

**UNIVERSIDADE  
ESTADUAL DE LONDRINA**

UNIVERSIDADE ESTADUAL DE LONDRINA  
CCE - Centro de Ciências Exatas  
DSTA - Departamento de Estatística  
Apostila de Estatística  
Prof. M.e Eng. Felinto Junior Da Costa

Londrina, 15 de março de 2023.



# Índice

	<b>9</b>
<b>1 Introdução histórica da estatística</b>	<b>11</b>
1.1 Primeiros levantamentos, estudos e publicações & Demografia e aritmética política	11
1.2 Visualização de dados & Estudos e primeiras publicações . . . . .	18
1.3 Nomes notáveis . . . . .	22
1.4 Revista Biometrika . . . . .	23
1.5 Eugenia . . . . .	24
<b>2 Introdução conceitual essencial</b>	<b>29</b>
2.1 Estatística descritiva . . . . .	29
2.2 Estatística inferencial . . . . .	30
2.3 Produção de conhecimento . . . . .	30
2.4 População (universo) & amostra . . . . .	32
2.5 Parâmetros e estatísticas . . . . .	32
2.6 Tipos de variáveis . . . . .	32
2.7 Indexação de dados ( $i$ ) . . . . .	34
2.8 Noções básicas sobre somatórios ( $\Sigma$ ) . . . . .	34
2.9 Análise combinatória: diagramas de árvore, permutações (arranjos) & combinações	36
2.10 Conectivos lógicos . . . . .	41
2.11 Leis de De Morgan . . . . .	42
2.12 Noções básicas para o uso de calculadora (Cassio fx-82MS) . . . . .	43
<b>3 Introdução à estatística descritiva</b>	<b>47</b>
3.1 Análise exploratória . . . . .	47
3.2 Dados brutos, em rol, diagrama de ramos & folhas e de dispersão unidimensional .	48
3.3 Sínteses numéricas descritivas . . . . .	51
3.4 Medidas de forma (assimetria & curtose) . . . . .	69
3.5 Apresentação tabular de dados . . . . .	72
3.6 Apresentação gráfica de dados . . . . .	87

<b>4 Introdução ao cálculo de probabilidades</b>	<b>95</b>
4.1 Introdução conceitual essencial . . . . .	96
4.2 Probabilidade . . . . .	105
4.3 Teorema de Bayes . . . . .	125
4.4 Teoremas da Teoria das probabilidades . . . . .	142
<b>5 Introdução a variáveis aleatórias</b>	<b>147</b>
5.1 Função discreta de distribuição de probabilidade . . . . .	147
5.2 Função de densidade de probabilidade . . . . .	151
5.3 Esperança e variância de uma variável aleatória discreta . . . . .	155
5.4 Esperança e variância de uma variável aleatória contínua . . . . .	158
<b>6 Introdução a modelos teóricos de probabilidade</b>	<b>159</b>
6.1 Modelos teóricos discretos . . . . .	159
6.2 Modelos teóricos do tempo de espera . . . . .	168
6.3 Modelos teóricos contínuos . . . . .	180
6.4 Tabelas . . . . .	208
<b>7 Introdução ao planejamento de pesquisas</b>	<b>213</b>
7.1 Planejamento de pesquisas . . . . .	215
7.2 Tipos de pesquisas . . . . .	216
7.3 Principais etapas de uma pesquisa: . . . . .	218
7.4 População . . . . .	219
7.5 Censo . . . . .	220
7.6 Amostra . . . . .	220
7.7 Planejamento do levantamento amostral . . . . .	221
7.8 Elaboração dos questionários . . . . .	221
7.9 Técnicas de amostragem . . . . .	223
7.10 Amostragem probabilística . . . . .	223
7.11 Amostragem não probabilística . . . . .	238
7.12 Dimensionamento de amostras . . . . .	239
<b>8 Introdução às estatísticas epidemiológicas</b>	<b>247</b>
8.1 Terminologia . . . . .	247
8.2 Medidas de risco, morte, associação e correlação . . . . .	250
8.3 Sobrevida . . . . .	254
8.4 Medidas de associação . . . . .	255
8.5 Intervalos de confiança . . . . .	265

<b>9 Introdução à distribuição das médias e diferenças entre médias amostrais e seus intervalos de confiança</b>	<b>271</b>
9.1 Distribuições amostrais . . . . .	271
9.2 Intervalos de confiança . . . . .	273
9.3 Distribuição das médias amostrais . . . . .	276
9.4 Distribuição das diferenças de médias amostrais independentes . . . . .	307
9.5 Distribuição das diferenças de médias amostrais dependentes . . . . .	324
<b>10 Introdução à distribuição das proporções amostrais e seus intervalos de confiança</b>	<b>327</b>
10.1 Conceito elementar de uma proporção . . . . .	327
10.2 Proporção amostral como uma variável Binomial . . . . .	328
10.3 Distribuição das proporções amostrais . . . . .	329
10.4 Intervalo de confiança para proporções amostrais . . . . .	335
<b>11 Introdução a testes de hipóteses</b>	<b>341</b>
11.1 Epistemologia . . . . .	341
11.2 Histórico . . . . .	343
11.3 Conceitos iniciais . . . . .	347
11.4 Efeito do limite central . . . . .	348
11.5 Erro global . . . . .	348
11.6 Diretrizes gerais de um teste de hipóteses . . . . .	350
11.7 Formulação e estruturação de um teste de hipóteses . . . . .	351
11.8 Natureza dos erros envolvidos em um teste de hipóteses . . . . .	351







## Módulo 1

# Introdução histórica da estatística

### 1.1 Primeiros levantamentos, estudos e publicações & Demografia e aritmética política

1086

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 1.1).

1602

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? [...]”

1603

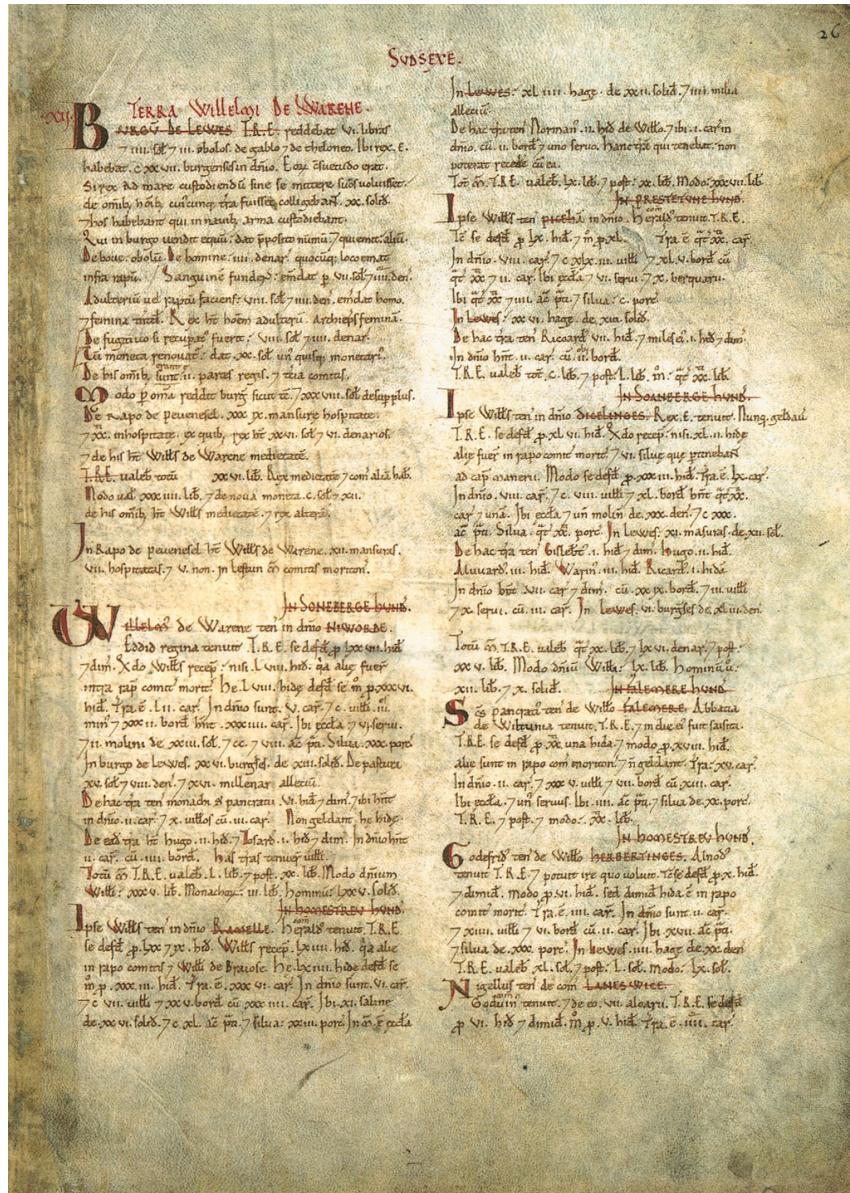


Figure 1.1: Domesday Book

## 1.1. PRIMEIROS LEVANTAMENTOS, ESTUDOS E PUBLICAÇÕES & DEMOGRAFIA E ARITMÉTICA POLÍTICA

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 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 1.2).

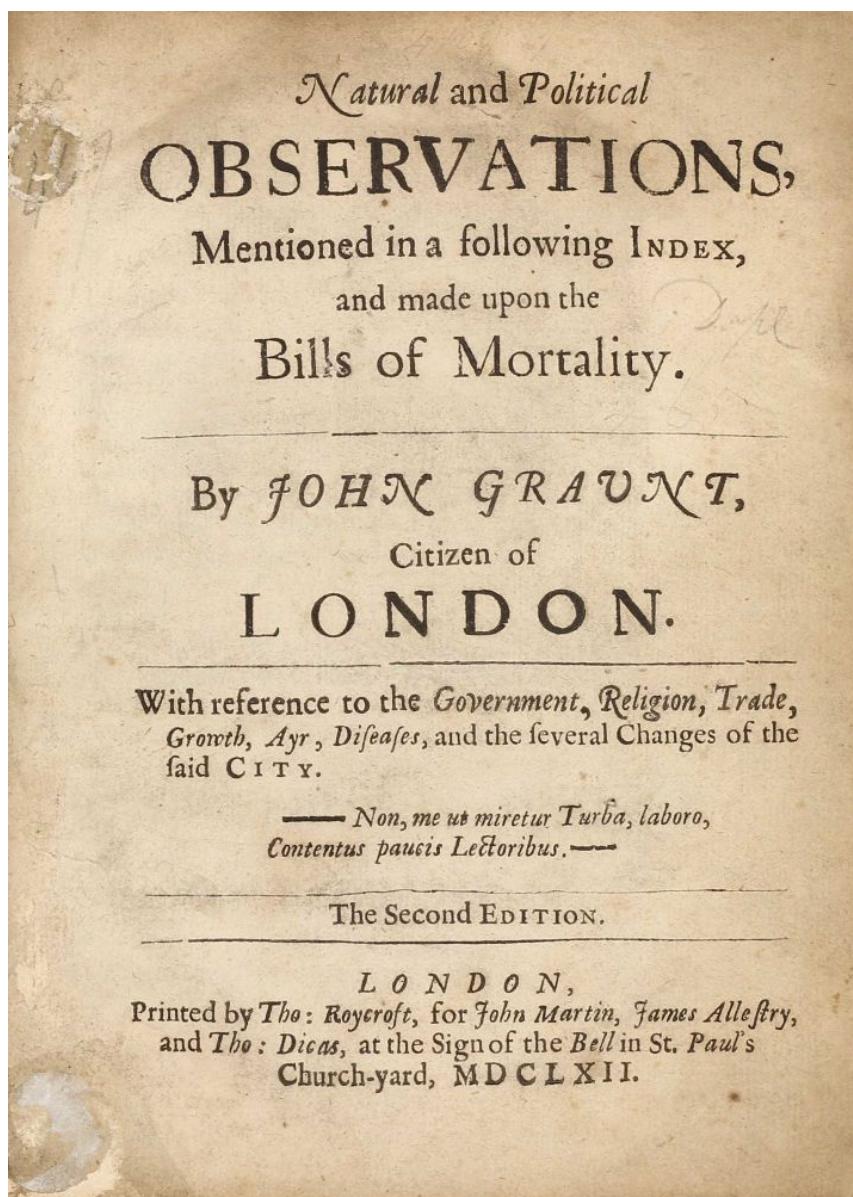


Figure 1.2: Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality (ed. de 1662)

1660

Herman Conring (1606-1681), professor de filosofia, medicina e política da Universidade de Helmstadt (atual Alemanha), criou um curso de Ciência política em 1660, que descrevia e examinava as questões fundamentais do Estado. Nele a **estatística** passou a ser considerada como uma disciplina autônoma que tinha por objetivo a descrição das coisas do Estado.

1687

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 1.3).

O Capitão John Graunt e William Petty instituiram na Inglaterra um novo ramo de estudos denominado *Political arithmetic* (Aritmética política)

1693

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 1.4).

1749

Com um sentido não relacionado com números ou matemática, a palavra **estatística** parece ter sido proposta pela primeira vez 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 1.5).

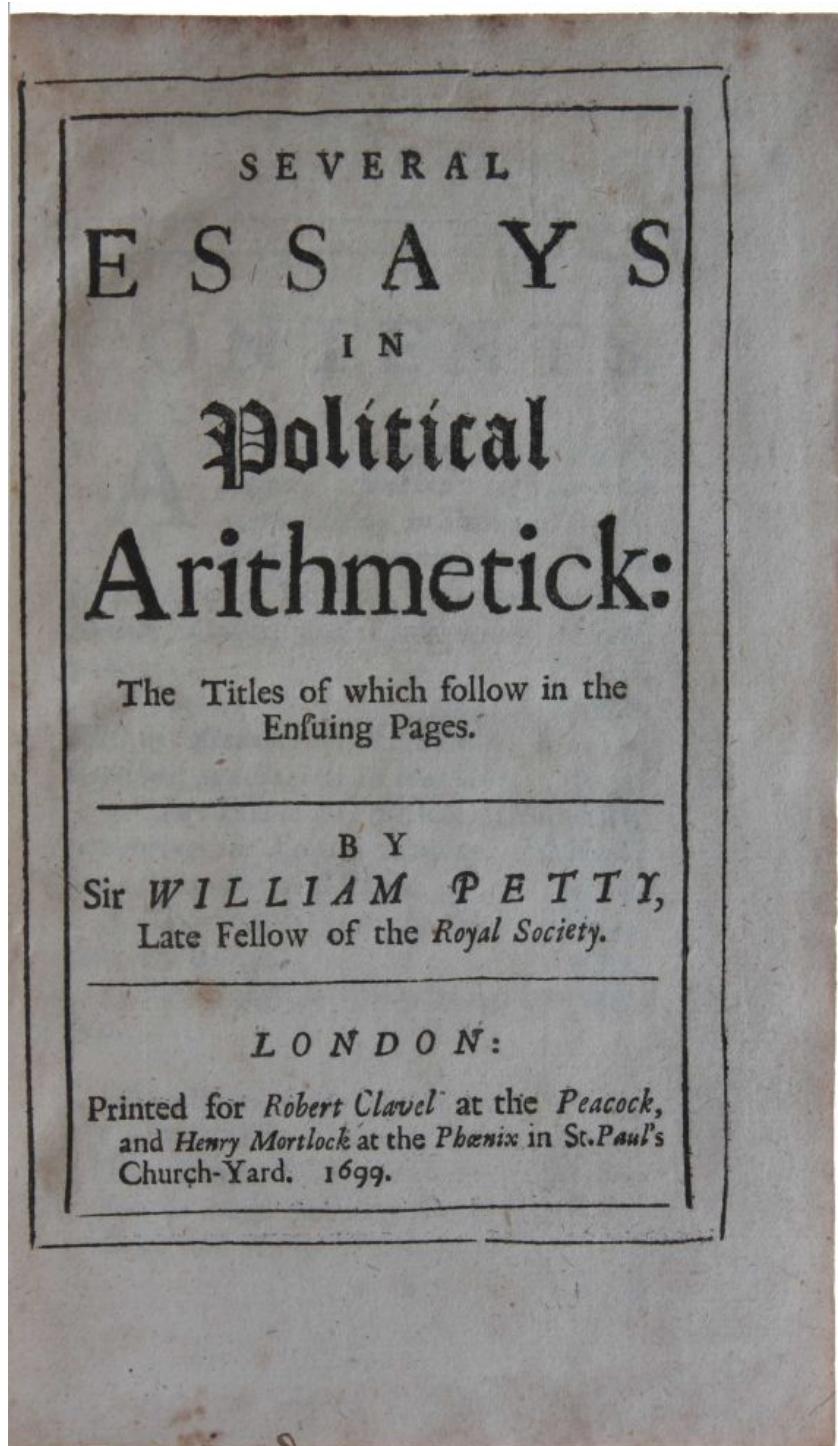


Figure 1.3: Several Essays in Political Arithmetick (ed. de 1699)

Age. Curt.	Per- sons.	Age. Curt.	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	463	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. Curt.	Per- sons.	Age. Curt.	Per- sons											
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	66	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 1.4: Halley's life table (1693)

1771

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 (mais uma vez num sentido não associado a números ou matemática, Figura 1.6).

1790

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 1.7).

1854

O médico inglês (considerado por alguns como o “pai” da epidemiologia moderna) John Snow



HAB Wolfenbüttel, M: Sf 3

Figure 1.5: Abriß der neuen Staatswissenschaft der vornehmen Europäischen Reiche und Republiken (1749)

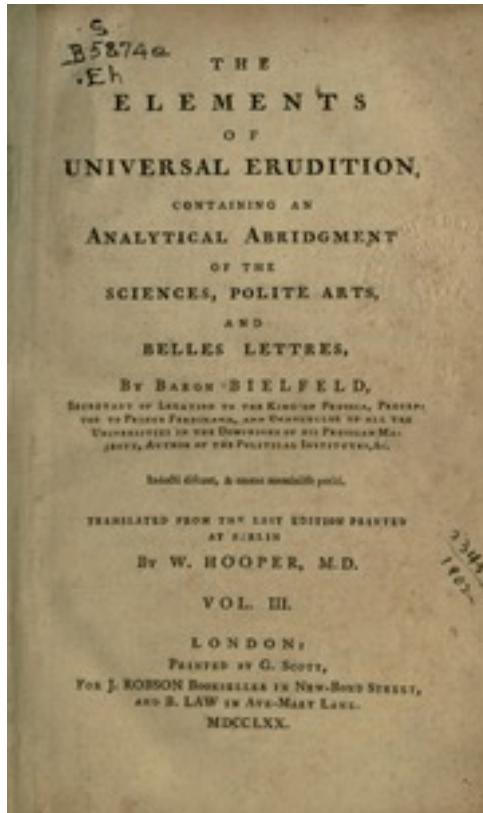


Figure 1.6: The Elements of Universal Erudition (1771)

(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 1.8).

## 1.2 Visualização de dados & Estudos e primeiras publicações

1765

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 1.9).

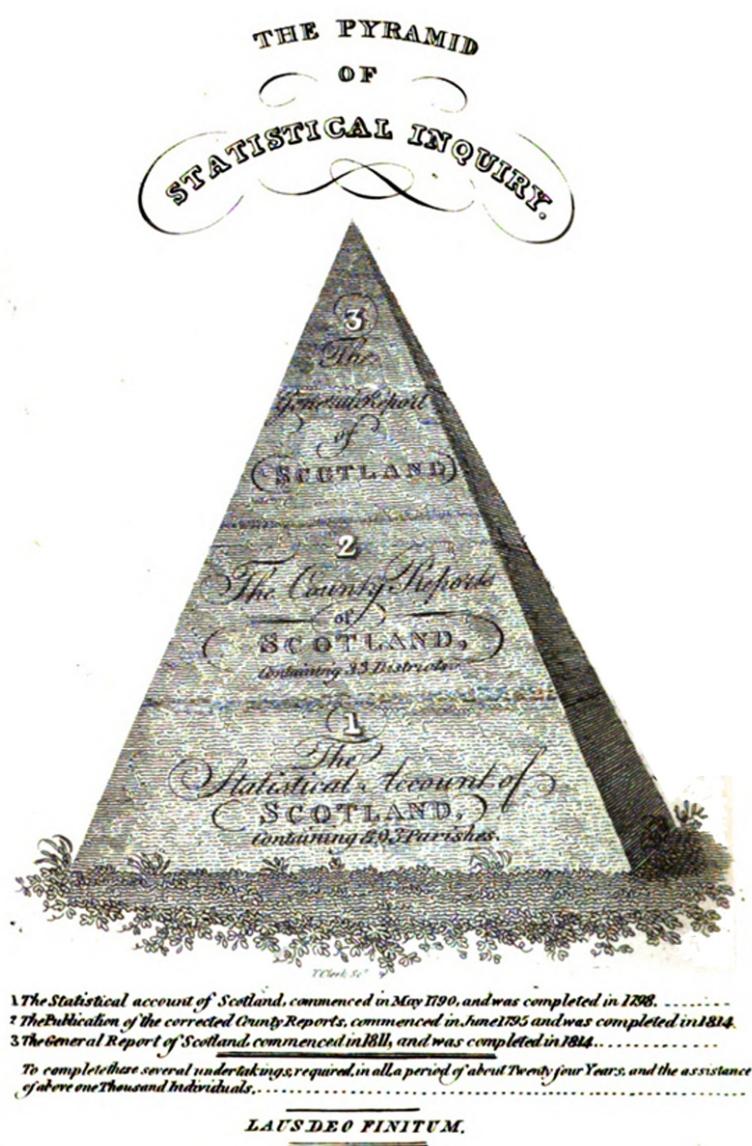


Figure 1.7: The Pyramid of Statistical Enquiry (1814)

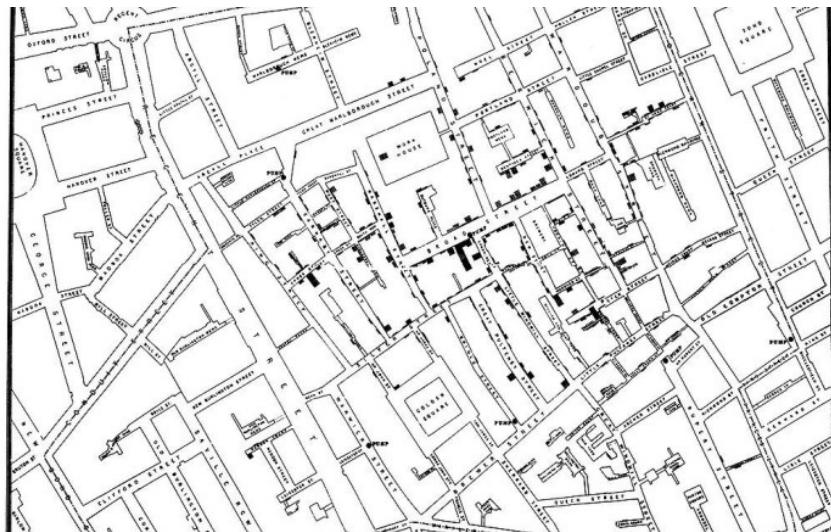


Figure 1.8: Mapa dos casos de cólera (1854)

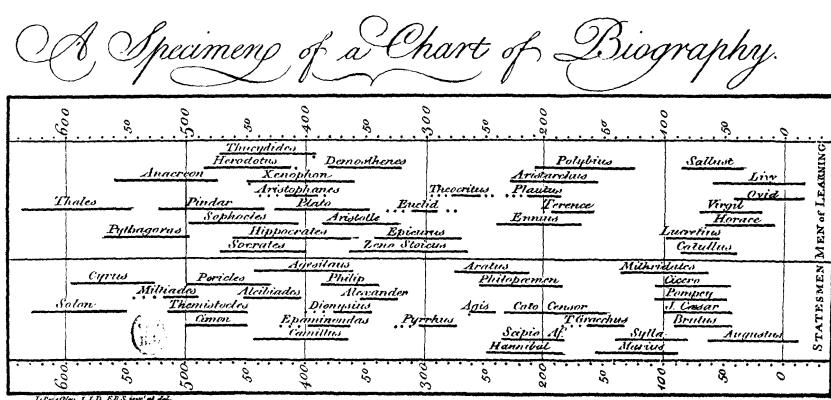


Figure 1.9: Expectativa de vida de diversas pessoas (1765)

1786

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 1.10); e,
- em 1801, o gráfico de setores (Figura 1.11).

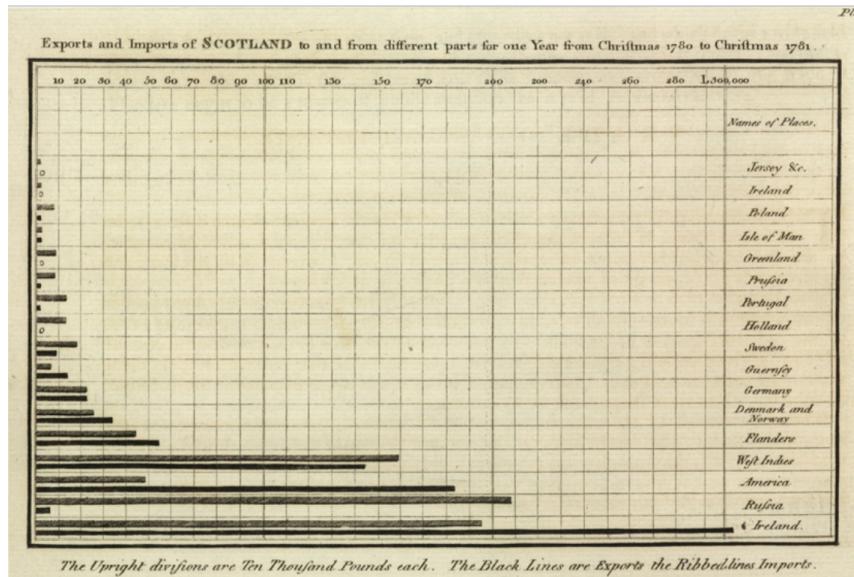


Figure 1.10: Commercial and Political Atlas (Atlas Comercial e Político de 1786): cada barra representa as exportações e importações da Escócia para 17 países em 1781

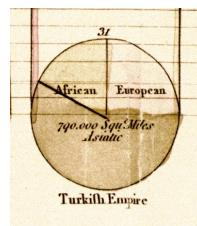


Figure 1.11: Statistical Breviary (Breviário Estatístico de 1801): proporção da extensão do Império Turco em diferentes regiões do mundo: Ásia, Europa e África, antes de 1789

1856

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 1.12 e 1.13.

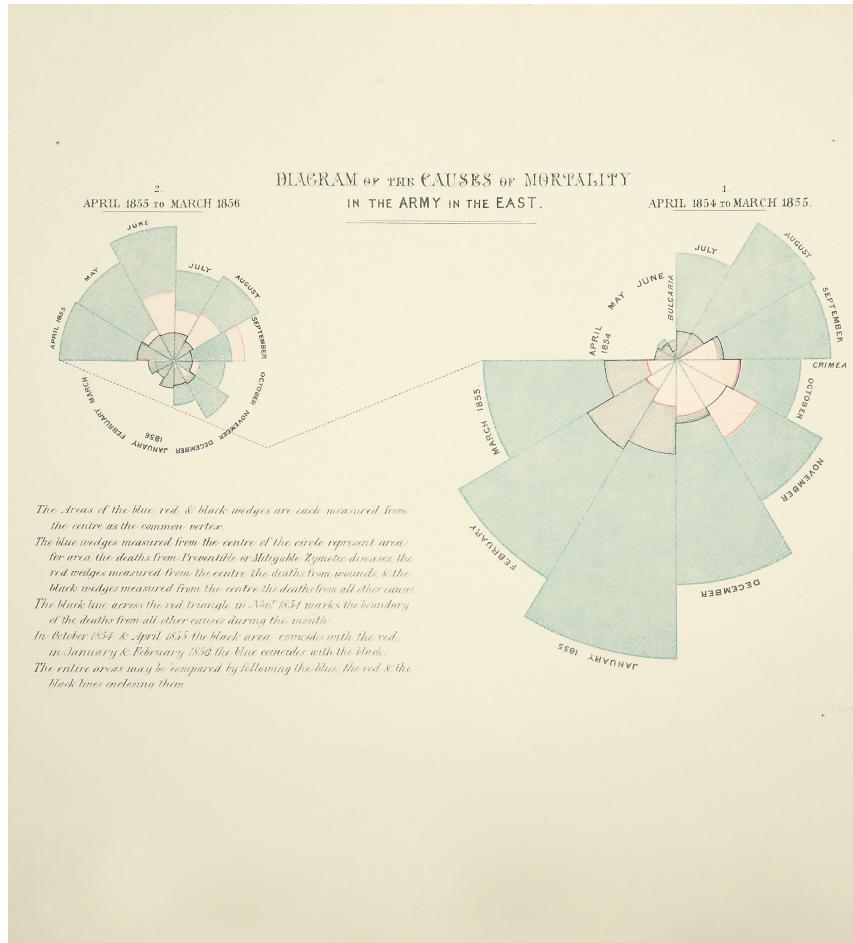


Figure 1.12: Esse diagrama (coxcomb) feito durante a Guerra da Crimeia foi dividido igualmente em 12 setores, representando os meses do ano, com a área sombreada do setor de cada mês proporcional à taxa de mortalidade naquele mês. Seu sombreamento com código de cores indicava a causa da morte em cada área do diagrama

### 1.3 Nomes notáveis

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

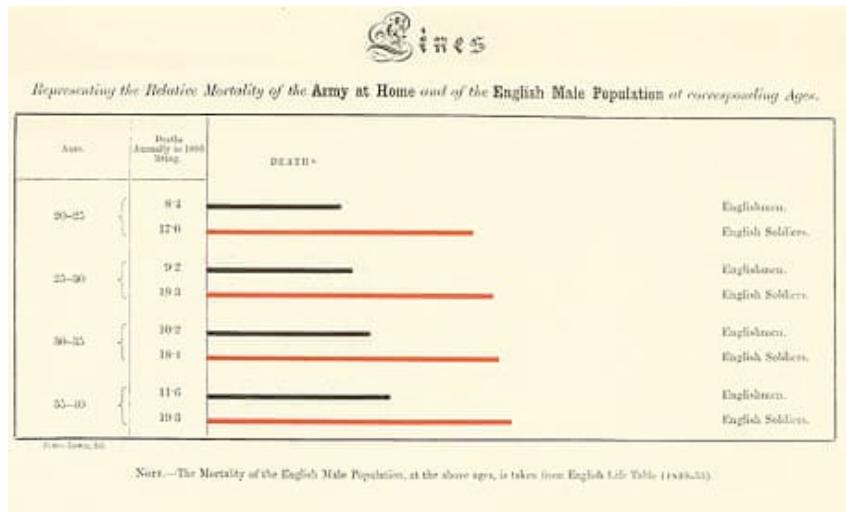


Figure 1.13: Gráfico de barras de Florence Nightingale mostrando as diferenças de mortalidade entre soldados britânicos e a população masculina inglesa geral (civis)

- Niccolò Fontana Tartaglia (Veneza à época, hoje Itália: 1499-1557)
- Girolamo Cardano (Pávia à época, hoje Itália: 1501-1576)
- Galileu Galilei (Florença 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)

## 1.4 Revista Biometrika

“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](#)).

## 1.5 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, significa bem-nascido.

Este livro caiu em domínio público e pode ser lido na íntegra online. A caracterização original de Eugenia 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)

Há poucos anos alguns grupos sociais viram no trabalho e opiniões de Fisher endossos ao colonialismo, à supremacia branca e à eugenia.

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

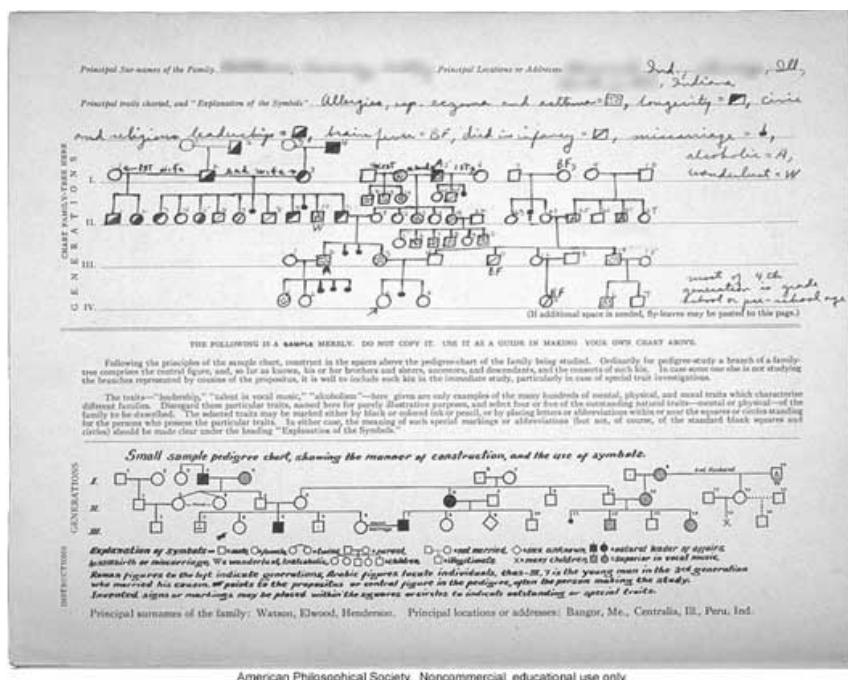


Figure 1.14: Gráfico de linhagens para alergias

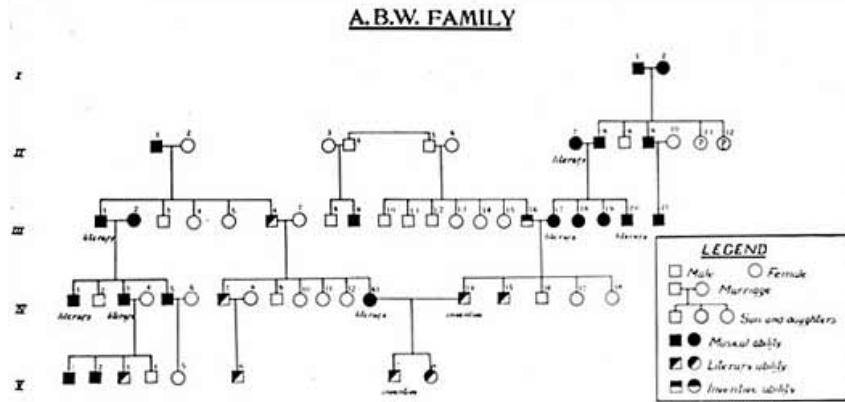


Figure 1.15: Gráfico de linhagens para aptidão musical

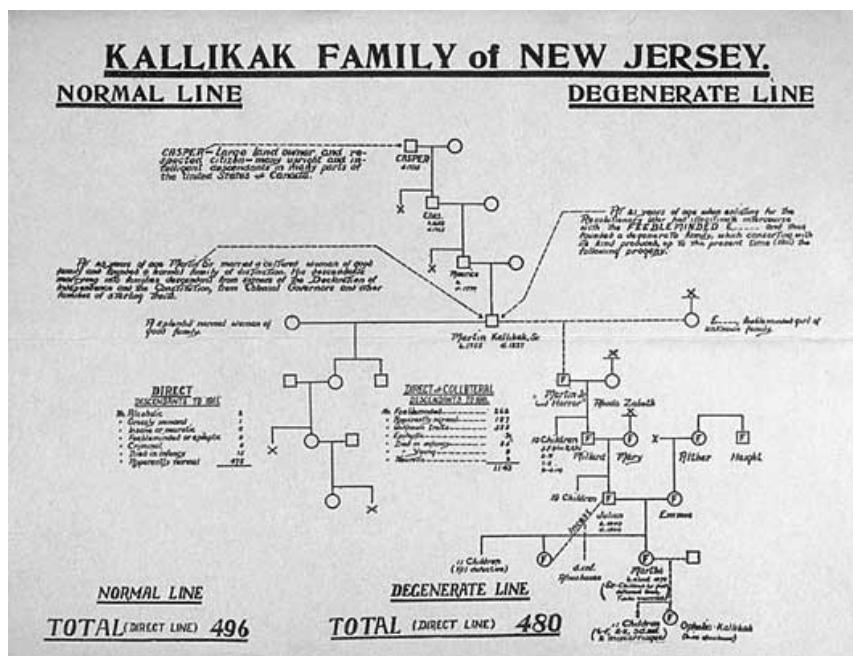


Figure 1.16: Linhas "normais" e "degeneradas" da família Kallikak (New Jersey)

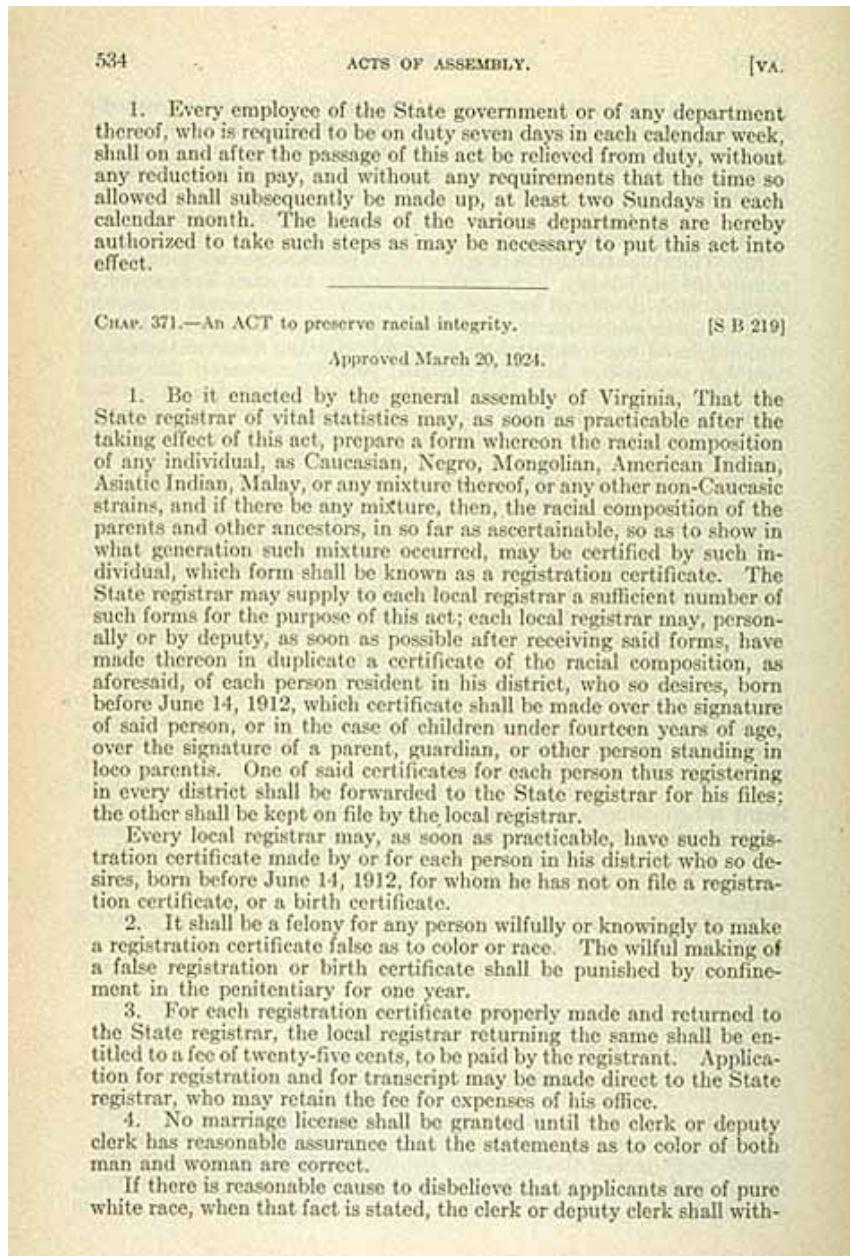


Figure 1.17: Lei da Inegridade Racia (Virginia, EUA, 1924)



Figure 1.18: Licença para casamento

## Módulo 2

# Introdução conceitual essencial

“Estatística é um conjunto de métodos que se destina a possibilitar a tomada de decisões, face às incertezas[...]'”

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

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 como, por exemplo:

(A)

- Hipótese (ideia, teoria, conjectura): “Hoje será um dia como outro qualquer.”
- Dedução: “Meu carro estará estacionado na garagem, no local de costume.”
- Dados (informação, fatos): “Meu carro não está lá!”
- Inferência: “Alguém deve tê-lo levado.”

(B)

- Hipótese (ideia, teoria, conjectura): “Meu carro foi roubado!”
- Dedução: “Meu carro não estará no local de costume.”

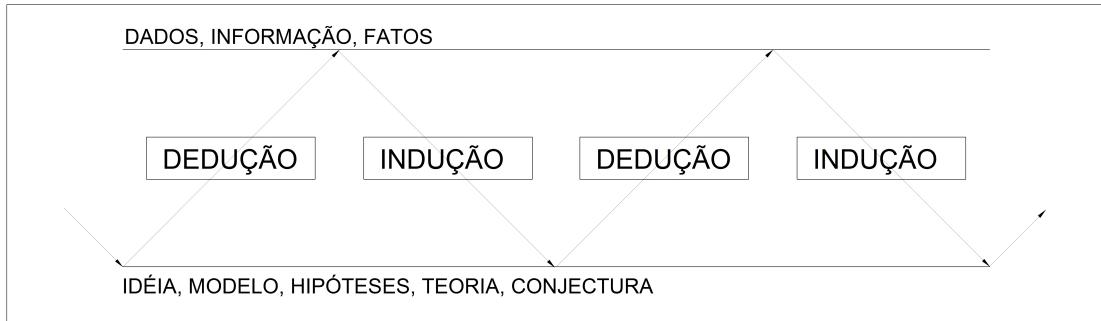


Figure 2.1: Fluxograma elementar de um processo de aprendizagem

- Dados (informação, fatos): “Meu carro está lá!”
- Inferência: “Alguém deve tê-lo levado e devolvido.”

(C)

- Hipótese (ideia, teoria, conjectura): “Um ladrão pegou e trouxe de volta.”
- Dedução: “Meu carro foi arrombado.”
- Dados (informação, fatos): “Meu carro está intacto e o alarme está desligado.”
- Inferência: “Alguém que tenha as chaves deve tê-lo levado.”

(D)

- Hipótese (ideia, teoria, conjectura): “Minha esposa usou meu carro.”
- Dedução: “Ela provavelmente deixou um bilhete.”
- Dados (informação, fatos): “Sim, aqui está o bilhete.”
- Inferência: “Minha hipótese estava correta.”

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;
- custo;
- precisão.

## 2.4 População (universo) & amostra



Figure 2.2: 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 poúlação (universo) e letras do alfabeto latino para aquelas relativas à amostra extraída:

Característica estudada	Notação populacional	Notação amostral
Número de elementos	$N$	$n$
Média	$\mu$	$\bar{x}$
Variância	$\sigma^2$	$s^2$
Desvio padrão	$\sigma$	$s$
Proporção	$\Pi$	$p$

## 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.

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

Figure 2.3: Alfabeto grego

## 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.

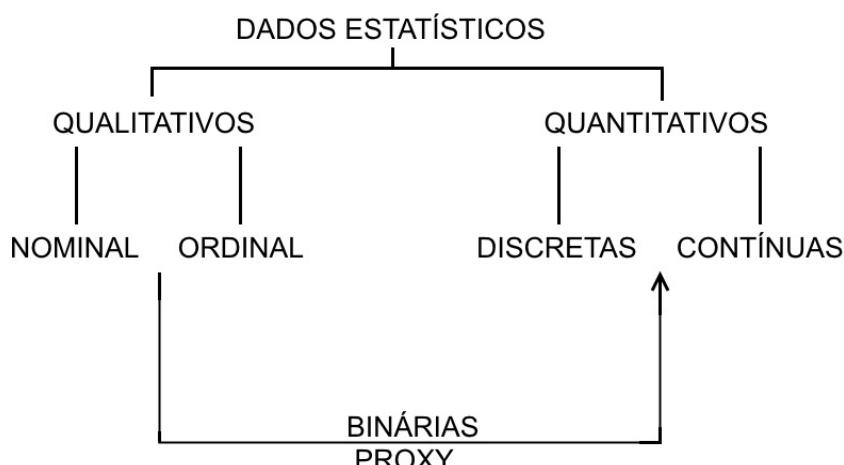


Figure 2.4: Tipos e codificações de variáveis

## 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 separatiz), cada um de seus valores estaria a ocupar uma *posição* indicada pelo índice  $i$ :

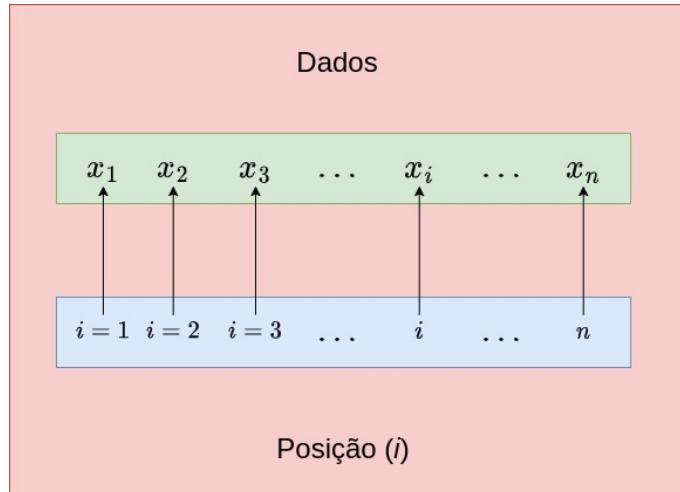


Figure 2.5: Entendendo a indexação de dados

## 2.8 Noções básicas sobre somatórios ( $\Sigma$ )

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

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:

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

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  $\Sigma$  (**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  $\Sigma$  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  $\Sigma$  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.

## 2.9 Análise combinatória: diagramas de árvore, permutações (arranjos) & combinações

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

Se um evento  $E$  pode ser decomposto em eventos sequenciais  $E_1, E_2, E_3, \dots, E_n$  e existem  $P_1$  possibilidades distintas de ocorrer  $E_1$ ,  $P_2$  possibilidades distintas de ocorrer  $E_2$  e assim sucessivamente, então o número total de possibilidades do evento  $E$  ocorrer é dado por:

$$P_1 \cdot P_2 \cdot \dots \cdot P_n$$

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.

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

## 2.9. ANÁLISE COMBINATÓRIA: DIAGRAMAS DE ÁRVORE, PERMUTAÇÕES (ARRANJOS) & COMBINAÇÕES

Um diagrama como ilustrado na Figura 2.6 (denominado **diagrama de árvore** em virtude de sua aparência) geralmente é usado para explicar o princípio acima

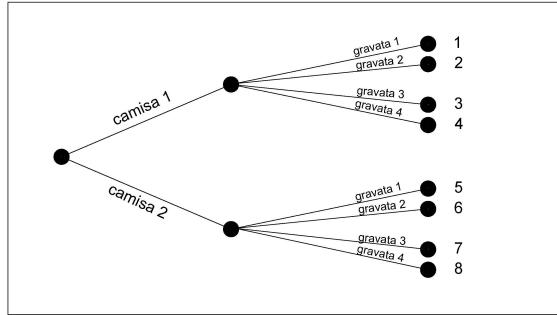
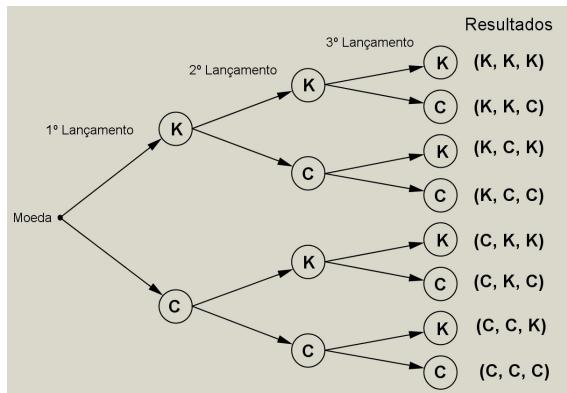


Figure 2.6: Diagrama de árvore

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 2.7):



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.

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:

$$n_1 + n_2 = 5$$

### 2.9.1 Permutações ou arranjos

O conceito de uma permutação (arranjo) refere-se a uma relação de  $n$  objetos distintos que serão agrupados  $p$   $p$  ( $p < n$ ). Nos agrupamentos possíveis considera-se a ordem dos elementos; sendo assim, qualquer mudança na ordem dos elementos em um agrupamento constitui um novo agrupamento: **agrupamentos que possuem os mesmos objetos em ordem distinta são considerados agrupamentos distintos**.

- Simples: não ocorre a repetição de um elemento no agrupamento; e,
- Com repetição: os elementos que compõem o conjunto podem aparecer repetidos; ou seja, um agrupamento pode apresentar elementos iguais.

O número de permutações (arranjos) **sem a repetição** de um mesmo elemento no agrupamento, formados por  $p$  elementos selecionados de um conjunto de  $n$  objetos distintos será:

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

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 **sem repetição**?

## 2.9. ANÁLISE COMBINATÓRIA: DIAGRAMAS DE ÁRVORE, PERMUTAÇÕES (ARRANJOS) & COMBINAÇÕES

$$n = 7$$

$$p = 3$$

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

O número de permutações (arranjos) **com repetição** de um mesmo elemento no agrupamento, formados por  $p$  elementos selecionados de um conjunto de  $n$  objetos distintos será:

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

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**?

$$n = 7$$

$$p = 3$$

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

### 2.9.2 Combinações

Em uma *permutação* consideramos que a **ordem\*** que os objetos assumem nos agrupamentos os tornam diferentes uns dos outros. Por exemplo, abc\*\* é uma agrupamento distinto de bca numa permutação.

Em muitos problemas, entretanto, estamos interessados somente na seleção ou escolha dos objetos \*\*sem que a ordem assumida pelos objetos nos agrupamentos os tornem diferentes uns dos outros\*.

Tais seleções são chamadas de *combinações*. Por exemplo, **abc** e **bca** são consideradas uma mesma combinação.

O conceito de uma combinação refere-se a uma relação de  $n$  objetos distintos que serão agrupados  $p$  a  $p$  ( $p < n$ ) sem repetição de qualquer objeto em um mesmo agrupamento. Os agrupamentos que possuem os mesmos objetos em ordem diferente **não são considerados agrupamentos distintos**.

- Simples: não ocorre a repetição de elementos no agrupamento; e,
- Com repetição: os elementos que compõem o agrupamento podem aparecer repetidos; ou seja, ocorre a repetição de um mesmo elemento em um agrupamento.

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

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

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

$$n = 8$$

$$p = 3$$

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

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: 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).

$$n = 3$$

$$p = 4$$

$$C_{(n+p-1,p)} = \frac{(3+4-1)!}{4!(+3-1)!} = 15$$

### 2.9.3 Observações acerca de alguns fatoriais

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

## 2.10 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  $\sim$  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.

## 2.11 Leis de De Morgan

Augustus de Morgan foi um matemático e lógico indiano.



Figure 2.8: 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”’ (ou seja, transformá-las em uma disjunção). Considerando as proposições “p” e “q” teremos:

- $\sim(p \cap q) = (\sim p) \cup (\sim q)$ ; 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:

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

## 2.12 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**.

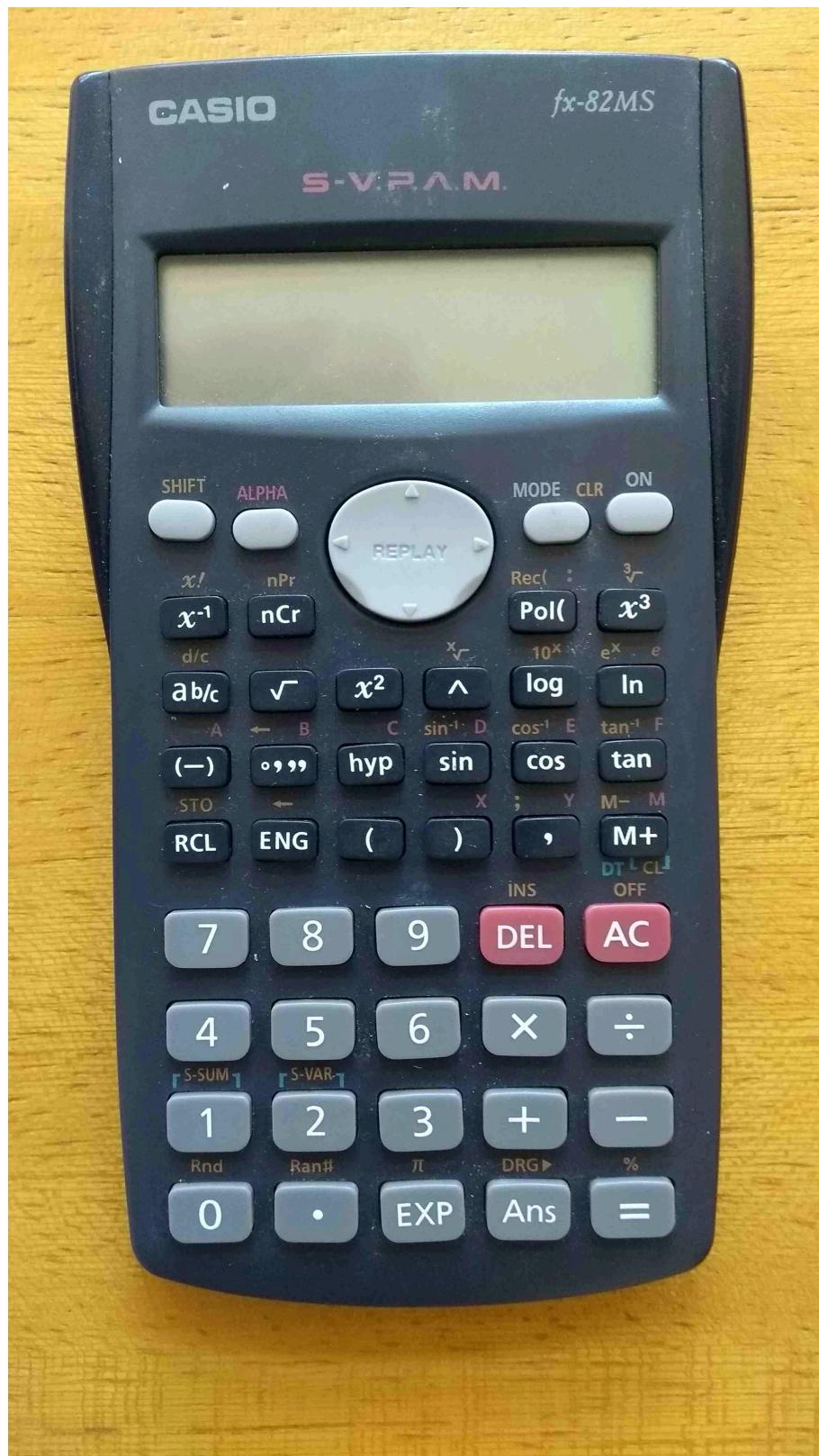


Figure 2.9: Calculadora Cassio

## Módulo 3

# Introdução à estatística descritiva

Do prefácio da tradução do livro de Jack Levin (Estatística aplicada às ciências humanas), Sérgio Francisco Costa diz 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 1.66 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 | 11223333345667889
##    18 | 233456
##    19 | 35
##    20 | 0
```

À esquerda do traço vertical (os ramos) são apresentadas frações das medidas das alturas (no caso, décimetros) 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",
           pch=20, at=0.5,
           main="Gráfico de dispersão unidimensional",
           col="blue",cex=1,
           xlab="Alturas dos estudantes (m)",
           ylab="Quantidades observadas (un)")
```

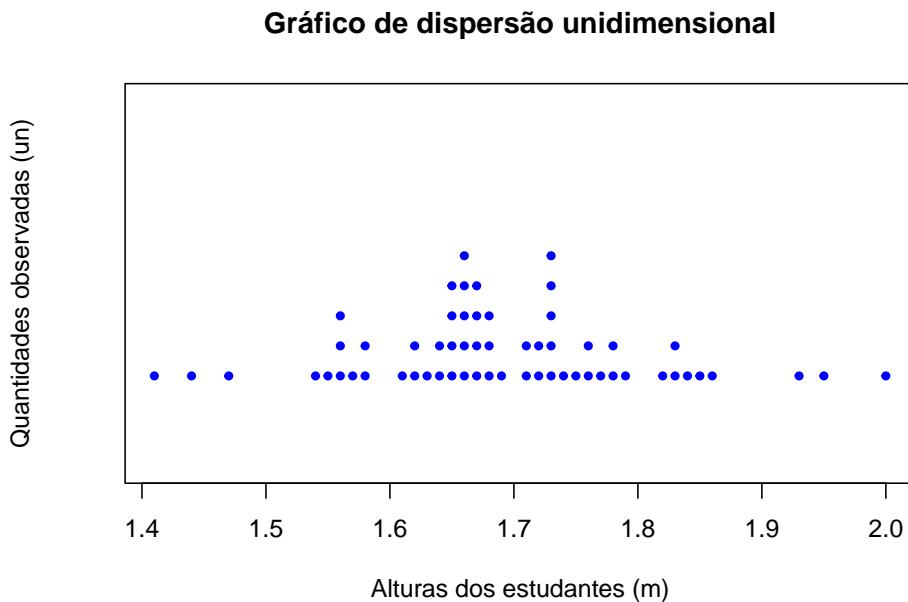


Figure 3.2: Gráfico de dispersão unidimensional (stripchart)

### 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:

- 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)
```

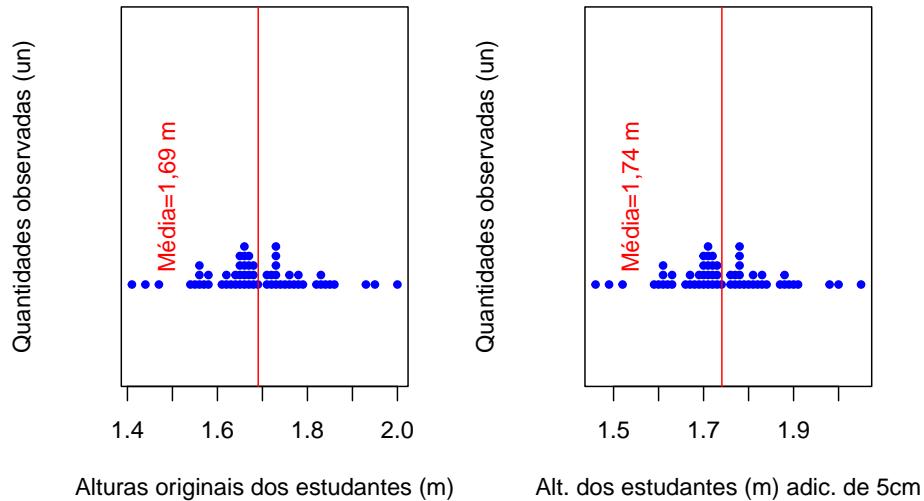


Figure 3.3: Mudanças na média pela adição (subtração) de uma constante  $k = 0.05$

- 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)
```

- 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.

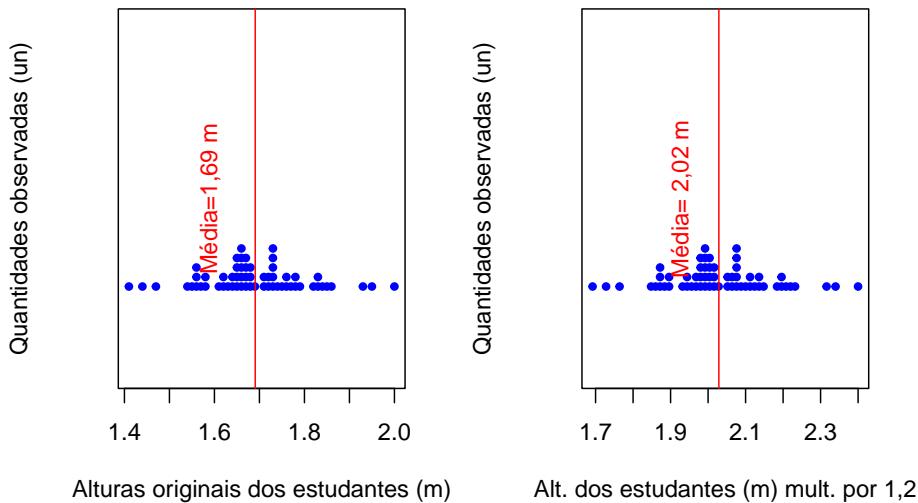


Figure 3.4: Mudanças na média pela multiplicação (divisão) de uma constante  $k = 1.2$

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   3   1   2   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)
```

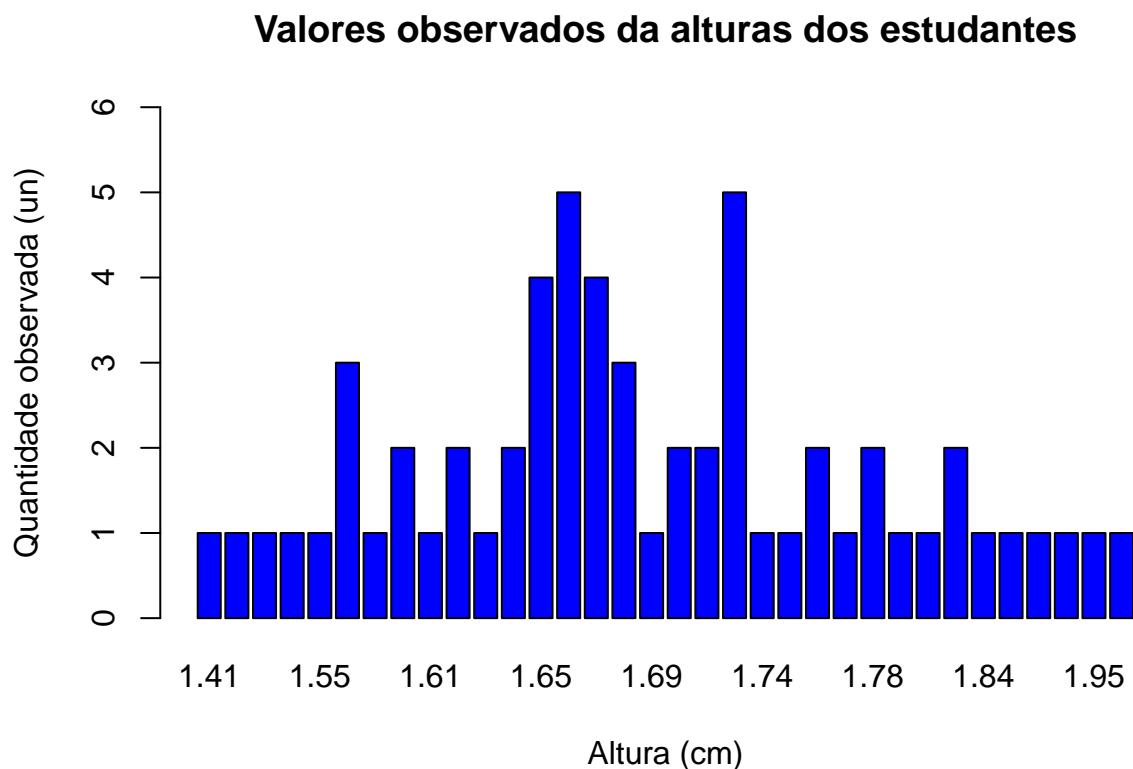


Figure 3.5: Bimodal: 1,66 m e 1,73 m

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

```
# 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.

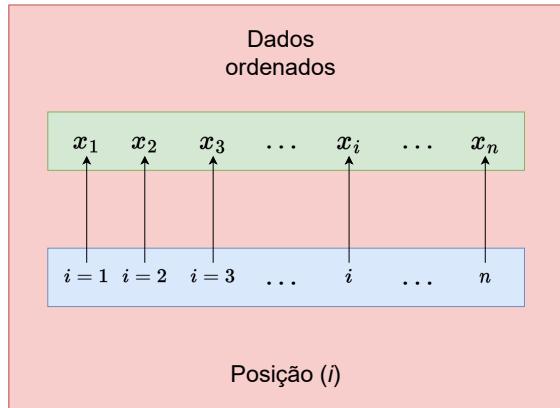


Figure 3.6: Entendendo a indexação de dados

Por essa razão, a mediana é uma medida 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, 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.

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

$$Md = l_{inf} + \left[ \frac{\left( \frac{n}{2} - F_{(i_{md}-1)} \right)}{n_{md}} \right] \times \Delta_i$$

onde:

-  $l_{inf}$ : limite inferior da **classe mediana**: a classe que contém o elemento de ordem  $\frac{n}{2}$ ; -  $F_{(i_{md}-1)}$ : é a frequência absoluta acumulada até a **classe anterior à classe mediana**; -  $n_{md}$ : é a frequência absoluta da **classe mediana**; e, -  $\Delta_i$ : é o intervalo de cada classe.

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 1.65 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)

```

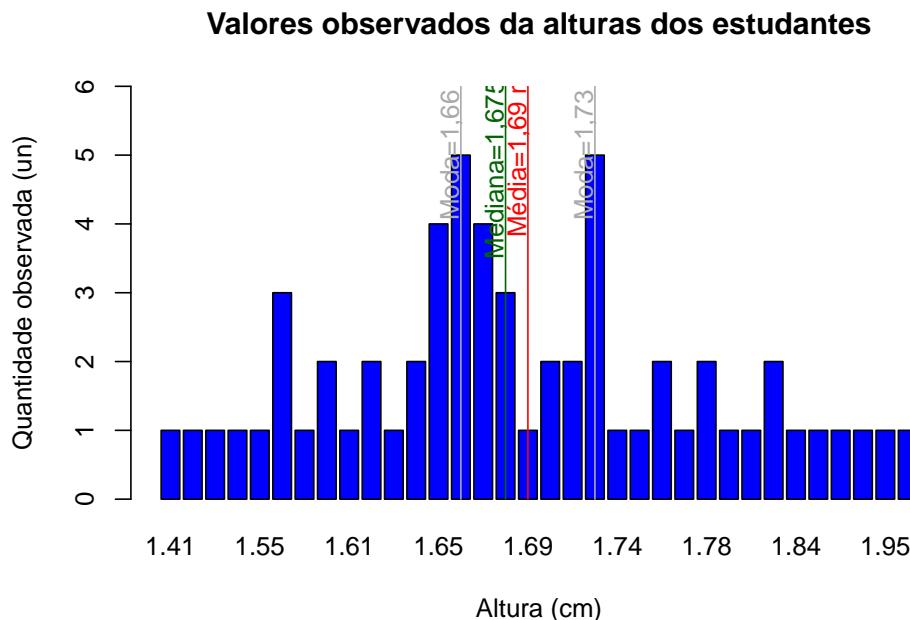


Figure 3.7: Valores observados das alturas dos estudantes e as posições da média, moda e mediana

### 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)

### 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.

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  $\nu$ . A perda de *um* grau de liberdade deve-se à necessidade de se substituir a média populacional desconhecida ( $\mu$ ) 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  $(n - 1)\}$ .

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).

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 1.65 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. Os quantis mais expressivos 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.

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  $q_p$ . Um importante gráfico que mais adiante será exposto em detalhes é o *Boxplot* que, além da mediana, para sua confecção necessitamos de duas outras separatrizes: o 1º e 3º quartis. \

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 desresível. De modo geral, para se calcular a posição  $L$  de um quantil qualquer de ordem  $p$  em um rol de dados pode-se usar a seguinte regra empírica:

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

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.

Assim, para a determinação dos quartis, o valor de  $p$  seria:

- para o *primeiro quartil* ( $Q_1$ ):  $L_{q_{0,25}} = \frac{25}{100} \times (n + 1)$ ;
- para o *segundo quartil* (a mediana ou  $Q_2$ ):  $L_{q_{0,50}} = \frac{50}{100} \times (n + 1)$ ; ou,

- para o *terceiro quartil* ( $Q_3$ ):  $L_{q_{0,75}} = \frac{75}{100} \times (n + 1)$ .

Novamente podemos nos deparar com **duas situações possíveis** para o valor calculado para a posição  $L$ :

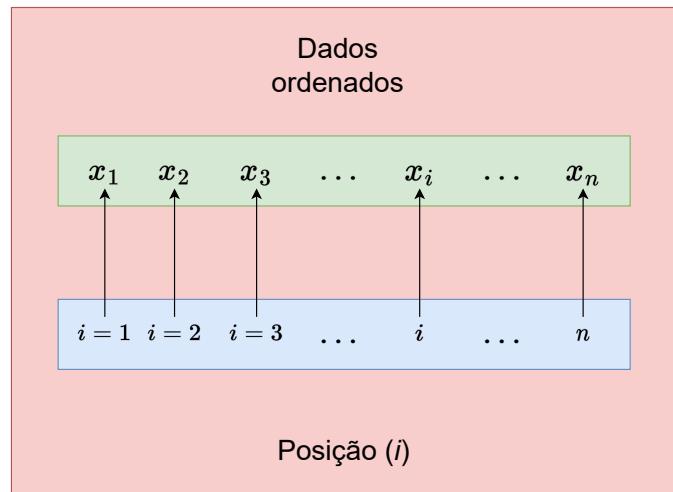


Figure 3.9: Entendendo a indexação de dados

- se valor calculado da **posição L** for um inteiro, essa será a posição onde 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  $1^o$ ,  $2^o$  e  $3^o$  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):

- 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*:

$$\begin{aligned}
 L_{Q_1} &= \frac{p}{100} \times (n + 1) \\
 &= \frac{25}{100} \times (60 + 1) \\
 &= 0,25 * 61 \\
 &= 15,25
 \end{aligned}$$

- para o *segundo quartil*:

$$\begin{aligned}
 L_{Q_2} &= \frac{p}{100} \times (n + 1) \\
 &= \frac{50}{100} \times (60 + 1) \\
 &= 0,5 * 61 \\
 &= 30,5
 \end{aligned}$$

- para o *terceiro quartil*:

$$\begin{aligned}
 L_{Q_3} &= \frac{p}{100} \times (n + 1) \\
 &= \frac{75}{100} \times (60 + 1) \\
 &= 0,75 * 61 \\
 &= 45,75
 \end{aligned}$$

E os quartis serão:

$$-Q_1=1,635 \quad -Q_2=1,675 \quad -Q_3=1,755$$

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

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 *platocúrticas* 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.

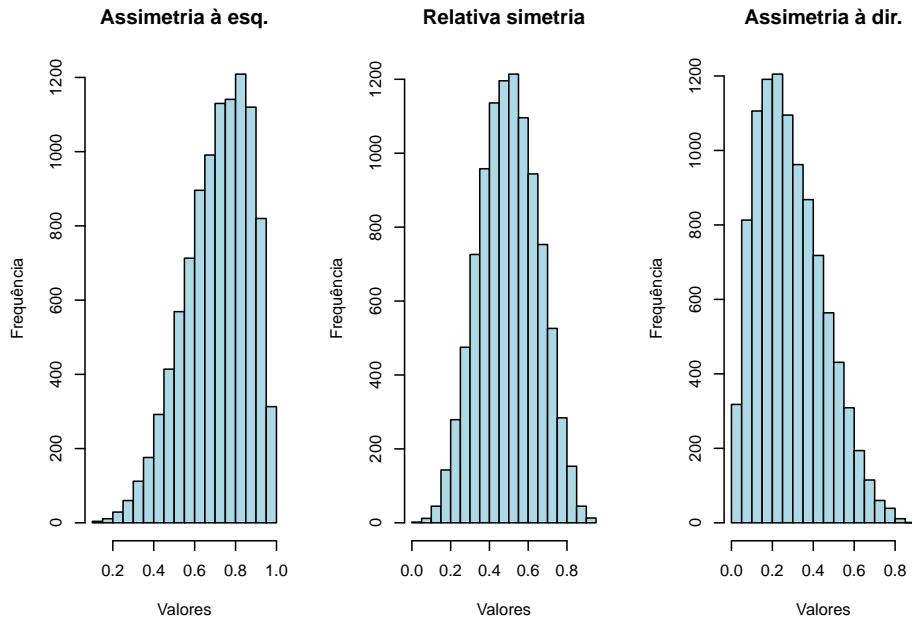


Figure 3.10: Diferentes formas na distribuição dos dados

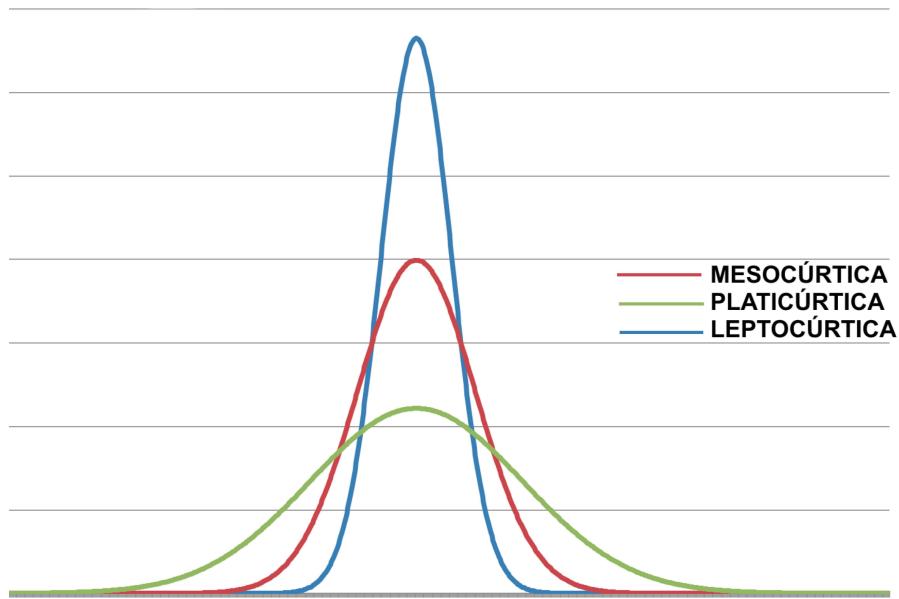


Figure 3.11: Diferentes aspectos de uma distribuição quanto à sua inclinação

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

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:

- $AS=0$ : distribuição simétrica;
- $AS<0$ : distribuição com assimetria negativa; e,
- $AS>0$ : distribuição com assimetria positiva.

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

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

Onde:

- $Q_3 = 3^o$  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 1.65 1.66 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 1.78 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

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 continente 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

A partir desse resumo, poderíamos eleborar a apresentação gráfica desses dados na forma de um *Gráfico de colunas* ou um *Gráfico de setores*.

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

barplot(tab_desembarque,
        main="Fig. 01: 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",
        xlab="",
        ylab="Quantidade observada (un)",
        ylim=c(0,6),
        col="blue",
        las=0,
        hor="FALSE")
```

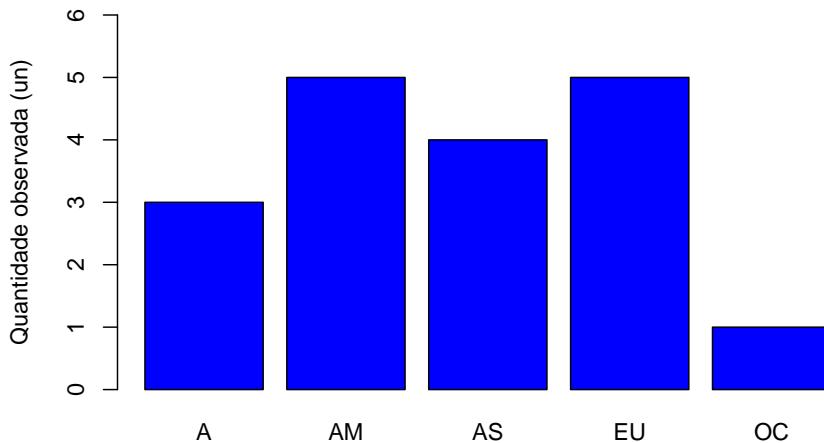
ou ainda assim:

```
library(scales)
library(ggplot2)

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

blank_theme=theme_minimal()+
```

**Fig. 01: Desembarques no terminal internacional A em Cumbica  
(10/10/2021: 8 h 00min às 12 h 00 min)**



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

Figure 3.12: Gráfico de barras dos dados observados no terminal de desembarque internacional do aeroporto

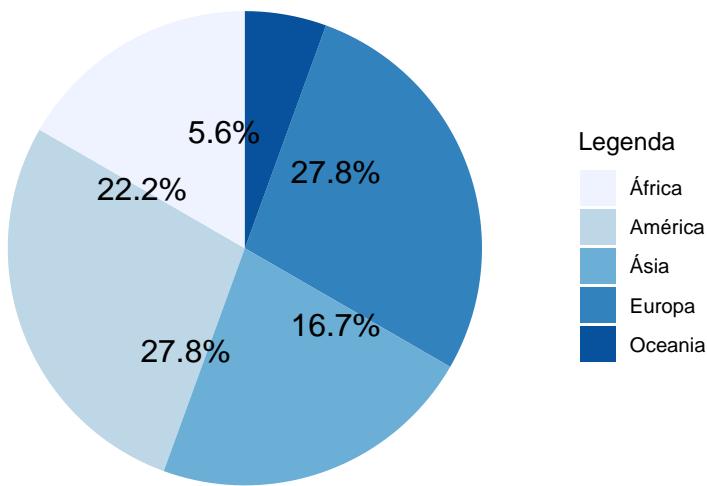
```
theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.border = element_blank(),
  panel.grid=element_blank(),
  axis.ticks = element_blank(),
  plot.title=element_text(size=14, face="bold")
)

ggplot(desembarques_classes, aes(x="", y=value, fill=group)) +
  blank_theme +
  scale_fill_brewer("Blues")+
  labs(title="Fig. 01: 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))
```

Outro exemplo de apresentação tabular onde são apresentadas as proporções 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.

**Fig. 01: Desembarques no terminal internacional A em Cunh**

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



Fonte: próprio autor

Figure 3.13: Gráfico de setores das alturas dos estudantes

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: Censo dos

EUA (2005)

Igualmente, a apresentação gráfica desses dados pode ser feita, por exemplo, por um *Gráfico de colunas* ou um *Gráfico de setores*.

```
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")
```

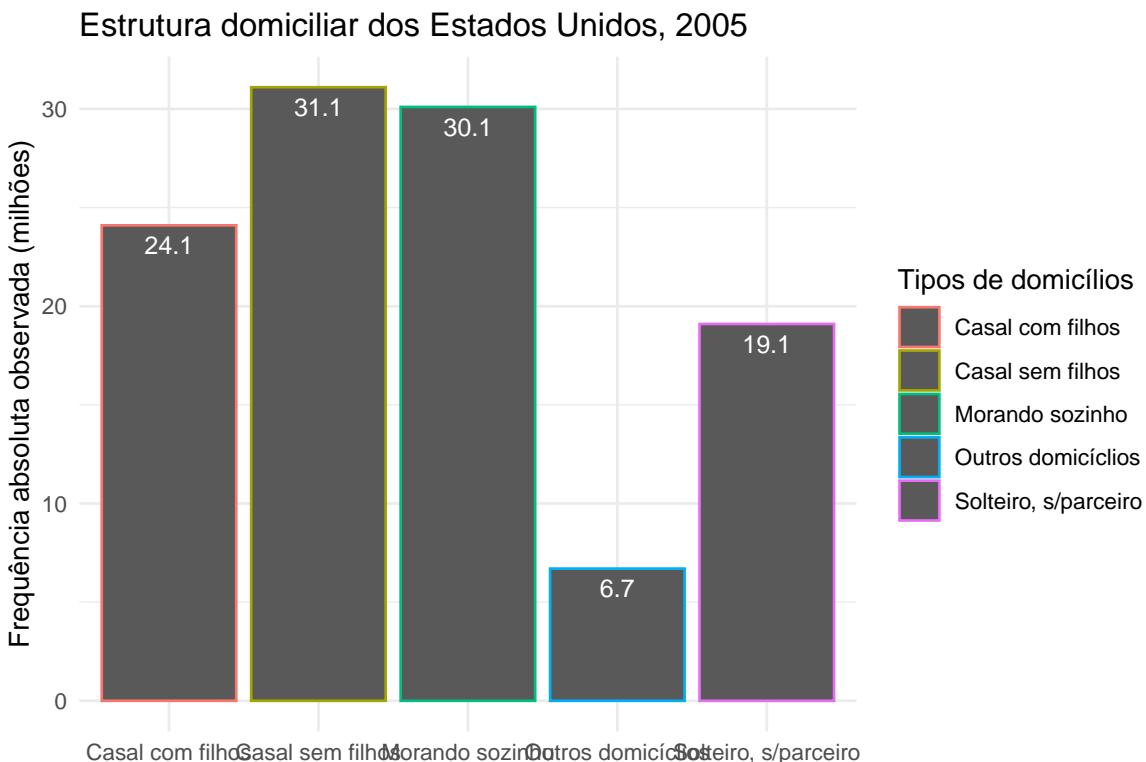


Figure 3.14: Gráfico de barras

```
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)
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

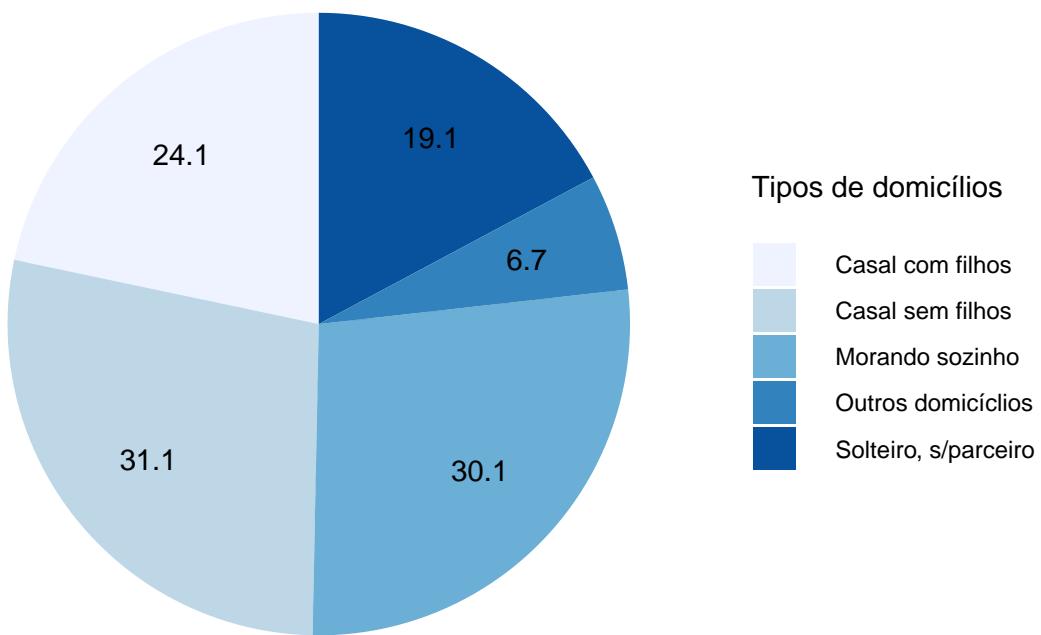


Figure 3.15: Gráfico de setores

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.

Table 3.4: 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, uma tabela com frequências relativas pode ser construída:

Table 3.5: 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

|

As representações gráficas são análogas às mostradas no exemplo anterior.

### 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	$25 \leq n \leq 220$	$k = \sqrt{n}$
Herbert Sturges	$135 \leq 572237$	$k = 1 + 3,3\log(n)^{(1)}$
Giuseppe Milone	$20 \leq 36315$	$k = -1 + 2\ln(n)^{(2)}$

- <sup>(1)</sup>: logarítmico na base 10; e
- <sup>(2)</sup>: 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.

Em nosso exemplo das alturas dos estudantes, a determinação do número de classes pelo critério da *raiz quadrada* ( $n=60$ ) sugere 8 classes (pelo critérios de Sturges  $k = 6,86 \sim 7$  e 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**:  $c = 0,08m$ .

As classes são então assim construídas:

- Limite inferior da 1<sup>a</sup> classe ( $LI_1$ ): valor mínimo observado; e,
- Limite superior da 1<sup>a</sup> 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\***:

$\dashv \circ - \bullet$

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  $c = 0,08m$  as classes ficam assim estabelecidas, tendo-se como ponto de partida o valor mínimo observado: 1,41 m - 1,49 m; 1,49 m - 1,57 m; 1,57 m - 1,65 m; 1,65 m - 1,73 m; 1,73 m - 1,81 m; 1,81 m - 1,89 m; 1,89 m - 1,97m; 1,97 - 2,05 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 ; 1,65 ; 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 ; 1,69 ; 1,71 ; 1,71 ; 1,72 ; 1,72 ; 1,73 ; 1,73 ; 1,73 ; 1,73 ; 1,73 ; 1,73 ; 1,74 ; 1,75 ; 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 assume a forma:

Classe	Frequência absoluta ( $f_i$ )
1,41 m $\leftarrow$ 1,57 m	8
1,57 m $\leftarrow$ 1,65 m	9
1,65 m $\leftarrow$ 1,73 m	21
1,73 m $\leftarrow$ 1,81 m	13
1,81 m $\leftarrow$ 1,89 m	6
1,89 m $\leftarrow$ 1,97 m	2
1,97 m $\leftarrow$ 2,05 m	1
Total	60

Alternativamente, caso adotássemos como ponto de partida (um pouco abaixo do valor mínimo observado) o valor de 1,40 m, a distribuição das classes seria: 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 ; 1,62 ; 1,63 ; 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 ; 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 ; 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 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 necessitaímos abrir mais uma classe apenas para incluí-lo.

Mas começando-se no valor mínimo observado (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 ; 1,56 ; 1,57 ; 1,58 ; 1,58 ; 1,61 ; 1,62 ; 1,62 ; 1,63 ; 1,64 ; 1,64 ; 1,65 ; 1,65 ; 1,65 ; 1,65 ; 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 ; 1,73 ; 1,74 ; 1,74 ; 1,75 ; 1,76 ; 1,76 ; 1,77 ; 1,78 ; 1,78 ; 1,79 ; 1,82 ; 1,83 ; 1,83 ; 1,84 ; 1,84 ; 1,85 ; 1,86 ; 1,93 ; 1,95 ; 2,00}

A tabela de distribuição de frequências com 6 classes assume a forma:

Classe	Frequência absoluta ( $f_i$ )
1,41 m - 1,51 m	3
1,51 m - 1,61 m	8
1,61 m - 1,71 m	23
1,71 m - 1,81 m	17
1,81 m - 1,91 m	6
1,91 m - 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 ( $\Delta_i$ ) 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 ( $x_i$ ) 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 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 ( $\Delta_i$ );
- Densidade relativa  $\delta_{fr_i} = \frac{fr_i}{\Delta_i}$ : o quociente da frequência relativa ( $fr_i$ ) pela amplitude ( $\Delta_i$ ) da classe.

Vejo como exemplo as tabelas abaixo:

Classe	Int. de valores	Alt. média	Freq. abs.	Freq. rel.	Freq. rel. (%)	Freq. acumulada	Freq. acum. (%)
1	1,41 ⊴ 1,51	1,46	( $f_i$ ) 3	( $fr_i$ ) 0,05	( $fr_i\%$ ) 5	( $fac_i$ ) 3	( $fac_i\%$ ) 5,00
2	1,51 ⊴ 1,61	1,56	8	0,13	13,33	11	18,33
3	1,61 ⊴ 1,71	1,66	23	0,38	38,33	34	56,66
4	1,71 ⊴ 1,81	1,76	17	0,28	28,34	51	85,00
5	1,81 ⊴ 1,91	1,86	6	0,10	10	57	95,00
6	1,91 ⊴ 2,01	1,96	3	0,05	5	60	100,00
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	( $\Delta_i$ ) 0,10	( $\delta_i$ ) 30	( $fr_i$ ) 0,05	( $\delta_{fr_i}$ ) 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:

$$\bar{x} = \frac{\sum_{i=1}^n n_i \cdot \bar{x}_i}{N}$$

### 3.5.4 Moda

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

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

onde:

- $l_{inf}$ : limite inferior da classe modal, **a classe de maior frequência absoluta**;
- $\Delta_1$  frequência absoluta da **classe modal** menos a frequência absoluta da **classe anterior**;
- $\Delta_2$  frequência absoluta da **classe modal** menos a frequência absoluta da **classe posterior**;
- $e$ ,
- $\Delta_i$  é o intervalo de cada classe.

### 3.5.5 Variância

Variância para dados agrupados:

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

Onde:

- $n_i$  é a frequência absoluta em cada classe  $i$ ; e,
- $\bar{x}_i$  é o valor médio de cada classe  $i$ .

## 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). Alguns dos tipos mais comuns são:

### 1. qualitativas

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

### 2. quantitativas

- ranking: barras;
- parte em relação ao todo: setores;
- dispersão unidimensional;
- distribuição: histograma e o *box plot*;
- correlação: pontos dispersos;
- *heat maps*; e,
- tendência: linha

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*.

### 3.6.1 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")
```

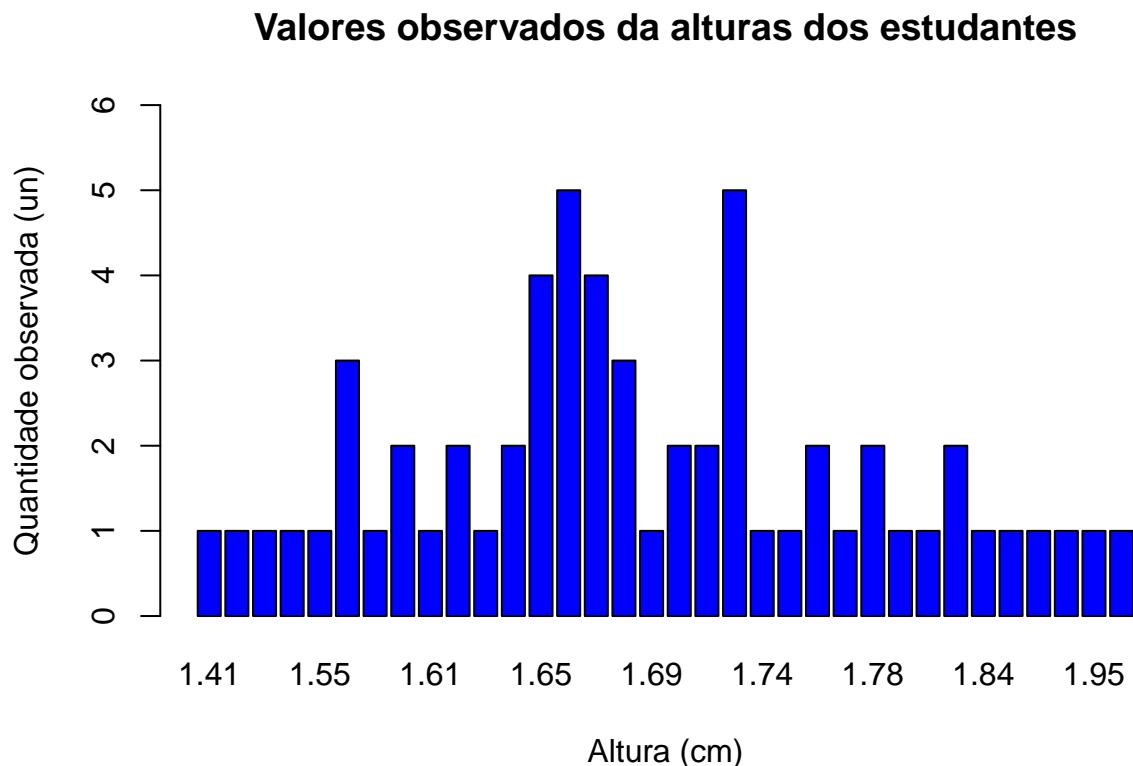


Figure 3.16: Gráfico de barras dos dados brutos: uma barra para cada observação e sua altura expressando o número de observações com esse valor

### 3.6.2 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:

```

h1=hist(alturas, breaks=seq(1.41 , 2.01 , 0.1), main= "Histograma das alturas dos estudantes", col
xlab="Classes de comprimento (cm)", ylab="Frequência absoluta observada (un)" , cex=0.7, ylim=c(0,
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)

```

**Histograma das alturas dos estudantes**

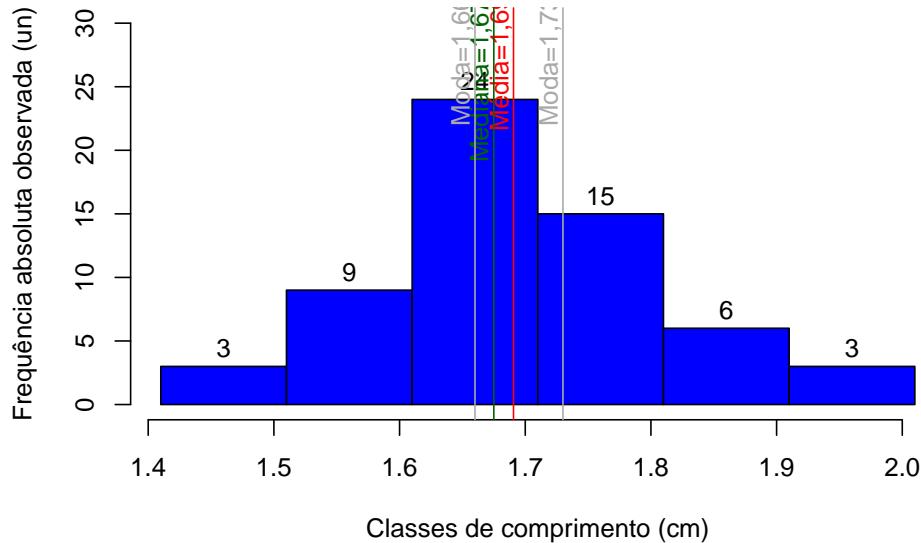


Figure 3.17: Histograma das alturas dos estudantes com as posições da média, moda e mediana

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*.

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

Num histograma de densidade, a altura de cada retângulo representa a densidade da ocorrência da *frequência relativa*.

```
h2=hist(alturas,breaks=seq(1.41 , 2.01 , 0.10), main= "Histograma das alturas dos estudantes", col="blue")
xlab="Classes de alturas (m)", ylab="Densidade da freq. relativa", prob="TRUE", ylim=c(0,5)
text(h2$midas,h2$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
```

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

- Base =  $\Delta_i$ ; e,\
- Altura =  $\frac{fr_i}{\Delta_i}$ .

### Histograma das alturas dos estudantes

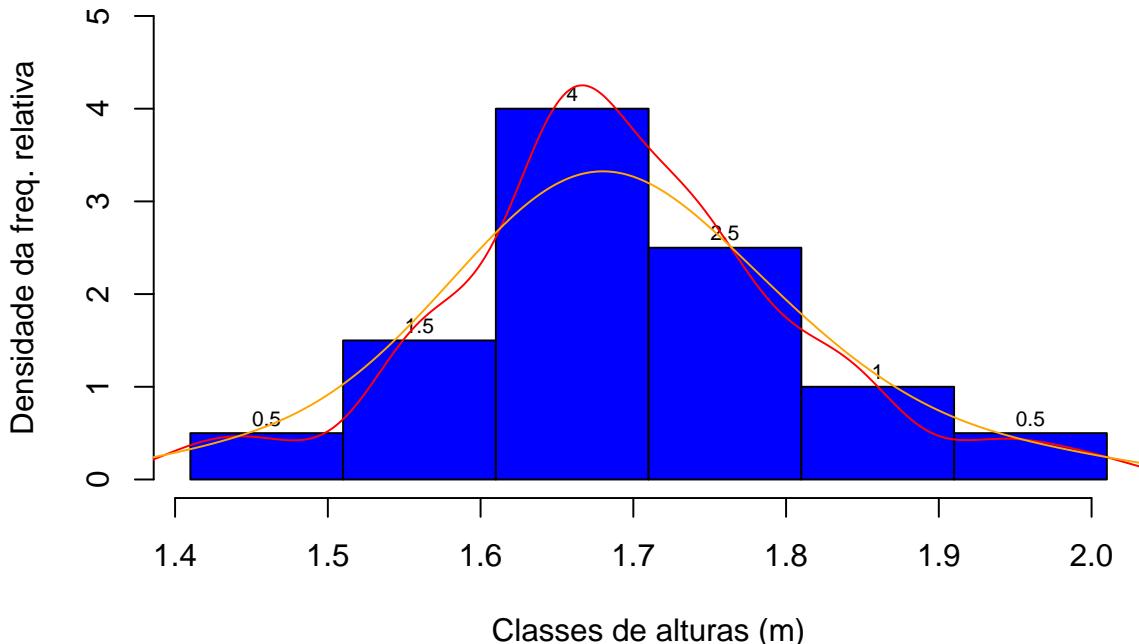


Figure 3.18: A linha vermelha é uma aproximação da Função de Densidade da frequêcia relativa de observação (a linha preta é a curva da função densidade de uma distribuição Normal com média e variâncias dadas pelos dados)

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%)

#### 3.6.3 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 ( $360^\circ$ ).

```
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))

```

### 3.6.4 Box-plot

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).

## Alturas dos estudantes

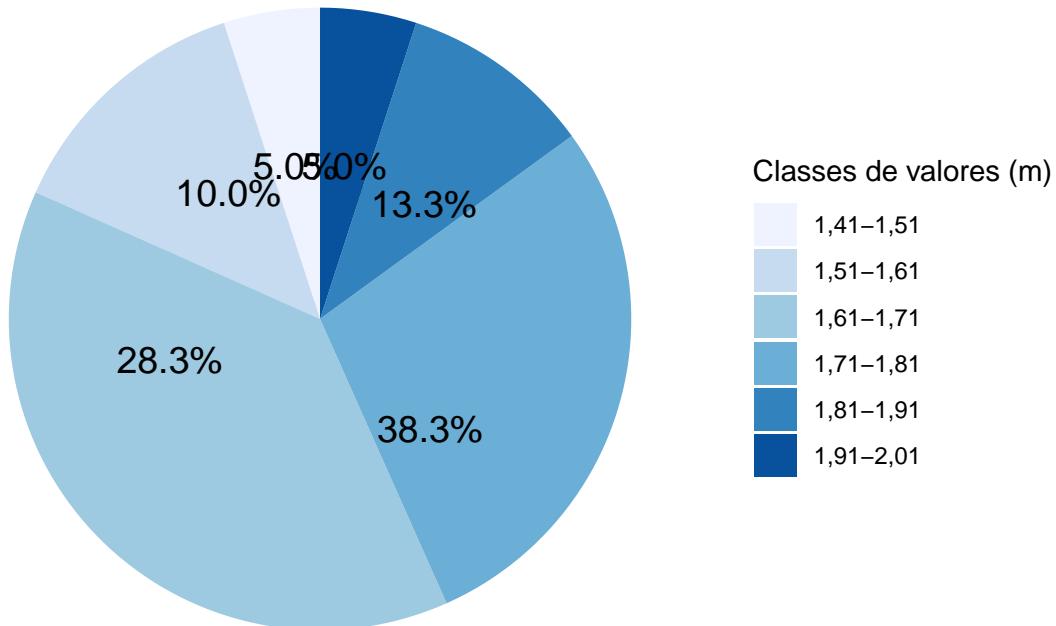


Figure 3.19: Gráfico de setores das alturas dos estudantes

```

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 \n(média=1,69 ; mediana=1,675 ; min=1,41 ; max=2,01)", ylim=c(1.3, 2.1))
lines( y=c(min, min), x=c(0.6,1), col="red")
text(x=0.60, y=min-0.05, "Valor mínimo observado", col = "red", srt=0)
lines( y=c(max,max), x=c(0.6,1), col="red")
text(x=0.60, y=max+0.05, "Valor máximo observado", col = "red", srt=0)

```

```

lines(y=c(med, med), x=c(1,1.4), col="red")
text(x=1.4 , y= med+0.02 , "Média", col = "red", srt=0)
lines(y=c(q1, q1), x=c(1, 1.4), col="red")
text(x=1.4, y=q1 -0.05, "Primeiro quartil (Q1)", col = "red", srt=0)
lines(y=c(q2, q2), x=c(0.6,1), col="red")
text(x=0.60 , y= q2 - 0.05, "Mediana (Q2)", col = "red", srt=0)
lines(y=c(q3, q3), x=c(1, 1.4), col="red")
text(x= 1.4 , y=q3 + 0.05, "Terceiro quartil (Q3)", col = "red", srt=0)
lines(y=c(li,li) , x=c(1.01,1.4) , col="blue", lty=2)
text(x=1.2, y=q1-1.5*iq-0.05 , "Limite inferior teórico (LI=1,455) ", col = "blue", srt=0)
lines(y=c(ls,ls) , x=c(1.01,1.4) , col="blue", lty=2)
text(x=1.2, y=q3+1.5*iq +0.05 , "Limite superior teórico (LS=1,935)" , col = "blue", srt=0)
lines(y=c(1.47,1.47) , x=c(0.85,0.95) , col="green", lty=2)
text(x=0.7, y=1.47 , "Última observação dentro do LI (x=1,47) " , col = "green", srt=0)
lines(y=c(1.93,1.93) , x=c(0.85,0.95) , col="green", lty=2)
text(x=0.7, y=1.93 , "Última observação dentro do LS (x=1,86) " , col = "green", srt=0)

```

**Boxplot do conjunto de dados de alturas**  
 median=1,675 ; min=1,41 ; máx=2,00 ; Q1=1,635 ; Q3=1,755 ; iq=0,12 , LI=1,455 , LS=1,935

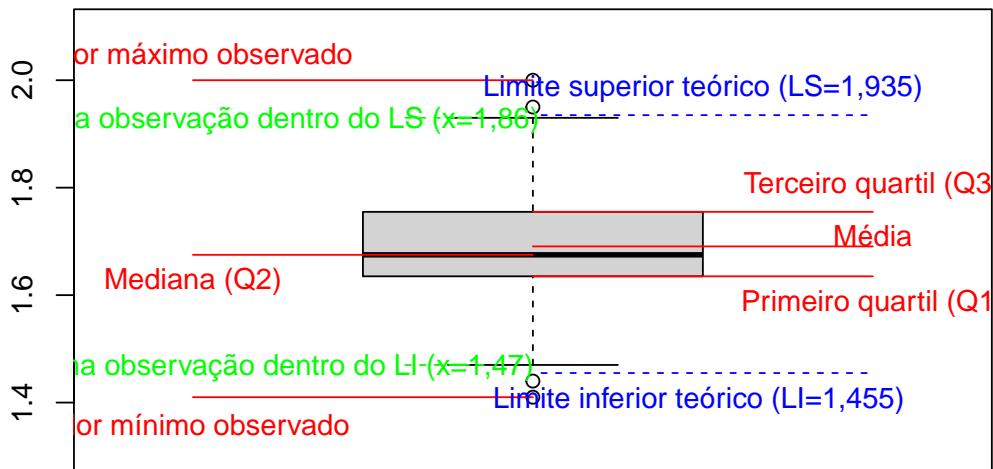


Figure 3.20: Box-plot de um rol de valores com Distribuição Normal (média 20 e variância 5

## Módulo 4

# Introdução ao cálculo de probabilidade

Seria bom começar um curso sobre teoria das probabilidades, dando uma definição de probabilidade concisa, simples e intuitiva, mas 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, definições simples são frequentemente enganosas ou, na melhor das hipóteses, tautológicas .

Por exemplo, poderíamos dizer que 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* porque usa o conceito de probabilidade, que é um sinônimo de possibilidade, chance, esperança, viabilidade, exequibilidade, expectativa, ...).

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

1. probabilidade refere-se a *experimentos aleatórios* e seus *eventos*; e,
2. probabilidade é um *número*.

O conceito clássico de probabilidade será a seguir apresentado e, ao final será abordado o conceito de probabilidade como uma função matemática alicerçada em alguns postulados (*conceito axiomático*).

## 4.1 Introdução conceitual essencial

### 4.1.1 Experimentos determinísticos e experimentos probabilísticos (aleatórios)

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

Ao contrário de um **experimento determinístico**, cujo resultado pode ser previamente determinado (como a reação de dois átomos de *H* com um átomo de *O* ou a distância percorrida - no vácuo sob velocidade constante e sem atrito - por um objeto  $S = V \times t$ ), o conceito de experimento aleatório é o que estabelece que seu resultado **não pode ser previsto com certeza**.

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.

### 4.1.2 Espaço amostral e seus elementos

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

Esses resultados são os elementos de um conjunto a que denominamos de *espaço amostral* e, usualmente o representamos pela letra grega maiúscula  $\Omega$ .

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

- 1- **apenas um** de seus elementos pode ocorrer ao se realizar o *experimento aleatório (resultado)*;
- e, 2- **ao menos um** dos possíveis resultados deverá ocorrer.

Tais propriedades são equivalentes a se dizer que os elementos do espaço amostral, os **resultados** listados como possibilidades de se verificar ao se realizar um **experimento aleatório** são **mutuamente exclusivos e exaustivos**.

Exemplos clássicos de experimentos aleatórios são o *lançamento de moedas*, *dados* ou extração de *cartas de um baralho*.

Os possíveis resultados como a face de uma moeda ou o número que um dado irá expor ao ser lançado, **embora não possam ser antecipados com certeza**, encontram-se limitados a um conjunto de todas as suas possibilidades, seu **espaço amostral**.

Para o lançamento de um dado:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

e para o lançamento de uma moeda

$$\Omega = \{\text{cara}, \text{coroa}\}$$

Um espaço amostral consiste então da enumeração (finita ou infinita) de todos os resultados possíveis de serem gerados em um experimento aleatório, generalizado como sendo o conjunto

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

**Cada um** dos possíveis resultados de um experimento aleatório com espaço amostral  $\Omega$  é chamado de um **elemento** desse espaço amostral e é denotado pela letra grega:  $\omega_n$ .

### 4.1.3 Evento de interesse (sucesso)

### 4.1.4 Evento

Denomina-se como **evento** um **subconjunto** finito do **espaço amostral** composto por um ou mais de seus elementos, e que **satisfazem (atendem)** ao enunciado definido no experimento aleatório desejado.

A expressão **evento de interesse** (ou sucesso) define, para o cálculo de probabilidades, a ocorrência de um resultado previamente definido no experimento aleatório.

Admita um **experimento aleatório** que consiste em se lançar um dado uma vez. Um **evento de interesse** pode ser definido como sendo obter um número par. A partir dessas condições, pode-se calcular-se a probabilidade de se obter **sucesso** no experimento aleatório; isto é, obter-se um número par ao se lançar um dado uma vez.

Admita um outro experimento aleatório que agora consiste em se lançar uma moeda duas vezes.

O **espaço amostral** desse experimento aleatório (**todos os possíveis resultados**) será um conjunto composto por quatro elementos:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$$

onde:

$$\begin{aligned}\omega_1 &= (\text{Cara}, \text{Coroa}) \\ \omega_2 &= (\text{Coroa}, \text{Cara}) \\ \omega_3 &= (\text{Cara}, \text{Cara}) \\ \omega_4 &= (\text{Coroa}, \text{Coroa})\end{aligned}$$

Se definirmos como **sucesso** nesse experimento aleatório obter-se

$$E = \{(\text{Cara}, \text{Cara})\}$$

,

dizemos que  $E$  é um **evento simples** pois é formado por apenas **um** elemento do espaço amostral.

Por outro lado, se definimos nosso sucesso como sendo obter

$$E_1 = \{(Cara, Coroa) \text{ ou } (Coroa, Cara)\}$$

$E_1$  será um **evento composto** pois é formado por **dois** elementos do espaço amostral.

Se codificarmos **Cara: 1** e **Coroa: 0**, podemos representar graficamente o espaço amostral  $\Omega$  do experimento aleatório e o **evento de sucesso**  $E_1$ , simultaneamente.

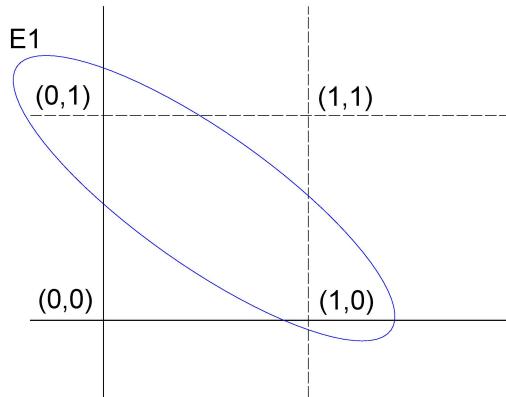


Figure 4.1: Representação gráfica do espaço amostral do experimento aleatório e do evento de interesse definido

Admita agora um outro experimento aleatório, estabelecido como a soma dos valores das faces de dois dados (ou um dado lançado duas vezes) aleatoriamente lançados. O espaço amostral desse experimento aleatório será um conjunto formado por 11 elementos.

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8, \omega_9, \omega_{10}, \omega_{11}\}$$

onde:

$$\begin{aligned}
\omega_1 &= 2 \\
\omega_2 &= 3 \\
\omega_3 &= 4 \\
\omega_4 &= 5 \\
\omega_5 &= 6 \\
\omega_6 &= 7 \\
\omega_7 &= 8 \\
\omega_8 &= 9 \\
\omega_9 &= 10 \\
\omega_{10} &= 11 \\
\omega_{11} &= 12
\end{aligned}$$

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 ( $f_i$ )
	(primeiro,segundo)		
2	(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	$\frac{5}{36}$
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	$\frac{1}{36}$

Se agora definimos nosso evento de interesse como sendo **obter uma soma ímpar**, nosso sucesso será verificado se ocorrer qualquer um desses elemetos do espaço amostral:

$$F = \{3; 5; 7; 9; 11\}$$

Nosso evento de interesse é um *evento composto* pois é formado por 5 elementos do *espaço amostral*  $\Omega$ .

Um evento de interesse  $G$  sobre o espaço amostral  $\Omega$  tal que

$$G = \Omega$$

expressa que qualquer um dos elementos de  $\Omega$  atende ao evento  $G$  e assim, a chance de ocorrência do evento  $G$  será absoluta. Esse tipo de evento é chamado de **evento certo**.

Se definirmos em evento de interesse  $I$  com um resultado não pertencente aos possíveis resultados representados no *espaço amostral*  $\Omega$ , como, por exemplo, 13, ou então um conjunto vazio  $\emptyset$ , esse evento será impossível de ocorrer. Esse tipo de evento é chamado de **evento impossível**.

Desse modo temos diferentes tipos de eventos de interesses:

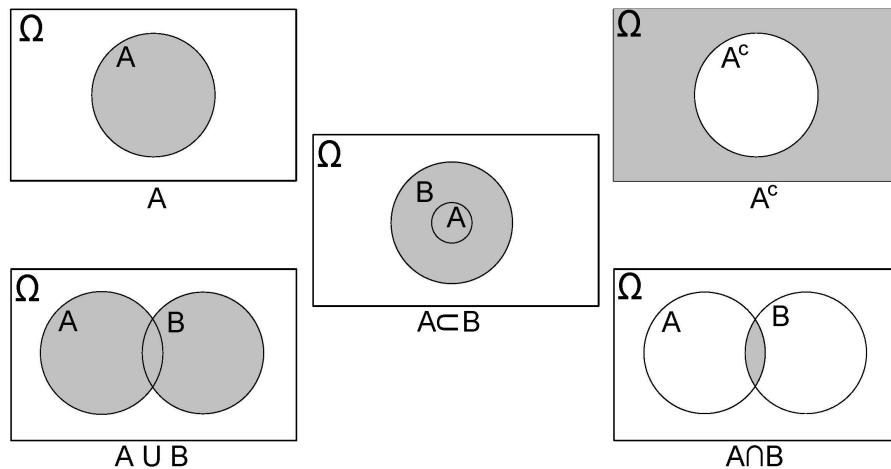
- 1- *simples*: composto por apenas um elemento do espaço amostral;
- 2- *composto*: composto por dois ou mais elementos do espaço amostral;
- 3- *certo*: composto por todos os elementos do espaço amostral;
- 4- *impossível*: composto por um elemento que não integra o espaço amostral.

#### 4.1.5 Operações com conjuntos e Diagramas de Venn

Em muitos dos problemas de probabilidade, o evento de interesse pode residir em **combinações de dois ou elementos** do conjunto que representa o espaço amostral. Uniões, interseções e complementos são alguns termos que, doravante, serão muito utilizados.

##### 4.1.5.1 União $A \cup B$

Sejam  $A$  e  $B$  dois subconjuntos finitos de um espaço amostral  $\Omega = \{1, 2, 3, 4, 5, 6\}$  tais que  $A = \{1, 2, 3\}$  e  $B = \{2, 4, 6\}$ .



### DIAGRAMAS DE VENN

Figure 4.2: Diagramas de Venn

Sua *união*, representada por  $A \cup B$ , é o subconjunto do espaço amostral  $\Omega$  que contém os elementos que pertençam a  $A$ , ou a  $B$  ou a ambos. Desse modo,  $A \cup B = \{1, 2, 3, 4, 6\}$  e o Diagrama de Venn correspondente será:

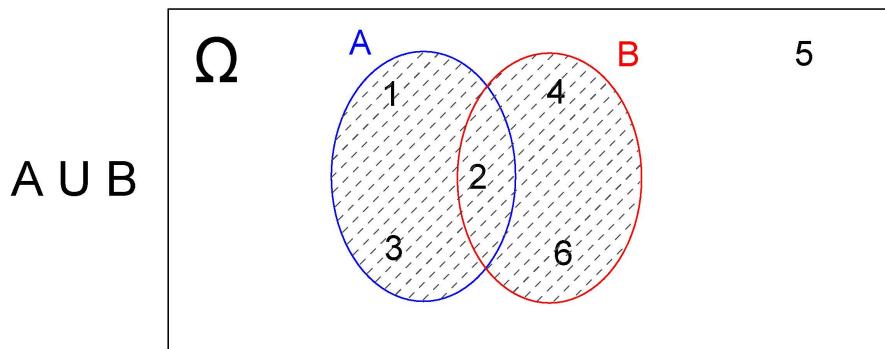
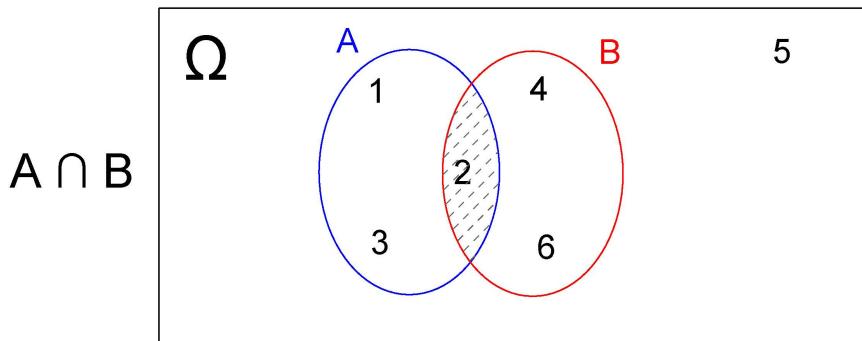


Figure 4.3: União:  $A \cup B$

#### 4.1.5.2 Interseção $A \cap B$

Sua *interseção*, representada por  $A \cap B$ , é o subconjunto do espaço amostral  $\Omega$  que contém todos os elementos que pertencem a **ambos simultaneamente**. Desse modo,  $A \cap B = \{2\}$  e o Diagrama de Venn correspondente será:

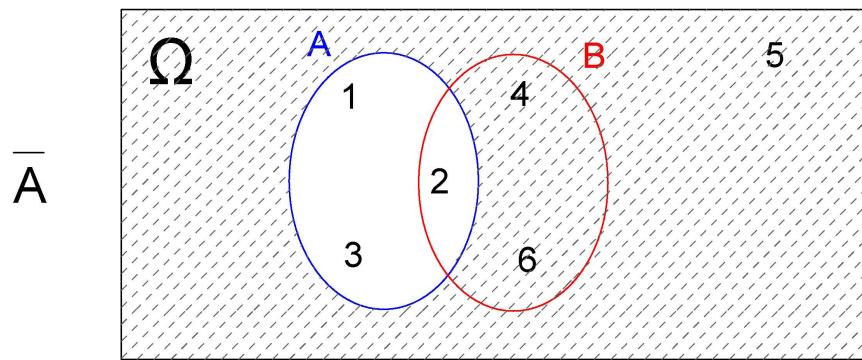
Figure 4.4: Interseção:  $A \cap B$ 

Caso não exista nenhum elemento na interseção (ela é vazia) tem-se :

$$A \cap B = \emptyset$$

#### 4.1.5.3 Complemento $A^c$

O complemento de  $A$ , representado por  $A^c$  (ou  $\bar{A}$ ), é o subconjunto do espaço amostral  $\Omega$  composto por todos os elementos que **não pertencem** a  $A$ . Desse modo,  $\bar{A} = \{4, 5, 6\}$  e o Diagrama de Venn correspondente será:

Figure 4.5: Complementar  $A^c$ 

O complemento} de  $B$ , representado por  $B^c$  (ou  $\bar{B}$ ), é o subconjunto do espaço amostral  $\Omega$  composto por todos os elementos que **não pertencem** a  $B$ . Desse modo,  $\bar{B} = \{1, 3, 5\}$  e o Diagrama de Venn correspondente será :

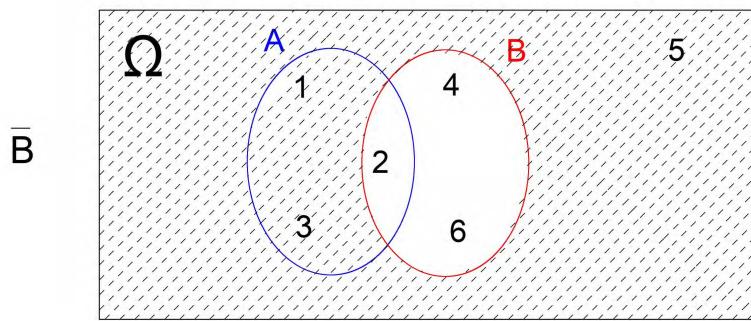


Figure 4.6: Complementar de B

#### 4.1.6 Eventos 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 esse espaço amostral é **uniforme** ou que seus elementos são **equiprováveis**.

No experimento de se lançar um dado e anotar o valor numérico de sua face todos os possíveis resultados apresentam a mesma probabilidade:  $\frac{1}{6}$ .

Já no experimento de se lançar dois dados e se anotar a soma dos valores numéricos de suas faces as probabilidades são diferentes.

Um significativo resultado é que a soma das probabilidades associadas a cada um desses possíveis resultados é um (1) (antecipando um dos postulados do conceito axiomático de probabilidade).

#### 4.1.7 Eventos independentes

Quando a possibilidade de ocorrência de um evento de interesse (sucesso) em um determinado experimento aleatório não é afetada pelo resultado **prévio** de outro diz-se que esses dois eventos são **independentes**. Caso contrário são ditos dependentes ou condicionados. Mais adiante esse conceito será introduzido de um modo mais detalhado.

#### 4.1.8 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 os eventos:  $E = \{(Cara)\}$  e  $F = \{(Coroa\}$  e assim sua interseção é vazia:

$$E \cap F = \emptyset$$

E por essa razão, se chamarmos de  $P(E)$  e  $P(F)$  as probabilidades de ocorrência desses resultados veremos que :

$$P(E) \cap P(F) = 0$$

#### 4.1.9 Eventos complementares

Definido um evento de interesse qualquer pode-se observar apenas dois resultados: **ocorrer ou não** o sucesso; ou seja, um ou outro deverá forçosamente ocorrer.

Chama-se de evento complementar ( $E^c$  ou  $\bar{E}$ ) a um evento ( $E$ ), aquele cuja probabilidade de sucesso seja:

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

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 Probabilidade

### 4.2.1 Introdução histórica

De acordo com alguns historiadores, a Teoria das probabilidades teve início como um ramo da Matemática com as célebres cartas entre Blaise Pascal (1623-1662) e Pierre de Fermat (1607-1665), após uma consulta feita por um nobre cavaleiro (Antoine Gombaud, o *Chevalier de Méré*) a Pascal, relacionadas a como repartir um montante apostado em um jogo de dados caso ele tenha que ser interrompido antes de sua conclusão. Todavia o estudo não formal remonta a alguns séculos atrás.

Probabilidade tem sido definida como sendo o estudo da frequência de aparição de um fenômeno em relação a todas as suas possíveis alternativas; ou seja, seu objeto é o estudo das possibilidades dos fenômenos aleatórios. O estudo das probabilidades possui, digamos assim, duas raízes históricas:

- 1- a solução de problemas relacionados a jogos; e,
- 2- a análise estatística de dados atuariais.



Figure 4.7: Astralagus (um dos ossos que compõem o calcanhar, usado no Egito antigo como um dado rudimentar)

#### 4.2.2 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*:

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 estabelecido sem a observação *a priori* dos resultados de experimento aleatório, reproduzindo de modo razoável a distribuição das frequências quando o experimento é diretamente observado.

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 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.

#### 4.2.3 Conceito frequentista ou *a posteriori*

Todavia, se realizarmos o experimento aleatório anterior algumas vezes apenas, tal regularidade poderá não ser, naturalmente, observada: 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.

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*.

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

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.

Essa é 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.

#### 4.2.4 Conceito axiomático

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  $\Omega$ ; isto é, uma função que associa uma quantidade  $P(\Omega)$  a cada elemento  $\omega \in \Omega$ .

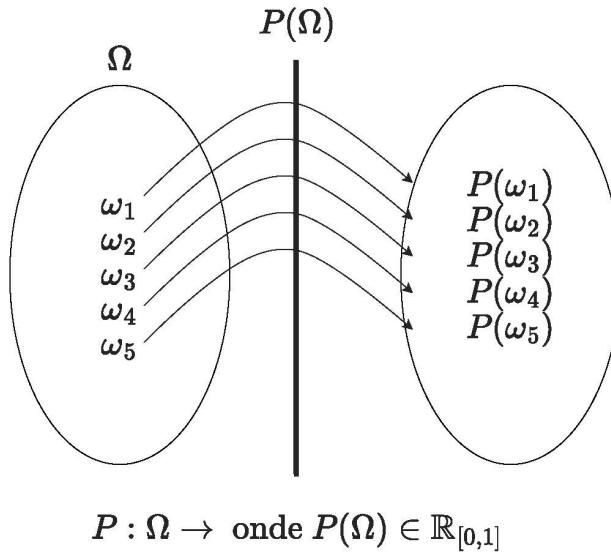


Figure 4.8: Representação gráfica da função  $P(\Omega)$

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.

#### 4.2.4.1 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.



Figure 4.9: Andrey Nikolaevich Kolmogorov (1903-1987)

$$0 \leq P(\Omega) \leq 1$$

#### 4.2.4.2 Postulado da certeza

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 Postulado da aditividade para eventos mutuamente exclusivos

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

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  $\omega_1$  e  $\omega_2$ , pode ser facilmente visto que é satisfeito 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  $\omega$  e  $P(\Omega)$  atende aos três axiomas anteriormente estabelecidos, diz-se que que ela é a **função de probabilidade** de  $\Omega$ .

#### 4.2.5 Regra geral da adição de probabilidades de eventos

Considerem agora a Tabela 4.3 de dupla entrada onde vemos a distribuição de alunos conforme seu sexo e o curso escolhido:

Table 4.3: Distribuição da quantidade de alunos segundo seu sexo e curso escolhido

Curso	Sexo		
		Masculino (M)	Feminino (F)
Matemática pura (M)	70	40	110
Matemática aplicada (A)	15	15	30
Estatística (E)	10	20	30

Curso	Sexo		
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 para a adição de probabilidades 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**:

$$P(A \cup B) = P(A) + P(B)$$

Exemplo: Seja o experimento aleatório de se lançar um dado (com seis faces) e observar o valor numérico da face que ficar exposta. Qual a probabilidade de se observar os valores **1 ou 4**?

Definindo os eventos de interesse:

- 1-  $E_1$  = sair face 1 ( $P(E_1) = \frac{1}{6}$ ); e,
- 2-  $E_4$  = sair face 4 ( $P(E_4) = \frac{1}{6}$ ).

Pede-se  $P(E_1 \cup E_4)$ .

Como  $E_1$  e  $E_4$  são \*eventos mutuamente exclusivos\*\*:  $E_1 \cap E_4 = \emptyset$  (portanto a probabilidade é zero), então  $P(E_1 \cup E_4) = P(E_1) + P(E_4) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$ .

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 adição de probabilidades 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:

$$\begin{aligned} P(E_A \cup E_C) &= P(E_A) + P(E_C) - P(E_A \cap E_C) \\ &= \frac{20}{10} + \frac{50}{100} - 0 \\ &= \frac{70}{100} \\ &= 0,70 \end{aligned}$$

#### 4.2.6 Probabilidade de eventos condicionados

Dois eventos  $A$  e  $B$  de 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  $B$ ,

e seu cálculo é feito pela **regra geral da probabilidade de dois eventos condicionados**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

sendo  $P(B) > 0$  e  $P(A) > 0$  nas expressões acima.

De modo geral, admita que os eventos  $E_1, E_2, \dots, E_n$  formam uma partição do espaço amostral.

Os eventos não têm interseções entre si e a união destes é igual ao espaço amostral e seja  $A$  um evento qualquer desse espaço.

Então a probabilidade de ocorrência desse evento será dada por:

$$\begin{aligned} P(A) &= P(A \cap E_1) + P(A \cap E_2) + \dots + P(A \cap E_n) \\ &= P(E_1) \times P(A|E_1) + P(E_2) \times P(A|E_2) + \dots + \\ &\quad P(E_n) \times P(A|E_n) \end{aligned}$$

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**)?

$$\begin{aligned}
 P(Fem|Est) &= \frac{P(Fem \cap Est)}{P(Est)} \\
 &= \frac{20}{30} = \frac{2}{3}
 \end{aligned}$$

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:

Do diagrama apresentado na Figura 4.10 podemos extrair imediatamente as probabilidades pedidas:

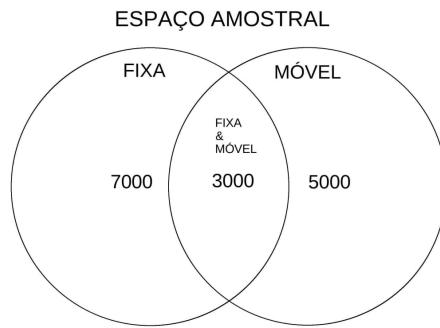


Figure 4.10: Diagrama de Venn do espaço amostral

- $P(F|M)$  (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 probabilidades 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 ( $P(T \cap F) \stackrel{?}{=} P(T) \times P(F)$ )

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

#### 4.2.7 Dependência e independência de eventos

Pela **regra geral da probabilidade de dois eventos condicionados**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Como a probabilidade de interseção não se altera ( $P(A \cap B) = P(B \cap A)$ ), podemos reescrever essas duas expressões:

$$P(A \cap B) = P(A|B) \times P(B)$$

$$P(A \cap B) = P(B|A) \times P(A)$$

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, ...).

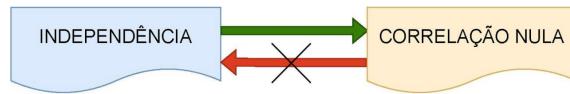


Figure 4.11: Independência implica em ausência de qualquer tipo de associação (a recíproca não se aplica)

#### 4.2.8 Regra geral do produto das probabilidades para eventos independentes

Se  $E_1, E_2, \dots, E_n$  são eventos totalmente 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:

$$\begin{aligned} P(Fab1) &= \frac{168}{252} \\ &= 0,6666 \end{aligned}$$

$$\begin{aligned} P(Aprov) &= \frac{240}{252} \\ &= 0,9523 \end{aligned}$$

$$\begin{aligned} P(Fab1 \cap Aprov) &= \frac{160}{252} \\ &= 0,6349 \end{aligned}$$

**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:

$$\begin{aligned}
 P(C_1 \cap C_2) &= 0,25 \times 0,40 \\
 &= 0,10.
 \end{aligned}$$



Figure 4.12: Thomas Bayes (1702 - 1761)

### 4.3 Teorema de Bayes

Pela **regra da probabilidade condicionada** temos que

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

e, de modo equivalente,

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Pela igualdade  $P(A \cap B) = P(B \cap A)$ , substituindo-se a segunda expressão na primeira chega-se a:

$$P(B|A) = \frac{P(A|B)P(A)}{P(B)}$$

uma **relação** entre duas probabilidades inversamente condicionadas conhecida como **Teorema de Bayes**.

Para um espaço amostral mais amplo, de modo geral consideremos, inicialmente o diagrama da Figura 4.13 onde  $\Omega$  é o espaço amostral de um experimento aleatório qualquer:

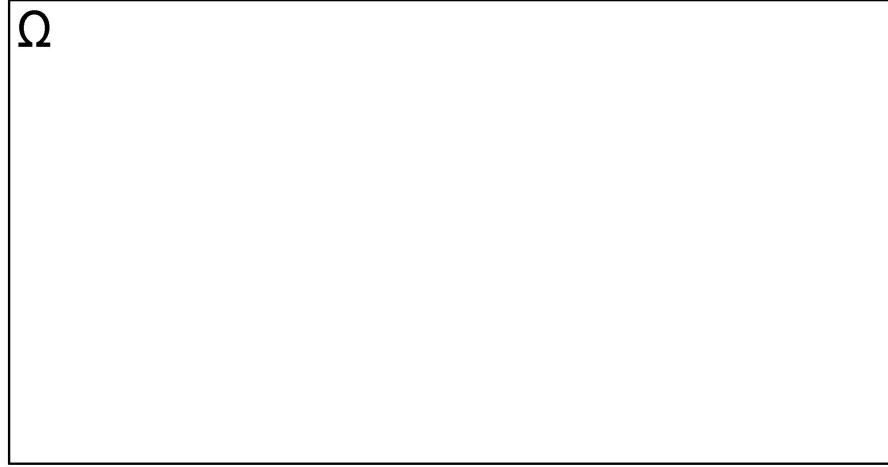


Figure 4.13: Espaço amostral

Admita que  $E_1, E_2, E_3$  e  $E_4$  formem a partição do espaço amostral  $\Omega$  (seus elementos são **mutuamente exclusivos**) como exposto na Figura 4.14

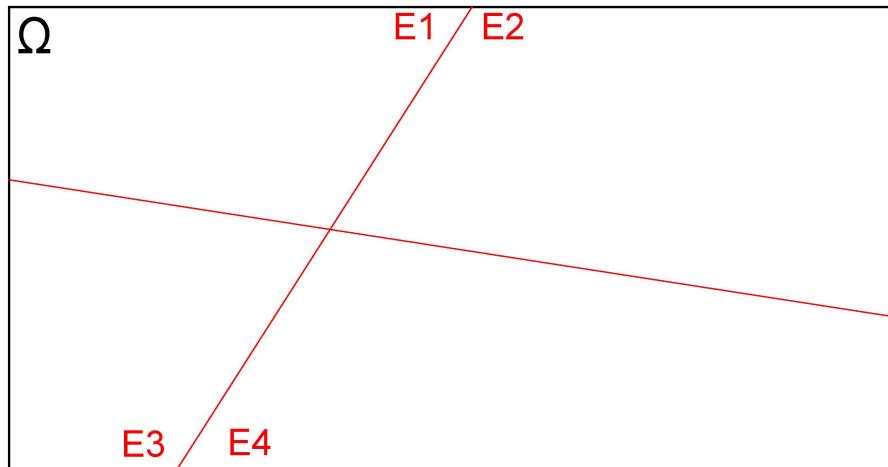


Figure 4.14: Espaço amostral e suas partições

E seja  $B$  um evento qualquer em  $\Omega$  como ilustrado na Figura 4.15

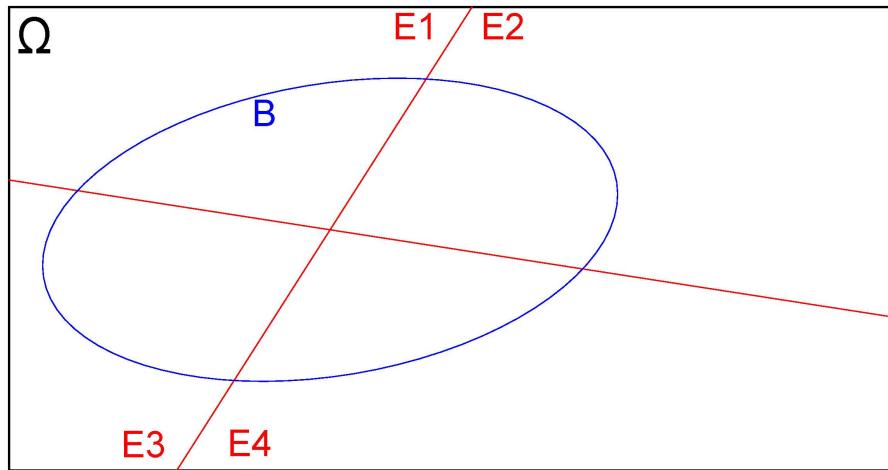


Figure 4.15: Evento definido sobre o espaço amostral

Delimitemos as interseções do evento  $B$  com as partições  $E_1, E_2, E_3$  e  $E_4$  do espaço amostral  $\Omega$ , como ilustrado na Figura 4.16

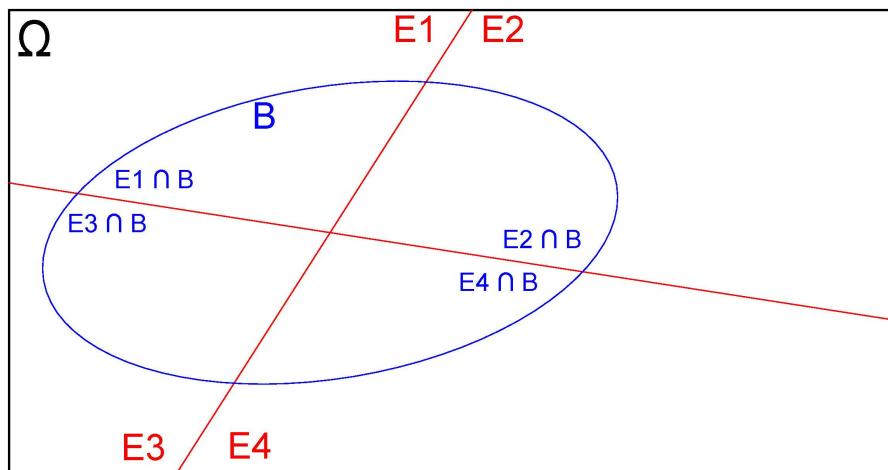


Figure 4.16: Interseções das partições do espaço amostral com o evento B

Isso pode ser estendido, em uma forma geral, para  $i = 1, \dots, n$  partições como ilustrado na Figura 4.17

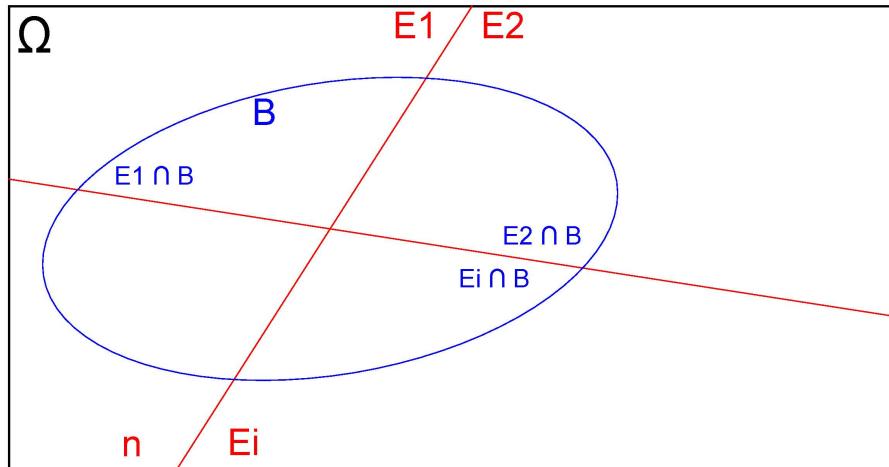


Figure 4.17: Interseções das  $n$  partições do espaço amostral com o evento  $B$

Na representação esquemática da Figura 4.17 podemos identificar:

- 1-  $E_1, E_2, \dots, E_i, \dots, E_n$  constituem-se em partições do espaço amostral  $\Omega$ ;
- 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  $\Omega$  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  $\Omega$

São conhecidas as probabilidades de ocorrência de cada um dos elementos do espaço amostral  $\Omega$ :

$$P(E_1); P(E_2); P(E_3); \dots; P(E_i); \dots; P(E_n)$$

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  $\Omega$ :

$$P(B) = P(E_1 \cap B) + P(E_2 \cap B) + \cdots + P(E_i \cap B) + \cdots + P(E_n \cap B)$$

$$P(B) = \sum_{i=1}^n P(E_i \cap B)$$

Pela **Regra do produto de eventos condicionados**, a probabilidade de ocorrência do evento  $B$  posto ter ocorrido um evento  $E_i$  é:

$$P(B|E_i) = \frac{P(E_i \cap B)}{P(E_i)}$$

$$P(E_i \cap B) = P(E_i) \times P(B|E_i)$$

com  $P(E) > 0$

Aplicando-se na expressão anteriormente desenvolvida da *probabilidade de ocorrência do evento  $B$*  teremos:

$$P(B) = P(E_1 \cap B) + P(E_2 \cap B) + \cdots + P(E_i \cap B) + \cdots + P(E_n \cap B)$$

$$P(B) = P(E_1) \times P(B|E_1) + P(E_2) \times P(B|E_2) +$$

$$\cdots + P(E_i) \times P(B|E_i) +$$

$$\cdots + P(E_n) \times P(B|E_n)$$

Portanto a **probabilidade total** do evento  $B$  em  $\Omega$  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  $\Omega$  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?

Nosso experimento aleatório é a medida do **incremento das vendas** de um produto de uma certa empresa que ela o considera ser **influenciado exclusivamente** por quatro eventos - ações que ela pode adotar ou sofrer - independentes indicados como sendo:

- 1- *marketing*;
- 2- propaganda;
- 3- flutuações na economia; ou,
- 4- efeitos sazonais.

Cada um deles possui uma **intensidade diferente**.

Da leitura do enunciado extraímos as probabilidades de ocorrência de cada um dos eventos influenciadores:

- Ação de *marketing* →  $P(E_1) = 0,40$ ;
- Ação de propaganda →  $\$P(E_{\{2\}}) = 0,30 \$$ ;
- Flutuações na economia →  $P(E_3) = 0,20$ ; ou,
- Sazonalidade →  $P(E_4) = 0,10$ .

As probabilidades de incremento das vendas ( $B$ ) pela ocorrência dos eventos causadores são (**posto ter ocorrido o evento  $E_i$** ):

- $P(B|E_1) = 0,07$  ;
- $P(B|E_2) = 0,075$ ;
- $P(B|E_3) = 0,03$ ; 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:

$$P(E_1|B) = \frac{P(E_1) \times P(B|E1)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]}$$

$$P(E_1|B) = \frac{0,40 \times 0,07}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)}$$

$$P(E_1|B) = 0,478$$

Probabilidade da empresa ter realizado propaganda, **posto** ter ocorrido um incremento nas vendas de seu produto:

$$P(E_2|B) = \frac{P(E_2) \times P(B|E2)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]}$$

$$P(E_2|B) = \frac{0,30 \times 0,075}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)}$$

$$P(E_2|B) = 0,385$$

Probabilidade da empresa ter ocorrido flutuações na economia, **posto** ter ocorrido um incremento nas vendas de seu produto:

$$P(E_3|B) = \frac{P(E_3) \times P(B|E3)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]}$$

$$P(E_3|B) = \frac{0,20 \times 0,03}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)}$$

$$P(E_3|B) = 0,103$$

Probabilidade da empresa ter ocorrido efeitos sazonais, **posto** ter ocorrido um incremento nas vendas de seu produto:

$$P(E_4|B) = \frac{P(E_4) \times P(B|E4)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]}$$

$$P(E_4|B) = \frac{0,10 \times 0,02}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)}$$

$$P(E_4|B) = 0,034$$

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**: 0,058.

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 4.18).

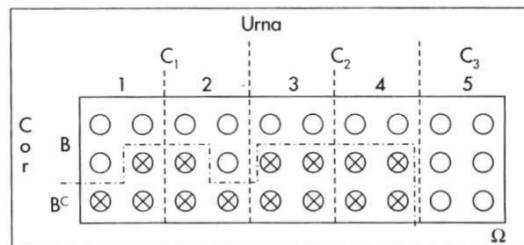


Figure 4.18: Cinco urnas cada uma com 6 bolas em cores de diferentes quantidades da cor branca

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:

$$\begin{aligned}
 P(C_1) &= \frac{2}{5} \\
 P(C_2) &= \frac{2}{5} \\
 P(C_3) &= \frac{1}{5} \\
 P(Branca|C_1) &= \frac{1}{2} \\
 P(Branca|C_2) &= \frac{1}{3} \\
 P(Branca|C_3) &= 1
 \end{aligned}$$

$$\begin{aligned}
 P(C_3|Branca) &= \frac{P(C_3) \times P(Branca|C_3)}{\sum_{i=1}^3 [P(C_i) \times P(Branca|C_i)]} \\
 P(C_3|Branca) &= \frac{0,20 \times 1,00}{(0,40 \times 0,50) + (0,40 \times 0,33) + (0,20 \times 1,00)} \\
 P(C_3|Branca) &= 0,375
 \end{aligned}$$

#### 4.3.1 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:

$$P(2^{\text{a}} \text{azul}) = \frac{5}{10} = \frac{1}{2}$$

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*:

$$P(A \cap B) = P(A) \times P(B)$$

$$\begin{aligned} P(1^a \text{ vermelha}) &= \frac{5}{10} = \frac{1}{2} \\ P(2^a \text{ azul}) &= \frac{5}{10} = \frac{1}{2} \\ P(1^a \text{ vermelha}, 2^a \text{ azul}) &= \frac{25}{100} = \frac{1}{4} \end{aligned}$$

Como  $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2}$ , os eventos são independentes.

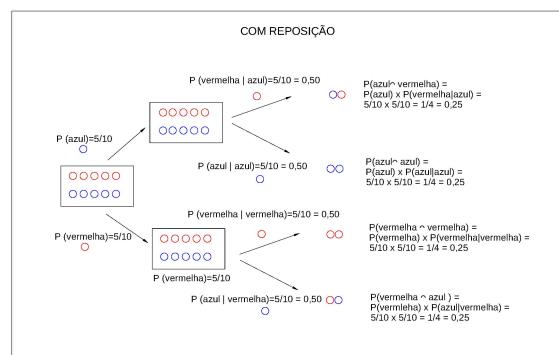


Figure 4.19: Ilustração do experimento aleatório sob a condição de reposição

### 4.3.2 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<sup>a</sup> 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<sup>a</sup> 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:

$$P(1^a \text{vermelha}) = \frac{2}{9} + \frac{5}{18} = \frac{1}{2}$$

3<sup>a</sup> 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<sup>a</sup> Etapa: analisar a possibilidade de se obter uma bola **vermelha** e em seguida **azul**:

- uma **vermelha** e depois outra **azul**:  $\frac{5}{18}$ ;

5<sup>a</sup> Etapa: Esses dois eventos são independentes?

Esses eventos serão *independentes se, e somente se*:

$$P(A \cap B) = P(A) \times P(B)$$

$$P(1^{\text{a}} \text{vermelha}) = \frac{2}{9} + \frac{5}{18} = \frac{1}{2}$$

$$P(2^{\text{a}} \text{azul}) = \frac{5}{18} + \frac{2}{9} = \frac{1}{2}$$

$$P(1^{\text{a}} \text{vermelha}, 2^{\text{a}} \text{azul}) = \frac{5}{18}$$

Como  $\frac{5}{18} \neq \frac{1}{2} \times \frac{1}{2}$ , os eventos **não são independentes**.

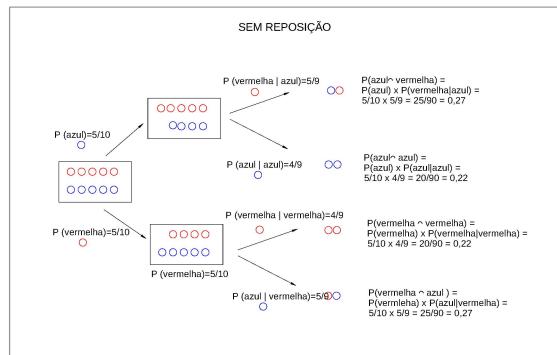


Figure 4.20: Ilustração do experimento aleatório sob a condição de não reposição

## 4.4 Teoremas da Teoria das probabilidades

### 4.4.1 Teorema 01

Se  $E$  é um evento num espaço discreto  $\Omega$ , 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 resultados (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:

$$P(E) = P(E_1) + P(E_2) + P(E_3) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

#### 4.4.2 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  $\Omega$ , 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:

$$\begin{aligned} P(E) &= P(E_1) + P(E_2) + P(E_3) + \dots + P(E_n) \\ P(E) &= \frac{1}{N} + \frac{1}{N} + \frac{1}{N} + \dots + \frac{1}{N} \\ P(E) &= \frac{n}{N} \end{aligned}$$

#### 4.4.3 Teorema 03

Se  $E$  e  $E^c$  são eventos complementares no espaço amostra  $\Omega$  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:

$$\begin{aligned} P(\Omega) &= 1 \\ 1 &= P(E \cup E^c) \\ 1 &= P(E) + P(E^c) \end{aligned}$$

#### 4.4.4 Teorema 04

$$P(\emptyset) = 0$$

Sendo  $\Omega$  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:

$$\begin{aligned} P(\Omega) &= P(\Omega \cup \emptyset) \\ P(\Omega) &= P(\Omega) + P(\emptyset) \\ P(\Omega) - P(\Omega) &= P(\emptyset) \\ P(\emptyset) &= 0 \end{aligned}$$

#### 4.4.5 Teorema 05

Se  $A$  e  $B$  são eventos em um mesmo espaço amostral  $\Omega$  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:

$$\begin{aligned} P(B) &= P(A) + P(A^c \cap B) \\ P(A) &= P(B) - P(A^c \cap B) \end{aligned}$$

#### 4.4.6 Teorema 06

A probabilidade de qualquer evento  $E$  em  $\Omega$  está compreendida entre  $0 \leq P(E) \leq 1$ .

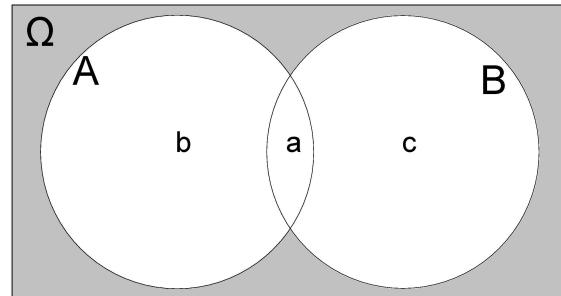
Estando  $\emptyset \subset E \subset \Omega$  e considerando-se o Teorema 5 tem-se:

$$P(\emptyset) \leq P(E) \leq P(\Omega) \quad 0 \leq P(E) \leq 1$$

#### 4.4.7 Teorema 07

Para dois eventos quaisquer em  $\Omega$ ,  $A$  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,
- $P(A^c \cap B) = c$ .

$$\begin{aligned} P(A \cup B) &= a + b + c \\ P(A \cup B) &= (a + b) + (c + d) - a \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

#### 4.4.8 Teorema 08

Para três eventos quaisquer em  $\Omega$ ,  $A$ ,  $B$  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

### 5.1 Função discreta de distribuição de probabilidade

Seja  $E$  um experimento aleatório e  $\Omega$  seu espaço amostral. Uma função ( $X$ ) que associe cada elemento  $\omega$  pertencente a  $\Omega$  a um número real  $X(\omega) = x$ , é denominada mais apropriadamente de função aleatória ou função estocástica.

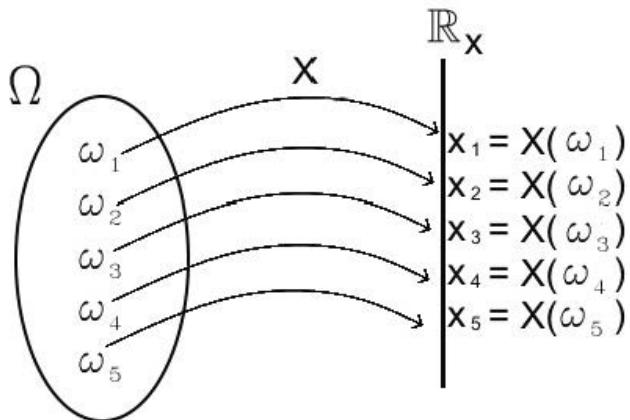


Figure 5.1: Função discreta de distribuição de probabilidade

Considere  $X$  uma variável aleatória discreta e suponha que os valores que ela pode assumir são dados por  $x_1, x_2, x_3, \dots$  dispostos em alguma ordem. Suponha que esses valores são assumidos tendo probabilidades de ocorrência dadas por:

$$P(X = x_k) = f(x_k)$$

Ponto amostral	(cara,cara)	(cara,coroa)	(coroa,cara)	(coroa,coroa)
$X$	2	1	1	0

com  $k = 1, 2, \dots$

Uma *função discreta de probabilidade* pode ser definida associando cada um dos possíveis valores da variável aleatória à sua probabilidade:

$$P(X = x) = f(x)$$

Para  $x = x_k$ ,

$$P(X = x_k) = f(x_k)$$

Para que uma *função  $f(x)$*  possa ser considerada uma **função (discreta ou contínua)** de distribuição de probabilidade, ela precisa necessariamente atender a:

$$0 \leq f(x_k) \leq 1$$

para qualquer  $x_k \in \Omega$ ; e também que

$$\sum_{k=1}^n f(x_k) = 1.$$

A probabilidade de ocorrência de um dos valores da variável aleatória deverá estar sempre compreendida entre  $0 \leq P(X = x_k) \leq 1$ : **postulado do intervalo**.

A soma das probabilidades de todos os possíveis valores que a variável aleatória poderá assumir deverá ser sempre 1: **postulado da probabilidade do evento certo**.

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 defina 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:

Função discreta de probabilidades	da variável	aleatória	X
$x_k$	0	1	2
$P(X = x_k) = f(x_k)$	$1/4$	$1/2$	$1/4$

As probabilidades de ocorrência de cada um desses eventos é:

$$P(\text{cara}, \text{cara}) = \frac{1}{4}$$

$$P(\text{cara}, \text{coroa}) = \frac{1}{4}$$

$$P(\text{coroa}, \text{cara}) = \frac{1}{4}$$

$$P(\text{coroa}, \text{coroa}) = \frac{1}{4}$$

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*.

$$P(X = 0) = P(\text{coroa}, \text{coroa}) = \frac{1}{4}$$

$$\begin{aligned} P(X = 1) &= P[(\text{cara}, \text{coroa}) \cup (\text{coroa}, \text{cara})] \\ &= P(\text{cara}, \text{coroa}) + P(\text{coroa}, \text{cara}) \\ &= \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{2} \end{aligned}$$

$$P(X = 2) = P(\text{cara}, \text{cara}) = \frac{1}{4}$$

Uma *função de distribuição cumulativa*  $F$  para uma variável aleatória  $X$  exprime a probabilidade de que a variável aleatória  $X$  assuma um valor inferior ou igual a determinado  $x$  e é definida por:

$$F(x) = P(X \leq x)$$

Propriedades:

$x_k$	0	1	2
$P(X = x_k) = f(x_k)$	1/4	1/2	1/4

- 1-  $0 \leq F(x) \leq 1$ ;
- 2-  $F(x)$  é não decrescente:  $F(x) \leq F(y)$  se  $x \leq y$ ;
- 3-  $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$ ;
- 4-  $F(+\infty) = \lim_{x \rightarrow \infty} F(x) = 1$

A função de probabilidade  $f$  para uma variável aleatória discreta  $X$  pode ser obtida de sua função de probabilidade cumulativa  $F$  pois para todo  $x$  em  $(-\infty, \infty)$ :

$$F(x) = P(X \leq x) = \sum_{u \leq n} f(u)$$

Equivale dizer que é a *soma sobre todos os valores u assumidos por X para os quais u ≤ 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:

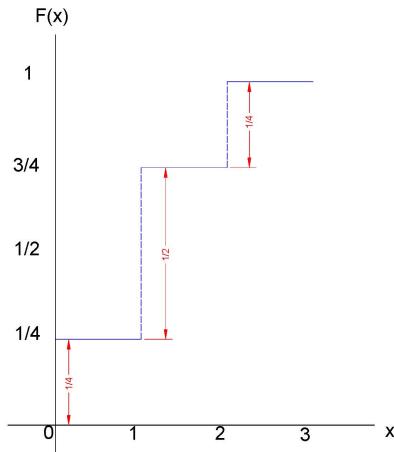
$$F(x) = \begin{cases} 0 & -\infty < x < x_1 \\ f(x_1) & x_1 \leq x < x_2 \\ f(x_1) + f(x_2) & x_2 \leq x < x_3 \\ \dots & \\ f(x_1) + \dots + f(x_n) & x_n \leq x < x_\infty \end{cases} \quad (5.1)$$

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:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4} & 0 \leq x < 1 \\ \frac{3}{4} & 1 \leq x < 2 \\ 1 & 2 \leq x \end{cases} \quad (5.2)$$

O gráfico de sua função de probabilidade cumulativa é:



- a) os "saltos" em 0, 1 e 2 são precisamente as probabilidades;
- b) o valor da função de probabilidade cumulativa em um inteiro é dado pelo degrau mais alto (continua à direita em 0);
- c) à medida que seguimos da esquerda para a direita (subimos) a função de probabilidade cumulativa permanece a mesma ou aumenta, assumindo valores de 0 a 1 (função monotonamente crescente)

Figure 5.2: Função de probabilidade cumulativa

## 5.2 Função de densidade de probabilidade

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.

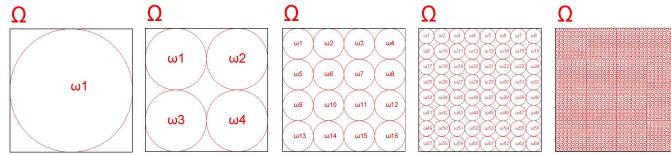


Figure 5.3: Diferentes espaços amostrais de um experimento aleatório (por razões gráficas desprezem o espaço fora dos círculos

Interpretem o último deles como um espaço amostral formado por  $\infty$  pontos amostrais.

Os eventos que compõem os quatro primeiros espaços amostrais são variável aleatória discretas.

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  $\Omega_1$ ,  $P(\omega_1) = 1$

Em  $\Omega_2$ ,  $P(\omega_1) = P(\omega_2) = P(\omega_3) = P(\omega_4) = 0,50$

Em  $\Omega_3$ ,  $P(\omega_1) = P(\omega_2) = \dots = P(\omega_{16}) = 0,0625$

Em  $\Omega_4$ ,  $P(\omega_1) = P(\omega_2) = \dots = P(\omega_{64}) = 0,015625$

Em  $\Omega_5$ ,  $P(\omega_n) \rightarrow 0$ , 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 com variáveis aleatórias contínuas não há sentido em se falar de uma *probabilidade pontual exata* (associada a um resultado específico).

Com variáveis aleatórias contínuas considera-se a probabilidade de realização de um *intervalo de valores* que ela assume e, ao estabelecermos sua função de probabilidade contínua ela apresentará as seguintes propriedades:

$$f(x) \geq 0$$

para todo  $x \in (-\infty, \infty)$

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

Se  $X$  é uma variável aleatória contínua então a probabilidade de que  $X$  assuma qualquer valor em particular é zero, enquanto que a *probabilidade intervalar* de que  $X$  esteja entre dois valores diferentes, digamos,  $a$  e  $b$  será dada por:

$$P(a < X < b) = \int_a^b f(x) dx$$

A interpretação gráfica de uma função de probabilidade de uma variável contínua é dada pela área sob a curva entre os limites de interesse:  $a$  e  $b$ .

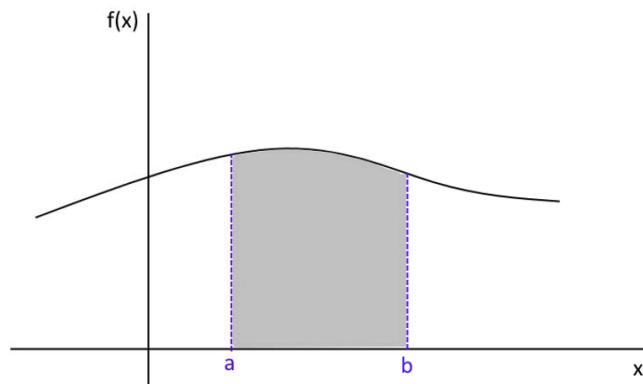


Figure 5.4: A área sob a curva de uma função de probabilidade de uma variável contínua entre dois valores quaisquer é a probabilidade de se observar valores entre esses dois pontos

Como  $f(x) \geq 0$ , essa curva estará acima do eixo  $x$  e a totalidade da área será igual a 1 posto que  $\int_{-\infty}^{\infty} f(x) dx = 1$ .

A função de probabilidade cumulativa:  $F(x) = P(X \leq x)$  assumirá igualmente a forma de uma curva, crescente, aumentando de 0 para 1.

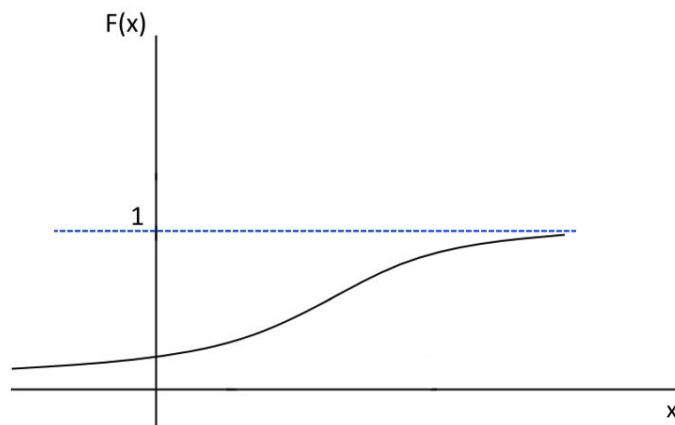


Figure 5.5: Função de probabilidade cumulativa

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  $0 \leq X \leq \frac{1}{2}$ .

$$f(x) = \begin{cases} 2x & \text{para } 0 \leq x \leq 1 \\ 0 & \text{fora desse intervalo} \end{cases} \quad (5.3)$$

A resolução deste exemplo será feita de um modo *geométrico*.

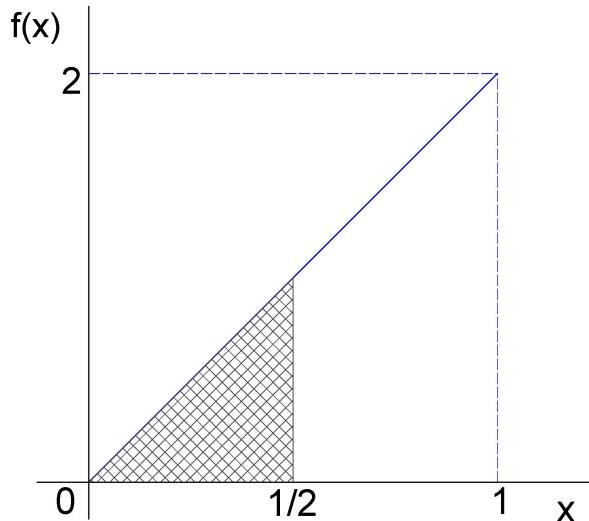


Figure 5.6: A probabilidade de se observar valores entre 0 e  $1/2$  é igual à área sob a função densidade de probabilidade entre esses dois valores

- (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:  $E(cX) = c.E(X)$  ( $c \in \mathbb{R}$ );
- 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:  $E(X.Y) = E(X).E(Y)$ .

A variância de uma variável aleatória qualquer  $X$ , denotada por  $Var(X)$ , será dada por:

$$\begin{aligned} Var(X) &= E(X^2) - [E(X)]^2 \\ Var(X) &= \sum_{i=1}^n [x_i - E(X)]^2 \cdot P_i \end{aligned}$$

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:  $Var(cX) = c^2.Var(X)$  ( $c \in \mathbb{R}$ );
- 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: \*

$x_i$	$P(X = x_i)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

Função discreta de distribuição de probabilidades de  $X$

$$E(X) = \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = 3,50$$

$$\begin{aligned} Var(X) &= (1 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + (2 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + \\ &\quad (3 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + (4 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + (5 - 3,50)^2 \cdot \left(\frac{1}{6}\right) + \\ &\quad (6 - 3,50)^2 \cdot \left(\frac{1}{6}\right) \\ &= 2,90 \end{aligned}$$

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

$$E(X) = (0.0, 10) + (1.0, 20) + 2.0, 30 + (3.0, 30) + (4.0, 10) = 2, 10 \text{ (caminhões por dia)}$$

$$\begin{aligned} Var(X) &= (0 - 2, 10)^2 \cdot 0, 10 + (1 - 2, 10)^2 \cdot 0, 20 + (2 - 2, 10)^2 \cdot 0, 30 + \\ &\quad (3 - 2, 10)^2 \cdot 0, 30 + (4 - 2, 10)^2 \cdot 0, 10 \\ &= 1, 29^1 \end{aligned}$$

<sup>1</sup>: (caminhões por dia)<sup>2</sup>

## 5.4 Esperança e variância de uma variável aleatória contínua

A esperança e a variância de uma variável aleatória contínua são dadas, respectivamente, por:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x) dx$$

# 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 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:  $X \sim Ber(p)$

$x_i$	Evento	$P(X = x_i)$
1	Sucesso	p
0	Fracasso	$q=1-p$
$\Sigma$	-	1

Para uma variável de Bernoulli:

- 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	$q=5/6$
$\Sigma$	-	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  $\epsilon$  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  $p$ .

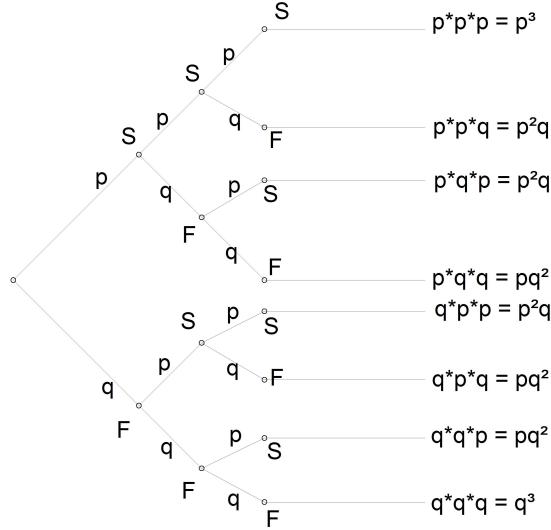
### 6.1.2 Binomial

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,  $P(X = 0) = 1 - p = q$ .



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 Geométrica 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,  $P(X \geq 2) = 0,3241 + 0,1852 + 0,0595 + 0,01020 + 0,00079 = 0,5797$

**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 Geométrica 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.

### 6.1.3 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**, ou em uma **determinada extensão, área ou volume**

Uma variável aleatória discreta  $X$  com Distribuição de *Poisson* é aquela que pode assumir **infinitos valores numeráveis** ( $k = 0, 1, 2, \dots, \infty$ ). 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  $e = 2,718$  (número irracional de Euler).

A esperança e a variância de uma variável aleatória discreta com Distribuição de *Poisson* são dados pelo seu parâmetro  $\lambda$  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-  $\lambda$  = 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  $\lambda$ );
- 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  $\lambda$  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):

$$\begin{aligned} P(X = 0) &= \frac{20^0 \cdot e^{-20}}{0!} = 2,061154e - 09 \\ P(X = 1) &= \frac{20^1 \cdot e^{-20}}{1!} = 4,122307e - 08 \\ P(X = 2) &= \frac{20^2 \cdot e^{-20}}{2!} = 4,122307e - 07 \end{aligned}$$

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 4,554699e - 07$$

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* ( $\lambda$ ) 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  $\lambda$  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:

$$\begin{aligned} P(X = k) &= \frac{\lambda^k \cdot e^{-\lambda}}{k!} \\ P(X = 4) &= \frac{6^4 \cdot e^{-6}}{4!} \\ &= 0,1338 \end{aligned}$$

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 Geométrica no qual  $n \rightarrow \infty$  e  $p$  é pequeno ( $n \geq 50$  e  $n.p \leq (5, 7)$ ). 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  $\lambda = n.p$ :

$$f(k) = P(X = k) = \frac{n.p^k \cdot e^{-n.p}}{k!}$$

## 6.2 Modelos teóricos do tempo de espera

As distribuições do tempo de espera é 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

Enquanto uma variável aleatória com distribuição Geométrica é uma variável que conta o número de sucessos ocorridos com a repetição de um experimento de Bernoulli (que apresenta duas possibilidades apenas) de modo independente, uma variável aleatória geométrica conta o número de tentativas até que **se verifique o primeiro sucesso**, atendendo também a:

- 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  $n - s$ imo 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:

$$\begin{aligned} f(X = x; p) &= P(X = j) = p \cdot (1 - p)^j \\ f(X = x; p) &= P(X = j) = p \cdot q^j \end{aligned}$$

O Modelo geométrico pode ser escrito sob uma “forma complementar”: o **número de tentativas  $n$  até se observar o primeiro sucesso**, agora com  $x = n = 1, 2, \dots$ .

$$f(X = x; p) = P(X = n) = p \cdot (1 - p)^{(n-1)}$$

$$f(X = x; p) = P(X = n) = p \cdot q^{(n-1)}$$

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:  $Var(X) = \frac{(1-p)}{p^2} = \frac{q}{p^2}$ .

Lembrando que uma variável aleatória Geométrica é uma 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 uma variável aleatória Geométrica é uma contagem do 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%.

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<sup>a</sup> e 6<sup>a</sup> 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 sexta 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 verificar o sucesso na 4<sup>a</sup> peça produzida (peça produzida com defeito) será:

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 3) &= (1 - 0,01)^3 \cdot 0,01 \\ P(X = 3) &= 0,009702 \end{aligned}$$

A probabilidade de verificar o sucesso na 5<sup>a</sup> peça produzida (peça produzida com defeito) será:

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 4) &= (1 - 0,01)^4 \cdot 0,01 \\ P(X = 4) &= 0,009605 \end{aligned}$$

A probabilidade de verificar o sucesso na 6<sup>a</sup> peça produzida (peça produzida com defeito) será:

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^k \cdot p \\ P(X = 5) &= (1 - 0,01)^5 \cdot 0,01 \\ P(X = 5) &= 0,009809 \end{aligned}$$

A probabilidade de termos uma peça **produzida com defeito** na quarta **OU** na quinta **OU** na sexta das peças produzidas será:

$$\begin{aligned} P(3 \leq j \leq 5) &= P(X = 3) + P(X = 4) + P(X = 5) \\ P(3 \leq j \leq 5) &= 0,009702 + 0,009605 + 0,009809 \\ P(3 \leq j \leq 5) &= 0,029116 \end{aligned}$$

A probabilidade de termos uma **peça defeituosa** na quarta **OU** na quinta **OU** na sexta das peças produzidas é de 2,9116%.

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 .s,** alinhamentos realizados).

Para o item (1) a probabilidade de termos a ocorrência de um sucesso (ou seja, um alinhamento ótico bem sucedido) na 4<sup>a</sup> montagem será:

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 3) &= (1 - 0,80)^3 \cdot 0,20 \\ P(X = 3) &= 0,0064 \end{aligned}$$

Para o item (2) considerando-se que as repetições são independentes, a probabilidade pedida será:

$$P(X = j)_{0 \leq j \leq 3} = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 0) &= (1 - 0,80)^0 \cdot 0,20 \\ P(X = 0) &= 0,80 \end{aligned}$$

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 1) &= (1 - 0,80)^1 \cdot 0,20 \\ P(X = 1) &= 0,16 \end{aligned}$$

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 2) &= (1 - 0,80)^2 \cdot 0,20 \\ P(X = 2) &= 0,032 \end{aligned}$$

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 3) &= (1 - 0,80)^3 \cdot 0,20 \\ P(X = 3) &= 0,0064 \end{aligned}$$

A probabilidade pedida é de:

$$\begin{aligned} P(X = j)_{0 \leq j \leq 3} &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ P(X = j)_{0 \leq j \leq 3} &= 0,9984 \end{aligned}$$

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:

$$P(X = j)_{3 \leq j \leq \infty} = 1 - P(X = 0) + P(X = 1) + P(X = 2)$$

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 0) &= (1 - 0,80)^0 \cdot 0,20 \\ P(X = 0) &= 0,80 \end{aligned}$$

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 1) &= (1 - 0,80)^1 \cdot 0,20 \\ P(X = 1) &= 0,16 \end{aligned}$$

$$\begin{aligned} f(X = x; p) &= P(X = j) \\ P(X = j) &= (1 - p)^j \cdot p \\ P(X = 2) &= (1 - 0,80)^2 \cdot 0,20 \\ P(X = 2) &= 0,032 \end{aligned}$$

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

Uma variável aleatória discreta que segue uma 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 é considerada a situação em que se modelam as probabilidades de se verificar mais de um evento de sucesso.

Ao se realizar repetidos experimentos de Bernoulli, uma variável aleatória Binomial Negativa modela as probabilidades relacionadas ao número de repetições necessárias para se observar  $r$  sucessos.

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 das demais (há independência);
- 3- a probabilidade de sucesso ( $p$ ) em cada repetição é constante;
- 4- consequentemente, a probabilidade de fracasso ( $q = 1-p$ ) em cada repetição também é constante; e,
- 5- o experimento aleatório prossegue até que sejam verificados  $r$  sucessos.

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 serão necessários para se observar  $r$  *Caras*?

Se arbitramos  $r = 3$  e observarmos a sequência:  $\{\text{Cara}, \text{Coroa}, \text{Coroa}, \text{Cara}, \text{Coroa}, \text{Coroa}, \text{Cara}\}$ , então  $n = 7$ : foram necessárias sete repetições até que três *Caras* fosse observadas.

A notação de uma variável aleatória Binomial Negativa é  $X \sim bn(p, r)$ , onde o parâmetro  $p$  ( $0 \leq p \leq 1$ ) indica a probabilidade individual de sucesso a cada repetição de Bernoulli e  $r$  o número total de sucessos desejado (estabelecido *a priori*).

Sua função discreta de probabilidade calcula a probabilidade de se observar um total de  $r$  sucessos (estabelecido *a priori*) após  $n$  de ensaios de Bernoulli realizados é a seguinte:

$$f(X = x; p; r) = P(X = n) = C_{r-1}^{n-1} \cdot p^r \cdot q^{(n-r)}$$

$$f(X = x; p; r) = \frac{(n-1)!}{(r-1)!.(n-r-2)!} \cdot p^r \cdot q^{(n-r)}$$

Pela razão óbvia de se necessitar no mínimo  $r$  tentativas para se obter  $r$  sucessos, a faixa de  $x = n = r, r + 1, r + 2\dots$ .

A esperança e a variância de uma variável aleatória discreta com Distribuição Binomial Negativa são:

- Esperança:  $E(X) = \frac{r}{p}$  ;
- Variância:  $Var(X) = \frac{r \times (1-p)}{p^2} = \frac{q \times r}{p^2}$ .

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 aleatório e em  $n$  tentativas a probabilidade de se observar  $k$  sucessos é medida pela sua função de distribuição discreta de probabilidades.

Uma variável aleatória Binomial Negativa é uma contagem do número de tentativas até se obter  $r$  sucessos; isto é, o número de sucessos  $r$  é predeterminado (fixo) e o número de tentativas é aleatório e a probabilidade de se observar  $r$  sucessos a cada  $n$  tentativas é calculada por sua função de distribuição discreta de probabilidades.

**Exemplo:** 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  $X \sim bn(p, r)$  onde  $p$  é a probabilidade individual de sucesso e  $r$  o total de sucessos.

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 *a priori*  $r = 4$ .

Pede-se a probabilidade de se observar **quatro** sucessos ( $r = 4$ ) em **dez** ( $n = 10$ ) transmissões.

A probabilidade de se obter  $r = 4$  sucessos ao se realizar  $n = 10$  tentativas é dada pela função discreta de probabilidade da variável aleatória Binomial Negativa:

$$\begin{aligned} f(X = x; p; r) &= P(X = n) = C_{r-1}^{n-1} \cdot p^r \cdot q^{n-r} \\ f(X = x; p; r) &= P(X = n) = \frac{(n-1)!}{(r-1)!.(n-r-2)!} \cdot p^r \cdot q^{n-r} \\ f(X = 10; p = 0,10; r = 4) &= P(X = 10) = \frac{(10-1)!}{(4-1)!.(10-4-2)!} \cdot 0,1^4 \cdot 0,9^{10-4} \\ P(X = 10) &= 0,004464104 \end{aligned}$$

A probabilidade de se observar 4 sucessos em 10 tentativas é de 0,4464104%.

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  $X \sim bn(p, r)$  onde  $p$  é a probabilidade individual de sucesso e  $r$  o total de sucessos.

Dados do problema:

- 1- a probabilidade de ocorrência de um sucesso é  $p = 0,70$ , e
- 2- o número de sucessos fixado *a priori* é  $r = 3$ .

Pede-se a probabilidade de se observar três sucessos em 5 arremessos  $n = 5$ .

A probabilidade de se obter  $r = 3$  sucessos ao se realizar  $n = 5$  tentativas é dada pela função discreta de probabilidade da variável Binomial Negativa:

$$\begin{aligned}
 f(X = x; p; r) &= P(X = n) = C_{r-1}^{n-1} \cdot p^r \cdot q^{n-r} \\
 f(X = x; p; r) &= P(X = n) = \frac{(n-1)!}{(r-1)!.(n-r-2)!} \cdot p^r \cdot q^{n-r} \\
 f(X = 5; p = 0,70; r = 3) &= P(X = 5) = \frac{(5-1)!}{(3-1)!.(5-3-2)!} \cdot 0,70^3 \cdot 0,9^{5-3} \\
 P(X = 5) &= 0,18522
 \end{aligned}$$

A probabilidade de Bob acertar 3 arremessos em 5 tentativas é de 18,522%.

{Exemplo: 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 fenômeno como  $X \sim bn(p, r)$  onde  $p$  é a probabilidade individual de sucesso e  $r$  o total de sucessos.

Dados do problema:

- a probabilidade de ocorrência de um sucesso é  $p = 0,5$ , e,
- o número de sucessos fixado *a priori* é  $r = 7$ .

Pede-se a probabilidade de se observar sete sucessos em doze ( $5+7$ ) tentativas  $n = 12$ .

A probabilidade de se obter  $r = 7$  sucessos ao se realizar  $n = 12$  tentativas é dada pela função discreta de probabilidade da variável Binomial Negativa:

$$\begin{aligned}
 f(X = x; p; r) &= P(X = n) = C_{r-1}^{n-1} \cdot p^r \cdot q^{n-r} \\
 f(X = x; p; r) &= P(X = n) = \frac{(n-1)!}{(r-1)!.(n-r-2)!} \cdot p^r \cdot q^{n-r} \\
 f(X = 5; p = 0,50; r = 7) &= P(X = 12) = \frac{(12-1)!}{(7-1)!.(12-7-2)!} \cdot 0,50^7 \cdot 0,50^{12-7} \\
 P(X = 5) &= 0,1128
 \end{aligned}$$

A probabilidade de se obter 7 sucessos em 12 tentativas é de 11,28%.

Exemplo: 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  $X \sim bn(p, r)$  onde  $p$  é a probabilidade individual de sucesso e  $r$  o total de sucessos.

Dados do problema:

- probabilidade de que a câmera montada no celular passe no teste é  $p = 0,80$ ; logo, a probabilidade de não passar será de  $(q = 1 - 0,80) = 0,20$ ;
- fica bem entendido 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*  $r = 2$  em  $n = 10$ ;
- no item (2) pede-se a probabilidade de se observar um número de sucessos também fixado *a priori* em  $r = 2$  mas agora em  $n \leq 4$  câmeras testadas; e,
- o valor esperado para o número de câmeras testadas ( $n = ?$ ) para que se observem  $r = 3$  sucessos.

A probabilidade de se obter  $r = 2$  sucessos ao se realizar  $n = 10$  tentativas é dada pela função discreta de probabilidade da variável Binomial Negativa:

$$\begin{aligned} f(X = x; p; r) &= P(X = n) = C_{r-1}^{n-1} \cdot p^r \cdot q^{n-r} \\ f(X = 10; p = 0,20; r = 2) &= P(X = 10) = C_{2-1}^{10-1} \cdot 0,20^2 \cdot 0,80^{10-2} \\ P(X = 10) &= 0,06039 \end{aligned}$$

A probabilidade de se obter  $r = 2$  sucessos em  $n = 10$  tentativas é de 6,039%.

As probabilidades de se obter  $r = 2$  sucessos ao se realizar  $n \leq 4$  tentativas é dada pela função discreta de probabilidade da variável Binomial Negativa aplicada a:

$$\begin{aligned}
 f(X = x; p; r) &= P(X = n) = C_{r-1}^{n-1} \cdot p^r \cdot q^{n-r} \\
 f(X = 2; p = 0,20; r = 2) &= P(X = 2) = C_{2-1}^{2-1} \cdot 0,20^2 \cdot 0,80^{2-2} \\
 P(X = 2) &= 0,04
 \end{aligned}$$

$$\begin{aligned}
 f(X = x; p; r) &= P(X = n) = C_{r-1}^{n-1} \cdot p^r \cdot q^{n-r} \\
 f(X = 3; p = 0,20; r = 2) &= P(X = 2) = C_{2-1}^{3-1} \cdot 0,20^2 \cdot 0,80^{3-2} \\
 P(X = 3) &= 0,064
 \end{aligned}$$

$$\begin{aligned}
 f(X = x; p; r) &= P(X = n) = C_{r-1}^{n-1} \cdot p^r \cdot q^{n-r} \\
 f(X = 4; p = 0,20; r = 2) &= P(X = 2) = C_{2-1}^{4-1} \cdot 0,20^2 \cdot 0,80^{4-2} \\
 P(X = 4) &= 0,0768
 \end{aligned}$$

A probabilidade de se obter  $r = 2$  sucessos em  $n \leq 4$  tentativas é de  $(0,032 + 0,064 + 0,0768)$  18,08%.

O valor esperado (esperança) do número de câmeras testadas para que se observem  $r = 3$  sucessos é dado

$$\begin{aligned}
 E(X) &= \frac{r}{p} \\
 E(X) &= \frac{3}{0,2} \\
 &= 15
 \end{aligned}$$

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 6.2

$$f(X = x) = \begin{cases} 2x & \text{para } 0 \leq x \leq 1 \\ 0, & \text{para qualquer outro } x \end{cases}$$

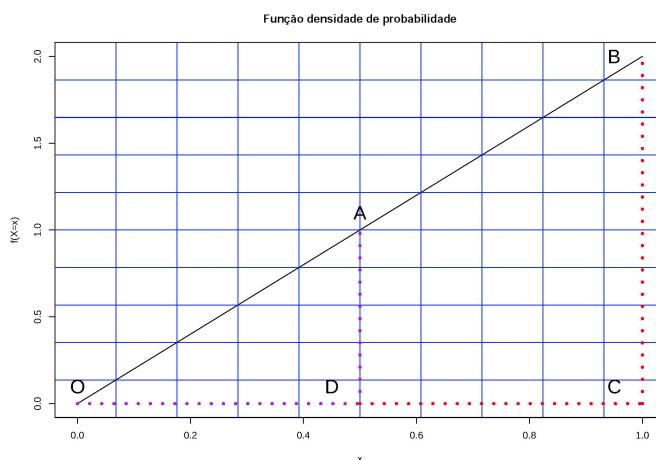


Figure 6.2: A área definida por (ODA) equivale à probabilidade de  $f(X = x)$  no intervalo  $0 \leq x \leq 0,50$  é notadamente menor que a área definida por (ABCD) equivalente à probabilidade de  $f(X = x)$  no intervalo  $0,5 \leq x \leq 1$ . Tendo os intervalos  $[0;0,50]$  e  $[0,50; 1,00]$  igual 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:  $Var(X) = \frac{(b-a)^2}{12}$ .

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:

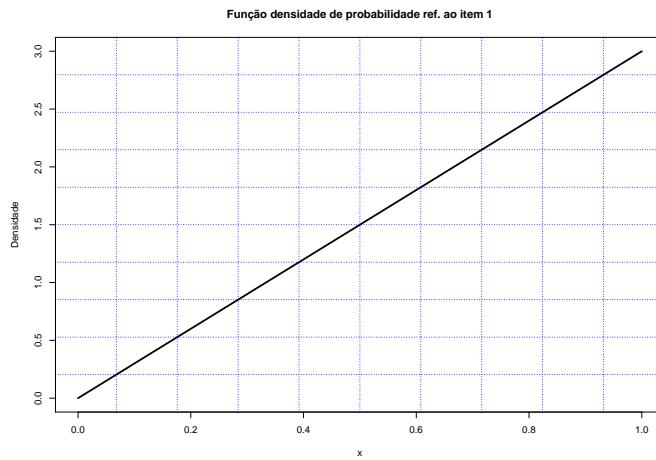


Figure 6.3: A área definida por  $f(x)$  no intervalo  $0 \leq x \leq 1$  é maior que 1. Por essa razão não pode ser uma fdp

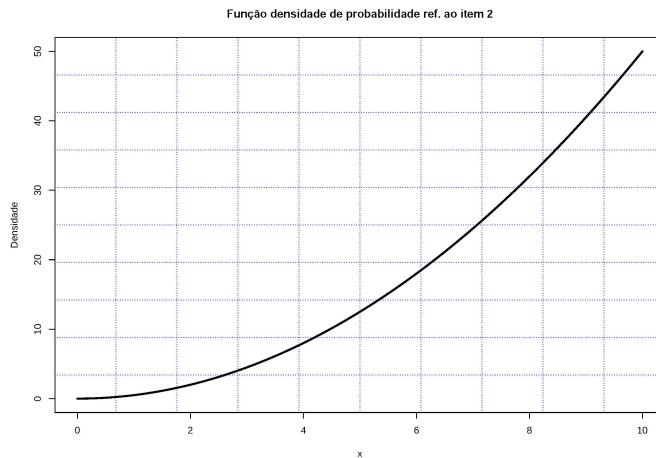


Figure 6.4: A área definida por  $f(x)$  no intervalo  $x \geq 0$  é maior que 1. Por essa razão não pode ser uma fdp

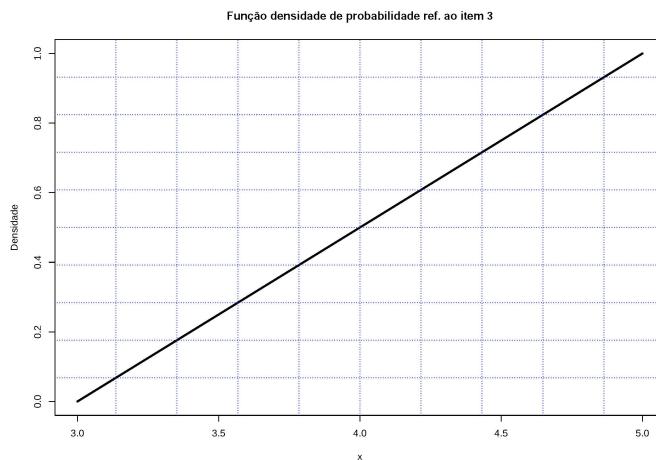


Figure 6.5: Os valores assumidos por  $f(x)$  são  $\geq 0$  e a área definida por  $f(x)$  o intervalo  $3 \leq x \leq 5$  é igual a 1. Por essa razão pode ser uma fdp

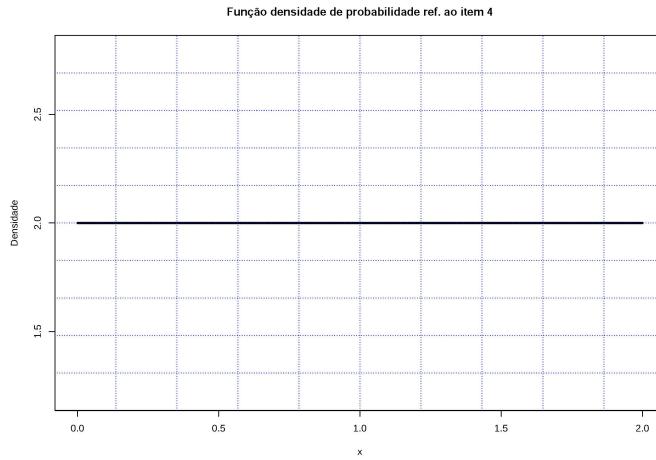


Figure 6.6: A área definida por  $f(x)$  no intervalo  $0 \leq x \leq 2$  é maior que 1. Por essa razão não pode ser uma fdp

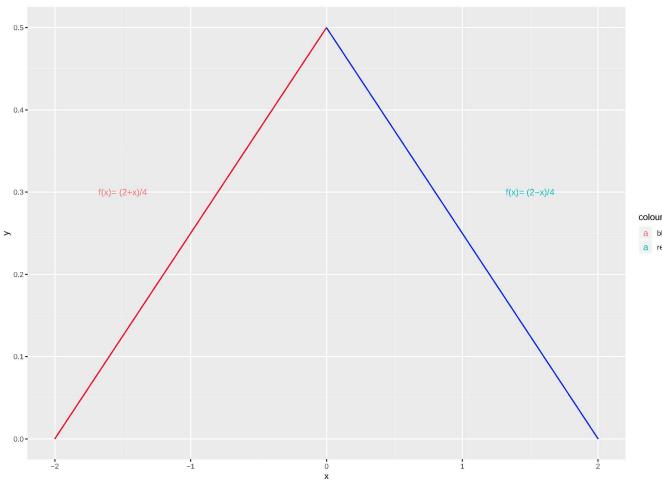


Figure 6.7: Os valores assumidos por  $f(x)$  são  $\geq 0$  e a área definida por  $f(x)$  nos intervalos  $-2 \leq x \leq 0$  e  $0 \leq x \leq 2$  é igual a 1. Pode ser uma fdp

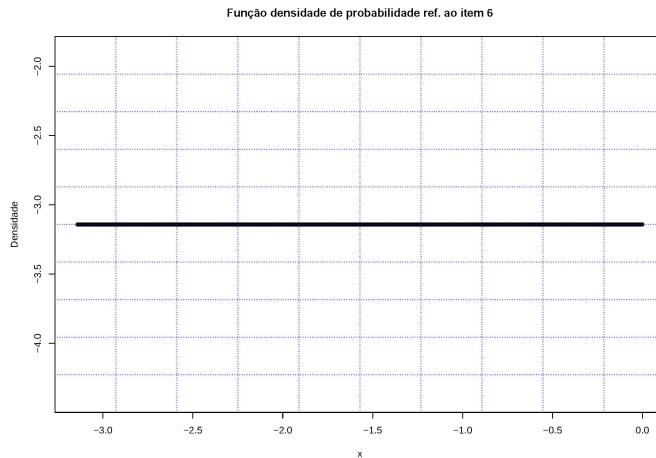


Figure 6.8: Os valores assumidos por  $f(x)$  são  $< 0$ . Por essa razão não pode ser uma fdp.

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:  $Var(X) = \frac{(70-50)^2}{12} = 33,33$ .

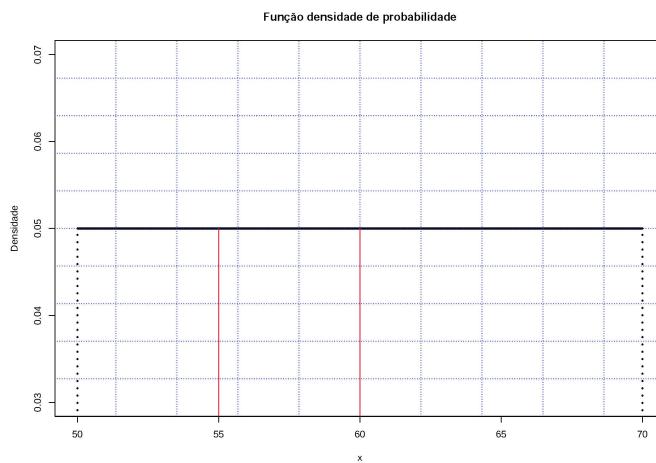


Figure 6.9: Os valores assumidos por  $f(x)$  são  $\geq 0$  e a área definida por  $f(x)$  no intervalo  $50 \leq x \leq 70$  é igual a 1. Por essa razão pode ser uma fdp. A probabilidade pedida equivale à área  $P(60 \leq x \leq 55) = (60 - 55).0,05 = 0,25$ .

### 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  $\lambda$ :  $X \sim Exp(\lambda)$ . Há duas parametrizações habituais.

Primeira parametrização:  $\lambda > 0$ : taxa e sua densidade de probabilidade é dada por:

$$f(X = x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & \text{para } x \geq 0 \\ 0, & \text{para } x < 0 \end{cases}$$

Segunda parametrização:  $\alpha = \frac{1}{\lambda}$ : escala 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 (esperança da variável).

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:

$$f(T = t) = \begin{cases} 0,20 \cdot e^{-0,20 \cdot t}, & \text{para } t \geq 0 \\ 0, & \text{para } x < 0 \end{cases}$$

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).

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 ( $\mu$ ) 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:

$$f(T = t) = \begin{cases} \frac{1}{100} \cdot \varepsilon^{-\frac{1}{100} \cdot t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

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]:

$$\begin{aligned} P(T > 150) &= 1 - P(0 < T < 150) = 1 - \int_0^{150} \alpha \cdot \varepsilon^{-\alpha \cdot x} dx \\ &= 1 - \varepsilon^{-\alpha \cdot x} \Big|_0^{150} \\ &= 1 - (\varepsilon^{-0,01 \cdot 0} - \varepsilon^{-0,01 \cdot 150}) \\ &= 1 - (1 - 0,22313) \\ &= 0,22313 \end{aligned}$$

A probabilidade de um fusível ter uma vida útil maior que 150 horas é de 0,22313.

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]:

$$\begin{aligned}
 P(T > 200) &= 1 - P(0 < T < 200) = 1 - \int_0^{200} \alpha \cdot e^{-\alpha \cdot x} dx \\
 &= 1 - e^{-\alpha \cdot x} \Big|_0^{200} \\
 &= 1 - (e^{-0,01 \cdot 0} - e^{-0,01 \cdot 200}) \\
 &= 1 - (1 - 0,1353) \\
 &= 0,1353
 \end{aligned}$$

A probabilidade de um fusível ter uma vida útil maior que 200 horas é de 0,1353.

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.

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 Abraham 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  $\mu$  e variância  $\sigma^2$  distribuídos segundo uma Curva Gaussiana é denotada por  $X \sim N(\mu, \sigma^2)$ , e sua função densidade de probabilidade é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

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:

$$\begin{aligned} f(z) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \\ f(z) &= 0,3989e^{-5z^2} \end{aligned}$$

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:

$$\begin{aligned} F(z) &= P(Z \leq z) \\ P(Z \leq z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{u^2}{2}} du \end{aligned}$$

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:

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 6.10: Tabela Z mostrando a probabilidade ao intervalo  $[0 ; 1,64]$  (quadro superior à esquerda explica onde a área se encontra)

A tabela Z possibilita:

- 1- encontrar a probabilidade (área) partindo de *score z*; e
- 2- encontrar o *score z*.

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  $\mu$ . 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 ; inf)=0,0505",
      geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue", lty=1, 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), annotate(geom="text", x=0.1, y=0.1, label="Probabilidade (área) =0,4495", angle=0, vjust=0, hjust=0), annotate(geom="text", x=2, y=0.05, label="Probabilidade (área) =0,0505", angle=0, vjust=0, hjust=0), theme_bw()
```

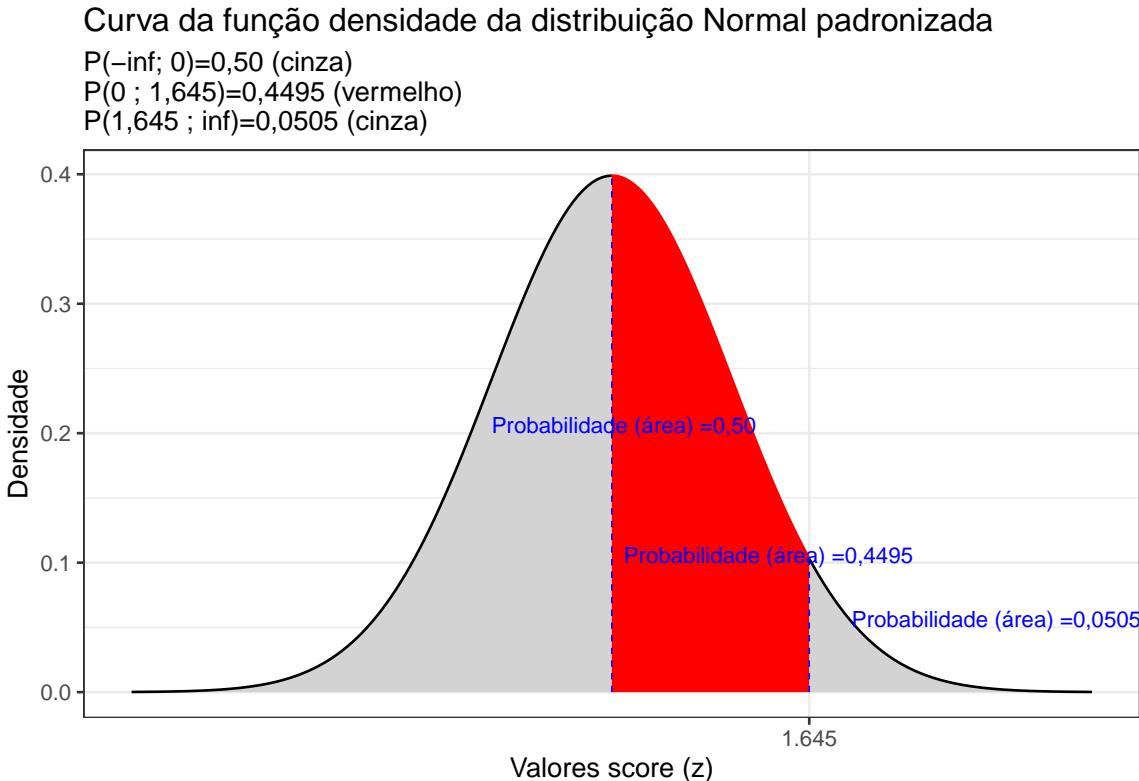


Figure 6.11: Curva da fdp da Distribuição Normal padronizada mostrando as áreas delimitadas pelo score z arbitrado (1,64)

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

$$P(X \geq 96,52) \equiv \text{rea}[96,52; +\infty]$$

**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:

$$\begin{aligned} X_1 &= 83,82 \\ Z_n &= \frac{X_n - \mu}{\sigma} \\ z_1 &= -1,40 \end{aligned}$$

E a probabilidade pedida equivale (graficamente) à área situada no intervalo  $[-\infty; -1,40]$  na curva da fdp distribuição Normal padronizada:

$$P(X \leq 83,82) = P(Z \leq -1,40) \equiv \text{rea}[-\infty; -1,40]$$

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 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 6.12 é a probabilidade pedida:

$$\begin{aligned} P(X \leq 83,82) &= 0,50 - 0,4192 \\ P(X \leq 83,82) &= 0,0808 \end{aligned}$$

```

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 ; inf)=0,5",
    geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue", lty=1, lwd=0.5)+
    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

$P(-\infty; -1,40) = 0,0808$  (vermelho)

$P(-1,40 ; 0) = 0,4192$  (cinza)

$P(0 ; \infty) = 0,50$  (cinza)

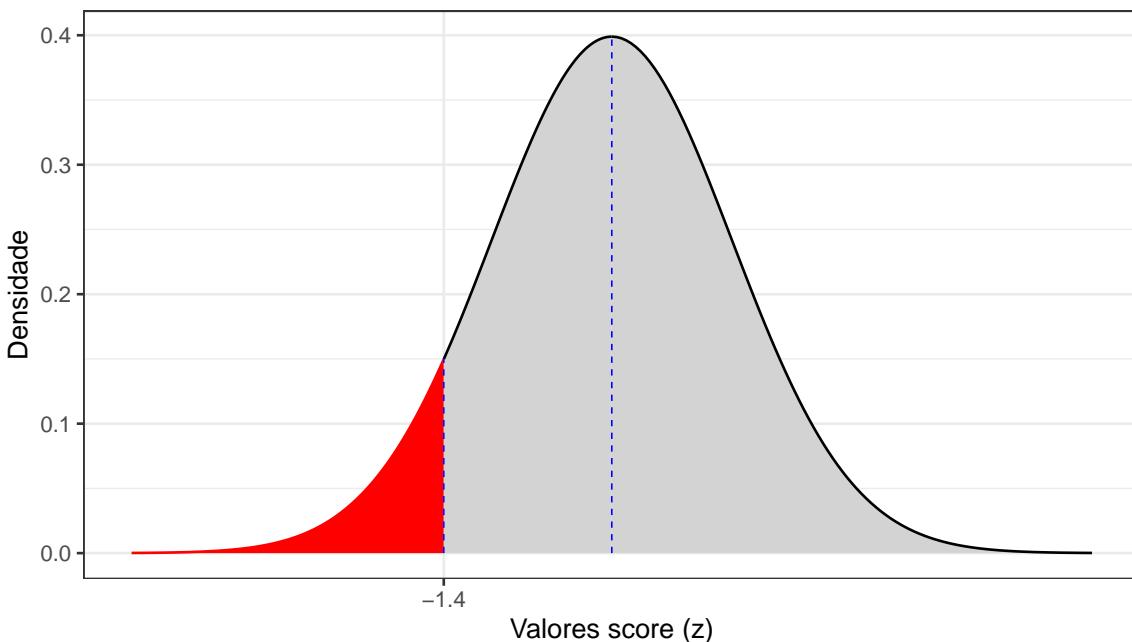


Figure 6.12: Curva da fdp da Distribuição Normal padronizada mostrando as áreas delimitadas pelo score z calculado (-1,40)

De modo análogo para a segunda questão 96,52 mm/ano) teremos:

$$\begin{aligned} X_2 &= 96,52 \\ Z_n &= \frac{X_n - \mu}{\sigma} \\ z_2 &= -0,40 \end{aligned}$$

E a probabilidade pedida equivale (graficamente) à área situada no intervalo  $[-0,40 ; \infty]$  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 6.13 é a probabilidade pedida:

$$P(X \geq 96,52) = 0,50 + 0,4192 = 0,6554$$

```
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 ; inf)=0,50",
    geom_segment(aes(x = z_desejado, y = 0, xend = z_desejado, yend = d_desejada), color="blue", lty=1, lwd=0.5),
    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**

$$\begin{aligned} P(-\infty; -0,40) &= 0,3446 \text{ (cinza)} \\ P(-0,40 ; 0) &= 0,1554 \text{ (vermelho)} \\ P(0 ; \infty) &= 0,50 \text{ (vermelho)} \end{aligned}$$

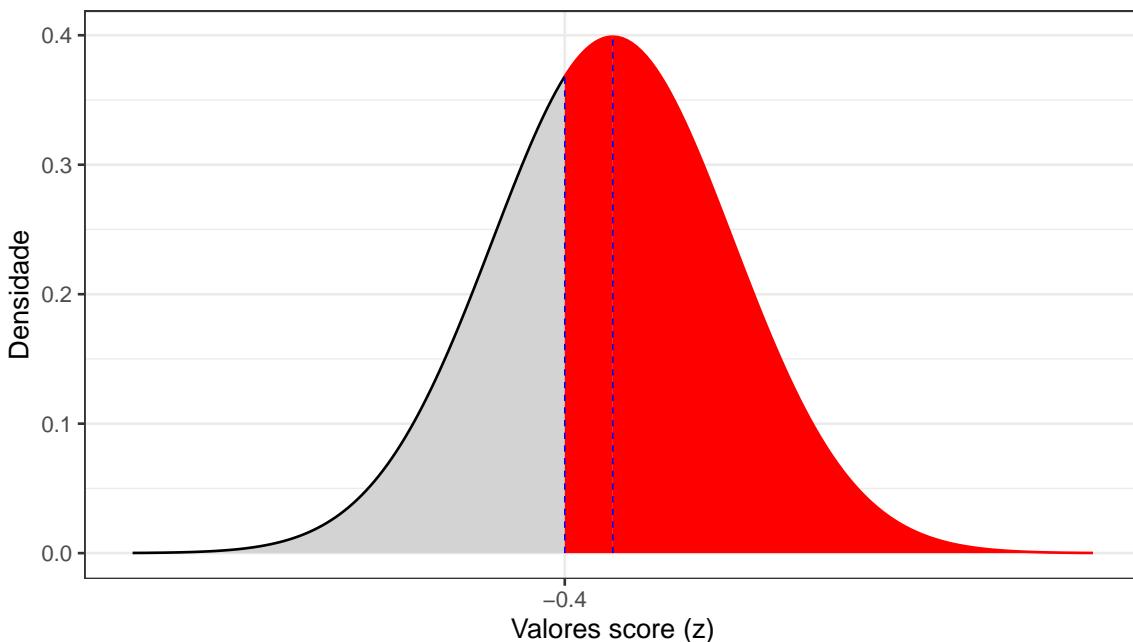


Figure 6.13: Curva da fdp da Distribuição Normal padronizada mostrando as áreas delimitadas pelo score z calculado (-0,40)

#### 6.3.4 Student “t”

Se uma variável aleatória  $T$  contínua com  $\nu$  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  $\sigma^2$  deseja-se estimar a média  $\mu$ .

Para grandes amostras ( $n \geq 30$ ) a distribuição amostral de  $\bar{X}$  é aproximadamente Normal, com média  $\mu$  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:  $\sigma$ ).

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**:  $\nu$ .

Toda estatística de teste que dependa de uma variável aleatória possui graus de liberdade ( $\nu$ ). 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  $\mu$  através da média amostral  $\bar{X}$ ; todavia, tivemos também que estimar sua variância  $\sigma^2$  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 como:

- ao intervalo fechado:  $[-t ; +t]$  (Figura 6.15);
- o intervalo aberto à esquerda:  $[-\infty ; t]$  (Figura 6.16); e,
- o intervalo aberto à direita:  $[t, \infty]$  (Figura 6.17).

Nas linhas horizontais lê-se os graus de liberdade  $\nu$  e nas colunas as áreas (probabilidades).

		Distribuição t de Student											
		Área contida nas duas caudas laterais (bicaudal) da distribuição t de Student											
gl/q	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	2,3534	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	2,2222	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 6.14: 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 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  $\mu = 0$ .

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",
       geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2), color="blue",
       annotate(geom="text", x=-0.1, y=0.2, label="Probabilidade (área) =0,90 \n(gl=10)", angle=0, vjust=0),
       annotate(geom="text", x=-3.5, y=0.1, label="Probabilidade (área) =0,05 \n(gl=10)", angle=0, vjust=0)
```

```
annotate(geom="text", x=2.5, y=0.1, label="Probabilidade (área) =0,05 \n(gl=10)", angle=0, vjust=0)
theme_bw()
```

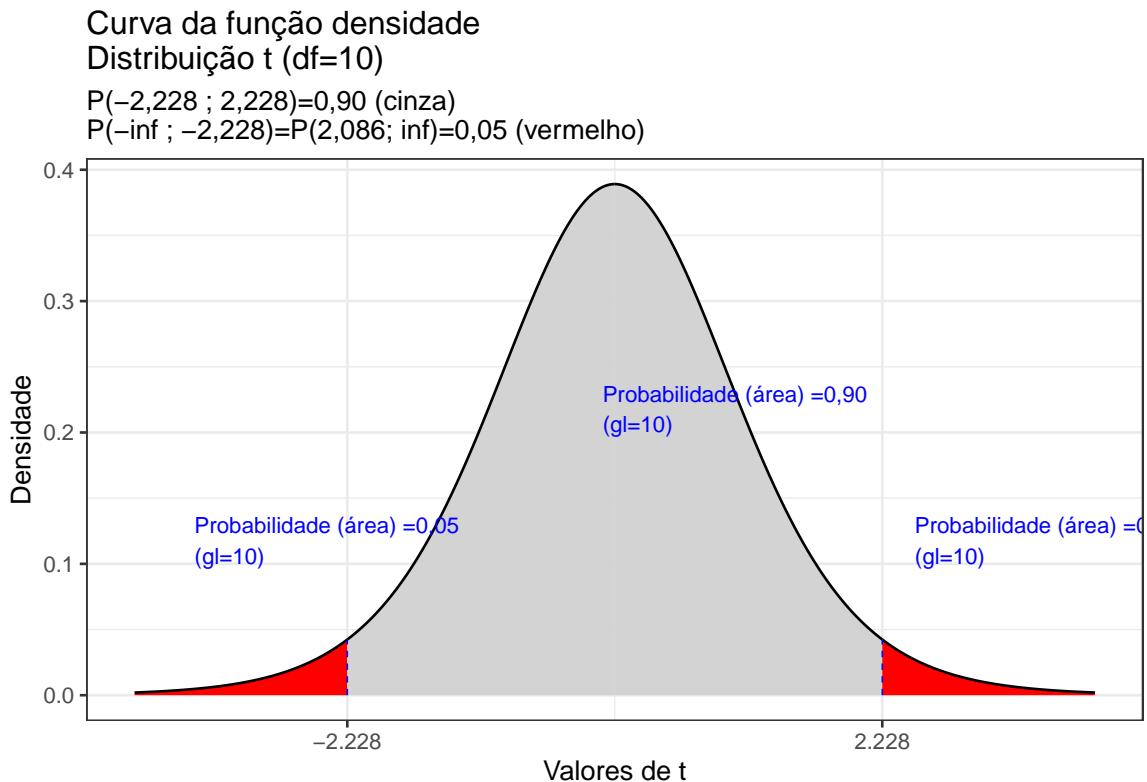


Figure 6.15: Curva da fdp da Distribuição Student para 10 graus de liberdade, mostrando as áreas delimitadas pelos valores  $+/-t$  ( $+/-2,28$ )

```
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),
            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=1, vj)
annotate(geom="text", x=-0.1, y=0.2, label="Probabilidade (área) =0,975 \n(gl=10)", angle=0, vj)
annotate(geom="text", x=-3.5, y=0.1, label="Probabilidade (área) =0,025 \n(gl=10)", angle=0, vj)
theme_bw()

```

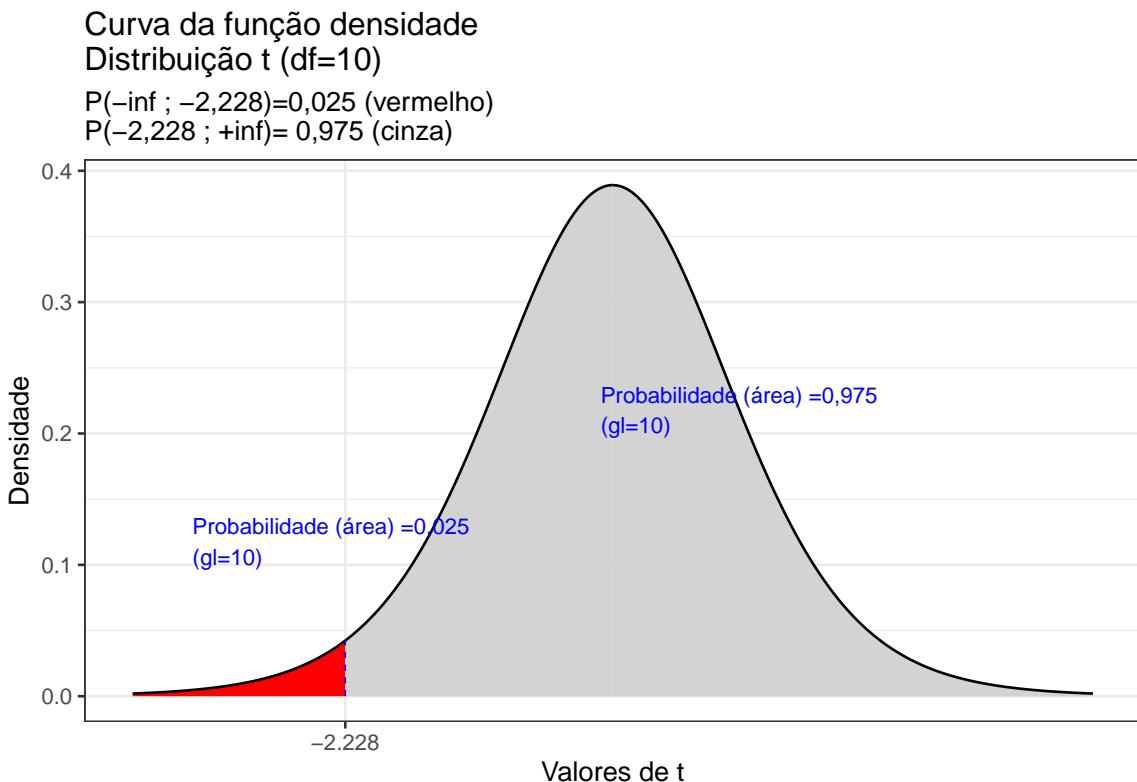


Figure 6.16: Curva da fdp da Distribuição Student para 10 graus de liberdade, mostrando as áreas delimitadas pelo valor  $-t$  (-2,28)

```

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=1)
annotate(geom="text", x=0, y=0.2, label="Probabilidade (área) =0,975 \n(gl=10)", angle=0, vjust=-1)
annotate(geom="text", x=2.5, y=0.1, label="Probabilidade (área) =0,025 \n(gl=10)", angle=0, vjust=1)
theme_bw()

```

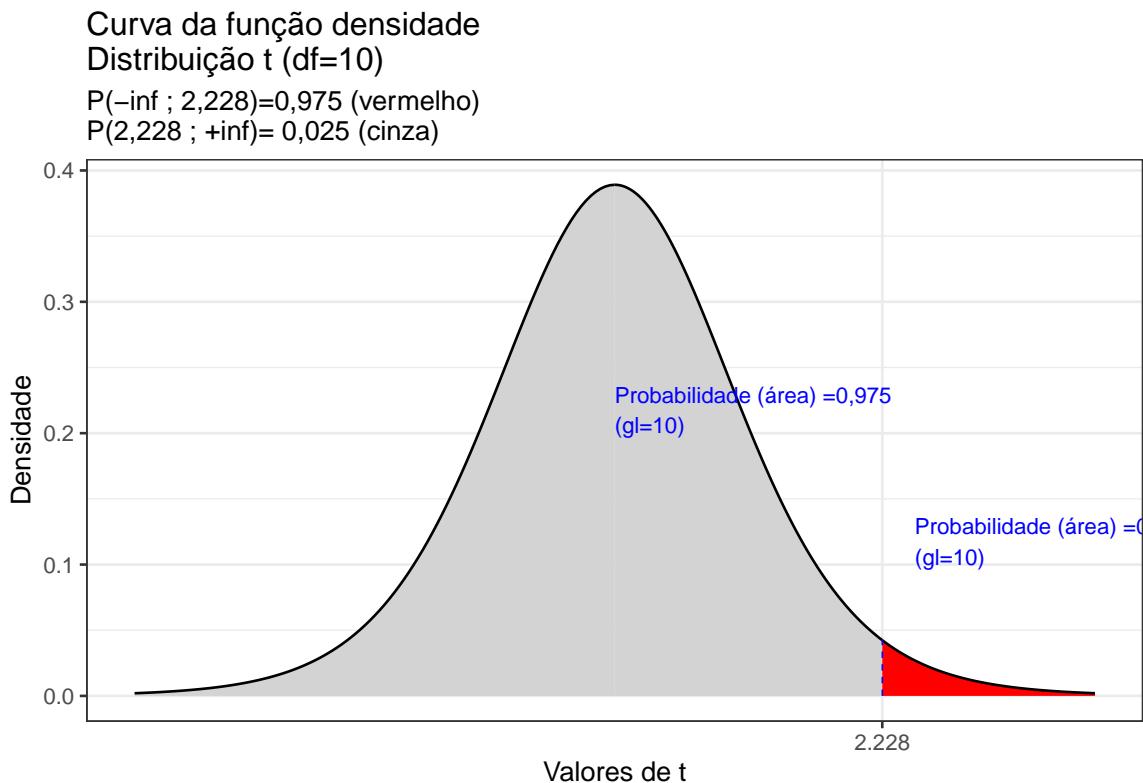


Figure 6.17: Curva da fdp da Distribuição Student para 10 graus de liberdade, mostrando as áreas delimitadas pelo valor  $-t$  (-2,28)

### 6.3.5 Qui-Quadrado

Considerem  $X_1, X_2, \dots, X_\nu$  como  $\nu$  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  $\chi^2$  possui seguinte fdp para  $x > 0$  (para  $x \leq 0, f(x) = 0$ ), com  $\nu$  graus de liberdade:

$$f(x) = \frac{1}{2^{\frac{\nu}{2}} \cdot \Gamma\left(\frac{\nu}{2}\right)} \cdot x^{\left(\frac{\nu}{2}\right)-1} e^{-\frac{x}{2}}$$

A função de probabilidade cumulativa é dada por:

$$P(\chi^2 \leq x) = \frac{1}{2^{\frac{\nu}{2}} \cdot \Gamma\left(\frac{\nu}{2}\right)} \int_{-\infty}^x u^{\left(\frac{\nu}{2}\right)-1} e^{-\frac{u}{2}} du$$

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  $\nu_1$  e  $\nu_2$  graus de liberdade, essa variável também seguirá a Distribuição Qui-quadrado com  $\nu_1 + \nu_2$  graus de liberdade
- É assimétrica e definda 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  $\nu_1$  e  $\nu_2$ , 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  $\nu_1$  e  $\nu_2$  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										
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

Parte inteira e primeira casa decimal de Z

 Professor Guru

professorguru.com.br

Figure 6.18: Tabela de valores “z” da Distribuição Normal padronizada

**AulasdeMatemática.com.br**  
Matemática | Estatística | Mat. Financeira | Rac. Lógico-Quantitativo

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 - USP  
Bach. Estatística - USP

**Distribuição t de Student**

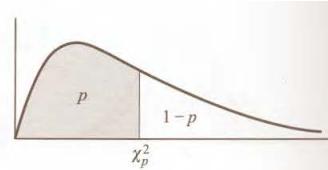
gl/q	Área contida nas duas caudas laterais (bicaudal) da distribuição t de Student											
	0,990	0,980	0,975	0,950	0,900	0,800	0,200	0,100	0,050	0,025	0,020	0,010
	0,995	0,990	0,9875	0,975	0,950	0,900	0,100	0,050	0,025	0,0125	0,010	0,005
1	0,0157	0,0314	0,0393	0,0787	0,1584	0,3249	3,0777	6,3138	12,7062	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	2,3534	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
a	0,0127	0,0254	0,0317	0,0635	0,1272	0,2566	1,3232	1,7207	2,0796	2,4138	2,5176	2,8314
21	0,0127	0,0254	0,0317	0,0635	0,1272	0,2566	1,3212	1,7171	2,0739	2,4055	2,5083	2,8188
22	0,0127	0,0254	0,0317	0,0634	0,1271	0,2564	1,3195	1,7139	2,0687	2,3979	2,4999	2,8073
s	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
d	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
e	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
l	0,0126	0,0253	0,0316	0,0633	0,1268	0,2557	1,3114	1,6991	2,0452	2,3638	2,4620	2,7564
i	0,0126	0,0253	0,0316	0,0632	0,1267	0,2556	1,3104	1,6973	2,0423	2,3596	2,4573	2,7500
b	0,0126	0,0253	0,0316	0,0632	0,1267	0,2555	1,3095	1,6955	2,0395	2,3556	2,4528	2,7440
e	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	0,1268	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
a	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
d	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.

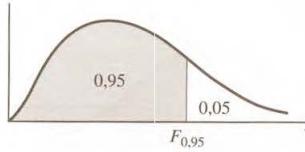
Figure 6.19: Tabela de valores “t” da Distribuição de Student

# Percentis $\chi_p^2$ da Distribuição Qui-Quadrado com $v$ Graus de Liberdade



$v$	$\chi_{0,005}^2$	$\chi_{0,01}^2$	$\chi_{0,025}^2$	$\chi_{0,05}^2$	$\chi_{0,10}^2$	$\chi_{0,25}^2$	$\chi_{0,50}^2$	$\chi_{0,75}^2$	$\chi_{0,90}^2$	$\chi_{0,95}^2$	$\chi_{0,975}^2$	$\chi_{0,99}^2$	$\chi_{0,995}^2$	$\chi_{0,999}^2$
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 6.20: 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	$\infty$
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	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
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
$\infty$	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 6.21: 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:

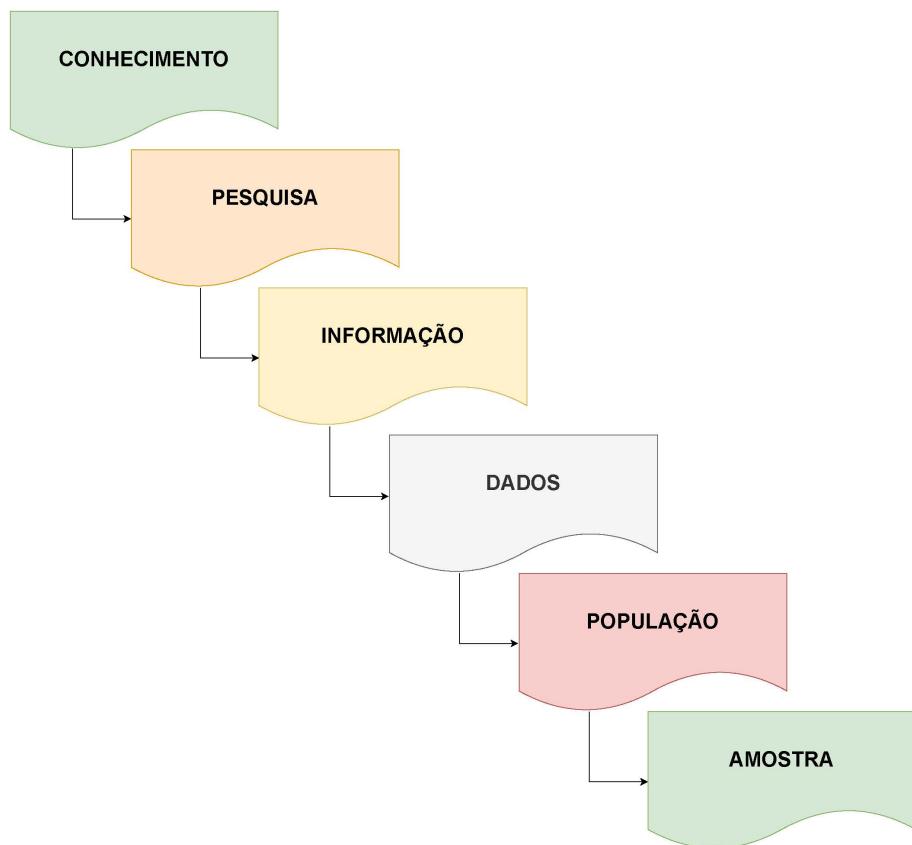


Figure 7.1: Representação esquemática do fluxo de informações da amostra à produção de conhecimento

Exemplo: Crianças socialmente isoladas assistem mais televisão do que crianças bem integradas a seus grupos?

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 cientificamente 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 permitem 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;
- em 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;
- Análise 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 ??:

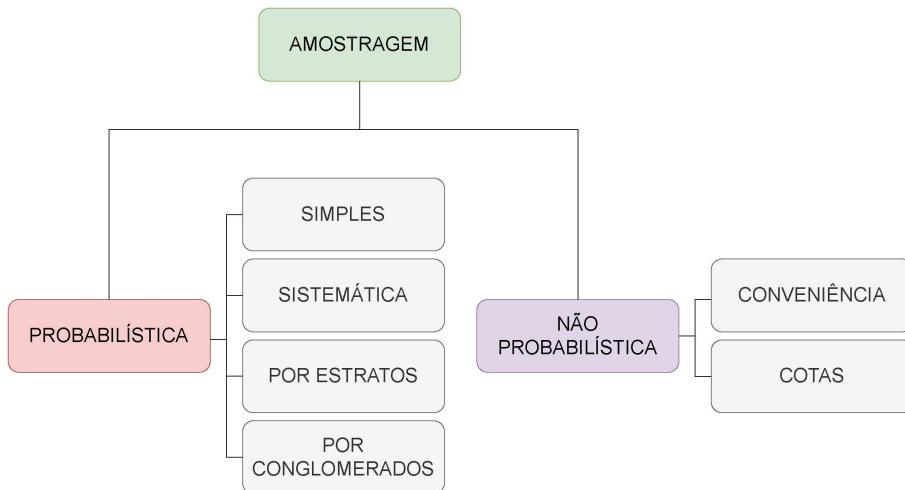


Figure 7.2: Principais procedimentos para se extrair uma amostra

## 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; isto é, a probabilidade de um elemento ser incluído na amostra é igual para todos.

Os elementos da população têm probabilidade conhecida e diferente de zero de serem selecionados para amostra (mas não necessariamente a mesma probabilidade).

A aleatorização visa assegurar que a informação extraída da amostra possa ser generalizada na população de origem.

### 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 deseje 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" "b" "e"

# Extração de uma amostra (com reposição) composta por 3 elementos do conjunto:
amostra_cr=sample(conjunto, 3, replace=TRUE)
amostra_cr

## [1] "a" "d" "e"
```

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

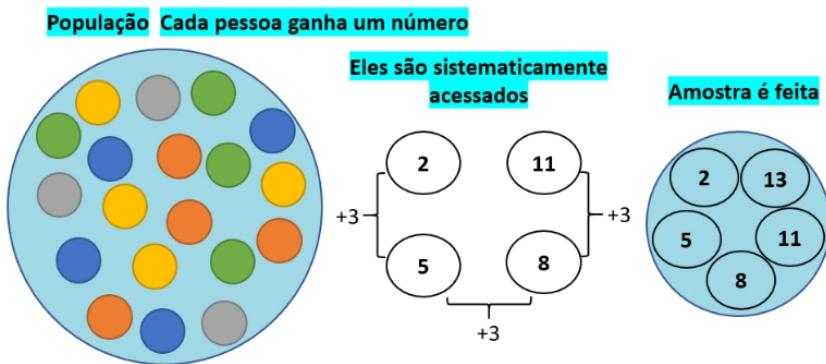


Figure 7.4: Amostra sistemática

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 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**.

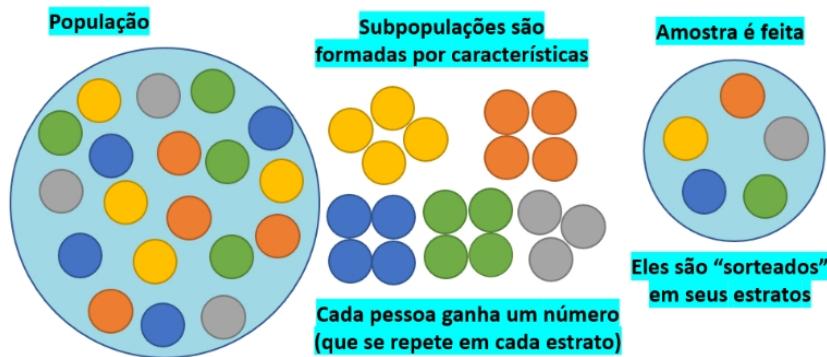


Figure 7.5: Amostra estratificada

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.

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 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  $n = n_1 + n_2 + \dots + n_k$ .

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 ( $N$ ) 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 extrair 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;

A proporção extraída de cada um dos estratos é constante:

$$\frac{n_p}{N_p} = \frac{n_f}{N_f} = \frac{n_e}{N_e} = 0,1125$$

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,
- $\sigma_i$  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  $\sigma_i$ ) 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. Pela descrição da população (estudantes universitário) observa-se que algumas das variáveis que habitualmente implicam em opiniões diferentes (escolaridade e idade) já não mais precisam ser consideradas. Um plano de estratificação de vários níveis pode ser estabelecido partindo-se da premissa de homogeneidade 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 é ilustrado na Figura 7.6.

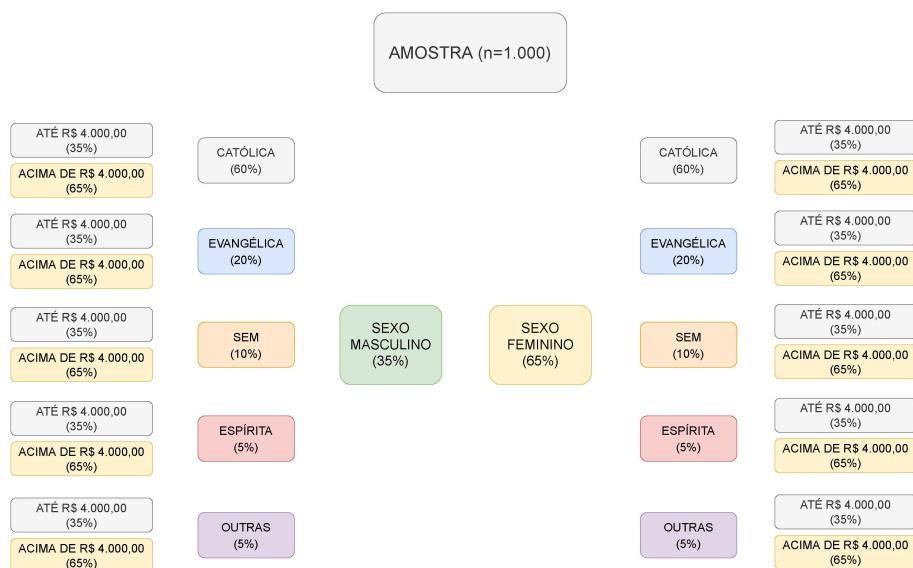


Figure 7.6: Plano de estratificação proporcional

#### 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 por 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.

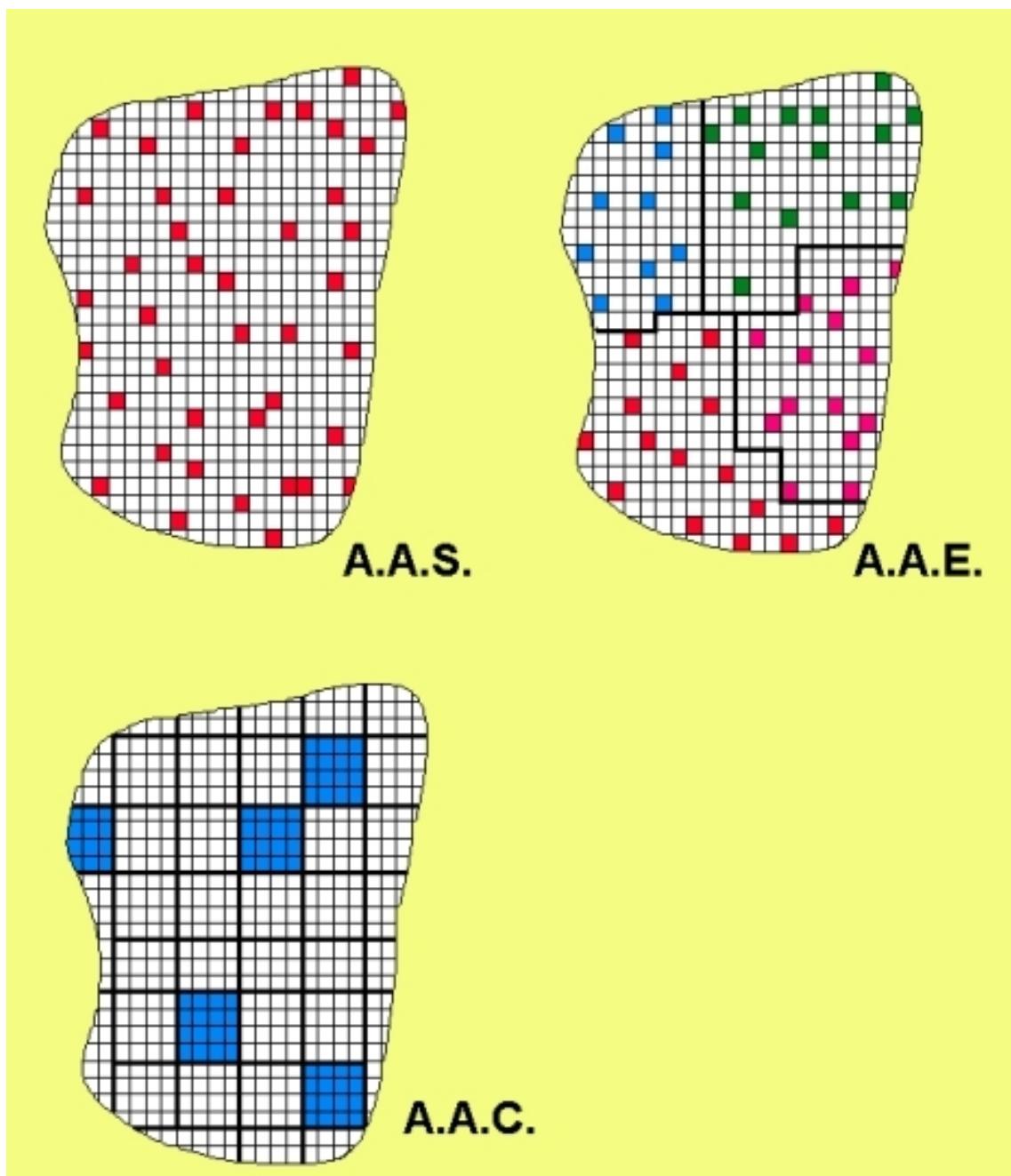


Figure 7.8: Ilustração comparativa dos principais modos de extração de amostras

## 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, inconsistentemente, 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*  $\alpha$  estabelecido *a priori* pelo pesquisador que se relaciona ao *nível de confiança* pretendido por  $(1 - \alpha)$ :

Table 7.2: Valores críticos de  $z_c$  correspondentes a alguns níveis de significância  $\alpha$ 

Níveis de significância	20%	10%	5%	1%	0,1 %
$z_c$	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** ( $N \leq 20.n_0$ ) e seguindo uma distribuição Normal:

$$n_0 = \left( \frac{z_c \cdot \sigma}{\varepsilon} \right)^2$$

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  $\alpha=5\%$ ,  $z_c = 1,96$ );
- $\sigma$  desvio padrão populacional obtido em estudos prévios; e,
- $\varepsilon$ : é o erro amostral, a máxima diferença entre  $\mu$  e  $\bar{x}$  que se decide tolerar.

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  $\sigma$  do peso para animais dessa idade é de 30 kg. Utilize um erro  $\varepsilon$  de 10 kg na estimativa e um nível de confiança  $(1 - \alpha)$  de 95%.

$$n_0 = \frac{Z^2 \cdot \sigma^2}{\varepsilon^2}$$

$$n_0 = \frac{1,96^2 \cdot 30^2}{10^2}$$

$$n_0 \sim 35$$

Se a população **não pode ser considerada infinita** ( $N \leq 20 \cdot n_0$ ) 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 ( $N \leq 20 \cdot n_0$ ) o tamanho da amostra seria:

$$n = \frac{N \cdot n_0}{N + n_0}$$

$$n = \frac{200 \cdot 35}{200 + 35}$$

$$n \sim 30$$

O conhecimento prévio do **desvio padrão populacional** ( $\sigma$ ) 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  $\sigma$ , 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:

$$n = \left( \frac{t_c \cdot s}{\varepsilon} \right)^2$$

com  $s$  calculado sobre a amostra piloto de  $n_0$  elementos e com  $t_c$  obtido em uma tabela considerando o nível de significância  $\alpha$  estabelecido e o tamanho da amostra piloto  $n_0$ .

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.

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	0,404	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	1,388	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	2,898	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	1,797	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	1,663	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 7.9: 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  $t_c$  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  $t_c = 1,96$  (distribuição de Student) e um valor crítico  $z_c = 1,96$  (distribuição Normal padrão).

### 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  $\varepsilon$  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  $\varepsilon$  é uma quantidade para **mais e para menos** da estimativa obtida de uma amostra de tamanho  $n$ , sob um certo nível de significância.

A expressão anterior considera que a variância populacional  $\sigma^2$  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  $\sigma^2$ ;
- se  $n < 30$ , adotar a variância amostral  $S^2$  como aproximação de  $\sigma^2$  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** ( $N \leq 20.n_0$ ):

$$n_0 = \frac{z_c^2 \cdot \pi \cdot (1 - \pi)}{\epsilon^2}$$

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  $\alpha=5\%$ ,  $z_c=1,96$ );
- $\pi$  é a proporção populacional;
- $\varepsilon$  é o erro amostral, a máxima diferença entre  $\pi$  e  $p$

Quando não se dispõe de nenhuma informação *a priori* sobre a proporção populacional ( $\pi$ ) a adoção do máximo valor possível ao produto:  $\pi \cdot (1 - \pi) = \frac{1}{4}$  assegura que o o tamanho de amostra obtido será suficiente para a estimativa qualquer que seja a proporção populacional  $\pi$ .

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** ( $N \leq 20n_0$ ) 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 de amostra ( $n_0$ ) suficiente para estimarmos a proporção da área com solo contaminado que necessita de certo tratamento de descontaminação, com precisão ( $\varepsilon$ ) 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 ( $\pi$ ), podemos exprimir o erro  $\varepsilon$  associado à estimativa da proporção ( $p$ ) obtida de uma amostra de tamanho  $n$  da proporção populacional, sob o critério mais conservador ( $\pi.(1 - \pi) = \frac{1}{4}$ )

$$\begin{aligned} \varepsilon &= z_c \cdot \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} \\ \varepsilon &= z_c \cdot \sqrt{\frac{\frac{1}{4}}{n}} \end{aligned}$$

em que  $\varepsilon$  é uma quantidade para **mais e para menos** da estimativa  $p$  obtida de uma amostra de tamanho  $n$  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* na proporção de eleitores a favor da liberação das pesquisas, a um nível de significância de 95% (46%; 54%).

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 Terminologia

- Epidemiologia

A epidemiologia é uma ciência médica que se concentra na distribuição e nos determinantes (fatores de risco) da frequência 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.

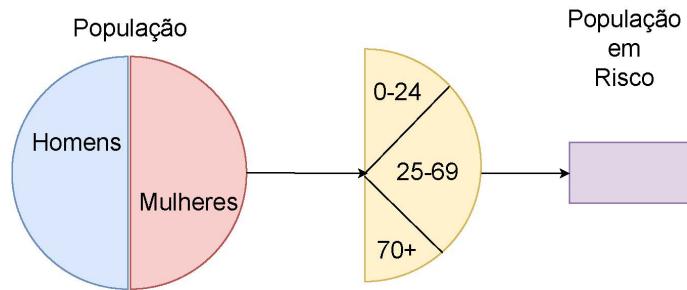


Figure 8.1: Adaptação: Basic Epidemiology: R. Bonita, R. Beaglehole, T Kjellström, 2006 (p. 17)

- 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 cardíaca coronariana, pois sabe-se que o consumo de café está associado com o uso de tabaco: as pessoas que bebem café são mais propensos 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.

## 8.2 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.

### 8.2.1 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:

$$I = \frac{\text{Número de novos casos de uma doença durante um determinado período de tempo}}{\text{Tamanho da população exposta ao risco nesse determinado período de tempo}} (\times 10^n)$$

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):

$$I = \frac{177 \text{ novos casos notificados de meningite no Maranhão em 2014}}{2.648.532 \text{ casos na população do Maranhão em 2014}} (\times 10^5) = \frac{4,41}{100.000}$$

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.

### 8.2.2 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:

$$P = \frac{\text{Número de casos existentes de doença em um determinado momento no tempo}}{\text{Tamanho da população em risco nesse mesmo momento no tempo}} (\times 10^n)$$

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:

$$P = \frac{89 \text{ casos de hipertensão na comunidade no dia 01/01/2014}}{3.500 \text{ indivíduos como população em risco na comunidade em 01/01/2014}} (\times 10^2) = \frac{2,54}{100}$$

Os dados sobre prevalência e incidência tornam-se muito mais úteis se convertidos em taxas!

### 8.2.3 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:

$$P = \text{Incidência} \times \text{Duração média da doença}$$

#### 8.2.4 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 <b>novos</b> casos de doença durante um determinado período de tempo	Número de casos <b>existentes</b> 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

#### 8.2.5 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:

$$IC = \frac{\text{Número de novos casos de uma doença durante um determinado período de tempo}}{\text{Tamanho da população em risco (exposta) livre (sem) da doença no começo de um determinado período de tempo}}$$

#### 8.2.6 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	-

### 8.2.7 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.

$$FC(\%) = \frac{\text{Número de mortes de casos diagnosticados da doença durante um determinado período de tempo}}{\text{Número de casos diagnosticados nesse período de tempo}} (\times 100)$$

## 8.3 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:

$$S = 1 - TM$$

### 8.3.1 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:

$$TM(\%) = \frac{\text{Número de mortes durante um determinado período de tempo}}{\text{Número de pessoas sob risco de morte nesse período de tempo}} (\times 10^n)$$

### 8.3.2 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.

## 8.4 Medidas de associação

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
	Presente	Ausente	
Exposto	(a)	(b)	(e)
Não exposto	(c)	(d)	(f)
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)

#### 8.4.1 Incidência observada de nascimentos com baixo peso entre mães expostas ao risco (fumantes)

$$\frac{(a)}{(e)} \times 100 = \frac{275}{2.419} \times 100 = 11,37\%$$

#### 8.4.2 Incidência observada de nascimentos com baixo peso entre mães não expostas ao risco (não fumantes)

$$\frac{(c)}{(g)} \times 100 = \frac{311}{4.807} \times 100 = 6,47\%$$

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

### 8.4.3 Prevalência de nascimentos com baixo peso na população estudada

$$\frac{(a) + (c)}{(e) + (g)} \times 100 = \frac{586}{7.226} \times 100 = 8,11\%$$

### 8.4.4 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.

Analizando-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:

$$\begin{aligned} RA &= \frac{(a)}{(e)} - \frac{(c)}{(f)} \\ &= \frac{275}{2.419} - \frac{311}{4.807} \\ &= 0,11368334 - 0,064697316 \\ &= 4,9\% \end{aligned}$$

### 8.4.5 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:

$$\begin{aligned}
 RR &= \frac{\frac{(a)}{(e)}}{\frac{(c)}{(f)}} \\
 &= \frac{\frac{275}{2.419}}{\frac{311}{4.807}} \\
 &= \frac{0,11368334}{0,064697316} \\
 &= 1,76
 \end{aligned}$$

#### 8.4.6 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.

Exemplo: tanto o tabagismo quanto a poluição do ar são causas de câncer de pulmão, mas a fração devida 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 é 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.

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:

$$\begin{aligned} FE &= \frac{\left( \frac{(a)}{(e)} - \frac{(c)}{(f)} \right)}{\frac{(c)}{(f)}} \\ &= \frac{\left( \frac{275}{2.419} - \frac{311}{4.807} \right)}{\frac{311}{4.807}} \\ &= \frac{(0,11368334 - 0,064697316)}{0,064697316} \\ &= 75,72\% \end{aligned}$$

Cerca de 75,72% dos casos de nascimentos de bebês com baixo peso é atribuível à exposição de mães ao fumo (mães fumantes).

#### 8.4.7 Odds ratio (Razão das chances)

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:

A chance (*odds*) de se observar o desfecho entre os expostos:

$$O_{exp} = \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{b}{a+b}\right)} = \frac{a}{b}$$

A chance (*odds*) de se observar o desfecho entre os não expostos:

$$O_{n.exp} = \frac{\left(\frac{c}{c+d}\right)}{\left(\frac{d}{c+d}\right)} = \frac{c}{d}$$

A razão das chances (*odds ratio* - OR) de exposição entre casos e controle:

$$OR = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)} = \frac{ad}{bc}$$

- OR (*odds ratio*) maior que 1: **fator de risco**;
- OR (*odds ratio*) menor que 1: **fator protetor**.

A razão de chances (*odds ratio*) exprime numericamente quantas vezes a exposição a um determinado fator de risco implica na possibilidade do desfecho estudado.

Analisando-se as incidências na Tabela vemos que a razão de chances é de:

$$\begin{aligned}
 OR &= \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)} \\
 &= \frac{\left(\frac{275}{2144}\right)}{\left(\frac{311}{4496}\right)} \\
 &= \frac{0,1282649}{0,0691726} \\
 &= 1,8542
 \end{aligned}$$

Uma razão de chances de 1,85 indica que uma gestante fumante terá 1,85 mais chances de ter um bebê com baixo peso no momento de seu nascimento do que uma gestante não fumante (alternativamente, para cada 1,85 bebês nascidos com peso abaixo do normal de mães fumantes, nasce 1 bebê com peso abaixo do normal de mãe não fumante).

Utilizando-se o mesmo grupo de dados, o valor obtido para a Razão de chances (*odds ratio* - OR) é geralmente maior do que aquele que se obtém através da fórmula tradicional da razão de risco (risco relativo - RR). Para os dados da Tabela, uma Razão de chances de 1,85 é uma aproximação razoável para um Risco relativo de 1,76.

À medida que o evento mensurado é mais raro esta aproximação torna-se progressivamente mais precisa.

#### 8.4.8 Correlação linear de Pearson

Em estatística, a expressão correlação se refere à relação existente entre variáveis, digamos  $X$  e  $Y$ . Essa correlação pode assumir padrões diferentes: linear, não linear (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.

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  $Y$ ;

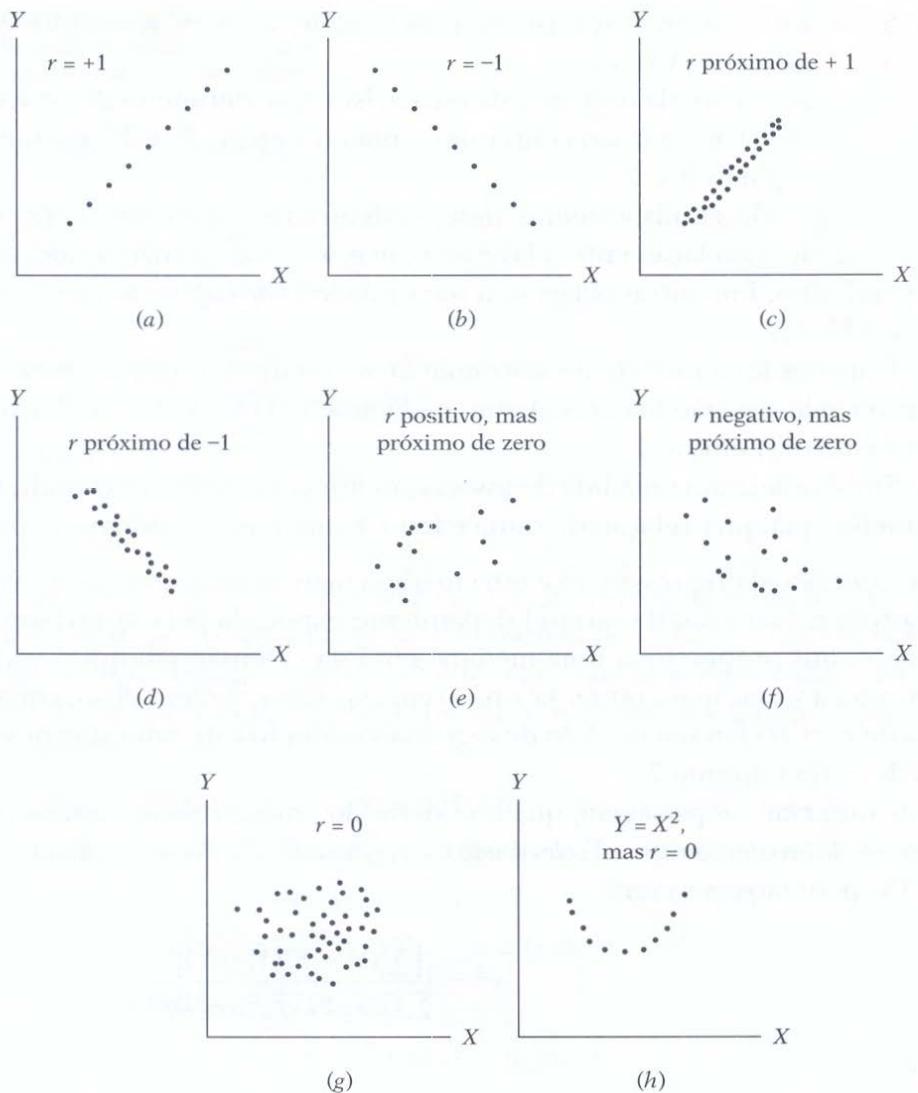


Figure 8.2: Diferentes diagramas de dispersão entre duas variáveis X e Y (Fonte: Introduction to Econometrics. Englewood Cliffs, 1978)

- em (B) e (D) vemos que uma variação inversa: quando o valor da variável  $X$  sofre um incremento, a variável  $Y$  sofre um decremento 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  $X \cdot Y$ .

$$r = \frac{\sum_{i=1}^n x_i \cdot y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sqrt{\left( \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right) \cdot \left[ \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \right]}}$$

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.

Simplificadamente podemos 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 decremento 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 8.5: Quadro de dados

$X$	$Y$
74	139
45	108
48	98
36	76
27	62
16	57

Assim, sendo  $n = 6$  observações segue-se:

Table 8.6: 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

$$\begin{aligned}
S_{xy} &= \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n} \\
&= 25172 - \frac{246 \cdot 540}{6} \\
&= 3032 \\
S_{xx} &= \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \\
&= 12086 - \frac{246^2}{6} \\
&= 2000 \\
S_{yy} &= \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \\
&= 53458 - \frac{540^2}{6} \\
&= 4858
\end{aligned}$$

Portanto:

$$\begin{aligned}
r &= \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}} \\
&= \frac{3032}{\sqrt{2000 \cdot 4858}} \\
&= 0,9727
\end{aligned}$$

## 8.5 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<sub>\_</sub>l) 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 diferença de proporções.

$$\log(IC_{(medida)}) = \log(medida) \pm [z_{(1-\frac{\alpha}{2})} \times EP(\log(medida))]$$

em que:

- $EP(\log(medida))$  é o erro padrão do logaritmo da medida e os valores mínimo e máximo do intervalo de confiança serão dados por  $\exp[\log((IC_{(medida)})]$ ;
- $\alpha$  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 8.7: Valores críticos  $z_c$  correspondentes a vários níveis de significância ( $\alpha$ )

Níveis de significância ( $\alpha$ )	0,10	0,05	0,01	0,005	0,002
Valores críticos de $z_c$ para testes unilaterais	-1,28 <b>ou</b> 1,28	-1,645 <b>ou</b> 1,645	-2,33 <b>ou</b> 2,33	-2,58 <b>ou</b> 2,58	-2,88 <b>ou</b> 2,88
Valores críticos de $z_c$ para testes bilaterais	-1,645 <b>e</b> 1,645	-1,96 <b>e</b> 1,96	-2,58 <b>e</b> 2,58	-2,81 <b>e</b> 2,81	-3,08 <b>e</b> 3,08

### 8.5.1 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:

$$\begin{aligned} EP(\log(RR)) &= \sqrt{\left[ \frac{1}{(a)} - \frac{1}{(a)+(b)} \right] + \left[ \frac{1}{(c)} - \frac{1}{(c)+(d)} \right]} \\ EP(\log(RR)) &= \sqrt{\left[ \frac{1}{(275)} - \frac{1}{2.419} \right] + \left[ \frac{1}{311} - \frac{1}{4.807} \right]} \\ EP(\log(RR)) &= \sqrt{0,006230374} \\ EP(\log(RR)) &= 0,078932718 \end{aligned}$$

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:

$$IC_{RR(1-\alpha=0,95)} = [1,49974; 2,04359]$$

### 8.5.2 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:

$$IC_{OR(1-\alpha=0,95)} = [1,560467; 2,193253]$$

### 8.5.3 Diferença de risco (Risco atribuível - RA)

Considere a estrutura dos dados presentes na Tabela para a estimativa 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:

$$\begin{aligned} EP(RA) &= \sqrt{\left[ \frac{a \times b}{(a+b)^3} \right] + \left[ \frac{c \times d}{(c+d)^3} \right]} \\ EP(RA) &= \sqrt{\left[ \frac{275 \times 2144}{(275+2.144)^3} \right] + \left[ \frac{311 \times 4.496}{(311+4.496)^3} \right]} \\ EP(RA) &= 0,007364887 \end{aligned}$$

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 Diferença de risco previamente calculada (0,049), um intervalo com nível de confiança de  $(1 - \alpha = 95\%)$  fica assim delimitado:

$$\begin{aligned} IC_{(RA)} &= RA \pm [z_{(1-\frac{\alpha}{2})} \times EP(RA)] \\ IC_{(RA)} &= 0,049 \pm [1,96 \times 0,007364887] \end{aligned}$$

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

### 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.

272 MÓDULO 9. INTRODUÇÃO À DISTRIBUIÇÃO DAS MÉDIAS E DIFERENÇAS ENTRE MÉDIAS AMOSTRAIS

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  $\mu$  (média) e  $\sigma^2$  (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.

## 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 ( $\sigma^2$ ) 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  $\alpha$  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 ( $\alpha$ ), 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 ( $\alpha$ ) **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) {
  plot(0, 0,
    type="n",
    xlim=c(mu-4,mu+4),
    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, " ",
    sub=paste0("Parâmetros da distribuição da população Normal ( \u03bc, \u03c3 ) = ( ", mu, ", ", sigma, " )"))
  abline(v=mu, col="blue")
  #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)
    sd = sd(x)
    li = media - zc * sd/(sqrt(n))
    ls = media + zc * sd/(sqrt(n))
    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")
    else points(media, i, col="black")
  }
}
```

```
N=100
n=30
mu=1.65
sigma=2
conf=0.95
IC.N(N, n, mu, sigma, conf)
```

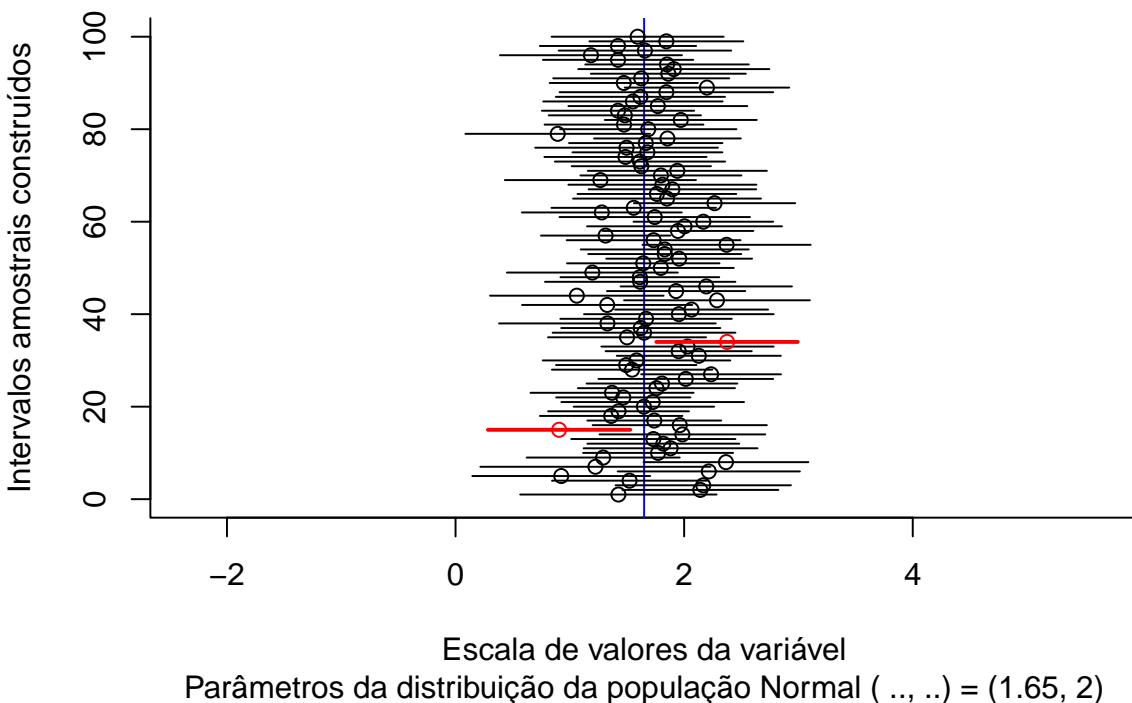
```
## Warning in title(...): conversion failure on 'Parâmetros da distribuição da
## população Normal ( , ) = (1.65, 2)' in 'mbcsToSbcs': dot substituted for <ce>

## Warning in title(...): conversion failure on 'Parâmetros da distribuição da
## população Normal ( , ) = (1.65, 2)' in 'mbcsToSbcs': dot substituted for <bc>

## Warning in title(...): conversion failure on 'Parâmetros da distribuição da
## população Normal ( , ) = (1.65, 2)' in 'mbcsToSbcs': dot substituted for <cf>

## Warning in title(...): conversion failure on 'Parâmetros da distribuição da
## população Normal ( , ) = (1.65, 2)' in 'mbcsToSbcs': dot substituted for <83>
```

### Intervalos com iguais níveis de confiança fixados em 95% (100 amostras de tamanho 30)



O gráfico acima expõe os intervalos de confiança:  $(1 - \alpha)=95\%$  produzidos para as 100 médias de amostras de tamanho 30 extraídas de uma população com parâmetros  $\mu : 1.65$  e  $\sigma : 2$ .

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

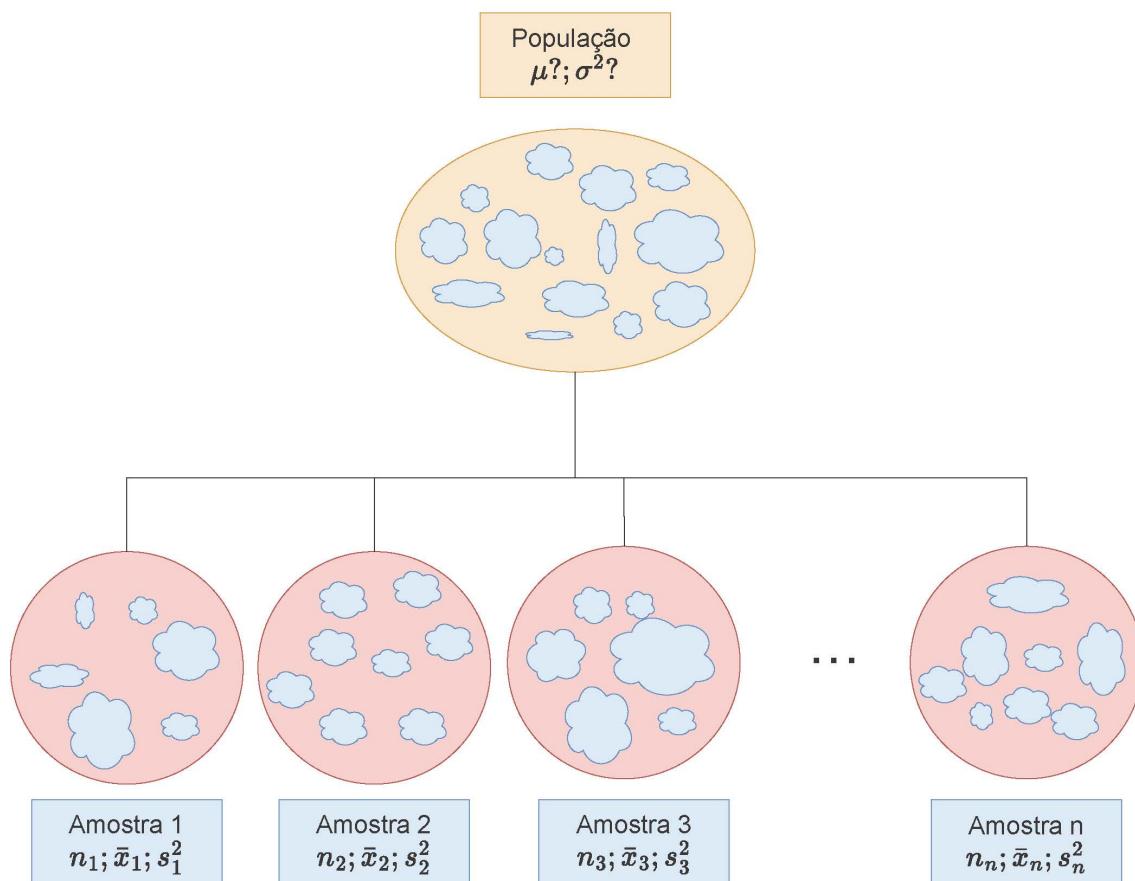


Figure 9.1: Ilustração esquemática de  $n$  amostras extraídas de uma mesma população de parâmetros  $\mu$  e  $\sigma$ , cada uma apresentando as respectivas estatísticas calculadas

Para estudarmos a distribuição das médias amostrais considerem uma população com parâmetros  $\mu$  (média) e  $\sigma^2$  (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.

A convergência da forma de distribuição e dos parâmetros dessa distribuição das médias amostrais são elucidadas pela **Lei dos Grandes Números** e pelo **Teorema Central do Limite**.

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  $\mu$  e variância  $\sigma^2$  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 9.2 e 9.4.

O **TCL** garante a aproximação da distribuição de  $\bar{X}$  a uma distribuição Normal com média  $\mu$  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**:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

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))
```

A Figura 9.2 mostra o histograma de uma amostra de 5000 elementos de uma população com Distribuição Uniforme de parâmetros  $v_{min} : 2$  e  $v_{max} : 6$ .

A Figura 9.3 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.

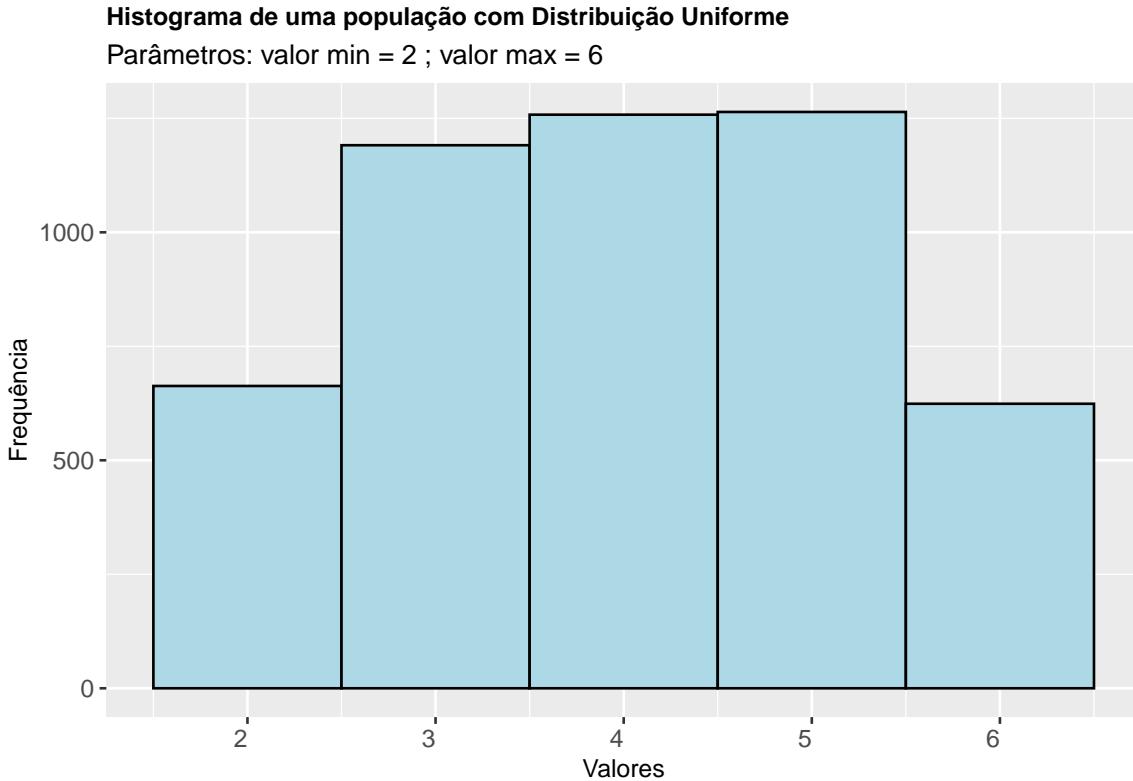


Figure 9.2: Histograma de uma população cuja característica de interesse segue uma Distribuição Uniforme

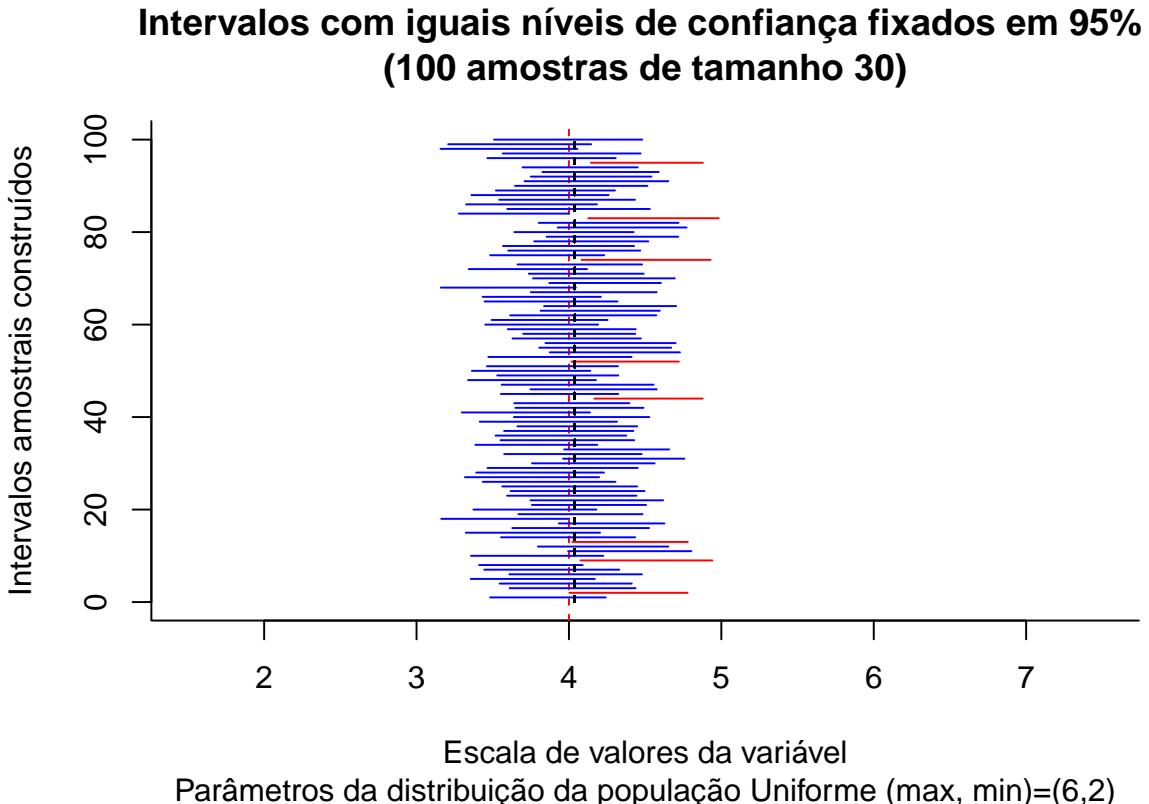


Figure 9.3: Intervalos de confiança construídos para diversas estimativas amostrais de uma população com Distribuição  $\sim N(\mu = \frac{\max - \min}{2}; \sigma^2 = \frac{1}{12}(\max - \min)^2)$

```

meu_titulo1=paste("Distribuição das médias de", N, "amostras de tamanho n=",n,"\\n população de origem")
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) +
  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 )

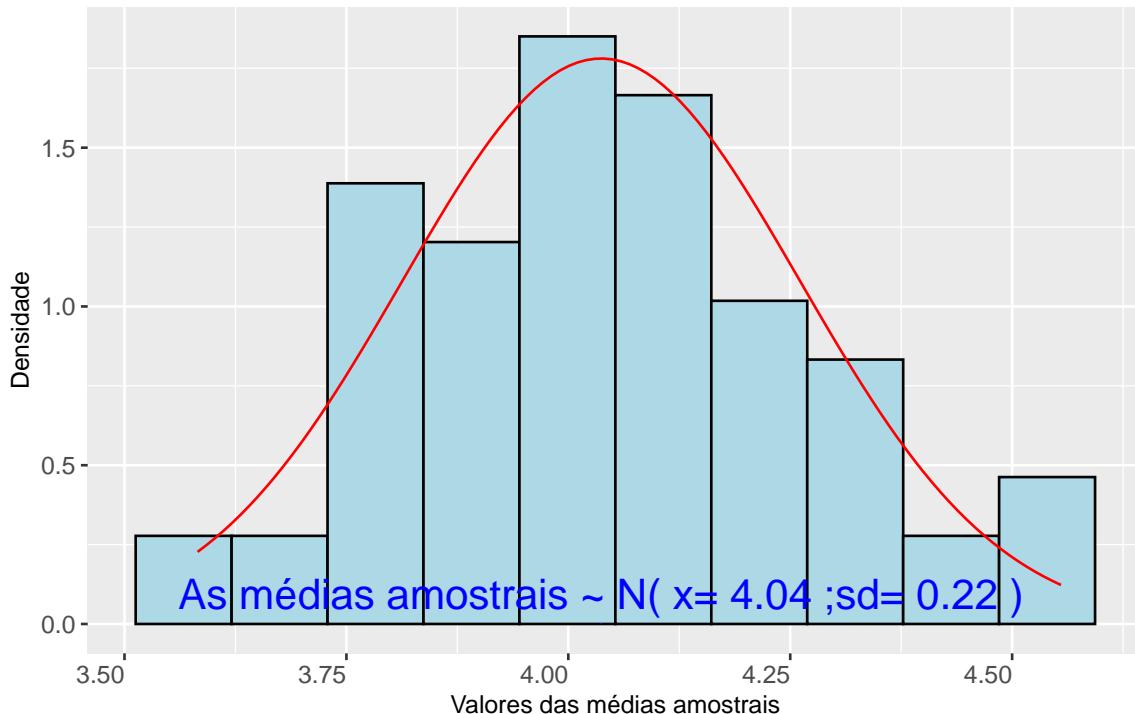


Figure 9.4: Histograma da distribuição das médias de amostras extraídas de uma população com Distribuição Uniforme mostra que as mesmas seguem uma Distribuição  $\sim N(\mu = \frac{\max - \min}{2}; \sigma^2 = \frac{1}{12}(\max - \min)^2)$

O histograma da Figura 9.4 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  $\sim N(\mu = \frac{v_{max}-v_{min}}{2}; \sigma^2 = \frac{1}{12}(v_{max} - v_{min})^2)$ .

Demostraçāo usando amostras extraídas de uma população com distribuição  $\sim N(\mu; \sigma)$

```
# Definindo os parāmetros 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))
```

A Figura 9.5 mostra o histograma de uma amostra de 5000 elementos de uma população com Distribuição Normal de parâmetros média= 80 e desvio padrão =4.

A Figura 9.6 expõe os intervalos sob nível de confiança de  $(1 - \alpha)=95\%$  produzidos para as 100 médias de amostras de tamanho 50 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")
meu_titulo2=paste("As médias amostrais ~ N( x\u0304=",round(mean(m),2),";sd=",round(sd(m),2),"")")

dados=as.data.frame(m)
```

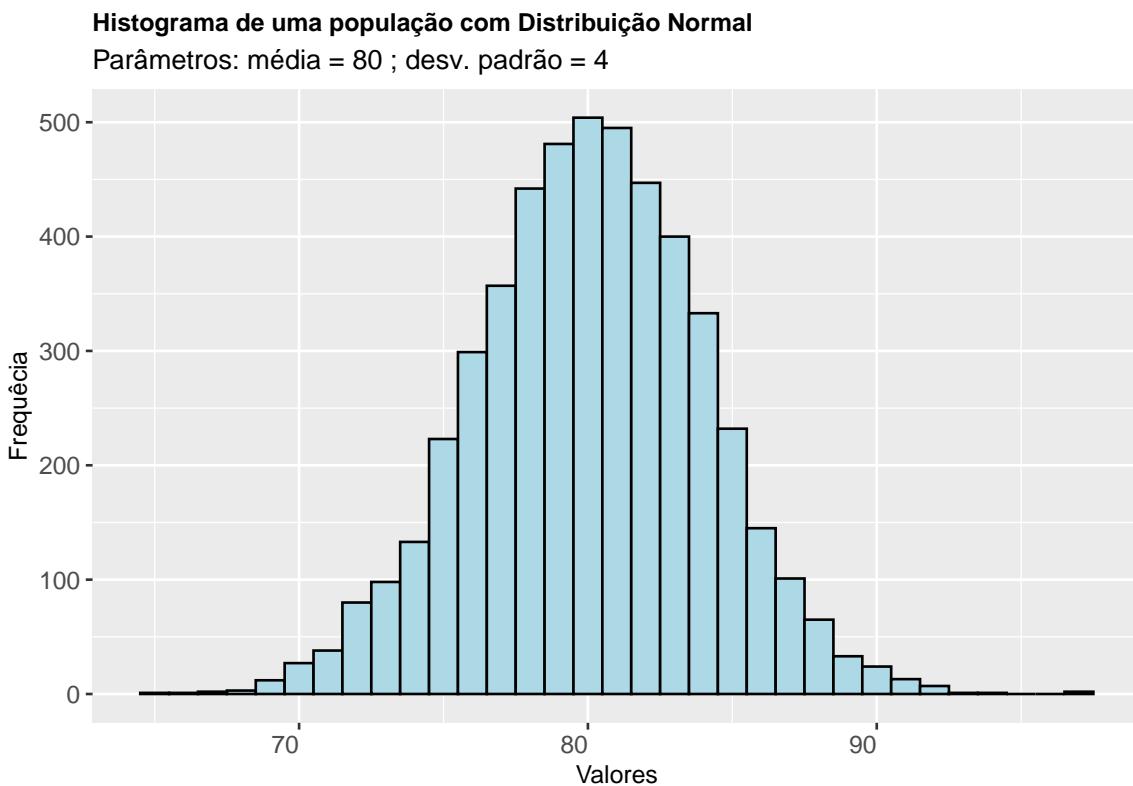


Figure 9.5: Histograma de uma população cuja característica de interesse segue uma Distribuição Normal

**Intervalos com iguais níveis de confiança fixados em 95%  
 (100 amostras de tamanho 50)**

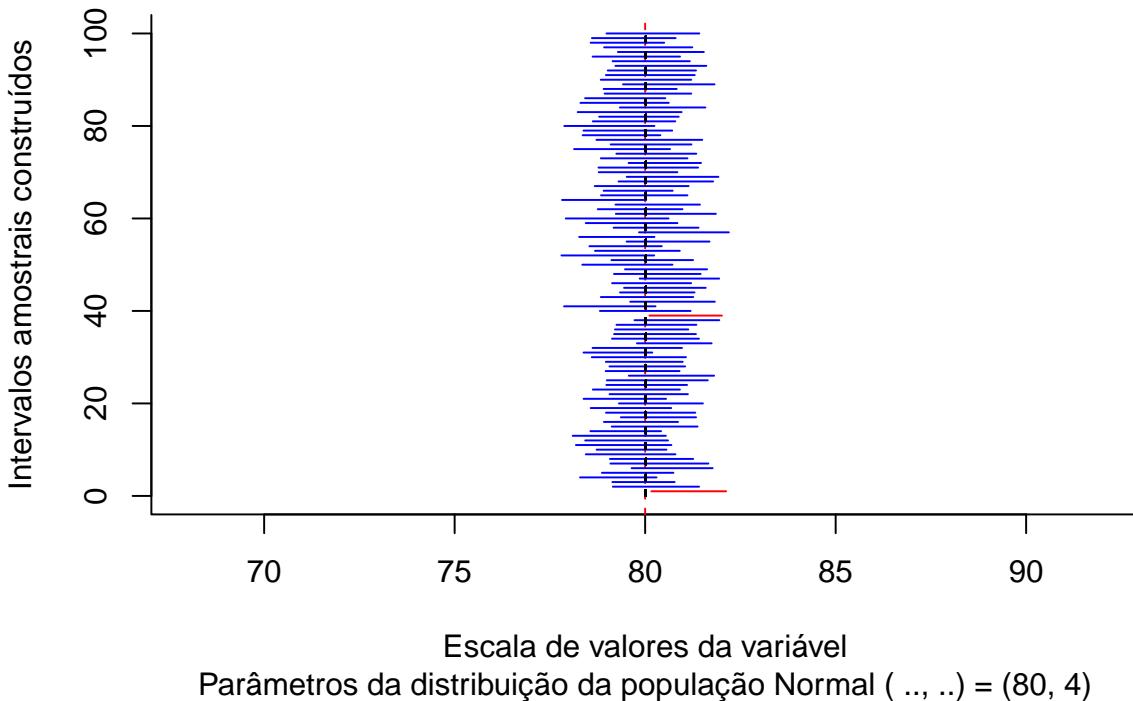


Figure 9.6: Intervalos de confiança construídos para diversas estimativas amostrais de uma população com Distribuição  $\sim N(\mu; \sigma)$

```
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,
  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))
```

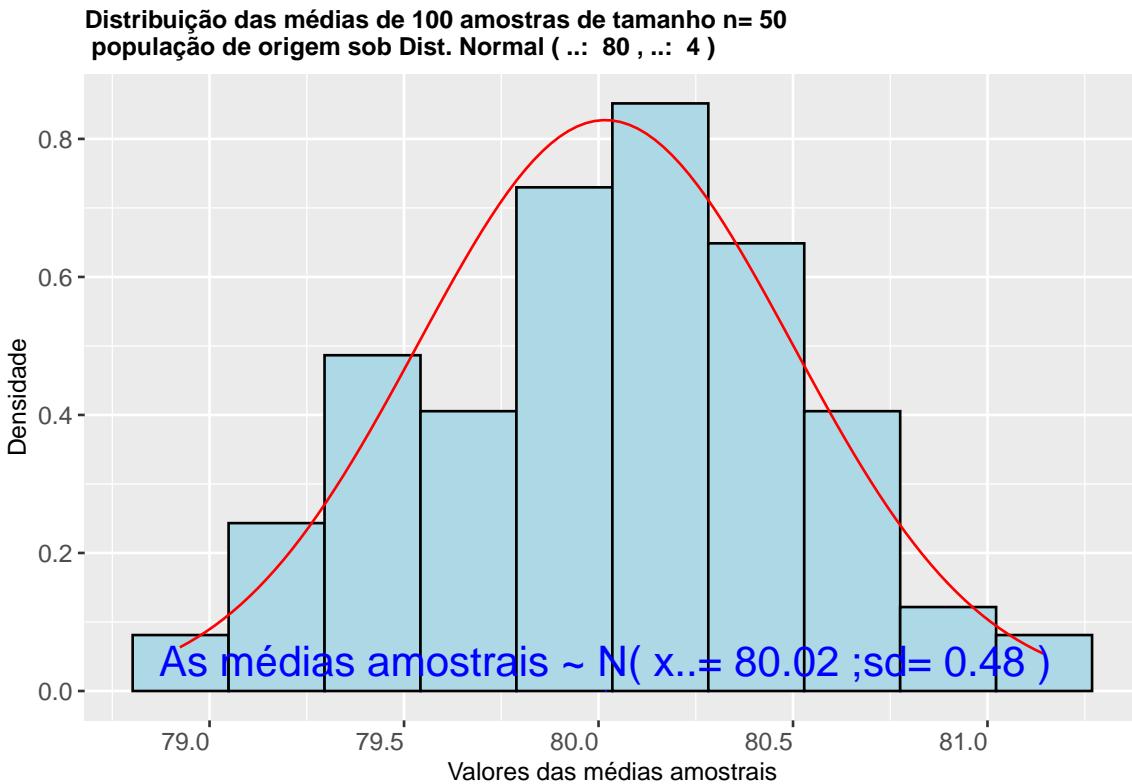


Figure 9.7: Histograma da distribuição das médias de amostras extraídas de uma população Normal mostra que as mesmas seguem uma Distribuição  $\sim N(\bar{x}=\mu; s=\frac{\sigma}{\sqrt{n}})$

O histograma da 9.7 ilustra que os valores das médias calculadas de 50 amostras extraídas de uma população com distribuição Normal  $\sim N(\mu, \sigma)$  seguem uma distribuição Normal  $\sim N(\mu = \mu; \sigma = \frac{\sigma}{\sqrt{n}})$ .

Corolário: se  $(X_1, X_2, \dots, X_n)$  for uma amostra aleatória simples da população  $X$  de média  $\mu$  e variância  $\sigma^2$  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  $\varepsilon$  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 ( $\sigma^2$ ) 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;
- $\mu$  é a média populacional;
- $\sigma$  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 ( $\sigma^2$ ).

Nessas situações, se o tamanho da amostra é grande (na prática  $n \geq 30$ ), podemos substituir  $\sigma$  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;
- $\mu$  é a média populacional;
- $S$  é o desvio padrão da amostra; e,
- $n$  é o tamanho da amostra.

Caso a variância populacional ( $\sigma^2$ ) 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  $\chi^2$  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;
- $\mu$  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 6.16)

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

em que:

- $\bar{X}$  é a média amostral;
- $\mu$  é a média populacional;
- $\sigma$  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  $\alpha$  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)"),
  geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1), color="blue"),
  geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2), color="blue"),
  annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="-z", angle=90, vjust=0, hjust=0),
  annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="z", angle=90, vjust=0, hjust=0),
  annotate(geom="text", x=z_desejado1-1.8, y=0.1, label="Intervalo aberto à esq. \n(probabilidade="),
  annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade="),
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade= (1-\u03b1)\u03b1"))
  theme_bw()

```

Na Figura 9.8 observa-se:

**Curva da função densidade  
Distribuição Normal Padrão**

$P(-z, z) = (1 - \alpha)$  em cinza (nível de confiança)  
 $P(-\dots; -z) = P(z; \dots) = \dots/2$  em vermelho

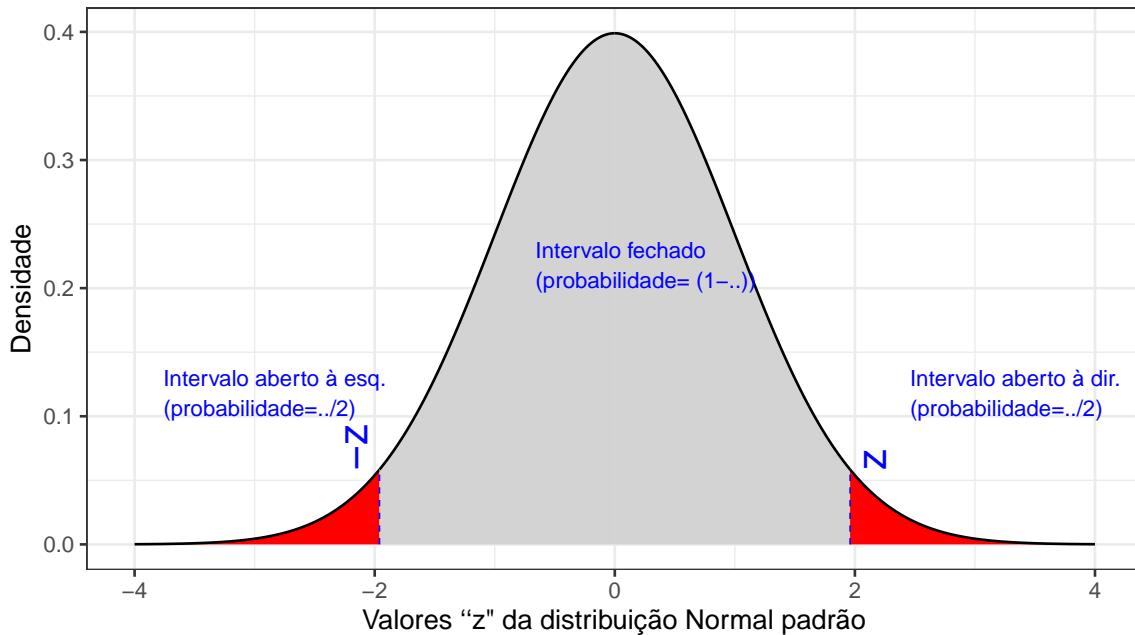


Figure 9.8: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores  $Z$  é inferior a  $\frac{\alpha}{2}$ , estabelecendo assim um intervalo com nível de confiança igual a  $(1 - \alpha)$

- o nível de significância  $\alpha$ ;
- o nível de confiança  $(1 - \alpha)$ ; e,
- o valor tabelado da estatística  $Z(z)$  para o nível de confiança fixado.

Assim,

$$\begin{aligned} P[-Z_{(1-\frac{\alpha}{2})} \leq Z \leq Z_{(1-\frac{\alpha}{2})}] &= (1 - \alpha) \\ P\left[-z_{(1-\frac{\alpha}{2})} \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P[\bar{x} - (z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}) \leq \mu \leq \bar{x} + (z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}})] &= (1 - \alpha) \end{aligned}$$

$$IC(\mu)_{(1-\alpha)} = [\bar{x} \pm z_c \cdot \frac{\sigma}{\sqrt{n}}]$$

Assim, se  $\bar{x}$  é usado como estimativa de  $\mu$ , 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  $\alpha$  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  $(1 - \alpha) = 95\%$ .

$$\begin{aligned} P[\bar{x} - (z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}) \leq \mu \leq \bar{x} + (z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}})] &= (1 - \alpha) \\ P[20.000 - (1,96 \cdot \frac{8.300}{\sqrt{15}}) \leq \mu \leq 20000 + (1,96 \cdot \frac{8.300}{\sqrt{15}})] &= 0,95 \\ P[20.000 - 4.200,38 \leq \mu \leq 20.000 + 4.200,38] &= 0,95 \end{aligned}$$

$$IC_{(1-\alpha=0,95)} = [\text{US\$ } 15.799,62; \text{US\$ } 24.200,38]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que, se extraímos 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  $\text{US\$ } 15.799,62; \text{ US\$ } 24.200,38$ , 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 6.17 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(−\u03b1; z)=(1−..) em cinza (nível de confiança)  \nP(z, + \u03b1)= \u03b1 em vermelho",
  annotate(geom="text", x=z_desejado1+3.5, y=d_desejada1, label="z", angle=90, vjust=0, hjust=0, color="black"),
  annotate(geom="text", x=z_desejado1+4.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade=..)", angle=90, vjust=0, hjust=0, color="blue"),
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à esq. \n(probabilidade=..)", angle=90, vjust=0, hjust=0, color="blue"),
  theme_bw())
```

### Curva da função densidade

#### Distribuição Normal Padrão

$P(-\alpha; z) = (1 - ..)$  em cinza (nível de confiança)  
 $P(z, + \alpha) = ..$ , em vermelho

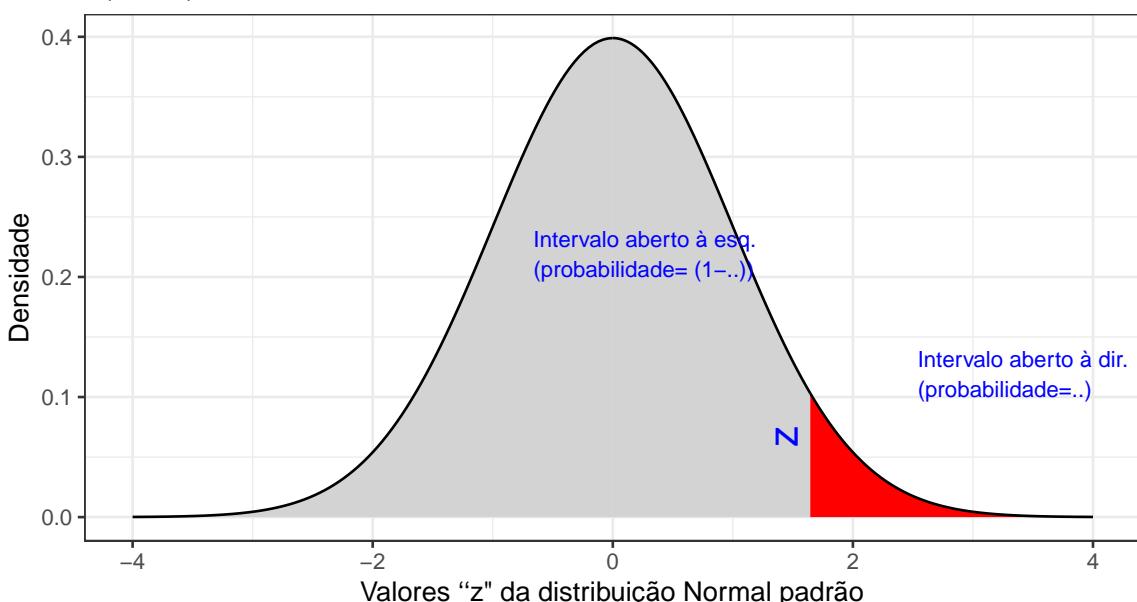


Figure 9.9: Região crítica, além da qual, a probabilidade associada aos valores  $Z$  é inferior a  $\alpha$ , delimitando assim, à esquerda, um intervalo aberto com nível de confiança igual a  $(1 - \alpha)$

A Figura 9.10 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,
  annotate(geom="text", x=z_desejado1-2, y=0.1, label="Intervalo aberto à esq. \n(probabilidade=\u03b1/2)",
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à dir. \n(probabilidade=\u03b1/2",
  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 9.11)

$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0, 1)$$

em que:

- $\bar{X}$  é a média amostral;
- $\mu$  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  $\alpha$  estabelecido.

### Curva da função densidade

#### Distribuição Normal Padrão

$P(-\dots; z) = \dots$ , em vermelho

$P(z, + \dots) = (1 - \dots)$  em cinza

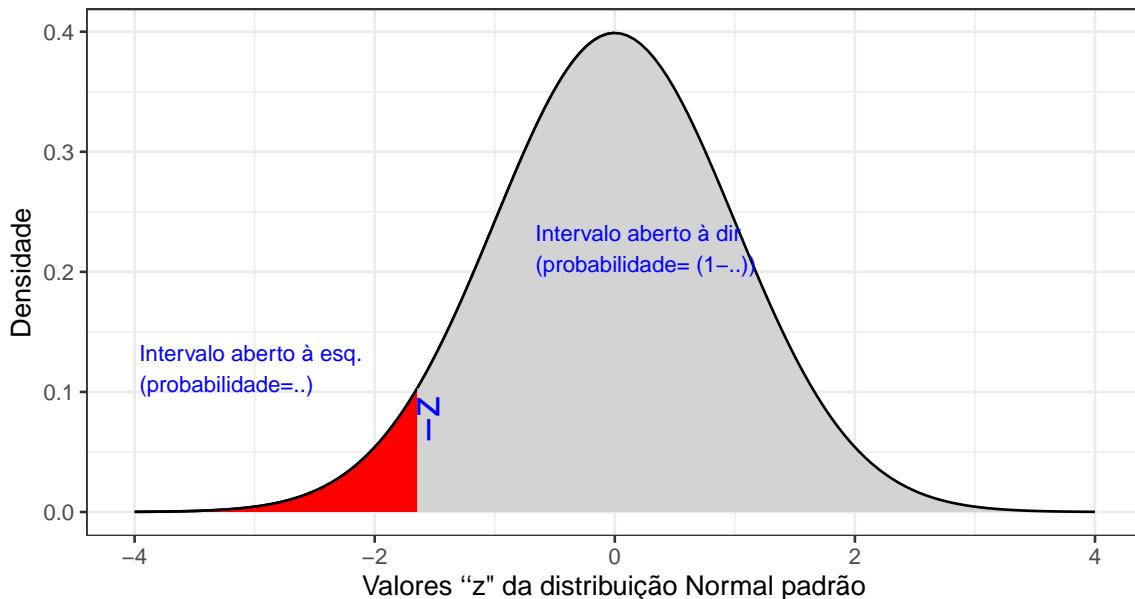


Figure 9.10: Região crítica, aquém da qual, a probabilidade associada aos valores  $Z$  é inferior a  $\alpha$ , delimitando assim, à direita, um intervalo aberto com nível de confiança igual a  $(1 - \alpha)$

```

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) em vermelho",
  geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1), color="blue"),
  geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2), color="blue"),
  annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="-z", angle=90, vjust=0, hjust=0),
  annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="z", angle=90, vjust=0, hjust=0),
  annotate(geom="text", x=z_desejado1-1.8, y=0.1, label="Intervalo aberto à esq. \n(probabilidade=../2)", angle=90, vjust=0, hjust=0),
  annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade=../2)", angle=90, vjust=0, hjust=0),
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade= (1-\u03b1))", angle=90, vjust=0, hjust=0),
  theme_bw())

```

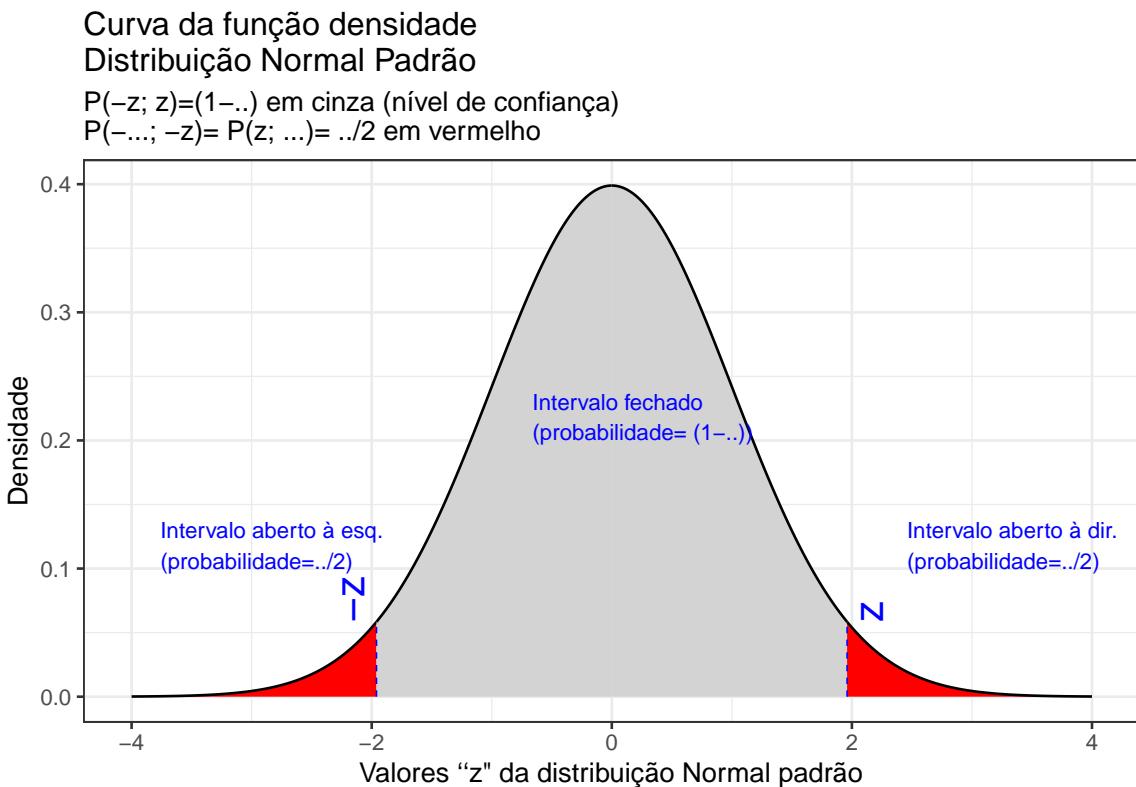


Figure 9.11: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores  $Z$  é inferior a  $\frac{\alpha}{2}$ , estabelecendo assim um intervalo com nível de confiança igual a  $(1 - \alpha)$

Na Figura 9.11 observa-se:

- o nível de significância  $\alpha$ ;
- o nível de confiança  $(1 - \alpha)$ ; e,
- o valor tabelado da estatística  $Z(z)$  para o nível de confiança fixado.

Assim,

$$\begin{aligned} P\left[-Z_{(1-\frac{\alpha}{2})} \leq Z \leq Z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P\left[-z_{(1-\frac{\alpha}{2})} \leq \frac{\bar{x} - \mu}{(\frac{S}{\sqrt{n}})} \leq z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P[\bar{x} - (z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}) \leq \mu \leq \bar{x} + (z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}})] &= (1 - \alpha) \end{aligned}$$

$$IC(\mu)_{(1-\alpha)} = [\bar{x} \pm z_c \cdot \frac{S}{\sqrt{n}}]$$

Assim, se  $\bar{x}$  é usado como estimativa de  $\mu$  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  $\alpha$  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  $(1 - \alpha) = 95\%$ .

$$\begin{aligned} P[\bar{x} - (z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}) \leq \mu \leq \bar{x} + (z_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}})] &= (1 - \alpha) \\ P[20.000 - (1,96 \cdot \frac{8.300}{\sqrt{60}}) \leq \mu \leq 20.000 + (1,96 \cdot \frac{8.300}{\sqrt{60}})] &= 0,95 \\ P[20.000 - 2.100,19 \leq \mu \leq 20.000 + 2.100,19] &= 0,95 \end{aligned}$$

$$IC_{(1-\alpha=0,95)} = [\text{US\$}17.899,81; \text{US\$}22.100,19]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que se extraímos 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  $\text{]US\$ 17.899,81; US\$ 22.100,19[}$ , 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 9.12 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(-\u2212z; z)=(1-\u03b1) em cinza (nível de confiança) \nP(z, + \u221e)= \u03b1 em vermelho (probabilidade de rejeição)",

  annotate(geom="text", x=z_desejado1+3.5, y=d_desejada1, label="z", angle=90, vjust=0, hjust=0,
  annotate(geom="text", x=z_desejado1+4.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade de rejeição)", angle=0, hjust=0, vjust=0),
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à esq. \n(probabilidade de aceitação)", angle=0, hjust=0, vjust=0),
  theme_bw()

```

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

$$P \left[ \mu \geq \bar{x} - z_c \cdot \frac{S}{\sqrt{n}} \right] = (1 - \alpha)$$

```

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",
            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") +
  labs(title="Curva da função densidade
    \nDistribuição Normal Padrão",
    subtitle = "P(-\u2212z; z)=(1-\u03b1) em cinza (nível de confiança) \nP(z, + \u221e)= \u03b1 em vermelho (probabilidade de rejeição)",

  annotate(geom="text", x=z_desejado1+3.5, y=d_desejada1, label="z", angle=90, vjust=0, hjust=0,
  annotate(geom="text", x=z_desejado1+4.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade de rejeição)", angle=0, hjust=0, vjust=0),
  annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à esq. \n(probabilidade de aceitação)", angle=0, hjust=0, vjust=0),
  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

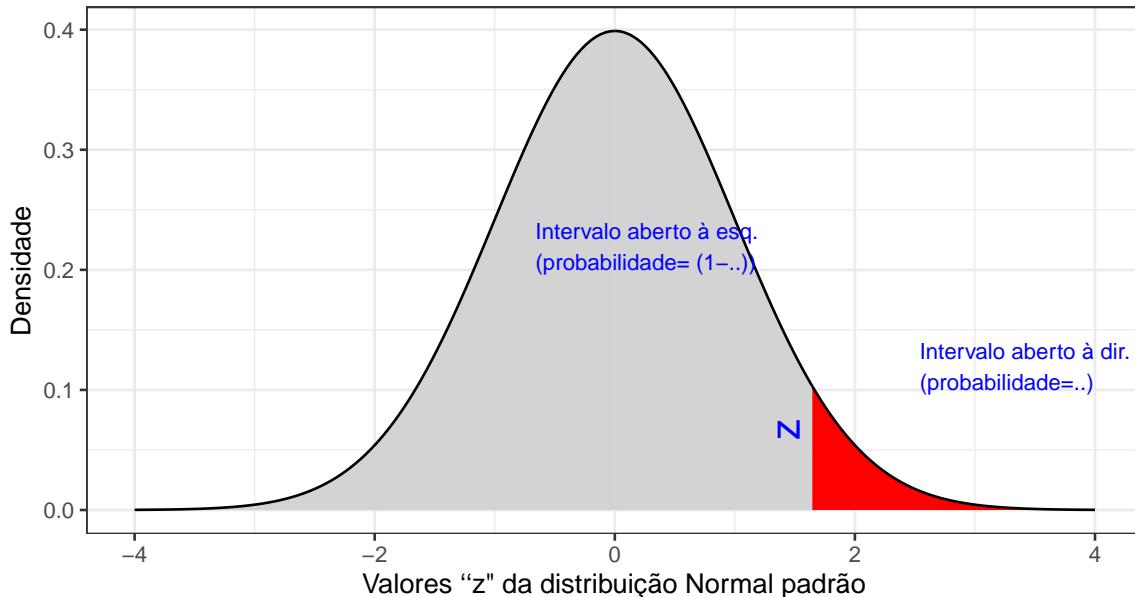


Figure 9.12: Região crítica, além da qual, a probabilidade associada aos valores  $Z$  é inferior a  $\alpha$ , delimitando assim, à esquerda, um intervalo aberto com nível de confiança igual a  $(1 - \alpha)$

```

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,
annotate(geom="text", x=z_desejado1-1.5, y=0.1, label="Intervalo aberto à esq. \n(probabilidade=",
annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo aberto à dir. \n(probabilidade=",
theme_bw()

```

### Curva da função densidade

#### Distribuição Normal Padrão

$P(-\dots; z) = \dots$ , em vermelho

$P(z, + \dots) = (1 - \dots)$  em cinza

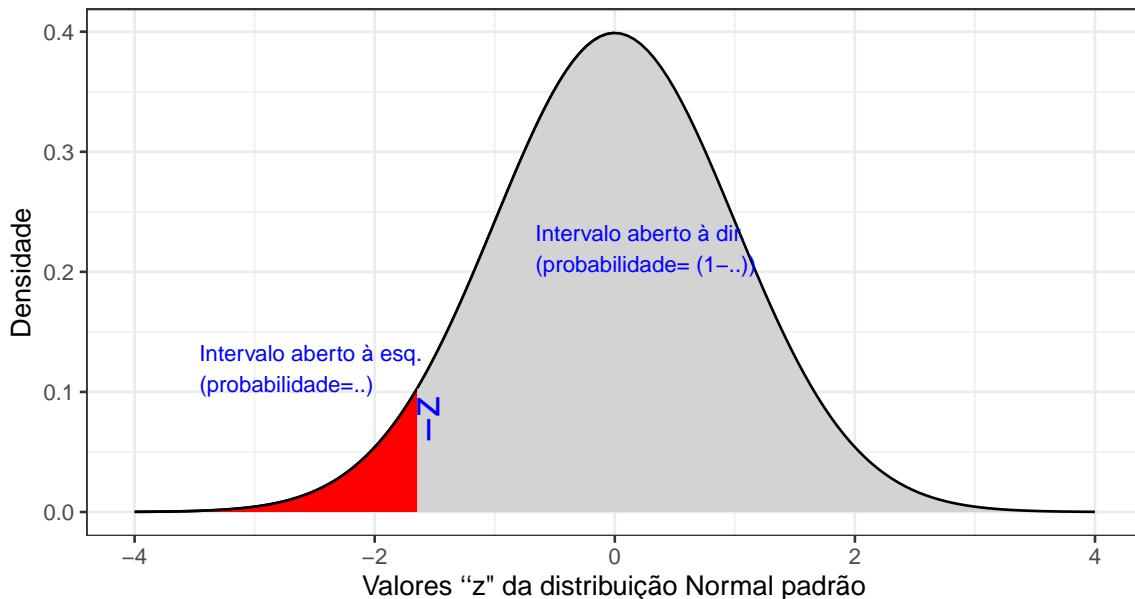


Figure 9.13: Região crítica, aquém da qual, a probabilidade associada aos valores  $Z$  é inferior a  $\alpha$ , delimitando assim, à direita, um intervalo aberto com nível de confiança igual a  $(1 - \alpha)$

#### 9.3.5 Intervalo de confiança para uma média amostral sob variância populacional desconhecida e amostras de qualquer tamanho (Figura 9.14)

$$T = \frac{(\bar{X} - \mu)}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

em que:

- $\bar{X}$  é a média amostral;
- $\mu$  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  $\alpha$  estabelecido.

```

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",
       geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1), color="blue"),
       geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2), color="blue"),
       annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="-t", angle=90, vjust=0, hjust=0),
       annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="t", angle=90, vjust=0, hjust=0),
       annotate(geom="text", x=t_desejado1-1.8, y=0.1, label="Intervalo aberto à esq. \n(probabilidade=",
       annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade=",
       annotate(geom="text", x=t_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade= (1-\u03b1)\u00b7"))

```

Na Figura 9.14 observa-se:

**Curva da função densidade  
Distribuição t**

$P(-t; t)=(1-\dots)$  em cinza (nível de confiança)  
 $P(-\dots; -t)=P(t; \dots)=\dots/2$  em vermelho

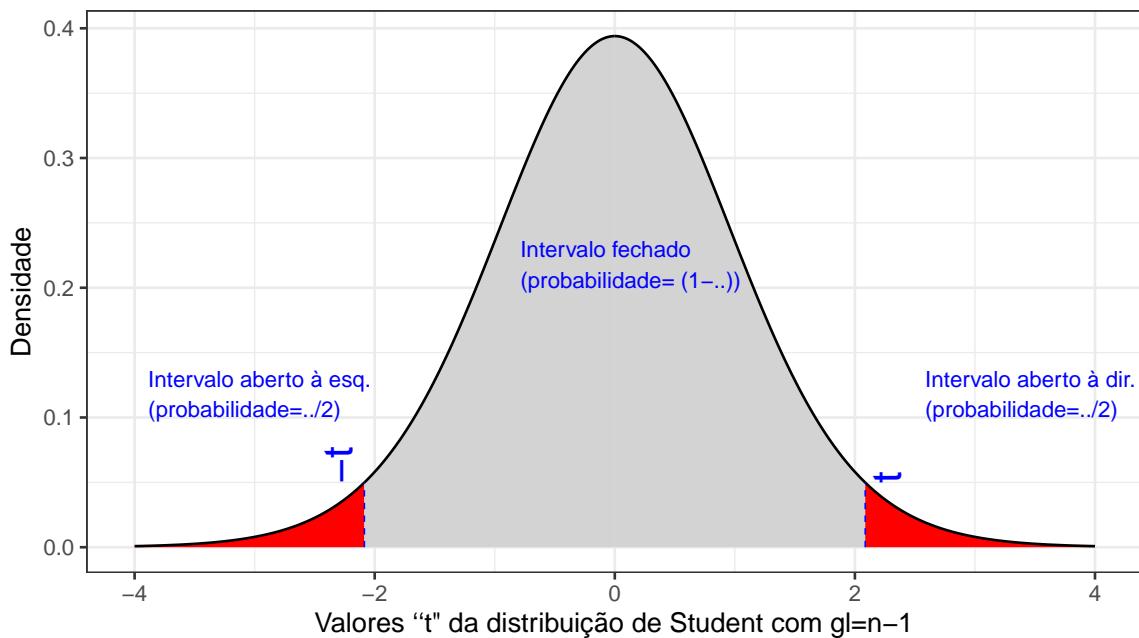


Figure 9.14: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores  $T$  ( $(n-1)$  graus de liberdade) é inferior a  $\frac{\alpha}{2}$ , estabelecendo assim um intervalo com nível de confiança igual a  $(1 - \alpha)$

- o nível de significância  $\alpha$ ;
- 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,

$$\begin{aligned} P\left[-T_{(1-\frac{\alpha}{2},(n-1))} \leq T \leq T_{(1-\frac{\alpha}{2},(n-1))}\right] &= (1 - \alpha) \\ P\left[-t_{(1-\frac{\alpha}{2},(n-1))} \leq \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{(1-\frac{\alpha}{2},(n-1))}\right] &= (1 - \alpha) \\ P[\bar{x} - (t_{(1-\frac{\alpha}{2},(n-1))} \cdot \frac{S}{\sqrt{n}}) \leq \mu \leq \bar{x} + (t_{(1-\frac{\alpha}{2},(n-1))} \cdot \frac{S}{\sqrt{n}})] &= (1 - \alpha) \\ IC(\mu)_{(1-\alpha)} &= [\bar{x} \pm t_{c_{(n-1)}} \cdot \frac{S}{\sqrt{n}}] \end{aligned}$$

Assim, se  $\bar{x}$  é usado como estimativa de  $\mu$  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  $\alpha$ ,  $(n-1)$  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,

## 304 MÓDULO 9. INTRODUÇÃO À DISTRIBUIÇÃO DAS MÉDIAS E DIFERENÇAS ENTRE MÉDIAS AMOSTRAIS

- 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  $(1 - \alpha) = 95\%$ .

$$P[\bar{x} - (t_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}) \leq \mu \leq \bar{x} + (t_{(1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}})] = (1 - \alpha)$$

$$P[20000 - (2,1448 \cdot \frac{8300}{\sqrt{15}}) \leq \mu \leq 20000 + (2,1448 \cdot \frac{8300}{\sqrt{15}})] = 0,95$$

$$P[20000 - 4596,41 \leq \mu \leq 20000 + 4596,41] = 0,95$$

$$IC_{(1-\alpha=0,95)} = [US\$15403,59; US\$24496,41]$$

Se quisermos ser rigorosos na interpretação do intervalo de confiança calculado podemos explicar que se extraímos 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 9.15 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  $T$  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(-\u221e, t)=(1-\u03b1) em cinza \nP(t, \u221e)= \u03b1 em vermelho "+geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1), color="blue",
       annotate(geom="text", x=t_desejado1+0.5, y=d_desejada1, label="t", angle=90, vjust=0, hjust=0,
       annotate(geom="text", x=t_desejado1+1, y=0.1, label="Intervalo aberto à esq. \n(probabilidade=\u03b1),
       annotate(geom="text", x=t_desejado1-2.5, y=0.2, label="Intervalo aberto \n(probabilidade= (1-\u03b1)

```

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  $T$  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)

```

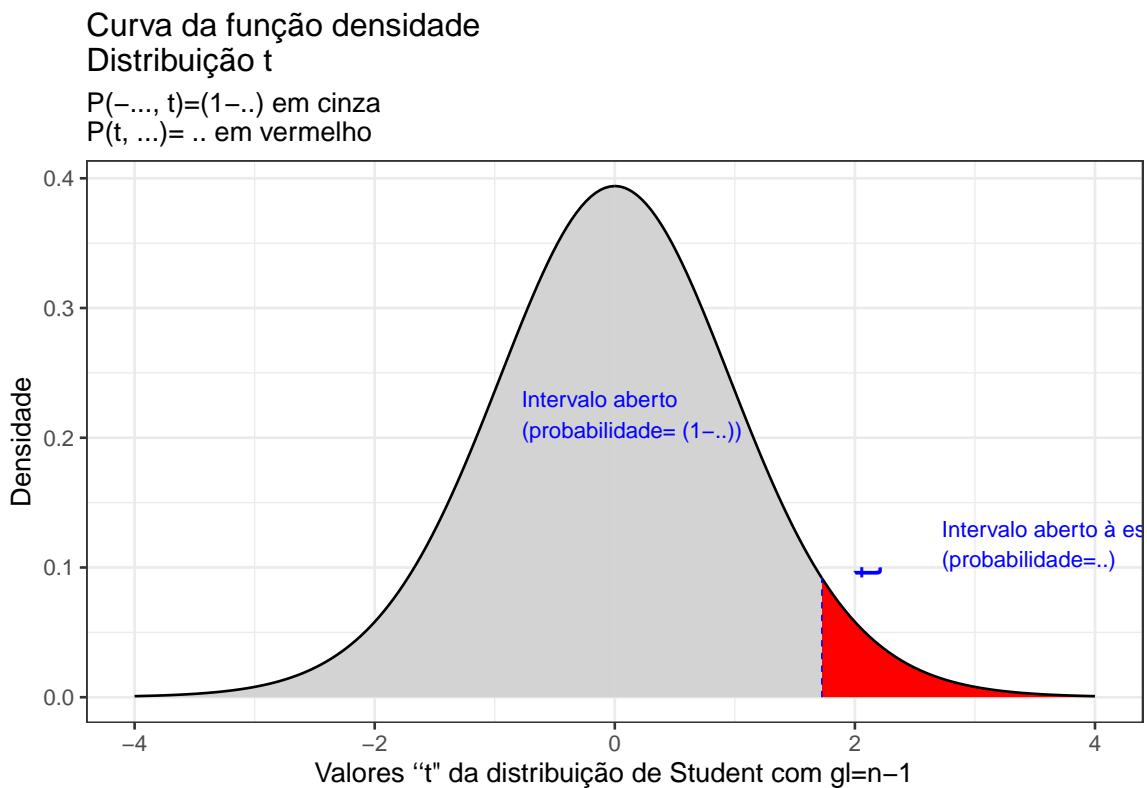


Figure 9.15: Região crítica, além da qual, a probabilidade associada aos valores  $T$  ( $(n-1)$  graus de liberdade) é inferior a  $\alpha$ , delimitando assim, à esquerda, um intervalo aberto com nível de confiança igual a  $(1 - \alpha)$

```

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",
               annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="-t", angle=90, vjust=0, hjust=0,
               annotate(geom="text", x=t_desejado1-2.5, y=0.1, label="Intervalo aberto à esq. \n(probabilidade=0.05)", angle=90, vjust=0, hjust=0),
               annotate(geom="text", x=t_desejado1+1, y=0.2, label="Intervalo aberto \n(probabilidade= (1-\u03b1)/2)", angle=90, vjust=0, hjust=0)

```

## 9.4 Distribuição das diferenças de médias amostrais independentes

Consideremos duas populações  $X$  e  $Y$  com médias  $\mu_1$  e  $\mu_2$  e variâncias  $\sigma_1^2$  e  $\sigma_2^2$ , 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:

Curva da função densidade  
Distribuição t

$P(-t, \dots) = (1 - \dots)$  em cinza  
 $P(-\dots; -t) = \dots$  em vermelho

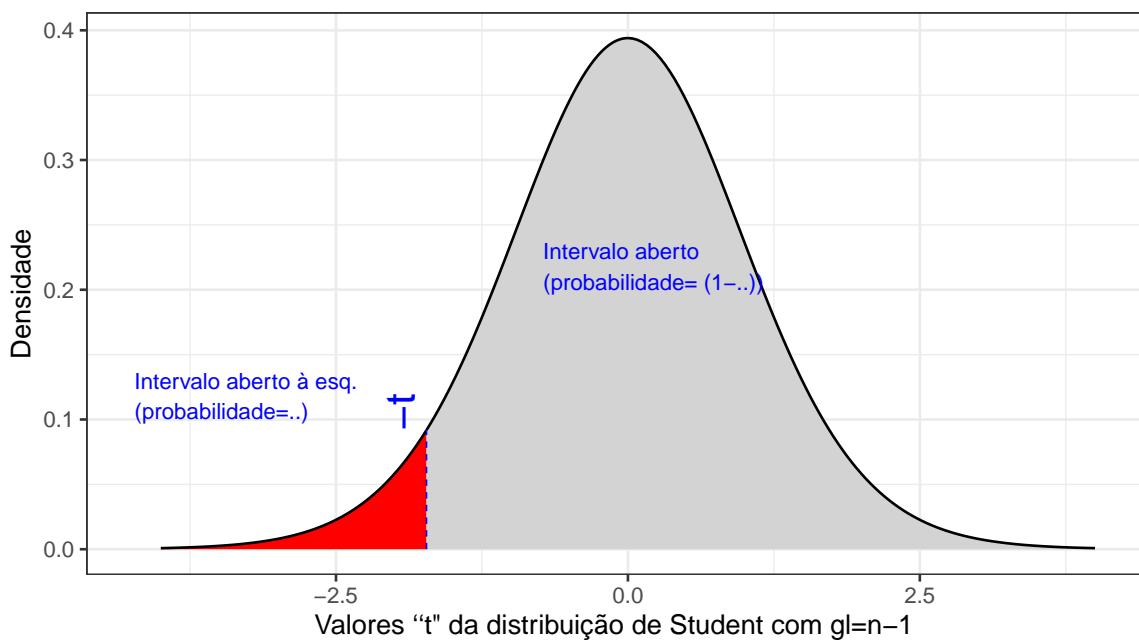


Figure 9.16: Região crítica, aquém da qual, a probabilidade associada aos valores  $T$  (( $n - 1$ ) graus de liberdade) é inferior a  $\alpha$ , delimitando assim, à direita, um intervalo aberto com nível de confiança igual a  $(1 - \alpha)$

$$\begin{aligned}\mu_{(X \pm Y)} &= \mu_1 \pm \mu_2 \\ \sigma^2_{(X \pm Y)} &= \sigma_1^2 + \sigma_2^2\end{aligned}$$

E a média e variância da soma ou diferença das distribuições amostrais das médias de  $X$  e  $Y$  é:

$$\begin{aligned}\mu_{(\bar{X} \pm \bar{Y})} &= \mu_1 \pm \mu_2 \\ \sigma^2_{(\bar{X} \pm \bar{Y})} &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\end{aligned}$$

#### 9.4.1 Intervalos de confiança para a diferença entre duas médias amostrais com variâncias 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  $\mu_1$  e  $\mu_2$ , e variâncias  $\sigma_1^2$  e  $\sigma_2^2$  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:

$$\begin{aligned}X &\sim N(\mu_1, \frac{\sigma_1^2}{\sqrt{n_1}}) \\ Y &\sim N(\mu_2, \frac{\sigma_2^2}{\sqrt{n_2}})\end{aligned}$$

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 9.17):

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

em que:

- $\bar{X}$  e  $\bar{Y}$  são as médias amostrais;
- $\mu_1$  e  $\mu_2$  são as médias populacionais;
- $\sigma_1^2$  e  $\sigma_2^2$  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",
geom_segment(aes(x = z_desejado1, y = 0, xend = z_desejado1, yend = d_desejada1), color="blue",
geom_segment(aes(x = z_desejado2, y = 0, xend = z_desejado2, yend = d_desejada2), color="blue",
annotate(geom="text", x=z_desejado1-0.1, y=d_desejada1, label="-z", angle=90, vjust=0, hjust=0,
annotate(geom="text", x=z_desejado2+0.3, y=d_desejada2, label="z", angle=90, vjust=0, hjust=0,
annotate(geom="text", x=z_desejado1-1.8, y=0.1, label="Intervalo aberto à esq. \n(probabilidade",
annotate(geom="text", x=z_desejado2+0.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade",
annotate(geom="text", x=z_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade= (1-\u03b1)
theme bw()

```

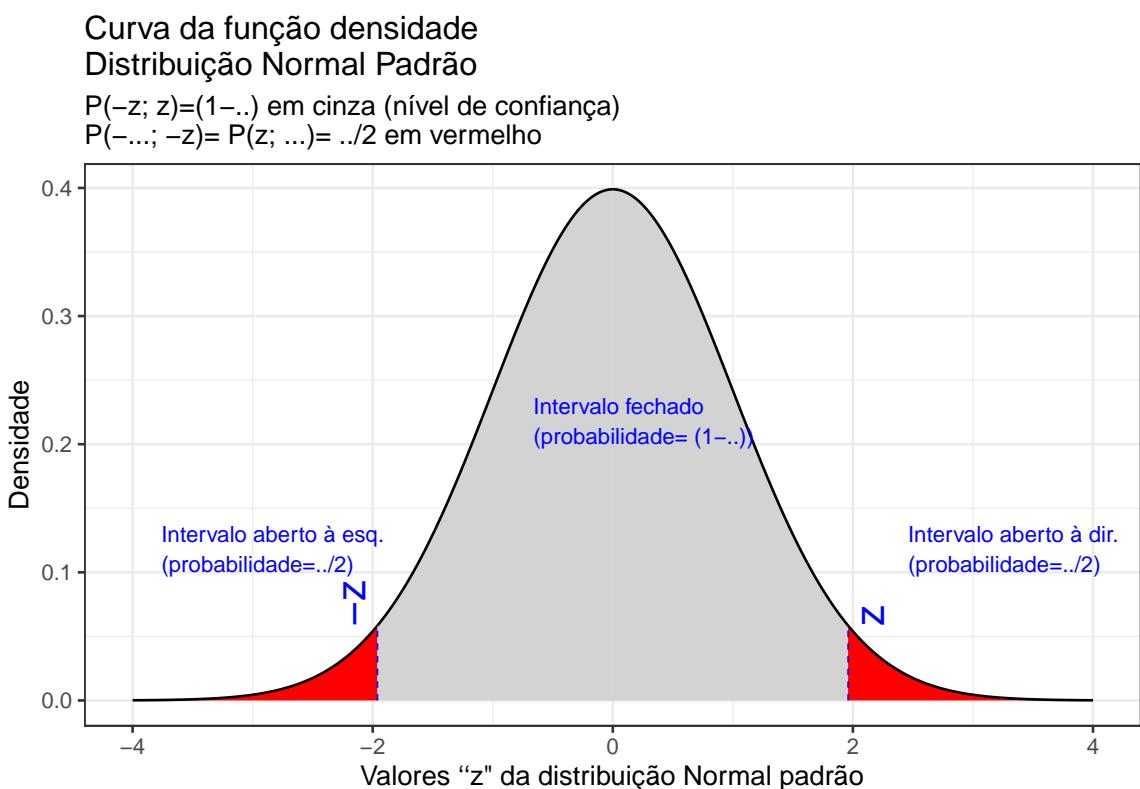


Figure 9.17: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores da estatística  $Z$  é inferior a  $\frac{\alpha}{2}$ , estabelecendo assim um intervalo com nível de confiança igual a  $(1-\alpha)$

Na Figura 9.17 observa-se:

- o nível de significância  $\alpha$ ;
  - o nível de confiança  $(1 - \alpha)$ ; e,
  - o valor tabelado da estatística  $Z(z)$  para o nível de confiança fixado.

312 MÓDULO 9. INTRODUÇÃO À DISTRIBUIÇÃO DAS MÉDIAS E DIFERENÇAS ENTRE MÉDIAS AMOSTRAIS

Assim,

$$\begin{aligned}
 P\left[-Z_{(1-\frac{\alpha}{2})} \leq Z \leq Z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\
 P\left[-z_{(1-\frac{\alpha}{2})} \leq \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{(1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\
 P[(\bar{x} - \bar{y}) - (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})] &= (1 - \alpha) \\
 IC(\mu_1 - \mu_2)_{(1-\alpha)} &= [(\bar{x} - \bar{y}) \pm z_c \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}]
 \end{aligned}$$

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  $(1 - \alpha) = 95\%$ .

$$\begin{aligned}
 P[(\bar{x} - \bar{y}) - (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})] &= (1 - \alpha) \\
 P[(\bar{x} - \bar{y}) - (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})] &= (1 - \alpha) \\
 P[10 - (1,96 \cdot \sqrt{\frac{25}{20} + \frac{9}{20}}) \leq (\mu_1 - \mu_2) \leq (10 + (1,96 \cdot \sqrt{\frac{25}{20} + \frac{9}{20}}))] &= 0,95 \\
 P[10 - (1,96 \times 1,3038) \leq (\mu_1 - \mu_2) \leq 10 + (1,96 \times 1,3038)] &= 0,95
 \end{aligned}$$

$$IC(\mu_1 - \mu_2)_{0,95} = [7; 13]$$

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:  $IC(\mu_1 - \mu_2)_{0,95} = [7; 13]$  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  $\mu_1$  e  $\mu_2$ , e variâncias  $\sigma_1^2$  e  $\sigma_2^2$  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:

$$\begin{aligned} X &\sim N(\mu_1, \frac{\sigma}{\sqrt{n_1}}) \\ Y &\sim N(\mu_2, \frac{\sigma}{\sqrt{n_2}}) \end{aligned}$$

## 314MÓDULO 9. INTRODUÇÃO À DISTRIBUIÇÃO DAS MÉDIAS E DIFERENÇAS ENTRE MÉDIAS AMOSTRAIS

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;
- $\mu_1$  e  $\mu_2$  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
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",
     geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1), color="blue"),
     geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2), color="blue"),
     annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="-t", angle=90, vjust=0, hjust=0),
     annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="t", angle=90, vjust=0, hjust=0),
     annotate(geom="text", x=t_desejado1-1.8, y=0.1, label="Intervalo aberto à esq. \n(probabilidade="),
     annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade="),
     annotate(geom="text", x=t_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade= (1-\u03b1)
  
```

Na Figura 9.18 observa-se:

- o nível de significância  $\alpha$ ;
- 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,

**Curva da função densidade  
Distribuição t (df=20)**

$P(-t; t) = (1 - \alpha)$  em cinza (nível de confiança)  
 $P(-t_{\alpha/2}; t_{\alpha/2}) = P(t_{\alpha/2}; -t_{\alpha/2}) = \alpha/2$  em vermelho

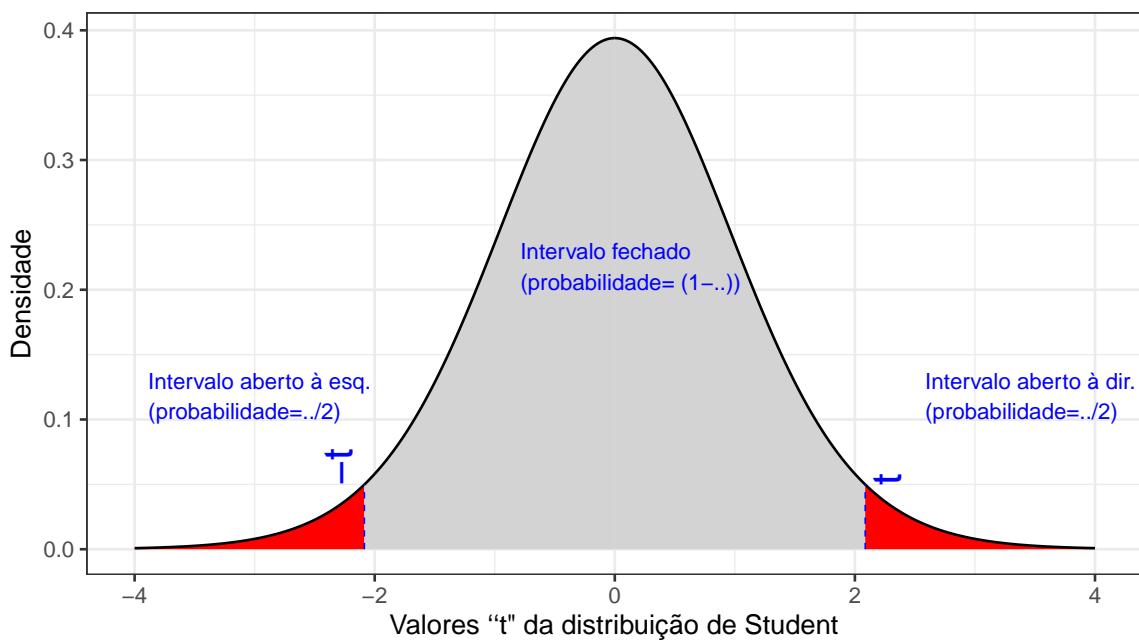


Figure 9.18: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores da estatística  $T$  ( $(n - 1)$  graus de liberdade) é inferior a  $\frac{\alpha}{2}$ , estabelecendo assim um intervalo com nível de confiança igual a  $(1 - \alpha)$

$$\begin{aligned}
 P\left[-T_{(n_1+n_2-2,1-\frac{\alpha}{2})} \leq T \leq T_{(n_1+n_2-2,1-\frac{\alpha}{2})}\right] &= (1-\alpha) \\
 P\left[-t_{(n_1+n_2-2,1-\frac{\alpha}{2})} \leq \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{(n_1+n_2-2,1-\frac{\alpha}{2})}\right] &= (1-\alpha) \\
 P[(\bar{x}-\bar{y}) - (t_{(n_1+n_2-2,1-\frac{\alpha}{2})} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}) \leq (\mu_1-\mu_2) \leq (\bar{x}-\bar{y}) + (t_{(n_1+n_2-2,1-\frac{\alpha}{2})} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})] &= (1-\alpha)
 \end{aligned}$$

$$IC(\mu_1 - \mu_2)_{(1-\alpha)} = [(\bar{x} - \bar{y}) \pm t_{c(n_1+n_2-2)} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}]$$

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 ( $\sigma_1^2$  e  $\sigma_2^2$ ) 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);
- $S_1^2 = 132,71$  e  $S_2^2 = 92,93$  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:

$$S_p = \sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}}$$

$$S_p = \sqrt{\frac{(4 - 1) \cdot 132,71 + (3 - 1) \cdot 92,93}{4 + 3 - 2}}$$

$$S_p = 10,81$$

$$P[(\bar{x} - \bar{y}) - (t_{(n_1+n_2-2, 1-\alpha)} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (t_{(n_1+n_2-2, 1-\alpha)} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})] = (1 - \alpha)$$

$$P[14 - (2,57 \cdot 10,81 \cdot \sqrt{\frac{1}{4} + \frac{1}{3}}) \leq (\mu_1 - \mu_2) \leq 14 + (2,57 \cdot 10,81 \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})] = 0,95$$

$$P[14 - 21,23 \leq (\mu_1 - \mu_2) \leq 14 + 21,23] = 0,95$$

$$IC(\mu_1 - \mu_2)_{0,95} = [-7,23; 35,23]$$

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.

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  $\mu_1$  e  $\mu_2$ , e variâncias  $\sigma_1^2$  e  $\sigma_2^2$  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:

$$\begin{aligned} X &\sim N(\mu_1, \frac{\sigma}{\sqrt{n_1}}) \\ Y &\sim N(\mu_2, \frac{\sigma}{\sqrt{n_2}}) \end{aligned}$$

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;
- $\mu_1$  e  $\mu_2$  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 ( $\nu$ ) 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,\u221e) em vermelho (probabilidade)", color="blue",
       geom_segment(aes(x = t_desejado1, y = 0, xend = t_desejado1, yend = d_desejada1), color="blue"),
       geom_segment(aes(x = t_desejado2, y = 0, xend = t_desejado2, yend = d_desejada2), color="blue"),
       annotate(geom="text", x=t_desejado1-0.1, y=d_desejada1, label="-t", angle=90, vjust=0, hjust=0),
       annotate(geom="text", x=t_desejado2+0.3, y=d_desejada2, label="t", angle=90, vjust=0, hjust=0),
       annotate(geom="text", x=t_desejado1-1.8, y=0.1, label="Intervalo aberto à esq. \n(probabilidade= 0.95)", angle=0, vjust=0, hjust=0),
       annotate(geom="text", x=t_desejado2+0.5, y=0.1, label="Intervalo aberto à dir. \n(probabilidade= 0.95)", angle=0, vjust=0, hjust=0),
       annotate(geom="text", x=t_desejado1+1.3, y=0.2, label="Intervalo fechado \n(probabilidade= (1-\u03b1)/2)", angle=0, vjust=0, hjust=0))

```

Na Figura 9.19 observa-se:

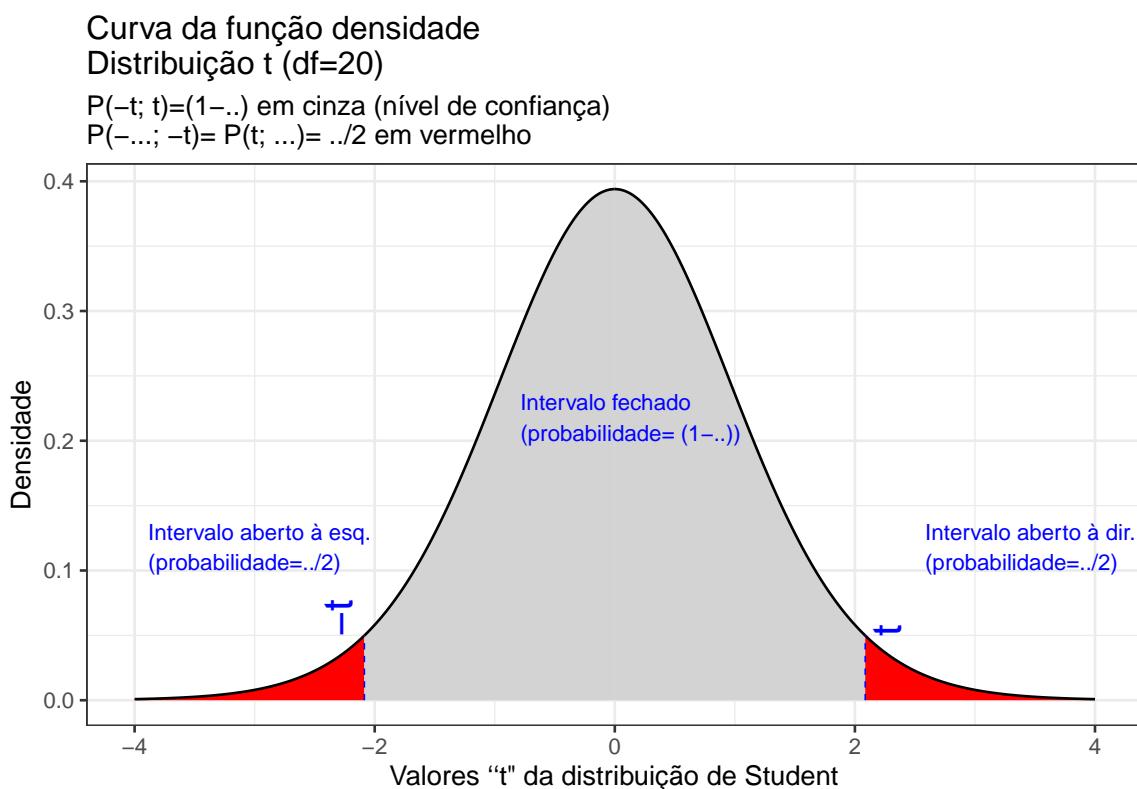


Figure 9.19: Regiões críticas, aquém e além das quais, a probabilidade associada aos valores da estatística  $T$  (com  $\nu$  graus de liberdade) é inferior a  $\frac{\alpha}{2}$ , estabelecendo assim um intervalo com nível de confiança igual a  $(1 - \alpha)$

## 322MÓDULO 9. INTRODUÇÃO À DISTRIBUIÇÃO DAS MÉDIAS E DIFERENÇAS ENTRE MÉDIAS AMOSTRAIS

- o nível de significância  $\alpha$ ;
- o nível de confiança  $(1 - \alpha)$ ; e,
- o valor tabelado da estatística  $T(t)$  sob  $\nu$  graus de liberdade para o nível de confiança fixado.

Assim,

$$\begin{aligned} P\left[-T_{(\nu, 1-\frac{\alpha}{2})} \leq T \leq T_{(\nu, 1-\frac{\alpha}{2})}\right] &= (1 - \alpha) \\ P\left[-t_{(\nu, 1-\frac{\alpha}{2})} \leq \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}} \leq t_{(\nu, 1-\frac{\alpha}{2})}}\right] &= (1 - \alpha) \\ P[(\bar{x} - \bar{y}) - (t_{(\nu, 1-\frac{\alpha}{2})} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}}) \leq (\mu_1 - \mu_2) \leq (\bar{x} - \bar{y}) + (t_{(\nu, 1-\frac{\alpha}{2})} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}})] &= (1 - \alpha) \end{aligned}$$

$$IC(\mu_1 - \mu_2)_{(1-\alpha)} = [(\bar{x} - \bar{y}) \pm t_{c(\nu)} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}}]$$

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 ( $\nu$ ) é dado por:

$$\begin{aligned}\nu &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} \\ \nu &= \frac{\left(\frac{2}{4} + \frac{14,40}{6}\right)^2}{\frac{\left(\frac{2}{4}\right)^2}{4-1} + \frac{\left(\frac{14,40}{6}\right)^2}{6-1}} \\ \nu &= \frac{2,90^2}{0,083 + 1,152} \\ \nu &= \frac{8,41}{1,23} = 6,83 \sim 7\end{aligned}$$

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 extraímos 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 da prova 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

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 ( $\mu_D$ ) 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:

$$S_D^2 = \sum_{i=1}^n \frac{(D_i - \bar{D})^2}{n-1}$$

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

$$\bar{D} = \sum_{i=1}^n D_i = -4,667$$

$$S_D^2 = \sum_{i=1}^n \frac{(Di - \bar{D})^2}{n-1} = 10,52354$$

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

### 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 Proporção amostral como uma variável Binomial

Admita a variável aleatória Binomial  $Y$  como sendo o número de sucessos observados em  $n$  tentativas de *Bernoulli*.

A proporção total de sucessos ao final das  $n$  repetições pode ser dada por  $p = \frac{Y}{n}$ .

Após um grande número de *repetições*, a proporção de sucessos observada ( $p$ ) irá se aproximar à probabilidade teórica ( $\pi$ ) da variável aleatória de *Bernoulli*. Assim, se  $\pi$  é a probabilidade de sucesso e  $\varepsilon$  é um número qualquer positivo, verifica-se:

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{Y}{n} - \pi\right| \geq \varepsilon\right) = 0$$

Considere a proporção populacional de uma característica presente em uma população e definida como:  $\pi = \frac{Y}{N}$ , onde:

- $Y$  é a variável aleatória que expressa a presença da alguma característica sob estudo nos indivíduos da população;
- $N$  é o tamanho da população; e,
- $\pi$  é a proporção populacional (habitualmente desconhecida).

Sendo  $n$  é o número de repetições dos ensaios de *Bernoulli* pode-se definir uma proporção amostral média  $\hat{p}$  em função do número de casos observados da característica em estudo (sucesso) pelo número de repetições realizadas:

$$\hat{p} = \frac{Y_n}{n}$$

Demonstra-se que  $\hat{p}$  é um bom estimador da proporção populacional  $\pi$  pois, quando  $n \rightarrow N$ ,  $Y_n \rightarrow X$  e  $\hat{p} \rightarrow \pi$ .

### 10.3 Distribuição das proporções amostrais

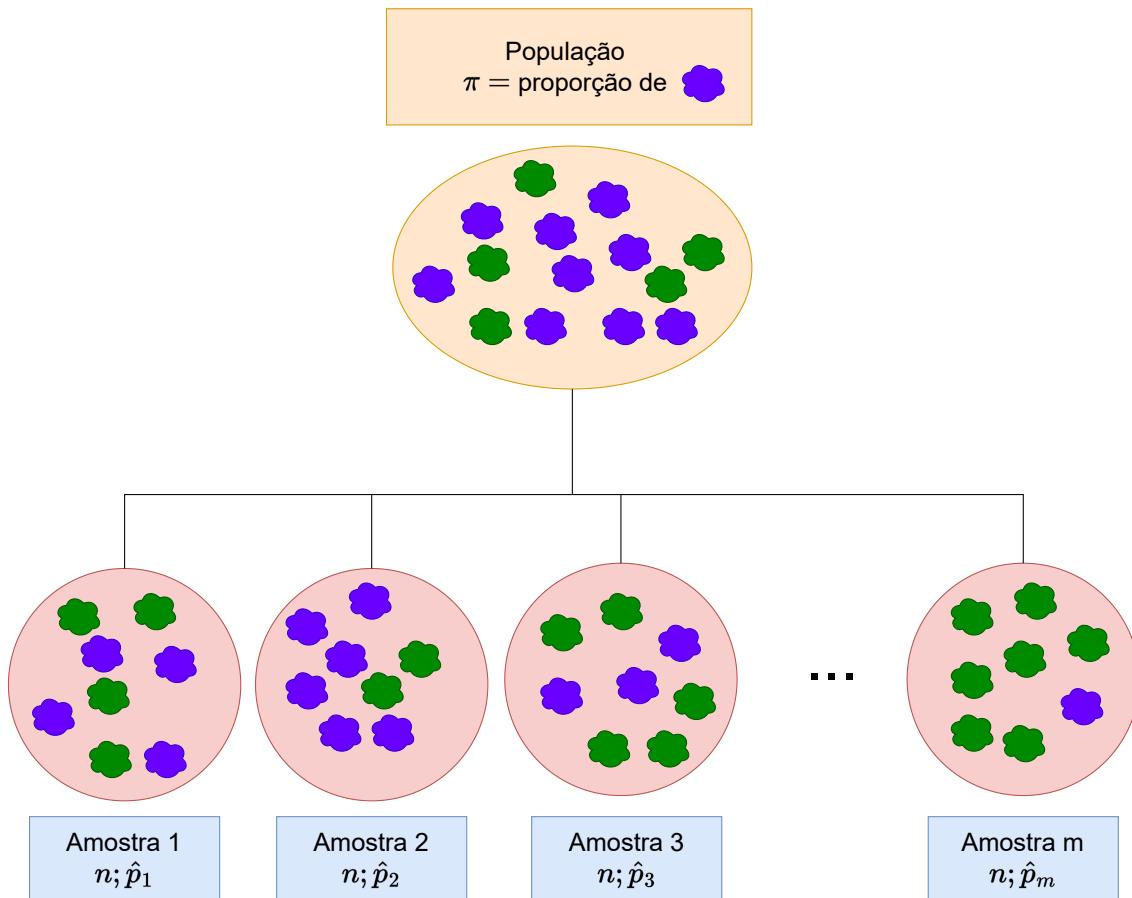


Figure 10.1: Ilustração de  $m$  amostras de mesmo tamanho ( $n$ ) extraídas de uma mesma população onde a característica de interesse se manifesta sob uma proporção populacional  $\pi$

Para estudarmos a distribuição das proporções amostrais ( $\hat{p}$ ) considerem uma população apresentando uma determinada característica de interesse com proporção  $\pi$ .

Essa característica de interesse assume apenas duas possibilidades em cada elemento da população: ela **pode ou não** estar presente. 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*.

Repetindo-se essa “extração” por  $n$  vezes, a probabilidade do número de elementos observados que apresentam essa característica de interesse pode ser estimada, por extensão, como uma variável Binomial. Dividindo-se esse número pelo número de repetições realizadas ( $n$ ) chaga-se a uma estimativa amostral  $\hat{p}$  da proporção populacional  $\pi$ .

Demonstra-se que para:

- um razoável número de repetições:  $n \geq 30$ ;
- de uma população onde a proporção  $\pi$  não é extrema: próximas a 0 ou 1; e tal que  $(n \cdot \pi)$  e  $(n \cdot (1 - \pi))$  sejam maiores que 5,

ao se repetir o experimento anotando cada proporção amostral  $\hat{p}$  verificada em cada repetição, após as  $n$  repetições de *Bernoulli*, o perfil da curva de distribuição dessas proporções amostrais torna-se razoavelmente simétrico à medida que o número ( $n$ ) de repetições cresce, para qualquer que seja a proporção populacional e oscila em torno desse valor ( $\pi$ ).

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:

$$E(Y) = n \cdot \pi \quad Var(Y) = n \cdot \pi \cdot (1 - \pi)$$

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  $\hat{p} = \mu$ :

$$E(\hat{p}) = E\left(\frac{Y}{n}\right) = \frac{1}{n} \cdot E(Y) = \frac{1}{n} \cdot n \cdot \pi = \pi$$

e a variância  $Var(\hat{p} = \frac{1}{n} \cdot \pi \cdot (1 - \pi))$ :

$$Var(\hat{p}) = Var\left(\frac{Y}{n}\right) = \frac{1}{n^2} \cdot Var(Y) = \frac{1}{n^2} \cdot n \cdot \pi \cdot (1 - \pi) = \frac{1}{n} \cdot \pi \cdot (1 - \pi).$$

Assim, a proporção amostral segue uma distribuição aproximadamente Normal com média  $\mu = \pi$  e variância  $\sigma^2 = \frac{\pi \cdot (1 - \pi)}{n}$ :

$$\hat{p} \sim N\left(\pi; \frac{\pi \cdot (1 - \pi)}{n}\right)$$

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 das 10.000 proporções amostrais extraídas de uma população com proporção  $\pi = p_1$ .

```
#####
# Considere uma população dicotômica de tamanho N_1 com dois tipos de elementos.
# A proporção de elementos com a propriedade A (sucesso) é p_1,
# enquanto que a proporção de elementos que não têm a propriedade A é (1-p_1)
#####

#número de amostras escolhido
N_1=10000

#proporção escolhida para a manifestação da característica (sim/não)
p_1=round(1/6,2)

#####
# Selecionando-se aleatoriamente um elemento desta população
# resulta em uma variável aleatória dicotômicas/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 re
# 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 frequência re
# a proporção observada s na amostra de tamanho `n_1` é também uma variável aleatória (p) com
# com distribuição altamente relacionada à variável X pois é a média de `n_1` ensaios de Bernoulli
#
# Repetindo-se sucessivamente `N_1` vezes extrações de tamanho `n_1`
```

## 332MÓDULO 10. INTRODUÇÃO À DISTRIBUIÇÃO DAS PROPORÇÕES AMOSTRAIS E SEUS INTERVALOS DE CONFIANÇA

```
# a anotando-se a proporção de sucesso em cada uma dessas amostras poderemos analisar como eles se comportam
# em relação à quantidade de elementos extraídos `n_1` (repetições de Bernoulli) e à verdadeira propriedade A
# com que a propriedade A se manifesta na população ( $\pi$ )
#
# Para `n_1` suficientemente grande (amostrado com reposição):
#  $n_1 * \pi > 5$  e  $n_1 * (1 - \pi) > 5$ 
# a distribuição de  $p$  pode ser aproximada pela distribuição Normal  $p \sim N(\mu, \sigma)$ 
# onde  $\mu$  e  $\sigma$  são aproximados por:
#  $\mu = E(p) = \pi$ 
#  $\sigma^2 = \sigma^2 = \pi * (1 - \pi) / n_1$ 
# sigma = sqrt[ pi * (1 - pi) / n_1 ]
#
#####
#
#tamanho escolhido para cada amostra (elementos sorteados)
n_1=10

#
#vetor com o número de sucessos observados (a frequência absoluta) nas  $N_1$  amostras de  $n_1$  elementos
suc_10rep=rbinom(n=N_1, 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
prop_10rep=suc_10rep/n_1
prop_10rep
dados_10=as.data.frame(prop_10rep)

#####
#
# O mesmo procedimento, mas agora com amostras com um maior número de elementos em cada uma
#####
#
#número de amostras escolhido
N_2=10000 #Mesma população de elementos binomiais de tamanho  $N_2$  com probabilidade  $p_1$ 

#
#tamanho escolhido para cada amostra (elementos sorteados)
n_2=100

#
#vetor com o número de sucessos observado (a frequência absoluta) nas  $N_2$  amostras de  $n_2$  elementos
suc_100rep=rbinom(n=N_2, size = n_2, prob = p_1)
suc_100rep

#
#vendo a proporção de sucessos (a frequência relativa) em cada uma das  $N_2$  amostras de  $n_2$  elementos
prop_100rep=suc_100rep/n_2
prop_100rep
dados_100=as.data.frame(prop_100rep)

#
#
#
#
#meu_titulo1=paste("Distribuição das frequências das proporções de sucesso observadas em \n",N_1, "
#                    "elementos dicotômicos extraídos (com reposição) da população","\n(proporção de
#meu_titulo2=paste("As proporções amostrais ~ \nN[x= \u03c0=",round(mean(dados_10$prop_10rep),2),",
#                    "]",sep=""))
#
#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 = mean(dados_10$prop_10rep), sd = sd(dados_10$prop_10rep)),
              colour="red") +
scale_y_continuous(name="",breaks = NULL) +
scale_x_continuous(name="Valores das proporções amostrais médias") +
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))

```

**Distribuição das frequências das proporções de sucesso observadas em 10000 amostras de n= 10 elementos dicotômicos extraídos (com reposição) da população (proporção de sucesso na população ..= 0.17 )**

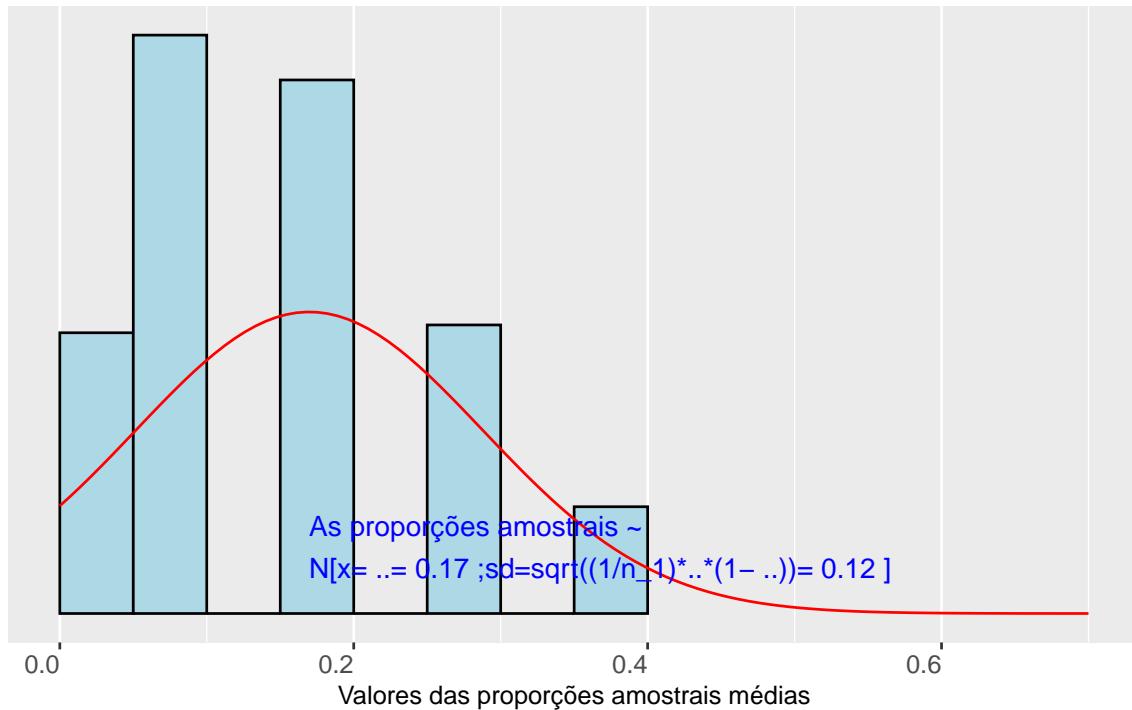


Figure 10.2: Distribuição das frequências das proporções de sucesso observadas em 10.000 amostras de tamanho n=10 elementos dicotômicos extraídos (com reposição) de uma população (a proporção de sucesso na população =1/6)

```

meu_titulo1=paste("Distribuição das frequências das proporções de sucesso observadas em \n",N_2,
                  "elementos dicotômicos extraídos (com reposição) da população","\n(proporção de",
                  meu_titulo2=paste("As proporções amostrais ~ \nN[x= \u03c0=\",round(mean(dados_100$prop_100rep),2),
ggsplot(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 = mean(dados_100$prop_100rep), sd = sd(dados_100$prop_100rep)),
              colour="red") +
scale_y_continuous(name="",breaks = NULL) +
scale_x_continuous(name="Valores das proporções amostrais médias") +
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=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 frequências das proporções de sucesso observadas em 10000 amostras de n= 100 elementos dicotômicos extraídos (com reposição) da população (proporção de sucesso na população ..= 0.17 )**

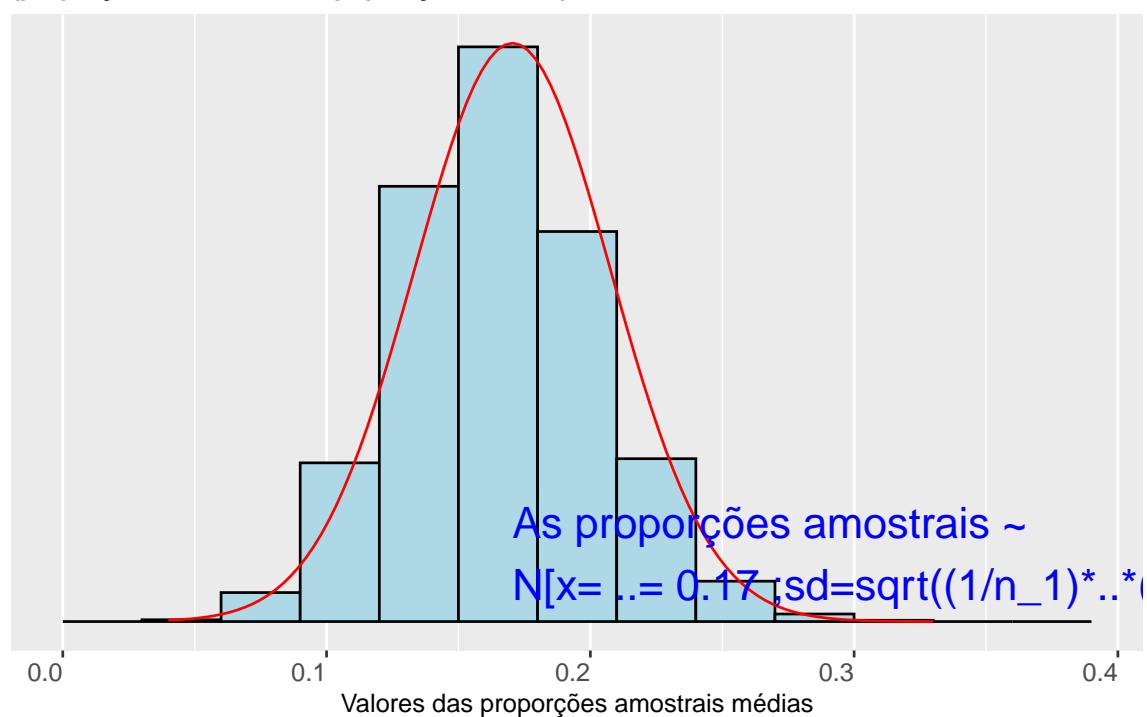


Figure 10.3: Distribuição das frequências das proporções de sucesso observadas em 10.000 amostras de tamanho n=100 elementos dicotômicos extraídos (com reposição) de uma população (a proporção de sucesso na população =1/6)

Definindo-se a estatística  $Z$  como a simples padronização da variável  $\hat{p}$ , temos 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  $\pi$  não próximo de 0 ou 1).

## 10.4 Intervalo de confiança para proporções amostrais

Podemos escrever o parâmetro ( $\pi$ ) 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  $\pi$  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:

- tamanhos amostrais  $n > 30$ ; e,
- $\hat{p}$  não muito próximos a 0 ou 1 tal que  $n \cdot \hat{p} \geq 15$  e  $n \cdot (1 - \hat{p}) \geq 15$  (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  $\pi$  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  $\pi$  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  $\pi$  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):

$$P[\pi \geq \hat{p} - z_{\alpha} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}] = (1 - \alpha)$$

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);
- $\pi$  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 ( $\pi$ ) sob um nível de confiança  $(1 - \alpha)$  de 0,95 estará delimitado:

$$\begin{aligned} \hat{p} - z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} &\leq \pi \leq \hat{p} + z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \\ 0,34 - 1,96 \cdot \sqrt{\frac{0,34 \cdot (1 - 0,34)}{400}} &\leq \pi \leq 0,34 + 1,96 \cdot \sqrt{\frac{0,34 \cdot (1 - 0,34)}{400}} \\ 0,2936 &\leq \pi \leq 0,3864 \end{aligned}$$

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);
- $\pi$  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 ( $\pi$ ) sob um nível de confiança ( $1 - \alpha$ ) de 0,99 estará delimitado:

$$\begin{aligned}\hat{p} - z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} &\leq \pi \leq \hat{p} + z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \\ 0,43 - 2,58 \cdot \sqrt{\frac{0,43 \cdot (1 - 0,43)}{2000}} &\leq \pi \leq 0,43 + 2,58 \cdot \sqrt{\frac{0,43 \cdot (1 - 0,43)}{2000}} \\ 0,4014 &\leq \pi \leq 0,4586\end{aligned}$$

#### 10.4.1 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  $\pi_X$  e  $\pi_Y$  a partir das proporções obtidas em duas amostras de razoável tamanho ( $n_X \geq 30$  e  $n_Y \geq 30$ ) e proporções amostrais

*há*  $p_X$  e  $p_Y$  não extremas (próximas 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, demonstran-se que se pode tomar as proporções amostrais  $\hat{p}_X$  e  $\hat{p}_Y$  como aproximações diretas das proporções populacionais  $\pi_X$  e  $\pi_Y$  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  $\pi_X$  e  $\pi_Y$  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:

$$P \left[ (\hat{p}_X - \hat{p}_Y) - z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \leq (\pi_X - \pi_Y) \leq (\hat{p}_X - \hat{p}_Y) + z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \right]$$

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 Epistemologia

Estritamente falando, todo o conhecimento fora da matemática e da lógica demonstrativa (um ramo da mesma) encontra-se baseado em conjecturas.

Naturalmente há inúmeros tipos de conjecturas, algumas das quais altamente respeitáveis e confiáveis como as expressas em certas leis gerais da física e da química, por exemplo.

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

Ao contrário, o método experimental (*indutivo*) é próprio das ciências naturais, que observam seus objetos e realizam experimentos.

O *raciocínio 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.

No caso das ciências naturais (física, química, biologia, etc.), o método é chamado *experimental e hipotético*.

Experimental porque se baseia em observações e em experimentos, tanto para formular quanto para verificar as teorias.

Hipotético porque os cientistas partem de hipóteses sobre os objetos que guiam os experimentos e a avaliação dos resultados.

O método experimental é hipotético-indutivo e hipotético-dedutivo.

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.

A lei científica obtida por via indutiva ou dedutiva permite descrever, interpretar e compreender um campo de fenômenos semelhantes e prever novos, a partir dos primeiros.

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.

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)

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.

Um teste de hipóteses refere-se, portanto, em uma metodologia quantitativa subsidiária em processos de decisão baseada na inferência estatística e de ampla aplicabilidade na experimentação e pesquisa, virtualmente, em qualquer área do conhecimento.

## 11.2 Histórico

Referências vagas a testes 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 (cf. Figura ??) sobre um procedimento teste de se isolar 200 ou 500 pacientes de um hospital com febre e pleurite em dois grupos iguais e aplicar a eles diferentes tratamentos e, ao final de um período de tempo verificar quantos funerais ocorreram num e no outro.

Em 1932 Karl Pearson se aposentou como 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:

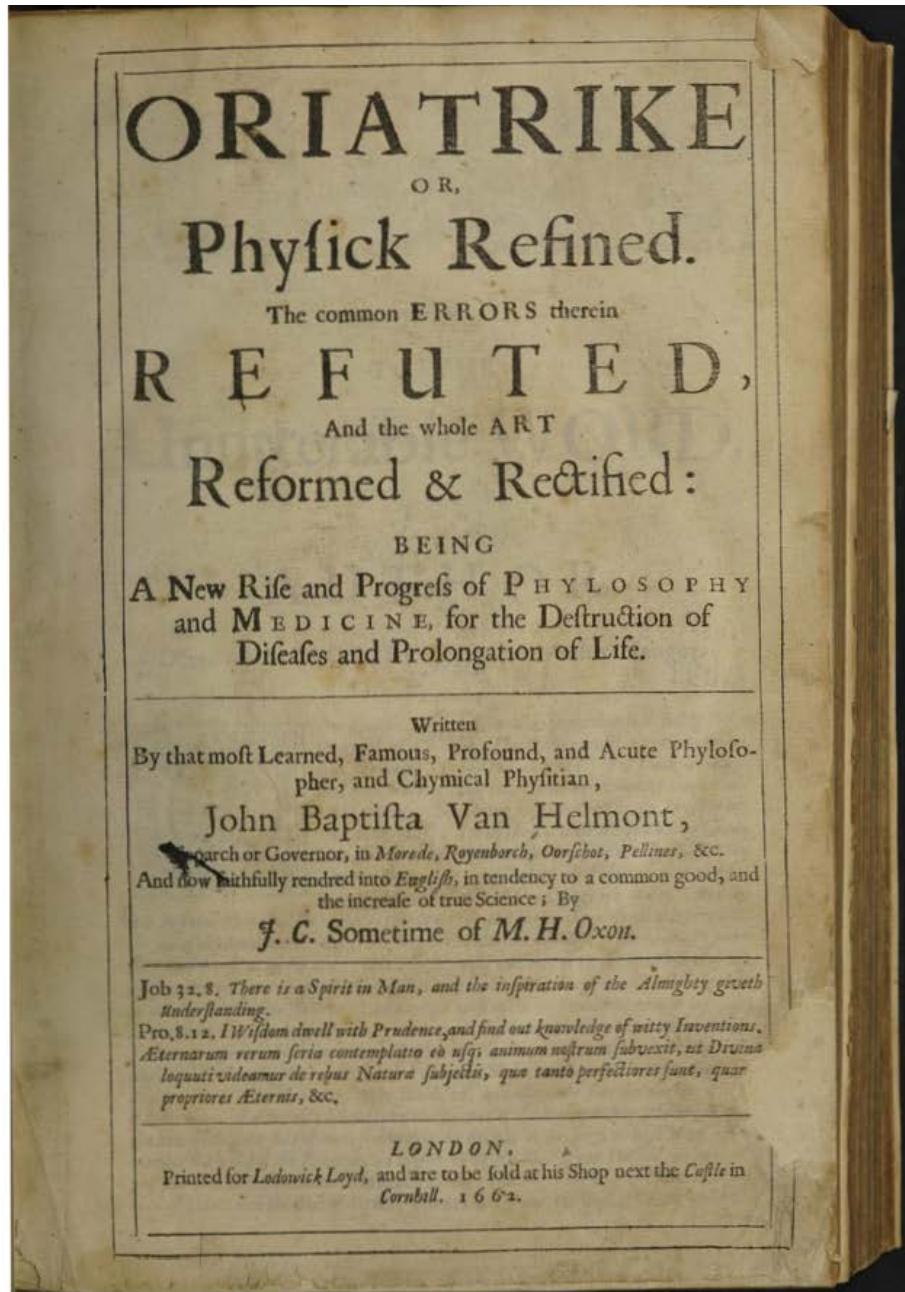


Figure 11.1: Oriatrike or, physick refined. The common errors therein refuted, and the whole art reformed and rectified: being a new rise and progress of phylosophy and medicine, for the destruction of diseases and prolongation of life (p. 526)



Figure 11.2: Tratamento mais utilizado à época (sangria)

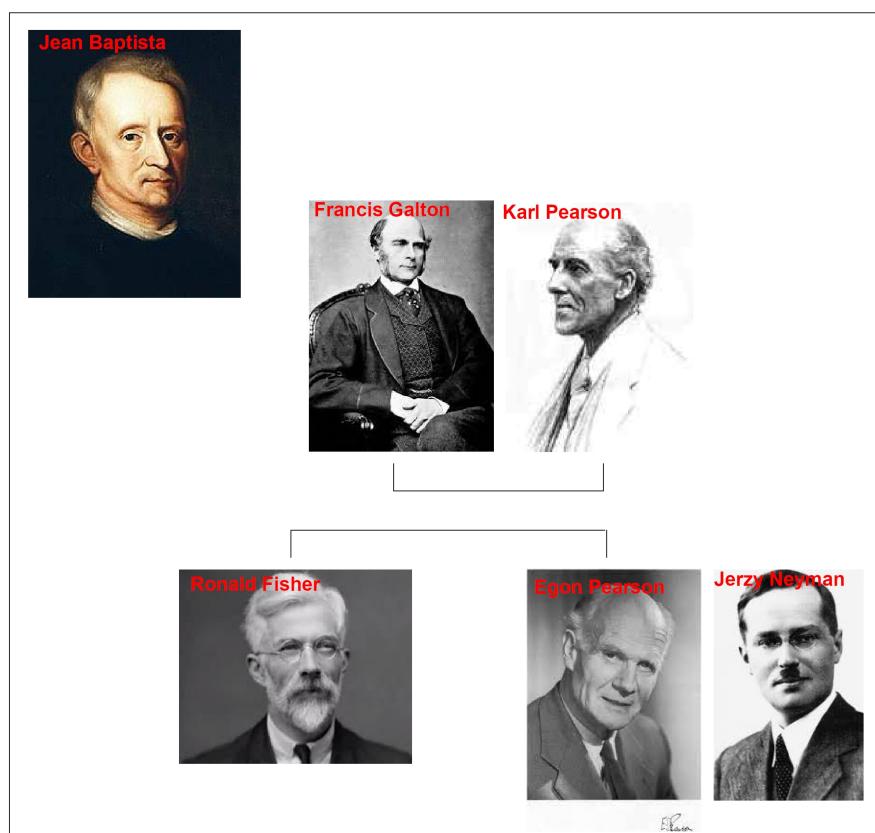


Figure 11.3: Personagens históricos

- 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*).

Testes de hipóteses quando feitos sobre os parâmetros da população são chamados de Testes paramétricos.

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.

“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 dar aos fatos a chance de refutar a *hipótese nula...*” (*The Design of Experiments*, Ronald Aylmer Fisher, 1935, p. 19)

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.

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.

### 11.3 Conceitos iniciais

A metodologia desenvolvida para a realização de um teste de hipóteses no fornece elementos auxiliares da decisão de rejeitar ou não, sob um prisma probabilístico, determinada conjectura acerca de um parâmetro da população estudada.

Ela nos possibilita associar um *nível de significância* ( $\alpha$ ) na tomada de decisão de modo a *minimizar* a chance de erro de se concluir pela *rejeição* de uma *hipótese verdadeira* previamente formulada\*.

Nível de significância ( $\alpha$ ) 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 nula quando ela é verdadeira.

Nível de confiança ( $1 - \alpha$ ) é a medida da confiabilidade de nossa conclusão no teste de hipóteses.

A *hipótese nula* é a hipótese que reflete a situação em que não há mudança, sendo, pois, uma hipótese conservadora. É aquela em que temos mais confiança (resultado de experiências passadas).

*Inicialmente* ela é assumida como verdadeira para, logo a seguir, ser confrontada com a informação amostral para se verificar a consistência 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).

## 11.4 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,  $P\{\lim_{n \rightarrow \infty} (\frac{X_1 + X_2 + \dots + X_n}{n}) = \mu\} = 1$

## 11.5 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,  $\varepsilon$  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  $\mu$  e variância  $\sigma^2$ .

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 comparam médias são geralmente robustos à não normalidade.

## 11.6 Diretrizes gerais de um teste de hipóteses

- o pesquisador deve delimitar o objeto de sua pesquisa;
- estabelecer um nível apropriado para a significância  $\alpha$  (em alguns campos do conhecimento níveis de significância muito reduzidos são impraticáveis);
- uma boa hipótese deve ser baseada em uma boa pergunta sobre o objeto do estudo;
- deve ser simples e específica; e,
- 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); e,
- 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 ( $\alpha$ ) é, em geral, de concordância com o parâmetro que se estuda da população (conservadora) e baseada em conhecimento prévio; e,
- a hipótese alternativa ( $H_1$ ) é contrária, oposta, antagônica à hipótese nula (novadora).

## 11.7 Formulação e estruturação de um teste de hipóteses

- 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  $\chi^2$ ;
- 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  $\alpha$ ;
- 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 Natureza dos erros envolvidos em um teste de hipóteses

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  $\alpha$ .

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  $\beta$ .

No quadro acima identificam-se:

Table 11.1: \*

Valor real do parâmetro (desconhecido)	Não rejeitar $H_0$
$H_0$ verdadeira	Decisão correta probabilidade associada= $(1 - \alpha)$
$H_0$ falsa	Erro do tipo II probabilidade associada= $\beta$

- $\alpha$ : 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);
- $\beta$ : 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 I: 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	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 medicamente “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 ( $\beta$ ) e, consequentemente, aumenta o poder do teste ( $1 - \beta$ );
- aumentar o nível de significância ( $\alpha$ ) 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 ( $\beta$ ).