



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, 22 de janeiro de 2023.

Contents

	7
1 - Introdução histórica da estatística	9
1.1 Primeiros levantamentos, estudos e publicações & Demografia e aritmética política	9
1.2 Visualização de dados & Estudos e primeiras publicações	19
1.3 Nomes notáveis	23
1.4 Revista Biometrika	24
1.5 Eugenia	25
2 - Introdução conceitual essencial	31
2.1 Estatística descritiva	31
2.2 Estatística inferencial	32
2.3 Produção de conhecimento	33
2.4 População (universo) & amostra	34
2.5 Parâmetros e estatísticas	35
2.6 Tipos de variáveis	35
2.7 Noções básicas sobre somatórios (Σ)}	36
2.8 Análise combinatória: diagramas de árvore, permutações (arranjos) & combinações	39
2.9 Conectivos lógicos	46
2.10 Leis de De Morgan	46
2.11 Noções básicas para o uso de calculadora (Cassio fx-82MS)	48

3 - Introdução à estatística descritiva	53
3.1 Análise exploratória	53
3.2 Dados brutos, em rol, diagrama de ramos & folhas e de dispersão unidimensional	54
3.3 Sínteses numéricas descritivas	56
3.4 Medidas de forma (assimetria & curtose)	70
3.5 Apresentação gráfica de dados	73
3.6 Apresentação tabular de dados quantitativos	79
3.7 Apresentação tabular de dados qualitativos	89
4 - Introdução ao cálculo de probabilidades	93
4.1 Introdução conceitual essencial	94
4.2 Probabilidade	104
4.3 Teorema de Bayes	123
4.4 Demonstração clássica de independência	137
4.5 Demonstração clássica de dependência	140
4.6 Teoremas da Teoria das probabilidades	145
5 - Introdução a variáveis aleatórias	153
6 - Introdução a modelos teóricos de probabilidade	155
7 - Introdução ao planejamento de pesquisas	157
8 - Introdução à estatística na epidemiologia	159
9 - Introdução à distribuição das médias amostrais, suas diferenças e seus intervalos de confiança	161
10 - Introdução à distribuição das proporções amostrais, suas diferenças e seus intervalos de confiança	163
11 - Introdução a testes de hipóteses	165
12 - Introdução ao modelo clássico de regressão linear simples	167

CONTENTS

7

13 - Introdução à análise multivariada: discriminante linear de Fisher **169**

14 - Introdução à estatística experimental: análise de variância (DIC e DBC) **171**

Chapter 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.

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).

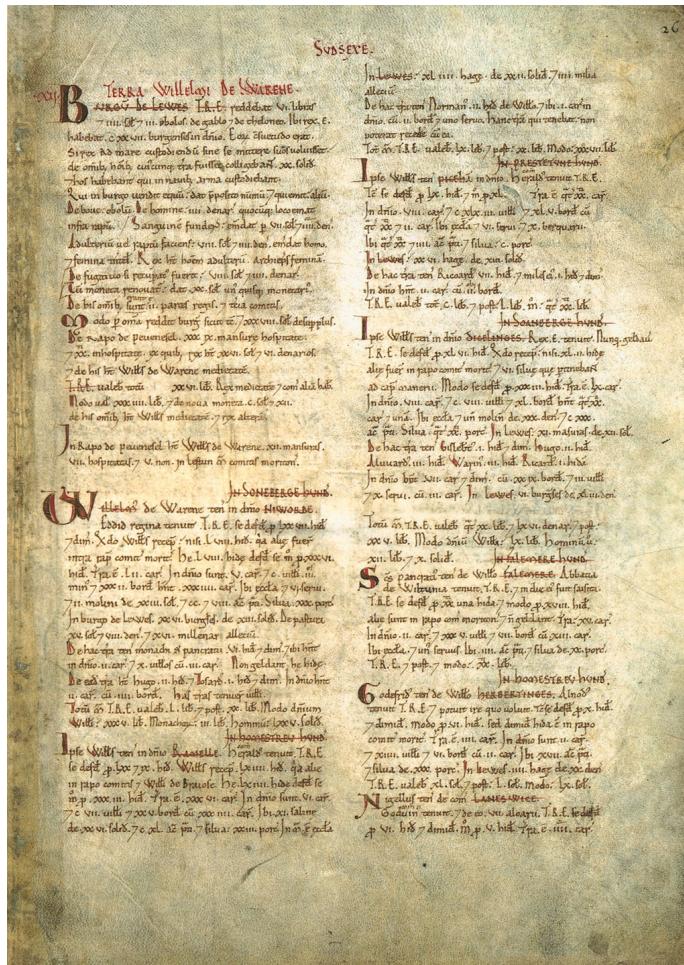


Figure 1.1: Domesday Book

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

“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

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.

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

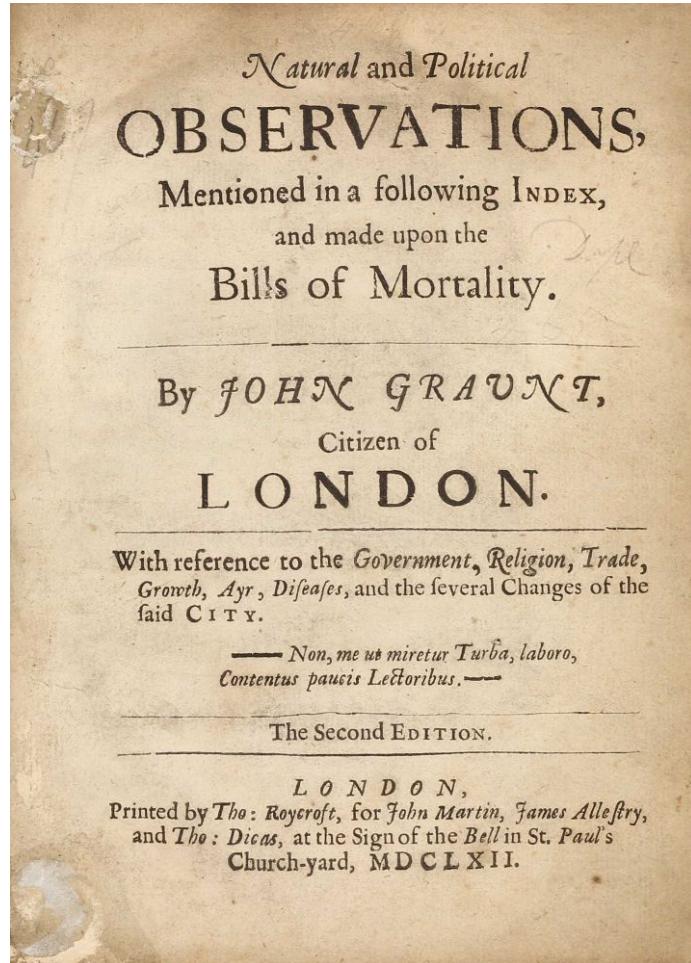


Figure 1.2: Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality (ed. de 1662)

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

Em 1687 o economista e filósofo inglês William Petty (1623-1687) publicou *Five Essays on Political Arithmetic* (Cinco ensaios sobre aritmética política), sugerindo ao governo inglês a criação de um departamento para registro de **estatísticas** vitais.

O Capitão John Graunt e William Petty instituiram na Inglaterra um novo ramo de estudos denominado de *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.

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).

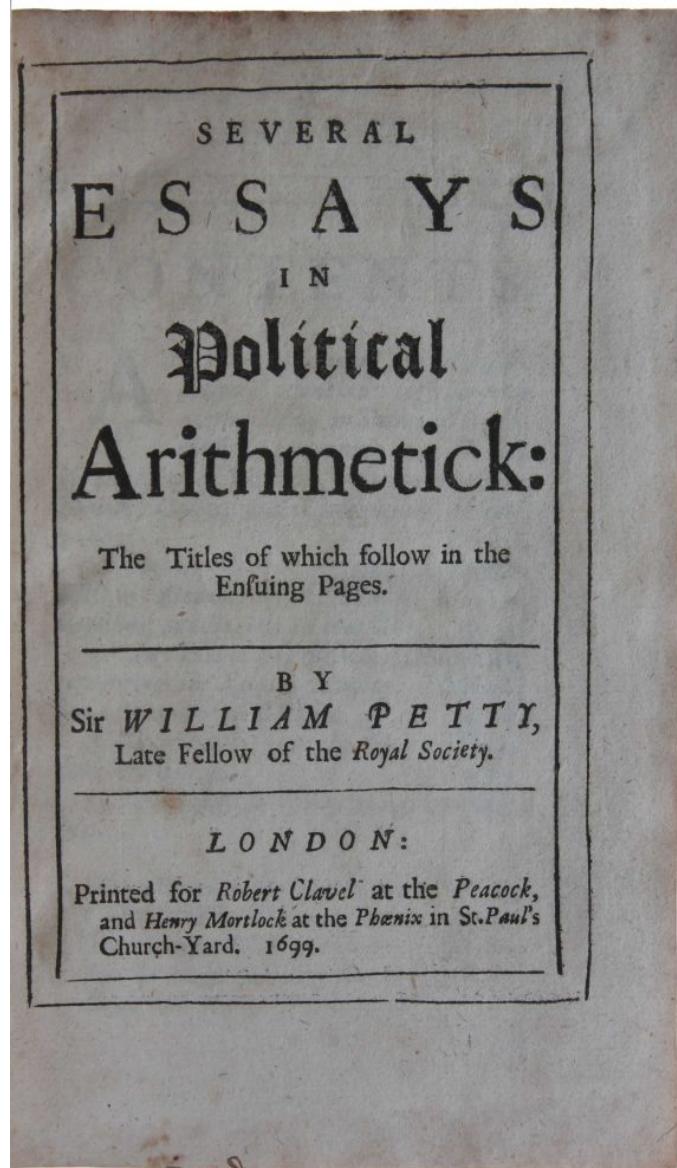


Figure 1.3: Several Essays in Political Arithmetick (ed. de 1699)

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

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).

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).



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

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

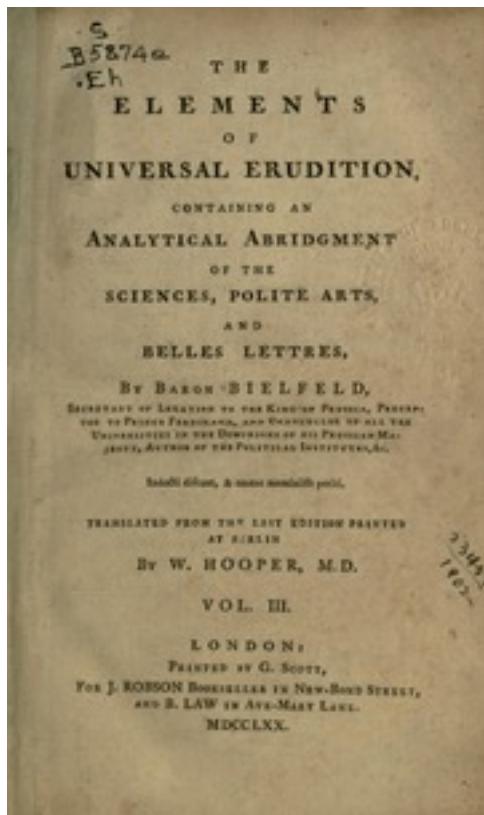


Figure 1.6: The Elements of Universal Erudition (1771)

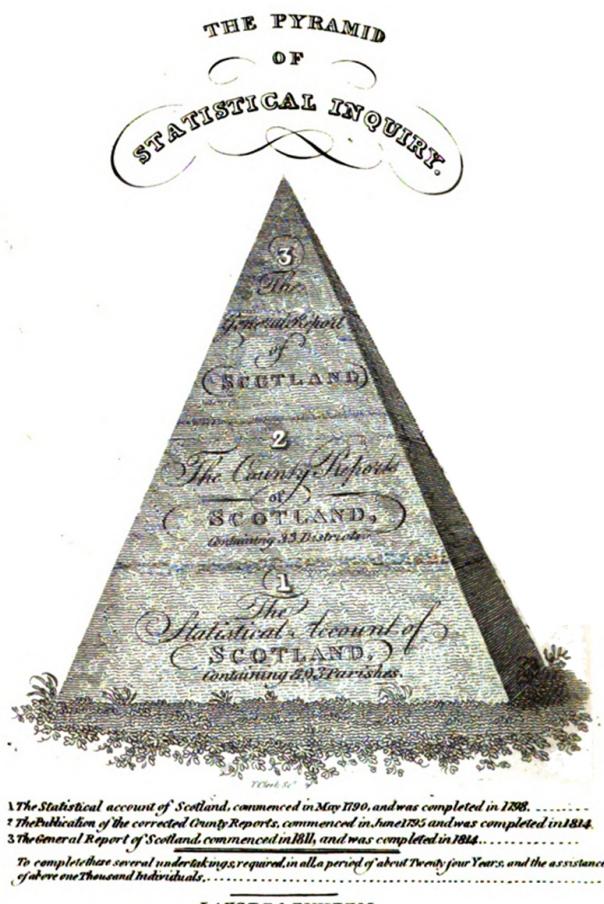


Figure 1.7: The Pyramid of Statistical Enquiry (1814)

1854

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

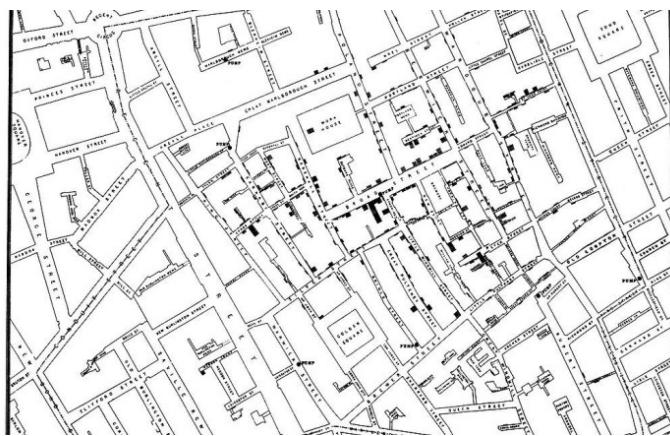


Figure 1.8: Mapa dos casos de cólera (1854)

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.

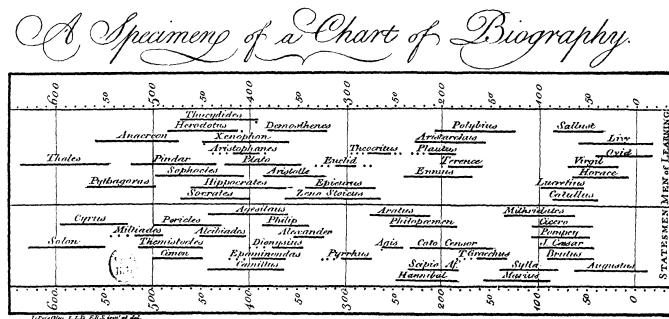


Figure 1.9: Expectativa de vida de diversas pessoas (1765)

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; e,
 - em 1801, o gráfico de setores.

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](#)).

1.2. VISUALIZAÇÃO DE DADOS & ESTUDOS E PRIMEIRAS PUBLICAÇÕES 23

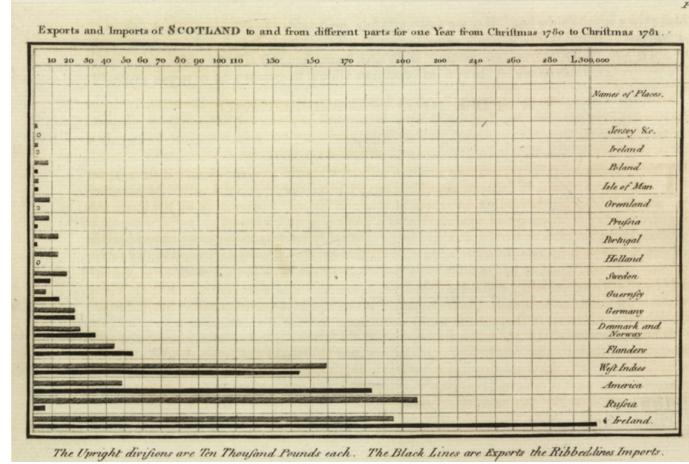


Figure 1.10: Commercial and Political Atlas (Atlas Comercial e Político de 1786): cada barra representa as exportações e importações da Escócia para 17 países em 1781

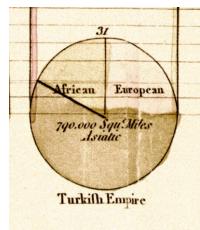


Figure 1.11: Statistical Breviary (Breviário Estatístico de 1801): proporção da extensão do Império Turco em diferentes regiões do mundo: Ásia, Europa e África, antes de 1789

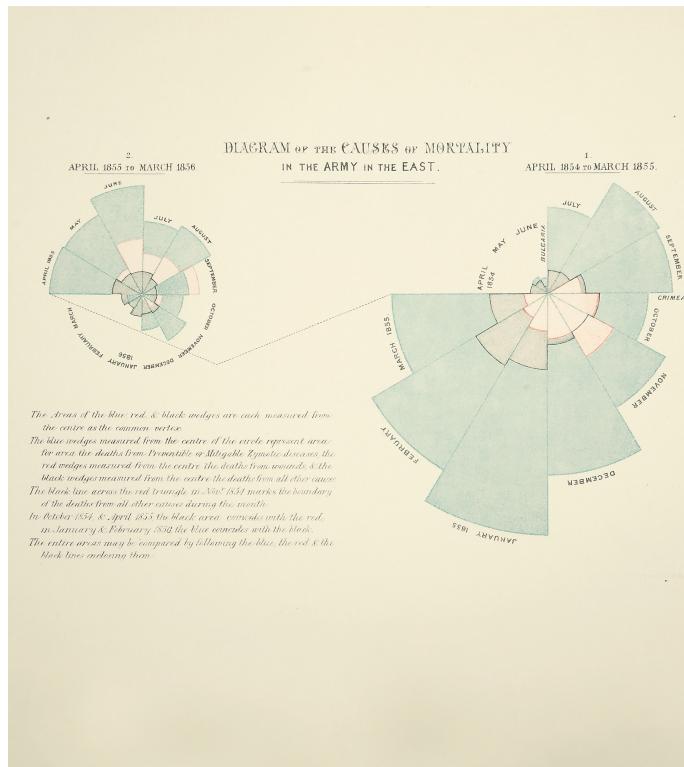


Figure 1.12: Esse diagrama (coxcomb) feito durante a Guerra da Crimeia foi dividido igualmente em 12 setores, representando os meses do ano, com a área sombreada do setor de cada mês proporcional à taxa de mortalidade naquele mês. Seu sombreamento com código de cores indicava a causa da morte em cada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

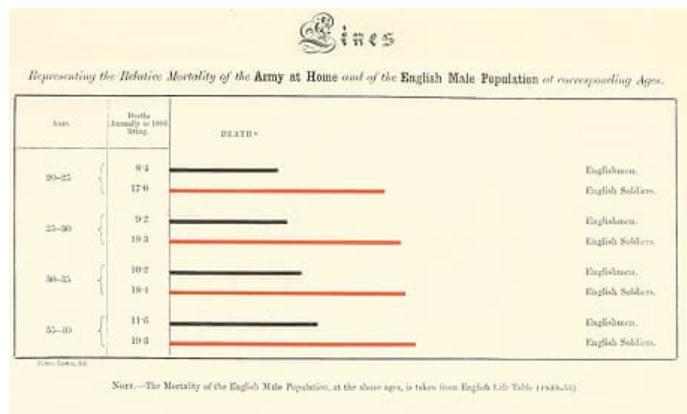


Figure 1.13: Gráfico de barras de Florence Nightingale mostrando as diferenças de mortalidade entre soldados britânicos e a população masculina inglesa geral (civis)

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](#)).

- 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 Quetelet (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 Mikhaïlovich 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.

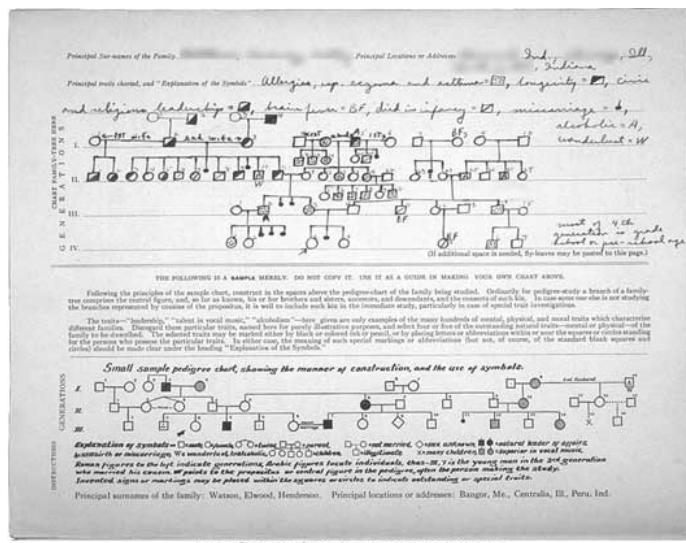


Figure 1.14: Gráfico de linhagens para alergias

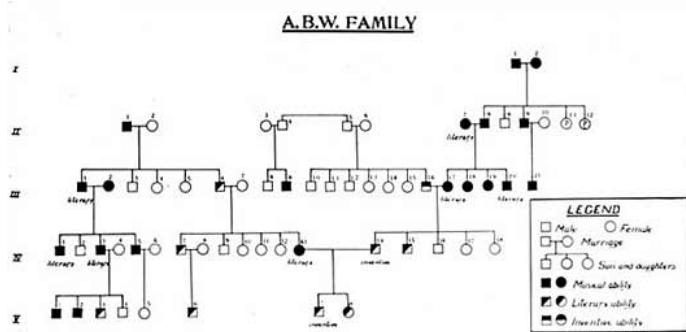


Figure 1.15: Gráfico de linhagens para aptidão musical

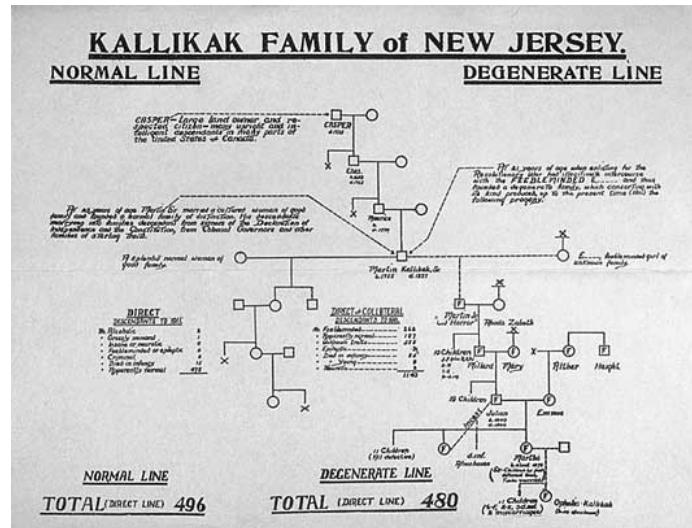


Figure 1.16: Linhas "normais" e "degeneradas" da família Kallikak (New Jersey)

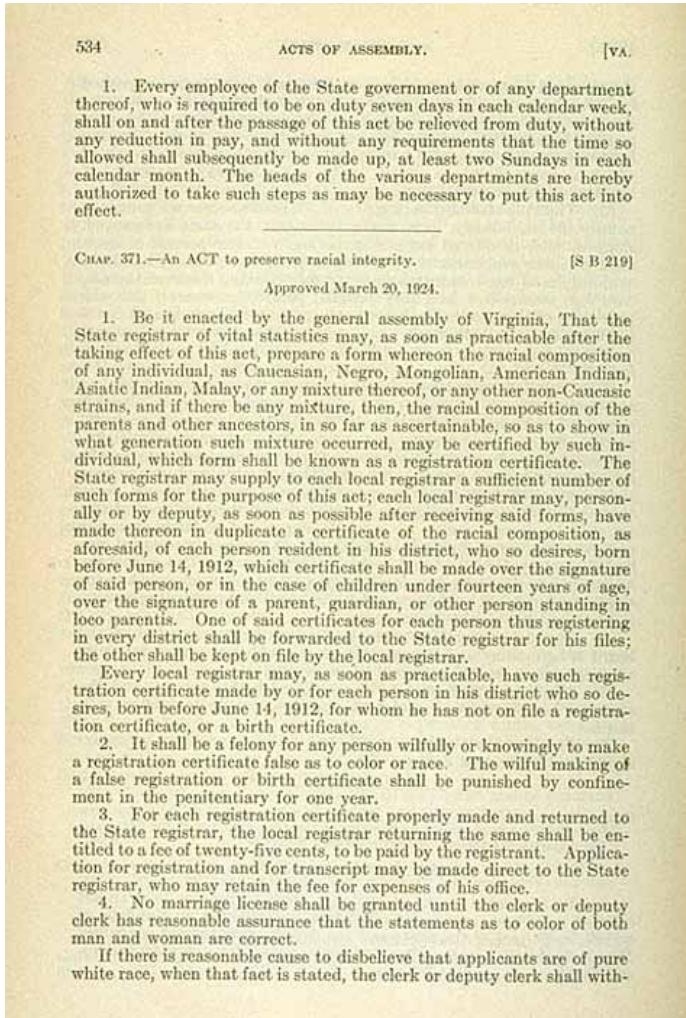


Figure 1.17: Lei da Inegridade Racia (Virginia, EUA, 1924)



Figure 1.18: Licença para casamento

Chapter 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

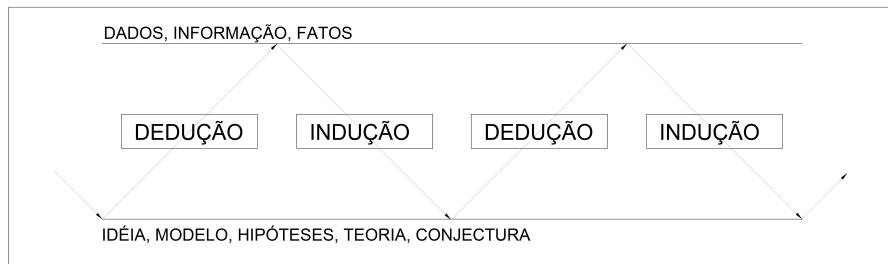


Figure 2.1: Fluxograma elementar de um processo de aprendizagem

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.”
- 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

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**).



Figure 2.2: Universo e amostra

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

2.5 Parâmetros e estatísticas

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

Característica estudada	Notação populacional	Notação amostral
Número de elementos	N	n
Média	μ	\bar{x}
Variância	σ^2	s^2
Desvio padrão	σ	s
Proporção	Π	p

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

Figure 2.3: Alfabeto grego

2.6 Tipos de variáveis

- discretas: são dados com um pouco menos de informação que os de natureza contínua mas possuem mais informação que dados qualitativos: número de andares de um prédio, de degraus de uma escada, número de filhos de um casal.

Variáveis qualitativas

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

Codificação de variáveis qualitativas

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

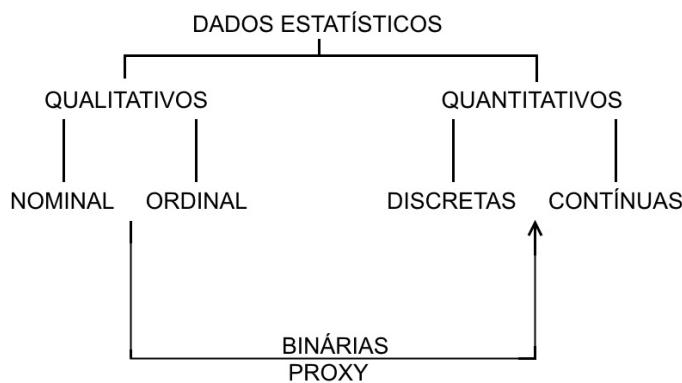


Figure 2.4: Tipos e codificações de variáveis

2.7 Noções básicas sobre somatórios (Σ)}

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

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

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

Acima da letra Σ 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

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

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

ou elevá-la ao quadrado:

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

Atenção para a diferença entre essas duas operações:

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

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**.

2.8 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.

Um diagrama como ilustrado na Figura 2.5 (denominado **diagrama de árvore** em virtude de sua aparência) geralmente é usado para explicar o princípio acima

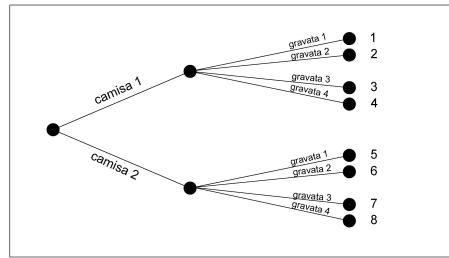


Figure 2.5: Diagrama de árvore

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 2.6):

Sejam os eventos mutuamente exclusivos E_1 com n_1 possibilidades distintas de ocorrer, E_2 com n_2 , ..., E_n com n_k ; então o número total de possibilidades de ocorrer **pelo menos um desses eventos** será dado por:

$$n_2 + n_2 + \dots + n_k$$

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

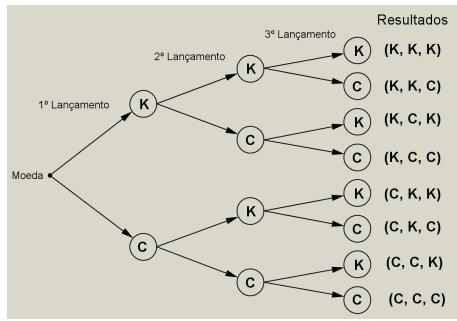


Figure 2.6: Diagrama de árvore

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:

$$n_1 + n_2 = 5$$

2.8.1 Permutações ou arranjos

O conceito de uma permutação (arranjo) refere-se a uma relação de n objetos distintos que serão agrupados p a 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?**

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

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

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**?

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

2.8.2 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?

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

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

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).

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

2.8.3 Observações acerca de alguns fatoriais

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

2.9 Conectivos lógicos

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

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

2.10 Leis de De Morgan

Augustus de Morgan foi um matemático e lógico indiano.



Figure 2.7: 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,
- $(p \cup q)^c = (p^c) \cap (q^c)$.

2.11 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.

2.11. NOÇÕES BÁSICAS PARA O USO DE CALCULADORA (CASSIO FX-82MS)51

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**.

2.11. NOÇÕES BÁSICAS PARA O USO DE CALCULADORA (CASSIO FX-82MS)53

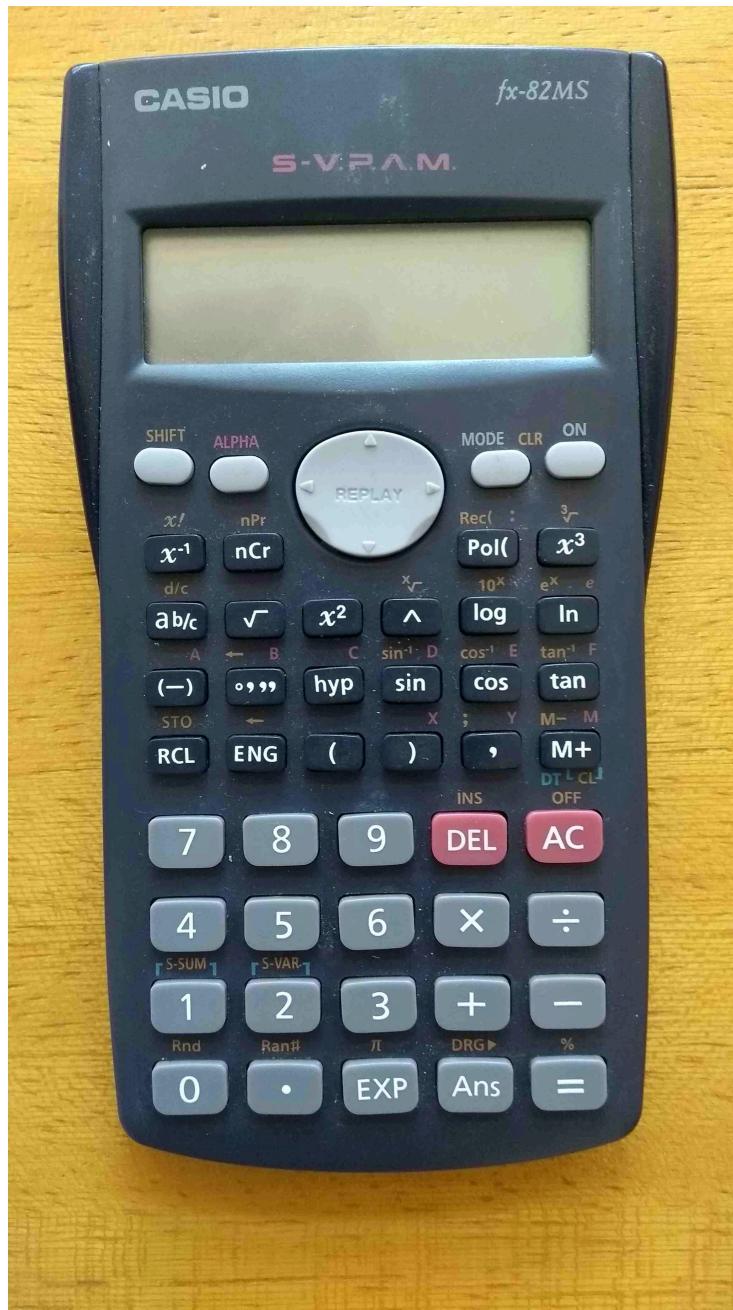


Figure 2.8: Calculadora Cassio

Chapter 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;



Figure 3.1: John Tukey (1915-2000)

- verificar os padrões de cada variável e eventuais associações entre duas ou mais delas; e,
- apresentação de tabelas e gráficos expositivos variados.

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)
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
```

3.2. DADOS BRUTOS, EM ROL, DIAGRAMA DE RAMOS & FOLHAS E DE DISPERSÃO UNIDIMENSIONAL 57

```
## [31] 1.66 1.68 1.61 1.73 1.72 1.83 1.69 1.84 1.66 1.78 1.54 1.74 1.56 1.66 1.56  
## [46] 1.62 1.55 1.86 1.44 1.67 1.76 1.79 1.75 1.41 1.65 1.58 1.93 1.57 1.71 1.58
```

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

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

```
sort(alturas)
```

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

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

- mínima; e,
- máxima dos estudantes.

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

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

```
stem(alturas)
```

```
##  
## The decimal point is 1 digit(s) to the left of the |  
##  
## 14 | 147  
## 15 | 45666788  
## 16 | 12234455556666677778889  
## 17 | 1122333345667889  
## 18 | 233456  
## 19 | 35  
## 20 | 0
```

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

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

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

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

Gráfico de dispersão unidimensional

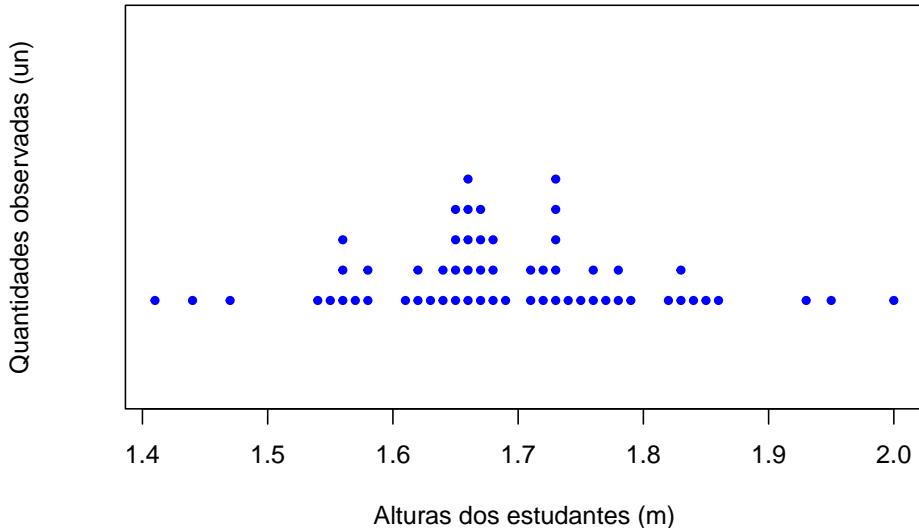


Figure 3.2: Gráfico de dispersão unidimensional (stripchart)

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 estatís-

tica 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%), decís (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:

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

Propriedades da média aritmética:

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

```
alturas_ad=alturas+0.05

par(mfrow=c(1,2))

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

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

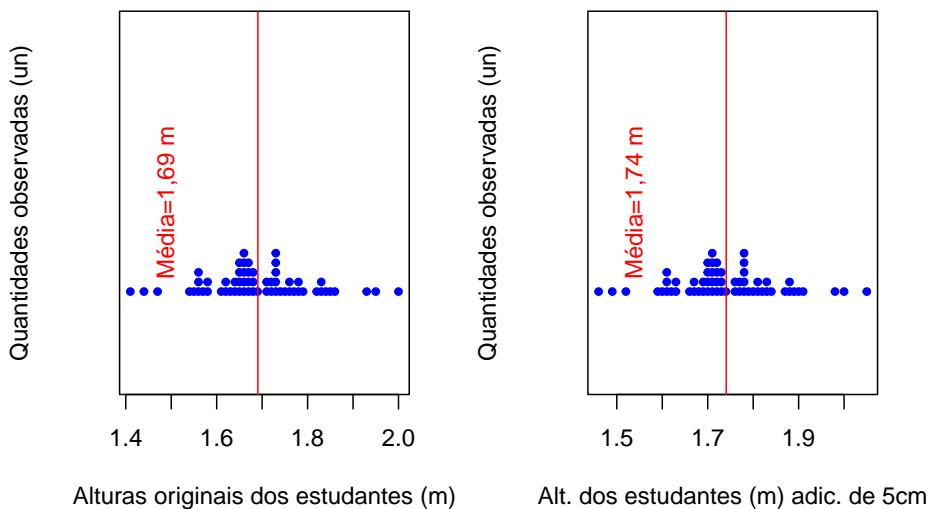


Figure 3.3: Mudanças na média pela adição (subtração) de uma constante $k = 0.05$

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

```
alturas_mult=alturas*1.2

par(mfrow=c(1,2))

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

stripchart(alturas_mult,method = "stack", at=0.5,
```

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

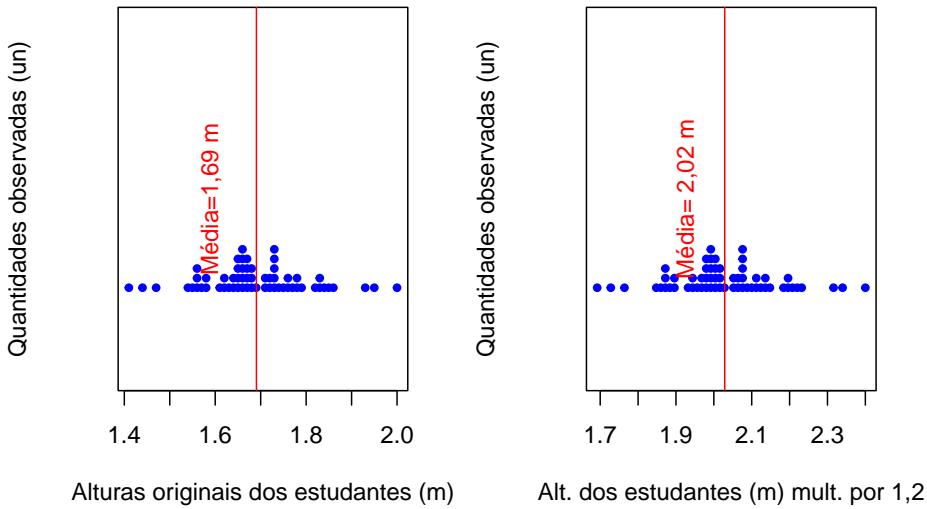


Figure 3.4: Mudanças na média pela multiplicação (divisão) de uma constante $k = 1.2$

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

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

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

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

3.3.1.2 Moda

Moda é o valor que ocorre com maior frequência na amostra. Uma amostra pode se apresentar como: - unimodal; - bimodal; - plurimodal; ou, - amodal.

```
tab_alturas=table(alturas)

tab_alturas

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

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

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

```
# 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 o valor do *i*-ésimo dado da amostra que ocupa a posição central na distribuição ordenada de modo crescente (ou decrescente), dividindo-a em duas partes de *quantidades de dados iguais*.

Valores observados da alturas dos estudantes

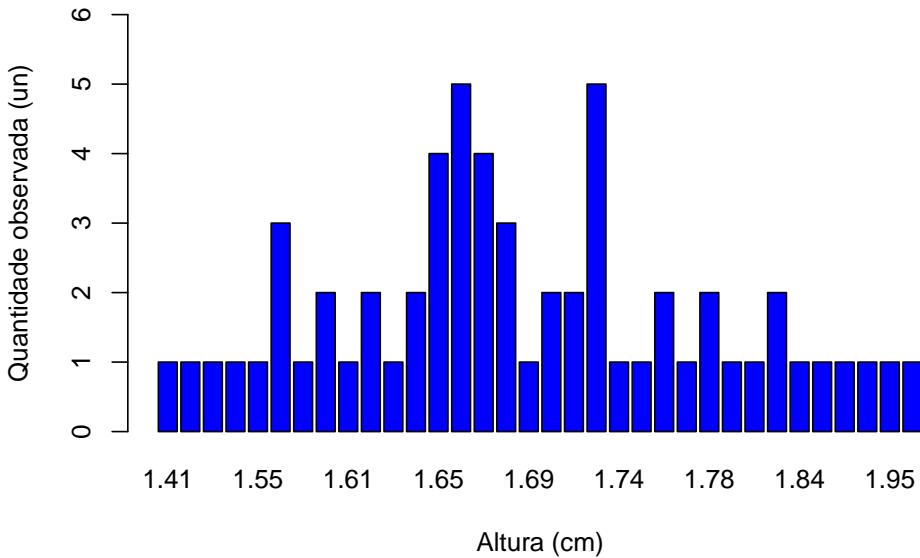


Figure 3.5: Bimodal: 1,66 m e 1,73 m

Sendo uma medida separatriz de 50%, equivale ao 2º quartil, ao 5º decil e ao 50º percentil.

1- amostra com um número **ímpar** (n) de elementos: a mediana será o valor do $\frac{n+1}{2}$ -ésimo} elemento (elemento na posição central) da amostra ordenada:

$$Md = x_i$$

onde:

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

2- amostra com um número **par** (n) de elementos: a mediana será a *média aritmética* dos elementos nas posições imediatamente anterior (i_{ant}) e posterior (i_{post}) à sua posição central virtual:

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

onde:

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

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,
- Δ_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 1.66 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
```

```
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)
```

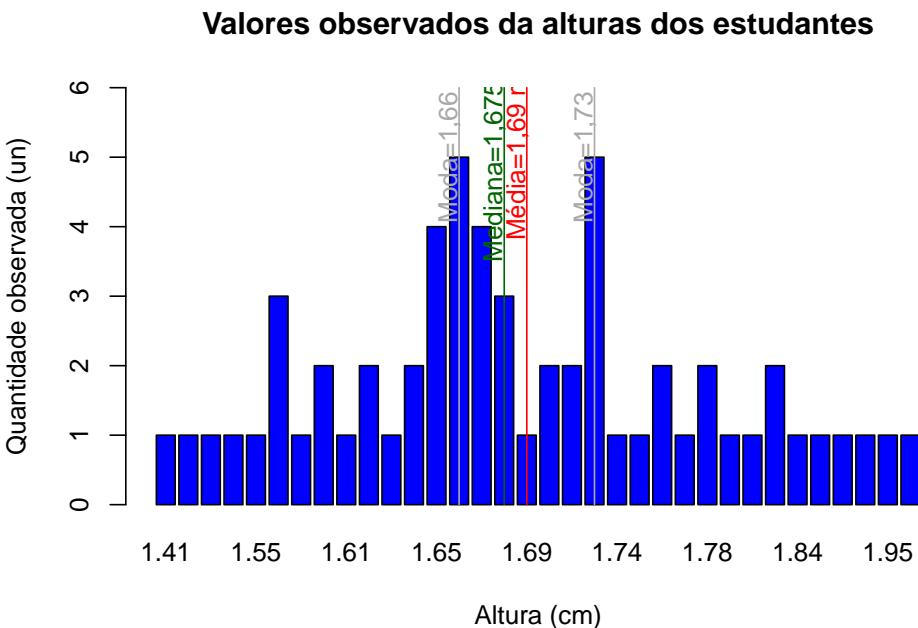


Figure 3.6: Valores observados das alturas dos estudantes e as posições da média, moda e mediana

```
h1=hist(alturas, breaks=seq(1.30 , 2.10 , 0.1), main= "Histograma das alturas dos estudantes",
       xlab="Classes de comprimento (cm)", ylab="Frequência absoluta observada (un)" , cex=0.7, ylim=c(0,6))
text(h1$midas,h1$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",
```

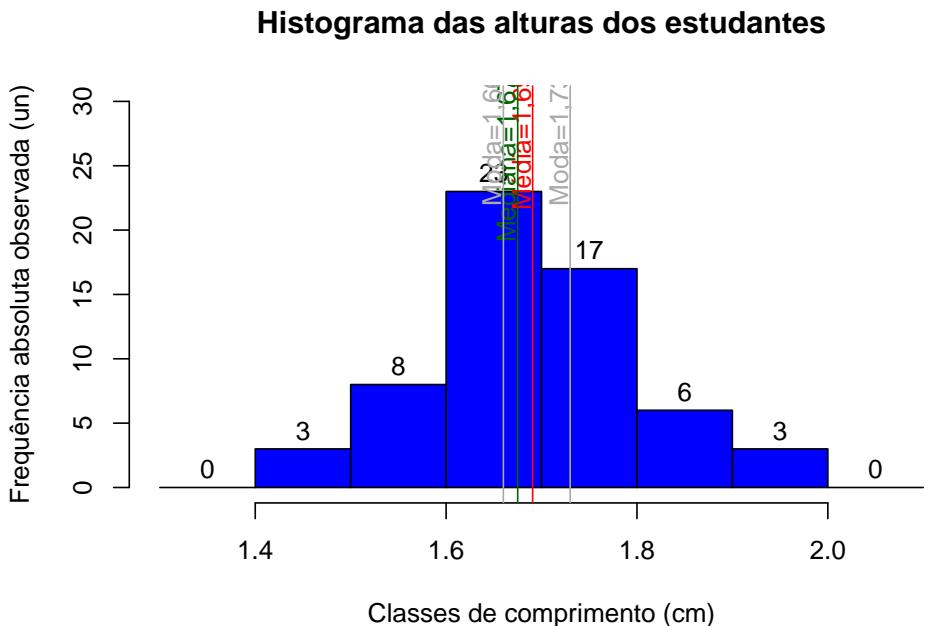


Figure 3.7: Histograma das alturas dos estudantes com as posições da média, moda e mediana

3.3.2 Medidas de dispersão (variabilidade)

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

- medidas absolutas: são expressas na mesma unidade de medida do fenômeno estudado.
 - amplitude total dos dados: diferença entre o maior e o menor dos valores observados; e,
 - variância e desvio padrão: é considerada a mais útil das medidas de dispersão.

Comparação entre medidas de posição

	Média	Mediana	Moda
Definição	$\bar{x} = \frac{\sum x}{n}$	Valor do meio	Valor mais frequente
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)

- medidas relativas: usadas para se comparar a variabilidade de duas ou mais distribuições, mesmo quando estas se refiram a diferentes fenômenos ou que sejam expressas em unidades diferentes.
 - coeficiente de variação; e,
 - unidades padronizadas.

3.3.2.1 Estimação da variância (e desvio padrão).

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

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

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

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

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

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

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

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

ou,

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

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

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

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
```

3.3.2.2 Coeficiente de variação.

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

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

3.3.2.3 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 1.66 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 1.74 1.75
## [46] 1.76 1.76 1.77 1.78 1.78 1.79 1.82 1.83 1.83 1.84 1.85 1.86 1.93 1.95 2.00
```

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

3.3.3 Medidas de subdivisão (separatrizes)

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

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

Para se calcular a **posição** L de um quantil de ordem p em um rol de dados, pode-se usar a seguinte regra:

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

Duas situações possíveis para a posição L : ser um número fracionário ou inteiro:

- se a posição **L for fracionária** deve-se fazer a média entre os dois valores que estão nas posições imediatamente anterior e imediatamente posterior à posição calculada;
- se a posição **L for um inteiro** essa será a posição do valor referente ao quantil desejado.

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.

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

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

- 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,
- $q_1 - x_1 \approx x_n - q_3$.

```

set.seed(1000)
y_normal=rnorm(1000 , 20 , 5 )
range(y_normal)

## [1] 3.191432 33.350358

quantile(y_normal)

##          0%         25%         50%         75%        100%
## 3.191432 16.685004 20.112604 23.220489 33.350358

hist(y_normal, prob=TRUE, breaks=20, ylab="", xlab="", yaxt="n", xaxt="n", main="Histograma")
curve(dnorm(x, mean(y_normal), sd(y_normal)), add=TRUE, col="darkblue", lwd=2)

abline(v=3.19, col="red")
text(3.19 -0.6, 0.02, "x(1)", col = "red", srt=90)
abline(v=16.68, col="red")
text(16.68 -0.6, 0.02, "1? Quartil", col = "red", srt=90)
abline(v=20.11, col="red")
text(20.11 - 0.6, 0.02, "2? Quartil", col = "red", srt=90)
abline(v=23.22, col="red")
text(23.22 - 0.6, 0.02, "3? Quartil", col = "red", srt=90)
abline(v=33.35, col="red")
text(33.35 -0.6, 0.02, "x(n)", col = "red", srt=90)

```

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*.

Histograma de uma variável com Distribuição Normal

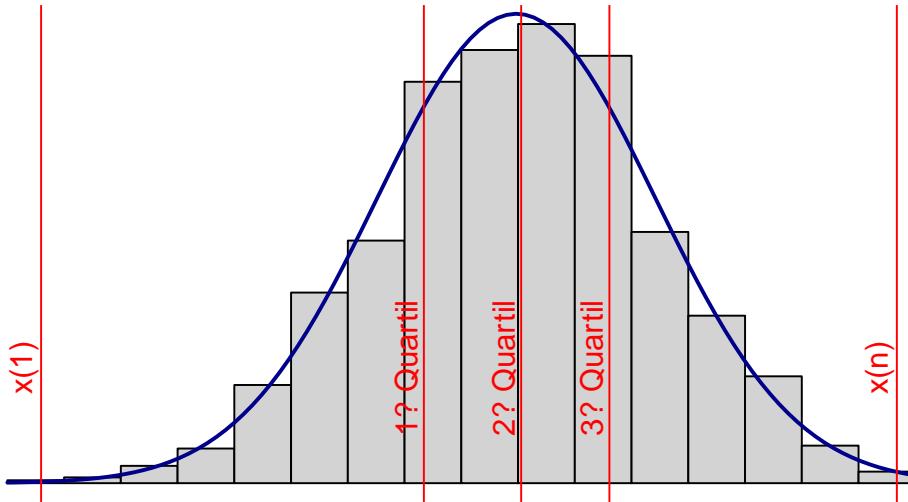


Figure 3.9: Histograma de uma variável com Distribuição Normal (média 20 e variância 5)

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

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

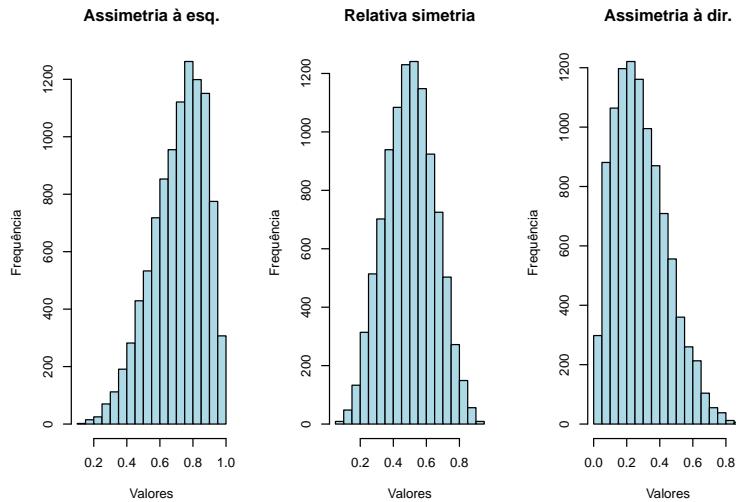


Figure 3.10: Diferentes formas na distribuição dos dados

mais plano são denominadas de *platicúrticas* e as com um cume agudo são denominadas *leptocúrticas*.

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

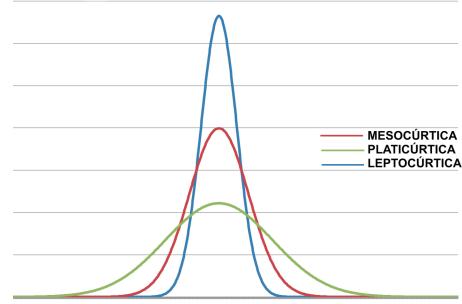


Figure 3.11: Diferentes aspectos de uma distribuição quanto à sua inclinação

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;
- $Q_1 = 1^o$ quartil;
- $P_{90} = 90^o$ percentil; e,
- $P_{10} = 10^o$ 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 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; 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*.

3.5.1 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")
```

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.

Se adotarmos arbitrariamente como classes para as alturas: 1,30-1,40; 1,40-1,50; 1,50-1,60; 1,60-1,70; 1,70-1,80; 1,80-1,90; 1,90-2,00; 2,00-2,10, o histograma terá esse aspecto:

Valores observados da alturas dos estudantes

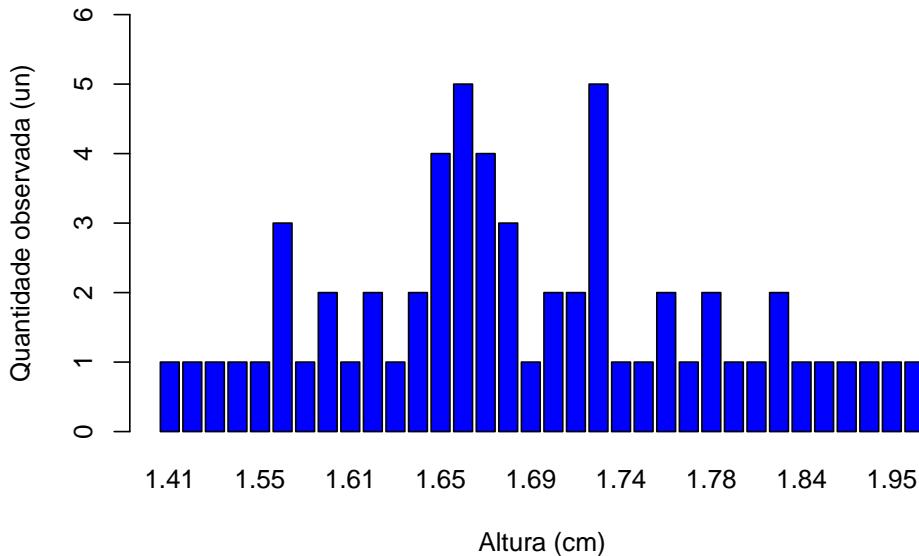


Figure 3.12: Gráfico de barras dos dados brutos: uma barra para cada observação e sua altura expressando o número de observações com esse valor

```
hist(alturas, breaks=seq(1.30 , 2.10 , 0.10), main= "Histograma das alturas dos estudantes", col="blue", xlab="Classes de altura (m)", ylab="Frequência observada (un)" , cex=0.7, ylim=c(0,30))
text(h1$midas,h1$counts,labels=h1$counts, adj=c(0.5, -0.5))
```

3.5.2 Setores

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

```
library(scales)
library(ggplot2)

alturas_classes=data.frame(
  group = c("1,40-1,50",
            "1,50-1,60",
            "1,60-1,70",
```

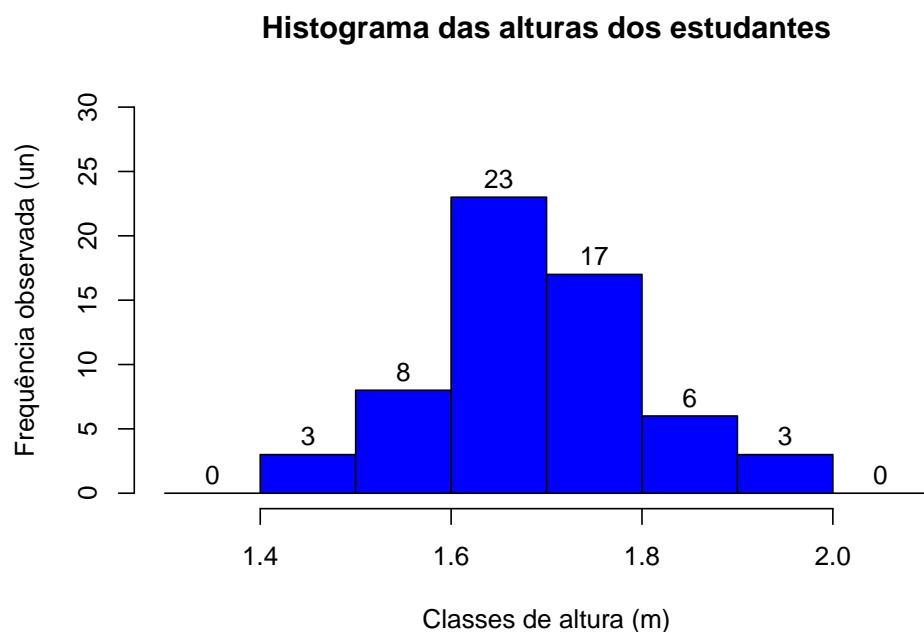


Figure 3.13: Histograma das alturas dos estudantes. Uma barra para cada classe de altura e sua altura expressando a quantidade de observações com valores dentro dessa classe (intencionalmente criamos duas classes sem nenhuma observação)

```

    "1,70-1,80",
    "1,80-1,90",
    "1,90-2,00"),
value = c(3,8,23,17,6,3)
)

bp=ggplot(alturas_classes, aes(x="", y=value, fill=group))+  

  geom_bar(width = 1, stat = "identity")
pie=bp + coord_polar("y", start=0)

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")
  )

pie +
  scale_fill_brewer("Blues")+
  blank_theme +
  theme(axis.text.x=element_blank()) +
  geom_text(aes(y = value/3 + c(0, cumsum(value)[-length(value)])),
            label = percent(value/100), size=5)+
  ggtitle("Alturas dos estudantes") +
  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 = "Classes de valores (m)",
                             label.position = "right",
                             title.position = "top", title.vjust = 1))

```

3.5.3 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: $q_3 - q_1$);

Alturas dos estudantes

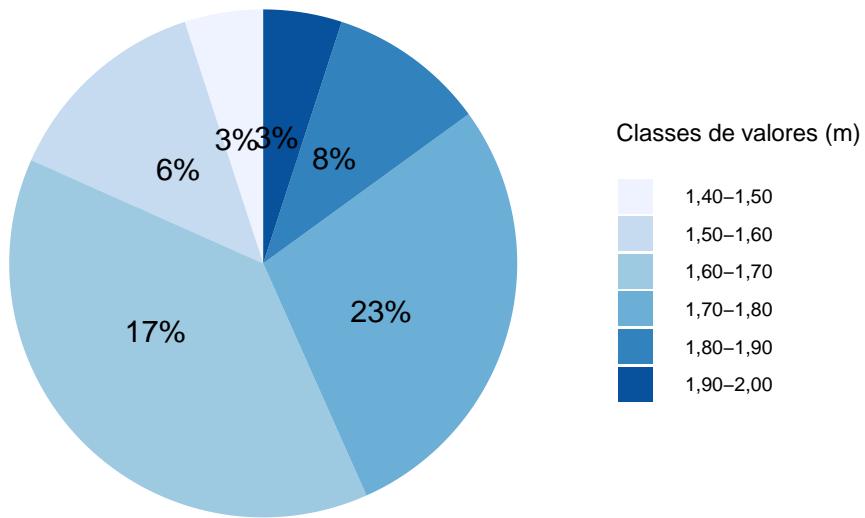


Figure 3.14: Gráfico de setores das alturas dos estudantes

- os limites superior: $LS = q_3 + 1,50d_q$, e inferior: $LI = q_1 - 1,50d_q$ (*bigodes*);
- as observações adjacentes aos limites: situadas entre o 1º quartil e o LI, e o 3º quartil e o LS; e,
- as observações exteriores aos limites: situadas abaixo do LI ou acima do LS que **podem ou não** ser *outliers* (dados atípicos).

```

res=summary(y_normal)
min=res[1]
q1=res[2]
q2=res[3]
med=res[4]
q3=res[5]
max=res[6]
dist=q3-q1

boxplot(y_normal, main="")
lines( y=c(min, min), x=c(0.6,1), col="red")
text(x=0.60, y=min -0.6, "Primeira observação", col = "red", srt=0)
lines( y=c(max,max), x=c(0.6,1), col="red")
text(x=0.60, y=max-0.6, "Última observação", col = "red", srt=0)
lines(y=c(q2, q2), x=c(0.6,1), col="red")

```

```

text(x=0.60 , y= q2 - 0.6, "Mediana", col = "red", srt=0)
lines(y=c(med, med), x=c(1,1.4), col="red")
text(x=1.4 , y= med - 0.6, "Média", col = "red", srt=0)
lines(y=c(q3, q3), x=c(1, 1.4), col="red")
text(x= 1.4 , y=q3 - 0.6, "Terceiro Quartil", col = "red", srt=0)
lines(y=c(q1, q1), x=c(1, 1.4), col="red")
text(x=1.4, y=q1 -0.6, "Primeiro Quartil", col = "red", srt=0)
lines(y=c(q1-1.5*dist, q1-1.5*dist) , x=c(1,1.4) , col="blue")
text(x=1.2, y=q1-1.5*dist-0.6 , "Obs. limitante inferior", col = "blue", srt=0)
lines(y=c(q3+1.5*dist, q3+1.5*dist) , x=c(1,1.4) , col="blue")
text(x=1.2, y=q3+1.5*dist -0.6 , "Obs. limitante superior", col = "blue", srt=0)

```

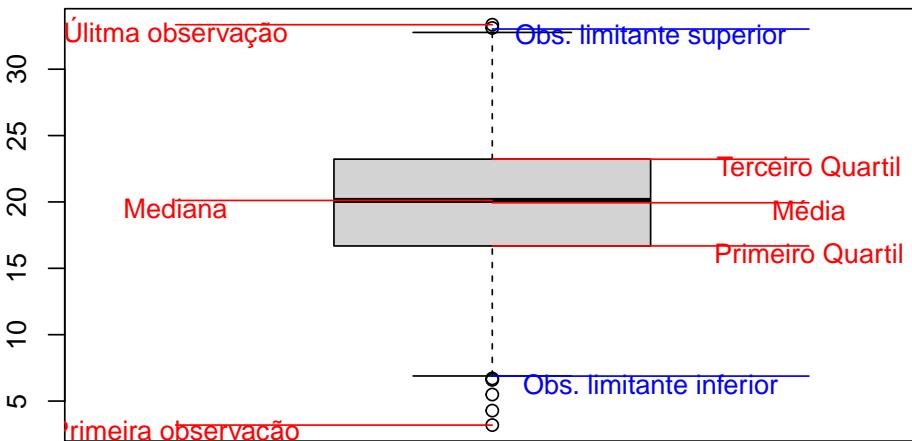


Figure 3.15: Box-plot de um rol de valores com Distribuição Normal (média 20 e variância 5)

3.6 Apresentação tabular de dados quantitativos

Ao se lidar com grandes conjuntos de dados a visualização da informação contida nos dados pode ficar comprometida. 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 ter:

- título que explique o que a tabela contém, local, data;

- cabeçalho com os nomes das variáveis;
- corpo formado pelos dados referentes às variáveis;
- fonte;
- uniformidade no número de casas decimais; e,
- todas as casas devem apresentar valores ou símbolos que expliquem a ausência da informação (NI, NE).

Trabalhos de natureza acadêmica ou científica deveriam obrigatoriamente seguir, quando publicados no Brasil, a norma vigente publicada pela ABNT: Associação Brasileira de Normas Técnicas.

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.

Com o propósito de mostrar alguns gráficos nos exemplos anteriores escolhemos, de modo arbitrário, agrupar seus valores (as alturas) 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*.

Devemos sempre ter certeza de que cada dado (medida ou observação) se enquadre em uma, e apenas uma, *classe* ou *categoria*.

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

Critério	Tamanho da amostra (<i>n</i>)	Fórmula
Raiz quadrada	$25 \leq n \leq 220$	$k = \sqrt{n}$
Herbert Sturges (<i>log</i>)	$135 \leq 572237$	$k = 1 + 3,3\log(n)$

Crítérico	Tamanho da amostra (n)	Fórmula
Giuseppe Milone (\ln)	$20 \leq 36315$	$k = -1 + 2\ln(n)$

Ao se escolher um número de classes deve-se ponderar para que:

- os intervalos das classes tenham, geralmente, a mesma amplitude;
- os intervalos: 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;
- 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.

Em nosso exemplo das alturas dos estudantes, a determinação do número de classes pelo critério da *raiz quadrada* ($n=60$) sugere 8 classes:

$$\begin{aligned} k &= \sqrt{n} \\ &= 7,74 \\ &\sim 8 \end{aligned}$$

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

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

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

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

A amplitude de cada classe é um valor fracionário que, se adotado, não irá tornar a visualização dos dados mais clara. Mesmo se adotássemos um número imediatamente maior ou menor de classes ($k=9$ ou $k=7$) esse problema persistiria.

Usando bom senso, adotaremos para como intervalo de classe o valor $c=0,10$ m e a primeira classe começando na altura de 1,40 m. O total de 6 classes (1,40 m a 2,00 m) cobre toda faixa de variação dos valores dos dados e é de rápida assimilação pelo leitor.

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***:

$\dashv \circ - \bullet$

Cada uma das classes terá os seguintes limites inferior e superior:

1,40 m \vdash 1,50 m
 1,50 m \vdash 1,60 m
 1,60 m \vdash 1,70 m
 1,70 m \vdash 1,80 m
 1,80 m \vdash 1,90 m
 1,90 m \vdash 2,00 m

Veja os dados em rol, onde a cor azul indica a mudança de classe com o progredir dos valores das alturas:

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 ; 1,71 ; 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 versão mais simplificada de uma *tabela de distribuição de frequências* (o número de observações nas classes) é de fácil construção, bastando contar o número de observações em cada classe:

Classe	Frequência (n_i)
1,40 m ⊂ 1,50 m	3
1,50 m ⊂ 1,60 m	8
1,60 m ⊂ 1,70 m	23
1,70 m ⊂ 1,80 m	17
1,80 m ⊂ 1,90 m	6
1,90 m ⊂ 2,00 m	3
Total	60

Tabelas de distribuição de frequências mais completas podem montadas agregando muitas informações adicionais em novas colunas.

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

- Classe i : é a simples identificação de cada classe;
- Amplitude (Δ_i) da classe i : a diferença entre o valor do limite superior e o do inferior de cada classe;
- Intervalo de valores da classe i (onde seu limite inferior **está contido** e o limite superior **não está contido**);
- Valor médio (\bar{x}_i) de cada classe i : a média aritmética entre os valores dos limites inferior e superior da classe considerada;
- 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);

- Frequênci a acumulada (fac_i) da classe i (ou frequênci a acumulada percentual, se assim apresentada): o númer o de observações com medidas contidas na classe i e nas anteriores a ela;
- Densidade ($\delta_i = \frac{f_i}{\Delta_i}$): o quociente do númer o de observações da classe (f_i) pela sua amplitude (Δ_i);
- Densidade $\delta_{fr_i} = \frac{fr_i}{\Delta_i}$: o quociente da frequênci a relativa (fr_i) pela amplitude (Δ_i) da classe.

Vejo como exemplo as tabelas abaixo:

	Int. de Classe	valores	Alt. média	Freq.	Freq. relativa	Freq. rel. (%)	Freq. acumu-lada	Freq. acum. (%)
			(\bar{x}_i)	(f_i)	(fr_i)	$(fr_i\%)$	(fac_i)	$(fac_i\%)$
1	1,40	1,50	1,45	3	0,05	5	3	5
2	1,50	1,60	1,55	8	0,13	13,33	11	18,33
3	1,60	1,70	1,65	23	0,38	38,34	34	56,57
4	1,70	1,80	1,75	17	0,28	28,33	51	84,87
5	1,80	1,90	1,85	6	0,10	10	57	94,57
6	1,90	2,00	1,95	3	0,05	5	60	99,87
Totais-				60	1,00	100,00	-	-

	Int. de Classe	valores	Freq.	Amplitude	Densidade	Freq. rel.	Dens. da freq. rel.
			(f_i)	(Δ_i)	(δ_i)	(fr_i)	(δ_{fr_i})
1	1,40	1,50	3	0,10	30	0,05	0,5
2	1,50	1,60	8	0,10	80	0,13	1,33
3	1,60	1,70	23	0,10	230	0,39	3,83
4	1,70	1,80	17	0,10	170	0,28	2,83

Classe	Int. de valores	Freq.	Amplitude	Densidade	Freq. rel.	Dens. da freq. rel.
5	1,80 ← 1,90	6	0,10	60	0,10	1
6	1,90 ← 2,00	3	0,10	30	0,05	0,5
Totais	-	60	-	-	1,00	-

3.6.1 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:

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

3.6.2 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**;
- Δ_1 frequência absoluta da **classe modal** menos a frequência absoluta da **classe anterior**;
- Δ_2 frequência absoluta da **classe modal** menos a frequência absoluta da **classe posterior**; e,
- Δ_i é o intervalo de cada classe.

3.6.3 Variância

Variância para dados agrupados:

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

Onde:

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

3.6.4 Histograma

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)).

```

h1=hist(alturas, breaks=seq(1.30 , 2.10 , 0.10), main= "Histograma das alturas dos estudantes",
       xlab="Classes de altura (m)", ylab="Frequência observada (un)" , cex=0.7, ylim=c(0,30))
text(h1$midas,h1$counts,labels=h1$counts, adj=c(0.5, -0.5))

```

Histograma das alturas dos estudantes

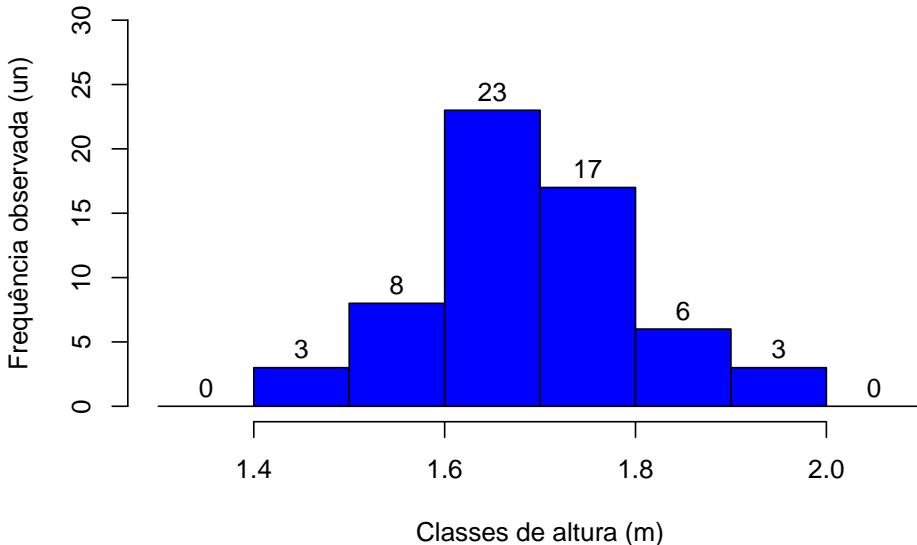


Figure 3.16: Histograma das alturas dos estudantes. Uma barra para cada classe de altura e sua altura expressando a quantidade de observações com valores dentro dessa classe (intencionalmente criamos duas classes sem nenhuma observação)

As informações da *Tabela de distribuição de frequências* dão origem a variados tipos de histogramas, como aquele feito com as frequências relativas:

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

```

h2=hist(alturas,breaks=seq(1.30 , 2.10 , 0.10), main= "Histograma das alturas dos estudantes",
       xlab="Classes de alturas (m)", ylab="Frequência relativa observada", 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), lty="dotted")

```

Histograma das alturas dos estudantes

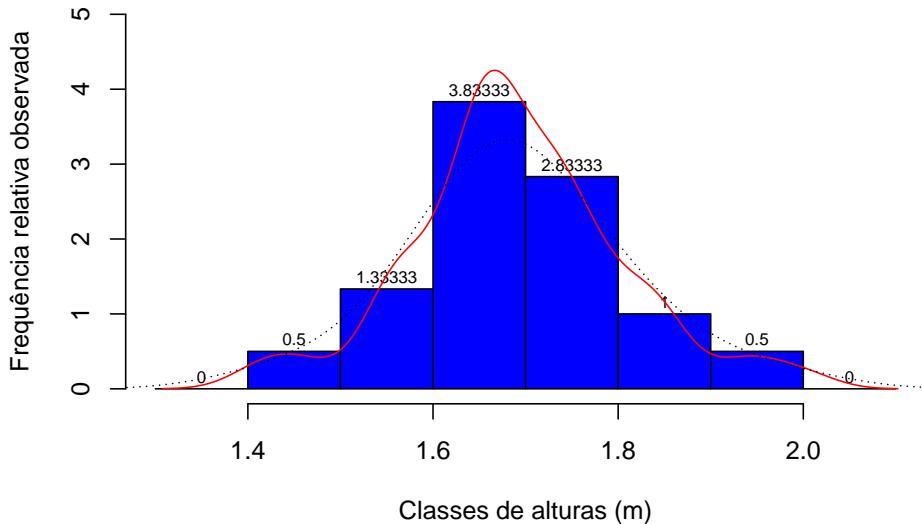


Figure 3.17: A linha vermelha é uma aproximação da Função de Densidade da frequência relativa de observação (a linhe preta é a curva da função densidade de uma distribuição Normal com média e variâncias dadas pelos dados

Uma aproximação para a **área sob a curva da Função de Densidade** pode ser soma das áreas de cada retângulo, onde cada um deles tem:

- Base = Δ_i ; e,\
- Altura = Densidade da proporção= $\frac{f_i}{\Delta_i}$.

Portanto, a área de cada retângulo é igual à proporção (f_i) da classe (i) e, assim, 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
```

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%)

3.7 Apresentação tabular de dados qualitativos

Frequentemente pesquisas são conduzidas tendo 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 proporções que expressam a frequência absoluta com que cada uma dessas variáveis (ou seus níveis) foi observada em relação ao total estudado.

Para essas situações, variados tipos de apresentações tabulares podem ser produzidos como a tabela abaixo, 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.

Tipo de família	Número (milhões)	Freq. abs.	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

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

```

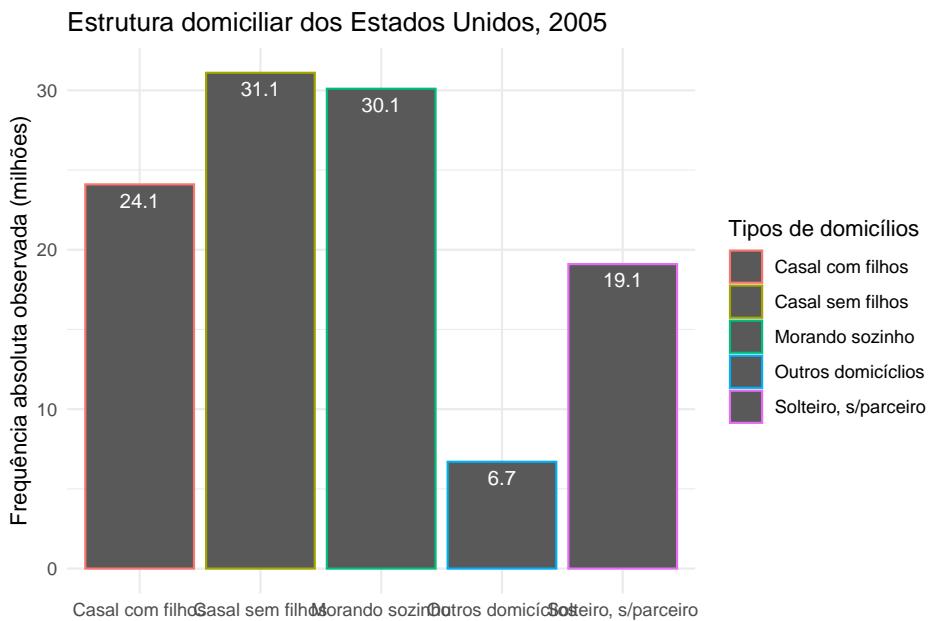


Figure 3.18: Gráfico de barras

```

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))

```

Outro tipo de apresentação tabular de dados qualitativos são as *Tabelas de Contingência*.

As tabelas de contingência são usadas para associar duas ou mais variáveis qualitativas (ou seus níveis) às 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.

Estrutura domiciliar dos Estados Unidos, 2005

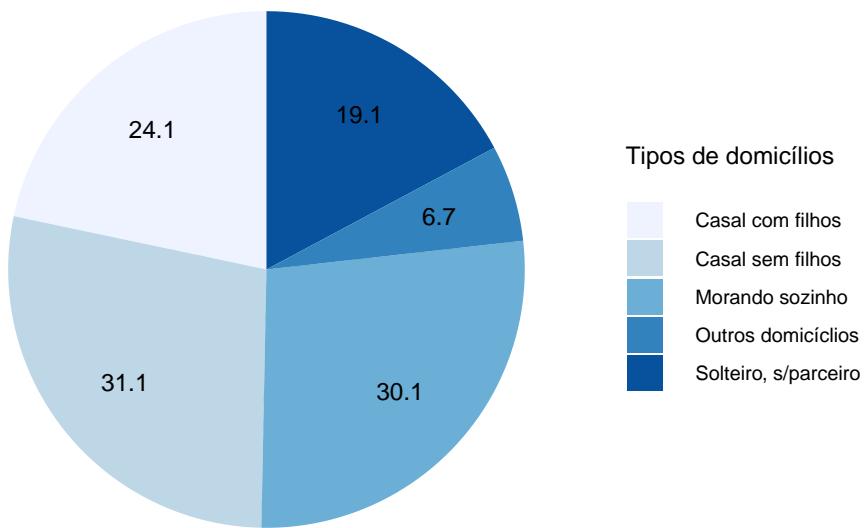


Figure 3.19: Gráfico de setores

Inclinação partidária

	Democrata	Republicano
Casal com filhos	762	468
Casal sem filhos	484	477
Total	1246	945

Inclinação partidária

	Democrata (milhões)	Republicano (milhões)	Total (milhões)
Casal com filhos	18,0	6,1	24,1
Casal sem filhos	29,1	2,0	31,1
Total	47,1	8,1	55,2

As representações gráficas são análogas às mostradas no exemplo anterior.

Chapter 4

- Introdução ao cálculo de probabilidade

“Not everything that can be counted counts and not everything that counts can be counted [...]” (Albert Einstein, 1879-1955)

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 Ω .

Para que Ω 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

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

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 Ω é chamado de um **elemento** desse espaço amostral e é denotado pela letra grega: ω_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

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

,

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 simultaneamente o espaço amostral Ω do experimento aleatório e o **evento de sucesso* E_1 de modo gráfico.

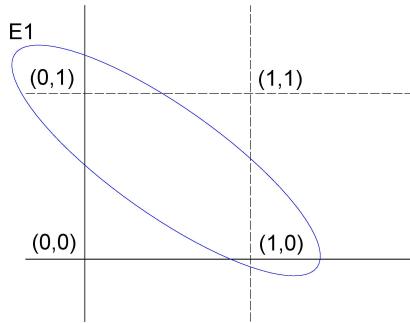


Figure 4.1: Representação gráfica do espaço amostral do experimento aleatório e do evento de interesse definido

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 abaixo 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	$\frac{1}{36}$

	Possíveis combinações de resultados nos lançamentos	Frequência (n_i)	Proporção (f_i)
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 elemertos 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* Ω .

Um evento de interesse G sobre o espaço amostral Ω tal que

$$G = \Omega$$

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

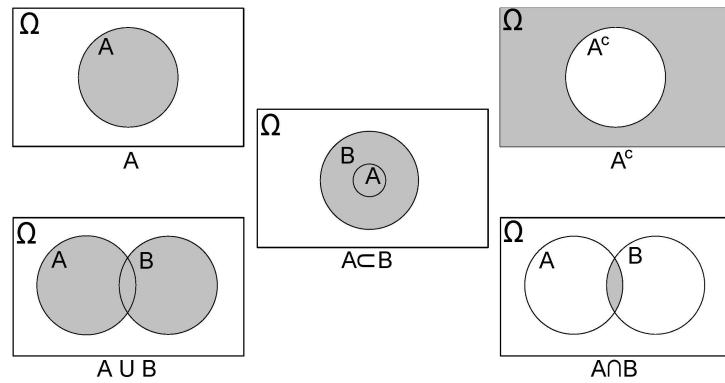
Se definirmos em evento de interesse I com um resultado não pertencente aos possíveis resutados representados no *espaço amostral* Ω , como, por exemplo, 13, ou então um conjunto vazio \emptyset , esse evento será impossível de ocorrer. Esse tipo de evento é chamado de **evento impossível**.

Desse modo temos diferentes tipos de eventos de inetersse:

- 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 & 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.



DIAGRAMAS DE VENN

Figure 4.2: Diagramas de Venn

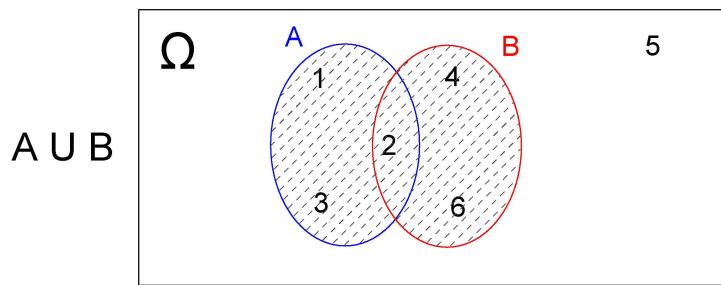
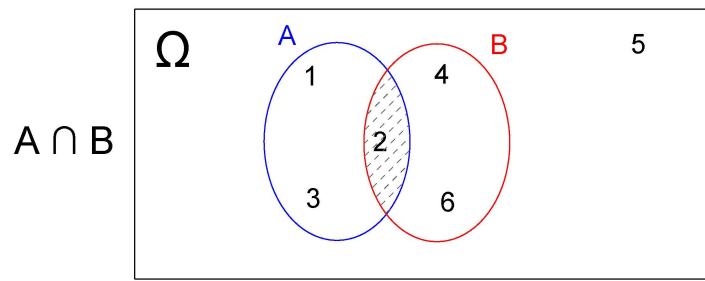
4.1.5.1 União $A \cup B$

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 $B = \{2, 4, 6\}$.

Sua *união*, representada por $A \cup B$, é o subconjunto do espaço amostral Ω 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:

4.1.5.2 Interseção $A \cap B$

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

Figure 4.3: União: $A \cup B$ 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 Ω composto por todos os elementos que **não pertencem** a A . Desse modo, $\bar{A} = \{4, 5, 6\}$ e o Diagrama de Venn correspondente será:

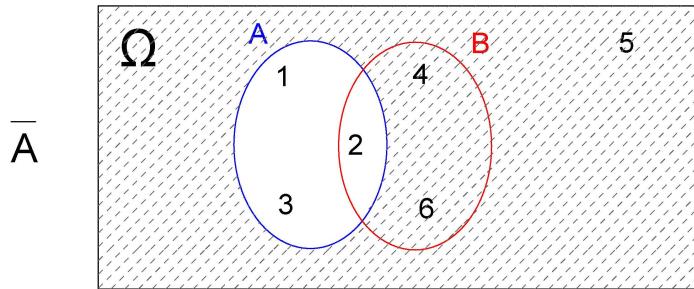


Figure 4.5: Complementar A^c

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

4.1.6 Eventos equiprováveis e não equiprováveis

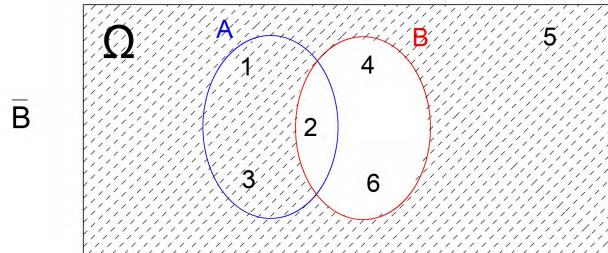


Figure 4.6: Complementar de B

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 calcâncar, 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 desses resultados poderá ser apresentada 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 ; 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.

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

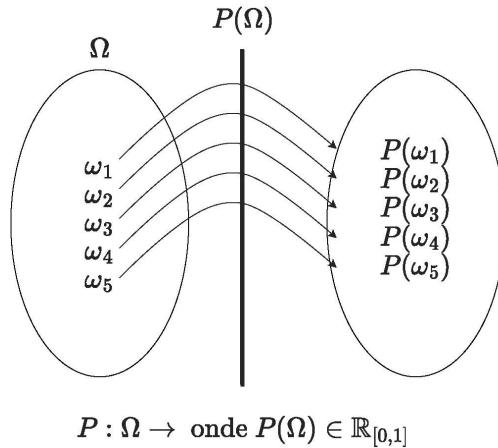


Figure 4.8: Representação gráfica da função $P(\Omega)$

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.

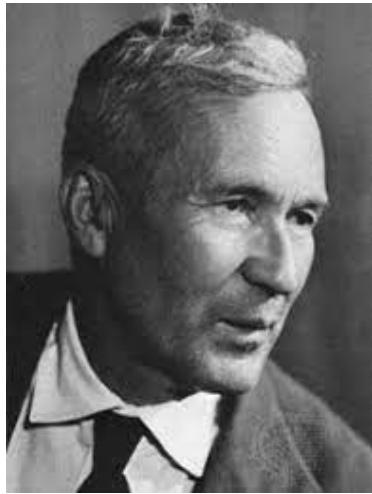


Figure 4.9: Andrey Nikolaevich Kolmogorov (1903-1987)

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 entã 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.

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

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 $\{i\} \cap \{j\} = \emptyset$ se $i \neq j$)

Tomando o terceiro postulado no caso mais simples, isto é, para **dois** eventos mutuamente exclusivos ω_1 e ω_2 , pode ser facilmente visto que é 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 ω e $P(\Omega)$ atende aos três axiomas anteriormente estabelecidos, diz-se que ela é a **função de probabilidade** de Ω .

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)	Total
Matemática pura (M)	70	40	110
Matemática aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Total	115	85	200

Essa tabela nos possibilita calcular a probabilidade de ocorrência de diversos eventos de interesse que desejemos estabelecer.

Exemplo: seja o experimento aleatório de se escolher, aleatoriamente, um estudante qualquer desses quatro cursos. Assim, se definimos nosso evento de interesse M como sendo **M:sexo masculino**, a probabilidade de sucesso (que o indivíduo sorteado aleatoriamente seja do sexo masculino) será:

$$P(M) = \frac{115}{200}$$

Exemplo: se nosso evento de interesse A como sendo $A : \text{curso de matemática aplicada}$, 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**:

$$1. P(A) = \frac{30}{200} \text{ (curso: matemática aplicada); e, } 2- P(M) = \frac{115}{200} \text{ (sexo masc).}$$

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**?

116 CHAPTER 4. - INTRODUÇÃO AO CÁLCULO DE PROBABILIDADES

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:

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

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:

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

Exemplo: Consideremos a Tabela 4.3 que apresenta o cruzamento do sexo dos alunos pelos seus respectivos cursos. Vamos definir os eventos *Fem* : **sexo feminino** e *Est* : **cursar estatística**. Como calcular a probabilidade condicionada de nosso evento de interesse $P(Fem|Est)$ (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:

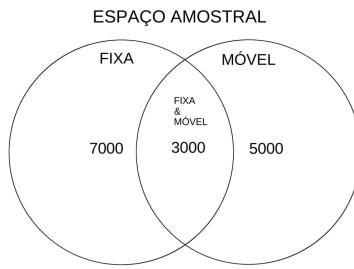


Figure 4.10: Diagrama de Venn do espaço amostral

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

- $P(F|M)$ (probabilidade de ter uma linha fixa sabendo que possui um telefone móvel); e,
- $P(M|F)$ (probabilidade de ter uma linha móvel sabendo que possui um telefone fixo);

$$\begin{aligned}
 P(F|M) &= \frac{n(MF)}{n(M)} \\
 &= \frac{3000}{8000} \\
 &= 0,375
 \end{aligned}$$

e

$$\begin{aligned} P(M|F) &= \frac{n(MF)}{n(F)} \\ &= \frac{3000}{10000} \\ &= 0,300 \end{aligned}$$

Mas também podemos calcular as probabilidades do modo como explicado no começo desta sessão. Definindo-se os eventos F : **telefone fixo** e M : **telefone móvel**, a primeira pergunta pede $P(F|M)$: probabilidade de ter um telefone fixo sabendo que ele tem um telefone móvel:

$$\begin{aligned} P(F|M) &= \frac{P(F \cap M)}{P(M)} \\ &= \frac{\frac{3000}{15000}}{\frac{8000}{15000}} \\ &= 0,375. \end{aligned}$$

A segunda pede $P(M|F)$: probabilidade de ter um telefone móvel sabendo que ele tem um telefone fixo:

$$\begin{aligned} P(M|F) &= \frac{P(M \cap F)}{P(F)} \\ &= \frac{\frac{3000}{15000}}{\frac{10000}{15000}} \\ &= 0,300 \end{aligned}$$

Exemplo: Considere a Tabela 4.5 onde são expostos os resultados de uma pesquisa relacionada ao gosto pela prática de tênis entre alunos e alunas. Definindo-se os eventos A : “gostar de tênis” e B : “ser do sexo feminino”, calcule as 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		
	Masculino (M)	Feminino (F)	Total
Gostam de tênis (T)	400	200	600
Não gostam de tênis (NT)	50	50	100
Total	450	250	700

4.2.7 Dependência & 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:

$$\begin{aligned} P(A \cap B) &= P(A|B) \times P(B) \\ P(A \cap B) &= P(B|A) \times P(A) \end{aligned}$$

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

Se os eventos A e B são guardam nenhuma relação de condicionamento eles são chamadas de **eventos independentes**. Equivale dizer que $P(A|B) = P(A)$ (ou $P(B|A) = P(B)$), a probabilidade de A não se altera pela prévia ocorrência de B (ou a de B pelo de A).

Portanto, **dois eventos são denominados independentes se, e somente se:**

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

Independência e correlação: se duas variáveis aleatórias são **independentes** não há associação de natureza alguma entre elas, **inclusive a linear**, um caso particular de correlação. Todavia uma **correlação linear nula** não implica em **independência** posto existirem várias outras formas outras de relacionamento (quadrática, cúbica, ...).



Figure 4.11: 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 verificada. 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 $E_1 = \text{sair face 1}$ e $E_5 = \text{sair face 5}$).

Solução:

Quando lançamos dois dados o resultado obtido em um deles (o valor numérico da face) **não condiciona ou altera** o resultado obtido no outro: os resultados são **são independentes**. Desse modo, sendo $P(E_1) = \frac{1}{6}$ e $P(E_5) = \frac{1}{6}$:

$$\begin{aligned} P(E_1 \cap E_5) &= \frac{1}{6} \times \frac{1}{6} \\ &= \frac{1}{36}. \end{aligned}$$

Exemplo: Uma empresa que compra produtos de dois fabricantes diferentes (**Fabricante 1** e **Fabricante 2**) adquiriu 168 unidades do primeiro e 84 do segundo. Sabendo que 8 unidades fabricadas pelo primeiro fornecedor não atenderam às especificações e apenas 4 do segundo, verifique se o fato de uma amostra ter atendido às especificações independe de ter sido produzida pelo **Fabricante 1**.

Solução:

Para a primeira verificação pedida defina os eventos $Fab1$: ter sido produzida pelo Fabricante 1, $Aprov$: ter atendido às especificações e $Fab2$: ter sido produzida pelo Fabricante 2. Na sequência podemos calcular as seguintes probabilidades:

$$\begin{aligned} P(Fab1) &= \frac{168}{252} \\ &= 0,6666 \\ P(Aprov) &= \frac{240}{252} \\ &= 0,9523 \\ P(Fab1 \cap Aprov) &= \frac{160}{252} \\ &= 0,6349 \end{aligned}$$

Se o fato de uma amostra ter sido aprovada **independe** de ter sido produzida pelo Fabricante 1 **então** $P(Aprov|Fab1) = P(Aprov)$:

$$\begin{aligned} P(Aprov|Fab1) &= \frac{P(Aprov \cap Fab1)}{P(Fab1)} \\ &= \frac{0,6349}{0,6666} \\ &= 0,9523. \end{aligned}$$

Como $P(Aprov|Fab1) = P(Aprov)$, verifica-se que o fato de uma amostra aleatoriamente sorteada entre as peças do fabricante 1 não condiciona sua aprovação.

Exemplo: A probabilidade de um consumidor (C_1) ficar satisfeito com o desempenho de certa marca de produto é de 25%. A probabilidade de um outro consumidor (C_2) ficar satisfeito com a mesma marca é de 40%. Admitamos que os dois consumidores irão consumir o produto num mesmo momento e de forma **independente (incomunicáveis)**. Qual a probabilidade de os **dois** consumidores ficarem satisfeitos simultaneamente?

Solução:

As probabilidades individuais dos consumidores 1 e 2 ficarem satisfeitos com o desempenho da marca do produto são:

$$P(C_1) = 0,25$$

$$P(C_2) = 0,40$$

A probabilidade de **ambos** ficarem satisfeitos, dado que o enunciado afirma que esses eventos são **independente** será:

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

4.3 Teorema de Bayes

Pela **regra da probabilidade condicionada** temos que

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

e, de modo equivalente,



Figure 4.12: Thomas Bayes (1702 - 1761)

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

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 Ω é o espaço amostral de um experimento aleatório qualquer:

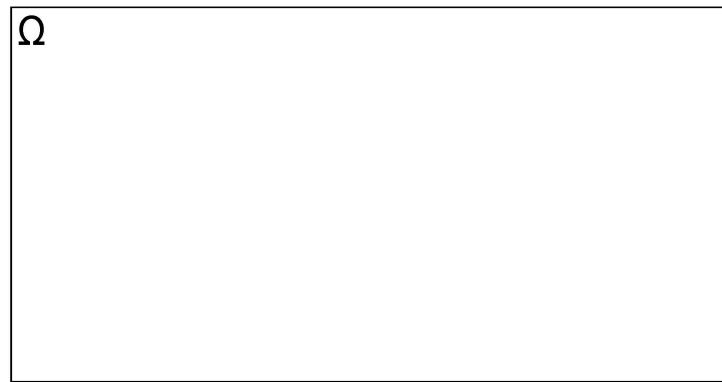


Figure 4.13: Espaço amostral

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

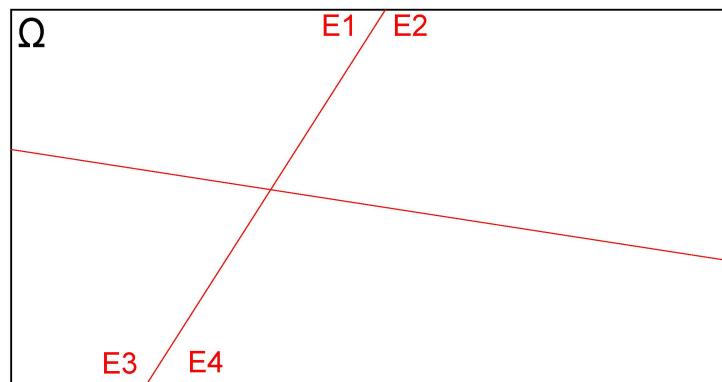


Figure 4.14: Espaço amostral e suas partições

E seja B um evento qualquer em Ω como ilustrado na Figura 4.15

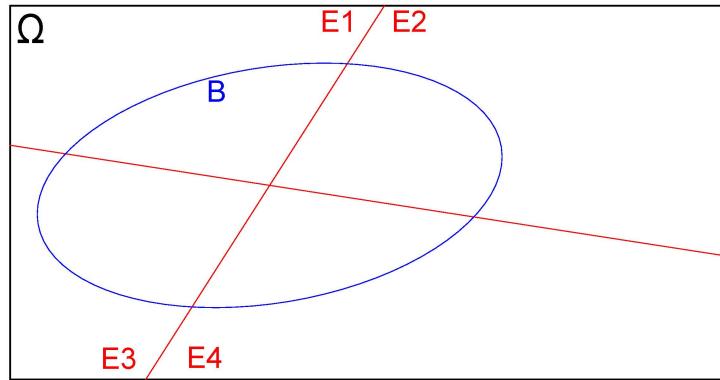


Figure 4.15: Evento definido sobre o espaço amostral

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

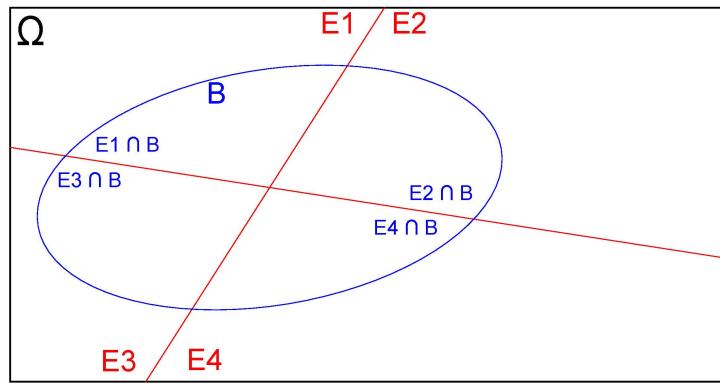


Figure 4.16: Interseções das partições do espaço amostral com o evento B

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

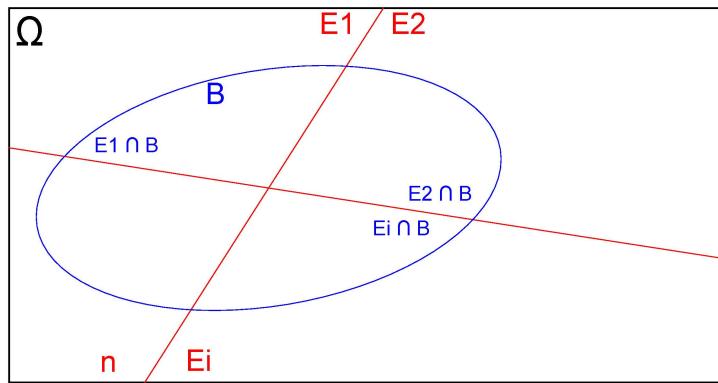


Figure 4.17: Interseções das n partições do espaço amostral com o evento B

Na representação esquemática da Figura 4.17 podemos identificar:

- \begin{itemize}
 - 1- $E_1, E_2, \dots, E_i, \dots, E_n$ constituem-se em partições do espaço amostral Ω ;
- 2- Todas as partições são mutuamente exclusivas: $E_i \cap E_j = \emptyset, \forall i \neq j$ (a interseção de quaisquer partições é vazia);
- 3- Sendo vazias as interseções entre quaisquer partições, o espaço amostral Ω será a simples união de todas elas: $\Omega = E_1 \cup E_2 \cup E_3 \cup E_4 \cup \dots \cup E_i \dots \cup E_n$; e,
- 4- B é um evento qualquer definido sobre as partições de Ω

São conhecidas as probabilidades de ocorrência de cada um dos elementos do espaço amostral Ω :

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

130 CHAPTER 4. - INTRODUÇÃO AO CÁLCULO DE PROBABILIDADES

e também as probabilidades do evento B condicionadas a cada elemento do espaço amostral:

$$P(B|E_1); P(B|E_2); \dots; P(B|E_i); \dots; P(B|E_n)$$

A *probabilidade de ocorrência* do evento B é dada pela soma das probabilidades de cada uma de suas interseções com os elementos do espaço amostral Ω :

$$\begin{aligned} P(B) &= P(E_1 \cap B) + P(E_2 \cap B) + \dots + P(E_i \cap B) + \dots + P(E_n \cap B) \\ P(B) &= \sum_{i=1}^n P(E_i \cap B) \end{aligned}$$

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

$$\begin{aligned} P(B|E_i) &= \frac{P(E_i \cap B)}{P(E_i)} \\ P(E_i \cap B) &= P(E_i) \times P(B|E_i) \end{aligned}$$

com $P(E) > 0$

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

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

Portanto a **probabilidade total** do evento B em Ω é dada pelo somatório:

$$P(B) = \sum_{i=1}^n [P(E_i) \cdot P(B|E_i)]$$

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

$$\begin{aligned}
 P(E_i|B) &= \frac{P(E_i \cap B)}{P(B)} \\
 P(E_i \cap B) &= P(B) \times P(E_i|B) \\
 P(B) &= \frac{P(E_i \cap B)}{P(E_i|B)}
 \end{aligned}$$

com $P(B) > 0$

Pela **igualdade** dos dois modos de se expressar a probabilidade total do evento B desenvolvidos:

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

e

$$P(B) = \sum_{i=1}^n [P(E_i) \cdot P(B|E_i)]$$

tem-se

$$\frac{P(E_i \cap B)}{P(E_i|B)} = \sum_{i=1}^n [P(E_i) \cdot P(B|E_i)]$$

Rearranjando-se em termos da expressão anterior para exprimir a probabilidade de ocorrência de um evento E_i posto ter ocorrido o evento B chegamos a:

$$P(E_i|B) = \frac{P(E_i \cap B)}{\sum_{i=1}^n [P(E_i) \cdot P(B|E_i)]}$$

Sendo

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

a expressão anterior pode ser reescrita como:

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

uma forma mais geral do **Teorema de Bayes**.

O Teorema de Bayes é também chamado de *Teorema da probabilidade a posteriori* ao nos permitir calcular $P(E_i|B)$ em termos da ocorrência $P(B|E_i)$

É, de certo modo, uma conjugação do *teorema na probabilidade total* e da *regra do produto* de probabilidades.

O denominador:

$$P(B) = \sum_{i=1}^n [P(E_i) \times P(B|E_i)]$$

é a denominada **probabilidade marginal** de ocorrência do evento B no espaço amostral Ω composto por n elementos (partições).

Na expressão do Teorema de Bayes:

- $P(E_k|B)$ é a denominada probabilidade *a posteriori* do evento E_k condicionada pela ocorrência anterior do evento B ;
- $P(E_k)$ é a denominada probabilidade *a priori* do evento E_k ;

- $P(B|E_k)$ é a denominada probabilidade *a posteriori* do evento B condicionada pela ocorrência anterior do evento E_k ;
- $P(E_i)$ é a denominada probabilidade *a priori* de cada evento E_i ;
- $P(B|E_i)$ é a denominada probabilidade *a posteriori* do evento B condicionada pela ocorrência anterior de cada evento E_i .

Exemplo: Constatou-se que o aumento nas vendas de um certo produto comercializado por uma empresa num mês pode ocorrer **sómente** 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* $\rightarrow P(E_1) = 0,40$;
- Ação de propaganda $\rightarrow P(E_2) = 0,30$ \$;
- Flutuações na economia $\rightarrow P(E_3) = 0,20$; ou,
- Sazonalidade $\rightarrow P(E_4) = 0,10$.

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:

136 CHAPTER 4. - INTRODUÇÃO AO CÁLCULO DE PROBABILIDADES

“Qual a probabilidade de ter ocorrido cada um dos quatro eventos (E_1, E_2, E_3, E_4) **posto** (dado) ter ocorrido}** um incremento nas vendas”?

Calculemos para cada um deles usando o Teorema de Bayes:

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

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

$$\begin{aligned} P(E_1|B) &= \frac{P(E_1) \times P(B|E1)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]} \\ P(E_1|B) &= \frac{0,40 \times 0,07}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)} \\ P(E_1|B) &= 0,478 \end{aligned}$$

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

$$\begin{aligned} P(E_2|B) &= \frac{P(E_2) \times P(B|E2)}{\sum_{i=1}^4 [P(E_i) \times P(B|E_i)]} \\ P(E_2|B) &= \frac{0,30 \times 0,075}{(0,40 \times 0,07) + (0,30 \times 0,075) + (0,20 \times 0,03) + (0,10 \times 0,02)} \\ P(E_2|B) &= 0,385 \end{aligned}$$

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 (urnas tipo C_3) possui 6 bolas brancas em seu interior (cf. Figura 4.18).

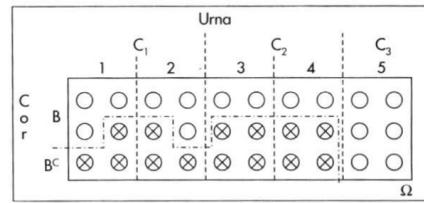


Figure 4.18: Cinco urnas cada uma com 6 bolas em cores de diferentes quantidades da cor branca

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:

$$P(C_1) = \frac{2}{5}$$

$$P(C_2) = \frac{2}{5}$$

$$P(C_3) = \frac{1}{5}$$

$$P(Branca|C_1) = \frac{1}{2}$$

$$P(Branca|C_2) = \frac{1}{3}$$

$$P(Branca|C_3) = 1$$

$$P(C_3|Branca) = \frac{P(C_3) \times P(Branca|C_3)}{\sum_{i=1}^3 [P(C_i) \times P(Branca|C_i)]}$$

$$P(C_3|Branca) = \frac{0,20 \times 1,00}{(0,40 \times 0,50) + (0,40 \times 0,33) + (0,20 \times 1,00)}$$

$$P(C_3|Branca) = 0,375$$

4.4 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:

140 CHAPTER 4. - INTRODUÇÃO AO CÁLCULO DE PROBABILIDADES

Há 10 bolas das quais 5 são **vermelhas**. A probabilidade de se retirar uma bola **vermelha** será:

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

- Probabilidade em se retirar uma bola **azul** em segundo lugar:

O enunciado do experimento assegura que após a retirada da primeira bola ela é **devolvida** ao sacola; por essa razão, ao se retirar a segunda bola, há novamente 10 bolas no total, das quais 5 são **azuis**. A probabilidade de se retirar uma bola **azul** será:

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

- 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*:

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

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

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

$$P(1^a \text{ vermelha}, 2^a \text{ azul}) = \frac{25}{100} = \frac{1}{4}$$

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

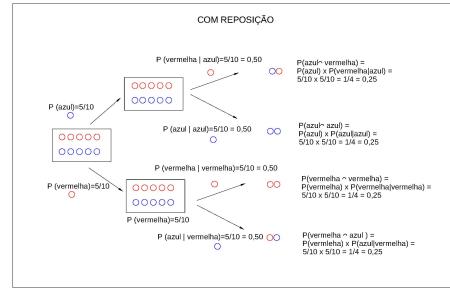


Figure 4.19: Ilustração do experimento aleatório sob a condição de reposição

4.5 Demonstraçāo clássica de dependēcia

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 a **não a colocamos de volta na**

sacola. Em seguida, removemos outra bola aleatória da bolsa e registramos sua cor.

- 1- Qual é a probabilidade de a primeira bola ser **vermelha** ?
- 2- Qual é a probabilidade de a segunda bola ser **azul**?
- 3- Qual é a probabilidade de a primeira bola ser **vermelha** e a segunda bola **azul**?
- 4- A primeira bola retirada foi uma bola **vermelha** e a segunda bola **azul**; esses eventos foram *independentes*?

Solução:

1^a Etapa: analisar todos os possíveis resultados

- Probabilidade da primeira bola retirada ser **vermelha** e a segunda ser **azul**:

Ao se retirar duas bolas do sacola há quatro possíveis combinações de resultados.
Nós podemos obter:

- uma **vermelha** e depois outra **vermelha**;
- uma **vermelha** e depois uma **azul**;
- uma **azul** e depois uma **vermelha** ; ou,
- uma **azul** e depois outra **azul**.

Queremos saber a probabilidade do segundo resultado após termos obtido uma bola **vermelha** na primeira seleção.

Como existem 5 bolas **vermelhas** e 10 bolas no total, existem $\frac{5}{10}$ maneiras de obter uma bola **vermelha** primeiro.

Entretanto, nessa nova situação, nós não colocamos a primeira bola de volta, então haverá apenas 4 bolas **vermelhas** e 5 bolas **azuis** na sacola.

- Haverá $\frac{4}{9}$ maneiras de obter uma segunda bola **vermelha** se a primeira bola for **vermelha**. Isso significa que existem: $\frac{5}{10} \times \frac{4}{9} = \frac{20}{90}$ maneiras de se obter uma bola **vermelha** em primeiro lugar e uma bola **vermelha** em segundo. Então, a probabilidade associada será de $\frac{2}{9}$;

- Haverá $\frac{5}{9}$ maneiras de obter uma segunda bola **azul** se a primeira bola for **vermelha**. Isso significa que existem: $\frac{5}{10} \times \frac{5}{9} = \frac{25}{90}$ maneiras de se obter uma bola **vermelha** em primeiro lugar e uma bola **azul** em segundo. Então, a probabilidade associada será de $\frac{5}{18}$;

- Haverá $\frac{5}{9}$ maneiras de obter uma segunda bola **vermelha** se a primeira bola for **azul**. Isso significa que existem: $\frac{5}{10} \times \frac{5}{9} = \frac{25}{90}$ maneiras de se obter uma bola **azul** em primeiro lugar e uma bola **vermelha** em segundo. Então, a probabilidade associada será de $\frac{5}{18}$.

- Haverá $\frac{4}{9}$ maneiras de obter uma segunda bola **azul** se a primeira bola for **azul**. Isso significa que existem: $\frac{5}{10} \times \frac{4}{9} = \frac{20}{90}$ maneiras de se obter uma bola **azul** em primeiro lugar e uma bola **azul** em segundo. Então, a probabilidade associada será de $\frac{2}{9}$;

Resumo das probabilidades calculadas:

- 1 -uma **vermelha** e depois outra **vermelha** : $\frac{2}{9}$;
- 2- uma **vermelha** e depois uma **azul**: $\frac{5}{18}$;
- 3- uma **azul** e depois uma **vermelha** : $\frac{5}{18}$; e,
- 4- uma **azul** e depois outra **azul**: $\frac{2}{9}$.

2^a Etapa: analisar a possibilidade de se obter uma bola **vermelha** na primeira extração:

- uma **vermelha** e depois outra **vermelha** : $\frac{2}{9}$;
- uma **vermelha** e depois uma **azul**: $\frac{5}{18}$.

A probabilidade total de se obter uma bola **vermelha** na primeira extração será:

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

3^a Etapa: analisar a possibilidade de se obter uma bola **azul** na segunda extração:

- uma **vermelha** e depois uma **azul**: $\frac{5}{18}$;
- uma **azul** e depois outra **azul**: $\frac{2}{9}$.

146 CHAPTER 4. - INTRODUÇÃO AO CÁLCULO DE PROBABILIDADES

A probabilidade total de se obter uma bola azul na segunda extração será:

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

4^a Etapa: analisar a possibilidade de se obter uma bola vermelha e em seguida azul:

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

5^a Etapa: Esses dois eventos são independentes?

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

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

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

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

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

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

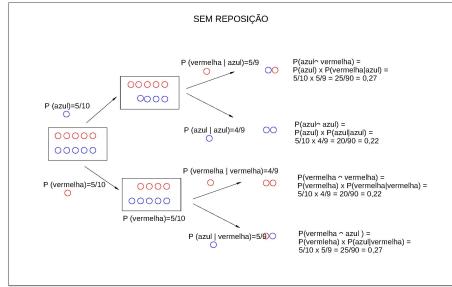


Figure 4.20: Ilustração do experimento aleatório sob a condição de não reposição

4.6 Teoremas da Teoria das probabilidades

4.6.1 Teorema 01

Se E é um evento num espaço discreto Ω , então $P(E)$ é igual à soma das probabilidades de ocorrência de todos os elementos do espaço amostral que satisfazem ao evento de interesse E .

Sejam E_1, E_2, E_3, \dots a sequência finita ou infinita de eventos que satisfazem ao evento de interesse E . Assim, $E = E_1 \cup E_2 \cup E_3 \dots$. Como E_1, E_2, E_3, \dots são eventos **mutuamente exclusivos**, pelo terceiro postulado das probabilidades teremos:

$$P(E) = P(E_1) + P(E_2) + P(E_3) + \dots$$

Exemplo: Experimento: lançamento de uma moeda duas vezes

Espaço amostral dos possíveis eventos (resultados): $\Omega = \{(cara, cara), (cara, coroa), (coroa, cara), (coroa, coroa)\}$

- Evento de interesse E : obter ao menos uma *cara*
- Eventos que satisfazem: $E_1 = \{(cara, cara)\}$; $E_2 = \{(cara, coroa)\}$; $E_3 = \{(coroa, cara)\}$

A probabilidade de E ($P(E)$) será a soma das probabilidades dos eventos que o satisfazem:

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

4.6.2 Teorema 02

Se um experimento aleatório pode ter N resultados possíveis e equiprováveis e um evento E pode ter n resultados que o satisfazem, então $P(E) = \frac{n}{N}$.

Sejam $E_1, E_2, E_3, \dots, E_N$ os resultados do espaço amostral Ω , cada um deles equiprovável ($P(E_i) = \frac{1}{N}$). Se E é a união de n desses eventos **mutuamente exclusivos}, pelo terceiro postulado das probabilidades teremos:

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

4.6.3 Teorema 03

Se E e E^c são eventos complementares no espaço amostra Ω então $P(E^c) = 1 - P(E)$.

Sendo os eventos E e E^c **mutuamente exclusivos** e também sendo $E \cup E^c = \Omega$, considerando-se que $P(\Omega) = 1$, pelos segundo e terceiro postulados tem-se:

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

4.6.4 Teorema 04

$$P(\emptyset) = 0$$

Sendo Ω e \emptyset são **mutuamente exclusivos** e, como de acordo com a definição de um espaço vazio $\Omega \cup \emptyset = \Omega$, pelo terceiro postulado tem-se:

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

4.6.5 Teorema 05

Se A e B são eventos em um mesmo espaço amostral Ω e $A \subset B$ então $P(A) \leq P(B)$.

Se $A \subset B$ então pode-se escrever: $B = A \cup (A^c \cap B)$ (verifica-se pelo correspondente diagrama de Venn).

Como A e $A^c \cap B$ são **mutuamente exclusivos**, pelo terceiro postulado tem-se:

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

4.6.6 Teorema 06

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

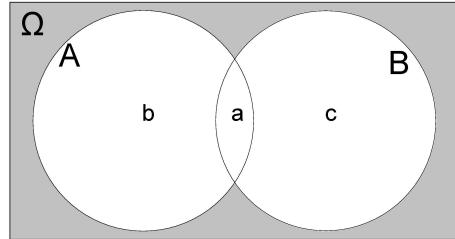
Estando $\emptyset \subset E \subset \Omega$ e considerando-se o Teorema 5 tem-se:

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

4.6.7 Teorema 07

Para dois eventos quaisquer em Ω , A e B tem-se que: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Sejam as seguintes probabilidades para esses eventos **mutuamente exclusivos**:



- $P(A \cap B) = a$;
- $P(A \cap B^c) = b$; e,
- $P(A^c \cap B) = c$.

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

4.6.8 Teorema 08

Para três eventos quaisquer em Ω , A , B e C tem-se que:

$$\begin{aligned} P(A \cup B \cup C) &= \\ &= P(A) + P(B) + P(C) - \\ &\quad P(A \cap B) - P(A \cap C) - P(B \cap C) + \\ &\quad P(A \cap B \cap C) \end{aligned}$$

Escrevendo-se $A \cup B \cup C$ como $A \cup (B \cup C)$ e usando o Teorema 7 duas vezes (uma para $P[A \cup (B \cup C)]$ e a outra para $P(B \cup C)$) tem-se:

$$\begin{aligned} P(A \cup B \cup C) &= P[A \cup (B \cup C)] \\ P(A \cup B \cup C) &= P(A) + P(B \cup C) - P[A \cap (B \cup C)] \\ P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(B \cap C) - P[A \cap (B \cup C)] \end{aligned}$$

Pela lei distributiva tem-se:

$$\begin{aligned} P[A \cap (B \cup C)] &= P[(A \cap B) \cup (A \cap C)] \\ P[A \cap (B \cup C)] &= P(A \cap B) + P(A \cap C) - P[(A \cap B) \cap (A \cap C)] \\ P[A \cap (B \cup C)] &= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \end{aligned}$$

Chega-se a :

$$\begin{aligned} P(A \cup B \cup C) = \\ P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + \\ P(A \cap B \cap C) \end{aligned}$$

Chapter 5

- Introdução a variáveis aleatórias

Chapter 6

- Introdução a modelos teóricos de probabilidade

Table 6.1: 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

Chapter 7

- Introdução ao planejamento de pesquisas

Table 7.1: 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

Chapter 8

- Introdução à estatística na epidemiologia

Table 8.1: 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

Chapter 9

- Introdução à distribuição das médias amostrais, suas diferenças e seus intervalos de confiança

Chapter 10

- Introdução à distribuição das proporções amostrais, suas diferenças e seus intervalos de confiança

Chapter 11

- Introdução a testes de
hipóteses

Chapter 12

- Introdução ao modelo
clássico de regressão linear
simples

Chapter 13

- Introdução à análise multivariada: discriminante linear de Fisher

Chapter 14

- Introdução à estatística experimental: análise de variância (DIC e DBC)