostos da população. Pode ser interpretado como a probabilidade de um indivíduo exposto apresentar o desfecho relativa à de um indivíduo não exposto também apresentar.

- razão de risco maior que 1: **fator de risco**;
- razão de risco menor que 1: **fator protetor**.

Analizando-se as incidências na Tabela vemos que a razão de risco de nascimento de bebês com baixo peso entre mães fumantes e não fumantes é de:

$$\begin{aligned}
 RR &= \frac{\frac{(a)}{(e)}}{\frac{(c)}{(f)}} \\
 &= \frac{\frac{275}{2.419}}{\frac{311}{4.807}} \\
 &= \frac{0,11368334}{0,064697316} \\
 &= 1,76
 \end{aligned}$$

Interpretação: As crianças de mães tabagistas têm aproximadamente 1,76 vezes mais risco de nascer com baixo peso em comparação com as de mães não tabagistas.

8.7.5 Risco atribuível proporcional (Fração etiológica - FE)

Quando se acredita que uma determinada exposição é um fator de risco de uma determinada doença, a fração atribuível é a proporção da doença na população específica que seria eliminada se a exposição fosse evitada. As frações etiológicas (frações relacionadas à origem da doença) são úteis para avaliar as prioridades da ação de saúde pública.

O Risco atribuível proporcional (fração etiológica) é, assim, a proporção de todos os casos que podem ser atribuídos diretamente a uma exposição específica. Pode ser determinado pelo quociente da diferença de riscos das incidências pela incidência entre a população exposta.

Esta medida é útil para determinar a importância relativa das exposições para toda a população. É a proporção pela qual a taxa de incidência do desfecho em toda a população seria reduzido se a exposição fosse eliminada.

Observe como calcular o Risco atribuível proporcional (Fração etiológica - FE):

$$FE = \frac{I_e - I_o}{I_e} \times 100$$

- I_e : é a incidência da doença no grupo exposto;
- I_o : é a incidência da doença no grupo não exposto.

Analizando-se as incidências na Tabela vemos que o risco atribuível proporcional de nascimento de bebês com baixo peso entre mães fumantes é de:

$$\begin{aligned} FE &= \frac{\left(\frac{(a)}{(e)} - \frac{(c)}{(f)} \right)}{\frac{(a)}{(e)}} \\ &= \frac{\left(\frac{275}{2.419} - \frac{311}{4.807} \right)}{\frac{275}{2.419}} \\ &= \frac{(0,11368334 - 0,064697316)}{0,11368334} \\ &= 43,09\% \end{aligned}$$

Interpretação: 43,09% dos nascimentos de bebês com baixo peso podem ser atribuídos diretamente ao hábito tabagista das mães.

8.8 Odds ratio (Razão das chances) em estudos de casos e controles

Em estudos de caso-controle os pacientes são incluídos de acordo com a **presença ou não do desfecho**.

Geralmente são definidos um grupo de casos (com o desfecho) e outro de controles (sem o desfecho) e avalia-se uma eventual exposição, **no passado** a potenciais fatores de risco nestes dois grupos.

Devido ao fato de que o **delineamento** deste tipo de estudo baseia-se no **próprio desfecho**, não se pode estimar diretamente a incidência do desfecho de acordo com a **presença ou ausência** da exposição, como é usual em **estudos de coorte**.

Isto se deve ao fato de que a proporção **casos/controles** (ou **desfecho/não-desfecho**) é determinada pelo próprio pesquisador (a proporção não é a mesma observada na população toda com possibilidade de exposição). Assim, a ocorrência de desfechos no grupo total estudado não é regida pela **história natural** da doença e depende de quantos casos e controles o pesquisador selecionou.

Apesar de não se poder estimar diretamente as incidências da doença (desfecho) entre **expostos e não-expostos** em estudos de caso-controle, é possível, entretanto, obter-se uma aproximação da Razão de risco (risco relativo - RR).

Se se o desfecho for suficientemente raro na população (10% ou menos), a Razão de risco (risco relativo - RR) pode ser **estimada aproximadamente** em estudos de caso-controle através da Razão de chances (*odds ratio* - OR) de exposição entre casos e controle:

Table 8.5: Casos e controles classificados em relação à exposição ao fator de risco

Fator	Grupos	
	Casos (com o desfecho)	Controles (sem o desfecho)
Exposto	(a)	(b)
Não exposto	(c)	(d)

Grupo dos casos: a partir das proporções dos elementos desse grupo que foram ou não expostos ao fator (P_e, P_0):

$$P_e = \frac{\text{casos expostos}}{\text{total de casos}} P_0 = \frac{\text{casos não expostos}}{\text{total de casos}}$$

a chance (*odds*) de se observar o desfecho entre os casos é a divisão dessas proporções:

$$Odds_{casos} = \frac{P_e}{P_0} Odds_{casos} = \frac{\text{casos expostos}}{\text{casos não expostos}} Odds_{casos} = \frac{a}{c}$$

Grupo dos controles: a partir das proporções dos elementos desse grupo que foram ou não expostos ao fator (P_e, P_0):

$$P_e = \frac{\text{controles expostos}}{\text{total de controles}} P_0 = \frac{\text{controles não expostos}}{\text{total de controles}}$$

a chance (*odds*) de se observar o desfecho entre os controles é a divisão dessas proporções:

$$Odds_{controles} = \frac{P_e}{P_0} Odds_{controles} = \frac{\text{controles expostos}}{\text{controles não expostos}} Odds_{controles} = \frac{b}{d}$$

A razão das chances (*odds ratio - OR*) de exposição entre casos e controles fica sendo

$$OR = \frac{Odds_{casos}}{Odds_{controles}}$$

- OR (*odds ratio*) maior que 1: **fator de risco**: a exposição ao fator aumenta a chance do desfecho
- OR (*odds ratio*) menor que 1: **fator protetor**: a exposição ao fator reduz a chance do desfecho.

A razão de chances (*odds ratio*) exprime numericamente quantas vezes a exposição a um determinado fator de risco implica na possibilidade do desfecho estudado.

Exemplo: tanto o tabagismo quanto a poluição do ar são causas de câncer de pulmão, mas a fração devido ao fumo é geralmente muito maior do que a devida ao ar poluição. Apenas em comunidades com prevalência de tabagismo muito baixa e severos índices de poluição, esta seria a provável de ser a principal causa de câncer de pulmão. Assim, em muitos países, controle do tabagismo deve ter prioridade nos programas de prevenção do câncer de pulmão.

Table 8.6: Casos e controles classificados em relação à exposição ao fator de risco

Fator	Grupos	
	Casos (com câncer)	Controles (sem câncer)
Exposto ao tabaco	200 (a)	50 (b)
Não exposto ao tabaco	100 (c)	150 (d)

A chance (*odds*) de se observar o desfecho entre os casos :

$$Odds_{casos} = \frac{\text{casos expostos}}{\text{casos não expostos}} Odds_{casos} = \frac{a}{c} Odds_{casos} = \frac{200}{100} = 2$$

Interpretação: entre as pessoas com câncer (casos), a chance de terem sido expostos ao tabaco é 2 vezes maior do que a chance de não terem sido expostos. Ou seja, é muito provável que tenham sido expostas.

A chance (*odds*) de se observar o desfecho entre os controles:

$$Odds_{controles} = \frac{\text{controles expostos}}{\text{controles não expostos}} Odds_{controles} = \frac{b}{d} Odds_{controles} = \frac{50}{150} = 0,33$$

Interpretação: entre as pessoas sem câncer (controles), a chance de terem sido expostos ao tabaco é 1/3 da chance de não terem sido expostas. Ou seja, é pouco provável que tenham sido expostas.

A razão das chances (*odds ratio - OR*) de exposição entre casos e controle:

$$OR = \frac{Odds_{casos}}{Odds_{controles}} OR = \frac{2}{0,33} = 6,06$$

Interpretação: uma *odds ratio* (OR) de aproximadamente 6,06 significa que a chance de uma pessoa exposta ao tabagismo desenvolver câncer de pulmão é cerca de 6 vezes maior do que a de uma pessoa não exposta, o que indica uma forte associação entre tabagismo e câncer de pulmão.

8.9 Correlação linear de Pearson

Em estatística, a expressão **correlação** se refere à relação existente entre variáveis, digamos X e Y .

Essa relação pode assumir diferentes relações funcionais que, basicamente podem ser:

- linear: positiva ou negativa
- não linear: logarítmica, cíclica (periódica), quadrática, cúbica

A correlação existente entre valores observados de uma mesma variável, digamos X em diferentes momentos de tempo $X_{(t_i-1)}, X_{(t_i)}$ é denominada autocorrelação.

É preciso sempre ter em mente que uma **correlação** estatística, por si só, não implica logicamente em **causação**. Para atribuir uma relação de causa-efeito deve-se lançar mão de considerações *a priori* ou teóricas acerca do objeto do estudo.

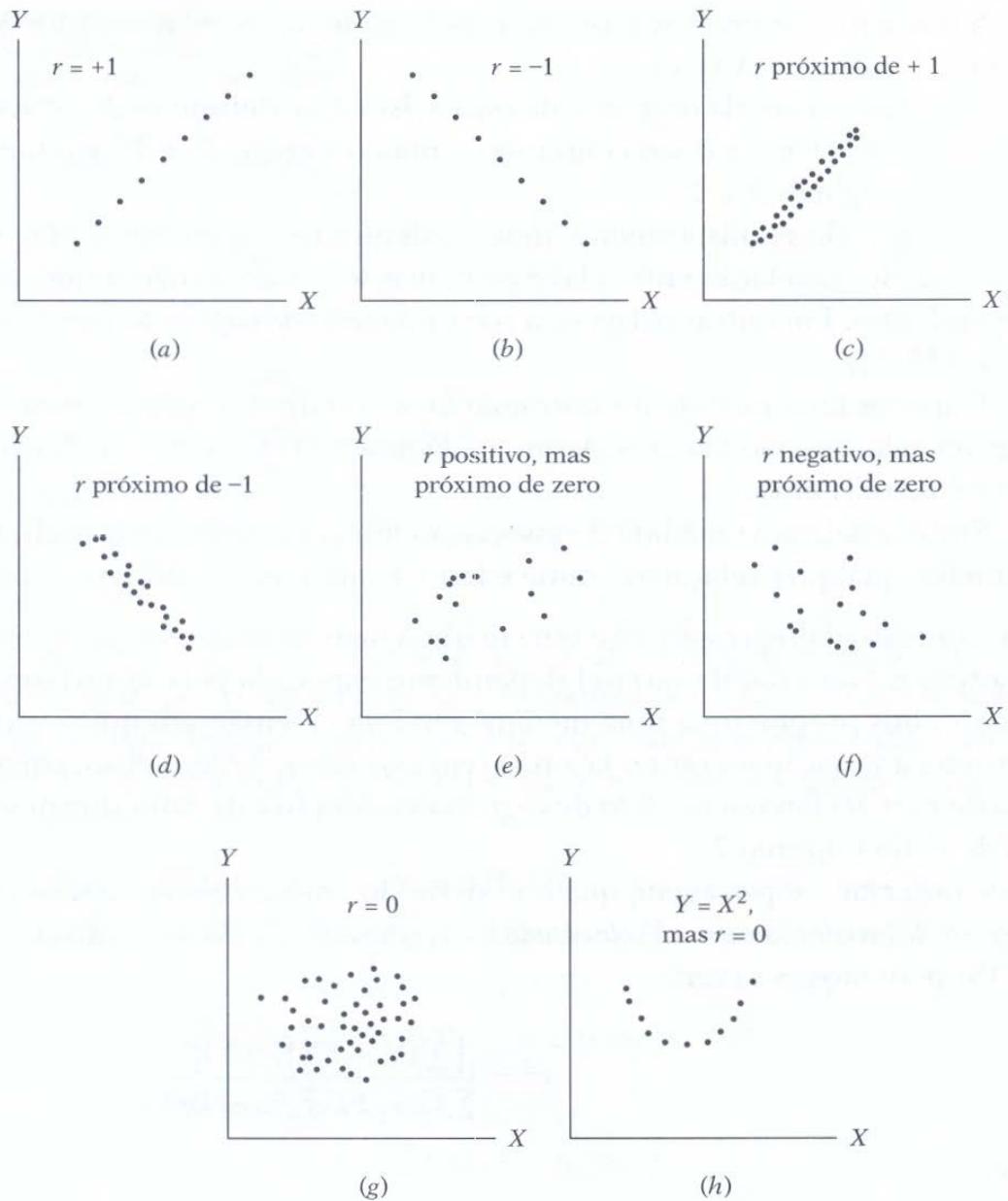


Figure 8.6: Diferentes diagramas de dispersão entre duas variáveis X e Y (Fonte: Introduction to Econometrics. Englewood Cliffs, 1978)

Em (A), (B), (C) e (D) parece-nos que a relação observada entre as variáveis X e Y pode ser expressa por uma função linear (uma reta):

- em (A) e (C) vemos que a variação de ocorre no mesmo sentido: quando o valor da variável X sofre um incremento, também assim ocorre, em algum grau, na variável Y ;
- em (B) e (D) vemos que uma variação inversa: quando o valor da variável X sofre um incremento, a variável Y sofre um decremente em algum grau;
- em (A) e (B) parece-nos que uma função linear exprimiria uma relação entre as variáveis X e Y de modo exato quando comparada a (C) e (D).

Em (G) não se vislumbra um padrão linear no comportamento das variáveis X e Y e em (H) o padrão de comportamento observado entre as variáveis X e Y sugere haver uma boa relação, todavia não **linear**.

O cálculo do **Coeficiente de correlação linear de Pearson (r)** envolve diversos somatórios dos valores das variáveis X , Y , seus quadrados e também de seu produto $X.Y$.

$$r = \frac{\sum_{i=1}^n x_i \cdot y_i - \frac{\sum_{i=1}^n x_i}{n} \cdot \frac{\sum_{i=1}^n y_i}{n}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right) \cdot \left[\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \right]}}$$

Na expressão acima:

- x_i : é o i -ésimo valor observado de X ;
- y_i : é o i -ésimo valor observado de Y ; e,
- n é o número de pares de valores observados.

Outra apresentação de sua fórmula de estimação é:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Na expressão acima:

- x_i : é o i -ésimo valor observado de X ;

- y_i : é o *i*-ésimo valor observado de Y ;
- \bar{x} : é o valor médio das observações x ; e,
- \bar{y} : é o valor médio das observações y .

Simplificadamente podemos também exprimir r na forma abaixo:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

em que:

$$\begin{aligned} S_{xy} &= \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} \\ S_{xx} &= \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \\ S_{yy} &= \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \end{aligned}$$

O coeficiente de correlação de Pearson quantifica a **intensidade** das relações lineares entre x e y e não estabelece *per si* nenhuma relação de causalidade.

É apenas uma medida da associação linear entre duas variáveis e, portanto, não tem sentido usá-lo na quantificação de relações que não o sejam.

O coeficiente de correlação linear de Pearson tem uma **faixa limitada de variação** e é simétrico; isto é, a correlação linear observada entre X e Y é a mesma que a medida entre Y e X .

$$-1 \leq r \leq 1$$

- se $r > 0$ dizemos que há uma relação linear positiva entre as variáveis estudadas: para um incremento na primeira variável observa-se também um incremento na segunda;
- se $r < 0$ a relação linear é negativa: um incremento em uma das variáveis é acompanhado por um decremente na outra; e,

- quando $r = 0$ não há **relação linear** entre as variáveis consideradas.

Exemplo: considere as medidas obtidas de duas variáveis no quadro abaixo.

Table 8.7: Quadro de dados

X	Y
74	139
45	108
48	98
36	76
27	62
16	57

Table 8.8: Quadro auxiliar para cálculo do coeficiente de correlação linear (r)

X	Y	$x_i \cdot y_i$	x_i^2	y_i^2
74	139	10286	5476	19321
45	108	4860	2025	11664
48	98	4704	2304	9604
36	76	2736	1296	5776
27	62	1674	729	3844
16	57	912	256	3249
246	540	25172	12086	53458

Assim, sendo $n = 6$ observações segue-se:

$$\begin{aligned}
 S_{xy} &= \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} \\
 &= 25172 - \frac{246 \cdot 540}{6} \\
 &= 3032 \\
 S_{xx} &= \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \\
 &= 12086 - \frac{246^2}{6} \\
 &= 2000 \\
 S_{yy} &= \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \\
 &= 53458 - \frac{540^2}{6} \\
 &= 4858
 \end{aligned}$$

Portanto:

$$\begin{aligned}
 r &= \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}} \\
 &= \frac{3032}{\sqrt{2000 \cdot 4858}} \\
 &= 0,9727
 \end{aligned}$$

8.10 Intervalos de confiança

As técnicas para obter intervalos de confiança para estimativas amostrais de riscos relativos e *odds ratio* que serão apresentadas estão descritas no livro *Statistics with Confidence* (Douglas Altman _et a_1) e, embora se constituam em aproximações para grandes amostras, são estimativas razoáveis para pequenos estudos.

Através de uma transformação logarítmica, obtém-se uma curva com forma aproximadamente Normal e assim esses intervalos podem ser delimitados a partir da função densidade de probabilidade da distribuição Normal padronizada.

Para o intervalo de confiança da estimativa amostral da diferença de risco (risco atribuível) a proposição se encontra no artigo *Statistical algorithms in Review Manager 5* de Jonathan J. Deeks e Julian P. T. Higgins e está baseada na distribuição da diferença de proporções.

$$\log(IC_{(medida)}) = \log(medida) \pm [z_{(1-\frac{\alpha}{2})} \times EP(\log(medida))]$$

em que:

- $EP(\log(\text{medida}))$ é o erro padrão do logaritmo da medida e os valores mínimo e máximo do intervalo de confiança serão dados por $\exp[\log((IC_{(\text{medida})})]$;
- α é o nível de significância tolerado e, por conseguinte, $(1 - \alpha)$ o nível de confiança pretendido; e,
- e os valores de $|z_{(1-\frac{\alpha}{2})}|$ poderão ser obtidos em uma tabela da distribuição Normal padronizada, sendo os mais usuais:

Table 8.9: Valores críticos z_c correspondentes a vários níveis de significância (α)

Níveis de significância (α)	0,10	0,05	0,01	0,005	0,002
Valores críticos de z_c para testes unilaterais	-1,28 ou 1,28	-1,645 ou 1,645	-2,33 ou 2,33	-2,58 ou 2,58	-2,88 ou 2,88
Valores críticos de z_c para testes bilaterais	-1,645 e 1,645	-1,96 e 1,96	-2,58 e 2,58	-2,81 e 2,81	-3,08 e 3,08

8.10.1 Razão de risco (Risco relativo - RR)

Considere a estrutura dos dados presentes na Tabela para a estimação dos erros padrão a seguir.

$$EP(\log(RR)) = \sqrt{\left[\frac{1}{(a)} - \frac{1}{(a) + (b)} \right] + \left[\frac{1}{(c)} - \frac{1}{(c) + (d)} \right]}$$

O erro padrão do Risco Relativo - RR para os dados da Tabela poderá ser assim estimado:

$$EP(\log(RR)) = \sqrt{\left[\frac{1}{(a)} - \frac{1}{(a) + (b)} \right] + \left[\frac{1}{(c)} - \frac{1}{(c) + (d)} \right]}$$

$$EP(\log(RR)) = \sqrt{\left[\frac{1}{(275)} - \frac{1}{2.419} \right] + \left[\frac{1}{311} - \frac{1}{4.807} \right]}$$

$$EP(\log(RR)) = \sqrt{0,006230374}$$

$$EP(\log(RR)) = 0,078932718$$

Para um nível de confiança de 95% (nível de significância de 0,05%) extraímos o valor crítico de $z_{(1-\frac{\alpha}{2})}$ da Tabela ($z_c = |1,96|$).

A partir do Risco relativo previamente calculado (1,76), um intervalo com nível de confiança de $(1 - \alpha = 95\%)$ fica assim delimitado:

$$\begin{aligned}\log(IC_{(RR)}) &= \log(RR) \pm [z_{(1-\frac{\alpha}{2})} \times EP(\log(RR))] \\ \log(IC_{(RR)}) &= \log(1,76) \pm (1,96 \times 0,078932718) \\ \log(IC_{(RR)}) &= 0,565313809 \pm 0,154708127 \\ \text{Limite superior } IC_{(RR)} &= \exp(0,7147081) \\ &= 2,04359 \\ \text{Limite inferior } IC_{(RR)} &= \exp(0,4052919) \\ &= 1,49974\end{aligned}$$

Assim, o intervalo com nível de confiança $(1 - \alpha)$ estabelecido em 95% para a estimativa amostra do Risco relativo (RR) calculada em 1,76 é:

$$IC_{RR(1-\alpha=0,95)} = [1,49974; 2,04359]$$

8.10.2 Razão de chances (*odds ratio* - OR)

Considere a estrutura dos dados presentes na Tabela para a estimação dos erros padrão a seguir.

$$EP(\log(OR)) = \sqrt{\frac{1}{(a)} + \frac{1}{(b)} + \frac{1}{(c)} + \frac{1}{(d)}}$$

O erro padrão da Razão das chances (*odds ratio* - OR) para os dados da Tabela poderá ser assim estimado:

$$\begin{aligned}EP(\log(OR)) &= \sqrt{\frac{1}{(a)} + \frac{1}{(b)} + \frac{1}{(c)} + \frac{1}{(d)}} \\ EP(\log(OR)) &= \sqrt{\frac{1}{275} + \frac{1}{2.144} + \frac{1}{311} + \frac{1}{4.496}} \\ EP(\log(OR)) &= \sqrt{0,007540636} \\ EP(\log(OR)) &= 0,08683683\end{aligned}$$

Para um nível de confiança de 95% (nível de significância de 0,05%) extraímos o valor de $z_{(1-\frac{\alpha}{2})}$ da Tabela ($z_c = |1,96|$).

A partir da Razão das chances previamente calculada (1,85), um intervalo com nível de confiança de $(1 - \alpha = 95\%)$ fica assim delimitado:

$$\begin{aligned}\log(IC_{(OR)}) &= \log(OR) \pm [z_{(1-\frac{\alpha}{2})} \times EP(\log(OR))] \\ \log(IC_{(OR)}) &= \log(1,85) \pm (1,96 \times 0,08683683) \\ \log(IC_{(OR)}) &= 0,6151856 \pm 0,1702002 \\ \text{Limite superior } IC_{(OR)} &= \exp(0,7853858) \\ &= 2,193253 \\ \text{Limite inferior } IC_{(OR)} &= \exp(0,4449854) \\ &= 1,560467\end{aligned}$$

Assim, o intervalo com nível de confiança $(1 - \alpha)$ estabelecido em 95% para a estimativa amostra da Razão de chances (OR) calculada em 1,85 é:

$$IC_{OR(1-\alpha=0,95)} = [1,560467; 2,193253]$$

8.10.3 Diferença de risco (Risco atribuível - RA)

Considere a estrutura dos dados presentes na Tabela para a estimação dos erros padrão a seguir.

$$EP(RA) = \sqrt{\left[\frac{a \times b}{(a+b)^3} \right] + \left[\frac{c \times d}{(c+d)^3} \right]}$$

$$IC_{(RA)} = RA \pm [z_{(1-\frac{\alpha}{2})} \times EP(RA)]$$

O erro padrão da Diferença de Risco - RA para os dados da Tabela poderá ser assim estimado:

$$EP(RA) = \sqrt{\left[\frac{a \times b}{(a+b)^3} \right] + \left[\frac{c \times d}{(c+d)^3} \right]}$$

$$EP(RA) = \sqrt{\left[\frac{275 \times 2144}{(275+2.144)^3} \right] + \left[\frac{311 \times 4.496}{(311+4.496)^3} \right]}$$

$$EP(RA) = 0,007364887$$

Para um nível de confiança de 95% (nível de significância de 0,05%) extraímos o valor de $z_{(1-\frac{\alpha}{2})}$ da Tabela ($z_c = |1,96|$).

A partir da Diferença de risco previamente calculada (0,049), um intervalo com nível de confiança de $(1 - \alpha = 95\%)$ fica assim delimitado:

$$IC_{(RA)} = RA \pm [z_{(1-\frac{\alpha}{2})} \times EP(RA))]$$

$$IC_{(RA)} = 0,049 \pm [1,96 \times 0,007364887]$$

Limite superior = 0,06343518
Limite inferior = 0,03456482

Assim, o intervalo com nível de confiança $(1 - \alpha)$ estabelecido em 95% para a estimativa amostras da Diferença de risco (RA) calculada em 4,9% é:

$$IC_{RA(1-\alpha=0,95)} = [3,46\%; 6,34\%]$$

Módulo 9

Introdução à distribuição das médias e diferenças entre médias amostrais e seus intervalos de confiança

A finalidade de uma amostra é obter uma estimativa do valor de um ou mais parâmetros de uma população.

Observa-se que os valores amostrais repetidamente extraídos de modo aleatório de uma mesma população variam de uma para outra amostra e também em relação ao verdadeiro parâmetro dessa população; todavia, demonstra-se que essa variabilidade pode ser descrita por meio de distribuições de probabilidade.

Distribuições de probabilidade quando usadas para esse propósito são denominadas de distribuições amostrais e permitem responder para cada amostra o quanto próxima está a estatística amostral do verdadeiro parâmetro populacional. Essa resposta depende fundamentalmente de três fatores:

- a estatística que está sendo utilizada: diferentes estatísticas requerem diferentes distribuições de probabilidade para modelar sua variabilidade;
- o tamanho da amostra que implica de modo inverso na variabilidade entre as amostras;
- a variabilidade existente na própria população sob estudo e amostragem.

9.1 Distribuições amostrais

Parâmetro é toda medida numérica descritiva de uma população. Quando essas medidas são calculadas sobre amostras extraídas de uma população passam a ser denominadas como estatísticas da população de origem. A média, a mediana, a variância, a proporção amostrais, assim como outras estatísticas amostrais, são exemplos de variáveis aleatórias (v.a.) uma vez que seus valores sofrem variação a cada amostra extraída.

Considere uma população com N elementos da qual se deseja extrair todas as possíveis amostras de tamanho n . Para cada amostra extraída pode-se calcular uma mesma medida descritiva como, por exemplo, a média (ou a

variância, proporção ...). O conjunto dos valores resultantes nos permite analisar como as estimativas amostrais se distribuem em comparação ao parâmetro que estão a estimar.

Essas distribuições são denominadas *distribuições amostrais*. O estudo das *distribuições amostrais* é um elemento fundamental na *inferência estatística* posto possibilitar o estabelecimento de *intervalos de confiança* relacionados ao valor de um *parâmetro* que se deseja inferir, a partir de uma estatística proveniente de uma única amostra.

O processo de extração de amostras pode ser *com* ou *sem* reposição. A extração *com* reposição assegura a independência entre os eventos e, eventos independentes são mais facilmente analisados.

O quantidade possível de amostras de tamanho n extraídas de uma população de tamanho N é dado por :

- com reposição: N^n ; e,
- sem reposição: $C_{(N,n)}$

Mais adiante veremos que processos de extração de amostras de tamanho n , *sem* reposição de populações finitas com parâmetros μ (média) e σ^2 (variância) a esperança da v.a. de sua média amostral ainda é dada por:

$$E(\bar{X}) = \mu$$

mas sua variância deve ser corrigida de:

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

para:

$$Var(\bar{X}) = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1} \right)$$

em que $(\frac{N-n}{N-1})$ é denominado como fator de correção para populações finitas.

Para ilustrar o conceito de distribuição das médias amostrais considere uma situação onde uma empresa produz lâmpadas e a vida útil média, em horas, dessas lâmpadas segue uma distribuição Normal tal que $VU \sim N(1600, 120)$.

Usando conceitos já explicados em uma unidade anterior podemos determinar o tamanho amostral em função de:

- um erro máximo: $\varepsilon=20$ horas;
- um nível de significância estabelecido: $\alpha=0,05$; e,
- e alguma informação sobre a medida da variabilidade da variável em estudo: $\sigma=120$ horas (no caso, o desvio padrão populacional).

Flutuação dos valores médios obtidos em 100 amostras de tamanho 10
(nível de confiança: 0.95; erro: 20)

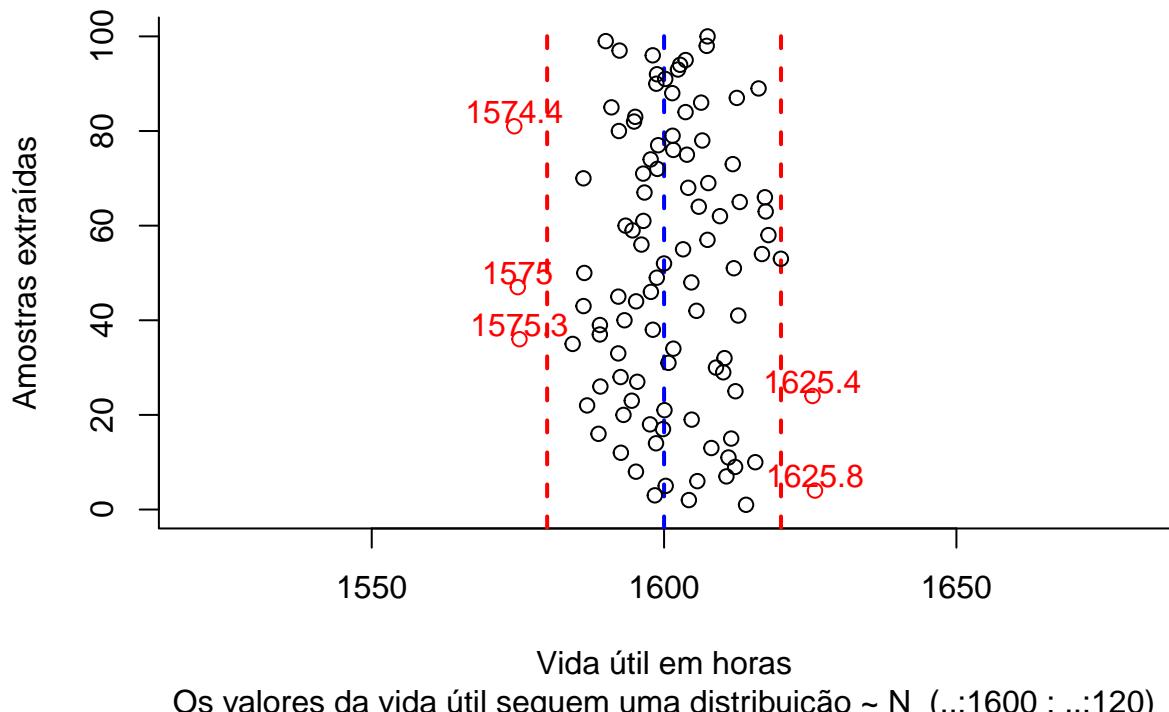


Figure 9.1: Flutuação dos valores médios para diversas amostras extraídas de uma mesma população distribuição $\sim N(\mu; \sigma)$

```
##          mu media      erro   li   ls
## 1    1600  1614  14.021556 1594 1634
```

352 MÓDULO 9. INTRODUÇÃO À DISTRIBUIÇÃO DAS MÉDIAS E DIFERENÇAS ENTRE MÉDIAS AMOSTRAIS E S

```
## 2 1600 1604 4.244245 1584 1625
## 3 1600 1598 -1.578456 1578 1619
## 4 1600 1626 25.809023 1605 1647
## 5 1600 1600 0.276911 1581 1620
## 6 1600 1606 5.690202 1586 1625
## 7 1600 1611 10.666920 1589 1632
## 8 1600 1595 -4.808723 1576 1615
## 9 1600 1612 12.149271 1591 1633
## 10 1600 1616 15.598953 1596 1635
## 11 1600 1611 11.013773 1591 1631
## 12 1600 1593 -7.389086 1574 1611
## 13 1600 1608 8.088597 1587 1629
## 14 1600 1599 -1.375825 1579 1619
## 15 1600 1611 11.481095 1593 1630
## 16 1600 1589 -11.215643 1568 1609
## 17 1600 1600 -0.183236 1582 1618
## 18 1600 1598 -2.405192 1578 1617
## 19 1600 1605 4.723879 1585 1624
## 20 1600 1593 -6.943322 1573 1613
## 21 1600 1600 0.067207 1580 1620
## 22 1600 1587 -13.161497 1568 1605
## 23 1600 1594 -5.524390 1573 1616
## 24 1600 1625 25.382529 1607 1644
## 25 1600 1612 12.207270 1592 1633
## 26 1600 1589 -10.897761 1570 1608
## 27 1600 1595 -4.584704 1574 1617
## 28 1600 1593 -7.435134 1574 1612
## 29 1600 1610 10.118402 1592 1628
## 30 1600 1609 8.880703 1588 1630
## 31 1600 1601 0.748728 1580 1621
## 32 1600 1610 10.335819 1589 1631
## 33 1600 1592 -7.838182 1572 1612
## 34 1600 1602 1.589119 1583 1620
## 35 1600 1584 -15.634651 1566 1602
## 36 1600 1575 -24.731040 1553 1598
## 37 1600 1589 -10.988834 1569 1609
## 38 1600 1598 -1.923153 1578 1618
## 39 1600 1589 -10.974759 1570 1608
## 40 1600 1593 -6.771013 1573 1613
## 41 1600 1613 12.691198 1591 1634
## 42 1600 1606 5.538798 1587 1624
## 43 1600 1586 -13.780478 1566 1606
## 44 1600 1595 -4.788150 1573 1618
## 45 1600 1592 -7.824716 1572 1612
## 46 1600 1598 -2.252233 1577 1618
## 47 1600 1575 -25.018984 1556 1594
## 48 1600 1605 4.666170 1583 1626
## 49 1600 1599 -1.239502 1577 1620
## 50 1600 1586 -13.641067 1567 1605
## 51 1600 1612 11.935098 1590 1634
## 52 1600 1600 -0.008568 1581 1619
## 53 1600 1620 19.979301 1599 1641
## 54 1600 1617 16.741732 1596 1637
## 55 1600 1603 3.250203 1584 1622
## 56 1600 1596 -3.868109 1578 1614
## 57 1600 1607 7.422455 1587 1628
```

```

## 58 1600 1618 17.851255 1595 1640
## 59 1600 1595 -5.388713 1575 1615
## 60 1600 1593 -6.583345 1574 1612
## 61 1600 1596 -3.542529 1574 1619
## 62 1600 1610 9.560425 1589 1630
## 63 1600 1617 17.377624 1599 1636
## 64 1600 1606 5.934449 1587 1625
## 65 1600 1613 12.941336 1595 1631
## 66 1600 1617 17.249112 1598 1636
## 67 1600 1597 -3.349552 1575 1619
## 68 1600 1604 4.128451 1584 1624
## 69 1600 1608 7.566495 1588 1627
## 70 1600 1586 -13.793145 1568 1604
## 71 1600 1596 -3.569164 1578 1615
## 72 1600 1599 -1.152451 1579 1619
## 73 1600 1612 11.748755 1591 1633
## 74 1600 1598 -2.297246 1579 1616
## 75 1600 1604 3.907768 1583 1625
## 76 1600 1602 1.579803 1582 1621
## 77 1600 1599 -1.010979 1578 1620
## 78 1600 1607 6.535436 1586 1627
## 79 1600 1601 1.456434 1581 1622
## 80 1600 1592 -7.722695 1574 1611
## 81 1600 1574 -25.618151 1556 1593
## 82 1600 1595 -5.131870 1575 1614
## 83 1600 1595 -4.936042 1574 1616
## 84 1600 1604 3.674917 1581 1626
## 85 1600 1591 -9.001679 1571 1611
## 86 1600 1606 6.341766 1589 1623
## 87 1600 1612 12.442371 1593 1632
## 88 1600 1601 1.401626 1583 1620
## 89 1600 1616 16.160007 1598 1635
## 90 1600 1599 -1.324410 1579 1618
## 91 1600 1600 0.202577 1579 1621
## 92 1600 1599 -1.206964 1578 1619
## 93 1600 1602 2.378023 1583 1622
## 94 1600 1603 2.742079 1582 1624
## 95 1600 1604 3.688408 1584 1624
## 96 1600 1598 -1.937796 1577 1619
## 97 1600 1592 -7.615215 1573 1611
## 98 1600 1607 7.289609 1588 1627
## 99 1600 1590 -9.977968 1570 1610
## 100 1600 1607 7.437296 1588 1627

```

Observa-se no gráfico acima que algumas das amostras (em vermelho), numa proporção igual ao nível de significância estabelecido quando do dimensionamento (5%), geram médias (amostrais) se afastam do valor médio na população mais que o erro estabelecido (20 h).

9.2 Intervalos de confiança

Um *intervalo de confiança (IC)* pode ser entendido com a faixa de valores delimitada por um mínimo e um máximo, calculados como função direta de um *nível de confiança* e da *variabilidade* e inversa da *tamanho amostral*.

$$\text{estimativa amostral} \pm \text{confiana.} \sqrt{\frac{\text{variabilidade}}{n}}$$

Raramente se dispõe de informação a respeito da variabilidade (σ^2) da população estudada. Assim, a variabilidade populacional será frequentemente incorporado na expressão acima, com ligeiras modificações, na forma de sua estimativa amostral (S^2).

De certo modo, um intervalo de confiança reflete uma estimativa objetiva da (im)precisão e do tamanho da amostra de determinada pesquisa e, assim, podemos considerá-lo como uma medida da qualidade da amostra e da pesquisa.

O *nível de confiança* é designado pela quantidade $(1 - \alpha)$ na qual α é denominado de *nível de significância*, uma medida da probabilidade de erro.

Dependendo do *nível de confiança* que escolhemos os limites superior e inferior do intervalo mudam para uma mesma estimativa amostral. Os intervalos de confiança mais utilizados na literatura são os de 90%, 95%, 99% e menos de 99,9%.

O *intervalo de confiança* de 95% é tradicionalmente o intervalo mais utilizado na literatura e isso está relacionado ao *nível de significância* estatística ($P < 0,05$) geralmente mais aceito.

Quanto menor for a *amplitude* de um intervalo, maior será a *precisão* da estimativa. Todavia, somente estudos com amostras razoavelmente *grandes* resultarão em um intervalo de confiança estreito, indicando simultaneamente com alta precisão e alto grau de confiança a estimativa do parâmetro.

Intervalos de confiança podem ser construídos a quase todas as quantidades estatísticas e suas diferenças (quando se procura estudar se há ou não diferenças entre os parâmetros de duas populações) como, por exemplo:

- médias;
- proporções; e,
- variâncias.

Um *intervalo de confiança* estabelecido sob certa probabilidade **não** deve ser interpretado como sendo a *faixa* de valores, delimitada por um mínimo e máximo, entre os quais o *parâmetro* da população (o qual se estima ou sobre o qual se infere) se insere.

Mas **sim** que, extraíndo-se um grande número de amostras de igual tamanho e da mesma população, e construindo-se para cada uma dessas amostras um intervalo de confiança de um mesmo nível de significância (α), observaremos que uma determinada proporção desses intervalos, chamada de nível de confiança ($1 - \alpha$) **irá, de fato, conter** o *parâmetro* sobre o qual se estima ou sobre o qual se infere. Por conseguinte, uma proporção desses intervalos chamada de nível de significância (α) **não irá** conter o verdadeiro valor do parâmetro populacional.

Assim, $(1 - \alpha)$ traduz o grau de confiança que se tem que um intervalo de confiança, calculado sobre uma estatística advinda de uma particular amostra de tamanho n da variável aleatória X , inclua o verdadeiro valor do parâmetro da população:

```
IC.N = function (N, n, mu, sigma, conf) {
  dados=data.frame()
  plot(0, 0,
    type="n",
    xlim=c(mu-0.4*mu,mu+0.4*mu),
    ylim=c(0,N),
    bty="l",
    xlab="Escala de valores da variável",
    ylab="Intervalos amostrais construídos",
    main=paste0("Intervalos com iguais níveis de confiança fixados em ", 100*conf, "%"
      ↪ `n(",N," amostras de tamanho ",n,")"),
    sub=paste0("Parâmetros da distribuição da população Normal ( \u03bc, \u03c3 ) = "
      ↪ (" ,mu, " , sigma, ")"))
  abline(v=mu, col='red', lwd=2, lty=2)
  #axis(1, at = c(mu-1*mu, mu, mu+1*mu))
  zc = qnorm(1-((1-conf)/2))
  #sigma.xbarra = sigma/sqrt(n)
  for (i in 1:N) {
    x = rnorm(n, mu, sigma)
    media = mean(x)
    erro= media-mu
    sd = sd(x)
    li = media - zc * sd/(sqrt(n))
    ls = media + zc * sd/(sqrt(n))
    temp=cbind(mu, media, erro, li, ls)
    dados=rbind(dados, temp)
    plotx = c(li,ls)
    ploty = c(i,i)
    if (li > mu | ls < mu) lines(plotx,ploty, col="red", lwd=2, lend=0)
    else lines(plotx,ploty, lend=0)
    if (li > mu | ls < mu) points(media, i, col="red", cex=1)+text(y=i+3,x=media,
      ↪ labels=round(media,1), cex=1, col='red')
```

```

    else points(media, i, col="black", cex=1)
}
colnames(dados)=c("mu", "media", "erro", "li", "ls")
return(dados)
}

```

```

N=100
n=64
mu=9.421
sigma=4.1681
conf=0.95
IC.N(N, n, mu, sigma, conf)

```

```

## Warning in title(...): conversion failure on 'Parâmetros da distribuição da
## população Normal ( , ) = (9.421, 4.1681)' in 'mbcsToSbcs': dot substituted
## for <ce>

```

```

## Warning in title(...): conversion failure on 'Parâmetros da distribuição da
## população Normal ( , ) = (9.421, 4.1681)' in 'mbcsToSbcs': dot substituted
## for <bc>

```

```

## Warning in title(...): conversion failure on 'Parâmetros da distribuição da
## população Normal ( , ) = (9.421, 4.1681)' in 'mbcsToSbcs': dot substituted
## for <cf>

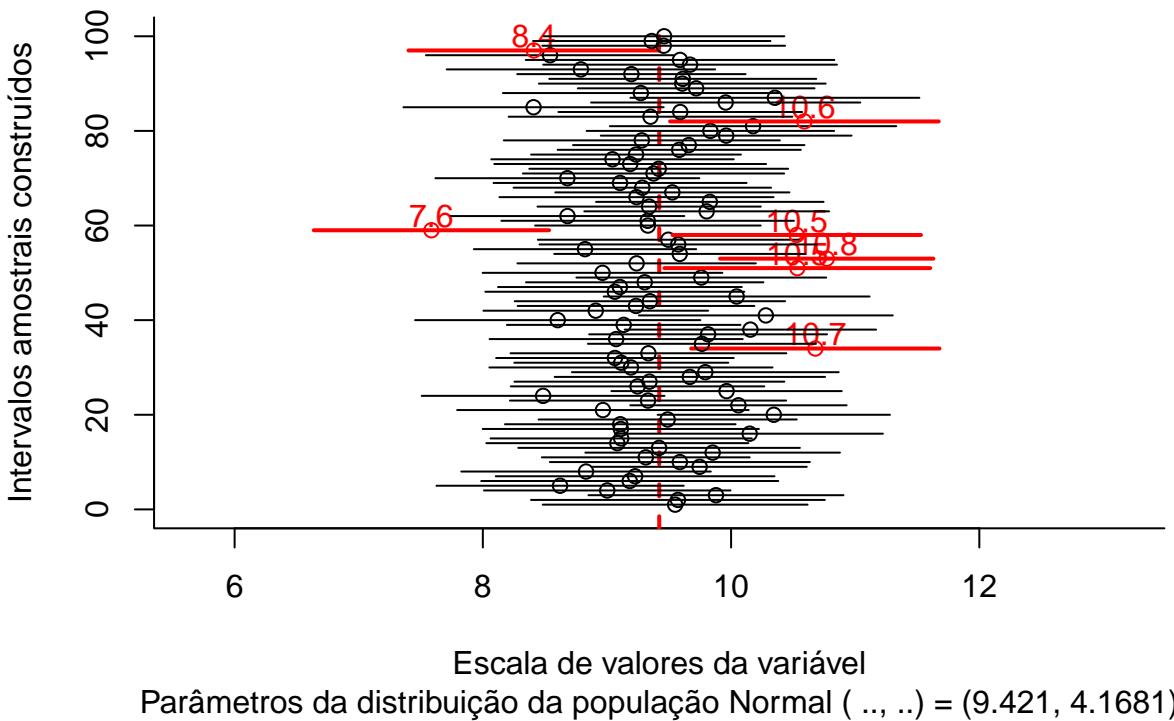
```

```

## Warning in title(...): conversion failure on 'Parâmetros da distribuição da
## população Normal ( , ) = (9.421, 4.1681)' in 'mbcsToSbcs': dot substituted
## for <83>

```

**Intervalos com iguais níveis de confiança fixados em 95%
(100 amostras de tamanho 64)**



```

##      mu   media     erro    li    ls
## 1  9.421  9.549  0.128381 8.482 10.616
## 2  9.421  9.572  0.150613 8.388 10.755
## 3  9.421  9.878  0.457184 8.851 10.906
## 4  9.421  9.002 -0.418846 8.009  9.995
## 5  9.421  8.622 -0.798800 7.626  9.618
## 6  9.421  9.183 -0.237608 7.987 10.380
## 7  9.421  9.227 -0.194332 8.104 10.350
## 8  9.421  8.831 -0.589727 7.826  9.837
## 9  9.421  9.746  0.325376 8.883 10.610
## 10 9.421  9.587  0.166049 8.539 10.635
## 11 9.421  9.313 -0.108343 8.475 10.150
## 12 9.421  9.851  0.430478 8.826 10.877
## 13 9.421  9.420 -0.001146 8.285 10.554
## 14 9.421  9.084 -0.337437 8.028 10.139
## 15 9.421  9.114 -0.306795 8.062 10.167
## 16 9.421 10.150  0.728882 9.077 11.223
## 17 9.421  9.112 -0.309328 7.999 10.224
## 18 9.421  9.107 -0.313727 8.178 10.036
## 19 9.421  9.489  0.067915 8.449 10.529
## 20 9.421 10.344  0.922999 9.408 11.280
## 21 9.421  8.968 -0.453488 7.793 10.142
## 22 9.421 10.059  0.638140 9.187 10.931
## 23 9.421  9.331 -0.090248 8.218 10.444
## 24 9.421  8.486 -0.935462 7.508  9.464
## 25 9.421  9.963  0.542440 9.035 10.892
## 26 9.421  9.246 -0.174557 8.225 10.268
## 27 9.421  9.342 -0.078885 8.255 10.429
## 28 9.421  9.668  0.246720 8.578 10.757

```

```

## 29 9.421 9.792 0.371135 8.717 10.867
## 30 9.421 9.193 -0.227708 8.053 10.334
## 31 9.421 9.115 -0.305765 8.253 9.977
## 32 9.421 9.064 -0.356875 8.107 10.021
## 33 9.421 9.334 -0.087086 8.223 10.445
## 34 9.421 10.679 1.257823 9.678 11.679
## 35 9.421 9.765 0.344268 8.847 10.684
## 36 9.421 9.074 -0.347093 8.054 10.094
## 37 9.421 9.815 0.393679 8.856 10.773
## 38 9.421 10.156 0.734616 9.143 11.168
## 39 9.421 9.135 -0.286419 8.194 10.076
## 40 9.421 8.603 -0.817802 7.454 9.752
## 41 9.421 10.280 0.858960 9.257 11.303
## 42 9.421 8.910 -0.511381 8.005 9.815
## 43 9.421 9.234 -0.187005 8.280 10.188
## 44 9.421 9.346 -0.075411 8.256 10.435
## 45 9.421 10.045 0.623695 8.973 11.116
## 46 9.421 9.064 -0.357466 8.020 10.107
## 47 9.421 9.104 -0.316720 8.122 10.087
## 48 9.421 9.304 -0.116914 8.348 10.261
## 49 9.421 9.760 0.338974 8.754 10.766
## 50 9.421 8.964 -0.456975 7.998 9.930
## 51 9.421 10.534 1.112683 9.463 11.605
## 52 9.421 9.239 -0.182273 8.278 10.199
## 53 9.421 10.771 1.349575 9.911 11.630
## 54 9.421 9.586 0.165248 8.575 10.598
## 55 9.421 8.823 -0.597882 7.927 9.719
## 56 9.421 9.574 0.152798 8.456 10.692
## 57 9.421 9.494 0.073152 8.443 10.545
## 58 9.421 10.529 1.107680 9.528 11.530
## 59 9.421 7.586 -1.835186 6.638 8.534
## 60 9.421 9.330 -0.091428 8.421 10.239
## 61 9.421 9.328 -0.093068 8.149 10.507
## 62 9.421 8.682 -0.739102 7.740 9.623
## 63 9.421 9.803 0.382144 8.817 10.789
## 64 9.421 9.340 -0.081385 8.440 10.239
## 65 9.421 9.829 0.407510 8.912 10.745
## 66 9.421 9.238 -0.183062 8.133 10.342
## 67 9.421 9.527 0.105730 8.584 10.470
## 68 9.421 9.285 -0.136303 8.247 10.322
## 69 9.421 9.105 -0.315666 8.086 10.125
## 70 9.421 8.681 -0.740013 7.618 9.744
## 71 9.421 9.375 -0.046094 8.322 10.427
## 72 9.421 9.417 -0.003932 8.374 10.460
## 73 9.421 9.187 -0.233901 8.092 10.282
## 74 9.421 9.045 -0.376289 8.068 10.022
## 75 9.421 9.234 -0.187042 8.389 10.079
## 76 9.421 9.582 0.161175 8.602 10.562
## 77 9.421 9.659 0.237803 8.726 10.591
## 78 9.421 9.281 -0.140243 8.169 10.393
## 79 9.421 9.960 0.539401 8.950 10.971
## 80 9.421 9.833 0.412139 8.836 10.830
## 81 9.421 10.177 0.756015 9.023 11.331
## 82 9.421 10.591 1.169510 9.508 11.673
## 83 9.421 9.351 -0.070201 8.208 10.494
## 84 9.421 9.590 0.169313 8.609 10.572

```

```
## 85 9.421 8.409 -1.012443 7.360 9.457
## 86 9.421 9.956 0.535418 8.871 11.041
## 87 9.421 10.353 0.931823 9.188 11.517
## 88 9.421 9.273 -0.148017 8.162 10.384
## 89 9.421 9.719 0.297756 8.765 10.672
## 90 9.421 9.607 0.185855 8.451 10.762
## 91 9.421 9.611 0.189749 8.535 10.687
## 92 9.421 9.197 -0.224475 8.277 10.116
## 93 9.421 8.790 -0.630786 7.709 9.872
## 94 9.421 9.669 0.248437 8.486 10.853
## 95 9.421 9.589 0.167722 8.345 10.832
## 96 9.421 8.541 -0.879776 7.541 9.541
## 97 9.421 8.411 -1.010130 7.401 9.421
## 98 9.421 9.457 0.036034 8.479 10.435
## 99 9.421 9.361 -0.060348 8.404 10.317
## 100 9.421 9.460 0.039245 8.491 10.429
```

O gráfico acima expõe os intervalos de confiança: $(1 - \alpha)=95\%$ produzidos para as 100 médias de amostras de tamanho 64 extraídas de uma população com parâmetros $\mu : 9.421$ e $\sigma : 4.1681$.

A proporção de intervalos amostrais que não contém o verdadeiro valor do parâmetro populacional pode ser visualmente inspecionada pelas linhas em vermelho.

Intervalos de confiança bilaterais: intervalos delimitados por dois valores: mínimo e máximo, para a proporção amostral, dentro do qual todos os valores possuem um mesmo nível de confiança de ocorrência.

Intervalos de confiança unilaterais: intervalos delimitados apenas em um de seus lados, nos quais todos os valores possuem um mesmo nível de confiança. Podem ser limitados à direita por um valor máximo ou limitados à esquerda por um valor mínimo.

9.3 Distribuição das médias amostrais e seus intervalos de confiança

Para estudarmos a distribuição das médias amostrais considerem uma população com parâmetros μ (média) e σ^2 (variância).

A distribuição das médias amostrais expressa como se distribuem os valores dessa estatística calculada para todas as possíveis amostras de tamanho n extraídas de uma população cujo valor desse parâmetro é desconhecido.

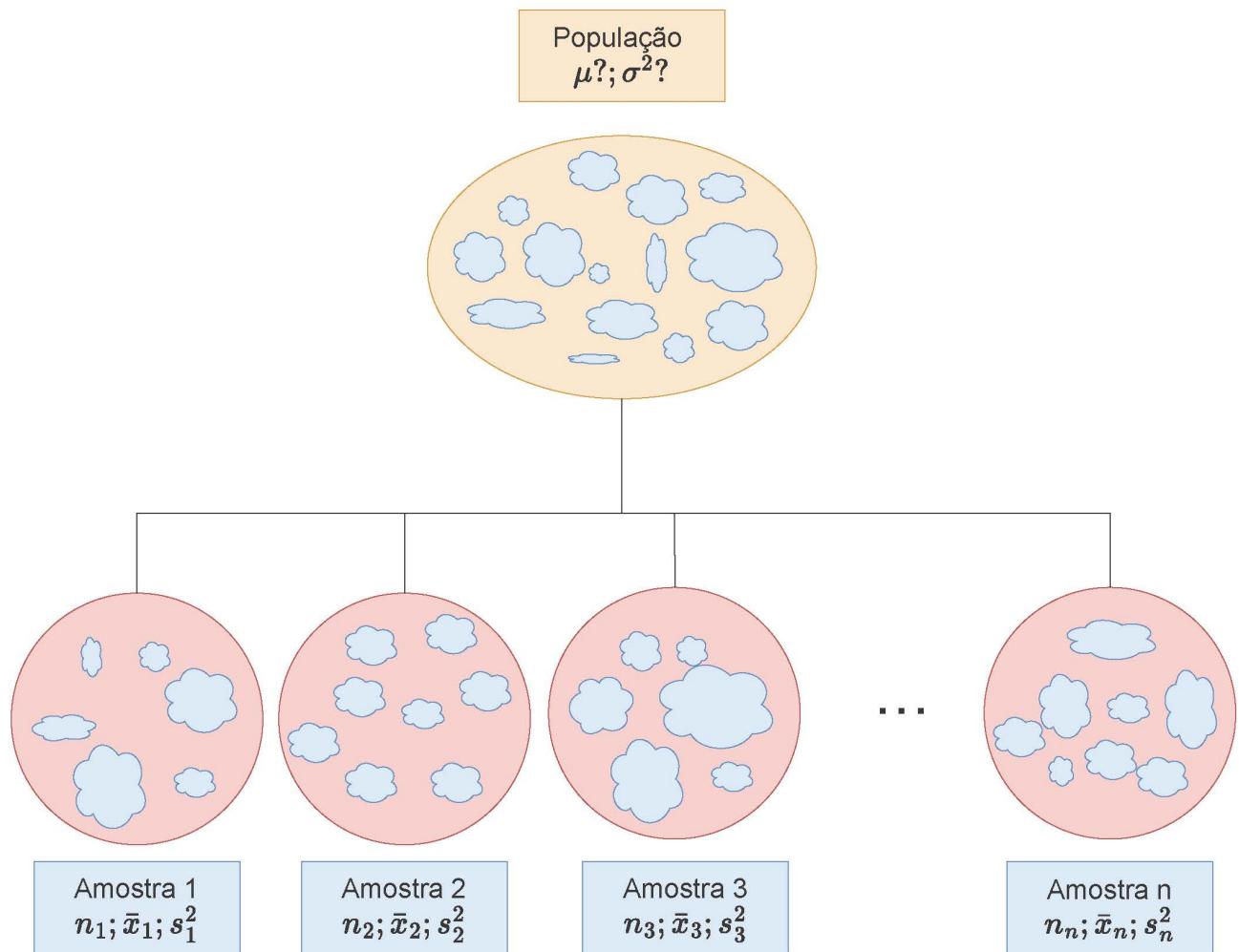


Figure 9.2: Ilustração esquemática de n amostras extraídas de uma mesma população de parâmetros μ e σ , cada uma apresentando as respectivas estatísticas calculadas

A convergência da forma de distribuição e dos parâmetros da distribuição das médias amostrais são elucidadas pelas **Leis (fraca e forte) dos Grandes Números** e pelo **Teorema Central do Limite** (George Pólya, 1920).

De acordo com a teoria, pelo uso de simulações computacionais consegue-se ilustrar que para uma amostra de tamanho n (onde x_1, x_2, \dots, x_n são os valores assumidos das variáveis aleatórias X_1, X_2, \dots, X_n) em amostras extraídas de uma população infinita de tamanho N com média μ e variância σ^2 a distribuição das médias amostrais (v.a. \bar{X}) segue uma distribuição com os média $= \mu$ e variância $= \frac{\sigma^2}{n}$ pois:

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n} \cdot \{E(X_1) + E(X_2) + \dots + E(X_n)\} \\ &= \left(\frac{1}{n}\right) \cdot \{\mu + \mu + \dots + \mu\} = \frac{n \cdot \mu}{n} = \mu \end{aligned}$$

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2} \cdot \{Var(X_1) + Var(X_2) + \dots + Var(X_n)\} \\ &= \left(\frac{1}{n^2}\right) \cdot \{\sigma^2 + \sigma^2 + \dots + \sigma^2\} = n \cdot \frac{\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Equivale afirmar que, **independentemente** da forma de distribuição da população de origem da qual são extraídas as amostras, a distribuição dos valores da variável aleatória \bar{X} tenderá a seguir uma distribuição $\sim N(\mu; \frac{\sigma^2}{n})$ à medida que n , o tamanho da amostra aumenta, como ilustrado nas Figuras 9.3 e 9.5.

O **TCL** garante a aproximação da distribuição de \bar{X} a uma distribuição Normal com média μ e variância $\frac{\sigma^2}{n}$ quando n é grande, independentemente da distribuição da população de origem. Na prática, essa aproximação é usada quando $n \geq 30$.

Portanto, para populações **infinitas** ou amostragem **com reposição**:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Demostração usando amostras extraídas de uma população com distribuição $\sim U(v_{min}; v_{max})$

```
# Definindo os parâmetros e a amostra
min_1=2
max_1=6
NN=5000
pop_1=runif(NN, min=min_1, max=max_1)
df=as.data.frame(pop_1)

# A distribuição da população ilustrada em um histograma
ggplot(df, aes(x=pop_1)) +
  geom_histogram( binwidth=1,color="black", fill="lightblue")+
  scale_y_continuous(name="Frequência") +
  scale_x_continuous(name="Valores")+
  labs(title= paste("Histograma de uma população com Distribuição Uniforme"),
       subtitle = paste("Parâmetros: valor min =",min_1,"; valor max =", max_1))+
```

theme(plot.title = element_text(size = 10, face = "bold"),
 axis.text.x = element_text(angle=0, hjust=1, size=10),
 axis.text.y = element_text(angle=0, hjust=1, size=10),
 axis.title.x = element_text(size = 10),
 axis.title.y = element_text(size = 10))

Histograma de uma população com Distribuição Uniforme

Parâmetros: valor min = 2 ; valor max = 6

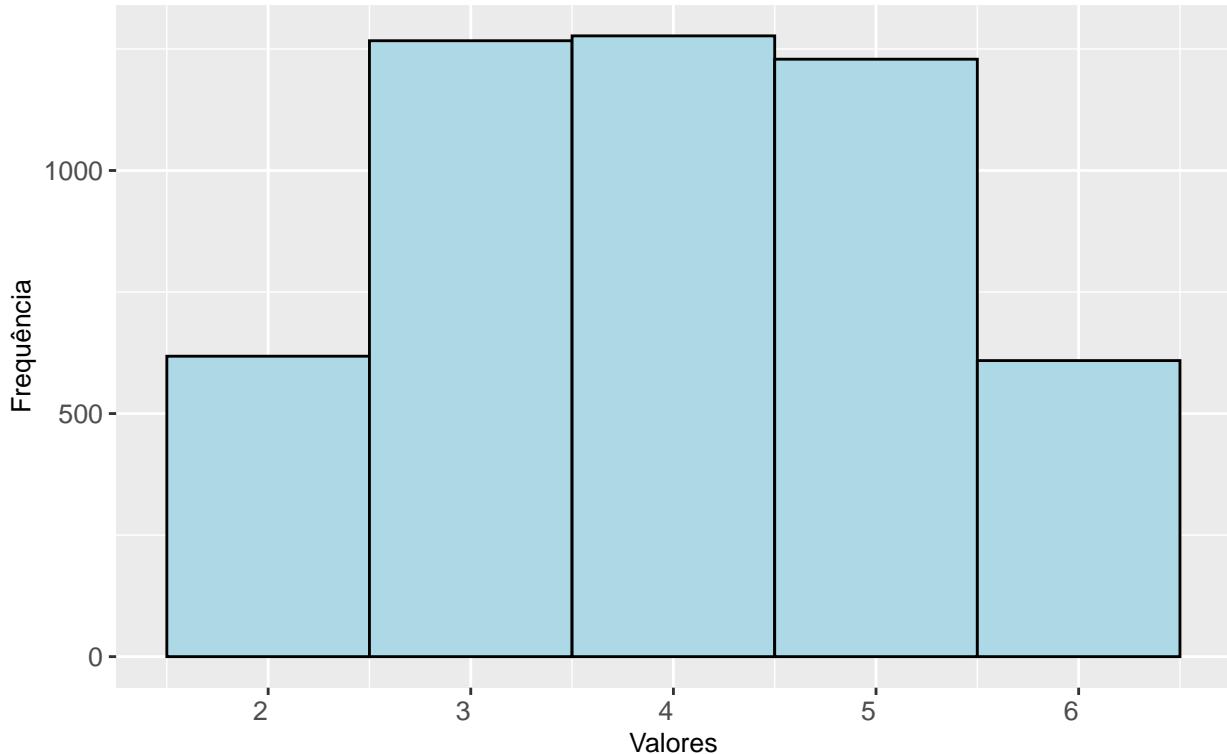


Figure 9.3: Histograma de uma população cuja característica de interesse segue uma Distribuição Uniforme

A Figura 9.3 mostra o histograma de uma amostra de 5000 elementos de uma população com Distribuição Uniforme de parâmetros $v_{min} : 2$ e $v_{max} : 6$.

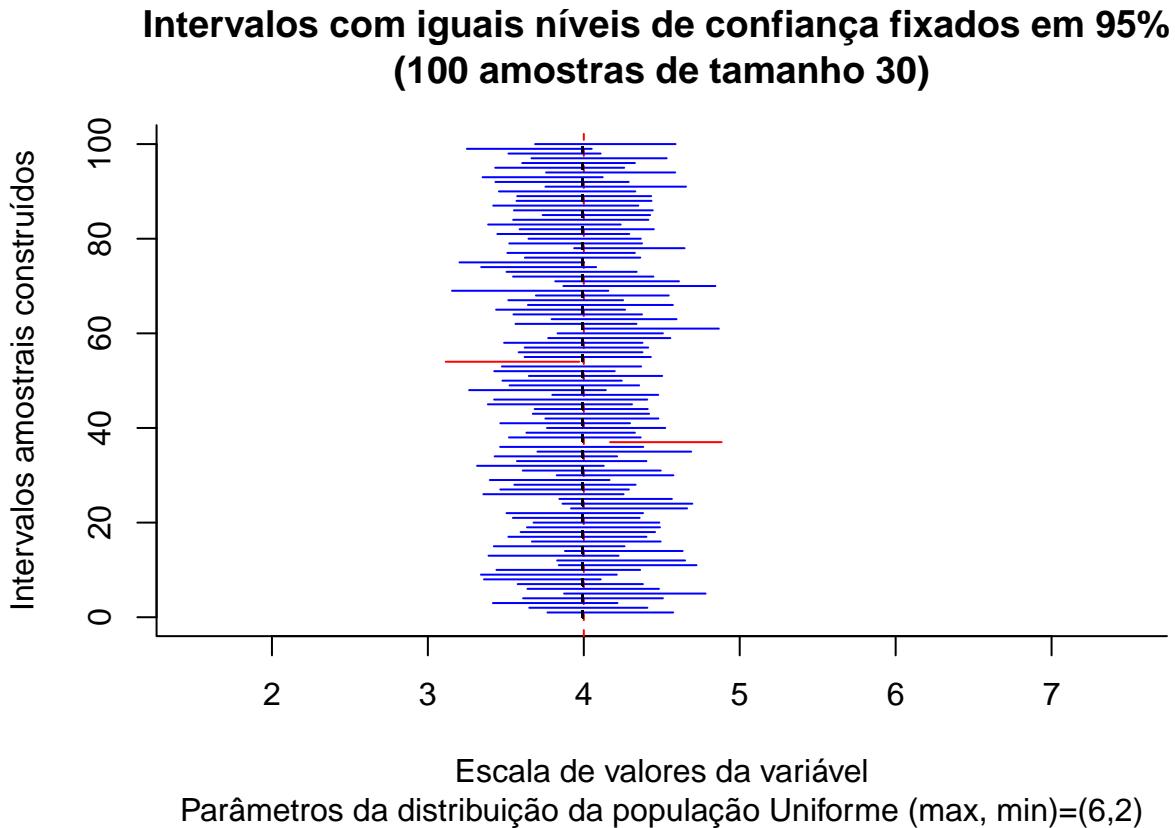


Figure 9.4: Intervalos de confiança construídos para diversas estimativas amostrais de uma população com Distribuição $\sim N(\mu = \frac{max-min}{2}; \sigma^2 = \frac{1}{12}(max-min)^2)$

A Figura 9.4 expõe os intervalos sob nível de confiança de $(1 - \alpha)=95\%$ produzidos para as 100 médias de amostras de tamanho 30 extraídas de uma população Uniforme com parâmetros $v_{max} : 6$ e $v_{min} : 2$ e, conforme assegura o **TCL**, o valor médio das médias amostrais (linha tracejada preta) converge assintoticamente para a média da população de origem (linha tracejada em vermelho) com o incremento do tamanho das amostras.

```
meu_titulo1=paste("Distribuição das médias de", N, "amostras de tamanho n=",n,"\\n população",  
"de origem sob Dist. Unif. (min: ", min_1, "; max: ", max_1, ")")  
meu_titulo2=paste("As médias amostrais ~ N( x=",round(mean(m),2),";sd=",round(sd(m),2),")")  
  
dados=as.data.frame(m)  
ggplot(dados, aes(m)) +  
  geom_histogram(aes(y = stat(density)), bins=10, fill="lightblue", col="black") +  
  geom_area(stat = "function",
```

```

fun = dnorm,
args = list(mean=mean(m), sd=sd(m)),
fill = NA,
colour="red") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores das médias amostrais") +
labs(title=meu_titulo1) +
geom_segment(aes(x = mean(m), y = 0, xend = mean(m), yend = max(dnorm(m))), color="blue",
  lty=2, lwd=0.3) +
annotate(geom="text", x=mean(m), y=max(dnorm(m)),
  label=meu_titulo2, angle=0, vjust=-0.5, hjust=0.5, color="blue", size=6) +
theme(plot.title = element_text(size = 10, face = "bold"),
  axis.text.x = element_text(angle=0, hjust=1, size=10),
  axis.text.y = element_text(angle=0, hjust=1, size=10),
  axis.title.x = element_text(size = 10),
  axis.title.y = element_text(size = 10))

```

**Distribuição das médias de 100 amostras de tamanho n= 30
população de origem sob Dist. Unif. (min: 2 ; max: 6)**

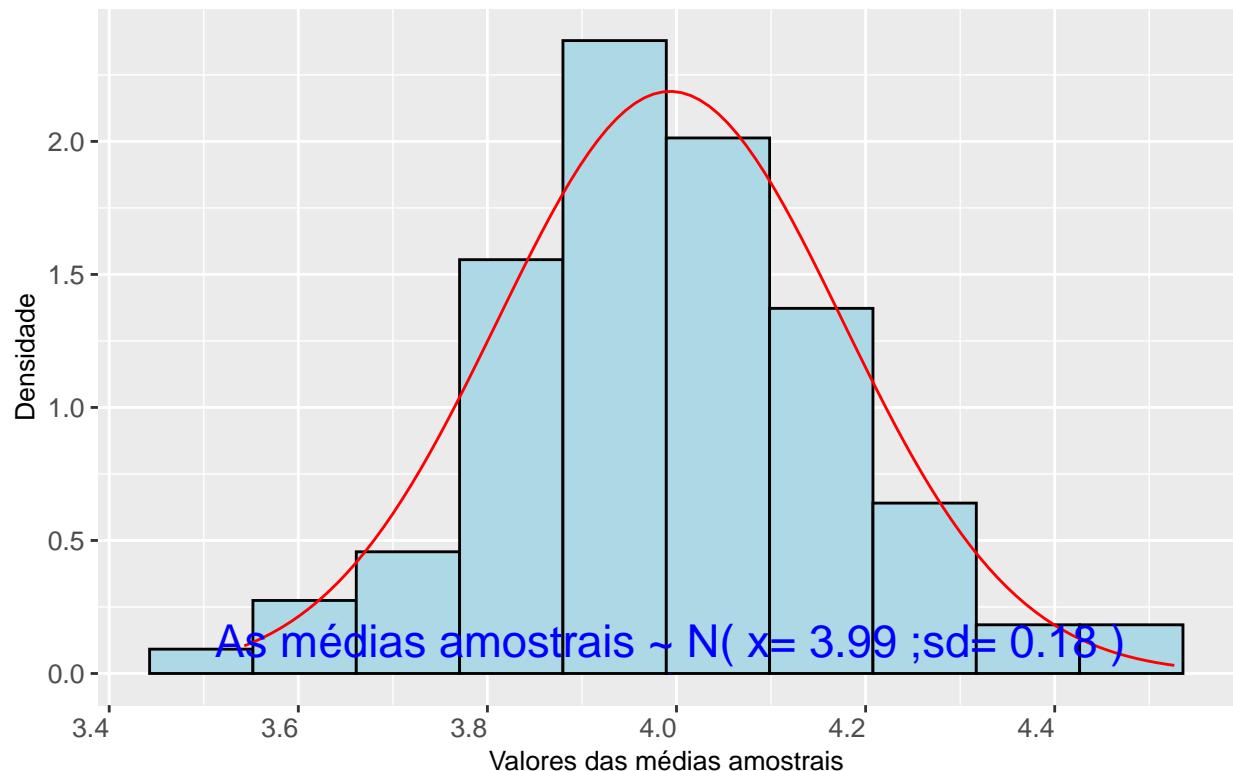


Figure 9.5: Histograma da distribuição das médias de amostras extraídas de uma população com Distribuição Uniforme mostra que as mesmas seguem uma Distribuição $\sim N(\mu = \frac{max-min}{2}; \sigma^2 = \frac{1}{12}(max - min)^2)$

O histograma da Figura 9.5 ilustra que os valores das médias calculadas de 30 amostras extraídas de uma população com distribuição Uniforme $\sim U(v_{min}, v_{max})$ seguem uma distribuição Normal $\sim N(\mu = \frac{v_{max}-v_{min}}{2}; \sigma^2 = \frac{1}{12}(v_{max} - v_{min})^2)$.

Demostração usando amostras extraídas de uma população com distribuição $\sim N(\mu; \sigma)$

```
# Definindo os parâmetros e a amostra
media=80
desvio=4
NN=5000
pop_2=rnorm(n=NN, mean = media, sd = desvio)

df=as.data.frame(pop_2)

# A distribuição da população ilustrada em um histograma
ggplot(df, aes(x=pop_2)) +
  geom_histogram( binwidth=1,color="black", fill="lightblue")+
  scale_y_continuous(name="Frequêcia") +
  scale_x_continuous(name="Valores")+
  labs(title= paste("Histograma de uma população com Distribuição Normal"),
       subtitle = paste("Parâmetros: média =",media,"; desv. padrão =", desvio))+
```

theme(plot.title = element_text(size = 10, face = "bold"),
 axis.text.x = element_text(angle=0, hjust=1, size=10),
 axis.text.y = element_text(angle=0, hjust=1, size=10),
 axis.title.x = element_text(size = 10),
 axis.title.y = element_text(size = 10))

Histograma de uma população com Distribuição Normal

Parâmetros: média = 80 ; desv. padrão = 4

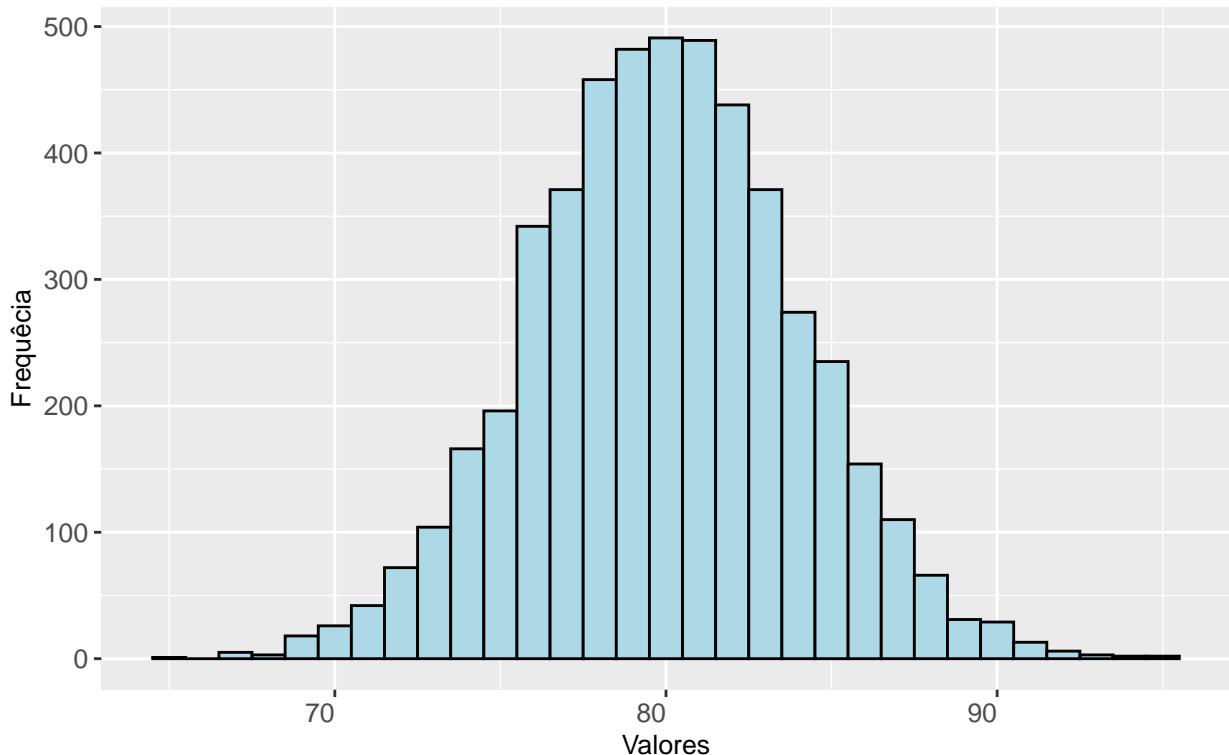


Figure 9.6: Histograma de uma população cuja característica de interesse segue uma Distribuição Normal

A Figura 9.6 mostra o histograma de uma amostra de 5000 elementos de uma população com Distribuição Normal de parâmetros média= 80 e desvio padrão =4.

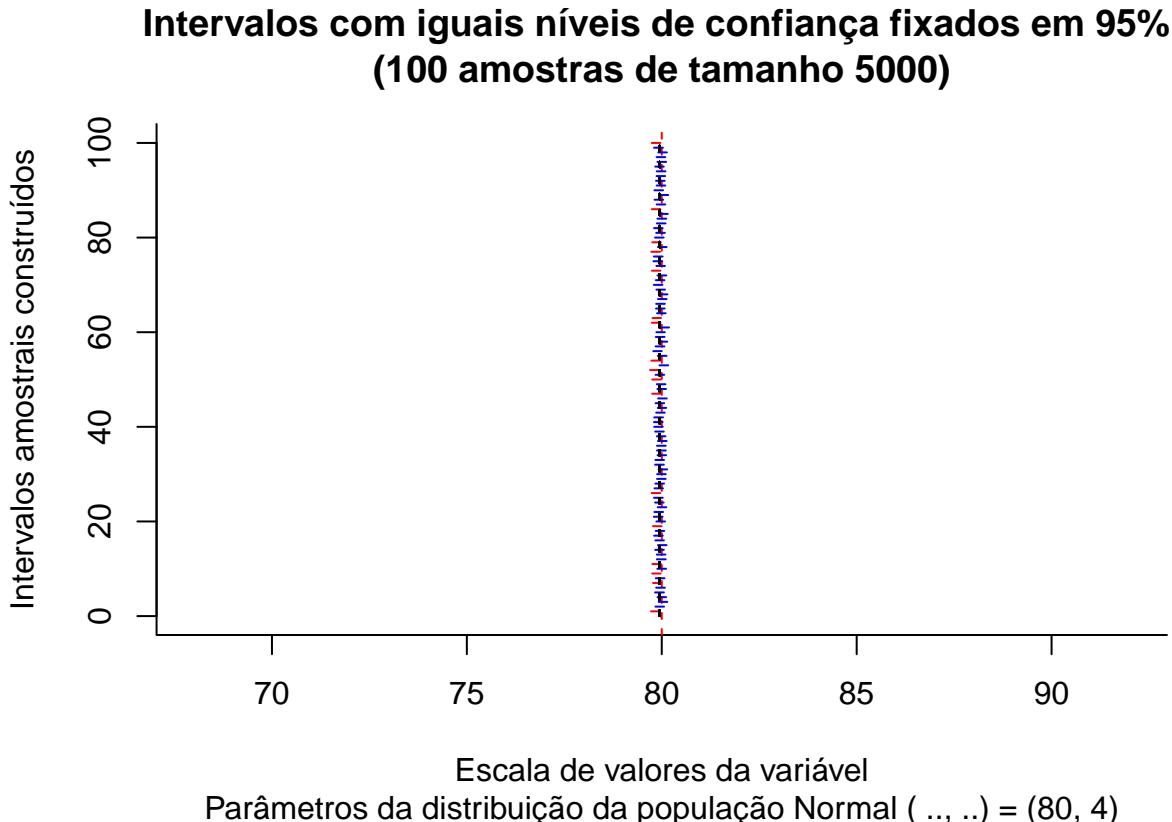


Figure 9.7: Intervalos de confiança construídos para diversas estimativas amostrais de uma população com Distribuição $\sim N(\mu; \sigma)$

A Figura 9.7 expõe os intervalos sob nível de confiança de $(1 - \alpha)=95\%$ produzidos para as 100 médias de amostras de tamanho 5000 extraídas de uma população Uniforme com parâmetros $v_{max} : 6$ e $v_{min} : 2$ e, conforme assegura o **TCL**, o valor médio das médias amostrais (linha tracejada preta) converge assintoticamente para a média da população de origem (linha tracejada em vermelho) com o incremento do tamanho das amostras.

```

meu_titulo1=paste("Distribuição das médias de", N, "amostras de tamanho n=",n,"\\n população"
                   "de origem sob Dist. Normal ( \u03bc: ", media, ", \u03c3: ", desvio, ")")
meu_titulo2=paste("As médias amostrais ~ N(
                   "x\u0304=",round(mean(m),2),";sd=",round(sd(m),2),")")

dados=as.data.frame(m)
ggplot(dados, aes(m)) +
  geom_histogram(aes(y = stat(density)), bins=10, fill="lightblue", col="black") +

```

```

geom_area(stat = "function",
           fun = dnorm,
           args = list(mean=mean(m), sd=sd(m)),
           fill = NA,
           colour="red") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores das médias amostrais") +
labs(title=meu_titulo1) +
geom_segment(aes(x = mean(m), y = 0, xend = mean(m), yend = max(dnorm(m))), color="blue",
             lty=2, lwd=0.3) +
annotate(geom="text", x=mean(m), y=max(dnorm(m)),
         label=meu_titulo2, angle=0, vjust=-0.5, hjust=0.5, color="blue",size=6) +
theme(plot.title = element_text(size = 10, face = "bold"),
      axis.text.x = element_text(angle=0, hjust=1, size=10),
      axis.text.y = element_text(angle=0, hjust=1, size=10),
      axis.title.x = element_text(size = 10),
      axis.title.y = element_text(size = 10))

```

**Distribuição das médias de 100 amostras de tamanho n= 5000
população de origem sob Dist. Normal (... 80 , ..: 4)**

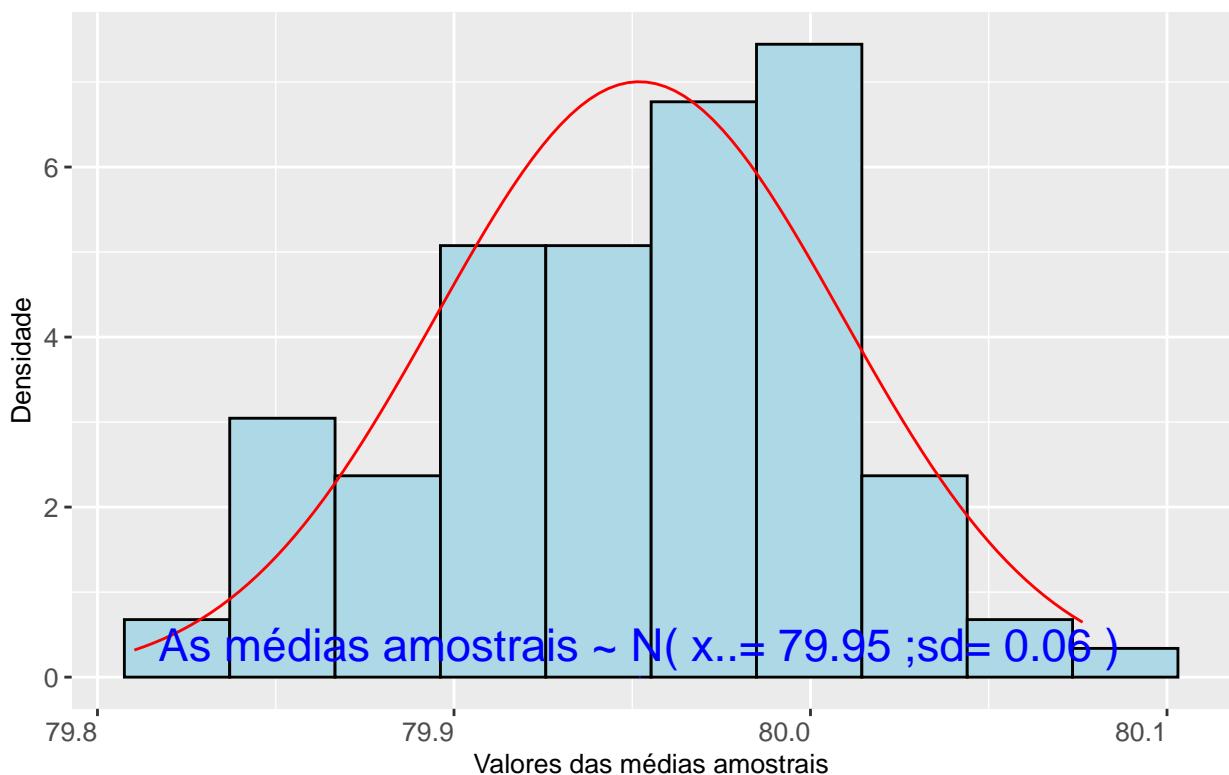


Figure 9.8: Histograma da distribuição das médias de amostras extraídas de uma população Normal mostra que as mesmas seguem uma Distribuição $\sim N(\bar{x} = \mu; s = \frac{\sigma}{\sqrt{n}})$

O histograma da Figura 9.8 ilustra que os valores das médias calculadas de 5000 amostras extraídas de uma população com distribuição Normal $\sim N(\mu, \sigma)$ seguem uma distribuição Normal $\sim N(\mu = \mu; \sigma = \frac{\sigma}{\sqrt{n}})$.

Sendo o erro amostral expresso como: $\varepsilon = \bar{X} - \mu$, o histograma abaixo ilustra que os valores dos erros calculados de 5000 amostras extraídas de uma população com distribuição Normal $\sim N(\mu, \sigma)$ seguem uma distribuição Normal $\sim N(\mu = \mu; \sigma = \frac{\sigma}{\sqrt{n}})$.

```
N=100
n=50
mu=80
sigma=4
conf=0.95
matriz=IC.Na(N, n, mu, sigma, conf)
```

Intervalos com iguais níveis de confiança fixados em 95% (100 amostras de tamanho 50)

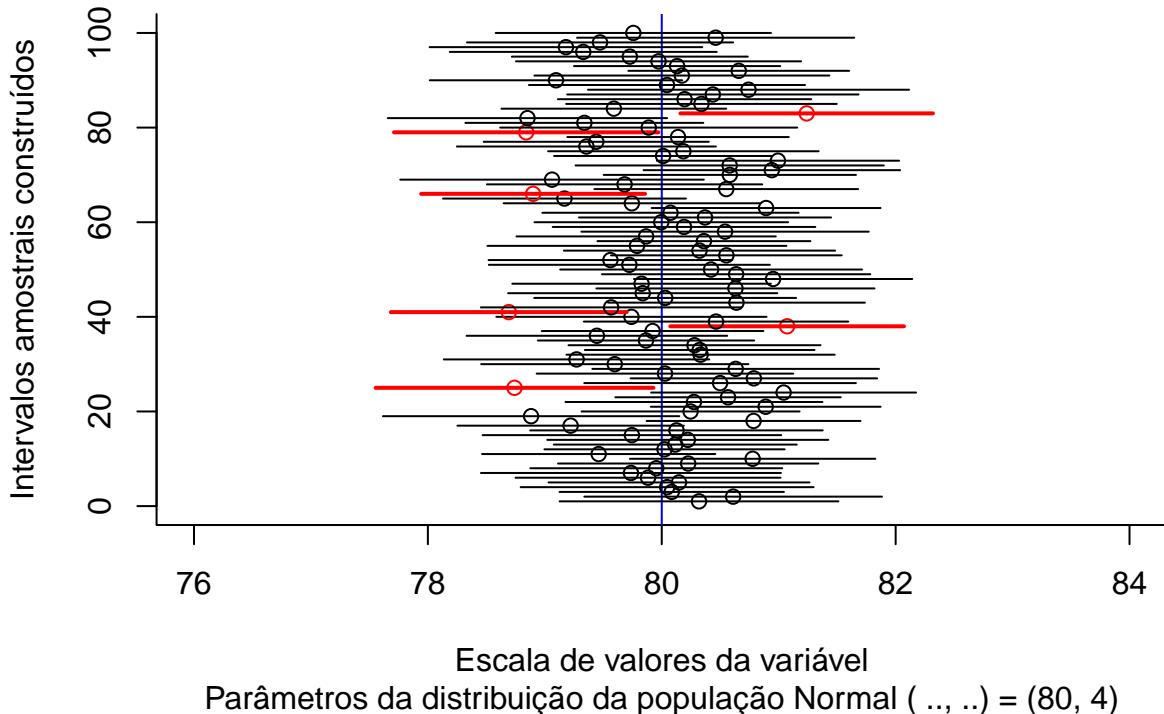


Figure 9.9: Histograma da distribuição dos erros de amostras de tamanho n , extraídas de uma população com distribuição $\sim N(\mu; \sigma)$ mostra que os mesmos seguem uma distribuição $\sim N(0; s = \frac{\sigma}{\sqrt{n}})$

```
erro_min=min(matriz$erro)
erro_max=max(matriz$erro)

meu_titulo1=paste("Distribuição dos erros de", N, "amostras de tamanho n=",n,"\\n extraídas
                   de uma população Normal ( \u03bc: ", mu, ", \u03c3: ", sigma, ")")
meu_titulo2=paste("Os erros amostrais ~ N( x\u0304=,round(mean(matriz$erro),2),"\u22480 ;
                   sd=",round(sd(matriz$erro),2)," \u2248 \u03c3/sqrt(n))")
```

```
ggplot(matriz, aes(x=erro)) +
  geom_histogram(aes(y = stat(density)), bins=round(sqrt(N),0), fill="lightblue",
                 col="black") +
  geom_area(stat = "function",
            fun = dnorm,
            args = list(mean=mean(matriz$erro), sd=sd(matriz$erro)),
            fill = NA,
            colour="red") +
  scale_y_continuous(name="Frequência") +
  scale_x_continuous(name="Valores dos erros amostrais", limits=c(-2,2)) +
  labs(title=meu_titulo1) +
  annotate(geom="text",
          label=meu_titulo2, x=-0.7,y= 0.9,
          angle=0, vjust=-0.5, hjust=0.5,
          color="blue",size=4) +
  theme(plot.title = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(angle=0, hjust=1, size=10),
        axis.text.y = element_text(angle=0, hjust=1, size=10),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10))
```

Distribuição dos erros de 100 amostras de tamanho n= 50 extraídas de uma população Normal (..: 80 , ..: 4)

Os erros amostrais $\sim N(\bar{x} = 0.05 \sim 0 ; \text{sd} = 0.58 \sim \sqrt{n})$

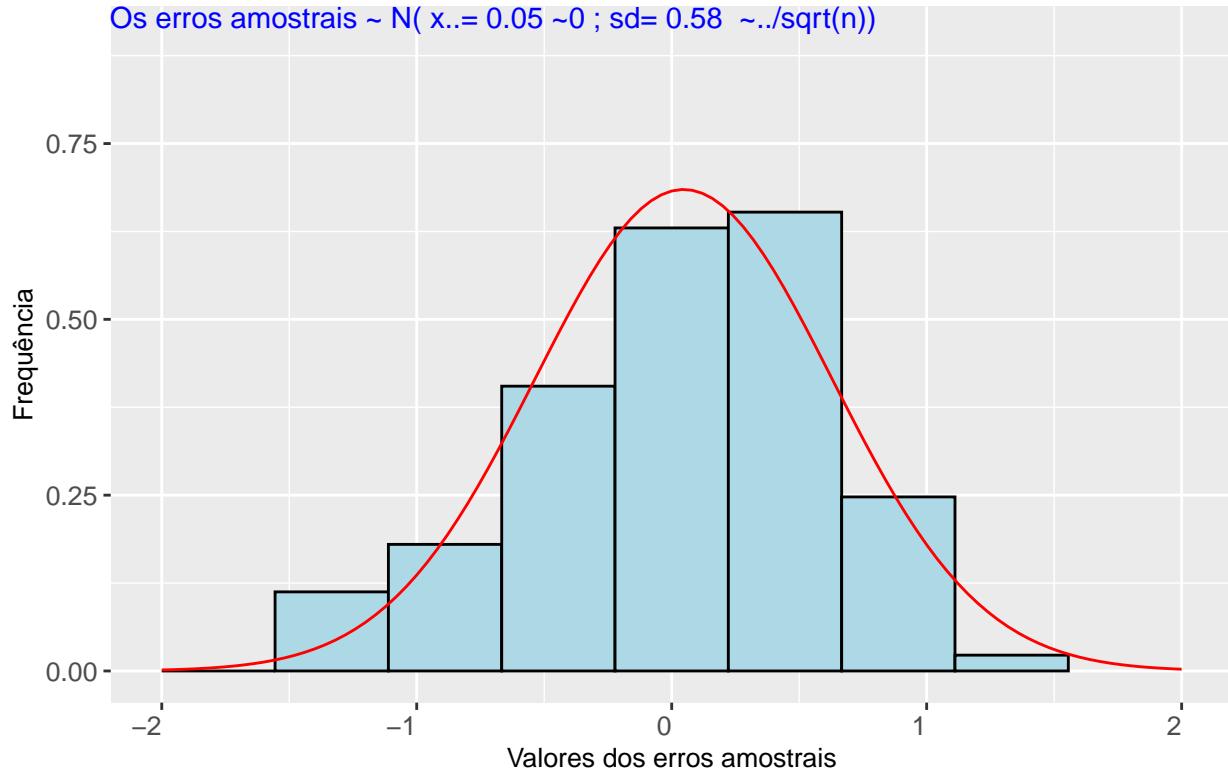


Figure 9.10: Histograma da distribuição dos erros de amostras de tamanho n, extraídas de uma população com distribuição $\sim N(\mu; \sigma)$ mostra que os mesmos seguem uma distribuição $\sim N(0; s = \frac{\sigma}{\sqrt{n}})$

Corolário: se (X_1, X_2, \dots, X_n) for uma amostra aleatória simples da população X de média μ e

variância σ^2 conhecida, e $\bar{X} = \frac{(X_1 + X_2 + \dots + X_n)}{n}$, tal que $n \geq 30$, então a estatística Z pode ser definida, bem como sua correspondente distribuição:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Uma vez que a estatística $Z \sim N(0, 1)$ (ela “decorre” da padronização da variável aleatória \bar{X}) as probabilidades para os intervalos desejados de valores Z podem ser facilmente encontrados em tabelas, como mais adiante se verá na constução de intervalos de confiança.

9.3.1 Fator de correção para populações finitas

Se amostras de tamanho n *sem reposição* são extraídas de uma população finita de tamanho N aplica-se o fator de correção para populações finitas ($\sqrt{\frac{(N-n)}{(N-1)}}$) junto ao desvio padrão das expressões do erro máximo ε anteriormente expostas:

$$\begin{aligned}\varepsilon &= (\bar{x} - \mu) = z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{(N-n)}{(N-1)}} \\ &= (\bar{x} - \mu) = z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{(N-n)}{(N-1)}} \\ &= (\bar{x} - \mu) = (t_{(1-\frac{\alpha}{2}, (n-1))}) \cdot \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{(N-n)}{(N-1)}}\end{aligned}$$

Portanto, para populações *finitas* com amostragem *sem reposição* (com $n < N$):

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n} \cdot \frac{(N-n)}{(N-1)})$$

9.3.2 Intervalo de confiança para médias amostrais

Se, por alguma razão, a variância populacional (σ^2) é conhecida, podemos utilizar \bar{X} como estimador pontual da média.

Assim, X seguirá uma distribuição Normal tal que:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Segue também que a estatística Z , como antes definida, seguirá uma distribuição Normal tal que:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

com:

- \bar{X} é a média da amostra;
- μ é a média populacional;
- σ é o desvio padrão populacional; e,
- n é o tamanho da amostra extraída.

Entretanto, a situação mais usual é aquela na qual não termos informação alguma sobre a variância populacional (σ^2).

Nessas situações, se o tamanho da amostra é grande (na prática $n \geq 30$), podemos substituir σ na estatística Z por S : substituir o desvio padrão populacional pelo desvio padrão da amostra extraída, sem que o erro cometido com esta substituição seja grande.

Com tal substituição, a estatística Z e passa a ser tal que:

$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0, 1)$$

em que:

- \bar{X} é a média amostral;
- μ é a média populacional;
- S é o desvio padrão da amostra; e,
- n é o tamanho da amostra.

Caso a variância populacional (σ^2) não seja conhecida e o tamanho da amostra **não possa** ser admitido como grande ($n < 30$) e sendo o estimador da variância amostral assim definido:

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X}_1)^2$$

Definindo-se a variável $Y = \frac{(n-1) \cdot s^2}{\sigma^2}$ tem uma distribuição χ^2 com $(n-1)$ graus de liberdade tal que:

$$Y = \frac{(n-1) \cdot s^2}{\sigma^2} \sim \chi^2_{(n-1)},$$

e considerando-se que Z é tal que:

$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0, 1)$$

segue a estatística T e sua correspondente distribuição, denominada por t de *Student*:

$$T = \frac{Z}{\sqrt{\frac{Y}{(n-1)}}} \sim t_{(n-1)}.$$

Para essa situação na qual a variância populacional não é conhecida e o tamanho amostral é pequeno, com alguma manipulação chega-se à estatística T e sua correspondente distribuição:

$$T = \frac{(\bar{X} - \mu)}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

em que:

- \bar{X} é a média amostral;
- μ é a média populacional;
- S é o desvio padrão da amostra; e,
- n é o tamanho da amostra; e,
- $(n - 1)$ é uma quantidade denominada como *graus de liberdade*.

As probabilidades associadas a um intervalo para um determinado valor da estatística “t” da distribuição de *Student* encontram-se tabeladas para variados graus de liberdade, como mais adiante se verá na constução de intervalos de confiança.

9.3.3 Intervalo de confiança bilateral para uma média amostral sob variância populacional conhecida (Figura 6.17)

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

em que:

- \bar{X} é a média amostral;
- μ é a média populacional;
- σ é o desvio padrão populacional;
- n é o tamanho da amostra; e,
- Z é a estatística a ser calculada para a construção do intervalo de confiança sob o nível de significância α estabelecido.

```

alfa=0.05

prob_desejada1=alfa/2
z_desejado1=round(qnorm(prob_desejada1),4)
d_desejada1=dnorm(z_desejado1, 0, 1)

prob_desejada2=1-alfa/2
z_desejado2=round(qnorm(prob_desejada2),4)
d_desejada2=dnorm(z_desejado2, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``z'' da distribuição Normal padrão") +
  labs(title=
      "Curva da função densidade \nDistribuição Normal Padrão",
       subtitle = "P(-z, z)=(1-\u03b1) em cinza (nível de confiança) \nP(-\u221e; -z)= P(z;
       \u221e)= \u03b1/2 em vermelho ")+
  geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3)+
  geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="-z", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
  annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="z", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
  annotate(geom="text", x=z_desejado1-1.8, y=0.1, label="Intervalo aberto à esq.
           \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Intervalo aberto à dir.
           \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade=
           (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+
```

theme_bw()

Curva da função densidade
Distribuição Normal Padrão

$P(-z, z) = (1 - \alpha)$ em cinza (nível de confiança)
 $P(-z; z) = P(z; -z) = \alpha/2$ em vermelho

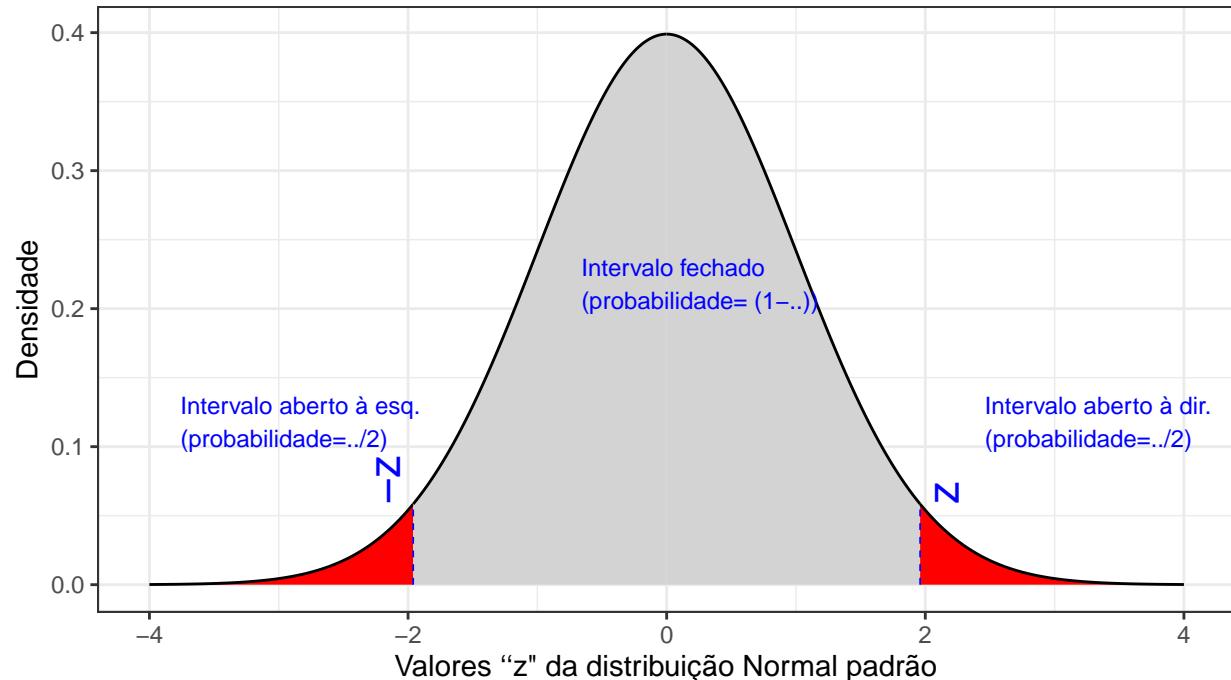


Figure 9.11: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores Z é inferior a $\frac{\alpha}{2}$, estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

Na Figura 9.11 observa-se:

- o nível de significância α ;
- o nível de confiança $(1 - \alpha)$; e,
- o valor tabelado da estatística $Z(z)$ para o nível de confiança fixado.

Assim,

$$\begin{aligned} P[-Z_{(1-\frac{\alpha}{2})} \leq Z \leq Z_{(1-\frac{\alpha}{2})}] &= (1 - \alpha) \\ P\left[-z_{(1-\frac{\alpha}{2})} \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P[\bar{x} - (z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}) \leq \mu \leq \bar{x} + (z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}})] &= (1 - \alpha) \end{aligned}$$

$$IC(\mu)_{(1-\alpha)} = [\bar{x} \pm z_c \cdot \frac{\sigma}{\sqrt{n}}]$$

Assim, se \bar{x} é usado como estimativa de μ , podemos afirmar estar $100.(1 - \alpha)\%$ confiantes de que o erro não excederá $(z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}})$.

A quantidade $\varepsilon = (\bar{x} - \mu) = z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}$ é chamada de Erro máximo da estimativa ao se arbitrar um nível de confiança α para um determinado tamanho amostral.

Exemplo: As vendas de 15 lojas de uma região do país apresentam uma média igual a US\$ 20.000,00. Sabendo-se que as vendas de todas as lojas da região é uma variável aleatória que segue uma distribuição Normal, com desvio padrão igual a US\$ 8.300,00, construa o intervalo de confiança para a média ao nível de confiança de 95%.

Dados do problema:

- o tamanho da amostra: $n = 15$;
- a média amostral: $\bar{x} = \text{US\$ } 20.000$;
- o desvio padrão populacional: $\sigma = \text{US\$ } 8.300$;
- nível de confiança: $(1 - \alpha) = 0,95$; e,
- valor extraído da tabela $z = 1,96$ correspondente ao nível de confiança estipulado $(1 - \alpha) = 95\%$.

$$\begin{aligned} P[\bar{x} - (z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}) \leq \mu \leq \bar{x} + (z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}})] &= (1 - \alpha) \\ P[20.000 - (1,96 \cdot \frac{8.300}{\sqrt{15}}) \leq \mu \leq 20000 + (1,96 \cdot \frac{8.300}{\sqrt{15}})] &= 0,95 \\ P[20.000 - 4.200,38 \leq \mu \leq 20.000 + 4.200,38] &= 0,95 \end{aligned}$$

$$IC_{(1-\alpha=0,95)} = [US\$15.799,62; US\$24.200,38]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que, se extrairmos um grande número de amostras de tamanho 15 dessa população, e para todas elas calcularmos intervalos de confiança como o acima definido, a proporção desses intervalos onde poderemos encontrar a média populacional de vendas será de 0,95 (95 intervalos em 100).

De uma forma mais sintética, podemos afirmar que o intervalo aleatório $[US\$ 15.799,62; US\$ 24.200,38]$, é um intervalo de confiança a 95% para a média de vendas.

De forma mais corrente, *embora menos correta* em termos teóricos, é usual afirmar que, com 95% de confiança a média de vendas se situa entre os valores US\$ 15.799,62 e US\$ 24.200,38.

Intervalos de confiança unilaterais para uma média amostral sob variância populacional conhecida.

A Figura 6.18 ilustra um intervalo de confiança unilateral limitado à direita por um valor máximo, dde tal sorte que a probabilidade associada ao intervalo de valores da estatística Z inferiores a esse limitante é

$$P \left[\mu \leq \bar{x} + z_c \cdot \frac{\sigma}{\sqrt{n}} \right] = (1 - \alpha)$$

```
prob_desejada=0.95
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-4, 0),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``z'' da distribuição Normal padrão") +
  geom_area(stat = "function",
            fun = dnorm,
```

```

fill = "lightgrey",
xlim = c(0, z_desejado),
colour="black")+
geom_area(stat = "function",
  fun = dnorm,
  fill = "red",
  xlim = c( z_desejado, 4),
  colour="black")+
labs(title=
  "Curva da função densidade
  \nDistribuição Normal Padrão",
  subtitle = "P(-\u2221e; z)=(1-\u03b1) em cinza (nível de confiança)  \nP(z, + \U2221e)=
  \u03b1, em vermelho ")+
annotate(geom="text", x=z_desejado1+3.5, y=d_desejada1, label="z", angle=90, vjust=0,
  hjust=0, color="blue",size=6)+
annotate(geom="text", x=z_desejado1+4.5, y=0.1, label="Intervalo aberto à dir.
  \n(probabilidade=\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à esq.
  \n(probabilidade= (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Curva da função densidade

Distribuição Normal Padrão

$P(-\dots; z)=(1-\dots)$ em cinza (nível de confiança)
 $P(z, + \dots)= \dots$, em vermelho

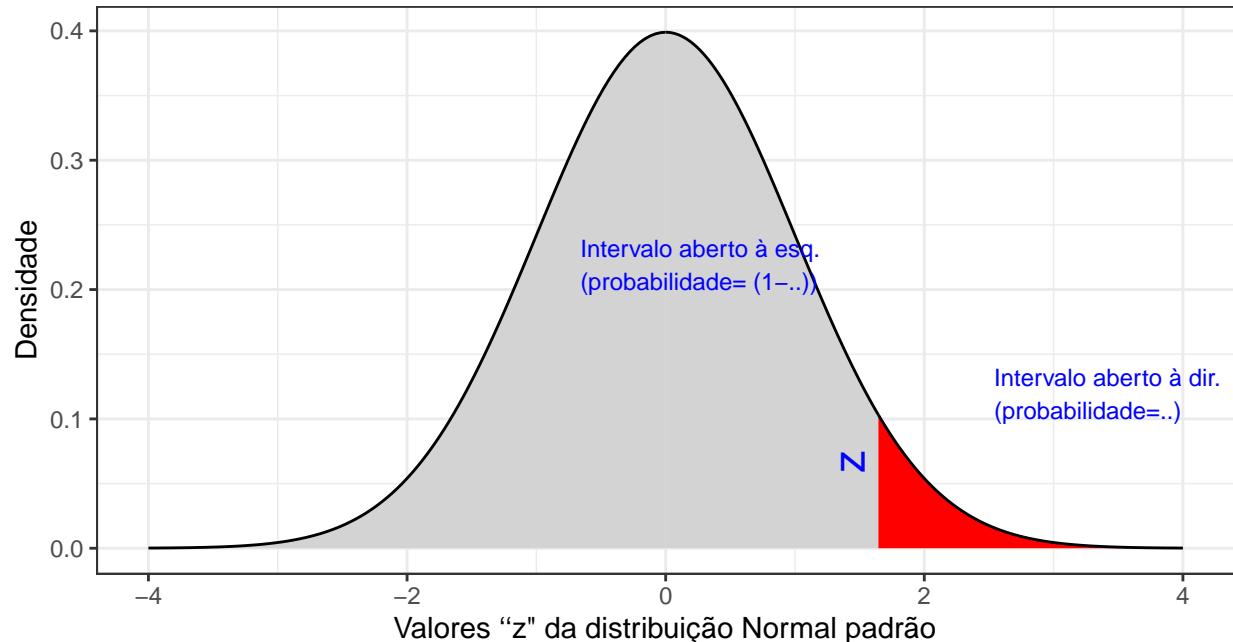


Figure 9.12: Região crítica, além da qual, a probabilidade associada aos valores Z é inferior a α , delimitando assim, à esquerda, um intervalo aberto com nível de confiança igual a $(1 - \alpha)$

A Figura 9.13 ilustra um intervalo de confiança unilateral limitado à esquerda por um valor mínimo, de tal sorte

que a probabilidade associada ao intervalo de valores da estatística Z superiores a esse limitante é

$$P \left[\mu \geq \bar{x} - z_c \cdot \frac{\sigma}{\sqrt{n}} \right] = (1 - \alpha)$$

```
prob_desejada=0.05
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-4, 0),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``z'' da distribuição Normal padrão") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado),
            colour="black")+
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c( z_desejado, 4),
            colour="black")+
  labs(title=
    "Curva da função densidade
    \nDistribuição Normal Padrão",
    subtitle = "P(-\u221e; z)=\u03b1, em vermelho \nP(z, + \u221e)= (1-\u03b1) em cinza")+
  annotate(geom="text", x=z_desejado1+0.5, y=d_desejada1, label="-z", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
  annotate(geom="text", x=z_desejado1-2, y=0.1, label="Intervalo aberto à esq.
    \n(probabilidade=\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à dir.
    \n(probabilidade= (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  theme_bw()
```

9.3.4 Intervalo de confiança para uma média amostral sob variância populacional desconhecida mas amostras não tão pequenas: $n \geq 30$ (Figura 9.14)

$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0, 1)$$

Curva da função densidade

Distribuição Normal Padrão

$P(-\dots; z) = \dots$, em vermelho

$P(z, + \dots) = (1 - \dots)$ em cinza

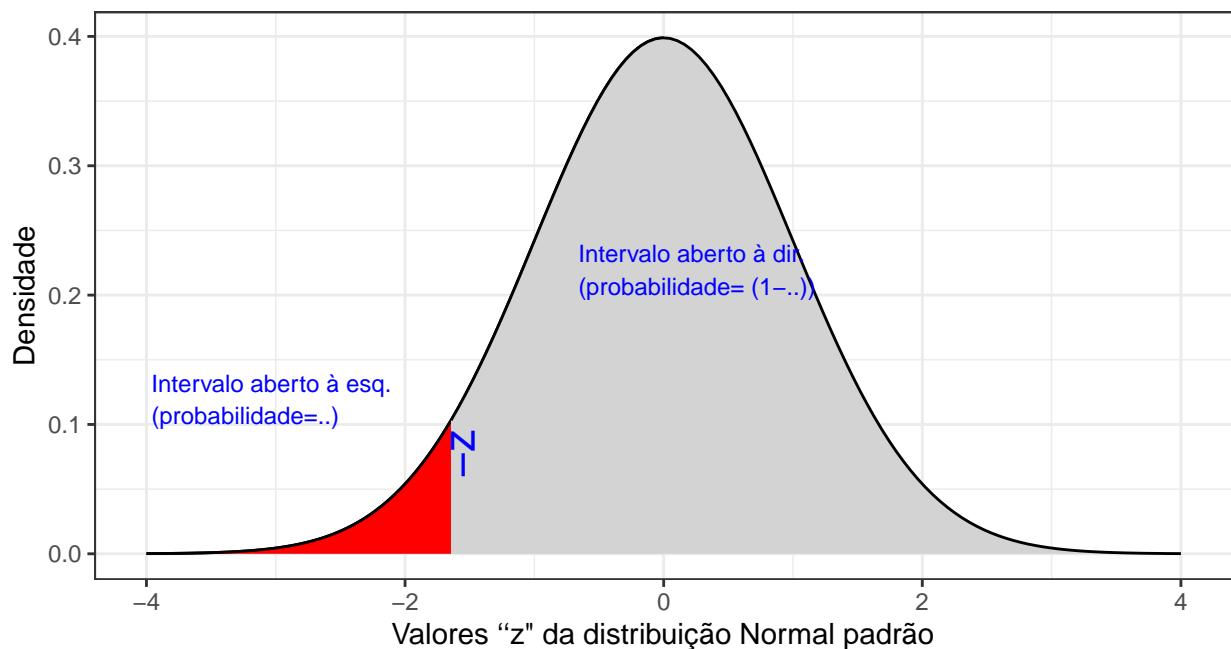


Figure 9.13: Região crítica, aquém da qual, a probabilidade associada aos valores Z é inferior a α , delimitando assim, à direita, um intervalo aberto com nível de confiança igual a $(1 - \alpha)$

em que:

- \bar{X} é a média amostral;
- μ é a média populacional;
- S é o desvio padrão amostral;
- n é o tamanho da amostra; e,
- Z é a estatística a ser calculada para a construção do intervalo de confiança sob o nível de significância α estabelecido.

```
alfa=0.05
```

```
prob_desejada1=alfa/2
z_desejado1=round(qnorm(prob_desejada1),4)
d_desejada1=dnorm(z_desejado1, 0, 1)

prob_desejada2=1-alfa/2
z_desejado2=round(qnorm(prob_desejada2),4)
d_desejada2=dnorm(z_desejado2, 0, 1)
```

```
ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``z'' da distribuição Normal padrão") +
  labs(title=
      "Curva da função densidade \nDistribuição Normal Padrão",
       subtitle = "P(-z; z)=(1-\u03b1) em cinza (nível de confiança) \nP(-\u221e; -z)= P(z;
      \u221e) = \u03b1/2 em vermelho") +
```

```

geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1),
  ↵ color="blue", lty=2, lwd=0.3)+
geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2),
  ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="-z", angle=90, vjust=0,
  ↵ hjust=0, color="blue",size=6)+
annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="z", angle=90, vjust=0,
  ↵ hjust=0, color="blue",size=6)+
annotate(geom="text", x=z_desejado1-1.8, y=0.1, label="Intervalo aberto à esq.
  ↵ \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Intervalo aberto à dir.
  ↵ \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade=
  ↵ (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Curva da função densidade Distribuição Normal Padrão

$P(-z; z) = (1 - \alpha)$ em cinza (nível de confiança)
 $P(-\dots; -z) = P(z; \dots) = \alpha/2$ em vermelho

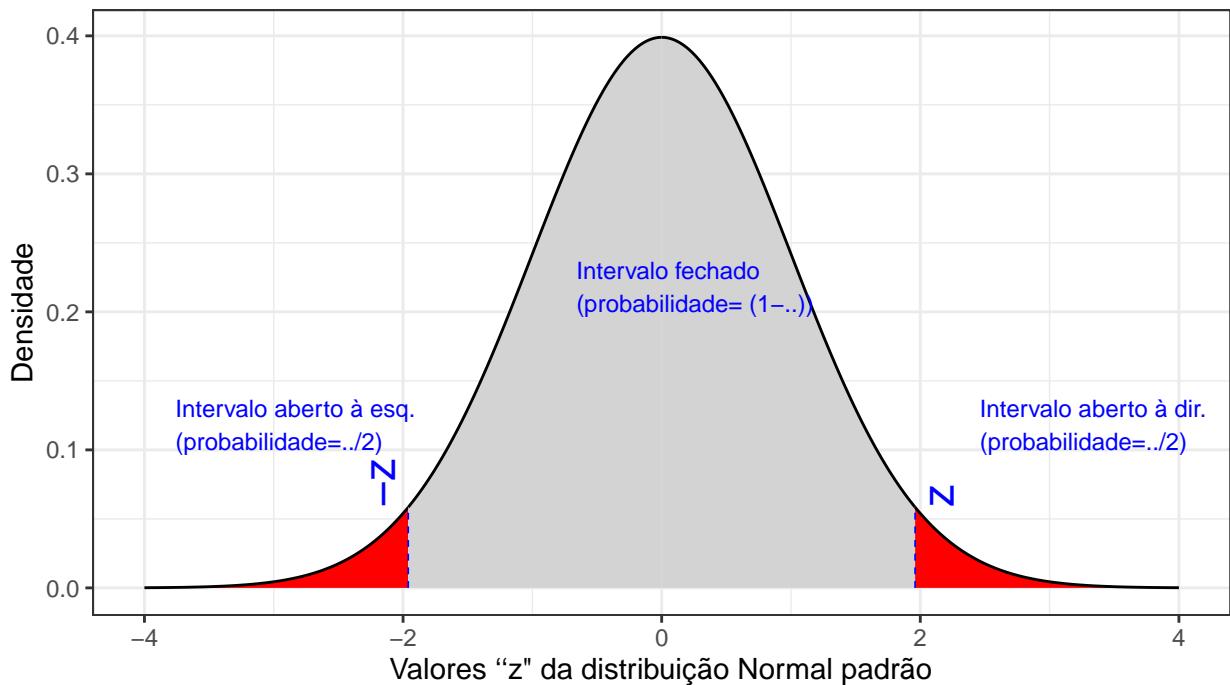


Figure 9.14: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores Z é inferior a $\frac{\alpha}{2}$, estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

Na Figura 9.14 observa-se:

- o nível de significância α ;

- o nível de confiança $(1 - \alpha)$; e,
- o valor tabelado da estatística $Z(z)$ para o nível de confiança fixado.

Assim,

$$\begin{aligned} P\left[-Z_{(1-\frac{\alpha}{2})} \leq Z \leq Z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P\left[-z_{(1-\frac{\alpha}{2})} \leq \frac{\bar{x} - \mu}{(\frac{S}{\sqrt{n}})} \leq z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P[\bar{x} - (z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}) \leq \mu \leq \bar{x} + (z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}})] &= (1 - \alpha) \end{aligned}$$

$$IC(\mu)_{(1-\alpha)} = [\bar{x} \pm z_c \cdot \frac{S}{\sqrt{n}}]$$

Assim, se \bar{x} é usado como estimativa de μ podemos afirmar estar $100(1 - \alpha)\%$ confiantes de que o erro não excederá $(z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}})$.

A quantidade $\varepsilon = (\bar{x} - \mu) = z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}$ é chamada de Erro máximo da estimativa ao se arbitrar um nível de confiança α para um determinado tamanho amostral.

Exemplo: As vendas de 60 lojas de uma região do país apresentam uma média igual a US\$ 20.000,00 e desvio padrão de US\$ 8.300,00. Construa o intervalo de confiança para a média ao nível de confiança de 95%.

Dados do problema:

- o tamanho da amostra: $n = 60$;
- a média amostral: $\bar{x} = \text{US\$}20.000$;
- o desvio padrão amostral: $s = \text{US\$}8.300$;
- nível de confiança: $(1 - \alpha) = 0,95$; e,
- valor extraído da tabela $z = 1,96$ correspondente ao nível de confiança estipulado $(1 - \alpha) = 95\%$.

$$\begin{aligned}
 P[\bar{x} - (z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}) \leq \mu \leq \bar{x} + (z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}})] &= (1 - \alpha) \\
 P[20.000 - (1,96 \cdot \frac{8.300}{\sqrt{60}}) \leq \mu \leq 20.000 + (1,96 \cdot \frac{8.300}{\sqrt{60}})] &= 0,95 \\
 P[20.000 - 2.100,19 \leq \mu \leq 20.000 + 2.100,19] &= 0,95
 \end{aligned}$$

$$IC_{(1-\alpha=0,95)} = [US\$17.899,81; US\$22.100,19]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que se extrairmos um grande número de amostras de tamanho 60 dessa população, e para todas elas calcularmos intervalos de confiança como o acima definido, a proporção desses intervalos onde poderemos encontrar a média populacional de vendas será de 0,95 (95 intervalos em 100).

De uma forma mais sintética, podemos afirmar que o intervalo aleatório $]US\$ 17.899,81; US\$ 22.100,19[$, é um intervalo de confiança a 95% para a média de vendas.

De forma mais corrente, *embora menos correta* em termos teóricos, é usual afirmar que, com 95% de confiança a média de vendas se situa entre os valores US\$ 17.899,81 e US\$ 22.100,19.

Intervalos de confiança unilaterais para uma média amostral sob variância populacional desconhecida mas amostras não tão pequenas: $n \geq 30$.

A Figura 9.15 ilustra um intervalo de confiança unilateral limitado à direita por um valor máximo, de tal sorte que a probabilidade associada ao intervalo de valores da estatística Z inferiores a esse limitante é

$$P\left[\mu \leq \bar{x} + z_c \cdot \frac{S}{\sqrt{n}}\right] = (1 - \alpha)$$

```

prob_desejada=0.95
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-4, 0),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``z'' da distribuição Normal padrão") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado),
            colour="black")+
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c( z_desejado, 4),
            colour="black")+
  labs(title=
    "Curva da função densidade
    \nDistribuição Normal Padrão",
    subtitle = "P(-\u221e; z)=(1-\u03b1) em cinza (nível de confiança) \nP(z, + \u221e)=
    \u03b1, em vermelho ")+
  annotate(geom="text", x=z_desejado1+3.5, y=d_desejada1, label="z", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
  annotate(geom="text", x=z_desejado1+4.5, y=0.1, label="Intervalo aberto à dir.
    \n(probabilidade=\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à esq.
    \n(probabilidade= (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  theme_bw()

```

A Figura 9.16 ilustra um intervalo de confiança unilateral limitado à esquerda por um valor mínimo, de tal sorte que a probabilidade associada ao intervalo de valores da estatística Z superiores a esse limitante é

$$P \left[\mu \geq \bar{x} - z_c \cdot \frac{S}{\sqrt{n}} \right] = (1 - \alpha)$$

```

prob_desejada=0.05
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",

```

Curva da função densidade

Distribuição Normal Padrão

$P(-\dots; z) = (1-\dots)$ em cinza (nível de confiança)
 $P(z, + \dots) = \dots$, em vermelho

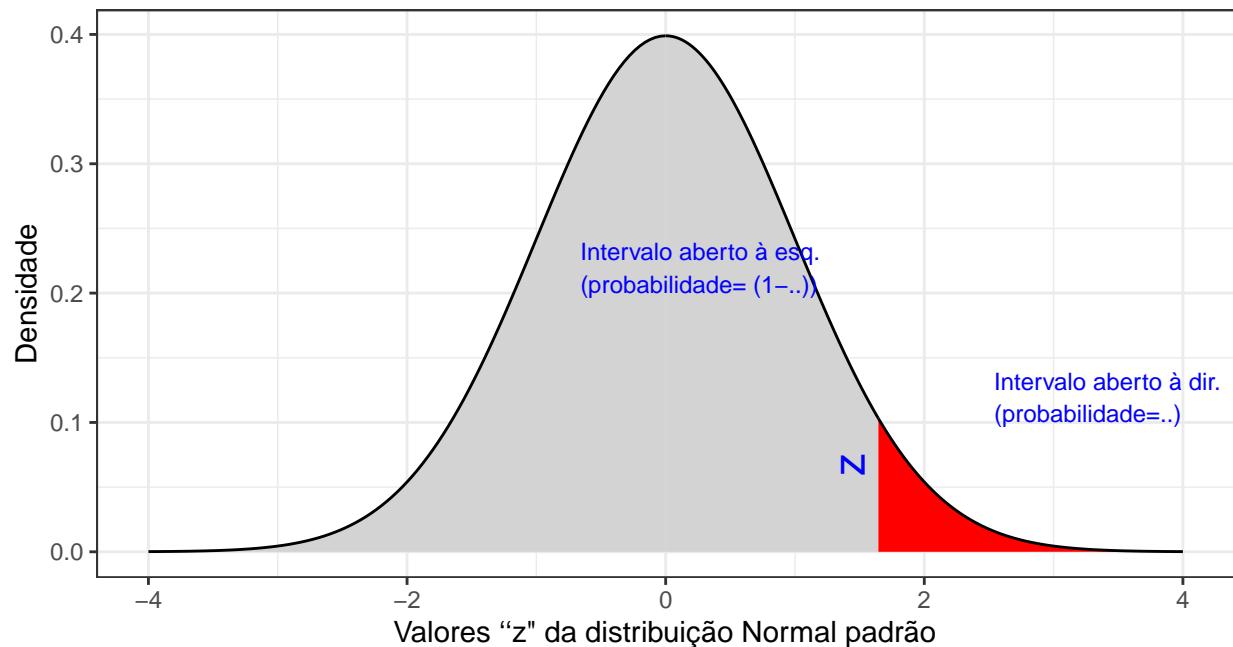


Figure 9.15: Região crítica, além da qual, a probabilidade associada aos valores Z é inferior a α , delimitando assim, à esquerda, um intervalo aberto com nível de confiança igual a $(1 - \alpha)$

```

    fun = dnorm,
    fill = "lightgrey",
    xlim = c(-4, 0),
    colour="black") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores ``z'' da distribuição Normal padrão") +
geom_area(stat = "function",
    fun = dnorm,
    fill = "red",
    xlim = c(-4, z_desejado),
    colour="black")+
geom_area(stat = "function",
    fun = dnorm,
    fill = "lightgrey",
    xlim = c( z_desejado, 4),
    colour="black")+
labs(title=
  "Curva da função densidade
  \nDistribuição Normal Padrão",
  subtitle = "P(-\u221e; z)=\u03b1, em vermelho \nP(z, + \u221e)= (1-\u03b1) em cinza")+
annotate(geom="text", x=z_desejado1+0.5, y=d_desejada1, label="-z", angle=90, vjust=0,
  ↵ hjust=0, color="blue",size=6)+
annotate(geom="text", x=z_desejado1-1.5, y=0.1, label="Intervalo aberto à esq.
  ↵ \n(probabilidade=\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à dir.
  ↵ \n(probabilidade= (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

9.3.5 Intervalo de confiança para uma média amostral sob variância populacional desconhecida e amostras de qualquer tamanho (Figura 9.17)

$$T = \frac{(\bar{X} - \mu)}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

em que:

- \bar{X} é a média amostral;
- μ é a média populacional;
- S é o desvio padrão amostral;
- n é o tamanho da amostra; e,
- T é a estatística a ser calculada para a construção do intervalo de confiança sob o nível de significância α estabelecido.

Curva da função densidade

Distribuição Normal Padrão

$P(-\dots; z) = \dots$, em vermelho

$P(z, + \dots) = (1 - \dots)$ em cinza

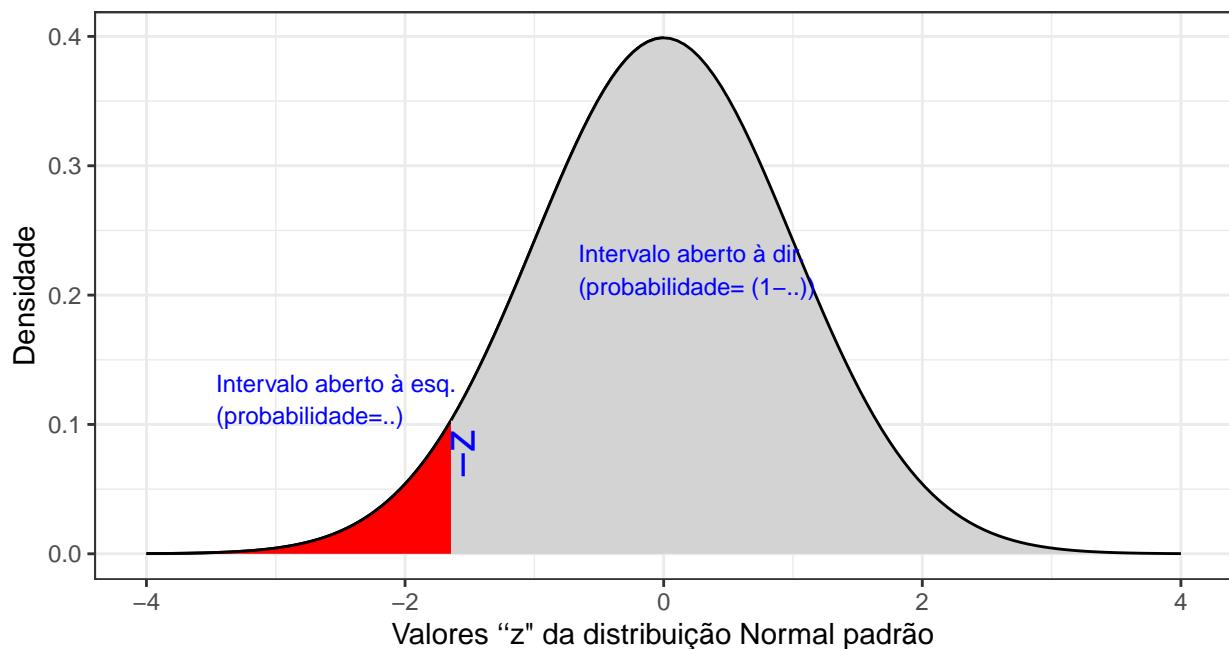


Figure 9.16: Região crítica, aquém da qual, a probabilidade associada aos valores Z é inferior a α , delimitando assim, à direita, um intervalo aberto com nível de confiança igual a $(1 - \alpha)$

```

alfa=0.05

prob_desejada1=alfa/2
df=20
t_desejado1=round(qt(prob_desejada1,df ),4)
d_desejada1=dt(t_desejado1,df)

prob_desejada2=1-alfa/2
df=20
t_desejado2=round(qt(prob_desejada2, df),4)
d_desejada2=dt(t_desejado2,df)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(-4, t_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(t_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(0, t_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(t_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``t'' da distribuição de Student com gl=n-1") +
  labs(title= "Curva da função densidade \nDistribuição t ",
       subtitle = "P(-t; t)=(1-\u03b1) em cinza (nível de confiança) \nP(-\u221e; -t)= P(t;
                  \u221e)= \u03b1/2 em vermelho ")+
  geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3)+
  geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="-t", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
  annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="t", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
  annotate(geom="text", x=t_desejado1-1.8, y=0.1, label="Intervalo aberto à esq.
           \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Intervalo aberto à dir.
           \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
```

```
annotate(geom="text", x=t_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade=\n(1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+ theme_bw()
```

Curva da função densidade Distribuição t

$P(-t; t) = (1 - \alpha)$ em cinza (nível de confiança)
 $P(-\infty; -t) = P(t; \infty) = \alpha/2$ em vermelho

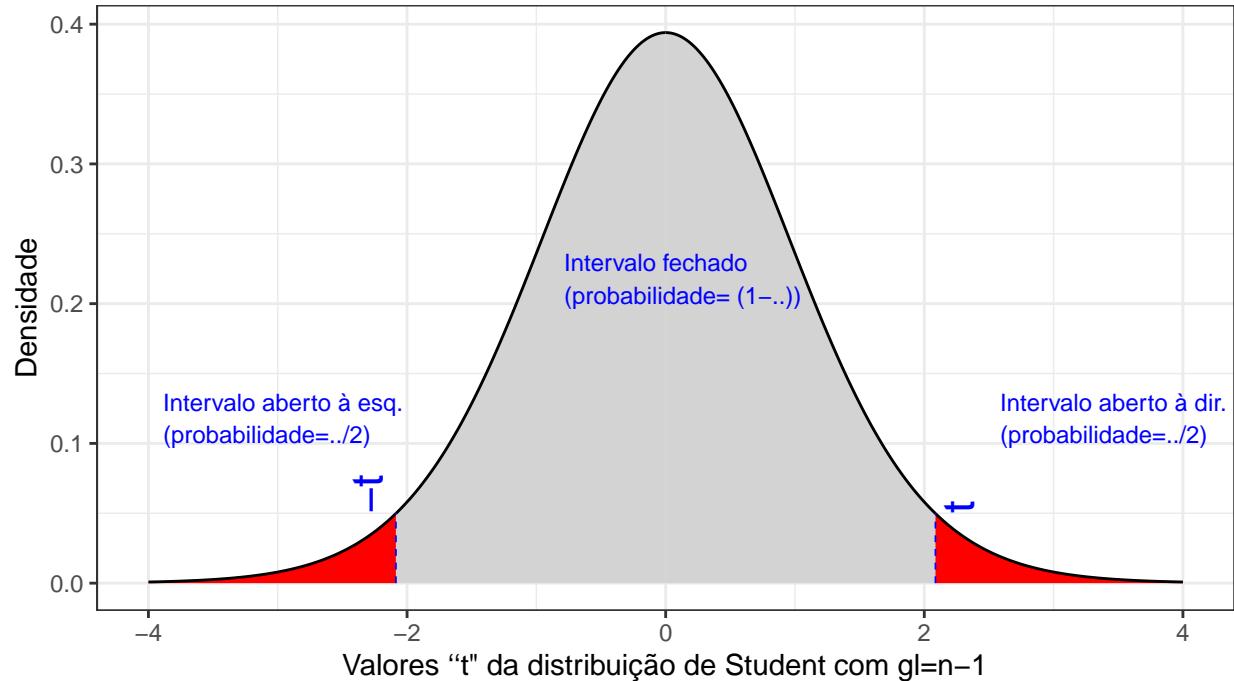


Figura 9.17: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores T ($(n - 1)$ graus de liberdade) é inferior a $\frac{\alpha}{2}$, estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

Na Figura 9.17 observa-se:

- o nível de significância α ;
- o nível de confiança $(1 - \alpha)$; e,
- o valor tabelado da estatística $T(t)$ sob $n - 1$ graus de liberdade para o nível de confiança fixado.

Assim,

$$\begin{aligned}
P[-T_{(1-\frac{\alpha}{2},(n-1))} \leq T \leq T_{(1-\frac{\alpha}{2},(n-1))}] &= (1 - \alpha) \\
P\left[-t_{(1-\frac{\alpha}{2},(n-1))} \leq \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{(1-\frac{\alpha}{2},(n-1))}\right] &= (1 - \alpha) \\
P[\bar{x} - (t_{(1-\frac{\alpha}{2},(n-1))} \cdot \frac{S}{\sqrt{n}}) \leq \mu \leq \bar{x} + (t_{(1-\frac{\alpha}{2},(n-1))} \cdot \frac{S}{\sqrt{n}})] &= (1 - \alpha) \\
IC(\mu)_{(1-\alpha)} &= [\bar{x} \pm t_{c_{(n-1)}} \cdot \frac{S}{\sqrt{n}}]
\end{aligned}$$

Assim, se \bar{x} é usado como estimativa de μ podemos afirmar estar $100(1 - \alpha)\%$ confiantes de que o erro não excederá $(t_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}})$.

A quantidade $\varepsilon = (\bar{x} - \mu) = (t_{(1-\frac{\alpha}{2},(n-1))} \cdot \frac{S}{\sqrt{n}})$ é chamada de Erro máximo da estimativa ao se arbitrar um nível de confiança α , $(n-1)$ graus de liberdade e um determinado tamanho amostral.

Exemplo: As vendas de 15 lojas de uma região do país apresentam uma média igual a US\$ 20.000,00 e desvio padrão de US\$ 8.300,00. Construa o intervalo de confiança para a média ao nível de confiança de 95%.

Dados do problema:

- o tamanho da amostra: $n = 15$;
- a média amostral: $\bar{x} = \text{US\$}20.000$;
- o desvio padrão amostral: $s = \text{US\$}8.300$;
- nível de confiança: $(1 - \alpha) = 0,95$; e,
- valor extraído da tabela da distribuição de *Student* sob $(n-1 = 15-1 = 14)$ graus de liberdade $t_c = 2,1448$ associado ao nível de confiança estipulado $(1 - \alpha) = 95\%$.

$$\begin{aligned}
P[\bar{x} - (t_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}) \leq \mu \leq \bar{x} + (t_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}})] &= (1 - \alpha) \\
P[20000 - (2,1448 \cdot \frac{8300}{\sqrt{15}}) \leq \mu \leq 20000 + (2,1448 \cdot \frac{8300}{\sqrt{15}})] &= 0,95 \\
P[20000 - 4596,41 \leq \mu \leq 20000 + 4596,41] &= 0,95
\end{aligned}$$

$$IC_{(1-\alpha=0,95)} = [US\$15403,59; US\$24496,41]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que se extrairmos um grande número de amostras de tamanho 15 dessa população, e para todas elas calcularmos intervalos de confiança como o acima definido, a proporção desses intervalos onde poderemos encontrar a média populacional de vendas será de 0,95 (95 intervalos em 100).

De uma forma mais sintética, podemos afirmar que o intervalo aleatório $]US\$ 15.403,59; US\$ 24.496,41[$, é um intervalo de confiança a 95% para a média de vendas.

De uma forma mais corrente, *embora menos correta* em termos teóricos, é usual afirmar que, com 95% de confiança a média de vendas se situa entre os valores US\$ 15.403,59 e US\$ 24.496,41.

Intervalos de confiança unilaterais para uma média amostral sob variância populacional desconhecida e amostras de qualquer tamanho

A Figura 9.18 ilustra um intervalo de confiança unilateral limitado à direita por um valor máximo, de tal sorte que a probabilidade associada ao intervalo de valores da estatística T inferiores a esse limitante é

$$P \left[\mu \leq \bar{x} + t_{c_{(n-1)}} \cdot \frac{S}{\sqrt{n}} \right] = (1 - \alpha)$$

```
alfa=0.95
prob_desejada1=alfa
df=20
t_desejado1=round(qt(prob_desejada1,df ),4)
d_desejada1=dt(t_desejado1,df )

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c( t_desejado1, 4),
            colour="black") +
```

```

geom_area(stat = "function",
           fun = dt,
           args=list(df),
           fill = "lightgrey",
           xlim = c(0, t_desejado1),
           colour="black") +
geom_area(stat = "function",
           fun = dt,
           args=list(df),
           fill = "lightgrey",
           xlim = c(-4, 0),
           colour="black") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores ``t'' da distribuição de Student com gl=n-1") +
labs(title= "Curva da função densidade \nDistribuição t",
     subtitle = "P(-\u2212t)=(1-\u03b1) em cinza \nP(t, t)=\u03b1 em vermelho "+)
geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1),
             color="blue", lty=2, lwd=0.3) +
annotate(geom="text", x=t_desejado1+0.5, y=d_desejada1, label="t", angle=90, vjust=0,
         hjust=0, color="blue",size=6) +
annotate(geom="text", x=t_desejado1+1, y=0.1, label="Intervalo aberto à esq.",
         \n(probabilidade=\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3) +
annotate(geom="text", x=t_desejado1-2.5, y=0.2, label="Intervalo aberto \n(probabilidade=
         (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3) + theme_bw()

```

Curva da função densidade

Distribuição t

$P(-\dots, t)=(1-\dots)$ em cinza

$P(t, \dots)=\dots$ em vermelho

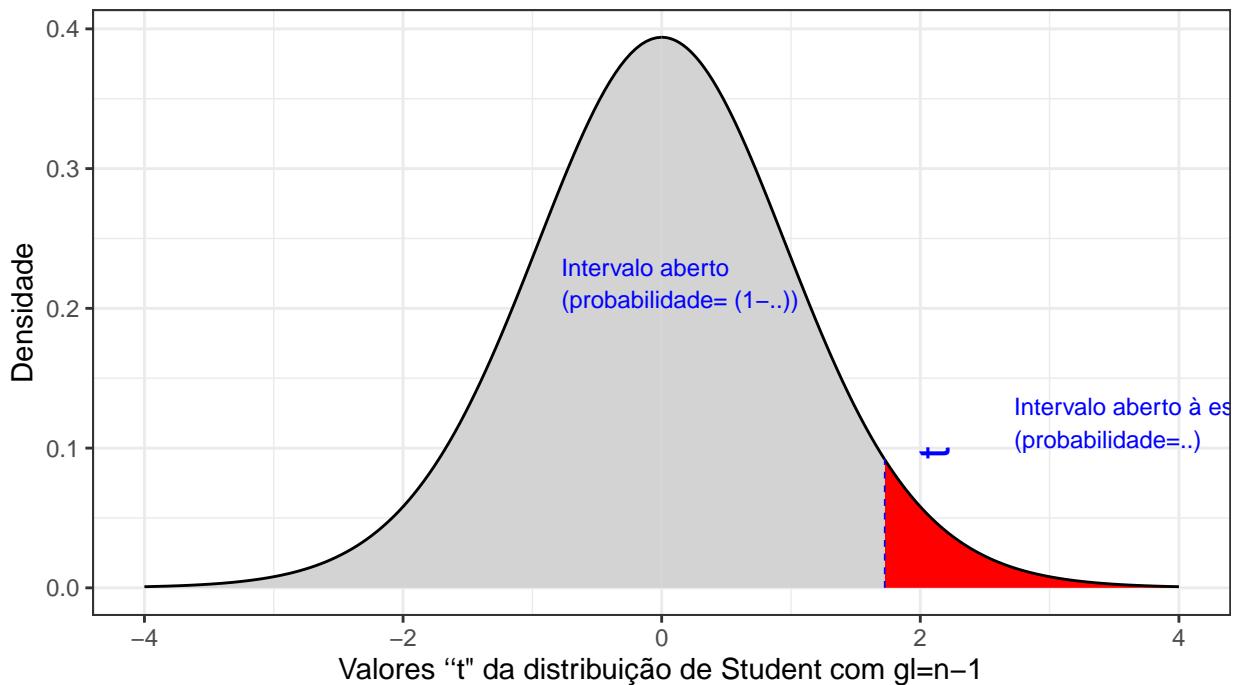


Figure 9.18: Região crítica, além da qual, a probabilidade associada aos valores T ($(n - 1)$ graus de liberdade) é inferior a α , delimitando assim, à esquerda, um intervalo aberto com nível de confiança igual a $(1 - \alpha)$

A Figura 9.19 ilustra um intervalo de confiança unilateral limitado à esquerda por um valor mínimo, de tal sorte que a probabilidade associada ao intervalo de valores da estatística T superiores a esse limitante é

$$P \left[\mu \geq \bar{x} - t_c \cdot \frac{S}{\sqrt{n}} \right] = (1 - \alpha)$$

```

alfa=0.05
prob_desejada1=alfa
df=20
t_desejado1=round(qt(prob_desejada1,df ),4)
d_desejada1=dt(t_desejado1,df)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(-4, t_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(t_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(0, 4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``t'' da distribuição de Student com gl=n-1") +
  labs(title= "Curva da função densidade \nDistribuição t ",
       subtitle = "P(-t, \u221e)=(1-\u03b1) em cinza \nP(-\u221e; -t)= \u03b1 em vermelho
                  ") +
  geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3) +
  annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="-t", angle=90, vjust=0,
           hjust=0, color="blue",size=6) +
  annotate(geom="text", x=t_desejado1-2.5, y=0.1, label="Intervalo aberto à esq.
             \n(probabilidade=\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=t_desejado1+1, y=0.2, label="Intervalo aberto \n(probabilidade=
             (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3) + theme_bw()

```

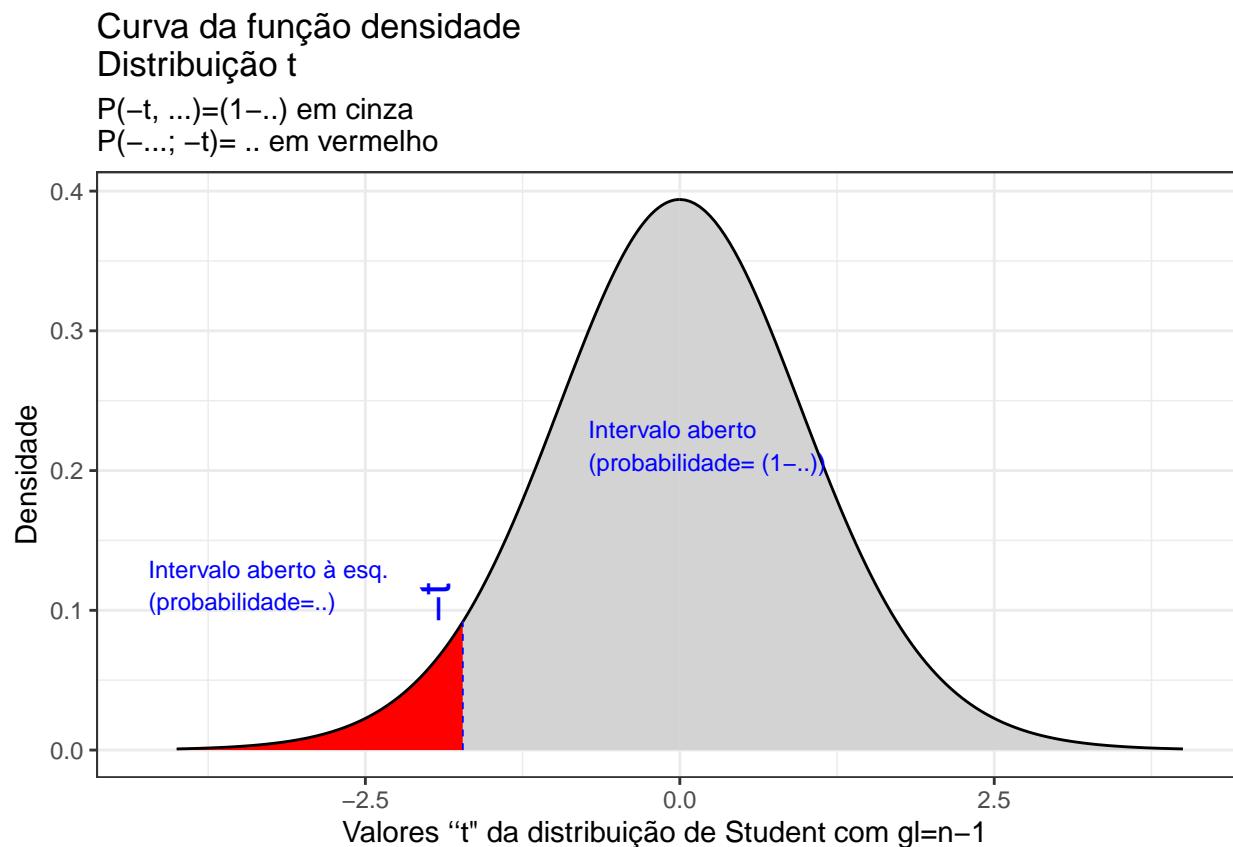


Figure 9.19: Região crítica, aquém da qual, a probabilidade associada aos valores T ($(n-1)$ graus de liberdade) é inferior a α , delimitando assim, à direita, um intervalo aberto com nível de confiança igual a $(1 - \alpha)$

9.4 Distribuição das diferenças de médias amostrais independentes e seus intervalos de confiança

Consideremos duas populações X e Y com médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2 , respectivamente.

Conforme seções anteriores, as médias amostrais \bar{X} e \bar{Y} são duas variáveis aleatórias tais que:

$$\begin{aligned}\bar{X} &\sim N(\mu_1, \frac{\sigma_1^2}{n_1}) \\ \bar{Y} &\sim N(\mu_2, \frac{\sigma_2^2}{n_2})\end{aligned}$$

Pode-se demonstrar, pelas propriedades da esperança e da variância, que a média e a variância de uma variável aleatória (população) que resulta da soma ou diferença de duas outras, X e Y , é:

$$\begin{aligned}\mu_{(X \pm Y)} &= \mu_1 \pm \mu_2 \\ \sigma_{(X \pm Y)}^2 &= \sigma_1^2 + \sigma_2^2\end{aligned}$$

E a média e variância da soma ou diferença das distribuições amostrais das médias de X e Y é:

$$\begin{aligned}\mu_{(\bar{X} \pm \bar{Y})} &= \mu_1 \pm \mu_2 \\ \sigma_{(\bar{X} \pm \bar{Y})}^2 &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\end{aligned}$$

9.4.1 Intervalos de confiança para a diferença entre duas médias amostrais com variância populacionais conhecidas

Se $(X_1, X_2, \dots, X_{n_1})$ e $(Y_1, Y_2, \dots, Y_{n_2})$ forem amostras aleatórias simples das populações X e Y com médias μ_1 e μ_2 , e variâncias σ_1^2 e σ_2^2 conhecidas, e $\bar{X} = \frac{(X_1 + X_2 + \dots + X_{n_1})}{n_1}$ e $\bar{Y} = \frac{(Y_1 + Y_2 + \dots + Y_{n_2})}{n_2}$, então:

9.4. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS INDEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

$$X \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$$

$$Y \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

Demonstra-se que a diferença entre \bar{X} e \bar{Y} é tal que:

$$\bar{X} - \bar{Y} \sim N((\mu_1 - \mu_2), \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

Demonstra-se que a estatística Z pode ser assim definida, bem como sua correspondente distribuição (cf.Figura 9.20):

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

em que:

- \bar{X} e \bar{Y} são as médias amostrais;
- μ_1 e μ_2 são as médias populacionais;
- σ_1^2 e σ_2^2 são as variâncias populacionais; e,
- n_1 e n_2 são os tamanhos das amostras

```
alfa=0.05

prob_desejada1=alfa/2
z_desejado1=round(qnorm(prob_desejada1),4)
d_desejada1=dnorm(z_desejado1, 0, 1)

prob_desejada2=1-alfa/2
z_desejado2=round(qnorm(prob_desejada2),4)
```

```
d_desejada2=dnorm(z_desejado2, 0, 1)
```

```
ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``z'' da distribuição Normal padrão") +
  labs(title=
      "Curva da função densidade \nDistribuição Normal Padrão",
      subtitle = "P(-z; z)=(1-\u03b1) em cinza (nível de confiança) \nP(-\u221e; -z)= P(z;
      \u221e)= \u03b1/2 em vermelho")+
  geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3)+
  geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="-z", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
  annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="z", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
  annotate(geom="text", x=z_desejado1-1.8, y=0.1, label="Intervalo aberto à esq.
  \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Intervalo aberto à dir.
  \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade=
  (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  theme_bw()
```

Na Figura 9.20 observa-se:

- o nível de significância α ;

9.4. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS INDEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

Curva da função densidade Distribuição Normal Padrão

$P(-z; z) = (1 - \alpha)$ em cinza (nível de confiança)
 $P(-z; z) = P(z; z) = \alpha/2$ em vermelho

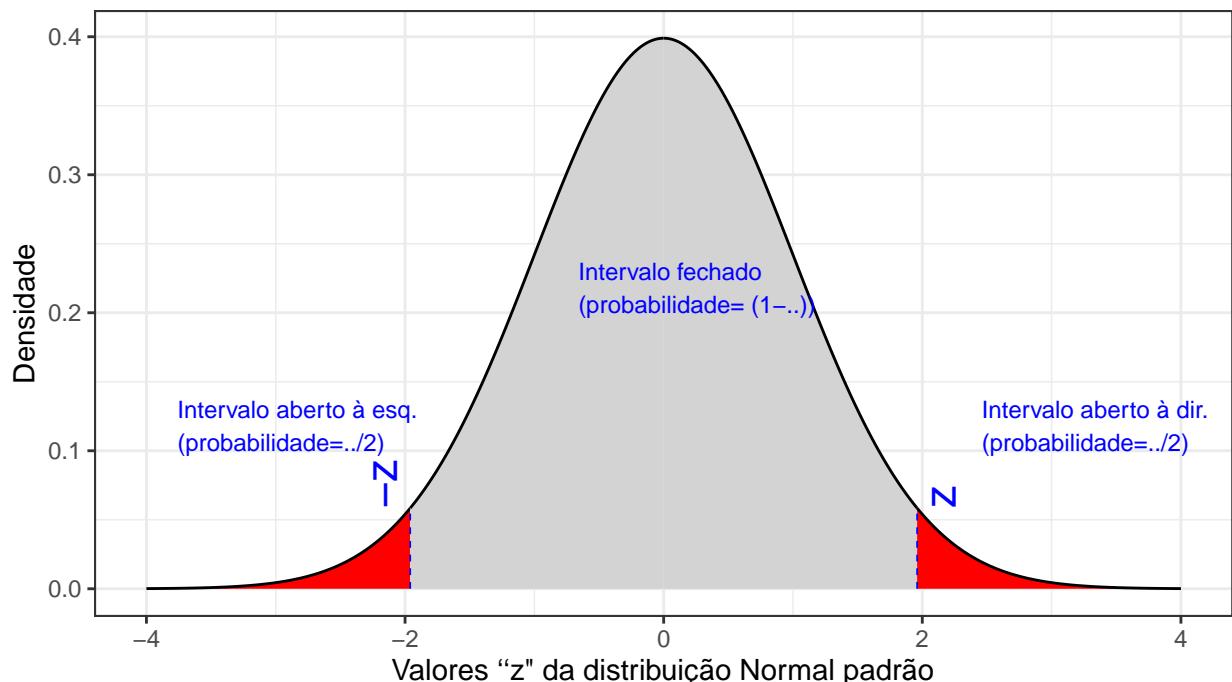


Figure 9.20: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores da estatística Z é inferior a $\frac{\alpha}{2}$, estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

- o nível de confiança $(1 - \alpha)$; e,
- o valor tabelado da estatística $Z(z)$ para o nível de confiança fixado.

Assim,

$$\begin{aligned} P\left[-Z_{(1-\frac{\alpha}{2})} \leq Z \leq Z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P\left[-z_{(1-\frac{\alpha}{2})} \leq \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P[(\bar{x} - \bar{y}) - (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})] &= (1 - \alpha) \end{aligned}$$

$$IC(\mu_1 - \mu_2)_{(1-\alpha)} = [\bar{x} - \bar{y}] \pm z_c \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Exemplo: Uma empresa possui duas filiais (A e B). Uma amostra das vendas de 20 dias forneceu uma venda média diária de 40 unidades dessa peça a filial A e de 30 unidades da mesma peça para a filial B. Os desvios padrão das vendas diárias dessa peça são de 5 e 3, respectivamente. Admitindo que a distribuição diária das vendas dessa peça siga uma distribuição Normal, qual o intervalo de confiança para a diferença de médias das vendas nas duas filiais com um nível de confiança de 95%?

Dados do problema:

- $\bar{X} = 40$ e $\bar{Y} = 30$ são as médias amostrais (vendas médias diárias nas filiais A e B, respectivamente);
- $\sigma_1^2 = 25$ e $\sigma_2^2 = 9$ são as variâncias populacionais;
- $n_1 = n_2 = 20$ são os tamanhos das amostras; e,
- valor extraído da tabela $z = 1,96$ correspondente ao nível de confiança estipulado $(1 - \alpha) = 95\%$.

9.4. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS INDEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

$$\begin{aligned}
 P[(\bar{x} - \bar{y}) - (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})] &= (1 - \alpha) \\
 P[(\bar{x} - \bar{y}) - (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})] &= (1 - \alpha) \\
 P[10 - (1,96 \cdot \sqrt{\frac{25}{20} + \frac{9}{20}}) \leq (\mu_1 - \mu_2) \leq (10 + (1,96 \cdot \sqrt{\frac{25}{20} + \frac{9}{20}}))] &= 0,95 \\
 P[10 - (1,96 \times 1,3038) \leq (\mu_1 - \mu_2) \leq 10 + (1,96 \times 1,3038)] &= 0,95
 \end{aligned}$$

$$IC(\mu_1 - \mu_2)_{0,95} = [7; 13]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que, se extraímos um grande número de amostras dessas mesmas dimensões das vendas dessa peça nas duas empresas, e para cada uma delas calcularmos suas médias e as diferenças entre elas, e calcularmos os intervalos de confiança como o acima definido, a proporção desses intervalos onde podemos encontrar a diferença das médias de vendas dessa peça da filial A para a filial B será de 0,95 (95 intervalos em 100).

De uma forma mais sintética podemos afirmar que, o anterior intervalo aleatório [7 ; 13], é um intervalo de confiança a 95% para a diferença das médias de vendas dessa peça nas duas empresas

De uma forma mais corrente, *embora menos correta* em termos teóricos, é usual afirmar que, com 95% de confiança a diferença das médias de vendas dessa peça da filial A para a filial B se situa entre os valores 7 e 13.

Uma segunda observação se faz pertinente e se refere à natureza dos dados analisados e a forma de apresentação do resultado. Por serem dados discretos, o intervalo de confiança deverá ser apresentado em igual forma, sem ultrapassar os limites estabelecidos. Isto posto: $IC(\mu_1 - \mu_2)_{0,95} = [7; 13] \text{ peças}$.

9.4.2 Intervalos de confiança para a diferença entre duas médias amostrais com variâncias populacionais desconhecidas mas admitidas iguais

Se $(X_1, X_2, \dots, X_{n_1})$ e $(Y_1, Y_2, \dots, Y_{n_2})$ forem amostras aleatórias simples das populações X e Y com médias μ_1 e μ_2 , e variâncias σ_1^2 e σ_2^2 desconhecidas porém iguais ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), e $\bar{X} = \frac{(X_1 + X_2 + \dots + X_{n_1})}{n_1}$ e $\bar{Y} = \frac{(Y_1 + Y_2 + \dots + Y_{n_2})}{n_2}$, então:

$$X \sim N(\mu_1, \frac{\sigma}{\sqrt{n_1}})$$

$$Y \sim N(\mu_2, \frac{\sigma}{\sqrt{n_2}})$$

Demonstra-se que a estatística T pode ser assim definida, bem como sua correspondente distribuição (cf. Figura ??):

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

em que:

- \bar{X} e \bar{Y} são as médias amostrais;
- S_1^2 e S_2^2 são as variâncias amostrais;
- μ_1 e μ_2 são as médias populacionais;
- S_p é um desvio padrão amostral ponderado para as duas amostras;
- n_1 e n_2 são os tamanhos das amostras;

O desvio padrão ponderado S_p é dado por:

$$S_p = \sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}}$$

```
alfa=0.05
prob_desejada1=alfa/2
df=20
t_desejado1=round(qt(prob_desejada1,df ),4)
d_desejada1=dt(t_desejado1,df)

prob_desejada2=1-alfa/2
```

9.4. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS INDEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

```

df=20
t_desejado2=round(qt(prob_desejada2, df),4)
d_desejada2=dt(t_desejado2,df)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(-4, t_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(t_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(0, t_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(t_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``t'' da distribuição de Student") +
  labs(title= "Curva da função densidade \nDistribuição t (df=20)",
       subtitle = "P(-t; t)=(1-\u03b1) em cinza (nível de confiança) \nP(-\u221e; -t)= P(t;
       ↵ \u221e)= \u03b1/2 em vermelho ")+
  geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3)+
  geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="-t", angle=90, vjust=0,
           ↵ hjust=0, color="blue",size=6)+
  annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="t", angle=90, vjust=0,
           ↵ hjust=0, color="blue",size=6)+
  annotate(geom="text", x=t_desejado1-1.8, y=0.1, label="Intervalo aberto à esq.
           ↵ \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Intervalo aberto à dir.
           ↵ \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=t_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade=
           ↵ (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+ theme_bw()

```

Na Figura 9.21 observa-se:

Curva da função densidade

Distribuição t (df=20)

$P(-t; t) = (1 - ..)$ em cinza (nível de confiança)

$P(-...; -t) = P(t; ...) = .. / 2$ em vermelho

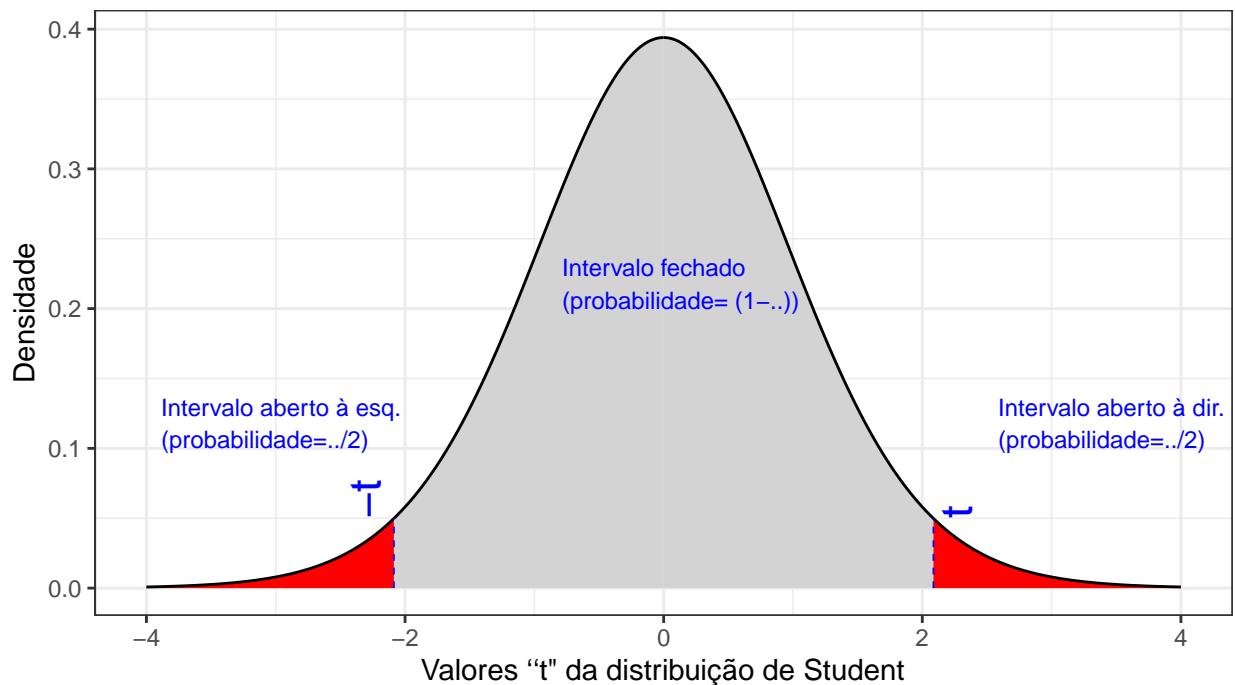


Figure 9.21: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores da estatística T ($(n - 1)$ graus de liberdade) é inferior a $\frac{\alpha}{2}$, estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

9.4. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS INDEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

- o nível de significância α ;
- o nível de confiança $(1 - \alpha)$; e,
- o valor tabelado da estatística $T(t)$ sob $(n_1 + n_2 - 2)$ graus de liberdade para o nível de confiança fixado.

Assim,

$$P\left[-T_{(n_1+n_2-2,1-\frac{\alpha}{2})} \leq T \leq T_{(n_1+n_2-2,1-\frac{\alpha}{2})}\right] = (1 - \alpha)$$

$$P\left[-t_{(n_1+n_2-2,1-\frac{\alpha}{2})} \leq \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{(n_1+n_2-2,1-\frac{\alpha}{2})}\right] = (1 - \alpha)$$

$$P[(\bar{x} - \bar{y}) - (t_{(n_1+n_2-2,1-\frac{\alpha}{2})} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (t_{(n_1+n_2-2,1-\frac{\alpha}{2})} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})] = (1 - \alpha)$$

$$IC(\mu_1 - \mu_2)_{(1-\alpha)} = [(\bar{x} - \bar{y}) \pm t_{c(n_1+n_2-2)} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}]$$

Exemplo: De uma grande turma extraiu-se uma pequena amostra de quatro notas de uma prova: 64, 66, 89, 77. De uma outra turma, extraiu-se uma outra amostra, independente, de três notas: 56, 71, 53. Se for razoável admitir que as variâncias das duas turmas (σ_1^2 e σ_2^2) sejam iguais, qual seria o intervalo de confiança para a diferença observada entre essas médias, a um nível de confiança de 95%?

Dados do problema:

- $\bar{X} = 74$ e $\bar{Y} = 60$ são as médias calculadas sobre as duas amostras (notas nas turmas);
- $S_1^2 = 132,67$ e $S_2^2 = 93$ são as variâncias calculadas sobre as duas amostras;
- $n_1 = 4$ e $n_2 = 3$ são os tamanhos das amostras;
- $n_1 + n_2 - 2 = 5$ são os graus de liberdade; e,
- $t = 2,57$ o valor tabelado da estatística para um nível de significância $\alpha = 5\%$ e graus de liberdade $gl = 5$.

$$P[(\bar{x} - \bar{y}) - (t_{(n_1+n_2-2, 1-\frac{\alpha}{2})} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (t_{(n_1+n_2-2, 1-\frac{\alpha}{2})} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})] = (1 - \alpha)$$

O desvio padrão ponderado S_p é dado por:

$$\begin{aligned} S_p &= \sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}} \\ S_p &= \sqrt{\frac{(4 - 1) \cdot 132,67 + (3 - 1) \cdot 93}{4 + 3 - 2}} \\ S_p &= 10,81 \end{aligned}$$

$$\begin{aligned} P[(\bar{x} - \bar{y}) - (t_{(n_1+n_2-2, 1-\frac{\alpha}{2})} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (t_{(n_1+n_2-2, 1-\frac{\alpha}{2})} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})] &= (1 - \alpha) \\ P[14 - (2,57 \cdot 10,81 \cdot \sqrt{\frac{1}{4} + \frac{1}{3}}) \leq (\mu_1 - \mu_2) \leq 14 + (2,57 \cdot 10,81 \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})] &= 0,95 \\ P[14 - 21,23 \leq (\mu_1 - \mu_2) \leq 14 + 21,23] &= 0,95 \end{aligned}$$

$$IC(\mu_1 - \mu_2)_{0,95} = [-7,23; 35,23]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que, se extraímos um grande número de amostras dessas mesmas dimensões das vendas dessa peça nas duas empresas, e para cada uma delas calcularmos suas médias e as diferenças entre elas, e calcularmos os intervalos de confiança como o acima definido, a proporção desses intervalos onde podemos encontrar a diferença das médias de vendas dessa peça da filial A para a filial B será de 0,95 (95 intervalos em 100).

De uma forma mais sintética podemos afirmar que o intervalo aleatório [-7,23; 35,23], é um intervalo de confiança a 95% para a diferença das médias das notas dessas provas nas duas turmas.

De uma forma mais corrente, *embora menos correta* em termos teóricos, é usual afirmar que, com 95% de confiança a diferença das médias das notas da primeira turma para a segunda turma se situa entre os valores -7,23 e 35,23.

9.4. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS INDEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

Uma importante conclusão pode ser extraída ao se analisar um pouco mais atentamente o intervalo calculado [-7,23 ; 35,23]. Vê-se que encontra-se dentro desse intervalo o valor 0 indicando que a diferença entre as médias amostrais pode ser zero sob esse nível de confiança, o que equivale dizer que sob esse nível de confiança não se pode afirmar existir diferença significativa (i.e. sob o nível de significância) entre as médias das notas dessas duas turmas.

9.4.3 Intervalos de confiança para a diferença entre duas médias amostrais com variâncias populacionais desconhecidas e desiguais

Se $(X_1, X_2, \dots, X_{n_1})$ e $(Y_1, Y_2, \dots, Y_{n_2})$ forem amostras aleatórias simples das populações X e Y com médias μ_1 e μ_2 , e variâncias σ_1^2 e σ_2^2 desconhecidas porém iguais ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), e $\bar{X} = \frac{(X_1 + X_2 + \dots + X_{n_1})}{n_1}$ e $\bar{Y} = \frac{(Y_1 + Y_2 + \dots + Y_{n_2})}{n_2}$, então:

$$X \sim N(\mu_1, \frac{\sigma}{\sqrt{n_1}})$$

$$Y \sim N(\mu_2, \frac{\sigma}{\sqrt{n_2}})$$

Demonstra-se que a estatística T pode ser assim definida, bem como sua correspondente distribuição (cf. Figura ??):

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu$$

em que:

- \bar{X} e \bar{Y} são as médias das amostras extraídas;
- μ_1 e μ_2 são as médias populacionais;
- n_1 e n_2 são os tamanhos das amostras; e,
- S_1^2 e S_2^2 são as variâncias das amostras.

O número de graus de liberdade (ν) é dado por:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

```

alfa=0.05

prob_desejada1=alfa/2
df=20
t_desejado1=round(qt(prob_desejada1,df ),4)
d_desejada1=dt(t_desejado1,df)

prob_desejada2=1-alfa/2
df=20
t_desejado2=round(qt(prob_desejada2, df),4)
d_desejada2=dt(t_desejado2,df)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(-4, t_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(t_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(0, t_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(t_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores ``t'' da distribuição de Student") +
  labs(title= "Curva da função densidade \nDistribuição t (df=20)",
       subtitle = "P(-t; t)=(1-\u03b1) em cinza (nível de confiança) \nP(-\u221e; -t)= P(t;
                  \u221e)= \u03b1/2 em vermelho ")+
  geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3)+
  geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="-t", angle=90, vjust=0,
           hjust=0, color="blue",size=6)+
```

9.4. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS INDEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

```
annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="t", angle=90, vjust=0,
        hjust=0, color="blue",size=6)+
annotate(geom="text", x=t_desejado1-1.8, y=0.1, label="Intervalo aberto à esq.
        \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Intervalo aberto à dir.
        \n(probabilidade=\u03b1/2)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=t_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade=
        (1-\u03b1))", angle=0, vjust=0, hjust=0, color="blue",size=3)+ theme_bw()
```

Curva da função densidade

Distribuição t (df=20)

$P(-t; t) = (1 - \alpha)$ em cinza (nível de confiança)
 $P(-\dots; -t) = P(t; \dots) = \dots/2$ em vermelho

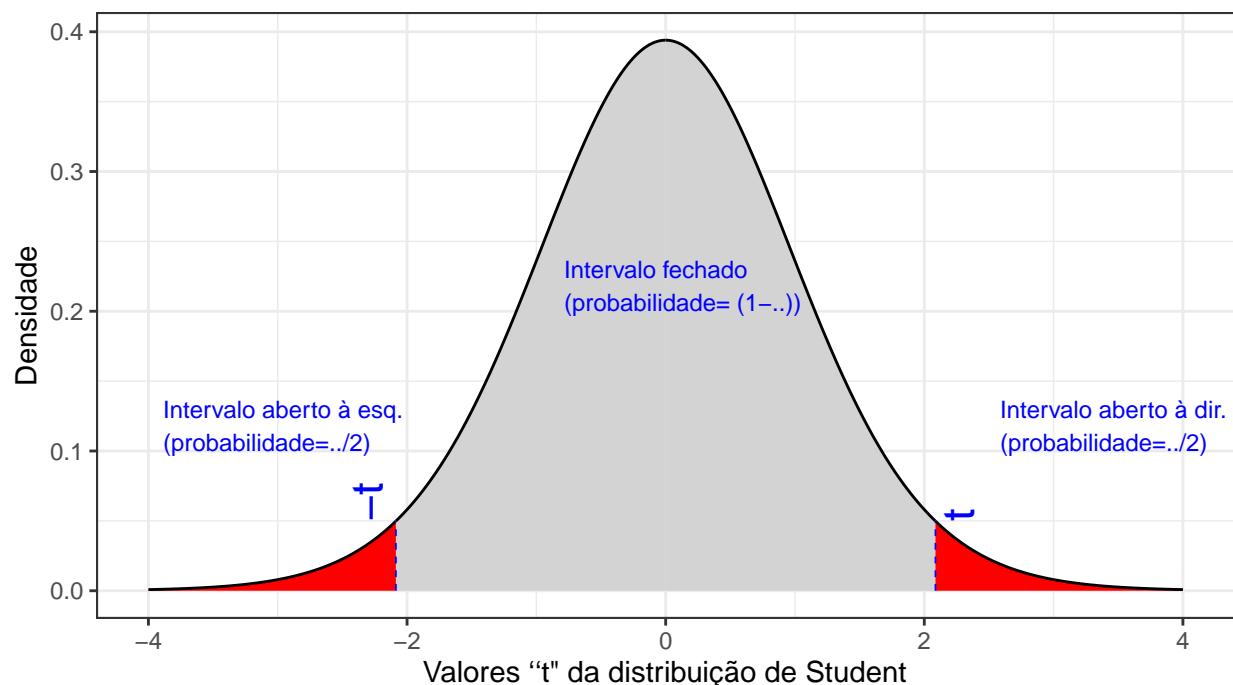


Figure 9.22: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores da estatística T (com ν graus de liberdade) é inferior a $\frac{\alpha}{2}$, estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

Na Figura 9.22 observa-se:

- o nível de significância α ;
- o nível de confiança $(1 - \alpha)$; e,
- o valor tabelado da estatística $T(t)$ sob ν graus de liberdade para o nível de confiança fixado.

410 MÓDULO 9. INTRODUÇÃO À DISTRIBUIÇÃO DAS MÉDIAS E DIFERENÇAS ENTRE MÉDIAS AMOSTRAIS E SISTÉMATICAS

Assim,

$$P\left[-T_{(\nu, 1-\frac{\alpha}{2})} \leq T \leq T_{(\nu, 1-\frac{\alpha}{2})}\right] = (1 - \alpha)$$

$$P\left[-t_{(\nu, 1-\frac{\alpha}{2})} \leq \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}} \leq t_{(\nu, 1-\frac{\alpha}{2})}\right] = (1 - \alpha)$$

$$P[(\bar{x} - \bar{y}) - (t_{(\nu, 1-\frac{\alpha}{2})} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (t_{(\nu, 1-\frac{\alpha}{2})} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}})] = (1 - \alpha)$$

$$IC(\mu_1 - \mu_2)_{(1-\alpha)} = [(\bar{x} - \bar{y}) \pm t_{c(\nu)} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}}]$$

Exemplo: De uma pequena classe do curso de ensino médio tomou-se uma amostra de 4 provas de matemática, obtendo-se um valor médio de 81 sob uma variância de 2. Outra amostra, de 6 provas de biologia, forneceu um valor médio de 77 sob uma variância de 14,4. Qual seria o intervalo de confiança para a diferença observada entre essas médias, sob um nível de confiança de 95%?

Dados do problema:

Dados do problema:

- $\bar{X} = 81$ e $\bar{Y} = 77$ são as médias calculadas sobre as duas amostras (notas nas turmas);
- $S_1^2 = 2$ e $S_2^2 = 14,40$ são as variâncias calculadas sobre as duas amostras; e,
- $n_1 = 4$ e $n_2 = 6$ são os tamanhos das amostras.

O número de graus de liberdade (ν) é dado por:

9.4. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS INDEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

$$\begin{aligned}\nu &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} \\ \nu &= \frac{\left(\frac{2}{4} + \frac{14,40}{6}\right)^2}{\frac{\left(\frac{2}{4}\right)^2}{4-1} + \frac{\left(\frac{14,40}{6}\right)^2}{6-1}} \\ \nu &= \frac{2,90^2}{0,083 + 1,152} \\ \nu &= \frac{8,41}{1,23} = 6,83 \sim 7\end{aligned}$$

Portanto, $t = 2,36$ é o valor tabelado da estatística para um nível de significância $\alpha = 5\%$ e graus de liberdade $gl = 7$.

$$\begin{aligned}P[(\bar{x} - \bar{y}) - (t_{(\nu, 1-\frac{\alpha}{2})} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (t_{(\nu, 1-\frac{\alpha}{2})} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})] &= (1 - \alpha) \\ P[4 - (2,36 \cdot \sqrt{\frac{2}{4} + \frac{14,40}{6}}) \leq (\mu_1 - \mu_2) \leq 4 + (2,36 \cdot \sqrt{\frac{2}{4} + \frac{14,40}{6}})] &= 0,95 \\ P[4 - (2,36 \cdot 1,70) \leq (\mu_1 - \mu_2) \leq 4 + (2,36 \cdot 1,70)] &= 0,95 \\ P[4 - 4,01 \leq (\mu_1 - \mu_2) \leq 4 + 4,01] &= 0,95\end{aligned}$$

$$IC(\mu_1 - \mu_2)_{0,95} = [-0,01; 8,01]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que, se extrairmos um grande número de amostras dessas mesmas dimensões das notas dessas provas nas duas turmas, e para cada uma delas calcularmos suas médias e as diferenças entre elas, e calcularmos os intervalos de confiança como o acima definido, a proporção desses intervalos onde podemos encontrar a diferença das notas das provas de matemática para a prova de biologia será de 0,95 (95 intervalos em 100).

De uma forma mais sintética podemos afirmar que, o anterior intervalo aleatório $[-0,01; 8,01]$, é um intervalo de confiança a 95% para a diferença das médias das notas dessas provas nas duas turmas.

De uma forma mais corrente, *embora menos correta* em termos teóricos, é usual afirmar que, com 95% de confiança a diferença das médias das notas da prova de matemática para a prova de biologia situa entre os valores -0,01 e 8,01.

Uma importante conclusão pode ser extraída ao se analisar um pouco mais atentamente o intervalo calculado [-0,01 ; 8,01]. Vê-se que encontra-se dentro desse intervalo o valor 0 indicando que a diferença entre as médias amostrais pode ser zero sob esse nível de confiança, o que equivale dizer que sob esse nível de confiança não se pode afirmar existir diferença significativa (i.e. sob o nível de significância) entre as médias dessas notas.

9.5 Distribuição das diferenças de médias amostrais dependentes e seus intervalos de confiança

Na prática temos algumas situações onde as populações não são independentes com, por exemplo, em situações onde as amostras são extraídas de uma mesma população em dois momentos distintos (antes e depois de algum fato), ou como numa situação de comparação inter laboratorial, onde dois laboratórios medem a mesma peça, as medidas entre os laboratórios não são independentes. Nestes casos diz-se que os dados são pareados.

Considere duas amostras dependentes (X_1, \dots, X_n) e (Y_1, \dots, Y_n) . O pareamento das observações será considerado tomando-se $(X_1, Y_1), \dots, (X_n, Y_n)$ e as diferenças serão tomadas a cada par $D_i = X_i - Y_i$, para $i = 1, \dots, n$.

Assim obtemos uma amostra (D_1, \dots, D_n) , resultante das diferenças entre os valores de cada par. A variável aleatória será admitida tal que

$$D \sim N(\mu_D, \sigma_D^2)$$

O parâmetro da média dessa distribuição (μ_D) será estimado a partir da própria amostra das diferenças, tal que:

$$\mu_D = \bar{D} = \sum_{i=1}^n D_i$$

e a variância populacional desconhecida será aproximada por:

$$S_D^2 = \sum_{i=1}^n \frac{(D_i - \bar{D})^2}{n-1}$$

9.5. DISTRIBUIÇÃO DAS DIFERENÇAS DE MÉDIAS AMOSTRAIS DEPENDENTES E SEUS INTERVALOS DE CONFIANÇA

Demonstra-se que a estatística T pode ser assim definida, bem como sua correspondente distribuição

$$T = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} \sim t_{(n-1)}$$

Assim,

$$IC(\mu_D)_{(1-\alpha)} = [\bar{D} \pm t_{c(n-1)} \cdot \sqrt{\frac{S_D^2}{n}}]$$

Exemplo: Determinar o intervalo de confiança sob um nível de confiança de 95% para a diferença de médias do resultados dos testes de um grupo de 15 alunos submetidos a um vídeo instrutivo tais que a primeira amostra foi tomada antes de assistirem ao vídeo e a segunda depois, mediante a aplicação de um novo teste, similar ao primeiro.

Aluno	Primeira nota (X)	Segunda nota (Y)
1	74	80
2	64	74
3	79	83
4	90	92
5	89	96
6	94	98
7	55	59
8	75	77
9	88	93
10	66	78
11	70	75
12	60	59
13	59	61
14	67	70
15	69	74

$$\bar{D} = \sum_{i=1}^n D_i = -4,667$$

$$S_D^2 = \sum_{i=1}^n \frac{(D_i - \bar{D})^2}{n-1} = 10,52354$$

Sendo o valor crítico tabelado da estatística para um nível de significância $\alpha = 5\%$ e graus de liberdade $gl = (n - 1) = 14$ igual a 1,761, o intervalo de confiança será:

$$\begin{aligned} IC(\mu_D)_{(1-\alpha)} &= [\bar{D} \pm t_{c(n-1)} \cdot \sqrt{\frac{S_D^2}{n}}] \\ IC(\mu_D)_{(1-\alpha)} &= [-4,667 \pm 1,761 \cdot \sqrt{\frac{10,52354}{15}}] \\ IC(\mu_D)_{(1-\alpha)} &= [-5,396; -3,937] \end{aligned}$$

Sendo negativos os valores desse intervalo de confiança deduz-se que a **primeira nota** é menor que a **segunda nota** ($X - Y < 0$) e assim, o vídeo que os alunos assistiram melhorou sua compreensão do assunto e seu desempenho no segundo teste (similar ao primeiro). Caso o valor “zero” estivesse contemplado nesse intervalo, a interpretação seria de que não há diferença estatisticamente significativa nas notas dos alunos nos dois testes (o vídeo não os ajudou em coisa alguma).

Módulo 10

Introdução à distribuição das proporções amostrais e seus intervalos de confiança

A finalidade de uma amostra reside em obter uma estimativa do valor de um ou mais parâmetros associados a uma população. Verifica-se que, ao se extrair repetidamente valores amostrais de forma aleatória da mesma população, estes variam de uma amostra para outra, assim como em relação ao verdadeiro parâmetro dessa população. No entanto, é possível demonstrar que essa variabilidade pode ser caracterizada por meio de distribuições de probabilidade.

Quando utilizadas com esse propósito, essas distribuições de probabilidade são chamadas de distribuições amostrais. Elas permitem avaliar, para cada amostra, quão próximo está o valor da estatística amostral em relação ao verdadeiro parâmetro da população. A resposta a essa questão depende essencialmente de três fatores:

- A estatística específica que está sendo empregada: diferentes estatísticas demandam diferentes distribuições de probabilidade para modelar sua variabilidade.
- O tamanho da amostra, que exerce uma influência inversa na variabilidade entre os valores amostrais.
- A variabilidade intrínseca da população em estudo e do processo de amostragem.

10.1 Conceito elementar de uma proporção

O conceito básico de proporção remete à razão entre duas grandezas. Vejam os exemplos:

- segundo dados demográficos de 2012 (IBGE), a cidade de Recife possui proporcionalmente mais mulheres que homens;

- em 18 dias de campanha, somente 25,09% do público-alvo se vacinou contra gripe no País, segundo dados divulgados pelo Ministério da Saúde. De 17 de abril, quando a imunização foi iniciada, até 5 de maio, 13,6 milhões de brasileiros procuraram os postos de saúde para se vacinar.

Na primeira afirmação, a ideia de proporcionalidade advém do quociente do número habitantes do sexo feminino pelo numero total de habitantes naquele ano ($\frac{827.885}{1.537.704} = 0,5384$). Já na segunda, a afirmação resulta do quociente do número de brasileiros vacinados pelo total da população-alvo ($\frac{13.600.000}{54.200.000} = 0,2509$).

10.2 Distribuição das proporções amostrais

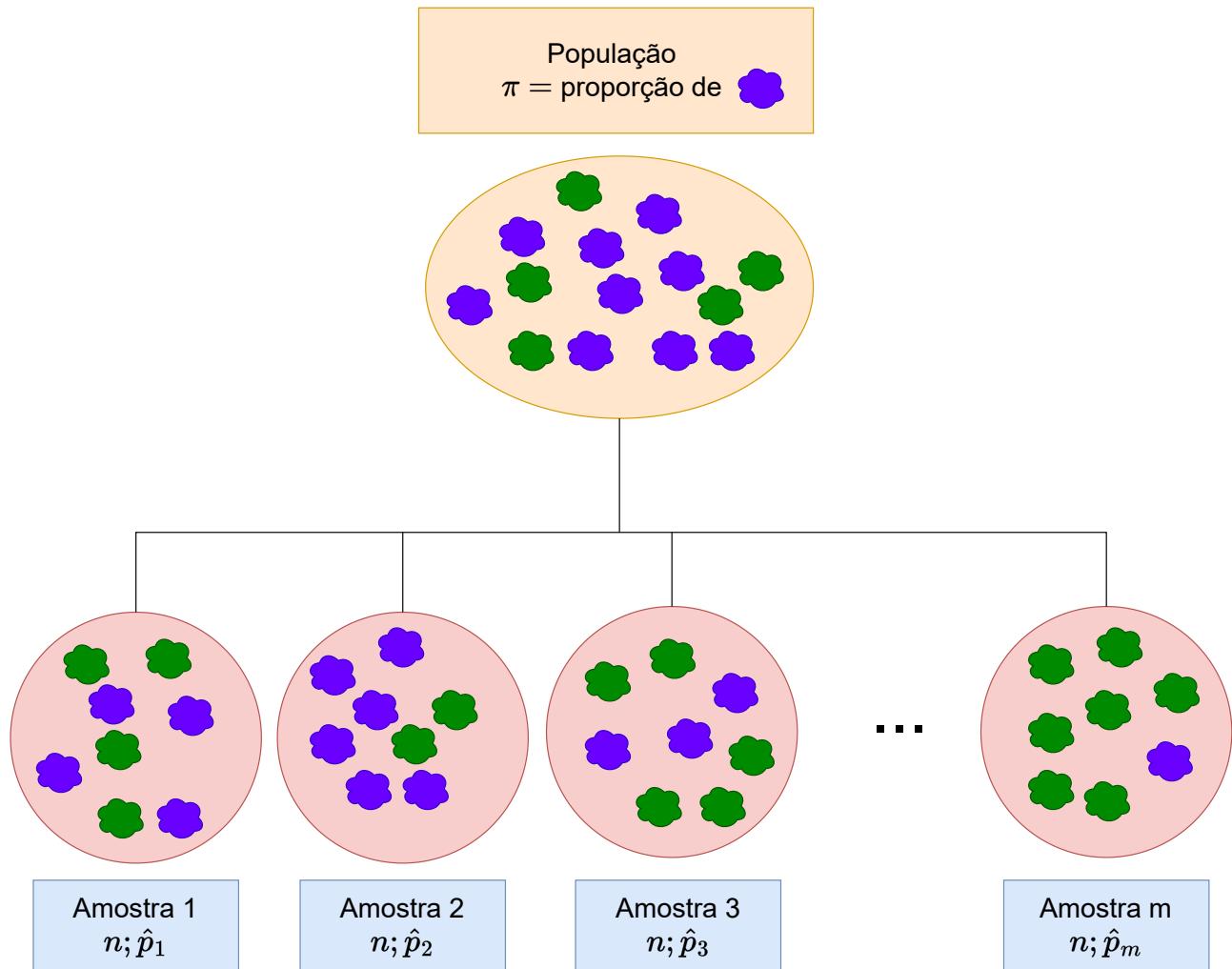


Figure 10.1: Ilustração de m amostras de mesmo tamanho (n) extraídas de uma mesma população onde a característica de interesse se manifesta sob uma proporção populacional π

Para estudarmos a distribuição das proporções amostrais (\hat{p}) considerem uma população apresentando uma determinada característica de interesse com proporção π . Essa característica de interesse assume apenas duas possibilidades em cada elemento da população: ela **pode ou não** estar presente:

$$X_i = \begin{cases} 1, & \text{se o i-ésimo elemento é portador da característica} \\ 0, & \text{se o i-ésimo elemento não é portador da característica} \end{cases}$$

Assim, ao se escolher ao acaso um elemento da população, a probabilidade dessa característica estar presente pode ser estimada seguindo o modelo teórico de uma variável de *Bernoulli* e assim $X_i \sim Ber(\pi)$ e, como tal, $E(X) = \pi$ e $Var(X) = \pi(1 - \pi)$.

Repetindo-se essa “extração” por n vezes podemos definir a variável aleatória Y_n como sendo o número de sucessos observados em n repetições de *Bernoulli*:

$$Y_n = X_1 + \cdots + X_n$$

e assim, $Y_n \sim Bin(n, \pi)$ e a proporção amostral observada de sucessos ao final das n repetições será a média:

$$\hat{p} = \frac{Y}{n} = \frac{X_1 + \cdots + X_n}{n}$$

em que \hat{p} é uma estimativa amostral da proporção populacional π .

Demonstra-se que para:

- um razoável número de repetições: $n \geq 30$;
- de uma população onde a proporção π não é extrema: próximas a 0 ou 1; e tal que $(n \cdot \pi)$ e $(n \cdot (1 - \pi))$ sejam maiores que 15 (alguns autores consideram limites mais brandos, iguais a 10 ou ainda a 5),

ao se repetir o experimento anotando-se as proporções amostrais \hat{p} obtida em cada uma das n repetições de *Bernoulli*, o perfil da curva de distribuição dessas proporções amostrais torna-se razoavelmente simétrico à medida que o número n de repetições de *Bernoulli* cresce, para qualquer que seja a proporção populacional, e oscila em torno de π .

Pelo Teorema de *DeMoivre e Laplace* (anteriores ao Teorema do Limite Central), demonstra-se que, para um grande número de repetições (n), o *valor esperado* e a *variância* das proporções amostrais são:

$$\begin{aligned}E(Y) &= n \cdot \pi \\Var(Y) &= n \cdot \pi \cdot (1 - \pi)\end{aligned}$$

e a distribuição das proporções amostrais será aproximadamente Normal com parâmetros $\mu = n \cdot \pi$ e $\sigma^2 = n \cdot \pi \cdot (1 - \pi)$:

$$Y \sim N(n \cdot \pi; n \cdot \pi \cdot (1 - \pi))$$

Uma vez que a proporção amostral está definida como: $\hat{p} = \frac{Y_n}{n}$ segue-se que o valor esperado $\hat{p} = \mu$:

$$\begin{aligned}E(\hat{p}) &= E\left(\frac{Y}{n}\right) \\&= \frac{1}{n} \cdot E(Y) \\&= \frac{1}{n} \cdot n \cdot \pi \\&= \pi\end{aligned}$$

e a variância $Var(\hat{p} = \frac{1}{n} \cdot \pi \cdot (1 - \pi))$:

$$\begin{aligned}
 Var(\hat{p}) &= Var\left(\frac{Y}{n}\right) \\
 &= \frac{1}{n^2} \cdot Var(Y) \\
 &= \frac{1}{n^2} \cdot n \cdot \pi \cdot (1 - \pi) \\
 &= \frac{1}{n} \cdot \pi \cdot (1 - \pi)
 \end{aligned}$$

Assim, as proporções amostrais se distribuem de modo aproximadamente Normal sob uma média $\mu = \pi$ e com uma variância $\sigma^2 = \frac{\pi \cdot (1 - \pi)}{n}$:

$$\hat{p} \sim N\left(\pi; \frac{\pi \cdot (1 - \pi)}{n}\right)$$

10.2.1 Simulações ilustrativas da aproximação da distribuição das proporções amostrais pela distribuição Normal

Para exemplificar considere o lançamento de um dado de seis faces,. A probabilidade de que uma certa face caia voltada para cima é de $\frac{1}{6} = 0,167$. Se lançarmos esse dado um número crescente de vezes e anotarmos a proporção delas em que a face escolhida caiu voltada para cima comprova-se que o valor esperado das proporções amostrais aproxima-se da proporção populacional.

As Figuras 10.2 (tamanho de cada amostra $n = n_1$) e 10.3 (tamanho de cada amostra $n = n_2$) mostram o perfil assumido pela distribuição de 100 proporções amostrais obtidas de uma população que apresenta uma proporção $\pi = p_1$ da característica de interesse.

```

#####
# Considere uma população cuja característica de interesse (A) se manifesta de modo
# dicotômico:
# sim/não, sob uma probabilidade p_1 e (1-p_1).
# A probabilidade de se obter um elemento com a característica de interesse
# - ao se sortear aleatoriamente um indivíduo qualquer - pode ser modelada como uma variável
# de Bernoulli.
# A probabilidade de se observar a característica de interesse ao se
# repetir a amostragem (com reposição) por n_1 (n_2) vezes pode ser modelada como uma
# variável binomial (repetição de um experimento de Bernoulli n_1/n_2 vezes)
# Repetindo-se esses experimentos binomiais por N vezes, as proporções amostrais de
# elementos com a característica de interesse (sucesso) nas N amostras obtidas será

```

420MÓDULO 10. INTRODUÇÃO À DISTRIBUIÇÃO DAS PROPORÇÕES AMOSTRAIS E SEUS INTERVALOS DE CONFIDÊNCIA

```

# dada pelo número de elementos de cada conjunto nas n_1 (n_2) repetições dividido por
# n_1 (n_2).
# Desse modo, obtemos N proporções de amostras de tamanho n_1 (n_2)
#
#
# Selecionando-se aleatoriamente um elemento desta população
# resulta em uma variável aleatória dicotômica/Bernoulli que assume
# o valor 1 caso o elemento selecionado possua a propriedade A (sucesso)
# e assume o valor 0 caso não possua a propriedade A.
#
# A retirada (com reposição) de `n_1` elementos dessa população poderemos observar a
# frequência absoluta com que a propriedade A (sucesso) se manifesta na amostra,
# a qual pode ser expressa como uma variável aleatória (X) que segue o modelo teórico
# Binomial de probabilidade.
#
# A frequência relativa, o quociente entre o número de sucessos por `n_1` expressa a
# proporção com que a propriedade "A" foi observada na 'amostra' de tamanho `n_1` é também
# uma variável aleatória (p) com distribuição altamente relacionada à variável X pois é a
# média de `n_1` ensaios (repetições) de Bernoulli.
#
# Repetindo-se sucessivamente `N` vezes extrações de tamanho `n_1`
# a anotando-se a proporção de sucesso em cada uma dessas amostras poderemos analisar como
# eles se distribuem em relação à quantidade de elementos extraídos `n_1` (repetições de
# Bernoulli)
# e à verdadeira proporção com que a propriedade A se manifesta na população (pi)
#
# Demonstra-se que:
# para `n_1` suficientemente grande (repetições de Bernoulli com reposição);]
# n_1 * pi > 5 e
# n_1*(1-pi) < 5
# a distribuição de p pode ser aproximada pela distribuição Normal
# tal que p ~N(mu,sigma)
# onde mu e sigma são aproximados por:
# mu = E(p) = pi
# sigma^2 = sigma^2*p >>> sigma = sqrt[ p*(1-p)/(n_1) ]
#
#####
#
# Proporção escolhida para a manifestação da característica: sim/não (probabilidade de cada
# evento de Bernoulli)
p_1=round(1/6,2)

# Número de amostras
N=100

# Tamanho escolhido para cada amostra: repetições de Bernoulli
n_1=10

# Vetor com o número de sucessos observados (a frequência absoluta) nas N amostras de n_1
# elementos dicotómicos (repetições de Bernoulli, sob uma probabilidade individual de
# sucesso igual a p_1)

suc_10rep=rbinom(n=N, size = n_1, prob = p_1)
suc_10rep

```

```

# Vendo a proporção de sucessos (a frequência relativa) em cada uma das  $N_1$  amostras de  $n_1$ 
# → elementos dicotômicos
prop_10rep=suc_10rep/n_1
mean(prop_10rep) # ~  $\pi$ 
sd(prop_10rep) # ~  $\sqrt{\pi * (1 - \pi) / n_1}$ 

# Dataframe com as  $N$  proporções amostrais sob  $n_1$ 
dados_10=as.data.frame(prop_10rep)

#####
# O mesmo procedimento, mas agora com amostras com um maior número de elementos em cada uma
#####

# Tamanho escolhido para cada amostra: repetições de Bernoulli
n_2=100

# Vetor com o número de sucessos observados (a frequência absoluta) nas  $N$  amostras de  $n_2$ 
# → elementos dicotômicos (repetições de Bernoulli, sob uma probabilidade individual de
# → sucesso igual a  $p_1$ )
suc_100rep=rbinom(n=N, size = n_2, prob = p_1)
suc_100rep

# Vendo a proporção de sucessos (a frequência relativa) em cada uma das  $N_1$  amostras de  $n_1$ 
# → elementos dicotômicos
prop_100rep=suc_100rep/n_2
mean(prop_100rep) # ~  $\pi$ 
sd(prop_100rep) # ~  $\sqrt{\pi * (1 - \pi) / n_2}$ 

# Dataframe com as  $N$  proporções amostrais sob  $n_2$ 
dados_100=as.data.frame(prop_100rep)

```

```

meu_titulo1=paste("Distribuição de frequência das proporções de sucesso observadas em \n",N,
# → "amostras de n=", n_1, "elementos dicotômicos extraídos (com reposição) da
# → população", "\n(proporção de sucesso na população \u03c0=", p_1,")")
meu_titulo2=paste("As proporções amostrais ~ \nN(\u03bc=
# → \u03c0=",round(mean(dados_10$prop_10rep),3)," ; \u03c3 =sqrt(\u03c0*(1-
# → \u03c0)/n)=",round(sd(dados_10$prop_10rep),3)," )")

ggplot(dados_10, aes(x = prop_10rep)) +
  geom_histogram(aes(y =..density..),
                 breaks = seq(0, 0.4, by = 0.05),
                 colour = "black",
                 fill = "lightblue") +
  stat_function(fun = dnorm,
                args = list(mean = p_1, sd = sqrt(p_1*(1-p_1)/n_1)),
                colour="red") +
  scale_y_continuous(name="",breaks = NULL) +
  scale_x_continuous(name="Valores das proporções amostrais") +
  #labs(title=meu_titulo1)+
```

```
annotate(geom="text", x=mean(prop_10rep), y=max(dnorm(prop_10rep)),
         label=meu_titulo2, angle=0, vjust=0, hjust=0, color="blue", size=4) +
theme(plot.title = element_text(size = 10, face = "bold"),
      axis.text.x = element_text(angle=0, hjust=1, size=10),
      axis.text.y = element_text(angle=0, hjust=1, size=10),
      axis.title.x = element_text(size = 10),
      axis.title.y = element_text(size = 10))
```

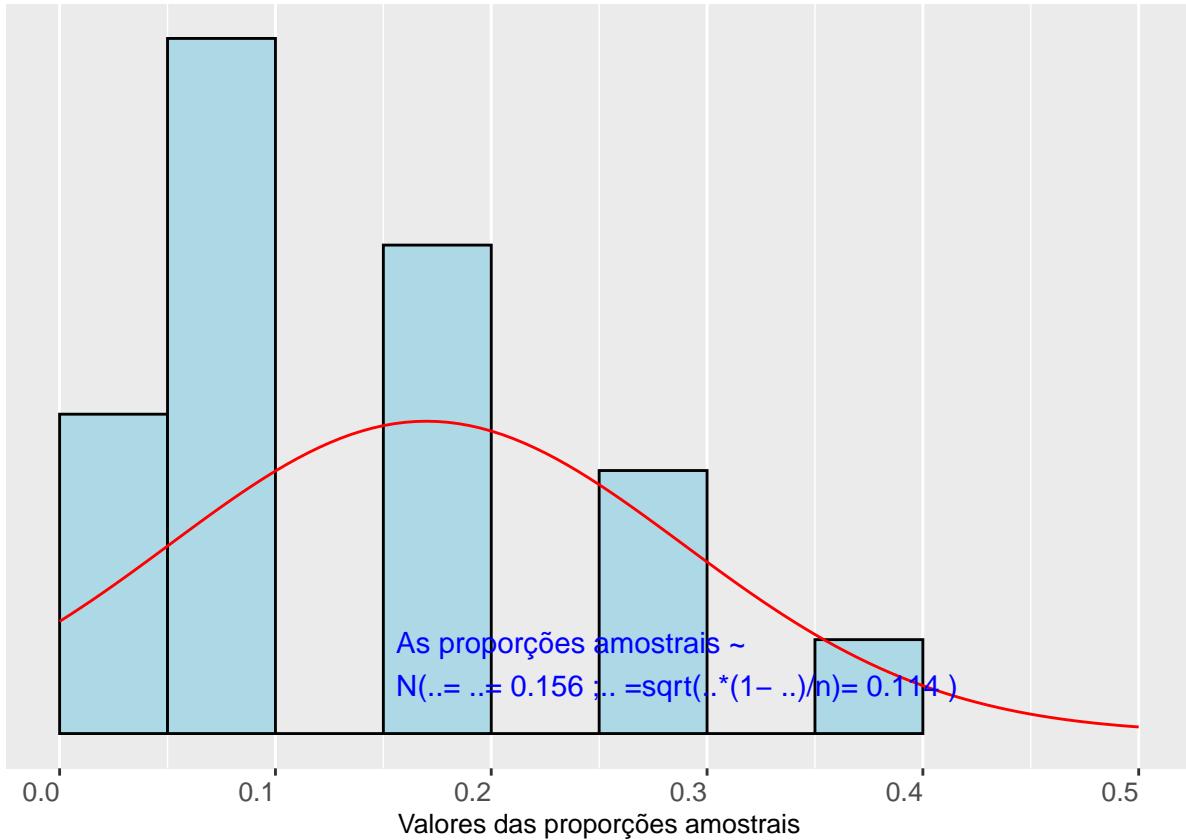


Figure 10.2: Distribuição das frequências das proporções de sucesso observadas em 100 amostras de tamanho $n=10$ elementos dicotômicos extraídos (com reposição) de uma população (a proporção de sucesso na população é $=1/6$)

```
meu_titulo1=paste("Distribuição de frequências das proporções de sucesso observadas em",
                   "\n",N, "amostras de n=", n_2, "elementos dicotômicos extraídos (com reposição) da",
                   "população", "\n(proporção de sucesso na população \u03c0=", p_1,")")

meu_titulo2=paste("As proporções amostrais ~ \nN(\u03bc=",
                   "\u03c0",",round(mean(dados_100$prop_100rep),3),";\u03c3=sqrt(\u03c0*(1-
                   "\u03c0)/n)",",round(sd(dados_100$prop_100rep),3),")")

ggplot(dados_100, aes(x = prop_100rep)) +
  geom_histogram(aes(y = ..density..),
                 breaks = seq(0, 0.4, by = 0.03),
                 colour = "black",
                 fill = "lightblue") +
```

```

stat_function(fun = dnorm,
              args = list(mean = p_1,
                          sd = sqrt(p_1*(1-p_1)/n_2)),
              colour="red") +
scale_y_continuous(name="",breaks = NULL) +
scale_x_continuous(name="Valores das proporções amostrais") +
#labs(title=meu_titulo1) +
annotate(geom="text", x=mean(prop_100rep), y=max(dnorm(prop_100rep)),
         label=meu_titulo2, angle=0, vjust=0, hjust=0, color="blue",size=4) +
theme(plot.title = element_text(size = 10, face = "bold"),
      axis.text.x = element_text(angle=0, hjust=1, size=10),
      axis.text.y = element_text(angle=0, hjust=1, size=10),
      axis.title.x = element_text(size = 10),
      axis.title.y = element_text(size = 10))

```

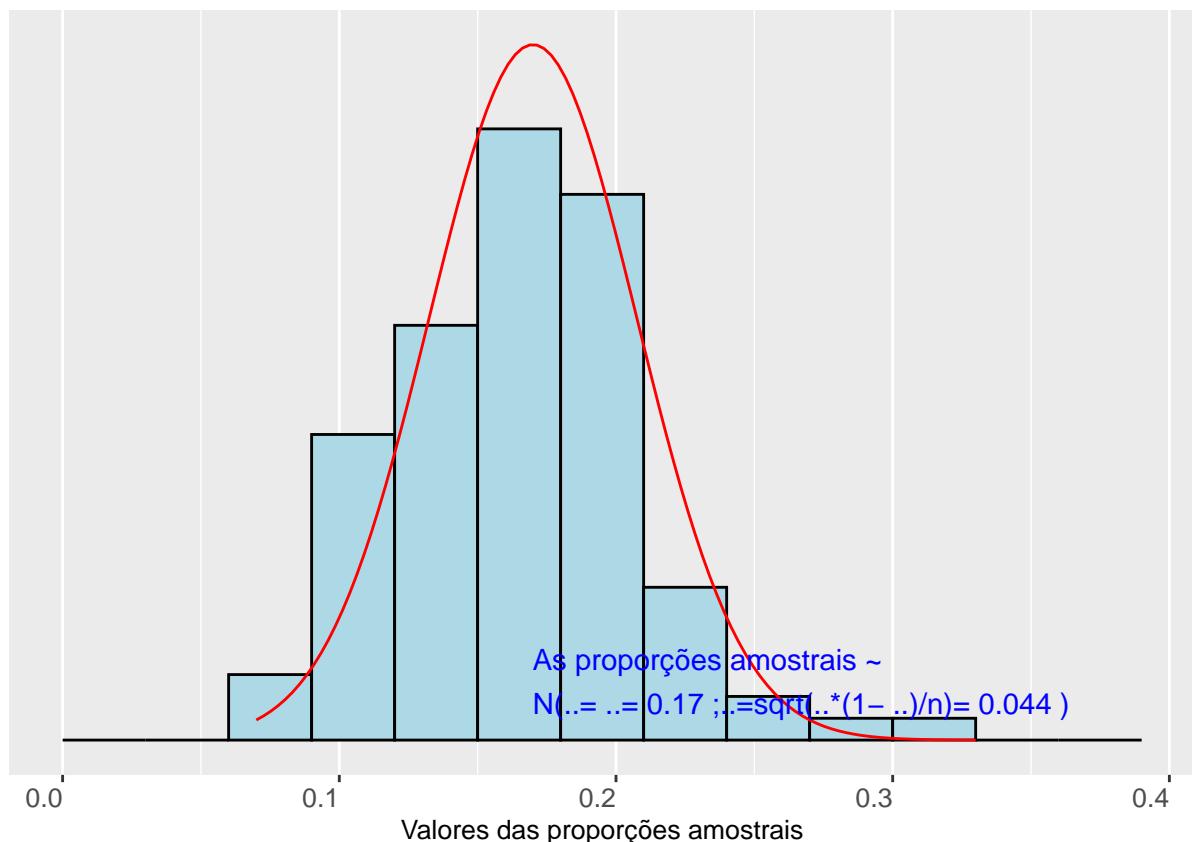


Figure 10.3: Distribuição das frequências das proporções de sucesso observadas em 100 amostras de tamanho $n=100$ elementos dicotômicos extraídos (com reposição) de uma população (a proporção de sucesso na população é $=1/6$)

10.3 Pobabilidades associadas à observação de uma proporção amostral \hat{p}

Ao se definir a estatística Z como a simples padronização da variável \hat{p} vemos que esta seguirá uma distribuição normal com média 0 e desvio-padrão 1:

$$Z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1)$$

Essa aproximação da distribuição de uma variável binomial (proporções amostrais \hat{p}) pela distribuição Normal será tanto mais simétrica e com perfil de um sino quanto vier a atender (n grande e π não próximo de 0 ou 1) e nos permite determinar probabilidades associadas a proporções amostrais.

Exemplo: um sistema de produção opera de tal maneira que 10% das peças produzidas são defeituosas. Suponha que os itens sejam vendidos em caixas com 100 unidades e calcule as probabilidades de que em uma caixa: - tenha mais do que 10% de defeituosas? - tenha menos do que 15% de defeituosas?

Dados do problema: $\pi = 0,10$ e $n = 100$.

Considerando que a proporção populacional π não é extrema (próxima a 0 ou 1) e $(n \cdot \pi)$ e $(n \cdot (1 - \pi))$ são maiores que 5, as proporções amostrais \hat{p} se distribuem, aproximadamente, do modo:

$$\hat{p} \sim N\left(\mu : \pi; \sigma^2 : \frac{\pi \cdot (1 - \pi)}{n}\right) \hat{p} \sim N\left(0, 10; \frac{0, 10 \cdot (1 - 0, 10)}{100}\right) \hat{p} \sim N(0, 10; 0, 0009)$$

Para se calcular as probabilidades de serem observadas proporções amostrais $\hat{p} > 0,10$ e $\hat{p} < 0,15$, basta-se mapear essas proporções amostrais à distribuição Normal padronizada. Assim, denotando-se uma variável aleatória (as proporções amostrais) $X \sim n(\mu : 0, 1; \sigma^2 : 0, 0009(\sigma : 0, 03))$ segue-se:

$$\begin{aligned}
P(\hat{p} > 0,10) &= P(X > 0,10) \\
&= P\left(\frac{X - 0,10}{0,03} > \frac{0,10 - 0,10}{0,03}\right) \\
&= P(Z > 0) \\
&= 0,50
\end{aligned}$$

e

$$\begin{aligned}
P(\hat{p} < 0,15) &= P(X < 0,15) \\
&= P\left(\frac{X - 0,15}{0,03} < \frac{0,15 - 0,10}{0,03}\right) \\
&= P(Z < 1,67) \\
&= 0,9525
\end{aligned}$$

10.4 A aleatoriedade das proporções amostrais e o tamanho amostral

No módulo ‘‘Introdução ao planejamento de pesquisas’’ explicamos que quando não se dispõe de nenhuma informação *a priori* sobre a proporção populacional (π) a adoção do máximo valor possível ao produto: $\pi.(1 - \pi) = \frac{1}{4}$ assegura que o o tamanho de amostra obtido será suficiente para a estimativa qualquer que seja a proporção populacional π . Trazendo a variável Z antes definida:

$$Z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1)$$

podemos reescrevê-la de modo a se obter o dimensionamento amostral em função do nível de confiança e um erro máximo estabelecidos:

$$z_{(1-\alpha)} = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} z_{(1-\alpha)} \cdot \sqrt{\frac{\pi(1-\pi)}{n}} = \hat{p} - \pi \frac{\pi(1-\pi)}{n} = \left(\frac{\varepsilon}{z_{(1-\alpha)}}\right)^2 n = \frac{z_{(1-\alpha)}^2}{\varepsilon^2} \cdot \pi(1-\pi)$$

Deste modo podemos simular a flutuação dos valores das proporções obtidas em sucessivas amostras, ilustrando simultaneamente as proporções amostrais observadas e a proporção das amostras que apresentam um erro amostral (ε) superior ao estipulado pelo nível de confiança ($1 - \alpha$).

Desconhecendo-se qualquer informação acerca da proporção populacional (π), a dimensão da amostra pode ser estipulada tomando-se o maior valor do produto $\pi(1 - \pi)$ como sendo igual a $\frac{1}{4}$ pois:

```
p <- seq(0, 1, by = 0.01)
y <- p * (1 - p)
plot(p, y, type = "l", xlab = "\u03c0", ylab = "\u03c0*(1- \u03c0)", main = "Possíveis
valores assumidos pelo produto: \u03c0*(1- \u03c0)")
```

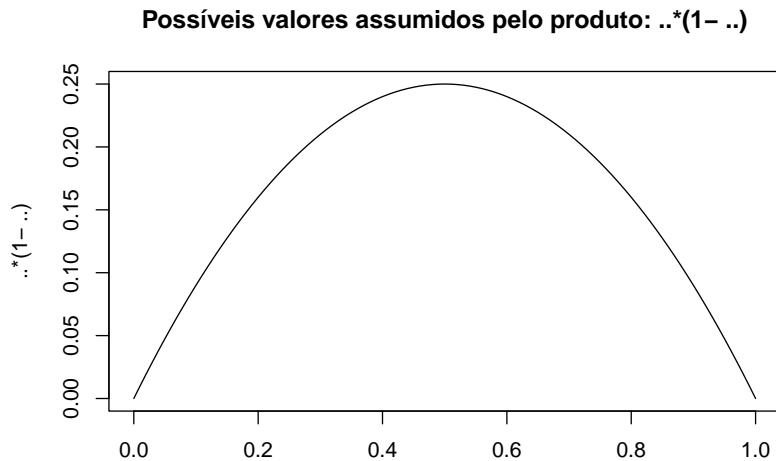


Figure 10.4: Possíveis valores assumidos pelo produto: $\pi(1 - \pi)$

Assim, a dimensão conservadora para a amostra será dada por:

$$n = \frac{z_{(1-\alpha)}^2}{\varepsilon^2} \cdot \frac{1}{4}$$

10.4.1 Simulações ilustrativas sobre as flutuações das proporções amostrais e o erro amostral fixado

As próximas figuras ilustram a flutuação das proporções amostrais obtidas de amostragens (com reposição) de elementos de uma população que apresentam a característica de interesse se manifestando de modo dicotômico, sob variados tamanhos amostrais (385, 210 e 100).

```
# Flutuação das proporções amostrais observadas

flut.N = function (N, n, p, conf, er) {
  qc = qnorm(1 - ((1 - conf) / 2))
  suc = rbinom(n = N, size = n, prob = p)
  prop_suc = suc / n
  dados = as.data.frame(prop_suc)
  names = c("Proporção amostral")
  colnames(dados) = names
  row.names(dados) = NULL
  meu_titulo01 = paste0("Flutuação das proporções amostrais \n", N, " amostras de tamanho ", n, "
    ↴ (dimensionamento sob um nível de confiança (1-\u03b1)= ", conf, " e um erro amostral
    ↴ \u03b5= ", er, " \nAs linhas verticais mostram a proporção populacional em azul (\u03c0=
    ↴ ", p, ") \ne os valores limites estabelecidos pelo erro arbitrado em vermelho (\u03c0
    ↴ +/-\u03b5= ", p, "+/-", er, ")")
  meu_titulo02 = paste0("Os valores das proporções amostrais seguem uma distribuição ~ N (
    ↴ \u03bc, \u03c3) = (", round(mean(dados$`Proporção amostral`), 4), ", ",
    ↴ round(sqrt(p * (1 - p) / n), 4), ")")

  plot(0, 0,
    type = "n",
    xlim = c(0.5 * min(dados$`Proporção amostral`), 1.1 * max(dados$`Proporção amostral`)),
    ylim = c(0, N),
    bty = "l",
    xlab = "Proporções amostrais observadas",
    ylab = "Amostras extraídas",
    main = "", #meu_titulo01
    sub = "") #meu_titulo02

  for (i in 1:N) {
    prop_amostral = dados$`Proporção amostral`[i]
    ploty = c(i, i)
    if (prop_amostral > p + er || prop_amostral < p - er)
      points(prop_amostral, i, col = "red", cex = 1) + text(y = i + 3, x = prop_amostral,
        ↴ labels = round(prop_amostral, 2), cex = 1, col = 'red')
    else
      points(prop_amostral, i, col = "black", cex = 1)
    segments(x0 = p, y0 = 0, x1 = p, y1 = N, col = "blue", lwd = 2, lty = 2)
    segments(x0 = p - er, y0 = 0, x1 = p - er, y1 = N, col = "red", lwd = 1, lty = 2)
    segments(x0 = p + er, y0 = 0, x1 = p + er, y1 = N, col = "red", lwd = 1, lty = 2)
  }
}
```

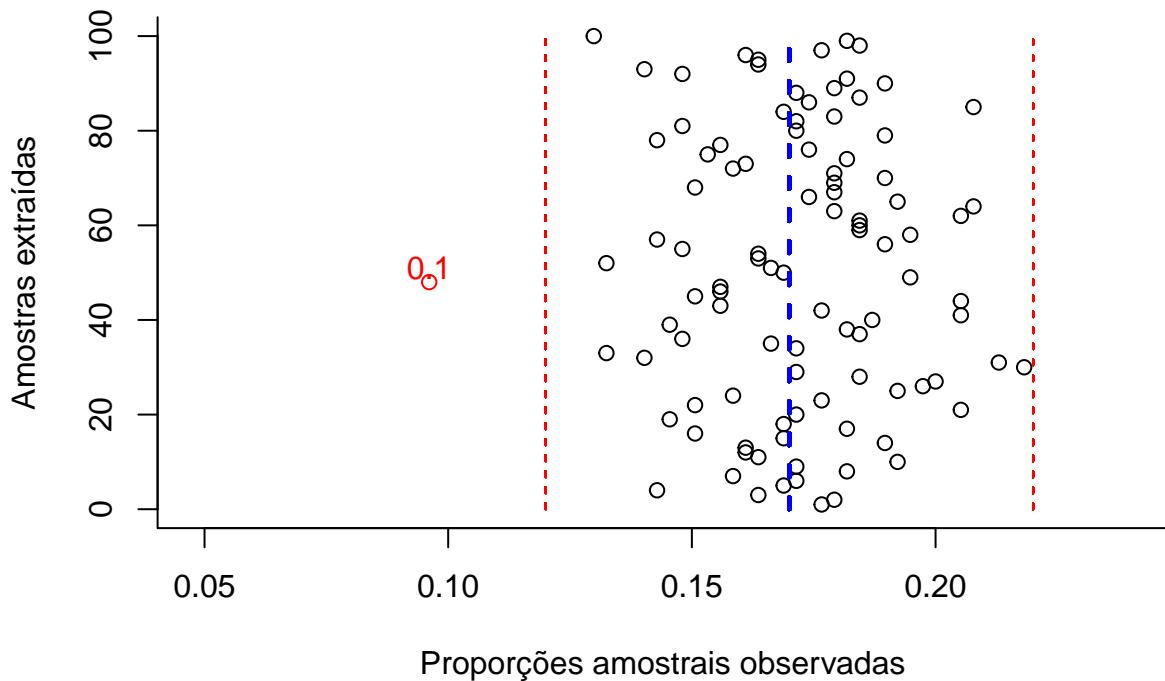


Figure 10.5: Flutuação das diversas proporções amostrais obtidas de amostragens cujo dimensionamento (385 elementos) foi estimado ignorando-se o conhecimento da proporção populacional () para um nível de confiança ($1-\alpha$)=0,95 e um erro amostral $\delta=0,05$ (em preto as proporções amostrais dentro da tolerância fixada e, em vermelho, as que aleatoriamente ultrapassam a tolerância fixada em $\pm\delta$).

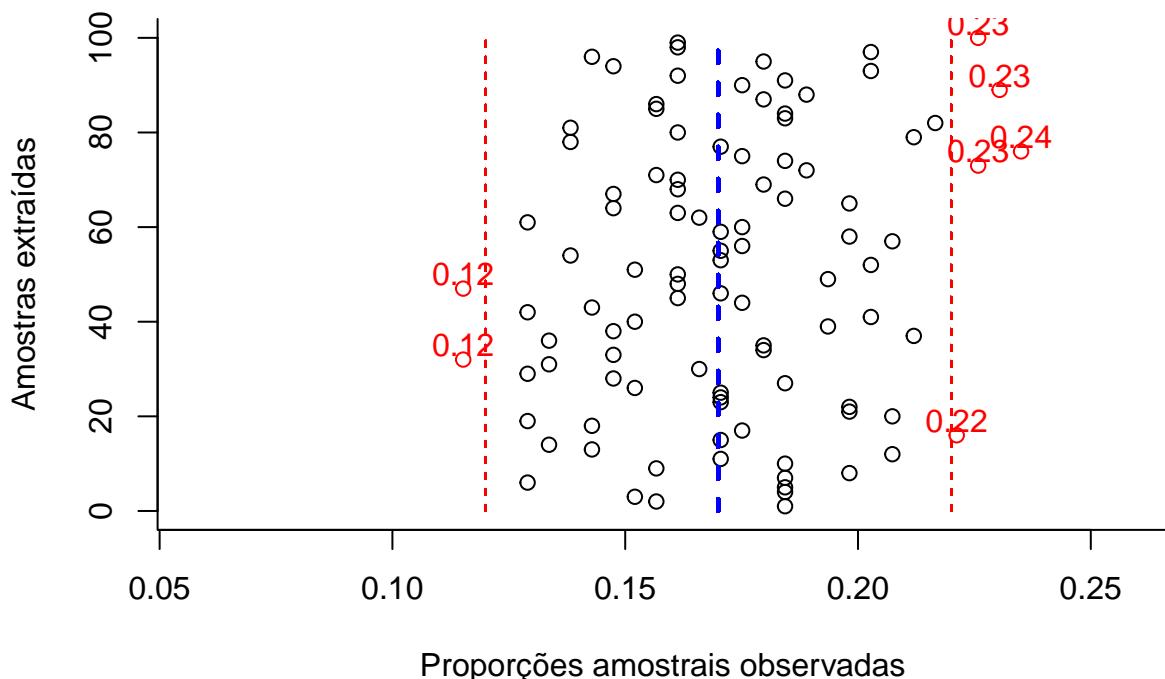


Figure 10.6: Flutuação das diversas proporções amostrais obtidas de amostragens cujo dimensionamento (217 elementos) foi estimado admitindo-se o conhecimento da proporção populacional () para um nível de confiança ($1-\alpha$)=0,95 e um erro amostral $\delta=0,05$ (em preto as proporções amostrais dentro da tolerância fixada e, em vermelho, as que aleatoriamente ultrapassam a tolerância fixada em $\pm\delta$).

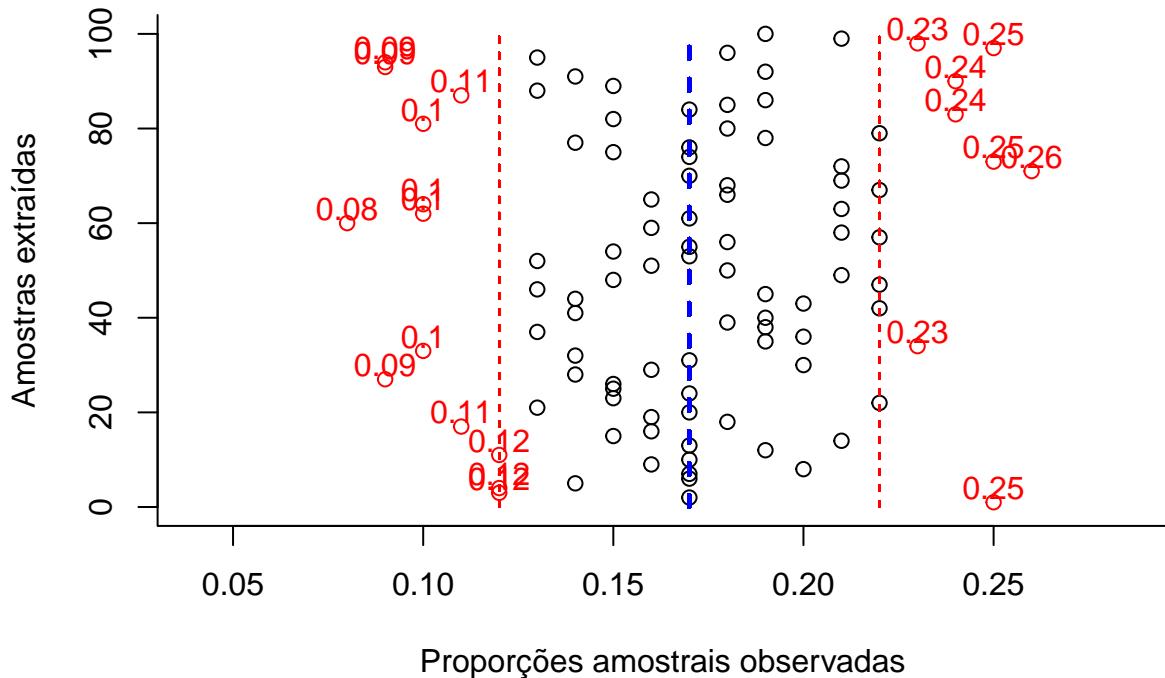


Figure 10.7: Flutuação das diversas proporções amostrais obtidas de amostragens cujo dimensionamento foi arbitrariamente fixado (100 elementos) para um nível de confiança ($1-\alpha$)=0,95 e um erro amostral $\delta=0,05$ (em preto as proporções amostrais dentro da tolerância fixada e, em vermelho, as que aleatoriamente ultrapassam a tolerância fixada em $\pm\delta$).

10.5 Intervalos de confiança para proporções amostrais

Podemos escrever o parâmetro (π) da proporção populacional em função da proporção amostral observada \hat{p} e de seu desvio padrão $\sigma_{\hat{p}}$:

$$Z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1),$$

ou

$$Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}}$$

com $Z \sim N(0, 1)$.

Assim,

$$\hat{p} - \pi = Z \cdot \sigma_{\hat{p}}$$

e

$$\pi = \hat{p} + Z \cdot \sigma_{\hat{p}}$$

Observa-se, todavia, que a variância da distribuição Normal da aproximação da distribuição das proporções amostrais é expressa em termos do parâmetro da proporção populacional π que não é conhecido:

$$\hat{p} \sim N[\pi; \frac{\pi \cdot (1 - \pi)}{n}]$$

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi(1 - \pi)}{n}}.$$

Demonstra-se que para:

- um razoável número de repetições: $n \geq 30$;
- de uma população onde a proporção π não é extrema: próximas a 0 ou 1; e tal que
- $(n \cdot \pi)$ e $(n \cdot (1 - \pi))$ sejam maiores que 15 (alguns autores consideram limites mais brandos, iguais a 10 ou ainda a 5),

Podemos tomar a proporção amostral \hat{p} como uma aproximação direta da proporção populacional π na expressão da variância da distribuição Normal que modela a distribuição das proporções amostrais sem que isso resulte em grande alteração na distribuição da variável Z .

Ou ainda, alternativamente, fazendo-se antes uma aproximação com correção de continuidade, onde definimos uma nova estimativa amostral da proporção populacional \hat{p}_c corrigida:

$$\hat{p}_c = \hat{p} + \frac{1}{2n}$$

se $\hat{p} < 0,50$,

ou

$$\hat{p}_c = \hat{p} - \frac{1}{2n}$$

se $\hat{p} > 0,50$.

As probabilidades associadas aos valores assumidos pela variável $Z \sim N(0, 1)$: **a área sob a curva**, encontram-se tabelados e podem ser utilizados para construir intervalos de confiança para o parâmetro da proporção populacional π associados a probabilidades desejadas.

$$P[\hat{p} - Z \cdot \sigma_{\hat{p}} < \pi < \hat{p} + Z \cdot \sigma_{\hat{p}}] = (1 - \alpha)$$

Assim (com \hat{p} ou \hat{p}_c) podemos construir *intervalos de confiança* em torno da proporção populacional π associados a um nível de significância estabelecido:

Bilaterais: intervalo delimitado por dois valores: mínimo e máximo, para a proporção amostral, dentro do qual todos os valores possuem um mesmo nível de significância:

$$P[\hat{p} - z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \leq \pi \leq \hat{p} + z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}] = (1 - \alpha)$$

Unilaterais: intervalos delimitados apenas em um de seus lados nos quais todos os valores possuem um mesmo nível de significância:

- Valor máximo (limitando à direita):

$$P[\pi \leq \hat{p} + z_\alpha \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}] = (1 - \alpha)$$

- Valor mínimo (limitando à esquerda):

$$P[\pi \geq \hat{p} - z_\alpha \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}] = (1 - \alpha)$$

```
# Intervalos de confiança das proporções amostrais observadas

IC.N = function (N, n, p, conf, er) {
  zc = qnorm(1 - ((1 - conf) / 2)) #Z=1,96
  suc = rbinom(n = N, size = n, prob = p)
  prop_suc = suc / n
  dados = as.data.frame(prop_suc)
  dados$lim_sup = dados$prop_suc + zc * sqrt(dados$prop_suc * (1 - dados$prop_suc) * (1 / n))
  dados$lim_inf = dados$prop_suc - zc * sqrt(dados$prop_suc * (1 - dados$prop_suc) * (1 / n))
  names = c("Proporção amostral", "lim superior", "lim inferior")
  colnames(dados) = names
  row.names(dados) = NULL
  meu_titulo001 = paste0("Intervalos com iguais níveis de confiança fixados em ", 100 * conf, "%"
    ↪ "\n(", N, " amostras de tamanho ", n, ") \nAs linhas verticais mostram a proporção
    ↪ populacional em azul (\u03c0: ", p, ") \ne a média das proporções amostrais em vermelho
    ↪ ( \u0070\u0302: ", round(mean(dados$`Proporção amostral`), 4), ".)")
  meu_titulo002 = paste0("Parâmetros da distribuição da população Normal aproximada ( \u03bc,
    ↪ \u03c3 ) = (", round(mean(dados$`Proporção amostral`), 4), ", ", round(sqrt(p * (1 - p) / n), 4)
    ↪ , ")")
```



```
plot(0, 0,
type = "n",
xlim = c(0.5 * min(dados$`lim inferior`), 1.1 * max(dados$`lim superior`)),
ylim = c(0, N),
bty = "l",
xlab = "Proporções amostrais observadas",
ylab = "Amostras extraídas",
main = "", #meu_titulo001
```

```

sub=""") #meu_titulo002

for (i in 1:N) {
  prop_amostral=dados$`Proporção amostral`[i]
  li = dados$`lim inferior`[i]
  ls = dados$`lim superior`[i]
  plotx = c(li,ls)
  ploty = c(i,i)
  if (li > p | ls < p) lines(plotx,ploty, col="red", lwd=2, lend=0)
  else lines(plotx,ploty, lend=0)
  if (li > p | ls < p) points(prop_amostral, i, col="red",
    + cex=1)+text(y=i+3,x=prop_amostral, labels=round(prop_amostral,1), cex=1, col='red')
  else points(prop_amostral, i, col="black", cex=1)
  segments(x0=mean(dados$`Proporção amostral`), y0=0, x1=mean(dados$`Proporção amostral`),
  + y1=N,col="red", lwd=2, lty=2)
  segments(x0=p , y0=0, x1=p ,y1=N,col="blue", lwd=2, lty=1)
}
}

```

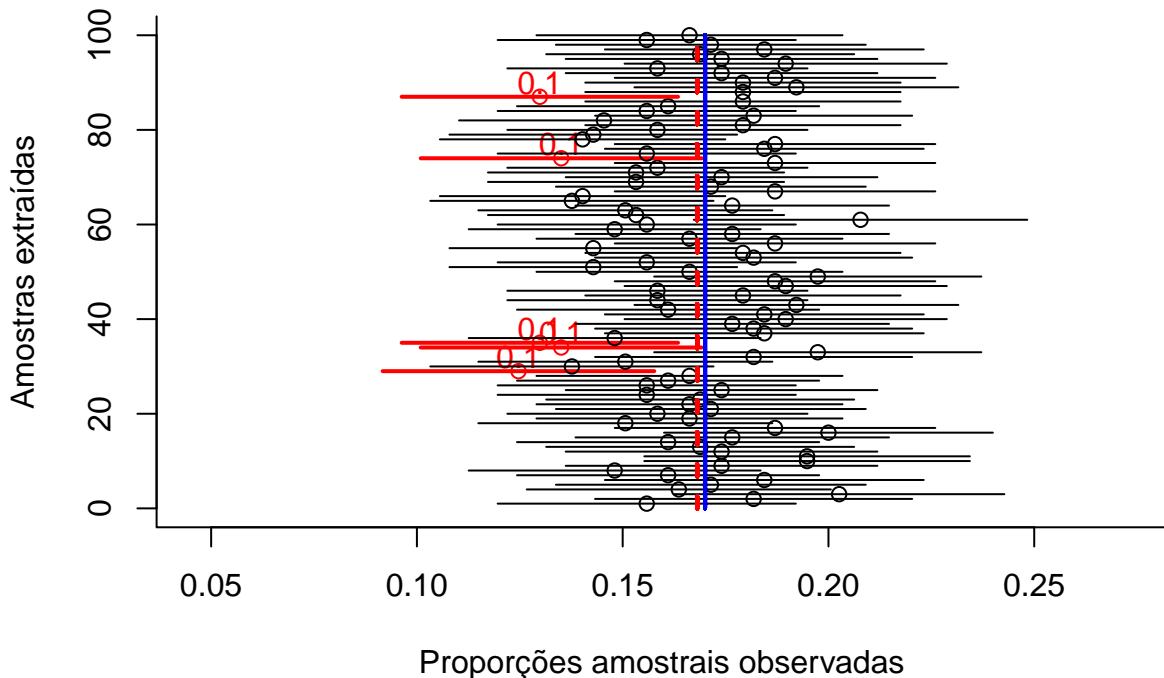


Figure 10.8: Intervalos de confiança construídos para as diversas proporções amostrais obtidas de amostragens (com reposição) de elementos de uma população que apresentam a característica de interesse se manifestando de modo dicotômico. O dimensionamento foi estimado ignorando-se o conhecimento da proporção populacional () para um nível de confiança (1-) e um erro amostral () estipulados: 385 elementos.

Exemplo: Em uma amostra aleatória, 136 pessoas de um grupo de 400 que receberam a vacina contra gripe, declararam haver sentido algum efeito colateral. Construa um intervalo com 95% de confiança para a verdadeira proporção populacional da ocorrência de efeitos colaterais vacinais .

Dados do problema:

- $\hat{p} = \frac{136}{400} = 0,34$ é a *proporção amostral* observada;
- o tamanho amostral ($n = 400$) é grande e a proporção amostral ($\hat{p} = 0,34$) não é extrema (próxima a zero ou um);
- π é a proporção populacional (desconhecida); e,
- para o nível de confiança solicitado ($(1 - \alpha) = 0,95$) temos da tabela $z_{(\frac{\alpha}{2})} = +/- 1,96$.

Um intervalo bilateral (fechado) para a proporção populacional desconhecida (π) sob um nível de confiança $(1 - \alpha)$ de 0,95 estará delimitado:

$$\begin{aligned}\hat{p} - z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} &\leq \pi \leq \hat{p} + z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \\ 0,34 - 1,96 \cdot \sqrt{\frac{0,34 \cdot (1 - 0,34)}{400}} &\leq \pi \leq 0,34 + 1,96 \cdot \sqrt{\frac{0,34 \cdot (1 - 0,34)}{n}} \\ 0,2936 &\leq \pi \leq 0,3864\end{aligned}$$

Exemplo: Em uma amostra aleatória de 2000 eleitores do Brasil constatou-se uma intenção de voto de 43% para um candidato à presidência. Realizada a eleição, deseja-se inferir qual o intervalo de variação da proporção populacional a um nível de confiança de 99%.

Dados do problema:

- $\hat{p} = 0,43$ é a *proporção amostral* observada;
- o tamanho amostral ($n = 2000$) é grande e a proporção amostral ($\hat{p} = 0,43$) não é extrema (próxima a zero ou um);
- π é a proporção populacional (desconhecida); e,
- para o nível de confiança solicitado ($(1 - \alpha) = 0,99$) temos da tabela $z_{(\frac{\alpha}{2})} = +/- 2,58$.

Um intervalo bilateral (fechado) para a proporção populacional desconhecida (π) sob um nível de confiança $(1 - \alpha)$ de 0,99 estará delimitado:

$$\begin{aligned} \hat{p} - z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} &\leq \pi \leq \hat{p} + z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \\ 0,43 - 2,58 \cdot \sqrt{\frac{0,43 \cdot (1 - 0,43)}{2000}} &\leq \pi \leq 0,43 + 2,58 \cdot \sqrt{\frac{0,43 \cdot (1 - 0,43)}{2000}} \\ 0,4014 &\leq \pi \leq 0,4586 \end{aligned}$$

10.5.1 Intervalos de confiança para a diferença entre duas proporções amostrais

Para a construção de um intervalo de confiança para a diferença de duas proporções populacionais π_X e π_Y a partir das proporções obtidas em duas amostras de razoável tamanho ($n_X \geq 30$ e $n_Y \geq 30$) e proporções amostrais \hat{p}_X e \hat{p}_Y não extremas (próximos a zero ou um) demonstra-se que a variável aleatória dessa diferença é tal que

$$Z = \frac{(\hat{p}_X - \hat{p}_Y) - (\pi_X - \pi_Y)}{\sqrt{\frac{\pi_X(1-\pi_X)}{n_X} + \frac{\pi_Y(1-\pi_Y)}{n_Y}}} \sim N(0, 1),$$

Sob as condições anunciadas, demonstra-se que se pode tomar as proporções amostrais \hat{p}_X e \hat{p}_Y como aproximações diretas das proporções populacionais π_X e π_Y na expressão da variância da distribuição Normal que modela a distribuição das diferenças das proporções amostrais sem que isso resulte em grande alteração na distribuição da variável Z .

$$Z = \frac{(\hat{p}_X - \hat{p}_Y) - (\pi_X - \pi_Y)}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}} \sim N(0, 1),$$

Assim podemos construir *intervalos de confiança* em torno da diferença das proporções populacionais π_X e π_Y associados a um nível de significância estabelecido:

Bilaterais: intervalo delimitado por dois valores: mínimo e máximo, para a proporção amostral, dentro do qual todos os valores possuem um mesmo nível de significância:

$$P \left[(\hat{p}_X - \hat{p}_Y) - z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \leq (\pi_X - \pi_Y) \leq (\hat{p}_X - \hat{p}_Y) + z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \right] = (1 - \alpha)$$

Unilaterais: intervalos delimitados apenas em um de seus lados nos quais todos os valores possuem um mesmo nível de significância:

- Valor máximo (limitando à direita):

$$P \left[(\pi_X - \pi_Y) \leq (\hat{p}_X - \hat{p}_Y) + z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \right] = (1 - \alpha)$$

- Valor mínimo (limitando à esquerda):

$$P \left[(\pi_X - \pi_Y) \geq (\hat{p}_X - \hat{p}_Y) - z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \right] = (1 - \alpha)$$

Módulo 11

Introdução a testes de hipóteses

11.1 Filosofia da ciência

Estritamente falando, todo o conhecimento fora da matemática, da lógica demonstrativa (um ramo da mesma) e da taxonomia encontra-se fundamentado em hipóteses (naturalmente há inúmeros tipos de hipóteses, mas as que estamos a nos referir são altamente confiáveis, como as expressas em certas leis gerais da física e da química como, por exemplo, a Lei de Hooke as Leis de Kepler dentre tantas outras).

O *raciocínio lógico demonstrativo* permeia as ciências até onde a matemática lhe suporta; todavia, em si (assim como também a matemática), é incapaz de gerar novos conhecimentos sobre o mundo que nos rodeia.

O *método lógico demonstrativo* é próprio para objetos que existem apenas *idealmente*, que são construídos inteiramente pelo nosso pensamento.

O *método hipotético experimental* é próprio das ciências naturais (física, química, biologia, etc.), que observam seus objetos e realizam experimentos.

Hipotético porque os cientistas partem de hipóteses sobre os objetos que guiam os experimentos e a avaliação dos resultados e *experimental* porque se baseia em observações e em experimentos, tanto para formular quanto para verificar as teorias.

O método hipotético experimental pode ser indutivo (fatos → lei geral) ou dedutivo (lei geral → fatos):

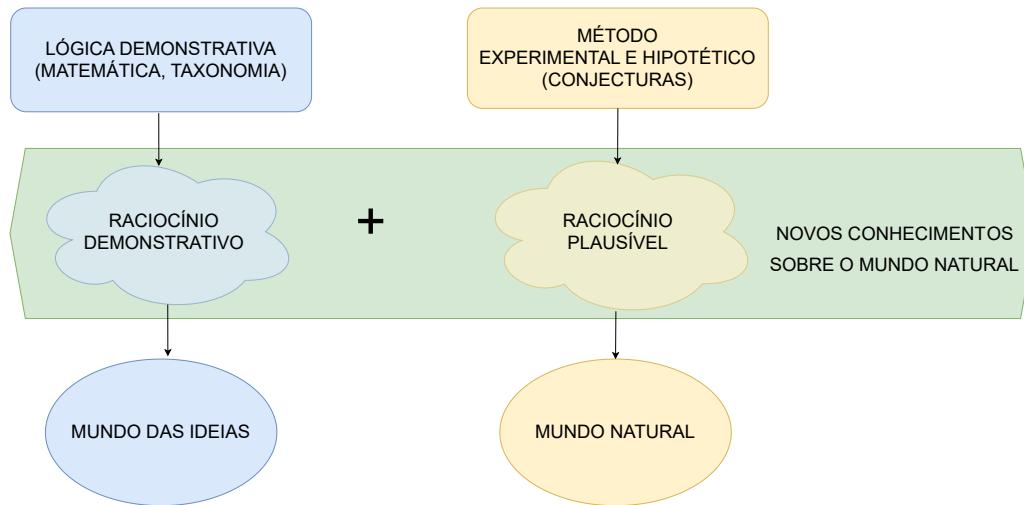


Figure 11.1: Método demonstrativo e Método experimental hipotético (George Polya, 1954)

Hipotético-indutivo porque o cientista observa inúmeros fatos variando as condições da observação; elabora uma hipótese e realiza novos experimentos (ou induções) para confirmar ou negar a hipótese; se esta não for negada, chega-se à lei do fenômeno estudado.

Hipotético-dedutivo porque tendo chegado à lei, o cientista pode formular novas hipóteses, deduzidas do conhecimento já adquirido, e com elas prever novos fatos, ou formular novas experiências, que o levam a conhecimentos novos.

Em muitos processos de investigação científica é frequente ao pesquisador formular perguntas que deverão ser apropriadamente respondidas.

- comparar esses resultados a outros valores; ou,
- comparar resultados obtidos pela aplicação de diferentes métodos/ou produtos (valores centrais, variabilidade, proporções) observados em diferentes amostras.

Uma hipótese é uma conjectura racional feita após um grande número de observações e experimentos; é uma tese que precisa ser confirmada ou verificada por meio de novas observações e experimentos.

Uma hipótese estatística é uma suposição feita sobre uma determinada característica de interesse de uma população sob estudo (um parâmetro) que subsiste (perdura, sobrevive, permanece incontestável) até que alguma informação sobre essa população seja estatisticamente significativa para contradizê-la.

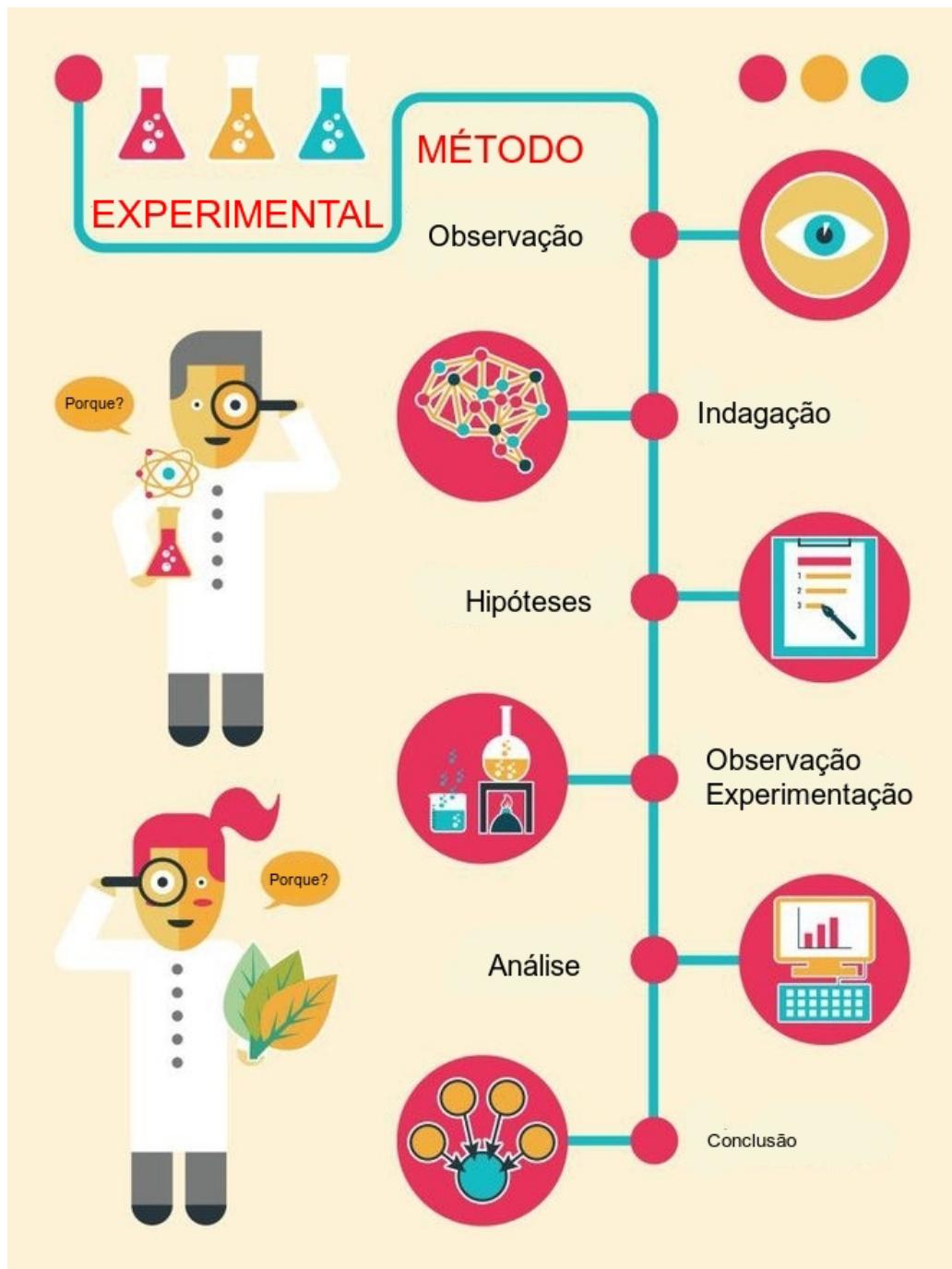


Figure 11.2: Método experimental hipotético

“A ciência não consegue provar coisa alguma. Ela pode apenas refutar as coisas’’ (Karl Popper)

Uma teoria científica é, portanto, transitória. Uma conjectura temporariamente sustentada que um dia poderá ser refutada e substituída por outra. Conclusões baseadas em raciocínios plausíveis são provisórias, ao contrário daquelas produzidas por raciocínios lógico demonstrativos.

Um teste de hipóteses refere-se, portanto, a um método quantitativo subsidiário em processos de decisão, baseado na inferência estatística e de ampla aplicabilidade na experimentação e pesquisa; virtualmente, em qualquer área do conhecimento.

11.2 História

Antigas referências relativas a testes de valores remontam aos séculos XVIII e XIX. Historicamente podemos retroceder a 1662, quando o médico flamengo Jean Baptista Van Helmont escreveu um desafio (*aposta de 300 florins*) em seu livro (Figura 11.4), sobre um *procedimento teste* que consistiria em se dividir 200 ou 500 pacientes com febre e pleurite em dois grupos iguais e aplicar a eles diferentes tratamentos: os habitualmente adotados pelos médicos da época e os seus próprios métodos. Ao final de um período de tempo (não foi especificado) verificar quantos *funerais* ocorreriam num e no outro (o livro foi publicado após sua morte, ocorrida em 1944, e não se tem registro sobre sua realização efetiva).

Outro registro histórico é o artigo publicado em 1710 na *Royal Society’s Philosophical Transactions* pelo médico escocês John Arbuthnot (1667-1735, Figura 11.5): *An argument for Divine Providence* ([link](#)).

Este artigo foi um marco na história da estatística; em termos modernos, ele realizou testes de hipóteses estatísticas, calculando o p-valor através de um teste de sinais e interpretou-o como estatisticamente significante e assim rejeitou a hipótese nula. Isso é creditado como “[...] o primeiro uso de testes de significância [...]” (in “Estatísticos do século”, David Bellhouse, 2001).

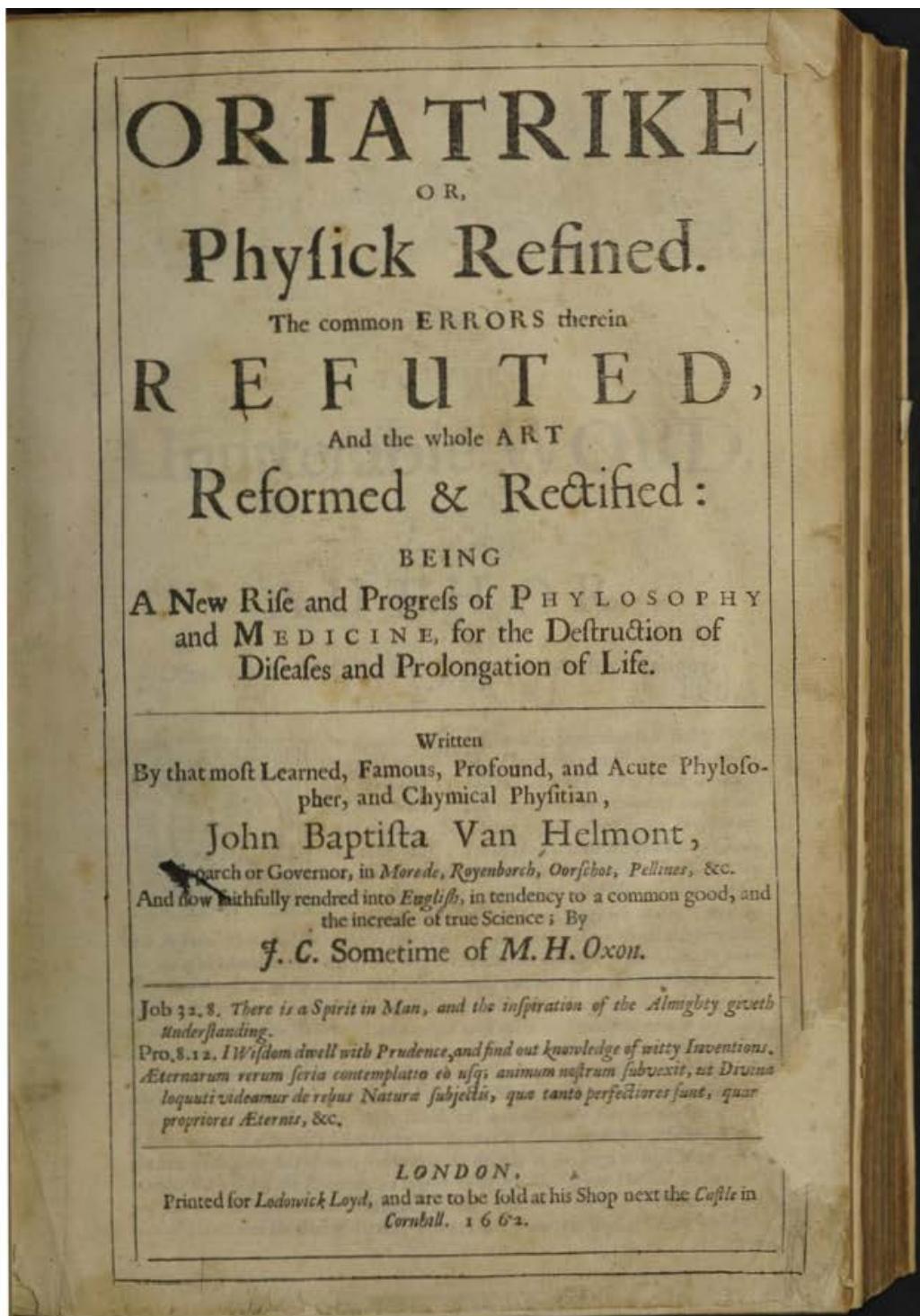


Figure 11.3: Oriatrike or, physick refined. The common errors therein refuted, and the whole art reformed and rectified: being a new rise and progress of phylosophy and medicine, for the destruction of diseases and prolongation of life (p. 526)



Figure 11.4: Tratamento mais utilizado à época (sangria)

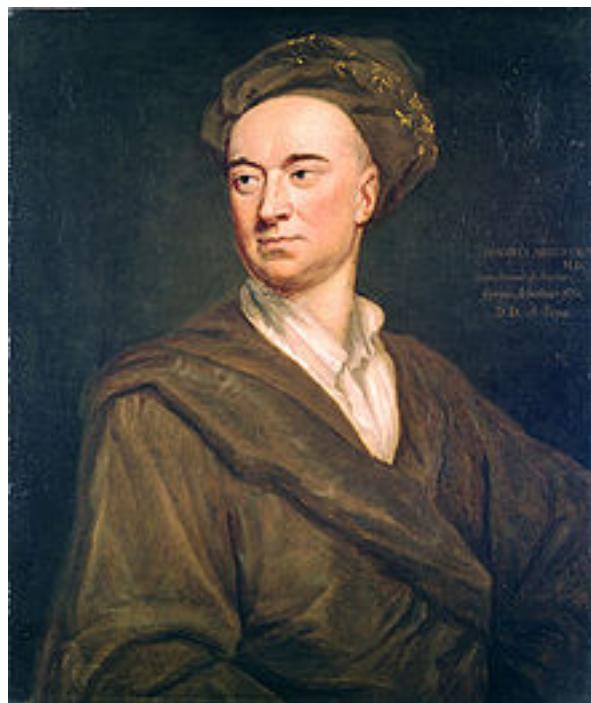


Figure 11.5: John Arbuthnot, FRS (1667-1735)

A estruturação dos testes de hipóteses, tal como são promovidos atualmente, é devida à metodologia empregada por alguns dos mais destacados cientistas da área do final do século XIX e começo do XX (Figura 11.6).

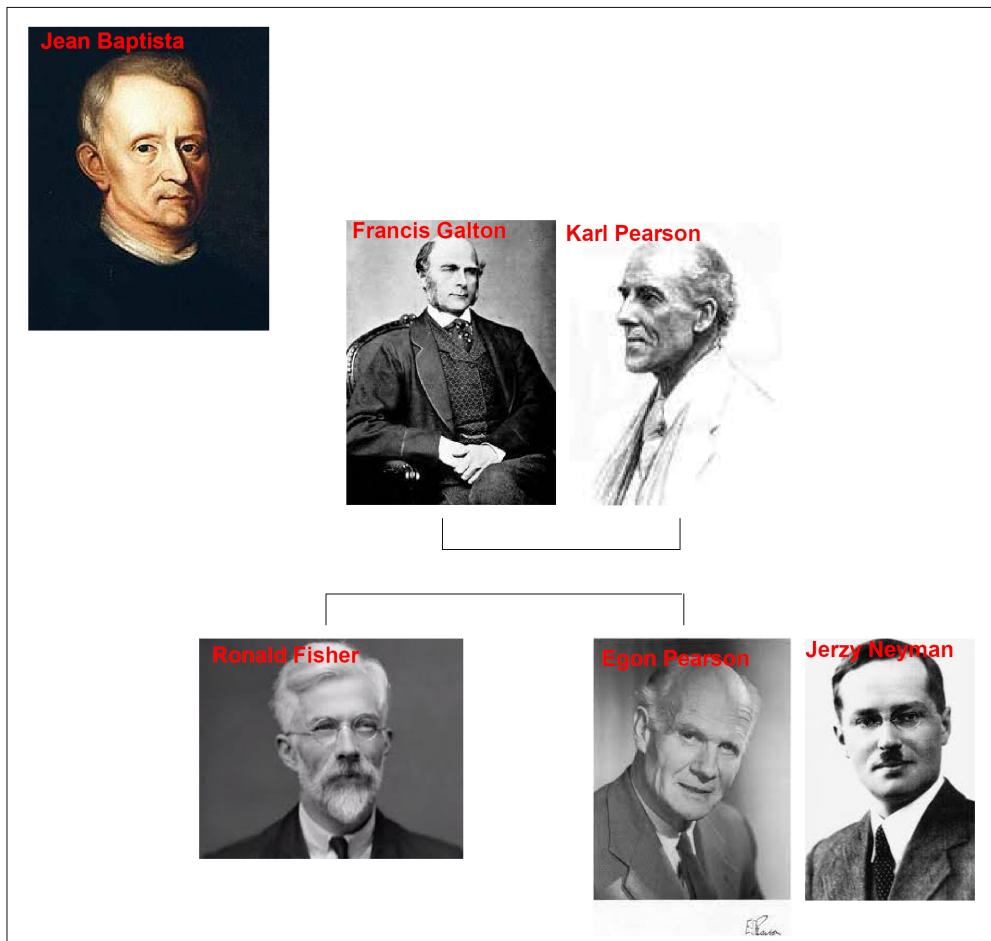


Figure 11.6: Personagens históricos

Em 1932 Karl Pearson se aposentou com professor da *University College London* e diretor do Laboratório Galton de eugenia. Apesar das objeções de Fisher, o laboratório de estatística foi dividido em dois departamentos. O Departamento de estatística (criado em 1901, o primeiro do gênero em uma universidade), assumido pelo filho mais novo de Karl, Egon; e o Laboratório de eugenia, assumido por seu sucessor na cadeira de Eugenia, Ronald Fisher.

O artigo de Henry F. Inman (*Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange From Nature*, 1994) registra uma intensa troca de correspondências entre Fisher e Pearson tendo por assunto suas diferenças conceituais matemáticas e estatísticas, pela contrariedade de Pearson ante a continuidade de Fisher em lecionar teoria estatística e até mesmo por espaço físico para os experimentos científicos de Fisher, ao remover material do Museu de eugenia deixado por Pearson.

O pensamento estatístico da primeira metade do século XXI tem seu interesse voltado à solução dos problemas de testes de hipóteses e sua formulação e filosofia, tal como hoje são conhecidos, foi em grande parte criada por Ronald Aylmer Fisher (1890-1962), Jerzy Neyman (1894-1981) e Egon Sharpe Pearson (1895-1980) no período compreendido entre 1915-1933:

- Estudo biológico realizado por Karl Pearson para tentar associar informações coletadas a distribuições de probabilidade apresentava os componentes básicos de um teste de hipóteses;
- Ronald Fisher (1925): *Statistical Methods for Research Workers*;
- George Waddel Snedecor (1940): *Statistical Methods*; e,
- Erich Leo Lehmann (1959): *Testing Statistical Hypotheses* condensando os estudos desenvolvidos em 1920 pelo filho de Pearson, Egon, e o matemático polonês, Jerzy Neyman (formulação de *Neyman-Pearson*).

11.3 Conceitos

A metodologia analisada na estruturação do método dos testes de hipóteses no fornece elementos auxiliares da decisão de rejeitar ou não - sob um prisma probabilístico - determinada conjectura postulada acerca de um parâmetro da população estudada.

A *conclusão* de um teste de hipóteses resume-se a: *aceitar* ou *rejeitar* uma hipótese. Muitos estatísticos não adotam a expressão *aceitar* uma hipótese preferindo, no lugar, usar a expressão *não rejeitar* a hipótese sob um certo nível de significância.

Por que essa distinção entre *aceitar* e *não rejeitar*?

Ao se usar a expressão *aceitar* pode haver uma pré-concepção de que a hipótese é universalmente verdadeira (lembrando que a conclusão encontra-se alicerçada simplesmente em uma amostra).

Utilizando-se a expressão *não rejeitar* salienta-se que a informação trazida pelos dados (a amostra) não foi suficientemente robusta para que pudéssemos abandonar essa hipótese em favor de uma outra.

Alguns dizem que os estatísticos não se perguntam qual a probabilidade de estarem *certos*; mas de não estarem *errados*.

Um *teste de hipóteses* guarda uma certa semelhança a um julgamento. Caso não haja indício forte o suficiente que comprove a culpa do acusado ele é declarado como inocente (mesmo que não o seja de fato). No contexto estatístico, os *indícios* que nos levam a rejeitar uma hipótese provêm da análise de informações observadas na amostra.

A *hipótese nula* (H_0) é a hipótese inicial, a que reflete a situação em que não há mudança. É pois uma *hipótese conservadora* (resultado de experimentos anteriores).

A *hipótese alternativa* (H_1) contradiz aquilo anunciado pela hipótese nula, é uma *hipótese inovadora*.

Inicialmente a hipótese nula ela é assumida como verdadeira para, logo a seguir, ser confrontada novas evidências amostrais para se verificar a sustentabilidade de sua afirmação:

- caso a informação amostral demonstre a consistência de hipótese nula tudo o que pode ser feito é se decidir por sua manutenção (falho na tentativa de se derrubar a hipótese conservadora); e,
- caso não seja, analisa-se quão improvável pode ser a informação amostral além de uma dúvida razoável ou mera coincidência (nível de significância).

“Em relação a qualquer experimento não devemos falar desta hipótese como a *hipótese nula*, e deve-se atentar que a *hipótese nula* nunca é provada ou estabelecida, mas é, possivelmente, refutada, no decorrer da experimentação. Todo experimento deve existir apenas para das aos fatos a chance de refutar a *hipótese nula...*” (*The Design of Experiments*, Ronald Aylmer Fisher, 1935, p. 19)

O objetivo de um teste de hipóteses é, pois, o de tomar uma decisão no sentido de verificar se existem razões para rejeitar ou não a hipótese nula. Esta decisão é baseada na informação disponível, obtida a partir de uma amostra, que se recolhe da população.

Teste de hipóteses nos possibilitam associar um *nível de significância* (α) como medida probabilística do erro que se pode incorrer ao se concluir pela *rejeição* de uma *hipótese verdadeira*, na tomada de decisão.

Nível de significância (α) é estabelecido pelo pesquisador (baseado tanto na expertise dele, quanto no campo a que o estudo pertence) antes do experimento ser realizado e corresponde ao grau do risco que se deseja incorrer ao se “rejeitar” uma hipótese verdadeira.

Nível de confiança ($1 - \alpha$) é a medida da confiabilidade de nossa conclusão no teste de hipóteses: “não rejeitar” uma hipótese verdadeira.

11.4 Natureza dos erros

Para introduzir os conceitos relacionados aos erros considere uma situação onde uma empresa produz lâmpadas e a vida útil média, em horas, dessas lâmpadas segue uma distribuição Normal tal que $VU \sim N(1600, 120)$.

Se não temos conhecimento algum sobre a real vida útil média dessas lâmpadas e alguém nos afirma que a vida útil é de 1.600 h, para confirmar ou não essa proposição (de um modo “científico”) devemos extrair uma amostra.

Usando conceitos já explicados em uma unidade anterior podemos determinar o tamanho amostral em função de:

- um erro máximo tolerado: $\varepsilon=20$ horas;
- um nível de significância estabelecido: $\alpha=0,05$; e,
- e alguma informação sobre a medida da variabilidade da variável em estudo: $\sigma=120$ horas (no caso, o desvio padrão populacional).

```
##      mu media     erro   li   ls
## 1 1600 1608  7.5927 1589 1626
## 2 1600 1582 -18.1055 1563 1601
## 3 1600 1602  1.8577 1582 1622
## 4 1600 1606  6.4590 1588 1625
## 5 1600 1611 10.5032 1590 1631
## 6 1600 1602  1.5498 1583 1620
## 7 1600 1610 10.2809 1592 1628
## 8 1600 1605  4.8077 1585 1625
## 9 1600 1584 -15.6798 1567 1602
## 10 1600 1611 10.9470 1592 1630
## 11 1600 1601  1.3122 1583 1620
## 12 1600 1607  7.1509 1588 1626
## 13 1600 1601  0.6101 1580 1622
## 14 1600 1606  5.8799 1588 1624
## 15 1600 1608  7.9382 1589 1627
```

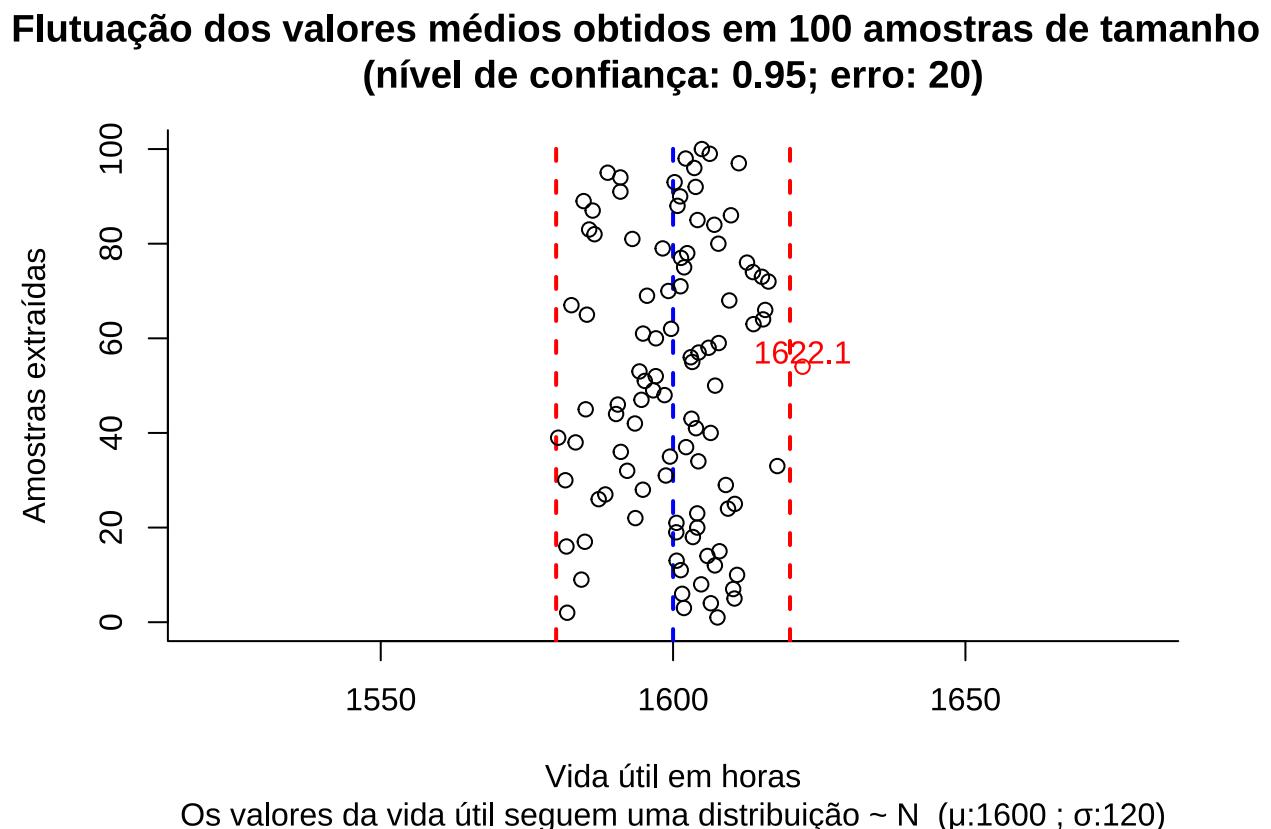


Figure 11.7: Flutuação dos valores médios para diversas amostras extraídas de uma mesma população distribuição $\sim N(\mu; \sigma^2)$

```
## 16 1600 1582 -18.2386 1560 1603
## 17 1600 1585 -15.0881 1567 1603
## 18 1600 1603 3.3909 1584 1623
## 19 1600 1601 0.5543 1582 1619
## 20 1600 1604 4.1425 1583 1626
## 21 1600 1601 0.5750 1582 1620
## 22 1600 1594 -6.4473 1573 1614
## 23 1600 1604 4.1334 1586 1623
## 24 1600 1609 9.3943 1589 1630
## 25 1600 1611 10.5605 1591 1630
## 26 1600 1587 -12.7224 1568 1606
## 27 1600 1588 -11.5858 1570 1607
## 28 1600 1595 -5.1665 1574 1615
## 29 1600 1609 9.0211 1589 1629
## 30 1600 1582 -18.4252 1562 1601
## 31 1600 1599 -1.2394 1579 1619
## 32 1600 1592 -7.8472 1573 1612
## 33 1600 1618 17.8246 1600 1636
## 34 1600 1604 4.3301 1586 1623
## 35 1600 1599 -0.5297 1580 1619
## 36 1600 1591 -8.9363 1569 1614
## 37 1600 1602 2.2254 1583 1622
## 38 1600 1583 -16.6711 1564 1603
## 39 1600 1580 -19.6498 1560 1601
## 40 1600 1606 6.4301 1587 1626
## 41 1600 1604 3.9207 1583 1625
## 42 1600 1593 -6.5274 1574 1613
## 43 1600 1603 3.1501 1583 1623
## 44 1600 1590 -9.7466 1571 1609
## 45 1600 1585 -14.9409 1567 1603
## 46 1600 1591 -9.4497 1572 1609
## 47 1600 1595 -5.4056 1573 1617
## 48 1600 1599 -1.4559 1581 1616
## 49 1600 1597 -3.3945 1579 1615
## 50 1600 1607 7.1934 1588 1627
## 51 1600 1595 -4.8362 1574 1616
## 52 1600 1597 -2.9590 1578 1616
## 53 1600 1594 -5.7470 1574 1615
## 54 1600 1622 22.1465 1603 1641
## 55 1600 1603 3.2787 1582 1624
## 56 1600 1603 3.0256 1582 1624
## 57 1600 1604 4.3715 1585 1624
## 58 1600 1606 6.0816 1588 1624
## 59 1600 1608 7.8030 1586 1630
## 60 1600 1597 -2.9318 1578 1616
## 61 1600 1595 -5.1310 1576 1613
## 62 1600 1600 -0.3395 1580 1620
## 63 1600 1614 13.7546 1595 1632
## 64 1600 1615 15.3934 1596 1634
## 65 1600 1585 -14.7379 1565 1606
## 66 1600 1616 15.7590 1596 1636
## 67 1600 1583 -17.3835 1563 1602
## 68 1600 1610 9.6190 1589 1630
## 69 1600 1596 -4.4557 1579 1612
## 70 1600 1599 -0.7982 1581 1617
## 71 1600 1601 1.2674 1581 1621
```

```

## 72 1600 1616 16.3181 1595 1638
## 73 1600 1615 15.1991 1593 1637
## 74 1600 1614 13.6631 1596 1631
## 75 1600 1602 1.8884 1582 1621
## 76 1600 1613 12.6427 1593 1632
## 77 1600 1601 1.3533 1580 1623
## 78 1600 1602 2.4236 1582 1622
## 79 1600 1598 -1.7740 1577 1619
## 80 1600 1608 7.7458 1585 1630
## 81 1600 1593 -6.9614 1575 1611
## 82 1600 1587 -13.4360 1568 1605
## 83 1600 1586 -14.3499 1569 1602
## 84 1600 1607 7.0568 1586 1628
## 85 1600 1604 4.1783 1584 1624
## 86 1600 1610 9.8978 1589 1631
## 87 1600 1586 -13.7489 1567 1606
## 88 1600 1601 0.7570 1584 1618
## 89 1600 1585 -15.2976 1564 1605
## 90 1600 1601 1.2083 1580 1623
## 91 1600 1591 -9.0059 1570 1612
## 92 1600 1604 3.8620 1584 1624
## 93 1600 1600 0.2658 1580 1621
## 94 1600 1591 -9.0057 1572 1610
## 95 1600 1589 -11.1806 1568 1610
## 96 1600 1604 3.6329 1585 1622
## 97 1600 1611 11.2440 1590 1633
## 98 1600 1602 2.1514 1582 1622
## 99 1600 1606 6.2768 1585 1627
## 100 1600 1605 4.9295 1586 1624

```

Observa-se que algumas das amostras, numa proporção igual ao nível de significância estabelecido quando do dimensionamento (5%), apresentam médias com valores que se afastam do valor médio populacional mais que o erro estabelecido (20 h).

Como já informado anteriormente, um teste de hipóteses é um método quantitativo e não se baseia, sobremaneira, em impressões pessoais ou outros achismos. Os cenários a seguir foram criados apenas para tentar estabelcer um paralelo entre a probabilidade de se obter médias amostrais muito destoantes da média populacional e uma “inclinação subjetiva” em se rejeitar uma afirmação.

Considere que a sua amostra em particular é uma das que não se afasta tanto do valor que lhe afirmaram (a vida útil das lâmpadas é de 1.600 h).

Nessa situação, talvez você não se “convencesse” de que a vida útil média fosse diferente daquilo que lhe informaram e, assim, não iria recusar a afirmação.

Agora considere que a sua amostra em particular é uma das que se afasta muito do valor que lhe afirmaram.

Nessa nova situação, certamente você iria “suspeitar” que a vida útil média é diferente daquilo que lhe informaram e assim, recusar a afirmação.

Na primeira decisão, você **não recusou uma afirmação que era, de fato, verdadeira**; ao passo que na segunda, você **rejeitou uma afirmação que era verdadeira** (lembrando que você **não sabia** que a vida útil média é, de fato, 1.600 h).

Como se vê no quadro abaixo, há **dois tipos de erros** envolvidos em um teste de hipóteses e suas consequências, muitas vezes, são bem diferentes.

- Erro do tipo I e
- Erro do tipo II.

Um *erro do tipo I* ocorre quando o pesquisador rejeita uma hipótese nula quando é verdadeira. A probabilidade (limitada pelo pesquisador) de se incorrer em um *erro do tipo I* é chamada de *nível de significância* e é frequentemente denotada pela letra grega α .

Um *erro do tipo II* ocorre quando o pesquisador não rejeita uma hipótese nula que é falsa. A probabilidade de cometer um *erro do tipo II*, também chamada de *poder do teste* e é frequentemente denotada pela letra grega β .

Table 11.1: Erros envolvidos na rejeição ou não da hipótese nula

Valor real do parâmetro (desconhecido)	Não rejeitar H_0	Rejeitar H_0
H_0 verdadeira	Decisão correta probabilidade associada = $(1 - \alpha)$	Erro do tipo I probabilidade associada = α
H_0 falsa	Erro do tipo II probabilidade associada = β	Decisão correta probabilidade associada = $(1 - \beta)$

No quadro acima identificam-se:

- α : a probabilidade associada ao cometimento de um *erro do tipo I*: rejeitar a hipótese nula sendo ela verdadeira (arbitrado pelo pesquisador, é denominado nível de significância do teste);

- β : a probabilidade associada ao cometimento de um *erro do tipo II*: não rejeitar a hipótese nula sendo esta falsa;
- $(1-\alpha)$: o nível de confiança estabelecido para a decisão, a probabilidade associada em **não se rejeitar a hipótese nula (H_0)** quando ela é, de fato, verdadeira; e,
- $(1-\beta)$: o *poder do teste*, a probabilidade associada em não se aceitar a hipótese nula (H_0) quando ela é, de fato, falsa.

Qual erro é o pior? Depende!

Por exemplo, se alguém testa a presença de alguma doença em um paciente, decidindo incorretamente sobre a necessidade do tratamento (ou seja, decidindo que a pessoa está doente), pode submetê-lo ao desconforto pelo tratamento (efeitos colaterais) além de perda financeira pela despesa incorrida.

Mas por outro lado, a falha em diagnosticar a presença da doença no paciente pode levá-lo à morte pela ausência de tratamento.

Outro exemplo clássico a ser citado seria o de condenar uma pessoa inocente ou libertar um criminoso.

Como não há uma regra clara sobre qual tipo de erro é o pior recomenda-se quando se usa dados para testar uma hipótese observar com muito cuidado as consequências que podem seguir os dois tipos de erros. Vários especialistas sugerem o uso de uma tabela como a abaixo para detalhar as consequências de um erro Tipo 1 e Tipo 2 em sua análise específica.

Table 11.2: Consequências da tomada de decisão face aos erros envolvidos

H_0 explicada	Erro tipo 1: rejeitar H_0 quando verdadeira	Erro tipo II: não rejeitar H_0 quando falsa
O medicamento “A” não alivia a Condição “B”	O medicamento “A” não alivia a Condição “B”, mas não é eliminado como opção de tratamento	O medicamento “A” alivia a condição “B”, mas é eliminado como opção de tratamento
Consequências	Pacientes com Condição “B” que recebem o Medicamento “A” não obtêm alívio. Eles podem experimentar piora da condição e/ou efeitos colaterais, até e incluindo a morte. A empresa produtora do medicamento pode enfrentar processos judiciais	Um tratamento viável permanece indisponível para pacientes com Condição “B”. Os custos de desenvolvimento são perdidos. O potencial lucro pela produção do medicamento “A” pela empresa é eliminado.

É desejável conduzir o teste de um modo a manter a probabilidade de ambos os tipos de erro em um mínimo.

- aumentar o tamanho amostral reduz a probabilidade associada ao cometimento de erro do tipo II (β) e, consequentemente, aumenta o poder do teste ($1 - \beta$);
- aumentar o nível de significância (α) tem implicação direta na probabilidade associada ao cometimento de erro do tipo I todavia reduz a probabilidade associada ao cometimento de erro do tipo II (β).

11.5 Recomendações gerais

- o pesquisador deve delimitar o objeto de sua pesquisa;
- uma boa hipótese deve ser baseada em uma boa pergunta sobre o objeto do estudo;
- deve ser simples e específica;
- deve ser formulada na fase propositiva da pesquisa e não após a coleta de dados (*post hoc*);
- enunciar as hipóteses: as hipóteses são apresentadas de tal maneira que sejam **mutuamente exclusivas** (o que afirmado por uma deve ser contradito pela outra);
- as hipóteses são comumente denominadas por hipótese nula (H_0) e hipótese alternativa (H_1);
- a hipótese nula (H_0) que será testada sob um nível de significância (α) é, em geral, de concordância com o parâmetro que se estuda da população (conservadora) e baseada em conhecimento prévio;
- a hipótese alternativa (H_1) é contrária, oposta, antagônica à hipótese nula (novadora); e,
- estabelecer um nível apropriado para a significância α (em alguns campos do conhecimento níveis de significância muito reduzidos são impraticáveis).

11.6 Efeito do limite central

Seja X_1, X_2, \dots uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada uma com média finita $\mu = E(X_i)$.

A Lei forte dos grandes números (teorema) demonstra que

$$\frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mu$$

quando $n \rightarrow \infty$.

Isto é, $P\{\lim_{n \rightarrow \infty} \left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \mu\} = 1$

11.6.1 Erro global

O erro global ($\varepsilon = X - \mu$) é um agregado de componentes. Uma medida (observação) obtida em um ensaio experimental específico pode estar sujeita a erros:

- analíticos;
- de amostragem (física, química, biológica, ...);
- processuais (produzido por falhas no cumprimento das configurações exatas das condições experimentais);
- erros devidos à variação de matérias-primas;
- medição (diferentes operadores de equipamentos ou equipamentos descalibrados).

Assim, ε será uma função linear de componentes $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ de erros. Se cada erro individual for relativamente pequeno, será possível aproximar o erro global como uma função linear dos componentes de erros, onde a são constantes:

$$\varepsilon = a_1 \varepsilon_1 + a_2 \varepsilon_2 + \dots + a_n \varepsilon_n$$

O Teorema do limite central afirma que, sob condições quase sempre satisfeitas no mundo real da experimentação, a distribuição de tal função linear de erros tenderá à uma distribuição Normal quando o número de seus componentes torna-se grande, **independentemente** da distribuição original da população de onde suas amostras geradoras se originaram.

Seja X_1, \dots, X_n uma sequência de variáveis aleatórias independentes e identicamente distribuídas, com média μ e variância σ^2 .

A distribuição assumirá um perfil

$$\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0, 1)$$

quando $n \rightarrow \infty$.

Assim, para $-\infty < a < \infty$,

$$P\left\{\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \rightarrow \mathcal{N}(0, 1)$$

quando $n \rightarrow \infty$.

Denotando-se de um modo alternativo, podemos então definir a estatística Z e sua correspondente distribuição como

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

Ou seja, Z é uma variável aleatória que segue a distribuição Normal com média zero e desvio-padrão unitário (Normal padronizada).

Em resumo: quando, como é habitual, um erro experimental é um agregado de vários erros de componentes, sua distribuição tende para a forma Normal, mesmo a distribuição dos componentes pode ser marcadamente não Normal;

A média da amostra tende a ser distribuída Normalmente, mesmo que as observações individuais em que se baseia não o sejam. Consequentemente, métodos estatísticos que dependam, não diretamente da distribuição das observações individuais, mas na distribuição das médias tendem a ser insensíveis ou robustos à não normalidade.

Procedimentos que compararam médias são geralmente robustos à não normalidade.

11.7 Estruturas das hipóteses

11.7.1 Interpretação gráfica dos níveis de significância/confiança

O delineamento de um teste de hipóteses inclui regras de decisão para se rejeitar ou não a hipótese nula.

Essas regras de decisão passam pela comparação dos valores calculados de uma estatística apropriada para o teste em curso com seus valores extremos, frequentemente obtidos em tabelas, os quais estão associados ao complemento de uma probabilidade (o nível de confiança) de ocorrência condizente ao nível de significância estabelecido na pesquisa.

Essa comparação é por demais facilitada se visualizada no gráfico da densidade de probabilidade da distribuição da estatística do teste, onde regiões (baseadas no nível de significância estabelecido) podem ser estabelecidas:

- testes bilaterais (*hipótese alternativa do tipo: diferente de*): a região é fechada, delimitada à esquerda e à direita por valores críticos de estatística do teste;
- testes unilaterais à direita (*hipótese alternativa do tipo: maior que*): a região é fechada à esquerda, delimitada por um valor crítico da estatística do teste e aberta à direita ($\text{ao} \rightarrow \infty$); e,
- testes unilaterais à esquerda (*hipótese alternativa do tipo: menor que*): a região é fechada à direita, delimitada por um valor crítico da estatística do teste e aberta à esquerda ($\rightarrow -\infty$).

No gráfico de densidade de probabilidade da estatística do teste temos uma primeira região frequentemente denominada de *região de não rejeição*: um intervalo de valores dentro do qual, se o valor calculado para a estatística de teste estiver contido, a hipótese nula não será rejeitada.

O intervalo de valores que delimitam a *região de não rejeição* é tal que a probabilidade dessa região é igual ao nível de confiança ($1 - \alpha$).

Se a estatística calculada para o teste estiver fora da faixa de valores delimitada na *região de não rejeição* a hipótese nula poderá ser rejeitada sob o nível de significância α estabelecido; ou seja, a probabilidade de se incorrer em um erro *Tipo I: rejeitar a hipótese nula quando ela é verdadeira* é igual a α .

Com a popularização dos programas estatísticos computacionais, a *probabilidade exata* associada ao valor calculado da estatística do teste passou ser neles apresentada de modo *default*, nominada pela expressão *valor p* (*p-Value*) que expressa uma probabilidade.

Para melhor entender o *valor-p* (*p-value*) suponha que o valor da estatística do teste seja igual a ζ . O *valor p* é o quantil associado (a probabilidade exata) a ζ na distribuição de probabilidade usada como referência. Se

o valor p for menor que o nível de significância (α) estipulado pelo pesquisador, rejeita-se a hipótese nula sob esse nível de significância de cometimento de um erro do tipo I.

11.7.2 Teste de hipóteses Bilateral

Nesse tipo de teste a *hipótese alternativa* é proposta como a dizer que o valor em teste é diferente daquele afirmado pela *hipótese nula* (conservadora):

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

em que μ é o valor conservador do parâmetro que se deseja testar frente ao valor alternativo μ_0 .

```
alfa=0.05

prob_desejada1=alfa/2
z_desejado1=round(qnorm(prob_desejada1),4)
d_desejada1=dnorm(z_desejado1, 0, 1)

prob_desejada2=1-alfa/2
z_desejado2=round(qnorm(prob_desejada2),4)
d_desejada2=dnorm(z_desejado2, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
```

```

scale_x_continuous(name="Valores da estatística calculada para o teste") +
  labs(title =
    "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
    \n teste",
    subtitle = "P((-val. crítc), (val. crít.))=(1-\u03b1) em cinza (nível de confiança)
    \n P(-\u221e; (-val. crític.))= P((val.crítc.); \u221e)= \u03b1/2 em vermelho ") +
  geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1),
    color="blue", lty=2, lwd=0.3) +
  geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2),
    color="blue", lty=2, lwd=0.3) +
  annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="-"(valor crítico),
    angle=90, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="(valor crítico)", angle=90,
    vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=z_desejado1-2, y=0.1, label="Região de rejeição da hipótese nula
    \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Região de rejeição da hipótese nula
    \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Região de não rejeição da hipótese
    \n nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  theme_bw()

```

Regiões críticas sob a curva da função densidade da distribuição apropriada ao teste

$P((-val. crítc), (val. crít.))=(1-\alpha)$ em cinza (nível de confiança)
 $P(-\infty; (-val. crític.))= P((val.crítc.); \infty)= \alpha/2$ em vermelho

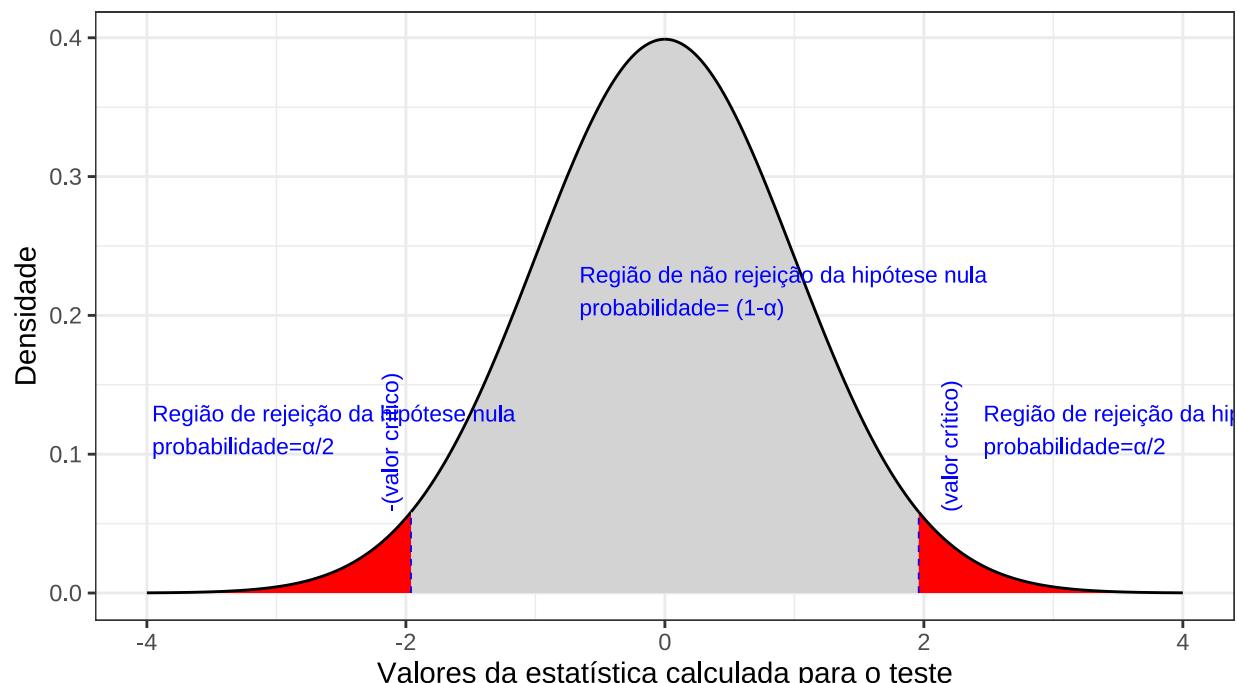


Figure 11.8: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores amostrais observados é inferior a $\frac{\alpha}{2}$, estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

Na Figura 11.8 observa-se:

- as regiões de rejeição da hipótese nula (subdivididas nos dois lados) sob a curva da função densidade de probabilidade da distribuição adequada ao teste com probabilidades iguais ao nível de significância α ;
- a região de não rejeição da hipótese nula (delimitada à esquerda e à direita) com probabilidade igual ao nível de confiança ($1 - \alpha$); e,
- os valores críticos da estatística do teste.

11.7.3 Teste de hipóteses Unilateral à esquerda

Nesse tipo de teste a *hipótese alternativa* é proposta como a dizer que o valor em teste não apenas é diferente, mas é menor do que aquele afirmado pela *hipótese nula* (conservadora):

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

em que μ é o valor conservador do parâmetro que se deseja testar frente ao valor alternativo μ_0 .

```
alfa=0.05
prob_desejada=alfa
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado),
            colour="black") +
```

```

geom_area(stat = "function",
           fun = dnorm,
           fill = "lightgrey",
           xlim = c(z_desejado,4),
           colour="black") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores da estatística calculada para o teste") +
labs(title=
  "Região crítica sob a curva da função densidade da \ndistribuição apropriada ao
  → teste",
  subtitle = "P( (-val. crít.),\u221e)=(1-\u03b1) em cinza (nível de confiança)
  → \nP(-\u221e; (-val. crít.))=\u03b1 em vermelho ")+
geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue",
  → lty=2, lwd=0.3)+
annotate(geom="text", x=z_desejado-0.1, y=d_desejada, label="-(valor crítico)", angle=90,
  → vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado-2, y=0.1, label="Região de rejeição da hipótese nula
  → \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado+1.3, y=0.2, label="Região de não rejeição da hipótese
  → nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Região crítica sob a curva da função densidade da distribuição apropriada ao teste

$P((-val. crít.),\infty)=(1-\alpha)$ em cinza (nível de confiança)
 $P(-\infty; (-val. crít.))=\alpha$ em vermelho

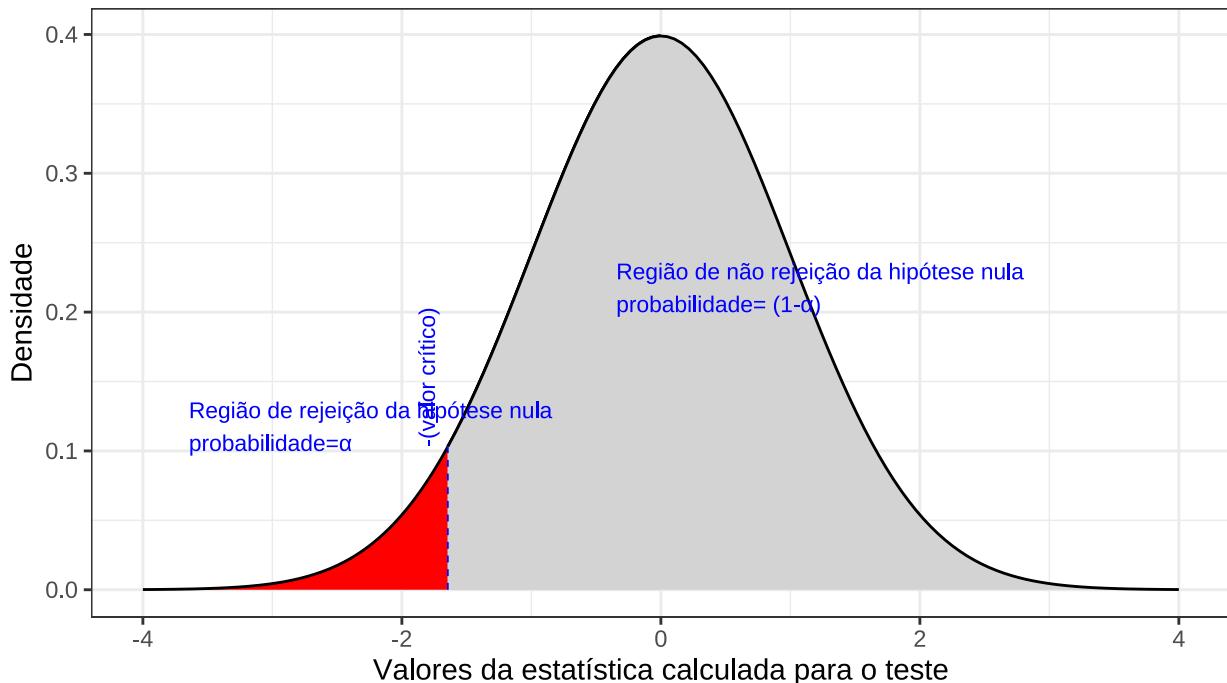


Figure 11.9: Região crítica aquém da qual a probabilidade associada aos valores amostrais observados é inferior a α , estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

Na Figura 11.9 observa-se:

- a região de rejeição da hipótese nula delimitada sob a curva da função densidade de probabilidade da distribuição adequada ao teste com probabilidade igual ao nível de significância α ;
- a região de não rejeição da hipótese nula (delimitada à esquerda) com probabilidade igual ao nível de confiança $(1 - \alpha)$; e,
- os valores críticos da estatística do teste.

11.7.4 Teste de hipóteses Unilateral à direita

Nesse tipo de teste a *hipótese alternativa* é proposta como a dizer que o valor em teste não apenas é diferente, mas é maior do que aquele afirmado pela *hipótese nula* (conservadora):

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

em que μ é o valor conservador do parâmetro que se deseja testar frente ao valor alternativo μ_0 .

```
alfa=0.95
prob_desejada=alfa
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-4, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores da estatística calculada para o teste") +
  labs(title=
      "Região crítica sob a curva da função densidade da \ndistribuição apropriada ao
      teste",
      subtitle = "P(-\U221e, (val. crít.))=(1-\u03b1) em cinza (nível de confiança)
      \nP((val.critic.); \U221e)= \u03b1 em vermelho "+
```

```

geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue",
  ↵ lty=2, lwd=0.3) +
annotate(geom="text", x=z_desejado+0.3, y=d_desejada, label="(valor crítico)", angle=90,
  ↵ vjust=0, hjust=0, color="blue",size=3) +
annotate(geom="text", x=z_desejado+0.5, y=0.1, label="Região de rejeição da hipótese nula
  ↵ \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  ↵ annotate(geom="text", x=z_desejado-2.5, y=0.2, label="Região de não rejeição da hipótese
  ↵ nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3) +
theme_bw()

```

Região crítica sob a curva da função densidade da distribuição apropriada ao teste

$P(-\infty, (\text{val. crít.})) = (1-\alpha)$ em cinza (nível de confiança)
 $P((\text{val. crít.}); \infty) = \alpha$ em vermelho

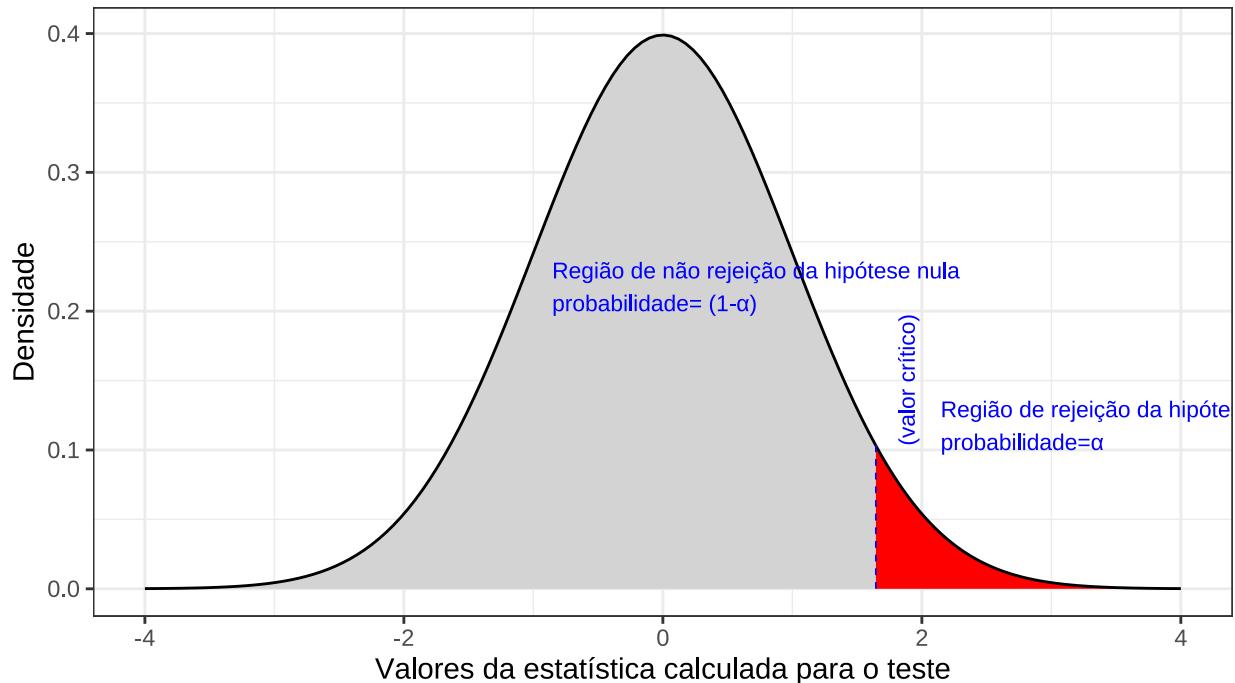


Figure 11.10: Região crítica além da qual a probabilidade associada aos valores amostrais observados é inferior a α , estabelecendo assim um intervalo com nível de confiança igual a $(1 - \alpha)$

Na Figura 11.10 observa-se:

- a região de rejeição da hipótese nula delimitada sob a curva da função densidade de probabilidade da distribuição adequada ao teste com probabilidade igual ao nível de significância α ;
- a região de não rejeição da hipótese nula (delimitada à direita) com probabilidade igual ao nível de confiança $(1 - \alpha)$; e,
- os valores críticos da estatística do teste.

11.8 Teste de uma média amostral

11.8.1 Cenários possíveis

- variância populacional (σ^2) *teoricamente conhecida*;
- variância populacional (σ^2) desconhecida, mas o tamanho da amostra (n) é grande: $n \geq 30(40)$; e,
- variância populacional (σ) desconhecida e as amostras de tamanho (n) reduzido: $n < 30$.

Estatística do teste para a primeira situação: variância populacional conhecida

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

em que:

- \bar{X} é a média observada na amostra;
- μ o valor (desconhecido) inferido à média populacional, a ser testado frente à média amostral observada;
- σ é o desvio padrão populacional; e,
- n é o tamanho da amostra.

Estatística do teste para a segunda situação: variância populacional desconhecida mas amostras grandes: $n \geq 30(40)$: S pode ser tomado como estimativa de σ :

$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

em que:

- \bar{X} é a média observada na amostra;
- μ o valor (desconhecido) inferido à média populacional a ser testado frente à média amostral observada;
- S é o desvio padrão amostral; e,
- n é o tamanho da amostra.

Estatística do teste para a terceira situação: variância populacional desconhecida e amostras pequenas: $n < 30$:

$$T = \frac{(\bar{X} - \mu)}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

em que:

- \bar{X} é a média observada na amostra;
- μ o valor (desconhecido) inferido à média populacional, a ser testado frente à média amostral;
- S é o desvio padrão amostral; e,
- n é o tamanho da amostra.

```
# Definição do eixo x
x <- seq(-4, 4, length.out = 100)

# Densidade da distribuição normal padrão
y_norm <- dnorm(x, mean = 0, sd = 1)

# Lista com diferentes graus de liberdade
df_list=c(1, 2, 4, 8, 20)

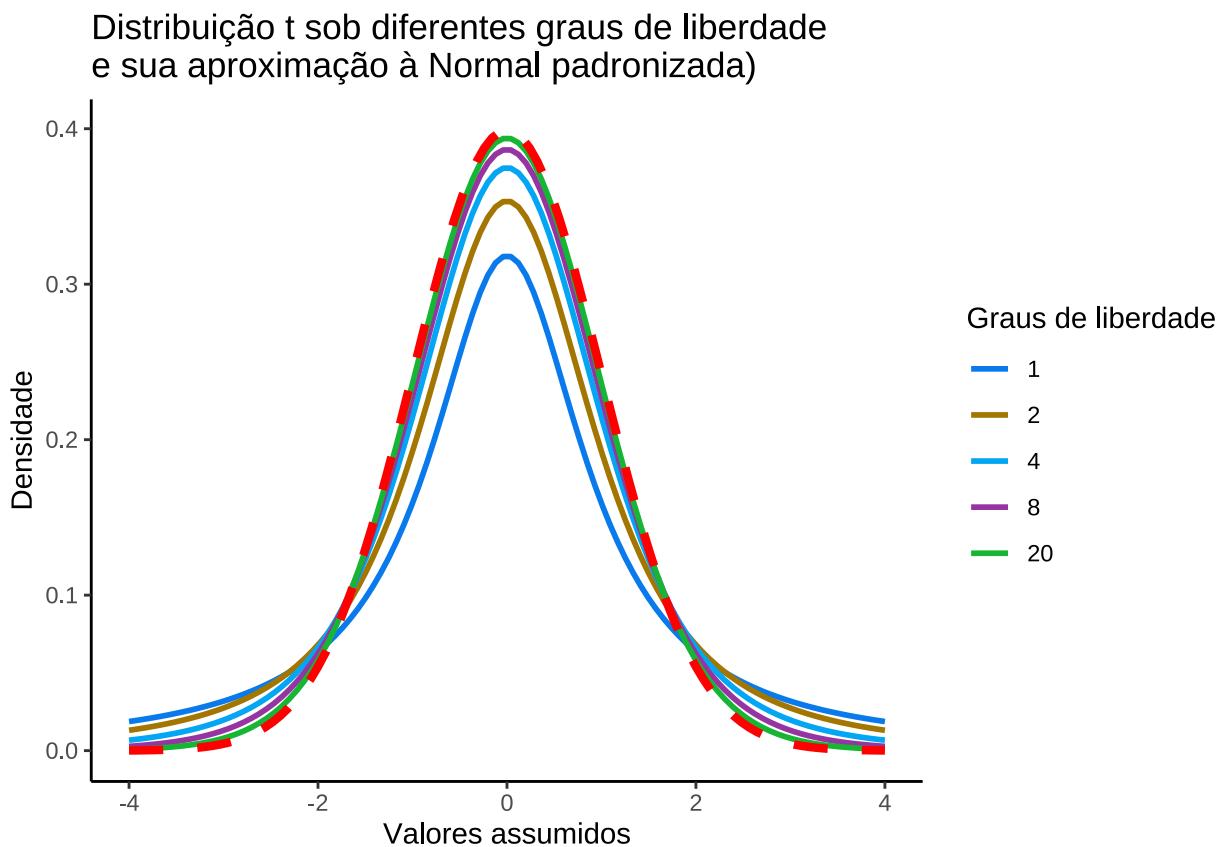
# Lista com cores para as curvas da distribuição t
colors=c("#097aeb", "#a37602", "#02a6f2", "#9635a1", "#16b533")
```

```

# Criação do data frame com todas as curvas
data=data.frame()
for (i in seq_along(df_list)) {
  df = df_list[i]
  y_t = dt(x, df)
  df_data = data.frame(x, y_t, df)
  data = rbind(data, df_data)
}

# Plotagem do gráfico
p = ggplot(data, aes(x = x)) +
  geom_line(aes(y = y_t, color = factor(df)), size = 1) +
  scale_color_manual(values = colors, name = "Graus de liberdade")+
  ggtitle("Distribuição t sob diferentes graus de liberdade \ne sua aproximação à Normal
  ↪ padronizada") +
  xlab("Valores assumidos") +
  ylab("Densidade") +
  theme_classic() +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), color = "red", size=1.5,
  ↪ linetype='dashed')
print(p)

```



11.8.2 Roteiro geral

- identificar o modelo de probabilidade do estimador do parâmetro da população que se estuda;
- identificar a estatística apropriada para o teste em razão das informações disponíveis acerca da população, do tamanho da amostra e sua independência:
 - escore médio;
 - proporção;
 - estatísticas T, Z, F, ou χ^2 ;
- determinar na curva de densidade de probabilidade do modelo da estatística de teste a(s) região(ões) crítica(s): faixa(s) de valores da estatística que nos levam à rejeição ou não da hipótese H_0 em função do nível de significância previamente arbitrado pelo pesquisador α ;
- calcular a estatística do teste apropriada para o parâmetro que se pretende inferir com base na amostra extraída;
- concluir com base nos resultados analisados: se o valor da estatística do teste pertence à(s) região(ões) crítica(s) de sua distribuição teórica, rejeitar H_0 ; caso contrário não há evidências estatisticamente significativas para rejeitá-la.

11.8.3 Probabilidade dos intervalos de confiança para os testes de hipóteses com o uso da estatística Z ($Z \sim \mathcal{N}(0, 1)$):

- Teste de hipóteses bilateral (tipo: diferente de):

$$\begin{aligned} P[|Z_{calc}| \leq Z_{tab(\frac{\alpha}{2})} | \mu = \mu_0] &= (1 - \alpha) \\ P(-Z_{tab(\frac{\alpha}{2})} \leq Z_{calc} \leq Z_{tab(\frac{\alpha}{2})}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à esquerda (tipo: menor que):

$$\begin{aligned} P[Z_{calc} \geq -Z_{tab(\alpha)} | \mu \geq \mu_0] &= (1 - \alpha) \\ P(Z_{calc} \geq -Z_{tab(\alpha)}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à direita (tipo maior que):

$$\begin{aligned} P[Z_{calc} \leq Z_{tab(\alpha)} | \mu \leq \mu_0] &= (1 - \alpha) \\ P(Z_{calc} \leq Z_{tab(\alpha)}) &= (1 - \alpha) \end{aligned}$$

11.8.4 Probabilidade dos intervalos de confiança para os testes de hipóteses com o uso da estatística T ($T \sim t_{(n-1)}$):

- Teste de hipóteses bilateral (tipo: diferente de):

$$\begin{aligned} P[|t_{calc}| \geq t_{tab(\frac{\alpha}{2}; n-1)} | \mu = \mu_0] &= (1 - \alpha) \\ P(-t_{tab(\frac{\alpha}{2}; n-1)} \leq t_{calc} \leq t_{tab(\frac{\alpha}{2}; n-1)}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à esquerda (tipo: menor que):

$$\begin{aligned} P[t_{calc} \geq -t_{tab(\alpha)} | \mu \geq \mu_0] &= (1 - \alpha) \\ P(t_{calc} \geq -t_{tab(\alpha; n-1)}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à direita (tipo: maior que):

$$P[t_{calc} \leq t_{tab(\alpha)} | \mu \leq \mu_0] = (1 - \alpha)$$

$$P(t_{calc} \leq t_{tab(\alpha; n-1)}) = (1 - \alpha)$$

Exemplo: O tempo de vida útil de uma amostra de 100 lâmpadas fluorescentes produzidas por uma fábrica foi calculado resultando em uma vida útil média de 1570 h sob um desvio padrão de 120 h. Seja μ é o tempo de vida útil das lâmpadas produzidas pela empresa. Teste a hipótese de $\mu = 1600h$ contra a hipótese alternativa de $\mu \neq 1600h$ sob um nível de significância $\alpha = 0,05$.

O problema nos pede um teste bilateral (tipo: diferente de):

$$\begin{cases} H_0 : \mu = 1.600 \\ H_1 : \mu \neq 1.600 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a vida útil média das lâmpadas a 1.600 h., fazendo então valer a hipótese alternativa que afirma ser a vida útil das lâmpadas **diferente de** 1.600 h.

Pelo enunciado do problema a variância populacional σ^2 é desconhecida mas, como a amostra é de grande tamanho ($n=100$) podemos tomar S como uma estimativa de σ e a estatística do teste fica definida como sendo:

$$Z = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

Extraindo os dados do problema:

- $\bar{X} = 1570h$ é a média amostral;
- $\mu_0 = 1600$ o valor (desconhecido) inferido à média populacional a ser testado frente à média amostral;
- $S = 120h$ é o desvio padrão amostral; e,

- $n = 100$ é o tamanho da amostra.

Calculando-se o valor da estatística do teste:

$$z_{calc} = \frac{1570 - 1600}{\frac{120}{\sqrt{100}}} = -2,50$$

Da tabela da distribuição Normal reduzida obtemos o valor crítico bicaudal: $|z_{crit}| = 1,96$. Pelo cálculo, a estatística do teste é $z_{calc} = -2,50$.

```
alfa=0.05

prob_desejada1=alfa/2
z_desejado1=round(qnorm(prob_desejada1),4)
d_desejada1=dnorm(z_desejado1, 0, 1)

prob_desejada2=1-alfa/2
z_desejado2=round(qnorm(prob_desejada2),4)
d_desejada2=dnorm(z_desejado2, 0, 1)

z_calculado=-2.5
d_calculado=dnorm(z_calculado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de z", breaks = c(z_desejado1,z_desejado2)) +
  labs(title=
      "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
      teste",
      subtitle = "P(-1,96, 1,96)=(1-\u03b1) em cinza (nível de confiança=0,95) \nP(-\U221e;
      -1,96)= P(1,96; \U221e)= \u03b1/2 em vermelho (nível de significância/2=0,025)
      ")+
```

```

geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1),
  ↵ color="blue", lty=2, lwd=0.3)+
geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2),
  ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="valor crítico=-1,96",
  ↵ angle=90, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="valor crítico=1,96",
  ↵ angle=90, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado1-2, y=0.1, label="Região de rejeição da hipótese nula
  ↵ \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Região de rejeição da hipótese nula
  ↵ \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Região de não rejeição da hipótese
  ↵ nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
  ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do
  ↵ teste=-2,5", angle=90, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Regiões críticas sob a curva da função densidade da distribuição apropriada ao teste

$P(-1,96, 1,96) = (1-\alpha)$ em cinza (nível de confiança=0,95)

$P(-\infty; -1,96) = P(1,96; \infty) = \alpha/2$ em vermelho (nível de significância/2=0,025)

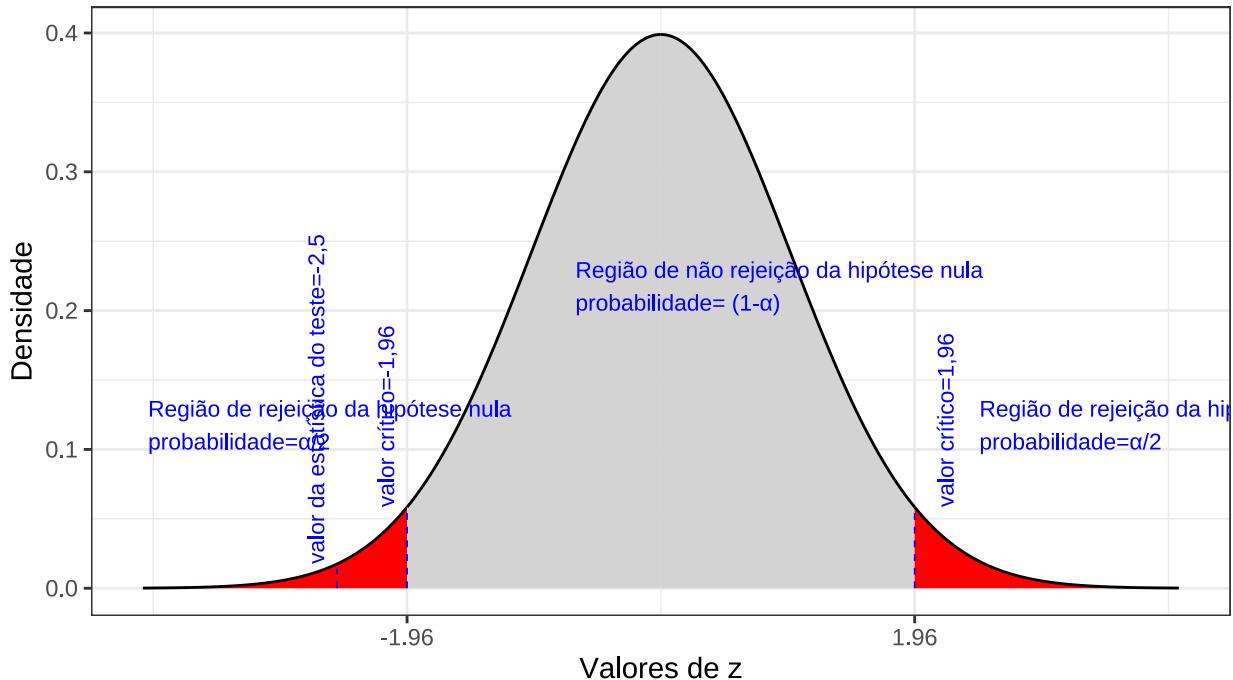


Figure 11.11: Regiões de rejeição da hipótese nula para o teste bilateral (tipo: diferente de) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valores críticos da estatística do teste: $z_{crit} = \pm 1,96$. O valor calculado da estatística ($z_{calc} = -2,50$) situa-se na faixa de significância do teste, possibilitando a rejeição da hipótese nula sob aquele nível de confiança

Conclusão: Os resultados obtidos na análise estatística realizada nos permitem rejeitar a hipótese de que a duração média populacional das lâmpadas seja igual a 1600h sob um nível de confiança de 95%. A vida útil média das lâmpadas é **diferente** de 1600h (Figura 11.11).

Podemos ainda realizar testes de hipóteses unilaterais ($\mu < \mu_0$ ou $\mu > \mu_0$).

Teste unilateral à esquerda (tipo: menor que)

$$\begin{cases} H_0 : \mu \geq 1.600 \\ H_1 : \mu < 1.600 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a vida útil média das lâmpadas igual ou superior a 1.600 h., fazendo então valer a hipótese alternativa que afirma ser a vida útil das lâmpadas **menor que** 1.600 h.

Da tabela da distribuição Normal reduzida obtemos o valor crítico monocaudal: $z_{crit} = -1,64$. Pelo cálculo, a estatística do teste é $z_{calc} = -2,50$.

```
alfa=0.05
prob_desejada=alfa
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

z_calculado=-2.5
d_calculado=dnorm(z_calculado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
```

```

xlim = c(z_desejado,0),
       colour="black") +
geom_area(stat = "function",
       fun = dnorm,
       fill = "lightgrey",
       xlim = c(0, z_desejado),
       colour="black") +
geom_area(stat = "function",
       fun = dnorm,
       fill = "lightgrey",
       xlim = c(z_desejado,4),
       colour="black") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores da estatística calculada para o teste") +
labs(title =
  "Região crítica sob a curva da função densidade da \ndistribuição apropriada ao
  ↵ teste",
  subtitle = "P( -1,64,\U221e,)=(1-\u03b1) em cinza (nível de confiança=0,95)
  ↵ \nP(-\U221e; -1,64)=\u03b1 em vermelho (nível de significância=0,05)")+
geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue",
  ↵ lty=2, lwd=0.3)+
annotate(geom="text", x=z_desejado-0.1, y=d_desejada, label="valor crítico=-1,64", angle=90,
  ↵ vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado-2.5, y=0.1, label="Região de rejeição da hipótese nula
  ↵ \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado+1, y=0.2, label="Região de não rejeição da hipótese nula
  ↵ \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
  ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do
  ↵ teste=-2,5", angle=90, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Conclusão: Os resultados obtidos na análise estatística realizada nos permitem rejeitar a hipótese de que a duração média populacional das lâmpadas seja igual ou superior a 1600h sob um nível de confiança de 95%. A vida útil média é **menor que** 1600h (Figura 11.12).

Teste unilateral à direita (tipo: maior que)

$$\begin{cases} H_0 : \mu \leq 1.600 \\ H_1 : \mu > 1.600 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a vida útil média das lâmpadas igual ou inferior a 1.600 h., fazendo então valer a hipótese alternativa que afirma ser a vida útil das lâmpadas **maior que** 1.600 h.

Região crítica sob a curva da função densidade da distribuição adequada ao teste

$P(-1,64, \infty) = 1 - \alpha$ em cinza (nível de confiança = 0,95)
 $P(-\infty; -1,64) = \alpha$ em vermelho (nível de significância = 0,05)

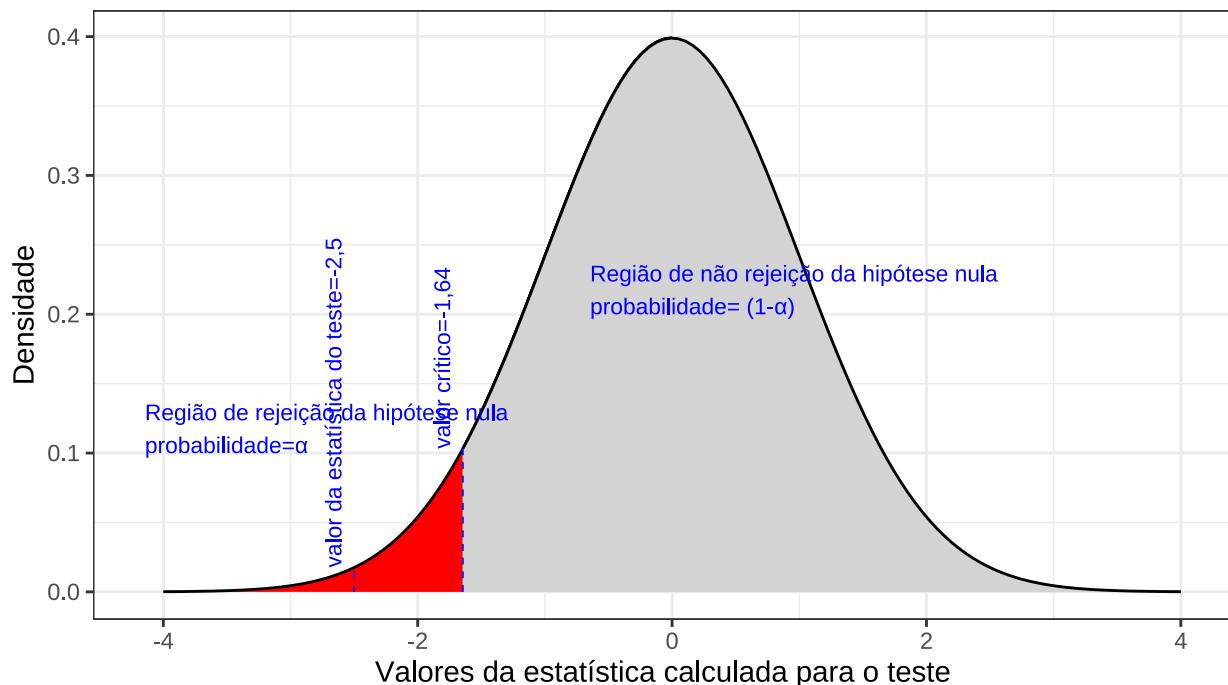


Figure 11.12: Região de rejeição da hipótese nula para o teste unilateral à esquerda (tipo: menor que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $z_{crit} = -1,64$. O valor calculado da estatística ($z_{calc} = -2,50$) situa-se na faixa de significância do teste, possibilitando a rejeição da hipótese nula sob aquele nível de confiança

Da tabela da distribuição Normal reduzida obtemos o valor crítico moncaudal: $z_{crit} = 1,64$. Pelo cálculo, a estatística do teste é $z_{calc} = -2,50$.

```

alfa=0.95
prob_desejada=alfa
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

z_calculado=-2.5
d_calculado=dnorm(z_calculado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-4, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores da estatística calculada para o teste") +
  labs(title=
      "Região crítica sob a curva da função densidade da \ndistribuição apropriada ao
       \teste",
       subtitle = "P( -1,96,\U221e,)=(1-\u03b1) em cinza (nível de confiança=0,95)
                  \nP(-\U221e; -1,96)=\u03b1 em vermelho (nível de significância=0,05) ")+
  geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue",
               lty=2, lwd=0.3)+
  annotate(geom="text", x=z_desejado-0.1, y=d_desejada, label="valor crítico=-1,64", angle=90,
           vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado+1, y=0.1, label="Região de rejeição da hipótese nula
           \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado-2.5, y=0.2, label="Região de não rejeição da hipótese
           nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do
           teste=-2,5", angle=90, vjust=0, hjust=0, color="blue",size=3)+

theme_bw()

```

Conclusão: Os resultados obtidos na análise estatística realizada não nos permitem rejeitar a hipótese de que a duração média populacional das lâmpadas seja igual ou inferior a 1600h sob um nível de confiança de 95%. A vida útil média é maior que 1600h (Figura 11.12).

Região crítica sob a curva da função densidade da distribuição adequada ao teste

$P(-1,96, \infty) = 1-\alpha$ em cinza (nível de confiança=0,95)
 $P(-\infty; -1,96) = \alpha$ em vermelho (nível de significância=0,05)

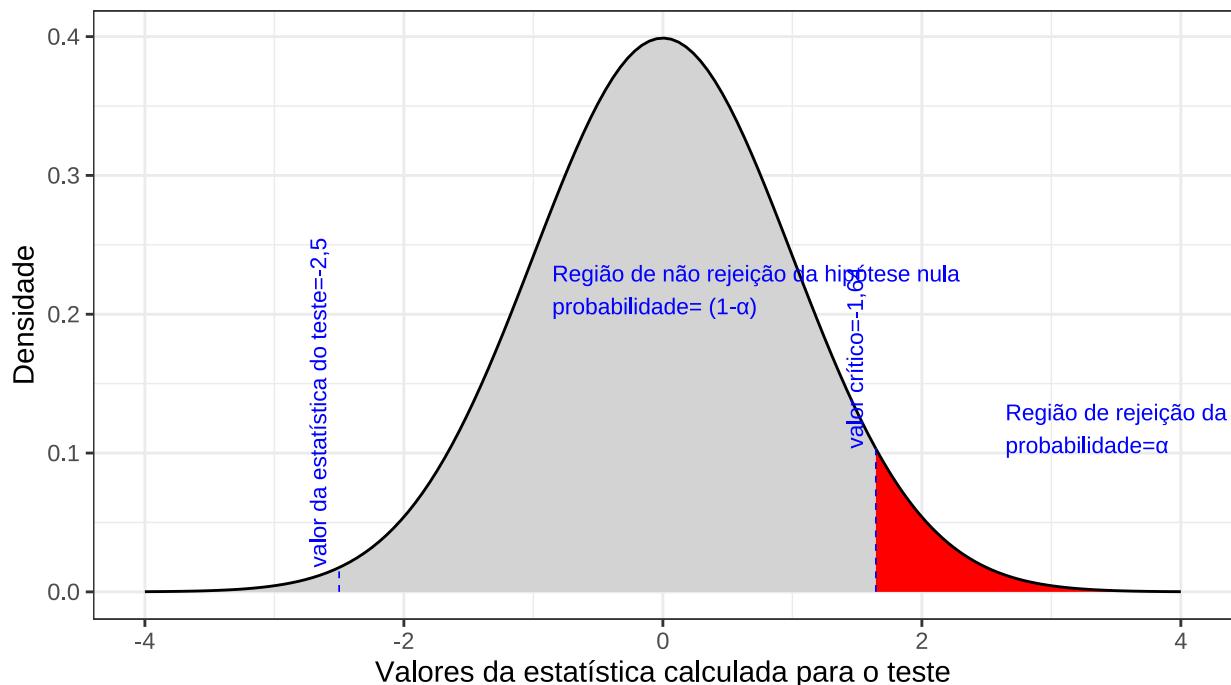


Figure 11.13: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: maior que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $z_{crit} = 1,64$. O valor calculado da estatística ($z_{calc} = -2,50$) situa-se na faixa de não significância do teste, não possibilitando a rejeição da hipótese nula sob aquele nível de confiança

Exemplo: De um universo Normal com parâmetros média e variância (μ e σ^2) desconhecidos, retirou-se uma amostra aleatória composta por 9 observações que apresentou as seguintes sínteses numéricas: $\bar{X} = 4$ e $S^2 = 2,2$. Proceda ao seguinte teste de hipóteses, a um nível de significância: $\alpha = 0,05$, de que a média populacional é igual a 5.

O problema nos pede um teste bilateral (tipo: diferente de):

$$\begin{cases} H_0 : \mu = 5 \\ H_1 : \mu \neq 5 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a média igual a 5, fazendo então valer a hipótese alternativa que afirma ser a média **diferente de 5**.

Pelo enunciado do problema a variância populacional σ^2 é desconhecida e a amostra é pequena ($n=9$). Nessa situação, a estatística do teste fica definida como sendo:

$$T = \frac{(\bar{X} - \mu_0)}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$$

Extraindo os dados do problema:

- $\bar{x} = 4$ é a média amostral;
- $\mu_0 = 5$ o valor (desconhecido) inferido à média populacional, a ser testado frente à média amostral;
- $s = \sqrt{2,2} = 1,48$ é o desvio padrão da amostra extraída;
- $n = 9$ é o tamanho da amostra extraída;

Calculando-se o valor da estatística do teste:

$$t_{calc} = \frac{(X - \mu_0)}{\frac{s}{\sqrt{n}}} = -2,02$$

Da tabela “t’’ de Student obtemos o valor crítico bicaudal: $|t_{tab(\frac{\alpha}{2},(n-1)}| = 2,306$. Pelo cálculo a estatística do teste é $t_{calc} = -2,02$.

```
alfa=0.05

prob_desejada1=alfa/2
df=8
t_desejado1=round(qt(prob_desejada1,df ),df)
d_desejada1=dt(t_desejado1,df)

prob_desejada2=1-alfa/2
df=8
t_desejado2=round(qt(prob_desejada2, df),df)
d_desejada2=dt(t_desejado2,df)

t_calculado=-2
d_calculado=dt(t_calculado,df)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(-4, t_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(t_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(0, t_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(t_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
```

```

scale_x_continuous(name="Valores de t", breaks = c(t_desejado1, t_desejado2)) +
  labs(title =
    "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
    ← teste",
    subtitle = "P(-2,306, 2,306)=(1-\u03b1) em cinza (nível de confiança=0,95)
    ← \nP(-\U221e; -2,306)= P(2,306; \U221e)= \u03b1/2 em vermelho (nível de
    ← significância/2=0,025)") + geom_segment(aes(x = t_desejado1, y = 0, xend =
    ← t_desejado1, yend = d_desejada1), color="blue", lty=2, lwd=0.3) +
  geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2),
    ← color="blue", lty=2, lwd=0.3) +
  annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="valor crítico=-2,306",
    ← angle=90, vjust=0, hjust=0, color="blue", size=3) +
  annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="valor crítico=2,306",
    ← angle=90, vjust=0, hjust=0, color="blue", size=3) +
  annotate(geom="text", x=t_desejado1-2, y=0.1, label="Região de rejeição da hipótese nula
    ← \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue", size=3) +
  annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Região de rejeição da hipótese nula
    ← \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue", size=3) +
  annotate(geom="text", x=t_desejado1+2, y=0.2, label="Região de não rejeição da hipótese
    ← nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue", size=3) +
  geom_segment(aes(x = t_calculado, y = 0, xend = t_calculado, yend = d_calculado),
    ← color="blue", lty=2, lwd=0.3) +
  annotate(geom="text", x=t_calculado-0.1, y=d_calculado, label="valor da estatística do
    ← teste=-2.02", angle=90, vjust=0, hjust=0, color="blue", size=3) +
  theme_bw()

```

Conclusão: Os resultados obtidos na análise estatística realizada não nos permitem rejeitar a hipótese de que a média populacional seja igual a 5 sob um nível de confiança de 95% (Figura 11.14).

```

# Dados do problema
n=9
media_amostral=4
var_amostral=2.2
media_populacao=5
alfa=0.05

# Estatística de teste
t=(media_amostral - media_populacao) / sqrt(var_amostral / n)

# Graus de liberdade
df=n - 1

# Valor-p à esquerda
p_valor_1=pt(-abs(t), df, lower.tail = TRUE)

# Valor-p à direita
p_valor_2=pt(abs(t), df, lower.tail = FALSE)

# p-valor
p_valor=p_valor_1+p_valor_2

```

Regiões críticas sob a curva da função densidade da distribuição adequada ao teste

$P(-2,306, 2,306) = (1-\alpha)$ em cinza (nível de confiança=0,95)
 $P(-\infty; -2,306) = P(2,306; \infty) = \alpha/2$ em vermelho (nível de significância/2=0,025)

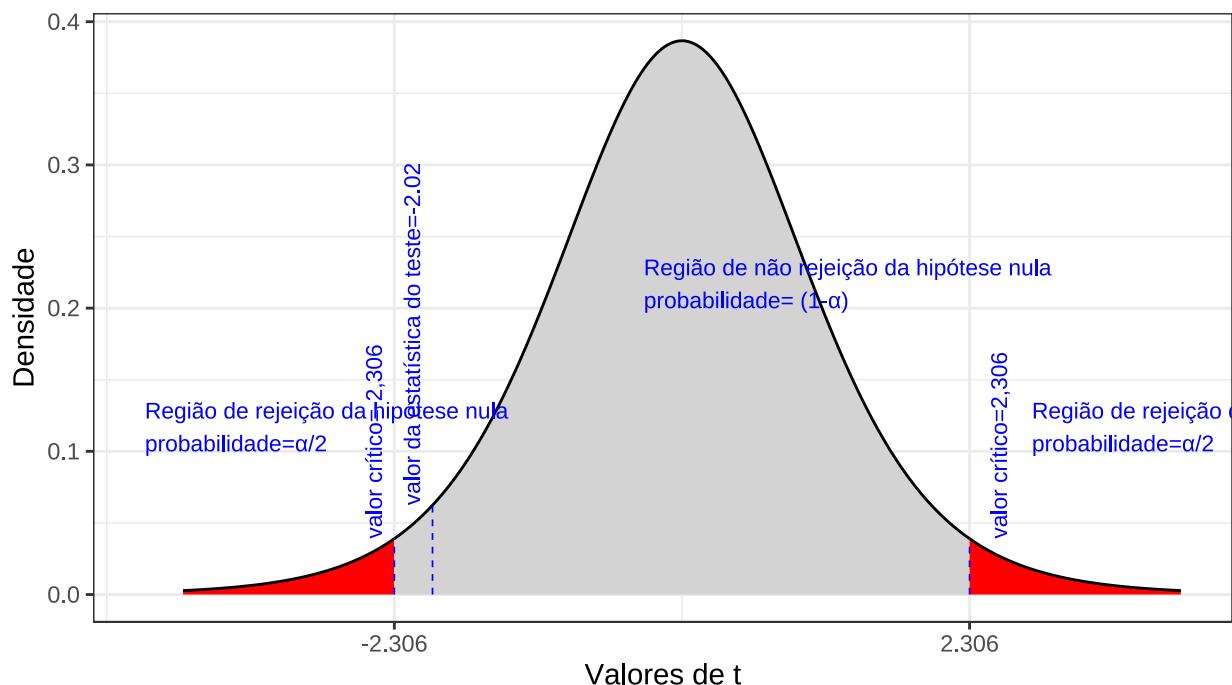


Figure 11.14: Regiões de rejeição da hipótese nula para o teste bilateral (tipo: diferente de) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valores críticos da estatística do teste: $t_{crit} = \pm 2,306$. O valor calculado da estatística ($t_{calc} = -2,02$) situa-se na faixa de significância do teste, possibilitando a rejeição da hipótese nula sob aquele nível de confiança

```
# Ou
p_valor <- 2 * pt(-abs(t), df)

# Decisão e conclusão
if (p_valor < alfa) {
  cat("Os dados amostrais trazidos à análise nos permitem rejeitar, sob o nível de
      ↵ significância estabelecido de", alfa, "de se cometer um erro do tipo I, a hipótese
      ↵ nula ( $H_0$ ) que afirma ser a média populacional igual a", media_populacao, ". A média
      ↵ populacional é diferente.")
} else {
  cat("Os dados amostrais trazidos à análise não nos permitem rejeitar, sob o nível de
      ↵ confiança de", 1-alfa, ", a hipótese nula ( $H_0$ ). A média populacional é igual a",
      ↵ media_populacao, ".")
}

## Os dados amostrais trazidos à análise não nos permitem rejeitar, sob o nível de confiança de 0.95 , a
```

> Teste unilateral à esquerda (tipo: menor que)

$$\begin{cases} H_0 : \mu \geq 5 \\ H_1 : \mu < 5 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a média igual ou maior a 5, fazendo então valer a hipótese alternativa que afirma ser a média **menor que** 5.

Da tabela “t” de Student obtemos o valor crítico monocaudal: $|t_{tab,\alpha,(n-1)}| = -1,86$. Pelo cálculo a estatística do teste é $t_{calc} = -2,02$.

```
alfa=0.05
prob_desejada=alfa
df=8
t_desejado=round(qt(prob_desejada,df ),4)
d_desejada=dt(t_desejado,df)

t_calculado=-2
d_calculado=dt(t_calculado,df)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
```

```

fill = "red",
xlim = c(-4, t_desejado),
colour="black") +
geom_area(stat = "function",
  fun = dt,
  args=list(df),
  fill = "lightgrey",
  xlim = c(t_desejado,4),
  colour="black") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores de t", breaks = c(t_desejado)) +
labs(title =
  "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
  → teste",
  subtitle = "P(-1,86, \U221e)=(1-\u03b1) em cinza (nível de confiança=0,95)
  → \nP(-\U221e; -1,86)= \u03b1 em vermelho (nível de significância=0,05)")+
geom_segment(aes(x = t_desejado, y = 0, xend = t_desejado, yend = d_desejada),
  color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=t_desejado-0.1, y=d_desejada, label="valor crítico=-1,86",
  angle=90, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=t_desejado-2, y=0.1, label="Região de rejeição da hipótese nula
  → \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=t_desejado+1.5, y=0.2, label="Região de não rejeição da hipótese
  → nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
geom_segment(aes(x = t_calculado, y = 0, xend = t_calculado, yend = d_calculado),
  color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=t_calculado-0.1, y=d_calculado, label="valor da estatística do
  → teste=-2.02", angle=90, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Conclusão: sob um nível de confiança de 95%, face aos dados trazidos à análise podemos rejeitar a hipótese de que a média seja de no mínimo a 5 (Figura 11.15).

Caso estabelecessemos um nível de confiança $(1 - \alpha) \geq 0,9611277$ (ou tivéssemos uma informação amostral $x \geq 4.080639$), a hipótese nula **não** seria rejeitada: a média populacional é maior ou igual a 5.

```

# Dados do problema
n=9
media_amostral=4
var_amostral=2.2
media_populacao=5
alfa=0.05

# Estatística de teste
t=(media_amostral - media_populacao) / sqrt(var_amostral / n)

# Graus de liberdade
df=n - 1

```

Regiões críticas sob a curva da função densidade da distribuição adequada ao teste

$P(-1,86, \infty) = (1-\alpha)$ em cinza (nível de confiança=0,95)
 $P(-\infty; -1,86) = \alpha$ em vermelho (nível de significância=0,05)

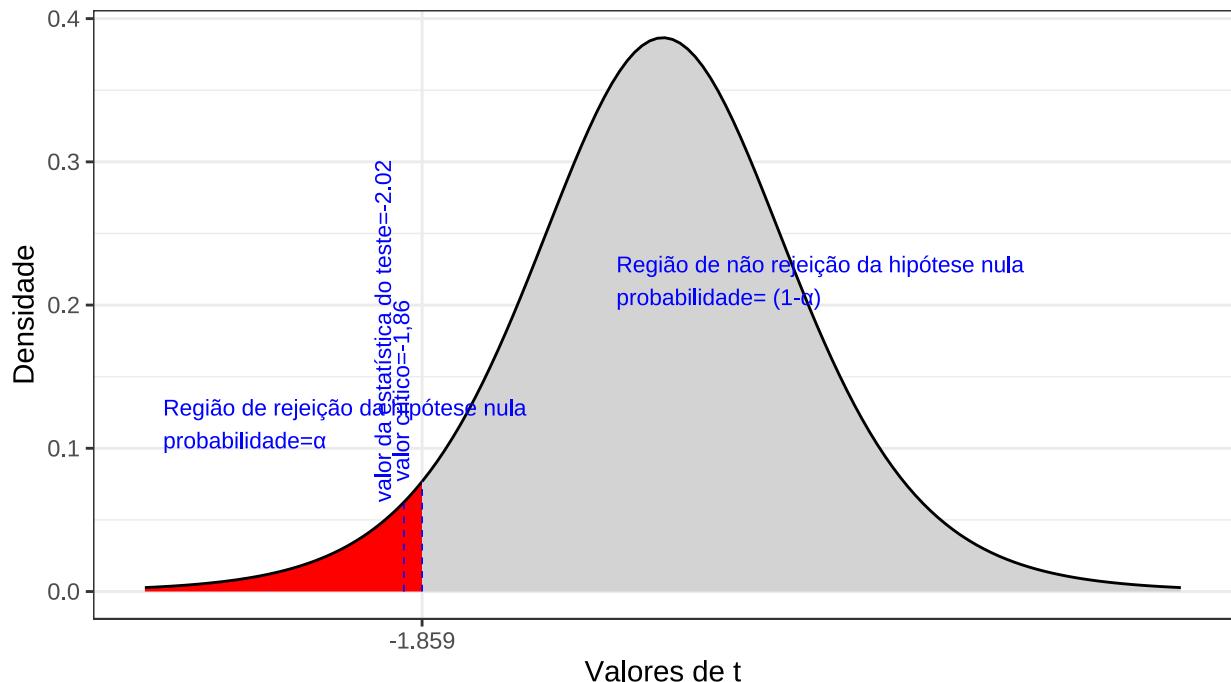


Figure 11.15: Região de rejeição da hipótese nula para o teste unilateral à esquerda (tipo: menor que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $t_{crit} = -1,86$. O valor calculado da estatística ($t_{calc} = -2,02$) situa-se na faixa de significância do teste possibilitando a rejeição da hipótese nula sob aquele nível de confiança

```

# Valor-p à esquerda
p_valor=pt(t, df)

# Decisão e conclusão
if (p_valor < alfa) {
  cat("Os dados amostrais trazidos à análise nos permitem rejeitar, sob o nível de
      ↵ significância estabelecido de", alfa , "de se cometer um erro do tipo I, a hipótese
      ↵ nula (H0) que afirma ser a média populacional maior ou igual a ", media_populacao,".A
      ↵ média populacional é menor.")
} else {
  cat("Os dados amostrais trazidos à análise não nos permitem rejeitar, sob o nível de
      ↵ confiança de", 1-alfa , "a hipótese nula (H0). A média populacional é maior ou igual
      ↵ a", media_populacao,".")
}

## Os dados amostrais trazidos à análise nos permitem rejeitar, sob o nível de significância estabeleci

```

Teste unilateral à direita (tipo: maior que)

$$\begin{cases} H_0 : \mu \leq 5 \\ H_1 : \mu > 5 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a média igual ou menor a 5, fazendo então valer a hipótese alternativa que afirma ser a média **maior que** 5.

Da tabela “t” de Student obtemos o valor crítico monocaudal: $|t_{tab, \alpha, (n-1)}| = 1,86$. Pelo cálculo a estatística do teste é $t_{calc} = -2,02$.

```

alfa=0.95
prob_desejada=alfa
df=8
t_desejado=round(qt(prob_desejada,df ),4)
d_desejada=dt(t_desejado,df)

t_calculado=-2
d_calculado=dt(t_calculado,df)

```

```

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(-4, t_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(t_desejado,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de t", breaks = c(t_desejado)) +
  labs(title =
      "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
      ← teste",
      subtitle = "P(-\u2212e; 1,86)=(1-\u03b1) em cinza (nível de confiança=0,95) \nP(1,86;
      ← \U2212e)= \u03b1 em vermelho (nível de significância=0,05) ")+
  geom_segment(aes(x = t_desejado, y = 0, xend = t_desejado, yend = d_desejada),
               color="blue", lty=2, lwd=0.3)+  

  annotate(geom="text", x=t_desejado-3, y=0.1, label="Região de não rejeição da hipótese
      ← nula \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+  

  annotate(geom="text", x=t_desejado, y=0.1, label="Região de rejeição da hipótese nula
      ← \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+  

  geom_segment(aes(x = t_calculado, y = 0, xend = t_calculado, yend = d_calculado),
               color="blue", lty=2, lwd=0.3)+  

  annotate(geom="text", x=t_calculado-0.1, y=d_calculado, label="valor da estatística do
      ← teste=-2.02", angle=90, vjust=0, hjust=0, color="blue",size=3)+  

  theme_bw()

```

Conclusão: sob um nível de confiança de confiança de 95%, face aos dados trazidos à análise não podemos rejeitar a hipótese de que a média seja inferior a 5 (Figura 11.16).

Caso estabelecessemos um nível de confiança $(1 - \alpha) \geq 0,9611277$ (ou tivéssemos uma informação amostral $\bar{x} \geq 5.919361$), a hipótese nula **seria** rejeitada: a média populacional é maior que 5.

```

# Dados do problema
n=9
media_amostral=4
var_amostral=2.2
media_populacao=5
alfa=0.95

# Estatística de teste
t=(media_amostral - media_populacao) / sqrt(var_amostral / n)

```

Regiões críticas sob a curva da função densidade da distribuição adequada ao teste

$P(-\infty; 1,86) = (1-\alpha)$ em cinza (nível de confiança=0,95)
 $P(1,86; \infty) = \alpha$ em vermelho (nível de significância=0,05)

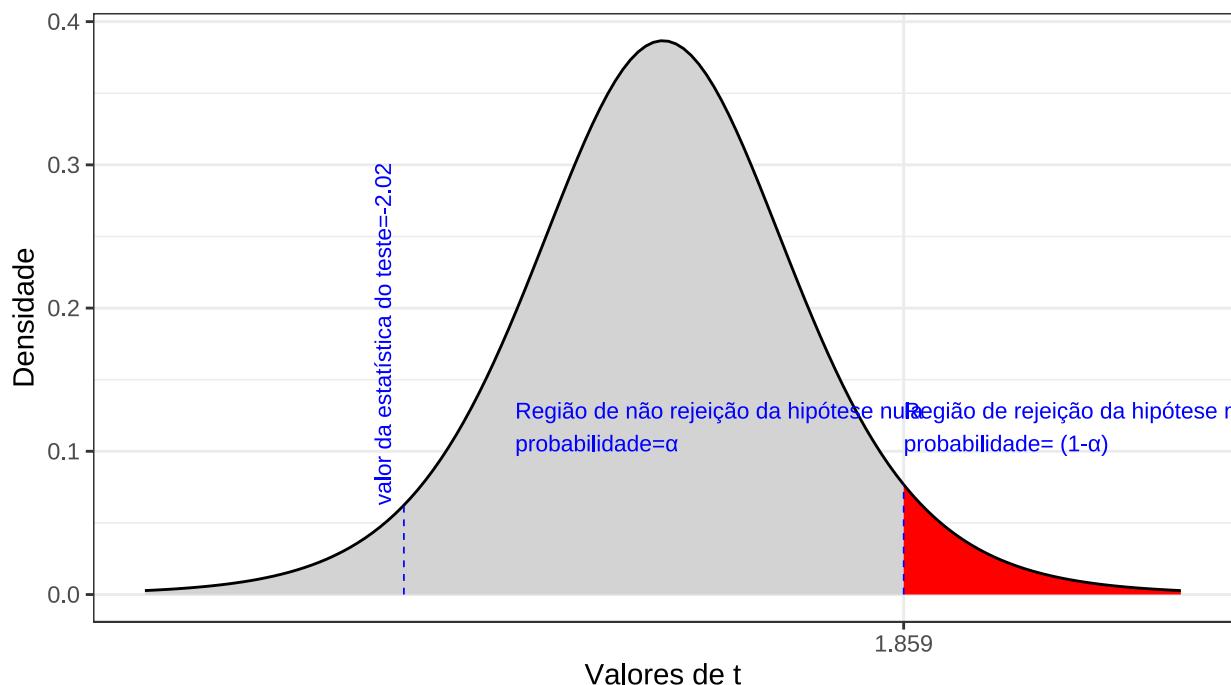


Figura 11.16: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: maior que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $t_{crit} = 1,86$. O valor calculado da estatística ($t_{calc} = -2,02$) situa-se na faixa de não significância do teste, não possibilitando a rejeição da hipótese nula sob aquele nível de confiança

```

# Graus de liberdade
df=n - 1

# Valor-p à direita
p_valor=pt(-t, df)

# Decisão e conclusão
if (p_valor < alfa) {
  cat("Os dados amostrais trazidos à análise nos permitem rejeitar, sob o nível de
      ↵ significância estabelecido de", alfa , "de se cometer um erro do tipo I, a hipótese
      ↵ nula ( $H_0$ ) que afirma ser a média populacional menor ou igual a", media_populacao,". A
      ↵ média populacional é maior que",media_populacao,"." )
} else {
  cat("Os dados amostrais trazidos à análise não nos permitem rejeitar, sob o nível de
      ↵ confiança de", 1-alfa , "a hipótese nula ( $H_0$ ). A média populacional é menor ou igual
      ↵ a", media_populacao,"." )
}

## Os dados amostrais trazidos à análise não nos permitem rejeitar, sob o nível de confiança de 0.05 ,a

```

11.9 Teste de médias amostrais independentes de duas populações Normais

Pelo Teorema Limite Central, para tamanhos amostrais n suficientemente grandes a média amostral \bar{X} tem distribuição aproximadamente Normal, com média μ e variância $\frac{\sigma^2}{n}$, independente da distribuição da população, onde μ e σ^2 são a média e a variância populacionais.

- grandes: $n \geq 30(40)$; e
- pequenas: $n < 30$.

Situações possíveis:

- Variâncias populacionais conhecidas ou não conhecidas mas com amostras de grande tamanho;

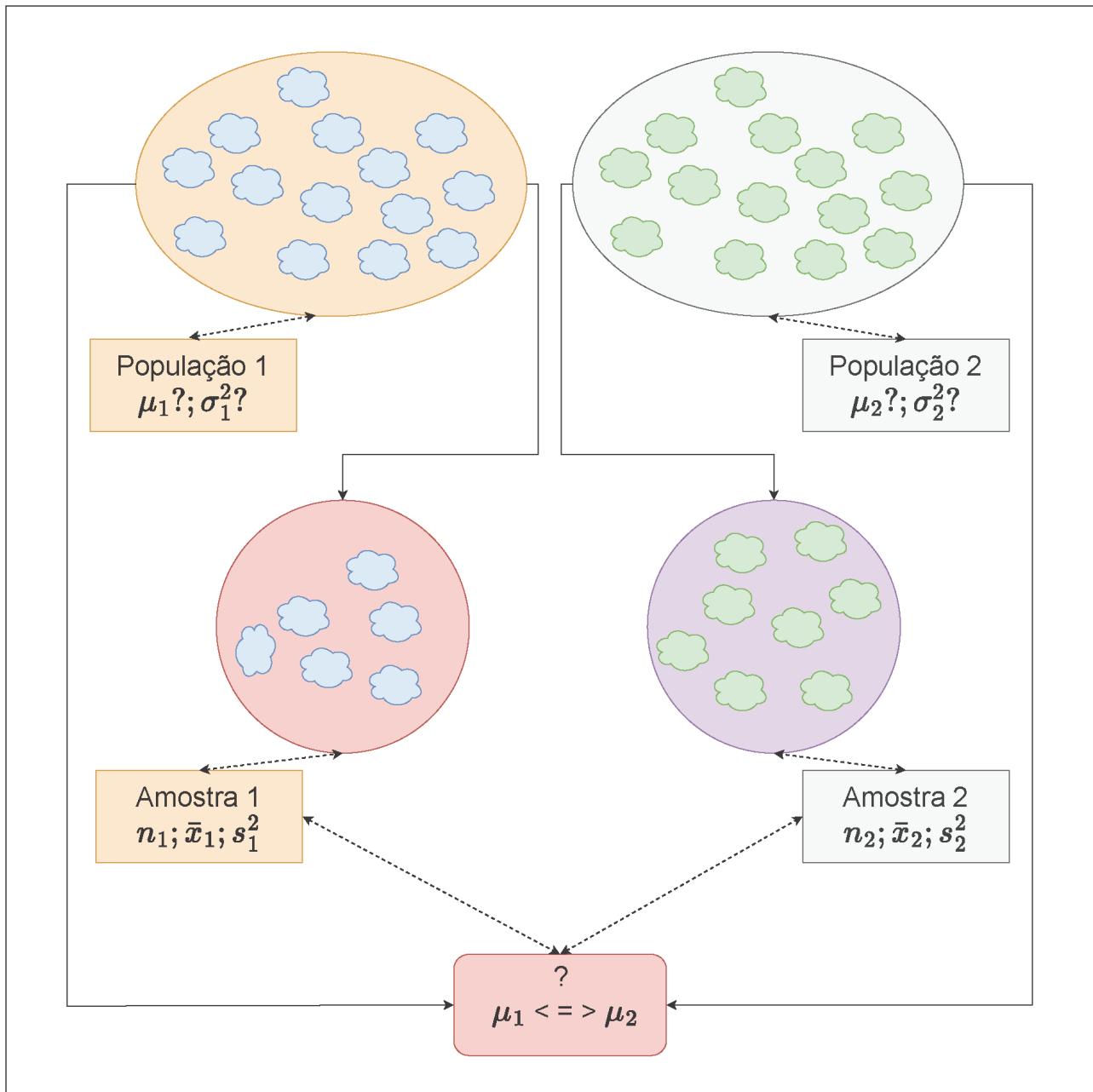


Figure 11.17: Visão esquemática das amostras de duas populações

- Variâncias populacionais desconhecidas:
 - Variâncias populacionais admitidas iguais; ou,
 - Variâncias populacionais quaisquer.

Os valores assumidos pelas características de nosso interesse nas populações são tais que:

$$X_1 \sim \mathcal{N}(\mu_1; \sigma_1^2)$$

e

$$X_2 \sim \mathcal{N}(\mu_2; \sigma_2^2)$$

Ao se extrair duas amostras, os valores amostrais assumidos por essas características serão duas variáveis aleatórias tais que:

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1; \frac{\sigma_1^2}{n_1}\right)$$

e

$$\bar{X}_2 \sim \mathcal{N}\left(\mu_2; \frac{\sigma_2^2}{n_2}\right).$$

É de nosso particular interesse definir uma variável aleatória expressa como a diferença das variáveis \bar{X}_1 e \bar{X}_2 .

Segue-se assim (por serem independentes) que

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

11.9.1 As estruturas possíveis dos testes de hipóteses relacionados às suas médias serão:

Teste bilateral (tipo: diferente de)

$$\begin{cases} H_0 : \mu_1 - \mu_2 = \Delta_0 \\ H_1 : \mu_1 - \mu_2 \neq \Delta_0 \end{cases}$$

Teste unilateral à esquerda (tipo: menor que)

$$\begin{cases} H_0 : \mu_1 - \mu_2 \geq \Delta_0 \\ H_1 : \mu_1 - \mu_2 < \Delta_0 \end{cases}$$

Teste unilateral à direita (tipo: maior que)

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq \Delta_0 \\ H_1 : \mu_1 - \mu_2 > \Delta_0 \end{cases}$$

Os valores assumidos pelas diferenças amostrais são tais que:

$$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

para

- amostras Normais: n_1 e n_2 qualquer;
- amostras sob outras distribuições, desde que: n_1 e $n_2 \geq 30(40)$:
- $Z_{tab(\frac{\alpha}{2})}$ ou $Z_{tab(\alpha)}$: valores da distribuição Normal padronizada para o nível de significância pretendido no teste (bilateral ou unilateral); e,
- $Z_{calc} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1)$

em que:

- Δ_0 é o valor inferido à diferença das médias populacionais μ_1 e μ_2 , usualmente 0 (igualdade);
- σ_1^2 é a variância da população 1;
- σ_2^2 é a variância da população 2;
- \bar{x}_1, n_1 são a média e o tamanho da amostra 1; e,
- \bar{x}_2, n_2 são a média e o tamanho da amostra 2.

11.9.2 Testes de hipóteses para as médias de duas populações com variâncias conhecidas (ou não conhecidas mas o tamanho das amostras é grande)

Probabilidade dos intervalos de confiança para os testes de hipóteses com o uso da estatística Z ($Z \sim \mathcal{N}(0, 1)$):

- Teste de hipóteses bilateral (tipo: diferente de):

$$P[|Z_{calc}| \leq Z_{tab(\frac{\alpha}{2})} | \mu_1 = \mu_2] = (1 - \alpha)$$

$$P(-Z_{tab(\frac{\alpha}{2})} \leq Z_{calc} \leq Z_{tab(\frac{\alpha}{2})}) = (1 - \alpha)$$

- Teste de hipóteses unilateral à esquerda (tipo: menor que):

$$P[Z_{calc} \geq -Z_{tab(\alpha)} | \mu_1 \geq \mu_2] = (1 - \alpha)$$

$$P(Z_{calc} \geq -Z_{tab(\alpha)}) = (1 - \alpha)$$

- Teste de hipóteses unilateral à direita (tipo maior que):

$$P[Z_{calc} \leq Z_{tab(\alpha)} | \mu_1 \leq \mu_2] = (1 - \alpha)$$

$$P(Z_{calc} \leq Z_{tab(\alpha)}) = (1 - \alpha)$$

Nas figuras 11.8, 11.9 e 11.10 observam-se:

- as regiões de rejeição da hipótese nula (subdivididas nos dois ou em apenas um dos lados) sob a curva da função densidade de probabilidade da distribuição adequada ao teste com probabilidades iguais ao nível de significância α ;
- a região de não rejeição da hipótese nula (delimitada à esquerda e à direita ou apenas em um dos lados) com probabilidade igual ao nível de confiança $(1 - \alpha)$; e,
- os valores críticos da estatística do teste.

Exemplo: Duas máquinas são usadas para encher garrafas plásticas com um volume líquido de 16oz. Os volumes de enchimento podem ser admitidos como normais, tendo desvios padrão iguais a $\sigma_1 = 0,020\text{oz}$ e $\sigma_2 = 0,025\text{oz}$. O departamento de engenharia da fábrica deseja saber a um nível de significância de $\alpha = 0,01$ se ambas as máquinas enchem um mesmo volume e para isso coletou uma amostra de 10 garrafas enchidas por cada uma das máquinas cf. tabela abaixo:

As variâncias populacionais σ_1^2 e σ_2^2 são conhecidas e as populações seguem uma distribuição Normal. A estatística do teste é:

Table 11.3: Enchimento de duas máquinas

Máquina 01	Máquina 02		
16,03	16,01	16,02	16,03
16,04	15,96	15,97	16,04
16,05	15,98	15,96	16,02
16,05	16,02	16,01	16,01
16,02	15,99	15,99	16,00

$$z_{calc} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

tal que tal que Z ($Z \sim \mathcal{N}(0, 1)$), em que:

- μ_1, μ_2 são as médias das populações em teste;
- $\sigma_1^2 = 0,020^2, \sigma_2^2 = 0,025^2$ são as variâncias das populações em teste;
- $\bar{x}_1 = 16,015, n_1 = 10$ são a média e o tamanho da amostra 1;
- $\bar{x}_2 = 16,005, n_2 = 10$ são a média e o tamanho da amostra 2; e,
- o nível de significância estabelecido para o teste é $\alpha = 0,01$.

O problema nos pede um teste bilateral (tipo: diferente de):

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

Se z_{calc} for tal que:

$$-z_{tab(\frac{\alpha}{2})} \leq z_{calc} \leq z_{tab(\frac{\alpha}{2})}$$

não se rejeita a hipótese nula sob o nível de significância estabelecido. Da tabela da distribuição Normal padronizada obtemos o valor crítico bicaudal: $|Z_{tab(\frac{\alpha}{2})}| = 2,57$. Pelo cálculo, a estatística do teste é $z_{calc} = 0,98773$.

```

alfa=0.01

prob_desejada1=alfa/2
z_desejado1=round(qnorm(prob_desejada1),4)
d_desejada1=dnorm(z_desejado1, 0, 1)

prob_desejada2=1-alfa/2
z_desejado2=round(qnorm(prob_desejada2),4)
d_desejada2=dnorm(z_desejado2, 0, 1)

z_calculado=0.98773
d_calculado=dnorm(z_calculado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de z", breaks = c(z_desejado1,z_desejado2)) +
  labs(title=
      "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
       teste",
      subtitle = "P(-2,57, 2,57)=(1-\u03b1) em cinza (nível de confiança=0,99) \nP(-\U221e;
      \u2248 -2,57)= P(2,57; \U221e)= \u03b1/2 em vermelho (nível de significância/2=0,005)
      ") +
  geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3)+
  geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="valor crítico=-2,57",
           angle=90, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="valor crítico=2,57",
           angle=90, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado1-1.5, y=0.1, label="Região de rejeição da hipótese nula
           \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3)+
```

```

annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Região de rejeição da hipótese nula
  ↵ \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado1+2, y=0.2, label="Região de não rejeição da hipótese
  ↵ nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
  ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do te
  ↵ teste=0,9877", angle=90, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Regiões críticas sob a curva da função densidade da distribuição adequada ao teste

$P(-2,57, 2,57) = (1-\alpha)$ em cinza (nível de confiança=0,99)

$P(-\infty; -2,57) = P(2,57; \infty) = \alpha/2$ em vermelho (nível de significância/2=0,005)

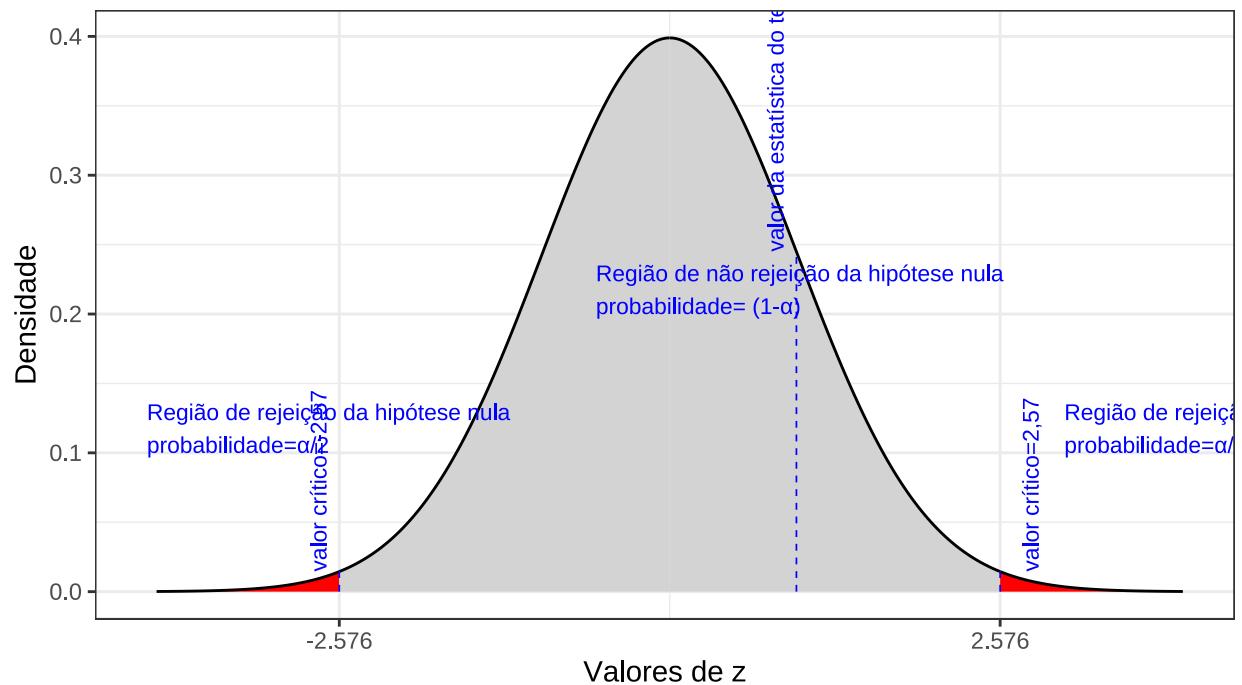


Figure 11.18: Regiões de rejeição da hipótese nula para o teste bilateral (tipo: diferente de) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valores críticos da estatística do teste: $z_{crit} = \pm 2,57$. O valor calculado da estatística ($z_{calc} = 0,987$) não nos possibilita a rejeição da hipótese nula sob aquele nível de confiança

Conclusão: Os resultados obtidos pela análise estatística de comparação de médias das duas amostras colhidas de garrafas de plástico enchidas por duas máquinas diferentes 1 e 2 não nos permitem rejeitar a hipótese de que suas médias sejam iguais sob um nível de confiança de 99% (Figura 11.18).

Podemos ainda realizar testes de hipóteses para as diferenças entre as médias observadas ($\mu_1 < \mu_2$ ou $\mu_1 > \mu_2$). As conclusões derivadas desses testes deverão indicar que as médias não diferem entre si ao nível de significância dos testes chegando assim, por outras vias (agora não se rejeitando a hipótese nula), à mesma conclusão do teste de igualdade das médias antes realizado.

Teste unilateral à esquerda (tipo: menor que)

Nessa situação postula-se que a diferença da média 1 **para** a média 2 é **no mínimo 0** (o que equivale dizer que a média 1 é **no mínimo igual** à média 2):

$$\begin{cases} H_0 : \mu_1 - \mu_2 \geq 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{cases}$$

Da tabela da distribuição Normal padronizada obtemos o valor crítico monocaudal: $Z_{tab(\alpha)} = -2,33$. Pelo cálculo, a estatística do teste é $Z_{calc} = 0,98773$.

```
alfa=0.01
prob_desejada=alfa
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

z_calculado=0.98773
d_calculado=dnorm(z_calculado, 0, 1)
```

```
ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-4, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado,4),
            colour="black") +
```

```

scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores de z", breaks = c(z_desejado)) +
labs(title =
  "Região crítica sob a curva da função densidade da \ndistribuição apropriada ao
  → teste",
  subtitle = "P( -2,33,\u221e,)=(1-\u03b1) em cinza (nível de confiança=0,99)
  → \nP(-\u221e; -2,33)=\u03b1 em vermelho (nível de significância=0,01) ")+
geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue",
  ↵ lty=2, lwd=0.3)+
annotate(geom="text", x=z_desejado-0.1, y=d_desejada, label="valor crítico=-2,33", angle=90,
  ↵ vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado-2, y=0.1, label="Região de rejeição da hipótese nula
  ↵ \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado+1, y=0.2, label="Região de não rejeição da hipótese nula
  ↵ \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
  ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do
  ↵ teste=0,98773", angle=90, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Região crítica sob a curva da função densidade da distribuição apropriada ao teste

$P(-2,33, \infty) = (1-\alpha)$ em cinza (nível de confiança=0,99)
 $P(-\infty; -2,33) = \alpha$ em vermelho (nível de significância=0,01)

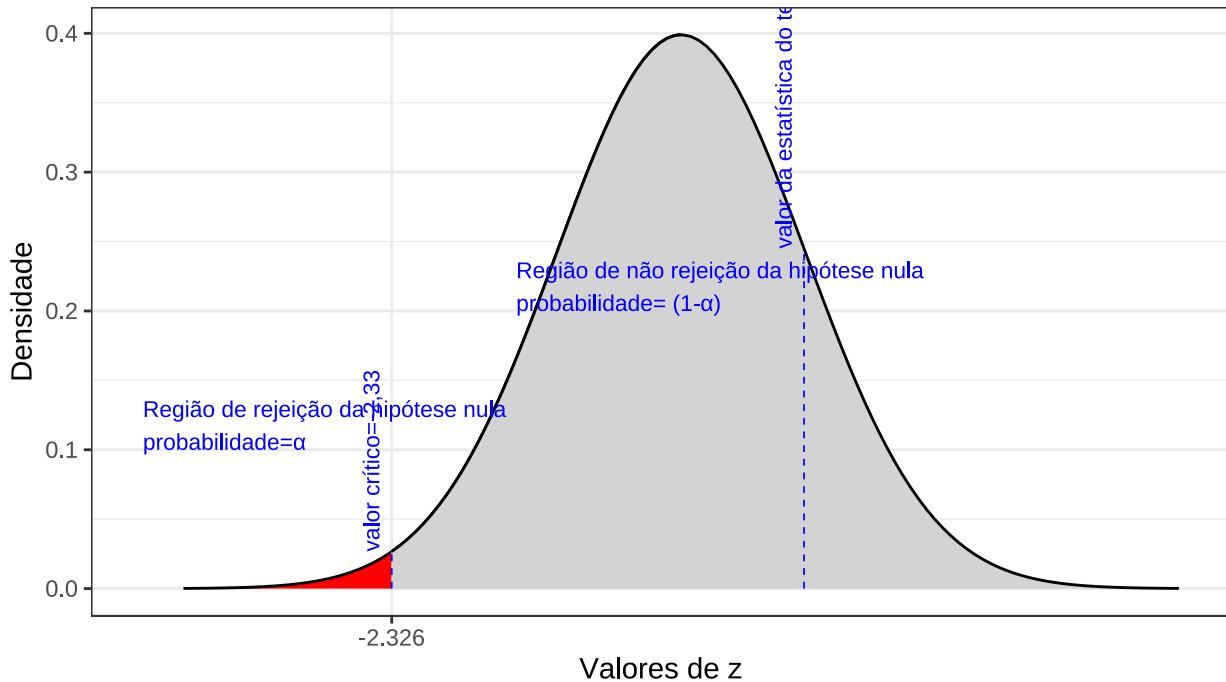


Figure 11.19: Regiões de rejeição da hipótese nula para o teste unilateral à esquerda (tipo: menor que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valor crítico da estatística do teste: $z_{crit} = -2,33$. O valor calculado da estatística ($z_{calc} = 0,98773$) não nos possibilita a rejeição da hipótese nula sob aquele nível de confiança

Conclusão: Os resultados obtidos pela análise estatística de comparação de médias das duas amostras colhidas de garrafas de plástico enchidas por duas máquinas diferentes 1 e 2 não nos permitem rejeitar a hipótese de que a média de enchimento da máquina 1 seja no mínimo igual à da máquina 2 sob um nível de confiança de 99% (Figura 11.19).

Teste unilateral à direita (tipo: maior que)

Nessa situação postula-se que a diferença da média 1 **para** a média 2 é **no máximo 0** (o que equivale dizer que a média 1 é **no máximo igual** à média 2):

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

Da tabela da distribuição Normal padronizada obtemos o valor crítico monocaudal: $Z_{tab(\alpha)} = -2,33$. Pelo cálculo, a estatística do teste é $Z_{calc} = 0,98773$.

```
alfa=0.99
prob_desejada=alfa
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

z_calculado=0.98773
d_calculado=dnorm(z_calculado, 0, 1)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-4, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de z", breaks = c(z_desejado)) +
  labs(title=
      "Região crítica sob a curva da função densidade da \ndistribuição apropriada ao
      → teste",
      subtitle = "P(- \U221e; 2,33)=(1-\u03b1) em cinza (nível de confiança=0,99) \nP(2,33
      → ; \U221e)=\u03b1 em vermelho (nível de significância=0,01)")+
  geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue",
               lty=2, lwd=0.3)+
```

```

annotate(geom="text", x=z_desejado-0.1, y=d_desejada, label="valor crítico=-1,88", angle=90,
  ↵ vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado, y=0.1, label="Região de rejeição da hipótese nula
  ↵ \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado-3, y=0.2, label="Região de não rejeição da hipótese nula
  ↵ \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
  ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do
  ↵ teste=0,98773", angle=90, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Região crítica sob a curva da função densidade da distribuição apropriada ao teste

$P(-\infty; 2,33) = 1 - \alpha$ em cinza (nível de confiança=0,99)

$P(2,33 ; \infty) = \alpha$ em vermelho (nível de significância=0,01)

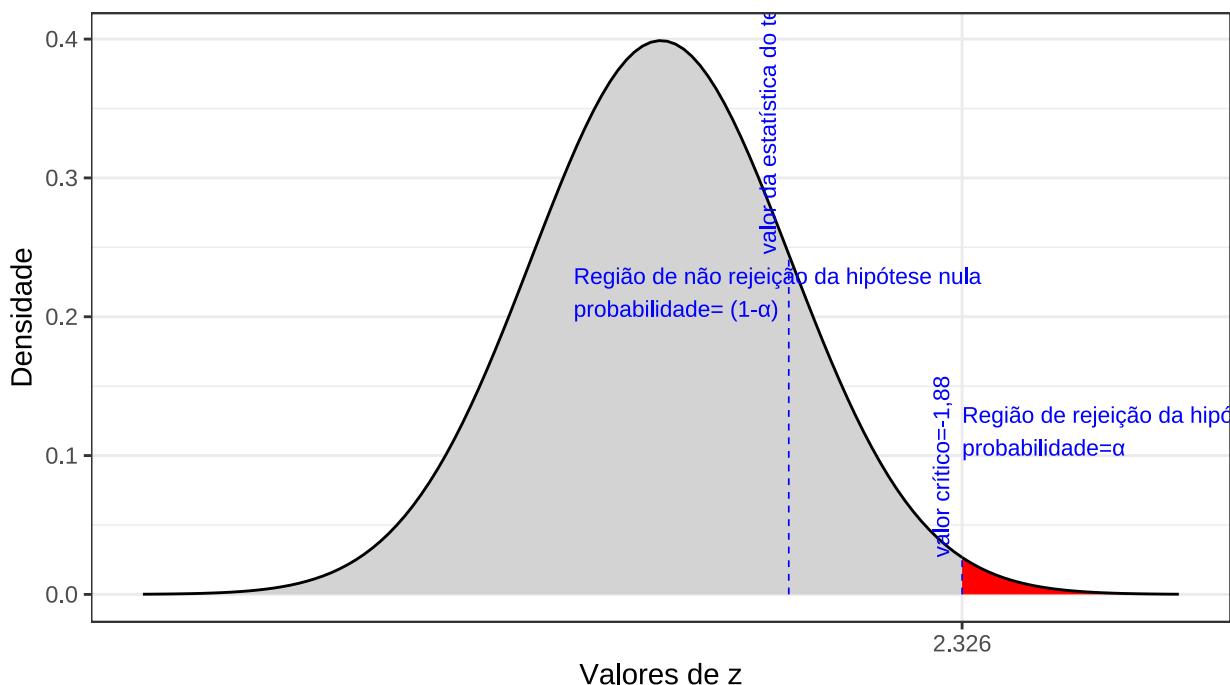


Figure 11.20: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: maior que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $z_{crit} = 2,33$. O valor calculado da estatística ($z_{calc} = 0,98773$) não nos possibilita a rejeição da hipótese nula sob aquele nível de confiança

Conclusão: Os resultados obtidos pela análise estatística de comparação de médias das duas amostras colhidas de garrafas de plástico enchidas por duas máquinas diferentes 1 e 2 não nos permitem rejeitar a hipótese de que a média de enchimento da máquina 1 seja no máximo igual à da máquina 2 sob um nível de confiança de 99% (Figura 11.20).

Pelo teste unilateral à esquerda conclui-se que $\mu_1 \geq \mu_2$; pelo teste unilateral à direita conclui-se que $\mu_1 \leq \mu_2$.

Sob o nível de significância estabelecido conclui-se que $\mu_1 = \mu_2$.

11.9.3 Testes de hipóteses para as médias de duas populações Normais com variâncias desconhecidas mas iguais: teste “t” homocedástico ($\sigma_1^2 = \sigma_2^2 = ?$)

Probabilidade dos intervalos de confiança para os testes de hipóteses com o uso da estatística t ($T \sim t_{(n_1+n_2-2)}$). Os valores assumidos pelas diferenças amostrais são tais que

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{S_c \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

em que:

- Δ_0 usualmente é 0 (igualdade);
- $\sigma_1^2 = \sigma_2^2 = \sigma^2$ são as variâncias populacionais desconhecidas, mas admitidas iguais (homogêneas);
- \bar{x}_1, S_1^2, n_1 são a média, a variância e o tamanho referentes à amostra 1;
- \bar{x}_2, S_2^2, n_2 são a média, a variância e o tamanho referentes à amostra 2; e,
- S_c^2 é a variância conjunta ou ponderada.

Condições:

- amostras Normais (n_1 e n_2 qualquer);
- amostras sob outras distribuições (desde que n_1 e $n_2 \geq 30$);
- a utilização da estatística “t” para n_1 e $n_2 \geq 30$ apenas pressupõe que S_c seja um estimador suficientemente bom para σ_i ; e,
- $t_{tab}(\frac{\alpha}{2}; n_1+n_2-2)$ ou $t_{tab}(\alpha; n_1+n_2-2)$: o quantil associado na distribuição “t” de Student ao nível de significância pretendido no teste, com $(n_1 + n_2 - 2)$ graus de liberdade.

A variância conjunta (ou variância ponderada) S_c^2 a ser utilizada no cálculo da estatística do teste é definida como:

$$S_c^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}$$

Probabilidade dos intervalos de confiança para os testes de hipóteses com o uso da estatística t ($T \sim t_{(n_1+n_2-2)}$)

- Teste de hipóteses bilateral (tipo: diferente de):

$$\begin{aligned} P[|t_{calc}| \geq t_{tab(\frac{\alpha}{2}; n_1+n_2-2)} | \mu_1 = \mu_2] &= (1 - \alpha) \\ P(-t_{tab(\frac{\alpha}{2}; n_1+n_2-2)} \leq t_{calc} \leq t_{tab(\frac{\alpha}{2}; n_1+n_2-2)}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à esquerda (tipo: menor que):

$$\begin{aligned} P[t_{calc} \geq -t_{tab(\alpha)} | \mu_1 \geq \mu_2] &= (1 - \alpha) \\ P(t_{calc} \geq -t_{tab(\alpha; n_1+n_2-2)}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à direita (tipo: maior que):

$$P[t_{calc} \leq t_{tab(\alpha)} | \mu_1 \leq \mu_2] = (1 - \alpha) P(t_{calc} \leq t_{tab(\alpha; n_1+n_2-2)}) = (1 - \alpha)$$

Nas figuras 11.8, 11.9 e 11.10 observam-se:

- as regiões de rejeição da hipótese nula (subdivididas nos dois ou em apenas um dos lados) sob a curva da função densidade de probabilidade da distribuição adequada ao teste com probabilidades iguais ao nível de significância α ;
- a região de não rejeição da hipótese nula (delimitada à esquerda e à direita ou apenas em um dos lados) com probabilidade igual ao nível de confiança $(1 - \alpha)$; e,
- os valores críticos da estatística do teste.

11.9.3.1 Teste “F” para a razão de duas variâncias

Para se verificar se a consideração de igualdade das variâncias é estatisticamente sustentável pode-se recorrer ao teste “F” de sua razão. Estrutura do teste:

$$\begin{cases} H_0 : \sigma_1^2 - \sigma_2^2 = \delta \\ H_1 : \sigma_1^2 - \sigma_2^2 \neq \delta \end{cases}$$

em que, usualmente, $\delta = 0$ (igualdade).

Tendo-se $\frac{\sigma_2^2}{\sigma_1^2} = \frac{\sigma_1^2}{\sigma_2^2} = 1$ na Hipótese nula (H_0) pela pressuposição da igualdade, F_{calc} será dado por:

$$f_{calc} = \left(\frac{S_1^2}{S_2^2} \right) \cdot \left(\frac{\sigma_1^2}{\sigma_2^2} \right) \sim F_{(n_1-1), (n_2-1)}$$

A Hipótese nula será rejeitada se:

$$f_{calc} \geq f_{((n_1-1), (n_2-1), 1-\frac{\alpha}{2})}$$

ou

$$f_{calc} \leq f_{((n_1-1), (n_2-1), \frac{\alpha}{2})}$$

em que $f_{(n_1-1), (n_2-1)}$ são os quantis de ordem α (pelo lado esquerdo da curva) e $(1 - \frac{\alpha}{2})$ (pelo lado direito da curva) da Distribuição F (Ronald Fisher e George Waddel Snedecor) com graus de liberdade: $(n_1 - 1)$ são os graus de liberdade (GL) no numerador e $(n_2 - 1)$ são os graus de liberdade (GL) no denominador (em concordância com a razão utilizada $(\frac{S_1}{S_2})$).

Em razão da limitação das tabelas torna-se interessante relembrar a propriedade:

$$f_{((n_1-1),(n_2-1),\alpha)} = \frac{1}{f_{((n_1-1),(n_2-1),(1-\frac{\alpha}{2}))}}$$

Regiões de rejeição da hipótese nula (Figura 11.21):

```

prob_desejada1=0.025
prob_desejada2=0.975

df1=3
df2=50

f_desejado1=round(qf(prob_desejada1,df1, df2), 4)
f_desejado2=round(qf(prob_desejada2,df1, df2), 4)

d_desejada1=df(f_desejado1,df1, df2)
d_desejada2=df(f_desejado2,df1, df2)

f_test_1=ggplot(data.frame(x = c(0, 6)), aes(x)) +
  stat_function(fun = df,
    geom = "area",
    fill = "red",
    xlim = c(0,f_desejado1),
    colour="black",
    args = list(
      df1 = df1,
      df2 = df2
    ))+
  stat_function(fun = df,
    geom = "area",
    fill = "lightgrey",
    xlim = c(f_desejado1, f_desejado2),
    colour="black",
    args = list(
      df1 = df1,
      df2 = df2
    ))+
  stat_function(fun = df,
    geom = "area",
    fill = "red",
    xlim = c(f_desejado2,6),
    colour="black",
    args = list(
      df1 = df1,
      df2 = df2
    ))+
  scale_y_continuous(name="Densidade") +

```

```
#scale_x_continuous(name="Valores score (f)", breaks = c(f_desejado1, f_desejado2))+  
scale_x_continuous(name="Valores score (f)")+  
labs(title="Curva da função densidade \nDistribuição F",  
subtitle = "P(f crítico 1, f crítico 2)=(1-\u03b1) em cinza (nível de confiança) \nP(0; f  
↳ crítico 1)= P(f crítico 2; \u221e)= \u03b1/2 em vermelho (nível de significância/2)  
↳ ")+  
geom_segment(aes(x = f_desejado1, y = 0, xend = f_desejado1, yend = d_desejada1),  
↳ color="blue", lty=2, lwd=0.3)+  
geom_segment(aes(x = f_desejado2, y = 0, xend = f_desejado2, yend = d_desejada2),  
↳ color="blue", lty=2, lwd=0.3)+  
annotate(geom="text", x=f_desejado1+0.2, y=0.2, label="f crítico 1", angle=90, vjust=0,  
↳ hjust=0, color="blue", size=4)+  
annotate(geom="text", x=f_desejado2-0.2, y=0.2, label="f crítico 2", angle=90, vjust=0,  
↳ hjust=0, color="blue", size=4)+  
annotate(geom="text", x=f_desejado1+1, y=0.4, label="Zona de não rejeição \n(para f  
↳ calculado)", angle=0, vjust=0, hjust=0, color="blue", size=3)+  
annotate(geom="text", x=f_desejado2+1, y=0.2, label="Zona de rejeição \n(para f  
↳ calculado)", angle=0, vjust=0, hjust=0, color="blue", size=3)+  
annotate(geom="text", x=f_desejado1-1, y=0.2, label="Zona de rejeição \n(para f  
↳ calculado)", angle=0, vjust=0, hjust=0, color="blue", size=3)+  
theme_bw()
```

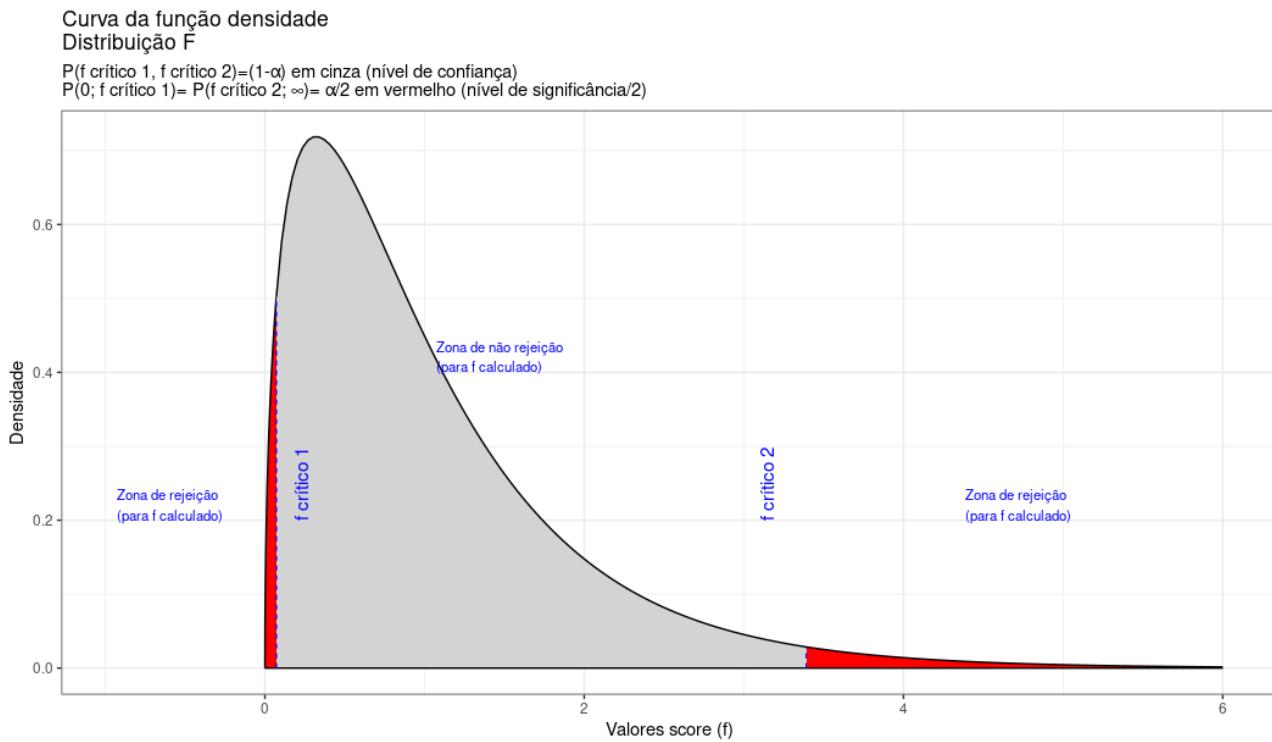


Figure 11.21: Regiões de rejeição da hipótese nula para o teste bilateral (tipo: diferente de) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valores críticos da estatística do teste: f_{crit1} e f_{crit2} para o nível de significância pretendido (α dividido em ambas as caudas) e $(df_1; df_2)$ graus de liberdade. A curva não é simétrica e assim, os valores críticos são diferentes

Uma regra prática permite reverter o teste bilateral em um teste unilateral à direita se tomarmos o maior valor (f_{calc} maior que 1, portanto) de f_{calc} dentre as possíveis razões:

$$f_{calc} = \left(\frac{S_1^2}{S_2^2}\right) \cdot \left(\frac{\sigma_1^2}{\sigma_2^2}\right) \sim F_{(n_1-1), (n_2-1)}$$

ou

$$f_{calc} = \left(\frac{S_2^2}{S_1^2}\right) \cdot \left(\frac{\sigma_2^2}{\sigma_1^2}\right) \sim F_{(n_2-1), (n_1-1)}$$

em que:

- $F_{tab(\alpha, n_1-1, n_2-1)}$ é o quantil de ordem α da Distribuição “F” (Ronald Fisher e George Waddel Snedecor) com graus de liberdade $(n_1 - 1)$ no numerador e $(n_2 - 1)$ no denominador (em concordância com a razão utilizada: $\frac{S_1^2}{S_2^2}$); ou,
- $(n_2 - 1)$ são os graus de liberdade (GL) no numerador e $(n_1 - 1)$ são os graus de liberdade (GL) no denominador (em concordância com a razão utilizada: $\frac{S_2^2}{S_1^2}$).

Região de rejeição da hipótese nula (Figura 11.22):

```
prob_desejada1=0.95

df1=3
df2=50

f_desejado1=round(qf(prob_desejada1, df1, df2), 4)
d_desejada1=df(f_desejado1, df1, df2)

df_test_2=ggplot(data.frame(x = c(0, 6)), aes(x)) +
  stat_function(fun = df,
    geom = "area",
    fill = "lightgrey",
    xlim = c(0,f_desejado1),
    colour="black",
    args = list(
      df1 = df1,
      df2 = df2
    ))+
```

```

stat_function(fun = df,
              geom = "area",
              fill = "red",
              xlim = c(f_desejado1,6),
              colour="black",
              args = list(
                df1 = df1,
                df2 = df2
              ))+
scale_y_continuous(name="Densidade") +
#scale_x_continuous(name="Valores score (f)", breaks = c(f_desejado1, f_desejado2))+
scale_x_continuous(name="Valores score (f)")+
labs(title="Curva da função densidade \nDistribuição F",
subtitle = "P(0; f crítico 1)=(1-\u03b1) em cinza (nível de confiança) \nP(f crítico ;
↪ \U221e)= \u03b1 em vermelho (nível de significância)")+
geom_segment(aes(x = f_desejado1, y = 0, xend = f_desejado1, yend = d_desejada1),
↪ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=f_desejado1+0.1, y=d_desejada1, label="f crítico 1", angle=90,
↪ vjust=0, hjust=0, color="blue",size=4)+
annotate(geom="text", x=f_desejado1+1, y=d_desejada1, label="Zona de rejeição \n(para f
↪ calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=f_desejado1-1, y=d_desejada1, label="Zona de não rejeição \n(para
↪ f calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+

theme_bw()

```

Exemplo: A Secretaria de Educação de um município deseja saber se o desempenho dos alunos de duas diferentes escolas municipais na disciplina de matemática pode ser considerado igual a um nível de significância de $\alpha = 0,05$. Verifique antes de as variâncias são **iguais**. Para tanto ministrou um mesmo teste a 10 alunos de cada uma delas e obteve os seguintes notas:

Table 11.4: Notas em matemática de duas escolas

Escola 01		Escola 02	
78	83	85	79
84	79	75	88
81	75	83	94
78	85	87	87
76	81	80	82

- Teste de hipóteses para a igualdade das variâncias:

$$\begin{cases} H_0 : \sigma_1^2 - \sigma_2^2 = \delta \\ H_1 : \sigma_1^2 - \sigma_2^2 \neq \delta \end{cases}$$

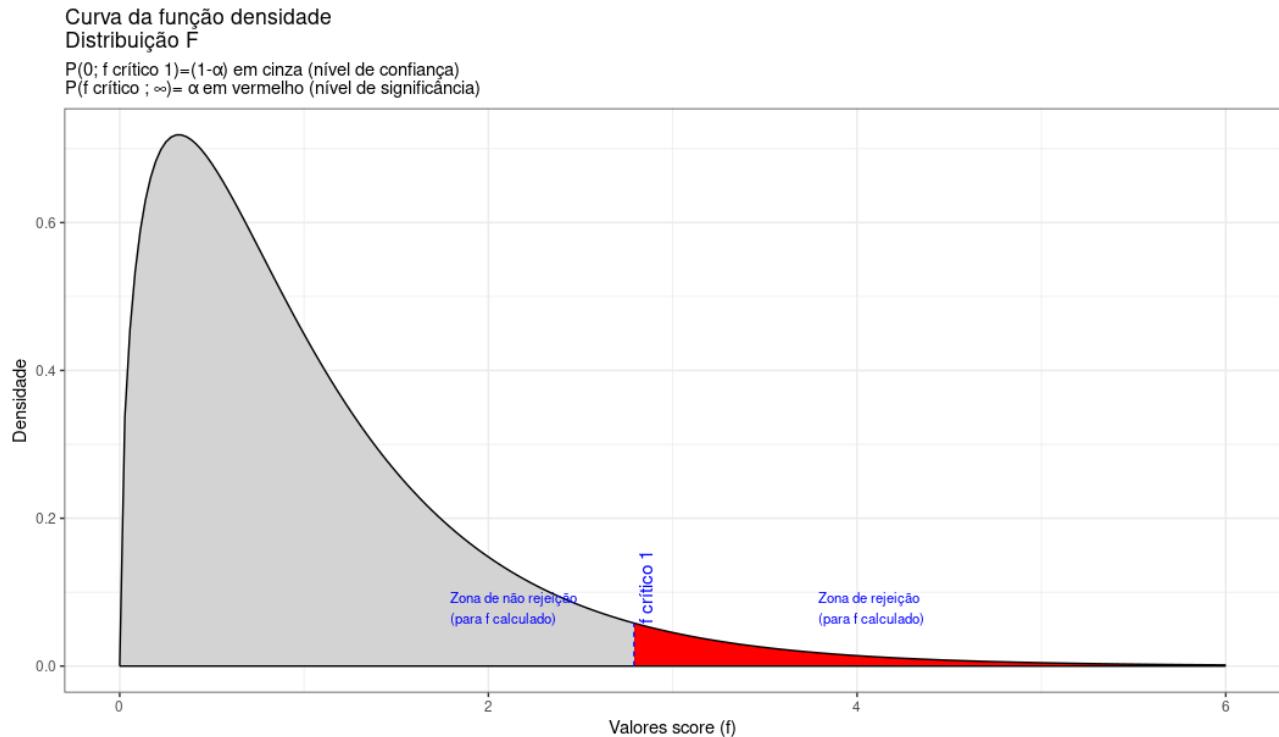


Figure 11.22: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: menor que): a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: f_{crit} para o nível de significância pretendido (α em uma cauda) e $(df_1; df_2)$ graus de liberdade.

em que, usualmente, $\delta = 0$ (igualdade). Se $\sigma_1^2 = \sigma_2^2$, então $\frac{\sigma_1^2}{\sigma_2^2} = 1$.

$$F_{cal} = \frac{S_2^2}{S_1^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} = 2,56$$

$$F_{critico(\alpha, n_1-1, n_2-1)} = F_{tab(5\%, 9, 9)} = 3,18$$

```
prob_desejada1=0.95
df1=9
df2=9

f_desejado1=round(qf(prob_desejada1,df1, df2), 4)
d_desejada1=df(f_desejado1,df1, df2)

f_calculado=2.56
d_calculado=df(f_calculado,df1, df2)
```

```

f_test_3=ggplot(data.frame(x = c(0, 6)), aes(x)) +
  stat_function(fun = df,
    geom = "area",
    fill = "lightgrey",
    xlim = c(0,f_desejado1),
    colour="black",
    args = list(
      df1 = df1,
      df2 = df2
    ))+
  stat_function(fun = df,
    geom = "area",
    fill = "red",
    xlim = c(f_desejado1,6),
    colour="black",
    args = list(
      df1 = df1,
      df2 = df2
    ))+
  scale_y_continuous(name="Densidade") +
#scale_x_continuous(name="Valores score (f)", breaks = c(f_desejado1, f_desejado2))+ 
scale_x_continuous(name="Valores score (f)")+
  labs(title="Curva da função densidade \nDistribuição F",
  subtitle = "P(0; 3,18 1)=(1-\u03b1) em cinza (nível de confiança=0,95) \nP(3,18 ; \U221e)=
  \u03b1 em vermelho (nível de significância=0,05) ")+
  geom_segment(aes(x = f_desejado1, y = 0, xend = f_desejado1, yend = d_desejada1),
  color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=f_desejado1+0.1, y=d_desejada1, label="F crítico 1=3,18",
  angle=90, vjust=0, hjust=0, color="blue",size=4)+
  annotate(geom="text", x=f_desejado1+1, y=d_desejada1, label="Zona de rejeição \n(para F
  calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=f_desejado1-2, y=d_desejada1, label="Zona de não rejeição \n(para
  F calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  geom_segment(aes(x = f_calculado, y = 0, xend = f_calculado, yend = d_calculado),
  color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=f_calculado+0.1, y=d_calculado, label="f calculado=2,56",
  angle=90, vjust=0, hjust=0, color="blue",size=4)+ 
  theme_bw()

```

O valor calculado da estatística de teste ($F_{calc} = 2,56$) situa-se na região não significante do teste, não permitindo a rejeição da hipótese nula de que as variâncias sejam iguais sob o nível de confiança estabelecido. Não se pode rejeitar a hipótese de que as variâncias sejam iguais a um nível de significância de 5% (Figura 11.23).

Estrutura do teste:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

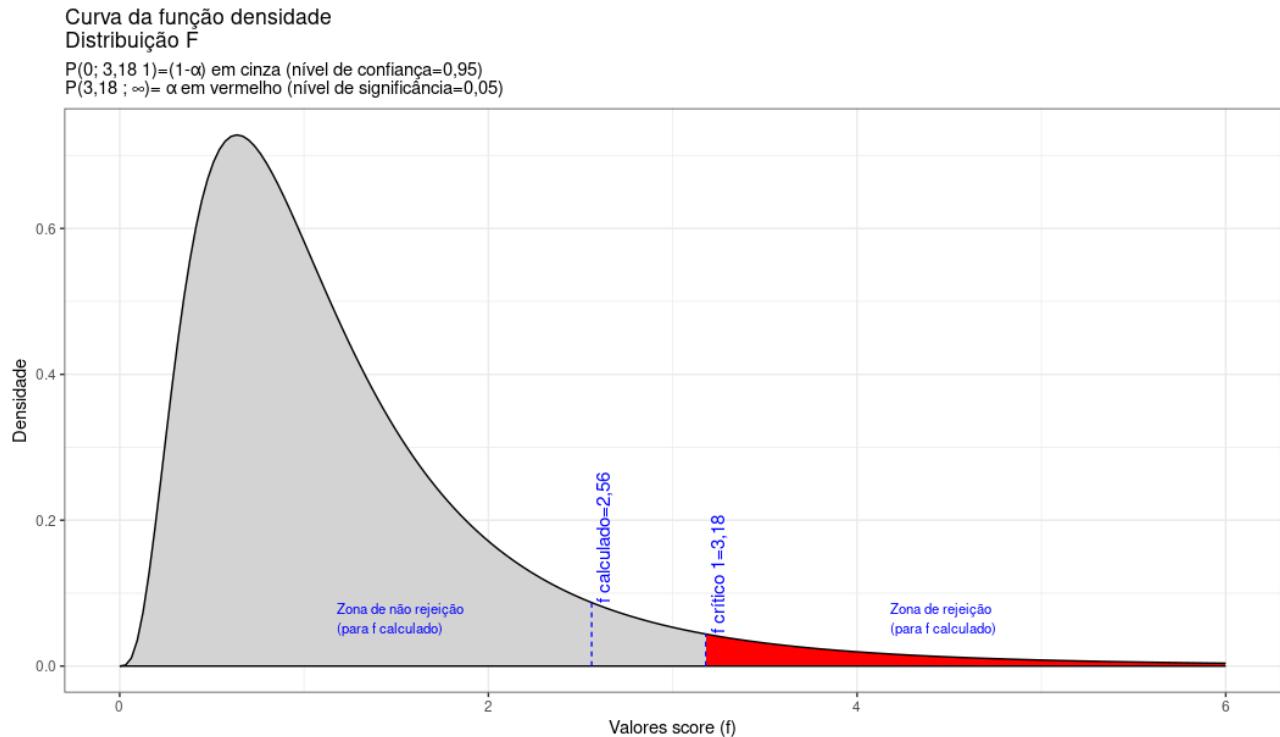


Figure 11.23: O valor calculado da estatística de teste ($F_{calc} = 2,56$) situa-se na região não significante do teste, não permitindo a rejeição da hipótese nula de que as variâncias são iguais sob o nível de confiança estabelecido.

Variâncias populacionais desconhecidas mas estatisticamente iguais. Nada se sabe sobre a distribuição da população e amostras de reduzido tamanho.

$$S_c^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}$$

é a variância conjunta ponderada, em que:

- μ_1, μ_2 são as médias das populações em teste;
- $\sigma_1^2 = \sigma_2^2 = \sigma^2$ são as variâncias das populações em teste, desconhecidas e estatisticamente iguais;
- $\bar{x}_1 = 80, S_1^2 = 3,366^2, n_1 = 10$ são a média, a variância e o tamanho referentes à amostra 1;
- $\bar{x}_2 = 84, S_2^2 = 5,395^2, n_2 = 10$ são a média, a variância e o tamanho referentes à amostra 2;
- $t_{tab}(\frac{\alpha}{2}; n_1 + n_2 - 2)$: o quantil associado na distribuição “t” de **Student** ao nível de significância pretendido no teste, com $(n_1 + n_2 - 2)$ graus de liberdade.

$$\begin{aligned} S_c^2 &= 20,2180 \\ S_c &= 4,4964 \end{aligned}$$

Estatística do teste:

$$t_{calc} = \frac{(\bar{x}_1 - \bar{x}_2)}{S_c \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t_{cal} = -1,9892$$

Teste bilateral:

$$t_{tab(\frac{\alpha}{2}; n_1 + n_2 - 2)} < t_{calc} < t_{tab(\frac{\alpha}{2}; n_1 + n_2 - 2)}$$

$$|t_{tab(\frac{\alpha}{2}; n_1 + n_2 - 2)}| = |t_{tab(2.5\%; 18)}| = 2,101$$

```
alfa=0.05

prob_desejada1=alfa/2
df=8
t_desejado1=round(qt(prob_desejada1,df ),df)
d_desejada1=dt(t_desejado1,df)

prob_desejada2=1-alfa/2
df=8
t_desejado2=round(qt(prob_desejada2, df),df)
d_desejada2=dt(t_desejado2,df)

t_calculado=-1.9892
d_calculado=dt(t_calculado,df)

ggplot(NULL, aes(c(-4,4))) +
```

```

geom_area(stat = "function",
           fun = dt,
           args=list(df),
           fill = "red",
           xlim = c(-4, t_desejado1),
           colour="black") +
geom_area(stat = "function",
           fun = dt,
           args=list(df),
           fill = "lightgrey",
           xlim = c(t_desejado1,0),
           colour="black") +
geom_area(stat = "function",
           fun = dt,
           args=list(df),
           fill = "lightgrey",
           xlim = c(0, t_desejado2),
           colour="black") +
geom_area(stat = "function",
           fun = dt,
           args=list(df),
           fill = "red",
           xlim = c(t_desejado2,4),
           colour="black") +
scale_y_continuous(name="Densidade") +
scale_x_continuous(name="Valores de t", breaks = c(t_desejado1, t_desejado2)) +
labs(title =
  "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
  ← teste",
  subtitle = "P(-2,101, 2,101)=(1-\u03b1) em cinza (nível de confiança=0,95)
  ← \nP(-\U221e; -2,101)= P(2,101; \U221e)= \u03b1/2 em vermelho (nível de
  ← significância/2=0,025)") + geom_segment(aes(x = t_desejado1, y = 0, xend =
  ← t_desejado1, yend = d_desejada1), color="blue", lty=2, lwd=0.3) +
geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2),
  ← color="blue", lty=2, lwd=0.3) +
annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="valor crítico=-2,101",
  ← angle=90, vjust=0, hjust=0, color="blue",size=3) +
annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="valor crítico=2,101",
  ← angle=90, vjust=0, hjust=0, color="blue",size=3) +
annotate(geom="text", x=t_desejado1-2, y=0.1, label="Região de rejeição da hipótese nula
  ← \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3) +
annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Região de rejeição da hipótese nula
  ← \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3) +
annotate(geom="text", x=t_desejado1+2, y=0.2, label="Região de não rejeição da hipótese
  ← nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3) +
geom_segment(aes(x = t_calculado, y = 0, xend = t_calculado, yend = d_calculado),
  ← color="blue", lty=2, lwd=0.3) +
annotate(geom="text", x=t_calculado-0.1, y=d_calculado, label="valor da estatística do
  ← teste=-1,9892", angle=90, vjust=0, hjust=0, color="blue",size=3) +
theme_bw()

```

Conclusão: Os resultados obtidos pela análise estatística de comparação de médias das duas amostras colhidas das notas de testes de matemáticas realizados em duas escolas diferentes (escola 1 e escola 2) não nos permitem

Regiões críticas sob a curva da função densidade da distribuição adequada ao teste

$P(-2,101, 2,101) = (1-\alpha)$ em cinza (nível de confiança=0,95)
 $P(-\infty; -2,101) = P(2,101; \infty) = \alpha/2$ em vermelho (nível de significância/2=0,025)

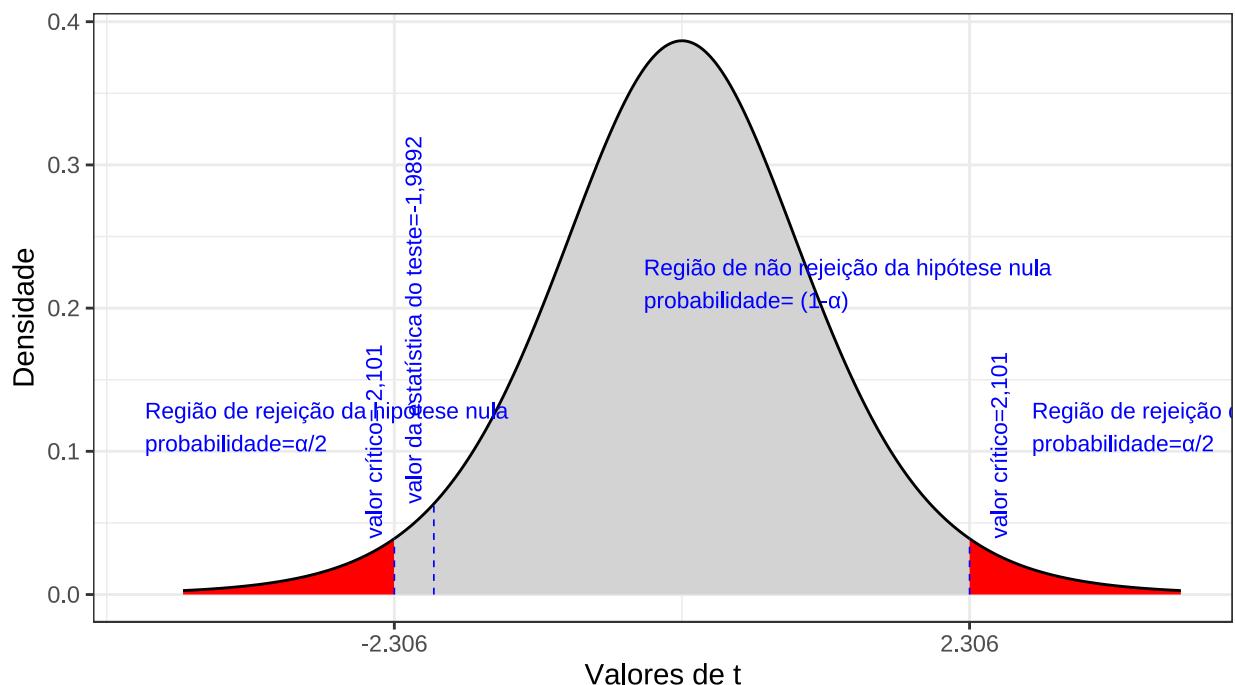


Figure 11.24: Regiões de rejeição da hipótese nula para o teste bilateral (tipo: diferente de) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valores críticos da estatística do teste: $t_{crit} = \pm 2,101$. O valor calculado da estatística ($t_{calc} = -1,9892$) situa-se na faixa de não significância do teste, impossibilitando a rejeição da hipótese nula sob aquele nível de confiança

rejeitar a hipótese de que suas médias sejam iguais a um nível de confiança de 5% (Figura 11.24).

11.9.4 Teste de hipóteses para as médias de duas populações Normais com variâncias desconhecidas e desiguais: teste ““t”’ heterocedástico ($\sigma_1^2 \neq \sigma_2^2 = ?$)

Probabilidade dos intervalos de confiança para os testes de hipóteses com o uso da estatística t ($T \sim t_\nu$). Os valores assumidos pelas diferenças amostrais são tais que

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$$

em que:

- Δ_0 usualmente é 0 (igualdade);
- \bar{x}_1, s_1^2, n_1 são a média, a variância e o tamanho referentes à amostra 1;
- \bar{x}_2, s_2^2, n_2 são a média, a variância e o tamanho referentes à amostra 2; e,
- a aproximação dos graus de liberdade (ν) é dada por uma combinação linear de variâncias de amostras independentes (Welch-Satterhwaite, 1946)

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

Condições:

- amostras Normais (n_1 e n_2 qualquer);
- amostras sob outras distribuições (desde que n_1 e $n_2 \geq 30$);

- $t_{tab(\frac{\alpha}{2};\nu)}$ ou $t_{tab(\alpha;\nu)}$: o quantil associado na distribuição “t” de Student ao nível de significância pretendido no teste, com ν graus de liberdade.

Probabilidade dos intervalos de confiança para os testes de hipóteses com o uso da estatística t ($T \sim t_\nu$)

- Teste de hipóteses bilateral (tipo: diferente de):

$$\begin{aligned} P[|t_{calc}| \geq t_{tab(\frac{\alpha}{2};\nu)} | \mu_1 = \mu_2] &= (1 - \alpha) \\ P(-t_{tab(\frac{\alpha}{2};\nu)} \leq t_{calc} \leq t_{tab(\frac{\alpha}{2};\nu)}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à esquerda (tipo: menor que):

$$\begin{aligned} P[t_{calc} \geq t_{tab(\alpha;\nu)} | \mu_1 \geq \mu_2] &= (1 - \alpha) \\ P(t_{calc} \geq t_{tab(\alpha;\nu)}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à direita (tipo: maior que):

$$\begin{aligned} P[t_{calc} \leq t_{tab(\alpha;\nu)} | \mu_1 \leq \mu_2] &= (1 - \alpha) \\ P(t_{calc} \leq t_{tab(\alpha;\nu)}) &= (1 - \alpha) \end{aligned}$$

Exemplo: a Secretaria de Educação de um município deseja saber se o desempenho dos alunos de duas diferentes escolas municipais na disciplina de matemática pode ser considerado igual a um nível de significância de $\alpha = 0,05$ (verifique antes se as variâncias podem ser admitidas como iguais). Para tanto ministrou um mesmo teste a 10 alunos de cada uma delas e obteve os seguintes notas:

Table 11.5: Desempenho dos alunos de duas escolas

Escola 01		Escola 02	
68	94	85	79
51	100	75	88
50	75	83	94
81	70	87	87
100	20	80	82

Estrutura do teste:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

- Teste de hipóteses bilateral (tipo: diferente de):

$$P(-t_{tab(\frac{\alpha}{2};\nu)} \leq t_{calc} \leq t_{tab(\frac{\alpha}{2};\nu)}) = (1 - \alpha)$$

As variâncias populacionais não são conhecidas e o tamanho das amostras é reduzido.

Teste de hipóteses para a igualdade das variâncias:

$$\begin{cases} H_0 : \sigma_1^2 - \sigma_2^2 = \delta & \text{usualmente } \delta = 0 \text{ (igualdade)} \\ H_1 : \sigma_1^2 - \sigma_2^2 \neq \delta \end{cases}$$

Se $\sigma_1^2 = \sigma_2^2$, então $\frac{\sigma_1^2}{\sigma_2^2} = 1$. O maior valor de F_{calc} é dado por:

$$F_{cal} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_1^2}{\sigma_2^2} = 22,056$$

e o valor crítico é

$$F_{tab(\alpha, n_1-1, n_2-1)} = F_{tab(5\%, 9, 9)} = 3,18$$

```
prob_desejada1=0.95
```

```
df1=9
df2=9
```

```
f_desejado1=round(qf(prob_desejada1, df1, df2), 4)
d_desejada1=df(f_desejado1, df1, df2)
```

```
f_calculado=22.056
d_calculada=df(f_calculado, df1, df2)
```

```
f_test_4=ggplot(data.frame(x = c(0, 25)), aes(x)) +
  stat_function(fun = df,
                geom = "area",
                fill = "lightgrey",
                xlim = c(0,f_desejado1),
                colour="black",
                args = list(
                  df1 = df1,
                  df2 = df2
                ))+
  stat_function(fun = df,
                geom = "area",
                fill = "red",
                xlim = c(f_desejado1,25),
                colour="black",
                args = list(
                  df1 = df1,
                  df2 = df2
                ))+
  scale_y_continuous(name="Densidade") +
  #scale_x_continuous(name="Valores score (f)", breaks = c(f_desejado1, f_desejado2))+ 
  scale_x_continuous(name="Valores score (f)")+
  labs(title="Curva da função densidade \nDistribuição F",
       subtitle = "P(0; 22,056)=(1-\u03b1) em cinza (nível de confiança) \nP(22,056 ; \u221e)=
       \u03b1 em vermelho (nível de significância)")+
  geom_segment(aes(x = f_desejado1, y = 0, xend = f_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=f_desejado1+0.1, y=d_desejada1, label="F crítico", angle=90,
           vjust=0, hjust=0, color="blue", size=4)+
```

```

geom_segment(aes(x = f_calculado, y = 0, xend = f_calculado, yend = d_desejada),
  color="blue", lty=2, lwd=0.3) +
annotate(geom="text", x=f_calculado+0.1, y=d_desejada1, label="F calculado", angle=90,
  vjust=0, hjust=0, color="blue",size=4) +
annotate(geom="text", x=f_desejado1+5, y=d_desejada1, label="Zona de rejeição \n(para F
  calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3) +
annotate(geom="text", x=f_desejado1-2.5, y=d_desejada1, label="Zona de não rejeição
  \n(para F calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3) +
theme_bw()

```

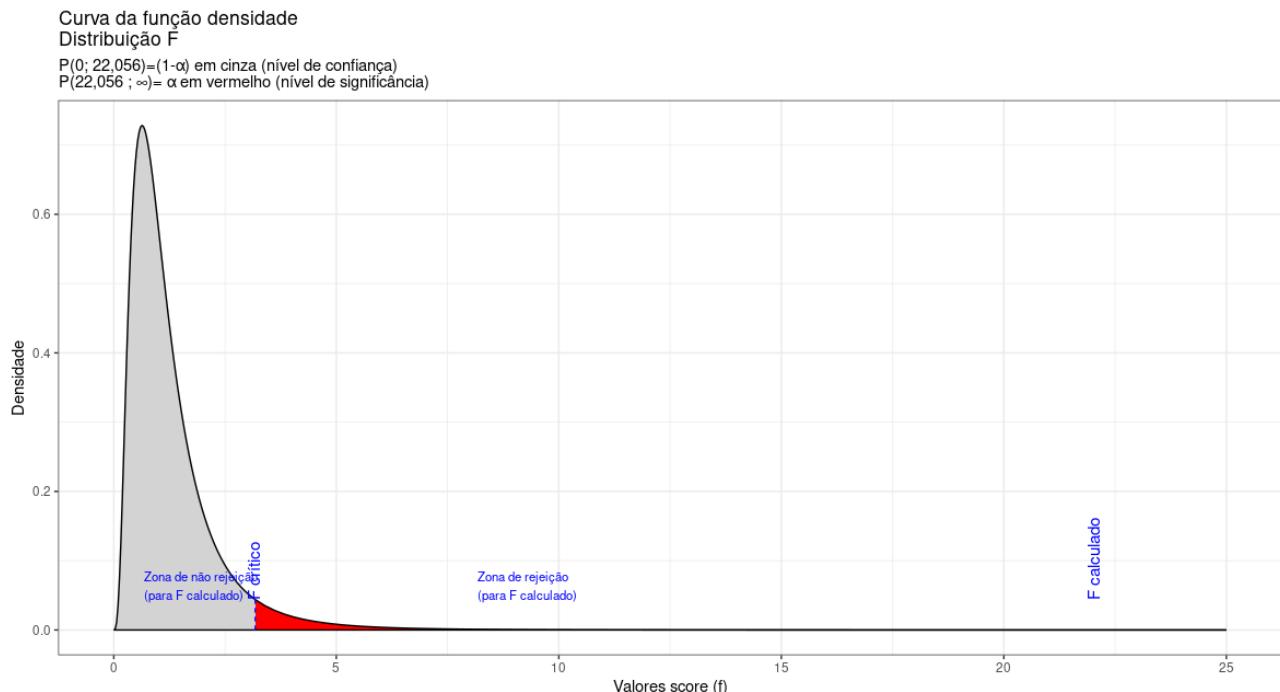


Figure 11.25: O valor calculado da estatística de teste ($F_{calc} = 3,18$) situa-se na região significante do teste, permitindo a rejeição da hipótese nula de que as variâncias sejam iguais sob o nível de confiança estabelecido.

Conclusão: não se pode aceitar a hipótese de que as variâncias sejam iguais a um nível de significância de 5% (cf. figura 11.25).

Estatística do teste: $T \sim t_{(\nu)}$ considerando que as variâncias populacionais não podem ser, estatisticamente, admitidas como iguais:

$$t_{calc} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

em que:

- μ_1, μ_2 são as médias das populações em teste;
- $\bar{x}_1 = 70, 90, S_1^2 = 25, 339^2, n_1 = 10$ são a média, a variância e o tamanho amostral 1;
- $\bar{x}_2 = 84, S_2^2 = 5, 395^2, n_2 = 10$ são a média, a variância e o tamanho amostral 2;
- $t_{tab}(\frac{\alpha}{2}; \nu)$ ou $t_{tab}(\alpha; \nu)$: o quantil associado na distribuição “t” de Student ao nível de significância pretendido no teste, com graus de liberdade (ν).

A aproximação dos graus de liberdade (ν) é dada por uma combinação linear das variâncias de amostras independentes (equação de Welch-Satterhwaite, 1946):

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2-1}} = 10$$

(aproximar o resultado para o inteiro superior mais próximo).

Cálculo da estatística do teste:

$$t_{calc} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = -1,599$$

Da tabela ‘t’ de Student obtemos o valor crítico bicaudal da estatística:

$$|t_{tab}(\frac{\alpha}{2}; \nu)| = |t_{tab}(0,025; 10)| = 2,22$$

```
alfa=0.05
```

```
prob_desejada1=alfa/2
df=8
t_desejado1=round(qt(prob_desejada1,df ),df )
d_desejada1=dt(t_desejado1,df )
```

```

prob_desejada2=1-alfa/2
df=8
t_desejado2=round(qt(prob_desejada2, df),df)
d_desejada2=dt(t_desejado2,df)

t_calculado=-1.599
d_calculado=dt(t_calculado,df)

ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(-4, t_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(t_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(0, t_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(t_desejado2,4),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de t", breaks = c(t_desejado1, t_desejado2)) +
  labs(title=
      "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
       teste",
      subtitle = "P(-2,22, 2,22)=(1-\u03b1) em cinza (nível de confiança=0,95) \nP(-\U221e;
       -2,22)= P(2,22; \U221e)= \u03b1/2 em vermelho (nível de significância/2=0,025)
       ") +
  geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend =
       d_desejada1), color="blue", lty=2, lwd=0.3) +
  geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2),
               color="blue", lty=2, lwd=0.3) +
  annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="valor crítico=-2,101",
           angle=90, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="valor crítico=2,101",
           angle=90, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=t_desejado1-2, y=0.1, label="Região de rejeição da hipótese nula
           \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Região de rejeição da hipótese nula
           \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=t_desejado1+2, y=0.2, label="Região de não rejeição da hipótese
           nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3) +

```

```
geom_segment(aes(x = t_calculado, y = 0, xend = t_calculado, yend = d_calculado),
  color="blue", lty=2, lwd=0.3)+  

annotate(geom="text", x=t_calculado-0.1, y=d_calculado, label="valor da estatística do  

  teste=-1,599", angle=90, vjust=0, hjust=0, color="blue",size=3)+  

theme_bw()
```

Regiões críticas sob a curva da função densidade da distribuição apropriada ao teste

$P(-2,22, 2,22) = 1 - \alpha$ em cinza (nível de confiança=0,95)

$P(-\infty; -2,22) = P(2,22; \infty) = \alpha/2$ em vermelho (nível de significância/2=0,025)

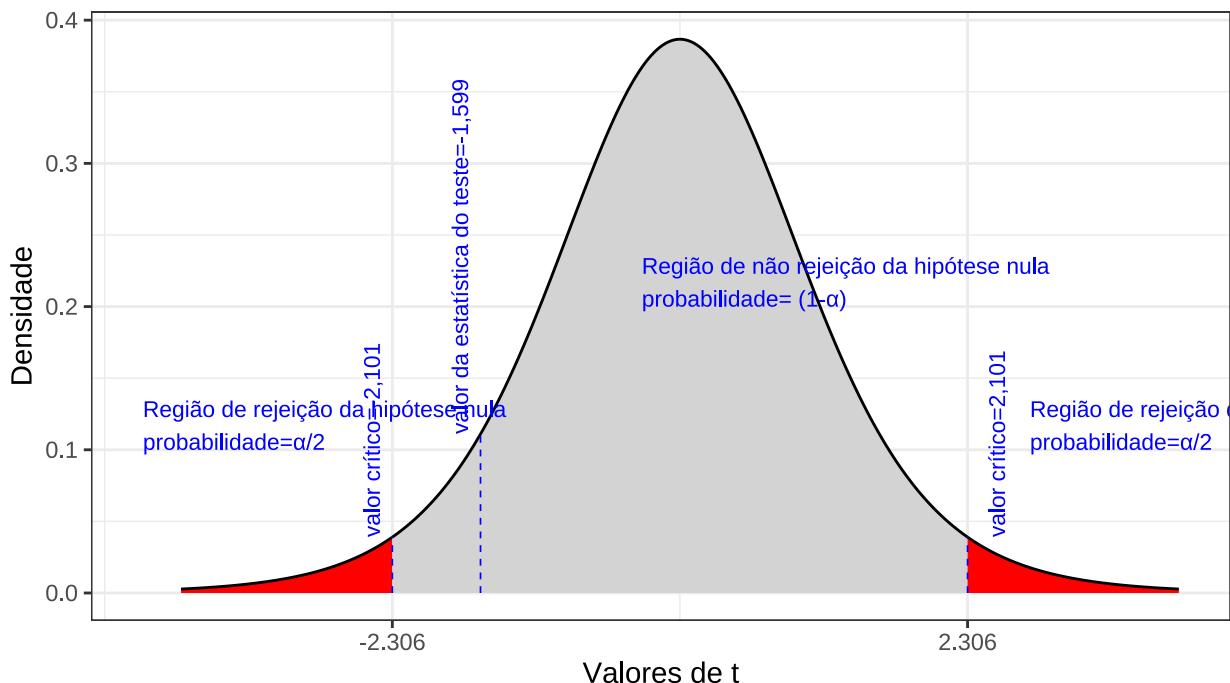


Figure 11.26: Regiões de rejeição da hipótese nula para o teste bilateral (tipo: diferente de) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valores críticos da estatística do teste: $t_{crit} = \pm 2,22$. O valor calculado da estatística ($t_{calc} = -1,599$) não se situa na faixa de significância do teste, não nos permitindo a rejeição da hipótese nula sob aquele nível de confiança

Conclusão: Os resultados obtidos pela análise estatística de comparação de médias das duas amostras colhidas das notas de testes de matemáticas realizados em duas escolas diferentes (1 e 2) não nos permitem rejeitar a hipótese de que suas médias sejam iguais a um nível de confiança de 5% (cf. figura 11.26).

11.10 Teste de uma proporção amostral

A aproximação de uma população sob distribuição Binomial pela distribuição Normal pode ser realizada desde que atendidas às seguintes condições:

- a amostra é colhida de modo aleatório, os ensaios são independentes e com probabilidade de “sucesso” constante;
- se a amostra é colhida sem reposição, o tamanho da população deve ser ao menos 10 (20) vezes o tamanho da amostra ($N \geq 10, 20 \cdot n$);
- tamanho de amostra deve ser de ao menos 30 ($n \geq 30$);
- a proporção populacional não extrema (próxima a 0 ou 1);
- o número de “sucessos” deve ser de ao menos 5 ($n \cdot \pi_0 \geq 5$); e,
- o número de “fracassos” deve ser de ao menos 5 ($n \cdot (1 - \pi_0) \geq 5$).

11.10.1 Estruturas possíveis para as hipóteses

Teste bilateral (tipo: diferente de)

$$\begin{cases} H_0 : \pi = \pi_0 \\ H_1 : \pi \neq \pi_0 \end{cases}$$

Teste unilateral à esquerda (tipo: menor que)

$$\begin{cases} H_0 : \pi \geq \pi_0 \\ H_1 : \pi < \pi_0 \end{cases}$$

Teste unilateral à direita (tipo: maior que)

$$\begin{cases} H_0 : \pi \leq \pi_0 \\ H_1 : \pi > \pi_0 \end{cases}$$

Estatística do teste:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0, 1)$$

em que:

onde:

- p é a proporção observada na amostra, uma estimativa da proporção populacional π ;
- π_0 o valor (desconhecido) inferido à proporção populacional, a ser testado frente à proporção amostral; e,
- n : é o tamanho da amostra.

11.10.2 Probabilidade dos intervalos de confiança para os testes de hipóteses com o uso da estatística Z ($Z \sim \mathcal{N}(0, 1)$):

- Teste de hipóteses bilateral (tipo: diferente de):

$$\begin{aligned} P[|Z_{calc}| \leq Z_{tab(\frac{\alpha}{2})} | \pi = \pi_0] &= (1 - \alpha) \\ P(-Z_{tab(\frac{\alpha}{2})} \leq Z_{calc} \leq Z_{tab(\frac{\alpha}{2})}) &= (1 - \alpha) \end{aligned}$$

- Teste de hipóteses unilateral à esquerda (tipo: menor que):

$$P[Z_{calc} \geq Z_{tab(\alpha)} | \pi \geq \pi_0] = (1 - \alpha)$$

$$P(Z_{calc} \geq Z_{tab(\alpha)}) = (1 - \alpha)$$

- Teste de hipóteses unilateral à direita (tipo maior que):

$$P[Z_{calc} \leq Z_{tab(\alpha)} | \pi \leq \pi_0] = (1 - \alpha)$$

$$P(Z_{calc} \leq Z_{tab(\alpha)}) = (1 - \alpha)$$

Nas figuras 11.8, 11.9 e 11.10 observam-se:

- as regiões de rejeição da hipótese nula (subdivididas nos dois ou em apenas um dos lados) sob a curva da função densidade de probabilidade da distribuição adequada ao teste com probabilidades iguais ao nível de significância α ;
- a região de não rejeição da hipótese nula (delimitada à esquerda e à direita ou apenas em um dos lados) com probabilidade igual ao nível de confiança $(1 - \alpha)$; e,
- os valores críticos da estatística do teste.

Exemplo: Um relatório de uma companhia afirma que 40% de toda a água obtida a partir de poços artesianos no nordeste é salobra. Há muita controvérsia sobre essa informação, alguns dizem que a proporção é maior, outros que é menor. Para dirimir essa dúvida, 400 poços foram sorteados e observou-se em 120 deles que a água era salobra. Qual seria a conclusão a um nível de significância de 3%?

O problema nos pede um teste bilateral (tipo: diferente de):

$$\begin{cases} H_0 : \pi = 0,40 \\ H_1 : \pi \neq 0,40 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a proporção dos poços com água salobra é de 40%, fazendo então valer a hipótese alternativa que afirma ser **diferente de 40%**.

Verificação das condições:

- nada se afirmou sobre o tamanho da população para se verificar: $N \geq 10n$;
- tamanho de amostra $n \geq 30$: nossa amostra é de 400 poços;
- proporção populacional não extrema (próxima a 0 ou 1): a afirmação é de que $\pi = 0,40$; e,
- $(n \cdot \pi)$ e $(n \cdot (1 - \pi))$ são maiores que 5 (160 e 240, respectivamente).

Assim, a estatística do teste fica definida como sendo:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0, 1)$$

em que:

- $p = 0,30$ é a proporção amostral, uma estimativa da proporção populacional π ;
- $\pi_0 = 0,40$ é o valor (desconhecido) inferido à proporção populacional, a ser testado frente à proporção amostral; e,
- $n = 400$: é o tamanho da amostra.

Da tabela da distribuição Normal padronizada obtemos o valor crítico bicaudal: $|Z_{tab(\frac{\alpha}{2})}| = 2,17$. Pelo cálculo, a estatística do teste é $z_{calc} = -4,082$.

```

alfa=0.03

prob_desejada1=alfa/2
z_desejado1=round(qnorm(prob_desejada1),4)
d_desejada1=dnorm(z_desejado1, 0, 1)

prob_desejada2=1-alfa/2
z_desejado2=round(qnorm(prob_desejada2),4)
d_desejada2=dnorm(z_desejado2, 0, 1)

z_calculado=-4.082
d_calculado=dnorm(z_calculado, 0, 1)

ggplot(NULL, aes(c(-5,5))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-5, z_desejado1),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado1,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado2),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado2,5),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de z", breaks = c(z_desejado1,z_desejado2)) +
  labs(title=
      "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
       teste",
      subtitle = "P(-2,17, 2,17)=(1-\u03b1) em cinza (nível de confiança=0,97) \nP(-\U221e;
      -2,17)= P(2,17; \U221e)= \u03b1/2 em vermelho (nível de significância/2=0,015)
      ") +
  geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1),
               color="blue", lty=2, lwd=0.3) +
  geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2),
               color="blue", lty=2, lwd=0.3) +
  annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="valor crítico=-2,17",
           angle=90, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="valor crítico=2,17",
           angle=90, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=z_desejado1-1.5, y=0.1, label="Região de rejeição da hipótese nula
           \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3) +
  annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Região de rejeição da hipótese nula
           \nprobabilidade=\u03b1/2", angle=0, vjust=0, hjust=0, color="blue",size=3) +

```

```

annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Região de não rejeição da hipótese
    ↵ nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
    ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do
    ↵ teste=-4,082", angle=90, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Regiões críticas sob a curva da função densidade da distribuição apropriada ao teste

$P(-2,17, 2,17) = (1-\alpha)$ em cinza (nível de confiança=0,97)

$P(-\infty; -2,17) = P(2,17; \infty) = \alpha/2$ em vermelho (nível de significância/2=0,015)

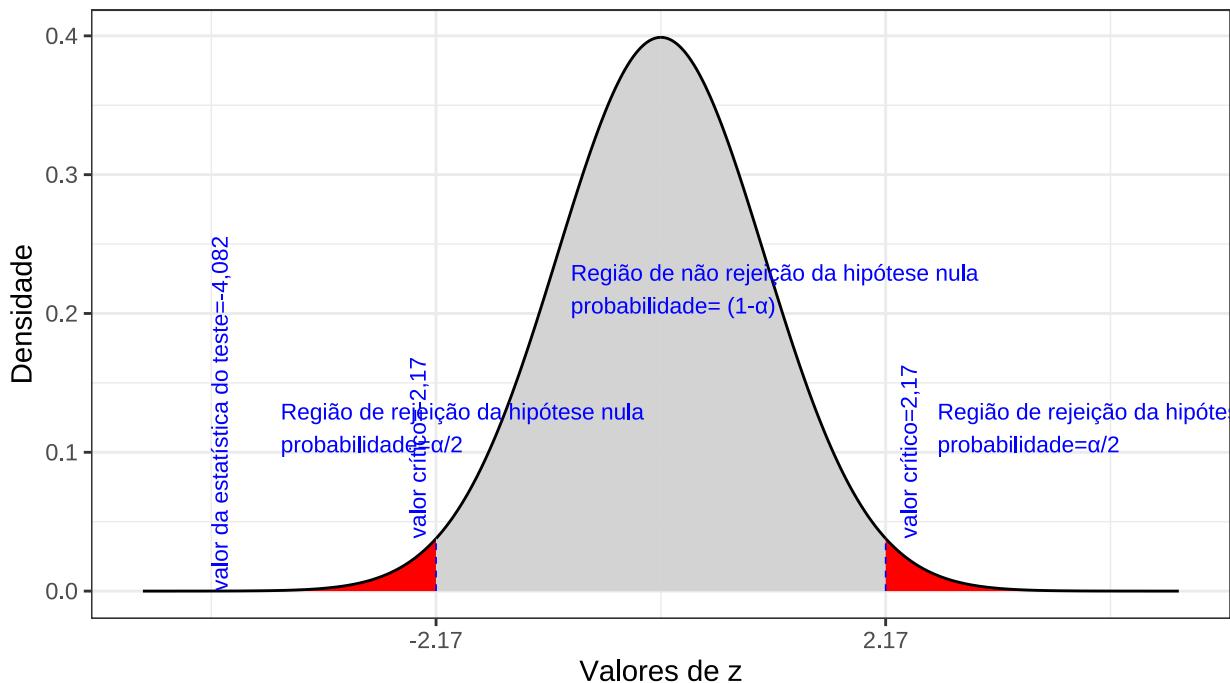


Figure 11.27: Regiões de rejeição da hipótese nula para o teste bilateral (tipo: diferente de) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valores críticos da estatística do teste: $z_{crit} = \pm 2,17$. O valor calculado da estatística ($z_{calc} = -4,082$) situa-se na faixa de significância do teste, possibilitando a rejeição da hipótese nula sob aquele nível de confiança

Conclusão: Os resultados obtidos na análise estatística realizada nos permitem rejeitar a hipótese de que a proporção de poços com água salobra é de 40% sob um nível de confiança de 97%. A proporção de poços com água salobra no Nordeste é **diferente** de 40% (Figura 11.25).

Teste unilateral à esquerda (tipo: menor que)

$$\begin{cases} H_0 : \pi \geq 0,40 \\ H_1 : \pi < 0,40 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a proporção igual ou maior a 40%, fazendo então valer a hipótese alternativa que afirma ser a proporção menor que 40%.

Da tabela obtemos o valor crítico monocaudal: $Z_{tab(\alpha)} = -1,88$. Pelo cálculo, a estatística do teste é $Z_{calc} = -4,082$.

```
alfa=0.03
prob_desejada=alfa
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

z_calculado=-4.082
d_calculado=dnorm(z_calculado, 0, 1)

ggplot(NULL, aes(c(-5,5))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(-5, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado,0),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(0, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(z_desejado,5),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de z", breaks = c(z_desejado)) +
  labs(title=
      "Região crítica sob a curva da função densidade da \ndistribuição apropriada ao
      \teste",
      subtitle = "P( -1,88,\U221e,)=(1-\u03b1) em cinza (nível de confiança=0,97)
      \nP(-\U221e; -1,88)=\u03b1 em vermelho (nível de significância=0,03) ")+
  geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue",
               lty=2, lwd=0.3)+
```

```

annotate(geom="text", x=z_desejado-0.1, y=d_desejada, label="valor crítico=-1,88", angle=90,
  ↵ vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado-2, y=0.1, label="Região de rejeição da hipótese nula
  ↵ \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
annotate(geom="text", x=z_desejado+1, y=0.2, label="Região de não rejeição da hipótese nula
  ↵ \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
  ↵ color="blue", lty=2, lwd=0.3)+
annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do
  ↵ teste=-4,082", angle=90, vjust=0, hjust=0, color="blue",size=3)+
theme_bw()

```

Região crítica sob a curva da função densidade da distribuição apropriada ao teste

$P(-1,88, \infty) = 1 - \alpha$ em cinza (nível de confiança=0,97)
 $P(-\infty; -1,88) = \alpha$ em vermelho (nível de significância=0,03)

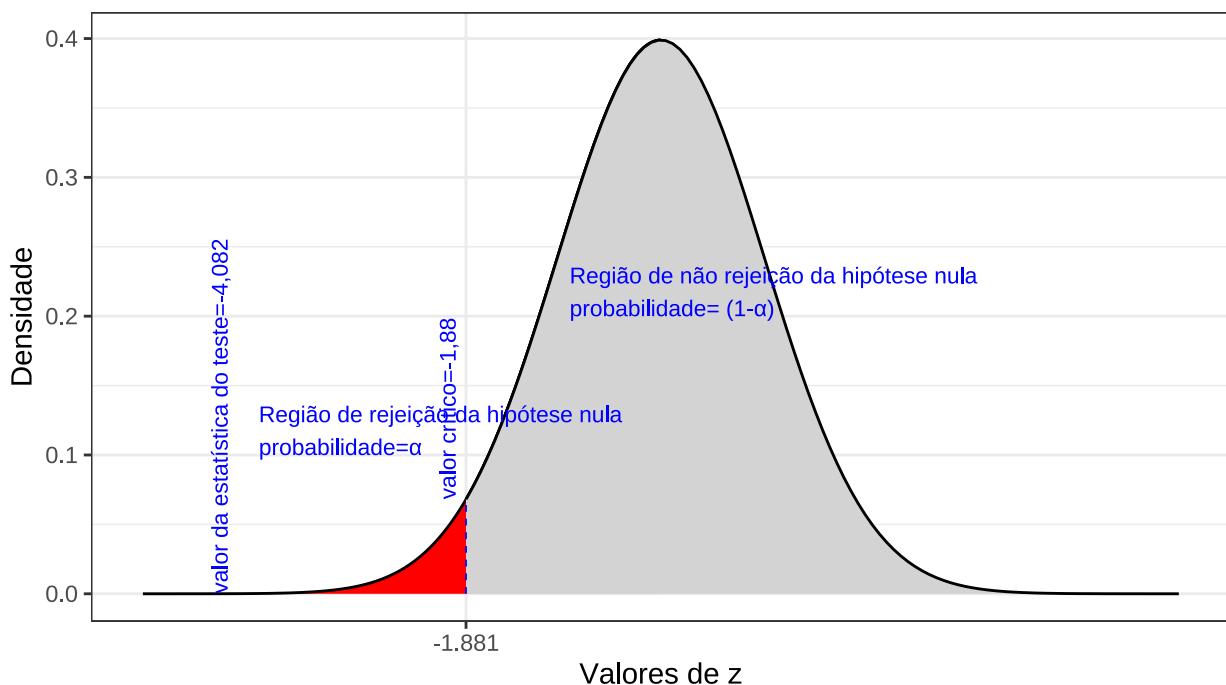


Figure 11.28: Regiões de rejeição da hipótese nula para o teste unilateral à esquerda (tipo: menor que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelos valor crítico da estatística do teste: $z_{crit} = -1,88$. O valor calculado da estatística ($z_{calc} = -4,082$) situa-se na faixa de significância do teste, o que nos permite a rejeição da hipótese nula sob aquele nível de confiança

Conclusão: Os resultados obtidos na análise estatística realizada nos permitem rejeitar a hipótese de que a proporção de poços com água salobra é de 40% sob um nível de confiança de 97%. A proporção de poços com água salobra no Nordeste é **menor que** de 40% (Figura 11.26).

Teste unilateral à direita (tipo: maior que)

$$\begin{cases} H_0 : \pi \leq 0,40 \\ H_1 : \pi > 0,40 \end{cases}$$

Iremos verificar se a informação amostral obtida nos permite rejeitar a hipótese nula que afirma ser a proporção igual ou menor a 40%, fazendo então valer a hipótese alternativa que afirma ser a proporção **maior que** 40%.

Da tabela obtemos o valor crítico monocaudal: $Z_{tab(\alpha)} = 1,88$. Pelo cálculo, a estatística do teste é $Z_{calc} = -4,082$.

```
alfa=0.97
prob_desejada=alfa
z_desejado=round(qnorm(prob_desejada),4)
d_desejada=dnorm(z_desejado, 0, 1)

z_calculado=-4.082
d_calculado=dnorm(z_calculado, 0, 1)

ggplot(NULL, aes(c(-5,5))) +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "lightgrey",
            xlim = c(-5, z_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dnorm,
            fill = "red",
            xlim = c(z_desejado,5),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de z", breaks = c(z_desejado)) +
  labs(title =
      "Região crítica sob a curva da função densidade da \ndistribuição apropriada ao
       teste",
      subtitle = "P( -\U221e; 1,88)=(1-\u03b1) em cinza (nível de confiança=0,97) \nP(1,88;
       \U221e)=\u03b1 em vermelho (nível de significância=0,03) ")+
  geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue",
               lty=2, lwd=0.3)+
  annotate(geom="text", x=z_desejado-0.1, y=d_desejada, label="valor crítico=-1,88", angle=90,
           vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado+1, y=0.1, label="Região de rejeição da hipótese nula
           \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=z_desejado-2.5, y=0.2, label="Região de não rejeição da hipótese
           nula \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
```

```
geom_segment(aes(x = z_calculado, y = 0, xend = z_calculado, yend = d_calculado),
  ↵ color="blue", lty=2, lwd=0.3) +
annotate(geom="text", x=z_calculado-0.1, y=d_calculado, label="valor da estatística do
  ↵ teste=-4,082", angle=90, vjust=0, hjust=0, color="blue",size=3) +
theme_bw()
```

Região crítica sob a curva da função densidade da distribuição apropriada ao teste

$P(-\infty; 1,88) = (1-\alpha)$ em cinza (nível de confiança=0,97)
 $P(1,88; \infty) = \alpha$ em vermelho (nível de significância=0,03)

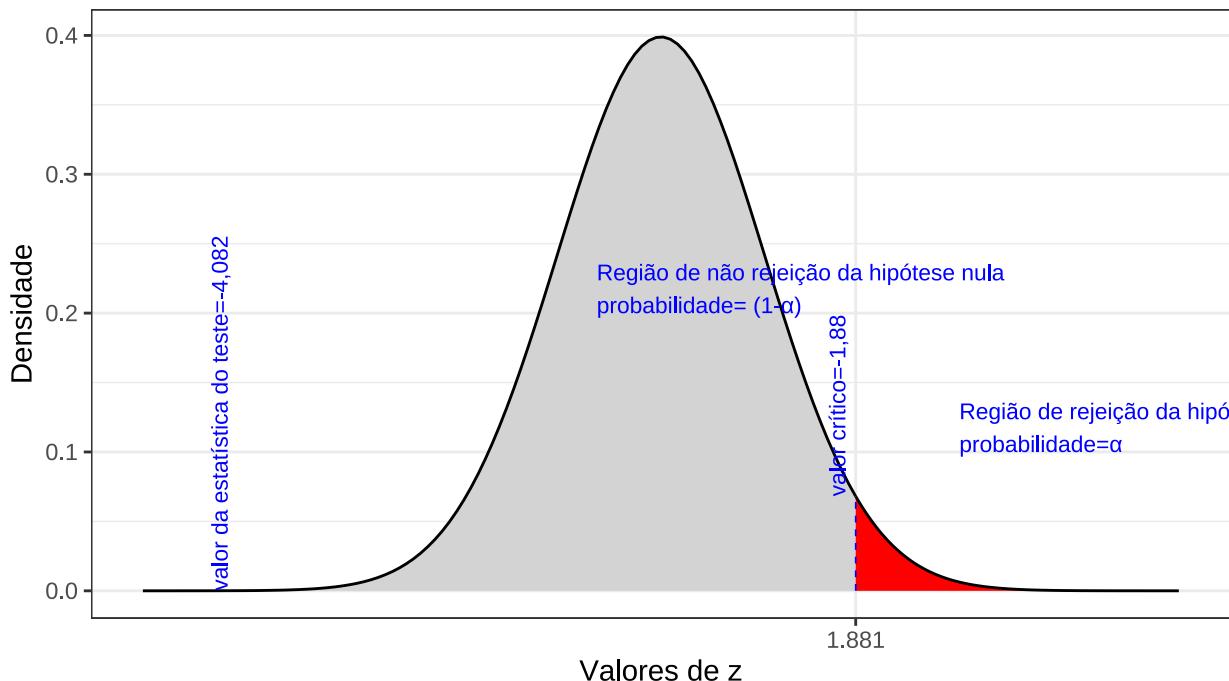


Figure 11.29: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: maior que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $z_{crit} = 1,88$. O valor calculado da estatística ($z_{calc} = -4,082$) situa-se na faixa de não significância do teste, não possibilitando a rejeição da hipótese nula sob aquele nível de confiança

Conclusão: Os resultados obtidos na análise estatística realizada não nos permitem rejeitar a hipótese de que a proporção de poços com água salobra seja menor ou igual a 40% sob um nível de confiança de 97%. (cf. Figura 11.27).

11.11 Testes não paramétricos

Um teste não paramétrico (às vezes chamado de teste livre de distribuição) não assume nada sobre a distribuição subjacente (por exemplo, que os dados vêm de uma distribuição Normal). Isso não equivale a dizer que não se saiba nada sobre a população de origem. Geralmente significa que se sabe que os dados populacionais não são de uma distribuição Normal.

Tipos de testes não paramétricos

- Teste de sinal;
- Teste de Sinal de Wilcoxon;
- Teste de Friedman;
- Teste de Mann-Whitney;
- Teste de Kruskal Wallis; e,
- Teste qui-quadrado.

Há um conjunto importante de testes de hipóteses que possibilita a análise de frequências que ocorrem nas classes de um fator.

Esses testes de hipóteses são muitas vezes referenciados como testes qui-quadrado porque a estatística do teste possui, de modo assintótico, distribuição qui-quadrado.

Embora esses testes se enquadrem em categorias distintas, compartilham algumas características comuns:

- Em cada situação considera-se a amostra aleatória, gerada por um ou mais experimentos multinomiais, independentes, de uma ou mais populações multinomiais. Obviamente, a população Bernoulli e a população binomial são casos particulares.
- A amostra aleatória é formada pelas frequências observadas nas classes, definidas pela classificação de cada uma das unidades de observação de acordo com um ou mais critérios de interesse. Em todas as situações, a estatística do teste envolve a comparação entre **frequências observadas** e **frequências esperadas**, obtidas sob a hipótese de nulidade. Na essência, o teste qui-quadrado verifica hipóteses sobre as probabilidades e utiliza a **discrepância** existente entre as **frequências observadas** e as **frequências esperadas** para concluir sobre elas. Basicamente, dispõe-se de observações (contidas na amostra) sobre uma ou mais populações e busca-se determinar de qual população multinomial essa amostra veio. A hipótese de nulidade especifica a população de interesse.
- Se as probabilidades não forem completamente especificadas, algumas probabilidades (e, consequentemente, frequências esperadas) deverão ser estimadas pelos dados, reduzindo os graus de liberdade da distribuição limite.
- Como mencionado, a distribuição limite da estatística do teste é a distribuição qui-quadrado. Uma regra usualmente exigida para uma boa aproximação da distribuição qui-quadrado é que a **frequência esperada seja maior ou igual a 5**. Evidentemente, quanto maiores forem as frequências esperadas, melhor será a aproximação.

Testes paramétricos exigem que a variável seja numérica e várias hipóteses relativas aos parâmetros sejam satisfeitas, tais como que os dados tenham uma distribuição Normal (ou a sigam assintoticamente) ou ainda,

em alguns casos que, suas variâncias sejam homogêneas (homocedasticidade) e as amostras tenham um certo tamanho ou frequência observada mínimos.

Testes não paramétricos não assumem nenhum tipo de distribuição e são menos **exigentes**, podendo também trabalhar com variáveis não numéricas. Como regra geral, opta-se por testes não paramétricos quando:

- os valores observados forem extraídos de populações que não possuem uma aproximação com a distribuição Normal;
- as populações de origem não possuem homogeneidade de variâncias (heterocedasticidade); e,
- as variáveis em estudo não apresentem medidas intervalares que possibilitem o cálculo de estatísticas tais como a média e desvios.

11.11.1 Teste Qui-quadrado para verificação da independência (homogeneidade)

O Teste Qui-quadrado de homogeneidade (ou independência) é um teste estatístico aplicado a dados categóricos para avaliar quão provável é que qualquer diferença observada nas proporções observadas entre os vários níveis de uma variável categórica em populações diferentes (ou níveis de uma segunda variável categórica) seja simples decorrência do acaso; ou seja, o teste Qui-quadrado é geralmente usado verificar quão homogêneas são entre si as frequências observadas não havendo, portanto, diferença estatisticamente significativa entre as populações (ou variáveis).

Diferenças entre o teste Qui-quadrado de homogeneidade e de independência:

- **Teste Qui-quadrado de homogeneidade:** selecionamos uma amostra de elementos de cada uma das populações e distribuímos os elementos de cada uma dessas amostras segundo as categorias da variável estudada; e,
- **Teste Qui-quadrado de independência:** distribuímos uma amostra de n elementos de apenas uma população segundo as categorias da primeira variável categórica A e as da segunda variável categórica B .

Esse tipo de investigação equivale à realização de Teste de Hipóteses onde a hipótese nula que pressupõe que existe homogeneidade (independência) na distribuição das contagens observadas em cada uma das categorias da variável nas populações amostradas (ou níveis da outra variável, no teste Qui-quadrado de Independência) será confrontada com a hipótese alternativa, de que não são homogêneas (dependência) e as flutuações não são podem ser atribuídas ao acaso.

Desse modo o foco será buscar evidência estatística robusta o suficiente que confirmem que as frequências observadas entre as diferentes populações (ou níveis da outra variável, no teste Qui-quadrado de Independência) podem ser consideradas homogêneas (independentes) sob um dado nível de significância α .

Consideremos para isso a tabela genérica para a realização do Teste Qui-quadrado onde em cada célula (habitualmente chamada de *casela*) temos uma frequência (uma quantidade) observada na Tabela a seguir.

Table 11.6: Tabela $(r \times s)$ de frequências observadas

Populações (ou uma segunda variável categórica) Variável categórica	B_1	B_2	...	B_s	Total
A_1	$n_{(1,1)}$	$n_{(1,2)}$...	$n_{(1,s)}$	$n_{(1,..)}$
A_2	$n_{(2,1)}$	$n_{(2,2)}$...	$n_{(2,s)}$	$n_{(2,..)}$
...
A_r	$n_{(r,1)}$	$n_{(r,2)}$...	$n_{(r,s)}$	$n_{(r,..)}$
Totais	$n_{(.,1)}$	$n_{(.,2)}$...	$n_{(.,s)}$	$n_{(.,.)}$

Notação utilizada na tabela:

- r é o número de linhas da tabela;
- s é o número de colunas da tabela;
- i indexa a i -ésima linha da tabela;
- j indexa a j -ésima coluna da tabela;
- $n_{i,j}$ indica o elemento localizado na casela situada na i -ésima linha e j -ésima coluna;
- $n_{(1,..)}$ indica o último elemento da primeira linha;
- $n_{(.,1)}$ indica o último elemento da primeira coluna;
- $n_{(.,.)}$ indica o último elemento simultaneamente das linhas e colunas da tabela.

Quantas observações devemos ter em cada casela da tabela acima para que as proporções observadas de A e B sejam consideradas estatisticamente homogêneas (independentes)?

Se A e B forem independentes então $P(A_i \cap B_j) = P(A_i) \times P(B_j)$.

O número esperado de observações com as características (A_i e B_j) entre as $n_{(.,.)}$ observações - sob a hipótese de homogeneidade (independência) da distribuição das contagens observadas entre das variáveis (ou da variável nas populações) - em cada casela deverá ser:

$$\begin{aligned} E_{(i,j)} &= n_{(.,.)} \times p_{(i,j)} \\ &= n_{(.,.)} \times p_{(i,.)} \times p_{(.j)} \\ &= n_{(.,.)} \times \frac{n_{(i,.)}}{n_{(.,.)}} \times \frac{n_{(.j)}}{n_{(.,.)}} \end{aligned}$$

Assim, o valor esperado - sob a hipótese de homogeneidade (independência) da distribuição das contagens observadas entre as variáveis (ou da variável nas populações) A e B - em cada célula deverá ser:

$$E_{(i,j)} = \frac{n_{(i,.)} \times n_{(.j)}}{n_{(.,.)}}$$

Em que:

- $E_{(i,j)}$ é o valor esperado na casela (i, j) ;
- $n_{(i,.)}$ é o total observado na linha i ;
- $n_{(.j)}$ é o total observado na coluna j ; e,
- $n_{(.,.)}$ é o total geral observado.

Para a aplicação do teste χ^2 exige-se que:

- preferencialmente as amostras sejam grandes ($n \geq 30$);

- no máximo 20% das caselas tenham uma frequência esperada **menor** que 5; e,
- em nenhuma casela a frequência esperada pode ser menor que 1.

A estatística (X) do Teste Qui-quadrado de homogeneidade (independência) baseia-se na diferença (diferença) entre as contagens observados e as contagens esperadas sob a suposição de homogeneidade (independência) pode ser definida da seguinte maneira:

$$X = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{(i,j)} - E_{(i,j)})^2}{E_{(i,j)}} \sim \chi^2_{((r-1) \times (s-1))}$$

e sua correspondente distribuição:

$$X \sim \chi^2_{((r-1) \times (s-1))}$$

A **hipótese nula** postula que não há associação: as variáveis são independentes. A flutuação observada nas contagens é devida apenas a fatores puramente aleatórios.

A **hipótese alternativa** a contradiz, afirmando existir algum fator não aleatório (alguma forma de associação) que resulta na distribuição não homogênea entre as contagens observadas: há dependência entre as variáveis.

$$\begin{cases} H_0 : \text{as variáveis são independentes (a flutuação nas contagens é aleatória)} \\ H_1 : \text{as variáveis não são independentes (há alguma associação)} \end{cases}$$

A distribuição de referência que permite julgar se um determinado valor da estatística X pode ser considerado grande o suficiente para rejeitar H_0 em favor de H_1 é a chamada distribuição Qui-quadrado: χ^2 .

Formulação do teste:

- teste de hipóteses unilateral à direita (tipo: maior que):

$$P[X_{calc} \leq \chi^2_{tab(\alpha; (r-1) \times (s-1))} | IND] = (1 - \alpha)$$

$$P(X_{calc} \leq \chi^2_{tab(\alpha; (r-1) \times (s-1))}) = (1 - \alpha)$$

A região de não rejeição da hipótese nula pode ser vista na Figura 11.30.

```

prob_desejada=0.95
r=4
s=3
df=(r-1)*(s-1)

q_desejado=round(qchisq(prob_desejada,df), 4)
d_desejada=dchisq(q_desejado,df)

ggplot(data.frame(x = c(0, 30)), aes(x)) +
  stat_function(fun = dchisq,
                geom = "area",
                fill = "lightgrey",
                xlim = c(0,q_desejado),
                colour="black",
                args=list(df=df) )+
  stat_function(fun = dchisq,
                geom = "area",
                fill = "red",
                xlim = c(q_desejado,30),
                colour="black",
                args = list(df = df))+ 
  scale_y_continuous(name="Densidade") +
#scale_x_continuous(name="Valores score (f)", breaks = c(f_desejado1, f_desejado2))+ 
  scale_x_continuous(name="Valores score (X)")+
  labs(title="Curva da função densidade \nDistribuição Qui-quadrado",
       subtitle = "P(0; x crítico)=(1-\u03b1) em cinza (nível de confiança) \nP(x crítico ;
       \u2192 \u03b1= \u03b1 em vermelho (nível de significância)")+
  geom_segment(aes(x = q_desejado, y = 0, xend = q_desejado, yend = d_desejada),
               color="blue", lty=2, lwd=0.3)+ 
  annotate(geom="text", x=q_desejado+0.5, y=d_desejada, label="x crítico", angle=90,
           vjust=0, hjust=0, color="blue",size=4)+ 
  annotate(geom="text", x=q_desejado+5, y=d_desejada, label="Zona de rejeição \n(para x
           calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+ 
  annotate(geom="text", x=q_desejado-5, y=d_desejada, label="Zona de não rejeição \n(para x
           calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+ 
  theme_bw()

```

Exemplo: verifique a independência (homogeneidade) nas contagens da intenção de voto de quatro

**Curva da função densidade
Distribuição Qui-quadrado**

$P(0; x \text{ crítico}) = (1-\alpha)$ em cinza (nível de confiança)
 $P(x \text{ crítico} ; \infty) = \alpha$ em vermelho (nível de significância)

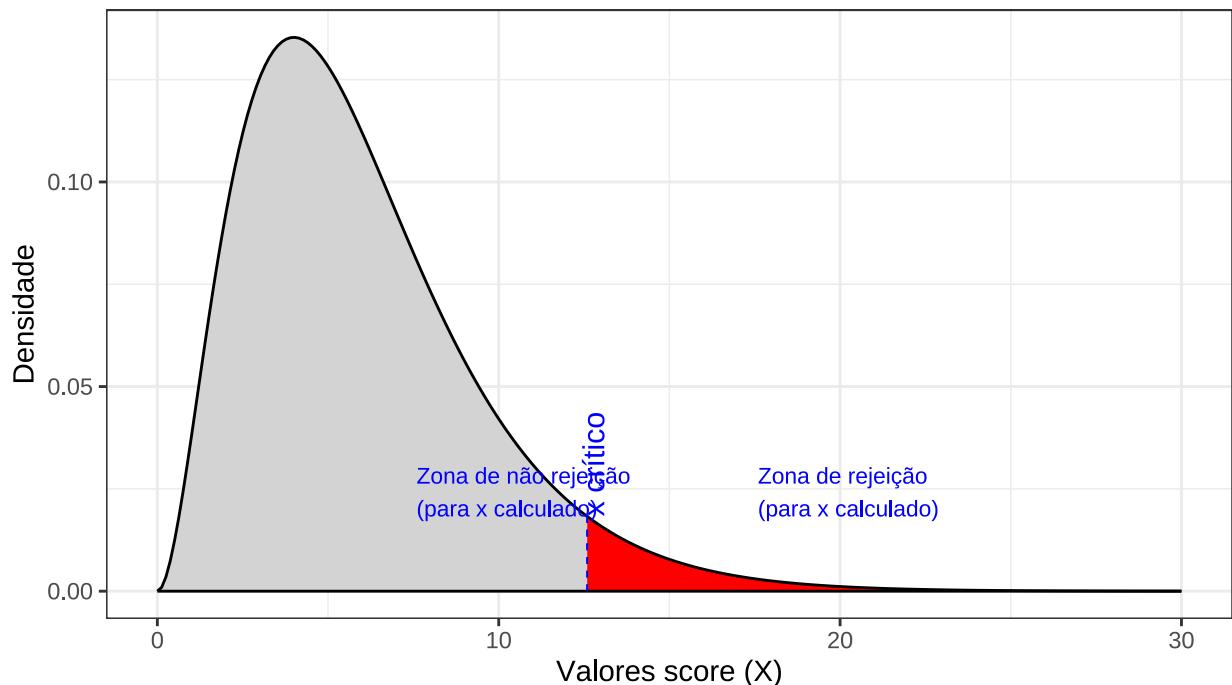


Figure 11.30: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: menor que): a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: x_{crit} para o nível de significância pretendido (α em uma cauda) e (df) graus de liberdade.

candidatos distintos em amostras de três diferentes bairros, partindo das informações consolidadas na tabela abaixo.

Table 11.7: Pesquisa sobre intenção de votos nos bairros “A”, “B” e “C”

Candidato	Bairros			Total
	“A”	“B”	“C”	
Candidato “A”	70	44	86	200
Candidato “B”	50	30	45	125
Candidato “C”	10	6	34	50
Candidato “D”	20	20	85	125
Totais	150	100	250	500

estrutura das hipóteses para o teste a um nível de significância: 0,05

$$\begin{cases} H_0 : \text{as contagens são homogêneas} \\ H_1 : \text{as contagens não são homogêneas} \end{cases}$$

Equivale dizer que há independência entre a escolha de um ou outro candidato e o bairro em questão (não há relação entre um determinado bairro e um determinado candidato)

Estatística do teste e sua distribuição:

$$X = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{(i,j)} - E_{(i,j)})^2}{E_{(i,j)}} \sim \chi^2_{((r-1) \times (s-1))}$$

Cálculo da frequência esperada em cada casela ($E_{(i,j)}$):

$$E_{(i,j)} = \frac{n_{(i,.)} \times n_{(.j)}}{n_{(..)}}$$

$$\frac{\text{soma da linha } i \times \text{soma da coluna } j}{\text{total de observações}}$$

As frequências esperadas em cada casela (i, j) serão calculadas pela fórmula acima seguir e estão apresentadas na tabela a seguir, em conjunto com as frequências observadas.

Table 11.8: Pesquisa sobre intenção de voto nos bairros “A”, “B” e “C”: frequências observadas (e entre parênteses e negrito as frequências esperadas)

Candidato	Bairros			Total
	“A”	“B”	“C”	
Candidato “A”	70 (60)	44 (40)	86 (100)	200
Candidato “B”	50 (37,5)	30 (25)	45 (62,5)	125
Candidato “C”	10 (15)	6 (10)	34 (25)	50
Candidato “D”	20 (37,5)	20 (25)	85 (62,5)	125
Totais	150	100	250	500

Nenhuma casela teve frequência esperada menor que 1 nem tampouco observou-se casela com frequência inferior a 5.

Cálculo da estatística do teste:

$$X = \sum_{i=1}^4 \sum_{j=1}^3 \frac{(O_{(i,j)} - E_{(i,j)})^2}{E_{(i,j)}} = 37,88$$

Da tabela χ^2 para o total de graus de liberdade $((r - 1) \times (s - 1)) = (4 - 1) \times (3 - 1) = 6$ obtemos o valor crítico da estatística do teste ($\chi^2_{crit(6)} = 12,60$).

```
prob_desejada=0.95
r=4
s=3
df=(r-1)*(s-1)

q_desejado=round(qchisq(prob_desejada,df), 4)
d_desejada=dchisq(q_desejado,df)

q_calculado=37.88
d_calculado=dchisq(q_calculado,df)
```

```

ggplot(data.frame(x = c(0, 50)), aes(x)) +
  stat_function(fun = dchisq,
    geom = "area",
    fill = "lightgrey",
    xlim = c(0,q_desejado),
    colour="black",
    args=list(df=df) )+
  stat_function(fun = dchisq,
    geom = "area",
    fill = "red",
    xlim = c(q_desejado,40),
    colour="black",
    args = list(df = df))+ 
  scale_y_continuous(name="Densidade") +
  #scale_x_continuous(name="Valores score (f)", breaks = c(f_desejado1, f_desejado2))+ 
  scale_x_continuous(name="Valores score (X)")+
  labs(title="Curva da função densidade \nDistribuição Qui-quadrado",
  subtitle = "P(0; 12,60)=(1-\u03b1) em cinza (nível de confiança) \nP(12,60 ; \U221e)=
  \u03b1 em vermelho (nível de significância)")+
  geom_segment(aes(x = q_desejado, y = 0, xend = q_desejado, yend = d_desejada),
  ↵ color="blue", lty=2, lwd=0.3)+ 
  annotate(geom="text", x=q_desejado+0.5, y=d_desejada, label="x crítico=12,60", angle=90,
  ↵ vjust=0, hjust=0, color="blue",size=4)+ 
  annotate(geom="text", x=q_desejado+5, y=d_desejada, label="Zona de rejeição \n(para x
  ↵ calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+ 
  annotate(geom="text", x=q_desejado-5, y=d_desejada, label="Zona de não rejeição \n(para x
  ↵ calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+ 
  geom_segment(aes(x = q_calculado, y = 0, xend = q_calculado, yend = d_calculado),
  ↵ color="blue", lty=2, lwd=0.3)+ 
  annotate(geom="text", x=q_calculado+0.5, y=d_calculado, label="x calculado=37,88",
  ↵ angle=90, vjust=0, hjust=0, color="blue",size=4)+ 
  theme_bw()

```

Conclusão: face aos dados trazidos à análise rejeitamos a proposição de que a preferência por um determinado candidato não esteja de algum modo associada ao bairro pesquisado sob um nível de significância de 5% (a probabilidade de cometimento de um erro tipo I. Há alguma relação entre a preferência por um ou outro candidato e os bairros (Figura 11.31). .

11.11.2 Correção de continuidade em tabelas 2x2

Em tabelas de dimensão 2x2, especialmente quando as amostras não forem muito grandes, recomenda-se aplicar a chamada correção de continuidade de Yates, que consiste em reduzir 0,5 unidade nas diferenças absolutas entre as frequências observadas e esperadas:

**Curva da função densidade
Distribuição Qui-quadrado**

$P(0; 12,60) = (1-\alpha)$ em cinza (nível de confiança)
 $P(12,60 ; \infty) = \alpha$ em vermelho (nível de significância)

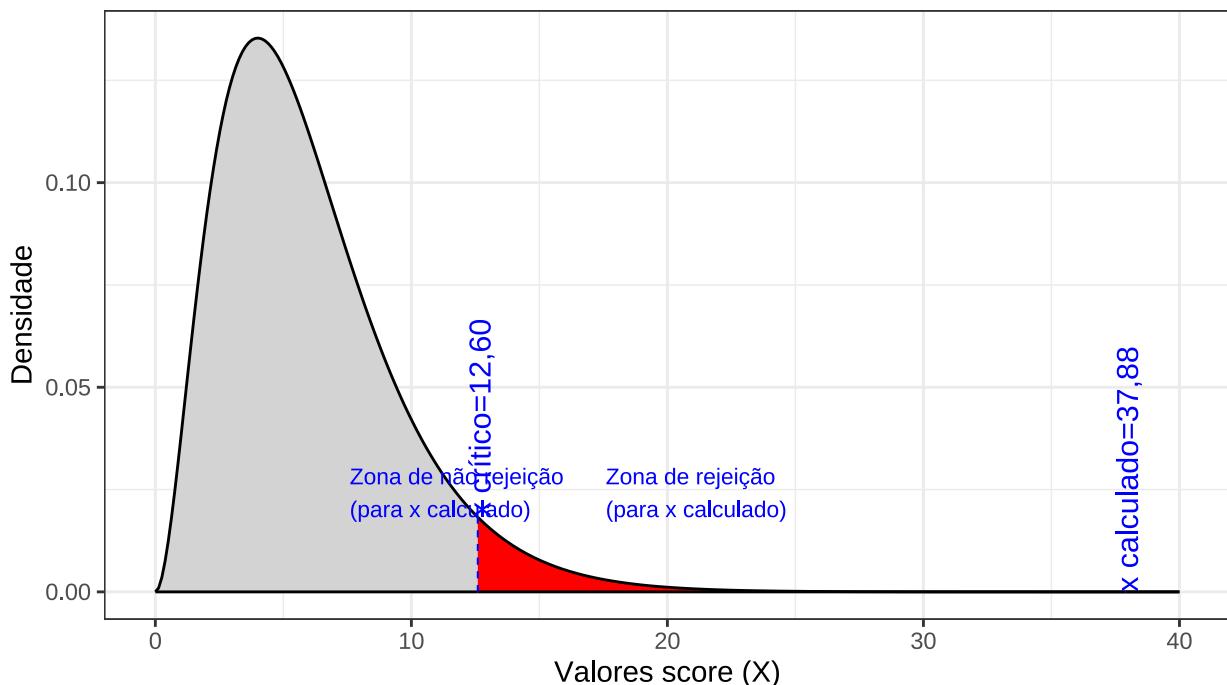


Figure 11.31: Região de rejeição da hipótese nula para o teste uniletaral à direita (tipo: menor que): a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $x_{crit} = 12,60$ para o nível de significância pretendido ($\alpha = 0,05$ em uma cauda) e ($df = 6$) graus de liberdade.

$$X = \sum_{i=1}^r \sum_{j=1}^s \frac{(|O_{(i,j)} - E_{(i,j)}| - 0,5)^2}{E_{(i,j)}}$$

Ou seja, em cada casela, depois de calculada a diferença entre a frequência observada e a frequência esperada, tomamos o módulo dessa operação (isto é, despreza-se o sinal \pm) e reduz-se esse valor em 0,5 unidade para, em seguida, elevamos ao quadrado e então dividir-se pela frequência esperada da célula.

11.11.3 Coeficiente de contingência de Pearson (modificado: C^*) }

Como vimos, a aplicação do teste qui-quadrado permite verificar se existe associação entre duas variáveis, com base em um conjunto de observações. A intensidade dessa associação pode ser quantificada por coeficientes que têm por objetivo medir a força da associação entre duas variáveis categorizadas. Um deles é o chamado coeficiente de contingência de Pearson modificado (uma correção em razão da dimensão da tabela).

Um coeficiente de associação, aplicado a uma tabela de contingência, produz um valor numérico que descreve se os dados se aproximam mais de uma situação de independência ($C^* = 0$) ou de uma situação de associação ou dependência perfeita ($C^* = 1$).

$$C^* = \sqrt{\frac{k \times X^2}{(k-1) \times (n + X^2)}}$$

em que:

- k é o menor valor entre o número de linhas (l) e de colunas (c) da tabela;
- n é o número de elementos da tabela; e,
- X^2 : valor calculado da estatística do teste qui-quadrado.

Exemplo: no exercício resolvido anteriormente ($X^2 = 37,88$ e uma tabela 3×4 com 500 observações) teremos o seguinte valor para o coeficiente de contingência modificado (C^*):

$$\begin{aligned}
C^* &= \sqrt{\frac{k \times X^2}{(k-1) \times (n + X^2)}} \\
&= \sqrt{\frac{3 \times 37,88}{(3-1) \times (500 + 37,88)}} \\
&= \sqrt{\frac{113,64}{(2) \times (537,88)}} \\
&= \sqrt{0,105637} \\
&= 0,325
\end{aligned}$$

11.11.4 Teste Qui-quadrado para verificação da qualidade do ajuste a uma distribuição teórica de probabilidade

O teste de ajuste de qui-quadrado é um teste não paramétrico usado para descobrir como o valor observado de um dado fenômeno é significativamente diferente do valor esperado.

No teste de ajuste do qui-quadrado, o termo qualidade de ajuste (*goodness-of-fit*) é usado para comparar a distribuição da amostra observada com uma distribuição teórica de probabilidade esperada. O teste de ajuste do qui-quadrado determina quão bem a distribuição teórica (como Normal, binomial ou Poisson) se encaixa na distribuição empírica.

No teste de ajuste do qui-quadrado, os dados da amostra são divididos em intervalos. Em seguida, os números de pontos que se enquadram no intervalo são comparados, com o número esperado de pontos em cada intervalo. Considere-se a seguinte tabela com as observações agrupadas em classes.

Table 11.9: Dados observados agrupados em classes

ID	Classes	Frequência observada (f_{obs_i})	Frequência teórica esperada (f_{esp_i})	$\frac{(f_{obs_i} - f_{esp_i})^2}{f_{esp_i}}$
1	$lim_{inf} \leftarrow lim_{sup}$	f_{obs_1}	f_{esp_1}
2	$lim_{inf} \leftarrow lim_{sup}$	f_{obs_2}	f_{esp_2}
...
k	$lim_{inf} \leftarrow lim_{sup}$	f_{obs_k}	f_{esp_k}	
Totais	-	$\sum_{i=1}^k f_{obs_i}$	-	$X_{calc} = \sum_{i=1}^k \frac{(f_{obs_i} - f_{esp_i})^2}{f_{esp_i}}$

A frequência esperada em cada classe, sob a suposição de que os dados seguem uma distribuição Normal: $X \sim \mathcal{N}(\mu, \sigma)$ é dada por:

$$\begin{aligned} f_{esp_i} &= P[lim_{inf_i} \leq X \leq lim_{sup_i}] \times \sum_{i=1}^k f_{obs_i} \\ &= P\left[\frac{(lim_{inf_i} - \mu)}{\sigma} \leq Z \leq \frac{(lim_{sup_i} - \mu)}{\sigma}\right] \times \sum_{i=1}^k f_{obs_i} \end{aligned}$$

Há de se considerar duas situações: μ e σ conhecidos, ou estimados a partir dos dados da amostra.

Caso sejam conhecidos, demonstra-se que $X_{calc} \sim \chi^2_{(k-1)}$; na outra situação, se forem estimados a partir da amostra (usando-se \bar{x} e s) então, igualmente, tem-se que $X_{calc} \sim \chi^2_{(k-1-2)}$, apenas com a perda de dois graus de liberdade pelas estimativas feitas.

A estatística do teste qui-quadrado de qualidade de ajuste baseia-se na distância entre as frequências observadas e as frequências esperadas sob a distribuição de probabilidade considerada e pode então ser definida, bem como o teste de hipóteses, da seguinte maneira:

$$X_{calc} = \sum_{i=1}^k \frac{(f_{obs_i} - f_{esp_i})^2}{f_{esp_i}}$$

Demonstra-se que para uma amostra grande e com classes com frequências esperadas ($f_{esp_i} \geq 5$) que $X_{calc} \sim \chi^2(k-1)$ e o correspondente teste de hipóteses assume a estrutura seguinte:

$$\begin{cases} H_0 : X \text{ segue o modelo teórico proposto} \\ H_1 : X \text{ não segue o modelo proposto} \end{cases}$$

Formulação do teste:

- Teste de hipóteses unilateral à direita (tipo: maior que):

$$P[X_{calc} \leq \chi^2_{tab(\alpha; (k-1))} | X \sim \mathcal{N}] = (1 - \alpha)$$

$$P(X_{calc} \leq \chi^2_{tab(\alpha; (k-1))}) = (1 - \alpha)$$

```

prob_desejada=0.95
r=4
s=3
df=(r-1)*(s-1)

q_desejado=round(qchisq(prob_desejada,df), 4)
d_desejada=dchisq(q_desejado,df)

ggplot(data.frame(x = c(0, 30)), aes(x)) +
  stat_function(fun = dchisq,
                geom = "area",
                fill = "lightgrey",
                xlim = c(0,q_desejado),
                colour="black",
                args=list(df=df) )+
  stat_function(fun = dchisq,
                geom = "area",
                fill = "red",
                xlim = c(q_desejado,30),
                colour="black",
                args = list(df = df))+ 
  scale_y_continuous(name="Densidade") +
  #scale_x_continuous(name="Valores score (f)", breaks = c(f_desejado1, f_desejado2))+ 
  scale_x_continuous(name="Valores score (X)")+
  labs(title="Curva da função densidade \nDistribuição Qui-quadrado",
       subtitle = "P(0; x crítico)=(1-\u03b1) em cinza (nível de confiança) \nP(x crítico ;
       \u2192 \U221e)= \u03b1 em vermelho (nível de significância)")+
  geom_segment(aes(x = q_desejado, y = 0, xend = q_desejado, yend = d_desejada),
               color="blue", lty=2, lwd=0.3)+ 
  annotate(geom="text", x=q_desejado+0.5, y=d_desejada, label="x crítico", angle=90,
           vjust=0, hjust=0, color="blue",size=4)+ 
  annotate(geom="text", x=q_desejado+5, y=d_desejada, label="Zona de rejeição \n(para x
           calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+ 
  annotate(geom="text", x=q_desejado-8, y=d_desejada, label="Zona de não rejeição \n(para x
           calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+ 
  theme_bw()

```

Exemplo: deseja-se verificar a afirmação de que a porcentagem de cinzas (material estranho ao produto) contidas em café torrado e moído produzido por certa empresa de torrefação segue uma

**Curva da função densidade
Distribuição Qui-quadrado**

$P(0; x \text{ crítico}) = (1-\alpha)$ em cinza (nível de confiança)
 $P(x \text{ crítico} ; \infty) = \alpha$ em vermelho (nível de significância)

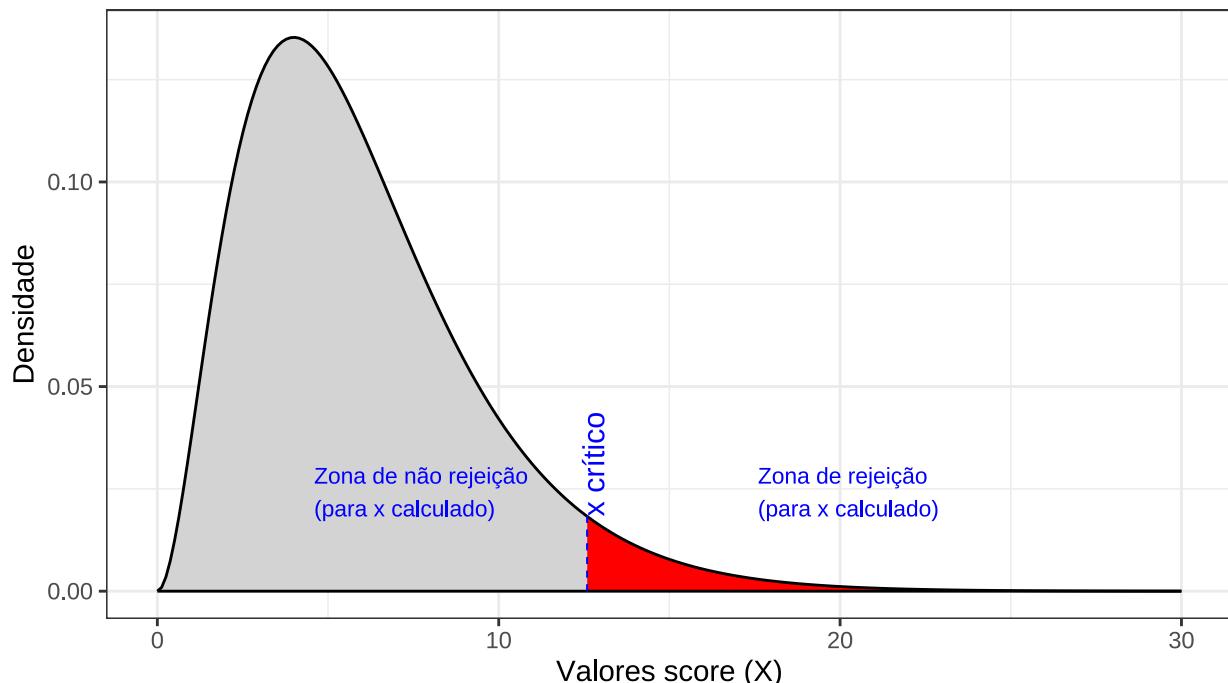


Figure 11.32: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: menor que): a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: x_{crit} para o nível de significância pretendido (α em uma cauda) e (df) graus de liberdade.

distribuição Normal. Os dados abaixo representam a quantidade percentual desse material encontradas em 250 amostras analisadas em laboratório. Faça um teste qui-quadrado de adequação das frequências observadas a essa distribuição com um nível de significância $\alpha = 0.04$.

Table 11.10: Análise da presença de cinzas em café torrado e moído

ID (k)	Cinzas de material estranho (%)	Frequência observada (f_{obs_i})
1	9,50 ⊢ 10,50	2
2	10,50 ⊢ 11,50	5
3	11,50 ⊢ 12,50	16
4	12,50 ⊢ 13,50	42
5	13,50 ⊢ 14,50	69
6	14,50 ⊢ 15,50	51
7	15,50 ⊢ 16,50	32
8	16,50 ⊢ 17,50	23
9	17,50 ⊢ 18,50	9
10	18,50 ⊢ 19,50	1
Totais		250

Análise do problema: verificar se as frequências observadas nas classes diferem das que seriam esperadas se a distribuição dessa variável seguisse uma distribuição Normal com parâmetros μ e σ (não informados pelo enunciado do problema).

Essa omissão nos força a utilizar a média e o desvio padrão amostrais (\bar{x} e S) como suas estimativas.

Isso irá nos impor a perda adicional de mais dois graus de liberdade na estatística do teste: $\chi^2_{(k-1-2)}$.

Para dados agrupados em classes a média e a variância são calculados por:

$$\sum_{i=1}^k \frac{\bar{x}_i \cdot f_{obs_i}}{n} = 14,512$$

e

$$S^2 = \frac{\sum_{i=1}^k (\bar{x}_i - \bar{x})^2 \times f_{obs_i}}{n-1} = 2,701$$

Na sequência, calculam-se as frequências esperadas para cada classe sob a premissa de Normalidade. Abaixo mostramos o cálculo para a primeira classe:

$$\begin{aligned}
 f_{esp_i} &= P[lim_{inf_i} \leq X \leq lim_{sup_i}] \cdot \sum_{i=1}^k f_{obs_i} \\
 &= P[9,50 \leq X \leq 10,50] \times 250 \\
 &= P\left[\frac{(lim_{inf_i} - \mu)}{\sigma} \leq Z \leq \frac{(lim_{sup_i} - \mu)}{\sigma}\right] \times \sum_{i=1}^k f_{obs_i} \\
 &= P\left[\frac{(9,50 - 14,512)}{\sqrt{2,701}} \leq Z \leq \frac{(10,50 - 14,512)}{\sqrt{2,701}}\right] \times 250 \\
 &= P\left[\frac{(9,50 - 14,512)}{\sqrt{2,701}} \leq Z \leq \frac{(10,50 - 14,512)}{\sqrt{2,701}}\right] \times 250 \\
 &= P[-3,0496 \leq Z \leq -2,4412] \times 250 \\
 &= (0,4989 - 0,4927) \times 250 \\
 &= (0,0062) \times 250 \\
 &= 1,55
 \end{aligned}$$

Table 11.11: Análise da presença de cinzas em café torrado e moído

ID (k)	Cinzas de material estranho (%)	Frequência observada (f_{obs_i})	Frequência teórica esperada (f_{esp_i})	$\frac{(f_{obs_i} - f_{esp_i})^2}{f_{esp_i}}$
1	9,50 ⊢ 10,50	2	1,543559	
2	10,50 ⊢ 11,50	5	6,525845	
3	11,50 ⊢ 12,50	16	19,25203	
4	12,50 ⊢ 13,50	42	39,648	
5	13,50 ⊢ 14,50	69	57,01595	
6	14,50 ⊢ 15,50	51	57,26207	
7	15,50 ⊢ 16,50	32	40,16374	
8	16,50 ⊢ 17,50	23	19,67134	
9	17,50 ⊢ 18,50	9	6,725776	
10	18,50 ⊢ 19,50	1	1,604656	
Totais		250	-	-

As frequências esperadas para as classes 1 e 10 são menores que 5 ($f_{esp_i} \geq 5$) impondo que essas duas classes sejam agrupadas às classes imediatamente adjacentes.

Table 11.12: Análise da presença de cinzas em café torrado e moído

ID (k)	Cinzas de material estranho (%)	Frequência observada (f_{obs_i})	Frequência teórica esperada (f_{esp_i})	$\frac{(f_{obs_i} - f_{esp_i})^2}{f_{esp_i}}$
1-2	9,50 – 11,50	7	8,069404	0,141724
3	11,50 – 12,50	16	19,25203	0,549329
4	12,50 – 13,50	42	39,648	0,139525
5	13,50 – 14,50	69	57,01595	2,518900
6	14,50 – 15,50	51	57,26207	0,684808
7	15,50 – 16,50	32	40,16374	1,659374
8	16,50 – 17,50	23	19,67134	0,563255
9-10	17,50 – 19,50	10	8,330432	0,334611
Totais		250	-	6,591525

Estrutura do teste: teste de hipóteses unilateral à direita (tipo: maior que):

$$\begin{cases} H_0 : X \sim \mathcal{N}(\bar{x}, S) \\ H_1 : X \text{ não segue o modelo proposto} \end{cases}$$

A hipótese nula postula que a variável X segue a distribuição Normal ($X \sim \mathcal{N}(\bar{x}, S)$)

Estatística do teste:

$$x_{calc} = \sum_{i=1}^k \frac{(f_{obs_i} - f_{esp_i})^2}{f_{esp_i}} = 6,59$$

Valor crítico da estatística de teste $\chi^2_{(\alpha), (k-1-2)}$:

$$\chi^2_{(0,04), (8-1-2)} = 11,64$$

```
prob_desejada=0.96
df=5

q_desejado=round(qchisq(prob_desejada,df), 4)
d_desejada=dchisq(q_desejado,df)

q_calculado=round(6.59, 4)
d_calculada=dchisq(q_calculado,df)
```

```

ggplot(data.frame(x = c(0, 30)), aes(x)) +
  stat_function(fun = dchisq,
    geom = "area",
    fill = "lightgrey",
    xlim = c(0,q_desejado),
    colour="black",
    args=list(df=df) )+
  stat_function(fun = dchisq,
    geom = "area",
    fill = "red",
    xlim = c(q_desejado,30),
    colour="black",
    args = list(df = df))+ 
  scale_y_continuous(name="Densidade") +
  #scale_x_continuous(name="Valores score (f)", breaks = c(f_desejado1, f_desejado2))+ 
  scale_x_continuous(name="Valores score (X)")+
  labs(title="Curva da função densidade \nDistribuição Qui-quadrado",
  subtitle = "P(0; 11,64)=0,96 em cinza (nível de confiança) \nP(11,64 ; \U221e)= 0,04 em
  ↵ vermelho (nível de significância)")+
  geom_segment(aes(x = q_desejado, y = 0, xend = q_desejado, yend = d_desejada),
  ↵ color="blue", lty=2, lwd=0.3)+ 
  annotate(geom="text", x=q_desejado+0.5, y=d_desejada, label="x crítico=11,64", angle=90,
  ↵ vjust=0, hjust=0, color="blue",size=4)+ 
  annotate(geom="text", x=q_desejado+5, y=d_desejada, label="Zona de rejeição \n(para x
  ↵ calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+ 
  annotate(geom="text", x=q_desejado-8, y=d_desejada, label="Zona de não rejeição \n(para x
  ↵ calculado)", angle=0, vjust=0, hjust=0, color="blue",size=3)+ 
  geom_segment(aes(x = q_calculado, y = 0, xend = q_calculado, yend = d_calculada),
  ↵ color="blue", lty=2, lwd=0.3)+ 
  annotate(geom="text", x=q_calculado+0.5, y=d_calculada, label="x calculado=6,59",
  ↵ angle=90, vjust=0, hjust=0, color="blue",size=4)+ 
  theme_bw()

```

Conclusão:

O resultado do teste de hipóteses realizado com as amostras trazidas à análise não nos permite rejeitar a afirmação de que os seus valores procedem de uma distribuição Normal ($X \sim \mathcal{N}(\bar{x}=14,512, S=1,6435)$) a um nível de significância de 4% (Figura 11.33).

11.11.5 Teste de significância para as médias de duas populações dependentes

O Teste “t” emparelhado é usado quando dados das duas amostras são colhidas de um mesmo indivíduo (ensaio clínico) ou em uma mesma unidade experimental (experimento agronômico) havendo, portanto, dependência

**Curva da função densidade
Distribuição Qui-quadrado**

$P(0; 11,64)=0,96$ em cinza (nível de confiança)
 $P(11,64 ; \infty)= 0,04$ em vermelho (nível de significância)

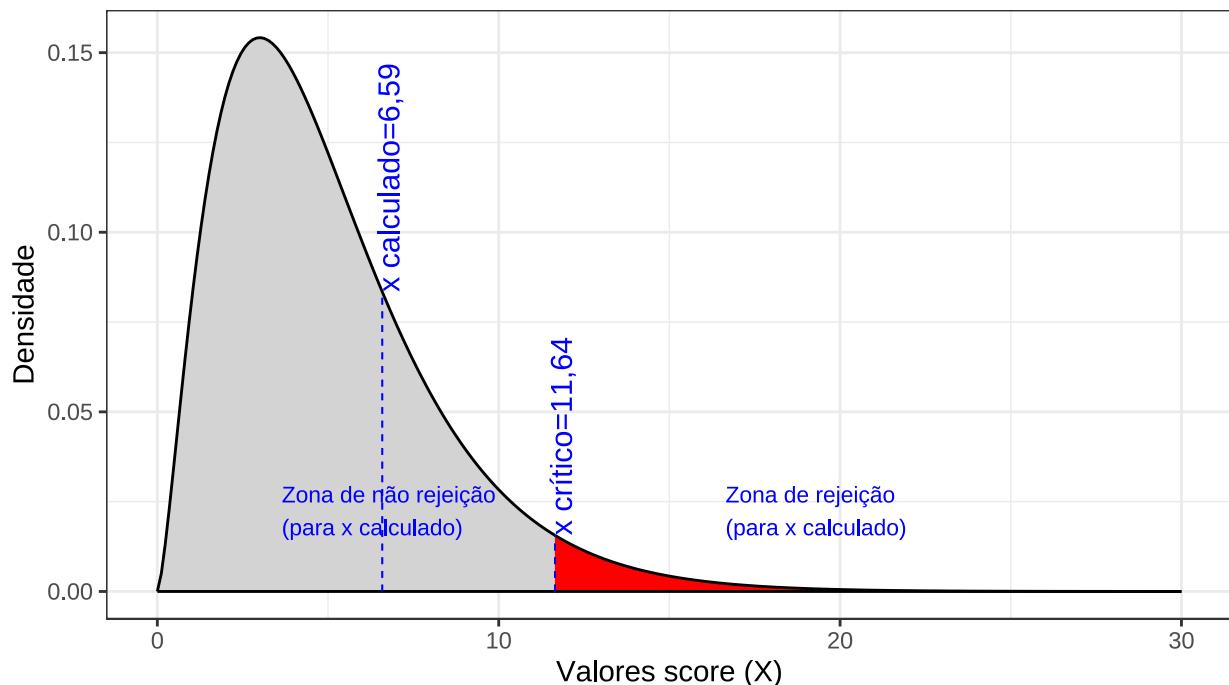


Figure 11.33: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: menor que): a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $x_{crit} = 11,64$ para o nível de significância pretendido (α em uma cauda) e (df) graus de liberdade.

entre os valores observados.

As possíveis estruturas dos testes de hipóteses para duas médias dependentes (amostras emparelhadas) são:

- Teste de hipóteses bilateral (tipo: diferente de):

$$\begin{cases} H_0 : \mu_{\text{dif}} = \Delta_0 \\ H_1 : \mu_{\text{dif}} \neq \Delta_0 \end{cases}$$

- Teste de hipóteses unilateral à esquerda (tipo: menor que):

$$\begin{cases} H_0 : \mu_{\text{dif}} \geq \Delta_0 \\ H_1 : \mu_{\text{dif}} < \Delta_0 \end{cases}$$

- Teste de hipóteses unilateral à direita (tipo: maior que):

$$\begin{cases} H_0 : \mu_{\text{dif}} \leq \Delta_0 \\ H_1 : \mu_{\text{dif}} > \Delta_0 \end{cases}$$

em que:

- Δ_0 é, usualmente, 0 (as médias são iguais); e,
- $\mu_{\text{dif}} = \mu_1 - \mu_2$ é a diferença entre os pares de observações;

Estatística do teste para amostras Normais (n_1 e n_2 quaisquer) ou amostras de outras distribuições, mas desde que n_1 e $n_2 \geq 30$:

- $t_{cal} = \frac{\sqrt{n} \cdot (\bar{x}_{dif} - \Delta_0)}{S_{dif}}$
- \bar{x}_{dif} : valor médio das diferenças entre as observações (amostra)
- S_{dif} : desvio padrão das diferenças entre as observações (amostra)
- $t_{tab(\frac{\alpha}{2}; n-1)}$ ou $t_{tab(\alpha; n-1)}$: o quantil associado na distribuição “t” de Student ao nível de significância pretendido no teste, com $(n - 1)$ graus de liberdade.

Formulação dos testes com a estatística T ($T \sim t_{(n-1)}$):

- Teste de hipóteses bilateral (tipo: diferente de):

$$\begin{aligned} P[|t_{calc}| \geq t_{tab(\frac{\alpha}{2}; n-1)} | \mu_{dif} = 0] &= (1 - \alpha) \\ P(-t_{tab(\frac{\alpha}{2}; n-1)} \leq t_{calc} \leq t_{tab(\frac{\alpha}{2}; n-1)}) &= (1 - \alpha) \end{aligned}$$

As regiões de rejeição (regiões críticas) da hipótese nula podem ser vistas na Figura 11.14.

- Teste de hipóteses unilateral à esquerda (tipo: menor que):

$$\begin{aligned} P[t_{calc} \geq t_{tab(\alpha; n-1)} | \mu_{dif} = 0] &= (1 - \alpha) \\ P(t_{calc} \geq t_{tab(\alpha; n-1)}) &= (1 - \alpha) \end{aligned}$$

A região de rejeição (região crítica) da hipótese nula pode ser vista na Figura 11.15.

- Teste de hipóteses unilateral à direita (tipo: maior que):

$$\begin{aligned} P[t_{\text{calc}} \leq t_{\text{tab}(\alpha; n-1)} | \mu_{\text{dif}} = 0] &= (1 - \alpha) \\ P(t_{\text{calc}} \leq t_{\text{tab}(\alpha; n-1)}) &= (1 - \alpha) \end{aligned}$$

A região de rejeição (região crítica) da hipótese nula pode ser vista na Figura 11.16.

Exemplo: Uma empresa precisa tomar a decisão de adquirir uma nova máquinas de usinagem. Contudo, o fornecedor apresentou dois modelos (A e B) de preços diferentes. Para tomar a decisão, convocou 5 de seus funcionários mais experientes e os despachou para a fábrica, que os treinou a executar a mesma tarefa em ambas as máquinas. A tabela abaixo apresenta os tempos gastos pelos funcionários em ambas as máquinas (cf. tabela ??). No nível de significância de 10% podemos afirmar que a tarefa realizada na máquina *A* demora mais que na máquina *B*?

Table 11.13: Tempo necessário para usinagem de uma mesma peça em duas máquinas diferentes, por 5 operadores diferentes

Funcionário	Máquina A (h)	Máquina B (h)
A	80	75
B	72	70
C	65	60
D	78	72
E	85	78

O enunciado do problema deixa bastante claro que as medidas, os tempos gastos para a realização da tarefa nas máquinas A e B foram tomados no mesmo grupo de funcionários, de tal sorte que não nos é possível afirmar que há independência. O Teste “t” é usado quando dados das duas amostras são colhidas de um mesmo sujeito, havendo, portanto dependência entre as amostras. A tabela a seguir apresenta as diferenças de tempo de usinagem entre as máquinas, para cada operador.

Estrutura do teste: teste de hipóteses unilateral à direita (tipo: maior que):

$$\begin{cases} H_0 : \mu_{\text{dif}}(\mu_A - \mu_B) \leq 0 \\ H_1 : \mu_{\text{dif}}(\mu_A - \mu_B) > 0 \end{cases}$$

Table 11.14: Diferenças nos tempos de usinagem

Funcionário	Diferença: A-B (h)
A	5
B	2
C	5
D	6
E	7
Média	5,00
Desvio padrão	1,8708

A hipótese nula afirma que o tempo médio μ_A é igual ou menor que o tempo médio μ_B ; já a hipótese alternativa, contrariamente, afirma que o tempo médio μ_A é maior que o tempo médio μ_B . Estatística do teste:

$$t_{cal} = \frac{\sqrt{n} \times (\bar{x}_{dif})}{S_{dif}}$$

$$t_{calc} > t_{tab(\alpha;(n-1))}$$

em que:

- $n = 5$;
- $t_{tab(0,10;(5-1))} = 1,533$ é o quantil associado na distribuição “t” de Student no nível de significância pretendido no teste e com $(n - 1)$ graus de liberdade (valor crítico monocaudal);
- $t_{cal} = \frac{\sqrt{n} \cdot (\bar{x}_{dif})}{S_{dif}} = 5,97$;
- $\bar{x}_{dif} = 5,00$ é o valor médio das diferenças entre as observações amostrais;
- $S_{dif} = 1,87$: desvio padrão das diferenças entre as observações amostrais.

```
alfa=0.90
prob_desejada=alfa
df=4
t_desejado=round(qt(prob_desejada,df ),4)
d_desejada=dt(t_desejado,df)

t_calculado=5.97
d_calculado=dt(t_calculado,df)
```

```

ggplot(NULL, aes(c(-7,7))) +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "lightgrey",
            xlim = c(-7, t_desejado),
            colour="black") +
  geom_area(stat = "function",
            fun = dt,
            args=list(df),
            fill = "red",
            xlim = c(t_desejado,7),
            colour="black") +
  scale_y_continuous(name="Densidade") +
  scale_x_continuous(name="Valores de t", breaks = c(t_desejado)) +
  labs(title=
      "Regiões críticas sob a curva da função densidade da \ndistribuição apropriada ao
      ← teste",
      subtitle = "P(-\U221e; 1,53)=(1-\u03b1) em cinza (nível de confiança=0,90) \nP(1,53;
      ← \U221e)= \u03b1 em vermelho (nível de significância=0,10) ")+
  geom_segment(aes(x = t_desejado, y = 0, xend = t_desejado, yend = d_desejada),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=t_desejado-0.1, y=d_desejada, label="Valor crítico da estatística
      ← do teste=1,53", angle=90, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=t_desejado-3, y=0.1, label="Região de não rejeição da hipótese
      ← nula \nprobabilidade=\u03b1", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  annotate(geom="text", x=t_desejado+1, y=0.1, label="Região de rejeição da hipótese nula
      ← \nprobabilidade= (1-\u03b1)", angle=0, vjust=0, hjust=0, color="blue",size=3)+
  geom_segment(aes(x = t_calculado, y = 0, xend = t_calculado, yend = d_calculado),
               color="blue", lty=2, lwd=0.3)+
  annotate(geom="text", x=t_calculado-0.1, y=d_calculado, label="Valor da estatística do
      ← teste=5,97", angle=90, vjust=0, hjust=0, color="blue",size=3)+
  theme_bw()

```

Conclusão:

O resultado do teste de hipóteses realizado com as amostras trazidas à análise não nos permite suportar a afirmação de que o tempo médio para a realização da tarefa na máquina *A* seja menor ou igual ao tempo médio gasto na máquina *B* a um nível de significância de 10%. O tempo médio na máquina *A* é maior (Figura 11.34).

11.12 Fluxograma auxiliar para escolha da estatística do teste de hipóteses

Regiões críticas sob a curva da função densidade da distribuição apropriada ao teste

$P(-\infty; 1,53) = (1-\alpha)$ em cinza (nível de confiança=0,90)

$P(1,53; \infty) = \alpha$ em vermelho (nível de significância=0,10)

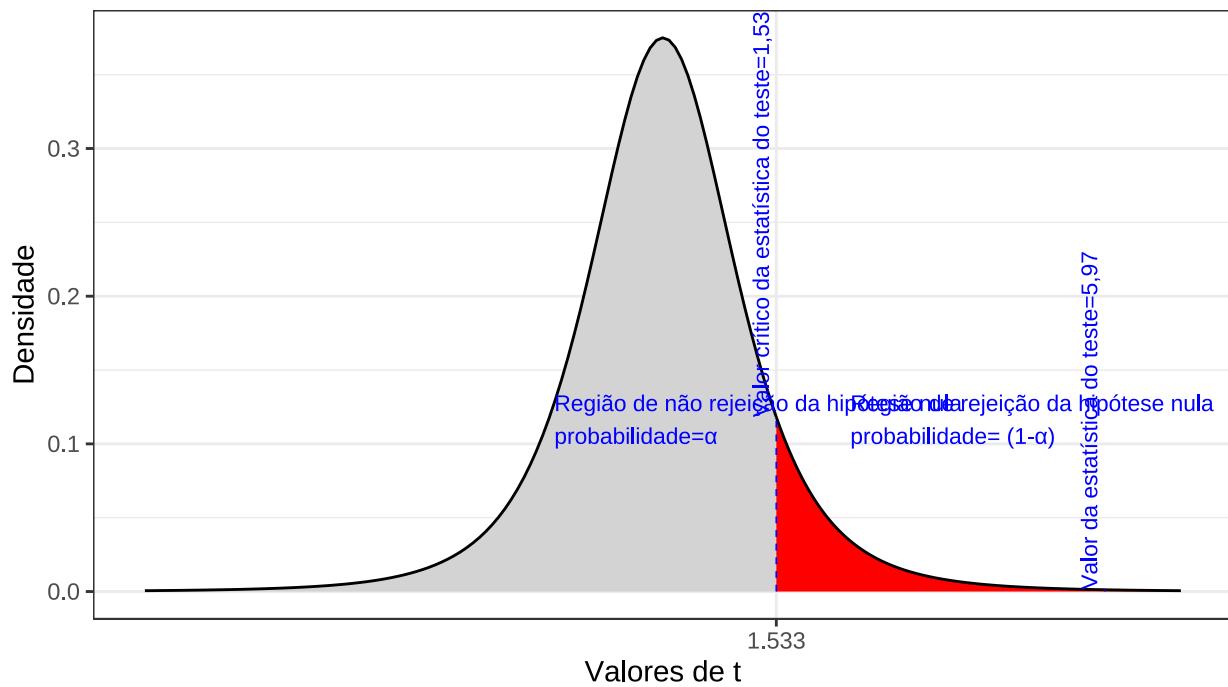


Figure 11.34: Região de rejeição da hipótese nula para o teste unilateral à direita (tipo: maior que) realizado: a região de não rejeição da hipótese nula (região de não significância do teste) está delimitada pelo valor crítico da estatística do teste: $t_{crit} = 1,53$. O valor calculado da estatística ($t_{calc} = 5,97$) situa-se na faixa de significância do teste, não possibilitando a rejeição da hipótese nula sob aquele nível de confiança

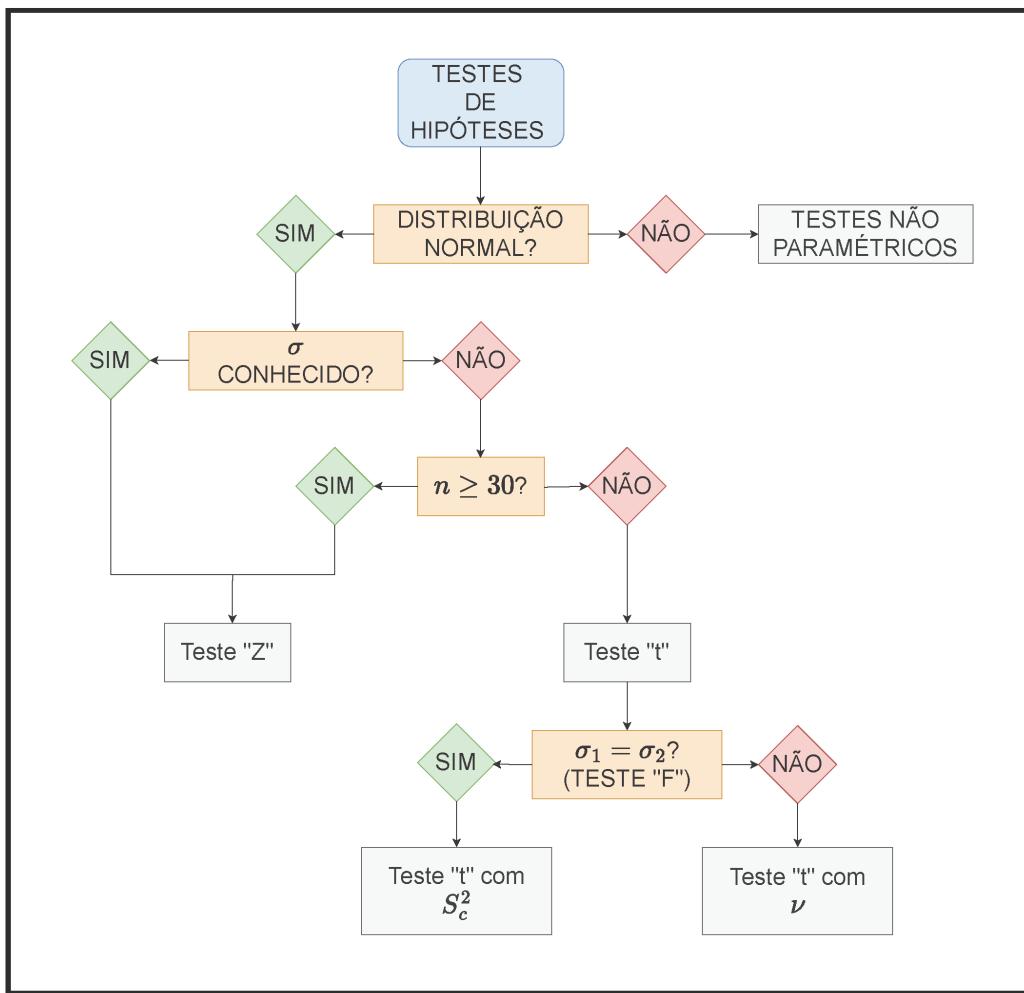
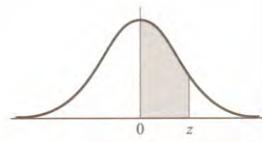


Figure 11.35: Fluxograma auxiliar para escolha da estatística do teste de hipóteses

11.13 Tabelas

Áreas sob a Curva Normal Padrão de 0 a z



z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0754
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2258	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0,7	0,2580	0,2612	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2996	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Figure 11.36: Tabela Normal padronizada

Tabela 2
Distribuição t-Student

Valores de t, segundo os graus de liberdade (ϕ) e o valor de α

Monocaudal, α	0,25	0,10	0,05	0,025	0,01	0,005
Bicaudal, α	0,50	0,20	0,10	0,05	0,02	0,01
ϕ						
1	1,000	3,078	6,314	12,706	31,821	63,657
2	0,816	1,886	2,920	4,303	6,965	9,925
3	0,765	1,638	2,353	3,182	4,541	5,841
4	0,741	1,533	2,132	2,776	3,747	4,604
5	0,727	1,476	2,015	2,571	3,365	4,032
6	0,718	1,440	1,943	2,447	3,143	3,707
7	0,711	1,415	1,895	2,365	2,998	3,499
8	0,706	1,397	1,860	2,306	2,896	3,355
9	0,703	1,383	1,833	2,262	2,821	3,250
10	0,700	1,372	1,812	2,228	2,764	3,169
11	0,697	1,363	1,796	2,201	2,718	3,106
12	0,695	1,356	1,782	2,179	2,681	3,055
13	0,694	1,350	1,771	2,160	2,650	3,012
14	0,692	1,345	1,761	2,145	2,624	2,977
15	0,691	1,341	1,753	2,131	2,602	2,947
16	0,690	1,337	1,746	2,120	2,583	2,921
17	0,689	1,333	1,740	2,110	2,567	2,898
18	0,688	1,330	1,734	2,101	2,552	2,878
19	0,688	1,328	1,729	2,093	2,539	2,861
20	0,687	1,325	1,725	2,086	2,528	2,845
21	0,686	1,323	1,721	2,080	2,518	2,831
22	0,686	1,321	1,717	2,074	2,508	2,819
23	0,685	1,319	1,714	2,069	2,500	2,807
24	0,685	1,318	1,711	2,064	2,492	2,797
25	0,684	1,316	1,708	2,060	2,485	2,787
26	0,684	1,315	1,706	2,056	2,479	2,779
27	0,684	1,314	1,703	2,052	2,473	2,771
28	0,683	1,313	1,701	2,048	2,467	2,763
29	0,683	1,311	1,699	2,045	2,462	2,756
∞	0,674	1,282	1,645	1,960	2,326	2,576

Figure 11.37: Tabela da distribuição t de Student

Tabela 4
Tabela F – 0,05

Valores de F para $\alpha = 5\%$, segundo o número de graus de liberdade do numerador (ϕ_1) e do denominador (ϕ_2)

$\phi_2 \backslash \phi_1$	1	2	3	4	5	6	7	8	9	10
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
∞	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88

Figure 11.38: Tabela da distribuição F de Fisher-Snedecor (5%)

TABELA IV

Distribuição do Qui-Quadrado - χ_n^2 Os valores tabelados correspondem aos pontos x tais que: $P(\chi_n^2 \leq x)$

n	P($\chi_n^2 \leq x$)												
	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975	0,99	0,995
1	3,93E-05	0,000157	0,000982	0,003932	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,345	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,545	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	14,562	18,358	22,718	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	24,478	29,936	34,800	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,443	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	61,698	69,334	77,577	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	71,145	79,334	88,130	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	80,625	89,334	98,650	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	90,133	99,334	109,141	118,498	124,342	129,561	135,807	140,170

Figure 11.39: Tabela da distribuição Qui-quadrado

ALFABETO GREGO

SÍMBOLOS		
NOME DA LETRA	MAIÚSCULA	MINÚSCULA
Alfa	A	α
Beta	B	β
Gama	Γ	γ
Delta	Δ	d
<i>Epsilon</i>	E	ε
Zeta	Z	ζ
Eta	H	h
Téta	Θ	θ
Iota	I	ι
Capa	K	κ
Lambda	Λ	λ
Mu(mi)	M	μ
Nu(ni)	N	ν
Csi	X	ξ
Omicron	O	\circ
Pi	P	ρ
Ró	P	ρ
Sigma	S	s
Tau	T	t
Upsilon(ipsilon)	Y	u
Fi	F	j
Chi(qui)	X	χ
Psi	Ψ	ψ
Omega	W	ω

Figure 11.40: Alfabeto grego

Módulo 12

Introdução à Correlação Linear de Pearson e Regressão Linear Simples

“Essentially, all models are wrong, but some are useful [...]” (George Edward Pelham Box, 1919 - 2013)

12.1 Contexto histórico

Sir Francis Galton (1822-1911), antropólogo e meteorologista inglês, propôs no artigo escrito em conjunto com J. D. Hamilton Dickson (*Family Likeness in Stature*) apresentado à *Royal Society of London* em 21 de janeiro de 1886, expressar por uma função uma relação que observou entre estaturas de pais e seus filhos e descendentes.

Nesse artigo, Galton verificou que, embora houvesse uma tendência de que pais mais altos tivessem filhos altos (e pais mais baixos, filhos mais baixos), a estatura média de crianças nascidas de pais com dada altura tendia a **regredir** à altura média da população como um todo. Nas palavras de Galton isso seria uma **regressão à mediocridade**: pais mais altos que a estatura média têm filhos mais baixos que eles

“Each peculiarity in a man is shared by his kinsman but, on the average, in a less degree[...]”

A *Lei da Regressão* de Galton foi referendada por Karl Pearson (*On the Laws of Inheritance*, 1903) poucos anos depois, quando analisou os dados de milhares de registros de estatura, tamanho do antebraço e da palma.

Em latim o prefixo *co* remete ao significado *colaboração*, *união* ou até *simultaneidade*. Correlação significa, portanto, uma relação mútua entre dois termos, uma correspondência.

Em *Correlations and their Measurement, chiefly from Anthropometric Data*, apresentado à *Royal Society of London* em dezembro de 1888, ele observou aquilo que viria a conceituar como *co-relação* ou *correlação de estrutura*.

Galton afirmou ao analisar o tamanho do braço com o da perna de um indivíduo que que dois órgãos são ditos serem correlacionados quando a variação de um é acompanhada, na média, pela variação para mais ou menos do outro:

- se a correlação fosse alta, uma pessoa com um braço longo teria também uma perna longa;
 - se a correlação fosse moderada, o comprimento da perna não seria tão longo e,
 - se não houvesse correlação, o comprimento de sua perna seria o comprimento médio desse membro na população.
-

“...Assim, ele naturalmente atingiu uma linha de regressão reta com variabilidade constante para todas as matrizes de um caractere para um dado caractere de um segundo. Talvez fosse melhor para o progresso do cálculo correlacional que este simples caso especial fosse exposto primeiro: é tão facilmente compreendido pelo iniciante[...]”

Houve um momento que Johann Carl Friedrich Gauss considerou sua descoberta (1795) da regressão estatística como “trivial”. O método dos mínimos quadrados parecia tão óbvio para Carl Friedrich Gauss que ele imaginou não ter sido o primeiro a usá-lo. Ele não declarou publicamente sua descoberta até alguns anos depois (*Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, 1809), quando seu contemporâneo Adrien-Marie Legendre (*Nouvelles méthodes pour la détermination des orbites des comètes*, 1805) publicou o método. Quando Gauss sugeriu que ele o havia usado antes deu-se partida a uma das mais famosas disputas de antecedência na história da ciência. Gauss acabaria recebendo a maior parte do crédito como fundador da regressão, mas não sem uma briga.

12.2 Conceitos

Em estatística, a expressão *correlação* faz referência à relação existente entre variáveis, digamos X e Y que pode assumir diferentes padrões: linear ou não linear (quadrática, cúbica, exponencial ...).

A *correlação* existente entre valores de uma mesma variável, digamos X (em diferentes momentos de tempo (X_{t_i}, X_{t_j}) ou espaço (X_{s_i}, X_{s_j}) é denominada *autocorrelação*.

12.2.1 Correlação linear *versus* regressão

a análise de correlação tem como principal objetivo medir a força ou o grau de associação linear entre as duas variáveis.

na análise de regressão linear o objetivo primário é expressar matematicamente uma relação linear entre duas variáveis de modo a possibilitar obter estimativas de uma para um valor não amostrado da outra, construir intervalos de confiança para essas estimativas e testar variadas hipóteses.

12.2.2 Correlação *versus* causação

Embora a análise de regressão lide com o comportamento de uma variável em relação a outra(s), isso não implica necessariamente em causação. É preciso levar em conta que uma relação estatística *por si só* não implica logicamente uma causação. Para atribuir uma relação de causação deve-se lançar mão de considerações *a priori* ou teóricas.

Considerem a correlação existente entre a altura dos alunos de 6 a 17 anos e as notas médias anuais obtidas em matemática. Naturalmente não é o incremento que os alunos sofrem em suas alturas na fase de crescimento que causa a melhora nas notas; mas sim processos biológicos e comportamentais que resultam em melhorias na capacidade cognitiva.

12.3 Diagrama de dispersão

Descrito pela primeira vez por Francis Galton (*Regression Towards Mediocrity in Hereditary Stature*, 1886), os diagramas de dispersão (*scatterplot*) ou gráficos de dispersão são representações de dados de duas (tipicamente) ou mais variáveis que são organizadas em um gráfico. O gráfico de dispersão utiliza coordenadas cartesianas para exibir valores de um conjunto de dados. Os dados são exibidos como uma coleção de pontos, cada um com o valor de uma variável determinando a posição no eixo horizontal e o valor da outra variável determinando a posição no eixo vertical (em caso de duas variáveis).

Considerem as simulações da dispersão de alguns valores de duas variáveis X e Y . Vemos que em alguns casos nos parece ser razoável tentar exprimir qualquer tipo de relação entre os valores de X e Y ; todavia, há situações onde claramente vemos alguma forma de relação.

Essas formas bem poderiam ser expressas, aproximadamente, por diferentes funções como:

- lineares (retas) ou
- não lineares (curvas).

Vemos também que essas formas de associação entre os valores de X e Y podem ser diretas ou inversamente proporcionais (“positiva” ou “negativa”). Estamos particularmente interessados em quantificar o grau da relação dos valores de X e Y nos padrões lineares.

(SIMULADOR 1)

12.4 Coeficiente de correlação linear de Pearson

O mais importante aspecto da correlação linear é a medida de sua intensidade, expressa pelo coeficiente de correlação linear (ou coeficiente de correlação produto momento de Pearson).

A notação adotada para o coeficiente de correlação linear de Pearson depende dos dados analisados: se são dados amostrais ou populacionais: - população: pela letra grega ρ (“rô”) - amostra: pela letra latina r

Cálculo do coeficiente de correlação amostral r :

$$r = \frac{\sum_{i=1}^n x_i \cdot y_i - \frac{\sum_{i=1}^n x_i}{n} \cdot \frac{\sum_{i=1}^n y_i}{n}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right) \cdot \left[\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \right]}}$$

em que x_i : é o iésimo valor observado da variável X , y_i : é o iésimo valor observado da variável Y , n é o número de pares de valores observados.

Ou, simplificadamente:

$$r = \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}}$$

em que $S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n}$, $S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$, $S_{yy} = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$ e n é o número de pares de valores observados.

- o coeficiente de correlação linear de Pearson tem uma faixa limitada de variação: $-1 \leq r \leq 1$,
 - é simétrico; isto é, a correlação linear observada entre X e Y é a mesma que a medida entre as variáveis Y e X ,
 - é apenas uma medida da associação linear entre duas variáveis e, portanto, não tem sentido usá-lo na quantificação de relações que não o sejam,
 - a possibilidade de uma **correlação linear negativa** virá do resultado do *numerador* (S_{xy}), pois no denominador temos duas somas de quadrados,
 - o coeficiente de correlação mede apenas a **intensidade** das relações lineares entre x e y e não estabelece *per si* nenhuma relação de causação.
-

- se $r > 0$ dizemos que há uma relação linear positiva entre as variáveis estudadas: para um incremento na primeira variável observa-se também um incremento na segunda;
 - se $r < 0$ a relação linear é negativa: um incremento em uma das variáveis é acompanhado por um decremente na outra; e,
 - se $r = 0$, então não há uma **relação linear** entre as variáveis consideradas.
-

O cálculo do coeficiente de correlação linear de Pearson assemelha-se a uma *análise de variância*

$$y - \bar{y} = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Elevando-se ao quadrado ambos os termos, para todos os valores observados, teremos:

$$\sum_{i=1}^n (y_i - \bar{y}_i)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

A quantidade à esquerda mede a variação total dos y (*Soma de quadrados total*); à direita temos a *Soma de quadrados da regressão* e a *Soma de quadrados dos resíduos* e,

$$r = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}}$$

A definição acima de r nos diz que $100 \cdot r^2$ é a *percentagem da variação total* dos y que está sendo explicada por sua regressão linear com x .

Exemplo 1: Um jornal deseja verificar a eficácia de seus anúncios na venda de carros usados e para isso realizou um levantamento de todos os seus anúncios e informações dos resultados obtidos pelas empresas que o contrataram e dele extraiu uma pequena amostra. A tabela a seguir mostra o número de anúncios e o correspondente número de veículos vendidos por 6 empresas que usaram apenas este jornal como veículo de propaganda. Existe alguma relação linear entre as variáveis? Construa o diagrama de dispersão e calcule o coeficiente de correlação linear.

Sendo $n = 6$ temos:

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} = 25172 - \frac{246 \cdot 540}{6} = 3032 S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = 12086 - \frac{246^2}{6} = 2000 S_{yy} = \sum_{i=1}^n y_i^2 -$$

Portanto:

$$r = \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}} = \frac{3032}{\sqrt{2000 \cdot 4858}} = 0,9727$$

12.5 Teste de hipóteses para a correlação linear na população

O coeficiente de correlação populacional ρ sempre é estimado a partir do coeficiente de correlação amostral r . Para se realizar inferências concernentes a ρ a partir de r temos que ter o conhecimento da distribuição amostral dos coeficientes de correlação linear r .

Para se testar a existência de correlação na população um teste de hipóteses na estrutura seguinte (bilateral) pode ser proposto:

$$\begin{cases} H_0 : \rho = 0, \text{ ie. a correlação linear entre X e Y é nula} \\ H_1 : \rho \neq 0, \text{ ie. a correlação linear entre X e Y não é nula} \end{cases}$$

Lembrando que um *teste de hipóteses* guarda uma certa semelhança a um julgamento: caso não haja indício algum que comprove a culpa do acusado ele é declarado inocente.

Seguindo essa analogia, o *índicio ou evidência* que nos permitirá rejeitar a hipótese nula virá de uma *evidência amostral*.

A quantificação da relevância da *evidência amostral* virá de uma estatística calculada (t_{calc}) a partir do coeficiente de correlação amostral r e o tamanho amostral n , que será comparado a um valor limite tabelado (t_{tab}) da correspondente distribuição da variável aleatória T :

A estatística do teste é:

$$t_{calc} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} T \sim t_{(n-2)}$$

Rejeita-se a hipótese nula (H_0) se o valor da estatística for tão extremo que se verifique:

$$t_{calc} \leq t_{tab}[\frac{\alpha}{2};(n-2)] \text{ ou } t_{calc} \geq t_{tab}[1-\frac{\alpha}{2};(n-2)]$$

em que t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao *nível de significância* pretendido (α) com $(n - 2)$ graus de liberdade. O número de graus de liberdade irá determinar qual curva da família dessa distribuição será utilizada, por essa razão, as tabelas apresentam-se individualizadas por nível de significância e graus de liberdade.

As curvas da família “t” possuem simetria em relação a um eixo vertical central. O valor tabelado dessa estatística acha-se associado à área sob ela pois é uma função densidade de probabilidade: a totalidade da área sob essa curva é igual a 1 (probabilidade de 100%).

Assim, se consultarmos em uma tabela o valor “t” para um nível de significância α qualquer, correspondente assim a um nível de confiança de $(1 - \alpha)$, qualquer veremos que ele será igual, *em módulo*, ao valor “t” no outro extremo dessa curva.

Por essa razão muitas tabelas apresentam valores dessa estatística sob os títulos de *mono-caudal* ou *bicaudal* pois estão apresentando os valores para um determinado nível de significância (α): área sob a curva, situado apenas em um lado (ou *subdividido* nos dois ramos da curva nas tabelas chamadas “bilaterais”).

O teste de hipótese que iremos realizar é um *teste bilateral*; assim, o gráfico apropriado para se decidir pela rejeição ou não da hipótese nula assume a forma mostrada nessa simulação.

(SIMULADOR 2 COM t)

12.5.1 Outros testes de hipóteses sobre a correlação linear na população

Outros tipos de testes só podem ser realizados através da estatística ζ (zeta) de Fisher. A transformação Z proposta por Fisher produz uma estatística que possui distribuição aproximadamente Normal. Para essa situação a estatística a ser utilizada é dada por:

$$\zeta = \frac{1}{2} \cdot \ln \frac{(1+r)}{(1-r)}$$

que possui uma distribuição aproximadamente Normal, com média e desvio padrão:

$$\mu_\zeta = \frac{1}{2} \cdot \ln \frac{(1+\rho_0)}{(1-(\rho_0))} \text{ e } \sigma_\zeta = \frac{1}{\sqrt{n-3}}.$$

Transformando-se ζ em unidades padrão (pela subtração de μ_ζ e divisão por σ_ζ), chega-se à estatística tabelada $z = (Z - \mu_\zeta) \cdot \sqrt{n-3}$.

Exemplo 2: Faça o teste de hipóteses para a correlação linear ρ a partir da correlação amostral r calculada no exercício dos anúncios de veículos, sob um nível de significância (α) de 0,05.

No exercício referido obtivemos um valor para a correlação linear de Pearson de $r = 0,9727$. A partir desse valor podemos calcular o valor de nossa estatística t_{calc} para o teste:

$$t_{calc} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = 8,38$$

Rejeitaremos a hipótese nula (H_0) se:

$$t_{calc} \leq t_{tab}[\frac{\alpha}{2};(n-2)] \text{ ou } t_{calc} \geq t_{tab}[1-\frac{\alpha}{2};(n-2)]$$

Da tabela extraímos o valor de nossa estatística de comparação a um nível de significância $\alpha = 5\%$ e, para um tamanho amostral $n = 6$, temos como graus de liberdade $n - 2 = 4$ ($t_{tab} = 2,776$). Vê-se que o valor calculado da estatística “t” encontra-se além dos limites estabelecidos pela estatística de comparação (t_{tab}) para um nível de significância de $\alpha = 5\%$

(SIMULADOR 2 COM t)

12.6 Regressão linear simples

12.6.1 Introdução

Considerem a proposição de John Maynard Keynes para a relação entre o consumo e a renda, onde ele postulava haver uma relação positiva entre ambos: uma mudança em uma das variáveis iria alterar a outra. Seu modelo funcional para essa relação, com Y sendo as despesas de consumo e X a renda, é:

$$Y = \alpha + \beta \cdot X$$

Esse modelo admite que a verdadeira relação entre Y e X seja uma linha reta e que a observação Y para cada nível de X seja uma variável aleatória. Assim, o valor esperado de Y para cada valor de X é:

$$Y_i = E(Y|X_i) = \alpha + \beta \cdot X_i$$

Nesse modelo, α e β são parâmetros desconhecidos da relação estabelecida entre as duas populações:

- α : intercepto (um consumo mínimo é observado mesmo nas situações em que a renda é nula, em razão de programas de assistência governamental).
- β : inclinação (a propensão média do crescimento do consumo com o incremento da renda).

É um modelo puramente teórico, de limitada aplicabilidade prática, pois pretende exprimir por uma relação exata (*determinística*) o consumo e a renda, quando se sabe que grande parte das relações entre duas variáveis não são exatas.

Entretanto, ao se fixar um único valor para a variável explicativa, observa-se que há flutuações nos valores observados da variável explicada. Essa inexatidão, esse desvio do valor observado Y_i em relação ao seu valor esperado, pode ser expresso da seguinte maneira:

$$\varepsilon_i = Y_i - E(Y|X_i)$$

em que $E(Y|X_i)$ é denominado componente sistemático ou determinístico, representando o *gasto médio* de todas as famílias com um mesmo nível de renda, e ε_i é denominado termo de erro ou distúrbio estocástico. O termo de erro pode ser admitido como um *substituto* para todas as demais variáveis omitidas ou negligenciadas no modelo e que podem afetar Y .

Um modelo de regressão pode ser *linear* nas variáveis ou nos parâmetros.

Uma função $Y = f(X)$ é dita linear em X se X tiver um expoente igual a 1 e não estiver multiplicado ou dividido por outra variável.

- a função $Y = \alpha + \beta \cdot X$ é dita linear em β se β tiver um expoente de 1 e não estiver multiplicado ou dividido por qualquer outro parâmetro.

A função $E(Y|X) = \alpha + \beta \cdot X^2$ não é linear em X , pois X está elevado ao quadrado.

- mas é linear nos parâmetros, pois, para $X = 3$, temos $E(Y|X = 3) = \alpha + 9 \cdot \beta$.

Das duas interpretações de linearidade, a *linearidade nos parâmetros* é a relevante para a formulação da teoria da regressão (a linearidade nas variáveis pode ou não ocorrer).

No contexto deste curso, o modelo será linear tanto nos parâmetros quanto na variável.

Admitindo-se que $E(Y|X_i)$ seja linear em X_i , podemos reescrever o modelo original na forma que incorpora o erro aleatório:

$$Y_i = E(Y|X_i) = \alpha + \beta \cdot X_i Y_i = E(Y|X_i) + \varepsilon_i Y_i = \alpha + \beta \cdot X_i + \varepsilon_i$$

em que α é o intercepto da reta, representando o valor esperado da variável Y quando $X = 0$, β é a inclinação da reta, representando a variação esperada de Y para um aumento unitário em X_i , $(\alpha + \beta \cdot X_i)$ é a parte explicada pelo modelo e ε_i é o termo de erro ou distúrbio estocástico.

Nessa função:

- Y : variável dependente (também chamada de explicada, prevista, regressando, resposta, endógena, saída, controlada) — aqui, representando o *consumo*.
- X : variável independente (também chamada de explicativa, previsora, regressor, estímulo, exógena, entrada, controle) — aqui, representando a *renda*.

Se o termo de erro ε_i representa todas aquelas variáveis omitidas no modelo (mas que, coletivamente, afetam Y), por que não formular um modelo de regressão com o máximo de variáveis possíveis?

- **Embasamento teórico vago:** A teoria existente suporta com certeza apenas algumas variáveis; o termo de erro ε_i serve como um substituto para todas as variáveis excluídas no modelo.
 - **Princípio da parcimônia:** Um modelo mais simples que explique bem a relação é preferível.
 - **Forma funcional equivocada:** Em gráficos de dispersão, é mais fácil inferir a relação entre duas variáveis do que com muitas.
 - **Limitação na quantidade de observações:** Muitas variáveis exigem mais observações para garantir a precisão do modelo.
-

Sendo inviável, e muitas vezes impossível, construir um *modelo populacional*, focamos o estudo em uma parte dessa população: uma *amostra*.

Um modelo funcional estimado com base em uma *amostra* apresenta **estimativas** dos parâmetros da função que descreve a população de origem (os quais são desconhecidos). Por isso, adota-se uma notação diferente para a **função de regressão amostral** em sua forma *estocástica*:

$$\hat{Y} = a + b \cdot X$$

em que \hat{Y} é um estimador de $E(Y|X)$, a é uma estimativa do parâmetro α e b é uma estimativa do parâmetro β .

Para um determinado valor de $X = x_i$, temos uma observação amostral $Y = y_i$ que pode ser expressa pela **função de regressão amostral** como:

$$y_i = \hat{y}_i + e_i \quad y_i = a + b \cdot x_i + e_i$$

em que \hat{y}_i é o valor estimado de Y_i para um determinado X_i , e_i é o erro amostral, que representa a diferença entre o valor observado y_i e o valor estimado \hat{y}_i .

Mas, como estimar a e b ?

12.6.2 Método dos mínimos quadrados

Na literatura estatística há vários métodos de estimação dos parâmetros de um modelo de regressão linear, dentre os quais:

- Método dos momentos (creditado a Karl Pearson-1895, Ronald Aylmer Fisher-1925, Neyman e Egon Pearson-1928, publicado por Lars Peter Hansen-1982);
 - Método da máxima verossimilhança (creditado a Johann Carl Friedrich Gauss, Pierre-Simon Laplace, Thorvald N. Thiele e Francis Ysidro Edgeworth, popularizado por Ronald Aylmer Fisher, 1912-1922); e,
 - Método dos mínimos quadrados (creditado a Johann Carl Friedrich Gauss-1795, publicado por Adrien-Marie Legendre-1805, Friedrich Robert Helmert-1872).
-

12.6.2.1 Contexto histórico

Desde tempos remotos as pessoas têm se interessado pelo problema de escolher o melhor valor único (médio) para resumir as informações fornecidas por várias observações, cada uma sujeita a erro.

O problema de se estimar as constantes na equação da linha reta que melhor se ajusta a três ou mais pontos não colineares no plano (x, y) cujas coordenadas são pares de valores associados de duas variáveis relacionadas: X e Y remonta a Galileu Galilei (1632).

Credita-se Johann Carl Friedrich Gauss como o desenvolvedor das bases fundamentais do Método dos mínimos quadrados, em 1795, quando Gauss tinha apenas dezoito anos.

Mas o Método dos mínimos quadrados foi publicado pela primeira vez por Adrien-Marie Legendre (1752-1833) em 1805: *Nouvelles méthodes pour la détermination des orbites des comètes*.

Alguns demonstradores:

- Robert Adrain (1775-1843) em 1808: *Research concerning the probabilities of the errors which happen in making observations*
- Johann Carl Friedrich Gauss (1777-1855) em 1809: *Theoria motus corporum coelestium*
- Pierre-Simon Laplace (1749-1827) em 1810: *Theorie analytique des Probabilité* - Johann Carl Friedrich Gauss (1777-1855) em 1823: *Theoria combinationis observationum erroribus obnoxiae*
- James Ivory (1765-1842) em 1825: *On the Method of the Least Squares*.

Para o modelo $y_i = a + b \cdot x_i$ na simulação mostrada:

- **problema:** determinar as constantes a e b da equação de uma linha reta que melhor se ajusta a três ou mais pontos não colineares
- **solução:** minimizar a soma dos quadrados dos resíduos como mostrado na simulação.

$$\sum_{i=1}^n e_i^2 \rightarrow 0$$

A grande vantagem do método dos mínimos quadrados é que ele é um método puramente geométrico, e não faz nenhuma suposição sobre a distribuição dos dados ou dos erros (resíduos).

Em outras palavras, ele é aplicado sem se preocupar com a natureza probabilística dos erros (resíduos). O objetivo é apenas ajustar a melhor reta possível para um conjunto de pontos de dados

(SIMULADOR 3)

Matematicamente, a partir da igualdade:

$$\sum_{i=1}^n [y_i - \hat{y}]^2 = \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

a solução passar por derivar-se em relação a: $a|b$ fixo, e em relação b: $b|a$ fixo, igualando-se a zero:

$$\frac{\delta}{\delta a} \sum_{i=1}^n [y_i - (ax_i + b)]^2 = 2 \cdot \sum_{i=1}^n (y_i - ax_i - b) (-x_i) = 0 \frac{\delta}{\delta b} \sum_{i=1}^n [y_i - (ax_i + b)]^2 = 2 \cdot \sum_{i=1}^n (y_i - ax_i - b) (-1) = 0$$

Após algumas manipulações algébricas obtemos as seguintes expressões para as estimativas: a e b :

$$\begin{aligned} b \cdot n + a \cdot \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ b \cdot \sum_{i=1}^n x_i + a \cdot \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i \cdot y_i \end{aligned}$$

chegando-se ao **estimador** para a :

$$a = \frac{n \cdot (\sum_{i=1}^n x_i y_i) - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

e ao **estimador** para b :

$$b = \frac{(\sum_{i=1}^n x_i^2) \cdot (\sum_{i=1}^n y_i^2) - (\sum_{i=1}^n x_i y_i) \cdot (\sum_{i=1}^n x_i)}{n \cdot (\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

Se definirmos S_{xy} e S_{xx} como sendo:

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n}$$

e

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$$

então podemos escrever:

$$b = \frac{S_{xy}}{S_{xx}} \text{ e } a = \bar{y} - b \cdot \bar{x}$$

Uma vez que

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \text{ e } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

o estimador a pode ser reescrito na forma:

$$a = \frac{\sum_{i=1}^n y_i - b \cdot \sum_{i=1}^n x_i}{n}$$

Exemplo 3: Um jornal deseja verificar a eficácia de seus anúncios na venda de carros usados e para isso realizou um levantamento de todos os seus anúncios e informações dos resultados obtidos pelas empresas que o contrataram e dele extraiu uma pequena amostra. A tabela abaixo mostra o número de anúncios e o correspondente número de veículos vendidos por 6 empresas que usaram apenas este jornal como veículo de propaganda. Obtenha a equação de regressão linear simples e estime o número de carros vendidos para um volume de 70 anúncios?

Sendo $n = 6$, $\bar{y} = 90$ e $\bar{x} = 41$:

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} = 25172 - \frac{246 \cdot 540}{6} = 3032 S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = 12086 - \frac{246^2}{6} = 2000 S_{yy} = \sum_{i=1}^n y_i^2 -$$

As estimativas dos parâmetros do modelo serão:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{3032}{2000} = 1,5160$$

e

$$a = \bar{y} - b \cdot \bar{x} = 90 - 1,5160 \cdot 41 = 27,844$$

e o modelo toma a seguinte forma $\hat{y} = 27,844 + 1,5160 \cdot x$. Para um volume de anúncios de 70 veiculações teremos, em média, 134 carros vendidos.

12.7 Modelo de regressão linear sob erros Normais

Embora o método dos mínimos quadrados forneça estimativas para a e b , ele **não nos diz nada sobre a incerteza dessas estimativas**.

Não podemos fazer inferências estatísticas tais como construir intervalos de confiança ou realizar testes de hipóteses, a menos que façamos suposições adicionais sobre os erros do modelo.

Para realizar inferências estatísticas, introduzimos um **modelo de regressão linear com erro normal**, que assume:

- os erros (ε_i) são variáveis aleatórias Normalmente distribuídas com média zero e variância constante (σ^2): $\varepsilon_i \sim N(0, \sigma^2)$
 - os erros são independentes entre si
 - a relação entre Y_i e X_i é linear, descrita pela equação $Y_i = \alpha + \beta X_i + \varepsilon_i$
-

12.7.1 Propriedades dos Estimadores sob Erro Normal

Demonstra-se que, para um modelo $Y_i = \alpha + \beta \cdot X_i + \varepsilon_i$ que:

- b é um estimador não tendencioso do parâmetro β com:

$$E(b) = \beta \text{ e } Var(b) = \frac{\sigma^2}{S_{xx}}$$

- a é um estimador não tendencioso do parâmetro α com:

$$E(a) = \alpha \text{ e } Var(a) = \sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} \right)$$

- $\hat{\sigma}^2$ é um estimador não tendencioso de σ^2 :

$$\hat{\sigma}^2 = \text{QMR} = \frac{S_{yy} - b \cdot S_{xy}}{n - 2}$$

Assim as variâncias dos estimadores a e b serão,

$$s_b = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{\text{QMRES}}{S_{xx}}}$$

$$s_a = \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} = \sqrt{\text{QMRES} \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

lembrando que:

$$S_{yy} = \sum (Y_i - \bar{Y})^2 S_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y}),$$

e $n - 2$ representa os graus de liberdade, já que dois parâmetros (α e β) são estimados.

12.7.2 Implicações da Normalidade

A normalidade dos resíduos ε_i garante que os estimadores a e b também sejam Normalmente distribuídos, o que é fundamental para realizar testes de hipóteses e construir intervalos de confiança nos modelos de regressão linear.

Isso permite o uso de distribuições de referência, como as distribuições t e F , especialmente em amostras pequenas, onde a variância dos estimadores não pode ser assumida como conhecida com precisão.

Na estimação de um modelo de regressão linear simples com erro Normal (na forma $Y = \beta_0 + \beta_1 X + \varepsilon$) muitas premissas preliminarmente como válidas deverão ser efetivamente verificadas a posteriori, na chamada etapa de diagnóstico do modelo, de modo a que a condução de inferências com esse modelo sejam dotada de razoável segurança.

Essas premissas podem ser classificadas em quatro categorias:

- linearidade da relação entre a variável preditora X e a variável resposta Y : o valor esperado da variável resposta é uma função linear da variável preditora
- Normalidade: $\varepsilon_i \sim N(0, \sigma_i^2)$
- independência estatística dos resíduos: $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0, i \neq j$ e, em particular, nenhuma correlação entre erros de observações sucessivas no caso de dados provenientes de uma série; e,
- homogeneidade da variância dos resíduos (homocedasticidade): $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_i^2$ quando analisada frente aos valores estimados pelo modelo (\hat{Y}), a variável preditora (X) ou o tempo de coleta nos casos de dados provenientes de uma série

Se qualquer uma dessas premissas for violada então uma conclusão científica baseada em resultados advindos desse modelo de regressão poderá estar seriamente comprometida. As violações desses pressupostos não podem ser detectadas pelas estatísticas de resumo do modelo que usualmente se dipõe logo após sua estimativa: estatísticas t , F dos testes de significância ou então o coeficiente de determinação R^2 .

Assim, é sobretudo fundamental examinar mais aprofundadamente o modelo de modo a se assegurar com razoável confiança de sua adequação aos dados antes de se avançar com seu uso. A esse exame denominamos diagnóstico do modelo.

12.7.3 Linearidade na relação entre a variável preditora X e a variável resposta Y :

A violação da linearidade é extremamente graves pois um modelo ajustado a dados não lineares leva a previsões equivocadas não somente para valores situados além das fronteiras amostrais (como se usualmente observa) mas também para valores próximos ao seu centro.

Uma técnica gráfica para se verificar a linearidade da relação é através de dois gráficos:

- valores observados em relação aos valores estimados; ou/e,
- resíduos contra valores estimados (ou valores observados).

Os padrões desejados nos gráficos acima deve assemelhar-se a:

- pontos dispersos de modo aproximadamente simétrico em torno de uma linha diagonal; e,
 - pontos dispersos de modo aproximadamente simétrico em torno de uma linha horizontal, com uma variância aproximadamente homogênea.
-

Relações não lineares devem ser tratadas por meio da aplicação de uma transformação não linear adequada ao padrão da relação na variável resposta ou no variável preditora.

Para dados estritamente positivos com uma relação não linear a transformação com a função logaritmo pode ser uma opção. Se uma a transformação com o uso da função logaritmo é aplicada apenas à variável resposta isso equivalente a assumir que ela cresce (ou decai) exponencialmente como uma função da variável preditora.

Outra possibilidade a considerar é adicionar outra variável preditora na forma de uma função não linear como, por exemplo, nos padrões de dispersão que mostrem uma curva parabólica onde pode fazer sentido regredir Y em função de X e X^2 .

Finalmente, a relação não linear observada pode decorrer da omissão de outra(s) variáveis importantes que explicam ou corrigem o padrão não linear quando então modelos de regressão linear múltipla devem ser estudados.

12.7.4 Homogeneidade da variância de ε (homocedasticidade):

A violação da homogeneidade de variância dos resíduos (heterocedasticidade) resulta numa estimativa imprecisa do verdadeiro desvio padrão dos erros das estimativas e acarreta em intervalos de confiança irreais: são mais amplos ou mais estreitos do que deveriam ser, e resultam em elevada imprecisão nas inferências feitas com estatísticas baseadas na variância (t , F).

Com variância constante (homocedasticidade) temos que $Var(\varepsilon|X_i) = \sigma^2$; todavia o que se observa em muitas situações é que a variância está relacionada de algum modo funcional com a média ($\sigma^2 = f(X)$) e, assim:

$$\begin{aligned} Var(\varepsilon_i|X_i) &= \sigma_i^2 \\ E(\varepsilon_i^2) &= \sigma_i^2 \end{aligned}$$

Na presença de heterocedasticidade nos resíduos, os estimadores de mínimos quadrados continuam sendo não viesados e consistentes, mas perdem eficiência. Equivale a dizer que haverá um outro estimador para os parâmetros do modelo que terá uma variância menor e menos tendencioso:

$$Var(b^*) < Var(b)$$

Uma técnica gráfica para se verificar a homocedasticidade dos resíduos é através dos gráficos:

- resíduos contra valores estimados; ou,
- resíduos contra a variável preditora

Os padrões desejados nos gráficos acima deve assemelhar-se a pontos dispersos de modo aproximadamente simétrico em torno de um eixo horizontal e que não exibam, sistematicamente, nenhum padrão de crescimento ou decaimento na amplitude visual de sua dispersão como nas imagens abaixo:

A heterocedasticidade pode ser um subproduto de uma violação significativa das premissas de linearidade e/ou independência, caso em que todas essas violações podem ser conjuntamente corrigidas com a aplicação de uma transformação de potência na variável dependente que terá como objetivos:

- linearizar o ajuste tanto quanto possível; e/ou,
- estabilizar a variância dos resíduos.

Algum cuidado e discernimento é requerido pois esses dois objetivos podem conflitar entre si. Geralmente opta-se em estabilizar a variância dos resíduos primeiramente para, só então analisar linearização das relações.

As transformações sugeridas pela família Box-Cox (1964) em função dos valor que maximizam a verissimilhança perfilada são:

- se $\lambda=-2 \rightarrow \frac{1}{Y^2}$
 - se $\lambda=-1 \rightarrow \frac{1}{Y}$
 - se $\lambda=-0,5 \rightarrow \frac{1}{\sqrt{Y}}$
 - se $\lambda=0 \rightarrow \log(Y)$
 - se $\lambda=0,50 \rightarrow \sqrt{Y}$
 - se $\lambda=1 \rightarrow Y$
 - se $\lambda=2 \rightarrow Y^2$
-

Gráficos dos valores absolutos dos resíduos (ou do quadrado dos resíduos pois os sinais dos resíduos não são significativos para o propósito desse exame) contra a variável preditora X ou em relação aos valores ajustados também são úteis para o diagnóstico da heterocedasticidade da variância dos resíduos.

Esses gráficos são recomendados quando não há muitas observações no conjunto de dados pois a plotagem dos resíduos absolutos ou seus quadrados coloca as informações sobre a alteração das suas magnitudes acima da linha horizontal do zero o que facilita a inspeção visual de possíveis alterações de sua magnitude em relação a outra variável adotada no gráfico.

12.7.4.1 Testes para verificação da a homogeneidade da variância:

- teste de Park;
 - teste de Bartlett;
 - teste de Levene;
 - teste de Brown-Forsythe;
 - teste de Breuch-Pagan;
-

12.7.5 Inconsistência de observações (outliers)

Outliers são observações extremas afastadas das demais observações que formam a amostra e sua identificação deve ser feita já na análise descritiva que antecede todo estudo estatístico.

Essas observações podem ser resultado dos mais variados erros de medição (observadores diferentes, equipamentos descalibrados, instrumentos de medição diversos) quando então, nessa hipótese e confirmado o erro de registro, devem ser descartados com discernimento.

Todavia na maior parte dos experimentos a identificação desse tipo de erro na etapa descritiva não é possível e, nessas situações, a análise dos resíduos gerados pelo modelo na estimativa de cada observação é a principal ferramenta.

A principal razão para sua identificação é que esses pontos extremos podem ter grande repercussão e exercer grande influência nas estimativas do modelo. Uma observação é influente se uma pequena modificação em seu valor ou sua exclusão do modelo produz alterações significativas nas estimativas dos parâmetros.

Uma técnica gráfica para se verificar a presença observações outliers é através dos gráficos:

- resíduos contra valores estimados; e/ou,
- resíduos contra a variável preditora

A plotagem de resíduos estudentizados é particularmente útil para distinguir as observações cujos resíduos distem muitos desvios padrão da média zero.

Os padrões desejados nos gráficos acima deve assemelhar-se a pontos dispersos de modo aproximadamente simétrico em torno do eixo horizontal zero, que não exibam, sistematicamente, nenhum padrão de crescimento ou decaimento na amplitude visual de sua dispersão. Uma regra comum para amostras grandes ($n > 30$) é considerar resíduos estudentizados com desvios padrão em valor absoluto de quatro ou mais desvios padrão serem outliers.

12.7.6 Pontos influentes com capacidade de alavanca (leverage):

Os elementos h_{ii} da diagonal da matriz de projeção (H) tem importante papel no diagnóstico de pontos influentes. Há diferentes opiniões sobre os valores críticos para essa medida:

- $h_{ii} > 2\frac{p}{n}$ (Hoaglin, D. C. and Welsch, R. E, 1978. The hat matrix in regression and ANOVA)
- $h_{ii} > 3\frac{p}{n}$ onde p é o número de parâmetros estimados no modelo ($\hat{\beta}_0$ e $\hat{\beta}_1$: 2 para uma regressão linear simples).

David Sam Jayakumar e A. Sulthan (Exact distribution of Hat Values and Identification of Leverage Points, 2014) propuseram a distribuição teórica exata para os valores da diagonal da matriz de projeção link de acesso ao recurso.

12.7.6.1 DFBeta:

A estatística $DFBeta$ indica o quanto cada coeficiente de regressão $\hat{\beta}_j$ se altera em unidades de desvio padrão quando a i -ésima observação for removida:

$$DFBeta_{(j,i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_i^2 C_{(jj)}}}$$

onde $C_{(jj)}$ é o j -ésimo elemento da diagonal da matriz $(X^t X)^{-1}$ e:

$$S_i^2 = \frac{(n-p-1)QMRes - \hat{\varepsilon}_i(1-h_{ii})}{(n-p)}$$

Valores superiores a $|DFBeta_{(ji)}| > \frac{2}{\sqrt{n}}$ requerem exame mais detalhado.

12.7.6.2 DFFits:

A estatística $DFFits$ indica a influência da i -ésima observação medindo o quanto os valores preditos se modificam, em unidades de desvio padrão, se aquela observação for removida:

$$DFFits = \frac{\hat{Y} - \hat{Y}_i}{\sqrt{S_i^2 h_{ii}}}$$

Valores superiores a $|DFFits| > 2\sqrt{\frac{p}{n}}$ requerem exame mais detalhado.

12.7.6.3 Distância de Cook:

A estatística proposta por Denis R. Cook mede a influência de um determinado dado da amostra no que tange a quanto ele está afetando a linha de regressão, sendo medida pelo quanto a linha de regressão se alteraria caso esse dado fosse removido da análise: ele exerce um destacado impacto da estimativa dos parâmetros do modelo. A influência na locação (afastamento de alguma observação da vizinhança do resto dos dados) pode ser investigada pelo gráfico feito das distâncias de Cook contra os valores ajustados.

Há vários critérios para se definir um valor limite para a estatística de Cook:

- $D_i > 1$: Cook e Weisberg, 1982 e Chatterjee, Hadi e Price, 2000;
 - duas vezes a média das distâncias de Cook;
 - $\frac{4}{n} < D_i < 1$: Bollen et al, 1990; e,
 - o valor crítico do quantil da distribuição F para uma significância igual a 0.5 com $df1=p$ e $df2=n-p$.
-

12.7.7 Independência

Quando as observações da amostra são independentes o que se espera é que seus resíduos apresentem-se aleatoriamente dispersos em torno da linha horizontal (zero) quando dispostos na sequência em que foram coletadas. O que se pretende aqui é verificar se há correlação serial entre as observações.

A autocorrelação pode ser definida como a correlação entre integrantes de séries de observações ordenadas no tempo (como as séries temporais) ou no espaço (como nos dados de corte transversal) quando então os resíduos de duas observações guardam correlação diferente de zero entre si:

$$\text{cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j | x_i, x_j) \neq 0 \\ i \neq j$$

A correlação serial pode decorrer:

- inércia: quando os efeitos na alteração da variável X demoram a se manifestar na variável Y (muito comum em dados econômicos);
- forma funcional do modelo incorreta;
- variáveis importantes foram omitidas.

A verificação da independência resíduos $\hat{\varepsilon}$ pode ser verificada informalmente através de vários modos gráficos dentre os quais destacam-se:

- resíduos contra o tempo ou ordem no qual as observações foram realizadas; e,
- observações contra o tempo ou ordem no qual foram realizadas (um gráfico sequencial).

O que se espera é que nenhuma relação funcional seja percebida. Há ferramentas estatísticas apropriadas para se analisar dados provenientes de séries.

12.7.8 Normalidade

A Normalidade dos resíduos $\hat{\varepsilon}$ pode ser verificada informalmente através de vários modos gráficos dentre os quais descam-se:

- pela comparação de suas frequências às frequências esperadas de uma distribuições Normal: 68%: ± 1 desvio padrão; 90%: ± 1.65 desvio padrão; 95%: ± 1.96 desvio padrão;
- gráficos de caixas;
- histogramas;
- gráficos dos quantis teóricos da distribuição Normal padronizada contra os quantis amostrais dos resíduos (QQ plot);
- gráfico com envoltória simulada dos resíduos (Brian David Ripley em Modelling Spatial Patterns, 1977).

Se os valores de uma amostra provêm de uma distribuição Normal, então os valores das estatísticas de ordem contruídas com os resíduos e os Z_i correspondentes obtidos da distribuição Normal padrão são linearmente relacionados e, assim, o gráfico dos valores deve ter o aspecto aproximado de uma reta.

Todavia observam-se que alguns aspectos desse gráfico diferentes de uma reta que sugerem ausência de Normalidade têm como provável causa:

- “S”: indica distribuições com caudas muito curtas, isto é, distribuições cujos valores estão muito próximos da média;
- “S invertido”: indica distribuições com caudas muito longas e, portanto, presença de muitos valores extremos; e,
- “J” e “J invertido”: indicam distribuições assimétricas, positivas e negativas, respectivamente.

A análise do modelo com respeito à Normalidade de seus resíduos é, em muitos aspectos, mais difícil do que para as outras verificações.

A menos que o tamanho da amostra seja muito grande ($n \sim 300$) a variação aleatória impõe sérias dificuldades para se estudar a natureza da distribuição de probabilidade da variável em estudo. Outros tipos de desvios podem também afetar a distribuição dos resíduos como quando a função é inadequada ou quando a variância não é constante. Assim, pequenos desvios dos resíduos em relação à distribuição Normal podem ser tolerados pois não causam problemas sérios na estimação do modelo.

12.7.8.1 Testes para Normalidade dos resíduos:

Para uma análise formal da Normalidade há vários testes definidos:

- K^2 de D'agostino (Ralph D'agostino);
 - Jarque-Bera (Carlos Jarque e Anil K. Bera);
 - Anderson-Darling (Theodore Wilbur Anderson e Donald Alan Darling);
 - Cramer-von Mises (H. Cramer e R.E. von Mises);
 - Lilliefors (Hubert W. Lilliefors);
 - Shapiro-Francia (Samuel Sandford Shapiro e S. Francia);
 - X^2 de Karl Pearson;
 - Shapiro-Wilk (Samuel Sandford Shapiro e Martin Bradbury Wilk);
 - Kolmogorov-Smirnov (Andrey Kolmogorov e Nikolai Smirnov); e,
 - teste de correlação linear entre os resíduos padronizados ordenados e os quantis teóricos da distribuição Normal padronizada;
-

12.7.9 Variáveis omitidas do modelo

Caso os dados sob análise possuam mais variáveis preditoras é prudente plotar um gráfico dos resíduos contra cada uma delas para que eventuais efeitos na variável resposta sejam descartados.

O objetivo desta análise adicional é determinar se há quaisquer outras variáveis que possam contribuir na explicação da variável resposta e assim, o padrão visual dos resíduos não pode diferir do padrão apresentado quando se plotam os resíduos contra a variável incorporada no modelo, não só na aleatoriedade de sua dispersão mas também nas frequências ou concentrações mostradas acima ou abaixo da linha base (zero).

12.8 Teste de significância (global) do modelo

O modelo $\hat{Y} = a + b \cdot X$ pode ser decomposto em duas partes:

- variação explicada: $a + b \cdot X$
- variação residual: $\hat{Y} - Y$, a diferença entre um valor estimado e o realmente observado.

Se a variação explicada for significativamente superior à variação residual, teremos um bom indicativo de existe regressão linear entre as variáveis X e Y e o modelo a está explicando razoavelmente bem.

Essa verificação é realizada pela **análise de variâncias**.

Sendo SQTOTAL = SQREG - SQRES, em que:

$$SQRES = S_{yy} - b \cdot S_{xy}S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} S_{yy} = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

A verificação da existência ou não de regressão linear na população é necessário testar o parâmetro β e, para tanto, propomos as seguintes hipóteses:

$$\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq 0 \end{cases}$$

Usualmente $\beta_0 = 0$, indicando não haver regressão na população.

A estatística calculada (F_{calc}) será comparada a uma estatística F_{tab} tabelada da Distribuição “F” (Ronald Aylmer Fisher-George Waddel Snedecor).

F_{tab} é o quantil de ordem α da Distribuição “F” (Ronald Aylmer Fisher-George Waddel Snedecor) com graus de liberdade 1, $(n - 2)$ (numerador e denominador, respectivamente).

Rejeita-se a hipótese nula (H_0) se:

$$F_{calc} = \frac{QMREG}{QMRES} \geq F_{tab[1,(n-2);\alpha]}$$

em um teste unilateral à direita: $(\alpha) \in \text{right tail}$.

Vejam nessa simulação o gráfico da função densidade de probabilidade “F” (Ronald Aylmer Fisher-George Waddel Snedecor) com graus de liberdade no numerador e denominador: 1, $(n - 2)$ e nível de significância (α) ∈ right tail.

SIMULADOR 4

Exemplo 4 Uma indústria farmacêutica vende um remédio para aliviar os sintomas do resfriado. Após dois anos de operação ela coletou as informações trimestrais de vendas desse produto e despesas com sua propaganda. Estime um modelo de regressão linear simples e teste a existência da regressão pela ANOVA a um nível de significância de 5%

Sendo $n = 8$, $\bar{y} = 16$ e $\bar{x} = 7,50$, calculamos:

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} = 1101 - \frac{60 \cdot 128}{8} = 141 S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = 522 - \frac{60^2}{8} = 72 S_{yy} = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

As estimativas dos parâmetros do modelo serão:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{141}{72} = 1,9583 a = \bar{y} - b \cdot \bar{x} = 16 - 1,9583 \cdot 7,50 = 1,3125$$

O modelo toma a seguinte forma:

$$\hat{y} = 1,3125 + 1,9583 \cdot x$$

Conclusão: frente ao resultado da análise dos dados rejeita-se a hipótese sob um nível de significância de 5%.
(SIMULADOR 4)

12.9 Teste de hipóteses para o coef. angular β

O teste de hipóteses para o coeficiente angular β pode ser proposto da forma que se segue:

$$\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0 \end{cases}$$

Usualmente fazemos $\beta_0 = 0$, indicando não haver regressão.

Estatística do teste:

$$t_{calc} = \frac{b - \beta_0}{s_b}$$

Rejeita-se a hipótese nula (H_0) se:

$$t_{calc} \leq t_{tab[\frac{\alpha}{2};(n-2)]} \text{ ou } t_{calc} \geq t_{tab[1-\frac{\alpha}{2};(n-2)]}$$

em um teste bilateral: $(\frac{\alpha}{2}) \in \text{left tail}; (\frac{\alpha}{2}) \in \text{right tail}$.

Sendo t_{tab} o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido (α) com $(n - 2)$ graus de liberdade. O número de graus de liberdade irá determinar qual curva da família dessa distribuição será utilizada, por essa razão, as tabelas apresentam-se na forma de linhas (graus de liberdade) e colunas (nível de significância).

Vejam nessa simulação o gráfico da função densidade de probabilidade “t” de Student (William Sealy Gosset, 1876-1937) com graus de liberdade: $(n - 2)$ e nível de significância: $(\frac{\alpha}{2}) \in \text{left tail}; (\frac{\alpha}{2}) \in \text{right tail}$.

(SIMULADOR 2 COM t)

12.10 Teste de hipóteses para o coef. angular α

O teste de hipóteses para o coeficiente linear α pode ser proposto da forma que se segue:

$$\begin{cases} H_0 : \alpha = \alpha_0 \\ H_1 : \alpha \neq \alpha_0 \end{cases}$$

Usualmente $\alpha_0 = 0$ indicando que a regressão passa pela origem.

Estatística do teste:

$$t_{calc} = \frac{\alpha - \alpha_0}{s_\alpha}$$

Rejeita-se a hipótese nula (H_0) se:

$$t_{calc} \leq t_{tab[\frac{\alpha}{2};(n-2)]} \text{ ou } t_{calc} \geq t_{tab[1-\frac{\alpha}{2};(n-2)]}$$

em um teste bilateral: $(\frac{\alpha}{2}) \in \text{left tail}; (\frac{\alpha}{2}) \in \text{right tail}$.

Sendo t_{tab} o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido (α) com $(n - 2)$ graus de liberdade. O número de graus de liberdade irá determinar qual curva da família dessa distribuição será utilizada, por essa razão, as tabelas apresentam-se na forma de linhas (graus de liberdade) e colunas (nível de significância).

Vejam nessa simulação o gráfico da função densidade de probabilidade “t” de Student (William Sealy Gosset, 1876-1937) com graus de liberdade: $(n - 2)$ e nível de significância: $(\frac{\alpha}{2}) \in \text{left tail}; (\frac{\alpha}{2}) \in \text{right tail}$.

(SIMULADOR 2 COM t)

12.11 Coeficiente de determinação R^2

O coeficiente de determinação amostral (R^2) é uma medida estatística que informa o quanto da variação observada na variável Y está sendo explicada no modelo pela relação linear estabelecida com a variável X .

$$R^2 = \frac{\text{variação explicada}}{\text{variação total}} R^2 = \frac{b \cdot S_{xy}}{S_{yy}}$$

Exemplo 5: O faturamento de uma loja durante o período de janeiro a gosto de 2010 é dado pela tabela abaixo (milhares de R\$). Construa um modelo, calcule a correlação existente, teste a existência da regressão pela ANOVA, a correlação linear obtida, as estimativas de seus coeficientes a e b de seus coeficientes α e β , a um nível de significância de 5%

Sendo $n = 8$, $\bar{Y} = 27$ e $\bar{x} = 4,5$, calculamos:

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} = 1063 - \frac{36 \cdot 216}{8} = 91 S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = 204 - \frac{36^2}{8} = 42 S_{yy} = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

As estimativas dos parâmetros do modelo serão:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{91}{42} = 2,166 a = \bar{y} - b \cdot \bar{x} = 27 - 2,166 \cdot 4,50 = 17,253$$

E o modelo toma a seguinte forma:

$$\hat{y} = 17,253 + 2,166 \cdot x$$

O coeficiente de correlação linear de Pearson é:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{91}{\sqrt{42 \cdot 202}} = 0,9880$$

Conclusão: frente ao resultado da análise dos dados rejeitamos a hipótese nula sob um nível de significância de 5%.

(SIMULADOR 4)

Teste de hipóteses para a correlação linear ρ :

$$\begin{cases} H_0 : \rho = \rho_0 \\ H_1 : \rho \neq 0 \end{cases}$$

com $\rho_0 = 0$. Estatística do teste:

$$t_{calc} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,9880 \cdot \sqrt{6}}{\sqrt{1-0,9880^2}} = 15,668$$

Rejeita-se a hipótese nula (H_0) se o valor da estatística for tão extremo que se verifique:

$$t_{calc} \leq t_{tab}[\frac{\alpha}{2};(n-2)] \text{ ou } t_{calc} \geq t_{tab}[1-\frac{\alpha}{2};(n-2)]$$

$$t_{tab}[\frac{\alpha}{2};(n-2)] = t_{tab}[\frac{0.05}{2};(6)] = -2,44 \quad t_{tab}[1-\frac{\alpha}{2};(n-2)] = t_{tab}[1-\frac{0.05}{2};(6)] = 2,44$$

Conclusão: frente ao resultado da análise dos dados rejeitamos a hipótese nula sob um nível de significância de 5%.

(SIMULADOR 2 COM t)

Teste de hipóteses para o coeficiente angular β :

$$\begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0 \end{cases}$$

com $\beta_0 = 0$. Estatística do teste:

$$t_{calc} = \frac{b - \beta_0}{s_b}$$

com:

$$s_b = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{\text{QMRES}}{S_{xx}}}$$

$$t_{calc} = 15,5491$$

Rejeita-se a hipótese nula (H_0) se:

$$t_{calc} \leq t_{tab}[\frac{\alpha}{2};(n-2)] \text{ ou } t_{calc} \geq t_{tab}[1-\frac{\alpha}{2};(n-2)]$$

$$t_{tab}[\frac{\alpha}{2};(n-2)] = t_{tab}[\frac{0.05}{2};(6)] = -2,44 \quad t_{tab}[1-\frac{\alpha}{2};(n-2)] = t_{tab}[1-\frac{0.05}{2};(6)] = 2,44$$

Conclusão: frente ao resultado da análise dos dados rejeita-se a hipótese nula sob um nível de significância de 5%.

(SIMULADOR 2 COM t)

Teste de hipóteses para o coeficiente linear α :

$$\begin{cases} H_0 : \alpha = \alpha_0 \\ H_1 : \alpha \neq \alpha_0 \end{cases}$$

com $\alpha_0 = 0$. Estatística do teste:

$$t_{calc} = \frac{a - \alpha_0}{s_a}$$

com

$$s_a = \sqrt{\text{QMRES} \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

$$t_{calc} = 24,5268$$

Rejeita-se a hipótese nula (H_0) se:

$$t_{calc} \leq t_{tab[\frac{\alpha}{2};(n-2)]} \text{ ou } t_{calc} \geq t_{tab[1-\frac{\alpha}{2};(n-2)]}$$

$$t_{tab[\frac{\alpha}{2};(n-2)]} = t_{tab[\frac{0.05}{2};(6)]} = -2,44 \quad t_{tab[1-\frac{\alpha}{2};(n-2)]} = t_{tab[1-\frac{0.05}{2};(6)]} = 2,44$$

Conclusão: frente ao resultado da análise dos dados rejeita-se a hipótese nula sob um nível de significância de 5%.

(SIMULADOR 2 COM t)

O coeficiente de determinação será:

$$R^2 = \frac{\text{variação explicada}}{\text{variação total}} R^2 = \frac{b \cdot S_{xy}}{S_{yy}} = 0,9758$$

O coeficiente de determinação amostral (R^2) é uma medida estatística que informa, em termos percentuais, o quanto da variação observada na variável Y está sendo explicada no modelo pela relação linear estabelecida com a variável X . No exemplo em tela, 97,58%.

12.12 Intervalos de confiança

Um *intervalo de confiança (IC)* pode ser entendido como uma **faixa de valores bastante específica** para uma estatística calculada dentro da qual, sob alguma confiança, podemos afirmar se localizar o valor do parâmetro estimado.

Essa faixa pode ser fechada ou aberta (delimitada apenas por dois ou apenas um valor, respectivamente):

- intervalos de confiança bilaterais: intervalos delimitados por dois valores: mínimo e máximo, dentro do qual todos os valores possuem um mesmo nível de confiança de ocorrência;
 - intervalos de confiança unilaterais: intervalos delimitados apenas em um de seus lados, nos quais todos os valores possuem um mesmo nível de confiança (limitados à direita por um valor máximo ou limitados à esquerda por um valor mínimo).
-

A amplitude de um intervalo de confiança é uma função diretamente proporcional a um *nível de confiança* e à *variabilidade* da população amostrada (quanto maior a variabilidade e/ou o nível de confiança, maior sua amplitude) e inversamente proporcional ao *tamanho amostral* (quanto maior o tamanho da amostra, menor sua amplitude).

$$\text{amplitude} = \text{estimativa amostral} \pm f(\text{confiança}, \text{variabilidade}, \frac{1}{n})$$

Como raramente se dispõe de informação a respeito da variabilidade da característica estudada na população, esse valor é considerado na expressão acima de modo estimado por uma amostra.

Um **intervalo de confiança** reflete uma estimativa objetiva da (im)precisão acarretada pelo tamanho da amostra e, assim, podemos considerá-lo como uma medida da qualidade da pesquisa.

O **nível de confiança** associado ao intervalo é designado pela quantidade $(1 - \alpha)$, sendo α denominado de **nível de significância**: uma medida da probabilidade de erro.

Dependendo do **nível de confiança** que escolhemos, os limites do intervalo mudam para uma **mesma** estimativa amostral. Os **níveis de confiança** mais utilizados na literatura são os de 90%, 95% e 99%.

Assim, $(1 - \alpha)$ traduz o grau de confiança que se tem em que uma *particular amostra* de tamanho n da variável aleatória X dê origem a um intervalo de valores (o intervalo de confiança) que compreenda o verdadeiro valor do parâmetro sobre o qual se estima ou sobre o qual se infere.

Vejam a simulação onde contruímos um grande número de intervalos de confiança calculados sob as mesmas condições (mesma população amostrada, mesmo tamanho amostral (n) e nível de significância α).

(SIMULADOR 5)

Nela podemos observar que uma determinada proporção desses intervalos (aproximadamente igual ao nível de confiança $1 - \alpha$), conterá o *parâmetro* sobre o qual se estima e se deseja inferir.

12.12.1 Intervalos de confiança nos modelos de regressão linear simples

Intervalo de confiança para a resposta média do modelo (equivale a dizer a resposta fornecida pelo modelo ajustado para **valores observados**)

Intervalo de predição para novas observações (equivale a dizer a resposta fornecida pelo modelo ajustado para **valores não observados**)

Intervalo de confiança para as estimativas dos parâmetros do modelo (o modelo ajustado apresenta meras **estimativas**: **a** e **b**, dos parâmetros desconhecidos: α e β).

12.12.1.1 Intervalo de confiança para a resposta média do modelo sob um nível de significância α

$$IC = \hat{y}_0 \pm t_{tab}[\frac{\alpha}{2};(n-2)] \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

em que:

$$\hat{\sigma} = \sqrt{QMRES} = \sqrt{\frac{SQRES}{(n-2)}} = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{(n-2)}}$$

e \hat{y}_0 é o valor médio estimado para um x_0 pertencente à amostra e t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido com $(n-2)$ graus de liberdade. O número de graus de liberdade irá determinar qual curva da família dessa distribuição será utilizada, por essa razão, as tabelas apresentam-se individualizadas por nível de significância e graus de liberdade.

(SIMULADOR 2 COM t)

12.12.1.2 Intervalo de predição para novas observações sob um nível de significância α

$$IC = \hat{y}_0 \pm t_{tab}[\frac{\alpha}{2};(n-2)] \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

em que

$$\hat{\sigma} = \sqrt{QMRES} = \sqrt{\frac{SQRES}{(n-2)}} = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{(n-2)}}$$

e \hat{y}_0 é o valor predito para um x_0 não pertencente à amostra e t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido com $(n-2)$ graus de liberdade. O número de graus de liberdade irá determinar qual curva da família dessa distribuição será utilizada, por essa razão, as tabelas apresentam-se individualizadas por nível de significância e graus de liberdade.

(SIMULADOR 2 COM t)

12.12.1.3 Intervalo confiança para a estimativa a do parâmetro α sob um nível de significância α

$$a \pm t_{tab}[\frac{\alpha}{2};(n-2)] \cdot \hat{\sigma} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

em que

$$\hat{\sigma} = \sqrt{QMRES} = \sqrt{\frac{SQRES}{(n-2)}} = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{(n-2)}}$$

e a é a estimativa do parâmetro α e t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido com $(n-2)$ graus de liberdade. O número de graus de liberdade irá determinar qual curva da família dessa distribuição será utilizada, por essa razão, as tabelas apresentam-se individualizadas por nível de significância e graus de liberdade.

(SIMULADOR 2 COM t)

12.12.1.4 Intervalo confiança para a estimativa b do parâmetro β sob um nível de significância α

$$b \pm t_{tab}[\frac{\alpha}{2},(n-2)] \cdot \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

em que

$$\hat{\sigma} = \sqrt{QMRES} = \sqrt{\frac{SQRES}{(n-2)}} = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{(n-2)}}$$

e b é a estimativa do parâmetro β e t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido com $(n-2)$ graus de liberdade. O número de graus de liberdade irá determinar qual curva da família dessa distribuição será utilizada, por essa razão, as tabelas apresentam-se individualizadas por nível de significância e graus de liberdade.

SIMULADOR 2

Exemplo 6: Um jornal deseja verificar a eficácia de seus anúncios na venda de carros usados e para isso realizou um levantamento de todos os seus anúncios e informações dos resultados obtidos pelas empresas que o contrataram e dele extraiu uma pequena amostra. A tabela abaixo mostra o número de anúncios e o correspondente número de veículos vendidos por 6 empresas que usaram apenas este jornal como veículo de propaganda. Obtenha a equação de regressão linear simples. Qual a estimativa de vendas do modelo para um volume de 36 anúncios? Qual a previsão do número de carros vendidos para um volume de 70 anúncios? Quais os intervalos (estimativa, predição e para os regressores do modelo) sob um nível de significância de 5%

Trazendo os resultados já calculados em exemplos anteriores:

com $n = 6$, $\bar{y} = 90$ e $\bar{x} = 41$ calcula-se

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} = 25172 - \frac{246 \cdot 540}{6} = 3032 S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = 12086 - \frac{246^2}{6} = 2000 S_{yy} = \sum_{i=1}^n y_i^2 -$$

As estimativas dos parâmetros do modelo serão:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{3032}{2000} = 1,5160 \\ a = \bar{y} - b \cdot \bar{x} = 90 - 1,5160 \cdot 41 = 27,844$$

E o modelo toma a seguinte forma:

$$\hat{y} = 27,844 + 1,5160 \cdot x$$

O *valor médio* estimado para um volume de anúncios de 36 veiculações é de 82 carros vendidos. O intervalo de confiança para a *resposta média* do modelo: $IC[\mu(x_0 = 36)]$ sob um nível de significância α será

$$\hat{y}_0 \pm t_{tab}[\frac{\alpha}{2};(n-2)] \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

em que

$$\hat{\sigma} = \sqrt{QMRES} = \sqrt{\frac{SQRES}{(n-2)}} = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{(n-2)}} = 8,0853$$

$\hat{y}_0 = 82$ é o valor médio estimado para o valor observado $x_0 = 36$ (um dado pertencente à amostra) e t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido ($\alpha = 5\%$) com $(n-2) = 4$ graus de liberdade ($t_{tab} = 2,77$).

Assim, $IC[\mu(x = 36)]_{(\alpha=5\%)} = (72,5201; 91,4799)$

(SIMULADOR 2 COM t)

O *valor predito* para um volume de anúncios de 70 veiculações é de 134 carros vendidos. O intervalo de predição para novas observações $IP[Y(x_0)]$ com nível de significância α será:

$$\hat{y}_0 \pm t_{tab}[\frac{\alpha}{2};(n-2)] \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

em que

$$\hat{\sigma} = \sqrt{QMRES} = \sqrt{\frac{SQRES}{(n-2)}} = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{(n-2)}} = 8,0853$$

$\hat{y}_0 = 134$ é o valor predito para um valor não observado $x_0 = 70$ e t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido ($\alpha = 5\%$) com $(n-2) = 4$ graus de liberdade ($t_{tab} = 2,77$).

Assim, $IP[Y(x_0)]_{(\alpha=5\%)} = (105,7845; 162,2155)$

12.13 (SIMULADOR 2 COM t)

Intervalo de confiança para a estimativa a do parâmetro α do modelo sob um nível de significância α :

$$a \pm t_{tab}[\frac{\alpha}{2};(n-2)] \cdot \hat{\sigma} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

em que

$$\hat{\sigma} = \sqrt{QMRES} = \sqrt{\frac{SQRES}{(n-2)}} = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{(n-2)}} = 8,0853$$

$a = 27,844$ é a estimativa do parâmetro α e t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido ($\alpha = 5\%$) com $(n-2) = 4$ graus de liberdade ($t_{tab} = 2,77$).

Assim, $IC(a)_{(\alpha=5\%)} = (5,3676; 50,3204)$.

(SIMULADOR 2 COM t)

Intervalo de confiança para a estimativa b do parâmetro β do modelo sob um nível de significância α :

$$b \pm t_{tab}[\frac{\alpha}{2},(n-2)] \cdot \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

em que

$$\hat{\sigma} = \sqrt{QMRES} = \sqrt{\frac{SQRES}{(n-2)}} = \sqrt{\frac{S_{yy} - b \cdot S_{xy}}{(n-2)}} = 8,0853$$

e $b = 1,5160$ é a estimativa do parâmetro β e t_{tab} é o quantil associado na distribuição “t” de Student (William Sealy Gosset, 1876-1937) ao nível de significância pretendido ($\alpha = 5\%$) com $(n-2) = 4$ graus de liberdade ($t_{tab} = 2,77$).

Assim, $IC(b)_{(\alpha=5\%)} = (1,0152; 2,0168)$.

(SIMULADOR 2 COM t)

12.14 Verificações gráficas (visuais) das premissas do MMQO

A análise dos resíduos de um modelo de regressão linear simples é parte fundamental para que se avalie se o modelo produzido representa de forma acurada a realidade estudada.

- Linearidade no parâmetro: deve-se esperar que a relação entre a variável dependente (Y) e a variável independente (X) possa ser representada por uma função linear};

- pela análise dos gráficos dos resíduos padronizados no eixo y pelos valores estimados e pela variável independente no eixo x . Em geral, valores próximos à linha horizontal representam observações bem estimadas pelo modelo. Os pontos acima e abaixo são observações superestimadas ou subestimadas pelo modelo. A premissa de linearidade é apoiada pelo padrão de distribuição dos pontos, que deve indicar uma razoável igualdade acima e abaixo da linha. Padronizam-se os resíduos brutos pela Divisão de cada um deles pelo desvio padrão; ou seja: $d_i = \frac{e_i}{\hat{\sigma}} = \frac{e_i}{\sqrt{QMRES}}$
-

- independência dos resíduos, com valor médio zero e estejam normalmente distribuídos: ($\varepsilon \sim N(0, \sigma^2)$):
 - pela análise do histograma dos resíduos padronizados, com o propósito de se verificar se sua distribuição guarda semelhança com a da curva normal
 - pela comparação das frequências relativas acumuladas dos resíduos padronizados para os intervalos de $(-1; +1)$, $(-1,64; +1,64)$, $(-1,96; +1,96)$ com as probabilidades da distribuição normal nesses mesmos intervalos (68%, 95% e 99%)
 - pela análise do gráfico dos resíduos padronizados ordenados pelos quantis da distribuição normal padronizada, que deve se aproximar da bisetriz do primeiro quadrante
 - a variância residual seja sempre constante (homocedástica) para todas as observações, isto é, $VAR(\varepsilon) = E(\varepsilon^2) = \sigma^2$
 - ausência de autocorrelação entre os termos de erros:
 - pela análise do gráfico dos resíduos padronizados pelos valores estimados \hat{y} , que deve apresentar pontos dispostos aleatoriamente sem padrão aparente;
 - mensuração das variáveis: assume-se que as variáveis foram medidas sem erro;
 - correta especificação do modelo: todas as variáveis independentes teoricamente relevantes foram incluídas no modelo e nenhuma irrelevante foi mantida;
 - ausência de multicolinearidade.
-

12.15 Verificações adicionais

- Análise de pontos com elevada capacidade de alavancar o modelo. A alavancagem mede o quanto uma observação x_i contribui para a predição de \hat{y}_i pelo modelo. Um ponto é considerado alavancado (*leverage*) quando este exerce uma forte influência no seu valor ajustado, sem com isso afetar a estimativa dos parâmetros do modelo. De modo análogo à distância de Cook, há diversos critérios para estabelecer um valor crítico para os *hat values*: h_{ii} :
 - $h_{ii} > 2p/n$ (Hoaglin e Welsch, 1978),
 - $h_{ii} > 3p/n$.
- Pontos discrepantes (*outliers*): A discrepancia pode ser medida pela distância residual. Entretanto, os resíduos não são uma medida completa da discrepancia. Para tanto basta-se imaginar casos onde uma observação possua elevada alavancagem que arraste o modelo inteiro em sua direção, resultando em pequenos resíduos. Uma forma de isolar esses pontos é dividindo seu resíduo por $1-h_{ii}$, obtendo-se a partir dessa expressão os resíduos *studentizados*.

- influentes: A estatística distância de Cook mede a influência de um determinado dado da amostra no que tange a quanto ele está afetando a linha de regressão, sendo medida pelo quanto a linha de regressão se alteraria caso esse dado fosse removido da análise: ele exerce um destacado impacto na estimativa dos parâmetros do modelo. A influência na locação (afastamento de alguma observação da vizinhança do resto dos dados) pode ser investigada pelo gráfico feito das distâncias de Cook contra os valores ajustados. Há vários critérios para se estabelecer um valor limite para a estatística de Cook:\
- $D_i > 1$ (Cook e Weisberg, 1982);
- duas vezes a média das distâncias de Cook;
- $4/n < D_i < 1$ (Bollen et al, 1990); ou,
- o valor crítico do quantil da distribuição F para uma significância igual a 0.5 com $df1=p$ e $df2=n-p$.

Exemplo 7: Um jornal deseja verificar a eficácia de seus anúncios na venda de carros usados e para isso realizou um levantamento de todos os seus anúncios e informações dos resultados obtidos pelas empresas que o contrataram e dele extraiu uma pequena amostra. A tabela abaixo mostra o número de anúncios e o correspondente número de veículos vendidos por 6 empresas que usaram apenas este jornal como veículo de propaganda. Estime os parâmetros de um modelo de regressão linear simples de X por Y verifique os pressupostos subjacentes ao método utilizado. Faça a análise dos resíduos e identifique possíveis *outliers* .

Trazendo o modelo estimado anteriormente: $\hat{y} = 27,844 + 1,5160 \cdot x$

12.15.0.1 Roteiro básico para uma análise de regressão linear simples

- Definir o problema de pesquisa, selecionar a variável dependente e identificar a variável independente; ou seja, proceder a especificação do modelo. Aqui o pesquisador deve definir qual é a relação esperada entre a variável dependente e a independente;
- Maximizar o número de observações no sentido de aumentar o poder estatístico, a capacidade de generalização e reduzir toda sorte de problemas associados a estimativa de parâmetros populacionais a partir de dados amostrais com n reduzido;
- Estimar um modelo;
- Verificar em que medida os dados disponíveis satisfazem os pressupostos da análise de regressão de mínimos quadrados ordinários. Como procedimento padrão, o pesquisador deve reportar as técnicas utilizadas para corrigir eventuais violações (transformações, re-codificações, aumento de n , etc.);
- Interpretar os resultados, caso o modelo seja validado.

12.15.0.2 Homocedasticidade: transformações para estabilização da variância

Quando se observa que a distribuição gráfica dos resíduos não se mostra homocedástica, muitas vezes é útil aplicar uma transformação de Box-Cox para estabilizarmos a variância (torná-la constante independentemente do valor do resíduo).

Considerando X_1, \dots, X_n os dados originais, a transformação de Box-Cox consiste em encontrar um λ tal que os dados transformados Y_1, \dots, Y_n se aproximem de uma distribuição normal.

O modelo passa a assumir a forma: $Y^\lambda = X \cdot \beta + \varepsilon$ com Y_λ sendo:

$$\frac{Y^\lambda - 1}{\lambda} \text{ se } \lambda \neq 0 \\ \ln(Y_i) \text{ se } \lambda = 0$$

12.15.0.3 Transformações para linearização das relações

Algumas vezes as relações observadas entre a variável dependente e a independente não se mostram diretamente lineares.

Relações não-lineares podem ser linearizadas pela aplicação de transformações aos dados:

- Função hiperbólica: $Y = \frac{X}{a \cdot X - b}$, pela forma transformada: $\frac{1}{Y} = a - \frac{b}{X}$
 - Função exponencial: $Y = a \cdot e^{b \cdot X}$, pela forma transformada: $\ln(Y) = \ln(a) + b \cdot X$
 - Função potência: $Y = a \cdot X^b$, pela forma transformada: $\ln(Y) = \ln(a) + b \cdot \ln(X)$
-

12.15.0.4 Tabelas

12.15.0.5 Resolução do sistema de equações matriciais

Seja a matriz Y das observações realizadas na variável dependente Y_i (dimensão $n \times 1$):

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Seja a matriz X das observações realizadas na variável independente X_i (dimensão $n \times 2$):

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Seja a matriz β dos parâmetros a serem estimados (dimensão: 2×1):

$$\beta = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

Seja a matriz e dos termos aleatórios (dimensão: $n \times 1$), não correlacionados, com média zero e variância constante:

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Então podemos escrever o seguinte sistema matricial:

$$Y = X \cdot \hat{\beta} + e$$

A minimização da soma dos quadrados dos resíduos pode ser realizada fazendo-se:

$$\Sigma(e_i)^2 = e^T \cdot e$$

O sistema acima tomará a forma:

$$e = Y - X \cdot \hat{\beta}$$

$$e^T \cdot e = (Y - X \cdot \hat{\beta})^T \cdot (Y - X \cdot \hat{\beta})$$

Expandindo:

$$Y^T \cdot Y - 2 \cdot \hat{\beta}^T \cdot X^T \cdot Y + \hat{\beta}^T \cdot X^T \cdot X \cdot \hat{\beta}$$

Minimizando os resíduos, obtemos a equação normal:

$$(X^T \cdot X) \cdot \hat{\beta} = X^T \cdot Y$$

Multiplicando ambos os lados por $(X^T \cdot X)^{-1}$:

$$(X^T \cdot X)^{-1} \cdot (X^T \cdot X) \cdot \hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

Por fim, considerando-se que $(X^T \cdot X)^{-1} \cdot (X^T \cdot X) = I$, obtemos a solução para $\hat{\beta}$:

$$\hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

Módulo 13

Orientações Gerais

13.1 Informações administrativas

13.1.1 Regimento geral da UEL

O Regimento Geral da Universidade Estadual de Londrina está disponível nesse link

R E G I M E N T O

G E R A L

D A U E L

Figure 13.1: Regimento Geral da UEL

Art. 60. É vedado o abono de falta.

Art. 61. Considerar-se-á aprovado nas disciplinas especiais o estudante que obtiver média final igual ou superior a seis (6,0) ou conceito equivalente e freqüência de, no mínimo, setenta e cinco por cento (75%).

Figure 13.2: Artigos 60 e 71 do Regimento Geral da UEL

13.1.2 Amparos e apoios na UEL

13.1.3 Tutoriais para os estudantes da graduação da UEL

Tutoriais para os estudantes da graduação da Universidade Estadual de Londrina estão disponíveis nesse link

AMPAROS e APOIOS

A Universidade dispõe de diversas resoluções para amparar legalmente os estudantes em razão de dificuldades vivenciadas ao longo da graduação: licença médica, trancamento de matrícula, fracionamento de série, entre outras.

Além desses, os estudantes com deficiência contam com amparos especializados como tempo adicional para avaliações, suporte de intérprete de Libras, ledores e cuidadores, conforme as especificidades de cada deficiência, e devem receber atenção diferenciada dos docentes em caso de dificuldades com os conteúdos, mediante supervisão e/ou monitoria acadêmica.

Também são amparados para suplementação de estudos e aceleração nos casos de altas habilidades/superdotação.

Além dos amparos legais, a instituição disponibiliza apoios pedagógicos, médicos, psicológicos e sociais, destacando-se:

1. SEBEC (Serviço de Bem-Estar à Comunidade), para suportes sociais e de saúde mental;
2. DASC (Divisão de Assistência à Saúde da Comunidade), com serviços na área médica; e
3. LABTED (Laboratório de Tecnologia Educacional), que oferece cursos para elaboração e apresentação de seminários acadêmicos com técnica e adequação didática.

Figure 13.3: Amparos e apoios na UEL

Tutoriais (Portal do Estudante de Graduação)

CONHEÇA O PORTAL DO ESTUDANTE DE GRADUAÇÃO
www.uel.br/portaldoestudante

Serviços – Requerimentos

Como requerer Segunda Chamada



Acesse o tutorial no formato PDF



Acesse o tutorial em vídeo

Figure 13.4: Tutoriais para os estudantes da graduação da UEL

13.2 Programas de atividade acadêmica

13.2.1 Relações Públicas: 6STA003 - Estatística

13.2.2 Química: 1STA001 - Introdução à estatística

13.2.3 Farmácia: 2STA010 - Elementos de bioestatística

13.2.4 Administração de Empresas: 7EMA003 - Estatística

13.2.5 Ciências Contábeis: 6STA016 - Estatística

13.2.6 Ciências Econômicas: 2STA004 - Estatística econômica

13.2.7 Engenharia Civil: 2STA016 - Estatística e probabilidades

13.2.7.1 Horários:

Turma 1000: segundas e sextas-feiras (10 h 15 min - 11 h 55 min) Turma 2000: segundas e sextas-feiras (8 h 20 min - 10 h 00 min)

13.2.7.2 Locais: CTU

Turma 1000: segundas-feiras: sala 1012 e sextas-feiras: sala 1015 Turma 2000: segundas-feiras: sala 1015 e sextas-feiras: sala 1014

13.2.7.3 Ementa contida na Resolução Resolução CEPE/CA 073/2022:

Técnicas de amostragem. Medidas de posição e dispersão. Introdução à probabilidade. Variáveis aleatórias discretas, contínuas e suas principais distribuições de probabilidade. Inferência sobre médias, variâncias e proporções. Noções de planejamento de experimentos aplicados à Engenharia.

13.2.7.4 Conteúdo programático:

1. Introdução à pesquisa científica, análise exploratória e descritiva de dados e assuntos correlacionados: noções sobre a produção de conhecimento por meio da pesquisa científica, conceitos de população e parâmetros, amostra e estatísticas, tipos de variáveis, indexação e somatório de dados, combinações, conectivos lógicos, conjuntos, diagramas de Venn, operações com conjuntos, coleta de dados das alturas dos alunos da turma para análise exploratória de dados, apresentação dos dados brutos e em rol, apresentações gráficas elementares, sínteses numéricas de posição (média, moda), sínteses numéricas de dispersão (máximo, mínimo, amplitude, variância e coeficiente de variação), medidas separatrizes (percentis, decis, quartis), mediana, apresentação tabular de dados, média, moda e variância para dados agrupados em tabelas de frequências, apresentações gráficas para variáveis qualitativas (barras, setores) e quantitativas (histograma)
2. Introdução ao cálculo de probabilidades, variáveis aleatórias e distribuições teóricas de probabilidade: conceitos essenciais de experimentos aleatórios e experimentos determinísticos, a variável aleatória, o conjunto de possíveis resultados do experimento aleatório (espaço amostral e seus elementos), eventos simples e compostos (representações com diagramas de Venn), conceitos de probabilidade: (1) clássico (a priori); (2) frequentista (a posteriori); (3) conceito axiomático: a probabilidade como uma função, probabilidade da união de eventos,

probabilidade de eventos condicionados e independentes, introdução ao teorema de Bayes, funções de distribuição e densidade de probabilidade, modelo teórico discreto de Bernoulli, binomial e de Poisson, modelo teórico Normal.

3. Introdução às distribuições amostrais e testes de hipóteses paramétricos: distribuição das médias amostrais e a construção de intervalos de confiança, distribuição das proporções amostrais e a construção de intervalos de confiança, distribuição das variâncias amostrais e a construção de intervalos de confiança, teste de hipóteses sobre a média de uma população, teste de hipóteses sobre a proporção de uma população, teste de hipóteses sobre a variância de uma população.
4. Introdução ao planejamento de pesquisas e levantamentos amostrais: tipos de levantamentos amostrais (probabilísticos e não probabilísticos), levantamento amostral aleatório simples e sistematizado, levantamento amostral aleatório estratificado e por conglomerados, dimensionamento de amostras para inferências sobre médias e proporções.

13.2.7.5 Bibliografia básica:

1. ROSS, Sheldon. Probabilidade: um curso moderno com aplicações. 8 ed. Porto Alegre: Bookman, 2010. 606 p.
2. BUSSAB, W. O., MORETTIN, P. Estatística básica, 8^a ed., São Paulo: Saraiva, 2013.

13.2.7.6 Procedimentos de ensino

1. O processo de ensino será composto por um conjunto de atividades presenciais teóricas expositivas e resolução de exercícios em sala de aula a partir do dia 14/10/2024, nos dias e horários determinados no Sistema UEL.
2. Todo material utilizado (textos, slides, listas de exercícios) será disponibilizado na sala de aula virtual a ser criada na Plataforma *Google Classroom*, de adesão compulsória por parte do discente, por meio de convite enviado ao seu *email* registrado no Sistema UEL.
3. **Toda** comunicação entre discentes e o docente se dará por meio de postagens na plataforma *Google Classroom*.
4. As datas das prova escrita presencial, do seminário e do exame final indicadas no cronograma a seguir **poderão** ser alteradas tanto em razão de **situações não previstas no atual calendário acadêmico** quanto do bom progresso das atividades didáticas, **mediante comunicação aos alunos com devida antecipação**.

13.2.7.7 Formas e critérios de avaliação:

Durante o semestre serão realizadas as seguintes atividades avaliativas:

- duas (2) **provas escritas presenciais** (P1, P2) referentes ao conteúdo das aulas e valendo de zero (0) a dez (10) pontos cada uma;
- duas (2) atividades (A1, A2) no formato de **listas de exercícios** disponibilizadas na plataforma *Google Classroom*, valendo de zero (0) a dez (10) pontos cada uma. As atividades A1 e A2 poderão ser compostas **por mais de uma lista de exercícios cada uma** sendo, nesse caso, atribuída a média aritmética das notas de todas as listas que compõem cada uma das atividades:

$$A1 = \frac{(L_{1,1} + \dots + L_{1,n})}{n} \text{ e } A2 = \frac{(L_{2,1} + \dots + L_{2,m})}{m}$$

A **média final** (*MF*) será calculada pela seguinte expressão que atribui peso 4 para as provas escritas presenciais e peso 1 para as listas de exercícios:\

$$MF = \frac{4(P1) + 4(P2) + 1(A1) + 1(A2)}{10}$$

Ao final da disciplina **haverá exame final conforme estabelecido no Regimento da UEL (Art. 59).**

13.2.8 Ciência de dados e Inteligência Artificial: 2STA011 - Probabilidade

13.2.8.1 Horários:

segundas e quartas-feiras (21 h 10 min - 22 h 50 min)

13.2.8.2 Locais: CCE

Sala 03 e XX

13.2.8.3 Ementa contida na Resolução CEPE/CA 060/2022:

Probabilidade e propriedades. Probabilidade condicional e independência. Variáveis aleatórias discritas e principais modelos de distribuição discretas. Variáveis aleatórias contínuas e principais modelos de distribuições contínuas. Processo de *Poisson*. Cadeias de *Markov*. Simulação de Monte Carlo. Uso de programa estatístico.

13.2.8.4 Conteúdo programático:

1. Módulo 1: Probabilidade
 - Introdução à Probabilidade: aspectos históricos, experimentos aleatórios e determinísticos, espaços e eventos, conceitos.
 - Probabilidade da adição de eventos (disjuntos e não disjuntos).
 - Probabilidade Condicional de eventos e independência de eventos.
 - Teorema de Bayes.
 - Simulações usando a linguagem R.
 - Variáveis aleatórias discretas e contínuas (função distribuição e de densidade de probabilidade, esperança e variância).
 - Principais modelos teóricos discretos e contínuos de probabilidade.
2. Módulo 2: Processos estocásticos
 - Aplicações do processo de Poisson, de simulações de Monte Carlo e cadeias de Markov.

13.2.8.5 Bibliografia básica:

1. ROSS, Sheldon. Probabilidade: um curso moderno com aplicações. 8 ed. Porto Alegre: Bookman, 2010. 606 p.
2. MEYER, Paul L. Probabilidade: Aplicações à Estatística. 2 ed. Rio de Janeiro: LTC, 2010. 426 p.

13.2.8.6 Procedimentos de ensino

1. O processo de ensino será composto por um conjunto de atividades presenciais teóricas expositivas e resolução de exercícios em sala de aula a partir do dia 14/10/2024, nos dias e horários determinados no Sistema UEL.
2. Todo material utilizado (textos, slides, listas de exercícios) será disponibilizado na sala de aula virtual a ser criada na Plataforma *Google Classroom*, de adesão compulsória por parte do discente, por meio de convite enviado ao seu *email* registrado no Sistema UEL.
3. **Toda** comunicação entre discentes e o docente se dará por meio de postagens na plataforma *Google Classroom*.
4. As datas das prova escrita presencial, do seminário e do exame final indicadas no cronograma a seguir **poderão** ser alteradas tanto em razão de **situações não previstas no atual calendário acadêmico** quanto do bom progresso das atividades didáticas, **mediante comunicação aos alunos com devida antecipação**.

13.2.8.7 Formas e critérios de avaliação:

Durante o semestre serão realizadas as seguintes atividades avaliativas:

- uma (1) **prova escrita presencial (P)** referente ao conteúdo das aulas do módulo 1 valendo de zero (0) a dez (10) pontos;
- uma (1) apresentação em sala na forma de **seminário (S)** com tema a ser determinado e referente ao conteúdo das aulas do módulo 2, valendo de zero (0) a dez (10) pontos;
- uma (1) atividade no formato de **listas de exercícios (L)** disponibilizadas na plataforma *Google Classroom*, referente ao conteúdo das aulas do módulo 1 e valendo de zero (0) a dez (10) pontos. A atividade L poderá ser composta **por mais de uma lista de exercícios** sendo, nesse caso, atribuída a média aritmética das notas de todas as listas que compõem essa atividade:

$$L = \frac{(L_{1,1} + \dots + L_{1,n})}{n}$$

A **média final (MF)** será calculada pela seguinte expressão que atribui peso 4 para a prova escrita presencial, peso 5 para o seminário e peso 1 para a lista de exercícios:

$$MF = \frac{4(P) + 5(S) + 1(L)}{10}$$

Ao final da disciplina **haverá exame final conforme estabelecido no Regimento da UEL (Art. 59)**.