

Universidade Federal de Jataí
Instituto de Ciências Exatas e
Tecnológicas - ICET
Coordenação de Matemática

Probabilidade e Estatística

Professor:

Gecirlei Francisco da Silva
gecirlei@ufj.edu.br

Novembro/2023

Sumário

1	Introdução	1
1.1	O que é Estatística?	3
1.2	Como se divide a Estatística?	3
1.3	Estatística e Computação	4
2	Organização, Descrição e Exploração dos Dados	7
2.1	Definições Importantes	7
2.2	Classificação dos Dados	9
2.3	Variável Qualitativa Nominal	10
2.3.1	Gráficos Circulares (Gráficos de Setores, ou Pizza, ou Torta)	12
2.3.2	Gráfico de Barras (ou Diagrama de Barras)	13
2.3.3	Gráfico de Barras Compostas ou Remontadas	13
2.3.4	Gráfico de Pareto	15
2.4	Variável Qualitativa Ordinal	17
2.5	Variável Quantitativa Discreta	19
2.6	Variável Quantitativa Contínua	22
2.7	Exercícios sobre representação gráfica e tabular de dados qualitativos e quantitativos	24
2.8	Medidas de Localização e Variação	28
2.8.1	Medidas de Localização	28
2.8.2	Medidas de Variação	30
2.9	Análise Gráfica de Variáveis Quantitativas	36
2.9.1	Histograma	36
2.9.2	Gráfico de Pontos	37
2.9.3	Box Plot ou Esquema dos cinco números	38
2.9.4	Gráfico de Dispersão	40
2.10	Exercícios	41

3	Probabilidades	45
3.1	Resenha Histórica	45
3.2	Introdução	46
3.3	Revisão das Técnicas de Contagem	47
3.3.1	Princípio Fundamental da Contagem	48
3.3.2	Diagrama da Árvore	49
3.3.3	Arranjo e/ou Permutação	49
3.3.4	Combinação	51
3.4	Definição e Propriedades	52
3.5	Probabilidade Condicional e Independência	56
3.5.1	Probabilidade Condicional e Teorema da multiplicação	56
3.5.2	Independência	60
3.6	Teorema de Bayes	61
4	Variáveis Aleatórias	65
4.1	Variáveis Aleatórias	65
4.2	Função de Distribuição Acumulada	65
4.3	Variável Aleatória Discreta	68
4.4	Relação entre a Função de Distribuição Acumulada e a Distribuição de Probabilidade Discreta	69
4.5	Variável Aleatória Contínua	71
4.6	Relação entre a Função de Distribuição Acumulada e a Função Densidade de Probabilidade Contínua	71
4.7	Esperança de Variáveis Aleatórias	72
4.7.1	Esperaça de Variáveis Aleatórias Discretas	72
4.7.2	Esperaça de Variáveis Aleatórias Contínuas	73
4.7.3	Propriedades da Esperança	75
4.8	Variância de Variáveis Aleatórias	75
4.8.1	Propriedades da Variância de uma Variável Aleatória	77
4.9	Exercícios	77
5	Modelos Probabilísticos Discretos	79
5.1	Distribuição Binomial	79
5.1.1	Valor Esperado e Variância	82
5.1.2	Exercícios	83
5.2	Distribuição de Poisson	84

5.2.1	Valor Esperado e Variância	84
5.2.2	Exercícios	85
5.3	Distribuição Geométrica	86
5.3.1	Valor Esperado e Variância	87
5.3.2	Exercícios	88
5.4	Distribuição Hipergeométrica	88
5.4.1	Valor Esperado e Variância	90
5.4.2	Exercícios	90
6	Modelos Probabilísticos Contínuos	93
6.1	Distribuição Uniforme	93
6.1.1	Valor Esperado e Variância	94
6.1.2	Exercícios	95
6.2	Distribuição Normal	95
6.2.1	Valor Esperado e Variância	100
6.2.2	Exercícios	101
7	Estimação	105
7.1	Introdução	105
7.2	Definições	105
7.3	Estimação pontual	105
7.4	Estimação Intervalar	106
7.5	Exercícios	107
8	Teste de Hipóteses	113
8.1	Introdução	113
8.1.1	Definições	113
8.1.2	Classificação dos testes	116
8.2	Teste de hipóteses para a média de populações com distribuição Normal	117
8.3	Teste de hipóteses para a diferença entre médias de duas populações com distribuição Normal	121
8.4	Teste de hipóteses para a diferença entre médias de duas populações com observações pareadas	127
8.5	Introdução ao Teste Qui-Quadrado	129
8.5.1	O Teste qui-quadrado de independência	130
8.5.2	O Teste qui-quadrado de homogeneidade	133

8.5.3	Tabela da distribuição qui-quadrado	136
8.5.4	Exercícios	137
8.5.5	Exercícios Propostos	139

Capítulo 1

Introdução

Em pesquisas dentro de quaisquer área do conhecimento são geradas uma grande quantidade de informações (Dados: um, ou mais, conjunto de valores ou não) e na maioria das vezes, não existe uma análise, das mesmas, para tentar elucidar dúvidas e até mesmo tomar decisões. Diante disso, surge a necessidade em conhecer ferramentas que possam ajudar nesse sentido.

As informações podem vir de pesquisas em Controle de Qualidade, Controle de Produção, Gerenciamento, Estudos geográficos, Estudos químicos, Testes psicológicos, Controle de Estoque, Pesquisa de Mercado, Propaganda e Marketing, etc., com isso, gerando um grande banco de dados que precisam ser lidos e analisados.

A estatística, importante não somente dentro da área científica, mas também na área prática, oferece uma grande opção de ferramentas em análise de dados. Nas últimas décadas, ela vem assumido um papel muito importante e não pode ser encarada apenas como mais uma disciplina, pois se trata principalmente de uma ferramenta que auxilia no raciocínio e análise das informações obtidas.

Além do que foi comentado, seguem abaixo alguns pontos indicando a importância da estatística em uma pesquisa:

1. Fornece meios de planejar experimentos com resultados mais significantes;
2. Permite uma descrição mais exata do fenômeno;
3. Força um raciocínio e procedimentos mais exatos;
4. Ajuda a resumir um conjunto de resultados de modo claro e conveniente;
5. Ajuda a obter conclusões gerais, podendo-se inclusive concluir com certo grau de certeza;
6. Ajuda a prever o que acontecerá, sob certas condições laboratoriais;
7. Ajuda a analisar e identificar fatores causais em eventos complexos e confusos, e
8. etc..

Para entender melhor

- **Método científico** é a "ferramenta" utilizada pelos cientistas para separarem as verdades das mentiras e das ilusões. O método científico permite que os cientistas resolvam problemas complicados fazendo uso de uma série de etapas menores:

1. Observar algo do universo e com isso, **identificar o problema**, ou seja, identificar o que precisa ser estudado;
2. Desenvolver uma **pesquisa**. Este é o processo de colecionar informação e dados sobre o tópico a ser estudado;
3. Criar uma **hipótese**, que seja consistente com o que você observou, ou seja, criar uma idéia sobre a solução do problema, baseado no conhecimento e na pesquisa;
4. Fazer os **experimentos** necessários para testar as hipóteses e com isso, fazer previsões. Este é o momento de colecionar dados sobre condições controladas e repetidas. Um experimento é um conjunto de ações e observações, realizados de forma a resolver um problema ou uma pergunta particular, para aceitar ou não uma hipótese ou uma pesquisa relacionadas ao **fenômeno**. Um fenômeno é um acontecimento observável, particularmente, algo especial;
5. Organizar e **analisar os dados** coletados usando cartas, gráficos e tabelas.
6. **Concluir** fazendo um teste das previsões através de experiências e/ou por observações adicionais e modificar a hipótese com base nos resultados.
7. Repetir as etapas 3 e 4 até que não haja nenhuma discrepância entre a teoria e a experiência e/ou a observação.

Quando a consistência é obtida a hipótese transforma-se em uma teoria e esta, fornece um conjunto coerente de proposições que explicam uma classe de fenômenos. Uma teoria é então uma estrutura dentro do que as observações são explicadas e as previsões são feitas.

Exemplo 1.1. *Imagine um engenheiro de produção responsável pela produção de uma peça de computador. Em determinado momento da produção, o engenheiro suspeita que a peça está fora das especificações. Nesse momento há indícios de falha no sistema produtivo, problema esse que dará ao profissional, oportunidade para formular hipóteses sobre o que poderia ter acontecido. Dentre elas, destacamos:*

- a matéria prima utilizada na fabricação da peça;
- os equipamentos utilizados;
- possíveis falhas humanas;

- *uso de método inadequado;*
- *problemas com meio ambiente;*
- *etc..*

Algumas vezes o pesquisador não tem hipóteses claras, e passa a levantar informações e questões à partir da análise de dados observados. É o que ocorre nesse caso.

Exemplo 1.2. *Um engenheiro de materiais, foi contratado por uma indústria para estudar os tipos de polímeros adequados para a confecção de canos plásticos(PVC), procurando aumentar a durabilidade com menor custo.*

Então à partir desse momento, o objetivo é buscar as condições ideais para se ter um produto com grande durabilidade e menor custo. Neste caso, temos um trabalho de investigação, em que as informações são obtidas nas peças(corpos de prova) e no ambiente relacionado, e o investigador é o engenheiro.

1.1 O que é Estatística?

Estatística é a ciência matemática que trata sobre a coleta, análise, interpretação ou explanação, e a apresentação de dados. É aplicável a uma grande variedade de disciplinas acadêmicas: das ciências físicas e sociais até as humanas. Também é muito empregada nas tomadas de decisões em todas as áreas de negócio e de governo.

1.2 Como se divide a Estatística?

Métodos estatísticos podem ser usados para resumir ou descrever uma coleção de dados: isto é chamado de estatística descritiva. Além disso, padrões de comportamento observados nos dados podem ser modelados de modo a quantificar a aleatoriedade e incerteza dos dados, para inferir sobre o processo ou população que está sendo estudada: isto é chamado inferência estatística. De modo geral, podemos dizer que a estatística pode ser dividida em três áreas:

- **Estatística Descritiva:** um conjunto de técnicas destinadas a descrever e resumir os dados, a fim de que possamos tirar conclusões a respeito de características de interesse;
- **Probabilidade:** teoria matemática utilizada para se estudar a incerteza oriunda de fenômenos de caráter aleatório;
- **Inferência Estatística:** é o estudo de técnicas que possibilitam a extrapolação, a um grande conjunto de dados (população), das informações e conclusões obtidas a partir de subconjuntos de valores, usualmente de dimensão muito menor (Amostra).

A figura 1.1 ilustra essas áreas.

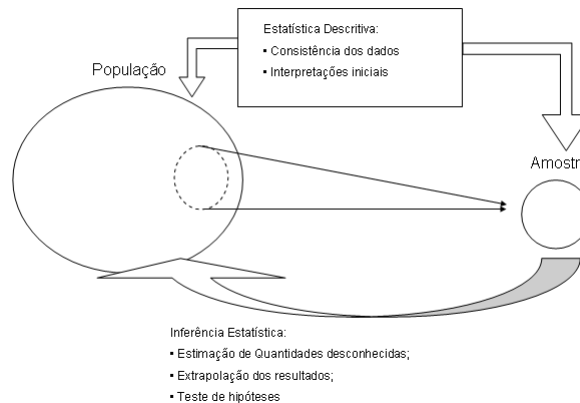


Figura 1.1: Divisão da Estatística

A estatística também apresenta outras áreas específicas de estudos: Amostragem, Planejamento de Experimento, Confiabilidade, Séries Temporais e etc.. Dentro da área de Amostragem, são definidas as formas para selecionarmos as amostras que serão estudadas e com isso, tiradas as inferências.

1.3 Estatística e Computação

Em seu artigo "Computers - The Second Revolution in Statistics", Yates (1966) revela que, para ele, a primeira revolução na Estatística veio com a introdução das máquinas de calcular. De fato, tanto as contribuições de Karl Pearson como as de R. A. Fisher, no desenvolvimento teórico da Estatística, não teriam ocorrido não fosse o precioso auxílio prestado pelas máquinas de calcular. Nas décadas de 40 e de 50, as máquinas de calcular manuais e elétricas tornaram-se comuns. Entretanto, faltava qualquer capacidade de programação, só trazida pelos computadores eletrônicos que acarretaram grande economia de tempo e de mão-de-obra.

Atualmente, um estatístico que não usa o computador é como uma espécie em extinção, cada vez mais raro de ser encontrado. Contudo, a realização de qualquer operação com um computador requer a existência de um programa apropriado, como por exemplo, o "Statistical Analysis System (SAS)", o "Statistical Package for Social Siences (SPSS)", o "Genstat", poderoso programa orientado primariamente para a análise de dados de experimentos planejados e para técnicas de análise multidimensional, o "R (Um software estatístico livre)" e vários outros.

A chegada de computadores pessoais cada vez mais poderosos foi decisiva e fez com que a Estatística se tornasse mais acessível aos pesquisadores dos diferentes campos de atuação. Atualmente, os equipamentos e softwares permitem a manipulação de grande quantidade de dados, o que veio a dinamizar o emprego dos métodos estatísticos. Hoje, a utilização da estatística está disseminada nas universidades, nas empresas privadas e públicas. Gráficos e tabelas são apresentados na exposição de

resultados das empresas. Dados numéricos são usados para aprimorar e aumentar a produção. Censos demográficos auxiliam o governo a entender melhor sua população e a organizar seus gastos com saúde, educação, saneamento básico, infraestrutura etc. Com a velocidade da informação, a estatística passou a ser uma ferramenta essencial na produção e disseminação do conhecimento. O grau de importância atribuído à estatística é tão grande que praticamente todos os governos possuem organismos oficiais destinados à realização de estudos estatísticos.

Os "cérebros eletrônicos", como foram chamados inicialmente os computadores, têm feito verdadeiras maravilhas, a ponto dos entusiastas da Inteligência Artificial acreditarem que, com o tempo, será possível duplicar qualquer atividade da mente humana, já que esta é também uma máquina. Entretanto, outros argumentam que o processo criativo da mente humana é de natureza diferente e jamais será reproduzido numa máquina. O uso intensivo dos computadores afastou o estatístico do escrutínio inteligente dos dados, com conseqüências maléficas, se não forem utilizados com sabedoria, pois como diz Yates "os computadores são bons serventes, mas, maus mestres".

Como referência citada, segue: YATES, F. Computers-the second revolution in statistics. *Biometrics*, Washington, DC, v. 22, p. 223-251, 1966.

Capítulo 2

Organização, Descrição e Exploração dos Dados

Neste capítulo, vamos supor que a pesquisa já tenha passado pelos passos: Definição dos objetivos, Formulação de Questões ou Hipóteses, Planejamento da Pesquisa e Coleta dos Dados. Os dados foram coletados, por um estudo observacional(amostragem) ou por um estudo experimental, e se encontram em alguma planilha. Na grande maioria das vezes se trata de uma massa de dados incompreensível, sem uma aparente estrutura, e precisam ser urgentemente "entendidos". Para que os mesmos sejam organizados, descritos formalmente de modo que se possa explora-los procurando indícios de padrões ou características interessantes que possivelmente indiquem possíveis tendências, e mesmo relatar ou expor características dos mesmos a outras pessoas, utilizamos as ferramentas da Estatística descritiva ou análise exploratória de Dados. Tais ferramentas consistem da leitura e resumo dos dados utilizando tabelas, gráficos, estatísticas e esquemas. As ferramentas descritivas devem fornecer resultados simples, atrair a atenção, ser auto-explicativas, de fácil compreensão e confiáveis. O maior interesse, depois de obtidos os dados, é saber como os dados estão se comportando. Uma descrição dos mesmos com tais propriedades, deve dar uma idéia global, sobre o conjunto de dados, de como os valores das variáveis observadas estão se distribuindo entre os indivíduos, e se houver, indicar tendências.

2.1 Definições Importantes

Nesta seção, apresentamos algumas definições que são utilizadas em Estatística.

1. **FENÔMENO ALEATÓRIO:** é a situação ou acontecimento cujos resultados não podem ser previstos com certeza. Por exemplo, as condições climáticas do próximo domingo não podem ser estabelecidas com total acerto;
2. **COLETA DE DADOS:** É a fase operacional. É o registro sistemático de dados, com um

objetivo determinado. Se divide em:

- (a) **Dados primários:** quando são publicados pela própria pessoa ou organização que os haja recolhido. Ex: tabelas do censo demográfico do IBGE;
- (b) **Dados secundários:** quando são publicados por outra organização. Ex: quando determinado jornal publica estatísticas referentes ao censo demográfico extraídas do IBGE;

OBS: É mais seguro trabalhar com fontes primárias. O uso da fonte secundária traz o grande risco de erros de transcrição.

- (c) **Coleta Direta:** quando é obtida diretamente da fonte. Ex: Empresa que realiza uma pesquisa para saber a preferência dos consumidores pela sua marca. Se divide em:

- **coleta contínua:** registros de nascimento, óbitos, casamentos;
- **coleta periódica:** recenseamento demográfico, censo industrial;
- **coleta ocasional:** registro de casos de dengue.
- **Coleta Indireta:** É feita por deduções a partir dos elementos conseguidos pela coleta direta, por analogia, por avaliação, indícios ou proporcionalização.

3. **APURAÇÃO DOS DADOS:** Resumo dos dados através de sua contagem e agrupamento. É a condensação e tabulação de dados;
4. **APRESENTAÇÃO DOS DADOS:** Há duas formas de apresentação, que não se excluem mutuamente. A apresentação tabular, ou seja é uma apresentação numérica dos dados em linhas e colunas distribuídas de modo ordenado, segundo regras práticas fixadas pelo Conselho Nacional de Estatística. A apresentação gráfica dos dados numéricos constitui uma apresentação geométrica permitindo uma visão rápida e clara do fenómeno;
5. **ANÁLISE E INTERPRETAÇÃO DOS DADOS:** A última fase do trabalho estatístico é a mais importante e delicada. Está ligada essencialmente ao cálculo de medidas e coeficientes, cuja finalidade principal é descrever o fenómeno (estatística descritiva);
6. **DADO ESTATÍSTICO:** é um ou mais conjuntos de valores numéricos ou não, e é considerado a matéria-prima sobre a qual iremos aplicar os métodos estatísticos;
7. **POPULAÇÃO:** é o conjunto total de elementos portadores de, pelo menos, uma característica comum;
8. **AMOSTRA:** é uma parcela representativa da população que É EXAMINADA com o propósito de tirarmos conclusões sobre essa população. É um subconjunto da população;

9. **PARÂMETROS:** São valores singulares que existem na população e que servem para caracterizá-la. Para definirmos um parâmetro devemos examinar toda a população. Ex: Os alunos do 3º período do curso de computação da UFG - Campus Jataí têm em média 1,73 metros de altura. É um símbolo que representa uma característica da população. No caso do exemplo, o símbolo que representa a média da população é μ . Os parâmetros são representados por letras gregas, tais como, μ , θ , α , σ e etc.;
10. **ESTIMADOR:** São valores singulares que existem na amostra e que servem para caracterizá-la. É uma combinação dos elementos da amostra, construída com a finalidade de representar, ou estimar, um parâmetro de interesse na população. É um símbolo que representa uma característica da amostra. O símbolo que representa a média amostral da altura dos alunos do 3º período de computação é $\hat{\mu} = \bar{x}$. Os parâmetros são representados por letras gregas, acrescidas de um chapéu, tais como, $\hat{\mu}$, $\hat{\theta}$, $\hat{\alpha}$, $\hat{\sigma}$ e etc.;
11. **ESTIMATIVA:** é um valor numérico assumido pelo estimador. É calculado com o uso da amostra.

2.2 Classificação dos Dados

Característica como peso, idade, sexo, altura, entre outras, são definidas variáveis. Assim, a variável altura assume os valores (em metros) 1,60; 1,58; ... e a variável sexo assume valores masculino e feminino. Claramente tais variáveis têm naturezas diferentes no que tange aos possíveis valores que podem assumir. Dessa forma, a figura 2.1 apresenta os vários tipos de variáveis.

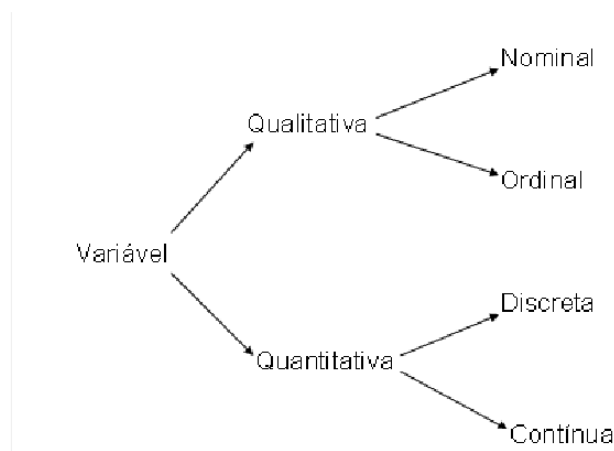


Figura 2.1:

Definimos:

- **Variável Qualitativa:** é quando os possíveis valores assumidos representam atributos e/ou qualidades;
- **Variável Quantitativa:** é quando os possíveis valores assumidos representam números;

Nas próximas seções, além de apresentarmos cada tipo de variável, vamos introduzir algumas técnicas utilizadas para a organização e interpretação dos dados. Para o estudo e descrição do comportamento de todos os indivíduos, com relação a uma ou mais variáveis em estudo, pode-se inicialmente verificar os valores encontrados no conjunto de dados, ordena-los com relação a essa variável, e depois descreve-los segundo o número de vezes (frequência) que ocorreu cada valor em particular, ou tipo de característica. A isto, denominamos de distribuição de frequência. Além dessa técnica, introduziremos alguns tipos de gráficos.

2.3 Variável Qualitativa Nominal

Uma variável Qualitativa é dita Nominal quando não é possível estabelecer uma ordem natural entre seus valores. Como exemplos, temos:

- Variável Sexo: apresenta como valores masculino e feminino;
- Variável Fuma: apresenta como valores sim ou não;
- etc..

Considere o exemplo 2.1 e observe o procedimento utilizado para resumir os dados, o que de certa forma, melhora a apresentação dos mesmos, bem como facilita sua interpretação.

Exemplo 2.1. *Suponha uma pesquisa de mercado para estudar a aceitabilidade de um possível produto a ser lançado; registrando-se entre muitas variáveis, o Tipo de Ocupação do entrevistado. Considerando-se como possíveis resultados dessa variável: Empregado / Desempregado / Patrão / Autônomo / Estudante / outros .*

Considere que a amostra coletada foi de 18 indivíduos, com os seguintes resultados: Des , Aut , Est , Emp , Emp , Des , Pat , Aut , Emp , Des , Est , Emp , Out , Aut , Emp , Est , Emp , Emp.

O que fazer diante desses dados?

Solução

O primeiro passo para descrever as ocorrências, é contar o número de vezes que cada tipo de ocupação (representa o número de categorias ou classes) ocorre, colocando os resultados em uma tabela. A tabela 2.1 apresenta a distribuição de frequência para a variável ocupação.

OCUPAÇÃO	FREQUÊNCIA = n_i
Autônomo	3
Desempregado	3
Empregado	7
Estudante	3
Patrão	1
Outros	1
TOTAL	18 = n

Tabela 2.1:

Outras formas são: usar a frequência relativa ou proporção dada por $fr = \frac{n_i}{n}$, ou a frequência percentual dada por $f\% = \left(\frac{n_i}{n}\right) * 100$. As proporções ou percentagens são muito úteis para a comparação entre resultados de pesquisas distintas, ou grupos distintos dentro da mesma pesquisa.

A tabela 2.2 apresenta um resumo mais completo dos dados.

Ocupação	Frequência n_i	Proporção ou Frequência relativa	Frequência percentual %	Frequência percentual Acumulada (%)
Autônomo	3	$3/18 = 0,1666$	$0,1666 \times 100 = 16,7$	16,7
Desempregado	3	0,1666	16,7	33,4
Empregado	7	0,3888	38,9	72,3
Estudante	3	0,1666	16,7	89,0
Patrão	1	0,0555	5,5	94,5
Outros	1	0,0555	5,5	100
TOTAL	18 = n	1	100	

Tabela 2.2:

Os métodos gráficos mais usados quando tratamos com variáveis qualitativas Nominais são,

- Gráficos Circulares (setores, ou pizza ou torta) é o mais usado
- Gráficos de barras;
- Gráficos de Colunas ou bastão;
- Gráficos de Barras Compostas;
- Gráficos Pictóricos;
- Gráfico de Pareto.

2.3.1 Gráficos Circulares (Gráficos de Setores, ou Pizza, ou Torta)

Como o próprio nome diz, o gráfico de Pizza é dividido em setores, cujas áreas são proporcionais às porcentagens das categorias observadas no conjunto de dados. A área total do círculo representa o conjunto total de dados.

Para fazer essa associação, consideremos o ângulo interno de um setor representando a porcentagem da respectiva categoria observada no conjunto de dados. Assim, o ângulo total do círculo, 360° , representa a porcentagem total de todas as categorias observadas, ou seja, 100%.

Isso é feito por uma regra de 3 simples:

100% está para 360° , assim como a frequência em % está para o ângulo. Ou ainda, n está para 360° , assim como, n_i está para o ângulo. O que dá a fórmula:

$$\alpha_i^\circ = \frac{n_i \cdot 360}{n}$$

Com isso, temos que $\alpha_1^\circ + \alpha_2^\circ + \dots + \alpha_k^\circ = 360^\circ$, com k sendo o número de categorias ou classes.

Exemplo 2.2. *Suponha uma amostra de 13 gerentes de grandes empresas, em que se tenha perguntado qual o tipo de formação ideal para um profissional da área de computação. Assim, para os 13 gerentes entrevistados, a tabela 2.3 resume a frequência para os tipos de formações.*

Formação	n_i	ângulo ($^\circ$)
Graduado	7	$\alpha_1 = 193,84$
Técnico	4	$\alpha_1 = 110,77$
Indiferente	2	$\alpha_1 = 55,39$
Total	13	360

Tabela 2.3:

Os ângulos foram calculados da seguinte forma:

$$\alpha_1 = \frac{7 \cdot 360}{13} = 193,84$$

$$\alpha_2 = \frac{4 \cdot 360}{13} = 110,77$$

$$\alpha_3 = \frac{2 \cdot 360}{13} = 55,39$$

A figura 2.2 ilustra o gráfico de pizza para este exemplo.

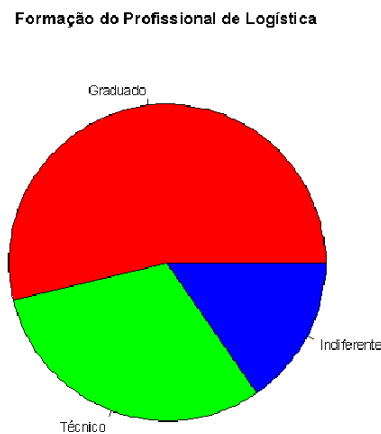


Figura 2.2:

2.3.2 Gráfico de Barras (ou Diagrama de Barras)

Para a confecção de um gráfico de barras, constrói-se um eixo horizontal ou vertical, e em intervalos apropriados, nesse eixo, coloca-se retângulos cujas alturas representam, proporcionalmente, as frequências das características observadas da variável em estudo.

Exemplo 2.3. Considere o exemplo que fala sobre a formação dos profissionais da área de computação. A figura 2.3 ilustra o gráfico de barras nessa situação.

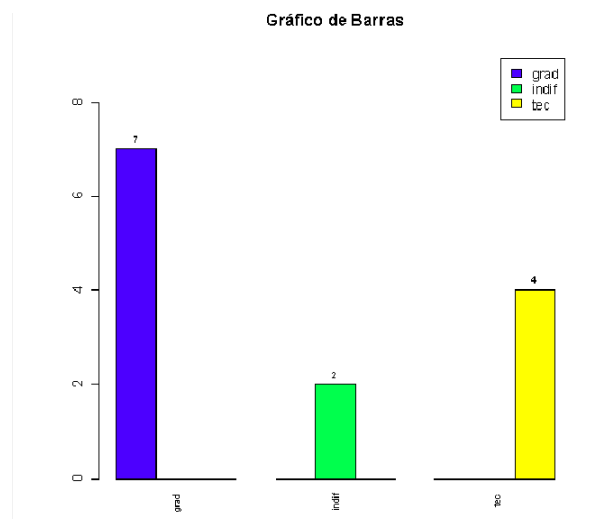


Figura 2.3:

2.3.3 Gráfico de Barras Compostas ou Remontadas

Para comparar dois ou mais grupos (fatores ou tratamentos), podemos construir um só gráfico composto de vários gráficos, um para cada grupo.

Exemplo 2.4. A tabela 2.4 mostra o número de toneladas de trigo e de milho produzidos na fazenda

GFarm, durante os anos de 1995 a 2005.

Ano	Trigo (t)	Milho (t)
1995	200	75
1996	185	90
1997	225	100
1998	250	85
1999	240	80
2000	195	100
2001	210	110
2002	225	105
2003	250	95
2004	230	110
2005	235	100

Tabela 2.4:

Faça um gráfico de barras para representar a produção de trigo e milho.

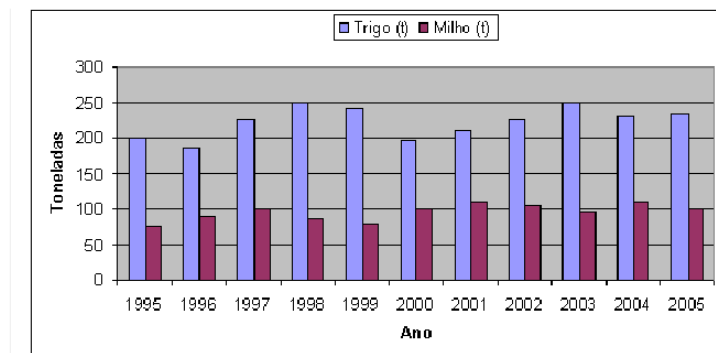


Figura 2.4:

Exercício 2.1. Segundo o IBGE (1988), a distribuição dos suicídios ocorridos no Brasil em 1986, segundo a causa atribuída, foi a seguinte: 263 por alcoolismo, 198 por dificuldade financeira, 700 por doença mental, 189 por outro tipo de doença, 416 por desilusão amorosa e 217 por outras causas. Pede-se:

- A tabela de distribuição de frequência desses dados;
- Sua representação gráfica.

2.3.4 Gráfico de Pareto

É um gráfico de barras verticais que dispõe a informação de forma a tornar evidente e visual a priorização de temas. A informação assim disposta também permite o estabelecimento de metas numéricas viáveis de serem alcançadas.

Problemas relativos à qualidade aparecem sob a forma de perdas (itens defeituosos e seus custos). É extremamente importante esclarecer o modo de distribuição destas perdas. O Gráfico de Pareto surge exatamente como ferramenta ideal para identificar quais itens são responsáveis pela maior parcela das perdas onde quase sempre são poucas as "vitais" e muitas as "triviais" (Princípio de Pareto). Então, se os recursos forem concentrados na identificação das perdas "vitais", e estas puderem ser identificadas, torna-se possível a eliminação de quase todas as perdas, deixando as "triviais" para solução posterior.

O princípio de Pareto estabelece que se forem identificados, por exemplo, cinquenta problemas relacionados à qualidade, a solução de apenas cinco ou seis destes problemas já poderá representar uma redução de 80 ou 90 % das perdas que a empresa vem sofrendo devido à ocorrência de todos os problemas existentes.

Construindo um Diagrama de Pareto

1. Selecione os problemas a serem comparados e estabeleça uma ordem através de:
 - Brainstorming - Exemplo: Qual é o nosso maior problema de qualidade no departamento de compras?
 - Utilização de dados existentes - Exemplo: Verificar os registros da qualidade do departamento de compras ao longo do último mês.
2. Selecione um padrão de comparação com unidade de medida - Exemplo: Custo mensal, frequência de ocorrência;
3. Especificar o período de tempo em que os dados serão coletados - Exemplo: Uma semana, um mês;
4. Coletar os dados necessários para cada categoria - Exemplo: Defeito A ocorreu X vezes ou defeito C custou Y;
5. Compare a frequência ou custo de cada categoria com relação a todas as outras categorias - Exemplo: Defeito A ocorreu 75 vezes, defeito B ocorreu 107 vezes, defeito C ocorreu 42 vezes ou defeito A custa 75 reais mensalmente, defeito B custa 580 reais mensalmente;
6. Liste as categorias da esquerda para direita no eixo horizontal em ordem decrescente de frequência ou custo. Os itens de menor importância podem ser combinados na categoria outros, que é colocada no extremo direito do eixo, com a última barra;

7. Acima de cada categoria desenhe um retângulo cuja a altura representa a frequência ou custo daquela categoria;
8. A partir do topo da maior barra e da esquerda para a direita, ascendendo, uma linha pode ser adiciona representando a frequência acumulada das categorias.

Exemplo 2.5. *Numa central telefônica de uma grande empresa, havia a sensação de saturação do sistema utilizado. Para melhor representar o que ocorria foi realizado um acompanhamento com as telefonistas que teriam que responder quais os tipos de problemas e quais os números de ocorrência de cada um e lançá-los na Lista de Verificação (tabela 2.5).*

Nº	Tipo de Defeito	Nº de Ocorrências	% Acumulado
1	Linha ruidosa	250	49
2	Linha aberta	110	70
3	Alarme	85	86
4	Não responde	45	95
5	Não toca	25	100
Total Geral		515	100

Tabela 2.5:

A figura 2.5 ilustra o gráfico de pareto para o problema de telefone.

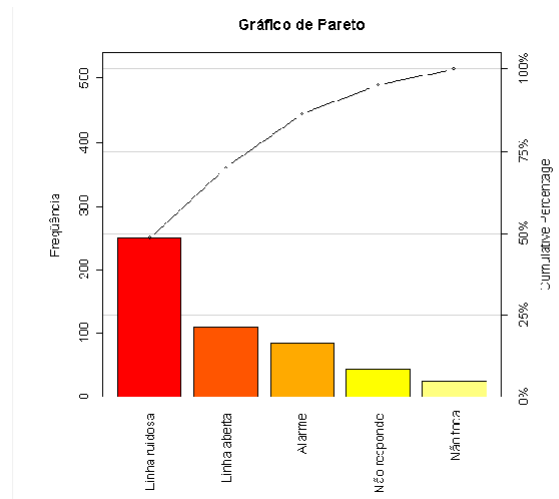


Figura 2.5:

Como é possível notar pelo gráfico o defeito "Linha ruidosa" (defeito nas uniões dos fios telefônicos ou emendas mal feitas) representa 49% de todos os defeitos ocorridos no período e que os dois maiores defeitos "Linha ruidosa" e "Linha Aberta" (deixar o telefone fora do gancho) representam juntos 70% de todos os defeitos. Corrigindo estes dois defeitos teremos uma melhoria de 70% no sistema.

Exercício 2.2. *Uma empresa de Embalagens de Papelão têm os seus defeitos mostrados na Tabela 2.6.*

Defeitos Principais	Mês				Total
	1	2	3	4	
Números Trocados	7	10	6	5	28
Perfurada	1	0	2	0	3
Caracteres Errados	6	8	5	9	28
Impressão Ilegível de Dados	0	1	1	0	2
Amassada	1	1	0	2	4
Rasgada	0	0	1	1	2
Outros	0	0	1	0	1
Total	15	20	16	17	68

Tabela 2.6: Tabela de Defeitos em Embalagens de Papelão

Faça um gráfico de pareto para ilustrar as ocorrências dos defeitos.

2.4 Variável Qualitativa Ordinal

Uma variável Qualitativa é dita Ordinal quando é possível estabelecer uma ordem natural entre seus valores. Como exemplos, temos:

- Variável Escolaridade: apresenta como valores 6^a, 7^a e 8^a série;
- Variável Tamanho: apresenta como valores pequeno, médio ou grande;
- Variável Classe Social: apresenta como valores baixa, média ou alta;
- etc..

Suponha que uma das questões de interesse em uma pesquisa de mercado seja a escolaridade dos indivíduos entrevistados. A tabela a seguir fornece a distribuição de frequências da escolaridade dos entrevistados.

Exemplo 2.6. *Suponha uma pesquisa de mercado para estudar a aceitabilidade de um possível produto a ser lançado; registrando-se entre muitas variáveis, a escolaridade do entrevistado. Considerando-se como possíveis resultados dessa variável: Analfabeto / 1^o grau / 2^o grau / superior.*

Considere que a amostra coletada foi de 18 indivíduos, com os seguintes resultados: 2grau , analf , analf , 2grau , 1grau , 1grau , 1grau , 2grau , sup , 1grau , 1grau , 2grau , 2grau , 2grau , 1grau, analf , 2grau , 2grau.

O que fazer diante desses dados?

Solução

O primeiro passo para descrever as ocorrências, é contar o número de vezes que cada categoria de escolaridade ocorre, colocando os resultados em uma tabela. A tabela 2.7 apresenta a distribuição de frequência para a variável escolaridade.

Escolaridade	FREQUÊNCIA = n_i
Analfabeto	3
1º Grau	6
2º Grau	8
Superior	1

Tabela 2.7:

A tabela 2.8 apresenta a frequência relativa ou proporção e a frequência percentual dos dados.

Escolaridade	Frequência n_i	Proporção ou Frequência relativa	Frequência percentual %	Frequência percentual Acumulada (%)
analfabeto	3	$3/18 = 0,1666$	$0,1666 \times 100 = 16,7$	16,7
1º Grau	6	0,3333	33,3	50
2º Grau	8	0,4445	44,4	94,4
Superior	1	0,0556	5,6	100
Total	18	1	100	

Tabela 2.8:

Os métodos gráficos mais usados quando tratamos com variáveis qualitativas Nominais são,

- Gráficos de barras, é o mais usado;
- Gráficos de Barras Compostas;
- Gráficos de Colunas ou bastão;
- Gráficos Circulares (setores, ou pizza ou torta);
- Gráficos Pictóricos

O gráfico de pizza, no caso da variável ser qualitativa ordinal, é feito da mesma forma que no caso da nominal, no entanto, é preferível que haja ordem na posição dos setores segundo a ordem crescente das categorias. A figura 2.6 apresenta o gráfico de pizza para a variável Escolaridade.

O gráfico de barras para a variável qualitativa ordinal, é feito de forma igual à variável nominal, só que os valores assumidos pela variável (categorias), devem ser colocados em ordem no eixo adequado. A figura 2.7 apresenta o gráfico de barras para a variável Escolaridade.

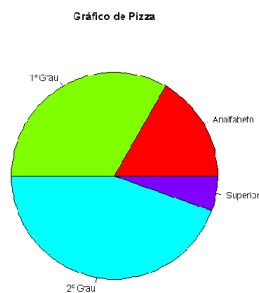


Figura 2.6:

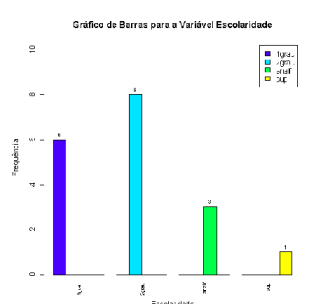


Figura 2.7:

2.5 Variável Quantitativa Discreta

Uma variável Quantitativa é dita Discreta quando seus valores são provenientes de contagens, diante disso, assume valores numéricos inteiros. Como exemplos, temos:

- Variável Número de Filhos: apresenta como valores 0, 1, 2 e etc.;
- Variável Número de alunos presentes às aulas de estatística : apresenta como valores 0, 1, 2, etc.;
- Variável Número de caminhões de cargas que passam por determinada rodovia: apresenta como valores 0, 1, 2, etc.;
- etc..

Um dos caminhos mais utilizados para resumir os dados brutos é a tabela de frequências. É um tipo de tabela que condensa uma coleção de dados conforme as frequências ou repetições dos valores das variáveis.

Diante da necessidade de se utilizar uma tabela de frequência, vamos apresentar algumas informações úteis para sua confecção.

1. **Distribuição de frequência SEM INTERVALOS DE CLASSE:** É a simples condensação dos dados conforme as repetições de seu valores. Para um tamanho razoável de categorias esta

distribuição de frequência é inconveniente, já que exige muito espaço. Veja exemplo abaixo:

Dados	Frequência
41	3
42	2
43	1
44	1
45	1
46	2
50	2
51	1
52	1
54	1
57	1
58	2
60	2
Total	20

2. **Distribuição de frequência COM INTERVALOS DE CLASSE:** Quando o tamanho da amostra é elevado, é mais racional efetuar o agrupamento dos valores em vários intervalos de classe.

Classes	Frequências
$41 \vdash 45$	7
$45 \vdash 49$	3
$49 \vdash 53$	4
$53 \vdash 57$	1
$57 \vdash 61$	5
Total	20

3. **CLASSE:** são os intervalos de variação da variável e é simbolizada por i e o número total de classes simbolizada por k . Ex: na tabela anterior $k = 5$ e $49 \vdash 53$ é a 3 classe, onde $i = 3$;
4. **LIMITES DE CLASSE:** são os extremos de cada classe. O menor número é o limite inferior de classe (l_i) e o maior número, limite superior de classe (L_i). Ex: em $49 \vdash 53$, temos que $l_3 = 49$ e $L_3 = 53$. O símbolo \vdash representa um intervalo fechado à esquerda e aberto à direita. Com isso, dizemos que o valor 53 não pertence a classe 3 e sim a classe 4 representada por $53 \vdash 57$;
5. **AMPLITUDE DO INTERVALO DE CLASSE:** é obtida através da diferença entre o limite superior e inferior da classe e é simbolizada por $h_i = L_i - l_i$. Ex: na tabela anterior $h_i = 53 - 49 = 4$. Obs: Na distribuição de frequência c/ classe o h_i será igual em todas as classes (de preferência);
6. **AMPLITUDE TOTAL DA DISTRIBUIÇÃO:** é a diferença entre o limite superior da última classe e o limite inferior da primeira classe. $AT = L(max) - l(min)$. Ex: na tabela anterior $AT = 61 - 41 = 20$;

7. **AMPLITUDE TOTAL DA AMOSTRA:** é a diferença entre o valor máximo e o valor mínimo da amostra. Onde $AA = X_{max} - X_{min}$. Em nosso exemplo $AA = 60 - 41 = 19$;
8. **PONTO MÉDIO DE CLASSE:** é o ponto que divide o intervalo de classe em duas partes iguais. Ex: em $49 \vdash 53$ o ponto médio $x_3 = (53 + 49)/2 = 51$, ou seja $x_3 = (l_3 + L_3)/2$.

Método prático para construção de uma Distribuição de Frequências

- 1º) Organize os dados brutos em ordem crescente;
- 2º) Calcule a amplitude amostral AA . No nosso exmplo: $AA = 60 - 41 = 19$;
- 3º) Calcule o número de classes através da "Regra de Sturges":

Tamanho da Amostra (n)	nº de classes
3 \vdash 5	3
6 \vdash 11	4
12 \vdash 22	5
23 \vdash 46	6
47 \vdash 90	7
91 \vdash 181	8
182 \vdash 362	9

Obs: Qualquer regra para determinação do número de classes da tabela não nos leva a uma decisão final; esta vai depender, na realidade de um julgamento pessoal, que deve estar ligado à natureza dos dados.

No nosso exemplo: $n = 20$ dados, então ,a princípio, a regra sugere a adoção de $k = 5$ classes.

- 4º) Decidido o número de classes, calcule então a amplitude do intervalo de classe $h > AA/k$. No nosso exemplo: $AA/k = 19/5 = 3,8$. Obs: Como $h > AA/k$ é um valor ligeiramente superior para haver folga na última classe, então vamos Utilizar $h = 4$;
- 5º) Nosso primeiro intervalo de classe possui como limite inferior (l_i) o menor elemento da amostra, e o limite superior (L_i) é calculado da seguinte forma: $L_i = l_i + h$. Portanto, para nosso exemplo, a primeira classe será representada por $l_i = 41$ e $L_i = 41 + 4 = 45$, ou seja, $41 \vdash 45$. As classes seguintes respeitarão o mesmo procedimento. O primeiro elemento das classes seguintes sempre serão formadas pelo último elemento da classe anterior.

Exercício 2.3. A tabela 2.9 apresenta os dados referentes ao capital de giro mensal (em mil reais) de 81 empresas do mesmo setor. Diante disso,

- a) Construa uma tabela de distribuição de frequências para os dados e determine todas as frequências, inclusive as frequências acumuladas;
- b) Com base nas frequências obtidas no item a) e sabendo que o nível mínimo admissível pelo mercado é de 32 mil reais, quantas empresas estão operando abaixo do mínimo e qual o percentual correspondente? Supondo normal a faixa de 44 a 56 mil reais e ainda conforme as mesmas frequências, quantas empresas estão operando dentro da normalidade e qual o percentual correspondente?

8	10	15	23	33	34	35	37	42	45	45	47	48	48	50	50	50	51
53	54	54	55	57	58	58	60	60	60	60	60	61	61	61	62	62	63
63	64	64	64	65	65	66	66	66	66	66	66	67	67	67	68	69	70
70	70	71	71	71	71	73	73	74	74	75	75	75	75	76	76	77	77
79	81	81	83	85	85	86	88	92									

Tabela 2.9: Capital de Giro de Empresas

Exercício 2.4. De acordo com a tabela 2.10, calcule as frequências percentuais, construa um gráfico de barras múltiplas para os pacientes que sobreviveram ou não e um gráfico de setores para os pacientes que não sobreviveram.

Faixas de Idade	Sobrevivência	
	sim	não
Menos de 50 anos	11	6
De 50 a 70 anos	18	8
Mais de 70 anos	15	9

Tabela 2.10:

2.6 Variável Quantitativa Contínua

Características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores fracionais fazem sentido. Usualmente devem ser medidas através de algum instrumento. Exemplos: peso (balança), altura (régua), tempo (relógio), pressão arterial, idade. Como valores temos,

- Variável altura: apresenta como valores 1,56; 1,65, 1,92 m e etc.;
- Variável peso: apresenta como valores 48,5; 56,8; 95,4 Kg e etc.;
- Variável Idade: apresenta como valores 5,65; 12,9; 32,5 anos e etc.;
- etc..

A análise dos dados oriundos de variáveis quantitativas contínuas se dá, em parte, como no caso discreto. Começamos com a tabela de distribuição de frequências e depois estrapolamos para alguns cálculos de medidas de tendência central e dispersão.

A construção da tabela de frequência segue como explicado no caso contínuo. Acompanhe o exemplo a seguir.

Exemplo 2.7. A tabela 2.11 apresenta o salário de 36 pessoas de uma empresa multinacional instalada em Jataí. Construa uma tabela de frequência para esses dados.

4	6,26	7,59	8,74	9,77	11,06	13,23	14,71	17,26
4,56	6,66	7,44	8,95	9,8	11,59	13,6	15,99	18,75
5,25	6,86	8,12	9,13	40,53	12	13,85	16,22	19,4
5,73	7,39	8,46	9,35	10,76	12,79	14,69	16,61	23,3

Tabela 2.11:

Solução

Primeiramente, devemos colocar os dados em ordem crescente para obtermos os valores máximo e mínimo do conjunto. Com isso, temos condições de calcular a amplitude da amostra. Neste caso, temos que a amplitude é dada por $AA = 23,3 - 4 = 19,3$.

Em seguida, considerando a regra de Sturges definimos o número de classes a ser utilizado na tabela. Considerando que temos 36 elementos na amostra, o número de classes a ser utilizado é $k = 6$.

A amplitude da classe é dada pelo cálculo de AA/k , lembrando que sempre pegamos um valor ligeiramente superior a este resultado. Neste caso, temos que a amplitude é de 3,5.

A tabela 2.12 apresenta a tabela de frequência para a variável salário.

Classe	f_i	fr_i	%	% acumulada
4,00 – 7,5	8	0,2222	22,22	22,22
7,50 – 11	12	0,3333	33,33	55,56
11,0 – 14,5	7	0,1944	19,44	75,00
14,5 – 18	6	0,1667	16,67	91,67
18,0 – 21,5	2	0,0556	5,56	97,22
21,5 – 25	1	0,0278	2,78	100,00
Total	36	1,0000	100,00	

Tabela 2.12:

Diante de uma tabela dessa, podemos tirar várias interpretações. Uma delas é que +/- 33% dos funcionários ganham entre 7,5 e 11 salários. Ou ainda, que 91,67 % dos funcionários ganham até

18 salários. Podemos dizer também que a empresa paga para a metade dos seus funcionários até 11 salários.

2.7 Exercícios sobre representação gráfica e tabular de dados qualitativos e quantitativos

Nesta seção, apresentamos alguns exercícios sobre a representação gráfica e tabular de dados qualitativos e quantitativos.

Exercício 2.5. *Classifique as variáveis abaixo e apresente uma justificativa para sua opção.*

- *marcas de computadores;*
- *capacidade do HD;*
- *tamanho da tela dos monitores;*
- *tempo gasto no uso da internet;*
- *numero de erros de digitação por pagina de documentos digitados;*
- *tamanho dos usuários de computadores;*

Exercício 2.6. *Uma pesquisa foi realizada com 64 consumidores de produtos de informática. O objetivo é verificar qual a opção de marca de monitores dentre cinco marcas dadas como opção. A tabela abaixo apresenta os resultados da pesquisa. Faça uma representação tabular e gráfica do resultados e apresente uma interpretação.*

Accer	Samsung	Accer	AOC	AOC	Samsung	toshiba	Samsung
LG	toshiba	toshiba	AOC	toshiba	Accer	Samsung	LG
Samsung	Accer	Samsung	AOC	Accer	Samsung	Samsung	Samsung
toshiba	LG	AOC	Accer	Samsung	LG	Accer	LG
AOC	LG	LG	Samsung	toshiba	Samsung	toshiba	Samsung
Samsung	LG	AOC	LG	LG	toshiba	Samsung	Samsung
AOC	LG	Samsung	Samsung	Samsung	LG	Accer	Accer
LG	toshiba	Accer	LG	Samsung	toshiba	Samsung	LG

Exercício 2.7. *Uma pesquisa foi feita com usuários de informática para saber o tipo e velocidade de internet que eles tinham disponível em casa. A tabela abaixo apresenta os resultados da pesquisa. Sabe-se que 11 pessoas entrevistadas não tem acesso a internet. Faça uma apresentação gráfica e apresente uma interpretação para a pesquisa. Além disso, diga qual a porcentagem de pessoas com acesso a internet. Qual a porcentagem dos entrevistados tem acesso ao serviço de banda larga?*

<i>Discada (Kbps)</i>	<i>ADSL (Kbps)</i>	<i>Radio (Kbps)</i>	<i>Cabo (Mbps)</i>
56	600	128	2
56	400	128	6
56	400	128	6
150	400	256	6
150	600	256	2
56	600	256	6
56	600	128	6
56	600	128	6
150	600	128	6
56	400		2
56	600		6
56	600		6
	400		
	400		

Exercício 2.8. Uma pesquisa foi realizada com empresas de faturamento de até R\$ 500.000,00. Um dos objetivos é verificar o número de computadores existentes nessas empresas. A tabela abaixo apresenta o resultados da pesquisa. Qual a porcentagem de empresas que tem até 4 computadores? Faça uma representação gráfica dos resultados. Faça uma interpretação geral.

2	1	2	2	1	1	1	1	2	4	7
2	2	1	2	1	2	3	1	2	6	6
1	5	1	2	1	1	4	1	2	7	5
2	3	4	3	1	1	3	1	3	1	4
1	2	3	3	2	3	2	2	2	2	3
1	4	2	5	3	4	2	2	2	2	2

Exercício 2.9. Uma pesquisa foi realizada com usuários de computadores com o objetivo de verificar o tempo gasto a frente do micro. A tabela abaixo apresenta o resultado da pesquisa. Faça uma tabela de frequência e diga qual a porcentagem dos entrevistados que gastam até 7,7 horas por dia na frente do micro? Faça uma representação gráfica e tire uma interpretação geral?

2,5	3	3,5	6	6,6	5,9	7,9	2,7	3,4	8
4	3	3,5	13	4,7	9,2	5,4	2,1	2,8	3,4
5	3,4	4	10,1	5,4	3,6	6,4	4,5	3,8	2,6
4,6	2,3	5	12,3	5,2	7,4	7,8	0,5	3	2,6
3,4	8,2	5,3	8,1	5,6	5,7	6,4	0,5	3,1	12,7
2	7	7	6,4	4,7	6,4	9,1	0,5	3,9	7,2
4	5,6	6,2	6,1	8,2	4,6	2	0,5	2	3,4
5	6,3	4,2	4,7	2,8	4,6	1,5	0,8	6	4,5

Exercício 2.10. Uma pesquisa foi realizada junto a 30 empresas que realizam a distribuição de produtos na cidade de Jataí. A classificação é feita da seguinte forma: MB representa "Muito Bom", "B" representa Bom, "MM" representa mais ou menos e "R" representa Ruim. Essa classificação é dada considerando produtos entregues dentro do prazo. Os dados estão apresentados na tabela 2.13.

MB	R	B	MM	MM	R
B	MM	B	MM	MM	R
MB	MM	B	R	R	MB
MM	R	MM	R	MM	B
R	R	MB	MB	MM	B

Tabela 2.13:

Dê a classificação desta variável, organize os dados em uma tabela de distribuição de frequência, faça um gráfico e apresente uma interpretação.

Exercício 2.11. Uma pesquisa foi realizada junto aos alunos do curso de computação da UFG campus jataí. Dos 100 alunos matriculados, foi retirada uma amostra de 30 e feita a seguinte pergunta: "Você vai atuar como profissional da área de computação quando concluir o curso?". Os dados estão apresentados na tabela 2.14.

sim	sim	não sei	sim	sim	nao sei
sim	sim	não sei	sim	sim	nao sei
não	sim	não sei	nao sei	sim	nao sei
não	sim	não	nao sei	sim	nao
não sei	não	sim	nao sei	nao	sim

Tabela 2.14:

Dê a classificação desta variável, organize os dados em uma tabela de distribuição de frequência, faça um gráfico e apresente uma interpretação. Sua opinião está entre a maioria?

Exercício 2.12. A empresa submarino montou uma central de relacionamento com o cliente de modo a verificar o nível de serviço prestado. A intenção é verificar dentre outras coisas, o prazo de entrega, suporte pós-venda, tempo de entrega e qualidade do produto. Diante disso, acompanhou 50 entregas e para a característica "qualidade dos produtos entregues" temos os dados apresentados na tabela 2.15.

Dê a classificação desta variável, organize os dados em uma tabela de distribuição de frequência, faça um gráfico e apresente uma interpretação.

Exercício 2.13. Uma pesquisa foi realizada por uma operadora logística, em 56 empresas de grande porte, afim de verificar em qual setor da empresa o custo era mais elevado. Dentre as opções no questionário temos: "T" representa a área de transporte, "I" representa o inventário, "A" representa

Sem defeito	Com defeito	Com defeito	Sem defeito	Sem defeito	Sem defeito	Sem defeito	Com defeito
Sem defeito	Com defeito	Sem defeito	Sem defeito	Sem defeito	Com defeito	Sem defeito	
Sem defeito	Com defeito	Sem defeito	Com defeito	Sem defeito	Sem defeito	Sem defeito	
Sem defeito	Com defeito	Sem defeito	Sem defeito	Sem defeito	Com defeito	Com defeito	
Sem defeito	Com defeito	Sem defeito	Sem defeito	Sem defeito	Sem defeito	Sem defeito	
Sem defeito	Sem defeito	Sem defeito	Sem defeito	Sem defeito	Sem defeito	Com defeito	
Sem defeito	Sem defeito	Sem defeito	Sem defeito	Com defeito	Com defeito	Sem defeito	

Tabela 2.15:

Armazenagem e "T&A" representa a área de processamento e administração. Os dados estão apresentados na tabela 2.16.

T	I	P&A	I	P&A	A	A	A
I	T	A	T	T	I	T	I
A	A	A	A	I	A	T	A
P&A	I	T	P&A	A	T	T	T
I	T	I	P&A	T	A	T	A
I	A	P&A	A	A	A	A	A
I	P&A	T	A	P&A	I	A	I

Tabela 2.16:

Dê a classificação desta variável, organize os dados em uma tabela de distribuição de frequência, faça um gráfico e apresente uma interpretação.

Exercício 2.14. Uma empresa produz 24 horas por dia, 7 dias na semana, utilizando 10 máquinas. No entanto, existem problemas que fazem com que as máquinas parem de produzir. Ao longo de uma semana de trabalho foram observadas 48 paradas e 7 defeitos foram identificados. Os defeitos foram codificados e enumerados de 1 até 7, e dessa forma transcritos na tabela 2.17.

1	5	3	3	6	5	2	1
1	4	4	5	2	5	3	4
3	3	3	6	2	4	4	2
2	5	7	5	3	4	4	5
3	5	6	5	3	3	6	4
4	3	7	6	3	4	5	3

Tabela 2.17: Tipos de defeitos observados em 48 paradas de máquinas

Dê a classificação desta variável, organize os dados em uma tabela de distribuição de frequência, faça um gráfico de pareto e apresente uma interpretação.

2.8 Medidas de Localização e Variação

2.8.1 Medidas de Localização

Média Aritmética

A média aritmética, ou simplesmente Média, é obtida dividindo-se a soma das observações pelo número delas. É um quociente geralmente representado pela letra μ ou pelo símbolo \bar{X} . Se tivermos uma série de N , ou n , valores de uma variável X , a média aritmética será determinada pela expressão:

$$\mu = \frac{(X_1 + X_2 + X_3 + X_4 + \dots + X_N)}{N} \quad \text{Média da População}$$

$$\bar{X} = \frac{(X_1 + X_2 + X_3 + X_4 + \dots + X_n)}{n} \quad \text{Média da Amostra}$$

Exemplo 2.8. Um aluno tirou as notas 5, 7, 9 e 10 em quatro provas. A sua média será $(5 + 7 + 9 + 10) / 4 = 7.75$.

Média Ponderada

Consideremos uma coleção formada por n números: x_1, x_2, \dots, x_n , de forma que cada um esteja sujeito a um peso [Nota: "peso" é sinónimo de "ponderação"], respectivamente, indicado por: p_1, p_2, \dots, p_n . A média ponderada desses n números é a soma dos produtos de cada um por seu peso, dividida por n , isto é:

$$\bar{X} = \frac{x_1 * p_1 + x_2 * p_2 + \dots + x_n * p_n}{p_1 + p_2 + \dots + p_n}$$

Exemplo 2.9. Um aluno fez um teste (peso 2) e uma prova (peso 3), tirando 10 no teste e 4 na prova. A sua média (ponderada) será $(10 * 2 + 4 * 3) / (2 + 3) = 6,4$.

OBS.: Se o teste e a prova tivessem mesmo peso (e não importa qual o valor do peso, importa apenas a relação entre os pesos), a média seria 7.

Mediana - Md

Valor que divide a distribuição em duas partes iguais, em relação à quantidade de elementos. Isto é, é o valor que ocupa o centro da distribuição, de onde conclui-se que 50% dos elementos ficam abaixo dela e 50% ficam acima.

Colocados em ordem crescente, a mediana (Med ou Md) é ou valor que divide a amostra, ou população, em duas partes iguais.

Em casos de populações (n) ímpares, a mediana será o elemento central $(n + 1)/2$. Para os casos de populações (n) pares, a mediana será o resultado da média simples dos elementos $n/2$ e $(n/2) + 1$.

Exemplo 2.10. Para a população: 1, 3, 5, 7, 9. A mediana é 5 (igual à média).

Exemplo 2.11. Para a população: 1, 2, 4, 10, 13. A mediana é 4 (enquanto a média é 6).

Exemplo 2.12. Para a população: 1,5; 2,3; 4,6; 7,2; 9,8 e 10. A mediana é $(4,6+7,2)/2$, que é 5,9.

Moda - Mo

É o valor que detém o maior número de observações, ou seja, o valor ou valores mais frequentes. A moda não é necessariamente única, ao contrário da média ou da mediana. É especialmente útil quando os valores ou observações não são numéricos, uma vez que a média e a mediana podem não ser bem definidas.

Exemplo 2.13. A moda de maçã, maçã, banana, laranja, laranja, laranja, pêssgo é laranja.

Exemplo 2.14. A série 1, 3, 5, 5, 6, 6 apresenta duas modas (bimodal): 5 e 6.

Exemplo 2.15. A série 1, 3, 2, 5, 8, 7, 9 não apresenta moda.

Exemplo 2.16. A série 11,5; 33,1; 21,5; 55,1; 21,5; 21,5 e 9 apresenta como moda o valor 21,5.

Quantis ou Quartis

Muitas medidas que resumem as propriedades do conjunto de dados usam os chamados "QUANTIS AMOSTRAIS", ("quantiles" em inglês, ou também chamados de "fractiles"). Estes termos são essencialmente equivalentes ao termo também comum "PERCENTIL". Um quantil amostral q_p é um número tendo a mesma unidade que o dado, o qual excede a proporção do dado dada pelo subscrito p , com $0 \leq p \leq 1$. O quantil amostral q_p pode ser interpretado aproximadamente como aquele valor do dado que excede um membro escolhido aleatoriamente do conjunto de dado, com probabilidade p . Analogamente, o quantil amostral q_p poderia ser interpretado como o $(p \times 100)$ é-simo percentil do conjunto de dados. A determinação dos quantis requer primeiro que os dados sejam ordenados. A notação utilizada comumente para designar os dados ordenados é a seguinte $(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, \dots, x_{(n)})$, onde $x_{(1)}$ é o valor mais baixo e $x_{(n)}$ o mais alto.

Alguns quantis são utilizados mais comumente como a mediana (ou $q_{0,5}$) ou o 50 percentil. Este é o valor do centro do conjunto de dados, no sentido que uma igual proporção de dados cai acima e abaixo deste valor. Outro uso tão comum quanto as medianas (2º quantil) são os quartis $q_{0,25}$ e $q_{0,75}$. Usualmente, são chamados de primeiro quartil e terceiro quartil. Estão localizados a meio caminho entre a mediana e os extremos $x_{(1)}$ e $x_{(n)}$.

Resumindo, dizemos que o primeiro quartil é o elemento da posição $(n+1)/4$ do conjunto de dados ordenados e o terceiro quartil é o elemento da posição $3(n+1)/4$, veja os exemplos.

Exemplo 2.17. Dada a série 1, 3, 2, 5, 8, 7, 9. Calcule o primeiro, segundo e terceiro quartil.

Solução

Os dados ordenados são: 1, 2, 3, 5, 7, 8 e 9. A posição do primeiro quartil é $(7+1)/4 = 2$. Logo, o elemento que ocupa essa posição é 2, o primeiro Quartil. A posição do terceiro quartil é $3(7+1)/4 = 6$. Logo, o elemento que ocupa essa posição é 8, o terceiro quartil. A mediana, neste caso é 5.

Visualização das medidas

A figura 2.8 apresenta visualmente as medidas de localização para um certo conjunto de dados. Neste caso, as medidas Média, Moda e Mediana são coincidentes, é o caso particular de um conjunto de dados com distribuição Normal (Gauss).

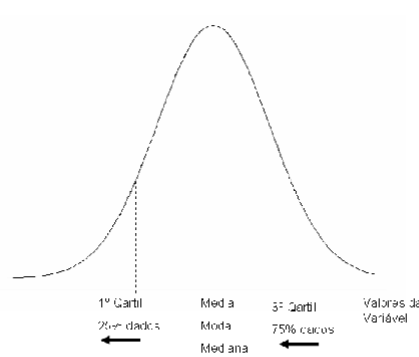


Figura 2.8:

2.8.2 Medidas de Variação

Dispersão ou Variabilidade: É a maior ou menor diversificação dos valores de uma variável em torno de um valor de tendência central (média ou mediana) tomado como ponto de comparação.

A média - ainda que considerada como um número que tem a faculdade de representar uma série de valores - não pode, por si mesma, destacar o grau de homogeneidade ou heterogeneidade que existe entre os valores que compõem o conjunto.

Exemplo 2.18. Consideremos os seguintes conjuntos de valores das variáveis $X = 70, 70, 70, 70, 70$; $Y = 68, 69, 70, 71, 72$ e $Z = 5, 15, 50, 120, 160$.

Observamos então que os três conjuntos apresentam a mesma média aritmética, ou seja, $350/5 = 70$.

Entretanto, é fácil notar que o conjunto X é mais homogêneo que os conjuntos Y e Z , já que todos os valores são iguais à média. O conjunto Y , por sua vez, é mais homogêneo que o conjunto Z , pois há menor diversificação entre cada um de seus valores e a média representativa.

Concluimos então que o conjunto X apresenta *DISPERSÃO NULA* e que o conjunto Y apresenta uma *DISPERSÃO MENOR* que o conjunto Z .

As principais medidas de variação são:

- Amplitude Total dos Dados;
- Distância Interquartílica;
- Desvio Médio;
- Variância;
- Desvio Padrão;
- Coeficiente de Variação.

Amplitude Total dos Dados

É a diferença entre o maior e o menor valor dos dados, mostrando o quanto os extremos estão "espalhados". É representado pelo cálculo

$$A_t = V_{\max} - V_{\min}$$

onde V_{\max} representa o maior valor dos dados e V_{\min} representa o menor valor dos dados. Esta medida leva em conta apenas dois valores e é muito sensível à valores extremos.

Exemplo 2.19. considere o conjunto de dados $A = \{12, 15, 16, 17, 20, 21\}$. A amplitude dos dados é dada por:

$$A_t = V_{\max} - V_{\min} = 21 - 12 = 9$$

Distância Interquartílica

É a diferença entre o terceiro e o primeiro quartil. Mostra a amplitude levando em conta 50% dos dados centrais, e não só os extremos. A figura 2.9 ilustra essa diferença.

Exemplo 2.20. Já vimos que, o primeiro e terceiro quartil da série 1, 3, 2, 5, 8, 7, 9 são dados por: $Q_1 = 2$ e $Q_3 = 8$. Com isso, a distância interquartílica é dada por

$$Q_3 - Q_1 = 8 - 2 = 6$$

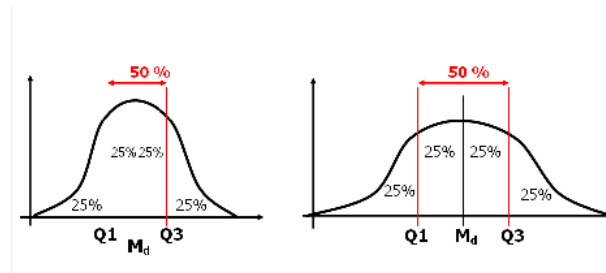


Figura 2.9:

Desvio Médio

O desvio de um valor observado é a sua "distância" da medida central, ou seja, a diferença entre seu valor e a média.

$$d_i = x_i - \bar{x}$$

A soma de todos os desvios é sempre zero. O desvio absoluto é o módulo do desvio (valor do desvio "sem o sinal"). O desvio médio é o somatório dos desvios absolutos divididos pelo número de observações, ou seja:

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

Exemplo 2.21. A tabela 2.18 apresenta as medições e o cálculo do desvio. A primeira coluna apresenta os dados, os quais possuem média 14,7, enquanto que, a segunda coluna apresenta o cálculo do desvio dos dados em relação a média. A terceira coluna apresenta o desvio absoluto. com isso, o desvio médio é dado por:

$$d_m = \frac{\sum_{i=1}^{12} |d_i|}{12} = \frac{3,7 + 2,7 + \dots + 0,3 + 1,3}{12} = 1,7$$

A figura 2.10 ilustra o cálculo do desvio.

Medições (x_i)	$d_i = (x_i - media)$	$ d_i $
11	-3,7	3,7
12	-2,7	2,7
13	-1,7	1,7
15	0,3	0,3
16	1,3	1,3
14	-0,7	0,7
13	-1,7	1,7
16	1,3	1,3
17	2,3	2,3
18	3,3	3,3
15	0,3	0,3
16	1,3	1,3

Tabela 2.18:

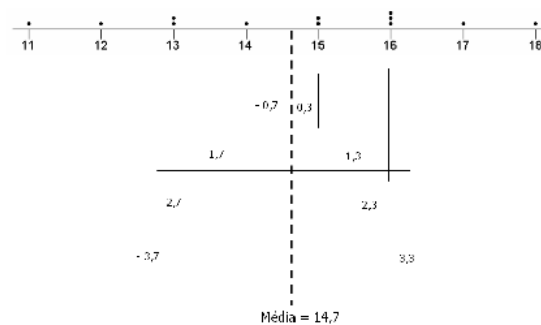


Figura 2.10:

Variância

É a soma dos quadrados dos desvios dividido pelo número de observações. Podemos calcular a variância da população e da amostra, da seguinte forma:

$$\sigma^2 = \sum_{i=1}^N \frac{(x_i - \bar{x})^2}{N} \quad \text{População (viciada)}$$

$$\sigma^2 = \sum_{i=1}^N \frac{(x_i - \bar{x})^2}{N-1} \quad \text{População (Não viciada)}$$

$$s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n} \quad \text{Amostra (viciada)}$$

OBS.: Elevando os desvios ao quadrado a influência de valores extremos é maior.

Exemplo 2.22. Considerando os dados da amostra utilizada no cálculo dos desvios no exemplo anterior, temos que a variância é dada por:

$$\begin{aligned} s^2 &= \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1} \\ &= \frac{(-3,7)^2 + (-2,7)^2 + \dots + (0,3)^2 + (1,3)^2}{12-1} \\ &= \frac{48,7}{11} = 4,42 \end{aligned}$$

Desvio Padrão

É a raiz da Variância, tendo a mesma unidade dos dados medidos. Podemos calcular o desvio padrão da população e da amostra, da seguinte forma:

$$\begin{aligned} \sigma &= \sqrt{\sum_{i=1}^N \frac{(x_i - \bar{x})^2}{N}} && \text{População (viciada)} \\ \sigma &= \sqrt{\sum_{i=1}^N \frac{(x_i - \bar{x})^2}{N-1}} && \text{População (Não viciada)} \\ s &= \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n}} && \text{Amostra (viciada)} \\ s &= \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}} && \text{Amostra (Não viciada)} \end{aligned}$$

Exemplo 2.23. No exemplo anterior, vimos que a variância é 4,42. Com isso, o desvio padrão é dado por:

$$s = \sqrt{4,42} = 2,1$$

Coefficiente de Variação

Uma pergunta que pode surgir é: O desvio padrão calculado é grande ou pequeno? Esta questão é relevante por exemplo, na avaliação da precisão de métodos.

Um desvio padrão pode ser considerado grande ou pequeno dependendo da ordem de grandeza da variável. Uma maneira de se expressar a variabilidade dos dados tirando a influência da ordem de grandeza da variável é através do coeficiente de variação, definido por:

$$CV = \frac{s}{\bar{x}}$$

O CV é:

- Interpretado como a variabilidade dos dados em relação à média. Quanto menor o CV mais homogêneo é o conjunto de dados;
- Adimensional, isto é, um número puro, que será positivo se a média for positiva; será zero quando não houver variabilidade entre os dados, ou seja, $s = 0$;
- Usualmente expresso em porcentagem, indicando o percentual que o desvio padrão é menor ($100\%CV < 100\%$) ou maior ($100\%CV > 100\%$) do que a média

Um CV é considerado baixo (indicando um conjunto de dados razoavelmente homogêneo) quando for menor ou igual a 25%. Entretanto, esse padrão varia de acordo com a aplicação. Por exemplo, em medidas vitais (batimento cardíaco, temperatura corporal, etc) espera-se um CV muito menor do que 25% para que os dados sejam considerados homogêneos.

Pode ser difícil classificar um coeficiente de variação como baixo, médio, alto ou muito alto, mas este pode ser bastante útil na comparação de duas variáveis ou dois grupos que a princípio não são comparáveis.

Exemplo 2.24. Em um grupo de pacientes foram tomadas as pulsações (batidas por minuto) e dosadas as taxas de ácido úrico (mg/100 ml). As médias e os desvios padrão foram:

Variável	\bar{x}	s
pulsação	68,7	8,7
ácido úrico	5,46	1,03

Os coeficientes de variação são: $CV_p = \frac{8,7}{68,7} = 0,127$ e $CV_{a.u.} = \frac{1,03}{5,46} = 0,232$, o que evidencia que a pulsação é mais estável do que o ácido úrico.

Exemplo 2.25. Em experimentos para a determinação de clorofila em plantas, levantou-se a questão de que se o método utilizado poderia fornecer resultados mais consistentes. Três métodos foram colocados à prova e 12 folhas de abacaxi foram analisadas com cada um dos métodos. Os resultados foram os seguintes:

Método (unidade)	\bar{x}	s	CV
1(100cm ³)	13,71	1,20	0,088
2(100g)	61,40	5,52	0,090
3(100g)	337,00	31,20	0,093

Note que as médias são bastante diferentes devido às diferenças entre os métodos. Entretanto, os três CV's são próximos, o que indica que a consistência dos métodos é praticamente equivalente, sendo que o método 3 mostrou-se um pouco menos consistente.

2.9 Análise Gráfica de Variáveis Quantitativas

2.9.1 Histograma

Na estatística, um histograma é uma representação gráfica da distribuição de frequências de uma massa de medições, normalmente um gráfico de barras verticais. É uma das Sete Ferramentas da Qualidade.

O histograma é um gráfico composto por retângulos justapostos em que a base de cada um deles corresponde ao intervalo de classe e a sua altura à respectiva frequência. Quando o número de dados aumenta indefinidamente e o intervalo de classe tende a zero, a distribuição de frequência passa para uma distribuição de densidade de probabilidades. A construção de histogramas tem caráter preliminar em qualquer estudo e é um importante indicador da distribuição de dados. Podem indicar se uma distribuição aproxima-se de uma função normal, como pode indicar mistura de populações quando se apresentam bimodais.

Exemplo 2.26. *150 peixes mortos foram encontrados vítimas de contaminação do rio e seus comprimentos foram medidos em milímetros. As medidas foram expressas na forma de tabela de frequência.*

<i>Comprimento do Peixe (mm)</i>	<i>Frequência</i>
<i>100 ┤ 110</i>	<i>7</i>
<i>110 ┤ 120</i>	<i>16</i>
<i>120 ┤ 130</i>	<i>19</i>
<i>130 ┤ 140</i>	<i>31</i>
<i>140 ┤ 150</i>	<i>41</i>
<i>150 ┤ 160</i>	<i>23</i>
<i>160 ┤ 170</i>	<i>10</i>
<i>170 ┤ 180</i>	<i>3</i>
<i>Total</i>	<i>150</i>

A figura 2.11 apresenta o histograma deste exemplo.

A figura 2.12 apresenta alguns modelos de histogramas com suas respectivas interpretações.

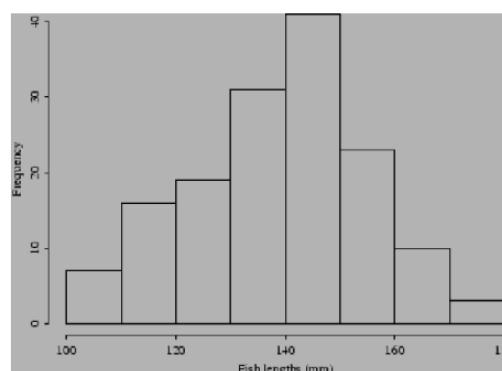


Figura 2.11:

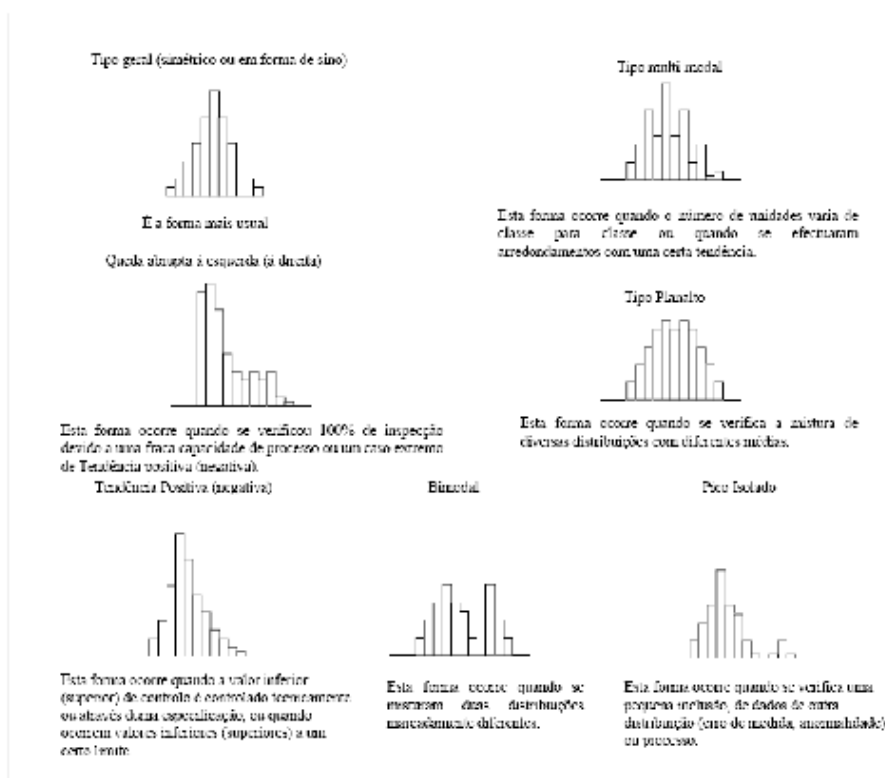


Figura 2.12:

2.9.2 Gráfico de Pontos

Uma representação alternativa ao histograma para a distribuição de frequências de uma variável quantitativa é o diagrama de pontos, como mostra a figura 2.13.

Neste gráfico, cada ponto representa uma observação com determinado valor da variável. Observações com mesmo valor são representadas com pontos empilhados neste valor.

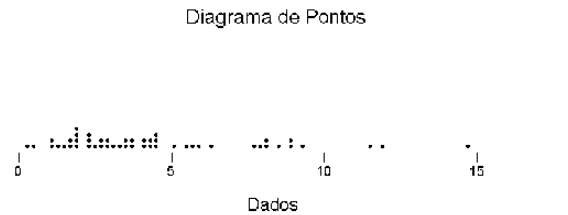


Figura 2.13:

2.9.3 Box Plot ou Esquema dos cinco números

O Box Plot foi inventado em 1977 pelo estatístico americano John Tukey. É a representação de um conjunto de valores, considerando apenas as seguintes medidas

1. **Limite Superior (LS):** É calculado por

$$LS = Q_3 + \frac{3}{2}(Q_3 - Q_1)$$

2. **Limite inferior (LI):** É calculado por

$$LI = Q_1 - \frac{3}{2}(Q_3 - Q_1)$$

3. **Quartil 1 ou Q_1 :** Deixa 25% dos dados abaixo dele;
4. **Quartil 2 ou Q_2 (Mediana):** Deixa 50% dos dados abaixo dele;
5. **Quartil 3 ou Q_3 :** Deixa 75% dos dados abaixo dele;

A figura 2.14 ilustra os componentes do Box Plot.

Em adição, o Box Plot pode indicar quais observações, se existir, são consideradas não usuais ou outliers. Aqueles valores que aparecerem acima ou abaixo do LS e LI são considerados valores estranhos e devem ser observados com mais carinho, pois, podem ser tanto peças importantes na interpretação dos dados quanto valores com problemas ao transcrever os dados!

O Box Plot também auxilia na verificação da dispersão e assimetria dos dados. Quanto maior a caixa central dos gráfico, maior será a variabilidade. Se a linha central estiver com uma tendência tanto para o lado do terceiro quartil quanto para o primeiro, podemos dizer que os dados tem uma tendência para serem assimétricos.

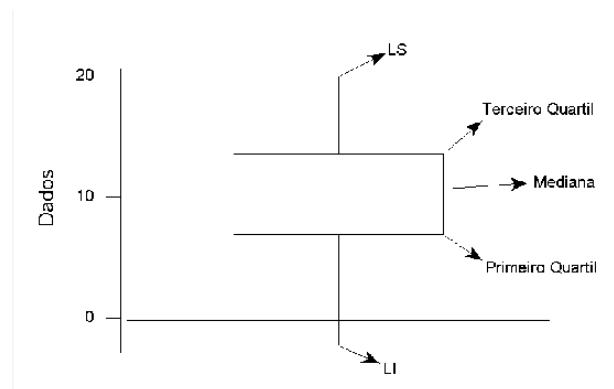


Figura 2.14:

Exemplo 2.27. Na figura 2.15 temos a representação de duas empresas de transporte que realizam o serviço de entrega de produtos dentro da cidade de Jataí. Estamos representando o tempo gasto pelas empresas para realizarem a mesma tarefa. Tire suas próprias conclusões.

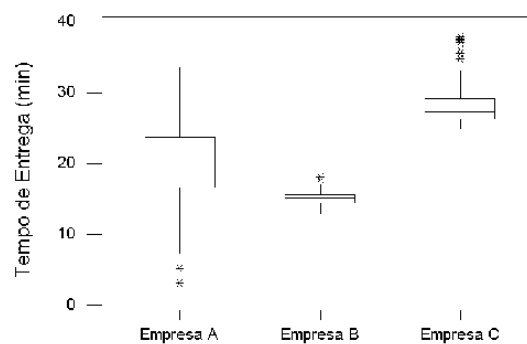


Figura 2.15:

Observe que a empresa A tem uma variação muito grande no tempo de entrega em comparação com as outras duas. A empresa B tem a menor variação entre as empresas e seu tempo de entrega é bem menor que as demais. A empresa C tem uma variação pequena no seu tempo de entrega, no entanto, verificamos uma certa assimetria nos dados, bem como um tempo de entrega maior. Percebemos que nas três empresas apareceram pontos considerados entranhos, outliers!

2.9.4 Gráfico de Dispersão

Um gráfico de dispersão constitui a melhor maneira de visualizar a relação entre duas variáveis quantitativas. É uma das sete ferramentas da qualidade. Coleta dados aos pares de duas variáveis (causa/efeito) para checar a existência real da relação entre essas variáveis.

A figura 2.16 apresenta a relação entre as variáveis Tempo e Temperatura.

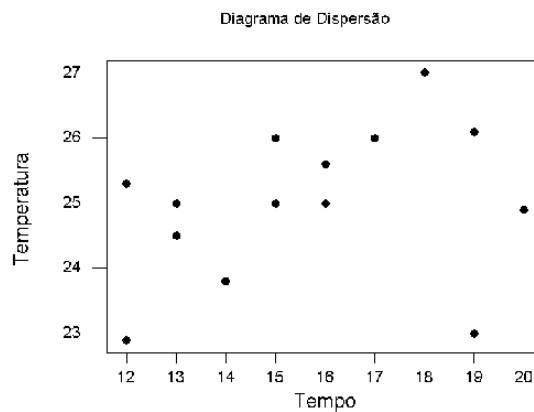


Figura 2.16:

Um gráfico de dispersão tem dois eixos de valores, mostrando um conjunto de dados numéricos ao longo do eixo horizontal (eixo X) e outro ao longo do eixo vertical (eixo Y). Ele combina esses valores em pontos de dados únicos e os exibe a intervalos irregulares, ou agrupamentos. Gráficos de dispersão são comumente usados para exibir e comparar valores numéricos, como dados científicos, estatísticos e de engenharia.

2.10 Exercícios

Exercício 2.15. *Quarenta embalagens plásticas de mel foram pesadas com precisão de decigramas. Os pesos são apresentados na tabela 2.19.*

31,1	31,7	33,9	33,4	34,2	34,5	35,2	34,3	34,7	34,7
32,0	31,9	35,7	33,8	35,1	33,9	35,5	34,6	34,7	34,6
35,4	32,1	35,0	32,6	31,8	33,6	35,7	34,7	35,4	33,9
32,8	33,6	34,2	30,7	31,3	32,3	32,8	34,8	35,2	34,0

Tabela 2.19:

Dê a classificação desta variável, organize os dados em uma tabela de distribuição de frequência e faça um gráfico. Depois, considerando todos os dados, calcule a média, o desvio padrão, a moda, a mediana, a amplitude, o 1º quartil, o 3º quartil, o coeficiente de variação e apresente uma interpretação. Apresente ainda um gráfico de boxplot e reforce sua interpretação.

Exercício 2.16. *Os números apresentados na tabela 2.20 representam as notas de 30 alunos matriculados na disciplina Estatística aplicada.*

5,5	3,0	4,0	4,5	7,0
6,5	3,5	4,5	3,0	7,5
4,5	0,0	4,5	3,5	4,5
7,0	9,0	6,0	4,0	5,0
8,0	9,5	4,5	4,5	4,5
2,5	2,0	5,0	6,0	4,5

Tabela 2.20:

Dê a classificação desta variável, organize os dados em uma tabela de distribuição de frequência e faça um gráfico. Depois, considerando todos os dados, calcule a média, o desvio padrão, a moda, a mediana, a amplitude, o 1º quartil, o 3º quartil, o coeficiente de variação e apresente uma interpretação. Apresente ainda um gráfico de boxplot e reforce sua interpretação.

Exercício 2.17. *A tabela 2.21 apresenta uma amostra de tarifas de transporte, realizadas por caminhões, para cargas de 2000 a 5000 kg, com origem em Jataí e tendo como destino várias cidades*

no brasil. A partir dos dados, construa um gráfico de dispersão e interprete-o. Existe algum relacionamento linear entre as variáveis?

Tarifa em R\$	Distância em Km
4,15	169
16,2	2220
9,11	1108
6,81	427
13,53	2197
9,84	1226
15,28	2685
6,92	465
9,51	936
8,03	751
7,8	848
12,77	1923
11,28	1004
7,8	657
8,24	955
8,4	801
13,38	1753
12,77	1998
10,69	1337
8,5	799

Tabela 2.21:

Como complemento ao gráfico, inclua uma reta vertical e outra horizontal nas respectivas médias das variáveis.

Exercício 2.18. A tabela 2.22 apresenta informações sobre pedido de produtos por canal de distribuição, durante o ano de 2007. Calcule a média, o desvio padrão, e o coeficiente de variação para cada canal. Apresente em um único gráfico o boxplot de cada canal e faça uma interpretação baseado nos resultados numéricos e visuais.

Exercício 2.19. A tabela 2.23 apresenta informações sobre a demanda mensal de três varejistas. Calcule a média, o desvio padrão e o coeficiente de variação para cada canal. Apresente em um único gráfico o boxplot de cada canal e faça uma interpretação baseado nos resultados numéricos e visuais.

C&I	Cliente	OEM
46307	24709	32007
55013	28023	33675
44683	21511	35761
54528	23487	33987
48492	29644	31626
42230	21204	32564
46709	24089	33078
50983	25958	34021
46792	26182	34123
65775	37272	32347
57932	33650	33690
47152	25482	32896

Tabela 2.22:

varejista 1	218	188	225	217	176	187	221	212	210	203	188	185
varejista 2	101	87	123	101	95	97	93	131	76	101	87	114
varejista 3	268	296	321	312	301	294	285	305	289	303	324	332

Tabela 2.23:

Capítulo 3

Probabilidades

3.1 Resenha Histórica

As probabilidades nasceram na Idade Média com os tradicionais jogos de azar e apostas que se efetuavam na Corte.

Os algebristas Italianos Pacioli, Cardano e Tartaglia (séc.XVI) fizeram as primeiras observações matemáticas relativas às apostas nos jogos de azar.

Porém, a verdadeira teoria relativa às probabilidades surgiu através da correspondência entre Blaise Pascal e seu amigo Pierre De Fermat, chegando estes à mesma solução do célebre problema da divisão das apostas em 1654, embora tivessem seguido caminhos diferentes.

Este problema foi posto a Pascal pelo Cavaleiro De Méré. Este Cavaleiro era considerado por alguns um jogador inveterado, por outros um filósofo e homem de letras.

Um fato curioso é que este problema era o mesmo que, sensivelmente, um século antes havia retido a atenção de Pacioli, Tartaglia e Cardano.

Gerolamo Cardano, médico e matemático Italiano, nascido em Pavia (1501-1576) escreveu o primeiro livro relativo às probabilidades "Liber de Ludo Alex" ("Livro dos jogos do azar"), embora este só tenha sido publicado em 1663.

Laplace publicou a obra da Teoria Analítica das Probabilidades, em 1812. Esta obra foi um importante tributo para o desenvolvimento dos conhecimentos nesta área, uma vez que reuniu as ideias descobertas até então, donde se salienta a famosa Lei de Laplace.

Laplace comentou as teorias de Pascal do seguinte modo:

"A teoria das probabilidades, no fundo, não é mais do que o bom senso traduzido em cálculo; permite calcular com exatidão aquilo que as pessoas sentem por uma espécie de instinto... É notável que tal ciência, que começou nos estudos sobre jogos de azar, tenha alcançado os mais altos níveis do conhecimento humano."

A teoria das probabilidades evoluiu de tal forma que no século XX possui uma axiomática própria

dentro da teoria matemática. Tal efeito deve-se sobretudo a Kolmogorov, que em 1933 adotou a nova definição de probabilidade que atualmente designamos por "Definição frequencista".

3.2 Introdução

Já vimos que para se obter informações sobre alguma característica da população, o tamanho amostral é de fundamental importância. Estudaremos agora a probabilidade, que é uma ferramenta usada e necessária para se fazer ligações entre a amostra e a população, de modo que a partir de informações da amostra se possa fazer afirmações sobre características da população. Assim, pode-se dizer que a probabilidade é a ferramenta básica da inferência. No dia-a-dia, usa-se o conceito de probabilidade como: "É pouco provável que amanhã chova" Provavelmente o candidato tal não se eleja. "A chance do Corinthians ser campeão é pequena. "Diminuiu a chance do paciente se recuperar. "etc...

São fenômenos ou eventos com resultados não completamente conhecidos a priori. Mesmo que a chance da ocorrência seja alta, os resultados não são conhecidos antes de ocorrer, mas de certa forma, mantem uma certa regularidade, o que permite determinar a chance de ocorrência; a Probabilidade.

Podemos classificar os fenômenos da natureza, ou criados pelo homem, em dois tipos: aleatórios (casuais) e não aleatórios (determinísticos). Trabalharemos com os aleatórios, os quais não sabemos o resultado a priori. No entanto, podemos listar os possíveis resultados do fenômeno aleatório, que formarão um conjunto denominado de Espaço Amostral (S). Ao estudarmos uma característica da qualidade de um processo (ou produto), o espaço amostral consiste de todos os valores possíveis que essa característica da qualidade pode assumir.

Exemplo 3.1. *Considerem os experimentos:*

- a) *Lançar um dado e observar a face que cair para cima. O espaço amostral é $S = \{1, 2, 3, 4, 5, 6\}$.*
- b) *Classificar um produto em conforme ou não conforme. Neste caso, o espaço amostral é $S = \{\text{Conforme}, \text{Não conforme}\}$.*
- c) *Contar o número de defeitos em uma peça pintada (por exemplo). Neste caso, os possíveis resultados são $S = \{0, 1, 2, 3, \dots\}$.*

Relacionado a um experimento, como acima, uma série de sentenças podem ser formuladas. Estas sentenças são denominadas Eventos.

Exemplo 3.2. *Consideremos o lançamento do dado no exemplo 3.1. Podemos definir vários eventos. Alguns são: $A = \text{"sair número par"}$, $B = \text{"sair número ímpar"}$, $C = \text{"sair número maior do que 3"}$. Esses eventos podem ser representados, respectivamente, pelos conjuntos: $A = \{2, 4, 6\}$, $B = \{1, 3, 5\}$*

e $C = \{4, 5, 6\}$. Considere o experimento de classificar a peça em conforme ou não, podemos definir como eventos, $A = \{\text{Conforme}\}$, $B = \{\text{Não conforme}\}$. Ao contarmos o número de defeitos em uma peça pintada, geralmente, estaremos interessados no evento $A = \{\text{Zero Defeito}\} = \{0\}$.

De uma forma geral, qualquer subconjunto de um espaço amostral será denominado Evento. Os eventos são denotados por letras maiúsculas (A, B, C, ...). Outro aspecto importante da teoria de probabilidade está na manipulação de eventos. Do ponto de vista prático, os eventos são as sentenças (perguntas) que podemos formular sobre nosso experimento. Assim, desejamos definir formas de manipular, ou seja, de operar estas sentenças. As três operações básicas são:

União (\cup) : A união de dois conjuntos quaisquer E e F conterá todos os elementos de E e de F, incluindo os elementos que sejam comum aos dois ou não.

Intersecção (\cap) : A intersecção de dois conjuntos quaisquer E e F conterá os elementos comuns a E e F.

Complementar (A^c) : O evento complementar ao evento A é o conjunto dos elementos do espaço amostral que não pertencem a A.

Exemplo 3.3. Consideremos o lançamento do dado no exemplo 3.2 . Temos:

$$a) A \cup B = \{1, 2, 3, 4, 5, 6\}$$

$$b) A \cap B = \{\} = \phi \quad \text{conjunto vazio}$$

$$c) A \cap C = \{4, 6\} \quad e \quad A \cup C = \{2, 4, 5, 6\}$$

$$d) C^c = \{1, 2, 3\}$$

Na terminologia da teoria de conjuntos, o conjunto vazio é o conjunto composto por nenhum elemento, que denotaremos por ϕ . Este conjunto está contido em qualquer outro evento do espaço amostral.

Para realizarmos o cálculo da probabilidade de ocorrência de certo **evento**, precisamos inicialmente definir o número de elementos do conjunto evento, além do número de elementos do conjunto **espaço amostral**, conforme veremos na definição de probabilidade. Em conjuntos com grande quantidade de resultados, podemos ter dificuldade em determinar o número de elementos, com isso, devemos utilizar alguma técnica de contagem.

3.3 Revisão das Técnicas de Contagem

Nesta seção, faremos uma breve revisão das técnicas de contagem: Princípio Fundamental da Contagem, Arranjo e Combinação.

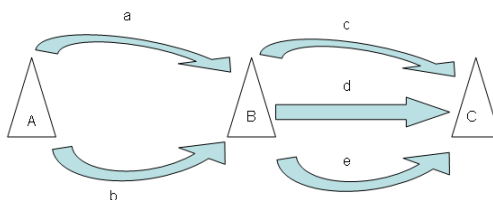
3.3.1 Princípio Fundamental da Contagem

Se um evento pode ocorrer de n_1 maneiras distintas e, a seguir, um segundo evento pode ocorrer de n_2 maneiras distintas, e assim sucessivamente, até um k -ésimo evento que pode ocorrer de n_k maneiras distintas, então o número de maneiras distintas em que os k eventos podem ocorrer sucessivamente é $n_1.n_2.....n_k$.

Exemplo 3.4. Desejamos ir da cidade A à cidade C. Os caminhos de A a C passam pela cidade B. Se há dois caminhos que ligam A a B e três caminhos que ligam B a C, de quantas maneiras podemos ir de A a C?

Solução

Considere o esquema apresentado na figura 3.4 .



Dessa forma, podemos montar o conjunto com as seguintes possibilidades:

$$S = \{(a, c); (a, d); (a, e); (b, c); (b, d); (b, e)\} \quad ,$$

o qual possui 6 elementos. Esse valor pode ser obtido multiplicando o número de possibilidades de A até B, 2, e o número de possibilidades de B até C, 3, ou seja,

$$\#S = 2 * 3 = 6$$

Exercício 3.1. Suponha que você está em dúvida sobre qual marca de carro comprar (VW, GM, Ford, Fiat) e para cada marca três modelos (Sport, Sedan, Adventure), de modo que qualquer combinação encaixe no seu orçamento e é de seu gosto. Quantas opções de compra você têm?

Exercício 3.2. A turma do 3º período de ciência da computação (UFG) tem 25 alunos. Um deles será escolhido para ser o representante da turma e outro para vice. Qual é o número de possíveis disposições das pessoas nas vagas?

Exercício 3.3. De quantas maneiras podemos responder a 10 perguntas de um questionário, cujas respostas para cada pergunta são: sim ou não?

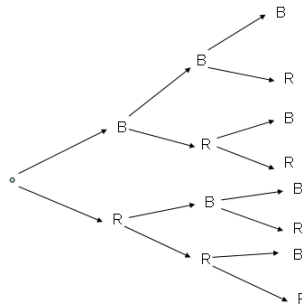
Exercício 3.4. Uma moeda é lançada 3 vezes. Qual o número de sequências possíveis de cara e coroa?

3.3.2 Diagrama da Árvore

É um esquema usado para enumerar todos os resultados possíveis de uma sequência de experimentos, onde cada um pode ocorrer em um número finito de maneiras.

Exemplo 3.5. Considere que um lote de 100 sacas de soja deve ser inspecionado, e que 3 sacas foram observadas e classificadas em boas e ruins. Quantifique os possíveis resultados desse experimento.

A figura 3.5 ilustra a árvore de possibilidades.



$$S = \left\{ \begin{array}{ll} (B, B, B) & (R, R, R) \\ (B, B, R) & (R, R, B) \\ (B, R, B) & (R, B, R) \\ (B, R, R) & (R, B, B) \end{array} \right\}$$

Com isso, temos que $\#S = 8$.

Exercício 3.5. Quais os resultados possíveis no lançamento de 2 moedas honestas?

Exercício 3.6. Quais os resultados possíveis no lançamento de 3 moedas honestas?

Exercício 3.7. Quais os resultados possíveis no lançamento de 2 dados honestos?

3.3.3 Arranjo e/ou Permutação

Um arranjo de um conjunto de n objetos, em dada ordem, é chamado de permutação dos objetos (tomados todos ao mesmo tempo). Um arranjo de quaisquer $r \leq n$ destes objetos, em dada ordem, é chamado de r -permutação ou permutação dos n objetos, tomados r a r .

Exemplo 3.6. Considere o conjunto de letras a, b, c e d . Então:

- a) $bdca, dcba$ e $acdb$ são permutações das 4 letras (tomadas todas ao mesmo tempo);
- b) bad, adb, cbd e bca são permutações das 4 letras, tomadas 3 a 3;

c) ad, cb, da, bd são permutações das 4 letras, tomadas 2 a 2.

A fórmula utilizada para o cálculo do número de arranjos é:

$$A_{n,r} = \frac{n!}{(n-r)!}$$

OBS.: A notação $n!$ representa um número, o qual chamamos de fatorial de n . O cálculo de $n!$ é dado por:

$$n! = n * (n-1) * (n-2) * \dots, 3 * 2 * 1.$$

Exemplo 3.7. a) $3! = 3 * 2 * 1 = 6$

b) $10! = 10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1 = 3628800$

c) $0! = 1$ (Definição)

Exemplo 3.8. Encontre o número de permutações de 6 objetos, a, b, c, d, e, e f, tomados 3 a 3. Em outras palavras, encontre o número de palavras de 3 letras, com letras distintas, que podem ser formadas com as 6 letras acima.

Solução

A primeira letra pode ser escolhida de 6 maneiras diferentes; seguindo isto, a segunda letra pode ser escolhida de 5 maneiras diferentes; e, ainda, a última letra pode ser escolhida de 4 maneiras diferentes. Com isso, o número de palavras é dado por: $6 * 5 * 4 = 120$.

Utilizando a fórmula, temos que $n = 6$ e $r = 3$, logo:

$$\begin{aligned} A_{6,3} &= \frac{6!}{(6-3)!} \\ &= \frac{6 * 5 * 4 * 3 * 2 * 1}{3 * 2 * 1} \\ &= 6 * 5 * 4 = 120 \end{aligned}$$

Exercício 3.8. Calcule:

a) $\frac{10!}{5!}$

b) $\frac{8!}{5! * 2!}$

c) $\frac{(n+1)!}{(n-1)!}$

Exercício 3.9. Em uma corrida com 12 participantes, de quantas maneiras distintas podemos ter as três primeiras colocações?

Exercício 3.10. Com oito pessoas que sabem dirigir, de quantas maneiras distintas conseguimos colocar 5 delas em um fusca?

Exercício 3.11. Um banco pede que cada cliente crie uma senha para se utilizar de seu sistema informatizado. Como essa senha deve ter 5 algarismos distintos, quantos são as possíveis senhas? E se pudesse haver repetição?

Exercício 3.12. Seja uma urna contendo 8 bolas. Ache o número de amostras ordenadas de tamanho 3 considerando:

a) Com reposição;

b) Sem reposição;

3.3.4 Combinação

Denominamos de $C_{n,r}$ e calculamos por

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

a combinação de n objetos, tomados r a r . Ou seja, de um conjunto com n elementos, retiramos um subconjunto com r elementos, sem nos preocuparmos com a ordem de escolha, com $C_{n,r}$ possibilidades. Observe o exemplo a seguir e veja a diferença entre combinação e Arranjo.

Exemplo 3.9. De quantas maneiras podemos selecionar um subconjunto com 3 elementos, do conjunto formado pelas letras a , b , c e d ?

Solução

A tabela 3.1 apresenta todas as possibilidades de se obter subconjuntos, com 3 elementos, de um conjunto contendo as letras a , b , c e d . Aqui, conseguimos identificar qual a diferença entre considerar ordem ou não. Podemos observar que ao considerarmos a ordem na formação dos subconjuntos, obtemos um número maior de possibilidades, enquanto que sem considerar a ordem, obtemos um número menor de possibilidades.

Com isso, temos que

$$A_{4,3} = \frac{4!}{(4-3)!} = 24$$

$$C_{4,3} = \frac{4!}{3!(4-3)!} = 4$$

	Combinações	Arranjos					
h!	abc	abc	acb	bac	bca	cab	cba
	abd	abd	adb	bad	bda	dab	dba
	acd	acd	adc	cad	cda	dac	dca
	bcd	bcd	bdc	cbd	cdb	dbc	dcb

Tabela 3.1:

Exercício 3.13. Com 10 espécies de frutas, quantos tipos de salada, contendo 6 espécies diferentes podem ser feitas?

Exercício 3.14. Numa reunião com 7 rapazes e 6 moças, quantas comissões podemos formar com 3 rapazes e 4 moças?

Exercício 3.15. Em um determinado jogo de baralho, todas as 52 cartas são distribuídas igualmente entre os 4 jogadores. Quantas são as possíveis distribuições das cartas?

Exercício 3.16. De quantas maneiras pode um professor escolher um ou mais estudantes dentre seis elegíveis?

3.4 Definição e Propriedades

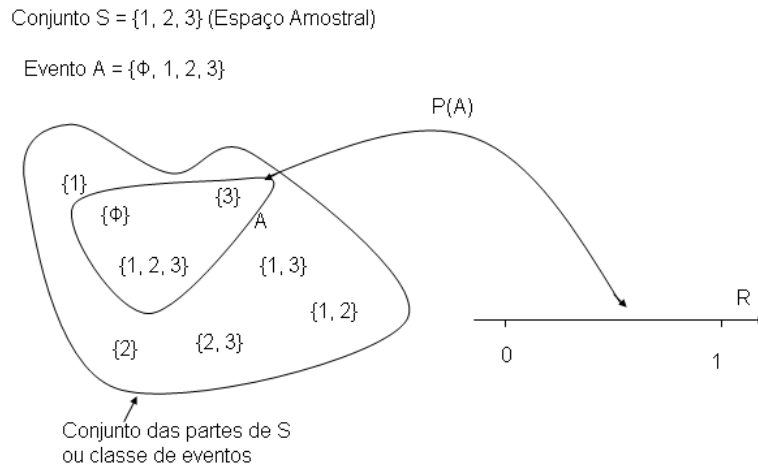
A probabilidade é uma forma de atribuímos “pesos” relativo a ocorrência dos eventos. A probabilidade, que denotaremos por P , é uma função que tem domínio na classe de eventos e tem como imagem números (pesos) entre 0 e 1, veja figura 3.4. Com isso, sejam S um espaço amostral, ξ a classe de eventos e P uma função de valor real definida em ξ . Então, P é chamada de função de probabilidade e $P(A)$, de probabilidade do evento A , se os seguintes axiomas valem:

1. Para todo evento A , $0 \leq P(A) \leq 1$;
2. $P(S) = 1$;
3. Se A_1, \dots, A_n, \dots são mutuamente exclusivos, isto é, $A_i \cap A_j = \emptyset$, $i \neq j$, então $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Se os elementos de um espaço amostral $S = e_1, e_2, \dots, e_n$ (finito) são equiprováveis, isto é, todos os elementos do espaço amostral tem o mesmo “peso” (probabilidade) de ocorrer, temos que

$$P(\{e_i\}) = \frac{1}{n}$$

Neste caso, podemos definir a probabilidade de um evento $E = \{e_1, \dots, e_k\}$, composto por k (com k menor ou igual que n) elementos, como sendo:



$$P(E) = \frac{\text{número de casos favoráveis a } E}{\text{número de casos possíveis de } S} = \frac{k}{n}$$

Exemplo 3.10. Considere o lançamento de um dado honesto. Neste caso, os elementos do espaço amostral $S = \{1, 2, 3, 4, 5, 6\}$ são equiprováveis, pois cada resultado tem a mesma chance de ocorrer, isto é,

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$$

Assim, temos que

$$P(A) = P(\{2, 4, 6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}$$

Com isso, obtemos que a probabilidade de ocorrer o evento A é igual ao número de elementos favoráveis a $A = \{2, 4, 6\}$ que é 3 (pois A tem 3 elementos) dividido pelo número de elementos no espaço amostral que é 6. Desta forma, para os eventos $A = \{2, 4, 6\}$, $B = \{1, 3, 5\}$ e $C = \{4, 5, 6\}$, obtemos

$$P(A) = \frac{3}{6} \quad , \quad P(B) = \frac{3}{6} \quad , \quad P(C) = \frac{3}{6}$$

$$P(A \cup B) = \frac{6}{6} = 1 \quad , \quad P(A \cap B) = \frac{0}{6} = 0$$

$$P(A \cup C) = \frac{4}{6} \quad , \quad P(A \cap C) = \frac{2}{6}$$

Uma propriedade importante para calcularmos a probabilidade de ocorrência de eventos associados ao experimento é a regra da soma (união) de dois eventos.

Regra da Soma: a probabilidade da união de dois eventos E e F pode ser calculada por

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Exemplo 3.11. Considere o exemplo 3.10. Queremos calcular $P(A \cup C)$. Temos

$$P(A \cup C) = P(A) + P(C) - P(A \cap C) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6}$$

Seguem alguns teoremas importantes de probabilidade:

1. Se ϕ é o conjunto vazio, então $P(\phi) = 0$;
2. Se A^C é o complemento de um evento A , então $P(A^C) = 1 - P(A)$;
3. Se $A \subset B$, então $P(A) \leq P(B)$;
4. Se A e B são dois eventos quaisquer, então $P(A - B) = P(A) - P(A \cap B)$;
5. Para quaisquer eventos A , B e C , temos

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \quad .$$

Exercício 3.17. Para ser membro de uma associação profissional é preciso ter o mestrado ou três anos de experiência como gerente sênior. Se existem 3 milhões de engenheiros de computação no Brasil, dos quais 20.000 possuem mestrado, 40.000 possuem três anos de experiência e 5.000 possuem ambos, qual a probabilidade de selecionarmos um engenheiro da computação, membro da associação profissional, dentre os 3 milhões, que possua mestrado ou experiência?

Solução

Sejam

S : Conjunto de todos os engenheiros de computação no Brasil;

A : Conjunto de engenheiros de computação com mestrado;

B : Conjunto de engenheiros de computação com 3 anos de experiência como gerente sênior;

$A \cap B$: Conjunto de engenheiros de computação com mestrado e experiência como gerente sênior;

com isso, temos que

$$P(A) = \frac{20000}{3000000} = 0,00667$$

$$P(B) = \frac{40000}{3000000} = 0,01334$$

$$P(A \cap B) = \frac{5000}{3000000} = 0,001667$$

Queremos saber

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\&= 0,00667 + 0,01334 - 0,001667 \\&= 0,0183\end{aligned}$$

Exercício 3.18. O experimento consiste em lançar dois dados e observar a diferença dos pontos das faces superiores. Determine o espaço amostral do experimento. (Sugestão: Use o diagrama de árvore).

Exercício 3.19. O experimento consiste em lançar três moedas e observar a diferença entre o número de caras e o número de coroas obtidos no lançamento. (Sugestão: Use o diagrama de árvore).

Exercício 3.20. O experimento consiste em retirar duas cartas de um baralho comum e anotar ordenadamente os naipes destas cartas. Determine o espaço amostral do experimento. (Sugestão: Use o diagrama de árvore).

Exercício 3.21. Uma urna contém duas peças defeituosas e três peças boas. Uma a uma as peças serão retiradas sem reposição, e analisadas. O experimento será encerrado quando as peças defeituosas forem identificadas. Determine o espaço amostral do experimento. (Sugestão: Use o diagrama de árvore).

Exercício 3.22. Com base no experimento realizado no exercício 3.18, calcule a probabilidade da diferença estar entre -3 e 3, inclusive.

Exercício 3.23. Com base no experimento realizado no exercício 3.19, calcule a probabilidade da diferença estar entre -3 e 3, exclusive.

Exercício 3.24. Com base no experimento realizado no exercício 3.20, calcule a probabilidade de retirar um naipe de copas na segunda retirada ou dois naipes de espada.

Exercício 3.25. Com base no experimento realizado no exercício 3.21, calcule a probabilidade das peças defeituosas serem identificadas na terceira retirada.

Exercício 3.26. Se há três pneus defeituosos em um lote de 20, e se escolhem quatro pneus do lote para uma inspeção, qual é a probabilidade de que um dos pneus defeituosos seja incluído?

Exercício 3.27. Extraem-se duas cartas de um baralho. quais as probabilidades de se obter:

a) Dois ases;

b) Duas cartas pretas;

c) Duas cartas de ouros?

Exercício 3.28. Seja C o evento "às 9:30 da manhã um médico está em seu consultório" e D é o evento "ele está no hospital", com $P(C) = 0,48$ e $P(D) = 0,27$. Calcule a probabilidade dele não estar nem no consultório e nem no hospital. (Sugestão: Use as leis de D'Morgan.

Exercício 3.29. Seja A o evento "um estudante fica em casa para estudar", e B o evento "ele vai ao cinema", com $P(A) = 0,64$ e $P(B) = 0,21$. Determine:

a) $P(A^C)$;

b) $P(A \cup B)$;

c) $P(A \cap B)$;

c) $P(A^C \cap B^C)$.

3.5 Probabilidade Condicional e Independência

3.5.1 Probabilidade Condicional e Teorema da multiplicação

A probabilidade de ocorrer um evento E dado que ocorreu um evento F é dada por:

$$P(E / F) = \frac{P(E \cap F)}{P(F)}$$

A figura 3.1 ilustra os conjuntos envolvidos no cálculo da Probabilidade Condicional. Observe que ao considerar o cálculo da condicional, mudamos o espaço amostral, antes era S , depois passa a ser F !

Exemplo 3.12. Na tabela 3.2 temos dados referentes a alunos matriculados nos cursos de computação e agronomia, da UFG.

		Sexo		total
		Homens	Mulheres	
Curso	Computação	150	80	230
	Agronomia	185	120	305
Total		335	200	535

Tabela 3.2: Número de Alunos Matriculados nos cursos de agronomia e computação, da UFG

Diante disso, se escolhermos ao acaso uma pessoa, calcule:

a) A probabilidade dela ser do curso de computação;

b) Se soubermos que a pessoa escolhida é do curso de computação, qual a probabilidade dela ser do sexo feminino;

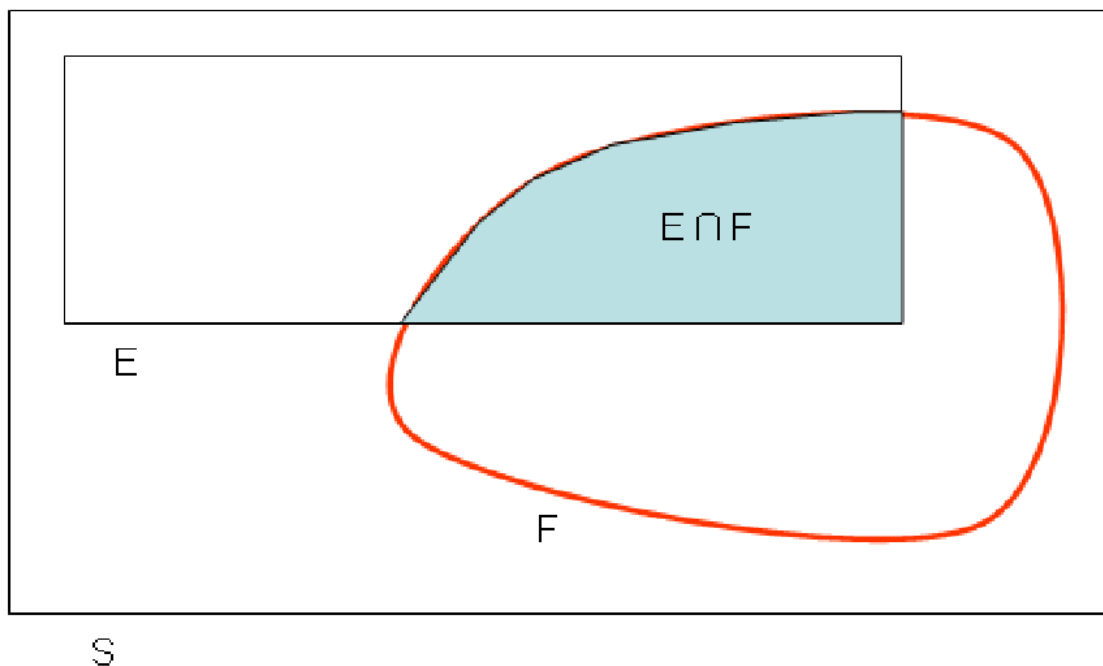


Figura 3.1:

c) A probabilidade dela ser do sexo feminino;

Solução

Ítem a):

Temos que o número de elementos do espaço amostral é 535 alunos. Seja A o evento "aluno é do curso de computação", o qual possui 230 elementos (alunos). Com isso,

$$P(A) = \frac{\text{número de elementos do conjunto } A}{\text{números de elementos do espaço amostral}} = \frac{230}{535} = 0,43$$

Ítem b):

Neste caso, queremos calcular a probabilidade de escolher uma pessoa do sexo feminino sabendo que ela é do curso de computação. Sejam A o evento "pessoa é do curso de computação", e B o evento "pessoa é do sexo feminino". Portanto, temos que calcular:

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

temos que:

$$P(A) = \frac{\text{número de elementos do conjunto } A}{\text{números de elementos do espaço amostral}} = \frac{230}{535} = 0,43$$

$$P(A \cap B) = \frac{\text{número de elementos do conjunto } A \cap B}{\text{números de elementos do espaço amostral}} = \frac{80}{535} = 0,15$$

Diante disso, temos que:

$$P(B / A) = \frac{P(B \cap A)}{P(A)} = \frac{0,15}{0,43} = 0,35$$

Ítem c):

Temos que o número de elementos do espaço amostral é 535 alunos. Seja A o evento "aluno é do sexo feminino", o qual possui 200 elementos (alunos). Com isso,

$$P(A) = \frac{\text{número de elementos do conjunto } A}{\text{números de elementos do espaço amostral}} = \frac{200}{535} = 0,37$$

A partir da fórmula utilizada para calcular a probabilidade condicional, obtemos o **teorema da multiplicação**, que é dado por

$$P(E \cap F) = P(F) \times P(E / F)$$

Com isso, concluímos que a probabilidade de ocorrência simultânea dos eventos E e F é igual a probabilidade de ocorrência do evento F (ou E) vezes a probabilidade de ocorrência do evento E (ou F) dado que ocorreu o evento F (ou E).

Exemplo 3.13. Em uma concessionária de máquinas agrícolas usadas, existem 10 colheitadeiras, sendo 7 boas (B) e 3 defeituosas (D), os defeitos não são perceptíveis e não comprometem o funcionamento da máquina. Se um comprador selecionar duas máquinas, ao acaso e sem reposição, qual a probabilidade dele escolher duas máquinas defeituosas?

Solução

$$P(D_1 \cap D_2) = P(D_1) \times P(D_2 / D_1) = \frac{3}{10} \times \frac{2}{9} = \frac{6}{90}$$

A figura 3.2 apresenta o diagrama de árvore considerando a situação acima. No diagrama, podemos observar todas as situações possíveis de escolha, bem como o cálculo da probabilidade.

Exercício 3.30. Considere uma urna contendo 12 bolas, das quais 8 são vermelhas e 4 são pretas. Qual a probabilidade de selecionar, ao acaso e sem reposição, 3 bolas e duas serem pretas?

Exercício 3.31. Considere um Barracão, com 1500 sacas de milho, o qual apresenta 8% de perda de produto. Um pequeno comprador obteve um desconto, de modo que ele pudesse pegar 4 sacas, selecionadas ao acaso e sem reposição. Qual a probabilidade dele pegar 3 sacas estragadas?

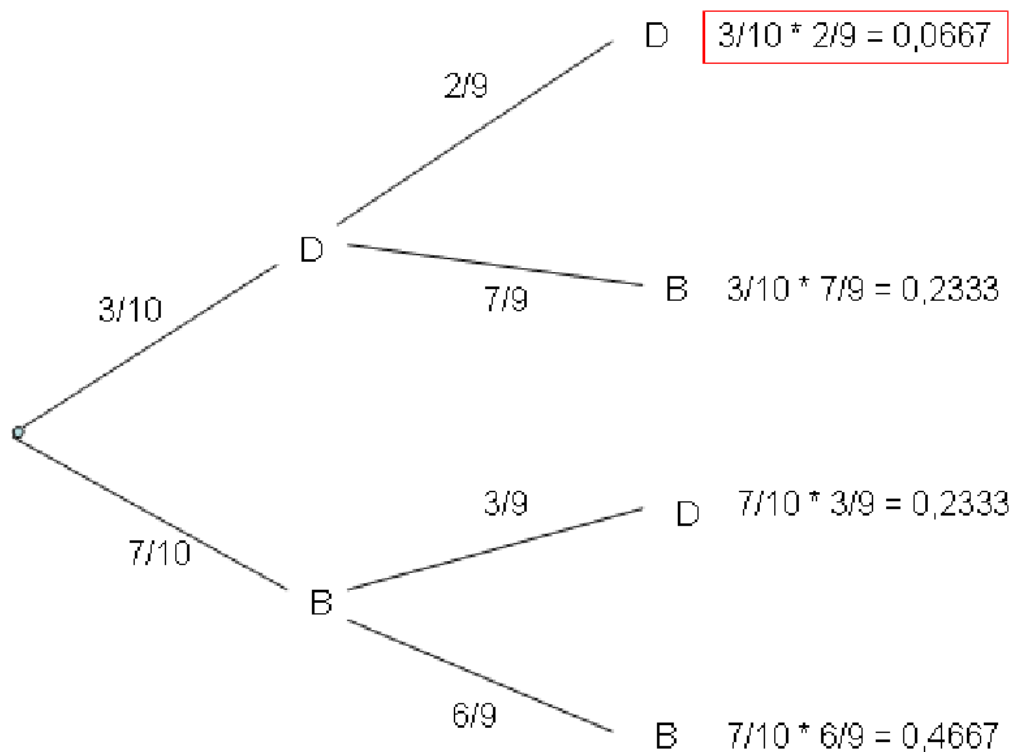


Figura 3.2:

Exercício 3.32. Um júri consiste de nove pessoas naturais do local e três naturais de outros estados. Se dois dos jurados são selecionados, aleatoriamente, para uma entrevista, qual é a probabilidade de serem ambos naturais de outro estado?

Exercício 3.33. Na turma do 3º período de computação composta de 16 alunos e 9 alunas, escolhe-se semanalmente, por sorteio, um deles para servir de assistente do professor de estatística. Qual é a probabilidade de ser escolhida uma aluna duas semanas seguidas, sem que a mesma aluna não possa servir duas semanas seguidas?

Exercício 3.34. Uma organização de pesquisa junto a consumidores estudou os serviços prestados dentro da garantia por 200 comerciantes de pneus na cidade de Jataí, obtendo os resultados resumidos na tabela 3.3.

Pneus	Bom serviço dentro da garantia	Serviço ruim dentro da garantia	Total
Firestone	64	16	80
Outras marcas	42	78	120
Total	106	94	200

Tabela 3.3:

Qual a probabilidade de escolher um comerciante do pneu firestone, dado que o selecionado presta bons serviços dentro da garantia?

Exercício 3.35. *Se cinco dentre 10 caminhões de entrega de uma companhia não atendem às exigências sobre emissão de fumaça, e se três deles são escolhidos aleatoriamente para uma inspeção, qual é a probabilidade de nenhum dos caminhões escolhidos atender às exigências regulamentares?*

Exercício 3.36. *Considere um processo que apresenta 8% de defeituosos. Duas peças são selecionadas ao acaso e classificadas em defeituosas ou não.*

- a) *Qual o espaço amostral associado ao experimento de selecionar duas peças e classificá-las?*
- b) *Qual a probabilidade de obtermos duas peças defeituosas?*

Exercício 3.37. *Considere um processo composto por duas etapas. A etapa I apresenta 5% de peças defeituosas, enquanto que a etapa II apresenta 9% de peças defeituosas. Qual a probabilidade do processo fornecer uma peça sem defeito?*

3.5.2 Independência

Dois eventos são independentes quando a ocorrência de um evento não influencia na ocorrência ou não do outro evento. Do ponto de vista probabilístico, definimos:

Independência: Dois eventos E e F são ditos “independentes” se

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

Exemplo 3.14. *Uma caixa contém 10 peças, sendo 7 boas (B) e 3 defeituosas (D). Retiramos duas peças, ao acaso e com reposição, para inspeção. Qual a probabilidade de se obter duas peças defeituosas?*

Resposta:

O experimento de realizar a primeira retirada tem como espaço amostral $S_1 = \{D_1; B_1\}$ e a segunda retirada tem como espaço amostral $S_2 = \{D_2; B_2\}$, onde D_i significa que retiramos uma peça Defeituosa na i -ésima retirada e B_i significa que retiramos uma peça Boa na i -ésima retirada, para $i = 1, 2$. Além disso, temos que

$$P(D_1) = P(D_2) = \frac{3}{10} \quad e \quad P(B_1) = P(B_2) = \frac{7}{10}$$

Pois as duas peças são retiradas ao acaso e com reposição, isto é, após retirarmos a primeira peça, esta é a resposta à caixa para que possamos efetuar a segunda retirada. Associamos ao experimento de retirar duas peças ao acaso e com reposição o espaço amostral

$$S = \{(D_1, B_2); (B_1, D_2); (D_1, D_2); (B_1, B_2)\} \quad .$$

Desde que a primeira e a segunda retiradas são executadas de forma independente, temos que

$$P[(D_1; D_2)] = P(D_1 \cap D_2) = P(D_1) \times P(D_2) = \frac{3}{10} \times \frac{3}{10} = \frac{9}{100}$$

Muitas vezes precisamos calcular a probabilidade da ocorrência de dois eventos simultaneamente. Para efetuarmos tal cálculo, introduzimos o conceito de probabilidade condicional.

Exercício 3.38. *Considere uma urna contendo 12 bolas, das quais 8 são vermelhas e 4 são pretas. Qual a probabilidade de selecionar, ao acaso e com reposição, 3 bolas e duas serem pretas?*

Exercício 3.39. *As probabilidades de um estudante ser aprovado em exames de estatística, de contabilidade ou de ambos são $P(E) = 0,70$, $P(C) = 0,80$ e $P(E \cap C) = 0,56$, respectivamente. Verifique se os eventos E e C são independentes.*

Exercício 3.40. *As probabilidades de chover ou nevar em determinada cidade no dia de Natal (C), no dia de Ano Novo (N) ou em ambos os dias são $P(C) = 0,60$, $P(N) = 0,60$ e $P(C \cap N) = 0,42$, respectivamente. Verifique se os eventos C e N são independentes.*

Exercício 3.41. *Extraíndo-se duas cartas de um baralho comum de 52 cartas, qual a probabilidade de serem ambas de ouros, se a extração é feita:*

- a) *Com reposição;*
- b) *Sem reposição?*

Exercício 3.42. *Uma loja de departamentos que fatura mensalmente as compras de seus clientes constatou que, se um cliente paga pontualmente em determinado mês, há 0,90 de probabilidade de ser pontual também no mês seguinte; entretanto, se um cliente não paga pontualmente em determinado mês, a probabilidade de ele ser pontual no mês seguinte é de apenas 0,40. Considerando uma análise em três meses, calcule:*

- a) *a probabilidade de um cliente pagar pontualmente durante os três meses;*
- b) *a probabilidade de um cliente não pagar pontualmente nos dois primeiros meses e, em seguida, pagar pontualmente?*

3.6 Teorema de Bayes

Suponha que os eventos A_1, A_2, \dots, A_n formam uma partição de um espaço amostral S , ou seja, os eventos A_i são mutuamente exclusivos e sua união é S . Seja B ser outro evento qualquer (Veja

ilustração na figura 3.3). Então,

$$\begin{aligned} B = S \cap B &= (A_1 \cup A_2 \cup \dots \cup A_n) \cap B \\ &= (A_1 \cap B) \cup (A_2 \cap B) \cup \dots \cup (A_n \cap B) \end{aligned}$$

onde os $A_i \cap B$ são também mutuamente exclusivos.

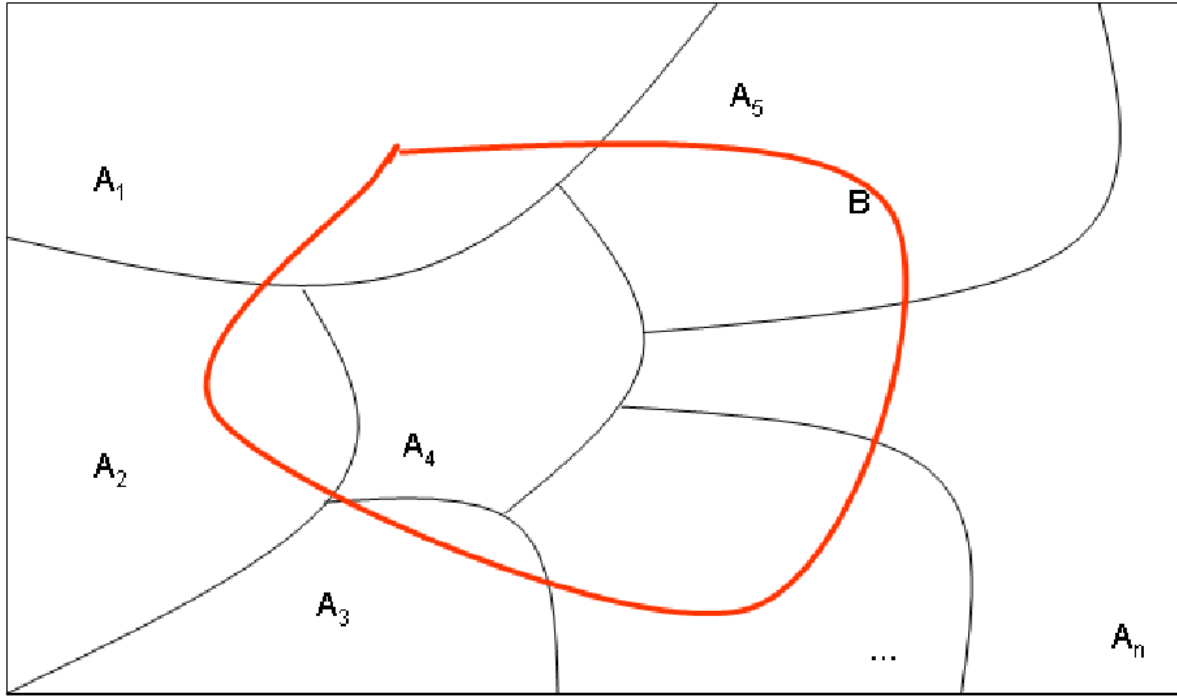


Figura 3.3:

Consequentemente,

$$\begin{aligned} P(B) &= P((A_1 \cap B) \cup (A_2 \cap B) \cup \dots \cup (A_n \cap B)) \\ &= P((A_1 \cap B)) + P((A_2 \cap B)) + \dots + P((A_n \cap B)) \end{aligned}$$

Assim, pelo teorema da multiplicação,

$$P(B) = P(A_1) P(B/A_1) + P(A_2) P(B/A_2) + \dots + P(A_n) P(B/A_n)$$

Por outro lado, para qualquer i , a probabilidade condicional de A_i dado B é definida como

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)}$$

Diante disso, definimos o teorema de Bayes como sendo

$$P(A_i/B) = \frac{P(A_i) P(B/A_i)}{P(A_1) P(B/A_1) + P(A_2) P(B/A_2) + \dots + P(A_n) P(B/A_n)}$$

Exemplo 3.15. Três máquinas, A , B e C , produzem 50%, 30% e 20%, respectivamente, do total de peças de uma fábrica. As porcentagens de produção defeituosa destas máquinas são 3%, 4% e 5%. Se uma peça é selecionada aleatoriamente, ache a probabilidade:

a) *dela ser defeituosa;*

b) *Dado que a peça selecionada é defeituosa, qual é a probabilidade dela ter sido produzida pela máquina A?*

Solução

Ítem a)

Seja D o evento em que uma peça é defeituosa. Então, temos que

$$\begin{aligned} P(D) &= P(A) P(D/A) + P(B) P(D/B) + P(C) P(D/C) \\ &= (0,50) (0,03) + (0,30) (0,04) + (0,20) (0,05) = 0,037 \end{aligned}$$

Ítem b)

O teorema de bayes para este caso é:

$$\begin{aligned} P(A/D) &= \frac{P(A) P(D/A)}{P(A) P(D/A) + P(B) P(D/B) + P(C) P(D/C)} \\ &= \frac{P(A) P(D/A)}{P(D)} \\ &= \frac{0,50) (0,03)}{0,037} = 0,4054 \end{aligned}$$

Exercício 3.43. *Um grande número de caixas de bombons são compostas de dois tipos, A e B. O tipo A contém 70 por cento de bombons doces e 30 por cento de bombons amargos, enquanto no tipo B essas porcentagens de sabor são inversas. Além disso, suponha-se que 60 por cento de todas as caixas de bombons sejam do tipo A, enquanto as restantes sejam do tipo B. Ao retirar um bombom de uma determinada caixa e experimentar, verificou-se que é de sabor doce, com isso, qual a probabilidade de que esse bombom tenha vindo da caixa A? e da caixa B?*

Exercício 3.44. *Num certo colégio, 4% dos homens e 1% das mulheres têm mais do que 1,60 m de altura. Além disso, 60% dos estudantes são mulheres. Ora, se um estudante é selecionado aleatoriamente e tem mais do que 1,60 m de altura, qual é a probabilidade de o estudante ser uma mulher?*

Exercício 3.45. *Uma companhia monta rádios cujas peças são produzidas em três de suas fábricas denominadas A₁, A₂ e A₃. Elas produzem, respectivamente, 15%, 35% e 50% do total. As probabilidades das fábricas A₁, A₂ e A₃ produzirem peças defeituosas são 0,01; 0,05 e 0,02, respectivamente.*

Uma peça é escolhida ao acaso do conjunto das peças produzidas. Essa peça é testada e verifica-se que é defeituosa. Qual a probabilidade que tenha sido produzida pela fábrica A_1 ? E pela A_2 ? E pela A_3 ?

Exercício 3.46. São dadas três urnas com as seguintes composições: a urna um tem três bolas brancas e cinco vermelhas, a urna dois tem quatro bolas brancas e duas vermelhas e a urna três uma bola branca e três vermelhas. Escolhe-se uma das três urnas de acordo com as seguintes probabilidades: urna um com probabilidade $\frac{2}{6}$, urna dois com probabilidade $\frac{3}{6}$ e urna três com probabilidade $\frac{1}{6}$. Uma bola é retirada de uma urna e verificou-se ser de cor branca, qual a probabilidade dela ter vindo da urna dois?

Exercício 3.47. Uma Empresa de digitação possui Quatro funcionários, aqui denominados de F_1 , F_2 , F_3 e F_4 . Eles produzem, respectivamente, 11%, 21%, 32% e 36% do total de itens digitados. As probabilidades dos digitadores F_1 , F_2 , F_3 e F_4 produzirem itens com erros de digitação são 0,01; 0,05; 0,07 e 0,08, respectivamente. Um item é escolhida ao acaso do conjunto de itens digitados. Esse item é analisado e verificou-se que apresenta erro de digitação. Qual a probabilidade que tenha sido digitado pelo Funcionário F_1 ? E pelo F_2 ? E pelo F_3 ?

Capítulo 4

Variáveis Aleatórias

4.1 Variáveis Aleatórias

Dado um fenômeno aleatório qualquer, com um certo espaço amostral, desejamos estudar a estrutura probabilística de quantidades associadas a esse fenômeno.

Por exemplo, ao descrever uma peça manufaturada podemos empregar apenas duas categorias: “defeituosa” ou “não defeituosa”. Mas, estamos na verdade interessados em atribuir um número real à cada resultado do experimento. Assim podemos atribuir o valor 1 às peças perfeitas e 0 às defeituosas. Então, podemos entender por variável aleatória uma função que associa à cada evento (neste exemplo os eventos são “defeituosa”, “não defeituosa”) um número real, que denotaremos pelas últimas letras (maiúsculas) do alfabeto: X, Y, Z .

Definição 4.1.1. Variável Aleatória

Consideremos um experimento e S o espaço amostral associado a esse experimento. Uma função X , que associa a cada elemento $s \in S$ um número real, $X(s)$, é denominada variável aleatória (v.a.). Ou seja, variável aleatória é um caractereístico numérico do resultado de um experimento.

Exemplo 4.1. *Considere três lançamentos independentes de uma moeda equilibrada.*

Seja C cara e K coroa.

O espaço amostral é $S = \{CCC, CCK, CKC, KCC, KKK, KKC, KCK, CKK\}$.

Podemos definir a variável aleatória X : “número de caras obtidas nos três lançamentos”.

□

4.2 Função de Distribuição Acumulada

A função de distribuição acumulada nos dá uma maneira de descrever como as probabilidades são associadas aos valores ou aos intervalos de valores de uma variável aleatória.

Definição 4.2.1. Função de Distribuição Acumulada

A função de distribuição acumulada de uma variável aleatória X é uma função que a cada número real x associa o valor:

$$F(x) = P[X \leq x].$$

A notação $[X \leq x]$ é usada para designar o conjunto $\{\omega \in S : X(\omega) \leq x\}$, isto é, denota a imagem inversa do intervalo $(-\infty, x]$ pela variável aleatória X .

O conhecimento da função de distribuição permite obter qualquer informação sobre a variável. Mesmo que a variável assuma valores apenas num subconjunto dos reais, a função de distribuição é definida em toda a reta. Ela é chamada de função de distribuição acumulada pois acumula as probabilidades dos valores inferiores ou iguais a x .

Exemplo 4.2. Consideremos o Exemplo 4.1. Vamos encontrar a função distribuição acumulada de X : “número de caras obtidas nos três lançamentos”.

Os valores que X pode assumir são: 0,1,2 e 3.

$$\begin{aligned} P(X = 0) &= P(\{KKK\}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}. \\ P(X = 1) &= P(\{KCC\}) + P(\{CKC\}) + P(\{CCK\}) = \frac{3}{8}. \\ P(X = 2) &= P(\{KKC\}) + P(\{KCK\}) + P(\{CKK\}) = \frac{3}{8}. \\ P(X = 3) &= P(\{KKK\}) = \frac{1}{8}. \end{aligned}$$

Assim temos que a função de distribuição acumulada de X é dada por

$$F(x) = \begin{cases} 0, & \text{se } x < 0; \\ 1/8, & \text{se } 0 \leq x < 1; \\ 1/2, & \text{se } 1 \leq x < 2; \\ 7/8, & \text{se } 2 \leq x < 3; \\ 1, & \text{se } x \geq 3. \end{cases}$$

Exemplo 4.3. O tempo de validade, em meses, de um óleo lubrificante num certo equipamento está sendo estudado. Seja $S = \{\omega \in \mathbb{R} : 6 < \omega \leq 8\}$. Uma variável de interesse é o próprio tempo de validade e, nesse caso, definimos $X(\omega) = \omega, \forall \omega \in S$.

A função de distribuição acumulada de X é dada por

$$F(x) = \begin{cases} 0, & \text{se } x < 6; \\ (x - 6)/2, & \text{se } 6 \leq x < 8; \\ 1, & \text{se } x \geq 8. \end{cases}$$

□

Proposição 4.2.1. Propriedades da Função de Distribuição Acumulada

A função de distribuição acumulada de uma variável aleatória X satisfaz as seguintes condições:

1. $0 \leq F(x) \leq 1$
2. $F(x)$ é não decrescente e contínua à direita
3. $\lim_{x \rightarrow -\infty} F(x) = 0$ e $\lim_{x \rightarrow \infty} F(x) = 1$

Exemplo 4.4. Para o lançamento de uma moeda, temos que $S = \{\text{cara}, \text{coroa}\}$ e que $P(\text{cara}) = P(\text{coroa}) = 1/2$. Definimos uma função X , variável aleatória, de S em \mathbb{R} da seguinte forma:

$$X(\omega) = \begin{cases} 1, & \text{se } \omega = \text{cara}; \\ 0, & \text{se } \omega = \text{coroa}. \end{cases}$$

Encontrar a função de distribuição acumulada de X .

Para obter a função de distribuição acumulada é conveniente separar os vários casos, de acordo com os valores da variável.

Para $x < 0$, $P(X \leq x) = 0$, uma vez que o menor valor assumido pela variável é 0. No intervalo $0 \leq x < 1$, temos $P(X \leq x) = P(X = 0) = 1/2$. E, para $x \geq 1$, vem $P(X \leq x) = P(X = 0) + P(X = 1) = 1$. Dessa forma, $F(x) = P(X \leq x)$ foi definida para todo x real. Assim, temos

$$F(x) = \begin{cases} 0, & \text{se } x < 0; \\ 1/2, & \text{se } 0 \leq x < 1; \\ 1, & \text{se } x \geq 1. \end{cases}$$

Note que as propriedades de função de distribuição são facilmente verificadas. A propriedade 1 é imediata. Para a propriedade 2 observe que, exceto nos pontos 0 e 1, F é contínua nos reais. Para os pontos 0 e 1 temos continuidade à direita, isto é,

$$F(0) = \lim_{x \rightarrow 0^+} F(x) \quad \text{e} \quad F(1) = \lim_{x \rightarrow 1^+} F(x).$$

Observe que $F(x)$ é não decrescente para todo x real e, assim, vale a propriedade 3.

□

A classificação das variáveis aleatórias é feita de acordo com os valores que assumem. Veremos agora dois tipos de variáveis aleatórias: variável aleatória discreta e variável aleatória contínua.

4.3 Variável Aleatória Discreta

Definição 4.3.1. Variável Aleatória Discreta

Seja X uma variável aleatória (v.a.). Se o número de valores possíveis de X for finito ou infinito enumerável, denominaremos X de variável aleatória discreta. Isto é, os possíveis valores de X podem ser postos em lista como x_1, x_2, \dots, x_n . No caso finito, a lista acaba, e no caso infinito numerável, a lista continua indefinidamente.

Exemplo 4.5. Suponha que peças saiam de uma linha de produção e sejam classificadas como defeituosas (D) e não-defeituosas (N). Admita que três peças, da produção diária, sejam escolhidas ao acaso e classificadas. O espaço amostral é dado por

$$S = \{DDD, DDN, DND, NDD, NND, NDN, DNN, NNN\}.$$

Nosso interesse é saber quantas peças defeituosas foram encontradas, não interessando a ordem que tenham ocorrido. Isto é, desejamos estudar a variável aleatória X , a qual atribui a cada resultado $s \in S$ o número de peças defeituosas. Conseqüentemente, o conjunto dos possíveis valores de X é $\{0, 1, 2, 3\}$, ou seja, X é uma variável aleatória discreta.

□

Definição 4.3.2. Função de Probabilidade

Seja X uma variável aleatória discreta. A cada possível resultado x_i associaremos um número $p(x_i) = P[X = x_i]$, denominado probabilidade de x_i . Os números $p(x_i)$, $i = 1, 2, \dots$, devem satisfazer às seguintes condições:

a) $p(x_i) \geq 0$ para todo i ,

b) $\sum_{i=1}^{\infty} p(x_i) = 1$.

A função p é denominada função de probabilidade da variável aleatória X .

Definição 4.3.3. Distribuição de Probabilidade

A coleção de pares $[x_i, p(x_i)]$, $i = 1, 2, \dots$ é algumas vezes denominada distribuição de probabilidade de X . Assim, podemos falar que a distribuição de probabilidades de uma variável aleatória discreta X , definida em um espaço amostral (S), é uma tabela que associa a cada valor de X sua probabilidade.

Exemplo 4.6. Considere que uma moeda é lançada duas vezes. Seja X a função definida no espaço amostral que é igual ao número de caras nos dois lançamentos (C - Cara e \bar{C} - Coroa).

Valores de X	Pontos amostrais	Probabilidades
0	\overline{CC}	$1/4$
1	$\overline{CC}, C\overline{C}$	$1/2$
2	CC	$1/4$

Tabela 4.1: Distribuição de probabilidade.

Temos na Tabela 4.2 a distribuição de probabilidade referente à variável aleatória X .

Os valores das probabilidades, na tabela acima, são obtidos da seguinte maneira:

$$\begin{aligned}
 P[X = 0] &= P(\overline{CC}) = \frac{1}{4} \\
 P[X = 1] &= P(\overline{CC}) + P(C\overline{C}) = \frac{1}{2} \\
 P[X = 2] &= P(CC) = \frac{1}{4}.
 \end{aligned}$$

□

O gráfico da distribuição de probabilidade é ilustrado na figura 4.1. A figura 4.2 ilustra a função distribuição acumulada.

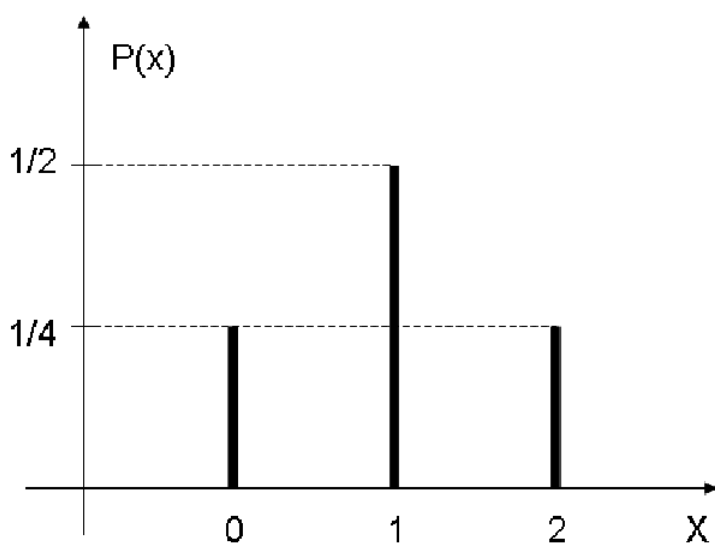


Figura 4.1: Gráfico da distribuição de probabilidade

4.4 Relação entre a Função de Distribuição Acumulada e a Distribuição de Probabilidade Discreta

Seja X uma variável aleatória discreta cuja distribuição de probabilidade associa aos valores

$$x_1, x_2, \dots, x_n$$

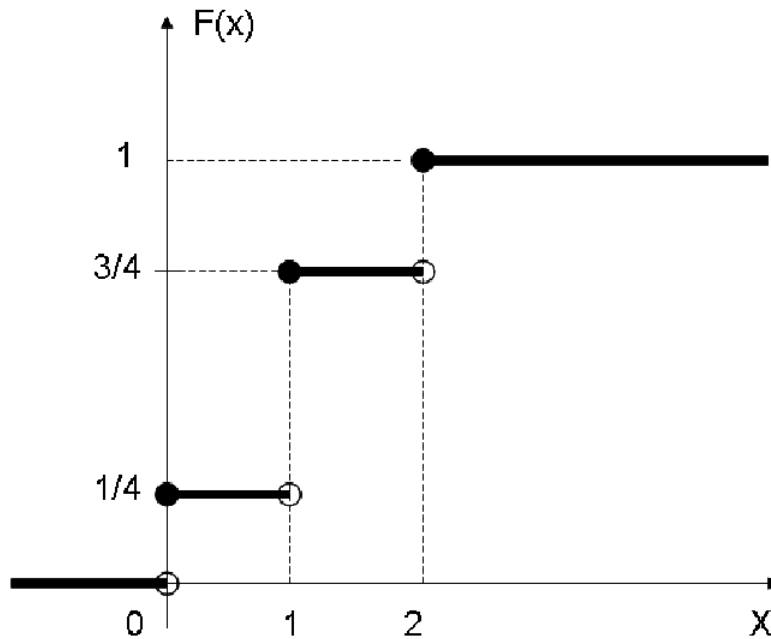


Figura 4.2: Gráfico da distribuição acumulada

as respectivas probabilidades

$$P[X = x_1], P[X = x_2], \dots, P[X = x_n].$$

Como os valores de X são mutuamente exclusivos, temos que a função de distribuição acumulada é dada por

$$F(x) = \sum_{i \in A_x} P[X = x_i], \text{ com } A_x = \{i : x_i \leq x\}.$$

Assim, dada a distribuição de probabilidade de uma variável aleatória discreta, conseguimos determinar sua função de distribuição acumulada.

Exemplo 4.7. Considere dois lançamentos independentes de uma moeda equilibrada. Com o espaço de probabilidade usual, defina X como sendo o número de caras nos dois lançamentos. Determine a função de distribuição acumulada de X .

A variável X é discreta e sua função de probabilidade será dada por

X	0	1	2
$p(x_i)$	1/4	1/2	1/4

A função de distribuição acumulada correspondente será:

$$F(x) = \begin{cases} 0, & \text{se } x < 0; \\ 1/4, & \text{se } 0 \leq x < 1; \\ 3/4, & \text{se } 1 \leq x < 2; \\ 1, & \text{se } x \geq 2. \end{cases}$$

4.5 Variável Aleatória Contínua

Definição 4.5.1. Variável Aleatória Contínua

Seja X uma variável aleatória. Suponha-se que o contradomínio (\mathbb{R}_x) de X seja um intervalo ou uma coleção de intervalos. Então diremos que X é uma variável aleatória contínua.

Exemplo 4.8. Um válvula eletrônica é instalada em um circuito e X , o período de tempo em que funcione, é observado.

Neste caso, X é uma variável aleatória contínua podendo tomar valores $x \geq 0$.

□

Definição 4.5.2. Função Densidade de Probabilidade

Seja X uma variável aleatória contínua. A função densidade de probabilidade f , indicada abreviamente por f_{dp} , é uma função que satisfaz às seguintes condições:

$$a) f(x) \geq 0 \text{ para todo } x \in \mathbb{R}_x;$$

$$b) \int_{\mathbb{R}_x} f(x)dx = 1.$$

Além disso, definimos para qualquer $c, d \in \mathbb{R}_x$, $c < d$,

$$P(c < X < d) = \int_c^d f(x)dx.$$

4.6 Relação entre a Função de Distribuição Acumulada e a Função Densidade de Probabilidade Contínua

Para uma variável aleatória contínua com densidade de probabilidade $f(x)$ podemos obter a função de distribuição acumulada $F(x)$ integrando a função densidade de probabilidade,

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(y)dy.$$

Se a densidade $f(x)$ for contínua no seu campo de definição, então decorre do teorema fundamental do cálculo que:

$$F'(x) = f(x).$$

Exemplo 4.9. Seja X uma variável contínua com f_{dp}

$$f(x) = \begin{cases} 2x, & 0 < x < 1; \\ 0, & \text{para quaisquer outros valores.} \end{cases}$$

Portanto, a função de distribuição acumulada é dada por

$$F(x) = \begin{cases} 0, & \text{se } x \leq 0; \\ \int_0^x 2s \, ds = x^2, & \text{se } 0 < x \leq 1; \\ 1, & \text{se } x > 1. \end{cases}$$

□

4.7 Esperança de Variáveis Aleatórias

4.7.1 Esperança de Variáveis Aleatórias Discretas

Definição 4.7.1. Valor Esperado para Variáveis Discretas

Seja X uma variável aleatória discreta, com valores possíveis x_1, \dots, x_n, \dots . Seja $p(x_i) = P[X = x_i]$, $i = 1, \dots, n, \dots$. Então, a esperança de X , também chamada de valor esperado de X , denotada por $E(X)$ ou μ_X , é definida como

$$E(X) = \mu_X = \sum_{i=1}^{\infty} x_i P[X = x_i]. \quad (4.1)$$

Este número é também denominado valor médio ou expectância ou esperança de X .

Observação: 1) Se X tomar apenas um número finito de valores, a expressão 4.1 se torna $E(X) = \sum_{i=1}^n x_i P[X = x_i]$. Isto pode ser considerado como uma média ponderada dos possíveis valores x_1, \dots, x_n . Se todos esses valores possíveis forem igualmente prováveis, $E(X) = (1/n) \sum_{i=1}^n x_i$ que representa a média aritmética simples ou usual dos n possíveis valores.

$$2) E(X^2) = \sum_{i=1}^{\infty} x_i^2 P[X = x_i].$$

Exemplo 4.10. Consideremos novamente o Exemplo 4.1. Vamos encontrar a esperança de X .

Pela equação 4.1 e com os cálculos feitos no Exemplo 4.2 temos que a esperança de X é dada por

$$\begin{aligned} E(X) &= \sum_{i=1}^3 x_i P(X = x_i) = 1P(X = 1) + 2P(X = 2) + 3P(X = 3) \\ &= \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{5}{4}. \end{aligned}$$

Exemplo 4.11. Consideremos o lançamento de um dado equilibrado. Consideremos a variável aleatória X : “número da face que caiu para cima”. Calcular o valor esperado de X .

Já vimos que os valores possíveis de X são $\{1, 2, 3, 4, 5, 6\}$ e que esses valores são equiprováveis.

Assim,

$$E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2}.$$

□

Observação: Este exemplo ilustra claramente que $E(X)$ não é o resultado que podemos esperar quando X for observado uma única vez. $E(X) = 7/2$ nem é mesmo um valor possível de X ! Esse valor na verdade significa que se jogássemos o dado um grande número de vezes e depois calculássemos a média aritmética dos vários resultados, esperaríamos que essa média ficasse mais próxima de $7/2$ quanto maior número de vezes o dado fosse jogado.

Exemplo 4.12. *Um fabricante produz peças tais que 10% delas são defeituosas e 90% são não-defeituosas. Se uma peça defeituosa for produzida, o fabricante perde R\$1,00, enquanto que uma peça não-defeituosa lhe dá um lucro de R\$5,00. Seja X a v.a. que representa o lucro líquido por peça. Calcule o valor esperado de X .*

Os possíveis valores de X são $\{-1, 5\}$. Além disso, $P[x = -1] = 0,1$ e $P[x = 5] = 0,9$. Portanto, o valor esperado é dado por

$$E(X) = -1(0,1) + 5(0,9) = \text{R\$}4,40.$$

Assim, supondo que um grande número de peças foi produzido, o fabricante espera ganhar cerca de R\$ 4,40 por peça, a longo prazo.

□

4.7.2 Esperança de Variáveis Aleatórias Contínuas

Definição 4.7.2. Valor Esperado para Variáveis Contínuas

Seja X uma variável aleatória contínua com função densidade de probabilidade f . Definimos valor esperado ou esperança matemática ou média de X por

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx, \quad (4.2)$$

desde que a integral esteja bem definida.

Observação: 1) Se a variável é limitada, o cálculo é feito sem ambiguidade e a existência do valor esperado está assegurada. No caso não limitado, podem aparecer situações indefinidas do tipo $\infty - \infty$, em que diremos que a esperança não existe. Assim, temos que $E(X)$ vai estar bem definida se a integral, em pelo menos um desses intervalos, for finita; isto é

$$\int_{-\infty}^0 xf(x) dx < \infty \quad \text{ou} \quad \int_0^{\infty} xf(x) dx < \infty.$$

2) A interpretação de $E(X)$ para o caso contínuo é similar ao mencionado para variáveis discretas.

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

Exemplo 4.13. Seja X o tempo (em minutos) durante o qual um equipamento elétrico é utilizado em carga máxima, em um certo período de tempo especificado. Então, X é uma variável aleatória contínua e sua fdp é dada por

$$f(x) = \begin{cases} \frac{1}{(1500)^2}x, & 0 \leq x \leq 1500; \\ \frac{-1}{(1500)^2}(x - 3000), & 1500 \leq x \leq 3000; \\ 0, & \text{para quaisquer outros valores.} \end{cases}$$

Calcular a esperança de X .

$$E(x) = \int_{-\infty}^{\infty} xf(x) dx = \frac{1}{(1500)(1500)} \left[\int_0^{1500} x^2 dx - \int_{1500}^{3000} x(x - 3000) dx \right] = 1500 \text{ minutos.}$$

□

Exemplo 4.14. Numa empresa, as previsões de despesa para o próximo ano foram calculadas como: R\$ 9, 10, 11, 12 e 13 bilhões. Supondo que as despesas do ano corrente sejam desconhecidas, as seguintes probabilidades foram atribuídas respectivamente: 30%, 20%, 25%, 5% e 20%. Seja X a variável aleatória “despesa referente ao ano i , $i = 1, \dots, 5$ ”. Assim, os possíveis valores de X são $\{9, 10, 11, 12, 13\}$. Qual é a distribuição de probabilidade para o próximo ano e qual o valor esperado das despesas para o próximo ano?

Distribuição de Probabilidade

Ano	Despesa(X)	$P(X)$
1	9	0,30
2	10	0,20
3	11	0,25
4	12	0,05
5	13	0,20
	Total:	1,00

Tabela 4.2: Distribuição de probabilidade.

O valor esperado das despesas é dado por

$$E(X) = 9 \times 0,30 + 10 \times 0,20 + 11 \times 0,25 + 12 \times 0,05 + 13 \times 0,20 = 10,65.$$

□

4.7.3 Propriedades da Esperança

Veremos agora algumas importantes propriedades do valor esperado de uma variável aleatória. Em cada caso, admitimos que todos os valores esperados mencionados, existem.

P1. Seja C uma constante e X uma variável aleatória. Então,

$$E(CX) = CE(X).$$

P2. Sejam X e Y duas variáveis aleatórias quaisquer. Então,

$$E(X + Y) = E(X) + E(Y).$$

P3. Sejam n variáveis aleatórias X_1, \dots, X_n . Então,

$$E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n).$$

P4. Sejam X e Y variáveis aleatórias independentes. Então,

$$E(XY) = E(X)E(Y).$$

4.8 Variância de Variáveis Aleatórias

Suponhamos que, para uma variável aleatória X , verificamos que $E(X) = 2$. Qual o significado disso? Como vimos acima, significa que se considerarmos um grande número de determinações de X , digamos x_1, \dots, x_n , ao calcularmos a média desses valores de X ela estará próxima de 2, se n for grande.

Suponhamos, por exemplo, que X representa a duração de vida de lâmpadas que estão sendo recebidas de um fabricante, e que $E(X) = 1000$ horas. Isto pode significar que a maioria das lâmpadas deve durar um período de tempo compreendido entre 900 horas e 1100 horas. Poderia significar também que as lâmpadas são formadas por dois tipos muito diferentes: cerca da metade são de muita boa qualidade e durarão aproximadamente 1400, enquanto que a outra metade são de muito má qualidade e durarão aproximadamente 600.

Assim, existe uma necessidade óbvia de se introduzir uma medida que possa distinguir entre essas duas situações.

Definição 4.8.1. Variância

Seja X a variável aleatória. Definimos a variância de X , denotada por $Var(X)$ ou σ^2 por

$$Var(X) = E[X - E(X)]^2 \quad (4.3)$$

A raiz quadrada positiva de $Var(X)$ é denominada o desvio-padrão de X , denotada por σ .

Observação: O número $Var(X)$ é expresso por unidades quadradas de X . Isto é, se X for medido em horas, então $Var(X)$ é expressa em (horas)².

O cálculo de $Var(X)$ pode ser simplificado com o auxílio do seguinte resultado.

Proposição 4.8.1.

$$Var(X) = E(X^2) - (E(X))^2.$$

Demonstração: Desenvolvendo $E[X - E(X)]^2$ e empregando as propriedades já estabelecidas de valor esperado, obtemos

$$\begin{aligned} Var(X) &= E[X - E(X)]^2 = E\{X^2 - 2XE(X) + [E(X)]^2\} \\ &= E(X^2) - 2E(X)E(X) + [E(X)]^2 = E(X^2) - [E(X)]^2. \end{aligned}$$

Exemplo 4.15. Suponhamos que X seja uma variável aleatória contínua com fdp

$$f(x) = \begin{cases} 1+x, & -1 \leq x \leq 0; \\ 1-x, & 0 \leq x \leq 1. \end{cases}$$

Calcular $Var(X)$.

Em virtude da simetria da fdp, $E(X) = 0$.

Além disso,

$$E(X^2) = \int_{-1}^0 x^2(1+x)dx + \int_0^1 x^2(1-x)dx = \frac{1}{6}.$$

Portanto, $Var(X) = \frac{1}{6}$.

□

Exemplo 4.16. Vamos calcular a variância da variável X definida no Exemplo 4.1.

$$E(X^2) = P(X=1) + 2^2P(X=2) + 3^2P(X=3) = \frac{3}{8} + 4 \times \frac{3}{8} + 9 \times \frac{1}{8} = \frac{24}{8} = 3.$$

Assim,

$$Var(X) = E(X^2) - [E(X)]^2 = 3 - \left(\frac{5}{4}\right)^2 = 3 - 1,56 = 1,44.$$

□

Exemplo 4.17. Calcule a variância da variável aleatória X referente ao Exemplo 4.14.

Temos pelo Exemplo 4.14 que $E(X) = 10,65$. Além disso,

$$E(X^2) = 9^2 \times 0,30 + 10^2 \times 0,20 + 11^2 \times 0,25 + 12^2 \times 0,05 + 13^2 \times 0,20 = 115,55.$$

Portanto, a variância de X é dada por

$$Var(X) = E(X^2) - [E(X)]^2 = 115,55 - 10,65^2 = 2,13.$$

□

4.8.1 Propriedades da Variância de uma Variável Aleatória

Existem várias propriedades importantes para a variância de uma variável aleatória, em parte análogas às aquelas expostas para o valor esperado.

P1. Se C for constante,

$$\text{Var}(X + C) = \text{Var}(X).$$

Observação: Esta propriedade é intuitivamente evidente, porque somar uma constante a um resultado X não altera sua variabilidade, que é aquilo que a variância mede.

P2. Se C for uma constante,

$$\text{Var}(CX) = C^2 \text{Var}(X).$$

P3. Se X e Y forem variáveis aleatórias *independentes*, então

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Observação: É importante compreender que a variância não é, em geral, aditiva, como o valor esperado. Com a hipótese complementar de independência, a P3 fica válida.

P4. Seja X_1, \dots, X_n variáveis aleatórias independentes. Então,

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n).$$

4.9 Exercícios

Exercício 4.1. De um lote que contém 25 peças, das quais 5 são defeituosas, 4 são escolhidas ao acaso. Seja X o número de defeituosas encontradas. Estabeleça a distribuição de probabilidade de X , quando:

- (a) As peças forem escolhidas com reposição.
- (b) As peças forem escolhidas sem reposição.

Exercício 4.2. O diâmetro de um cabo elétrico supõe-se ser uma variável aleatória contínua X , como fdp $f(x) = 6x(1 - x)$, $0 \leq x \leq 1$.

- (a) Verifique se essa expressão é realmente uma função densidade de probabilidade.
- (b) Obtenha a função de distribuição acumulada de X .
- (c) Determine um número b tal que $P(X < b) = 2P(X > b)$.

(d) Calcule $P(X \leq 1/2 \mid 1/3 < X < 2/3)$.

Exercício 4.3. A porcentagem de álcool ($100X$) em certo composto pode ser considerada uma variável aleatória X , $0 < X < 1$, tem a seguinte função densidade de probabilidade:

$$f(x) = 20x^3(1 - x), \quad 0 < x < 1.$$

(a) Estabeleça a expressão da função de distribuição de X .

(b) Calcule $P(X \leq 2/3)$.

Exercício 4.4. Os valores na Tabela 4.3 representam a distribuição de probabilidade de D , a procura diária de um certo produto. Calcule $E(D)$ e $\text{Var}(D)$.

d	1	2	3	4	5
$P(D = d)$	0,1	0,1	0,3	0,3	0,2.

Tabela 4.3: Distribuição de probabilidade - procura diária de um certo produto.

Capítulo 5

Modelos Probabilísticos Discretos

Agora iremos apresentar alguns dos principais modelos probabilísticos utilizados para descrever vários fenômenos ou situações que encontramos na natureza ou ainda experimentos por nós construídos.

Na prática, nossos experimentos consistem em medir etapas de um processo. Como resultados destas medições obtemos valores numéricos ou atributos, que caracterizam a performance do processo. Os resultados das medições, como já vimos, são denominados variáveis aleatórias.

Uma variável aleatória fica completamente caracterizada pela sua função de distribuição acumulada. No caso discreto, podemos também usar a função de probabilidade com o mesmo objetivo. Por essa razão, nos modelos para variáveis discretas que iremos apresentar, estaremos sempre indicando, na definição sua função de probabilidade.

5.1 Distribuição Binomial

Quando queremos classificar peças de um lote com 20 peças em defeituosas ou não, podemos contar o número de peças defeituosas e associar uma variável aleatória X que represente este número.

Esta variável pode assumir os seguintes valores: $0, 1, 2, \dots, 20$. Associamos estes valores a uma variável aleatória X , que é discreta, pois assume um número finito de valores. Definimos a função de probabilidade da variável aleatória X , como a probabilidade da variável X assumir o valor x . A função de probabilidade será denotada por $P[X = x]$.

Como o leitor deve ter notado, no exemplo dado cada elemento da população é classificado como “defeituosa” ou “não-defeituosa”. Generalizando, este modelo se aplica a situações em que os elementos observados são classificados como possuindo ou não determinada característica.

Primeiramente, para construir o modelo binomial vamos introduzir uma sequência de ensaios de Bernoulli.

Uma sequência de Bernoulli é definida por meio das seguintes condições:

- i. Em cada ensaio considera-se somente a ocorrência ou não-ocorrência de um certo evento que

será denominado sucesso (S) e cuja não-ocorrência será denominada fracasso (F).

ii. Os ensaios são independentes.

iii. A probabilidade de sucesso, que denotaremos por p , é a mesma para cada ensaio. A probabilidade de fracasso será denotada por $1 - p$.

Para um experimento que consiste na realização de n ensaios de Bernoulli, o espaço amostral pode ser considerado como o conjunto de n -uplas, em que cada posição há um sucesso (S) ou uma falha (F).

Pelas condições ii e iii vemos que a probabilidade de um ponto amostral com sucessos nos k primeiros ensaios e falhas nos $n - k$ ensaios seguintes é $p^k(1 - p)^{n-k}$. De fato, temos que

$$\begin{aligned} P(\text{ocorrência de sucesso nos } k \text{ primeiros ensaios e falhas nos } n - k \text{ seguintes}) &= \\ &= P(\underbrace{S; S; \dots; S}_k; \underbrace{F; \dots; F}_{n-k}) = P(S \cap S \cap \dots \cap S \cap F \dots \cap F) = \\ &= P(S) \times \dots \times P(S) \times P(F) \times \dots \times P(F) = \underbrace{p \times p \times \dots \times p}_k \underbrace{(1 - p) \times \dots (1 - p)}_{n-k} = p^k(1 - p)^{n-k}. \end{aligned}$$

Note que esta é a probabilidade de qualquer ponto com k sucessos e $n - k$ falhas. O número de pontos do espaço amostral que satisfaz essa condição é igual ao número de maneiras com que podemos escolher k ensaios para a ocorrência de sucesso dentre o total de n ensaios, pois nos $n - k$ restantes deverão ocorrer falhas. Este número é igual ao número de combinações de n elementos tomados k a k , ou seja $\binom{n}{k} = \frac{n!}{k!(n - k)!}$.

Decorre do que foi exposto que, para $k = 0, 1, \dots, n$:

$$P[X = k] = \binom{n}{k} p^k(1 - p)^{n-k}. \quad (5.1)$$

Definição 5.1.1. Distribuição Binomial

Seja X o número de sucessos obtidos na realização de n ensaios de Bernoulli independentes. Diremos que X tem distribuição binomial com parâmetros n e p , onde p é a probabilidade de sucesso em cada ensaio, se sua função de probabilidade for dada por 5.1. Usamos a notação $X \sim \text{Binomial}(n, p)$.

O número de sucessos X em n ensaios de Bernoulli pode ser representado por meio de variáveis aleatórias associadas a cada ensaio, que assumem valores 0 ou 1.

Seja $X_i = 1$ se ocorre sucesso no i -ésimo ensaio e $X_i = 0$ se ocorre falha, para $i = 1, 2, \dots, n$. Então X pode ser expresso da seguinte maneira:

$$X = X_1 + X_2 + \dots + X_n.$$

Como motivação, suponha que estamos interessados em retirar o número 4 ao lançar um dado. Se ocorrer o número 4 diremos que ocorreu “sucesso”, caso contrário, diremos que ocorreu “fracasso”. Assim, temos que

$$P(\text{sucesso}) = \frac{1}{6} \quad \text{e} \quad P(\text{fracasso}) = \frac{5}{6}.$$

Suponha agora que lançamos o dado 5 vezes. É claro que o resultado de um lançamento independe do anterior, do posterior ou de qualquer outro lançamento.

Digamos que estamos interessados em calcular a probabilidade de obter o número 4, duas vezes. Isso pode ocorrer de várias maneiras. Uma maneira é (a não ocorrência de 4 será denotada por 0):

$$4 \ 4 \ 0 \ 0 \ 0 \quad \text{com probabilidade} \quad \frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^3.$$

Uma outra maneira é

$$4 \ 0 \ 4 \ 0 \ 0 \quad \text{com probabilidade} \quad \frac{1}{6} \times \frac{5}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6} = \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^3.$$

Assim, qualquer seqüência contendo duas vezes o número 4 e três outros valores quaisquer tem a mesma probabilidade. Como qualquer uma dessas seqüências serve ao nosso interesse, a probabilidade procurada é a soma das probabilidades de todas as seqüências. Precisamos saber então quantas seqüências existem. A resposta é dada por:

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = 10$$

onde $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ (fatorial de 5).

Assim, temos que

$$P(\text{ocorrer o número 4 duas vezes}) = 10 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3.$$

Agora vamos generalizar esse resultado. Suponha um experimento com apenas dois resultados possíveis: “sucesso” e “fracasso”, tal que $P(\text{sucesso}) = p$ e $P(\text{fracasso}) = 1 - p = q$. Vamos repetir esse experimento n vezes. Estamos interessados em obter k sucessos e, conseqüentemente, $n - k$ fracassos. O número de sucessos a serem obtidos é uma variável aleatória X . Com isso,

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

onde $k = 0, 1, 2, \dots, n$ e

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Exemplo 5.1. Suponha que numa linha de produção a probabilidade de se obter uma peça defeituosa (sucesso) é $p = 0,1$. Toma-se uma amostra de 10 peças para serem inspecionadas. Qual a probabilidade de se obter:

- a) Uma peça defeituosa?
- b) Nenhuma peça defeituosa?
- c) Duas peças defeituosas?
- d) No mínimo duas peças defeituosas?
- e) No máximo duas peças defeituosas?

$$a) P(X = 1) = \binom{10}{1} (0,1)^1 (1-0,1)^{10-1} = \frac{10!}{1!(10-1)!} 0,1(0,9)^9 = 0,3874$$

$$b) P(X = 0) = \binom{10}{0} (0,1)^0 (1-0,1)^{10-0} = \frac{10!}{0!(10-0)!} (0,9)^{10} = 0,3486$$

$$c) P(X = 2) = \binom{10}{2} (0,1)^2 (1-0,1)^{10-2} = \frac{10!}{2!(10-2)!} (0,1)^2 (0,9)^8 = 0,1937$$

$$d) P(X \geq 2) = P(X = 2) + P(X = 3) + \dots + P(X = 9) + P(X = 10)$$

$$\text{ou seja, } P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 0,2639$$

$$e) P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0,9298.$$

□

5.1.1 Valor Esperado e Variância

Seja X um variável aleatória com distribuição Binomial (n,p) . O valor esperado representa o número médio de sucessos. Por definição, temos que

$$E(X) = \sum_{k=0}^n kP(X = k) = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = np.$$

$$E(X) = np$$

Por exemplo, para uma amostra de tamanho 10 e $p = 0.1$, obtemos que

$$E(X) = np = 10 \cdot 0,1 = 1$$

A variância $Var(X)$ é dada por

$$Var(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2 = np(1-p).$$

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

Para a mesma amostra de tamanho 10 e $p = 0.1$, temos

$$\text{Var}(X) = np(1 - p) = 10 \cdot 0,1 \cdot 0,9 = 0,9$$

e o desvio padrão é

$$\sigma_x = \sqrt{\sigma_x^2} = 0,9487.$$

5.1.2 Exercícios

Exercício 5.1. A taxa de imunização de uma vacina é de 80 %. Se um grupo de 20 pessoas foram vacinadas, calcule a probabilidade de:

- a) 12 terem sido imunizadas;
- b) pelo menos 5 terem sido imunizadas;
- c) não mais de cinco terem sido imunizadas.

Exercício 5.2. Em determinado trecho de rodovia, sabe-se que 30 % dos carros excedem o limite de velocidade de 100 km/h. Quinze carros passaram por uma câmera da polícia que estava escondida. Calcule a probabilidade de 7 carros ou mais estarem acima do limite de velocidade?

Exercício 5.3. Um experimento consiste em lançar um dado honesto e observar a ocorrência ou não da face superior igual a 6. Calcule a probabilidade de sair 3 faces superiores iguais a 6, em 8 lançamentos do dado.

Exercício 5.4. Uma prova consiste de 15 questões com respostas de verdadeiro ou falso. Calcule a probabilidade da prova ter 6 respostas verdadeiras. Calcule o número médio de respostas verdadeiras e o respectivo desvio padrão.

Exercício 5.5. Em uma linha de produção, cada componente eletrônico produzido tem a probabilidade de 0,04 de ser defeituoso. Calcule a probabilidade de que em 1000 componentes produzidos:

- a) 10 sejam defeituosos;
- b) no máximo 5 sejam defeituosos.

Exercício 5.6. Em uma família de 6 crianças, qual a probabilidade de que 5 delas sejam do sexo feminino?

5.2 Distribuição de Poisson

Na distribuição binomial quando o tamanho da amostra n é grande ($n \rightarrow \infty$) e p é pequeno ($p \rightarrow 0$), o cálculo da probabilidade

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

pode ser feito usando a seguinte expressão

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad (5.2)$$

onde $k = 0, 1, 2, 3, \dots$, $e = 2,718$ e $\lambda = np$.

Essa expressão é devido a Poisson e é muito usada para calcular probabilidades de ocorrências de defeitos “raros” em sistemas e componentes. O número de defeitos é a variável representada por X . Assim temos a seguinte definição:

Definição 5.2.1. Distribuição de Poisson

Uma variável X segue o modelo de Poisson de parâmetro λ , $\lambda > 0$, se sua função de probabilidade for dada por 5.2. Usamos a notação $X \sim \text{Poisson}(\lambda)$. O parâmetro λ indica a taxa de ocorrência por unidade de medida.

5.2.1 Valor Esperado e Variância

A média de X , que freqüentemente é chamada de *taxa de defeitos*, é dada por:

$$E(X) = \sum_{k=0}^{\infty} k P(X = k) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} = \lambda$$

$$E(X) = \lambda$$

A variância de X é dada por:

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \lambda$$

$$\text{Var}(X) = \lambda$$

Exemplo 5.2. Considere um processo que têm uma taxa de 0,2 defeitos por unidade. Qual a probabilidade de uma unidade qualquer apresentar:

- a. dois defeitos?
- b. um defeito?
- c. zero defeito?

Temos que $\lambda = 0,2$, então

$$\text{a. } P(X = 2) = \frac{e^{-0,2}(0,2)^2}{2!} = 0,0164;$$

$$\text{b. } P(X = 1) = \frac{e^{-0,2}(0,2)^1}{1!} = 0,1637;$$

$$\text{c. } P(X = 0) = \frac{e^{-0,2}(0,2)^0}{0!} = 0,8187.$$

Esse último valor, $P(X = 0)$, é chamado de “rendimento” do processo (ou produto).

□

5.2.2 Exercícios

Exercício 5.7. Suponha que é observado o número de chegadas a um caixa eletrônico de um banco durante um dia. Suponha ainda, que o número médio de pessoas que chegam no período de 15 minutos é 10. Calcule a probabilidade de 5 chegadas em 15 minutos.

Exercício 5.8. Certo estudo psiquiátrico demonstrou que a distribuição mensal de suicídios entre adolescentes de determinada comunidade, entre 1977 e 1987, obedecia uma Poisson com média igual a 2,75. Qual a probabilidade estimada de que num certo mês sorteado do ano ocorram 3 suicídios de adolescentes?

Exercício 5.9. Uma central telefônica tipo PABX recebe uma média de 5 chamadas por minuto. Qual a probabilidade deste PABX não receber nenhuma chamada durante um intervalo de 1 minuto? Qual a probabilidade de receber menos de 4 chamadas, em 1 minuto?

Exercício 5.10. Sabe-se que a média de consumo de determinada peça em um estoque é de 6 peças por mês. Calcule a probabilidade de em determinado mês serem consumidas 4 peças. Calcule a probabilidade de serem consumidas mais de 4 peças, em determinado mês.

Exercício 5.11. Em um determinado cruzamento de trânsito, na cidade de Jataí, ocorrem em média 7 acidentes por quinzena. Calcule:

- a) a probabilidade de que ocorram 9 acidentes em determinada quinzena;
- b) a probabilidade de que não ocorram mais de 5 acidentes em determinada quinzena;
- c) a probabilidade de que ocorram mais de 5 acidentes em uma determinada quinzena.

Exercício 5.12. Em média 17 pessoas compram um computador por mês em determinada loja. Para realizar seu planejamento, o gerente de vendas tem interesse de saber:

- a) qual a probabilidade de que ocorram 22 vendas no próximo mês?

- b) qual a probabilidade de que ocorram pelo menos 12 vendas?
- c) qual a probabilidade de vender menos de 10 computadores?

Exercício 5.13. Após a fabricação e montagem de um carro, o mesmo é passado por testes de qualidade. Em média são encontrados, entre defeitos simples e complexos, 7 defeitos. O gerente do setor de qualidade deseja saber:

- a) qual a probabilidade de encontrar 11 defeitos em um carro?
- b) qual a probabilidade de encontrar pelo menos 15 defeitos?
- c) qual a probabilidade de encontrar menos de 14 defeitos?

5.3 Distribuição Geométrica

Consideremos uma seqüência ilimitada de Bernoulli, com probabilidade de sucesso p em cada ensaio. Designemos sucesso por S e falha por F . Realizamos os ensaios até que ocorra o primeiro sucesso.

O espaço amostral para este experimento é o conjunto :

$$(S, FS, FFS, \dots, FF\dots S, \dots, FFFF\dots S, \dots).$$

Ou seja, um elemento típico desse espaço amostral é uma seqüência de comprimento n em que nas primeiras $n - 1$ posições temos F e na n -ésima temos S .

Definição 5.3.1. Distribuição Geométrica

Seja X a variável aleatória que dá o número de falhas até o primeiro sucesso. A variável X tem distribuição Geométrica com parâmetro p , $0 < p < 1$, se sua função de probabilidade é dada por:

$$P[X = j] = (1 - p)^j p, \quad j = 0, 1, \dots \quad (5.3)$$

Usaremos a notação $X \sim \text{Geo}(p)$.

O evento $[X = j]$ ocorre se, e somente se, ocorrem somente falhas nos j primeiros ensaios e sucesso no ensaio $(j + 1)$. A expressão 5.3 segue da independência dos ensaios.

A distribuição geométrica tem uma propriedade que serve para caracterizá-la no conjunto das distribuições discretas, que é expressa pela seguinte proposição:

Proposição 5.3.1. Se X é variável aleatória discreta com distribuição geométrica, então, para todo $j, k = 1, 2, \dots$ tem-se:

$$P[X \geq j + k | X \geq j] = P[X \geq k].$$

Este resultado reflete a falta de memória ou de desgaste da distribuição geométrica.

5.3.1 Valor Esperado e Variância

Vamos calcular $E(X)$ a partir da definição. No Cálculo de $E(X)$, utilizaremos uma expressão que vale a pena ser destacada, pois é de interesse geral.

Para todo número real x no intervalo $(0,1)$ consideremos a série geométrica cuja soma é dada a seguir:

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x} \quad (5.4)$$

Derivando-se ambos os membros da igualdade, temos:

$$\frac{d}{dx} \sum_{i=1}^{\infty} x^i = \sum_{i=1}^{\infty} i x^{(i-1)} = \frac{1}{(1-x)^2}. \quad (5.5)$$

Usando-se a definição de esperança, a equação 5.5 e tomando $x = 1 - p$ temos:

$$E(X) = \sum_{j=0}^{\infty} j(1-p)^j p = p \sum_{j=0}^{\infty} j(1-p)^j = p(1-p) \sum_{j=0}^{\infty} j(1-p)^{j-1} = \frac{p(1-p)}{p^2}.$$

Simplificando vem:

$$E(X) = \frac{1-p}{p}$$

Usando a expressão podemos calcular $E(X^2)$ e obter a variância de X . Sugerimos ao leitor que faça esse cálculo que fornecerá:

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

Exemplo 5.3. A duração (em centenas de horas) de um determinado componente eletrônico, foi modelada por uma distribuição geométrica com parâmetro $p=0,8$. Determine a probabilidade desse componente eletrônico:

a. durar menos de 400 horas;

b. durar mais de 500 horas.

Sabemos que $P[X = k] = (1-p)^k p$.

a. Assim,

$$\begin{aligned} P[X \leq 400 \text{ horas}] &= P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3] \\ &= (1-0,8)^0(0,8) + (1-0,8)^1(0,8) + (1-0,8)^2(0,8) + (1-0,8)^3(0,8) \\ &= 0,8 + 0,16 + 0,032 + 0,0064 = 0,9984 \end{aligned}$$

b. Já neste caso,

$$\begin{aligned} P[X \geq 500 \text{ horas}] &= 1 - P[X = 5] = 1 - (1-0,8)^5(0,8) \\ &= 1 - 0,999936 = 0,000064. \end{aligned}$$

□

Duração em horas(centenas)	Probabilidade	Acumulada
0	0,8000	0,8000
1	0,1600	0,9600
2	0,0320	0,9920
3	0,0064	0,9984
4	0,0013	0,9997
5	0,0003	0,9999

Tabela 5.1: Tabela de probabilidade da distribuição geométrica.

5.3.2 Exercícios

Exercício 5.14. *A probabilidade de encontrar aberto o sinal de trânsito é de 0,20. Calcule a probabilidade de que seja necessário passar 5 vezes ao local para encontrar pela 1ª vez o sinal aberto?*

Resp.: 0,08192

Exercício 5.15. *Um dado é lançado repetidamente. Calcule a probabilidade de sair o número 2 na sexto lançamento.*

Exercício 5.16. *Suponha que a probabilidade de um componente de computador ser defeituoso é de 0,2. Numa mesa de testes, uma batelada é posta à prova, um a um. Determine a probabilidade do primeiro componente defeituoso ser encontrado no sétimo componente testado.*

Exercício 5.17. *A probabilidade de que um bit, transmitido através de um canal de transmissão digital, seja recebido com erro é 0.1. Suponha que as transmissões são eventos independentes. Calcule a probabilidade de que um bit seja transmitido com erro somente na 6 transmissão.*

Resp.: 0,06

Exercício 5.18. *Joga-se uma moeda para cima e observa-se o resultado. Calcule a probabilidade de obter-se uma cara somente na quarta jogada.*

Exercício 5.19. *A probabilidade de uma caixa de litro de leite conter uma partícula grande de contaminação é de 0,01. Se for considerado que as caixas de leite sejam independentes, qual será a probabilidade de que exatamente 125 caixas necessitem ser analisadas antes que uma partícula grande de contaminação seja detectada.*

Resp.: 0,0029

5.4 Distribuição Hipergeométrica

A distribuição que iremos ver agora representa um modelo para amostragem sem reposição de uma população com um número finito de elementos, em que cada um pode ser de um de dois tipos. Se a

população tem N elementos, terá então M elementos de um tipo e $N - M$ de outro. Assim, podemos mostrar que a distribuição de probabilidade da variável aleatória X é dada por:

$$P[X = k] = \frac{\binom{M}{k} \binom{N - M}{n - k}}{\binom{N}{n}}, \quad (5.6)$$

sendo k inteiro e $\max\{0, n - (N - M)\} \leq k \leq \min\{M, n\}$.

Definição 5.4.1. Distribuição Hipergeométrica

Diremos que uma variável aleatória X tem distribuição Hipergeométrica de parâmetros M , N e n se sua função de probabilidade for dada pela expressão 5.6. Denotamos, $X \sim Hgeo(M, N, n)$

Por exemplo, consideremos uma urna contendo M bolas brancas e $N - M$ bolas vermelhas. Retira-se da urna n bolas sem reposição, isto é, após cada retirada a bola selecionada não é reposta na urna. Vamos designar X como sendo o número de bolas brancas entre as n bolas retiradas da urna. Para justificar os limites, notemos que o número de bolas brancas na amostra k é menor ou igual ao número de bolas brancas (M) e também menor ou igual ao número de bolas na amostra n , portanto menor ou igual ao menor deles. Se o tamanho da amostra n é menor ou igual ao número de bolas vermelhas $N - M$, então na amostra todas podem ser vermelhas e portanto $k = 0$. Se $n \geq (N - M)$, então mesmo que todas as $N - M$ vermelhas pertençam à amostra, haverá $n - (N - M)$ brancas na amostra. Por isso, $\max\{0, n - (N - M)\} \leq k \leq \min\{M, n\}$.

O espaço amostral para esse experimento é formado pelo conjunto das amostras não ordenadas de n bolas retiradas das N , ou o que é o mesmo, pelo conjunto das combinações de N elementos tomados n a n , cuja representação é igual a:

$$\binom{N}{n}.$$

Existem $\binom{M}{k}$ combinações de k bolas brancas retiradas das M existentes e $\binom{N - M}{n - k}$ combinações de $n - k$ vermelhas retiradas das $N - M$ vermelhas. Assim o número de combinações com k brancas e $n - k$ vermelhas é o produto:

$$\binom{M}{k} \binom{N - M}{n - k}$$

Mostramos assim a Distribuição de Probabilidade da Hipergeométrica. Dessa forma, considere uma urna contendo $N = 15$ bolas, das quais, $M = 10$ são bolas brancas e $N - M = 5$ são bolas

vermelhas. Se retirarmos uma amostra de tamanho $n = 7$ bolas, sem reposição, a probabilidade de encontrarmos 3 bolas brancas é dada por:

$$P[X = 3] = \frac{\binom{10}{3} \binom{5}{4}}{\binom{15}{7}} = \frac{120 * 5}{3003} = 0,20$$

5.4.1 Valor Esperado e Variância

Se X segue uma distribuição Hipergeométrica com parâmetros N , M e n , então a Esperança é dada por:

$$E(X) = n \frac{M}{N}$$

e a variância é dada por:

$$Var(X) = n \frac{M}{N} \frac{(N - M)}{N} \left(1 - \frac{n - 1}{N - 1} \right)$$

Exemplo 5.4. *Uma empresa fabrica um tipo de tomada que são embalados em lote de 25 unidades. Para aceitar o lote enviado por essa fábrica, o controle de qualidade da empresa tomou o seguinte procedimento: sorteia-se um lote e desse lote seleciona-se 8 tomadas para teste, sem reposição. Se for constatado no máximo duas tomadas defeituosas, aceita-se o lote fornecido pela fábrica. Se a caixa sorteada tiver 7 peças defeituosas, qual a probabilidade de se rejeitar o lote? $N=25$, $M = 7$ (n° de defeituosas) e $n=8$ (tamanho da amostra).*

$$P[\text{aceitar o lote}] = P[D \leq 2] = P[D = 0] + P[D = 1] + P[D = 2]$$

$$= \frac{\binom{7}{0} \binom{25-7}{8-0}}{\binom{25}{8}} + \frac{\binom{7}{1} \binom{25-7}{8-1}}{\binom{25}{8}} + \frac{\binom{7}{2} \binom{25-7}{8-2}}{\binom{25}{8}} = 0,0010069.$$

□

5.4.2 Exercícios

Exercício 5.20. *Uma batelada de peças contém 100 peças de tubo vindos de um fornecedor local e 200 peças de um fornecedor de tubos de um estado vizinho. Se quatro peças forem selecionadas, ao acaso e sem reposição, calcule:*

- a) a probabilidade de que elas sejam todas provenientes do fornecedor local; Resp.: 0,0119
- b) a probabilidade de que duas ou mais peças na amostra sejam proveniente do fornecedor local; Resp.: 0,408
- c) a probabilidade de que no mínimo, um peça na amostra seja proveniente do fornecedor local. Resp.: 0,804

Exercício 5.21. Uma firma compra lâmpadas por centenas. Examina sempre uma amostra de 15 lâmpadas para verificar se estão boas. Se uma centena inclui 12 lâmpadas queimadas, qual a probabilidade de se escolher uma amostra com pelo menos uma lâmpada queimada.

Resp.: 0,8747

Exercício 5.22. De um baralho com 52 cartas, retiram-se 8 cartas ao acaso, sem reposição. Qual a probabilidade de que 4 sejam figuras.

Resp.: 0,0601

Exercício 5.23. Pequenos motores são guardados em caixas de 50 unidades. Um inspetor de qualidade examina cada caixa, antes da posterior remessa, testando 5 motores. Se nenhum motor for defeituoso, a caixa é aceita. Se pelo menos um for defeituoso, todos os 50 motores são testados. Há 6 motores defeituosos numa caixa. Qual a probabilidade de que seja necessário examinar todos os motores dessa caixa?

Resp.: 0,4874

Capítulo 6

Modelos Probabilísticos Contínuos

Agora apresentaremos os modelos probabilísticos descritos por variáveis aleatórias que possuem uma densidade de probabilidade, ou seja, variáveis aleatórias contínuas. Cada modelo corresponde a uma família de distribuição de probabilidade, expressa por uma densidade de probabilidade que depende de um ou mais parâmetros.

6.1 Distribuição Uniforme

Definição 6.1.1. *Distribuição Uniforme*

Uma variável aleatória X tem distribuição Uniforme no intervalo $[a, b]$ se sua função densidade de probabilidade for dada por:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b; \\ 0, & \text{caso contrário;} \end{cases}$$

Vamos descrever um experimento cujo resultado nos dá uma variável com distribuição Uniforme no intervalo $(0, 2\pi)$. Consideremos um segmento de comprimento 2π . Vamos unir as duas pontas desse segmento e formar um círculo de raio unitário. O comprimento desse círculo é precisamente 2π . Vamos fixar um ponteiro no centro desse círculo e vamos então girá-lo, observando até que ele venha parar. Por razões de simetria, vemos que a chance do ponteiro parar de girar em qualquer arco do círculo é a mesma para qualquer arco de um comprimento dado. Seja X o comprimento do arco determinado pela origem e pelo ponto onde o ponteiro parar. Assim, temos que X é uma variável aleatória com distribuição Uniforme no intervalo $(0, 2\pi)$.

Se quisermos obter a distribuição Uniforme no intervalo $[a, b]$ basta fazermos $b - a = 2\pi r$, construir um círculo de raio $r = \frac{b-a}{2\pi}$ e proceder da maneira descrita.

Exemplo 6.1. A ocorrência de panes em qualquer ponto de uma rede telefônica de 7 km foi modelada por uma distribuição Uniforme no intervalo $[0, 7]$. Qual é a probabilidade de que uma pane venha a ocorrer nos primeiros 800 metros? E qual a probabilidade de que ocorra nos 3 km centrais da rede?

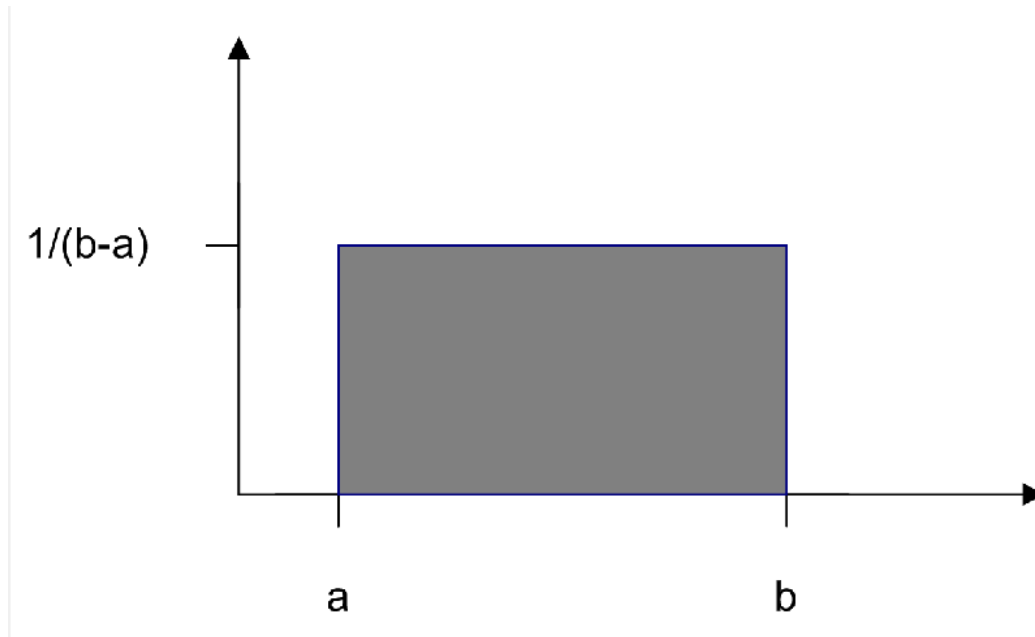


Figura 6.1: Gráfico da função densidade de probabilidade da distribuição Uniforme.

A função densidade da distribuição Uniforme é dada por $f(x) = \frac{1}{7}$, $0 \leq x \leq 7$. Assim, a probabilidade de ocorrer pane nos primeiros 800 metros é

$$P[X \leq 0,8] = \int_0^{0,8} f(x)dx = \frac{0,8 - 0}{7} = 0,1142.$$

E a probabilidade de ocorrer pane nos 3 Km centrais da rede é

$$P[2 \leq x \leq 5] = \int_2^5 f(x)dx = P[X \leq 5] - P[X \leq 2] = \frac{5}{7} - \frac{2}{7} = \frac{5-2}{7} = 0,4285.$$

□

6.1.1 Valor Esperado e Variância

Obtemos a esperança pela definição dada pela equação 4.2.

$$E(X) = \int x \frac{1}{b-a} dx = \frac{a+b}{2} \quad (6.1)$$

$$E(X) = \frac{a+b}{2}$$

Obtemos da mesma forma que

$$E(X^2) = \frac{1}{b-a} \int_a^b x^2 dx = \frac{a^2 + ab + b^2}{3}. \quad (6.2)$$

Com os valores dados por 6.1 e 6.2, obtemos a variância de X

$$Var(X) = E(X^2) - (E(X))^2 = \frac{(b-a)^2}{12}. \quad (6.3)$$

$$Var(X) = \frac{(b-a)^2}{12}$$

6.1.2 Exercícios

Exercício 6.1. *Várias linguagens de programação de computadores têm funções que geram números pseudo-aleatórios cuja distribuição é basicamente uniforme. Se uma função desse tipo gera números entre 0 e 2, qual a probabilidade de um número gerado estar entre 1 e 1,5? Qual a média esperada e o respectivo desvio padrão?*

Exercício 6.2. *Calcular a média e o desvio padrão de uma variável aleatória contínua X com distribuição uniforme no intervalo de 100 a 200.*

Exercício 6.3. *Com o bjetivo de verificar a resistência à pressão de água, os técnicos de qualidade de uma empresa inspecionam os tubos de PVC produzidos. Os tubos inspecionados têm 6 metros de comprimento e são submetidos a grandes pressões até o aparecimento do primeiro vazamento, cuja distância a uma das extremidades é anotada para fins de análise posterior. Escolhe-se um tubo ao acaso para ser inspecionado. Considerando a distribuição uniforme, calcule a probabilidade de que o vazamento esteja, no máximo, a um metro das extremidades.*

Resp.: 0,33

6.2 Distribuição Normal

Definição 6.2.1. Distribuição Normal

Uma variável aleatória contínua X tem distribuição Normal se sua função densidade de probabilidade for dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right], \quad x \in (-\infty, +\infty)$$

Usaremos a notação $X \sim N(\mu, \sigma^2)$.

A variação natural de muitos processos industriais é realmente aleatória. Embora as distribuições de muitos processos possam assumir uma variedade de formas, muitas variáveis observadas possuem uma distribuição de frequências que é, aproximadamente, uma distribuição de probabilidade Normal.

O gráfico da função densidade de probabilidade da distribuição Normal tem forma de “sino”, como mostra a Figura 6.2.

Para achar a área sob a curva normal devemos conhecer dois valores numéricos (também chamados de parâmetros), a média μ e o desvio padrão σ . O gráfico 6.3 mostra algumas áreas importantes:

Uma característica importante de uma variável aleatória X com distribuição Normal, com média e desvio padrão quaisquer, é que podemos reduzi-la a uma variável aleatória Z com distribuição Normal com média zero e variância 1. Ou seja, a partir de

$$X \sim N(\mu, \sigma^2)$$

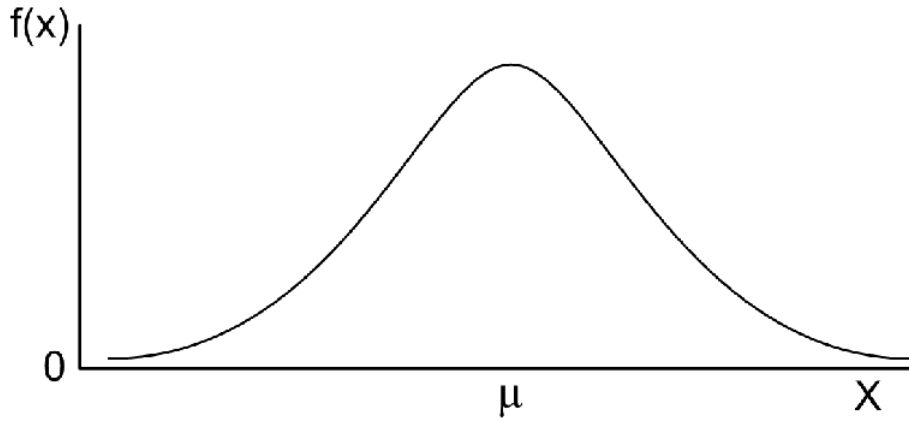


Figura 6.2: Gráfico de uma função densidade de uma variável com distribuição Normal.

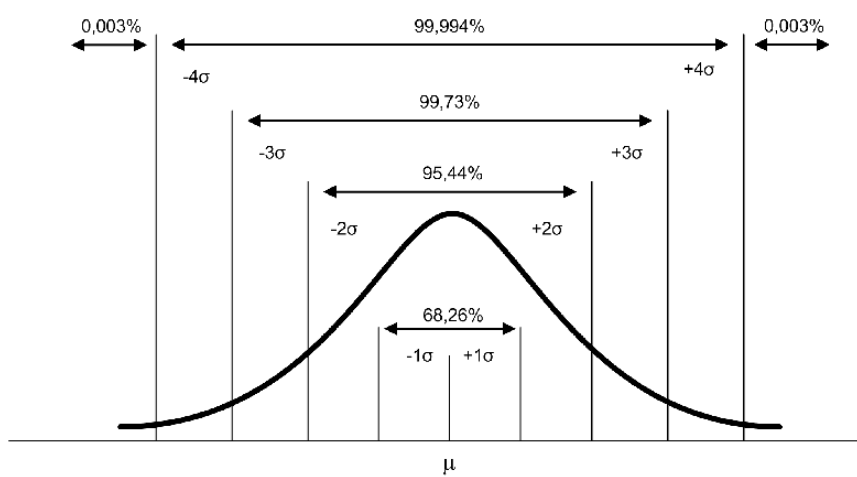


Figura 6.3: Áreas sob a Curva Normal

obtemos

$$\frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Assim, dizemos que a variável aleatória Z tem distribuição Normal padrão. Para cada valor de μ e/ou σ , temos uma distribuição. Mas para se calcular áreas específicas, se faz uso da distribuição Normal padronizada. Como vimos, esta distribuição tem média $\mu = 0$ e desvio padrão $\sigma = 1$, e está tabelada. Como a distribuição é simétrica em relação à média, a área à direita é igual a área à esquerda de μ . Assim, as tabelas fornecem áreas acima de valores não-negativos que vão desde 0.00 até 4.09, dependendo da tabela. A variável que tem distribuição Normal padronizada é denotada por Z . Veja o gráfico da curva normal padronizada na Figura 6.4.

Exemplo 6.2. Considere X uma variável aleatória Normal com média 11,15 e desvio-padrão 2,238.

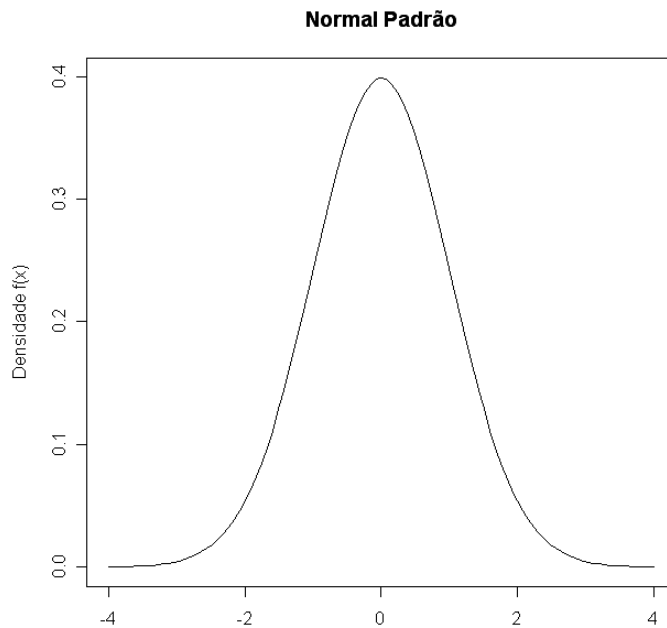


Figura 6.4: Distribuição Normal Padronizada

Calcule a probabilidade de X ser menor que 8,7.

Temos que $X \sim N(11,15; (2,238)^2)$. Daí, temos que $Z = \frac{X - 11,15}{2,238} \sim N(0,1)$.

Assim,

$$P[X < 8,7] = P\left[\frac{X - 11,15}{2,238} < \frac{8,7 - 11,15}{2,238}\right] = P[Z < -1,0947] = 0,1368 = 13,7\% \quad (6.4)$$

□

Exemplo 6.3. A área sob a curva para Z maior do que 1,00 é 0,1587. Ou seja, a probabilidade de Z ser maior do que 1 é 15,87%. Veja o gráfico na Figura 6.5

□

Exemplo 6.4. A área sob a curva para Z maior do que 1,19 é 0,1170, ou seja, a probabilidade de Z ser maior do que 1,19 é 11,70%. Veja o gráfico na Figura 6.6

□

Exemplo 6.5. A área sob a curva para Z menor do que 2,00 não é fornecida diretamente pela tabela. Então devemos encontrar a área para Z maior do que 2,00. Em seguida fazemos 1 menos a área encontrada e temos a área desejada.

A área sob a curva para Z maior do que 2,00 é 0,0228. A área desejada é $1 - 0,0228 = 0,9772$. Ou seja, a probabilidade de Z ser menor do que 2,00 é 97,72%. Veja o gráfico na Figura 6.7

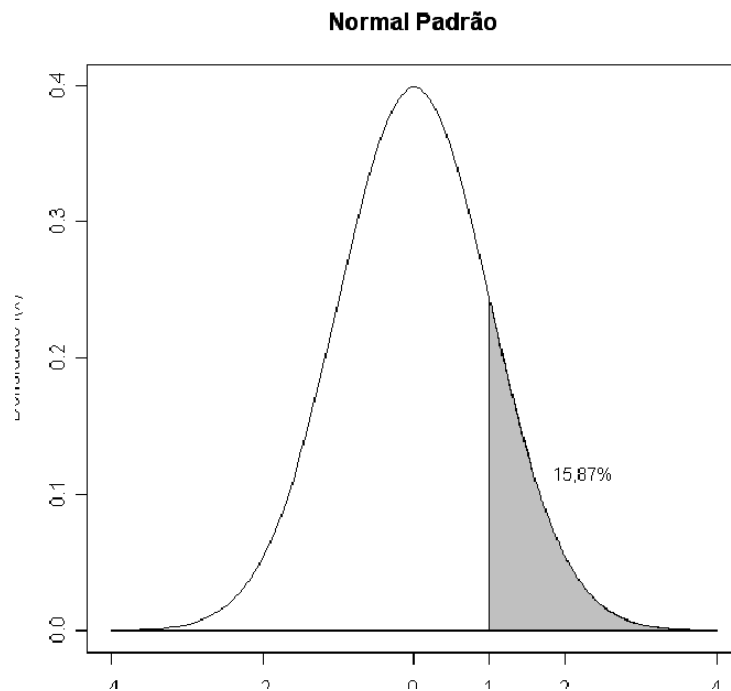


Figura 6.5: Área sob a curva normal

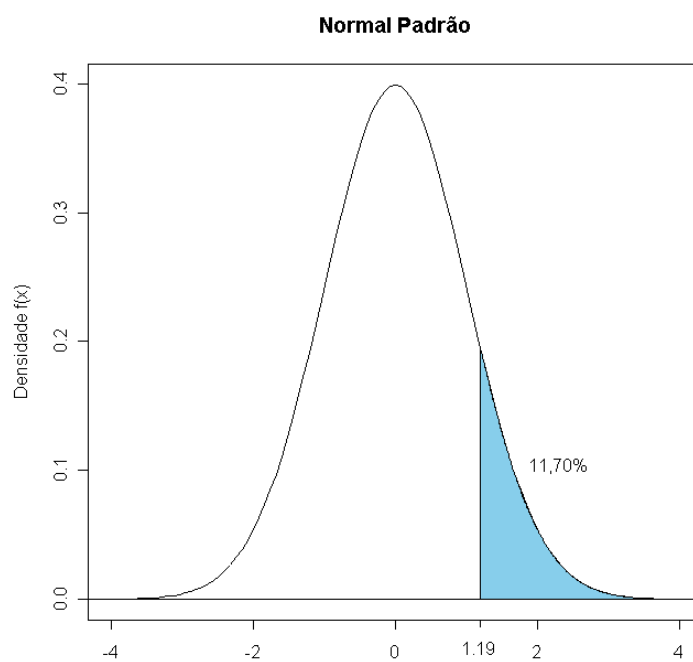


Figura 6.6: Área sob a curva normal

□

Exemplo 6.6. Suponha que a espessura da arruelas tenha distribuição normal com média 11,15 mm e desvio padrão 2,238 mm. Qual a porcentagem de arruelas que tem espessura entre 8,70 e 14,70?

Temos que encontrar dois pontos da distribuição normal padronizada. O primeiro ponto é:

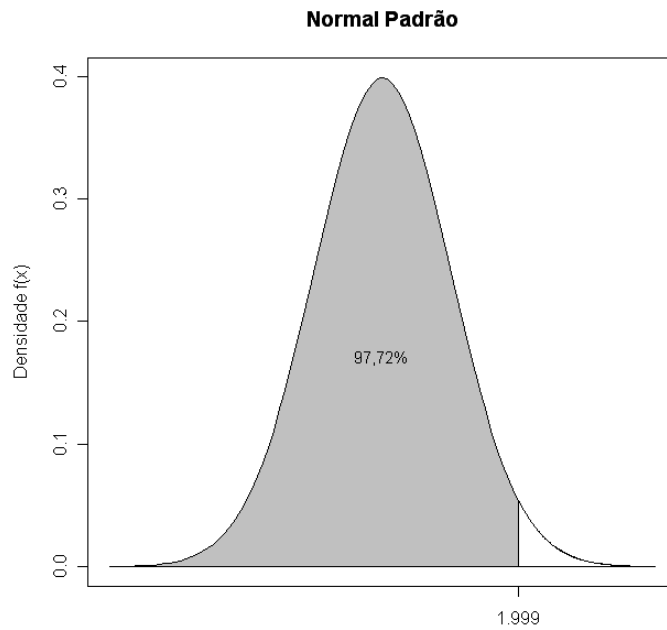


Figura 6.7: Área sob a curva normal

$$Z_1 = \frac{8,70 - 11,15}{2,238} = -1,09$$

A área para valores maiores do que -1,09 é 0,8621 ou 86,21%.

O segundo ponto é:

$$Z_2 = \frac{14,70 - 11,15}{2,238} = 1,58$$

A área para valores maiores do que 1,58 é 0,0571 ou 5,71%.

O que procuramos é a área entre Z_1 e Z_2 , como mostram os gráficos nas Figuras 6.8 e 6.9.

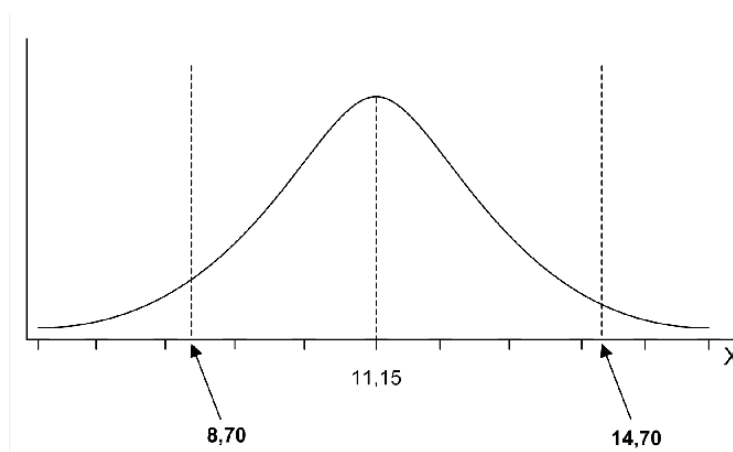


Figura 6.8: Área sob a curva normal

Portanto, fazemos:

$$0,8621 - 0,0571 = 0,8050$$

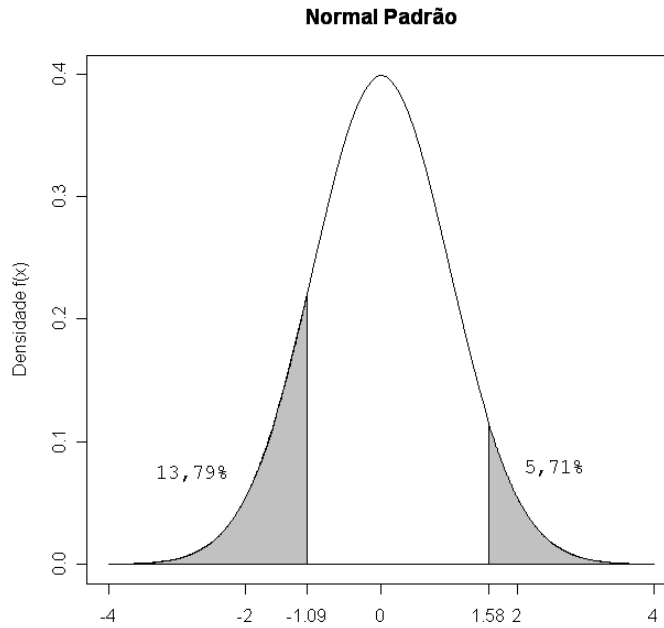


Figura 6.9: Área sob a curva normal

Ou seja, a porcentagem de arruelas com espessura entre 8,70 e 14,70 (limites de tolerância da especificação) é somente de 80,50%. Portanto, cerca de 19,50% das arruelas não atendem aos limites de especificações. Anteriormente, havíamos calculado esta porcentagem diretamente do histograma e o valor encontrado foi de 22%. A diferença entre os dois cálculos fica por conta da suposição de normalidade que fizemos.

□

6.2.1 Valor Esperado e Variância

Se X tem distribuição Normal, a média é dada por

$$\mu = E(X) = \int_{-\infty}^{\infty} f(x)dx \quad \text{e} \quad \mu \in (-\infty, +\infty).$$

A variância é dada por:

$$Var(X) = \sigma^2 = E(X^2) - [E(X)]^2 \quad \text{e} \quad \sigma^2 \in [0, +\infty).$$

Quando μ e σ são desconhecidos, como geralmente acontece, são substituídos por \bar{X} e s , respectivamente, valores obtidos de uma amostra, X_1, \dots, X_n , de X , .

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

6.2.2 Exercícios

Exercício 6.4. O tempo gasto no exame vestibular de uma universidade tem distribuição Normal, com média 120 min e desvio padrão 15 min. Sorteando um aluno ao acaso, qual é a probabilidade que ele termine o exame antes de 100 minutos? (Resp.: 0,0918) Qual deve ser o tempo de prova de modo a permitir que 95% dos vestibulandos terminem no prazo estipulado? (Resp.: 144,6 min.) Qual é o intervalo central de tempo, tal que 80% dos estudantes consigam completar o exame? (Resp.: 100,8 min. e 139,2 min.)

Exercício 6.5. Temos que a média de altura entre os homens, em determinado grupo de pessoas, é de 1,753 metros com desvio padrão de 0,07. calcule a probabilidade de selecionar de forma aleatória um homem e este ter a altura entre 1,7 e 1,9 metros. Calcule ainda a probabilidade dele ter ao menos 1,85 metros de altura.

Exercício 6.6. Uma empresa distribuidora de arroz, envaza sacos com peso médio de 5 kg de arroz, com variância de 0,01. Calcule:

- a) a probabilidade de selecionarmos um saco, e este conter mais de 5,3 Kg;
- b) a probabilidade de selecionarmos um saco, e este conter mais de 4,9 Kg e menos de 5,15 kg;
- c) a probabilidade de selecionarmos um saco, e este não conter mais de 4,92 kg.

Exercício 6.7. Latas de conservas são fabricadas por uma indústria com média de 250 g e desvio padrão de 5g. Uma lata é rejeitada pelo controle de qualidade dessa indústria se possuir peso menor que 235g. Calcule a probabilidade de uma lata ser rejeitada. (Resp : 0,0013)

Exercício 6.8. No enchimento de latas de cerveja, a quantidade de líquido colocado na garrafa é uma variável aleatória Normalmente distribuída de média 350 ml e desvio padrão 5 ml. Garrafas com menos de 340 ml são devolvidas para completar o enchimento. Calcular qual a porcentagem de garrafas devolvidas. (Resp : 2,27%)

Exercício 6.9. Seja X uma variável aleatória com distribuição Normal de média 22,3 e variância 1,5. Calcule:

- a) a probabilidade de X assumir valores inferiores a 18,89; (Resp : 0,0027)
- b) a probabilidade de X assumir valores superiores a 18,89; (Resp : 0,9973)
- c) a probabilidade de X assumir valores superiores a 21,15; (Resp : 0,8261)
- d) a probabilidade de X assumir valores entre 21,15 e 21,72. (Resp : 0,1440)

Exercício 6.10. *O consumo de gasolina por km rodado, para certo tipo de carro, em determinadas condições de teste, tem uma distribuição normal de média 100 ml e desvio padrão 5 ml. Calcular a probabilidade de um carro gastar de 95 a 110 ml (Resp : 0,8186). Calcule também, a probabilidade de consumir mais de 93 ml (Resp : 0,9192).*

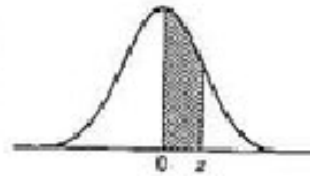
Exercício 6.11. *A vida útil média de lavadoras de pratos automáticas é de 1,5 anos, com desvio padrão de 0,3 anos. Se o tempo de vida distribue-se normalmente, que porcentagem das lavadoras vendidas necessitará de conserto antes de expirar o período de garantia de um ano? (Resp : 4,78%)*

Exercício 6.12. *O ginecologista do posto de saúde da vila fátima tem conhecimento, por registros anteriores, que a média de idade das mulheres gestantes que buscam seu atendimento é de 23,3 anos com desvio padrão de 1,35 anos. Além disso, considere que a variável idade seja ajustada pela distribuição normal. Com o intuito de ajudar o planejamento do posto de saúde, calcule:*

- a) *a probabilidade de que ao selecionar uma gestante de forma aleatória, ela tenha idade entre 21 e 24,6 anos; (Resp : 0,7880)*
- b) *a probabilidade de que ao selecionar uma gestante de forma aleatória, ela tenha idade entre 25,2 e 27,1 anos; (Resp : 0,0772)*
- c) *a probabilidade de que ao selecionar uma gestante de forma aleatória, ela tenha idade superior a 27,1 anos. (Resp : 0,0024)*
- d) *a probabilidade de que ao selecionar uma gestante de forma aleatória, ela não tenha idade superior a 21,8 anos. (Resp : 0,1333)*

ÁREAS UNDER THE NORMAL CURVE

An entry in the table is the proportion under the entire curve which is between $z = 0$ and a positive value of z . Areas for negative values of z are obtained by symmetry.



Second decimal place of z										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

From Paul G. Hoel, *Elementary Statistics*, 3rd ed., © 1971, John Wiley and Sons, Inc., New York, p. 287.

Figura 6.10: Tabela da distribuição Normal

Capítulo 7

Estimação

7.1 Introdução

Estimar o consumo médio de um automóvel, estimar o tempo médio que um funcionário leva para aprender uma nova tarefa, ou estimar a percentagem (proporção) de pessoas que irão consumir um produto que vai ser lançado no mercado, são exemplos de estimação. A estimação pode ser feita por dois processos:

- Estimação Pontual;
- Estimação Intervalar.

7.2 Definições

Veja algumas definições importantes:

- **Parâmetro:** São as quantidades da população, em geral desconhecidas e sobre as quais temos interesse;
- **Estimador:** É a combinação dos elementos da amostra, construída com a finalidade de representar, ou estimar, um parâmetro de interesse na população.
- **Estimativa:** São os valores numéricos assumidos pelos estimadores.

7.3 Estimação pontual

Na estimação pontual, estima-se o parâmetro θ desconhecido usando o valor de um estimador $\hat{\theta}$, o qual é designado por estimador pontual. A tabela 7.1 apresenta como exemplo alguns estimadores.

Parâmetro populacional	Exemplo de estimador pontual
Média (μ)	\bar{X}
Variância (σ^2)	S^2

Tabela 7.1:

7.4 Estimação Intervalar

A estimação intervalar consiste na determinação de um intervalo onde, com uma certa confiança (probabilidade), esteja o parâmetro θ desconhecido, levando-se em conta o seu estimador.

Assim, $P(L_1 < \theta < L_2) = 1 - \alpha$ significa que a probabilidade do intervalo aleatório (L_1, L_2) conter o valor exato de θ é $1 - \alpha$.

O intervalo (L_1, L_2) é designado por intervalo de confiança para o parâmetro θ , com um nível de confiança $1 - \alpha$. Depois de recolhida uma amostra aleatória, usam-se os valores observados dessa amostra, para calcular os valores observados das variáveis aleatórias L_1 e L_2 , que serão representados, respectivamente, por l_1 e l_2 . (l_1, l_2) é o intervalo de confiança concreto para aquela amostra. A figura 7.1 ilustra os intervalos para cada amostra.

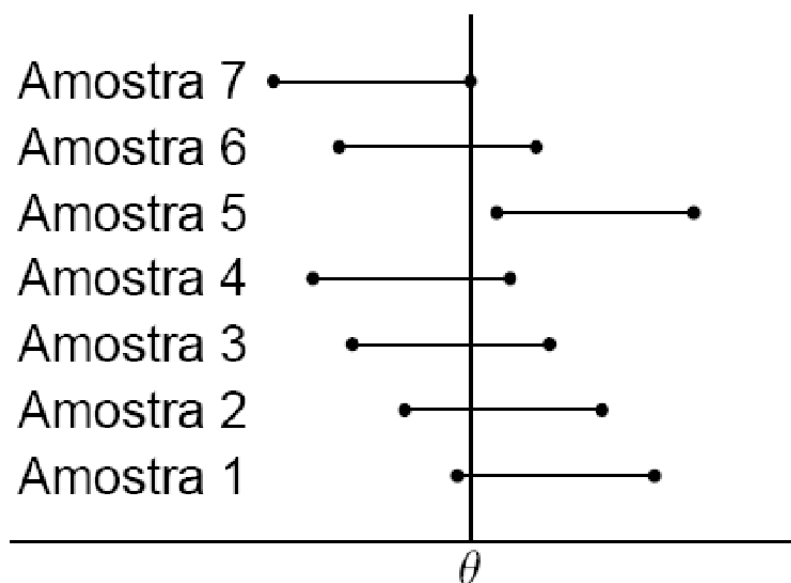


Figura 7.1:

Resumo dos Intervalos para a Média ou diferença entre médias de populações com distribuição Normal

Para uma amostra

1. Quando a variância é conhecida, o intervalo é dado por:

$$\bar{X} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

2. Quando a variância é desconhecida, o intervalo é dado por:

$$\bar{X} \pm t_{\frac{\alpha}{2}; n-1} * \frac{S}{\sqrt{n}}$$

Para duas amostras

1. Quando as amostras são independentes com variâncias conhecidas, o intervalo é dado por:

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

2. Quando as amostras são independentes com variâncias desconhecidas mas iguais, o intervalo é dado por:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}; n_1+n_2-2} * S_c * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Onde

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

7.5 Exercícios

Exercício 7.1. Calcule o intervalo de confiança para a média de uma distribuição $N(\mu; \sigma^2)$ em cada um dos casos a seguir:

média amostral (cm)	tamanho da amostra	desvio padrão da população (cm)	coeficiente de confiança (%)
170	100	15	95
165	184	30	85
180	225	30	70

Exercício 7.2. Um pesquisador está estudando a resistência de determinado material sob certas condições. Ele sabe que essa variável é normalmente distribuída com desvio padrão de duas unidades.

- a) Utilizando os valores 4,9; 7,0; 8,1; 4,5; 5,6; 6,8; 7,2; 5,7 e 6,2 unidades, obtidos de uma amostra aleatória de tamanho 9, determine o intervalo de 90% de confiança para a resistência média.
- b) Suponha que no item anterior o desvio padrão populacional não fosse conhecido. Como você procederia para obter o intervalo de confiança?

Exercício 7.3. De 50.000 válvulas fabricadas por uma companhia, retira-se uma amostra aleatória de tamanho 400, obtendo-se média amostral de 800 horas e desvio padrão amostral de 100 horas.

- a) Determine um intervalo de 90% de confiança para a vida média da população;
- b) Que tamanho deve ter a amostra para que seja de 95% a confiança na estimativa $800 \pm 7,84$?

Exercício 7.4. Medições do comprimento de 25 peças produzidas por uma máquina conduziram a uma média $\bar{x} = 140$ mm. Admita que cada peça tem comprimento aleatório com distribuição normal de valor esperado μ e desvio padrão $\sigma = 10$ mm, e que o comprimento de cada peça é independente das restantes. Construa um intervalo de confiança de 95% para a média da população.

Exercício 7.5. Um engenheiro civil mediu a resistência do cimento, considerando duas marcas. Das duas marcas, amostrou-se, de forma aleatória e independente 10 resistências, as quais estão apresentadas na tabela abaixo:

Marca A	3250	3268	4302	3184	3266	3297	3332	3502	3064	3116
Marca B	3094	3268	4302	3184	3266	3124	3316	3212	3380	3018

Ele assumiu que as amostras provêm de populações normais com desvio padrão igual a 353 e 133, respectivamente. Determine um intervalo de confiança de 95%, para a diferença entre as médias das duas populações.

Exercício 7.6. Para confrontar dois tipos de ceifeiras (máquinas) um campo de trigo foi dividido em secções longitudinais, e cada uma das secções adjacentes, tratadas por cada uma das máquinas. As produtividades alcançadas foram as seguintes:

Máquina A : 8,0; 8,4; 8,0; 6,4; 8,6; 7,7; 7,7; 5,6; 5,6; 6,2

Máquina B : 5,6; 7,4; 7,3; 6,4; 7,5; 6,1; 6,6; 6,0; 5,5; 5,5

Ao agricultor que experimenta as ceifeiras interessa averiguar se a produtividade média das duas máquinas pode se considerar igual. considere ($\alpha = 5\%$).

Exercício 7.7. Uma amostra aleatória de 50 empregados é tomada de uma linha de produção de 500. A média aritmética de horas extras trabalhadas por semana é cinco horas com desvio padrão de uma hora. Construa um intervalo de 99% de confiança para a média das horas extras trabalhadas por semana para toda a linha de produção.

Exercício 7.8. A média aritmética dos gastos com livros de uma amostra aleatória simples de 100 estudantes do primeiro ano de química é R\$70,0 com desvio padrão de R\$15,00. Construa um intervalo de 95% de confiança para o gasto médio de todos os estudantes.

Exercício 7.9. Uma amostra aleatória de 20 páginas datilografadas por um centro de datilografia mostrou que a média é de cinco erros por página com um desvio padrão de um erro. Estime, usando intervalos de 95 e 99% de confiança, a média da população e comente as diferenças em seus resultados.

Exercício 7.10. O gerente de uma empresa de transporte suspeita da afirmação de um vendedor de pneus de que o seu produto tem uma vida média de, ao menos, 28.000 Km. O gerente da empresa instala 40 desses pneus em seus caminhões, obtendo uma vida média de 27.563 Km, com desvio padrão de 1.348 Km. Construa um intervalo de confiança de 95% e conclua com relação a vida média afirmada pelo vendedor.

Exercício 7.11. Uma empresa avaliadora de imóveis está estudando as regiões central e parte alta da cidade de Jataí/GO. O objetivo principal é verificar se o preço médio, praticado para imóveis comerciais de um dado tamanho, é o mesmo nas duas regiões. De levantamentos anteriores, a empresa sabe que a área da parte alta apresenta uma heterogeneidade de preços imobiliários maior do que a região central, sendo os desvios padrões iguais a 0,82 UPC (Unidade Padrão de Construção) para a região da parte alta e 0,71 UPC para a região central. Foram realizados dois estudos, um em cada região, com amostras de tamanho 20 e 18 para a região central e parte alta, respectivamente. Das amostras, foram calculadas as médias que são 40,2 UPC para a região central e 36,7 UPC para a parte alta. Construa um intervalo de confiança de 95 % para a diferença das médias e verifique se o objetivo foi atingido.

Exercício 7.12. Duas técnicas de vendas são aplicadas por dois grupos de vendedores: a técnica A por 12 vendedores, e a técnica B, por 15 vendedores. Com resultados na tabela abaixo. Construa um intervalo de confiança com 95% para a diferença de média e verifique se existe diferença entre as técnicas. Informações adicionais permitem supor que as vendas sejam normalmente distribuídas, com variância comum, no entanto, desconhecida.

Dados	Técnica A	Técnica B
Média	68	76
variância	50	75

Exercício 7.13. Para estimar a renda semanal média de garçons de restaurantes em uma grande cidade, é colhida uma amostra da renda semanal de 75 garçons. A média e o desvio padrão amostrais encontrados são R\$227,00 e R\$15,00, respectivamente. Determine um intervalo de confiança, com coeficiente de confiança de 90%, para a renda média semanal.

Exercício 7.14. Seja X uma v.a. com distribuição normal de média desconhecida e variância igual a 36.

- Para uma amostra de tamanho 50, obtivemos média amostral 18,5. Construa um intervalo de confiança com coeficientes de confiança 91%, 96% e 99% para a média populacional;
- Para uma confiança de 94%, construa intervalos de confiança supondo três tamanhos de amostra 25, 50 e 100 (admita que todos forneceram média amostral igual a 18,5);

c) Comente sobre a precisão dos intervalos construídos em a) e b).

Exercício 7.15. O tempo de permanência de químicos recém formados no primeiro emprego, em anos, foi estudado considerando um modelo normal com média e variância desconhecidas. Deseja-se estimar a média populacional. Para uma amostra de 15 profissionais, a média obtida foi de 2,7 anos e o desvio padrão foi de 1,4 anos.

a) Encontre um intervalo para o tempo médio populacional de permanência com uma confiança de 90%;

b) Refaça o item a) considerando que a amostra era formada por 150 profissionais.

c) Comente sobre os resultados obtidos nos itens a) e b).

Exercício 7.16. A seguir encontra-se uma amostra de 10 árvores castanheiras todas com 8 anos de idade numa certa floresta. O diâmetro (polegadas) das árvores foram medidos à uma altura de 3 pés. Queremos encontrar um intervalo de confiança de 95% para o verdadeiro diâmetro médio de todas as árvores castanheiras dessa idade na floresta. Admita uma distribuição Normal.

19,4 21,4 22,3 22,1 20,1 23,8 24,6 19,9 21,5 19,1

Exercício 7.17. Os pulsos em repouso de 920 pessoas saudáveis foram tomados, e uma média de 72,9 batidas por minuto (bpm) e um desvio padrão de 11 bpm foram obtidos. Construa um intervalo de confiança de 95% para a pulsação média em repouso de pessoas saudáveis com base nesses dados.

Exercício 7.18. Para determinar o faturamento mensal das 500 maiores empresas de uma região, coletou-se uma amostra com o faturamento mensal de 60 dessas empresas. A média encontrada nessa amostra foi de R\$3.542,00. Sabendo-se que o desvio padrão do faturamento das 500 empresas é de R\$380,00 determinar o intervalo que deverá conter a média populacional. Utilize 68,26% como nível de confiança e considere uma população Normal.

Exercício 7.19. Uma empresa pretende treinar seus empregados em uma nova metodologia de trabalho e deseja saber quanto tempo precisará dispor para realizar esse treinamento. Para isso, selecionou 15 empregados para receber os novos conhecimentos. Na tabela 7.2 temos uma descrição de quanto tempo (em dias) cada empregado demorou para atingir o nível satisfatório. Adotar um nível de confiança de 95%.

Empregado	Tempo	Empregado	Tempo	Empregado	Tempo
1	52	6	59	11	54
2	44	7	50	12	58
3	55	8	54	13	60
4	44	9	62	14	62
5	45	10	46	15	63

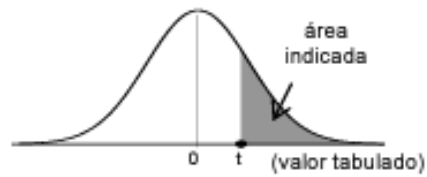
Figura 7.2: Dados referentes a dias utilizados de treinamento

Exercício 7.20. *Tintas para marcação em asfalto de rodovias são oferecidas em duas cores, branca e amarela. O tempo de secagem dessas tintas é de muito interesse e, especificamente, suspeita-se que a tinta amarela seca mais rápido do que a branca. Amostras foram obtidas para a marcação dos tempos de secagem em minutos, em condições reais das duas tintas.*

Branca 120 132 123 122 140 110 120 107

Amarela 126 124 116 125 109 130 125 117 129 120

Assumindo distribuição Normal, Obtenha o IC 95% para o tempo médio de secagem de cada tipo de tinta. Em seguida, Calcule o IC de 95% de confiança para a diferença entre as médias das duas populações.

IV Distribuição *t* de Student

g'	Área na cauda superior								
	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,660
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
35	0,682	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
45	0,680	1,301	1,679	2,014	2,412	2,690	2,952	3,281	3,520
50	0,679	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
z	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Figura 7.3: Tabela da distribuição Normal

Capítulo 8

Teste de Hipóteses

8.1 Introdução

Em geral, intervalos de confiança são a forma mais informativa de apresentar os principais resultados de um estudo. Entretanto, algumas vezes existe um particular interesse em decidir sobre a verdade ou não de uma hipótese específica (se dois grupos têm a mesma média ou não, ou se o parâmetro populacional tem um valor em particular ou não). Teste de hipóteses nos fornece a estrutura para que façamos isto.

Quando se colhe uma amostra da população, não se sabe com certeza se alguma medida estatística (como a média) dessa amostra corresponde realmente à população. Considerando que a amostra cobre somente uma fração do todo, não se espera, em geral, que essa medida da amostra corresponda exatamente à da população.

Entretanto, os testes de hipóteses são usados para verificar se as diferenças entre os valores da amostra e os valores da população são devidos ao acaso. Para isto, elabora-se uma hipótese sobre a população da qual a amostra foi retirada. Esta é chamada de hipótese nula, H_0 , pois, propõe que não exista diferença entre a amostra e a população, no aspecto que está sendo considerado. Em contrapartida, a hipótese alternativa, H_1 ou H_a , é formulada para testar a hipótese contrária à hipótese nula e pode ser tanto para mais quanto para menos.

8.1.1 Definições

Inicialmente, vamos apresentar algumas definições importantes:

- DECISÕES ESTATÍSTICAS

Na prática, somos chamados com muita frequência a tomar decisões acerca de populações, com base em informações de amostras, o que denominamos de Decisões Estatísticas.

- HIPÓTESES ESTATÍSTICAS

Ao tentarmos a fixação de decisões, é conveniente a formulação de suposições ou de conjecturas acerca das populações de interesse, que, em geral, consistem em considerações sobre parâmetros das mesmas. Essas suposições, que podem ser ou não verdadeiras, são denominadas de Hipóteses Estatísticas.

- HIPÓTESE NULA

É aquela Hipótese Estatística, prefixada, formulada sobre o parâmetro populacional estudado, com o único propósito de ser rejeitada ou invalidada. É representada por H_0 .

- HIPÓTESE ALTERNATIVA

São quaisquer hipóteses que difiram da Hipótese Nula. Utilizaremos uma hipótese alternativa, representada por H_1 .

- TESTE DE HIPÓTESES

Os processos que habilitam a decidir se aceitam ou rejeitam as hipóteses formuladas, ou determinar se a amostra observada difere, de modo significativo, dos resultados esperados, são denominados de Testes de Hipóteses ou Testes de Significância.

- ERROS DO TIPO I E TIPO II

Decisões possíveis	Estados possíveis	
	Ho verdadeira	Ho falsa
Aceitação de H_0	Decisão correta	Erro do tipo II
Rejeição de H_0	Erro do tipo I	Decisão correta

Qual, entre os dois tipos de erro, é o mais grave e que deve ser evitado? Façamos uma analogia com a decisão de um Juiz de Direito. O que será mais grave? Condenar um inocente ou absolver um culpado? É claro que será mais grave a condenação de um inocente. Rejeitar a hipótese nula sendo ela verdadeira equivale a condenar um inocente, logo, o Erro Tipo I é o mais grave e então, devemos minimizar a probabilidade de cometer este erro. Esta probabilidade chama-se Nível de Significância do Teste, e denotamos por α . Já a probabilidade do Erro Tipo II, denotado por β , não pode ser calculada, a menos que se especifique um valor alternativo para o parâmetro desconhecido, que está sob teste. O poder do teste é dado por $(1 - \beta)$.

- NÍVEL DE SIGNIFICÂNCIA

Ao testar uma hipótese estabelecida, a probabilidade máxima com a qual se sujeitaria a correr o risco de um erro do tipo I é denominada de Nível de Significância do Teste. Essa probabilidade, representada freqüentemente por α , é geralmente especificada antes da extração de quaisquer amostra, de modo que os resultados obtidos não influenciem na escolha.

Exemplo 8.1. Suponha que é feita uma auditoria na empresa do Sr Mario, a qual resulta numa acusação de infração fiscal. Obviamente, se o fiscal das finanças não conseguir juntar provas que sustentem a acusação, a empresa não será considerada culpada. Dessa forma, as hipóteses estatísticas são as seguintes:

$$\begin{cases} H_0 & : \text{ a empresa não cometeu uma infração fiscal} \\ H_1 & : \text{ a empresa cometeu uma infração fiscal} \end{cases}$$

Nesta situação podem ser tomadas duas decisões:

- Rejeitar a hipótese H_0 : a empresa é considerada culpada, isto é, aceita-se a hipótese H_1 como sendo verdadeira;
- Não rejeitar a hipótese H_0 : não se conseguiu provar a veracidade de H_1 e como tal, não se pode rejeitar a hipótese H_0 . Note que, isto não significa aceitar H_0 , significa tão só, que não há provas (não há evidência) para rejeitar esta hipótese. Por isso, é preferível dizer “não rejeitar H_0 ” a dizer, “aceitar H_0 ”.

Neste caso, há duas possibilidades de tomar uma decisão errada:

- a empresa é considerada culpada quando de fato não cometeu nenhuma infração fiscal: Rejeitar H_0 sendo H_0 verdadeira (Erro do Tipo I);
- não se rejeita a hipótese de a empresa ser inocente, quando de fato esta cometeu uma infração fiscal: Não rejeitar H_0 sendo H_0 falsa (Erro do Tipo II).

Exemplo 8.2. Testar uma hipótese quer dizer avaliar uma crença sobre a população. Esta crença a ser testada recebe o nome de hipótese nula e a hipótese alternativa será o oposto dela.

Por exemplo: um fiscal de controle de qualidade de uma empresa que produz milho em conserva quer saber se todas as latas de milho estão seguindo as especificações da embalagem em termos do peso, no caso: 250 gramas. A hipótese nula deste fiscal seria: o peso das latas é igual a 250g - esta é a crença que ele quer testar - e a hipótese alternativa é o oposto dela, ou seja, diferente de 250 g. Para realizar este estudo o fiscal pegaria uma amostra de latas de diferentes lotes e pesaria para avaliar se as máquinas ao despejarem o milho na lata estão reguladas o suficiente para colocar em média os 250 g de milho em cada latinha. Caso seja confirmada a sua hipótese ele não precisaria tomar outras providências, e com isso, a produção poderia continuar. Caso sua hipótese seja rejeitada, medidas de manutenção deveriam ser tomadas.

Exercício 8.1. Quando um diagnóstico médico é fornecido, qual dos erros é geralmente mais sério: um resultado falso positivo que diz que a pessoa tem a doença quando na verdade ela não tem ou um

resultado falso negativo, que diz que a pessoa não tem a doença quando na verdade ela tem? Imagine as situações onde a pessoa está fazendo parte de um screening para câncer de mama e, em outro, a pessoa realiza o teste para detectar anticorpos anti-HIV. Apresente as hipóteses nula e alternativa sobre a situação de saúde do paciente; fazendo uma analogia com teste de hipóteses, que tipo de erro (I ou II) seria cometido se o resultado do teste fosse falso positivo? Que tipo de erro (I ou II) seria cometido se o resultado do teste fosse falso negativo?

8.1.2 Classificação dos testes

Os testes de hipóteses podem ser identificados conforme os seguintes casos: Monocaudal a direita, Monocaudal a esquerda e Bicaudal. Como exemplo, considere as seguintes situações:

Caso 1: Monocaudal a direita: Supor que se deseja comparar a eficácia de uma nova droga (DN) com a eficácia de uma droga padrão, que vem sendo utilizada (DA).

$$\begin{cases} H_0 & : D_N = D_A \\ H_1 & : D_N > D_A \end{cases}$$

Caso 2: Monocaudal a esquerda: Se o estudo envolvesse a comparação de duas drogas onde a nova droga (DN) se propõe a reduzir os efeitos colaterais.

$$\begin{cases} H_0 & : D_N = D_A \\ H_1 & : D_N < D_A \end{cases}$$

Caso 3: Bicaudal: Está sendo lançada uma nova droga (DN) para depressão e deseja-se investigar se a droga inibe (ou provoca) o apetite, como efeito colateral. Assim, antes do estudo não se conhece o efeito da droga sobre o apetite dos pacientes.

$$\begin{cases} H_0 & : D_N = D_A \\ H_1 & : D_N \neq D_A \end{cases}$$

Os testes podem ainda ser classificados como paramétricos ou não paramétricos. Podemos entender isto como:

- **Testes Paramétricos:** São os testes onde consideramos suposições sobre a distribuição dos dados da população em análise, e além disso, suposições sobre os parâmetros de interesse; se alguma(s) destas suposições são violadas, então os testes tradicionais utilizados não têm rigor estatístico, e deverão ser evitados. Em sua substituição deveremos utilizar testes que não exigem o cumprimento de tais pressupostos. Estes testes designam-se por testes não paramétricos.

- **Testes Não Paramétricos:** Os testes não paramétricos não estão condicionados por qualquer distribuição de probabilidades dos dados em análise, sendo também conhecidos por testes de distribuição livre.

Tal como não é estatisticamente rigorosa a utilização de testes paramétricos quando não se cumprem os pressupostos necessários, também deverá ser evitada a utilização dos testes não paramétricos em situações em que prevalecem as condições de utilização dos testes paramétricos, pois estes (paramétricos) são mais potentes que os testes não paramétricos.

Após, fazermos comentários gerais sobre teste de hipóteses, vamos resumir as etapas aplicadas para a realização dos mesmos:

- 1º Determinar as hipóteses nula e alternativa que são apropriadas para a aplicação;
- 2º Selecionar a estatística de teste que será usada para decidir rejeitar ou não a hipótese nula;
- 3º Especificar o nível de significância α para o teste;
- 4º Usar o nível de significância para desenvolver a regra de decisão que indica os valores críticos da estatística de teste que levará a rejeição de H_0 ;
- 5º Coletar os dados amostrais e calcular a estatística de teste;
- 6º Comparar o valor da estatística do teste com o(s) valor(es) crítico(s) especificado(s) na regra de decisão para determinar se H_0 deve ser rejeitado;

8.2 Teste de hipóteses para a média de populações com distribuição Normal

Seja X uma variável aleatória com distribuição Normal com parâmetro de média μ e parâmetro de variância σ^2 . Nosso interesse agora é desenvolver um teste para o parâmetro de média e, para isto, as hipóteses a serem testadas são:

$$\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{array} \right.$$

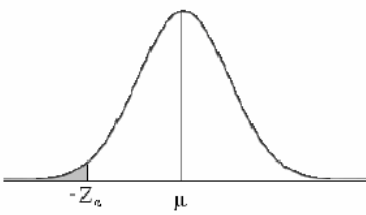
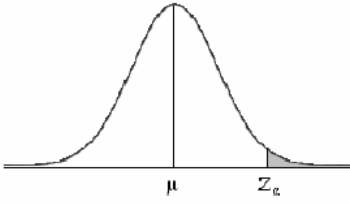
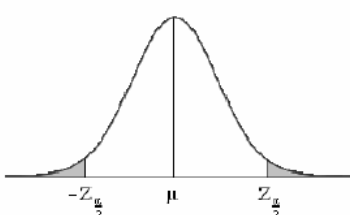
Para o caso de uma população, temos duas situações:

1. Variância conhecida ou tamanho de amostra grande.

Neste caso, a estatística de teste é dada por:

$$Z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Aqui, Z_{obs} tem distribuição normal padrão. Dessa forma, a região de rejeição da hipótese nula pode ser observada na figura abaixo. Na figura, Z_α vem da tabela da distribuição normal padrão.

Hipóteses	Rejeita-se H_0 (ao n.s = α)	Região de Rejeição de H_0
$H_0: \mu = \mu_0$ $H_A: \mu < \mu_0$	$Z_{obs} < -Z_\alpha$	
$H_0: \mu = \mu_0$ $H_A: \mu > \mu_0$	$Z_{obs} > Z_\alpha$	
$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$Z_{obs} < -Z_{\frac{\alpha}{2}}$ ou $Z_{obs} > Z_{\frac{\alpha}{2}}$	

Se Z_{obs} cair na região de rejeição, dizemos que rejeitamos a hipótese nula.

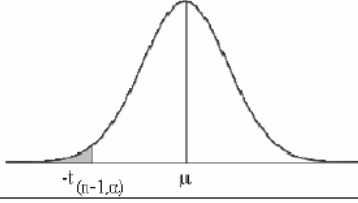
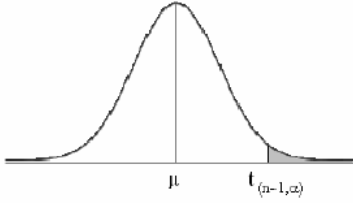
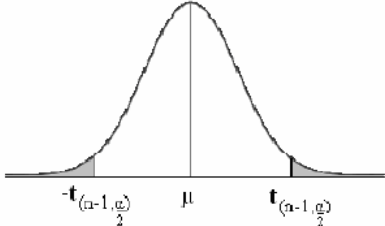
2. Variância desconhecida ou tamanho de amostra pequeno.

Neste caso, a estatística de teste é dada por:

$$T_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

Aqui, T_{obs} tem distribuição t-Student com $(n - 1)$ graus de liberdade. Dessa forma, a região de rejeição da hipótese nula pode ser observada na figura abaixo. Na figura, $t_{(n-1;\alpha)}$ vem da tabela da distribuição t-Student com $(n - 1)$ graus de liberdade e nível de significância α .

Se T_{obs} cair na região de rejeição, dizemos que rejeitamos a hipótese nula.

Hipóteses	Rejeita-se H_0 (ao n.s = α)	Região de Rejeição de H_0
$H_0: \mu = \mu_0$ $H_A: \mu < \mu_0$	$T_{obs} < -t_{(n-1; \alpha)}$	
$H_0: \mu = \mu_0$ $H_A: \mu > \mu_0$	$T_{obs} > t_{(n-1; \alpha)}$	
$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$T_{obs} < -t_{(n-1; \frac{\alpha}{2})}$ ou $T_{obs} > t_{(n-1; \frac{\alpha}{2})}$	

Exemplo 8.3. Sabe-se que em indivíduos sem problemas de visão, a pressão intra-ocular média é 20 mmHg. Um oftalmologista, desejando comprovar que há aumento na pressão intra-ocular em pacientes com glaucoma, selecionou 25 indivíduos portadores da doença e mediu, em cada um, a pressão intra-ocular. Obteve uma média amostral de 21,5 mmHg e um desvio-padrão de 2 mmHg. Os dados apoiam a conjectura do pesquisador, ao nível de significância de 5%?

Solução:

A hipótese nula do Oftalmologista é que a média (μ) da pressão intra-ocular, em pacientes saudáveis, é de 20 mmHG, no entanto, ele deseja mostrar que pacientes com glaucoma possuem uma pressão intra-ocular maior que 20, que neste caso, denominamos de hipótese alternativa. Neste caso, as hipóteses do teste unilateral são:

$$\begin{cases} H_0 & : \mu = 20 \\ H_1 & : \mu > 20 \end{cases}$$

Neste caso, como o desvio padrão ($S = 2$) veio da amostra, a estatística de teste é:

$$\begin{aligned} T_{obs} &= \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{21,5 - 20}{2/\sqrt{25}} \\ &= 3,75 \end{aligned}$$

Pela tabela da distribuição t-Student, com $25 - 1 = 24$ graus de liberdade e nível de significância de 0,05, temos que, $t_{24; 0,05} = 1,710882$. Logo, a região de aceitação é dada pelo intervalo $(-1,710882; 1,710882)$, com isso, como $T_{obs} = 3,75$ não pertence a essa região, dizemos que rejeitamos

a hipótese nula ao nível de significância de 0,05. Ou seja, ao nível de 5% de significância, pode-se concluir que a amostra fornece evidências estatísticas para dizer que há um aumento na pressão intra-ocular média em pessoas com glaucoma.

Exercício 8.2. Suponha que o tempo médio de terapia tradicional em pacientes com depressão seja de 2 anos. Admita ainda que se pretende testar um tipo de terapia alternativa cujo tempo de recuperação esperado seja menor que 2 anos. Realizou-se um experimento com 25 pacientes submetidos à nova terapia e obteve-se tempo médio de 1,5 anos e desvio padrão de 0,5 anos. Com base nestas informações teste a hipótese que a nova terapia apresenta tempo de recuperação inferior a 2 anos. Use significância de 5%.

Exercício 8.3. A média obtida através dos anos em um teste vocacional foi de 100 pontos. Com o objetivo de saber se a nova classe (calouros) é típica com respeito a vocação, tomou-se uma amostra de 50 alunos. O resultado foi uma média 95 com desvio padrão de 10. Pode-se afirmar, a um nível de significância de 5 %, que essa nova turma é igual às anteriores ?

Exercício 8.4. A média em dias de internação de crianças que sofreram acidente de trânsito e que não estavam usando o cinto de segurança é de 1,39 dias. Em um levantamento de 123 crianças que estavam usando o cinto, a média foi de 0,83 dias e desvio padrão de 0,16 dias. Podemos concluir que o uso do cinto diminui o tempo médio de internação? Adote = 5 %.

Exercício 8.5. A fim de acelerar o tempo que um analgésico leva para surtir efeito, um químico analista acrescentou certo ingrediente à fórmula original, que acusava um tempo médio de 43 minutos para fazer efeito. Em 49 observações com a nova fórmula, obteve-se um tempo médio de 41 minutos, com desvio padrão de 10 minutos. A nova fórmula é melhor, pior ou igual a anterior ? Adote = 5 %.

Exercício 8.6. A altura média dos estudantes da UFG é de 1,70 m. Em uma amostra casual de tamanho 25 foi estimada a média de 1,72 m e desvio padrão da amostra de 0,08 m. Pode-se considerar que a média amostral não difere da média da população? Adote = 5 %.

Exercício 8.7. A quantidade de calorias de um produto segue uma distribuição normal. Para a indústria, a média é 30, mas para os concorrentes este valor é diferente. Para avaliar o produto foi tirada uma amostra de tamanho 25, cujos valores são apresentados a seguir: 30,05; 29,38; 28,45; 31,22; 31,07; 34,44; 34,50; 34,48; 31,75; 30,59; 31,92; 31,76; 30,25; 33,28; 33,40; 31,46; 31,43; 32,92; 29,91; 33,63; 27,98; 33,07; 31,01; 29,85; 29,70. Faça um teste de hipóteses ao nível de 5% e verifique se a concorrência tem razão.

Exercício 8.8. A associação dos proprietários de indústrias metalúrgicas está preocupada com o tempo perdido em acidentes de trabalho, cuja média, nos últimos tempos, tem sido da ordem de 60 hora /homens por ano com desvio padrão de 20 horas/homem. Tentou-se um programa de prevenção

de acidentes e, após o mesmo, tomou-se uma amostra de 9 indústrias e mediu-se o número de horas/homem perdidas por acidente, que foi de 50 horas. Você diria, ao nível de 5% de melhoria?

Exercício 8.9. O tempo médio, por operário, para executar uma tarefa, tem sido 100 minutos. Introduziu-se uma modificação para diminuir este tempo, e, após certo período, sorteou-se uma amostra de 16 operários, medindo-se o tempo de execução gasto por cada um. O tempo médio da amostra foi 85 minutos com desvio padrão de 12 minutos. Este resultado evidencia uma melhora no tempo gasto para realizar a tarefa? Apresente as conclusões aos níveis de 5% e 1% de significância.

Exercício 8.10. Uma firma tem seguido a política de oferecer uma garantia de 2000 utilizações para determinado aparelho que comercializa. Este procedimento baseia-se em estudos levados a cabo no período inicial de produção, que indicavam um número médio de utilizações possíveis por aparelho de 2060, com uma variabilidade traduzida por $\sigma = 20$. Existindo indícios de que presentemente a situação pode ter mudado, pretende-se averiguar se continua a ser 2060 o número médio de utilizações por aparelho. Para o efeito foram seleccionados ao acaso e testados pela firma 10 aparelhos, os quais forneceram os seguintes valores: 2100; 2025; 2071; 2067; 2150; 2115; 2064; 2088; 1995; 2095. Suponha que o número de utilizações permitidas por aparelho comporta-se de forma aproximadamente normal. Faça um teste de hipóteses ao nível de 5% e conclua sobre a situação.

Exercício 8.11. As estaturas de 20 recém nascidos foram tomadas no Departamento de Pediatria do hospital das clínicas, cujos resultados são em cm: 41; 50; 52; 49; 49; 54; 50; 47; 52; 49; 50; 52; 50; 47; 49; 51; 46; 50; 49; 50. Diante disso, a) suponha inicialmente que a população das estaturas é normalmente distribuída com variância 2 cm². Teste a hipótese de que a média seja diferente de 50cm ($\alpha = 0,05$) b) Faça o mesmo teste para a média, mas agora desconhecendo a variância ($\alpha = 0,05$).

Exercício 8.12. Uma companhia de cigarros anuncia que o índice médio de nicotina dos cigarros que fabrica apresenta-se abaixo de 23 mg por cigarro. Um laboratório realiza 6 análises desse índice, obtendo: 27, 24, 21, 25, 26, 22. Sabe-se que o índice de nicotina se distribui normalmente, com variância igual a 4,86 mg². Pode-se aceitar, ao nível de 10%, a afirmação do fabricante?

8.3 Teste de hipóteses para a diferença entre médias de duas populações com distribuição Normal

Sejam, X uma variável aleatória com distribuição Normal com parâmetro de média μ_x e parâmetro de variância σ_x^2 e Y uma variável aleatória com distribuição Normal com parâmetro de média μ_y e parâmetro de variância σ_y^2 . Nosso interesse agora é desenvolver um teste para a diferença entre os

parâmetros de média das duas populações, para isto, as hipóteses a serem testadas são:

$$\begin{cases} H_0 : \mu_x = \mu_y & \text{ou} & \mu_x - \mu_y = 0 \\ H_1 : \mu_x \neq \mu_y & \text{ou} & \mu_x - \mu_y \neq 0 \end{cases}$$

ou

$$\begin{cases} H_0 : \mu_x = \mu_y & \text{ou} & \mu_x - \mu_y = 0 \\ H_1 : \mu_x > \mu_y & \text{ou} & \mu_x - \mu_y > 0 \end{cases}$$

ou

$$\begin{cases} H_0 : \mu_x = \mu_y & \text{ou} & \mu_x - \mu_y = 0 \\ H_1 : \mu_x < \mu_y & \text{ou} & \mu_x - \mu_y < 0 \end{cases}$$



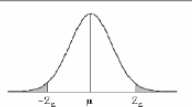
Temos duas situações:

1. Variâncias (σ_x^2 e σ_y^2) conhecidas.

Neste caso, a estatística de teste é dada por:

$$Z_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}$$

Aqui, Z_{obs} tem distribuição normal padrão. Dessa forma, a região de rejeição da hipótese nula pode ser observada na figura abaixo. Na figura, Z_α vem da tabela da distribuição normal padrão.

Hipóteses	Rejeita-se H_0 (ao n.s = α)	Região de Rejeição de H_0
$H_0: \mu_x = \mu_y$ $H_A: \mu_x < \mu_y$	$Z_{obs} < -Z_\alpha$	
$H_0: \mu_x = \mu_y$ $H_A: \mu_x > \mu_y$	$Z_{obs} > Z_\alpha$	
$H_0: \mu_x = \mu_y$ $H_A: \mu_x \neq \mu_y$	$Z_{obs} < -Z_{\frac{\alpha}{2}}$ ou $Z_{obs} > Z_{\frac{\alpha}{2}}$	

Se Z_{obs} cair na região de rejeição, dizemos que rejeitamos a hipótese nula.

Exemplo 8.4. A quantidade de um certo elemento no sangue varia segundo o sexo. Para os homens o desvio padrão é de 14,1 ppm e para as mulheres é 9,5 ppm. Amostras aleatórias de 75 homens e 50 mulheres forneceram média de 28 e 33 ppm respectivamente. Pode-se afirmar ao nível de 5% que a média de concentração do elemento no sangue é a mesma para ambos sexos?

Solução

Sejam, X a variável aleatória que representa a quantidade de elemento no sangue do homem e Y a variável aleatória que representa a quantidade de elemento no sangue da mulher. Com isso, temos que, as hipóteses a serem testadas são:

$$\begin{cases} H_0 & : \mu_x = \mu_y \quad \text{ou} \quad \mu_x - \mu_y = 0 \\ H_1 & : \mu_x \neq \mu_y \quad \text{ou} \quad \mu_x - \mu_y \neq 0 \end{cases}$$

Do enunciado temos, $n_x = 75$ e $n_y = 50$, $\bar{x} = 28$ e $\bar{y} = 33$, $\sigma_x^2 = 14,1^2 = 198,81$ e $\sigma_y^2 = 9,5^2 = 90,25$. Logo,

$$\begin{aligned} Z_{obs} &= \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}} \\ &= \frac{28 - 33}{\sqrt{198,81/75 + 90,25/50}} \\ &= -2,36 \end{aligned}$$

Pela tabela da distribuição normal, temos que $Z_\alpha = 1,96$ e com isso, como $Z_{obs} < -Z_\alpha$ (veja figura acima), ou seja, $-2,36 < -1,96$, logo, rejeitamos a hipótese nula e portanto, dizemos que a quantidade de elementos no sangue do homem e da mulher são diferentes.

2. Variâncias desconhecidas porém iguais.

Neste caso, a estatística de teste é dada por:

$$T_{obs} = \frac{\bar{x} - \bar{y}}{S_c / \sqrt{1/n_x + 1/n_y}}$$

onde

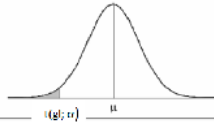
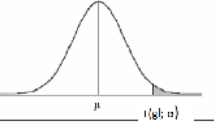
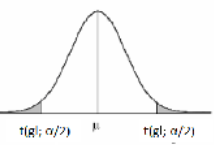
$$S_c = \sqrt{\frac{(n_x - 1) * S_x^2 + (n_y - 1) * S_y^2}{n_x + n_y - 2}}$$

Aqui, T_{obs} tem distribuição t-Student com $(n_x + n_y - 2)$ graus de liberdade. Dessa forma, a região de rejeição da hipótese nula pode ser observada na figura abaixo. Na figura, $t_{(gl;\alpha)}$ vem da tabela da distribuição t-Student com $(gl = n_x + n_y - 2)$ graus de liberdade e nível de significância α .

Se T_{obs} cair na região de rejeição, dizemos que rejeitamos a hipótese nula.

Exemplo 8.5. A quantidade de um certo elemento no sangue varia segundo o sexo. Amostras aleatórias de 75 homens e 50 mulheres forneceram média de 28 e 33 ppm e desvio padrão de 12,7 e 8,9 ppm, respectivamente. Pode-se afirmar ao nível de 5% que a média de concentração do elemento no sangue é a mesma para ambos sexos?

Solução

Hipóteses	Rejeita-se H_0 (ao n.s. = α)	Região de Rejeição de H_0
$H_0: \mu_X = \mu_Y$ $H_A: \mu_X < \mu_Y$	$T_{obs} < -t_{(g); \alpha}$	
$H_0: \mu_X = \mu_Y$ $H_A: \mu_X > \mu_Y$	$T_{obs} > t_{(g); \alpha}$	
$H_0: \mu_X = \mu_Y$ $H_A: \mu_X \neq \mu_Y$	$T_{obs} < -t_{(g); \alpha/2}$ ou $T_{obs} > t_{(g); \alpha/2}$	

Sejam, X a variável aleatória que representa a quantidade de elemento no sangue do homem e Y a variável aleatória que representa a quantidade de elemento no sangue da mulher. Com isso, temos que, as hipóteses a serem testadas são:

$$\begin{cases} H_0 : \mu_x = \mu_y & \text{ou} & \mu_x - \mu_y = 0 \\ H_1 : \mu_x \neq \mu_y & \text{ou} & \mu_x - \mu_y \neq 0 \end{cases}$$

Do enunciado temos, $n_x = 75$ e $n_y = 50$, $\bar{x} = 28$ e $\bar{y} = 33$, $S_x^2 = 12,7^2 = 161,29$ e $S_y^2 = 8,9^2 = 79,21$. Logo,

$$\begin{aligned} S_c &= \sqrt{\frac{(n_x - 1) * S_x^2 + (n_y - 1) * S_y^2}{n_x + n_y - 2}} \\ &= \sqrt{\frac{(75 - 1) * 161,29 + (50 - 1) * 79,21}{75 + 50 - 2}} \\ &= 11,3398 \end{aligned}$$

Com isso,

$$\begin{aligned} T_{obs} &= \frac{\bar{x} - \bar{y}}{S_c \sqrt{1/n_x + 1/n_y}} \\ &= \frac{28 - 33}{11,3398 \sqrt{1/75 + 1/50}} \\ &= -2,415 \end{aligned}$$

Pela tabela da distribuição t -Student com $gl = n_x + n_y - 2 = 75 + 50 - 2 = 123$ e $\alpha = 0,05$, temos que $t_{(gl; \alpha/2)} = 1,96$ e com isso, como $T_{obs} < -T_{(gl; \alpha/2)}$ (veja figura acima), ou seja, $-2,415 < -1,96$, rejeitamos a hipótese nula e portanto, dizemos que a quantidade de elementos no sangue do homem e da mulher são diferentes.

Exercício 8.13. Um estudo foi feito para verificar se existe diferença significativa entre as notas médias de alunos da escola pública e de alunos da escola particular na disciplina de matemática. Uma

amostragem feita com 10 alunos da escola pública e com 12 alunos da escola particular mostrou os seguintes resultados: *PUB* 60; 65; 45; 55; 50; 75; 80; 85; 70; 70; *PAR* 66; 81; 48; 45; 50; 55; 60; 65; 65; 75; 90; 58. Verifique, por meio do teste de hipóteses com 5% de significância, se a nota média da escola pública é igual a nota média da escola particular. Considere que as variâncias são desconhecidas, porém iguais.

Exercício 8.14. A tabela a seguir mostra um resumo dos resultados obtidos em uma pesquisa, que tinha como um dos objetivos, verificar a existência de diferença significativa entre o salário recebido por homens e por mulheres em um determinado emprego.

<i>Sexo</i>	<i>n</i>	<i>Média</i>	<i>Desvio Padrão</i>
<i>Masculino</i>	25	12	4
<i>Feminino</i>	22	10	4

Verificar se existe diferença significativa entre o salário médio recebido por pessoas do sexo masculino e por pessoas do sexo feminino. Use o teste de hipóteses para diferença entre médias, com 1% de significância.

Exercício 8.15. Em uma determinada pesquisa onde se utilizou uma amostra de 15 meninos e 10 meninas em idade pré-escolar, obteve-se peso médio de meninos de 20 kg, com desvio padrão de 5 kg e peso médio de meninas de 18 kg com desvio padrão de 3 kg. Usando uma significância de 1%, podemos dizer que existe diferença significativa entre peso de meninos e de meninas.

Exercício 8.16. Estuda-se o conteúdo de nicotina de duas marcas de cigarros (*A* e *B*), obtendo-se os seguintes resultados. *A*: 17; 20; 23; 20 *B*: 18; 20; 21; 22; 24 Admitindo que o conteúdo de nicotina das duas marcas tem distribuição normal e que as variâncias populacionais são iguais, com $\alpha = 0,05$, pode-se afirmar que existe alguma diferença significativa no conteúdo médio de nicotina entre as duas marcas?

Exercício 8.17. Uma clínica pretende comparar dois tipos de dietas. Com esse objectivo, escolheu aleatoriamente e independentemente, uma amostra de pacientes com excesso de peso e durante 10 semanas 30 desses pacientes foram sujeitos à dieta 1 e 27 à dieta 2. Após as 10 semanas, anotou-se o total de peso perdido (em kg) por cada paciente e obteve-se média de 9,3 quilos e desvio padrão de 2,4 quilos para a dieta 1 e média de 8,2 quilos e desvio padrão de 2,6 quilos para a dieta 2. Pode-se admitir que as duas dietas têm, em média, o mesmo efeito na perda de peso? Justifique a resposta considerando $\alpha = 0.05$.

Exercício 8.18. Um fabricante produz dois tipos de pneus. Para o pneu do tipo *A* o desvio padrão é de 2500 km e para o pneu do tipo *B* é de 3000 km. Uma cia de taxis testou 50 pneus do tipo *A* e 40 do tipo *B*, obtendo 24000 km de média para o tipo *A* e 26000 km para o tipo *B*. Adotando $\alpha = 4\%$ testar a hipótese de que a duração média dos dois tipos é a mesma.

Exercício 8.19. Um engenheiro desconfia que a qualidade de um material pode depender da matéria-prima utilizada. Há dois fornecedores de matéria-prima sendo usados. Testes com 10 observações de cada fornecedor indicaram que a média e desvio padrão do fornecedor 1 são 39 und e 7 und, respectivamente e para o fornecedor 2, 43 und e 9 und, respectivamente. Use um nível de significância de 5% e teste a hipótese do engenheiro.

Exercício 8.20. A altura média de 50 estudantes do sexo masculino que tiveram participação superior à média nas atividades atléticas colegiais era de 178,23 cm, com desvio padrão de 6,35 cm. Enquanto que os 50 que não mostraram nenhum interesse nessas atividades apresentaram a altura média de 175,45 cm, com desvio de 7,11 cm. Testar a hipótese dos estudantes do sexo masculino que participam de atividades atléticas serem diferente dos demais. Adote $\alpha = 5\%$.

Exercício 8.21. Pode-se concluir que as crianças nascidas em hospital particular são mais pesadas do que as crianças nascidas em hospital público? Adote $\alpha = 5\%$.

Hospital	Tamanho da amostra	Média (kg)	Desvio padrão (kg)
Particular	50	3,1	1,6
Público	50	2,7	1,4

Exercício 8.22. um estudo relata os resultados de um ensaio clínico aleatorizado, duplo-cego, realizado com o objetivo de comparar a tianeptina com o placebo. Participaram desse estudo pacientes de Belo Horizonte, Campinas e Rio de Janeiro. Sucintamente, o ensaio consistiu em administrar a droga a dois grupos de pacientes, compostos de forma aleatória, e quantificar a depressão através da escala de MADRS, em que os valores maiores indicam maior gravidade da doença. O escore foi obtido para cada paciente 7, 14, 21, 28 e 42 dias após o início do ensaio. Pelo planejamento adotado, os dois grupos não diferiam em termos de depressão no início do ensaio. Assim, uma evidência sobre o efeito da tianeptina é obtida comparando-se os dois grupos ao fim de 42 dias. A Tabela abaixo apresenta os escores finais dos pacientes dos dois grupos admitidos em Belo Horizonte.

Grupo	Escores
Placebo	6; 33; 21; 26; 10; 29; 33; 29; 37; 15; 2; 21; 7; 26; 13
Tianeptina	10; 8; 17; 4; 17; 14; 9; 4; 21; 3; 7; 10; 29; 13; 14; 2

Faça um teste e verifique se a média de depressão do grupo placebo difere do grupo que utilizou Tianeptina. Adote $\alpha = 5\%$.

8.4 Teste de hipóteses para a diferença entre médias de duas populações com observações pareadas

Aqui, vamos apresentar o procedimento para realizar testes de comparação de médias para dados pareados (amostras dependentes), obtidas de populações Normais. Para cada par definido, o valor da primeira amostra está claramente associado ao respectivo valor da segunda amostra.

Para observações pareadas, o teste apropriado para a diferença entre duas médias consiste em determinar primeiro a diferença (d) entre cada par de valores, e então testar a hipótese nula de que a média das diferenças na população é zero. Então, do ponto de vista de cálculo, o teste é aplicado a uma única amostra de valores d . Portanto, testamos as seguintes hipóteses:

$$\left\{ \begin{array}{l} H_0 : d = 0 \\ H_1 : d \neq 0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0 : d = 0 \\ H_1 : d > 0 \end{array} \right. \quad \text{ou} \quad \left\{ \begin{array}{l} H_0 : d = 0 \\ H_1 : d < 0 \end{array} \right.$$

Neste caso, a estatística de teste é dado por:

$$T_{obs} = \frac{\bar{d}}{S_d/\sqrt{n_d}}$$

onde \bar{d} e S_d é a média e o desvio padrão amostral da variável d , respectivamente.

Neste caso, T_{obs} tem distribuição t-Student com $(n - 1)$ graus de liberdade, com isso, o valor de $t_{(gl;\alpha)}$ é obtido na tabela desta mesma distribuição, e com isso, rejeita-se H_0 se T_{obs} cair na região de rejeição.

Exemplo 8.6. A tabela abaixo apresenta a Pressão sistólica em mmHg em 10 mulheres que não usavam Contraceptivo Oral (OC) no início da pesquisa (Antes) e quando usavam OC (Depois). Faça um teste para a diferença entre as médias de antes e depois e veja se a pressão sistólica é alterada após o tratamento. Adote $\alpha = 5\%$.

Paciente	Antes	Depois	$d_i = \text{Depois} - \text{Antes}$
1	115	128	13
2	112	115	3
3	107	106	-1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	-2
10	115	117	2
Total	1156	1204	48

Com isso, temos que

Variáveis: Pressão Arterial Sistólica

Grupos:

I - PAS antes do tratamento

II - PAS após o tratamento com lisinopril

μ_1 : Média da PAS no grupo I

μ_2 : Média da PAS no grupo II

Hipótese Estatísticas:

$$\begin{cases} H_0 & : d = 0 \\ H_1 & : d \neq 0 \end{cases}$$

Dos dados temos que $\bar{d} = 4,8$, $S_d = 4,566$ e $n = 10$. Logo,

$$\begin{aligned} T_{obs} &= \frac{\bar{d}}{S_d/\sqrt{n_d}} \\ &= \frac{4,8}{4,566/\sqrt{10}} \\ &= 3,32 \end{aligned}$$

Pela tabela da distribuição *t*-Student com $gl = 10 - 1 = 9$ e $\alpha/2 = 0,025$ temos que $t_{(gl;\alpha/2)} = 2,262$.

Como $T_{obs} > t_{(gl;\alpha/2)}$, ou seja, $3,32 > 2,262$, rejeitamos a hipótese nula, ou seja, temos evidências para dizer que a pressão sistólica é alterada após o uso de Contraceptivo Oral.

Exercício 8.23. Deseja-se avaliar a efetividade de uma dieta combinada com um programa de exercícios físicos na redução do nível sérico de colesterol. A tabela abaixo mostra os níveis de colesterol de 12 participantes no início e no final do programa.

Programa	Níveis de Colesterol
Início (x_1)	201; 231; 221; 260; 228; 237; 326; 235; 240; 267; 284; 201
Final (x_2)	200; 236; 216; 233; 224; 216; 296; 195; 207; 247; 210; 209

Realize um teste para verificar se o colesterol diminui após a dieta combinada com o programa de exercícios. Adote $\alpha = 5\%$.

Exercício 8.24. Dez cobaias foram submetidas ao tratamento de engorda com certa ração. Os pesos em gramas, antes e após o teste são dados a seguir (supõe-se que provenham de distribuições normais). A 1% de significância, podemos concluir que o uso da ração contribuiu para o aumento do peso médio dos animais?

Programa	Níveis de Colesterol
Antes	635; 704; 662; 560; 603; 745; 698; 575; 633; 669
Depois	640; 712; 681; 558; 610; 740; 707; 585; 635; 682

Exercício 8.25. Em um estudo de Terapia para vítimas de violação, um grupo, nesta experiência, recebeu aconselhamento de suporte. Para isto, mediram-se os sintomas de desordem de stress pós-traumáticos (PTSD) antes e depois da terapia. A tabela abaixo apresenta os valores obtidos no estudo. Faça um teste, com nível de significância de 0,05, e verifique se a mudança observada é suficiente para que a diferença possa ser considerada estatisticamente significativa?

Situação	Terapia para a PTSD
Antes	21; 24; 21; 26; 32; 27; 21; 25; 18
Depois	15; 15; 17; 20; 17; 20; 8; 19; 10

Exercício 8.26. Um método para avaliar a efetividade de uma droga é observar sua concentração em amostras de sangue ou urina em certos períodos de tempo após seu uso. Suponha que desejaríamos comparar a concentração de dois tipos de aspirinas (tipo A e B) na urina da mesma pessoa, 1 hora após ela ter tomado a droga. Uma dosagem específica da aspirina A é ministrada e, em seguida, é medida sua concentração na urina. Uma semana depois, após a primeira aspirina ser presumidamente eliminada do organismo, uma dosagem da aspirina B é ministrada na mesma pessoa e sua concentração na urina é medida. Os resultados desse experimento são apresentados na tabela a seguir

Aspirina	Concentração (mg %)
A	15; 26; 13; 28; 17; 20; 7; 36; 12; 18
B	13; 20; 10; 21; 17; 22; 5; 30; 7; 11

Pode-se afirmar que a média de proteína urinária sofreu alterações?

Exercício 8.27. O processo de cura de presunto inclui a imersão da peça numa solução de ácido sórbico. Numa fábrica de presunto registaram-se os resíduos de ácido sórbico, em partes por milhão, em 8 peças de presunto imediatamente depois de estas serem imersas na solução, e depois de 60 dias de cura:

Período	Resíduos de ácido sórbico
Antes da cura	224; 270; 400; 444; 590; 660; 1400; 680
Após 60 dias de cura	116; 96; 239; 329; 437; 597; 689; 576

Para um nível crítico de $\alpha = 5\%$ verifique se o processo de cura diminui o número de resíduos de ácido sórbico.

8.5 Introdução ao Teste Qui-Quadrado

Nesta seção, vamos apresentar uma introdução sobre o teste qui-quadrado que pode ser utilizado para verificar se duas variáveis podem ser independentes e/ou se duas variáveis podem ser consideradas

homogêneas. Neste sentido, precisamos entender quais são as hipóteses utilizadas em cada uma destas utilizações e quais as decisões que podem ser tomadas. O teste qui-quadrado é muito utilizado nas áreas de humanas e saúde, ou em áreas que trabalham com dados qualitativos.

O teste qui-quadrado (denotado por χ^2) é uma estatística criada por Karl Pearson em 1899. Seu principal objetivo é comparar frequências observadas e frequências esperadas. As frequências observadas surgem é claro da observação prática, já as frequências esperadas surgem da hipótese nula H_0 que está sendo considerada. O teste qui-quadrado pode ser desenvolvido para verificar a aderência de uma função de distribuição aos dados analisados, para verificar a independência e a homogeneidade entre duas variáveis. Aqui, vamos trabalhar com as duas últimas situações.

8.5.1 O Teste qui-quadrado de independência

É utilizado para verificar se duas variáveis possuem ou não uma relação de dependência uma com a outra. Analisemos alguns exemplos abaixo para entendermos melhor esta definição.

Exemplo 8.7. *Uma instituição financeira no período de 2006 a 2011 realizou um total de 230 parcelamentos, sendo que 136 foram para o sexo masculino e 94 para o sexo feminino. Deste total, apenas 92 foram pagos integralmente no primeiro acordo. A proporção foi a seguinte: 52 do sexo masculino e 40 do sexo feminino. O gerente da instituição deseja saber se realizar acordo e cumprir totalmente o acordo está relacionado com o sexo. A tabela 8.1 resume os valores envolvidos.*

Tabela 8.1:

Sexo	Acordos		Total
	Realizado (R)	Cumprido (C)	
Masculino	136	52	188
Feminino	94	40	134
Total	230	92	322

Vamos desenvolver um teste ao nível de 5%. As hipóteses a serem testadas são:

$$\begin{cases} H_0 & : \text{ Realizar e cumprir acordo independe do sexo} \\ H_1 & : \text{ Realizar e cumprir acordo depende do sexo} \end{cases}$$

Os dados experimentais podem sofrer variações de amostra para amostra, sendo assim, para podermos avaliar as diferenças entre as frequências observadas e as frequências esperadas, logo, a estatística de teste utilizada no teste qui quadrado é dada pela expressão 8.1.

$$\chi_o^2 = \frac{(f_o - f_e)^2}{f_e} \quad (8.1)$$

onde, f_o representa a frequência observada e f_e a frequência esperada.

Para calcularmos as frequências esperadas, vamos utilizar a expressão 8.2.

$$f_e = \frac{t_l * t_c}{t_g} \quad (8.2)$$

onde t_l representa o total da linha, t_c representa o total da coluna e t_g representa o total geral.

Com isso, apresentamos na tabela 8.2 os dados da tabela 8.1 acrescidos dos respectivos valores esperados. Na sequência, apresentamos a tabela 8.3 que auxiliará no cálculo do qui-quadrado.

Tabela 8.2:

Sexo	Acordos		Total
	Realizado (R)	Cumprido (C)	
Masculino	Observado = 136	Observado = 52	188
	Esperado = $(188*230)/322 = 134$	Esperado = $(188*92)/322 = 54$	
Feminino	Observado = 94	Observado = 40	134
	Esperado = $(94*230)/322 = 67$	Esperado = $(94*92)/322 = 27$	
Total	230	92	322

Tabela 8.3:

Observado (f_o)	Esperado (f_e)	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
136	134	2	4	0,0298
52	54	-2	4	0,0741
94	96	-2	4	0,0417
40	38	2	4	0,1053
322	322	0		$\chi_o^2 = 0,2509$

Para o teste de independência, o grau de liberdade é dado pelo produto do número de linhas (L) menos um e pelo número de colunas (C) menos um, isto na tabela inicial, assim: $GLIB=(L-1)(C-1)$. Note que a parte principal da tabela possui duas linhas (cada uma referente a um sexo) e duas colunas (a primeira referente a acordo realizado e a segunda acordo cumprido), sendo assim o grau de liberdade é dado por: $GLIB = (2-1)*(2-1) = 1$.

Com $\alpha = 5\%$ e $GLIB=1$, temos que, o qui-quadrado tabelado (χ_T^2), obtido pela tabela da distribuição Qui-Quadrado, é igual a 3,841. Portanto, como $\chi_o^2 < \chi_T^2$ dizemos que, não temos evidências para rejeitar H_0 ao nível de 5%, ou seja, as variáveis realizar e cumprir acordo independe do sexo.

Exemplo 8.8. Suponha que um distribuidor de água mineral deseja saber se há dependência no consumo de seus produtos e a região da cidade. A tabela 8.4 apresenta a frequência de galões de 20 litros vendidos conforme a região da cidade.

Tabela 8.4: Venda semanal - Galões de 20 l de água vendida versus região atendida

Marcas	Regiões atendidas				Total
	Leste	Oeste	Norte	Sul	
Pura	700	350	250	200	1500
Saborosa	600	300	200	400	1500
Refrescante	450	350	200	100	1100
Mariah	550	250	470	130	1400
Total	2300	1250	1120	830	5500

Vamos desenvolver um teste ao nível de 5%. As hipóteses a serem testadas são:

$$\begin{cases} H_0 & : \text{A preferência pela marca independe da região} \\ H_1 & : \text{A preferência pela marca depende da região} \end{cases}$$

Com o auxílio da expressão 8.2, montamos a tabela 8.5. Na sequência, apresentamos a tabela 8.6 que auxiliará no cálculo do qui-quadrado.

Tabela 8.5: Venda semanal - Galões de 20 l de água vendida versus região atendida

Marcas	Regiões atendidas				Total
	Leste	Oeste	Norte	Sul	
Pura	$f_o = 700$	$f_o = 350$	$f_o = 250$	$f_o = 200$	1500
	$f_e = 627$	$f_e = 341$	$f_e = 305$	$f_e = 226$	
Saborosa	$f_o = 600$	$f_o = 300$	$f_o = 200$	$f_o = 400$	1500
	$f_e = 627$	$f_e = 341$	$f_e = 305$	$f_e = 226$	
Refrescante	$f_o = 450$	$f_o = 350$	$f_o = 200$	$f_o = 100$	1100
	$f_e = 460$	$f_e = 250$	$f_e = 224$	$f_e = 166$	
Mariah	$f_o = 550$	$f_o = 250$	$f_o = 470$	$f_o = 130$	1400
	$f_e = 585$	$f_e = 318$	$f_e = 285$	$f_e = 211$	
Total	2300	1250	1120	830	5500

Com $\alpha = 5\%$ e $GLIB = (4 - 1) * (4 - 1) = 9$ temos que $\chi_T^2 = 16,9190$, e com isso, como $\chi_o^2 > \chi_T^2$ rejeitamos H_0 , ou seja, a preferência pela marca dependerá da região.

Tabela 8.6: Cálculo do Qui-quadrado

Observado (f_o)	Esperado (f_e)	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
700	627	73	5329	8,5
600	627	-27	729	1,16
450	460	-10	100	0,22
550	585	-35	1225	2,09
350	341	9	81	0,24
300	341	-41	1681	4,93
350	250	100	10000	40
250	318	-68	4624	14,54
250	305	-55	3025	9,92
200	305	-105	11025	36,15
200	224	-24	576	2,57
470	285	185	34225	120,09
200	226	-26	676	2,99
400	226	174	30276	133,96
100	166	-66	4356	26,24
130	211	-81	6561	31,09
				$\chi_o^2 = 434,69$

8.5.2 O Teste qui-quadrado de homogeneidade

Utilizamos o teste qui-quadrado de homogeneidade para verificar se diferentes populações têm as mesmas características, enquanto que no teste qui-quadrado de independência o objetivo era verificar se duas variáveis possuíam ou não uma relação de dependência uma com a outra. Para ilustrarmos este teste, veremos os exemplos a seguir.

Exemplo 8.9. *Suponha que tenhamos razões para crer que as notas obtidas por estudantes de escolas públicas, no exame vestibular para uma Universidade, sejam menores que as notas obtidas por estudantes de escolas particulares. Para testar essa hipótese, foram selecionadas duas amostras de estudantes que prestaram o vestibular, a tabela 8.7 apresenta a frequência observada para 4 intervalos de médias pré-definidas.*

Neste exemplo devemos testar se as duas populações (alunos de escola particular e pública) são homogêneas em relação aos intervalos de médias pré-definidos. Desta forma, as hipóteses a serem

Tabela 8.7: Número de alunos das escola Pública e Particular por faixa de notas

Escola	Faixa de notas				Total
	(0; 2,5]	(2,5; 5,0]	(5,0; 7,5]	(7,5; 10,0]	
Pública	15	22	18	3	58
Particular	6	10	20	6	42
Total	21	32	38	9	100

testadas são:

$$\begin{cases} H_0 & : \text{Notas de alunos da Escola Pública} = \text{Notas de alunos da Escola Particular} \\ H_1 & : \text{Notas de alunos da Escola Pública} \neq \text{Notas de alunos da Escola Particular} \end{cases}$$

A estatística de teste, ou seja, a expressão matemática utilizada no teste é a mesma expressão 8.1 utilizada no teste de independência, além disso, para calcularmos os valores esperados, vamos utilizar a expressão 8.2, também utilizada no teste de independência.

Com o auxílio da expressão 8.2, montamos a tabela 8.8. Na sequência, apresentamos a tabela 8.9 que auxiliará no cálculo do qui-quadrado.

Tabela 8.8: Número de alunos das escola Pública e Particular por faixa de notas

Escola	Faixa de notas				Total
	(0; 2,5]	(2,5; 5,0]	(5,0; 7,5]	(7,5; 10,0]	
Pública	$f_o = 15$	$f_o = 22$	$f_o = 18$	$f_o = 3$	58
	$f_e = 12$	$f_e = 19$	$f_e = 22$	$f_e = 5$	
Particular	$f_o = 6$	$f_o = 10$	$f_o = 20$	$f_o = 6$	42
	$f_e = 9$	$f_e = 13$	$f_e = 16$	$f_e = 4$	
Total	21	32	38	9	100

Com $GLIB = (2 - 1) * (4 - 1) = 3$ e $\alpha = 1\%$, temos que $\chi_T^2 = 11,345$, e com isso, como $\chi_o^2 < \chi_T^2$, não temos evidências para rejeitar H_0 , ou seja, as notas dos alunos da escola pública e particular são homogêneas.

Tabela 8.9: Cálculo do Qui-quadrado

Observado (f_o)	Esperado (f_e)	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
15	12	3	9	0,7500
22	19	3	9	0,4737
18	22	-4	16	0,7273
3	5	-2	4	0,8000
6	9	-3	9	1,0000
10	13	-3	9	0,6923
20	16	4	16	1,0000
6	4	2	4	1,0000
100	100	0		$\chi_o^2 = 6,4433$

Exemplo 8.10. Um epidemiologista resolve fazer uma pesquisa para ver como o consumo de fruta está associado ao câncer, para isso, foi selecionada uma amostra de 300 pessoas, sendo que metade tinha câncer e a outra metade não. Foi preparado um questionário cuja pergunta central era o consumo de fruta, os dados obtidos estão na tabela 8.10.

Tabela 8.10: Consumo de fruta versus câncer

Câncer	Consome Fruta		Total
	Não	Sim	
Sem	85	65	150
Com	95	55	150
Total	180	120	300

Neste exemplo queremos testar se as duas populações são homogêneas, para o nível de significância $\alpha = 1\%$. Para isso, estabelecemos as seguintes hipóteses:

$$\begin{cases} H_0 & : \text{O consumo de fruta é o mesmo em pessoas com e sem câncer} \\ H_1 & : \text{O consumo de fruta é diferente entre pessoas com e sem câncer} \end{cases}$$

Com o auxílio da expressão 8.2, montamos a tabela 8.11. Na sequência, apresentamos a tabela 8.12 que auxiliará no cálculo do qui-quadrado.

O valor de $GLIB = (2 - 1) * (2 - 1) = 1$. Além disso, pela tabela da distribuição Qui-quadrado e considerando $\alpha = 1\%$, temos que $\chi_T^2 = 6,635$. Portanto, $\chi_o^2 < \chi_T^2$ e neste caso, não temos evidências para rejeitar H_0 , ou seja, o consumo de fruta é o mesmo em pessoas com e sem câncer.

Tabela 8.11: Consumo de fruta versus câncer

Câncer	Consome Fruta		Total
	Não	Sim	
Sem	$f_o = 85$	$f_o = 65$	150
	$f_e = 90$	$f_e = 60$	
Com	$f_o = 95$	$f_o = 55$	150
	$f_e = 90$	$f_e = 60$	
Total	180	120	300

Tabela 8.12: Cálculo do Qui-quadrado

Observado (f_o)	Esperado (f_e)	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
85	90	-5	25	0,2777
95	90	5	25	0,2777
65	60	5	25	0,4166
55	60	-5	25	0,4166
300	300	0		$\chi_o^2 = 1,3886$

8.5.3 Tabela da distribuição qui-quadrado

A tabela 8.13 apresenta os valores da distribuição qui-quadrado para os valores de α igual a 1% e 5%.

Tabela 8.13: Tabela da distribuição qui-quadrado

GRAUS DE LIBERDADE	α	
	1%	5%
1	6,635	3,841
2	9,210	5,991
3	11,345	7,815
4	13,277	9,488
5	15,086	11,070
9	21,6660	16,9190
10	23,209	18,307
20	37,566	31,410
30	50,892	43,773

8.5.4 Exercícios

Exercício 8.28. *Verifique se existe dependência entre a renda e o número de filhos em famílias de uma cidade. Foram selecionadas de modo aleatório 250 e os resultados estão apresentados na tabela 8.14. Utilize $\alpha = 5\%$.*

Tabela 8.14: Renda em função do número de filhos

Renda (R\$)	Número de filhos				
	0	1	2	+ de 2	Total
menos de 2000	15	27	50	43	135
2000 a 5000	25	30	12	8	75
5000 ou mais	8	13	9	10	40
Total	48	70	71	61	250

Exercício 8.29. *Considere 1237 indivíduos adultos classificados segundo a pressão sangüínea (mm Hg) e o nível de colesterol (mg/100cm³). Dados apresentados na tabela 8.15. Verificar se existe independência entre essas variáveis. Utilize $\alpha = 5\%$.*

Tabela 8.15: Colesterol em função da pressão

Colesterol	Pressão			
	< 127	127 a 166	> 166	Total
< 200	117	168	22	307
200 a 260	204	418	63	685
> 260	67	145	33	245
Total	388	731	118	1237

Exercício 8.30. Um pesquisador deseja verificar se há dependência no consumo de seus chocolates e as cidades de sua região. Dados apresentados na tabela 8.16. Verificar se existe independência entre essas variáveis. Utilize $\alpha = 5\%$.

Tabela 8.16: Sabor do chocolate em função das cidades					
Sabor do chocolate	Cidades do Vale do Taquari				Total
	Lajeado	Santa Cruz	Estrela	Taquari	
Chocolate com caju	60	30	20	40	150
Chocolate com amendoim	45	35	20	10	110
Chocolate com flocos	55	25	47	13	140
Chocolate com passas	70	35	25	20	150
Total	230	125	112	83	550

Exercício 8.31. Um pesquisador deseja verificar se as variáveis marca de cigarro e sexo do fumante são dependentes. Para isso, colheu uma amostra de 200 fumantes (homens e mulheres) e os classificou em função de três marcas de cigarro: A, B e C. Dados apresentados na tabela 8.17. Verificar se existe independência entre essas variáveis. ($\alpha = 5\%$).

Tabela 8.17: Marca do cigarro em função do sexo

Sexo	Marca do cigarro			
	A	B	C	Total
Masc. (M)	20	70	30	120
Fem. (F)	40	15	25	80
Total	60	85	55	200

Exercício 8.32. Uma pesquisa com 674 pessoas avalia a relação entre a categoria socioprofissional (CSP) e a principal fonte de informação sobre os problemas ambientais. A tabela 8.18 apresenta os dados. Verifique a CSP é homogênea em relação as fontes de informação. ($\alpha = 5\%$).

Tabela 8.18: Categoria Socioprofissional em função da fonte de informação

Categoria	Fonte de informação					Total
	TV	Jornal	Rádio	Livros	Autarquia	
Agricultor	26	18	9	5	6	64
Quadro superior	19	49	4	16	3	91
Quadro médio	44	87	4	39	3	177
Operário	181	107	16	31	7	342
Total	270	261	33	91	19	674

Exercício 8.33. Foi realizado um inquérito sobre o consumo de vitaminas. Os resultados, expressos consoante o nível cultural dos indivíduos (em número de anos de escolaridade), são apresentados na tabela 8.19. Verifique se a frequência de consumo de vitaminas é homogênea em relação ao número de anos de escolaridade. ($\alpha = 5\%$).

Tabela 8.19: Consumo de vitamina em função dos anos de escolaridade

Consumo de	Anos de escolaridade				Total
	< 7	de 7 a < 12	de 12 a < 15	≥ 15	
Vitaminas					
Sim	46	392	290	249	977
Não	461	3305	1738	1059	6563
Total	507	3697	2028	1308	7540

Exercício 8.34. Um estudo nacional sobre patologia hepática deu os seguintes resultados: em 1995, houve 13607 hospitalizações das quais 8088 por hepatite alcoólica e 2772 por hepatite não alcoólica, o resto sendo por outras formas de hepatite. Em contrapartida, das 17310 hospitalizações feitas em 2001, 10428 foram por hepatite alcoólica e 2956 por outras formas. Verifique se a repartição dos diversos tipos de hepatite são homogêneas com relação aos dois anos analisados. ($\alpha = 5\%$).

8.5.5 Exercícios Propostos

Exercício 8.35. Verifique se existe independência entre o nível de renda e a opinião de eleitores referente a reforma da lei tributária. Mil eleitores em uma amostra aleatória são classificados como eleitores de baixa, média ou alta renda e se são a favor ou não da reforma. Os resultados estão apresentados na tabela 8.20. Utilize $\alpha = 5\%$.

Tabela 8.20: Nível de Renda em função da opinião de eleitores

Reforma dos impostos	Nível de Renda			
	Baixa	Média	Alta	Total
A favor	182	213	203	598
Contra	154	138	110	402
Total	336	351	313	1000

Exercício 8.36. Uma amostra aleatória de 500 eleitores foi selecionada segundo a afiliação política participante (Democratas, Republicanos e Independentes) e suas opiniões classificadas como a favor, contra e indecisos com relação a lei do aborto. Diante disso, verifique se as opiniões sobre a lei do aborto são as mesmas (homogêneas) dentro de cada afiliação política. Os resultados estão apresentados na tabela 8.21. Utilize $\alpha = 5\%$.

Tabela 8.21: Opinião em função da afiliação política

Lei do Aborto	Afiliação Política			
	Democratas	Republicanos	Independentes	Total
A favor	82	70	62	214
Contra	93	62	67	222
Indecisos	25	18	21	64
Total	200	150	150	500

Exercício 8.37. Um estudo realizado em uma fábrica está interessado em verificar se o número de itens produzidos com defeitos é o mesmo com relação ao turno de fabricação, neste caso, manhã, tarde e noite. Para isto, foram analisados 2835 itens produzidos, classificados em defeituosos ou não, durante um dia de produção. Os resultados estão apresentados na tabela 8.22. Utilize $\alpha = 5\%$.

Tabela 8.22: Classificação dos itens em função do turno de produção

Classificação dos itens	Turno de fabricação			
	Manhã	Tarde	Noite	Total
Defeituosos	45	55	70	170
Não Defeituosos	905	890	870	2665
Total	950	945	940	2835

Exercício 8.38. Uma amostra aleatória de 90 adultos é classificada de acordo com o gênero e o número de horas que eles assistem à televisão durante a semana. Os resultados estão apresentados na tabela 8.23. Verifique se o tempo gasto com televisão é independente do gênero do telespectador. Utilize $\alpha = 5\%$.

Tabela 8.23: Gênero em função do tempo assistindo televisão

Tempo	Gênero	
	Masculino	Feminino
Mais de 25 horas	15	29
Menos de 25 horas	27	19

Exercício 8.39. *Uma amostra aleatória de 200 homens casados, todos aposentados, foi classificada de acordo com o nível educacional e o número de filhos. Verifique se o tamanho da família é independente do nível de educação obtido pelo pai. Os resultados estão apresentados na tabela 8.24. Utilize $\alpha = 5\%$.*

Tabela 8.24: Nível educacional em função do número de filhos

Nível Educacional	Número de filhos		
	0-1	2-3	acima de 3
Elementar	14	37	32
Médio	19	42	17
Superior	12	17	10

Exercício 8.40. *Uma escola deseja verificar se a opinião dos pais, alunos, professores e comunidade local é a mesma com relação ao hábito de cantar o hino nacional todos os dias antes do início das aulas. Os resultados estão apresentados na tabela 8.25. Utilize $\alpha = 5\%$.*

Tabela 8.25: Opinião em função da categoria

Opinião	Categoria			
	Pais	Alunos	Professores	Comunidade local
A favor	65	30	40	34
Contra	42	66	33	42
Sem opinião	93	54	27	24

Exercício 8.41. *Crie um exemplo de aplicação para realizar um teste de independência ou homogeneidade. A tabela de dados deve ter 3 linhas e 3 colunas. Realize o teste ao nível de 5%.*