Universidade Federal de Uberlândia
Faculdade de Matemática
Curso de Licenciatura em Matemática a Distância

ESTATÍSTICA
Da educação básica ao ensino superior

Aurélia Aparecida de Araújo Rodrigues

Segunda Edição
Revista e Atualizada

UFU
2017

Copyright © 2015 by Aurélia Aparecida de Araújo Rodrigues

1ª edição 2015
2ª edição 2017

Rodrigues, Aurélia Aparecida de Araújo
ESTATÍSTICA: Da educação básica ao ensino superior / Aurélia Aparecida de Araújo Rodrigues. 2ª ed. - Uberlândia, MG : UFU, 2017.

185 p.:il.

Licenciatura em Matemática.

1. ESTATÍSTICA: Da educação básica ao ensino superior

Reitor
Valder Steffen Júnior

Coordenador UAB/CEAD/UFU
Maria Teresa Menezes Freitas

Conselho Editorial
Carlos Rinaldi - UFMT
Carmen Lucia Brancaglion Passos - UFScar
Célia Zorzo Barcelos - UFU
Eucidio Arruda Pimenta - UFMG
Ivete Martins Pinto - FURG
João Frederico Costa Azevedo Meyer - UNICAMP
Marisa Pinheiro Mourão - UFU

Edição
Centro de Educação a Distância
Comissão Editorial - CEAD/UFU

Diagramação
Equipe CEAD/UFU

PRESIDENTE DA REPÚBLICA
Michel Miguel Elias Temer

MINISTRO DA EDUCAÇÃO
José Mendonça Bezerra Filho

UNIVERSIDADE ABERTA DO BRASIL
DIRETORIA DE EDUCAÇÃO A DISTÂNCIA/CAPES
Carlos Cezar Modernel Lenuzza

UNIVERSIDADE FEDERAL DE UBERLÂNDIA - UFU
REITOR
Valder Steffen Júnior

VICE-REITOR
Orlando César Mantese

CENTRO DE EDUCAÇÃO A DISTÂNCIA
DIRETORA E REPRESENTANTE UAB/UFU
Maria Teresa Menezes Freitas

FACULDADE DE MÁTEMÁTICA – FAMAT – UFU
DIRETOR
Luís Antonio Benedetti

COORDENADOR DO CURSO DE LICENCIATURA EM MATEMÁTICA – PARFOR
Rogério de Melo Costa Pinto

COORDENAÇÃO DE TUTORIA
Janser Moura Pereira

EQUIPE DO CENTRO DE EDUCAÇÃO A DISTÂNCIA DA UFU - CEaD/UFU

ASSESSORA DA DIRETORIA
Sarah Mendonça de Araújo

EQUIPE MULTIDISCIPLINAR
Alberto Dumont Alves Oliveira
Darcius Ferreira Lisboa Oliveira
Dirceu Nogueira de Sales Duarte Júnior
Gustavo Bruno do Vale
Otaviano Ferreira Guimarães

SETOR DE FORMAÇÃO CONTINUADA
Marisa Pinheiro Mourão

APOIO PEDAGÓGICO
Alícia Felisbino Ramos
Ana Rafaella Ferreira Ramos
Giseli Vale Gatti
Maria Helena Cicci Romero

EQUIPE DE ESTAGIÁRIOS DO CEAD E DO CURSO DE MATEMÁTICA

REVISORA
Paula Godoi Arbex

SUMÁRIO

SUMÁRIO **5**
FIGURAS **8**
INFORMAÇÕES **10**
MÓDULO 1 ..... **11**
ESTATÍSTICA: DA EDUCAÇÃO BÁSICA AO ENSINO SUPERIOR **13**
I - TEXTO BÁSICO 13
1. Introdução **13**
1.1. Histórico **13**
1.2. O papel da estatística na metodologia científica **15**
1.3. Aplicações da estatística a situações cotidianas **19**
2. Análise exploratória de dados **19**
2.1 Introdução à análise de dados **19**
II - ATIVIDADES 19
III - ATIVIDADES 21
2.1.2. Apresentação e tabulação de dados **23**
2.1.3. Distribuição de frequências **25**
IV - ATIVIDADES 28
2.1.4. Análise gráfica **28**
2.2 Medidas estatísticas **36**
2.2.1 Medidas de tendência central: média aritmética, mediana, moda. **37**
2.2.2 Outras medidas de tendência central: **39**
2.2.3. Separatrizes – quartil, decil, percentil. **40**
2.2.4 Medidas de dispersão **44**
2.2.5.Medidas de assimetria e curtose **50**
2.2.6 Box-plot (diagrama de caixas) **54**
V - SÍNTESE DO MÓDULO 58
VI - REFERÊNCIAS 59
MÓDULO 2 ..... **61**
Da educação básica ao ensino superior **63**
I - TEXTO BÁSICO 63
3. Probabilidade e variáveis aleatórias **63**

3.1. Introdução à probabilidade - conceitos e propriedades **63**
3.1.1. Definição de probabilidade **65**
3.1.2. Axiomas da probabilidade **67**
3.2. Probabilidade condicional **69**
3.3. Independência de eventos **71**
3.4. Teorema do Produto **72**
3.5. Teorema da Soma **73**
3.6. Outros teoremas **74**
3.7. Teorema de Bayes **74**
3.8. Variáveis aleatórias unidimensionais discretas e contínuas **77**
3.8.1. Variável aleatória discreta **78**
3.8.2. Variável aletória contínua **80**
3.9. Esperança matemática e variância de variáveis aleatórias unidimensionais .. **82**
3.9.1. Esperança matemática de uma v.a **82**
3.9.2. Propriedades da média **82**
3.9.3. Variância de um ***v.a***. **82**
3.9.4. Propriedades da variância **83**
3.9.5. Esperança matemática e variância de uma variável aleatória contínua **83**
3.10. Distribuições discretas de probabilidade **84**
3.10.1. Distribuição uniforme discreta **84**
3.10.2. Distribuição de Bernoulli **85**
3.10.3. Distribuição binomial **86**
3.10.4. Distribuição de Poisson **88**
3.10.5. Distribuição Geométrica **90**
3.10.6. Distribuição Hipergeométrica **92**
3.11. Distribuições contínuas de probabilidade **93**
3.11.2. Distribuição exponencial **93**
3.11.3. Distribuição normal **94**
II - SÍNTESE 101
III - TABELAS 102
IV - REFERÊNCIAS 104
MÓDULO 3 105
ESTATÍSTICA: DA EDUCAÇÃO BÁSICA AO ENSINO SUPERIOR 107
I - TEXTO BÁSICO 107

4. Amostragem e Distribuições Amostrais **107**
4.1. Conceitos Básicos **107**
4.2. Parâmetros e Estimação pontual **108**
4.3. Distribuições amostrais **109**
4.3.1. Distribuição amostral da média **109**
4.3.2. O Teorema Central do Limite (TCL) **110**
4.3.3. Distribuição amostral da proporção **112**
4.3.4. Distribuição amostral da variância **112**
4.4. Outras técnicas de amostragem **113**
4.4.1. Amostragem Sistemática **113**
4.4.2. Amostragem por Conglomerados **113**
4.4.3. Amostragem estratificada **115**
5. Estimação por intervalo **119**
5.1 Conceitos Básicos **119**
5.2. Intervalos de Confiança para a média populacional **119**
5.2.1. Intervalos de Confiança para a média populacional, quando σ for conhecido **119**
5.2.2. Intervalos de Confiança para a média populacional, quando σ for desconhecido e $n < 30$. **122**
5.2.3. Intervalos de Confiança para a média populacional, quando σ for desconhecido e $n \geq 30$ **125**
5.3. Intervalos de Confiança para a proporção **126**
5.4. Intervalos de Confiança para a variância **128**
5.5. Tamanho da amostra **130**
II - SÍNTESE 132
III - TABELAS 133
IV - REFERÊNCIAS 137
MÓDULO 4 139
Teste de hipóteses, Correlação e Regressão e Aplicações 141
I - TEXTO BÁSICO 141
6. Testes de Hipóteses **141**
6.1. Introdução : conceitos fundamentais e tipo de erros **141**
6.1.1. Conceitos fundamentais. **141**
6.1.2. Tipo de erros **142**
6.1.6. Procedimento para se efetuar um teste de hipóteses **143**

6.2. Teste de hipótese para a média **143**
6.2.1. Teste de hipótese para a média, σ conhecido **143**
6.2.2. Teste de hipótese para a média, σ desconhecido e $n < 30$ **148**
6.2.3. Teste de hipótese para a média, σ desconhecido e $n \geq 30$ **152**
6.3. Teste de hipótese para a proporção **154**
6.4. Teste de hipóteses para independência **157**
7. Regressão e Correlação **162**
7.1 Diagrama de dispersão **162**
7.2 Coeficiente de Correlação de Pearson **164**
7.3. Regressão linear simples **167**
7.3.1 Modelo matemático **168**
7.3.2. Equação da reta do modelo regressão linear simples **168**
SÍNTESE 170
Tabelas 171
REFERÊNCIAS 175

FIGURAS

Gráfico 1.1 Porcentagem de estudantes que trabalham, UFU,2014 25
Quadro 2.1 Quadro ilustrativo 32
Tabela 2.1 Tabela ilustrativa..... 32
Gráfico 2.1 Histograma para a idade, em anos 37
Gráfico 2.2 Polígono de frequências para a idade, em anos..... 38
Tabela 2.6 Distribuição de frequências acumuladas para a idade, em anos..... 39
Gráfico 2.3 Ogiva para a idade, em anos 39
Gráfico 2.4 Porcentagem de suicidas.....42
Gráfico 2.5 Porcentagem de suicidas..... 43
Gráfico 2.6 Porcentagem de suicidas..... 43
Tabela 2.8: Taxa de mortalidade por câncer (mortes por 100 mil pessoas) 44
Gráfico 2.7: Taxa de mortalidade por câncer 44
Quadro 2.2: Posições das separatrizes 50
Tabela 2.9 Notas e média..... 52
Tabela 2.10 Notas, média e medidas de dispersão 54
Figura 2.2: Curva de uma distribuição assimétrica à direita..... 59
Figura 2.3: Curva de uma distribuição assimétrica à esquerda..... 59
Figura 2.4: Curva de uma distribuição leptocúrtica..... 60
Figura 2.5: Curva de uma distribuição mesocúrtica 60
Figura 2.6: Curva de uma distribuição platicúrtica..... 61
Figura 2.7: Comparação entre as curvas das distribuições mesocúrtica, leptocúrtica e platicúrtica 61
Figura 2.8 Box-plot..... 62
Figura 2.8 Distribuição simétrica..... 62
Figura 2.9 Distribuição assimétrica à direita..... 63
Figura 2.10 Distribuição assimétrica à esquerda..... 63
Gráfico 2.8 Box-plot para tempo de espera 64
Gráfico 2.9 Box-plot para tempo de espera 65
Quadro 3.1 Peso e pressão arterial 80
Quadro 3.2 Peso e pressão arterial 82
Figura 3.1: Ilustração do Teorema de Bayes..... 85

Gráfico 3.1: Representação de uma distribuição de probabilidade 88
Gráfico 3.2 Gráfico da $f(x) = 2x$, $0 \leq x \leq 1$ 90
Quadro 3.1: Exemplo da distribuição binomial 96
Quadro 3.2: Exemplo da distribuição geométrica 100
Figura 3.2: Curva da distribuição normal 104
Figura 3.3: Curva da distribuição normal, $\mu = 100$, $\sigma = 10$ 105
Figura 3.4: Curva da distribuição normal padrão 106
Figura 3.5: Curva da distribuição normal e normal padrão, $P(X \leq 110)$ e $P(Z \leq 1)$ 107
Figura 3.6: Curva da distribuição normal e normal padrão, $P(X \leq 87{,}5)$ e $P(Z \leq -1{,}25)$ 108
Figura 3.7: Curva da distribuição normal e normal padrão, $P(X \leq 13{,}5)$ e $P(Z \leq 1{,}5)$ 109
Figura 3.8: Curva da distribuição normal e normal padrão, $P(X > 13{,}5)$ e $P(Z > 1{,}5)$ 109
Quadro 3.1: Parâmetros e estimadores 118
Tabela 3.1: Valores de $\bar{X}$ para as 9 amostras aleatórias simples 119
Figura 4.1: Ilustração de amostragem por conglomerados 123
Figura 4.2: Exemplo de amostragem por conglomerados 124
Figura 4.3: Ilustração de amostragem estratificada 126
Figura 4.4: Exemplo de amostragem estratificada 127
Figura 4.5: Exemplo de amostragem estratificada proporcional 128
Figura 4.6: Ilustração da curva da distribuição normal 130
Figura 4.7: Ilustração de valores tabelados de z 132
Figura 4.8: Ilustração de valores tabelados de z 132
Figura 4.7: Ilustração da curva da distribuição *t* de Student 134
Figura 4.8: Comparação da curva da distribuição normal e da *t* de Student 134
Figura 4.11: Ilustração de valores tabelados de t 136
Tabela 4.1: Reprodução parcial da tabela C 136
Figura 4.12: Ilustração de valores tabelados de z 138
Figura 4.9: Ilustração da curva da distribuição normal 138
Figura 4.10: Ilustração da curva da distribuição qui-quadrado 141
Figura 4.15: Ilustração de valores tabelados da distribuição qui-quadradro 143
Tabela 4.2: Reprodução parcial da tabela D 143
Quadro 6.1: Erro tipo I e erro tipo II 153
Figura 6.1. Região crítica e valores críticos para teste bilateral 154

Figura 6.2. Região crítica e valor crítico para teste unilateral à esquerda 154

Figura 6.3. Região crítica e valor crítico para teste unilateral à direita 155

Quadro 6.2: Resumo do teste Z para média μ 156

Quadro 6.3: Exemplo do teste Z para média μ 157

Figura 6.4. Região crítica e valores críticos para teste bilateral 158

Figura 6.6. Região crítica e valor crítico para teste unilateral à direita 159

Quadro 6.4: Resumo do teste *t* para média μ 160

Quadro 6.5: Exemplo do teste *t* para média μ 161

Quadro 6.6: Exemplo do teste *t* para média μ 162

Quadro 6.7: Exemplo do teste *Z* para média μ 164

Quadro 6.8: Resumo do teste *Z* para a proporção *p* 165

Quadro 6.9: Exemplo do teste *Z* para a proporção *p* 166

Figura 6.7. Região crítica e valor crítico para teste qui-quadrado.....168

Tabela 6.2: Tabela de contingência *r X c*, com frequências esperadas.....169

Quadro 6.10: Resumo do teste qui-quadrado para independência 169

Tabela 6.3: Frequências observadas do desempenho em matemática, por turno ..170

Figura 7.1. Correlação linear positiva 172

Figura 7.2. Correlação linear negativa 173

Figura 7.3. Não há correlação.....173

Figura 7.4. Correlação não linear 174

Tabela 7.1. Horas de estudo, por semana, e a nota final.....176

Gráfico 7.1. Gráfico de dispersão para dados da Tabela7.1 176

Gráfico 7.2. Reta de regressão para dados da Tabela 7.1. .....

INFORMAÇÕES

Prezado(a) aluno(a),

Ao longo deste guia impresso você encontrará alguns “ícones” que lhe ajudarão a identificar as atividades.

Fique atento ao significado de cada um deles, isso facilitará a sua leitura e seus estudos.

Destacamos alguns termos no texto do Guia cujos significados serão importantes para sua compreensão. Para permitir sua iniciativa e pesquisa, não criamos um glossário, mas se houver dificuldade, interaja no *Fórum de Dúvidas*.

SOBRE A AUTORA

Aurélia Aparecida de Araújo Rodrigues possui graduação em Licenciatura em Matemática pela UFU – Universidade Federal de Uberlândia (1998), Mestrado em Estatística e Métodos Quantitativos pela UnB – Universidade de Brasília (2001) e Doutorado em Engenharia de Produção pela Puc-rio - Pontifícia Universidade Católica do Rio de Janeiro (2005). Desde 2006, é professora da Faculdade de Matemática da Universidade Federal de Uberlândia (FAMAT – UFU). Atualmente atua no ensino de estatística em cursos de graduação do ensino superior, nos quais leciona as disciplinas Estatistica, Bioestatística e Controle Estatístico de Qualidade.

INTRODUÇÃO

Prezado (a) aluno (a),

Os Parâmetros Curriculares Nacionais (PCN's) do Ministério da Educação, BRASIL (1997) e BRASIL (1998), descrevem os conteúdos de estatística, probabilidade e contagem que podem ser apresentados aos alunos do 1º ao 9º ano do ensino fundamental. A descrição desses conteúdos é feita em um bloco específico, no bloco Tratamento da Informação.

De acordo com os PCN's, a finalidade é fazer com que o aluno compreenda os procedimentos para coletar, organizar, comunicar dados do cotidiano, utilizando tabelas, gráficos e medidas estatísticas. Dessa forma, é importante, por exemplo, conhecer as medidas estatísticas como média, mediana e moda com o objetivo de fornecer elementos para representar dados. Com relação à probabilidade, a principal finalidade é a de que o aluno compreenda que muitos dos acontecimentos do cotidiano são de natureza aleatória. As noções de acaso e incerteza, que se manifestam intuitivamente, podem ser exploradas na escola, em situações em que o aluno realiza experimentos e observa eventos. Relativamente aos problemas de contagem, o objetivo é levar o aluno a lidar com situações que envolvam diferentes tipos de agrupamentos que possibilitem o desenvolvimento do raciocínio combinatório e a compreensão do princípio multiplicativo para sua aplicação no cálculo de probabilidades.

Os alunos do 2º ao 5º ano do ensino fundamental já exploram idéias básicas de estatística, e os do 6º e 7º ano devem ampliar essas noções, aprendendo também a formular questões pertinentes para um conjunto de informações, a elaborar algumas conjecturas e comunicar informações de modo convincente, a interpretar diagramas e fluxogramas. É importante entender como analisar e avaliar informações estatísticas e tomar decisões. No 6º e 7º ano, também amplia-se a exploração das possibilidades de quantificar o incerto. Com as noções elementares de probabilidade os alunos aprenderão a determinar as chances de ocorrência de alguns eventos (moedas, dados, cartas). Assim, poderão ir se familiarizando com o modo como a Matemática é usada para fazer previsões e perceber a importância da probabilidade na vida cotidiana.

O entendimento do Tratamento da Informação pode ser aprofundado no 8º e 9º ano, pois, os alunos têm melhores condições de desenvolver pesquisas sobre sua própria realidade e interpretá-la, utilizando-se de gráficos e algumas medidas estatísticas. As pesquisas sobre Saúde, Meio Ambiente, Trabalho e Consumo poderão fornecer contextos

em que os conceitos e procedimentos estatísticos ganham significados.

Os PCN's para Ensino Médio também recomendam o desenvolvimento de competências e habilidades referentes a Estatística, tais como, emprego e interpretação de gráficos, tabelas, expressões algébricas, previsão de tendências, com aplicações a contextos sócio-econômicos, científicos ou cotidianos.

As técnicas de contagem, citadas no bloco Tratamento da Informação dos PCN's, são tema de estudo na disciplina Fundamentos da Matemática Elementar II, portanto não serão abordadas aqui.

O objetivo desta disciplina é preparar os professores de matemática para ensinar estatística e probabilidade aos alunos do ensino fundamental e médio, proporcionando uma visão crítica da análise de dados.

Espera-se que ao final dessa disciplina, o professor tenha adquirido conhecimentos para preparar projetos de ensino que despertem o interesse dos alunos pelas situações que envolvem a coleta e análise de dados. As atividades do projeto devem estimular a aplicação da estatística às situações práticas e emprego de planilhas eletrônicas.

A presente disciplina está dividida em quatro módulos e a duração de cada módulo é de quatro semanas.

Quanto à metodologia, nesta disciplina teremos vídeo aulas, estudo da teoria do Livro Texto, resolução de Listas de Exercícios, onde se encontram os exercícios a serem entregues e outros para que o aluno pratique os conceitos estudados. Teremos também Atividades Avaliativas e Provas Presenciais.

Quanto ao sistema de avaliação, serão distribuídos 100 pontos, sendo 60 pontos de provas escritas em modo presencial e 40 pontos das atividades passadas pelo Ambiente Virtual de Aprendizagem (AVA).

Quanto ao cronograma, as 90 horas do curso são distribuídas nos módulos de acordo com o número de semanas, considerando 4 horas de atividades de estudo da teoria por semana, sendo necessário considerar para cada hora de estudo em teoria pelo menos uma hora de estudo através de exercícios. Esse esquema tem por finalidade assegurar um treino mínimo nos módulos.

Desejo a todos um bom estudo,

Professora Aurélia.

MÓDULO 1

Introdução à estatística e análise exploratória de dados

Conteúdos básicos do Módulo:

1. Contexto histórico
2. O papel da estatística na metodologia científica
3. Aplicações da estatística a situações cotidianas
4. Análise exploratória de dados

Objetivos do Módulo:

Ao final deste estudo, esperamos que você, aluno(a), possa:

- Entender a importância da Estatística para a sociedade, desde a antiguidade até os dias atuais.
- Identificar relações entre a Estatística e a produção do conhecimento científico.
- Empregar métodos estatísticos, tais como tabelas, gráficos e medidas estatísticas, e interpretá-los.
- Compreender como a Estatística se relaciona com o dia a dia dos alunos, resul tando em implicações para o ensino dessa disciplina em sala de aula.

I - TEXTO BÁSICO

1. Introdução

Conhecer o contexto de origem e evolução da Estatística permite compreender melhor a importância dessa disciplina para a ciência e as suas aplicações em diversas áreas. Nenhuma disciplina tem interagido tanto com as demais disciplinas em suas atividades como a Estatística.

1.1. Histórico

A prática de coletar e analisar dados surgiu antes da era cristã. A seguir serão destacados alguns fatos relacionados com a história da Estatística, os quais foram extraídos dos trabalhos de MEMÓRIA (2004), LOPES e MEIRELLES (2005) e POUBEL (2011).

5000 a.C. - No Egito, foram feitos registros de presos de guerra.

3000 a.C. - Na Babilônia, China e Egito foram realizados censos. Foram elaborados registros egípcios da falta de mão de obra para a construção de pirâmides.

2239 a.C. – Na China, foi feito um recenseamento chinês com fins agrícolas e comerciais.

Século XI – No Império romano, Guilherme, o conquistador, solicitou um levantamento das propriedades rurais dos conquistados anglo-saxões para se inteirar de suas riquezas.

Século XVI – Na Itália, ocorre a emergência da estatística descritiva. Houve interesse pela coleta de dados, principalmente por suas aplicações na administração pública. Na Igreja Católica Romana, a partir do Concílio de Trento (1545-1563), tornam-se obrigatórios os registros de batismo, casamentos e óbitos. Embora alguns matemáticos italianos, como Niccolò Fontana Tartaglia, Girolamo Cardano e Galileu Galilei, já tivessem se interessado por problemas de probabilidades relacionados com jogos de dados, foi na França que surgiu o cálculo de probabilidades com o objetivo de solucionar problemas relacionados com jogos de azar, propostos por Chevalier de Méré, em 1653. Mediante correspondências, Blaise Pascal e Pierre Fermat contribuíram com a solução desses problemas e com a formalização da teoria da probabilidade.

Século XVII – A estatística começa a ser usada com o objetivo de descrever bens do estado. Foi feita a primeira tentativa para se tirar conclusões a partir de dados numéricos, na obra Aritmética Política, hoje chamada de demografia. John Graunt , em 1962, analisou registros disponíveis sobre mortalidade em 1696, foi construída a primeira tábua de sobrevivência, usada para cálculo de seguros de vida, pelo astrônomo inglês Edmond Halley, sendo considerado o criador do cálculo atuarial.

Século XVIII – O alemão Gottfried Achenwall, em 1746, criou e empregou o termo *estatística*. Ocorreu a formalização de importantes teorias estatísticas, Pierre Simon de Laplace generalizou o teorema de Moivre em seu trabalho e obteve o Teorema central do limite. A distribuição normal e o método de mínimos quadrados, em 1795, foram temas de estudo de Carl Friedrich Gauss. A estatística aparece, pela primeira vez, como meio indutivo de investigação.

Século XIX – Surge a estatística médica, com o trabalho de William Farr. Surgem trabalhos na área de técnicas de correlação e ajustamento de curvas. Karl Pearson, em 1896, obteve equação para o cálculo do coeficiente de correlação, que é utilizado nos dias de hoje, e posteriormente, desenvolveu análise de regressão e teste de hipóteses qui-quadrado. Lambert Quetelet analisou dados sociais. Aplicação de métodos estatísticos aos problemas biológicos relacionados com a evolução e a hereditariedade (evolução de Darwin). Foram obtidos resultados valiosos para o desenvolvimento da Inferência Estatística.

No Brasil, em 1808, D. João VI solicita que sejam enviadas a Portugal informações estatísticas referentes à cultura, produção, consumo e exportação. De 1871 a 1876, foi feito o primeiro censo geral do Império em 23 volumes contendo 8546 tabelas.

Século XX- O período de 1900 e 1915 foi considerado de transição entre a visão original e a nova visão de estatística, sendo o início da estatística moderna. William Sealy Gosset, com o pseudônimo de *Student*, iniciou estudo sobre pequenas amostras, levando à formalização da distribuição de *t* de *Student*.

Em 1912, Ronald Aylmer Fisher, considerado por muitos o maior estatístico de todos os tempos, criou os métodos modernos da Análise e Delineamento de Experimentos. O período de 1925 a 1960 tornou-se a época áurea do pensamento estatístico, com importantes trabalhos na área de inferência estatística, planejamento de experimentos, levantamentos por amostragem e séries temporais.

Século XXI - Atualmente, há aumento da formalização da estatística, através de modelos matemáticos, e aumento do uso de recursos computacionais, devido aos avanços tecnológicos.

1.2. O papel da estatística na metodologia científica

Os trabalhos científicos que envolvem coleta e análise de dados devem empregar métodos estatísticos de forma sistemática para obter informações adequadas, ou seja, teremos generalizações, estimativas de parâmetros ou modelos de previsões representativos e confiáveis. Quando bem fundamentados, os resultados de pesquisas, acadêmicas ou não, terão validade e aceitação no meio científico.

Assim sendo, é necessário que os trabalhos científicos contemplem as seguintes fases do método estatístico:

1. Definição do Problema
2. Planejamento
3. Coleta de dados
4. Apuração dos dados
5. Apresentação dos dados.
6. Análise e interpretação dos dados

A seguir, faremos o detalhamento de cada fase do método estatístico.

1. Definição do Problema: consiste na formulação do problema que se deseja estudar, deve-se saber exatamente o que se pretende pesquisar.

Exemplo: Estudar a condição social dos estudantes da UFU. Então, há interesse pelo perfil dos estudantes (idade, sexo, renda, profissão, endereço e outros).

2. Planejamento: determina o procedimento necessário para resolver o problema.

"Como levantar as informações desejadas?"

"Que dados deverão ser obtidos?"

"Como se deve obtê-los?"

Muitas vezes, é possível usar dados de banco de dados já existentes (IBGE, empresas...).

É nesta fase que se decide o tamanho da amostra, se a pesquisa será censitária por amostragem.

Quando for o caso, será formulado o questionário, que deve ser sucinto e possuir perguntas claras.

Outros elementos importantes desta fase são: **cronograma das atividades, estimativa dos custos envolvidos, análise das informações disponíveis, a forma como serão organizados os dados, tamanho da amostra.**

PARE E PENSE: o planejamento é a fase mais importante do trabalho que envolve coleta e análise de dados.

3. Coleta de dados: obtenção das informações através do contato direto do pesquisador (entrevistador) com os indivíduos da amostra (pessoas ou produtos).

$\Rightarrow$ É muito comum o uso de dados vindos de outras fontes, por exemplo, o uso dos dados do IBGE.

4. Apuração dos dados: é a montagem de um banco de dados. Os dados chegam ao analista de forma desorganizada.

Exemplo: considere uma situação fictícia e suponha que 400 estudantes da UFU foram entrevistados; destes, 200 alunos não trabalham, 120 trabalham e 80 não opinaram. Veja parte dos dados no Quadro 1.1.

Quadro 1.1 Banco de dados fictício, UFU, 2014.

Nome ou código	Idade	Sexo	Trabalha?
1	23	M	Sim
2	27	F	Não
3	18	M	Não opinou
⋮	⋮	⋮	⋮
400	20	M	Sim

5. Apresentação dos dados: os dados coletados podem ser apresentados em tabelas, gráficos e medidas estatísticas. Os dados são organizados através de contagem e agrupamento. Essa etapa é feita com auxílio de programas computacionais , tais como

softwares estatísticos ou planilhas eletrônicas.

a. **Tabelas:** apresentação numérica, segundo algumas regras fixadas pela ABNT (Associação Brasileira de Normas Técnicas).

Tabela 1.1 Número de estudantes que trabalham, de acordo com faixa etária e sexo, UFU, 2014.

Idade	Homens	Mulheres	TOTAL
< 20 anos	15	8	23
20 ⊢ 30 anos	20	10	30
30 ⊢ 40 anos	40	20	60
≥ 40 anos	5	2	7
TOTAL	80	40	120

b. **Gráfico:** permite uma visualização rápida e clara dos dados .

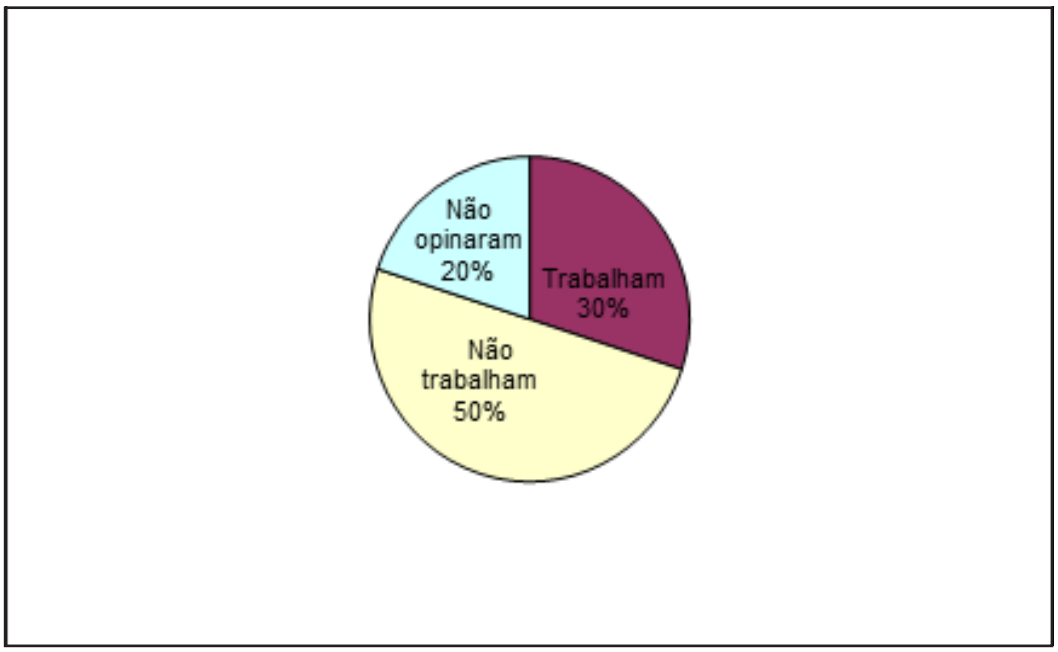


Gráfico 1.1 Porcentagem de estudantes que trabalham, UFU,2014

c. **Medidas estatísticas**: mostra características do conjunto de dados empregando apenas um valor numérico.

Exemplo: a idade média dos estudantes entrevistados é 23,4 anos.

Os diversos tipos de tabelas, gráficos e medidas estatísticas serão apresentados e detalhados nos capítulos seguintes.

6. Análise e interpretação dos dados: é fundamental analisar os gráficos, tabelas e medidas estatísticas.

No Quadro 1.1, observa-se que 400 estudantes da UFU foram entrevistados. De acordo com a Tabela 1.1, tem-se que 120 estudantes trabalham, dos quais 80 são homens e 40 são mulheres.

De acordo com o Gráfico 1.1, dos 400 estudantes entrevistados, 50% não trabalham, 30% trabalham e 20% não opinaram.

Todos os gráficos, tabelas e medidas estatísticas devem ser seguidos de uma análise, por mais óbvias que sejam as informações contidas neles.

Em alguns casos, há mais de uma análise possível. Neste caso, cabe ao pesquisador escolher a análise que permite destacar o resultado que ele julga ser o mais importante.

Ao se interpretar um gráfico ou uma tabela, o pesquisador deve-se limitar a emitir as informações neles contidas. Deve-se evitar dar opiniões pessoais sobre o resultado obtido**.**

Frequentemente, em alguns contextos, após fases 1 a 6, é possível empregar métodos estatísticos específicos para generalizar resultados de dados amostrais (amostragem e Inferência estatística) ou para obter modelos de previsão para dados em séries de tempo (Séries temporais) ou para analisar mais de uma variável simultaneamente (análise multivariada) ou para estudar a relação entre variáveis (modelos de regressão) ou outros. Alguns desses métodos serão estudados aqui.

1.3. Aplicações da estatística a situações cotidianas

Sabe-se que nos dias de hoje, com o advento da computação e da internet, as mais diversas instituições públicas e privadas geram bancos de dados enormes. Podemos citar, como exemplo de geradores de dados, as instituições financeiras (bancos e cartões de crédito), o comércio (lojas de departamento e empresas de telefonia) e órgãos do governo (Ministério da Saúde, Ministério da Educação, IBGE) entre outros.

Além disso, os resumos de dados estão presentes na TV, nos jornais e nas revistas. Daí a importância de todo cidadão ter conhecimento de estatística e condições de fazer uma leitura consciente e crítica do conteúdo que é divulgado nos meios de comunicação. Assim sendo, a estatística é tema de ensino nos programas do ensino fundamental e médio e superior.

No dia a dia, é comum o emprego da estatística em debates e notícias que tratam de economia, política, esportes, educação, saúde, alimentação, moradia, meteorologia, pesquisas de opinião, pesquisas de intenção de votos, meio ambiente, sociais, engenharia, entre outros.

Neste curso, será proposta a produção de um projeto de ensino de estatística para um dos tópicos estudados na disciplina, o qual deverá considerar a relação entre os métodos estatísticos com as situações cotidianas.

2. Análise exploratória de dados

2.1 Introdução à análise de dados

Proponho, inicialmente, algumas questões para reflexão.

II - ATIVIDADES

Pergunta: O que é estatística?

Pergunta: A estatística é importante nos dias de hoje? Por quê?

Pergunta: Como é o ensino de estatística no ensino fundamental e médio?

Pare aqui e reflita antes de prosseguir. Faça as anotações no espaço após cada pergunta.

A abrangência do emprego da Estatística dificulta conceituá-la; em geral, é satisfatório dizer que:

Estatística é o conjunto de métodos usados para obter, organizar e analisar informações numéricas, de modo que se possa descrever situações, tomar decisões, generalizar resultados, obter modelos de previsões e outros.

Para fins didáticos, divide-se a Estatística Básica em três áreas, as quais serão estudadas aqui.

a) **Estatística descritiva:** organiza e resume os dados, que podem ser muito complexos, por meio de números, tabelas e gráficos.

Exemplo: taxas de acidente, salário médio.

b) **Probabilidade:** lida com situações influenciadas por fatores relacionados com o acaso.

Exemplo: probabilidade de chover no final de semana.

c) **Inferência estatística:** generaliza informações amostrais para a população.

Exemplo: seleciona alguns cintos de segurança da marca A para testar a resistência. Esses testes são autodestrutivos e não devem ser aplicados a todos os cintos produzidos. Portanto, a resistência média da amostra será generalizada para todos os cintos de segurança da marca A. Outros tópicos que serão estudados aqui são: técnicas de

amostragem, teste qui-quadrado e regressão linear simples.

As áreas da Estatística que não serão estudadas aqui são: planejamento de experimentos, séries temporais, análise multivariada, regressão linear múltipla, regressão não linear, controle estatístico de qualidade e outros.

2.1.1. Conceitos Básicos

Antes de prosseguirmos com o estudo dos métodos estatísticos, é importante conhecer alguns conceitos importantes e que são base para o nosso estudo.

População é o conjunto constituído por elementos que apresentam pelo menos uma característica em comum.

Exemplo: moradores de Uberlândia, clientes do banco X, televisores da marca Y.

Amostra é um subconjunto da população.

Muitas vezes, analisar todos os elementos de uma população gastaria muito tempo e seria muito caro.

Nem sempre a população é constituída de pessoas, tem-se também população de animais, vegetais, itens de produção, alimentos, medicamentos e outros.

III - ATIVIDADES

Pergunta : Por que amostrar?

Em outras situações, há inviabilidade prática, ou seja, não é possível acessar todos os elementos da população. Por exemplo, seria impossível coletar informações de todos os diabéticos de uma cidade, visto que muitos diabéticos desconhecem serem portadores de diabetes.

Em uma **Amostra Aleatória Simples,** todos os elementos têm a mesma probabilidade de serem selecionados. Na prática, atribui-se um número a cada elemento da população e o emprego de um sorteio é uma forma de obter uma amostra aleatória simples.

Para obter uma amostra representativa, é necessário utilizar técnicas de amostragem. E a técnica de amostragem mais simples é a amostragem aleatória simples.

Quando a população for grande, é possível utilizar uma rotina computacional para identificar os elementos que vão compor a amostra aleatória.

Se os dados da amostra não foram obtidos de forma aleatória, então temos um **Estudo de Caso.** Por exemplo, um estudo com a turma do primeiro ano não é representativo para a escola toda.

Quando temos uma amostra aleatória, ela é representativa. Então, as informações dessa amostra podem ser estendidas para a população. Essa propriedade é importantíssima para a Inferência Estatística.

Além disso, é bom lembrar que quando todos os elementos da população fazem parte da amostra, temos um **Censo.**

Quando a amostra não for aleatória, ela não é representativa para a população. Então, as informações dessa amostra não podem ser estendidas para a população.

Em estatística, a característica de interesse de um estudo recebe o nome de **variável.** Formalmente, dizemos que a **variável** é a característica que vai ser observada, medida ou contada nos elementos da população ou amostra, e que pode variar, ou seja, assumir um valor diferente de elemento para elemento. Há mais de um tipo de variável.

As **variáveis quantitativas** são sempre numéricas. As variáveis quantitativas são classificadas em:

- **Variável contínua**: pode assumir qualquer valor real (positivo).

Exemplo: peso = 60,5 kg e altura= 1,72 m.

Outros exemplos de variáveis contínuas são: tempo, comprimento, velocidade, peso, nota

- **Variável discreta:** assume valores inteiros (resultam da contagem de itens). **Exemplo**: número de alunos fumantes, números de alunos aprovados, número de geladeiras defeituosas, número de celulares vendidos, número de filhos por casal, número de livros.

As **variáveis qualitativas** (ou categóricas) envolvem categorias pré-definidas. As variáveis qualitativas são classificadas em:

- **Variável nominal:** as categorias não apresentam ordenação natural.

Exemplo: Cor dos olhos: ( ) castanhos ( ) verdes ( ) azuis ( ) outras

- **Variável ordinal:** as categorias possuem ordenação natural.

Exemplo: classe social: ( ) baixa ( ) média ( ) alta.

2.1.2. Apresentação e tabulação de dados

Os quadros e as tabelas são formas assemelhadas de apresentação dos dados, com as informações dispostas em linhas e colunas. Conforme Milone (2004), os quadros não resumem informações, apenas registram, razão pela qual os valores que os compõem não podem ser relacionados entre si. As tabelas, ao contrário, premitem totalizar linha e colunas e estabelecer proporções em várias direções, conforme a necessidade de estudo.

Esteticamente, os quadros são completamente contornados por traços e as tabelas são delimitadas apenas superior e inferiormente. Veja ilustração nos Quadros 1.1 e 2.1 e nas Tabelas 1.1 e 2.1. As linhas e as colunas podem ser destacadas, mas há que se tomar cuidado pois, o excesso de linhas divisórias tende a dificultar a leitura das informações.

Quadro 2.1 Quadro ilustrativo

Tabela 2.1 Tabela ilustrativa

Fonte:

Os elementos fundamentais da tabela são: **título**, **cabeçalho**, **coluna indicadora** e **corpo**. O **título** aponta o conteúdo, local de ocorrência e data; o **cabeçalho** explica o conteúdo das colunas, a **coluna indicadora** detalha as linhas, o **corpo** mostra os dados.

Na Tabela 2.1, o cabeçalho é:

Idade	Homens	Mulheres	TOTAL

Na Tabela 2.1, a coluna indicadora é:

Idade
< 20 anos
20 ⊢ 30 anos
30 ⊢ 40 anos
≥ 40 anos
TOTAL

Outros elementos que devem constar na tabela, quando for o caso são: **fonte** e **nota.** A **fonte** cita o informante; as **notas** esclarecem o conteúdo e indicam a metodologia adotada na obtenção ou elaboração da informação.

O título da tabela deve ser colocado acima da tabela e a fonte, a nota e a chamada devem ficar abaixo da tabela, veja Tabela 2.1. Os títulos das tabelas e gráficos devem ser numerados em ordem cronológica.

É importante informar que no *site* do IBGE encontram-se as normas da ABNT utilizadas para a apresentação dos dados estatísticos: http://biblioteca.ibge.gov.br/visualizacao/livros/liv23907.pdf

2.1.3. Distribuição de frequências

O banco de dados com grande quantidade de observações é cansativo e não dá uma visão rápida da situação em estudo. Então, surge a necessidade de organizar os dados em uma tabela, que é chamada de **distribuição de frequências**.

Considere um conjunto de dados simples para a idade de 10 alunos: 23, 34, 28, 23, 25, 32, 24, 38, 23, 32. Essas idades podem ser agrupadas de duas maneiras diferentes. Em uma **Distribuição de frequências para dados agrupados**, temos a Tabela 2.2:

Tabela 2.2 Distribuição de frequências para a idade, em anos

Idade (X_i)	Frequência (f_i)
23	3
24	1
25	1
28	1
32	2
34	1
38	1
Soma	10

Onde f_i: **frequência absoluta** (número de vezes que o elemento aparece na amostra)

n= 10 é o número de elementos do conjunto de dados

Toda tabela deve ser seguida de uma interpretação. Por exemplo, de acordo com a Tabela 2.2, três alunos possuem 23 anos.

Podemos, também, obter **Distribuição de frequências para dados agrupados em classes** para a idade dos alunos, veja a tabela 2.3:

Tabela 2.3 Distribuição de frequências para a idade, em anos

Idade (X_i)	Frequência (f_i)
23 ⊢ 28	5
28 ⊢ 33	3
33 ⊢⊣ 38	2
Soma	10

⊢ : indica intervalo fechado à esquerda

⊢⊣ indica intervalo fechado à esquerda e à direita

Interpretação: cinco alunos possuem idade de 23 a 27,99 anos (pois o intervalo da primeira classe não inclui valor 28), três alunos possuem de 28 a 32,99 anos e dois alunos possuem de 33 a 38 anos.

Quando o conjunto de dados for numérico, não há um regra para escolher entre a formação de grupos ou de classes. Teoricamente os dois são corretos. No entanto, é de bom senso escolher aquele que resume mais os dados. No exemplo anterior, observa-se pouca repetição de valores (idades). Veja que, na Tabela 2.2 temos sete grupos e na Tabela 2.3 temos três classes. Portanto, a distribuição de frequências para dados agrupados em classes é mais adequada, pois resume mais os dados.

Quando o conjunto de dados for numérico e houver pouca repetição de valores, recomenda-se o agrupamento dos dados em classes, conforme Tabela 2.3.

Quando o conjunto de dados for numérico e houver muita repetição de valores , recomenda-se a formação de grupos, conforme Tabela 2.2.

A distribuição de frequências pode ser apresentada em termos de frequência relativa, em que : $F_i = \frac{f_i}{n}$

Tabela 2.4 Distribuição de frequências para a idade, em anos

Idade (X_i)	Frequência (f_i)
23 ⊢ 28	0,5 ou 50%
28 ⊢ 33	0,3 ou 30%
33 ⊢ 38	0,2 ou 20%
Soma	1 ou 100%

Interpreta-se que 50% dos alunos possuem idade de 23 a 27,99 anos, 30% dos alunos possuem de 28 a 32,99 anos e 20% dos alunos possuem de 33 a 38 anos.

Procedimento para a construção de uma distribuição de frequências em classes

1. **Dados brutos:** Considere a idade de 25 pessoas.

32, 26, 39, 25, 28, 16, 32, 26, 21, 28, 30, 33, 21, 15, 26, 40, 25, 39, 22, 21, 18, 34, 23, 15, 27.

2. **Rol**: é o arranjo dos dados em ordem crescente.

15, 15, 16, 18, 21, 21, 21, 22, 23, 25, 25, 26, 26, 26, 27, 28, 28, 30, 32, 32, 33, 34, 39, 39, 40.

3. **Amplitude (R):** é a diferença entre o maior e o menor valor observado.

Logo, $\mathbf{R = X_{máx} - X_{mín}}$ e $R = 40 - 15 = 25$.

4. **Número de classes (*k*):** há várias possibilidades para o cálculo do número de classes, apresentadas por Milone(2004). O **critério da raiz** pode ser usado quando $n \leq 400$. E o **critério de Sturges** pode ser usado para $16 \leq n \leq 573.23$. Como, $n = 25$, logo, pelo critério da raiz , $k = \sqrt{25} = 5$ classes.

5. **Amplitude das classes (*h*):** é dada pela razão entre a amplitude total (R) e o número de classes (k), ou seja, $h = R \div k$. Logo, $h = 25 \div 5 = 5$ **.**

6. Distribuição de frequências

Tabela 2.5 Distribuição de frequências para a idade, em anos

Idade (X_i)	Frequência (f_i)
15 ⊢ 20	4
20 ⊢ 25	5
25 ⊢ 30	8
30 ⊢ 35	5
35 ⊢ 40	3
Soma	25

IV - ATIVIDADES

Apresente uma interpretação para a Tabela 2.5.

Além de resumir dados numéricos, a distribuição de frequências dá uma idéia da distribuição dos dados. Através da distribuição de frequencias, é possível ver se há concentração de valores no meio ou nas extremidades ou se não há concentração de valores.

2.1.4. Análise gráfica

O gráfico é uma representação visual do comportamento dos dados. A partir da distribuição de frequência é possível obter o **histograma,** o **polígono de frequências** e o **gráfico de frequências acumulada.**

a) Histograma (Classes f_i)**:** é a representação gráfica de uma distribuição de frequência por meio de retângulos justapostos.

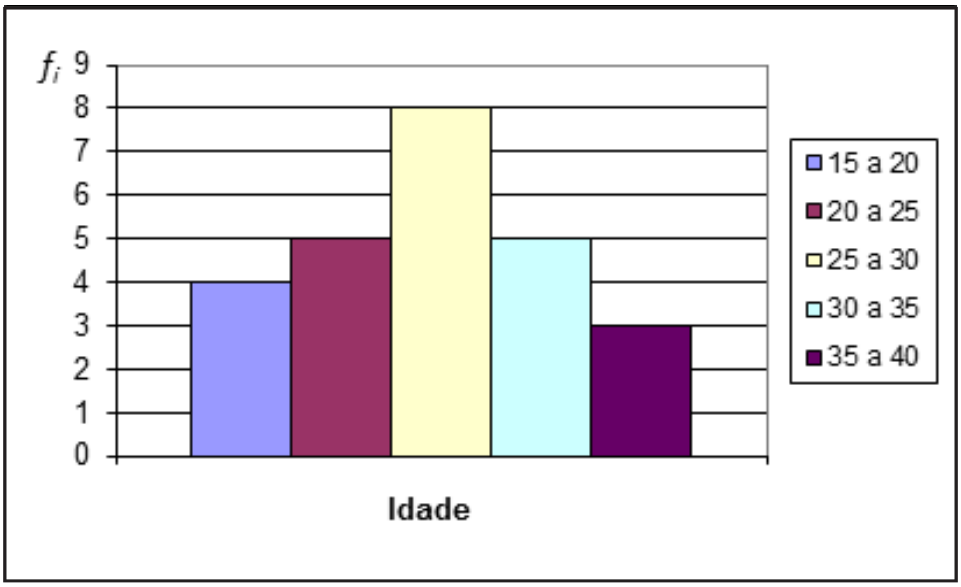


Gráfico 2.1 Histograma para a idade, em anos

Há mais de uma interpretação possível para o histograma.

Interpretação 1 : Quatro pessoas com idade de 15 a 19,99 anos, 5 pessoas com idade de 20 a 24,99 anos, 8 pessoas com idade de 25 a 29,99 anos, 5 pessoas com idade de 30 a 34,99 anos e 3 pessoas com idade de 35 a 40 anos.

O pesquisador pode interpretar todas as classes do histograma ou apenas a parte que julgar mais importante.

Interpretação 2: Apenas 3 pessoas tem 35 anos ou mais. Ou 22 pessoas têm menos de 35 anos.

O histograma tem a mesma função que a distribuição de frequência. Ele também dá uma idéia da distribuição dos dados. Através do histograma, é possível ver se há concentração de valores no meio ou nas extremidades ou se não há concentração de valores.

b) Polígono de frequências (M_i X f_i): é a representação gráfica de uma distribuição por meio de um polígono.

Seja M_i o ponto médio das classes (M_i) tal que M_i = é a média aritmética entre o limite

inferior($L_{inferior}$) e o limite superior ($L_{superior}$)da classe. Logo,

$$M_i = \frac{L_{inferior} + L_{superior}}{2}$$

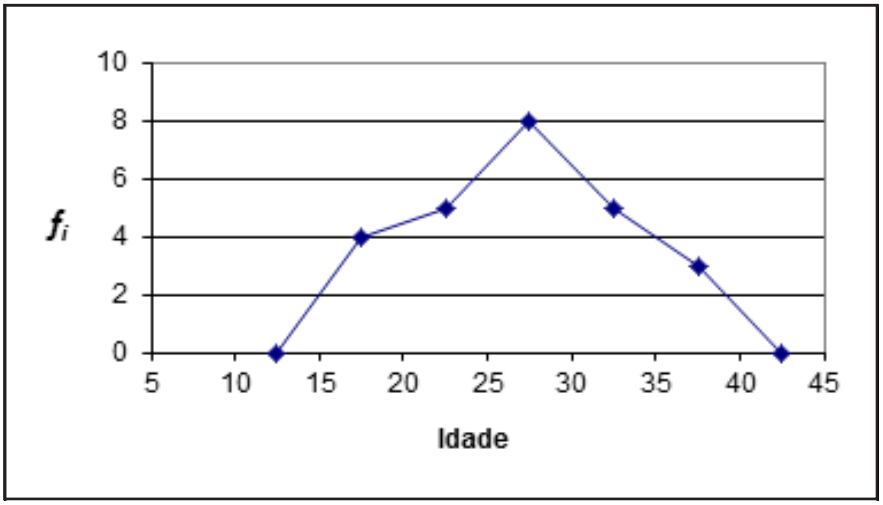


Gráfico 2.2 Polígono de frequências para a idade, em anos

No polígono de frequências, cada classe é representada pelo respectivo ponto médio.

Como temos 5 classes, então, temos 5 pontos suspensos. Para fechar o polígono de frequências, as linhas do gráfico devem tocar o eixo horizontal, passando pelo ponto médio da classe vizinha da primeira e da última classe.

Interpretação: Quatro pessoas com idade de 15 a 19,99 anos, 5 pessoas com idade de 20 a 24,99 anos, 8 pessoas com idade de 25 a 29,99 anos, 5 pessoas com idade de 30 a 34,99 anos e 3 pessoas com idade de 35 a 40 anos.

O polígono de frequências tem a mesma função que a distribuição de frequência e que o histograma. Ele também dá uma idéia da distribuição dos dados. Através do polígono de frequências, é possível ver se há concentração de valores no meio ou nas extremidades ou se não há concentração de valores.

c) Ogiva ou gráfico de frequência acumulada: ($classes \times fac$) ou ($classes \times Fac$)

Outras frequências podem ser obtidas a partir da frequência absoluta. Veja na Tabela 2.6, a *fac* é a frequência absoluta acumulada, a F_i é a frequência relativa ($F_i = \frac{f_i}{n}$) e a

Fac é a frequência relativa acumulada.

Na coluna fac, podemos dizer que 4 é o número de pessoas com menos de 20 anos, 9 é número de pessoas com menos de 25 anos e, assim, sucessivamente.

Na coluna Fac, podemos dizer que 16% das pessoas possuem menos de vinte anos, 36% das pessoas possuem menos de 25 anos, 68% das pessoas possuem menos de 30 anos e, assim, sucessivamente.

Tabela 2.6 Distribuição de frequências acumuladas para a idade, em anos

Idade (X_i)	f_i	*fac*	F_i	*Fac*
15 ⊢ 20	4	4	0,16 ou 16%	0,16
20 ⊢ 25	5	9	0,20	0,36
25 ⊢ 30	8	17	0,32	0,68
30 ⊢ 35	5	22	0,20	0,88
35 ⊢ 40	3	25	0,12	1
Soma	25		1	

A partir da frequência acumulada, podemos obter a ogiva. Veja o Gráfico 2.3.

A ogiva não mostra a distribuição dos dados. A ogiva mostra a frequência acumulada até um valor fixo do conjunto de dados.

A construção da ogiva tem início no limite inferior da primeira classe com frequência acumulada igual a zero.

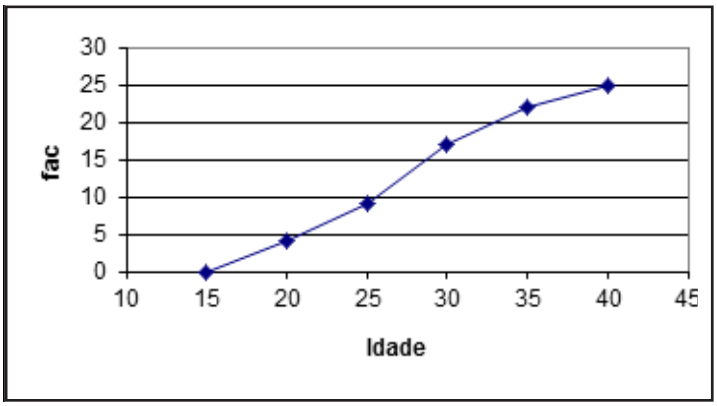


Gráfico 2.3 Ogiva para a idade, em anos

Intepretação 1: Nenhuma pessoa possui menos de 15 anos, 4 pessoas possuem menos de 20 anos, 9 pessoas possuem menos de 25 anos, 17 pessoas possuem menos de 30 anos, 22 pessoas possuem menos de 35 anos e 25 pessoas possuem menos de 40 anos.

Podem ser feitas interpretações parciais da ogiva, de acordo com o interesse do pesquisado.

Interpretação 2: "22 pessoas possuem menos de 35 anos" ou, equivalentemente, podemos dizer que "3 pessoas com 35 anos ou mais".

d) Ramo-e-folhas

O gráfico de ramo-e-folhas tem a mesma função que o histograma, ou seja mostra a distribuição dos dados.

O gráfico ramo-e-folhas é indicado quando a quantidade de dados for pequena.

A construção do ramo-e-folhas é mais simples do que a do histograma.

Procedimento para a construção do ramo-e-folhas:

1°) Identificar o menor e o maior valor ($X_{mín}$ e $X_{máx}$)

2°) Determinar uma escala, que depende da magnitude dos dados.

Exemplo: 2 , 5, 10, 20, 100, 200, 500, 1000....

3°) Acrescentar à direita da barra vertical o último algarismo de cada observação.

Não é necessário colocar os dados em ordem crescente.

Para o mesmo conjunto de dados, há mais de um ramo-e-folha possível.

Exemplo: 65<u>4</u>, 87<u>6</u>, 76<u>1</u>, 138<u>5</u>, 69<u>5</u>, 90<u>4</u>, 65<u>0</u>, 63<u>7</u>, 60<u>2</u>.

1°) $X_{mín}$ = 602 é $X_{máx}$ = 1385

2°) Escala: 100

6	4 5 0 7 2
7	1
8	6
9	4
10	
11	
12	
13	5

A folha 6 contabiliza os valores do intervalo 600 ⊢ 699 , a folha 7 contabiliza os valores do intervalo 700 ⊢ 799 e , assim, sucessivamente.

Interpretação: Há sete valores menores que 900 no conjunto de dados.

De acordo com o ramo-e-folhas, podemos observar que o conjunto de dados possui concentração de valores na extremidade inferior do intervalo de dados e há um valor atípico (muito distante dos demais).

Exemplo: 74, 78, 79, 81, 85, 86, 87, 87, 90, 91, 95.

1°) $X_{mín} = 74$ é $X_{máx} = 95$

2°) Escala: 10

3°)

7	4 8 9
8	1 5 6 7 7
9	0 1 5

A folha 7 contabiliza os valores do intervalo 70 ⊢ 79, a folha 8 contabiliza os valores do intervalo 80 ⊢ 89 e, assim, sucessivamente.

Interpretação: Há três valores menores que 80 no conjunto de dados.

Além disso, o ramo-e-folhas mostra concentração de valores na parte central do intervalo de dados.

Outra possibilidade:

2°) Escala: 5

3°)

7	4
7	8 9
8	1
8	5 6 7 7
9	0 1
9	5

A primeira folha 7 contabiliza os valores do intervalo 70 ⊢ 74 , a segunda folha 7 contabiliza os valores do intervalo 75 ⊢ 79 e, assim, sucessivamente.

Interpretação: Há três valores menores que 80 no conjunto de dados.

Além disso, o ramo-e-folhas mostra concentração de valores na parte central do intervalo de dados.

Os dois ramo-e-folhas do exemplo anterior mostram os dados em uma escala razoável. Expandindo-se a escala ainda mais do que na segunda possibilidade, os dados ficariam muito espalhados, prejudicando a visualização da concentração dos dados.

e) Gráficos de barras, gráfico de colunas e gráfico de setores

São usados para representar dados agrupados, quando a variável é qualitativa ou quando a variável é quantitativa discreta com repetição de valores.

Tabela 2.7: Suicidas segundo o sexo, Brasil, 1986

Sexo	Frequência (f_i)	Frequência relativa (F_i)
Masc	3562	0,75 ou 75%
Fem	1192	0,25 ou 25%
	4754	

O gráfico em colunas (Gráfico 2.4), o gráfico em barras (Gráfico 2.5) e o gráfico de setores (Gráfico 2.6) correspondem à distribuição de frequências da Tabela 2.7.

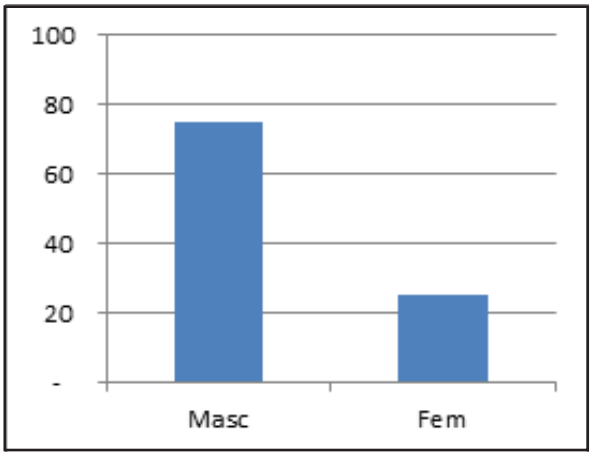


Gráfico 2.4 Porcentagem de suicidas

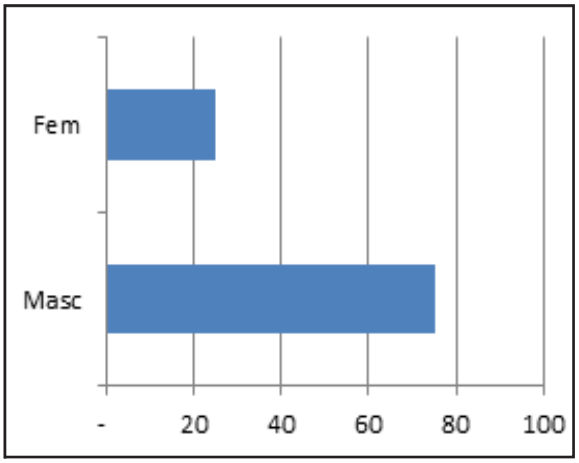


Gráfico 2.5 Porcentagem de suicidas

O processo de construção de um gráfico de setores (ou pizza) é simples, pois o setor de circunferência formado por um ângulo de 360° equivale a 100% da área da circunferência;assim para obter-se o setor cuja área representa uma determinada frequência, basta resolver uma regra de três simples, como apresentada a seguir:

360°	→	Total
x°	→	f_i

Como x° = 360*(f_i /Total),

Logo,

$$x° = 360 * F_i$$

Portanto, para a tabela acima, teremos:

x = 90° para suicidas do sexo feminino e x = 270° para suicidas do sexo masculino. Veja no Gráfico 2.6

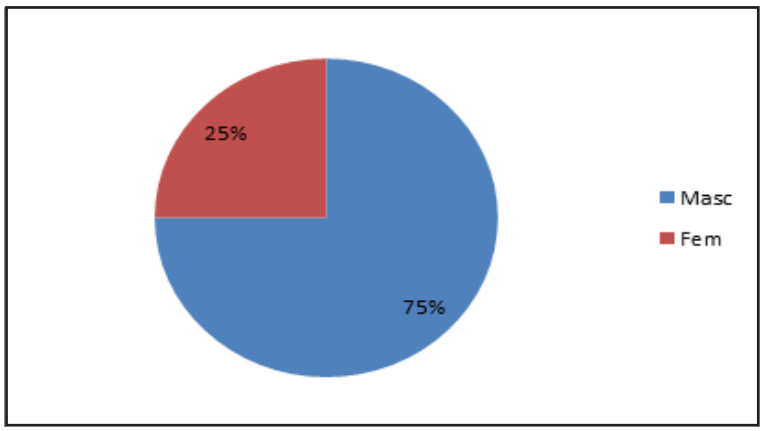


Gráfico 2.6 Porcentagem de suicidas

Interpretação: De acordo com os Gráficos 2.4, 2.5 e 2.6, 75% do suicidas são do sexo masculino e 25% são do sexo feminino.

f) Gráfico de linhas

Mostram o comportamento de um fenômeno em certo intervalo de tempo. Tais conjuntos de dados constituem as chamadas séries históricas.

Veja na tabela a seguir as taxas de mortalidade por câncer (mortes por 100.000 pessoas) nos EUA durante o período de 20 anos, de 1940 a 1960.

Tabela 2.8: Taxa de mortalidade por câncer (mortes por 100 mil pessoas)

Ano	1940	1945	1950	1955	1960
Mortes	120,3	134,0	139,8	146,5	149,2

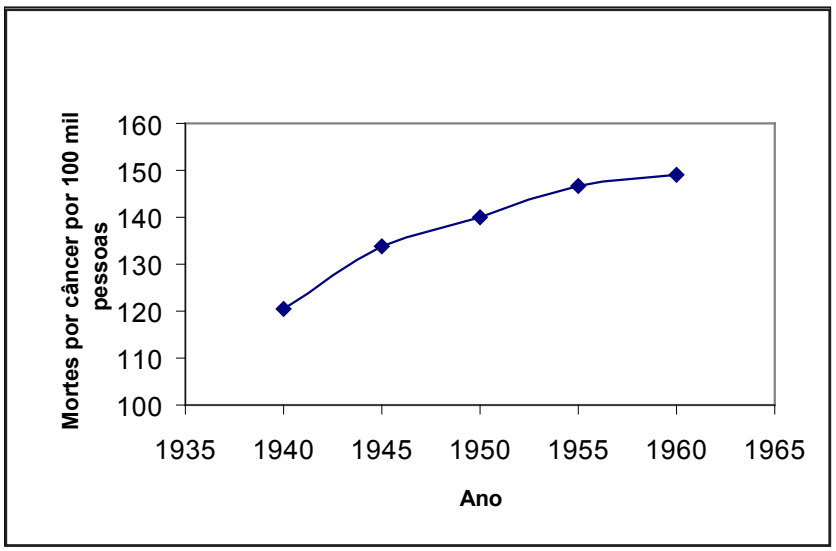


Gráfico 2.7: Taxa de mortalidade por câncer

O gráfico de linhas (Gráfico 2.7) mostra o crescimento acentuado na taxa de mortalidade por câncer, durante o período em estudo.

2.2 Medidas estatísticas

2.2.1 Medidas de tendência central: média aritmética, mediana, moda.

a. **Média aritmética simples** ou **Média** ($\bar{X}$): é a medida de tendência central mais importante.

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

Exemplo: 5, 7, 8, 8, 14. Então, n = 5 elementos, $\bar{X} = \frac{5+7+8+8+14}{5} = \frac{42}{5}$ = 8,4.

b. **Moda** (Mo): é o valor mais frequente em um conjunto de dados. A moda fornece rapidamente uma medida de tendência central.

Exemplo: 5, 7, 8, 8, 14. A Mo = 8. O conjunto de dados é **unimodal**.

Exemplo: 2, 5, 5, 6, 7, 7, 8. Mo = 5 e 7. O conjunto de dados é **bimodal**.

Exemplo: 2, 4, 6. Todos os elementos tem a mesma frequência. O conjunto de dados não tem moda, ele é **amodal**.

Exemplo: 2, 2, 4, 4, 6, 6. Todos os elementos tem a mesma frequência. O conjunto de dados é **amodal**.

c. **Mediana** (Md): é o elemento que ocupa a posição central de um conjunto de dados, uma vez que os elementos foram colocados em ordem crescente.

Procedimento para calcular a mediana:

1º) Colocar os dados em ordem crescente.

2º) Calcular a posição da mediana : P_{Md} = 0,5 (n +1).

3º) A mediana é o elemento da posição mediana: Md= $X_{P_{Md}}$.

Se n é ímpar, a mediana é o elemento que está no centro do conjunto de dados; se n é par, a mediana é dada pela média aritmética dos elementos centrais.

Exemplo: (*n* é impar)

1º) Dados em ordem crescente: 5, 7, 8, 8, 14, *n* = 5 elementos.

2º) Calcular a posição da mediana : $P_{Md} = 0{,}5\ (5 + 1) = 3$.

3º) A mediana é o elemento da 3ª posição: Md= X_3= 8.

Exemplo: (*n* é par)

1º) Dados em ordem crescente: 5, 7, 8, 9, 9, 14, *n* = 6 elementos.

2º) Calcular a posição da mediana : $P_{Md} = 0{,}5\ (6 + 1) = 3{,}5$.

3º) Então, a estimativa da mediana depende dos elementos da 3ª e 4ª posições:

Md = $\frac{X_3+X_4}{2} = \frac{8+9}{2}$ = 8,5.

Exemplo: Dados os registros diários das temperaturas diurnas da cidade XY: 0, 9, 9, 10

a) Calcular a média, a moda e a mediana das temperaturas. Interpretar o resultado.

b) Qual(is) medida(s) é(são) adequada(s) para representar a temperatura da cidade?

Resolução:

a) $\bar{X} = \frac{28}{4}$ = 7. **Interpretação**: A temperatura média da cidade XY é 7 ºC.

Mo = 9. **Interpretação**: De acordo com a moda, a temperatura da cidade XY é 9 ºC.

b) $Md = \frac{X_2+X_3}{2} = \frac{9+9}{2}$ = 2. **Interpretação**: De acordo com a mediana, a temperatura da cidade XY é 9 ºC.

A temperatura da cidade XY é mais bem representada pela moda ou mediana. Pois o valor zero é um **valor discrepante** (ou **atípico**). Os valores discrepantes subestimam ou superestimam os valores da média.

Não há regra fixa para se identificar a medida de tendência central apropriada a cada situação. De qualquer forma, é bom saber que:

A média é adequada quando não há valores discrepantes.

A moda é adequada quando há valores repetidos no conjunto de dados.

A mediana é adequada quando a distribuição dos dados for assimétrica (há concentração de valores nas extremidades).

2.2.2 Outras medidas de tendência central

a) Média ponderada ($\bar{X}_P$)

A média ponderada é usada quando as variáveis têm importâncias diferentes.

Por definição, a média ponderada dos números x_1, x_2, ..., x_n, com pesos p_1, p_2, ..., p_n é:

$$\bar{X}_p = \frac{x_1p_1 + x_2p_2 + \cdots + x_np_n}{p_1 + p_2 + \cdots + p_n}$$

Exemplo: A nota final na disciplina Português depende das notas da prova de gramática (peso 6) e da redação (peso 4). Qual é a nota final do aluno que teve nota 8 na prova de gramática e nota 10 na redação.

$$\bar{X}_p = \frac{8 * 6 + 10 * 4}{4 + 6} = 8{,}8$$

b) Média geométrica ($\bar{X}_G$) e Média harmônica ($\bar{X}_H$)

A média geométrica e a média harmônica têm aplicação na área de economia, sendo usadas nos cálculos de índices econômicos.

Por definição, a **média geométrica** dos números x_1, x_2, ..., x_n, com as respectivas frequências absolutas F_1, F_2, ..., F_n é:

$$\bar{X}_G = \sqrt[n]{x_1^{F_1} \cdot x_2^{F_2} \cdots x_n^{F_n}}$$

Exemplo: Encontrar a média geométrica de {3, 6, 12}.

Aqui $n = 3$, $F_1 = F_2 = F_3 = 1$ e $\bar{X}_G = \sqrt[3]{3^1 \cdot 6^1 \cdot 12^1} = 6$

A média geométrica é o centro de massa de um conjunto de dados sintetizável em uma progressão geométrica (MILONE, 2004).

Por definição, a **média harmônica** dos números x_1, x_2, ..., x_n, com as frequências absolutas

$$\bar{X}_H = \frac{n}{\frac{F_1}{x_1} + \frac{F_2}{x_2} + \cdots + \frac{F_n}{x_n}}$$

Exemplo: Encontrar a média harmônica de {3, 4, 6, 12}.

$$\bar{X}_H = \frac{4}{\frac{1}{3} + \frac{1}{4} + \frac{1}{6} + \frac{1}{12}} = \frac{4 \cdot 12}{10} = 4{,}8$$

A média harmônica define o centro de massa de um conjunto sintetizável em uma progressão harmônica. Como a série harmônica resulta do inverso dos valores de uma progressão aritmética, é definível como o inverso da média aritmética dos inversos dos valores (MILONE, 2004).

2.2.3. Separatrizes – quartil, decil, percentil.

Separatriz (ou **quantil**) é a medida de posição que divide o conjunto de dados em partes iguais.

As separatrizes são úteis para estabelecer quantidades ou porcentagens de itens menores que certo referencial e dar ideia da assimetria da distribuição dos dados.

Embora o total de partições seja um valor arbitrário definido pelo usuário, em função de objetivos e interesses específicos, é comum dividir o conjunto de dados em duas, quatro, dez e cem partes iguais.

As principais separatrizes são: mediana, quartil, decil, percentil.

a)Mediana (Md)

A mediana também pode ser usada como separatriz. Neste caso, a mediana divide o conjunto de dados em duas partes.

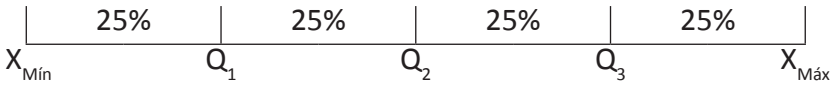
Exemplo: O conjunto de dados {5, 7, 8, 9, 9, 14}, com n = 6 elementos, possui Md = 8,5. Há várias interpretações possíveis:

1ª) 50% dos valores são menores ou iguais a 8,5 ou, equivalentemente, três valores são menores ou iguais a 8,5.

2ª) 50% dos valores são maiores do que 8,5 ou três valores são maiores que 8,5.

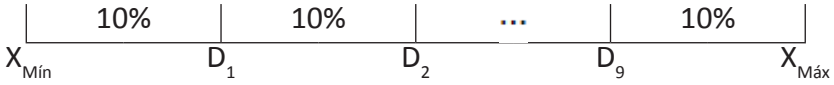
b) **Quartil (Q_i)**

Os quartis , Q_i's, com i = 1,2,3, dividem o conjunto de dados em quatro partes iguais.



c) Decil (D_i)

Os decis , D_i's, com i = 1 a 9, dividem o conjunto de dados em dez partes iguais.



d) Percentil (P_i)

Os percentis, P_i's, com i = 1 a 99, dividem o conjunto de dados em cem partes iguais.

Procedimento para cálculo das separatrizes:

1º) Colocar os dados em ordem crescente.

2º) Calcular a posição da separatriz.

Quadro 2.2: Posições das separatrizes

Quartil (Q_i)	Decil(D_i)	Percentil(P_i)
$P_{Q_1} = 0{,}25(n+1)$	$P_{D_1} = 0{,}1(n+1)$	$P_{P_1} = 0{,}01(n+1)$
$P_{Q_2} = 0{,}5(n+1)$	$P_{D_2} = 0{,}2(n+1)$	$P_{P_2} = 0{,}02(n+1)$
$P_{Q_3} = 0{,}75(n+1)$	⋮	⋮
	$P_{D_9} = 0{,}9(n+1)$	$P_{P_{99}} = 0{,}99(n+1)$

3º) A separatriz é o elemento da posição da separatriz.

$$Q_i = X_{P_{Q_i}}, \quad D_i = X_{P_{D_i}}, \quad P_i = X_{P_{P_i}}$$

Se posição da separatriz for um número fracionário, a estimativa da separatriz estará entre dois elementos. Logo, pode-se obter a separatriz calculando a média aritmética desses dois elementos.

PARE E PENSE: Na literatura, há vários métodos para calcular as separatrizes, todos fornecem valores próximos. A forma de interpreção das separatrizes não se altera.

Exercício: Obter os quartis do conjunto de dados: 27, 25, 20, 15, 30, 34, 28, 25.

1º) Em ordem crescente: 15, 20, 25, 25, 27, 28, 30, 34 $\rightarrow$ $n = 8$.

2º) Posições

1º Quartil	2º Quartil	3º Quartil
$P_{Q_1} = 0{,}25(9) = 2{,}25$	$P_{Q_2} = 0{,}5(9) = 4{,}5$	$P_{Q_3} = 0{,}75(9) = 6{,}75$

3º) Quartis

1º Quartil	2º Quartil	3º Quartil
$Q_1 = X_{2{,}25}$ $Q_1 = \frac{x_2 + x_3}{2} = \frac{20+25}{2}$ $Q_1 = 22{,}5$	$Q_2 = X_{4{,}5}$ $Q_2 = \frac{x_4 + x_5}{2} = \frac{25+27}{2}$ $Q_2 = 26$	$Q_3 = X_{6{,}75}$ $Q_3 = \frac{x_6 + x_7}{2} = \frac{28+30}{2}$ $Q_3 = 29$

Interpretação para 1º Quartil:

1ª) 25% dos valores são menores ou iguais a 22,5 ou, equivalentemente, dois valores são menores ou iguais a 22,5.

2ª) 75% dos valores são maiores do que 22,5 ou seis valores são maiores que 22,5.

Interpretação para 3º Quartil:

1ª) 75% dos valores são menores ou iguais a 29 ou, equivalentemente, seis valores são menores ou iguais a 29.

2ª) 25% dos valores são maiores do que 29 ou dois valores são maiores que 29.

Interpretação para 2º Quartil:

1ª) 50% dos valores são menores ou iguais a 26 ou, equivalentemente, quatro valores são menores ou iguais a 26.

2ª) 50% dos valores são maiores do que 26 ou quatro valores são maiores que 26.

A mediana e o 2º Quartil são iguais, ou seja, Md = Q_2.

Exercício: Obter o **6º decil** do conjunto de dados: 27, 25, 20, 15, 30, 34, 28, 25.

1º) Em ordem crescente: 15, 20, 25, 25, 27, 28, 30, 34 $\rightarrow$ $n = 8$.

2º) Posição: $P_{D_6} = 0{,}6(9) = 5{,}4$.

3º) O 6º decil é $D_6 = X_{5,4} = \frac{x_5 + x_6}{2} = \frac{27+28}{2} = 27{,}5$.

Interpretação para 6º decil:

1ª) 60% dos valores são menores ou iguais a 27,5 .

2ª) 40% dos valores são maiores do que 27,5.

Exercício: Obter o **87º percentil** do conjunto de dados: 27, 25, 20, 15, 30, 34, 28, 25.

1º) Em ordem crescente: 15, 20, 25, 25, 27, 28, 30, 34 → $n = 8$.

2º) Posição: $P_{P_{87}} = 0{,}87\ (9) = 7{,}83$

3º) O 87º percentil é $P_{87} = X_{7{,}83} = \frac{x_7 + x_8}{2} = \frac{30+34}{2} = 32$.

Interpretação para 87º percentil:

1ª) 87% dos valores são menores ou iguais a 32 .

2ª) 13% dos valores são maiores do que 32.

Mediana, 2º Quartil, 5º Decil e 50º Percentil são iguais, ou seja, Md = Q_2 = D_5 = P_{50}.

2.2.4 Medidas de dispersão

Nem sempre a medida de posição (média, mediana,...) é suficiente para descrever, de modo satisfatório, um conjunto de dados. É então, importante analisar as "distâncias" dos dados em relação à medida de posição. Essa "distância" é chamada de **dispersão** ou **variabilidade** dos dados.

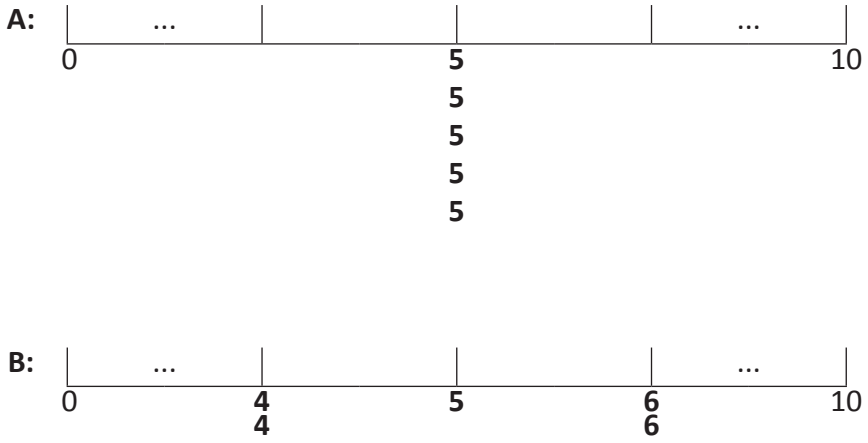
As **medidas de dispersão** medem a variabilidade dos dados, ou seja, a distância dos dados em relação à medida de posição.

Como ilustração, veja as notas das 5 provas de 4 alun

Tabela 2.9 Notas e média

	NOTAS					MÉDIA
ALUNO	P_1	P_2	P_3	P_4	P_5	$\bar{X}$
A	5	5	5	5	5	5
B	6	4	5	4	6	5
C	10	5	5	5	0	5
D	10	10	5	0	0	5

Intuitivamente, é possível observar que as notas do aluno A coincidem com a média. E que as notas do aluno B são mais distantes da média do que as notas do aluno A. Logo, as notas do aluno B possui maior variabilidade do que as notas do aluno A.



Assim sendo, surge a necessidade de definir medidas que calculem a variabilidade. As principais medidas de variabilidade são: amplitude, desvio médio, variância, desvio padrão e coeficiente de variação.

a) Amplitude (R): é a diferença entre o maior e o menor valor do conjunto de dados.

$$R = X_{máx} - X_{mín}.$$

Exemplo: Para as notas da Tabela 2.9, a amplitude das notas do aluno A é zero (R = 5 - 5 = 0), a amplitude das notas do aluno B é 2 (R = 6 - 4 = 2), a amplitude das notas dos alunos C e D é 10 (R = 10 - 0 = 10).

Quanto maior for a amplitude, maior é a variabilidade.

Interpretação: De acordo com a amplitude, as notas do aluno C têm variabilidade maior do que as notas do aluno B.

Interpretação: De acordo com a amplitude, as notas do aluno C e as notas do aluno D possuem a mesma variabilidade. **Será?**

Vantagem: A amplitude possui cálculo simples.

Desvantagem: A amplitude utiliza apenas dois valores do conjunto de dados. Portanto, há perda de informação da variabilidade.

b) **Desvio médio (DM):** é a média das diferenças dos valores do conjunto de dados em relação à sua média.

$$DM = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

O termo «diferença» dos valores em relação à média é equivalente ao termo «distância» dos valores em relação à sua média.

Exemplo: Para as notas da Tabela 2.9, o desvio médio das notas do aluno B é 0,8.

$$DM = \frac{|6-5| + |4-5| + |5-5| + |4-5| + |6-5|}{5}$$

$$DM = \frac{|1| + |-1| + |0| + |-1| + |1|}{5} = \frac{4}{5} = 0{,}8$$

Os desvios médios das notas dos alunos A, C e D estão na Tabela 2.10.

Tabela 2.10 Notas, média e medidas de dispersão

	NOTAS					MÉDIA	MEDIDAS DE DISPERSÃO				
ALUNO	P_1	P_2	P_3	P_4	P_5	$\bar{X}$	R	DM	S^2	S	CV
A	5	5	5	5	5	5	0	0	0	0	0
B	6	4	5	4	6	5	2	0,8	1	1	0,2
C	10	5	5	5	0	5	10	2	12,5	3,54	0,7
D	10	10	5	0	0	5	10	4	25	5	1

Quanto maior for o desvio médio, maior é a variabilidade.

Interpretação: De acordo com o desvio médio na Tabela 2.10, as notas do aluno C têm variabilidade maior do que as notas do aluno B.

Interpretação: De acordo com o desvio médio, na Tabela 2.10, as notas do aluno D têm variabilidade maior do que as notas do aluno C.

Vantagem: O desvio médio utiliza todos os valores do conjunto de dados. Portanto, não há perda de informação da variabilidade.

Desvantagem: O desvio médio não é muito utilizado na prática.

c) **Variância (S^2):** é a média dos quadrados das diferenças dos valores do conjunto de dados em relação à sua média.

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

Por motivos associados à teoria estatística, é usual utilizar (n – 1) em vez de n no denominador da equação acima.

Então, a equação que usaremos para calcular a variância é:

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

Essa é a equação utilizada pelas calculadoras e pelos programas estatísticos.

Se n for grande, $\sigma^2 = S^2$.

Exemplo: Para as notas da Tabela 29, a variância das notas do aluno B é.

$$S^2 = \frac{(6-5)^2 + (4-5)^2 + (5-5)^2 + (4-5)^2 + (6-5)^2}{5-1}$$

$$S^2 = \frac{(1)^2 + (-1)^2 + (0)^2 + (-1)^2 + (1)^2}{4} = 1$$

As variâncias das notas dos alunos A, C e D estão na tabela 2.10.

Quanto maior for a variância, maior é a variabilidade.

Interpretação: De acordo com a variância, na Tabela 2.10, as notas do aluno C têm variabilidade maior do que as notas do aluno B.

Interpretação: De acordo com a variância, na Tabela 2.10, as notas do aluno D têm variabilidade maior do que as notas do aluno C.

Vantagem: A variância utiliza todos os valores do conjunto de dados. Portanto, não há perda de informação da variabilidade. A variância é muito utilizada na prática.

Desvantagem: A variância possui valores muito altos, pois usa o quadrado das diferenças.

d) **Desvio padrão (S):** é raiz quadrada da variância.

$$S = \sqrt{S^2}$$

Exemplo: Para as notas da Tabela 2.9, o desvio padrão das notas do aluno B é.

$$S = \sqrt{1} = 1$$

Os desvios padrão das notas dos alunos A, C e D estão na Tabela 2.10.

PARE E PENSE: quanto maior for o desvio padrão, maior é a variabilidade.

Interpretação: De acordo com o desvio padrão, na Tabela 2.10, as notas do aluno C têm variabilidade maior do que as notas do aluno B.

Interpretação: De acordo com o desvio padrão, na Tabela 2.10, as notas do aluno D têm variabilidade maior do que as notas do aluno C.

Vantagem: O desvio padrão utiliza todos os valores do conjunto de dados. Portanto, não há perda de informação da variabilidade. O desvio padrão é muito utilizado na prática.

Desvantagem: Possui interpretação subjetiva. Em alguns casos, o pesquisador pode ficar em dúvida ao classificar a variabilidade como alta ou baixa.

e) **Coeficiente de Variação (CV):** é uma medida relativa de variabilidade, através da comparação da média com o desvio padrão.

$$CV = \frac{S}{\bar{x}} \text{ ou } CV = \frac{S}{\bar{x}} \times 100\%$$

Não há uma regra fixa para a classificação da variabilidade em função do coeficiente de variação. De qualquer forma, é razoável considerar que:

- CV < 0,3; então , a variabilidade é baixa.
- 0,3 ≤ CV ≤ 0,5; então a variabilidade é moderada
- CV > 0,5; então a variabilidade é alta.

Se CV < 0,3, então a média é representativa para o conjunto de dados.

Exemplo: Na Tabela 2.10, o coeficiente de variação das notas do aluno B é: $CV = \frac{1}{5} = 0{,}2$ ou 20%. Portanto a variabilidade das notas do aluno B é baixa.

Os coeficientes de variação das notas dos alunos A, C e D estão na tabela 2.10.

Quanto maior for o coeficiente de variação, maior é a variabilidade.

Interpretação: De acordo com o coeficiente de variação, na Tabela 2.10, as notas do aluno D tem variabilidade maior do que as notas do aluno C.

Vantagem: O coeficiente de variação utiliza todos os valores do conjunto de dados. Portanto, não há perda de informação da variabilidade. O coeficiente de variação é muito utilizado na prática.

Desvantagem: se a média for muito pequena, ela inflaciona o coeficiente de variação.

PARE E PENSE: É sempre muito importante analisar a média e uma medida de dispersão simultaneamente. A média só é representativa para um conjunto de dados que possui baixa variabilidade.

DESAFIO: Calcule o coeficiente de variação do grupo 1 e do grupo 2 e interprete.

Grupo 1: 3, 1 e 5 e **Grupo 2**: 55, 57 e 53.

Resulução no final do módulo.

A comparação da variabilidade de dois grupos com médias diferentes deve ser feita através do coeficiente de variação.

2.2.5.Medidas de assimetria e curtose

Conforme já dito anteriormente, o histograma ou a distribuição de frequências mostram a distribuição dos dados, a qual pode ser classificada em:

a) Distribuição simétrica

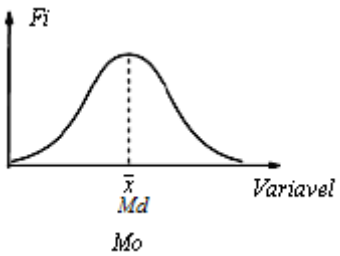


Figura 2.1: Curva de uma distribuição simétrica

Em uma distribuição simétrica, os dados estão concentrados no meio. Além disso,

$$Mo \cong Md \cong \bar{X}.$$

b) Distribuição assimétrica à direita

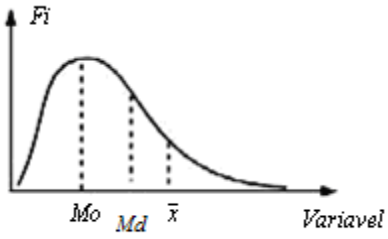


Figura 2.2: Curva de uma distribuição assimétrica à direita

Em uma distribuição assimétrica à direita, os dados estão concentrados à esquerda e poucos dados ficam à direita, formando uma cauda à direita. Além disso,

$$Mo \leq Md \leq \bar{X}.$$

C) Distribuição assimétrica à esquerda

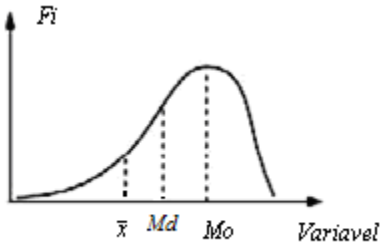


Figura 2.3: Curva de uma distribuição assimétrica à esquerda

Em uma distribuição assimétrica à direita, os dados estão concentrados à direita e poucos dados ficam à esquerda, formando uma cauda à esquerda. Além disso,

$$\bar{X} \leq Md \leq Mo.$$

d) Coeficientes de Assimetria

Permite classificar uma distribuição de dados quanto ao grau de assimetria.

Primeiro coeficiente de Pearson: quando dispomos da média, moda e desvio padrão.

$$As = \frac{\bar{X} - Mo}{S}$$

Segundo coeficiente de Pearson: usado quando conhecemos os quartis.

$$As = \frac{Q_3 + Q_1 - 2 * Q_2}{Q_3 - Q_1}$$

A interpretação dos coeficientes de Pearson é a seguinte:

Se $A_s = 0$, a distribuição é simétrica;

Se $A_s > 0$, a distribuição é assimétrica à direita;

Se $A_s < 0$, a distribuição é assimétrica à esquerda.

e) Coeficiente de curtose

A medida de curtose mede o grau de achatamento de uma distribuição, correspondendo a um indicador da forma dessa distribuição.

Esta medida quantifica a concentração ou dispersão dos valores de um conjunto de dados em relação às medidas de tendência central.

Uma distribuição é classificada quanto ao grau de achatamento como:

a) Leptocúrtica: os dados estão fortemente concentrados em torno de seu centro.

Figura 2.4: Curva de uma distribuição leptocúrtica

b) Mesocúrtica (normal): os dados estão razoavelmente concentrados em torno do seu centro.

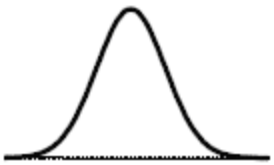


Figura 2.5: Curva de uma distribuição mesocúrtica

c) Platicúrtica: os dados estão bastante espalhados, afastando-se do centro.

Figura 2.6: Curva de uma distribuição platicúrtica

Esses gráficos podem ser comparados com:

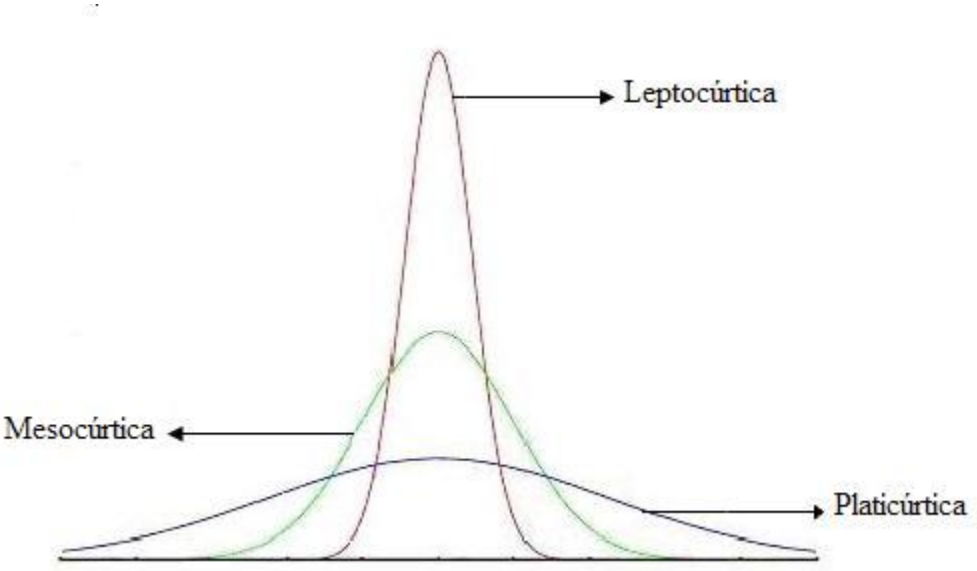


Figura 2.7: Comparação entre as curvas das distribuições mesocúrtica, leptocúrtica e platicúrtica

O **coeficiente de curtose** é dado por:

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

onde P_{90} é o 90º percentil e P_{10} é o 10º percentil.

A interpretação do coeficiente de curtose é:

Se $K = 0{,}263$, a distribuição é mesocúrtica;

Se $K > 0{,}263$, a distribuição é platicúrtica;

Se $K < 0{,}263$, a distribuição é leptocúrtica.

2.2.6 Box-plot (diagrama de caixas)

O Box-Plot é um gráfico que utiliza cinco números estatísticos: $x_{mín}$, Q_1, Md, Q_3, $x_{máx}$.

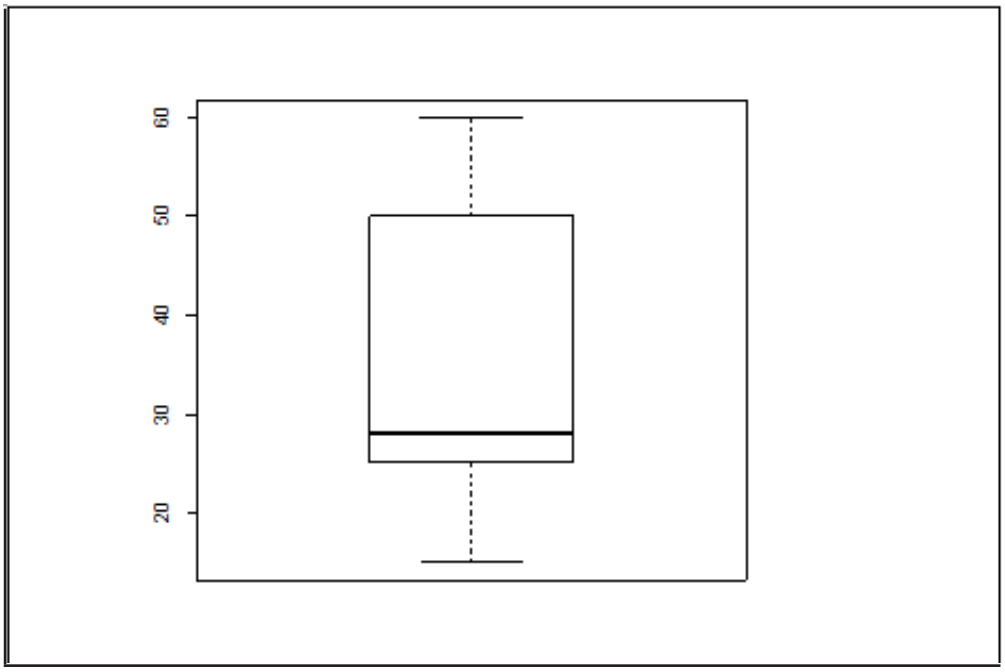


Figura 2.8 Box-plot

Na Figura 2.8 , $x_{mín} = x_1$ e $x_{máx} = x_n$.

O Box-Plot (diagrama de caixas) é uma representação gráfica abrangente, que mostra simultaneamente várias características importantes de um conjunto de dados, são elas.

1. Variabilidade: R = Xmáx - Xmin

2. Medidas de posição: Q_1, Md, Q_3

3. Assimetria

a) Distribuição simétrica

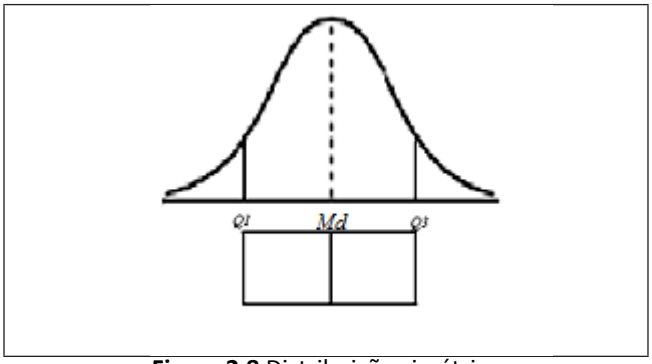


Figura 2.8 Distribuição simétrica

b) Distribuição assimétrica à direita

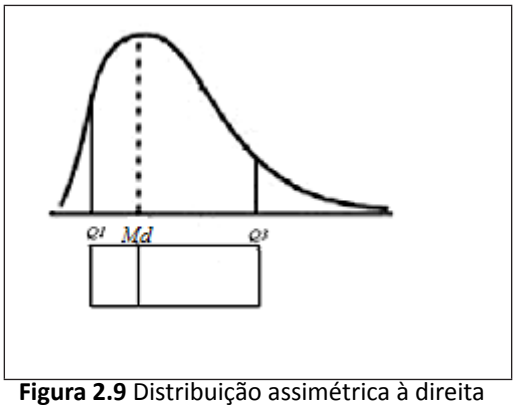


Figura 2.9 Distribuição assimétrica à direita

c) Distribuição assimétrica à esquerda

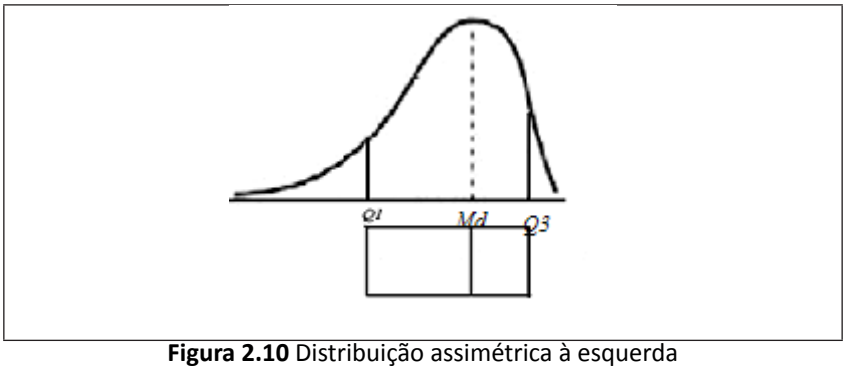


Figura 2.10 Distribuição assimétrica à esquerda

4. Valores atípicos (discrepantes ou *outliers*):

Os valores atípicos são muito distantes dos demais valores do conjunto de dados. Para a identificação desses valores, o box-plot emprega o seguinte critério:

Limite Superior: $LS = Q_3 + 1{,}5\ (Q_3 - Q_1)$

Limite inferior: $LI = Q_1 - 1{,}5\ (Q_3 - Q_1)$

Se um valor do conjunto de dados for maior que LS ou menor que LI, então ele é um valor atípico . O valor atípico é destacado no gráfico através de pontos.

Os limites LS e LI não aparecem no boxplot.

Após investigação, há três medidas possíveis para o valor atípico:

Corrigir o valor atípico. Quando forem detectadas falhas de medição ou de registros dos

valores, as quais podem ser corrigidas.

Excluir o valor atípico. Quando forem detectadas falhas de medição ou de registros dos valores, os quais NÃO podem ser corrigidos ou quando for verificado que o valor é raro e dificilmente ocorrerá novamente.

Manter o valor atípico. Quando for constatado que o valor é destoante, no entanto, não houve falha de medição é um valor com ocorrência possível

A decisão de excluir ou não um outlier do conjunto de dados deve ser amparada por justificativas de especialistas da área do tema em estudo. Nem sempre, um analista ou estatístico tem condições de tomar decisão sem as considerações desse especialista.

O box-plot é útil para a comparação estatística de dois conjuntos de dados.

Exemplo: Seja T= tempo de espera, em minutos, para atendimento em consultório médico, tal que T = (15, 20, 25, 25, 27, 28, 28, 30, 50, 50, 55, 60). O box-plot para o tempo de espera é:

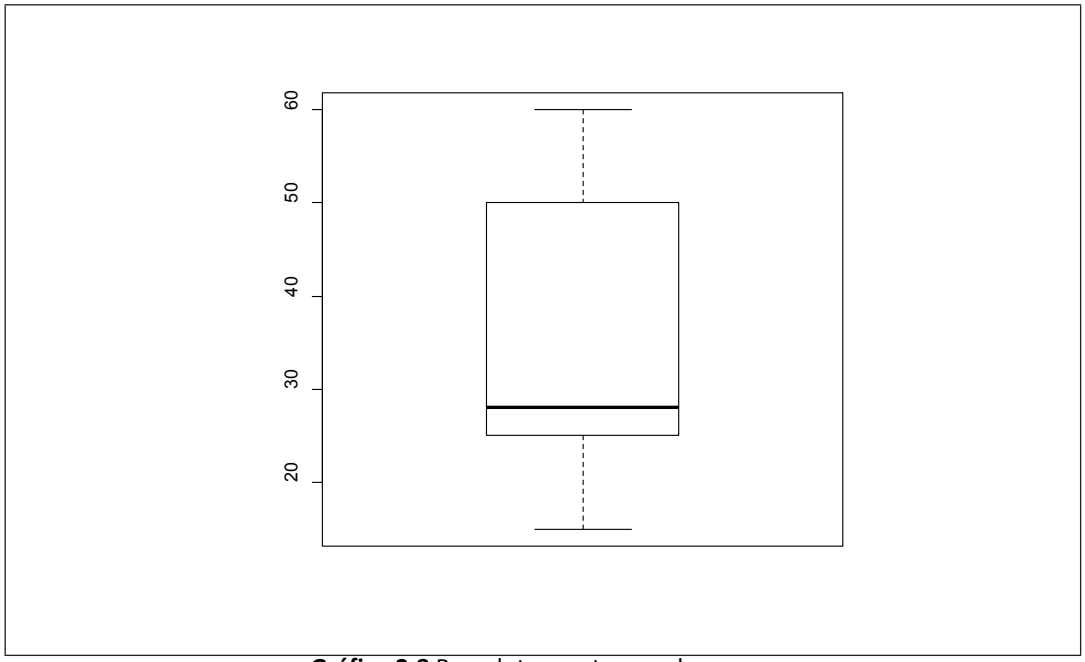


Gráfico 2.8 Box-plot para tempo de espera

Interpretação:

O tempo de espera varia de 15 a 60 minutos. Além disso, a amplitude R = 60 - 15 = 45 minutos.

Como Q_1 = 25, então, 25% dos pacientes esperaram 25 minutos ou menos pelo atendimento. Como Md = Q_2 = 28, então 50% dos pacientes esperaram 28 minutos ou menos pelo atendimento. Como Q_3 = 50, então 75% dos pacientes esperaram 50 minutos ou menos pelo atendimento.

A distribuição do tempo de espera é assimétrica à direita. Então, a mediana pode ser usada para representar o conjunto de dados. Portanto, o tempo de espera pelo atendimento do conjunto de dados é representado por 28 minutos.

Não há tempo de espera atípico no conjunto de dados analisados.

Exemplo: Seja T= tempo de espera, em minutos, para atendimento em consultório médico, tal que T = (15, 20, 25, 25, 27, 28, 28, 30, 50, 50, 55, 88). O box-plot para o tempo de espera é:

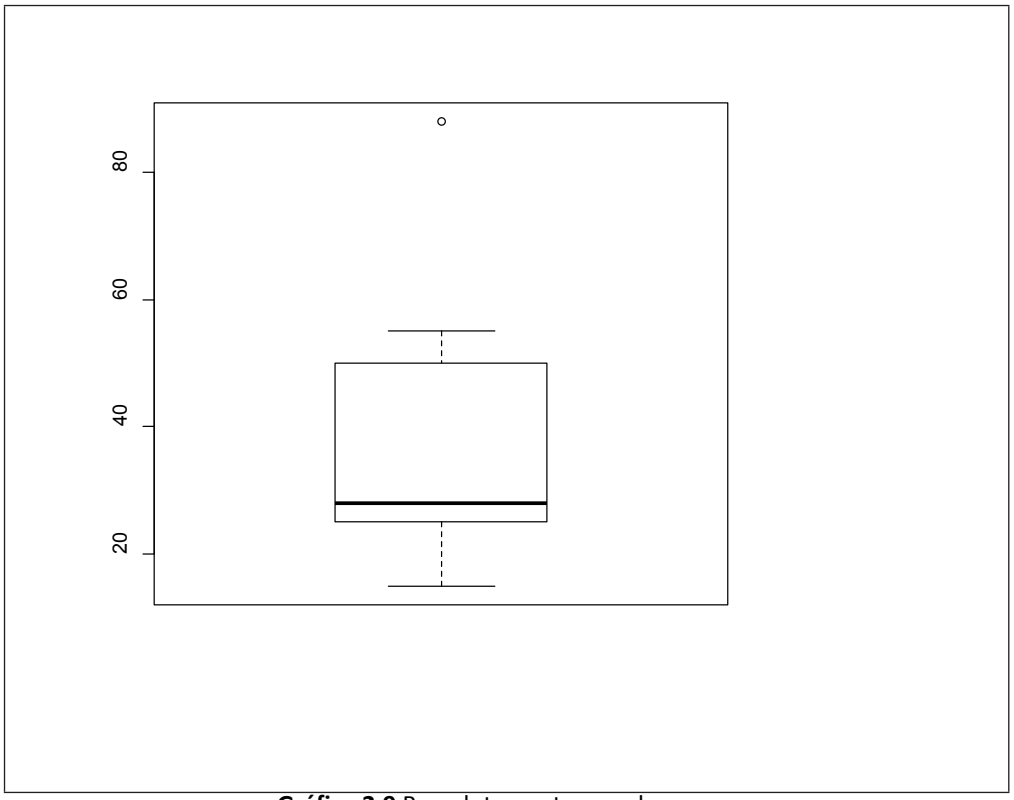


Gráfico 2.9 Box-plot para tempo de espera

Interpretação:

1. O tempo de espera varia de 15 a 88 minutos. Além disso, a amplitude R = 88 -15 = 73 minutos.

2. Como Q_1 = 25, então, 25% dos pacientes esperaram 25 minutos ou menos pelo atendimento. Como Md = Q_2 = 28, então 50% dos pacientes esperaram 28 minutos ou menos pelo atendimento. Como Q_3 = 50, então 75% dos pacientes esperaram 50 minutos ou menos pelo atendimento.

3. A distribuição do tempo de espera é assimétrica à direita. Então, a mediana pode ser usada para representar o conjunto de dados. Portanto, o tempo de espera pelo atendimento do conjunto é representado por 28 minutos.

4. O tempo de espera igual a 88 minutos é atípico e deve ser investigado.

Resolução: Para o grupo 1, a média é $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{9}{3} = 3$.

A variância é $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = \frac{(3-3)^2+(1-3)^2+(5-3)^2}{3-1} = 4$.

Logo, S = 2 e CV = $S/\bar{X}$ =2/3= 0,66 ou 66%. A variabilidade do grupo 1 é alta. Portanto, a média não é representativa.

Para o grupo 2, a média é $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{165}{3} = 55$.

A variância é $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = \frac{(55-55)^2+(57-55)^2+(53-55)^2}{3-1} = 4$.

Logo, S = 2 e CV = 2/55= 0,0364 ou 3,64%. A variabilidade do grupo 2 é baixa. Portanto, a média é representativa.

Além disso, de acordo com o coeficiente de variação, a variabilidade dos dados do grupo 1 é maior do que a do grupo 2.

V - SÍNTESE DO MÓDULO

Neste módulo, foram apresentados detalhadamente conceitos básicos e métodos estatísticos para a análise exploratória de dados. Com os métodos, grandes quantidades de dados podem ser resumidos em tabelas (distribuições de frequências), gráficos (histograma, ogiva, gráfico de colunas, gráficos de setores, gráfico de linhas, ramo-e-folhas, box-plot e outros) e medidas estatísticas (média, mediana, moda, variância, desvio-padrão, coeficiente de variação, percentil e outras). Cada método apresentado foi ilustrado com um exemplo e a respectiva interpretação. Foram discutidas as situações adequadas para o emprego, na prática, de cada método.

VI - REFERÊNCIAS

BRASIL, Secretaria de Educação Fundamental. **Parâmetros curriculares nacionais (2º ao 5º ano): Matemática**. Brasília : MEC/SEF, 1997.

BRASIL, Secretaria de Educação Fundamental. **Parâmetros curriculares nacionais (6º ao 9º ano): Matemática** . - Brasília: MEC/SEF, 1998.

BRASIL, Secretaria de Educação Fundamental. **Parâmetros curriculares nacionais (ensino médio): Matemática** . Brasília: MEC/SEF, 2000.

LOPES, Celi Espasandin; MEIRELLES, Elaine. O desenvolvimento da Probabilidade e da Estatística. **Anais do XVIII Encontro Regional de Professores de Matemática**. Campinas: LEM/IMECC/UNICAMP, 2005.

MEMÓRIA, José Maria Pompeu. **Breve história da estatística**. Brasília, DF : Embrapa Informação Tecnológica, 2004.

MILONE, Giuseppe. **Estatística Geral e Aplicada.** São Paulo: Pioneira Thomson Learning, 2004.

POUBEL, Martha Werneck . Um Estudo da História da Estatística: o 1º. Censo Demográfico. **Anais do IX Seminário Nacional de História da Matemática**, Aracaju, 2011.

MÓDULO 2

Da educação básica ao ensino superior

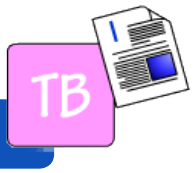
Conteúdos básicos do Módulo:

1. Conceitos básicos de probabilidade
2. Variáveis aleatórias
3. Distribuições discretas de probabilidade
4. Distribuições contínuas de probabilidade

Objetivos do Módulo:

Ao final deste estudo, esperamos que você, aluno(a), possa:

- Compreender o que é aleatoriedade
- Compreender o conceito de probabilidade e suas propriedades
- Aplicar o Teorema de Bayes
- Aplicar as distribuições discretas de probabilidade problemas reais.
- Aplicar as distribuições contínuas de probabilidade a problemas reais.



I - TEXTO BÁSICO

3. Probabilidade e variáveis aleatórias

É conveniente dispormos de uma medida que exprima a incerteza presente em afirmações do tipo “É possível que chova hoje” ou “Não há chance de ganhar o prêmio”, em termo de uma escala numérica que varie do impossível ao certo. Essa medida é a probabilidade.

3.1. Introdução à probabilidade - conceitos e propriedades

O conceito de probabilidade é fundamental para o estudo de situações em que há mais de um resultado possível, mesmo quando mantidas inalteradas as condições de sua realização. Por exemplo, jogando-se um dado, temos seis resultados possíveis. Embora não sejamos capazes de afirmar de antemão que resultado ocorrerá, temos condições de descrever o conjunto de todos os resultados possíveis.

A **aleatoriedade** está presente na natureza e se manifesta em diversas situações. Observe que os registros de temperaturas do ambiente se manifestam de forma aleatória o número de clientes, por dia, de uma loja também é aleatória.

PERGUNTA:

Que outros acontecimentos do seu cotidiano ocorrem de forma aleatória?

Antes de passarmos à definição de probabilidade, é necessário compreender os conceitos de **experimento, espaço amostral e evento.**

a) Experimento aleatório (E)

Um experimento que pode fornecer diferentes resultados, embora seja repetido toda vez da mesma maneira, é chamado **experimento aleatório.**

Alguns exemplos de experimentos aleatórios são:

E_1: jogar um dado e observar a face de cima

E_2: jogar uma moeda

E_3: retirar uma carta de um baralho

E_4: selecionar um aluno de uma turma de 30 alunos

A origem da teoria da probabilidade, no século XVII, está relacionada ao interesse em resultados de lançamentos de dados e moedas e jogos de cartas presentes nos jogos dos cassinos. Por isso, exemplos envolvento esses experimentos são comuns em livros didáticos da área de probabilidade.

b) Espaço Amostral (S)

É o conjunto de todos os resultados possíveis de um experimento.

$S_1 = \{1, 2, 3, 4, 5, 6\}$

$S_2 = \{Ca, Co\}$

$S_3 = \{52 \text{ cartas}\}$, ou seja, 52 possibilidades.

$S_4 = \{30 \text{ alunos}\}$, ou seja, 30 possibilidades.

c) Evento

É um subconjunto de um espaço amostral.

Por exemplo, considere o experimento E_1 do lançamento de um dado e o respectivo espaço amostral S_1. Logo, podemos definir o evento A como sendo a ocorrência de face par e o evento B para a ocorrência de face ímpar. Formalizando, teremos:

A = {2, 4, 6} e B = {1, 3, 5}

d) Eventos mutuamente exclusivos (eventos disjuntos)

São eventos que não ocorrem simultaneamente.

Exemplo:

E_1: lançamento de um dado

S_1= {1, 2, 3, 4, 5. 6}

A = {2, 4, 6}

B = {1, 3, 5}

C = {2}

Os eventos A e B são mutuamente exclusivos, pois face par e ímpar não podem ocorrer simultaneamente. Logo, $A \cap B = \emptyset$.

Os eventos A e C não são mutuamente exclusivos, pois a face dois ocorre no evento A e no evento C.

3.1.1. Definição de Probabilidade

a) Definição Clássica

Consideramos o caso em que se joga um dado repetidas vezes. O dado tem seis faces. Se o dado é homogêneo, equilibrado, jogando-o uma vez não há razão para dizermos que determinada face tenha preferência sobre as outras. Todos os seis resultados são igualmente possíveis. Então, a probabilidade de aparecer face 3, por exemplo, é 1/6. O evento que nos interessa consiste em “um” elemento, e o espaço amostral tem seis

elementos. A probabilidade é, pois, a relação do número de pontos do evento para o número de pontos do espaço amostral.

Ou seja, dado um espaço amostral S, se A é o evento de interesse, a probabilidade de A, representada por P (A), é dada por:

$$P(A) = \frac{Número\ de\ casos\ favoráveis\ ao\ evento\ A}{Número\ de\ casos\ possíveis}$$

A definição clássica de probabilidade se aplica quando os pontos do espaço amostral são equiprováveis.

Exemplo: Na jogada de um dado, qual a probabilidade de ocorrer face par?

Solução: P(A) = 3/6= ½.

Exemplo: Na jogada de um dado, qual a probabilidade de ocorrer face ímpar?

Solução: P(B) = 3/6= ½.

Podemos escrever o resultado da probabilidade em porcentagem: ½ = 0,5 = 50%.

b) Probabilidade e Frequência Relativa

A definição clássica de probabilidade só se aplica a espaços amostrais em que os eventos simples são igualmente possíveis. Esse é o caso da maioria das aplicações de probabilidades aos jogos de azar. Esses mesmos jogos, entretanto, repetidos inúmeras vezes, levaram a considerar a probabilidade de um evento como a frequência relativa, ou seja, como a proporção de vezes que um evento ocorre em uma série suficientemente grande de realizações de um experimento, em condições idênticas. Surgiu, então, uma nova definição de probabilidade, a definição *frequencial*.

Se A é o evento de interesse, a probabilidade de A é dada por:

$$P(A) = \frac{Número\ de\ vezes\ que\ A\ ocorreu}{Número\ total\ de\ repetições\ do\ experimento}$$

em que o número de repetições deve ser grande.

Por exemplo, se você repetir o experimento do lançamento de uma moeda 100 vezes, é esperado que ocorram aproximadamente 50 caras e 50 coroas.

Proponho que você execute essa experiência, e anote os resultados.

3.1.2. Axiomas da probabilidade

Se A e B são eventos do espaço amostral S, então P(A) e P(B), satisfazem aos axiomas:

i) $0 \leq P(A) \leq 1$

ii) $P(S) = 1$. "A probabilidade do espaço amostral é 1".

iii) Se $A \cap B = \emptyset$ (disjuntos), então $P(A \cup B) = P(A) + P(B)$.

Axiomas são afirmações que são aceitas como verdadeiras, mas não foram demonstradas.

Exemplo: Obtenha o espaço amostral para o lançamento de duas moedas.

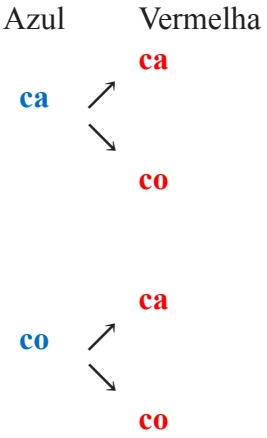
Solução:

S = {(ca,ca); (ca, co); (co,ca); (co, co)}

Para situações que envolvem lançamentos de duas ou mais moedas, considerar que as moedas são distintas, podendo ser distinguíveis por cores ou tamanhos diferentes.

Observe que (ca, co) ≠ (co, ca).

Esquematicamente, temos:

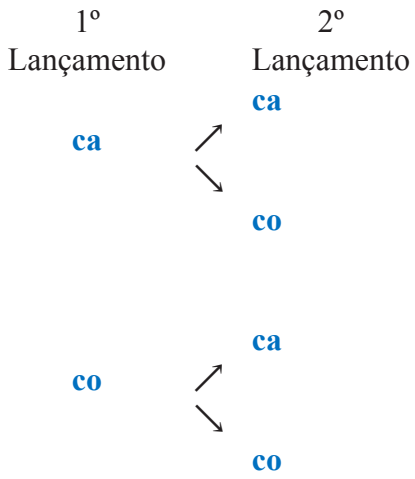


Exemplo: Obtenha o espaço amostral para uma moeda que será lançada duas vezes.

Solução:

S = {(ca,ca); (ca, co); (co,ca); (co, co)}

Esquematicamente, temos:



Exemplo: Dois dados são lançados simultaneamente.

a) Apresentar o espaço amostral para esse experimento.

b) Qual é a probabilidade de que a soma das faces seja 5?

Solução:

Os dados são distintos e o espaço amostral é:

S= {(1,1); (1,2); (1,3); (**1,4**); (1,5); (1,6);

(2,1); (2,2); (**2,3**); (2,4); (2,5); (2,6);

(3,1); (**3,2**); (3,3); (3,4); (3,5); (3,6);

(**4,1**); (4,2); (4,3); (4,4); (4,5); (4,6);

(5,1); (5,2); (5,3); (5,4); (5,5); (5,6);

(6,1); (6,2); (6,3); (6,4); (6,5); (6,6)}

Observe que (1,4) ≠ (4,1).

b) Evento A: soma das faces = 5. Logo, P (A) = 4/36.

Para situações que envolvem lançamentos de dois ou mais dados, considerar que os dados são distintos, podendo ser distinguíveis por cores ou tamanhos diferentes.

3.2. Probabilidade condicional

Seja o espaço amostral S = {1, 2, 3, 4, 5, 6}, que corresponde à face de um dado.

Pergunta 1: Qual é a probabilidade de ocorrer face 5?

Resposta: A = face 5. Então: P(A) = 1/6.

Pergunta 2: Qual é a probabilidade de sair face 5 sabendo-se que a face é ímpar?

Resposta:

1º modo: na pergunta 2, há uma informação adicional: sabe-se que a face do dado é ímpar. Essa informação reduz o espaço amostral, que passa a ser 1, 3, 5, o qual será usado para calcular a probabilidade de ocorrer face 5.

Então, A = {5} e B = {1,3,5}.

Logo, P(A|B) = 1/3.

Vemos, nesse exemplo, que a probabilidade de um evento A = {5} se modifica quando dispomos de informação sobre a ocorrência de outro evento associado, por exemplo, B = {ocorreu face ímpar}.

2º modo:

Por definição, a probabilidade do evento A, quando se sabe que o evento B ocorreu, é chamada **probabilidade condicional** de A dado B, denota-se por $P(A|B)$, e é calculada por:

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

No exemplo, temos S = {1, 2, 3, 4, 5, 6} e os eventos A, B e A∩B relacionados a ele, tal que A = {5} , B = {1,3,5} e A∩B={5}.

Então, P (B)= 3/6 e P (A∩B) = 1/6.

Logo,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{6} \cdot \frac{6}{3} = \frac{1}{3}.$$

Exemplo: Um grupo de pessoas foi classificado quanto ao peso e pressão arterial de acordo com as proporções do quadro a seguir.

Quadro 3.1 Peso e pressão arterial

	PESO			
PRESSÃO	Excesso	Normal	Deficiente	TOTAL
Alta	0,10	0,08	0,02	0,20
Normal	0,15	0,45	0,20	0,80
TOTAL	0,25	0,53	0,22	1,00

a) Qual a probabilidade de uma pessoa escolhida ao acaso nesse grupo ter pressão alta?

b) Se se verifica que a pessoa escolhida tem excesso de peso, qual a probabilidade de ela ter também pressão alta?

Esse exemplo foi extraído de Farias *et al*. (2003)

Solução:

a) Como a pessoa é escolhida ao acaso em um grupo em que 20% tem pressão alta, chamando A o evento "ter pressão alta", P (A) = 0,20 é a probabilidade pedida.

b) Chamaremos B o evento "ter excesso de peso". Ante a informação de que a pessoa tem excesso de peso, então estamos interessados na primeira coluna, P (B) = 0,25.

Além disso, o evento "ter pressão alta e excesso de peso" é o evento (A∩B) e P(A∩B) = 0,10.

Logo, podemos calcular a probabilidade condicional de A dado B.

$$P(A\,|B) = \frac{P(A \cap B)}{P(B)} = \frac{0{,}10}{0{,}25} = 0{,}40 \text{ ou } 40\%.$$

3.3. Independência de eventos

Dois eventos são independentes se a ocorrência de um deles não interfere na ocorrência do outro.

Formalizando, dados dois eventos A e B de um espaço amostral S, temos que:

$$P\,(A|B) = \mathrm{P}\,(\mathrm{A}) \iff \text{eventos A e B são indepententes}$$

Em outras palavras podemos dizer que:

A probabilidade condicional do evento A, dado o evento B, é igual à probabilidade de A se e somente se, os eventos A e B são independentes.

Exemplo: Um grupo de pessoas foi classificado quanto a peso e pressão arterial de acordo com as proporções do quadro a seguir.

Quadro 3.2 Peso e pressão arterial

	PESO			
PRESSÃO	Excesso	Normal	Deficiente	TOTAL
Alta	0,10	0,08	0,02	0,20
Normal	0,15	0,45	0,20	0,80
TOTAL	0,25	0,53	0,22	1,00

Os eventos 'excesso de peso' e 'pressão alta' são independentes?

Solução:

Chamando A o evento "ter pressão alta" e de B o evento "excesso de peso", temos que $P(A) = 0{,}20$ e $P(A|B) = 0{,}40$.

Como $P(A) \neq P(A|B)$, então A e B não são independentes.

3.4. Teorema do Produto

Da definição de probabilidade condicional, temos: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Se A e B são independentes, podemos dizer que $P(A|B) = P(A)$, a qual pode ser substituída na definição de probabilidade condicional.

Logo, $P(A) = \frac{P(A \cap B)}{P(B)}$. Portanto, $P(A \cap B) = P(A).P(B)$.

" Se A e B são eventos independentes, então $P(A \cap B) = P(A).\ P(B)$".

Exemplo: Uma moeda é jogada 2 vezes. Qual é a probabilidade de sair face cara nos dois lançamentos?

Solução:

A "face cara no segundo lançamento" independe da "face cara no primeiro lançamento".

O evento *A* "face cara no primeiro lançamento" tem probabilidade 1/2. Logo, $P(A) = ½$.

O evento *B* "face cara no segundo lançamento" tem probabilidade 1/2. Logo, $P(B) = ½$.

$P(A \cap B) = P(A).\ P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

3.5. Teorema da Soma

i) Se A e B forem disjuntos, então $P(A \cup B) = P(A) + P(B)$. (Axioma iii)

ii) Se A e B podem ocorrer ao mesmo tempo, então $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ "

Exemplo:

Uma urna possui duas bolas brancas, uma azul e uma vermelha. Qual é a probabilidade de uma bola colorida ser selecionada?

Solução:

O espaço amostral para as quatro bolas é S = {Branca1, Branca2, Azul, Vermelha}

O evento *A* "selecionar bola azul" tem probabilidade 1/4. Logo, $P(A) = 1/4$.

O evento *B* "selecionar bola vermelha" tem probabilidade 1/4. Logo, $P(B) = 1/4$.

O evento A e B são disjuntos, não podem ocorrer ao mesmo tempo.

O evento $A \cup B$ "selecionar bola azul ou vermelha", tem probabilidade

$P(A \cup B) = P(A) + P(B) = 1/4+1/4=2/4 = 1/2$.

Exemplo: Qual é a probabilidade de se extrair um ás ou uma carta de espada de um baralho?

Seja A o evento extração de um ás de um baralho e B a extração de uma carta de espada.

O evento *A* “extração de um ás” tem probabilidade 4/52. Logo, *P*(*A*) = 4/52.

O evento *B* “extração de uma espada” tem probabilidade 13/52. Logo, *P*(*B*) = 13/52.

Os eventos A e B não são disjuntos, podem ocorrer ao mesmo tempo.

Logo, *P*(A∩ B)= 1/52.

P(A ∪ B) = *P*(A) + *P*(B) - *P*(A ∩ B) =

= 4/52 + 13/52 - 1/52 =16/52= 4/13.

Exemplo: Em uma universidade, 50% dos alunos falam inglês, 20% francês e 5% os dois idiomas. Qual é a probabilidade de encontrar alunos que falem alguma língua estrangeira?

P(A) = 0,5, *P*(B)= 0,2 e *P*(A ∩ B)= 0,05

P(A ∪ B) = *P*(A) + *P*(B) - *P*(A ∩ B) =

=0,5 + 0,2 - 0,05 = 0,65.

3.6. Outros teoremas

i) Se Ø é um conjunto vazio, então *P* (Ø)= 0.

ii) Se $\bar{A}$ é o complemento do evento A, então *P*($\bar{A}$) = 1 – *P*(A).

iii) Se A ⊂ B, então *P*(A) ≤ *P*(B).

3.7. Teorema de Bayes

Sejam A_1, A_2, ..., A_n, *n* eventos mutuamente exclusivos tais que $A_1 \cap A_2 \cap \ldots \cap A_n = S$.

Sejam $P(A_i)$ as probabilidades conhecidas dos vários eventos, e *B* um evento qualquer de *S* tal que conhecemos todas as probabilidades condicionais P ($B|A_i$).

Então para cada *i*, teremos:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i \cap B)}{\sum_{i=1}^{n} P(A_i \cap B)} = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

Na equação acima, $P(B)$ é chamada de probabilidade total.

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

O Teorema de Bayes é uma generalização da probabilidade condicional.

Para ilustrar o Teorema de Bayes, consideremos o diagrama da Figura 3.1 a seguir:

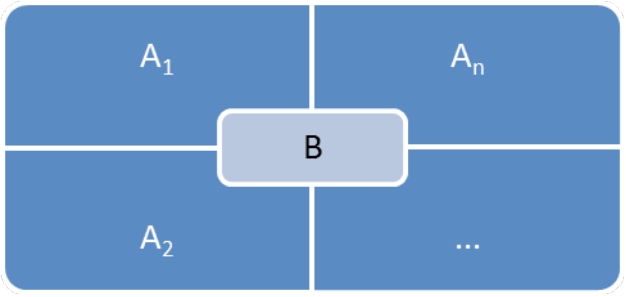


Figura 3.1: Ilustração do Teorema de Bayes.

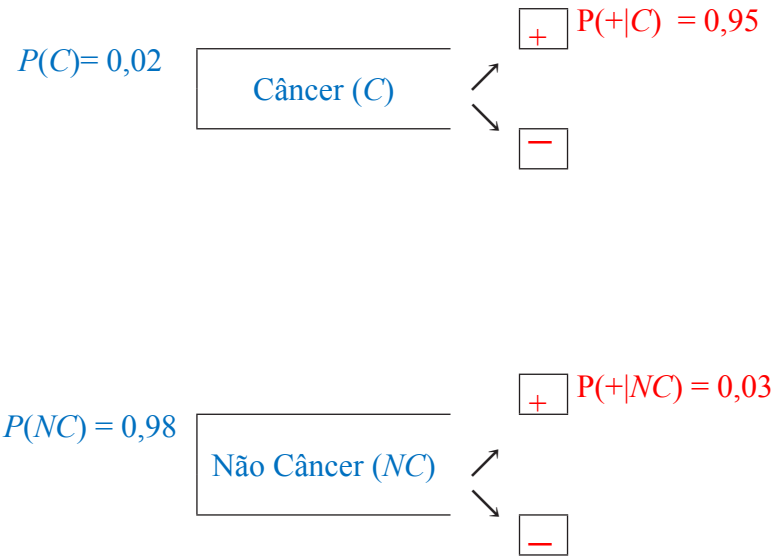
Exemplo: Suponha um teste para câncer em que 95% dos que tem a doença reagem positivamente, enquanto 3% dos que não tem a doença reagem positivamente. Suponha ainda que 2% dos pacientes de um hospital tenham câncer.

a) Qual é a probabilidade de um paciente escolhido ao acaso, e que reaja positivamente ao teste, ter de fato a doença?

Usaremos a seguinte notação:

C : evento “ter câncer”

+ : evento “ ter reação positiva”



Probabilidade procurada é:

$$P(C|+) = \frac{P(C \cap +)}{P(+)} = \frac{P(C)P(+|C)}{P(C)P(+|C) + P(NC)P(+|NC)}$$

$$P(C|+) = \frac{(0{,}02)(0{,}95)}{(0{,}02)(0{,}95) + (0{,}98)(0{,}03)}$$

$$P(C|+) = \frac{0{,}019}{0{,}048} = 0{,}396$$

b) Qual é a probabilidade de um paciente escolhido ao acaso, e que reaja negativamente ao teste, não ter de fato a doença?

Resolva este item e veja a solução no final do módulo

3.8. Variáveis aleatórias unidimensionais discretas e contínuas

Sejam *E* um experimento e *S* um espaço amostral associado ao experimento.

Por definição, uma função *X*, que associe a cada elemento $s \in S$ um número real, *X*(s), é denominada **variável aleatória** (v.a.).

Exemplo:

E = lançamento de duas moedas simultaneamente.

S = { (ca,ca), (ca, co), (co, co), (co, ca)}

X = número de caras obtidas

A cada evento simples do espaço amostral associamos um número, que é o valor assumido pela variável aleatória X.

Evento	caca	caco	coca	coco
X	2	1	1	0

Logo, temos que

P(X = 0) = 1/4 , P(X = 1) = 2/4 e P(X = 2) =1/4.

Poderíamos ter considerado a variável aleatória X = número de coroas obtidas.

Daqui pra frente, o termo variável aleatória será abreviado por *v.a*.

As *v.a*'s são usualmente representadas por letras maiúsculas e podem ser classificadas em discretas ou contínuas.

3.8.1. Variável aleatória discreta

Definição: Seja X uma variável aleatória. Se o número de valores possíveis de X for finito ou infinito numerável, denomina-se X de **variável aleatória discreta**.

A *v.a*. discreta X recebe valores inteiros.

Exemplo: X = *v.a*. número de alunos aprovados.

Então, numa turma de 15 alunos, o número de aprovados pode ser de 0 a 15.

Formalizando, X = 0, 1, 2, 3, ...,15.

a) Distribuição de probabilidade

Definição: A **distribuição de probabilidade** é o conjunto de todos os valores que podem ser assumidos pela *v.a*. discreta, com as respectivas probabilidades.

Exemplo: No lançamento de um dado, temos seis resultados possíveis e a sua distribuição de probabilidade é:

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

b) Representação gráfica de uma distribuição de probabilidade

A representação gráfica para a distribuição de probabilidade das faces de um dado é:

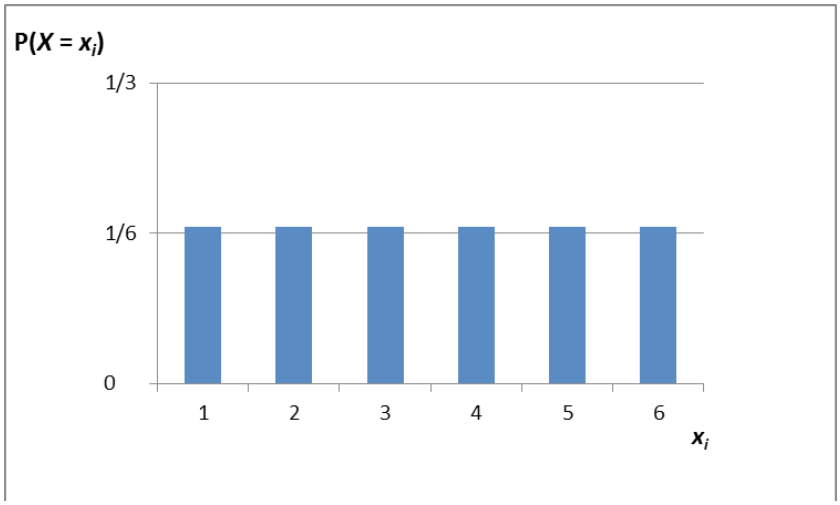


Gráfico 3.1: Representação de uma distribuição de probabilidade

c) Função de probabilidade

Definição: Para uma *v.a*. discreta X, com valores possíveis x_1, x_2, ..x_n, a função de probabilidade é $p(x_i) = P(X = x_i)$.

Seja $p(x_i)$ é definida como sendo uma probabilidade. Então, verificam-se as seguintes propriedades:

i) $p(x_i) \geq 0, \forall x_i$

ii) $\sum_{i=1}^{n} p(x_i) = 1$

Exemplo: para a face de um dado.

i) $p(x_i) = (1/6) \geq 0, \forall x_i$

ii) $\sum_{i=1}^{n} p(x_i) = p(x_1) + p(x_2) + p(x_3) + p(x_4) + p(x_5) + p(x_6)$

$$= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1$$

d) **Função de distribuição acumulada** (ou Função de distribuição)

Definição: a função de distribuição acumulada de uma *v.a.* discreta *X*, denotada por *F*(*X*), é : $F(X) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$

3.8.2. Variável aletória contínua

Definição: Uma variável aleatória que pode assumir qualquer valor numérico em um intervalo ou em uma coleção de intervalos é chamada **variável aleatória contínua.**

Portanto, a *v.a*. contínua *X* pode assumir valores contínuos (reais).

Exemplo: altura, temperatura, peso, tempo de espera numa fila.

Função densidade de probabilidade

Definição: Diz-se que *X* é uma *v.a.* contínua, se existir uma função *f*, denominada **função densidade de probabilidade** (fdp) de *X* que satisfaça às seguintes condições.

i) f(x) >=0

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

iii) para quaisquer *a*, *b*, com $-\infty < a < b < \infty$, teremos:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx.$$

Exemplo: suponha que a v.a. X seja contínua. Seja a fdp *f* dada por:

$$f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, caso\ contrário \end{cases}$$

i) Para qualquer valor no intervalo $0 < x < 1$, temos que $f(x) >= 0$, veja o gráfico 3.2.

x	$f(x)=2x$
0	2(0) = 0
0,5	2(0,5) = 1
1	2(1) = 2

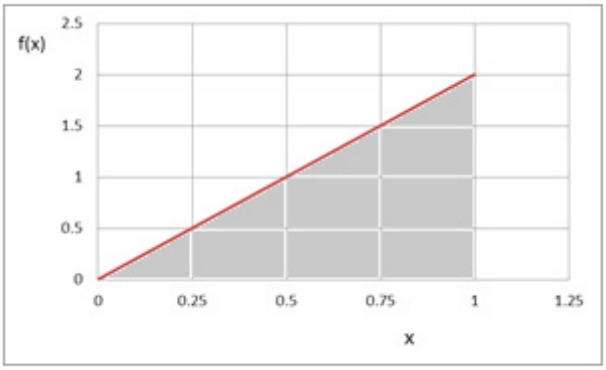


Gráfico 3.2 Gráfico da $f(x) = 2x$, $0 \leq x \leq 1$

ii) Agora, vamos calcular a integral da função no intervalo em que ela foi definida.

$$\int_0^1 2x dx = x^2]_0^1 = 1^2 - 0^2 = 1$$

Logo, $f(x) = 2x$ é uma função densidade de probabilidade.

A área sob a curva f(x) é igual a 1.

Então, a função f(x) pode ser usada para calcular probabilidade, conforme propriedade iii.

iii) Por exemplo, podemos calcular a probabilidade de $P(0 \leq X \geq 1/2)$.

Aqui $a = 0$ e $b = 1/2$. Logo, $P(0 \leq X \geq 1/2) = \int_0^{1/2} 2x dx = x^2]_0^{1/2} = (1/2)^2 - 0^2 = 1/4$.

3.9. Esperança matemática e variância de variáveis aleatórias unidimensionais

3.9.1. Esperança matemática de uma v.a

Definição: Esperança matemática (ou **média** ou **valor esperado**) de uma *v.a*. discreta é a soma de todos os produtos possíveis da *v.a*. pela respectiva probabilidade.

$$E(X) = \mu = \sum_{i=1}^{n} x_i p(x_i)$$

Exemplo: Lançamento de um dado.

x_i	$p(x_i)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

$$E(X) = \sum_{i=1}^{n} x_i p(x_i) = 1 \cdot \left(\frac{1}{6}\right) + 2.\left(\frac{1}{6}\right) + 3.\left(\frac{1}{6}\right) + 4.\left(\frac{1}{6}\right) + 5.\left(\frac{1}{6}\right) + 6.\left(\frac{1}{6}\right) = 21/6 = 3{,}5.$$

3.9.2. Propriedades da média

Sejam as variáveis aleatórias X e Y. Para as constantes a, b e c, temos as propriedades.

i) $E(c) = c$

ii) $E(cX) = c\,E(X)$

iii) $E(X \pm Y) = E(X) \pm E(Y)$

iv) $E(X \pm c) = E(X) \pm c$

v) $E\,(a + bX) = a + bE(X)$

3.9.3. Variância de um *v.a.*

A média é uma medida de posição de uma *v.a.*, então, é natural que procuremos uma medida de dispersão dessa variável em relação à média. Essa medida é a variância, denotada por σ^2 ou *Var*(*X*).

Definição: $\sigma^2 = Var\,(X) = E[(X - \mu)^2]$

Desenvolvendo, temos a equação que usaremos para calcular a variância**:**

$Var\,(X) = E(X^2) - [(E(X))^2]$, onde $E(X^2) = \sum_{i=1}^{n} x_i^2 p(x_i)$.

Exemplo: lançamento de um dado.

$E(X^2) = \sum_{i=1}^{n} x_i^2 p(x_i) = 1^2 \cdot \left(\frac{1}{6}\right) + 2^2.\left(\frac{1}{6}\right) + 3^2.\left(\frac{1}{6}\right) + 4^2.\left(\frac{1}{6}\right) + 5^2.\left(\frac{1}{6}\right) + 6^2.\left(\frac{1}{6}\right) = 91/6$

$E(X^2)$=15,167, $E(X) = 21/6 = 3{,}5$.

Logo, $Var(X) = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{546-441}{36} = \frac{105}{36} = \frac{35}{12}$.

3.9.4. Propriedades da variância

Seja a variável aleatória *X*. Para as constantes *a*, *b* e *c*, temos que:

i) $Var(c) = 0$

ii) $Var(cX) = c^2\, Var(X)$

iii) $Var(X \pm c) = Var(X)$

iv) $Var(a + bX) = b^2\, Var(X)$

3.9.5. Esperança matemática e variância de uma variável aleatória contínua

Definição: uma variável aleatória contínua X, com função densidade de probabilidade *f*(*x*), tem a esperança e variância definidas por:

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

$$Var\,(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

3.10. Distribuições discretas de probabilidade

Através da teoria da probabilidade e suas propriedades, é possível construir modelos probabilísticos (ou função de distribuições de probabilidade) que permitem calcular a probabilidade de ocorrência de eventos. Neste curso, estudaremos as principais distribuições, ou seja, distribuição uniforme discreta, a distribuição de Bernoulli, a distribuição binomial, a distribuição de Poisson, a distribuição geométrica e a distribuição hipergeométrica.

Para empregar as distribuições de probabilidade, o usuário deve identificar a variável aleatória (característica de interesse), os parâmetros conhecidos (por exemplo, média, a proporção, e outros), classificar a distribuição da variável aleatória e calcular as probabilidades

3.10.1. Distribuição uniforme discreta

Definição: se a variável aleatória *X* assume os valores $x_1, x_2, ..., x_n$ com igual probabilidade, então a **distribuição uniforme discreta** é dada por:

$$P(X = x) = \frac{1}{n}, \; x = x_1, x_2, ..., x_n$$

A média e a variância da distribuição uniforme discreta são:

$$E(X) = \mu = \frac{\sum_{i=1}^{n} x_i}{n} \text{ e } Var(X) = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}.$$

Exemplo: Quando selecionamos uma pessoa, aletoriamente, de um grupo de cinco pessoas, cada elemento do espaço amostral S = { 1, 2, 3, 4, 5} tem probabilidade 1/5. Portanto, temos uma distribuição uniforme.

$$P(X = x) = \frac{1}{5}, \; x = 1, 2, 3, 4, 5.$$

Distribuição uniforme discreta depende apenas do parâmetro *n*.

Exercício: Que outra situação prática poderia ser modelada com a distribuição uniforme discreta?

3.10.2. Distribuição de Bernoulli

A **tentativa de Bernoulli** consiste em realizar um experimento aleatório uma só vez e observar se certo evento ocorre ou não. Se o evento de interesse ocorre, ele é classificado por **sucesso**. Caso contrário, ele é classificado por **fracasso**. Além disso, o sucesso tem probabilidade *p* de ocorrência e o fracasso tem probabilidade (1 - *p*) de ocorrência, sendo *p* conhecida. Portanto, a variável aleatória *X* assume dois valores (0 e 1):

$$X = \begin{cases} 1, \text{se ocorre sucesso} \\ 0, \text{se ocorre fracasso} \end{cases}$$

Exemplos:

$$X = \begin{cases} 1, \text{ fumante} \\ 0, \text{não fumante} \end{cases}, \quad X = \begin{cases} 1, \text{cara} \\ 0, \text{coroa} \end{cases}, \quad X = \begin{cases} 1, \text{feminimo} \\ 0, \text{masculino} \end{cases}$$

Outro exemplo: O valor 1 é atribuído à característica de interesse no estudo.

Se um estudo for considerar os clientes insatisfeitos do restaurante Y, teremos:

$$X = \begin{cases} 1, \text{insatisfeito} \\ 0, \text{satisfeito} \end{cases}$$

Se o estudo for considerar os clientes satisfeitos do restaurante Y, teremos:

$$X = \begin{cases} 1, \text{satisfeito} \\ 0, \text{insatisfeito} \end{cases}$$

Definição: Se a variável aleatória *X* pode assumir os valores 1 ou 0, com probabilidade *p* e (1 - *p*) respectivamente, então a **distribuição de Bernoulli** é dada por:

$$P(X = x) = p^x(1-p)^{1-x}, \; x = 0, 1.$$

A **média** e a **variância** da distribuição de Bernoulli são:

$$E(X) = pE(X) = p \text{ e } Var(X) = p(1-p).$$

Exemplo: Seja *X* a face de uma moeda. Como $X = \begin{cases} 1, \text{cara} \\ 0, \text{coroa} \end{cases}$ e $p = 0{,}5$ é a probabilidade de ocorrência de cara e, então:

$$P(X = x) = 0{,}5^x(1-0{,}5)^{1-x}, \; x = 0, 1.$$

Da equação anterior temos que:

$P(X = 1) = 0{,}5^1(1-0{,}5)^{1-1} = 0{,}5$ é a probabilidade de ocorrer cara e

$P(X = 0) = 0{,}5^0(1-0{,}5)^{1-0} = 0{,}5$ é probabilidade de ocorrer coroa.

Distribuição de Bernoulli depende apenas do parâmetro *p*.

3.10.3. Distribuição binomial

A distribuição binomial é uma extensão da distribuição de Bernoulli. Enquanto, na distribuição de Bernoulli apenas um item (ou indivíduo) é classificado, na distribuição de Bernoulli são classificados *n* itens (ou indivíduos). Portanto, a variável aleatória X, que é o número de sucessos em *n* elementos, possui distribuição binomial.

Definição: Uma tentativa de Bernoulli pode resultar em um sucesso com probabilidade *p*, ou em um fracasso, com probabilidade (1-*p*). Então, a variável aleatória *X*, que é o número de sucessos em *n* tentativas independentes, possui **distribuição binomial** com parâmetros *n* e *p* :

$$P(X = x) = C_{n,x}p^x(1-p)^{n-x}, \text{ x = 0, 1, 2, ..., } n.$$

Em que $C_{n,x} = \frac{n!}{x!(n-x)!}$.

A **média** e a **variância** da distribuição binomial são:

$$E(X) = np \text{ e } Var(X) = np(1-p).$$

Exemplo: Suponha que 10% das pessoas sejam fumantes. Em um grupo de 4 pessoas, qual é a probabilidade de que apenas uma seja fumante?

A variável aleatória *X* é o número de fumantes, que possui distribuição binomial. Os parâmetros são *n*= 4 e *p* = 0,10.

Logo, a probabilidade de que o número de fumantes seja igual a 1, *P* (*X* = 1) , é:

$P(X = 1) = C_{4,1}(0,1)^1(0,9)^{4-1}$. Como, $C_{4,1} = \frac{4!}{1!(4-1)!} = \frac{4.3!}{3!} = 4$. Logo,

$P(X = 1) = 4(0,1)^1(0,9)^3 = 0,2916$ ou 29,16%.

Distribuição Binomial depende dos parâmetros *n* e *p*.

Exemplo: Suponha que 10% das pessoas sejam fumantes. Em um grupo de 4 pessoas,

a) qual é a probabilidade de que não haja fumantes?

b) qual é a probabilidade de que haja duas fumantes?

c) qual é a probabilidade de que haja três fumantes?

d) qual é a probabilidade de que haja quatro fumantes?

Esse exemplo é uma continuação do anterior. Logo:

X: número de fumantes.

Os parâmetros são $n = 4$ e $p = 0{,}10$ e $x = 0,1,2,3,4$.

Quadro 3.1: Exemplo da distribuição binomial

	x	$P(X = x)$	$C_{n,x}p^x(1-p)^{n-x}C_{n,x}p^x(1-p)^{n-x}$
a)	0	$P(X = 0)$	$C_{4,0}(0,1)^0(0,9)^4 = 0,6561$
	1	$P(X = 1)$	$C_{4,1}(0,1)^1(0,9)^3 = 0,2916$
b)	2	$P(X = 2)$	$C_{4,2}(0,1)^2(0,9)^2 = 0,0486$
c)	3	$P(X = 3)$	$C_{4,3}(0,1)^3(0,9)^1 = 0,0036$
d)	4	$P(X = 4)$	$C_{4,4}(0,1)^4(0,9)^0 = 0,0001$
		SOMA	**1**

PROPRIEDADES:

No **Quadro 3.1**, observamos algumas propriedades:

i) A soma das probabilidades é igual a 1, ou seja, $\sum_{i=1}^{n} p(x_i) = 1$

No exemplo anterior, $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1$,

ou seja, 0,6561 + 0,2916 + 0,0486 + 0,0036 + 0,0001 = 1

ii) A probabilidade acumulada é: P($X \leq x$) = $\sum_{x_i \leq x} p(x_i)$.

No exemplo anterior, $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$= 0{,}6561 + 0{,}2916 + 0{,}0486 = 0{,}9963$ ou 99,63%.

$P(X \leq 2)$ é a probabilidade de que o número de fumantes seja no máximo 2, no grupo de quatro pessoas.

iii) $P(X \geq x) = 1 - P(X < x)$

No exemplo anterior, $P(X \geq 2) = 1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)]$

$= 1 - [0{,}6561 + 0{,}2916]$

$= 1 - [0{,}9477] = 0{,}0523$ ou 5,23%

$P(X \geq 2)$ é a probabilidade de que o número de fumantes seja no mínimo 2, no grupo de quatro pessoas.

Em distribuições discretas, $P(X \leq x) \neq P(X < x)$.

No exemplo anterior,

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0{,}9963$

$P(X < 2) = P(X = 0) + P(X = 1) = 0{,}9477$

3.10.4. Distribuição de Poisson

Em algumas situações, será avaliado o número de ocorrências de um tipo de evento por intervalo. Esse intervalo pode ser referente ao tempo, comprimento, área, volume, peça ou outro. Como exemplo, temos as seguintes variáveis aleatórias:

X: Número de pacientes atendidos em uma clínica, por hora.

X: Número de pacientes atendidos por um médico, por dia.

X: Número de defeitos em piso de cerâmica, por m^2.

X: Numero de camisas com defeitos, por lote.

Definição: Se o número médio de contagens no intervalo for $\lambda > 0$, a variável aleatória *X*, que é igual ao número de contagens no intervalo, terá **distribuição de Poisson**, com parâmetro λ:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0, 1, 2, 3, 4,...$$

A **média** e a **variância** da distribuição de Poisson são:

$\boldsymbol{E(X) = \lambda E(X) = \lambda}$ e $\boldsymbol{Var(X) = \lambda Var(X) = \lambda}$.

PROPRIEDADES:

i) A soma das probabilidades é igual a 1, ou seja, $\sum_{i=1}^{n} p(x_i) = 1$

$P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + ... = 1$

ii) A probabilidade acumulada é: P(*X* ≤ *x*) = $\sum_{x_i \leq x} p(x_i)$.

P(*X* ≤ x) = $P(X = 0) + P(X = 1) + ... + P(X = x)$

iii) P(*X* ≥ *x*) = 1 – P(*X* < *x*).

Exemplo: Sabe-se que na prova final de matemática do sexto ano do ensino fundamental, há, em média, duas questões erradas por prova. Seleciona-se uma prova, e pergunta-se:

a) qual é a probabilidade de que não haja questão errada?

b) qual é a probabilidade de que haja três questões erradas?

c) qual é a probabilidade de que haja no máximo três questões erradas?

d) qual é a probabilidade de que haja no mínimo três questões erradas?

Resolução:

A variável aleatória é *X*: número de questões erradas, por prova, e possui distribuição de Poisson com parâmetro $\lambda = 2$.

a) $P(X = 0) = \frac{e^{-2}2^{0}}{0!} = \frac{(0{,}1353)1}{1} = 0{,}1353$ ou 13,53%.

b) $P(X = 3) = \frac{e^{-2}2^{3}}{3!} = \frac{(0{,}1353)8}{6} = 0{,}1804$ ou 18,04%.

c) $P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$$= \frac{e^{-2}2^{0}}{0!} + \frac{e^{-2}2^{1}}{1!} + \frac{e^{-2}2^{2}}{2!} + \frac{e^{-2}2^{3}}{3!}$$

= 0,1353+ 0,2707 +0,2707 + 0,1804

= 0,8571 ou 85,71%

d)$P(X \geq 3) = 1 - P(X < 3) =$

$= 1 - [P(X = 0) + P(X = 1) + P(X = 2)] =$

$= 1 - [0,1353 + 0,2707 + 0,2707]$

$= 1 - 0,6767 = 0,3233$ ou 32,33 %

No caso da distribuição de Poisson, temos que x = 0, 1, 2, ..., , ou seja, a variável aleatória *X* pode assumir infinitos valores.

Por isso, no item d) é obrigatório o emprego da propriedade P(X ≥ x) = 1 – P(X < x).

3.10.5. Distribuição Geométrica

Definição: Em uma série de tentativas independentes de Bernoulli, com probabilidade *p* de um sucesso, a variável aleatória *X* , que denota o número de tentativas até que o primeiro sucesso ocorra, tem uma **distribuição geométrica**, com parâmetro *p*:

$$P(X = x) = (1 - p)^{x-1}p, \; x = 1, 2, 3, 4...$$

A **média** e a **variância** da distribuição geométrica são:

$E(X) = 1/p$ e $Var(X) = (1 - p)/p^2$.

PROPRIEDADES:

i) A soma das probabilidades é igual a 1, ou seja, $\sum_{i=1}^{n} p(x_i) = 1$

$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + ... = 1$

ii) A probabilidade acumulada é: P(*X* ≤ *x*) $= \sum_{x_i \leq x} p(x_i)$.

$P(X \leq x) = P(X = 1) + P(X = 2) + ... + P(X = x)$

iii) P(X ≥ x) = 1 – P(X < x)

Distribuição Geométrica depende apenas do parâmetro *p*.

Exemplo: Um jovem casal deseja ter um filho do sexo masculino.

a) qual é a probabilidade de que o casal realize seu desejo na primeira tentativa?

b) qual é a probabilidade de que o casal realize seu desejo na terceira tentativa?

c) qual é a probabilidade de que o casal realize seu desejo com no máximo três tentativas?

Resolução:

A variável aleatória é X: número de tentativas até que ocorra menino, a qual possui distribuição geométrica com parâmetro $p = 1/2$ (probabilidade se nascer menino)

Quadro 3.2: Exemplo da distribuição geométrica

		$(X = x)$	$P(X = x) = (1-p)^{x-1}p$
a)	M	$X = 1$	$(1/2)^0(1/2) = 0,5$
	FM	$X = 2$	$(1/2)(1/2) = 0,25$
b)	FFM	$X = 3$	$(1/2)^2(1/2) = 0,125$
	FFFM	$X = 4$	$(1/2)^3(1/2) = 0,0625$
		⋮	
	FFFF....M	$X = x$	$(1 - p)^{x-1}p$

a) $P(X = 1) = (1/2)^0(1/2) = 0,5$ ou 50%

b) $P(X = 3) = (1/2)^2(1/2) = 0,125$ ou 12,5%

c) $P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0,5 + 0,25 + 0,125 = 0,875$ ou 87,5%.

Fique atento, os possíveis valores para a variável aleatória com distribuição geométrica se iniciam em um.

3.10.6. Distribuição Hipergeométrica

Definição: Considere um conjunto de *N* elementos, *r* dos quais apresentam determinada característica ($r \leq N$) e $N - r$ não apresentam esta característica. Extrai em *n* elementos ($n \leq N$), sem reposição. A variável aleatória *X*, que é o número de elementos que possuem a característica entre os *n* retirados, possui **distribuição hipergeométrica**:

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \; x = 0, 1, 2, \ldots, r.$$

A **média** e a **variância** da distribuição hipergeométrica são:

$$E(X) = n\left(\frac{r}{N}\right) \text{ e } \; Var(X) = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

PROPRIEDADES:

i) A soma das probabilidades é igual a 1, ou seja, $\sum_{i=1}^{n} p(x_i) = 1$

ii) A probabilidade acumulada é: P($X \leq x$) = $\sum_{x_i \leq x} p(x_i)$.

iii) P($X \geq x$) = 1 – P($X < x$).

O termo $\binom{r}{x}$ é uma combinação, $C_{r,x}$, ou seja: $\binom{r}{x} = C_{r,x} = \frac{r!}{x!(r-x)!}$. Logo, a distribuição geométrica pode ser equivalentemente reescrita:

$$P(X = x) = \frac{C_{r,x}C_{N-r,n-x}}{C_{N,n}}$$

Exemplo: Um eletricista comprou uma caixa com 10 lâmpadas, as quais são muito frágeis e, de acordo com o fabricante, 20% delas estão queimadas. Um eletricista seleciona 4 lâmpadas ao acaso, sem reposição, para instalá-las numa loja comercial. Qual é a probabilidade de que:

a) não haja lâmpadas queimadas?

b) no mínimo uma lâmpada esteja queimada?

c) Calcule o número esperado de lâmpadas queimadas para essa amostra.

Resolução:

A variável aleatória é X: número de lâmpadas queimadas, a qual possui distribuição hipergeométrica com parâmetros $N = 10$, $r = 2$, $n = 4$.

Logo, a variável aleatória pode assumir $x = 0, 1, 2$.

a) $P(X = 0) = \frac{C_{2,0}C_{8,4}}{C_{10,4}} = \frac{(1)(70)}{(210)} = 0{,}3333$ ou 33.33%.

b) $P(X \geq 1) = 1 - P(X < 1)$

$= 1 - P(X = 0) = 1 - 0{,}3333 = 0{,}6667$ ou 66,67%.

c) O número esperado de lâmpadas queimadas é: $\boldsymbol{E(X)} = n\left(\frac{r}{N}\right) = 4(0{,}2) = 0{,}8$.

3.11. Distribuições contínuas de probabilidade

Nesta seção, apresentaremos três distribuições contínuas bastante conhecidas: uniforme contínua, exponencial e normal.

3.11.1. Distribuição uniforme contínua

Definição: Se X é uma variável aleatória uniformemente distribuída sobre o intervalo [a, b], então sua função densidade é:

$$f(x) = \frac{1}{b-a} = p \text{ , } a \leq x \leq b$$

A **média** e a **variância** da distribuição uniforme contínua são:

$\boldsymbol{E(X)} = (a+b)/2$ e $\boldsymbol{Var(X)} = (b-a)^2/12$.

3.11.2. Distribuição exponencial

Definição: Uma variável X tem **distribuição exponencial** se sua função densidade de probabilidade é da forma:

$$f(x) = \begin{cases} \alpha e^{-\alpha x}, se\ x > 0\ e\ \alpha > 0 \\ 0, caso\ contrário \end{cases},$$

Onde α é o parâmetro da distribuição

A **média** e a **variância** da distribuição exponencial são:

$\boldsymbol{E(X)} = 1/\alpha$ e $\boldsymbol{Var(X)} = 1/\alpha^2$

São exemplos de variáveis com distribuição exponencial: tempo de espera em uma fila, tempo de sobrevivência de um paciente após o início do tratamento e tempo de vida de

um equipamento eletrônico (teoria da confiabilidade).

PROPRIEDADES:

$P(X \leq x) = 1 - e^{-\alpha x}$

$P(X > x) = e^{-\alpha x}$

Distribuição exponencial depende apenas do parâmetro α.

Exemplo: Se o tempo médio de espera pela entrega da refeição em um restaurante é uma distribuição exponencial com média igual a 10 minutos,

Qual é a probabilidade de espera superior a 10 minutos?

Qual é a probabilidade de espera inferior ou igual a 3 minutos?

Resolução:

A variável aleatória é *X*: tempo de espera pela entrega da refeição, a qual possui distribuição exponencial com média $E(X) = 1/\alpha = 10$. Logo, $\alpha = 0{,}1$.

$P(X > 10) = e^{-0,1(10)} = e^{-1} = 0{,}3679$ ou 36,79%.

$P(X \leq 3) = 1 - e^{-0,1(3)} = 1 - e^{-0,3} = 1 - 0{,}7408 = 0{,}2592$ ou 25,92%.

3.11.3. Distribuição normal

Nesta seção, iremos estudar a distribuição de probabilidade mais importante da estatística, a distribuição normal. A distribuição normal pode ser aplicada a diversas situações na natureza, na indústria, negócios e outros. As variáveis aleatórias pressão sanguínea, idade, altura, peso possuem distribuição normal.

Definição: Uma variável aleatória *X* tem **distribuição normal** se sua função densidade de probabilidade é da forma:

$$fx = f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \text{ para } -\infty < x < \infty.$$

com parâmetros μ e σ^2, em que $\sigma > 0$.

Observe que μ é a média populacional e σ^2 é variância populacional.

A **média** e a **variância** da distribuição normal são:

$\boldsymbol{E(X) = \mu}$ e $\boldsymbol{Var(X) = \sigma^2}$

NOTAÇÃO:

$X \sim N(\mu, \sigma^2)$: a variável aleatória *X* possui distribuição normal com média μ e variância σ^2.

PROPRIEDADES

A função *f(x)* da distribuição normal satisfaz as propriedades da definição de função densidade de probabilidade

i) $f(x) \geq 0$.

Exemplo: $\mu = 100$, $\sigma^2 = 100$,

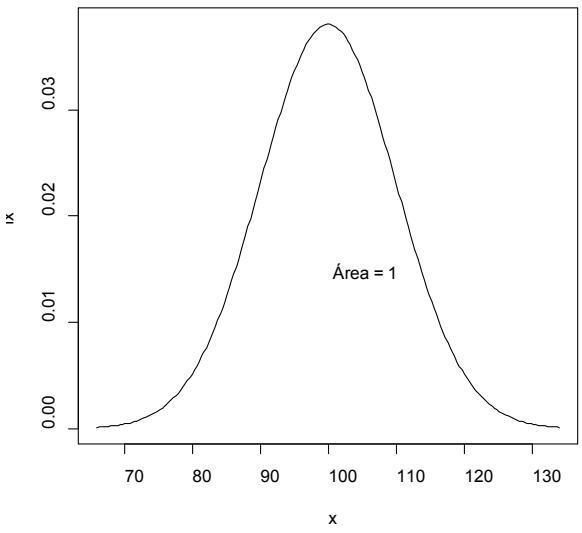


Figura 3.2: Curva da distribuição normal

Além disso, a curva do gráfico da *f(x)* da distribuição normal tem forma de sino.

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

Em outras palavras, a área total sob a curva normal é igual a um.

iii) para quaisquer *a*, *b*, com $-\infty < a < b < \infty$, teremos:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx = \text{área.}$$

Para a distribuição normal e demais distribuições contínuas, temos que:

$$P(a \leq X \leq b) = P(a < X < b)$$

Exemplo: Se μ = 100 , σ= 10, para calcular $P(X \leq 110)$, teremos:

$P(X \leq 110) = P(-\infty \leq X \leq 110) = \int_{-\infty}^{110} f(x)\,dx$ = área hachurada.

Observe que $a = -\infty$ e $b = 110$.

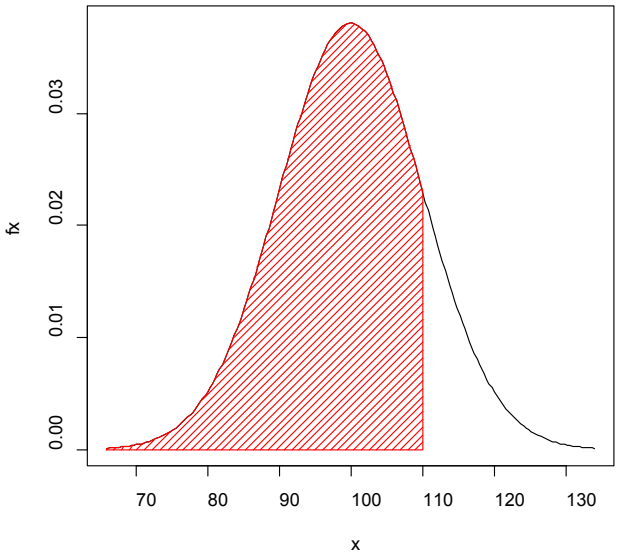


Figura 3.3: Curva da distribuição normal, μ = 100 , σ= 10

O cálculo de probabilidades através da equação $P(a \leq X \leq b) = \int_a^b f(x)\,dx$ é razoavelmente trabalhoso. Então, para a obtenção de probabilidades de variáveis aleatórias com distribuição normal, pode ser usada distribuição normal padrão. Veja a seguir.

iv) a curva da distribuição normal é simétrica em torno da média. No exemplo, μ = 100.

v) a curva da distribuição normal é assintótica ao eixo horizontal.

DISTRIBUIÇÃO NORMAL PADRÃO

Definição: Uma variável aleatória normal com μ = 0 e σ^2 = 1 é chamada de **variável aleatória normal padrão**. Uma variável aleatória normal padrão é denotada por *Z* e a distribuição normal padrão é:

$$fz = f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \text{ para } -\infty < z < \infty.$$

PROPRIEDADE:

Se *X* for uma variável aleatória normal com E(*X*) = μ e Var (*X*) = σ^2, então a variável aleatória

$$z = \frac{x - \mu}{\sigma}$$

será uma variável aleatória normal, com E(*z*) = 0 e Var(*z*) = 1.

NOTAÇÃO:

$Z \sim N(0,1)$: a variável aleatória *Z* possui distribuição normal com média 0 e variância 1.

Graficamente, temos:

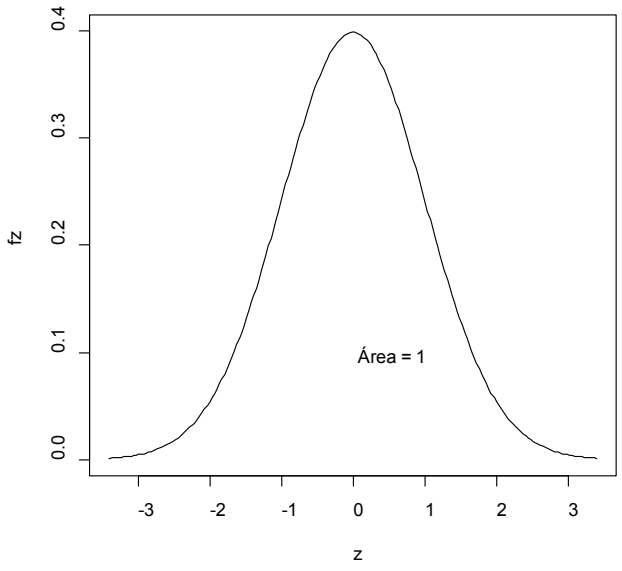


Figura 3.4: Curva da distribuição normal padrão

A área sob a curva da distribuição normal padrão também é um.

TABELA DA DISTRIBUIÇÃO NORMAL PADRÃO

A distribuição normal padrão é importante porque todas as distribuições normais podem ser transformadas numa **normal padrão,** a qual possui probabilidades tabeladas.

Seja $Z = \frac{x-\mu}{\sigma}$. A tabela da distribuição normal padrão fornece a probabilidade de Z tomar um valor menor ou igual a um valor fixo z_0:

$P(Z \leq z_0)$ = Área tabelada.

Veja Tabela A e B no final do módulo IV

PROPRIEDADE: $P(Z > z_0) = 1 - P(Z \leq z_0)$

A distribuição normal padrão é uma distribuição contínua, logo,

$$P(Z < z_0) = P(Z \leq z_0)$$

$$P(Z > z_0) = P(Z \geq z_0)$$

Exemplo: Se μ = 100, σ = 10, para calcular $P(X \leq 110)$, teremos:

Como x = 110, então o valor padronizado é: $z = \frac{110-100}{10} = 1$.

P (X ≤ 110) = $P(Z \leq 1)$ = 0,8413, conforme Tabela B , no final do módulo IV.

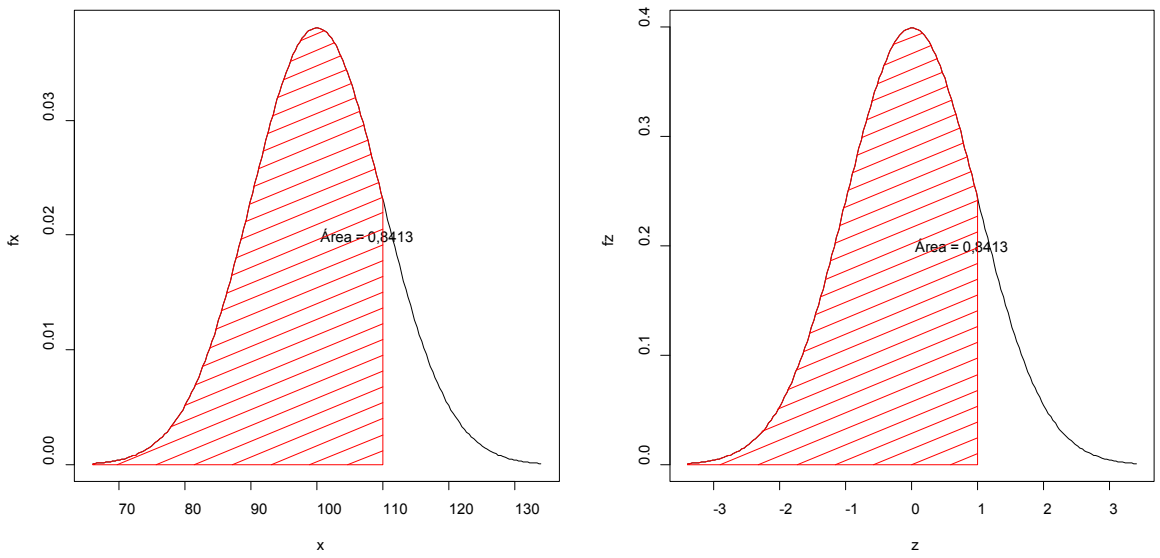


Figura 3.5: Curva da distribuição normal e normal padrão, $P(X \leq 110)$ e $P(Z \leq 1)$

Exemplo: Se μ = 100 , σ = 10, para calcular P ($X \leq 87{,}5$), teremos:

Como *x* = 87,5, então o valor padronizado é: $z = \frac{87{,}5-100}{10} = -1{,}25$.

P (*X* ≤ 87,5) = $P(z \leq -1{,}25)$ = 0,1056, conforme tabela A, no final do módulo IV.

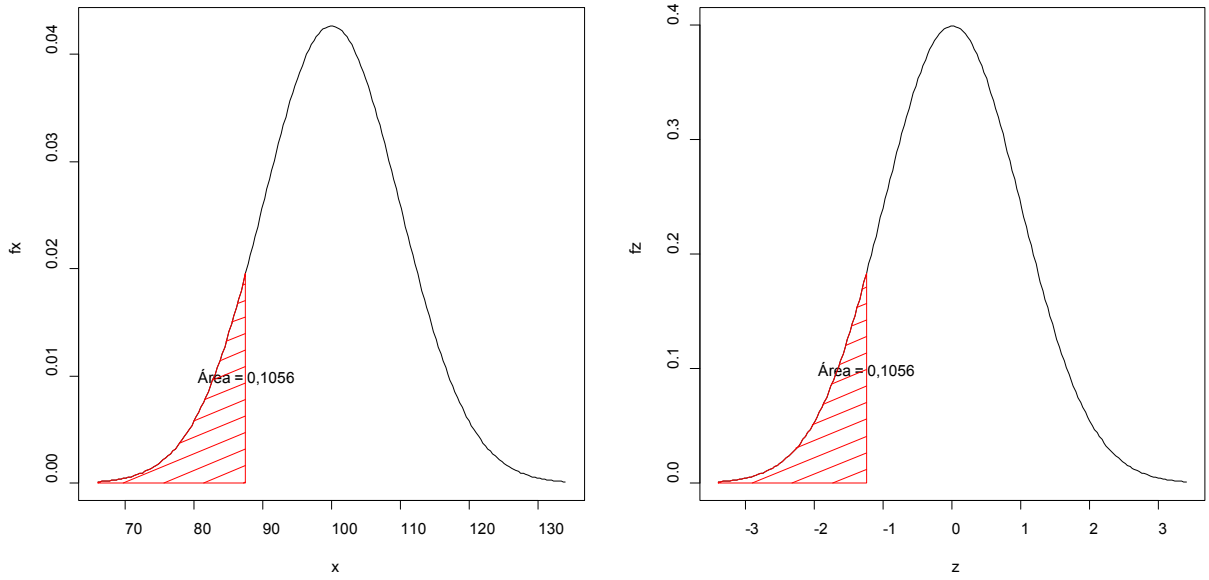


Figura 3.6: Curva da distribuição normal e normal padrão, *P* (*X* ≤ 87,5) e $P(Z \leq -1{,}25)$

Exemplo: A taxa de hemoglobina no sangue das pessoas saudáveis segue uma distribuição normal com média 12 e desvio padrão 1.

Qual é a probabilidade de se encontrar uma pessoa saudável com taxa de hemoglobina inferior a 13,5?

Qual é a probabilidade de se encontrar uma pessoa saudável com taxa de hemoglobina superior a 13,5?

Resolução:

A variável aleatória é *X*: taxa de hemoglobina no sangue, a qual possui distribuição normal, com média μ= 12 e σ = 1.

Para *P*($X \leq 13{,}5$), teremos:

Como *x* = 13,5, então o valor padronizado é: $z = \frac{13{,}5-12}{1} = 1{,}5$.

$P(X \leq 13,5) = P(Z \leq 1,5) = 0,9332$, conforme tabela B, no final do módulo IV.

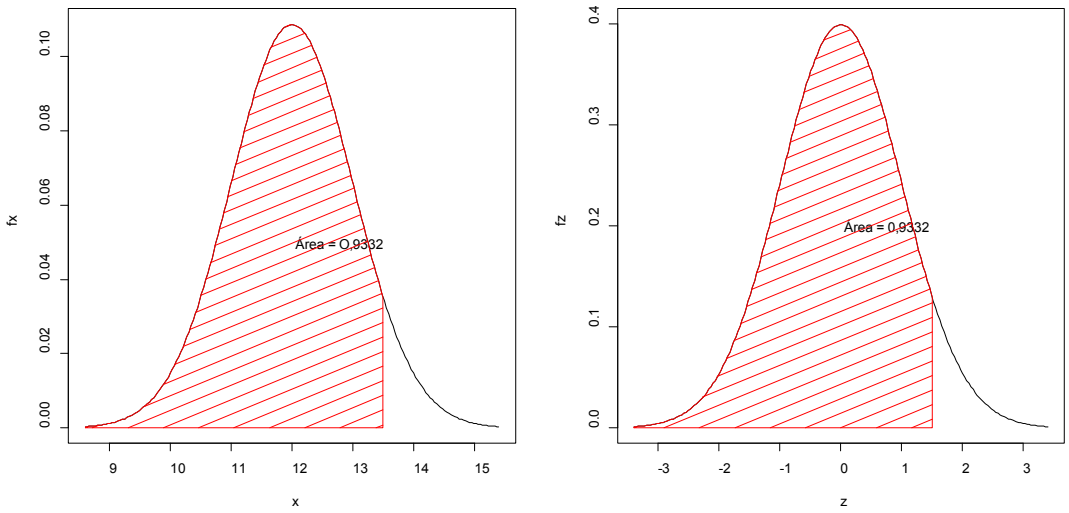


Figura 3.7: Curva da distribuição normal e normal padrão, $P(X \leq 13,5)$ e $P(Z \leq 1,5)$

Para $P(X \geq 13,5)$, teremos:

Como $x = 13,5$, então o valor padronizado é: $z = \frac{13,5-12}{1} = 1,5$.

$P(X > 13,5) = P(Z > 1,5) = 1 - P(Z \leq 1,5) = 1 - 0,9332 = 0,0668$.

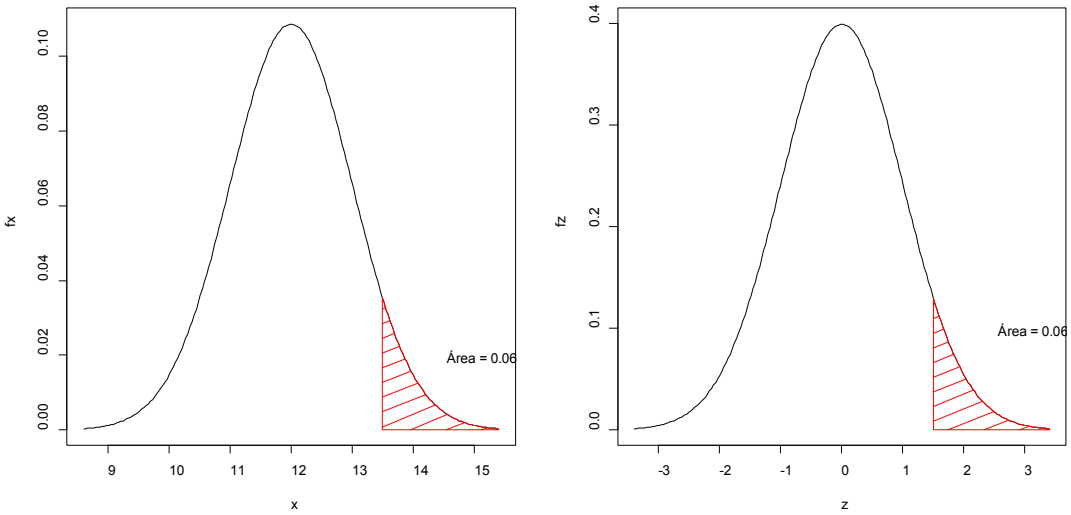


Figura 3.8: Curva da distribuição normal e normal padrão, $P(X > 13,5)$ e $P(Z > 1,5)$

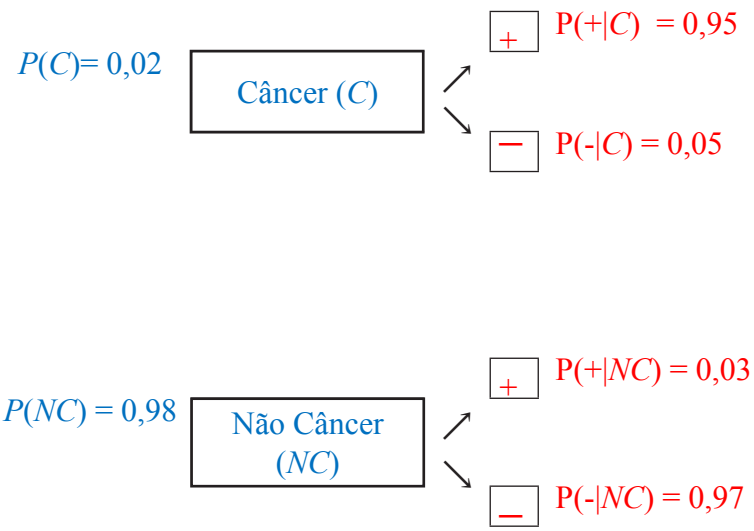
Nas Tabelas A e B da distribuição normal padrão, se $z_0 > 3,39$, assume $z_0 = 3,39$.

$$P(Z \leq 4,5) = 0,9999 = 1.$$

$$P(Z \leq 6) = 0,9999 = 1.$$

$$P(Z \leq -7,2) = 0$$

Resolução:



Usaremos a seguinte notação:

NC : evento «não ter câncer»

--: evento «ter reação negativa»

A probabilidade procurada é:

$$P(NC|-) = \frac{P(NC \cap -)}{P(-)} = \frac{P(NC)P(-|NC)}{P(C)P(-|C) + P(NC)P(-|NC)}$$

$$P(NC|-) = \frac{(0{,}98)(0{,}97)}{(0{,}02)(0{,}05) + (0{,}98)(0{,}97)}$$

$$P(NC|-) = \frac{0{,}9506}{0{,}9516} = 0{,}999$$

Esse exemplo foi extraído de FARIAS *et al*. (2003)

II - SÍNTESE

Neste módulo, foram estudados os conceitos básicos de probabilidade, a definição de variáveis aleatórias e suas propriedades, as distribuições discretas de probabilidade e as distribuições contínuas de probabilidade. Essas distribuições possuem ampla aplicação prática, pois no cotidiano há várias situações em que as características de interesse podem assumir mais de um valor.

IV - REFERÊNCIAS

FARIAS, A. A. ; CÉSAR,C. C.; SOARES, J. F. **Introdução à Estatística. 2 ed.** Rio de Janeiro: LTC, 2003.

LARSON, R; FABER, B. **Estatística Aplicada**. 4 ed. São Paulo: Pearson Prentice Hall, 2010.
MILONE, Giuseppe. **Estatística Geral e Aplicada**. São Paulo: Pioneira Thomson Learning, 2004.

MÓDULO 3

Inferência: Distribuição Amostral e Estimação

Conteúdos básicos do Módulo:

1. Distribuições amostrais
2. Teorema Central do Limite
3. Técnicas de amostragem
4. Estimação pontual
5. Estimação por intervalo

Objetivos do Módulo:

Ao final deste estudo, esperamos que você, aluno(a), possa:

- Relacionar estimadores amostrais e parâmetros populacionais
- Selecionar amostras aleatórias
- Obter estimativas pontuais para os parâmetros populacionais
- Obter estimativas por intervalo para os parâmetros populacionais

I - TEXTO BÁSICO

4. Amostragem e distribuições amostrais

Um dos objetivos da inferência estatística é obter estimativas para os parâmetros populacionais usando a informação numérica contida em uma amostra. Assim sendo, é necessário conhecer a relação entre as estimativas amostrais e os parâmetros populacionais. Essa relação é proveniente da distribuição amostral.

4.1. Conceitos básicos

No capítulo 2, definimos o que é **amostra** e o que é **população**. As definições são reapresentadas a seguir.

População é o conjunto constituído por elementos que apresentam pelo menos uma característica em comum.

Amostra é um subconjunto da população.

Muitas vezes, para obter medidas numéricas da população, deve-se recorrer a estimativas amostrais, por gastar menos tempo ou por ter custos mais baixos.

Exemplos:

Um fabricante de pneus desenvolveu um novo tipo de pneu, com maior durabilidade. Para estimar a durabilidade desse pneu, o fabricante selecionou uma amostra de 80 pneus, os quais foram testados. A durabilidade média amostral dos pneus foi de 45.670 km. Portanto, uma estimativa da durabilidade média para a população dos pneus é de 45.670 km.

Um partido precisa decidir se vai apoiar ou não um candidato a presidente. Para identificar a proporção de eleitores favoráveis ao candidato, foi selecionada uma amostra de 500 eleitores, dos quais 200 preferem o candidato. A estimativa da proporção da população de eleitores favoráveis ao candidato foi de 200/500 = 0,4.

4.2. Parâmetros e estimação pontual

As características numéricas de uma população, por exemplo, a média (μ), a variância (σ^2) e proporção (p), são chamadas **parâmetros.** E as características numéricas de uma amostra ($\bar{X}$, S^2 e $\hat{p}$) são chamadas de **estimadores pontuais.**

Os estimadores da média e da variância populacional ($\bar{X}$, S^2) já foram apresentados no capítulo 2. Veja resumo no quadro 3.1.

O estimador da proporção populacional *p* é dado por: $\hat{p} = \frac{X}{n}$, *n* é o número de elementos da amostra e *X* é o número de elementos da amostra que apresentam a característica de interesse.

Quadro 3.1: Parâmetros e estimadores

Parâmetro populacional	Estimador pontual
μ	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
σ^2	$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$
p	$\hat{p} = \frac{X}{n}$

* *n* é o número de elementos da amostra

No exemplo anterior.

$\bar{X}$ = 45.670, o valor “45.670” é uma **estimativa** para μ.

$\hat{p} = 0{,}4$, o valor é “0,4” é uma **estimativa** para *p*.

Para garantir que os estimadores pontuais forneçam estimativas representativas para a população, devemos usar uma amostra aleatória, a qual pode ser obtida através de técnicas de amostragem. Por exemplo, podemos usar a **Amostragem Aleatória Simples**, em que todos os elementos têm a mesma probabilidade de serem selecionados. Outras técnicas de amostragem serão apresentadas na última seção desse capítulo.

4.3. Distribuições amostrais

No capítulo 3, vimos que uma variável aleatória pode ser vista como uma descrição numérica do resultado de um experimento. Se considerarmos que o processo de escolher uma amostra aleatória simples é um experimento, a média amostral $\bar{X}$ é uma descrição numérica do resultado do experimento. Desse modo, a média amostral $\bar{X}$ é uma variável aleatória. Consequentemente, à semelhança do que ocorre com qualquer variável aleatória, $\bar{X}$ tem um valor médio, um desvio padrão e uma distribuição de probabilidade.

4.3.1. Distribuição amostral da média

Uma vez que os diversos valores possíveis de $\bar{X}$ são provenientes de diversas amostras aleatórias simples, a distribuição de probabilidade de $\bar{X}$ é chamada **distribuição amostral de $\bar{X}$**. Conhecer essa distribuição amostral e suas propriedades nos possibilitará fazer afirmações a respeito de quão próxima a média da amostra $\bar{X}$ está da média da população μ.

Exemplo: Seja uma turma constituída de três alunos (N = 3), os quais obtiveram notas 8, 9, e 10. Então, a média populacional das notas dessa turma é μ = 27/3 = 9 e a variância populacional é $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N} = \frac{(-1)^2+(0)^2+(1)^2}{3} = \frac{2}{3}$.

Suponha que se deseja selecionar dois alunos dessa turma (n = 2) para formar uma amostra e para a qual serão calculadas a média e a variância das notas. A seleção será sem reposição, então, temos 6 amostras possíveis, veja na tabela 3.1.

Tabela 3.1: Valores de $\bar{X}$ para as 6 amostras aleatórias simples

Amostra	Valores amostrais	$\bar{X}$
1	8 e 9	8,5
2	8 e 10	9
3	9 e 8	8,5
4	9 e 10	9,5
5	10 e 8	9
6	10 e 9	9,5

Podemos calcular a média de $\bar{X}$: E ($\bar{X}$) = $\frac{8,5+9+8,5+9,5+9+9,5}{6}$ = 9. Logo, E ($\bar{X}$) = μ.

A média de $\bar{X}$ é igual à média populacional.

A variância de $\bar{X}$ é Var ($\bar{X}$) = $\frac{(8,5-9)^2+(9-9)^2+(8,5-9)^2+(9,5-9)^2+(9-9)^2+(9,5-9)^2}{6}$ = 1/6.

Por outro lado, podemos observar que : Var ($\bar{X}$)= $\frac{(N-n)}{(N-1)}\frac{\sigma^2}{n}$.

Ao longo desse curso, assumiremos que o tamanho da população (*N*) é muito grande, e que $n/N \leq 0,05$, então Var ($\bar{X}$) = $\frac{\sigma^2}{n}$.

A variância de $\bar{X}$ depende da variância populacional e do tamanho da amostra.

Consequentemente, podemos calcular o desvio padrão de $\bar{X}$, que é chamado de **erro padrão da média,** *EP* ($\bar{X}$).

$$EP(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

O valor do erro padrão da média é útil para determinar a distância da média amostral em relação à média da população.

No exemplo anterior, a distribuição amostral de $\bar{X}$ será:

$\bar{x}$	8,5	9	9,5
P($\bar{X} = \bar{x}$)	2/6	2/6	2/6

Generalizando, temos que a **média** e o **erro padrão da distribuição amostral da média** são**:**

$$E(\bar{X}) = \mu \quad \text{e} \quad EP(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

4.3.2. O Teorema Central do Limite (TCL)

Se $x_1, x_2, \ldots, x_n$ for uma amostra aleatória de tamanho *n*, retirada de uma população com média μ e variância σ^2, e se $\bar{X}$ for a média da amostra, então:

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

Quando $n \to \infty$.

N (0,1) é distribuição normal padrão.

Este resultado é muito útil em estimação intervalar da média μ, no capítulo 5

Exemplo: Um auditor de banco declara que as contas de cartões de crédito são normalmente distribuídas com uma média de 2870 reais e um desvio padrão de 900 reais.

a) Qual é a probabilidade de que um titular de cartão de crédito aleatoriamente selecionado tenha uma conta menor que 2500 reais?

b) Você seleciona 25 titulares de cartões de crédito de forma aleatória. Qual é a probabilidade de que a média da conta deles seja menor que 2500?

Resolução:

A variável aleatória é *X*: valor da conta de um titular de cartão de crédito, a qual possui distribuição normal, com média μ= 2870 e σ = 900.

Para P ($X \leq 2500$), teremos:

Como *x* = 2500, então o valor padronizado é: $z = \frac{2500-2870}{900} = -0{,}41$.

P ($X \leq 2500$) = $P(Z \leq -0{,}41)$ = 0,3409 ou 34,09% , conforme tabela A, no final do módulo II.

Para *n* = 25 titulares, queremos a *P*($\bar{X} \leq 2500$).

Neste caso, devemos usar o Teorema Central do Limite e o valor padronizado é :

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Logo, $z = \frac{2500-2870}{900/25}$ = -2,06.

P($\bar{X} \leq 2500$) = $P(Z \leq -2{,}06)$ = 0,0197 ou 1,97%, conforme tabela A, no final do módulo IV.

4.3.3. Distribuição amostral da proporção

Seja a proporção *p* de uma população que apresenta certa característica. Então, a proporção de amostral é:

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

n é o número de elementos da amostra e *X* é o número de elementos da amostra que apresenta a característica de interesse.

A **média** e o **erro padrão** da **distribuição amostral da proporção** são**:**

$$E(\hat{p}) = p \quad \text{e} \quad EP(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

A distribuição amostral de $\hat{p}$ pode ser aproximada por meio de uma distribuição normal.

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

Este resultado será muito útil em estimação intervalar da proporção *p*, no capítulo 5.

4.3.4. Distribuição amostral da variância

A **média** e o **erro padrão** da **distribuição amostral da variância** são**:**

$$E(s^2) = \sigma^2 \quad \text{e} \quad EP(s^2) = \sqrt{\frac{2\sigma^2}{(n-1)}}$$

Resultado: Se S^2 é a variância de uma amostra aleatória de tamanho *n*, retirada de uma população normal, com variância σ^2, então a estatística

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

tem distribuição qui-quadrado com $\varphi = n - 1$ graus de liberdade.

Este resultado será muito útil em estimação intervalar da variância σ^2, no capítulo 5.

4.4. Outras técnicas de amostragem

4.4.1. Amostragem sistemática

Os elementos da população podem ser organizados em sequência. Seleciona-se aleatoriamente a posição do primeiro elemento da amostra e a posição dos demais elementos são calculadas pela razão ($r = N/n$)

____, _____, ____, _____,

1o 2o .... *n*-ésimo

Exemplo: Selecionar 5 alunos de uma turma de 20 alunos.

Como, $r = 20/5 = 4$. Então, os elementos que vão compor a amostra são:

(1º, 5º, 9 º, 13 º, 17 º) ou (2 º, 6 º, 10 º, 14 º, 18 º) ou (3 º, 7 º, 11 º, 15 º, 19 º) ou

(4 º, 8 º, 12 º, 16 º, 20 º)

4.4.2. Amostragem por conglomerados

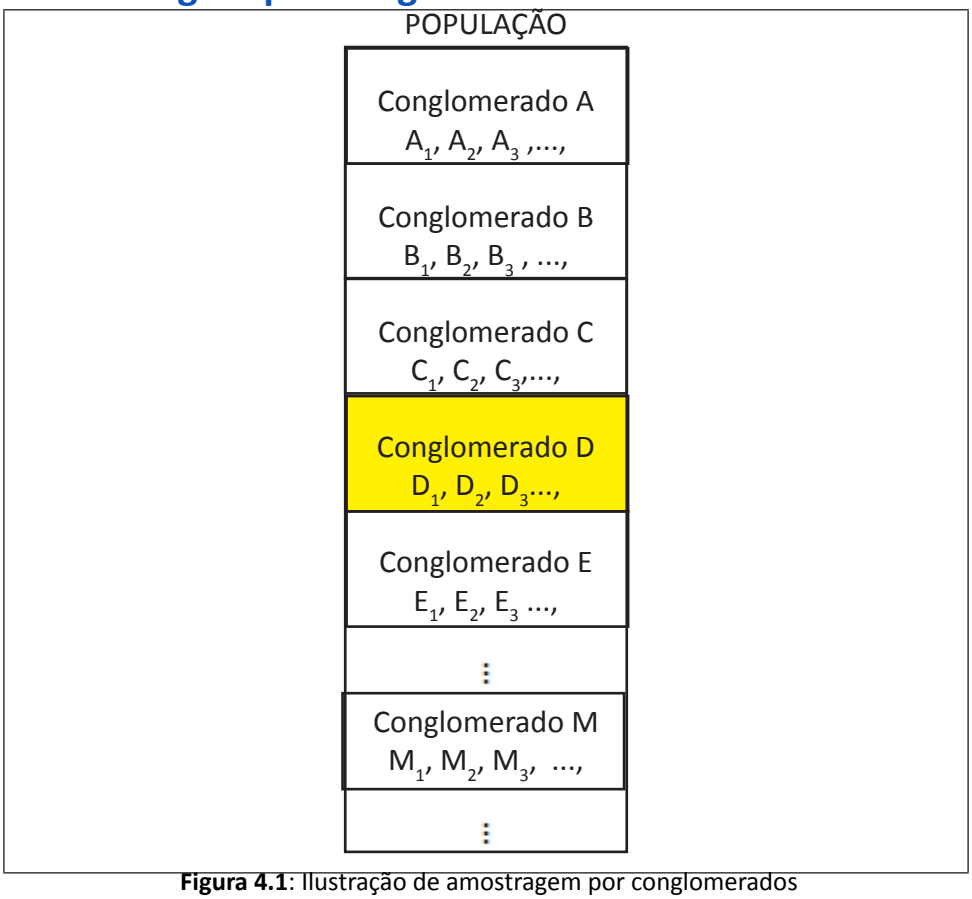


Figura 4.1: Ilustração de amostragem por conglomerados

Antes de iniciar uma amostragem, deve-se identificar as característica de interesse do estudo. Por exemplo, peso, idade, altura, renda, número de filhos, temperatura do ambiente, taxa de glicose e outros.

Divide-se a população em conglomerados (subgrupos).

Os elementos dos conglomerados são heterogêneos internamente (diferentes) e os conglomerados são homogêneos entre si (semelhantes).

Cada elemento da população pertence a um único conglomerado.

Seleciona-se aleatoriamente um (ou mais) conglomerado(s).

Todos os elementos do(s) conglomerado(s) selecionado(s) fará(ão) parte da amostra.

O conglomerado deve ser o mais representativo possível para a população.

Exemplo: Seja a população de alunos do 7º ano do ensino fundamental da Escola A. Cada uma das seis turmas corresponde a um conglomerado, se a característica de interesse for o "número de horas de estudo, por semana".

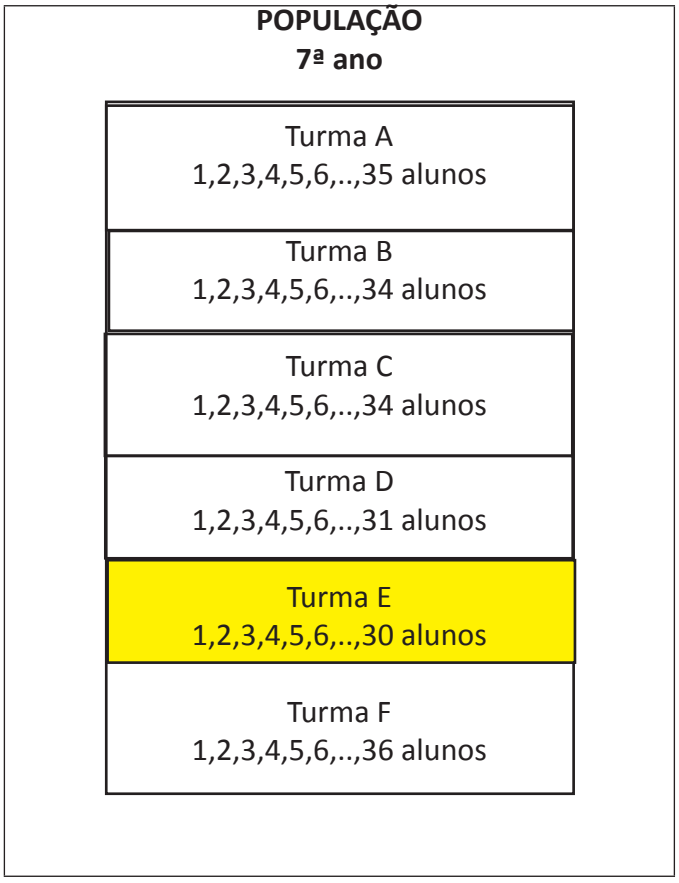


Figura 4.2: Exemplo de amostragem por conglomerados

Ao formar esses conglomerados, estamos considerando que o "número de horas de estudo, por semana" entre os conglomerados (turmas) são semelhantes entre si, e que dentro dos conglomerados (turma) o «número de horas de estudo, por semana» dos alunos é variado.

Ao proceder a seleção aleatória, a «turma E» foi selecionada, logo, os 30 alunos dessa turma farão parte da amostra.

4.4.3. Amostragem estratificada

Após identificar a característica de interesse do estudo, divide-se a população em estratos (subgrupos).

Os elementos dos estratos são homogêneos internamente (semelhantes) e os estratos são heterogêneos entre si (diferentes).

Cada elemento da população pertence a um único estrato.

Seleciona-se aleatoriamente elementos de cada estrato, de modo que todos os estratos tenham representantes na amostra. A amostra final será:

$$n = n_1 + n_2 + ... + n_h,$$

em que h é o número de estratos.

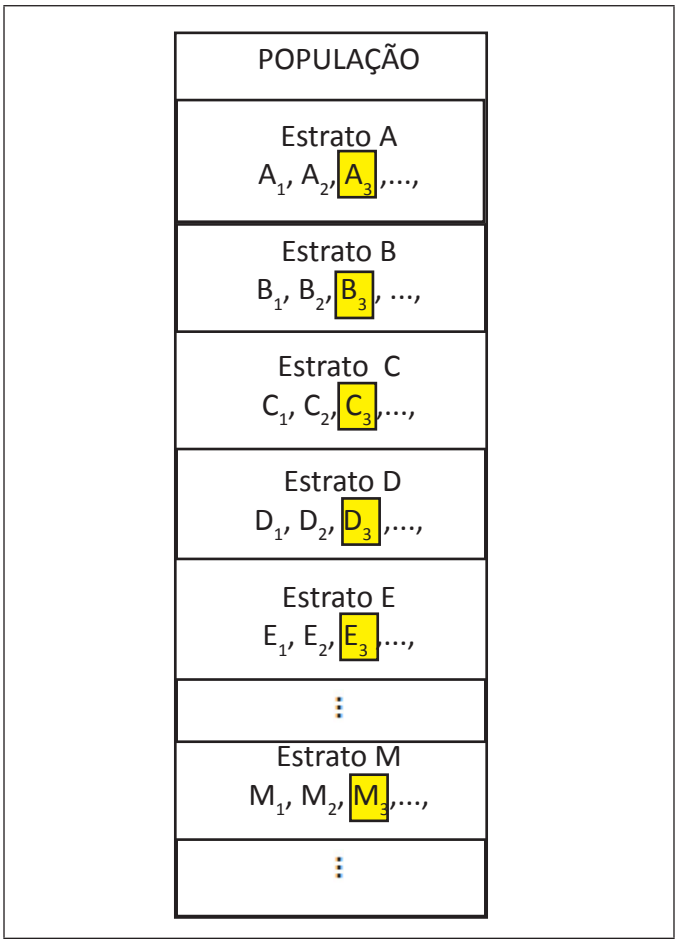


Figura 4.3: Ilustração de amostragem estratificada

Exemplo: Seja a população de todos os alunos do ensino fundamental da Escola A, cada ano corresponde a um estrato, se a característica de interesse for a "idade".

POPULAÇÃO
Ensino fundamental

1ºano 1,2,3,4,5,6,..,100 alunos	→ 6 anos
2ºano 1,2,3,4,5,6,..,100 alunos	→ 7 anos
3ºano 1,2,3,4,5,6,..,100 alunos	→ 8 anos
4ºano 1,2,3,4,5,6,..,100 alunos	→ 9 anos
5ºano 1,2,3,4,5,6,..,100 alunos	→10 anos
6ºano 1,2,3,4,5,6,..,200 alunos	→11 anos
7ºano 1,2,3,4,5,6,..,200 alunos	→ 12 anos
8ºano 1,2,3,4,5,6,..,200 alunos	→ 13 anos
9ºano 1,2,3,4,5,6,..,200 alunos	→ 14 anos

Figura 4.4: Exemplo de amostragem estratificada

Os estratos são formados considerando que as «idades» do 1º ao 9º estrato (ano) são variadas, e dentro de cada estrato (ano), as «idades» são semelhantes. Por exemplo, na figura 4.5 a executar a seleção aleatória (sorteio), desejamos selecionar 26 alunos através da amostragem proporcional.

POPULAÇÃO **Ensino fundamental**	Amostragem proporcional
1ºano 1,2,3,4,5,6,..,100 alunos	2 alunos
2ºano 1,2,3,4,5,6,..,100 alunos	2 alunos
3ºano 1,2,3,4,5,6,..,100 alunos	2 alunos
4ºano 1,2,3,4,5,6,..,100 alunos	2 alunos
5ºano 1,2,3,4,5,6,..,100 alunos	2 alunos
6ºano 1,2,3,4,5,6,..,200 alunos	4 alunos
7ºano 1,2,3,4,5,6,..,200 alunos	4 alunos
8ºano 1,2,3,4,5,6,..,200 alunos	4 alunos
9ºano 1,2,3,4,5,6,..,200 alunos	4 alunos

Figura 4.5: Exemplo de amostragem estratificada proporcional

Formalizando, temos que na **amostragem estratificada proporcional**, o número de elementos do estrato *h* é :

$$n_h = \left(\frac{N_h}{N}\right) * n$$

Na equação acima:

N é o tamanho da população

n é o tamanho da amostra

N_h é o tamanho populacional do estrato *h*.

Na amostragem estratificada uniforme, cada estrato tem a mesma quantidade de representantes na amostra, independente do tamanho populacional de cada estrato.

5. Estimação por intervalo

5.1 Conceitos Básicos

A **estimação por intervalo** ou **intervalo de confiança** procura determinar intervalos com limites aleatórios, que abranjam o valor do parâmetro populacional, com uma margem de segurança pré-fixada.

Exemplos: $8 \leq \mu \leq 10$ e $0{,}22 \leq p \leq 0{,}28$.

Vamos construir um intervalo de confiança para o parâmetro desconhecido com uma probabilidade de (1 - α)100% – chamado de **nível de confiança**- de que o intervalo contenha o verdadeiro parâmetro. Observem que (1 - α) pode ser igual a 99%, 90%, etc.

5.2. Intervalos de confiança para a média populacional

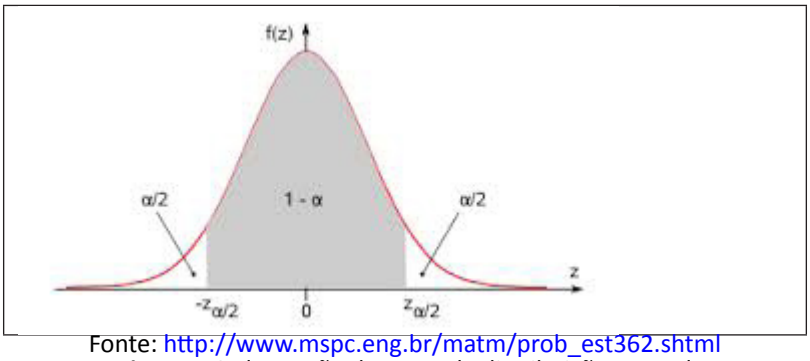
5.2.1. Intervalos de confiança para a média populacional, quando σ for conhecido

Seja x_1, ...x_n, uma amostra aleatória de tamanho *n* da variável aleatória $X \sim N(\mu, \sigma^2)$. Seja o estimador para μ.

Pelo Teorema Central do Limite, sabe-se que

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

Então, para $(1-\alpha)$ fixo, podemos construir:



Fonte: http://www.mspc.eng.br/matm/prob_est362.shtml
Figura 4.6: Ilustração da curva da distribuição normal

Logo, podemos dizer que:

$$P\left(-z_{\alpha/2} \leq z \leq z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$$

Utilizando propriedades algébricas, podemos isolar a média populacional μ da expressão acima.

$$P\left[-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \bar{X}-\mu \leq z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1-\alpha$$
$$P\left[-\bar{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq -\mu \leq -\bar{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1-\alpha$$

Multiplicando a inequação por (-1), temos:

$$P\left[\bar{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1-\alpha$$

O **intervalo de confiança da média populacional**, quando σ for conhecido é

$$IC(\mu)_{1-\alpha} = \bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

em que $(1-\alpha)$ é o coeficiente de confiança e $z_{\alpha/2}$ é o valor *z* que produz uma área de $\alpha/2$ na cauda à direita da distribuição normal padrão de probabilidade.

Interpretação:

A média populacional $\mu \in \left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right);\ \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$ com $(1-\alpha)100\%$ de confiança.

O nível de confiança $(1-\alpha)100\%$ significa que, se retirarmos um grande número de amostras de tamanho *n*, fixo, da população em estudo e para cada amostra construirmos um intervalo, então, o verdadeiro parâmetro populacional estará contidoem $(1-\alpha)100\%$ desses intervalos.

É ERRADO dizer que a média populacional $\mu \in \left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right);\ \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$ com $(1-\alpha)100\%$ de probabilidade.

Neste curso, usaremos os níveis de confiança de 90%, 95% e 99%. Os valores de $z_{\alpha/2}$ correspondentes a esses níveis de confiança estão a seguir:

Quadro 3.2: Valores para $Z_{\alpha/2}$

Caso	(1- α)	α	$\alpha/2$	$z_{\alpha/2}$
A	0,90	0,10	0,05	1,64 ou 1,65 ou 1,645
B	0,95	0,05	0,025	1,96
C	0,99	0,01	0,005	2,575 ou 2,58 ou 2,575

No caso B, colocamos na extremidade à direita da curva, a área 0,025.

Então, a área 0,975 (=1- 0,025) fica à esquerda da curva.

Finalmente, procura-se no interior da tabela B o valor 0,975, o qual é correspondente ao z_0 = 1, 96. Veja no esboço:

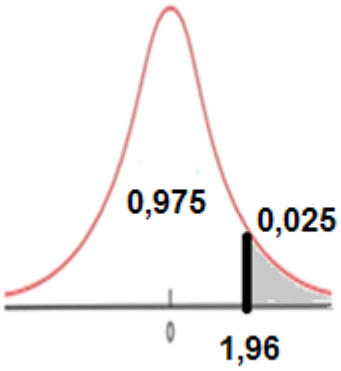


Figura 4.7: Ilustração de valores tabelados de z

No caso C, colocamos na extremidade à direita da curva, a área 0,005.

Então, a área 0,995 (=1- 0,005) fica à esquerda da curva.

Finalmente, procura-se no interior da tabela B o valor mais próximo de 0,995 que é 0,9949 (ou 0,9951), os quais são correspondentes ao z_0 = 2,57 (ou 2,58). Veja no esboço a seguir.

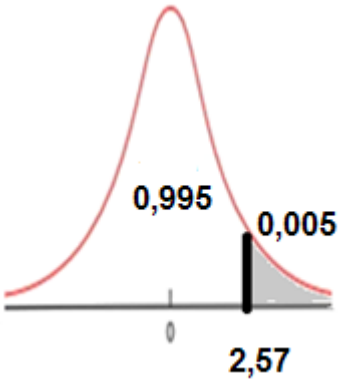


Figura 4.8: Ilustração de valores tabelados de z

Pode-se ainda, utilizar a média de 2,57 e 2,58, logo z_0 = 2,575.

Outros valores de $z_{\alpha/2}$ podem ser obtidos nas tabelas A e B.

Exemplo:

Uma máquina enche pacotes de café com uma variância igual a 100g^2. Uma amostra de 25 pacotes apresentou peso médio igual a 450 g.

a) Determine o intervalo de 95% de confiança para a média populacional.

n = 25; $\bar{X}$=450; $\sigma = \sqrt{100} = 10$; α = 0,05; $z_{0,025}$ = 1,96.

$$IC(\mu)_{1-\alpha} = \bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$IC(\mu)_{0,95} = 450 \pm 1{,}96\left(\frac{10}{\sqrt{25}}\right)$$

$$IC(\mu)_{0,95} = 450 \pm 3{,}92$$

Interpretação:

A verdadeira média populacional μ pertence ao intervalo [446,08; 453,92] com 95% de confiança.

b) Determine o intervalo de 90% de confiança para a média populacional.

n = 25; $\bar{X}$=450; $\sigma = \sqrt{100} = 10$; $z_{0,05} = 1{,}645$;

$$IC(\mu)_{1-\alpha} = \bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$IC(\mu)_{0,90} = 450 \pm 1{,}645\left(\frac{10}{\sqrt{25}}\right)$$

$$IC(\mu)_{0,90} = 450 \pm 3{,}3$$

Interpretação:

A verdadeira média populacional μ pertence ao intervalo [446,7; 453,3] com 90% de confiança.

5.2.2. Intervalos de confiança para a média populacional, quando σ for desconhecido e *n* < 30.

Na prática é comum não ser conhecido o valor de σ . Nesse caso, o intervalo de confiança é calculado utilizando-se uma outra estatística

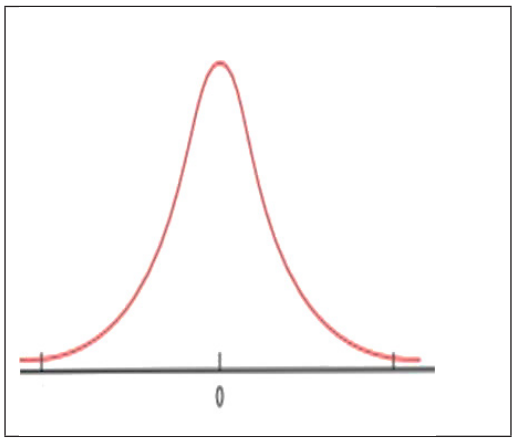
$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{\alpha,n-1}$$

s é o estimador do desvio padrão e

$t_{\alpha,n-1}$ refere-se à **distribuição *t* de Student** com φ = (n – 1) graus de liberdade (g.l) ,

n é o tamanho da amostra.

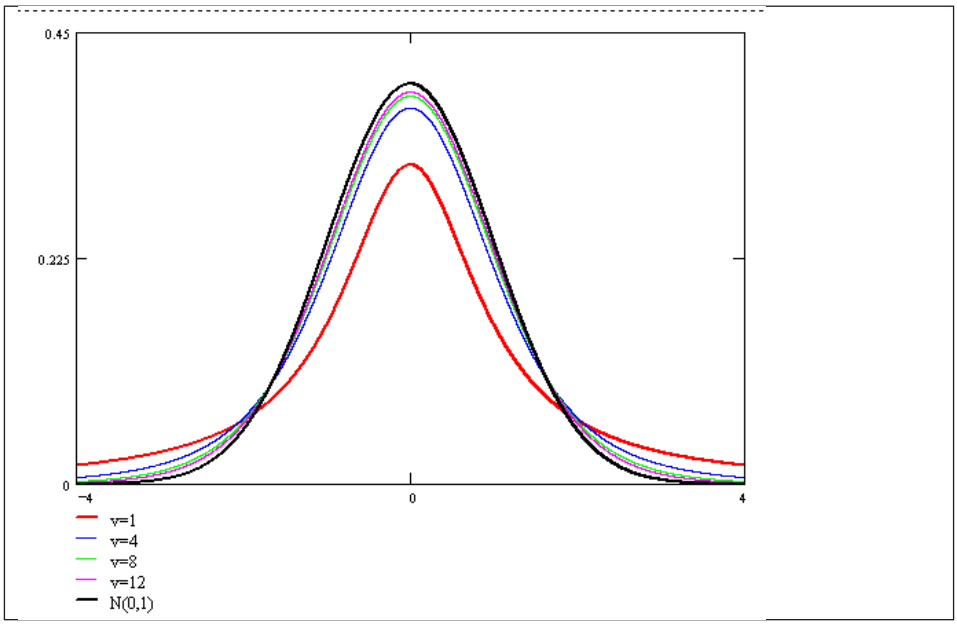
A forma da distribuição *t* de Student é parecida com a da normal.



Fonte: http://projetos-estatisticaspss.blogspot.com.br/2014/05/distribuicao-t-student.html
Figura 4.7: Ilustração da curva da distribuição *t* de Student

A distribuição *t* de Student é simétrica em relação a 0, mas apresenta caudas mais 'grossas', ou seja, maior variância do que a normal. Aumentando-se *n*, a distribuição *t* de Student tende para a normal.

Veja a figura abaixo:



Fonte: http://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_t_de_Student
Figura 4.8: Comparação da curva da distribuição normal e da *t* de Student

Logo, podemos dizer que:

$$P\left(-t_{\alpha/2} \leq T \leq t_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

Utilizando propriedades algébricas, podemos isolar a média populacional μ da expressão acima.

$$P\left[-t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \leq \bar{X} - \mu \leq t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

$$P\left[-\bar{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \leq -\mu \leq -\bar{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

Multiplicando a inequação por (-1), temos:

$$P\left[\bar{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right] = 1 - \alpha$$

O **intervalo de confiança da média populacional**, quando σ for desconhecido e $n < 30$ é

$$IC(\mu)_{1-\alpha} = \bar{X} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

em que $(1 - \alpha)$ é o coeficiente de confiança e $t_{\alpha/2}$ é o valor t que produz uma área de $\alpha/2$ na cauda à direita da distribuição t de student, com $\varphi = n - 1$ graus de liberdade.

Interpretação:

A média populacional $\mu \in \left[\bar{X} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right);\ \bar{X} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right]$ com $(1 - \alpha)100\%$ de confiança.

Exemplo: Uma amostra com as notas de 15 estudantes universitários selecionados aleatoriamente apresentou média 2,35 e desvio padrão 1,03. Determine o intervalo de 95% de confiança para a média populacional.

$n = 15$; $\bar{X} = 2{,}35$; $s = 1{,}03$; $\alpha = 0{,}05$.

Como $\varphi = n - 1 = 14$ graus de liberdade. Para $t_{\frac{\alpha}{2}} = t_{\frac{0{,}05}{2}} = t_{0.025} = 2{,}145$, conforme **Tabela C**.

$$IC(\mu)_{1-\alpha} = \bar{X} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$IC(\mu)_{0,95} = 2{,}35 \pm 2{,}145\left(\frac{1{,}03}{\sqrt{15}}\right)$$

$$IC(\mu)_{0,95} = 2{,}35 \pm 0{,}57$$

Interpretação:

A verdadeira média populacional μ pertence ao intervalo [1,78; 2,92] com 95% de confiança.

Ilustração da obtenção de $t_{0,025}$ = 2,145.

Como $\varphi\varphi$ = n -1 = 14 graus de liberdade e área à direita da curva é 0,025. Então, da tabela C, temos que $t_{0,025}$ = 2,145.

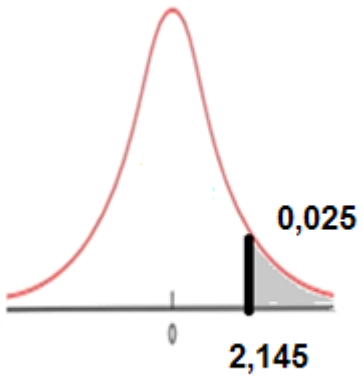


Figura 4.11: Ilustração de valores tabelados de t

Tabela 4.1: Reprodução parcial da tabela C

Graus de liberdade	Área da cauda direita					
	0,2	0,1	0,05	0,025	0,01	0,005
1						
2						
⋮	⋮	⋮	⋮	⋮	⋮	⋮
12						
13						
14				2,145		
15						

5.2.3. Intervalos de confiança para a média populacional, quando σ for desconhecido e $n \geq 30$

Quando $n \to \infty$, a distribuição t de student se aproxima da distribuição Normal. Logo ^{2} ð= s^2.

O **intervalo de confiança da média populacional**, quando σ for desconhecido e $n \geq 30$ é

$$IC(\mu)_{1-\alpha} = \bar{X} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

em que $(1-\alpha)$ é o coeficiente de confiança e $z_{\alpha/2}$ é o valor z que produz uma área de $\alpha/2$ na cauda à direita da distribuição normal padrão de probabilidade.

Exemplo 5: Uma amostra de 80 motoristas de determinado estado indica que um automóvel roda, em média, 22.000 km, com desvio padrão de 3800 km. Construa um intervalo de 98% de confiança para a rodagem anual média dos carros.

n = 80; $\bar{X}$ = 22.000; s = 3.800; α = 0,02; $z_{0,01}$ = 2.33

$$IC(\mu)_{1-\alpha} = \bar{X} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$IC(\mu)_{0,98} = 22.000 \pm 2,33\left(\frac{3800}{\sqrt{80}}\right)$$

$$IC(\mu)_{0,98} = 22.000 \pm 989,91$$

Interpretação: A verdadeira média populacional μ pertence ao intervalo [21.010,09; 22.989,91] com 98% de confiança.

Ilustração da obtenção de $z_{0,01}$ = 2,33.

$(1-\alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
0,98	0,02	0,01	2,33

Primeiro, colocamos na extremidade à direita da curva, a área 0,01.

Então, a área 0,99 (=1- 0,01) fica à esquerda da curva.

Finalmente, procura-se no interior da tabela B o valor mais próximo de 0,99 que é 0,9901, o qual correspondente ao z_0 = 2,33. Veja no esboço a seguir.

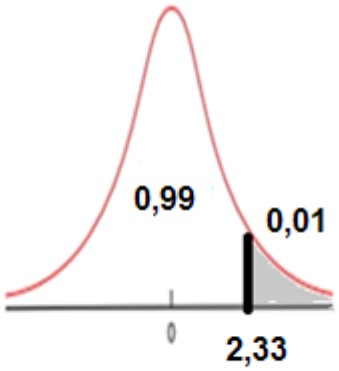


Figura 4.12: Ilustração de valores tabelados de z

5.3. Intervalos de Confiança para a proporção

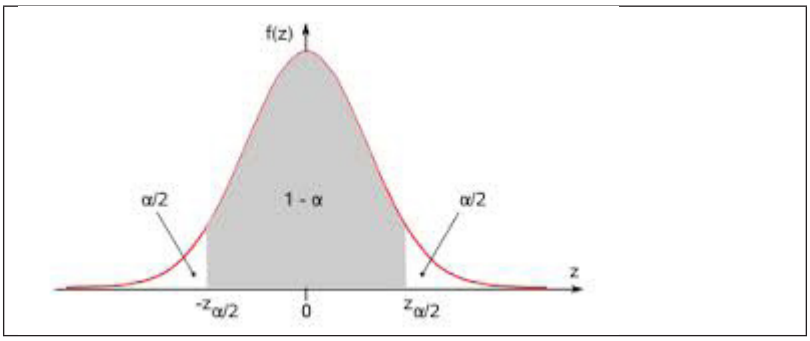
Seja *X* o número de elementos de uma amostra de tamanho *n* que apresentam a característica de interesse. Iremos agora estabelecer um intervalo de confiança para a proporção populacional *p.* Temos:

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

Quando *n* for suficientemente grande ($n \geq 30$), pelo Teorema Central do Limite, teremos que a distribuição amostral de $\hat{p}$ pode ser aproximada por meio de uma distribuição normal.

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

Então, para $(\mathbf{1} - \boldsymbol{\alpha})$ fixo, podemos construir:



Fonte: http://www.mspc.eng.br/matm/prob_est362.shtml
Figura 4.9: Ilustração da curva da distribuição normal

Logo, podemos dizer que:

$$P\left(-z_{\alpha/2} \leq z \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Utilizando propriedades algébricas, podemos isolar a proporção populacional p da expressão acima.

$$P\left(-z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \leq \hat{p} - p \leq z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\right) = 1 - \alpha$$

$$P\left(\hat{p} - z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \leq p \leq \hat{p} + z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\right) = 1 - \alpha$$

O **intervalo de confiança da proporção populacional**, quando n ≥ 30

$$IC(p)_{1-\alpha} = \hat{p} \pm z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

em que $(1-\alpha)$ é o coeficiente de confiança e $z_{\alpha/2}$ é o valor *z* que produz uma área de $\alpha/2$ na cauda à direita da distribuição normal padrão de probabilidade.

Interpretação:

A proporção populacional $p \in \left[\hat{p} - z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right);\ \hat{p} + z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\right]$

com $(1-\alpha)$100% de confiança.

Exemplo: Dois candidatos disputam as eleições para prefeito em uma cidade. 98 eleitores foram entrevistados e 53 deles pretendem votar no candidato A.

a) construir um intervalo de confiança de 95% de confiança para a proporção populacional *p* de eleitores que pretendem votar no candidato A

$\hat{p} = \frac{x}{n} = \frac{53}{98}$ = 0,54 e $(1-\hat{p})$ = 0,46 ; α = 0,05 e $z_{0,025}$ = 1,96

$$IC(p)_{1-\alpha} = \hat{p} \pm z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$IC(p)_{0,95} = 0{,}54 \pm 1{,}96\left(\sqrt{\frac{0{,}54\ (0{,}46)}{98}}\right)$$

$$IC(p)_{0,95} = 0{,}54 \pm 0{,}1 IC(p)_{0,95} = 0{,}54 \pm 0{,}1.$$

Interpretação: a verdadeira proporção populacional *p* pertence ao intervalo [0,44; 0,64] com 95% de confiança.

De modo análogo, podemos obter o intervalo de confiança para a proporção ***p***, quando *n* for pequeno.

O **intervalo de confiança da proporção populacional**, quando n < 30

$$IC(p)_{1-\alpha} = \hat{p} \pm t_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

em que $(1-\alpha)$ é o coeficiente de confiança e $t_{\alpha/2}$ é o valor *t* que produz uma área de $\alpha/2$ na cauda à direita da distribuição *t* de student, com φ = *n* -1 graus de liberdade.

5.4. Intervalos de Confiança para a variância

Sabe-se que a estatística

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

tem distribuição qui-quadrado com φ = *n* -1 graus de liberdade.

Veja a forma da distribuição qui-quadrado a seguir:

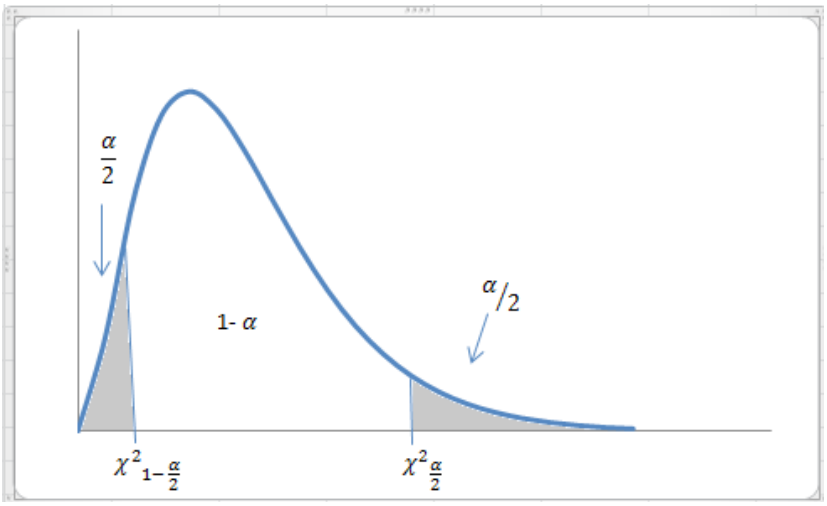


Figura 4.10: Ilustração da curva da distribuição qui-quadrado

Se $\chi^2_{\alpha/2}$ e $\chi^2_{1-\alpha/2}$ são valores da distribuição qui-quadrado que deixam áreas $\frac{\alpha}{2}$ e (1 - $\frac{\alpha}{2}$) respectivamente à direita, temos:

$$P\left[\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

Invertendo a equação teremos :

$$P\left[\frac{1}{\chi^2_{1-\alpha/2}} > \frac{\sigma^2}{(n-1)s^2} > \frac{1}{\chi^2_{\frac{\alpha}{2}}}\right] = 1 - \alpha$$

Logo:

$$P\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} > \sigma^2 > \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}\right] = 1 - \alpha$$

Ou:

$$P\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right] = 1 - \alpha$$

O **intervalo de confiança para a variância populacional** é:

$$IC(\sigma^2)_{1-\alpha} = \left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$$

em que $(1-\alpha)$ é o coeficiente de confiança, s^2 é a variância de uma amostra de tamanho n e $\chi^2_{\alpha/2}$ e $\chi^2_{1-\alpha/2}$ são os valores de χ que produz uma área de $\frac{\alpha}{2}$ e $\left(1-\frac{\alpha}{2}\right)$, respectivamente, à direita da distribuição qui-quadrado com φ = n -1 graus de liberdade.

Exemplo: Têm-se os seguintes pesos, em gramas, de 10 pacotes postais, remetidos por certa empresa:

46,4 46,1 45,8 47,0 46,1 45,9 45,8 46,9 45,2 46,0

Então $\bar{X}$ = 46,12 (média amostral), s^2 = 0,286 (variância amostral) e s = 0,535 (desvio padrão amostral). Admitindo ser normal a distribuição dos pesos, determinar um intervalo de 95% de confiança para a variância de todos os pacotes (população) expedidos pela empresa.

Resolução:

$\varphi = n - 1 = 9$ gl; $\chi^2_{0,025} = 19$ e $\chi^2_{0,975} = 2,7$, conforme **Tabela D.**

$$IC(\sigma^2)_{1-\alpha} = \left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$$

$$IC(\sigma^2)_{0,95} = \left[\frac{(n-1)s^2}{\chi^2_{0,025}}; \frac{(n-1)s^2}{\chi^2_{0,975}}\right]$$

$$IC(\sigma^2)_{0,95} = \left[\frac{9\ (0,286)}{19}; \frac{9(0,286)}{2,7}\right]$$

$$IC(\sigma^2)_{0,95} = [0,1355; 0,9533]$$

Interpretação: a verdadeira variância populacional σ^2 pertence ao intervalo $[0,1355; 0,9533]$ com 95% de confiança.

Todos os intervalos de confiança acima podem ser construídos apenas para dados provenientes de população normal.

Ilustração da obtenção de $\chi^2_{0,025}$ = 19 e $\chi^2_{0,975}$ = 2,7.

Se φ = *n* -1 = 9 graus de liberdade.

Quando área à direita da curva é 0,025. Então, da tabela D, temos $\chi^2_{0,025}$ = 19.

Quando área à direita da curva é 0,975. Então, da tabela D, temos $\chi^2_{0,975}$ = 2,7

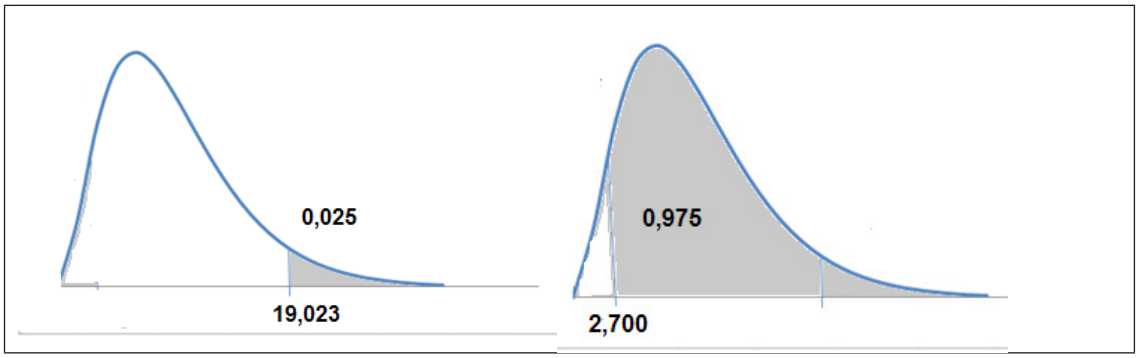


Figura 4.15: Ilustração de valores tabelados da distribuição qui-quadradro

Tabela 4.2: Reprodução parcial da tabela D

Graus de liberdade	Área da cauda direita							
	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1								
2								
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8								
9			2.700			19.023		
⋮			⋮			⋮		

5.5. Tamanho da amostra

Se o objetivo de um estudo for estimar a média populacional, então podemos calcular o tamanho da amostra para uma margem de erro (*E*) pré-fixada.

Seja a equação do intervalo de confiança $IC(\mu)_{1-\alpha} = \bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

Então, a margem de erro (*E*) deve ser no máximo $\left[z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$.

Logo,

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq E$$

Portanto, isolando *n* da equação anterior, temos:

$$n \geq \left[\frac{(z_{\alpha/2})\sigma}{E}\right]^2$$

Exemplo:

Em um estudo para a determinação do perfil dos alunos de um colégio, sabe-se de outros estudos em que a altura tem desvio padrão 0,3. Qual deve ser o tamanho da amostra para que tenhamos 95% de confiança em que o erro da estimativa da média das alturas não supere 0,05?

σ = 0,3; $z_{0,025}$ = 1,96 e *E*= 0,05

$$n \geq \left[\frac{(z_{0,025})\sigma}{E}\right]^2$$

$$n \geq \left[\frac{(1,96)0,3}{0,05}\right]^2$$

$n \geq [11,76]^2$ =138,30

Portanto, *n* ≥139 amostras.

Se o objetivo de um estudo for estimar a proporção populacional, então podemos calcular o tamanho da amostra para uma margem de erro (*E*) pré-fixada.

Seja a equação do intervalo de confiança

$$IC(p)_{1-\alpha} = \hat{p} \pm z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Então, a margem de erro (*E*) deve ser no máximo $z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$.

Logo,

$$z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \leq E$$

Portanto, isolando *n* da equação anterior, temos:

$$n \geq \left[\frac{(z_{\alpha/2})}{E}\right]^2 \hat{p}(1-\hat{p})$$

Exemplo 10: Sabe-se de um estudo preliminar em que 68% das famílias possuem televisor LCD. Se quisermos estimar a proporção populacional de famílias que possuem televisor LCD, qual o tamanho da amostra necessário para que tenhamos 95% de confiança em que o erro de nossa estimativa não seja superior a 0,02?

$\hat{p}$ = 0,68; $z_{0,025}$ = 1,96 e *E*= 0,02

$$n \geq \left[\frac{(z_{\alpha/2})}{E}\right]^2 \hat{p}(1-\hat{p})$$

$$n \geq \left[\frac{1,96}{0,02}\right]^2 0,68(0,32)$$

$n \geq 2089,83$.

Portanto, $n \geq 2090$.

II - SÍNTESE

Neste módulo, apresentamos os conceitos de amostragem aleatória e distribuições amostrais. Foram apresentados vários métodos de amostragem; amostragem aleatória simples, amostragem sistemática, amostragem por conglomerado, amostragem estratificada. Os dados amostrais podem ser usados para obtenção de estimativas por pontos e por intervalo para os parâmetros populacionais (média, proporção e variância).

IV - REFERÊNCIAS

FARIAS, A. A. ; CÉSAR,C. C.; SOARES, J. F. **Introdução à Estatística. 2 ed**. Rio de Janeiro: LTC, 2003.

LARSON, R; FABER, B. **Estatística Aplicada.** 4 ed. São Paulo: Pearson Prentice Hall, 2010.

MÓDULO 4

Teste de hipóteses, Correlação e Regressão e Aplicações

Conteúdos básicos do Módulo:

1. Teste de hipóteses
2. Correlação
3. Regressão linear simples

Objetivos do Módulo:

Ao final deste estudo, esperamos que você, aluno(a), possa:

- Verificar afirmações estatísticas
- Estabelecer a relação linear entre duas variáveis.

TESTE DE HIPÓTESES, CORRELAÇÃO E REGRESSÃO E APLICAÇÕES

I - TEXTO BÁSICO

6. Testes de hipóteses

No módulo anterior, começamos a estudar inferência estatística e vimos como obter um intervalo de confiança para um parâmetro populacional, por exemplo, para a média (μ), variância (σ^2) e proporção (p). Neste módulo, continuaremos estudando inferência estatística, mas, aqui, aprenderemos como utilizar um teste de hipóteses para testar se uma afirmação sobre um parâmetro é verdadeira ou não.

6.1. Introdução : conceitos fundamentais e tipo de erros

6.1.1. Conceitos fundamentais.

Teste de hipóteses é uma regra de decisão para aceitar ou rejeitar uma hipótese estatística com base nos dados amostrais.

Hipótese estatística é uma suposição quanto ao valor de um parâmetro da população, a qual será verificada por um teste. São exemplos de hipóteses estatísticas:

1) A altura média dos brasileiros é de 1, 65 m. Ou, usando a nomenclatura estatística, $\mu = 1{,}65$.

2) A altura média dos brasileiros é menor que 1, 65 m ($\mu < 1{,}65$).

3) 10% dos brasileiros são fumantes ($p = 0{,}1$).

Notação: H é a notação de hipótese estatística

A hipótese H_0 é a **hipótese nula**, que é a hipótese estatística que será testada. A hipótese nula contém a afirmação sobre o parâmetro populacional.

A hipótese H_1 é a **hipótese alternativa**, que é o complemento da hipótese nula.

Exemplos:

H_0: μ = 1,65 hipótese nula

H_1: $\mu \neq 1,65$ hipótese alternativa

H_0: $\mu = 1,65$ hipótese nula

H_1: $\mu < 1,65$ hipótese alternativa

H_0: $\mu = 1,65$ hipótese nula

H_1: $\mu > 1,65$ hipótese alternativa

Na hipótese nula, é comum o emprego do sinal de "=". Esta opção será adotada aqui. Mas, também podem ser usados " $\leq$ "ou " $\geq$ ".

6.1.2. Tipo de erros

Na hipótese alternativa podem ser utilizados os sinais de"$\neq$" ou "<" ou ">".

Num teste de hipótese, toma-se como referência a hipótese nula (H_0) e haverá duas decisões possíveis: aceitar H_0 ou rejeitar H_0.

Ao aceitar H_0, estamos, consequentemente, rejeitando H_1. Ao rejeitar H_0, estamos consequentemente aceitando H_1. Essa decisão está sujeita a erros, já que a hipótese H_0 pode ser falsa ou verdadeira. Assim, podemos rejeitar H_0, quando H_0 é verdadeira. Veja a seguir

Erro tipo I: consiste em rejeitar H_0 quando H_0 é verdadeira

Erro tipo II: consiste em aceitar H_0 quando H_0 é falsa

α**:**(nível de significância): é a probabilidade de se cometer o erro do tipo I.

ß: é a probabilidade de se cometer o erro tipo II

(1 - ß): é o poder do teste.

Quadro 6.1: Erro tipo I e erro tipo II

	DECISÃO	
REALIDADE	Aceitar H_0	Rejeitar H_0
H_0 é verdadeira	decisão correta **(1 - α)**	**Erro tipo I - α**
H_0 é falsa	**Erro tipo II - ß**	decisão correta - **(1 - ß)**

A situação ideal seria aquela em que ambas as probabilidades α e ß fossem zero. Entretanto à medida que se diminui α , ß tende a aumentar. Deve-se ter cuidado ao efetuar um teste de hipótese para que o erro mais importante a ser evitado seja o erro tipo I. Os níveis de significância mais usados são α = 0,01; 0,05 e 0,1.

6.1.6. Procedimento para se efetuar um teste de hipóteses

1) Identificar as hipóteses do teste (H_0 e H_1) .

2) Especificar o nível de significância (α)

3) Obter o valor crítico ($Z_{tabelado}$ ou $T_{tabelado}$ ou $\chi^2_{tabelado}$) e esboçar a região crítica (região de rejeição).

4) Calcular a estatística do teste.

5) Decisão. "Através da comparação da estatística do teste com o valor crítico da região crítica"

6) Interpretar a decisão de forma contextualizada.

6.2. Teste de hipótese para a média

A média de uma população é uma das características mais importantes. Assim sendo, frequentemente temos de tomar decisões a seu respeito**.**

6.2.1. Teste de hipótese para a média, σ conhecido

O **teste de hipóteses para a média** μ (ou **teste Z para média** μ) é um teste estatístico para a média populacional. O **teste Z para média** μ pode ser usado quando a população é normal e σ conhecido.

O desenvolvimento do teste Z para média μ depende das hipóteses que serão testadas. Se as hipóteses forem $H_0: \mu = \mu_0$ e $H_1: \mu < \mu_0$, então teremos **teste bilateral** com duas regiões de rejeição (região crítica) e dois valores críticos ($Z_{tabelado} = -z_{\frac{\alpha}{2}}$ e $z_{\frac{\alpha}{2}}$)

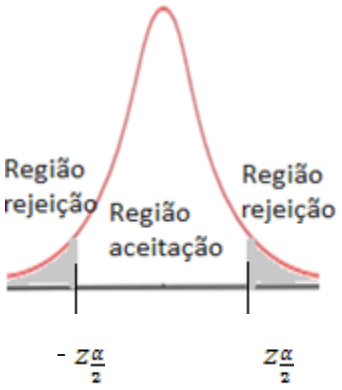


Figura 6.1. Região crítica e valores críticos para teste bilateral

Se as hipóteses forem $H_0: \mu = \mu_0$ e $H_1: \mu < \mu_0$, então teremos **teste unilateral à esquerda** com uma região de rejeição à esquerda (região crítica) e um valor crítico ($Z_{tabelado} = -z_{\alpha}$)

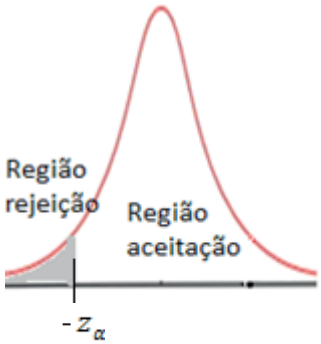


Figura 6.2. Região crítica e valor crítico para teste unilateral à esquerda

Se as hipóteses forem $H_0: \mu = \mu_0$ e $H_1: \mu > \mu_0$, então teremos **teste uniateral à direita** com uma região de rejeição à direita (região crítica) e um valor crítico ($Z_{tabelado} = z_\alpha$)

Figura 6.3. Região crítica e valor crítico para teste unilateral à direita

A **estatística do teste** de hipóteses da média populacional μ, com σ conhecido é:

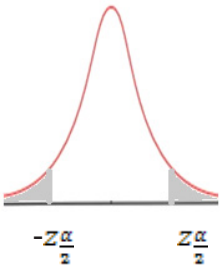
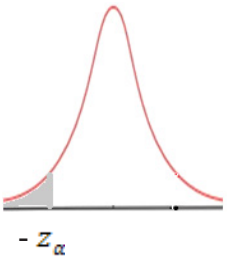
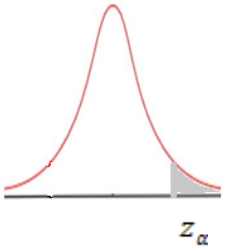
$$Z_{calculado} = Z_{calc} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

em que μ_0 é o valor de referência proveniente da afirmação a ser testada.

Finalmente, tome a decisão de rejeitar H_0 se Z_{calc} estiver na região de rejeição.

Veja o resumo teste Z para média μ no **Quadro 6.2**.

Quadro 6.2: Resumo do teste Z para média μ

Teste	Bilateral	Unilateral à esquerda	Unilateral à direita
1. Hipóteses	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$
2. α	fixar α	fixar α	fixar α
3. Região crítica			
4. Estatística do teste	$Z_{calc} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$Z_{calc} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$Z_{calc} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
5. Decisão: rejeita-se H_0	$z_{calc} < -z_{\alpha/2}$ ou $z_{calc} > z_{\alpha/2}$	$z_{calc} < -z_\alpha$	$z_{calc} > z_\alpha$

Exemplo:

O governo de um país afirma que a expectativa de vida (média) é maior do que 70 anos. Uma amostra aleatória com 100 registros de mortes durante o ano passado mostrou que a expectativa de vida é 71,8 anos . Assumindo um desvio padrão de 8,9 anos (conhecido de estudos anteriores), isso indica que a expectativa de vida média é maior do que 70 anos? Use nível de significância de 5%.

Resolução:

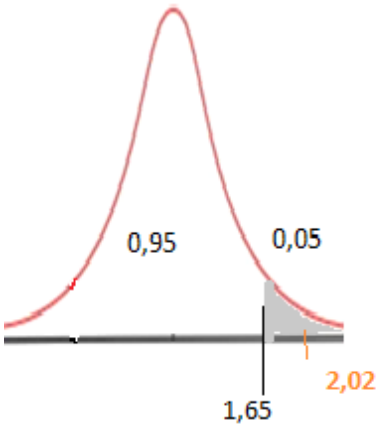
As informações em relação à amostra são : n = 100 e $\bar{x}$ = 71,8 anos

A informação em relação à população é: σ = 8,9 anos.

Queremos testar se $\mu > 70$ anos (afirmação do governo). Logo, o valor de referência é μ_0 = 70.

Como o nível de significância é 5%, então α = 0,05.

Quadro 6.3: Exemplo do teste Z para média μ

Teste	**Unilateral à direita**
1. Hipóteses	H_0: $\mu = 70$ H_1: $\mu > 70$ " afirmação do governo"
2. α	$\alpha = 0{,}05$
3. Região crítica	 $z_{0,05} = 1{,}65$, na tabela B da distribuição normal padrão
4. Estatística do teste	$Z_{calc} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{71{,}8 - 70}{8{,}9/\sqrt{100}} = 2{,}02$
5. Decisão:	Rejeita-se H_0: $\mu = 70$, ao nível de significância de 5%. Pois, 2,02 > 1,65 (2,02 está na região de rejeição de H_0).

Interpretação: Portanto, confirma-se que a expectativa de vida (média) é maior do que 70 anos, ao nível de significância de 5 %. A afirmação do governo está correta.

A decisão pode ser diferente se usarmos outros valores de α (nível de significância).

Na decisão e na interpretação, você deve informar o nível de significância utilizado no teste.

6.2.2. Teste de hipótese para a média, σ desconhecido e $n < 30$

O **teste *t* para média** μ é um teste para a média populacional e pode ser usado quando a população é normal e σ desconhecido.

O desenvolvimento do teste *t* para média μ também depende das hipóteses que serão testadas. Se as hipóteses forem H_0: $\mu = \mu$ e H_1: $\mu < \mu_0$, então, teremos **teste bilateral** com duas regiões de rejeição (região crítica) e dois valores críticos ($T_{tabelado} = -t_{\frac{\alpha}{2}}$ e $t_{\frac{\alpha}{2}}$)

Figura 6.4. Região crítica e valores críticos para teste bilateral

Na figura 6.4, $t_{\alpha/2}$ é o valor *t* que produz uma área de $\alpha/2$ na cauda à direita da distribuição *t* de student, com $\varphi = n$ -1 graus de liberdade, na tabela C.

Se as hipóteses forem H_0: $\mu = \mu_0$ e H_1: $\mu < \mu_0$, então teremos **teste unilateral à esquerda** com uma região de rejeição à esquerda (região crítica) e um valor crítico ($T_{tabelado} = -t_{\alpha}$)

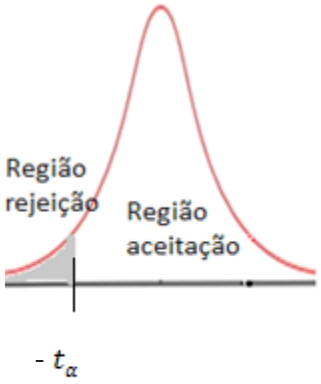


Figura 6.5. Região crítica e valor crítico para teste unilateral à esquerda

Na figura 6.5, - t_α é o valor simétrico ao valor de t_α , o qual produz uma área de α na cauda à direita da distribuição *t* de student, com $\varphi = n - 1$ graus de liberdade, na tabela C.

Se as hipóteses forem H_0: $\mu = \mu_0$ e H_1: $\mu > \mu_0$, então teremos **teste bilateral à direita** com uma região de rejeição à direita (região crítica) e um valor crítico ($T_{tabelado} = t_\alpha$)

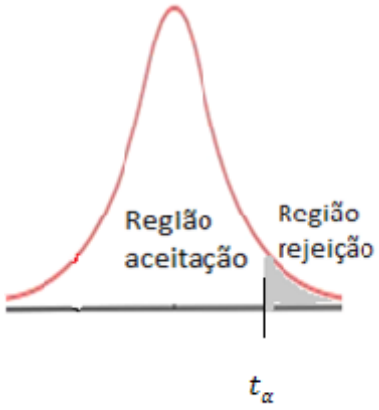


Figura 6.6. Região crítica e valor crítico para teste unilateral à direita

Na figura 6.6, t_α é o valor de t que produz uma área de α na cauda à direita da distribuição *t* de student, com $\varphi = n - 1$ graus de liberdade, na tabela C.

A **estatística do teste** de hipóteses da média populacional μ , com σ desconhecido e $n < 30$ é:

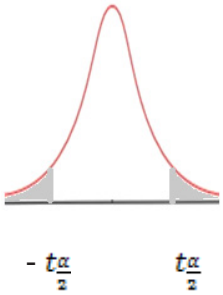
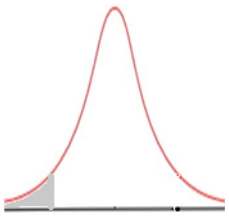
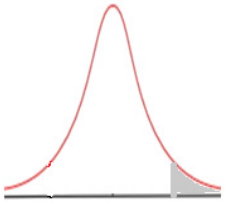
$$T_{calculado} = T_{calc} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

em que μ_0 é o valor de referência proveniente da afirmação a ser testada.

Finalmente, tome a decisão de rejeitar H_0 se T_{calc} estiver na região de rejeição.

Veja o resumo teste *t* para média μ no **Quadro 6.4**.

Quadro 6.4: Resumo do teste *t* para média μ

Teste	Bilateral	Unilateral à esquerda	Unilateral à direita
1. Hipóteses	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$
2. $\alpha\alpha$	fixar α	fixar α	fixar α
3. Região crítica	 $-t_{\frac{\alpha}{2}}$ $t_{\frac{\alpha}{2}}$	 $-t_\alpha$	 t_α
4. Estatística do teste	$T_{calc} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$T_{calc} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$T_{calc} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
5. Decisão: rejeita-se H_0	$T_{calc} < -t_{\alpha/2}$ ou $T_{calc} > t_{\alpha/2}$	$T_{calc} < -t_\alpha$	$T_{calc} > t\alpha$

Exemplo:

Uma indústria afirma que a média do nível do pH na água do rio mais próximo é de 6,8. Você seleciona 19 amostras de água e mede os níveis de pH de cada uma, as quais apresentam média amostral é 6,7 e desvio padrão 0,24. Ao nível de significância de 5%, verifique se a afirmação da indústria está correta.

Resolução:

As informações sobre a amostra são: $n = 19$ e $\bar{x} = 6,7$ e $s = 0,24$

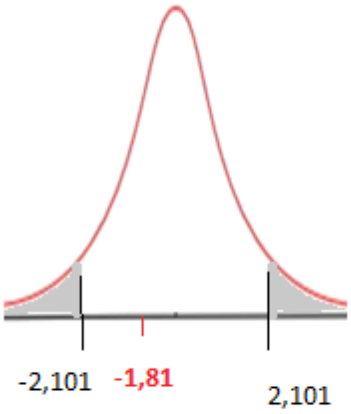
Não há informação sobre a população.

Para nível de significância é 5% ($\alpha = 0,05$), queremos testar se $\mu = 6,8$ (alegação da indústria). Logo, o valor de referência é $\mu_0 = 6,8$. Neste exemplo, podemos escolher

APENAS um teste: o teste bilateral ou o teste unilateral à esquerda.

Então, para o teste bilateral, teríamos:

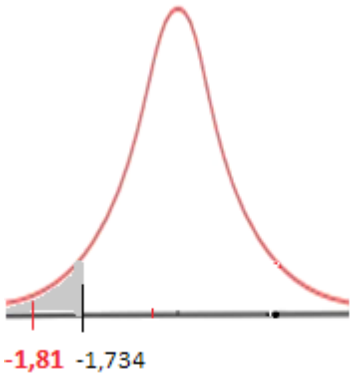
Quadro 6.5: Exemplo do teste *t* para média μ

Teste	**Bilateral**
1. Hipóteses	H_0: $\mu = 6,8$ "Afirmação da indústria" H_1: $\mu \neq 6,8$
2. α	$\alpha = 0,05$.
3. Região crítica	 $t_{\frac{\alpha}{2}} = t_{0,025} = 2,101$, com 18 graus de liberdade (na tabela C, da distribuição *t*, $\varphi = n$ -1 graus de liberdade)
4. Estatística do teste	$T_{calc} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{6,7 - 6,8}{0,24/\sqrt{19}} = -1,81$
5. Decisão	Aceita-se H_0: $\mu = 6,8$, ao nível de significância de 5%. Pois -2,101 $\leq$ -1,81 $\leq$ 2,101 (- 1,81 está na região de aceitação de H_0).

Interpretação: Portanto, confirma-se que a média do nível do pH na água do rio é de 6,8, ao nível de significância de 5%. A afirmação da indústria está correta.

E para o teste unilateral à esquerda, teríamos:

Quadro 6.6: Exemplo do teste *t* para média μ

Teste	**Unilateral à esquerda**
1. Hipóteses	H_0: $\mu = 6,8$ "Afirmação da indústria" H_1: $\mu < 6,8$
2. α	$\alpha = 0,05$.
3. Região crítica	 $-t_\alpha = -t_{0,05} = -1,734$, com 18 graus de liberdade (na tabela C, da distribuição *t*, $\varphi = n - 1$ graus de liberdade)
4. Estatística do teste	$T_{calc} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{6,7 - 6,8}{0,24/\sqrt{19}} = -1,81$
5. Decisão:	Rejeita-se H_0: $\mu = 6,8$, ao nível de significância de 5%. Pois, -1,81 < -1,734 (- 1,81 está na região de rejeição de H_0).

Interpretação: Portanto, a média do nível do pH na água do rio é menor que 6,8, ao nível de significância de 5%. A afirmação da indústria está errada.

6.2.3. Teste de hipótese para a média, σ desconhecido e $n \geq 30$

Quando $n \to \infty$, a distribuição *t* de Student se aproxima da distribuição normal. Logo $\sigma = s$.

A **estatística do teste** de hipóteses da média populacional μ , com σ desconhecido e $n \geq 30$ é:

$$Z_{calculado} = Z_{calc} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

em que $\sigma = s$ e μ_0 é o valor de referência proveniente da afirmação a ser testada.

A execução do teste *Z* para média μ, quando σ é desconhecido e $n \geq 30$ segue o procedimento apresentado no **Quadro 6.2**.

Exemplo:

Em um anúncio, uma pizzaria afirma que a média de seu tempo de entrega é menor que 30 minutos. Uma seleção aleatória de 36 tempos de entrega tem média amostral de 28,5 minutos e desvio padrão de 3,5 minutos. Você concorda com a afirmação da pizzaria ao nível de significância de 1%?

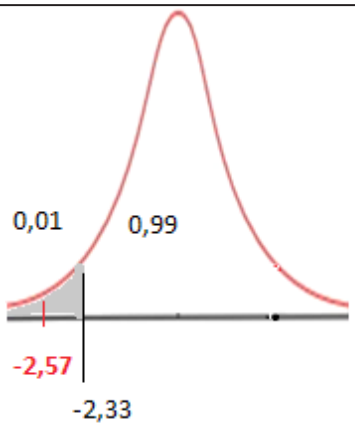
Resolução:

As informações sobre a amostra são: $n = 36$ e $\bar{x} = 28{,}5$ e $s = 3{,}5$.

Não há informação sobre a população, no entanto, como $n \geq 30$, o σ pode ser estimado por *s*, logo $\sigma = s = 3{,}5$.

Para nível de significância de 1% ($\alpha = 0{,}01$), queremos testar se $\mu < 30$ (alegação da pizzaria). Logo, o valor de referência é $\mu_0 = 28{,}5$. Então, teremos o teste unilateral à esquerda.

Quadro 6.7: Exemplo do teste *Z* para média μ

Teste	**Unilateral à esquerda**
1. Hipóteses	H_0: $\mu = 30$ H_1: $\mu < 30$ "afirmação da pizzaria"
2. α	$\alpha = 0{,}01$
3. Região crítica	 - $z_{0,01}$= -2,33, pois $z_{0,01}$= 2,33 na tabela B, da distribuição normal padrão.
4. Estatística do teste	$Z_{calc} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{28{,}5 - 30}{3{,}5/\sqrt{36}} = -2{,}57$
5. Decisão:	Rejeita-se H_0: $\mu = 30$ ao nível de significância de 1%. Pois, -2,57 < -2,33 (-2,57 está na região de rejeição de H_0).

Interpretação: Portanto, confirma-se a média de seu tempo de entrega é menor que 30 minutos, ao nível de significância de 1%. A afirmação da pizzaria está correta.

6.3. Teste de hipótese para a proporção

Seja *X* o número de elementos de uma população de tamanho *N* que apresenta a característica de interesse. A proporção populacional é $p = \frac{X}{N}$.

Seja *x* o número de elementos de uma amostra de tamanho *n* que apresenta a característica de interesse. A proporção amostral é $\hat{p} = \frac{x}{n}$.

O **teste Z para a proporção *p*** é um teste estatístico para uma proporção populacional *p*.

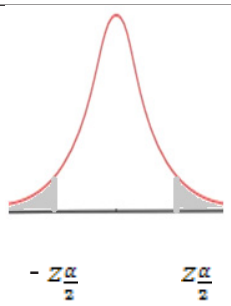
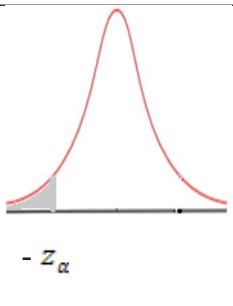
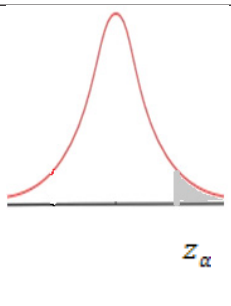
A **estatística do teste** de hipóteses para proporção, com $n \geq 30$ é:

$$Z_{calculado} = Z_{calc} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

em que p_0 é o valor de referência proveniente da afirmação a ser testada.

Veja o resumo **teste Z para a proporção *p*** no **Quadro 6.8**.

Quadro 6.8: Resumo do teste *Z* para a proporção *p*

Teste	Bilateral	Unilateral à esquerda	Unilateral à direita
1. Hipóteses	$H_0: p = p_0$ $H_1: p \neq p_0$	$H_0: p = p_0$ $H_1: p < p_0$	$H_0: p = p_0$ $H_1: p > p_0$
2. α	fixar α	fixar α	fixar α
3. Região crítica			
4. Estatística do teste	$Z_{calc} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$Z_{calc} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$Z_{calc} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
5. Decisão: rejeita-se H_0	$z_{calc} < -z\alpha_{/2}$ ou $z_{calc} > z\alpha_{/2}$	$z_{calc} < -z\alpha$	$z_{calc} > z\alpha$

Exemplo: Os médicos acreditam que um medicamento comumente prescrito para aliviar a tensão nervosa tem 60% de eficácia. O laboratório não concorda e decide investigar essa eficácia. Os resultados experimentais com o medicamento administrado em uma amostra aleatória de 100 adultos que sofrem de tensão nervosa mostraram que 70 deles sentiram alívio. Você concorda com os médicos, ao nível de significância de 5%?

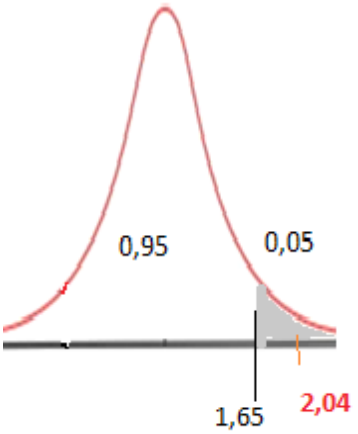
Resolução:

As informações da amostra são: n = 100 e $\hat{p} = \frac{X}{n} = \frac{70}{100}$ = 0,70

Queremos testar se p = 0,6 (afirmação dos médicos). Logo, o valor de referência é p_0 = 0,6.

Como nível de significância é 5%, então α = 0,05.

Quadro 6.9: Exemplo do teste Z para a proporção p

Teste	**Unilateral à direita**
1. Hipóteses	H_0: p = 0,6 "Alegação dos médicos" H_1: p > 0,6
2. α	α = 0,05
3. Região crítica	 $z_{0,05} = 1,65$, na tabela B da distribuição normal padrão
4. Estatística do teste	$Z_{calc} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0,7 - 0,6}{\sqrt{\frac{0,6(0,4)}{100}}} = 2,04$
5. Decisão:	Rejeita-se H_0: p = 0,6 , ao nível de significância de 5%. Pois, 2,04 > 1,65 (2,04 está na região de rejeição de H_0).

Interpretação: Portanto, um medicamento comumente prescrito para aliviar a tensão nervosa tem eficácia maior que 60%, ao nível de significância de 5%. A afirmação dos médicos está errada.

6.4. Teste de hipóteses para independência

Na seção 3, aprendemos que dois eventos são independentes se a ocorrência de um evento não afeta a probabilidade da ocorrência de outro evento. Por exemplo, os resultados do lançamento de dados e de moedas são independentes. Mas suponha que um médico pesquisador queira determinar se há uma relação entre o consumo de cafeína e o risco de ataque do coração ou que um professor queira saber se há uma relação entre o desempenho em matemática (notas A, B e C) e o turno do aluno (manhã e noite). Essas variáveis são independentes ou dependentes?

Aqui você vai aprender como usar o teste qui-quadrado para independência de modo a responder esse tipo de pergunta. Para fazer o teste qui-quadrado para independência, você vai usar amostras de dados que são organizados em uma tabela de contingência.

Uma **tabela de contingência *r* x *c*** mostra as frequências observadas para duas variáveis. As frequências observadas são arranjadas nas linhas ***r*** e nas colunas ***c***. A interseção de uma linha e uma coluna é chamada **célula.**

Tabela 6.1: Tabela de contingência $r \times c$

	Colunas			
Linhas	**1**	**2**		***c***
1	O_{11}	O_{12}	$\cdots$	O_{1c}
2	O_{21}	O_{22}	$\cdots$	O_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
r	O_{r1}	O_{r2}	$\cdots$	O_{rc}

O **teste qui-quadrado para independência** é usado para testar a independência de duas variáveis.

Para usar o teste qui-quadrado para independência, as seguintes condições devem ser satisfeitas:

As frequências observadas devem ser obtidas usando uma amostra aleatória.

Cada frequência esperada deve ser maior ou igual a 5.

A região crítica será aproximada pela distribuição qui-quadrado com $\varphi = (r-1)(c-1)$ graus de liberdade, em que r é o numero de linhas e c é o número de colunas da tabela de contingência.

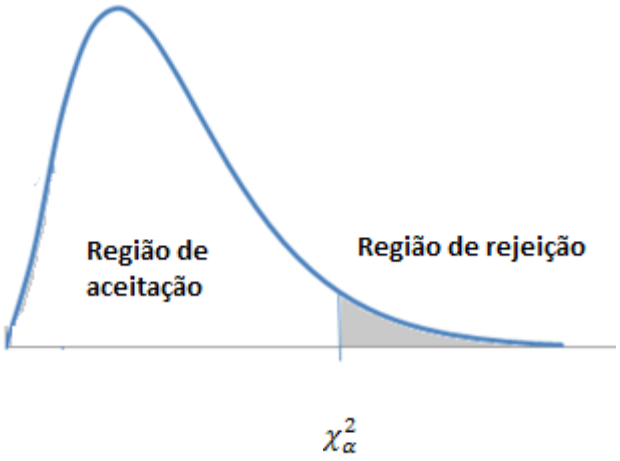


Figura 6.7. Região crítica e valor crítico para teste qui-quadrado

Na figura 6.7, α corresponde à área da cauda direita da curva da distribuição qui-quadrado.

A **estatística do teste** para o teste qui-quadrado de independência é

$$\chi^2_{calculdado} = \chi^2_{calc} = \sum \frac{(O-E)^2}{E}$$

Em que O representa as frequências observadas e E representa as frequências esperadas.

A frequência esperada de uma célula $E_{r,c}$, em uma tabela de contingência é:

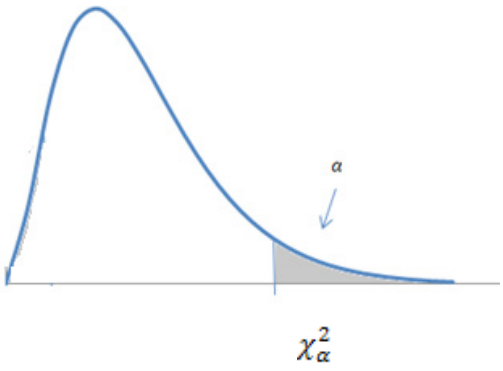
$$E_{r,c} = \frac{(Soma\ da\ linha\ r) \times (Soma\ da\ coluna\ c)}{Total\ da\ amostra}$$

Tabela 6.2: Tabela de contingência *r X c*, com frequências esperadas

Linhas	Colunas			
	1	2		*c*
1	$O_{11}(E_{11})$	$O_{12}(E_{12})$	⋯	$O_{1c}(E_{1c})$
2	$O_{21}(E_{21})$	$O_{22}(E_{22})$	⋯	$O_{2c}(E_{2c})$
⋮	⋮	⋮	⋱	⋮
r	$O_{r1}(E_{r1})$	$O_{r2}(E_{r2})$	⋯	$O_{rc}(E_{rc})$

Veja o resumo teste qui-quadrado para independência no **Quadro 6.10**.

Quadro 6.10: Resumo do teste qui-quadrado para independência

1. Hipóteses	H_0: As variáveis X e Y são independentes H_1: As variáveis X e Y não são independentes.
2. α	fixar α
3. Região crítica	
4. Estatística do teste.	$\chi^2_{calc} = \sum \frac{(O-E)^2}{E}$
5.Decisão: rejeita-se H_0	$\chi^2_{calc} > \chi^2_\alpha$

Exemplo: Os resultados de uma amostra aleatória de 237 alunos, que foram classificados quanto ao desempenho em matemática e turno em que frequentam a escola, estão apresentados na tabela de contingência. Com nível de significância de 5%, teste a hipótese de que o desempenho e o turno dos alunos são independentes.

Tabela 6.3: Frequências observadas do desempenho em matemática, por turno

	Desempenho em matemática			
Turno	**A**	**B**	**C**	**TOTAL**
Manhã	45	60	40	145
Noite	20	50	22	92
TOTAL	65	110	62	237

Resolução:

Na tabela de contingência, com $r = 2$ linhas e $c = 3$ colunas, identificamos as frequências observadas O_{rc}:

$O_{11} = 45$; $O_{12} = 60$; $O_{13} = 40$;

$O_{21} = 20$; $O_{22} = 50$; $O_{23} = 22$;

Agora, devemos encontrar as frequências esperadas:

$$E_{r,c} = \frac{(Soma\ da\ linha\ r) \times (Soma\ da\ coluna\ c)}{Total\ da\ amostra}$$

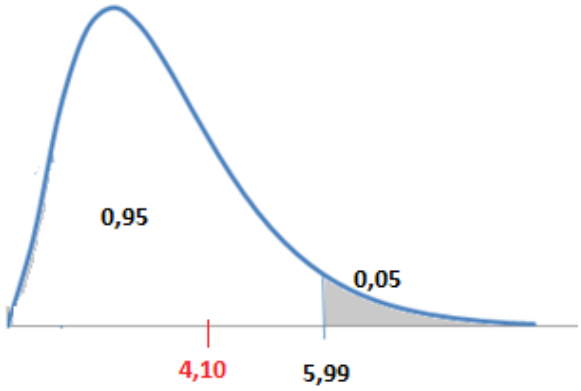
$$E_{1,1} = \frac{(145)(65)}{237} = 39{,}77; \quad E_{1,2} = \frac{(145)(110)}{237} = 67{,}30; \quad E_{1,3} = \frac{(145)(62)}{237} = 37{,}93$$

$$E_{2,1} = \frac{(92)(65)}{237} = 25{,}23; \quad E_{2,2} = \frac{(92)(110)}{237} = 42{,}70; \quad E_{2,3} = \frac{(92)(62)}{237} = 24{,}07$$

Na tabela de contingência a seguir, apresentamos as frequências observadas e a respectiva frequência esperada, em cada célula.

Tabela 6.4: Frequências observadas e esperadas do desempenho em matemática, por turno

	Desempenho em matemática			
Turno	**A**	**B**	**C**	**TOTAL**
Manhã	45 (39,77)	60(67,30)	40(37,93)	145
Noite	20(25,23)	50 (42,70)	22(24,07)	92
TOTAL	65	110	62	237

1. Hipóteses	H_0: O desempenho em matemática e o turno dos alunos <u>são</u> independentes H_1: O desempenho em matemática e o turno dos alunos <u>não são</u> independentes
2. α	$\alpha = 0{,}05$
3. Região crítica	 $\chi^2_\alpha = \chi^2_{0,05} = 5{,}99$, com $\varphi = (r\text{-}1)(c\text{-}1) = (2\text{-}1)(3\text{-}1) = 2$ graus de liberdade (na tabela D, da distribuição qui-quadrado)
4. Estatística do teste.	$\chi^2_{calc} = \sum \frac{(O-E)^2}{E}$ $\chi^2_{calc} = \frac{(45-39{,}77)^2}{39{,}77} + \frac{(60-67{,}3)^2}{67{,}3} + \frac{(40-37{,}93)^2}{37{,}93} + \frac{(20-25{,}23)^2}{25{,}23} + \frac{(50-42{,}7)^2}{42{,}7} + \frac{(22-24{,}07)^2}{24{,}07}$ $\chi^2_{calc} = 0{,}69 + 0{,}79 + 0{,}11 + 1{,}08 + 1{,}25 + 0{,}18$ $\chi^2_{calc} = 4{,}10$
5. Decisão:	Aceita-se H_0, ao nível de significância de 5%. Pois, 4,10 < 5,99 (4,10 está na região de aceitação de H_0).

Interpretação: O desempenho em matemática e o turno dos alunos são independentes, ao nível de significância de 5%.

7. Regressão e Correlação

Em algumas situações, interessa-se pelo estudo do comportamento de duas variáveis , *X* e *Y*, ou seja, deseja-se saber se existe relação entre elas. Por exemplo, altura (*X*) e Peso (*Y*).

Esse comportamento pode ser **observado** através do **diagrama de dispersão** e **medido** através do **coeficiente de correlação linear**.

7.1 Diagrama de dispersão

O **diagrama de dispersão** é o gráfico no qual cada par (x_i, y_i) é representado com um ponto plotado em um sistema bidimensional de coordenadas. A **variável independente** (ou explanatória) *X*, é medida pelo eixo horizontal, e a **variável dependente** (resposta) *Y* é medida pelo eixo vertical.

Um diagrama de dispersão pode ser usado para determinar se existe uma **correlação linear** (linha reta) entre duas variáveis.

Os diagramas de dispersão mostram diversos tipos de correlação.

a) Correlação linear positiva: os valores de *X* e *Y* aumentam.

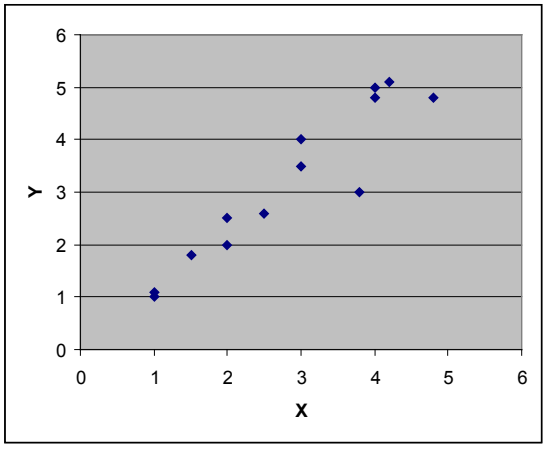


Figura 7.1. Correlação linear positiva

b) Correlação linear negativa: os valores de uma das variáveis aumentam enquanto os valores da outra decrescem.

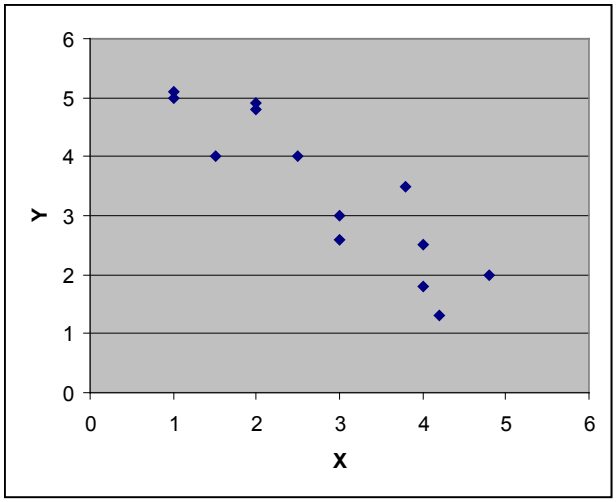


Figura 7.2. Correlação linear negativa

c) Não há correlação: não é possível identificar um comportamento para *X* e *Y*

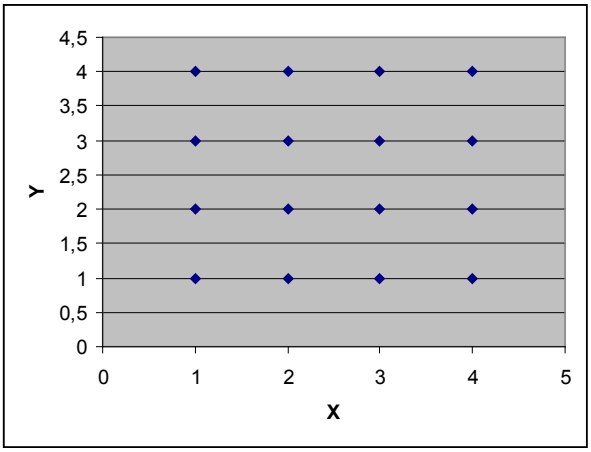


Figura 7.3. Não há correlação

d) Correlação não linear: a relação entre *X* e *Y* é do tipo não linear.

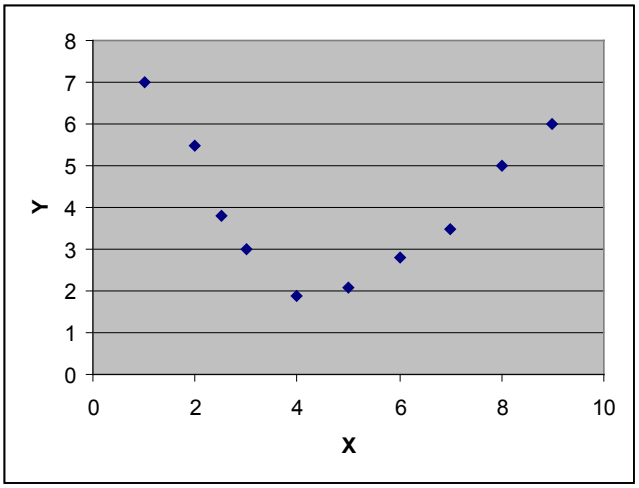


Figura 7.4. Correlação não linear

A correlação não linear não é tema de estudo nessa disciplina.

Desvantagem do diagrama de dispersão: leva a conclusões subjetivas sobre a correlação, as quais podem levar a equívocos.

7.2 Coeficiente de Correlação de Pearson

O **coeficiente de correlação de Pearson** (ou **coeficiente de correlação**) mede a relação linear entre duas variáveis

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

em que *n* é o número de pares ordenados

O coeficiente de correlação de Person pode ter valores no intervalo $-1 \leq r \leq 1$.

Se $r = 0$, então **não há correlação linear** entre as variáveis X e Y.

Se $r = 1$, então há **correlação linear positiva PERFEITA** entre as variáveis X e Y.

Se $r = -1$, então há **correlação linear negativa PERFEITA** entre as variáveis X e Y.

r = 0 indica que não há correlação linear. No entanto, existe a possibilidade de a correlação ser do tipo não linear, como situação da Figura 7.4.

Não há regra para interpretar outros valores do coeficiente de correlação. No entanto, é razoável considerar que:

r ≈ 0,9: correlação linear positiva FORTE

r ≈ 0,7: correlação linear positiva MODERADA

r ≈ 0,5: correlação linear positiva FRACA.

r ≈ -0,9: correlação linear negativa FORTE

r ≈ -0,7: correlação linear negativa MODERADA

r ≈ -0,5: correlação linear negativa FRACA

O numerador do coeficiente de correlação pode ser nomeado de S_{XY}

$$S_{XY} = \sum XY - \frac{\sum X \sum Y}{n}$$

E, no denominador, temos os termos S_{XX} e S_{YY}:

$$S_{XX} = \sum X^2 - \frac{(\sum X)^2}{n}$$

Logo, podemos escrever o coeficiente de correlação de outra maneira:

$$r = \frac{S_{XY}}{\sqrt{(S_{XX})(S_{YY})}}$$

Exemplo:

Um estudo quer verificar se há relação linear entre número de horas de estudo (X), por semana, e a nota final (Y) dos alunos, conforme mostrado na Tabela 7.1.

Tabela 7.1. Horas de estudo, por semana, e a nota final

Horas de estudo (X), por semana	Nota final (Y)
1	1
2	2
3	4
4	5
5	8

a) Obtenha o diagrama de dispersão e interprete.

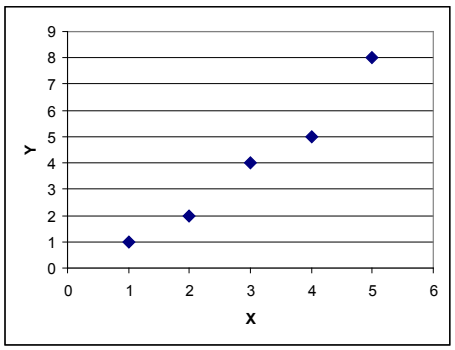


Gráfico 7.1. Gráfico de dispersão para dados da Tabela7.1

Interpretação: Suspeita-se que há correlação linear positiva.

b) Calcule o coeficiente de correlação (*r*) e interprete.

Tabela 7.2. Cálculo de $\sum X, \sum Y, \sum XY, \sum X^2$ e $\sum Y^2$

	X	*Y*	*XY*	X^2	Y^2
	1	1	1	1	1
	2	2	4	4	4
	3	4	12	9	16
	4	5	20	16	25
	5	8	40	25	64
Soma (Σ):	**15**	**20**	**77**	**55**	**110**

Para calcular o coeficiente de correlação devemos:

1. Encontrar o valor de *n*: *n* = 5. Pois, temos cinco pares de dados.
2. Encontrar a soma dos valores de *X*: $\sum X$ = 15.
3. Encontrar a soma dos valores de *Y*: $\sum Y$ = 20.
4. Encontrar a soma dos valores de XY: $\sum XY$ = 77.
5. Encontrar a soma dos valores de X^2: $\sum X^2$= 55.
6. Encontrar a soma dos valores de Y^2: $\sum Y^2$= 110.
7. Calcular o valor de $S_{XY} = \sum XY - \frac{\sum X \sum Y}{n} = 77 - \frac{(15)(20)}{5} = 17$.
8. Calcular o valor de $S_{XX} = \sum X^2 - \frac{(\sum X)^2}{n} = 55 - \frac{(15)^2}{5} = 10$.
9. Calcular o valor de $S_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 110 - \frac{(20)^2}{5} = 30$.
10. Calcular o coeficiente de correlação (*r*)

$$r = \frac{S_{XY}}{\sqrt{(S_{XX})(S_{YY})}} = \frac{17}{\sqrt{(10)(30)}} = 0{,}98$$

Interpretação: há correlação linear positiva **forte**.

No procedimento que calcula o coeficiente de correlação:

$\sum X^2 \neq (\sum X)^2$ pois , $55 \neq (15)^2$ e $\sum Y^2 \neq (\sum Y)^2$, pois $110 \neq (20)^2$

7.3. Regressão linear simples

Quando for identificado que há correlação **linear** entre duas variáveis, através do diagrama de dispersão e/ou coeficiente de correlação (*r*), torna-se necessário modelar e investigar a relação entre essas variáveis, ou seja, como a variável Y varia em função da

variável X.

$$Y = f(X) = \alpha + \beta X$$

$Y = \alpha + \beta X$ é a equação da reta , conforme estudado em álgebra.

Y é a variável dependente (ou variável resposta)

X é a variável independente (ou explicativa ou exógena).

Por exemplo, o peso de crianças (Y) varia conforme a idade (X).

7.3.1 Modelo matemático

Podemos obter o seguinte modelo para representar um conjunto de pontos:

$$\ddot{Y}_i = \alpha + \beta X_i + \varepsilon_i$$

em que $\ddot{Y}_i$ é o valor estimado de Y para um determinado valor de X.

Além disso, os valores de α e β são os coeficientes de regressão, tal que β fornece a inclinação da reta e α fornece a interseção da reta com o eixo *Y*.

Já ε_i é o erro aleatório, tal que $\varepsilon_i = \ddot{Y}_i - Y_i$

São necessárias algumas suposições para o erro ε_i:

os ε_i 's são variáveis aleatórias não-correlacionadas entre si.

Como $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$, os erros ε_i 's são variáveis aleatórias independentes e possuem distribuição normal: $\varepsilon_i \sim N(0, \sigma^2)$.

7.3.2. Equação da reta do modelo regressão linear simples

Consiste em estimar os parâmetros α e β da equação da reta de regressão: $\ddot{Y} = \alpha + \beta X$.

Através do **método de estimação dos mínimos quadrados** , obtém-se que

$$\beta = \frac{S_{XY}}{S_{XX}}$$

$$\alpha = \bar{Y} - \beta\bar{X}.$$

Na equação que calcula β, valores de S_{XY} e S_{XX} são aqueles usados no cálculo do coeficiente de correlação.

Na equação que calcula α, $\bar{Y}$ é a média dos valores de *Y*, ou seja, $\bar{Y} = \frac{\sum Y}{n}$ e $\bar{X}\bar{X}$ é a média dos valores de *X*, ou seja $\bar{X} = \frac{\sum X}{n}$.

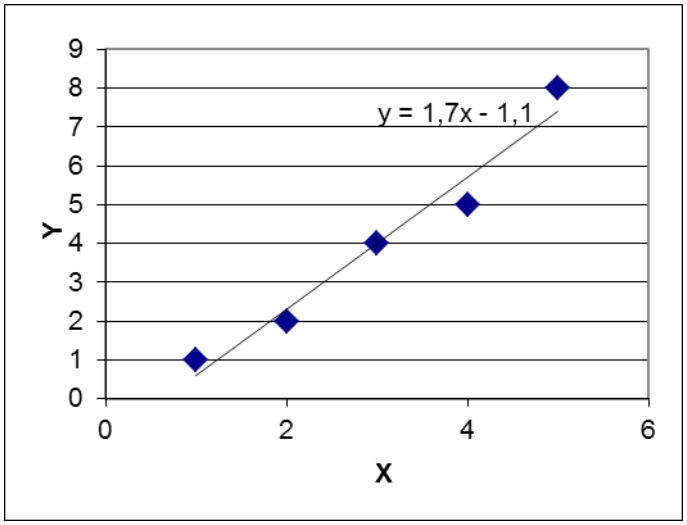
Exemplo: Obter o modelo de regressão para os dados da Tabela 7.1 do exemplo anterior.

Teremos $\beta = \frac{S_{XY}}{S_{XX}} = \frac{17}{10}$ = 1,7. Pois, do exemplo anterior, S_{XY} = 17 e S_{XX} = 10.

Teremos $\alpha = \bar{Y} - \beta\bar{X}$ = $4 - 1{,}7(3)$ = -1,1. Pois, $\bar{Y} = \frac{\sum Y}{n} = \frac{20}{5}$ = 4 e $\bar{X} = \frac{\sum X}{n} = \frac{15}{5} = 3$.

Logo, o modelo da reta de regressão é $\hat{Y} = -1{,}1 + 1{,}7X$.

Graficamente, teremos a reta de regressão:



Podemos usar valores de X do intervalo de dados da tabela 7.1 (1 a 5 horas), para obter estimativas da nota final (Y). Se X = 3,5 horas, $\hat{Y} = -1{,}1 + 1{,}7(3{,}5)$ = 4,85.

Interpretação: Estima-se que se o aluno estudar 3,5 horas, por semana, terá nota final 4,85.

Para fazer estimativas de Y, é recomendado não usar valores de X que extrapolam o intervalo usado na obtenção do modelo de regressão linear, pois a precisão da estimativa ficará comprometida.

Por exemplo, para X = 6, não devemos usar o modelo para estimativas.

Quando o coeficiente de correlação for moderado ou fraco, o modelo não deve ser usado para fazer estimativas de Y.

SÍNTESE

O teste de hipótese é um procedimento estatístico para verificar, com base em dados amostrais, se a afirmação a respeito de um parâmetro populacional deve ou não ser rejeitada. Na seção 6 deste módulo, apresentamos alguns testes que são frequentemente encontrados na prática, são eles: teste de hipóteses para a média, para a proporção e para a independência. Na sessão 7, apresentamos o diagrama de dispersão e o coeficiente de correlação de Pearson, para investigar a existência de relação linear entre duas variáveis e vimos como obter a equação da reta do modelo de regressão linear simples.

Tabelas

Tabela A : Distribuição normal padrão para valores negativos , P (Z ≤ 0)

Exemplo: P (Z ≤ - 1,96) = 0,0250

z_0	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003

Tabela B : Distribuição normal padrão para valores positivos , P (Z ≤ 0)

Exemplo: P (Z ≤ 1,96) = 0,9750

z_0	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997

Tabela C: Distribuição t de student

Exemplo: Para $t_{0.05}$ = 1,812, com 10 graus de liberdade.

Graus de liberdade	Área da cauda direita					
	0.2	0.1	0.05	0.025	0.01	0.005
1	1.376	3.078	6.314	12.706	31.821	63.657
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
10	0.879	1.372	1.812	2.228	2.764	3.169
11	0.876	1.363	1.796	2.201	2.718	3.106
12	0.873	1.356	1.782	2.179	2.681	3.055
13	0.870	1.350	1.771	2.160	2.650	3.012
14	0.868	1.345	1.761	2.145	2.624	2.977
15	0.866	1.341	1.753	2.131	2.602	2.947
16	0.865	1.337	1.746	2.120	2.583	2.921
17	0.863	1.333	1.740	2.110	2.567	2.898
18	0.862	1.330	1.734	2.101	2.552	2.878
19	0.861	1.328	1.729	2.093	2.539	2.861
20	0.860	1.325	1.725	2.086	2.528	2.845
21	0.859	1.323	1.721	2.080	2.518	2.831
22	0.858	1.321	1.717	2.074	2.508	2.819
23	0.858	1.319	1.714	2.069	2.500	2.807
24	0.857	1.318	1.711	2.064	2.492	2.797
25	0.856	1.316	1.708	2.060	2.485	2.787
26	0.856	1.315	1.706	2.056	2.479	2.779
27	0.855	1.314	1.703	2.052	2.473	2.771
28	0.855	1.313	1.701	2.048	2.467	2.763
29	0.854	1.311	1.699	2.045	2.462	2.756
30	0.854	1.310	1.697	2.042	2.457	2.750
31	0.853	1.309	1.696	2.040	2.453	2.744
32	0.853	1.309	1.694	2.037	2.449	2.738
33	0.853	1.308	1.692	2.035	2.445	2.733
34	0.852	1.307	1.691	2.032	2.441	2.728
35	0.852	1.306	1.690	2.030	2.438	2.724
40	0.851	1.303	1.684	2.021	2.423	2.704
50	0.849	1.299	1.676	2.009	2.403	2.678
100	0.845	1.290	1.660	1.984	2.364	2.626
200	0.843	1.286	1.653	1.972	2.345	2.601
?	0.842	1.282	1.645	1.960	2.327	2.576

Tabela D: Distribuição qui-quadrado

Exemplo: Para $\chi^2_{0,05}$ = 18,3, com 10 graus de liberdade.

Graus de liberdade	Área da cauda direita							
	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

REFERÊNCIAS

LARSON, R; FABER, B. **Estatística Aplicada**. 4 ed. São Paulo: Pearson Prentice Hall, 2010.