Introdução ao Estudo de Probabilidade e Estatística com auxílio do software R

Book · August 2016

CITATIONS
0

READS
4,399

3 authors:

Matheus Henrique Dal Molin Ribeiro
Federal University of Technology - Paraná/Brazil (UTFPR), Pato Branco, Brazil
18 PUBLICATIONS **22** CITATIONS
SEE PROFILE

João Victor do Pilar
Federal University of Santa Catarina
1 PUBLICATION **0** CITATIONS
SEE PROFILE

José Donizetti de Lima
Federal University of Technology - Paraná/Brazil (UTFPR)
79 PUBLICATIONS **64** CITATIONS
SEE PROFILE

Universidade Tecnológica Federal do Paraná
Câmpus Pato Branco
Matheus Henrique Dal Molin Ribeiro
João Victor do Pilar
José Donizetti de Lima

Introdução ao Estudo de Probabilidade e Estatística com auxílio do *software* R

Recursos Educacionais Abertos

Agosto de 2016

Sumário

Prefácio

A motivação para elaboração deste material, intitulado **Introdução ao Estudo de Probabilidade e Estatística com auxílio do *software* R**, está relacionada a necessidade de construção de um roteiro de estudos que seja norteador as práticas pedagógicas dos docentes que trabalham disciplinas de Probabilidade e Estatística. Nesse sentido, com a proposição e contemplação do Edital nº 015/2015 – PROGRAD - UTFPR, "Recurso Educacional Aberto produzido com fomento do Programa de Bolsas para o desenvolvimento de Recursos Educadionais Abertos (PIBEA) por meio do Programa de Bolsas de Fomento às Ações de Graduação da UTFPR"essa construção foi possível.

Os responsáveis pela elaboração e revisão deste material são os professores Matheus Henrique Dal Molin Ribeiro e José Donizetti de Lima lotados no Departamento de Matemática (DAMAT), Câmpus Pato Branco. Também participou efetivamente da construção o discente João Victor do Pilar, acadêmico do curso de Engenharia de Computação.

Algumas falhas sempre ocorrem. Nesse sentido, qualquer inconsistência nas informações apresentadas, por gentileza, entrar em contato pelo endereço de e-mail mribeiro@utfpr.edu.br

Esperamos que o texto redigido contribua para o enriquecimento do processo de construção do conhecimento e que o leitor consiga desenvolver suas potencialidades relacionadas aos tópicos abordados.

Os autores

Resumo

Esta apostila contempla a compilação de informações e exercícios baseados nas principais bibliografias, das disciplinas de Probabilidade e Estatística ofertadas na Universidade Tecnológica Federal do Paraná, Câmpus Pato Branco, tais como Barbetta, Reis e Bornia (2010), Devore (2006), Morettin (2010), Loesch (2012), Magalhães (2011) e Mello e Peternelli (2013) utilizadas nas disciplinas dessa universidade com resolução de atividades utilizando o *software* R, versão 3.3.2. No primeiro capítulo é apresentado o *software* R, utilizando a interface Studio, desde o processo para sua instalação até a exposição de diversos exemplos necessários para compreensão de como manipular informações por meio desse software. Já nos capítulos 2, 3 e 4 são contemplados conceitos de amostragem, organização de dados e medidas descritivas, sempre que possível com resolução de exemplos no software mencionado. Nos capítulos 5 e 6 são apresentados conceitos de probabilidade, desde a ideia de variável aleatória, até o conceito de diferentes distribuições de probabilidade discretas e contínuas. No capítulo 7 são apresentados tópicos inerentes ao estudo de intervalos de confiança e testes de hipóteses. Por fim, nos capítulos 8 e 9 são apresentados conceitos para a compreensão dos tópicos de Análise de Variância e Regressão Linear Simples. Todos os capítulos são finalizados com uma série de exercícios, para que o leitor possa praticar as teorias apresentadas.

Palavras-chaves: Estatística, probabilidade, *software* R.

Abstract

This apostille includes the compilation of information and exercises based on the main bibliographies of Probability and Statistics disciplines offered at the Federal Technological University of Paraná, Pato Branco, such as Barbetta, Reis & Bornia (2010), Devore (2006), Morettin (2010), Loesch (2012), Magalhães (2011) and Mello & Peternelli (2013) used in the disciplines the university with activities resolution using the software R, version 3.3.2. The first chapter presents the software R, using the Studio interface, since the process for its installation to display several examples needed to understand how to manipulate information through this software. Already in Chapters 2, 3 and 4 are included sampling concepts, data organization and descriptive measures, where possible with examples of the resolution mentioned software. In chapters 5 and 6 are presented probability concepts, since the idea of random variable, to the concept of different distributions of discrete and continuous probability. Chapter 7 presents themes inherent in the study of confidence intervals and hypothesis testing. Finally, in chapters 7 and 8 are presented concepts for the understanding of the topics of ANOVA and simple linear regression. All the chapters are completed with a series of exercises, where the reader can practice the theories presented.

Keywords: Statistics, probability, software R.

Capítulo 1

Conhecendo o R - Primeiros Passos

1.1 Contexto Histórico

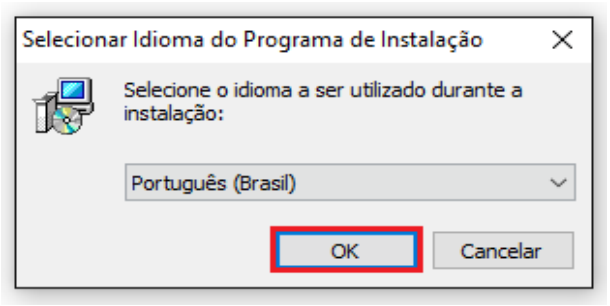
Segundo Mello e Peternelli (2013), o R tem suas raízes em duas linguagens anteriores: a linguagem S e a Scheme. Foi criado como um projeto de pesquisa pelo neozelandês Ross Ihaka e pelo canadense Robert Gentleman. Nos dias atuais está sob constante desenvolvimento por um grupo chamado *R Core Team*.

Vale ressaltar que o R não é apenas um programa estatístico, pois permite operações matemáticas, manipulação de vetores e matrizes, desenvolvimento de gráficos, entre outros. Ao contrário dos softwares proprietários (pagos), ao invés de contar com o suporte técnico oferecido pela empresa que mantém o software, o R conta com a colaboração de milhares de usuários ao redor do mundo (MELLO; PETERNELLI, 2013).

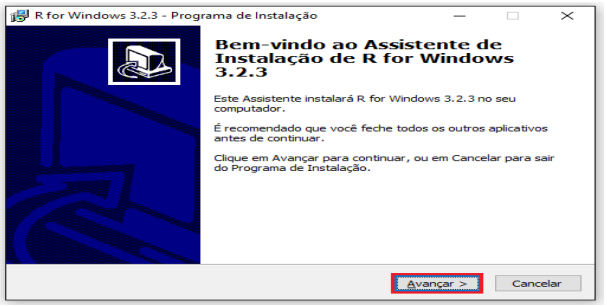
1.2 Instalação do R

O instalador do Software R (R Core Team, 2014) pode ser obtido no endereço: <https://www.r-project.org/>, no qual é apresentado em versões de acordo com os sistemas operacionais: Windows, Linux e MacOS X. Para iniciar o download basta acessar a aba “CRAN”, e escolher o local de disponibilização do programa mais próximo de onde você se encontra. Para o Brasil, existem atualmente cinco opções: Universidade Federal do Paraná (PR), Universidade Estadual de Santa Cruz, Fundação Oswaldo Cruz (RJ), Universidade de São Paulo (SP) e Universidade de São Paulo – Piracicaba (SP).

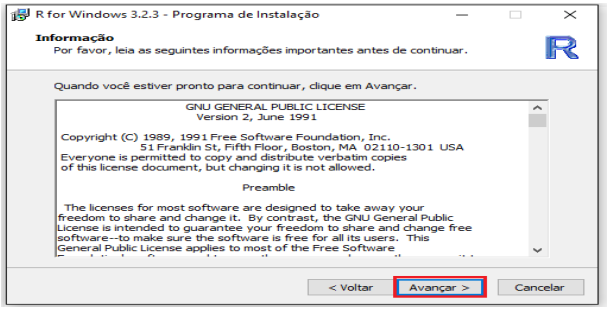
Após escolher o local, em “Download and Install R” selecione o sistema operacional. Para a instalação basta seguir as instruções, conforme Figura 1.



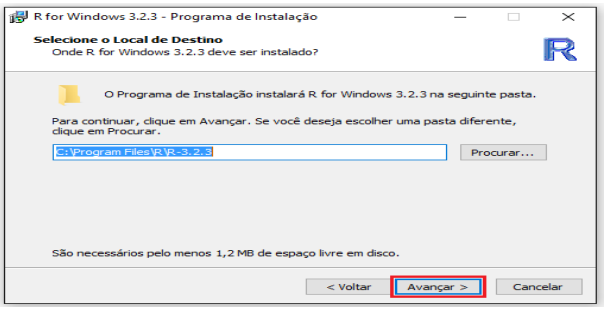
(a) Seleção do Idioma



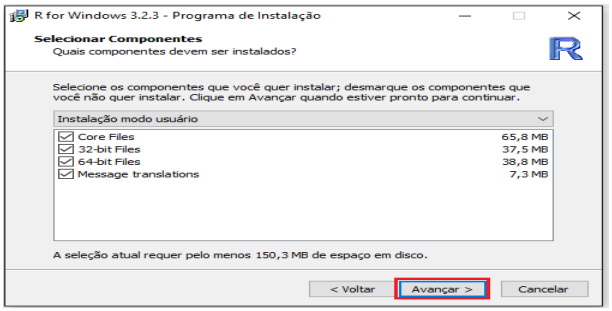
(b) Inicio da instalação



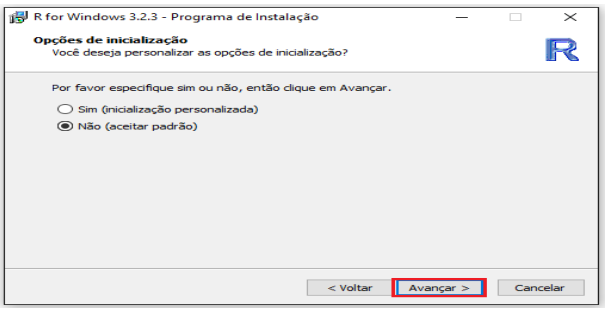
(c) Licença de instalação do .



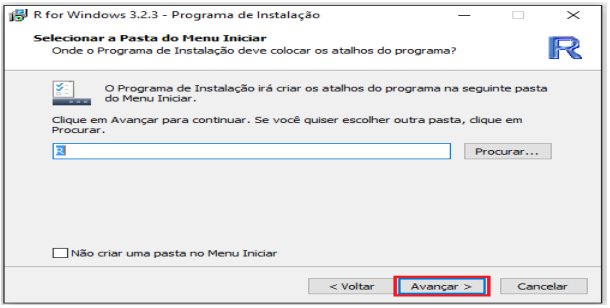
(d) Diretório de instalação



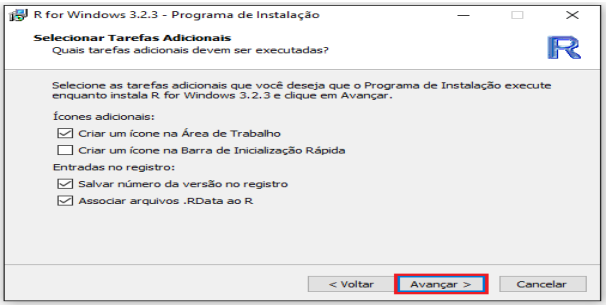
(e) Componentes



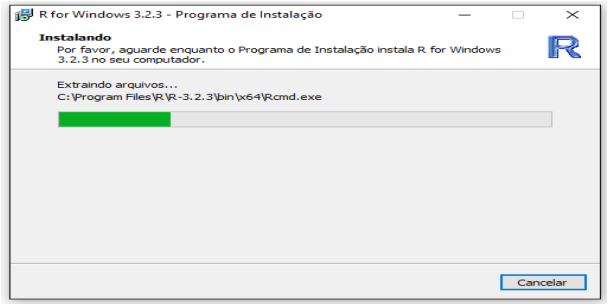
(f) Opções de inicialização



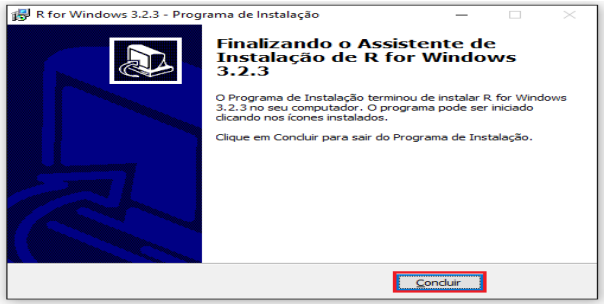
(g) Pasta para criação de atalho



(h) Tarefas que devem ser executadas



(i) Progresso da instalação

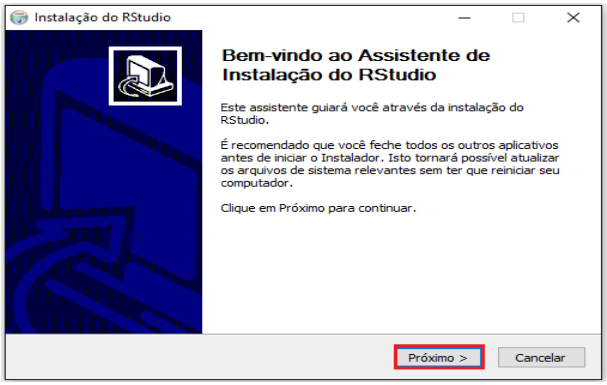


(j) Finalização

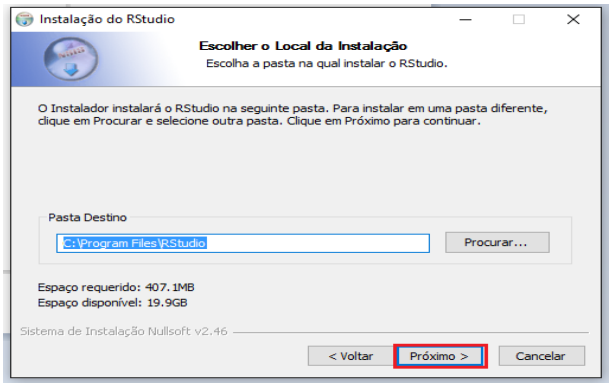
Figura 1 – Assistente de Instalação do R

1.3 Instalação do R-Studio

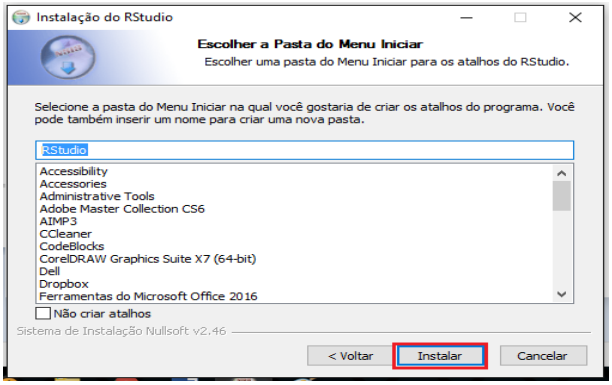
O R-Studio pode ser obtido no endereço: <https://www.rstudio.com/products/rstudio/download/>, no qual é apresentado em versões de acordo com os sistemas operacionais: Windows, Linux e MacOS X. Para iniciar o download, em "Installers for Supported Platforms", basta selecionar o link com a plataforma do computador. Para a instalação basta seguir as instruções, conforme é apresentado na Figura 2.



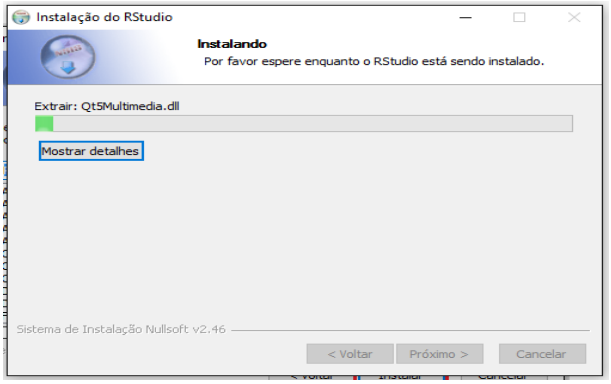
(a) Seleção do Idioma



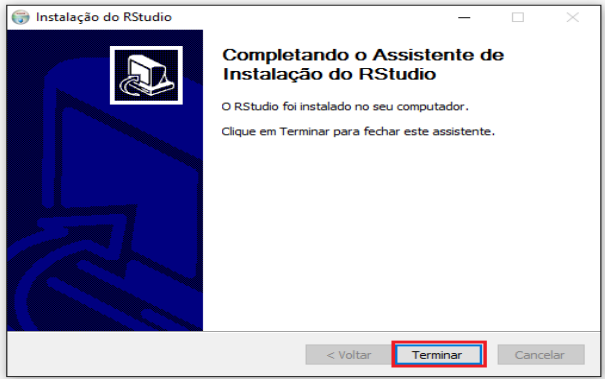
(b) Inicio da instalação



(c) Licença de instalação do R



(d) Diretório de instalação



(e) Componentes

Figura 2 – Assistente de Instalação do R-Studio

1.4 Interface Gráfica

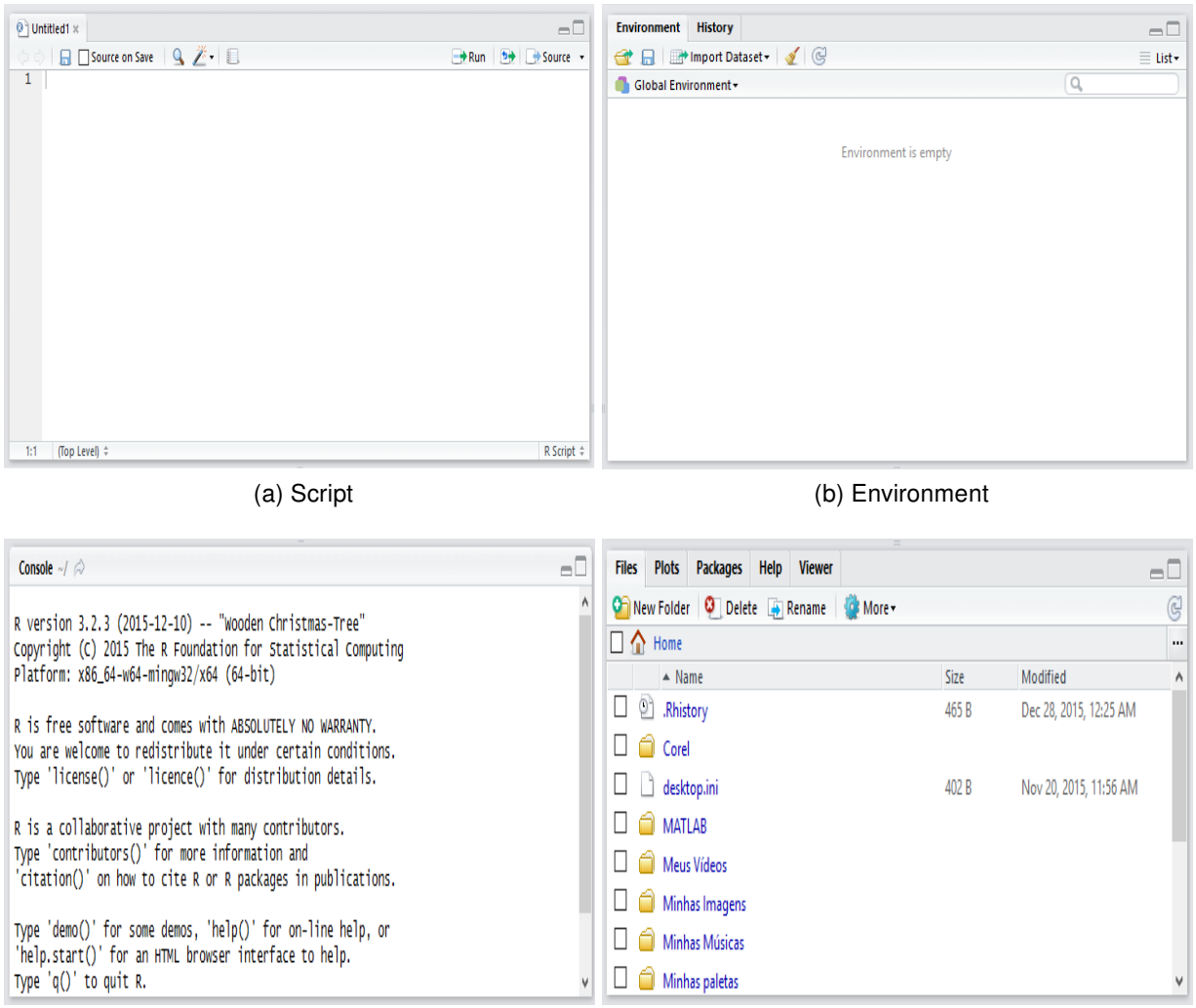
Para compreensão da interface gráfica do R, faremos uma breve descrição das telas que aparecem ao inicializar o programa. São elas:

a) Script: Está localizado no canto superior esquerdo do R-Studio. É o editor de texto no qual os códigos são escritos, e também onde os comentários podem ser colocados, e assim, são salvos.(Figura 3a).

b) Environment: Está localizado no canto superior direito. Os objetos criados e o histórico dos comandos podem ser acessados nessa aba (Figura 3b).

c) Console: Está localizado no canto inferior esquerdo. O console é o local em que os comandos são executados, e também onde aparecem os erros (Figura 3c).

d) Files, Plots, Packages, Help, Viewer: Ficam no canto inferior direito. Você pode explorar pastas e arquivos diretamente do RStudio na aba “Files”. Os gráficos que forem feitos apareceram na aba “Plots”. Os pacotes instalados em sua máquina estão listados em “Packages”. As ajudas das funções aparecem em “Help”. E o “Viewer” serve para visualização de páginas em HTML e JavaScript (Figura 3d).



(a) Script (b) Environment

(c) Console (d) Files, Plots, Packages, Help, Viewer

Figura 3 – R: Interface Gráfica do R-Studio

1.5 Entrada de dados

A entrada de dados no R pode ser desenvolvida de várias formas, entre as quais podemos destacar a importação de tabelas para o programa, bem como a entrada direta dos dados no **script**. Essas duas formas possuem suas diferenças e podem ser utilizadas quando forem pertinentes. Nas seções a seguir serão feitas descrições para essas duas variações da entrada de dados.

1.5.1 Entrada por intermédio de arquivos externos

Se os dados estiverem salvos em arquivos, sob forma de planilhas, tabelas, etc., deve-se fazer com que o R leia estes arquivos, transformando-os em um objeto. Para que o R reconheça o conjunto de dados do arquivo é necessário que as colunas sejam separadas. Caso isso não ocorra o R não conseguirá separar as colunas e emitirá uma

mensagem de erro. Um modo fácil de resolver este problema é salvar a planilha de dados com o formato (**.csv**) que utiliza virgula (**,**) como elemento separador das colunas. Porém, antes de iniciar a entrada de dados no R deve-se alterar a pasta de trabalho padrão em que o arquivo de dados **.csv** será salvo. Para isso, basta utilizar o comando:

```
setwd("Diretorio")
```

Para conferir o diretório atualizado, utiliza-se o comando:

```
getwd()
```

De posse da pasta de trabalho e do arquivo no formato **.csv** no diretório, procederemos com o seguinte comando:

```
dir()
```

Com esse comando o R irá verificar se há algum arquivo na pasta de trabalho. Como previamente havíamos salvo um arquivo **.csv**, sabemos que o R irá encontrar este arquivo no diretório especificado anteriormente.

Em seguida, devemos dar o comando para que o R carregue o arquivo **.csv** no **console** de trabalho. Para isso digite o comando:

```
carregar<-read.table("NomeArquivo.csv",header=T,sep=",",dec=".")
```

- **carregar**: é o objeto no qual os dados lidos serão reconhecidos pelo R
- **=**: sinal que atribui os dados lidos ao objeto carregar
- **read.table**: função que lê o arquivo do tipo **.csv**
- O parâmetro "**header**" nos permite indicar se o arquivo de dados (**data.frame**) tem ou não o nome nas colunas (título) na primeira linha de dados. O parâmetro "**sep**" permite indicar o tipo de separador dos dados presentes no arquivo. Finalmente o parâmetro "**dec**" permite indicar o caractere usado como separador de casas decimais dos números reais.

Caso o arquivo tenha título, podemos verificar o nome desses títulos por meio do comando:

```
names()
```

Podemos ver a dimensão do arquivo carregado por meio do seguinte comando:

```
dim()
```

Isto porque o R, agora, considera o arquivo carregado como uma matriz. Desta forma, podemos localizar linhas, colunas e elementos desta matriz. Para isso, utilizamos os comandos:

```
carregar[1,1]
carregar[1:5,]
```

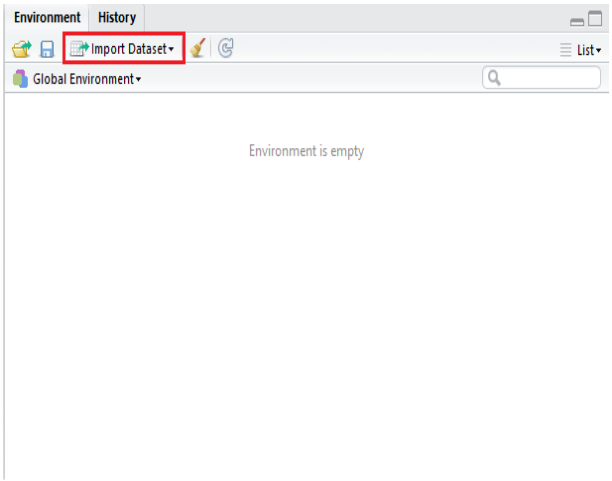
Para ilustrar os conceitos apresentados anteriores, considere o exemplo a seguir:

Exemplo 1 *Exemplo para importação de uma tabela salva em um diretório qualquer.*

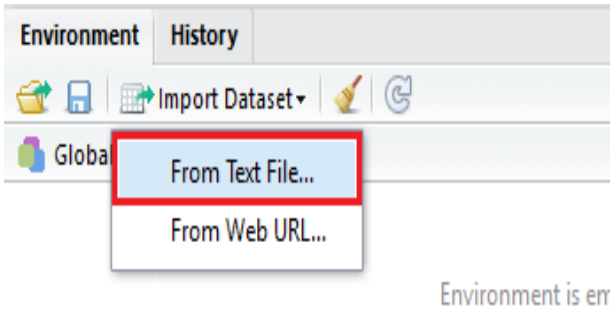
```
dir()
carregar<- read.table(file="TESTE.csv",header=T,dec=".")
carregar
```

Para facilitar, é possível importar uma tabela de dados de um arquivo sem a necessidade de utilizar códigos. Esse procedimento pode ser visualizado no Exemplo 2.

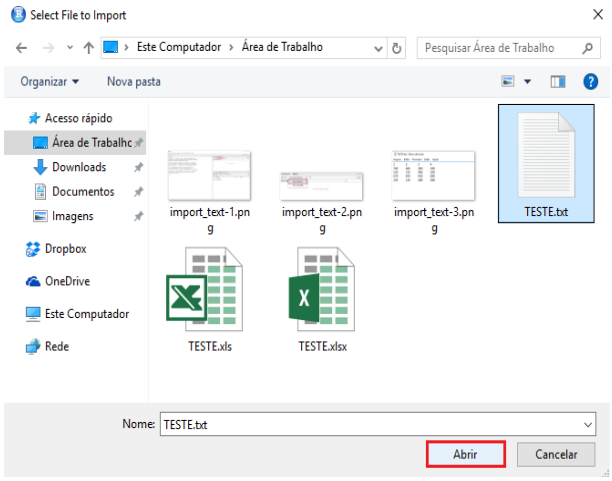
Exemplo 2 *Na tela "Environment", Basta clicar no botão Import Dataset e, em seguida, clicar na opção From Text File. Para finalizar o processo basta encontrar o arquivo desejado. É possível conferir os passos na figura 4.*



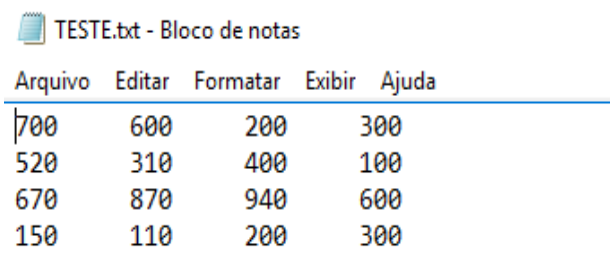
(a) Espaço Environment



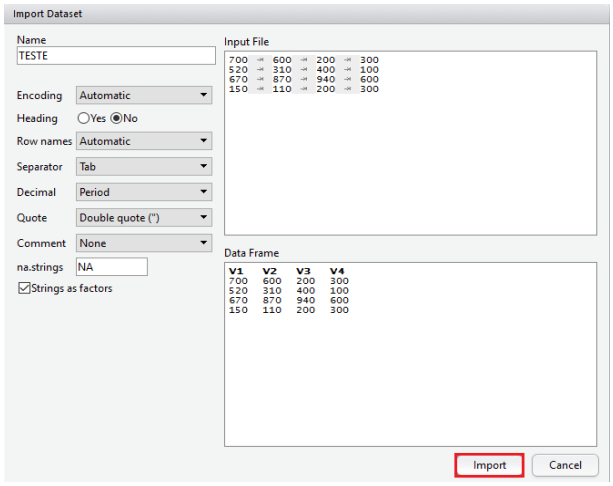
(b) Importar de texto



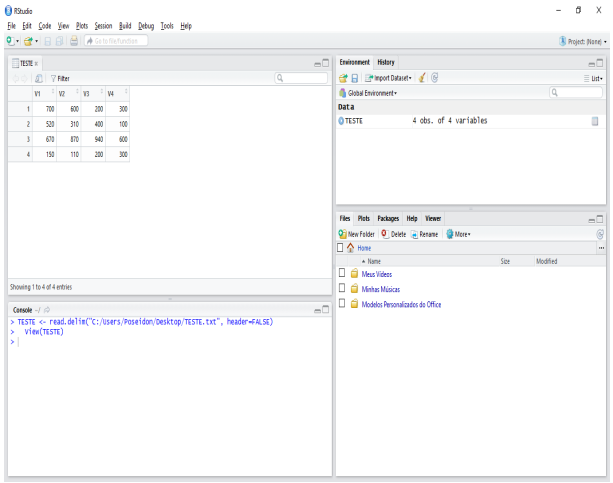
(a) Localização arquivo



(b) Arquivo de texto TEXTE.txt



(c) Visualizar importação



(d) Importação concluida

Figura 4 – Importação de um arquivo .txt

1.5.2 Entrada diretamente por intermédio do Script

Para inserir uma tabela com diversas informações, deve-se criar um data frame, como no exemplo 2:

Exemplo 3 *Exemplo para entrada de dados diretamente no script.*

```
carregar=read.table(header=TRUE, text="
mpg      motor      potencia      peso
18       307          130         3504
15       350          165         3693
18       318          150         3436")

carregar
```

1.6 Objetos

No R existem seis tipos de objetos: números, vetores, matrizes, arrays, data frames e listas, conforme ilustra a Figura 5. Pode-se entender objeto como uma caixinha na qual você pode guardar o que quiser. A partir daí todas as operações matemáticas podem ser feitas usando esses objetos. Isso torna as coisas mais simples. Para criar um objeto é só atribuir um valor a um nome, ou seja, quando se coloca um valor dentro de um objeto, este passa a existir automaticamente. Uma atribuição pode ser feita, basicamente, de duas maneiras, usando o sinal de = ou utilizando uma seta formada pela junção dos sinais de menor que e menos <-. Note que essa seta sempre deve levar o valor ao objeto, ou seja, sempre apontar para o objeto. Portanto, é possível usar a setinha em ambas as direções <- ou ->. Outro sinal muito útil na linguagem R é o sinal de comentário #, ou seja, a partir do sinal, o que for escrito não será interpretado como comando.

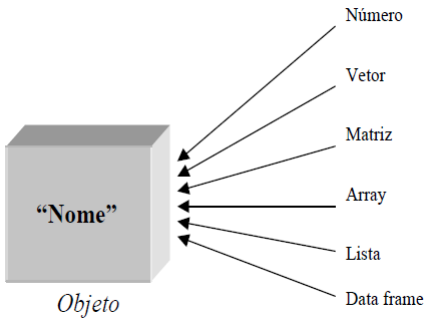


Figura 5 – Esquema de um objeto.

1.6.1 Número

É possível atribuir apenas um número a um objeto. Para verificar quanto vale cada objeto, apenas digite seu nome e tecle enter. O um entre colchetes se refere à primeira posição do vetor, ou seja, um número é entendido como um vetor de uma posição.

```
a<-10
a
```

1.6.2 Vetor

O vetor da linguagem R tem um significado um pouco diferente que o vetor da Matemática. Para o R, um vetor é qualquer conjunto unidimensional de valores. Esses valores podem ser números, strings (palavras) ou valores lógicos (F-falso e T-verdadeiro). Para se atribuir um conjunto de valores a um objeto pode-se usar o comando *c()*, no qual os valores vêm separados por vírgulas, dentro dos parênteses.

Exemplo 4 *Exemplo da criação de vetores*

```
x<-c("Masculino","Feminino")
x

y<-c(45.1, 55.3)
y
```

1.6.3 Array

Podemos definir arrays como um conjunto de elementos de dados, geralmente do mesmo tamanho e tipo de dados. Elementos individuais são acessados por sua posição no array. A posição é dada por um índice, também chamado de subscrição. O índice geralmente utiliza uma sequência de números naturais.

Um array é atribuído a um objeto pela função *array()*. Essa função tem como argumentos o conjunto de dados e as dimensões do array, nessa ordem.

Arrays podem ser de qualquer tipo, porém abordaremos apenas arrays numéricos, devido a sua grande importância para declaração de matrizes. Existem arrays unidimensionais e multidimensionais. Arrays numéricos unidimensionais nada mais são do que vetores. Já arrays numéricos multidimensionais podem ser usados para representação de matrizes. A sintaxe e um exemplo para criar um array são apresentados a seguir.

Sintaxe: x<- array(dados , dim= dimensao)

Exemplo 5 *Exemplo para criação de um array.*

```
x<- array(c(1:10), dim = c(2,5))
x
```

1.6.4 Matriz

Uma matriz é atribuída a um objeto pela função *matrix()*. Essa função tem como argumentos o conjunto de dados, o número de linhas e o número de colunas da matriz, nessa ordem. A sintaxe para construção de uma matriz no R e um exemplo são dados a seguire.

Sintaxe: A<- matrix(data = dados, nrow = m, ncol = n, byrow = Q),

sendo "m"o número de linhas, "n"o número de colunas. Se Q = 1 ativa disposição dos elementos por linhas e se Q = 0 mantém disposição dos elementos por colunas;

Exemplo 6 *Exemplo para criação de uma matriz*

```
A<-matrix(c(1:10),2,5,1)
A[1,]
A[2,4]
```

1.6.5 Data frame

Essa estrutura de dados é uma espécie de tabela, de estrutura bidimensional de dados. Números e strings podem fazer parte de um mesmo data frame. Além disso, podem ser dados nomes às colunas. Sua função é *data.frame()*.

Exemplo 7 *Exemplo para criação de um data frame.*

```
x<-c("Masculino","Feminino")
y<-c(45, 55)
dados<-data.frame(x,y)
dados
```

1.6.6 Lista

Uma lista é um conjunto de objetos de tamanhos e naturezas diferentes. Ela é regida pela função *list()*. Essa é a estrutura mais geral da linguagem R. Suas posições são designadas por números entre dois colchetes [[ ]].

Exemplo 8 *Construindo uma lista.*

```
Lista<-list(um.vetor=45:55,uma.matriz=matrix(1:10,2,5),
   um.dframe=data.frame(Atributo=c("Masculino","Feminino")
   ,Idade=c(45,55)));

Lista
```

1.7 Manipulação de Objetos

Em geral, é necessário, em muitas ocasiões, trabalhar com os objetos criados em R. Nesse caso, deve-se conhecer alguns operadores básicos descritos nas seções abaixo.

1.7.1 Operadores básicos

Expressões aritméticas podem ser construídas por intermédio dos operadores usuais apresentados na Tabela 1 serão mostrados abaixo.

Tabela 1 – Operações básicas

1	$\wedge$	Radiciação
2	/	Divisão à direita
3	$*$	Multiplicação
4	+	Adição
5	-	Subtração

1.7.2 Operadores relacionais e lógicos

O R também possui os operadores relacionais e operadores lógicos, os quais serão apresentados na Tabela 2.

Tabela 2 – Operadores lógicos e relacionais

Símbolo	**Descrição**
<	Menor
<=	Menor ou igual
>	Maior
>=	Maior ou igual
==	Igual (comparação)
!=	Diferente
&	AND
|	OR
!	NOT
TRUE ou 1	Valor booleano verdadeiro
FALSE ou 0	Valor booleano falso

1.7.3 Funções básicas

Existem algumas funções de uso constante, pertencentes aos pacotes básicos: algumas são apresentadas na Tabela 3.

Tabela 3 – Funções básicas

Função	**Descrição**
sum(x)	soma todos os elementos de um objeto x.
length(x)	retorna o comprimento de um objeto x.
rep(x,n)	repete o número x, n vezes.
seq(a,b,by=c)	gera uma sequência de números entre a e b, distantes c unidades.
table(x)	retorna tabela com frequências de ocorrência da cada elemento de x.

1.7.4 Funções matemáticas básicas

O R possui várias funções matemáticas já implementadas. Na Tabela 4 será listado as principais funções com sua respectiva descrição.

Tabela 4 – Funções matemáticas básicas

Função	**Descrição**
abs(x)	valor absoluto de x
log(x, b)	logaritmo de x com base b
log(x)	logaritmo natural de x
log10(x)	logaritmo de x com base 10
exp(x)	exponencial elevado a x
sin(x)	seno de x
cos(x)	cosseno de x
tan(x)	tangente de x
round(x, digits = n)	arredonda x com n decimais
ceiling(x)	arredondamento de x para o maior valor
floor(x)	arredondamento de x para o menor valor
length(x)	número de elementos do vetor x
sum(x)	soma dos elementos do vetor x
prod(x)	produto dos elementos do vetor x
max(x)	seleciona o maior elemento do vetor x
min(x)	seleciona o menor elemento do vetor x
range(x)	retorna o menor e o maior elemento do vetor x

É evidente que as operações apresentada na Tabela 3 podem ser aplicadas aos vetores criados anteriormente.

1.7.5 Operações e funções com Matrizes

A Tabela 5 apresenta as principais funções que podem ser aplicadas a matrizes.

Tabela 5 – Operações e funções com matrizes

Função	Descrição
A∗B	produto elemento a elemento de A e B
A%∗%B	produto matricial de A por B
B = t(A)	matriz transposta: $B = A^t$
B = solve(A)	matriz inversa: B = A^{-1}
x = solve(A, b)	resolve o sistema linear Ax = b
det(A)	retorna o determinante de A
diag(v)	retorna uma matriz diagonal no qual o vetor v é a diagonal
diag(A)	retorna um vetor que é a diagonal da matriz A
diag(n)	sendo n um inteiro, retorna uma matriz identidade de ordem n
eigen(A)	retorna os autovalores e autovetores de A
eigen(A)$values	retorna os autovalores de A
eigen(A)$vectors	retorna os autovetores de A

Para aplicar as funções mencionadas na Tabela 5, serão utilizados os objetos criados anteriormente, isto é, a matriz **(A)** e o array **(x)**, com uso dos exemplos a seguir.

Exemplo 9 *Calculando a transposta de uma matriz.*

```
B <- t(A)
B
```

Exemplo 10 *Resolvendo um sistema linear do tipo Ax=b.*

```
b<-array(c(0,1,5),dim=c(3,1));
C<-matrix(c(c(1,1,0),c(0,1,4),c(0:2)),3,3,1);
y<-solve(C,b)
y
```

Exemplo 11 *Multiplicando uma matriz pela sua inversa.*

```
Cinv<-solve(C)
Cinv\%*\%b
```

1.8 Help

O jeito mais fácil de se aprender a usar R é consultando constantemente seus tópicos de ajuda. Existem basicamente quatro tipos de ajuda no R:

a) help('função()'): Essa ajuda deve ser solicitada quando se sabe da existência de uma função (sabe-se seu nome exato), mas existe dúvidas em como usá-la. Se o pacote que contém essa função estiver instalado e carregado, será aberta a documentação da mesma para esclarecimentos;

b) help.search(' '): Quando se deseja investigar a existência de uma função, essa ajuda recebe uma palavra-chave (em Inglês) e retorna todas aquelas funções que contêm aquela palavra em sua documentação. A busca é feita nos pacotes existentes no computador em questão, ou seja, se uma busca não retornar nenhum resultado adequado, não significa que a função não exista. Significa que ela não existe, pelo menos, em seu computador;

c) Ajuda Html: Essa ajuda pode ser chamada pela barra de menu, no botão Ajuda (**Help**). Quando acionada, ela abre um documento em html que contém diversas informações sobre o R, sua linguagem, suas funções básicas, seus pacotes, seus autores, sua licença, perguntas mais freqüentes, entre outros;

d) RSiteSearch(' '): Quando conectado à internet, essa ajuda faz a busca de uma palavra-chave em todas as páginas da internet relacionadas com o R, principalmente aquelas páginas publicadas com as perguntas e respostas das listas de discussões do R. Existem diversos tipos de listas de discussões que podem ser encontradas na página do R. Nelas são tiradas dúvidas mais grave, são dadas sugestões para as novas versões do R, são desvendados e descoberto pequenos erros de programação etc. Elas colocam os usuários do R em contato com os estatísticos que fazem e mantêm o R.

Quando se deseja saber informações acerca de uma dada função existente deve-se digitar help("função") ou, simplesmente, ?função(). Caso se deseje saber se um tópico possui função no R, o comando deve ser: help.search("tópico").]

1.9 Exercícios de Aplicação

Para fixar os conceitos explicados anteriormente, serão descritos alguns exercícios para melhorar o aprendizado.

1. Criar um array de 2x7
2. Criar uma matriz de 4x4
3. Fazer a busca do elemento na linha 1, coluna 4
4. Fazer a busca de todos os elementos da linha 3
5. Fazer a busca de todos os elementos da coluna 4
6. Obter a matriz transposta
7. Obter a matriz inversa
8. Calcular o determinante da matriz

1.9.1 Gabarito

1.
```
A<-array(c(1:14), dim = c(2,7))
A
```

2.
```
A<- matrix(c(1:16),4,4,1)
A
```

3.
```
A[1,4]
```

4.
```
A[3,]
```

5.
```
A[,4]
```

6.
```
B<- t(A)
```

7.
```
B<-solve(A)
```

8.
```
x<-det(A)
x
```

Capítulo 2

Noções de Amostragem

2.1 Introdução

Ao conduzirmos uma pesquisa, na maioria das vezes não podemos trabalhar com a população como um todo, devido a uma série de fatores. Sendo assim, é necessário que seja escolhido uma parte desta população, a qual deve ser representativa, para que os resultados obtidos por meio das análises conduzidas com os elementos deste grupo sejam semelhantes aos que seriam obtidos se fosse analisada a população. O ato de retirar uma parte desta população é conhecido como amostragem. Neste capítulo estaremos interessados em trabalhar com algumas noções em relação aos conceitos de amostragem.

2.2 Conceitos Iniciais

Dada uma pesquisa, onde desejamos analisar as características das empresas do agronegócio do sul do país, porém não podemos analisar todas as empresas em questão, como selecionar uma parte desta população que seja representativa?

A maneira de obter uma amostra é tão importante e existem tantos modos de fazê-los.

Estes procedimentos constituem uma especialidade da estatística conhecida como amostragem. Os planos de amostragem podem ser agrupados em: **Planos probabilísticos** e **não probabilísticos**. No primeiro caso consiste atribuir a cada elemento da amostra uma probabilidade a priori de pertencer à amostra. No segundo grupo estão as mostras intencionais.

Por quê fazer amostragem? São 4 razões as razões:

- Economia;

- Tempo;
- Confiabilidade dos dados,
- Operacionalidade.

2.2.1 Amostragem com reposição dos elementos

Ao trabalharmos em diversas situações nos deparamos com diferentes cenários. Nesse sentido, ao selecionarmos elementos de uma população, a nossa amostragem pode ser feita com reposição dos elementos de um conjunto. Nesse caso, o número de amostras possíveis a serem obtidas, pode ser dado pela expressão a seguir,

$$T = N^n, \tag{2.1}$$

sendo *N* o número de elementos da população e *n* o número de elementos da amostra a ser obtida.

2.2.2 Amostragem sem reposição dos elementos

Em determinadas ocasiões é possível realizar amostragem sem reposição dos elementos, isto é, uma vez pertencente a amostra, ele não pode mais ser avaliado. A amostragem sem reposição é mais eficiente que a amostragem com reposição e reduz a variabilidade uma vez que não é possível retirar elementos extremos mais do que uma vez. Nessa situação, o número de amostras possíveis a serem obtidas, pode ser dado pela expressão a seguir,

$$T = \frac{N!}{n!(N-n)!}, \tag{2.2}$$

sendo *N* o número de elementos da população e *n* o número de elementos da amostra a ser obtida.

Exemplo 12 *Em um estudo sobre o consumo de combustível, definiu-se uma população composta por quatro ônibus de uma pequena companhia de transporte urbano. Os consumos dos ônibus (km/l), em condições padrões de teste eram 3,9, 3,8, 4,0 e 4,1. Determinar:*

a) *O número de amostras de dois elementos que podem ser obtidas com reposição;*

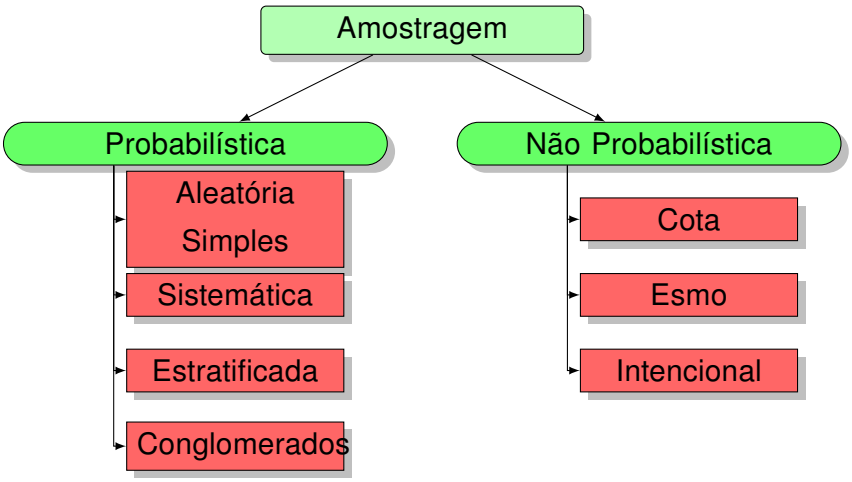
b) *O número de amostras de dois elementos que podem ser obtidas sem reposição;*

c) *Todas as possíveis amostras de dois elementos, escolhidos com reposição;*

d) *Todas as possíveis amostras de dois elementos, escolhidos sem reposição.*

2.2.3 Tipos de Amostragem

Podemos classificar a amostragem em não probabilística e probabilística (mais utilizada). Dentro da amostragem não probabilística temos a amostragem a esmo, intencional e cotas, para a amostragem probabilística existe a amostragem simples ou ocasional, sistemática, estratificada e por conglomerados. Por meio do diagrama a seguir, é possível identificar os tipos de amostragem.



2.2.4 Amostragens Probabilísticas

2.2.4.1 Amostragem Simples ou Ocasional

É o processo mais elementar e frequentemente utilizado. Todos os elementos da população têm igual probabilidade de serem escolhidos. Para uma população finita o processo deve ser sem reposição. Todos os elementos da população devem ser numerados. Para realizar o sorteio dos elementos é utilizada a Tabela de Números Aleatórios, ou um software estatístico.

Exemplo 13 *A Tabela a seguir refere-se ao número de vezes que uma empresa de insumos visita seus clientes de determinada, no período de 1 ano. Há 30 propriedades, deseja-se estudar a quantia que cada produtor compra de determinado adubo para sua lavoura, porém não dispomos de tempo para analisar todas as trinta propriedades. Para isso faça uma amostragem aleatória simples de 10 propriedades.*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Tabela 6 – Número de Visitas por propriedade

Determinar a amostra utilizando a tabela dos números aleatórios.

Solução: Considerando a Tabela de números aleatórios que está em anexo ao material, utilizando os dois últimos algarismos dos números de cada coluna, as propriedades selecionadas são as de número 13, 27, 06, 05, 02, 15, 21, 28, 04 e 03.

Para resolver no R, os comandos utilizados são os a seguir, que resultarão nas posições do elementos desejados.

```
set.seed(1212)
dados<-seq(1:30)
amostras<-sample(dados,10,replace=FALSE)
amostras
```

2.2.4.2 Amostragem Sistemática

Trata-se de uma variação da amostragem aleatória simples, quando a população está ordenada.

Sendo N o tamanho da população e n o tamanho da amostra desejada, define-se a quantidade $\frac{N}{n} = k$, chamado de intervalo de amostragem. Seleciona-se aleatoriamente um número entre 1,2,3,. . . ,k e se obtêm o valor *i*, que será o primeiro elemento. Os demais elementos poderão ser calculados utilizando a fórmula do termo geral de uma PA, dado pela expressão a seguir:

$$a_n = a_1 + (n-1)k \tag{2.3}$$

sendo *n* o Número total de elementos, a_1 o primeiro elemento, a_n o N-ésimo elemento e k a Razão da PA.

Exemplo 14 *Em um lote há 150 lâmpadas. O controle de qualidade deseja inspecioná-lo. Todavia, apenas algumas. Obter uma amostra desta população com 30 lâmpadas, utilizando amostragem sistemática. Identificar as lâmpadas selecionadas, sabendo que elas estão numeradas de 1 a 150.*

Solução *Determinando o valor de k temos que $k = \frac{150}{30} = 5$. Como k = 5, selecionaremos os números em um intervalo de 5. Nesse caso, as lâmpadas selecionadas. Escolha um número entre 1 e 5 e a partir disso, os elementos da amostra são compostos por meio da expressão para cálculo dos elementos de uma P.A.*

No R, temos os seguintes comandos para execução.

```
n <-30
N <-150
k <-N/n
set.seed (1212)
amostra<-sample(k,1)
amostra
amostras<-seq(amostra,N,k)
amostras
```

Caso o valor de k não seja inteiro, então deve-se usar o comando *ceiling(k)* para que ele assuma o próximo valor inteiro.

2.2.4.3 Amostragem Estratificada

A amostragem estratificada é mais uma variante da amostra aleatória simples, uma vez que após divisão da população alvo em subgrupos homogêneos chamados "estratos", a seguir se tira de forma aleatória uma amostra de cada estrato.

A Amostragem aleatória estratificada é utilizada, ao contrário das anteriores, quando a população inteira é reconhecida por certas características precisas, tais como a idade, o sexo, a incidência de uma condição de saúde, tudo isto para assegurar a melhor representatividade possível. Com efeito, quando os elementos da população estão divididos em grupos não sobrepostos, é mais fácil e mais eficiente escolher, independentemente, uma amostra aleatória simples dentro de cada um destes grupos, os quais são chamados estratos.

Esta forma de amostragem é uma das mais utilizadas, já que a maioria das populações têm estratos bem definidos: os homens e as mulheres; os alunos das escolas X, Y, Z; os operários pertencentes aos índices salariais 190, 195, etc.

O mais comum é utilizar-se a amostragem estratificada proporcional, que consiste em selecionar os elementos da amostra entre os vários estratos, em número proporcional ao tamanho de cada um dos estratos.

Como a população se divide em subconjuntos, convém que o sorteio dos elementos leve em consideração tais divisões, para que os elementos da amostra representem o número de elementos desses subconjuntos. Como exemplo observe a figura abaixo:

Exemplo 15 *Obter uma amostra estratificada de 10% da população para a pesquisa da estatura de 90 alunos de uma escola sendo que destes 54 sejam meninos e 36 sejam meninas.*

Para realizar essa amostragem estratificada no R, basta seguir os comandos a seguir.

```
set.seed(1212)
n <-0.10*(54+36)
N1<-54
N2<-36
N<-N1 + N2
N
f<-n/N
f
n1<-ceiling(f*N1)
n1
n2<-f*N2
n2
amostra<-cbind(n1, n2)
amostra.1<- sample(N1, n1, replace=F)
amostra.1
[1] 15  6 51 18 32  2
amostra.2<- sample(N2, n2, replace=F)
amostra.2
[1]  4  3 24
```

Após executar os comandos acima, os elementos da amostra serão os das posições obtidas. Nesse caso, cada conjunto foi ordenado de 1 ao número máximo de elementos e feita a amostragem.

Exemplo 16 *Temos uma população de 200 plantas geneticamente modificadas. Deseja-se estudar as alterações nas características de cada planta causadas por esta modificação genética. Deste total há 80 frutíferas e 120 flores. Obter através da amostragem estratificada uma amostra de 30% da população, de modo que tal amostra seja representativa.*

2.2.4.4 Amostragem por Conglomerados

Como regra geral, o número de elementos em um conglomerado deve ser pequeno em relação ao tamanho da população e o número de conglomerados razoavelmente grande. A amostragem por conglomerado é recomendada quando:

- Ou não se tem um sistema de referência listando todos os elementos da população.
- O custo da obtenção de informações cresce com o aumento da distância entre os elementos.

2.3 Exercícios de Aplicação

1. Em uma amostragem sistemática, de tamanho 40, de uma população de 2000 elementos, o primeiro elemento selecionado é o 36. Quais são os próximos dois elementos a serem escolhidos são?

2. Numa escola estão matriculados 280 meninos e 320 meninas (não existindo alunos irmãos), onde todos tem pelo menos um problema de visão. O diretor da escola, desejoso de conhecer os problemas de seus alunos e não dispondo de tempo para entrevistar todas as família, resolveu fazer um levantamento por amostragem proporcional estratificada, composta de 50 alunos. As famílias de quantos meninos serão entrevistas?

3. Uma amostragem sistemática será constituída de 6 elementos, para uma população de 360 elementos. O primeiro elemento sorteado para a amostra foi o 45. Se todos os elementos do universo são numerados de 1 a 360, qual será o terceiro elemento dessa amostra?

4. No total de 60 fazendas, 25 serão selecionados para investigação com relação ao desmatamento de áreas protegidas, organizar esta amostra em esquema de Amostragem Sistemática e retirar da população de fazendas que serão investigados.

5. Utilizando os dados do exercício anterior extraia uma amostra estratificada de tamanho n = 15. Escolha estes 15 hospitais utilizando a 5ª coluna a partir da 22ª linha.

6. Calcule o número de amostras que deverão ser sorteadas, dentro de cada estrato, de uma população de tamanho N = 2000, sendo o tamanho da amostra n = 80. A população esta distribuída em quatro extratos A =500, B = 1200, C = 200, D = 100.

7. Dentre 100 pessoas escaladas para sorteio de um júri, 30 são negras que estão no estrato 1 e 70 são brancas que estão no estrato 2. Obter uma amostra estratificada de tamanho n = 12.

2.3.1 Gabarito

1. 2º = 86 e 3º = 122.

2. 24 famílias.

3. 3º = 165

4. A = 20, B= 48, C= 8, D= 4

Capítulo 3

Estatística Descritiva

3.1 Introdução

A origem da palavra Estatística está associada à palavra latina STATUS (Estado). Há indícios de que 3000 anos A.C. já se faziam censos na Babilônia, China e Egito e até mesmo o 4o. livro do Velho Testamento faz referência a uma instrução dada a Moisés, para que fizesse um levantamento dos homens de Israel que estivessem aptos para guerrear. Usualmente, estas informações eram utilizadas para a taxação de impostos ou para o alistamento militar.

A estatística envolve técnicas para coletar, organizar, descrever, analisar e interpretar dados, ou provenientes de experimentos, ou vindos de estudos observacionais. Análise estatística de dados geralmente tem por objetivo a tomada de decisões, resoluções de problemas ou produção de conhecimento. Mas novos conhecimentos normalmente geram problemas de pesquisas, resultando em um processo iterativo.

Normalmente o termo estatística esta associado a números, tabelas, gráficos, mas a importância da estatística fica melhor representada por dois ingredientes comuns em nosso dia a dia:

- Dados
- Variabilidade

Para o engenheiro conhecer as propriedades físicas de um material, ele pode medir duas características. Se ele medir vários corpos de prova do mesmo material poderá encontrar valores diferentes. Fazer algumas indagações, sem analisar a variabilidade dos dados pode comprometer todo o processo.

Com a alta competitividade de hoje, para que uma empresa sobreviva, ela tem o desafio de adequar o produto ao cliente. A demanda de certo produto exige que certo material tenha um valor específico de dureza. Porém, como encontrar tal valor?

Por outro lado, adequar o produto ao cliente envolve saber o que o consumidor deseja. Mas os consumidores têm preferências diferentes, o que exige a realização de pesquisas observacionais, ou de levantamento. Como fazer esse levantamento?

Podemos dividir a estatística em três áreas: Estatística Descritiva, Probabilística e Estatística Inferencial.

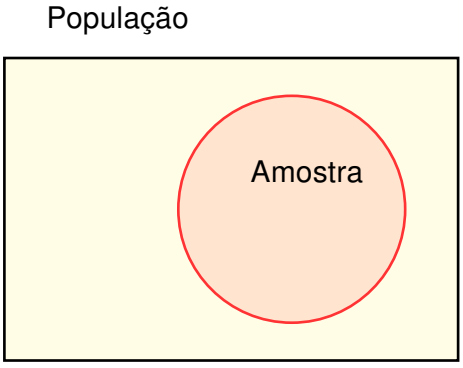
A estatística descritiva é a etapa inicial da análise utilizada para descrever e resumir os dados. A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes revigorou está área da estatística.

Inferência Estatística é, o estudo de técnicas fundamentada na teoria das probabilidades, que possibilitam a extrapolação a um grande conjunto de dados, das informações e conclusões obtidas a partir da amostra.

3.2 Conceitos Básicos

População: Conjunto de elementos que tem pelo menos uma característica em comum. Esta característica deve delimitar corretamente quais são os elementos da população que podem ser animados ou inanimados.

Amostra: Subconjunto não vazio da População.

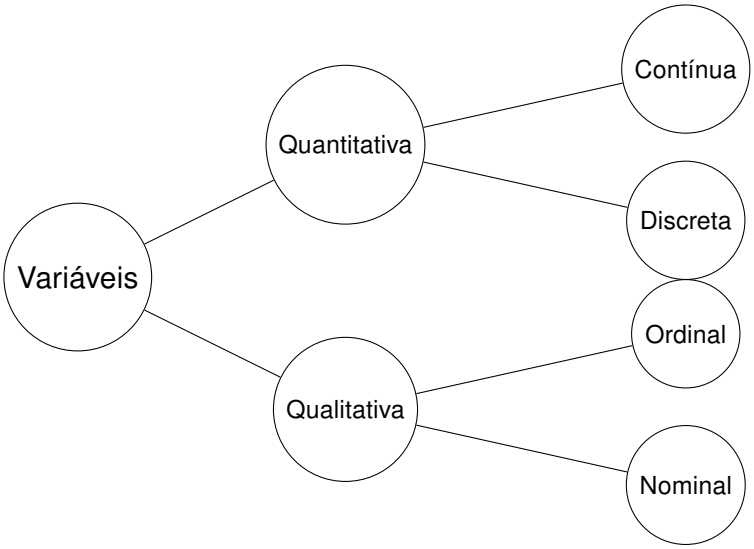


3.3 Apresentação dos Dados

Uma pesquisa estatística envolve um grande número de dados numéricos. Nosso objetivo é organizá-los, analisá-los e apresentá-los de maneira prática, a qual deve ser de fácil compreensão. Tanto a organização, análise e apresentação dos dados depende da variável em estudo.

3.3.1 Variáveis

Pode-se identificar as variáveis que podem estar presente em um estudo por meio do diagrama abaixo.



3.3.1.1 Variáveis Qualitativas

É uma variável que assume como possíveis valores, atributos ou qualidades. Também são denominadas variáveis categóricas.

3.3.1.2 Variáveis Qualitativas Nominais

Variável que assume como possíveis valores, atributos ou qualidades e estes não apresentam uma ordem natural de ocorrência.

Exemplo 17 *Profissão, tipo sanguíneo, sexo.*

3.3.1.3 Variáveis Qualitativas Ordinais

Assume como possíveis valores atributos ou qualidades e estes apresentam uma ordem natural de ocorrência.

Exemplo 18 *Solteiro(a), casado(a), viúvo(a).*

3.3.1.4 Variáveis Quantitativas

É uma variável que assume pode ser expressa por variáveis com níveis de medição intervalar ou de razão, isto é, os resultados assumem valores em uma escala numérica.

3.3.1.5 Variáveis Quantitativas Discretas

Variável que assume como possíveis valores números, em geral inteiros, formando um conjunto enumerável.

Exemplo 19 *Número colheitas em um determinado período.*

3.3.1.6 Variáveis Quantitativas Contínuas

Variável que assume como possíveis valores números em intervalos, em geral, resultantes de mensurações.

Exemplo 20 *Altura, peso, taxas.*

A partir do conhecimento das variáveis, o primeiro passo é organizá-las. Isso é feito por meio de tabelas de frequência, onde observa-se a ocorrência de cada uma das características que está em análise.

Os dados apresentados nas Tabelas 7 e 8 foram extraídas do material disponibilizado por **??**) e quando necessárias, adaptadas para o contexto deste material.

Exemplo 21 *Considere o conjunto de dados abaixo, o qual contém o setor de trabalho, estado civil e número de horas extras que um integrante de cada uma de 37 famílias que participaram de determinada pesquisa.*

Tabela 7 – Informações para pessoas de diferentes famílias, quanto setor de trabalho, estado civil e número de horas no trabalho.

Nº	SETOR	ESTADO CIVIL	HORAS EXTRAS
1	Fábrica	Casado	3
2	Fábrica	Solteiro	4
3	Fábrica	Separado	3
4	Laboratório	Casado	5
5	Fábrica	Casado	4
6	Lavoura	Solteiro	2
7	Lavoura	Casado	3
8	Fábrica	Solteiro	2
9	Almoxarifado	Separado	4
10	Lavoura	Solteiro	3
11	Administrativo	Casado	4
12	Fábrica	Casado	3
13	Administrativo	Solteiro	3

14	Lavoura	Separado	2
15	Administrativo	Casado	4
16	Almoxarifado	Solteiro	1
17	Fábrica	Casado	4
18	Lavoura	Casado	4
19	Administrativo	Solteiro	2
20	Lavoura	Casado	1
21	Fábrica	Solteiro	3
22	Administrativo	Separado	4
23	Laboratório	Solteiro	2
24	Administrativo	Solteiro	2
25	Laboratório	Solteiro	3
26	Administrativo	Casado	4
27	Fábrica	Casado	3
28	Laboratório	Solteiro	4
29	Lavoura	Solteiro	2
30	Laboratório	Casado	1
31	Administrativo	Solteiro	3
32	Fábrica	Casado	2
33	Lavoura	Solteiro	4
34	Almoxarifado	Casado	2
35	Fábrica	Casado	3
36	Administrativo	Solteiro	3
37	Lavoura	Casado	6

a) Classifique cada uma das variáveis do estudo (setor, estado civil e horas extras) de acordo com os tipos de variáveis mencionadas acima.

3.3.2 Tabelas de Frequência e Gráficos

Tabelas de frequência ou distribuição de frequências, resumem as informações presentes em um conjunto de dados. São construídas de acordo com o tipo de variável que é objeto de estudo.

Exemplo 22 *Considere os dados obtidos em uma pesquisa realizada com 20 estudantes,*

escolhidos aleatoriamente, de uma instituição de ensino superior.

Para as variáveis codificadas, tem-se que Q-01 representa o estado onde nasceu; Q-02 estado onde morava antes de ingressar na instituição atual; Q-03 Número de vestibulares realizados e reprovados; Q-04 número de irmãos e Q-05 comprimento do dedo maior da mão direita (medido por cinco alunos e feito o cálculo da média das 5 medidas).

3.3.2.1 Tabelas e Gráficos para variáveis qualitativas

De acordo com as classificações apresentadas na seção 3.3.1.1, Sexo, Q01, Q02 e classificação de peso, são caracterizadas como **variáveis qualitativas**. Para tal, é possível construir os seguintes gráficos e tabelas de frequência:

- Variável Sexo

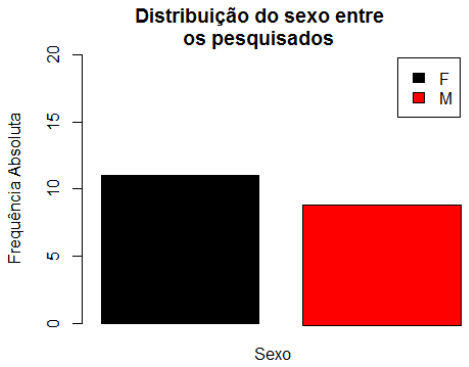


Figura 6 – Gráfico de Colunas

Modalidade	Freq. Abs.	Freq. Rel
Feminino	11	55%
Masculino	09	45%

Tabela 9 – Tabela de Frequência

- Variável Q01

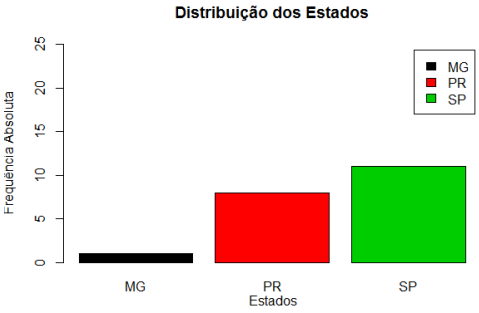


Figura 7 – Gráfico de Colunas

Modalidade	Freq. Abs.	Freq. Rel
MG	01	5%
PR	08	40%
SP	11	55%

Tabela 10 – Tabela de Frequência

Tabela 8 – Dados coletados para 20 alunos de uma instituição de ensino superior.

Unidade	Sexo	Idade	Q01	Q02	Q03	Q04	Altura	Peso	IMC	Classificação	Q05
1	Masculino	19	PR	PR	2	2	1,90	78,00	21,61	Peso normal	8,50
2	Feminino	19	PR	MT	7	2	1,76	58,00	18,72	Peso normal	8,50
3	Masculino	18	SP	SP	3	2	1,79	60,00	18,73	Peso normal	8,30
4	Masculino	18	SP	SP	4	1	1,85	65,00	18,99	Peso normal	9,00
5	Masculino	20	SP	SP	6	1	1,76	81,00	26,15	Sobrepeso	8.50
6	Masculino	19	PR	PR	1	1	1,73	66,40	22,19	Peso normal	7,30
7	Feminino	21	SP	SP	3	2	1,62	51,00	19,43	Peso normal	7,70
8	Feminino	19	SP	SP	3	1	1,69	52,00	18,21	Abaixo do peso	8,10
9	Masculino	19	MG	MG	6	2	1,85	72,00	21,04	Peso normal	9,00
10	Feminino	18	SP	SP	0	1	1,65	67,00	24,61	Peso normal	8,00
11	Feminino	19	PR	PR	5	1	1,60	53,00	20,70	Peso normal	7,50
12	Masculino	18	SP	SP	2	0	1,70	68,00	23,53	Peso normal	9,00
13	Masculino	19	SP	SP	3	3	1,69	70,00	24,51	Peso normal	9,00
14	Masculino	20	SP	SP	4	1	1,86	78,00	22,55	Peso normal	9,50
15	Feminino	18	PR	PR	5	2	1,63	54,00	20,32	Peso normal	7,00
16	Feminino	19	PR	PR	7	2	1,65	58,50	21,49	Peso normal	7,00
17	Feminino	21	SP	SP	8	2	1,64	49,00	18,22	Abaixo do peso	7,50
18	Feminino	18	PR	PR	2	1	1,74	55,00	18,17	Abaixo do peso	8,50
19	Feminino	16	PR	PR	0	2	1,70	57,00	19,72	Peso normal	8,00
20	Feminino	19	SP	SP	3	1	1,70	51,00	17,65	Abaixo do peso	8,00

- Variável Q02

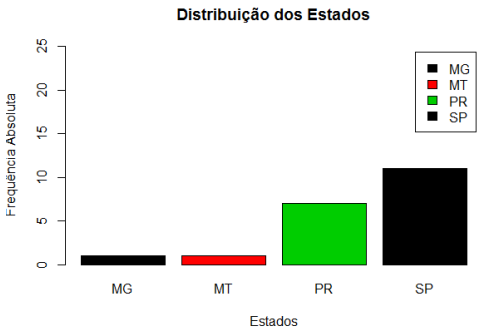


Figura 8 – Gráfico de Colunas

Modalidade	Freq. Abs.	Freq. Rel
MG	01	5%
MT	01	5%
PR	07	35%
SP	11	55%

Tabela 11 – Tabela de Frequência

- Classificação do Peso

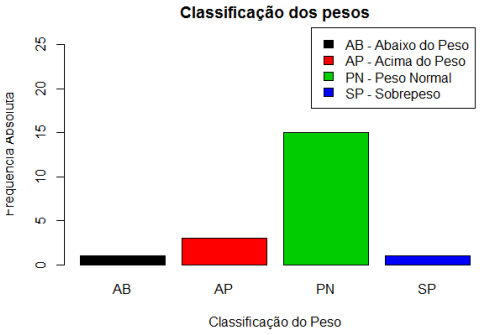


Figura 9 – Gráfico de Colunas

Modalidade	Freq. Abs.	Freq. Rel
AB	01	5%
AP	03	15%
PN	15	75%
SP	01	05%

Tabela 12 – Tabela de Frequência

Além das tabelas apresentadas anteriormente, é possível "cruzar"as informações das variáveis duas a duas, originando o gráfico e tabela a seguir.

- Classificação do Peso de acordo com o sexo

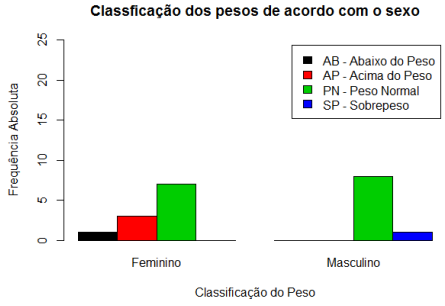


Figura 10 – Gráfico de Colunas

	Feminino	Masculino
AB	1	0
AP	3	0
PN	7	8
SP	0	1

Tabela 13 – Tabela de Frequência

Exemplo 23 *Para as variáveis qualitativas apresentadas na Tabela 7, construir a tabela de frequência e seus respectivos gráficos.*

Para construção dos gráficos da seção 3.3.2 no R, basta executar os comandos abaixo. Ressalta-se que a tabela de frequência deve ser construída no *software* de acordo com as necessidades do usuário, isto é, incluindo os elementos necessários. Ainda, ao passar os comandos a seguir para o *software*, extraia os espaços obsoletos existentes no código.

```
dat1<-read.table(header=TRUE, text="Sexo  Idade Estado_Civil  Q01
    Q02 Q03 Q04 Altura  Peso  IMC Classificacao Q05
Masculino 19  Solteiro  PR  PR  2  2  1.90  78.0  21.61 PN  8.5
Feminino  19  Solteiro  PR  MT  7  2  1.76  58.0  18.72 PN  8.5
Masculino 18  Solteiro  SP  SP  3  2  1.79  60.0  18.73 PN  8.3
Masculino 18  Solteiro  SP  SP  4  1  1.85  65.0  18.99 PN  9.0
Masculino 20  Solteiro  SP  SP  6  1  1.76  81.0  26.15 SP  8.5
Masculino 19  Solteiro  PR  PR  1  1  1.73  66.4  22.19 PN  7.3
Feminino  21  Solteiro  SP  SP  3  2  1.62  51.0  19.43 PN  7.7
Feminino  19  Solteiro  SP  SP  3  1  1.69  52.0  18.21 AP  8.1
Masculino 19  Solteiro  MG  MG  6  2  1.85  72.0  21.04 PN  9.0
Feminino  18  Solteiro  SP  SP  0  1  1.65  67.0  24.61 PN  8.0
Feminino  19  Solteiro  PR  PR  5  1  1.60  53.0  20.70 PN  7.5
Masculino 18  Solteiro  SP  SP  2  0  1.70  68.0  23.53 PN  9.0
Masculino 19  Solteiro  SP  SP  3  3  1.69  70.0  24.51 PN  9.0
Masculino 20  Solteiro  SP  SP  4  1  1.86  78.0  22.55 PN  9.5
Feminino  18  Solteiro  PR  PR  5  2  1.63  54.0  20.32 PN  7.0
Feminino  19  Solteiro  PR  PR  7  2  1.65  58.5  21.49 PN  7.0
Feminino  21  Solteiro  SP  SP  8  2  1.64  49.0  18.22 AP  7.5
Feminino  18  Solteiro  PR  PR  2  1  1.74  55.0  18.17 AP  8.5
Feminino  16  Solteiro  PR  PR  0  2  1.70  57.0  19.72 PN  8.0
Feminino  19  Solteiro  SP  SP  3  1  1.70  51.0  17.65 AB  8.0")

barplot(table(dat1$Sexo),beside = T, legend=TRUE, col=1:2, ylim=c
   (0,25),xlab="Sexo",ylab="Frequencia Absoluta",
main="Distribuicao do sexo entre os pesquisados")

barplot(table(dat1$Q01),beside =T, legend=TRUE, col=1:3, ylim=c
   (0,35),  legend.text = c("MG", "PR" , "SP"),xlab="Estados",ylab
   ="Frequencia Absoluta",
main="Distribuicao dos Estados")

barplot(table(dat1$Q02),beside = T, legend=TRUE, col=1:3, ylim=c
   (0,25),  legend.text = c("MG", "MT", "PR" , "SP"),xlab="Estados
   ",ylab="Frequencia Absoluta",
main="Distribuicao dos Estados")
```

```
barplot(table(dat1$Classificacao),beside = T, legend=TRUE, col
   =1:4, ylim=c(0,25),  legend.text = c("AB - Abaixo do Peso", "AP
    - Acima do Peso", "PN - Peso Normal", "SP - Sobrepeso"), xlab=
   "Classificacao do Peso",ylab="Frequencia Absoluta",
main="Classificacao dos pesos")

barplot(table(dat1$Classificacao,dat1$Sexo),beside=TRUE,legend=
   TRUE, col=1:4, ylim=c(0,25),  xlab="Classificacao do Peso",ylab
   ="Frequencia Absoluta",
main="Classficacao dos pesos de acordo com o sexo")

x<-ftable(dat1$Sexo, dat1$Classificacao)
barplot(x)
```

3.3.2.2 Tabelas e Gráficos para variáveis quantitativas

Como visto na seção 3.3.1, as variáveis quantitativas podem ser classificadas como discretas ou contínuas. Nesse sentido, existem diferentes formas de representar variáveis com essas características. Assim, as maneiras mais simples de apresentar dados, são por meio de tabelas de frequência, bem como por intermédio de gráficos associados a elas.

Uma tabela de frequência, para variáveis quantitativas discretas, deve ter os seguintes elementos:

a) Título;

b) Elemento que representa determinada ocorrência, representado por (x_i) ;

c) Frequência absoluta e frequência absoluta acumulada (f_i e F_i);

d) Frequência relativa e frequência relativa acumulada (r_i e R_i).

Para determinar os elementos da F_i, repete-se na primeira classe o valor de f_i. Para as classes abaixo, soma-se o valor de F_i imediatamente superior com seu respectivo f_i. O processo repete-se até completar todas as lacunas. O mesmo ocorre com os elementos de R_i, sempre somando com seus respectivos valores de r_i. O valor dos elementos da coluna r_i e R_i são calculados por meio das expressões;

$$r_i(\%) = \frac{f_i}{n} \times 100 \quad e \quad R_i(\%) = \frac{F_i}{n} \times 100 \tag{3.1}$$

sendo n o número total de elementos do conjunto de dados.

As variáveis (Tabela 8) número de vestibulares realizados e reprovados (Q03) e número de irmãos (Q04) são classificadas como discretas, isto é, são representadas por características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente são o resultado de contagens.

- Variável Q04.

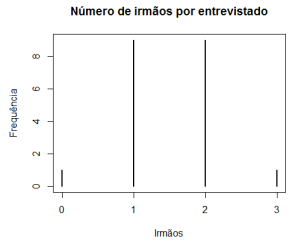


Figura 11 – Gráfico de bastões.

x_i	f_i	F_i	r_i	R_i
0	1	1	5%	5%
1	9	10	45%	50%
2	9	19	45%	95%
3	1	20	5%	100%

Tabela 14 – Tabela de Frequência

- Variável Q03.

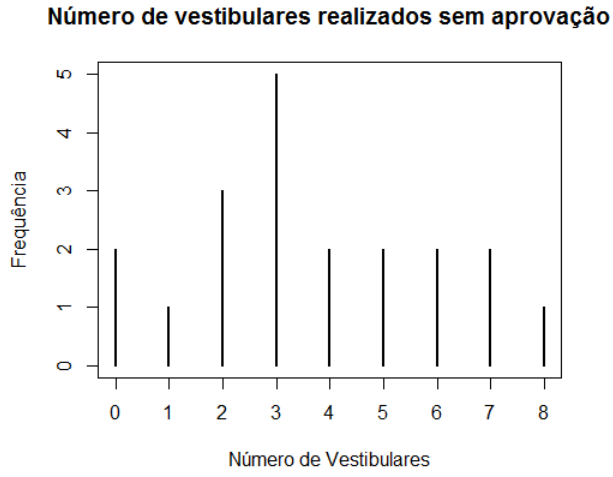


Figura 12 – Gráfico de bastões.

x_i	f_i	F_i	r_i	R_i
0	2	2	10%	10%
1	1	3	5%	15%
2	3	6	15%	30%
3	5	11	25%	55%
4	2	13	10%	65%
5	2	15	10%	75%
6	2	17	10%	85%
7	2	19	10%	95%
8	1	20	5%	100%

Tabela 15 – Tabela de Frequência

Nos dois exemplos acima, os códigos para execução do R são dados a seguir. Ressalta-se que as tabelas que podem ser construídas automaticamente, via função, no R, tem as duas primeiras colunas apenas.

```
Q03<-table(dat1$Q03)

plot(Q03,main="Numero de vestibulares realizados sem
   aprovacao",xlab="Numero de Vestibulares",ylab="
   Frequencia")
```

```
Q04<-table(dat1$Q04)

plot(Q04,main="Numero de irmaos por entrevistado",xlab="
   Irmaos",ylab="Frequencia")
```

Os gráficos apresentados nas Figuras 20 e 11 são formados por segmentos de retas perpendiculares ao eixo horizontal (eixo da variável), cujo comprimento corresponde à frequência absoluta ou relativa de cada elemento da distribuição. Suas coordenadas não podem ser unidas porque a leitura do gráfico deve tornar claro que não há continuidade entre os valores individuais assumidos pela variável em estudo.

As variáveis (Tabela 8) Idade, altura, IMC e Q05 são classificadas como contínuas, uma vez que são características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores fracionais fazem sentido. Nesse caso, a construção das tabelas de frequência torna-se mais complexa devido a característica de continuidade das variáveis.

Uma tabela de frequência para variáveis contínuas deve ter os seguintes elementos:

a) Título;

b) Intervalo de classe ($I_c = [L_i; L_s[$);

c) Média dos intervalos de classe (x_i);

d) Frequência absoluta e frequência absoluta acumulada (f_i e F_i);

e) Frequência relativa e frequência relativa acumulada (r_i e R_i).

Inicialmente, organiza-se os dados em ordem crescente. Um I_c é formado por dois extremos, os quais dependem de sua amplitude, que por sua vez é calculada com base nos dados coletados, bem como de acordo com o número de classes (k linhas) que serão utilizadas na construção da tabela. Existem diversas expressões que permitem calcular o número de classes para a Tabela, entre as quais destaca-se a seguinte:

$$k = \sqrt{n}. \tag{3.2}$$

Em geral, k não é um número natural. Nesse caso, opta-se pelo que melhor se adequar aos dados. Geralmente, arredonda-se o número para o inteiro imediatamente superior, o que será adotado nesse material.

Já a **amplitude** de um intervalo de classe, baseia-se na **amplitude total dos dados**, que é dada por

$$A = V_{Máximo} - V_{Mínimo} \tag{3.3}$$

em que $V_{Máximo}$ e $V_{Mínimo}$ representam o maior e menor valor presente no conjunto de dados, respectivamente. Uma vez determinados os valores de *A* e *k*, deve-se obter a amplitude do intervalo de classe, a qual é dada por meio da expressão

$$H = \frac{A}{k}. \tag{3.4}$$

Observação: Em determinadas situações o valor de H pertence ao conjunto dos número irracionais. Nesse caso, deve-se ajustar o valor de A, o qual será A', para que esse seja inteiro. Isso originará um erro (E = A' - A), que entre diferentes abordagens, pode ser acrescido no ao valor de L_s da última classe.

Após determinar os limites dos I_c, isto é, **limite inferior** (L_i) e **limite superior** (L_s), calcula-se a média de cada intervalo, digamos (x_i), por meio da expressão

$$x_i = \frac{L_i + L_s}{2}. \tag{3.5}$$

Observação: Deve-se tomar cuidado, pois a notação para média é a mesma utilizada para representar a ocorrência de um elemento da variável discreta.

Para aplicação da metodologia supracitada, considere o exemplo a seguir.

Exemplo 24 *A variável idade, apresentada na Tabela 8, é contínua. Nesse sentido, para construir sua Tabela de frequência, ou distribuição de frequência, tem-se que:*

a) *Organizar os dados em ordem crescente.*

16 18 18 18
18 18 18 19 19 19
19 19 19 19 19
19 20 20 21 21

b) *Calcular o valor de k.*

$$k = \sqrt{20} = 4,47 \approx 5$$

c) *Calcular o valor de A.*

$$A = 21 - 16 = 5$$

d) *Calcular o valor de H.*

$$H = \frac{5}{5} = 1$$

e) *Construção da tabela.*

Tabela 16 – Distribuição das idades dos entrevistados

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[16;17[	16,50	1	5%	1	5%
2	[17;18[	17,50	0	0%	1	5%
3	[18;19[	18,50	6	30%	7	35%
4	[19;20[	19,50	9	45%	16	80%
5	[20;21]	20,50	4	20%	20	100%

f) *Construção do Gráfico (Histograma)*

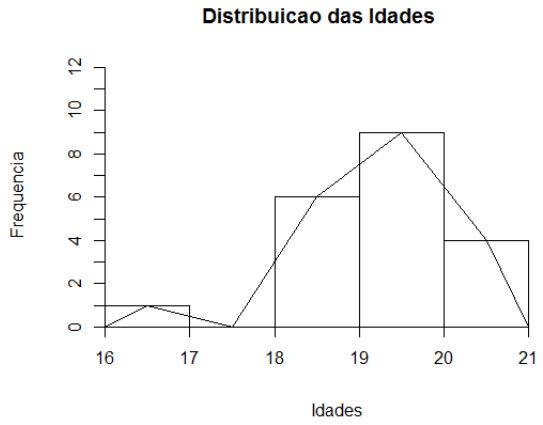


Figura 13 – Histograma para distribuição das idades.

g) *Comandos para construção da tabela e do histograma no R.*

```
classes<-seq(min(dat1$Idade), max(dat1$Idade),length.out
   =6)
h<-hist(dat1$Idade, ylim=c(0,12),freq=TRUE, axes=FALSE,
   breaks=classes, main="Distribuicao das Idades",xlab="
   Idades", ylab="Frequencia",right=FALSE)
axis(1, at = seq(16, 21, by = 1), pos = 0)
axis(2, at = seq(0, 12, by = 1), pos = 16)

lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$
   counts, 0), type = "l")

tabela<-table.freq(h)
tabela
```

Exemplo 25 *A variável altura, apresentada na Tabela 8, é contínua. Nesse sentido, para construir sua Tabela de frequência, ou distribuição de frequência, tem-se que:*

a) *Organizar os dados em ordem crescente.*

1,60 1,62 1,63 1,64 1,65 1,65 1,69 1,69 1,70 1,70 1,70 1,73 1,74 1,76 1,76 1,79 1,85 1,85 1,86 1,90

b) *Calcular o valor de k.*

$$k = \sqrt{20} = 4,47 \approx 5$$

c) *Calcular o valor de A.*

$$A = 1,90 - 1,60 = 0,30$$

d) *Calcular o valor de H.*

$$H = \frac{0,30}{5} = 0,06$$

e) *Construção da tabela.*

Tabela 17 – Distribuição das alturas dos entrevistados

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[1,60;1,66[	1,63	6	30%	6	30%
2	[1,66;1,72[	1,69	5	25%	11	55%
3	[1,72;1,78[	1,75	4	20%	15	75%
4	[1,78;1,84[	1,81	1	5%	16	80 %
5	[1,84;1,90]	1,87	4	20%	20	100%

f) *Construção do Gráfico (Histograma) - Figura 14*

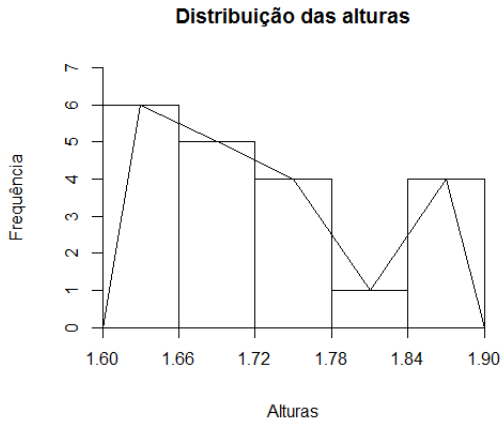


Figura 14 – Histograma para distribuição das alturas.

g) *Comandos para construção da tabela e do histograma no R.*

```
classes<-seq(min(dat1$Altura), max(dat1$Altura),length.
    out=6)
h<-hist(dat1$Altura, ylim=c(0,7),freq=TRUE, axes=FALSE,
    breaks=classes, main="Distribuicao das alturas",xlab="
    Alturas", ylab="Frequencia",right=FALSE)
axis(1, at = seq(1.60, 1.90, by = 0.06), pos = 0)
axis(2, at = seq(0, 8, by = 1), pos = 1.60)

lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$
    counts, 0), type = "l")

tabela<-table.freq(h)
tabela
```

Exemplo 26 *A variável peso, apresentada na Tabela 8, é contínua. Nesse sentido, para construir sua Tabela de frequência, ou distribuição de frequência, tem-se que:*

a) *Organizar os dados em ordem crescente.*

49 51 51 52 53 54 55 57 58 58,5 60 65 66,4 67 68 70 72 78 78 81

b) *Calcular o valor de k.*

$$k = \sqrt{20} = 4,47 \approx 5$$

c) *Calcular o valor de A.*

$$A = 81 - 49 = 32$$

d) *Calcular o valor de H.*

$$H = \frac{32}{5} = 6,4$$

e) *Construção da tabela.*

Tabela 18 – Distribuição dos pesos dos entrevistados

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[49,0;55,4[	52,2	7	35%	7	35%
2	[55,4;61,8[	58,6	4	20%	11	55%
3	[61,8;68,2[	65	4	20%	15	75%
4	[68,2;74,6[	71,4	2	10%	17	85 %
5	[74,6;81,0]	77,8	3	15%	20	100%

f) *Construção do Gráfico (Histograma)*

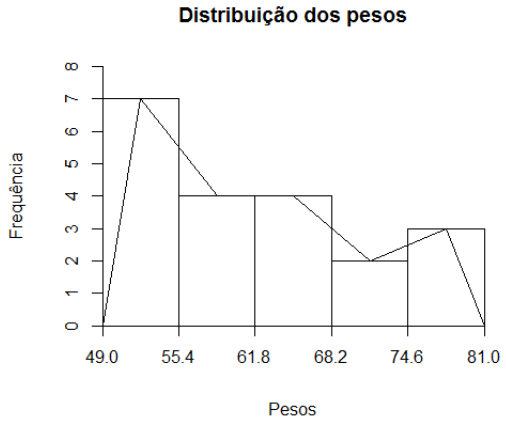


Figura 15 – Histograma para distribuição dos pesos.

g) *Comandos para construção da tabela e do histograma no R.*

```
classes<-seq(min(dat1$Peso), max(dat1$Peso),length.out=6)
h<-hist(dat1$Peso, ylim=c(0,8),freq=TRUE, axes=FALSE,
   breaks=classes, main="Distribuicao dos pesos",xlab="
   Pesos", ylab="Frequencia")
axis(1, at = seq(49, 81, by = 6.4), pos = 0)
axis(2, at = seq(0, 8, by = 1), pos = 49)
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$
   counts, 0), type = "l")

tabela<-table.freq(h)
tabela
stat.freq(h)
```

Os gráficos apresentados nas Figuras 13, 14 e 19 são gráficos de colunas justapostas, chamados de **histogramas**, que representam uma distribuição de frequência para dados contínuos ou uma variável discreta quando esta apresentar muitos valores distintos. No eixo horizontal são dispostos os limites das classes segundo as quais os dados foram agrupados enquanto que o eixo vertical corresponde às frequências absolutas ou relativas das mesmas.

Já as linhas (**polígonos de frequência**) inseridas nos histogramas, são gráficos de linha cuja construção é feita unindo-se os pontos de coordenadas de abscissas correspondentes aos pontos médios de cada classe e as ordenadas, às frequências absolutas ou relativas dessas mesmas classes. O polígono de frequência é um gráfico que deve ser fechado no eixo das abscissas. Então,para finalizar sua elaboração, deve-se acrescentar à distribuição, uma classe à esquerda e outra à direita, ambas com frequências zero. Tal procedimento permite que a área sob a linha de frequências seja igual à área do histograma. Uma das vantagens da aplicação de polígonos de frequências é que, por serem gráficos de linhas, permitem a comparação entre dois ou mais conjuntos de dados por meio da superposição dos mesmos.

Exemplo 27 *Construir a tabela de frequência e histograma para a variável IMC (Tabela 8).*

3.4 Exercícios de Aplicação

1. Diferenciar variáveis qualitativas de variáveis quantitativas e citar exemplos.
2. Descrever o que uma população e o que é uma amostra e citar exemplos.
3. Para que serve um histograma?
4. O que é um intervalo de classe?
5. Explique a diferença entre frequência absoluta e frequência relativa.
6. O que é frequência acumulada?
7. Os dados abaixo representam o lucro (em milhões de reais), de uma indústria metalúrgica, dos últimos 50 meses.

116	133	143	103	135	136	123	123	123	119
124	118	121	131	142	129	124	119	120	159
132	124	125	136	122	113	137	118	130	119
143	124	121	124	110	116	104	114	126	124
147	117	120	127	118	118	117	115	124	124

   O responsável pela área financeira deve realizar uma apresentação aos acionistas dessa empresa. Antes da apresentação essa pessoa deve saber quais as características dos dados, para que possa organizá-los da melhor forma possível. Nesse caso, supondo que você é essa pessoa,

   a) Classificar aos dados quanto o tipo de variável;

   b) Construir distribuição frequência;

   c) Construir o gráfico associado a essa variável.

8. Acidentes de trabalho ocorrem com frequência nas mais diversas ocasiões, como por exemplo em empresas da construção civil. A responsável pelo RH de uma empresa de pequeno porte fez o levantamento dos últimos 30 meses, do número de funcionários que tiveram algum acidente, obtendo os dados abaixo.

0	1	0	1	0	0	0	0	2	3	0	1	2	3	4
0	0	0	1	4	1	1	0	0	3	5	1	0	0	1

   Dada a necessidade de estabilizar esse cenário e propor ações preventivas, deve-se conhecer a estrutura dos dados. Nesse caso, deve-se:

   a) Classificar aos dados quanto o tipo de variável.

   b) Construir distribuição frequência.

c) Construir o gráfico associado a essa variável.

9. Um banco selecionou ao acaso 70 contas de pessoas físicas em uma agência, em determinado dia, obtendo os seguintes saldos em milhares de reais:

48 56 56 59 60 60 61 61 61 62 63 63 64 65 65 65 66 67 67 67 67 67
67 68 68 68 69 70 70 70 70 70 70 70 71 71 71 71 71 71 72 72 72 72 72
73 73 73 73 73 74 74 74 74 75 75 75 77 77 78 78 80 81 82 82 84 84 86 93

a) Classificar aos dados quanto o tipo de variável.

b) Construir distribuição frequência.

c) Construir o gráfico associado a essa variável.

10. Os dados abaixo representam o consumo de energia ao longo de 50 semanas, de uma empresa de TI.

74,8 74,0 74,7 74,4 75,9 76,8 74,3 74,9 77,0 75,1
75,0 74,6 72,9 72,9 73,6 76,8 74,8 74,4 73,8 75,1
75,3 73,4 74,7 73,4 74,2 74,9 74,5 77,1 74,6 74,8
76,4 73,2 76,5 75,6 73,5 76,2 74,7 76,0 75,8 77,3
76,3 74,1 75,0 76,0 74,7 75,2 77,7 74,7 73,3 74,3

O responsável pela administração deseja ter um panorama dos consumos. Nesse caso, faça o que se pede a seguir.

a) Classificar aos dados quanto o tipo de variável.

b) Construir distribuição frequência.

c) Construir o gráfico associado a essa variável.

11. Uma indústria embala peças em caixas com 100 unidades. O controle de qualidade selecionou 48 caixas na linha de produção e anotou em cada caixa o número de peças defeituosas. Obteve os seguintes dados:

2 0 0 4 3 0 0 1 0 0 1 1
2 1 1 1 1 1 1 0 0 0 3 0
0 0 2 0 0 1 1 2 0 2 0 0
0 0 0 0 0 0 0 0 0 0 1 0

a) Classificar aos dados quanto o tipo de variável.

b) Construir distribuição frequência.

c) Construir o gráfico associado a essa variável.

3.4.1 Gabarito

7. k = 7; A = 103-159 = 56; H = 8;

Tabela 19 – Distribuição do lucro obtido ao londo do tempo.

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[103;111[	107	3	6%	3	6%
2	[111;119[	115	11	2%	14	28%
3	[119;127[	123	21	42%	35	70%
4	[127;135[	131	6	12%	41.00	82%
5	[135;143[	139	5	10%	46	92%
6	[143;151[	147	3	6%	49	98%
7	[151;159]	155	1	2%	50	100%

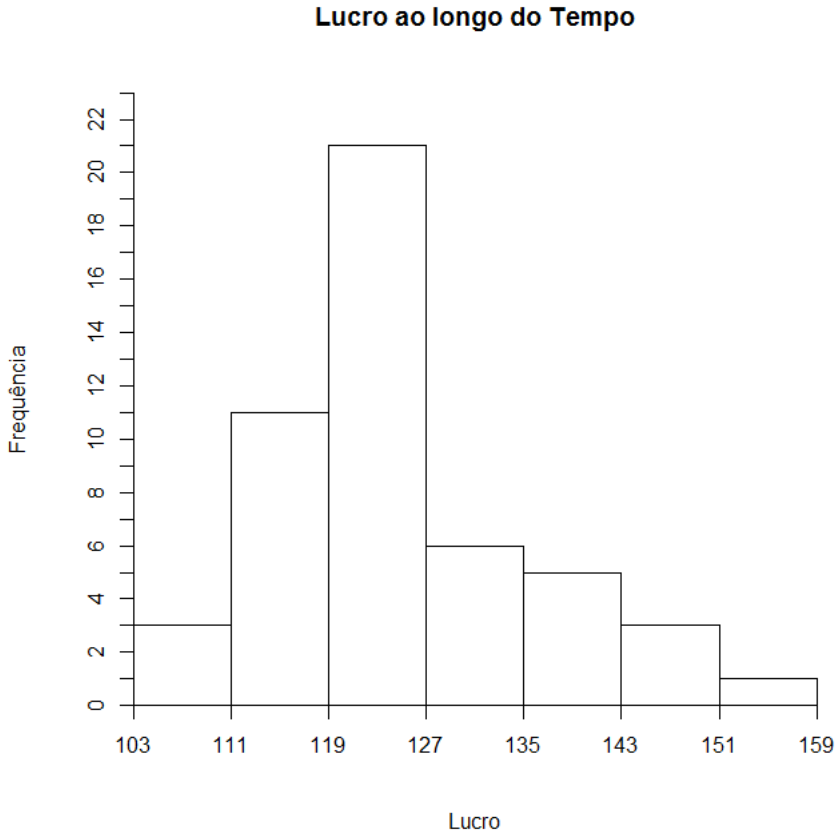


Figura 16 – Histograma para LUCRO obtido.

Tabela 20 – Distribuição para o número de acidentes de trabalho.

x_i	f_i	F_i	r_i	R_i
0	14	14	46,67%	46,67%
1	8	22	26,67%	73,33%
2	2	24	6,67%	80,01%
3	3	27	10%	90,01%
4	2	29	6,67%	96,67%
5	1	30	3,33%	100%

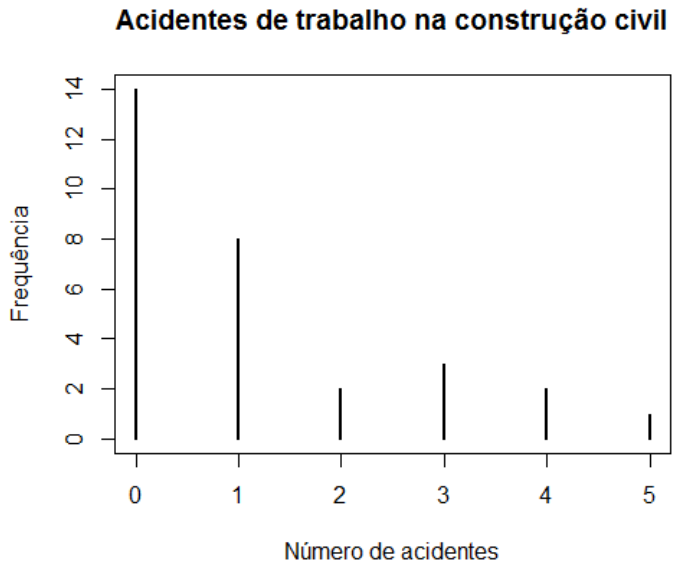


Figura 17 – Gráfico de bastões.

8.

9. k = 9; A = 93-48 = 45; H = 5;

Tabela 21 – Distribuição do saldo disponível em conta.

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[48;53[	50,50	1	1,40%	1	1,40%
2	[53;58[	55,50	2	2,90%	3	4,30%
3	[58;63[	60,50	7	12,90%	12	17,10%
4	[63;68[	65,50	14	21,40%	27	38,60%
5	[68;73[	70,50	22	34,30%	51	72,90%
6	[73;78[	75,50	14	15,70%	62	88,60%
7	[78;83[	80,50	6	5,70%	66	94,30%
8	[83;88[	85,50	3	4,30%	69	98,60%
9	[88;93]	90,50	1	1,40%	70	100,00 %

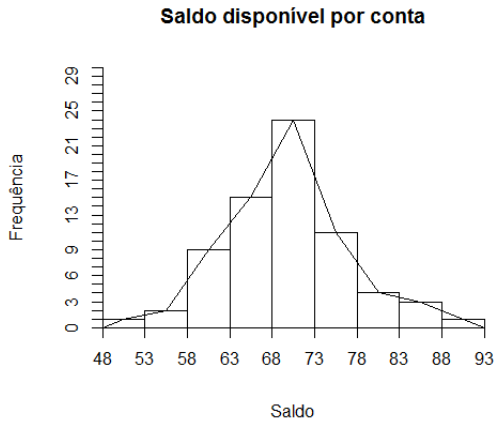


Figura 18 – Histograma para saldo disponível em conta.

10. k = 8; A = 77,7 - 72,9 = 4,8; H = 0,6;

Tabela 22 – Consumo de energia ao longo do tempo.

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[72,90;73,50[	73,20	6	14%	7	14%
2	[73,50;74,10[	73,80	4	8%	11	22%
3	[74,10;74,70[	74,40	9	26%	24	48%
4	[74,70;75,30[	75,00	15	22%	35	70%
5	[75,30;75,90[	75,60	3	6%	38	76%
6	[75,90;76,50[	76,20	6	12%	44	88%
7	[76,50;77,10[	76,80	4	8%	48	96%
8	[77,10;77,70[	77,40	3	4 %	50	100%

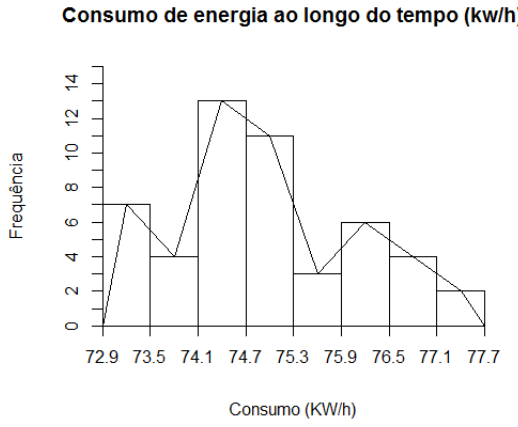


Figura 19 – Histograma para o consumo de energia ao longo do tempo.

11.

Tabela 23 – Distribuição das peças defeituosas

x_i	f_i	F_i	r_i	R_i
0	28	28	58,33%	58,33%
1	12	40	25%	83,33%
2	5	45	10,41%	93,74%
3	2	47	4,16%	97,9%
4	1	48	2,10%	100%

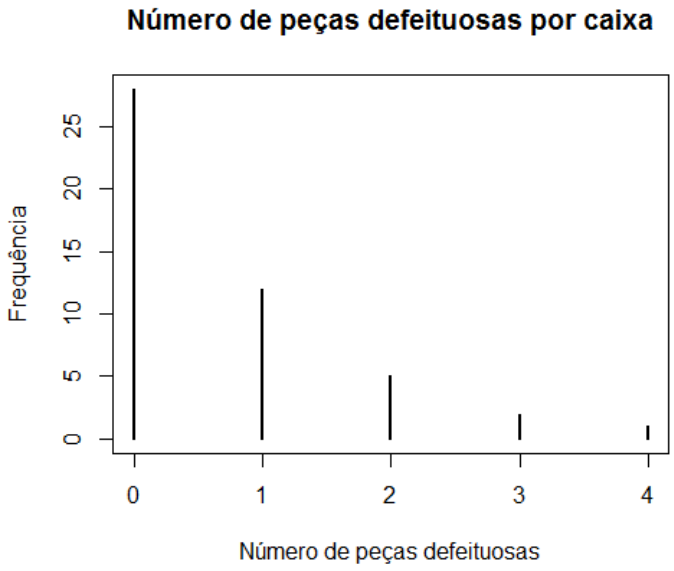


Figura 20 – Gráfico de bastões.

Capítulo 4

Medidas Descritivas

Em qualquer área do conhecimento, as informações costumam não ser constantes. Assim, é necessário que a medida que esteja associada a variável avaliada, a qual esteja indicado o valor em torno do qual se agrupam os dados, seja associada a uma medida que faça referência à **variabilidade** que reflita tal flutuação.

Portanto, o próximo passo e objetivo desta seção consistirá em definir alguns tipos de medidas que sintetizem os dados estudados, ou seja, a partir de um conjunto de informações organizadas em uma distribuição de frequência, ou até mesmo para os dados brutos (ou melhor, uma série de observações sem ordenação), pretendemos descrevê-los mediante duas ou três quantidades sintéticas.

Nesse sentido, podem ser examinadas várias características, sendo as mais comuns:

- A tendência central dos dados;
- A dispersão ou variação em relação a esse centro;
- Os dados que ocupam certas posições;
- Simetria dos dados;
- A forma na qual os dados se agrupam.

As medidas descritivas, com frequência utilizadas, são ou de tendência central, como por exemplo

- Média;
- Mediana;

- Moda,

e as medidas de dispersão ou variabilidade, como por exemplo

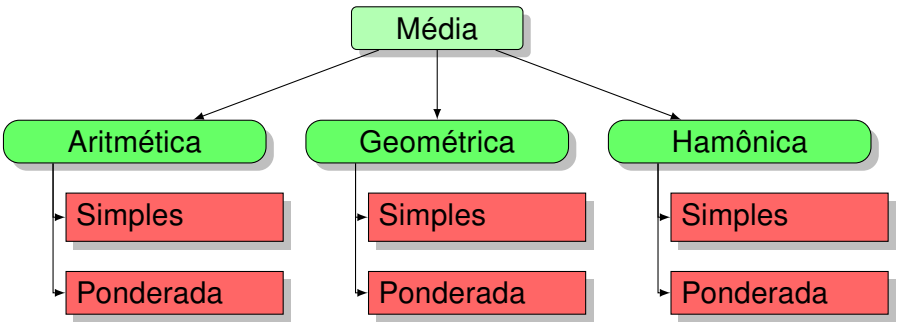
- Variância;
- Desvio padrão;
- Coeficiente de variação.

4.1 Medidas de Tendência Central ou Locação

Mostram um valor representativo em torno do qual os dados tendem a agrupar-se com maior ou menor frequência. São utilizadas para sintetizar em um único número o conjunto de dados observados.

4.1.1 Média

De acordo com o problema abordado, é possível utilizar diversas expressões que fornecem um valor médio, que sintetizará as informações coletas, levando em consideração as características do problema em estudo. Por meio do diagrama a seguir é possível identificar as diferentes nomenclaturas utilizadas para o cálculo de médias. Todavia, nos limitares ao uso das médias aritméticas simples e ponderadas.



4.1.1.1 Média Aritmética Simples para dados brutos

Para uma sequência numérica X: $x_1, x_2, \ldots, x_n$ a média aritmética simples, que designaremos por $\overline{x}$ onde:

$$\overline{x} = \frac{x_1 + \ldots + x_n}{n}. \tag{4.1}$$

Exemplo 28 *Calcular a média aritmética simples dos valores 2, 0, 5 e 3.*

4.1.1.2 Média Aritmética Ponderada para dados brutos

Para uma sequência numérica $X : x_1, x_2, \ldots, x_n$, em que cada valor possui um peso $p_1, p_2, ..., p_n$ respectivamente, a média aritmética ponderada, será calculada por:

$$\overline{x} = \frac{x_1 p_1 + \ldots + x_n p_n}{p_1 + \ldots + p_n}. \tag{4.2}$$

Exemplo 29 *Na disciplina de estatística foram feitas duas avaliações. A primeira delas com peso um (p_1=1) e a segunda com peso dois (p_2=2). Se um aluno tirou 4,5 na primeira prova e 7 na segunda prova, qual foi a sua média final?*

4.1.1.3 Média aritmética para dados tabelados

Ao calcular a média aritmética para dados tabelados, deve-se levar em conta que o número de elementos pertencentes a classe pode diferir, o que pode ser compreendido como um peso. Nesse caso, da expressão que fornece a média aritmética ponderada de um conjunto de dados brutos, tem-se

$$\overline{x} = \frac{x_1 p_1 + \ldots + x_n p_n}{p_1 + \ldots + p_n},$$

e adaptando-a para o caso de dados tabelados, surge a expressão a seguir,

$$\overline{x} = \frac{x_1 f_1 + \ldots + x_n f_n}{f_1 + \ldots + f_n}, \tag{4.3}$$

sendo x_i a média dos extremos do I_c da i-ésima linha e f_i a frequência absoluta da i-ésima linha.

Exemplo 30 *Para os dados 8, uma das variáveis contínuas, em que construiu-se a tabela de frequência para os dados coletados é o peso dos entrevistados. A tabela de frequência é dada a seguir.*

Tabela 24 – Dados do Exemplo 30

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[49,0;55,4[	52,2	7	35%	7	35%
2	[55,4;61,8[	58,6	4	20%	11	55%
3	[61,8;68,2[	65	4	20%	15	75%
4	[68,2;74,6[	71,4	2	10%	17	85 %
5	[74,6;81,0]	77,8	3	15%	20	100%

Por meio da expressão 4.3, segue o resultado para a média.

$$\overline{x} = \frac{52,2 \times 7 + 58,6 \times 4 + 65 \times 4 + 71,4 \times 2 + 77,8 \times 3}{7 + 4 + 4 + 2 + 3} = \frac{1236}{20} = 61,8$$

Com o resultado obtido, estima-se que o peso médio dos entrevistados é aproximadamente 61,8 m. Para resolver no R, basta executar os comandos a seguir, considerando que já foi feita a entrada de dados. Ressalta-se que é o mesmo conjunto de dados utilizado no Capítulo 3.

```
classes<-seq(min(dat1$Peso), max(dat1$Peso),length.out=6)
h<-hist(dat1$Peso, ylim=c(0,8),freq=TRUE, axes=FALSE,breaks=
   classes, main="Distribuicao dos pesos",xlab="Pesos", ylab="
   Frequencia")
axis(1, at = seq(49, 81, by = 6.4), pos = 0)
axis(2, at = seq(0, 8, by = 1), pos = 49)
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts,
   0), type = "l")

tabela<-table.freq(h)
tabela
stat.freq(h)
```

O comando stat.freq(h) retorna algumas medidas descritivas, entre as quais a média.

Exemplo 31 *O Nielsen Home Technology Report fornece informações sobre a tecnologia e aparelhos domésticos e sua utilização. Os dados a seguir correspondem ao tempo em horas do uso de computadores pessoais, durante uma semana, para uma amostra de 50 pessoas.*

a) *Organizar os dados em Rol;*

4,1 1,5 10,4 5,9 3,4 5,7 1,6 6,1 3,0 3,7
3,1 4,8 2,0 16,7 5,4 4,2 3,9 4,1 11,1 3,5
4,1 4,1 8,8 5,6 4,3 3,3 7,1 10,3 6,2 7,6
10,8 2,8 9,5 12,9 12,1 12,1 0,7 4,0 9,2 4,4 5,7
7,2 6,1 5,7 5,9 4,7 3,9 3,7 3,1 6,1 3,1

b) *Construir a tabela de distribuição de frequência para os dados;*

c) *Construir o gráfico para a distribuição de frequência do item anterior;*

d) *Calcular o tempo médio em horas gasto com uso dos computadores pessoais.*

Resolvendo no R, na ordem das perguntas, temos:

```
#a)
dados<-c(4.1, 1.5,10.4, 5.9, 3.4, 5.7, 1.6, 6.1, 3.0, 3.7,
3.1, 4.8, 2.0, 16.7, 5.4, 4.2, 3.9, 4.1, 11.1, 3.5,
4.1, 4.1, 8.8, 5.6, 4.3, 3.3, 7.1, 10.3, 6.2, 7.6,
10.8, 2.8, 9.5, 12.9, 12.1, 12.1, 0.7, 4, 9.2, 4.4,
5.7, 7.2, 6.1, 5.7, 5.9, 4.7, 3.9, 3.7, 3.1, 6.1, 3.1)

sort(dados)
```

```
#b) e c)
```

```
k<-sqrt(length(dados))
k1<-ceiling(k)
classes<-seq(min(dados), max(dados),length.out=(k1+1))
classes

X11()
h<-hist(dados, ylim=c(0,25),freq=TRUE, axes=FALSE,breaks=
   classes, main="Tempo de Uso do computador", xlab="Tempo em
    Horas", ylab="Frequencia",right=FALSE)
axis(1, at = seq(0.7, 16.7, by = 2), pos = 0)
axis(2, at = seq(0, 30, by = 1), pos = 0.7)

tabela<-table.freq(h)
tabela

#d)
descritivas<-stat.freq(h)
descritivas$mean
```

Como descrito no capítulo anterior, uma variável pode ser quantitativa discreta ou contínua. Nesse sentido, se for necessário calcular a média aritmética para os dados relacionados a uma variável quantitativa discreta, deve-se tomar cuidado que o valor que x_i assume está associado a ocorrência da variável analisada.

Exemplo 32 *Para exemplificar essa situação, considere a Tabela a seguir, resultado da organização dos dados para variável Q03 da 8.*

Tabela 25 – Tabela de Frequência

x_i	f_i	F_i	r_i	R_i
0	2	2	10%	10%
1	1	3	5%	15%
2	3	6	15%	30%
3	5	11	25%	55%
4	2	13	10%	65%
5	2	15	10%	75%
6	2	17	10%	85%
7	2	19	10%	95%
8	1	20	5%	100%

Nesse caso, a média aritmética é dada por

$$\overline{x} = \frac{0\times 2+1\times 1+2\times 3+3\times 5+4\times 2+5\times 2+6\times 2+7\times 2+8\times 1}{2+1+3+5+2+2+2+2+1} = \frac{74}{20} = 3,7$$

Com o resultado obtido, estima-se que o número médio de vestibulares prestados sem aprovação é aproximadamente 4.

Exemplo 33 *O responsável pelo RH de uma empresa de TI coletou informações sobre o número de funcionários admitidos ao longo dos últimos 20 meses, obtendo os seguintes dados.*

5 4 1 9 5 2 8 5 4 3 9 6 2 9 7 6 3 5 9 5 4

a) *Construir a tabela de distribuição de frequência para os dados;*

b) *Construir o gráfico para a distribuição de frequência do item anterior;*

c) *Calcular o número médio de funcionários admitidos no período avaliado.*

Para resolver no R, basta utilizar os comandos a seguir:

```
#a)
dados<-c(5,4,1,9,5,2,8,5,4,3,9,6,7,2,9,6,3,5,9,5,4)
Tabela<-data.frame(table(dados))
Tabela

#b)
plot(table(dados),main="Numero de funcionarios admitidos",
   xlab="Numero de Funcionarios",ylab="Frequencia")

#c)
mi<-c()
xi<-c()
fi<-c()

for(i in 1:9)
{
xi[i]<-Tabela$dados[i]
fi[i]<-Tabela$Freq[i]
mi[i]<-xi[i]*fi[i]
soma<-sum(mi)
media<-soma/length(dados)
}
media
```

4.1.2 Mediana

É um valor real que separa o conjunto de dados em duas partes deixando à sua esquerda o mesmo número de elementos que a sua direita. Portanto, a mediana é um valor que ocupa a posição central em uma série.

4.1.2.1 Mediana para dados brutos

Passos para determinação da mediana:

1º Organizar os dados em ordem crescente;

2º Observar se o número de dados da amostra é par ou ímpar.

- Se o n^o de elementos da amostra é ímpar, a mediana é o elemento de posição,

$$M_d = \left(\frac{n+1}{2}\right)^o. \tag{4.4}$$

- Se o n^o de elementos é par, a mediana é a média entre os elementos centrais, ou utilizar a expressão a seguir.

$$M_d = \frac{\left(\frac{n}{2}\right)^o + \left(\frac{n}{2}+1\right)^o}{2}. \tag{4.5}$$

Para aplicação dessa metodologia, considere os exemplos a seguir.

Exemplo 34 *Obter a mediana dos conjuntos A = {6,7,9,12,3} e B = {12,5,8,4,9,14}, onde os elementos de A referem-se ao número de dias que choveram durante 5 meses e para o conjunto B são observações do número de dias de chuva durante 6 meses.*

a) *Mediana para o conjunto A é 9;*

```
A<-c(6,7,9,12,3)
mediana<-median(A)
mediana
```

b) *Mediana para o conjunto B é a média entre os números 8 e 9, isto é, 8,5.*

```
B<-c(12,5,8,4,9,14)
mediana<-median(B)
mediana
```

4.1.2.2 Mediana para dados Tabelados

- Para dados discretos o termo mediano é obtido com os auxílios das expressões anteriores;
- Neste caso precisaremos determinar a posição do termo mediano, e consequentemente seu valor através da expressão abaixo

No caso de dados contínuos, a mediana pode ser obtida por meio da expressão 4.6.

$$M_d = L_i + \frac{\frac{n}{2} - F_{ant}}{f_{md}}.h \tag{4.6}$$

- L_i - Limite inferior do Ic da classe da mediana
- F_{ant} - Frequência absoluta acumulada anterior à classe da mediana;
- f_{md} - Frequência absoluta da classe da mediana;
- h - Amplitude do intervalo de classe.

Exemplo 35 *A tabela abaixo refere-se a um experimento que analisou o número de acidentes de trabalho em 40 diferentes obras na construção de edifícios. Complete-a e determine a mediana.*

Tabela 26 – Número de acidentes de trabalho ao longo do tempo

x_i	f_i	r_i	F_i	R_i
0	1			
1	0			
2	6			
3	4			
4	6			
5	6			
6	6			
7	5			
8	5			
9	1			

Exemplo 36 *Uma imobiliária gerencia o aluguel de residências particulares, segundo a tabela de frequência abaixo. Calcular o aluguel mediano para essa distribuição de valores.*

Tabela 27 – Dados do Exemplo 36

Classe	$[L_i; L_s[$	x_i	f_i
1	[0;200[	100	30
2	[200;400[	300	52
3	[400;600[	500	28
4	[600;800[	700	7
5	[800;1.000]	900	3

Entre as propriedades da mediana podemos destacar, a vantagem de não ser afetada por valores extremos, pois não depende dos valores que a variável assume, além da ordem das mesmas. Seu uso é adequado para distribuições assimétricas;

4.1.3 Moda

4.1.3.1 Moda para dados brutos

Chamaremos de moda qualquer máximo relativo da distribuição de frequência, ou seja, elemento que tiver a maior frequência. A moda tem alguns inconvenientes como a possibilidade de não ser determinada e não ser única em todas as situações.

Exemplo 37 *Para os conjuntos abaixo podemos observar a moda.*

a) *2,8,3,5,4,4,3,5,5,1 a Moda = 5 - Unimodal.*

b) *6,10,5,6,10,2 a Moda = 6 e 10 - Bimodal.*

c) 2,2,5,8,5,8 a Moda não existe $\Rightarrow$ Amodal

4.1.3.2 Moda para dados Tabelados

- Para dados discretos a moda é obtida por meio da definição.
- Neste caso precisaremos determinar a classe que contenha a moda, e consequentemente seu valor por intermédio da expressão abaixo.

$$M_o = I_{mo} + \frac{f_{mo} - f_{ant}}{2 * f_{mo} - (f_{post} + f_{ant})}.h \tag{4.7}$$

- I_{mo} - Limite inferior do Ic da classe modal;
- f_{post} - Frequência absoluta posterior à classe modal;
- f_{ant} - Frequência absoluta anterior à classe modal;
- f_{mo} - Frequência absoluta da classe modal;
- h - Amplitude do intervalo de classe.

sendo essa expressão proposta por Czuber, a qual recebe o nome de Moda de Czuber.

Exemplo 38 *A tabela de frequência abaixo refere-se aos valores de 54 notas fiscais emitidas na mesma data, selecionadas em uma loja de departamentos.*

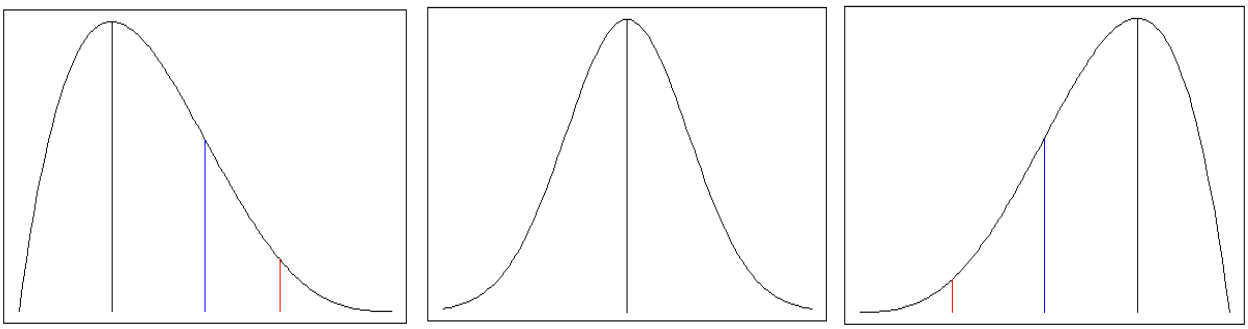
Complete a tabela acima e obtenha o valor mais frequente.

Tabela 28 – Tabela de frequência do Exemplo 38

Classe	$[L_i; L_s[$	f_i
1	[0;50[	5
2	[50;100[	28
3	[100;150[	12
4	[150;200[	2
5	[200;250]	1
6	[250;300]	1

4.1.4 Relação entre média, moda e mediana

A partir dos conceitos de medidas de tendência central, é possível estabelecer relações entre os valores obtidos para média, moda e mediana. Observe a Figura 21.



(a) Dist. Assimétrica à Direita (b) Dist. Simétrica (c) Dist. Assimétrica à Esquerda

Figura 21 – Relação entre média, moda e mediana

As Figuras 21a, 21b e 21c representam o formato que a distribuição dos dados pode assumir, isto é, uma formato assimétrico à direita, simétrico ou assimétrico à esquerda. Nessas situações, as medidas descritivas (média, moda, mediana) respeitam uma ordem de grandeza. No primeiro caso, assimetria à direita, temos que a moda (linha preta) é menor que menor que a mediana (linha azul), que por sua vez é inferior a média (linha vermelha). No caso de uma distribuição apresentar formato simétrico, ocorre que as três medidas citadas assumem o mesmo valor. Por fim, no caso assimétrico à esquerda, a média (linha vermelha) é menor que menor que a mediana (linha azul), que por sua vez é inferior moda (linha preta). Na prática, na grande maioria das vezes os dados possuem formatos assimétricos e sua análise deve ser baseada em modelos estatísticos que captem essa característica.

Em um contexto geral, em distribuições unimodais (distribuições com apenas uma moda), a mediana está frequentemente compreendida entre a média e a moda. Em distribuições que apresentam certa inclinação, é melhor analisarmos o conjunto de dados

utilizando a mediana. No entanto, em estudos estatísticos é mais comum resumirmos o conjunto de dados em torno da média.

4.2 Medidas de Dispersão ou Variabilidade

As medidas de dispersão medem a variabilidade dos dados em estudo. Permitem verificar se o conjunto de dados é homogêneo ou heterogêneo. São medidas estatísticas que medem a dispersão dos dados, em torno de um valor central. As medidas a serem trabalhadas são:

- Amplitude;
- Variância;
- Desvio padrão;
- Coeficiente de variação.

4.2.1 Amplitude

Amplitude total ou máxima é a diferença entre o maior e o menor valor de um conjunto de dados.

$$A = V_{màx} - V_{min} \tag{4.8}$$

Esta é uma medida que não fornece uma informação precisa acerca dos dados fornecidos.

4.2.2 Variância

É uma medida de dispersão que mede a variabilidade de um conjunto de dados. Quando desejamos obter a variância de uma população, o que geralmente não é possível pelo fato de desconhecermos toda a população, representamos variância por σ^2. Quando desejamos obter a variância amostral, representamos por s^2, sendo s^2 uma estimativa de σ^2. Ela é definida como a média das diferenças quadráticas de *n* valores em relação à sua média aritmética.

Quando os dados analisados são dados brutos, as expressões que fornecem o valor da variância são dados por:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{4.9}$$

sendo que x_i corresponde a cada observação e $\overline{x}$ a média do conjunto.

4.2.3 Propriedades da Variância

- Para qualquer distribuição a variância é sempre uma quantidade positiva .
- Se os valores das observações são todos iguais então a variância é zero.
- Variância de uma constante é zero.
- Se somarmos ou subtrairmos uma mesma constante de cada elemento de um conjunto de dados sua variância não se altera.
- Se multiplicarmos ou dividirmos a cada elemento de um conjunto de dados por uma mesma constante sua variância fica multiplicada ou dividida pelo quadrado da constante.

4.2.3.1 Variância para dados brutos

Para obtermos a variância de dados brutos, podemos utilizar as expressões apresentadas em 4.9.

Exemplo 39 *Um painel de economistas apresentou previsões para a economia dos Estados Unidos, referentes aos primeiros seis meses de 2007 (The Wall Street Journal, 2 de janeiro de 2007). As mudanças percentuais amostrais no produto interno bruto (PIB) previstas pelos 30 economistas são as seguintes:*

2,6	3,1	2,3	2,7	3,4	0,9	2,6	2,8	2,0	2,4
2,7	2,7	2,7	2,9	3,1	2,8	1,7	2,3	2,8	3,5
0,4	2,5	2,2	1,9	1,8	1,1	2,0	2,1	2,5	0,5

Qual foi é a variação no PIB de acordo com a opinião dos 30 economistas consultados?

Solução: Nesse caso, para obter a variância amostral, basta construir cada termo para ser incluso na expressão que fornece o valor de s^2. Primeiramente calcula-se a média das observações.

$$\overline{x} = \frac{2,6 + 3,1 + \ldots + 2,5 + 0,5}{30} = 2,3 \tag{4.10}$$

Posteriormente, é possível construir a tabela abaixo.

Tabela 29 – Cálculo das diferenças entre observações e média estimada.

	x_i	$\overline{x}$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	*2,6*	*2,3*	*0,3*	*0,09*
2	*3,1*	*2,3*	*0,8*	*0,64*
3	*2,3*	*2,3*	*0*	*0*
4	*2,7*	*2,3*	*0,4*	*0,16*
5	*3,4*	*2,3*	*1,1*	*1,21*
6	*0,9*	*2,3*	*-1,4*	*1,96*
7	*2,6*	*2,3*	*0,3*	*0,09*
8	*2,8*	*2,3*	*0,5*	*0,25*
9	*2*	*2,3*	*-0,3*	*0,09*
10	*2,4*	*2,3*	*0,1*	*0,01*
11	*2,7*	*2,3*	*0,4*	*0,16*
12	*2,7*	*2,3*	*0,4*	*0,16*
13	*2,7*	*2,3*	*0,4*	*0,16*
14	*2,9*	*2,3*	*0,6*	*0,36*
15	*3,1*	*2,3*	*0,8*	*0,64*
16	*2,8*	*2,3*	*0,5*	*0,25*
17	*1,7*	*2,3*	*-0,6*	*0,36*
18	*2,3*	*2,3*	*0*	*0*
19	*2,8*	*2,3*	*0,5*	*0,25*
20	*3,5*	*2,3*	*1,2*	*1,44*
21	*0,4*	*2,3*	*-1,9*	*3,61*
22	*2,5*	*2,3*	*0,2*	*0,04*
23	*2,2*	*2,3*	*-0,1*	*0,01*
24	*1,9*	*2,3*	*-0,4*	*0,16*
25	*1,8*	*2,3*	*-0,5*	*0,25*
26	*1,1*	*2,3*	*-1,2*	*1,44*
27	*2*	*2,3*	*-0,3*	*0,09*
28	*2,1*	*2,3*	*-0,2*	*0,04*
29	*2,5*	*2,3*	*0,2*	*0,04*
30	*0,5*	*2,3*	*-1,8*	*3,24*

Assim, tem-se que

$$s^2 = \frac{17,2}{30-1} = 0,5931,$$

o que implica que a variação entre os dados é de aproximadamente 0,6. Isto é, entre as previsões do PIB de acordo com os 30 economistas é de aproximadamente 0,6 milhões de dólares.

Para resolver no R, basta utilizar os comandos abaixo.

```
dados<-c(2.6 ,3.1 ,2.3 ,2.7 ,3.4 ,0.9 ,2.6 ,2.8 ,2.0 ,2.4,
2.7 ,2.7 ,2.7 ,2.9 ,3.1 ,2.8 ,1.7 ,2.3 ,2.8 ,3.5,
0.4 ,2.5 ,2.2 ,1.9 ,1.8 ,1.1 ,2.0 ,2.1 ,2.5 ,0.5)

variancia<-var(dados)
variancia
```

Exemplo 40 *Obter a variância da variável Q05 apresentada Tabela 8 e interpretar o resultado.*

4.2.3.2 Variância para dados tabelados

Para dados tabelados, devemos levar em consideração que os dados estão divididos em intervalos. Sendo assim, cada um contêm um número diferente de elementos. Portanto devemos considerar a frequência absoluta de cada um. Logo, a variância é obtida através da expressão abaixo:

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1} \tag{4.11}$$

sendo k o número de intervalos de classe, x_i a média de cada intervalo de classe e $\overline{x}$ a média do conjunto.

Exemplo 41 *A Tabela 33 apresenta os dados relativos ao peso dos 20 entrevistados apresentados na Tabela 8. Obter a variância dos pesos.*

Tabela 30 – Distribuição dos pesos dos entrevistados

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[49,0;55,4[	52,2	7	35%	7	35%
2	[55,4;61,8[	58,6	4	20%	11	55%
3	[61,8;68,2[	65	4	20%	15	75%
4	[68,2;74,6[	71,4	2	10%	17	85 %
5	[74,6;81,0]	77,8	3	15%	20	100%

Solução: Inicialmente deve-se obter o peso médio para as 20 observações. Nesse sentido, tem-se:

$$\overline{x} = \frac{52,2 \times 7 + 58,6 \times 4 + 65 \times 4 + 71,4 \times 2 + 77,8 \times 3}{7 + 4 + 4 + 2 + 3} = \frac{1236}{20} = 61,8$$

Por meio da Tabela abaixo, obtemos os termos da expressão que fornece s^2 para dados tabelados.

Tabela 31 – Cálculo da variância para dados tabelados.

x_i	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \times f_i$
52,20	-9,60	92,16	645,12
58,60	-3,20	10,24	40,96
65,00	3,20	10,24	40,96
71,40	9	92,16	184,32
77,80	16	256	768,00
			1679,36

Nesse sentido, tem-se que

$$s^2 = \frac{1679,36}{20 - 1} = 88,38kg$$

sendo que o resultado sugere alta variabilidade para os dados.

Para resolver no R, basta utilizar os comandos a seguir.

```
classes<-seq(min(dat1$Peso), max(dat1$Peso),length.out=6)
h<-hist(dat1$Peso, ylim=c(0,8),freq=TRUE, axes=FALSE,breaks=
   classes, main="Distribuicao dos pesos",xlab="Pesos", ylab="
   Frequencia")
axis(1, at = seq(49, 81, by = 6.4), pos = 0)
axis(2, at = seq(0, 8, by = 1), pos = 49)
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0)
   , type = "l")
```

```
tabela<-table.freq(h)
tabela
stat.freq(h)$variance
```

Exemplo 42 *Anualmente os funcionários de uma determinada empresa são submetidos a avaliação de desempenho. Foram coletados dados das avaliação de alguns funcionários dessa empresa, e apresentados abaixo:*

Tabela 32 – Notas do desempenho na avaliação anual

6 0 6,5 2 5 3,5 4 7 7,5 7 6 4,5 1,5 6,5 6
2 5 9,5 8,2 6,6 4,9 7,2 2,8 9,2 8
4 5,5 6,2 8,8 7 4,5 9,8 7,7 6,4 8,5 1,5 2,7 2,2 5,9 9 8,4 6,7 5,2 5 1

Construir a Tabela de distribuição de frequência e obter a variância para as notas.

Para resolver no R, basta utilizar os comandos a seguir:

```
dados1<-c(6, 0, 6.5, 2, 5, 3.5, 4, 7, 7.5, 7, 6, 4.5,
1.5, 6.5, 6, 2, 5, 9.5, 8.2, 6.6, 4.9, 7.2,5.1,
2.8, 9.2, 8, 4, 5.5, 6.2, 8.8, 7, 4.5, 9.8,5.2,
7.7, 6.4, 8.5, 1.5, 2.7, 2.2, 5.9, 9, 8.4, 6.7)

k<-sqrt(length(dados1))
k1<-ceiling(k)
classes<-seq(min(dados1), max(dados1),length.out=(k1+1))
classes
X11()
h<-hist(dados1, ylim=c(0,15),freq=TRUE, axes=FALSE,breaks=
   classes, main="Notas Anuais",
xlab="Notas", ylab="Frequencia",right=FALSE)
axis(1, at = seq(0, 9.8, by = 1.4), pos = 0)
axis(2, at = seq(0, 15,  by = 1),   pos = 0)
```

```
tabela<-table.freq(h)
tabela

descritivas<-stat.freq(h)
descritivas$variance
```

4.2.3.3 Desvio padrão

A variância não tem a mesma magnitude que as observações. Se quisermos que a medida de dispersão seja da mesma dimensão que as das observações, basta extrair a raiz quadrada. Portanto, defini-se os desvio padrão amostral (s) e desvio padrão populacional (σ^2) como:

$$\sigma = \sqrt{\sigma^2} \qquad s = \sqrt{s^2} \tag{4.12}$$

Exemplo 43 *Considerando a variância amostral obtida no exemplo 41, o desvio padrão associado é dado por*

$$s = \sqrt{88,38} = 9,40kg$$

o que sugere que há uma variação de 9,40 kg entre os pesos dos entrevistados.

4.2.3.4 Coeficiente de Variação

O coeficiente de varição é definido como a razão entre o desvio padrão e a média de um conjunto de dados.

$$CV = \frac{s}{\overline{x}}. \tag{4.13}$$

Ele diminui a dimensionalidade das variáveis e tem em conta a proporção existente entre as médias e o desvio padrão. Este coeficiente nos permite a comparar dois grupos distintos. Quanto maior o coeficiente de variação, maior é a variação entre os dados do grupo avaliado. Em geral, a variabilidade de um conjunto de dados depende da área de pesquisa. Contudo, alguns autores apresentam que se

a) CV $\leq$ 20% a amostra é homogênea;

b) CV $>$ 20% a amostra é heterogênea;

Exemplo 44 *Obter o coeficiente de variação para os dados do exemplo 41.*

Para calcular o desvio padrão no R, basta utilizar a função **sd(nome)**.

4.2.4 Medidas Separatrizes

São valores da variável caracterizados por superar uma certa porcentagem de observações de uma população ou amostra. Suponha que os rendimentos anuais de certo empregado, seja aproximadamente R$15.000,00 e que deseja-se saber como ele se situa dentro de seu grupo empresarial. Para fazer essa análise pode-se usar uma distribuição de porcentagens acumuladas como **percentil**. Ele é obtido dividindo-se a população, organizada em ordem crescente, em 100 partes iguais.

De um modo geral, os percentis, são números reais que dividem a sequência ordenada de dados em partes que contêm a mesma quantidade de elementos da série. É possível ter os seguintes múltiplos.

- Mediana - Divide um conjunto em dois grupos com 50% dos dados cada;
- Quartis - Divide o conjunto em quatro partes Iguais; (Q)
- Quintis - Cada parte fica com 20%; (K)
- Decis - Dividimos em 10 partes iguais o conjunto, onde cada parte fica com 10% dos dados.(D)
- Percentis - Dividimos um conjunto de dados em cem partes, onde cada uma ficará com 1% dos elementos.
- Os quartis, quintis e decis são múltiplos dos percentis.

Em síntese, determinar a mediana, um quartil, um quintil ou um percentil é equivalente a determinar o seu percentil equivalente.

4.2.4.1 Percentis para dados brutos

Para determinar estas medidas basta seguir os itens:

1º Calcular *i*% de n.

2º **a)** Se *i* for um número inteiro, *i* denotará a posição do *p*-ésimo percentil;

b) Se *i* não for um número inteiro, o *p*-ésimo percentil será a média dos valores que ocupam as posições *i* e *i*+1.

.

Exemplo 45 *Os dados a seguir referem-se ao gasto médio com material de limpeza de 12 empresas do setor elétrico no mês de janeiro.*

3.310, 3.355, 3.450, 3.480, 3.480, 3.490, 3.520, 3.540, 3.550, 3.650, 3.730, 3.925

O objetivo é determinar o 1º quartil ou 25º percentil, assim como o 85º percentil. Nesse caso, temos:

$$P_{25} = \left(\frac{25}{100}\right) * 12 = 3 \Rightarrow P_{25} = 3.450 \quad e \quad P_{85} = \left(\frac{85}{100}\right) * 12 = 10,2 \Rightarrow P_{85} = 3.640$$

Para obter os valores dos quartis, utilizando o R, basta usar os comandos abaixo.

```
x<-c(3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550,
   3650, 3730, 3925)
sort(x)
quantile(x,type=1)

0%  25%  50%  75% 100%
3310 3450 3490 3550 3925
```

Observação: Dependendo de cada autor, de acordo com a forma de definir o cálculo dos percentis, podemos ter variados resultados. Nesse sentido, deve-se tomar cuidado ao consultar outras bibliografias que não constem no plano de ensino.

Para determinar os percentis para dados agrupados em tabelas de frequência, quando a variável é discreta, repete-se o processo determinado acima, levando-se em conta a frequência absoluta acumulada.

4.2.4.2 Estatísticas de posição para variáveis contínuas

Como as medidas separatrizes são múltiplos dos percentis, todas são calculadas a partir da expressão abaixo;

$$P_i = I_i + \frac{\frac{i.n}{100} - F_{ant}}{f_i}.h \tag{4.14}$$

- I_i - Limite inferior da classe que contém o percentil *i*.
- F_{ant} - Frequência absoluta anterior à classe do percentil *i*;
- f_i - Frequência absoluta da classe do percentil *i*;
- h - Amplitude do intervalo de classe.

Exemplo 46 *A Tabela 33 apresenta os dados relativos ao peso dos 20 entrevistados apresentados na Tabela 8. Obter Q1, Q2, Q3 para a distribuição dos dados ta Tabela 33.*

Tabela 33 – Distribuição dos pesos dos entrevistados

Classe	$[L_i; L_s[$	x_i	f_i	r_i	F_i	R_i
1	[49,0;55,4[	52,2	7	35%	7	35%
2	[55,4;61,8[	58,6	4	20%	11	55%
3	[61,8;68,2[	65	4	20%	15	75%
4	[68,2;74,6[	71,4	2	10%	17	85 %
5	[74,6;81,0]	77,8	3	15%	20	100%

4.2.4.3 Gráfico Box-Plot

Um Box-plot é um gráfico diferente dos apresentados anteriormente. Possui formato de caixa, e a partir dele podemos observar a variabilidade do conjunto de dados, assim como fazer comparação com outros conjuntos.

Para construí-lo precisamos dos quartis 1, 2 e 3, assim como de uma medida chamada de intervalo interquartil I_q, que é obtida calculando-se a diferença do Q_1 com Q_3. Vejamos um exemplo.

Exemplo 47 *A partir dos dados do exemplo 45, temos que*

$$Q_1 = 3.450 \quad , \quad Q_3 = 3.550,$$

assim, o intervalo interquartil I_q é dado por

$$I_q = 3.550 - 3.450 = 100$$

Uma vez obtidos $Q_1, Q_2 Q_3$ e I_q, a primeira aresta do box-plot (de baixo para cima) é a linha onde se encontra o Q_1, a linha central é a mediana do conjunto de dados ou Q_2, a segunda aresta de cada caixa é a linha onde se encontra o Q_3.

As linhas pontilhadas até os extremos são chamados de bigodes, os quais são obtidos calculando-se $Q_1 - 1,5 I_q$ e $Q_3 + 1,5 I_q$. Os pontos que estão fora destes extremos são chamados de *outliers*. Para os dados do exemplo 45, o box-plot é apresentado na Figura 22.

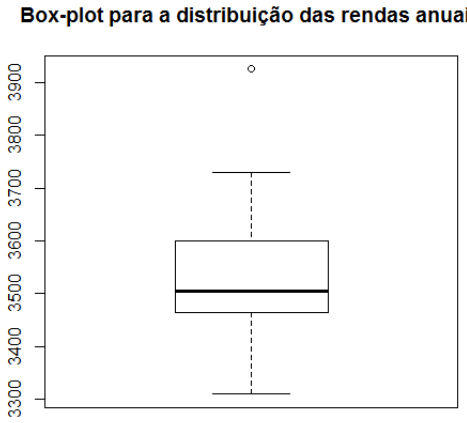


Figura 22 – Box-plot para distribuição dos rendimentos anuais.

Exemplo 48 *Ao longo de 40 dias foi anotada a metragem diária de construção em uma obra de grande porte. Construir um box-plot para esses dados.*

48, 58, 56, 63, 52, 50, 59, 51, 59, 38, 57, 56, 69, 66, 55, 49, 61, 49, 47, 58, 49, 52, 55, 60, 52, 54, 57, 67, 66, 60, 59, 50, 45, 57, 64, 56, 57, 60, 53, 58

Para construção do box-plot, utilizando o R, basta utilizar os comandos a seguir.

```
x1<-c(48, 58, 56, 63, 52, 50, 59, 51, 59, 38, 57, 56, 69, 66,
    55, 49, 61, 49, 49, 52, 55, 60, 52, 54, 57, 67, 66, 60,
   59, 50, 45, 57, 64, 56, 57, 60, 47, 58, 53, 58)
quantile(x1,type=1)
medias1<-mean(x1)

boxplot(x1,type=1,main='Metragem construida diariamente')
points(1,medias1, col=2, pch=19)
```

4.3 Exercícios de Aplicação

1. Qual a diferença entre medidas de tendência central em relação as medidas de dispersão?

2. Qual é a medida de variabilidade adequada para se comparar a variabilidade de diferentes conjuntos de dados?

3. O que é a mediana de um conjunto?

4. O número de acidentes de trabalho em uma linha de produção foi registrado ao longo de 30 dias consecutivos.

0	1	0	1	0	0	0	0	2	3	0	1	2	3	4
0	0	0	1	4	1	1	0	0	3	5	1	0	0	1

   Dada a necessidade de estabilizar esse cenário e propor ações preventivas, deve-se conhecer a estrutura dos dados. Nesse caso, deve-se:

   a) Obter as medidas de tendência central.

   b) Obter as medidas de dispersão.

5. Uma loja de departamentos, selecionou um grupo de 54 notas fiscais, durante um dia, e obter o seguinte quadro:

Tabela 34 – Distribuição dos valores para as notas fiscais

Classe	$[L_i; L_s[$	f_i
1	[1.000,00;1.200,00[	2
2	[1.200,00;1.400,00[	6
3	[1.400,00;1.600,00[	10
4	[1.600,00;1.800,00[	5
5	[1.800,00;2.000,00]	2

   a) Obter as medidas de tendência central. Qual a relação entre a média, moda e mediana?

   b) Obter as medidas de dispersão.

   c) Construir um box-plot para os dados dessa tabela.

6. A escassez de candidatos para ocupar cargos de gerência em empresas de TI ten exigido que os estados paguem salários mais altos e ofereçam vantagens extras para atrair e manter os funcionários nesses cargos. Os dados a seguir mostram o salário-base (em milhares de dólares) anual registrado em 20 estados na área de Nova York.

187 184 174 185 175 172 202 197 165 208 215 164 162 175 172 156 172 175 170 183

a) Construa uma tabela de distribuição de frequência.

b) Obtenha as medidas de tendência central. Qual a relação entre a média, moda e mediana?

c) Obtenha as medidas de dispersão.

d) Construir um box-plot para os dados dessa tabela.

4.3.1 Gabarito

4. a. $\overline{x}$ = 1,14, mediana = 1, moda = 0;

b. s^2 = 2,12, s = 1,46, cv = 1,28

5. a. $\overline{x}$ = 1492, mediana = 1490, moda = 1488. Distribuição assimétrica a direita.

b. s^2 = 44933,33, s = 211,97, cv = 0,1420

c. Q_1 = 1341, Q_3 = 1630.

6. a.

Tabela 35 – Distribuição dos salários observados

Classe	$[L_i; L_s[$	f_i	r_i	F_i	R_i
1	[158;167,5[	4	20%	4	20%
2	[167,5;177[	8	40%	12	60%
3	[177;186,5[	3	15%	15	75%
4	[186,5;196[	1	5%	16	80%
5	[196;205,5]	2	10%	18	90%
6	[205,5;215]	2	10%	20	100%

b. $\overline{x}$ = 179,38, mediana = 174,62, moda = 171,72. Distribuição assimétrica a direita.

c. s^2 = 236,31, s = 15,37, cv = 0,0857

d. Q_1 = , Q_3 = .

Capítulo 5

Elementos de Probabilidade

5.1 Introdução

Em pleno século XXI o volume de informação presente em nosso cotidiano cresce constantemente. Em diversas situações, somos obrigados a tomar decisões, as quais geram incerteza. Assim consciente ou inconscientemente, a probabilidade é usada por qualquer indivíduo que toma decisão em situações de incerteza, bem como de aleatoriedade. O estudo da probabilidade surgiram a mais de 300 anos atrás e as aplicações iniciais do ocorreram em função de jogos de azar, no século XVI (DEVORE, 2006). Contudo, foi apenas no século XX, em que por meio dos estudos de A. Kolmogorov, foram elaborados axiomas matemáticos, os quais constituem a base da Teoria das probabilidades.

O uso dessa teoria indica a existência de um elemento de acaso, ou de incerteza, quanto à ocorrência ou não de um evento. Por exemplo, se lançarmos uma moeda para o ar, de modo geral não podemos afirmar se vai sair cara ou coroa. A probabilidade nos indicará uma medida de quão provável é a ocorrência de determinado evento.

Neste capítulo estamos interessados em compreender conceitos básicos para o cálculo de probabilidades, os quais são base para cálculos mais elaborados.

5.2 Desenvolvimento axiomático

5.2.1 Experimento Aleatório

Definição 1 *É um processo cujos resultados podem apresentar variações mesmo quando realizadas em condições uniformes, isto é, o resultado está sujeito à incerteza.*

Exemplo 49 *Em um lançamento de dados podemos identificar os resultado após o lança-*

mento, porém não podemos identificar os resultados de um experimento sem antes realizá-lo.

Exemplo 50 *Se examinarmos três fusíveis em sequência e anotarmos o resultado de cada exame, o resultado do experimento é qualquer sequência de N e D de 3 elementos.*

5.2.2 Espaço Amostral

Definição 2 *É o conjunto formado por todos os possíveis resultados de um experimento aleatório.*

Exemplo 51 *Para o experimento aleatório anterior, o espaço amostral é formado pelos elementos*

$$\Omega = \{NNN, NND, NDN, DNN, NDD, DNN, DDN, DDD\}$$

Exemplo 52 *Dois postos de gasolina estão localizados em uma determinada interseção. Cada um possui seis bombas. Considere o experimento em que o número de bombas em uso em determinada hora do dia é determinado para cada posto. Um resultado experimental especifica quantas bombas estão em uso no primeiro posto e quantas são usadas no segundo posto. Um resultado possível é (2,2), isto é, para o primeiro e segundo posto há duas bombas em uso. Determinar o espaço amostral.*

5.2.3 Eventos

Definição 3 *Chamamos de evento, todo subconjunto do espaço amostral Ω. Assim, se E é um evento de Ω, então E está contido em Ω. Chamamos Ω de evento certo e $\emptyset$ de evento impossível.*

Exemplo 53 *No lançamento de um dado, observando o número da face superior, podemos descrever alguns eventos.*

a) *A: Obter número par;*

b) *B: Obter número menor que 3;*

c) *C: Obter número maior que 5;*

d) *D: Obter número zero;*

e) *E: Obter número menor que 7.*

Exemplo 54 *A Kentucky Power & Light Company (KP&L) está iniciando um projeto idealizado para aumentar a capacidade de geração de energia emuma de suas usinas no norte de Kentucky. O projeto se divide em duas etapas, ou passos, sequenciais: etapa 1 (projeto) e etapa 2 (construção). Não obstante cada etapa estar programada e ser controlada o mais cuidadosamente possível, a administração não é capaz de prever o tempo exato necessário para o término de cada fase do projeto. Uma análise preliminar apresentou que para execução da primeira etapa seriam necessários 2, 3 ou 4 meses. Enquanto para executar a segunda etapa seriam necessários 6, 7 ou 8 meses. A meta da administração da empresa é que a obra seja realizada em até 10 meses, em função da necessidade crítica de energia elétrica adicional.*

a) *Quais são as possíveis combinações de prazos?*

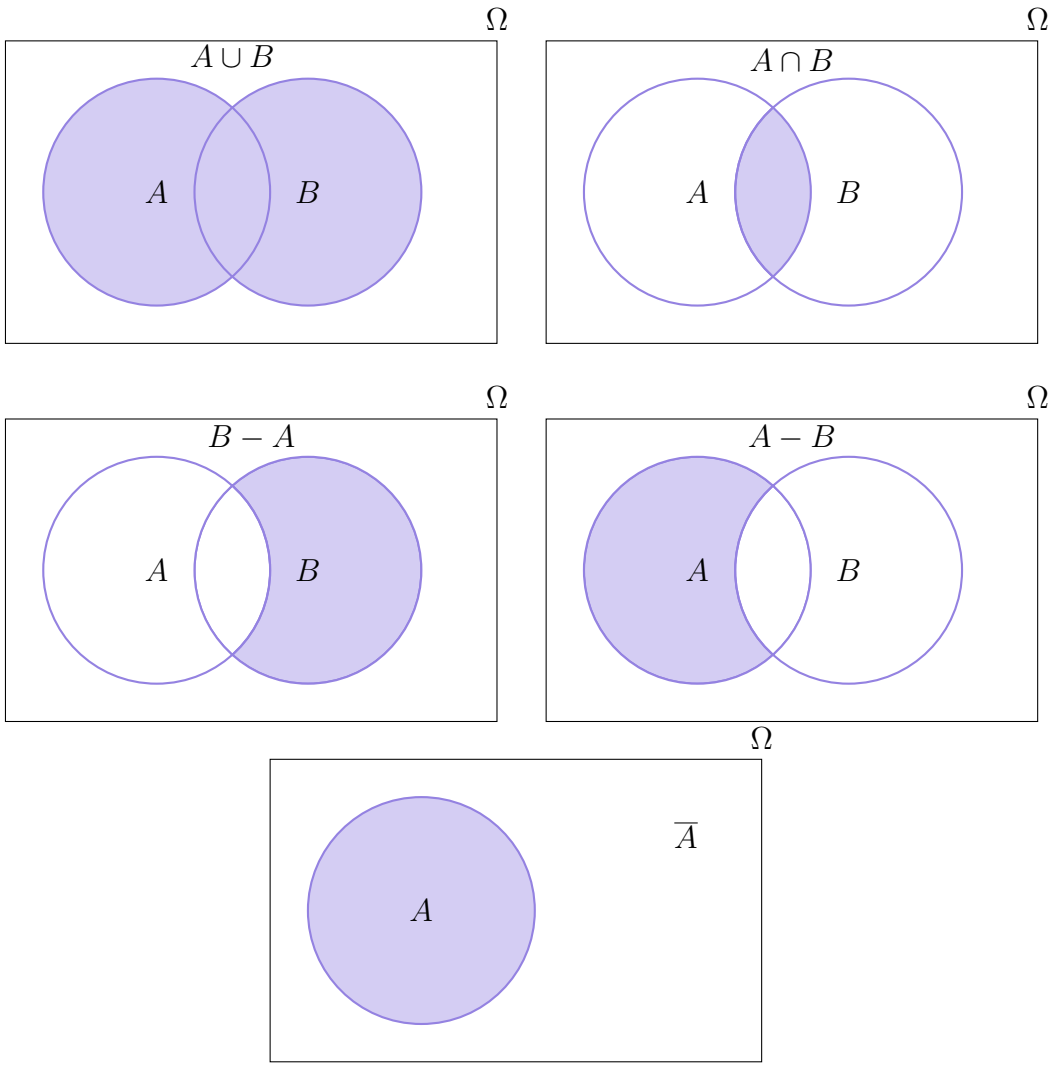
b) *Quais prazos combinados resultam na execução da obra em um tempo inferior a 10 meses?*

c) *Quais prazos combinados resultam na execução da obra em exatamente 10 meses?*

d) *Quais prazos combinados resultam na execução da obra em um tempo superior a 10 meses?*

5.2.4 Operações Entre Eventos

A teoria das probabilidades está associada diretamente a teoria de conjuntos. Nesse sentido, para definir operações entre eventos, recorre-se as operações usuais entre conjuntos, isto é, união, interseção e complemento. Nesse sentido, sejam os eventos A, B e complementar de A, digamos $\overline{A}$, em um espaço amostral Ω, por meio de Diagramas de Venn temos:



Exemplo 55 *Para as informações do exemplo 54, considerando os diagramas acima apresentados, determine:*

a) $B \cap C$*;*

b) $B \cup C$*;*

c) $\overline{D}$*;*

5.2.5 Definições de Probabilidade, Axiomas e interpretações.

De acordo com Magalhães (2011), na literatura podemos encontrar três definições de probabilidade: Clássica, Geométrica e Frequentista.

Na definição clássica todos os elementos de um espaço amostral possuem a mesma chance de acontecerem. Seja um evento A de interesse, associado a um espaço amostral Ω. Então a probabilidade de ocorrência do evento A, será a razão entre o número de elementos do evento de interesse com o número de elementos do espaço amostral.

$$P(A) = \frac{n(A)}{n(\Omega)}$$

Considerando que o espaço amostral pode ser não enumerável, então o conceito de probabilidades se aplicará ao comprimento de intervalos, medida de áreas ou similares, dando origem a probabilidade geométrica.

$$P(A) = \frac{Comprimento \quad de \quad A}{Comprimento \quad de \quad \Omega}$$

Por fim, na definição frequêntista, deve-se considerar o limite das frequências relativas como o valor da probabilidade. Assim, seja n_A o número de ocorrências de A em n repetições independentes do experimento em questão, temos:

$$P(A) = \lim_{n \longrightarrow \infty} \frac{n_A}{n}$$

Em síntese, isso significa que a medida que são realizados os experimentos a probabilidade de ocorrência de um evento determinado se aproxima do verdadeiro valor a medida que o número de realizações tende ao infinito.

Para isso, considere um experimento que consiste em lançar uma moeda 10, 50, 100 e 1000 vezes, e observar o número de ensaios em que o resultado é cara. Os resultados podem ser verificados na Figura 23. Observe que a medida que o número de ensaios cresce a probabilidade acumulada da ocorrência de cara converge para sua verdadeira ocorrência, isto é 0,5.

Considerando que as definições anteriores são úteis para o cálculo de diversos problemas práticos e teóricos, é necessário enunciar uma série de axiomas para que se

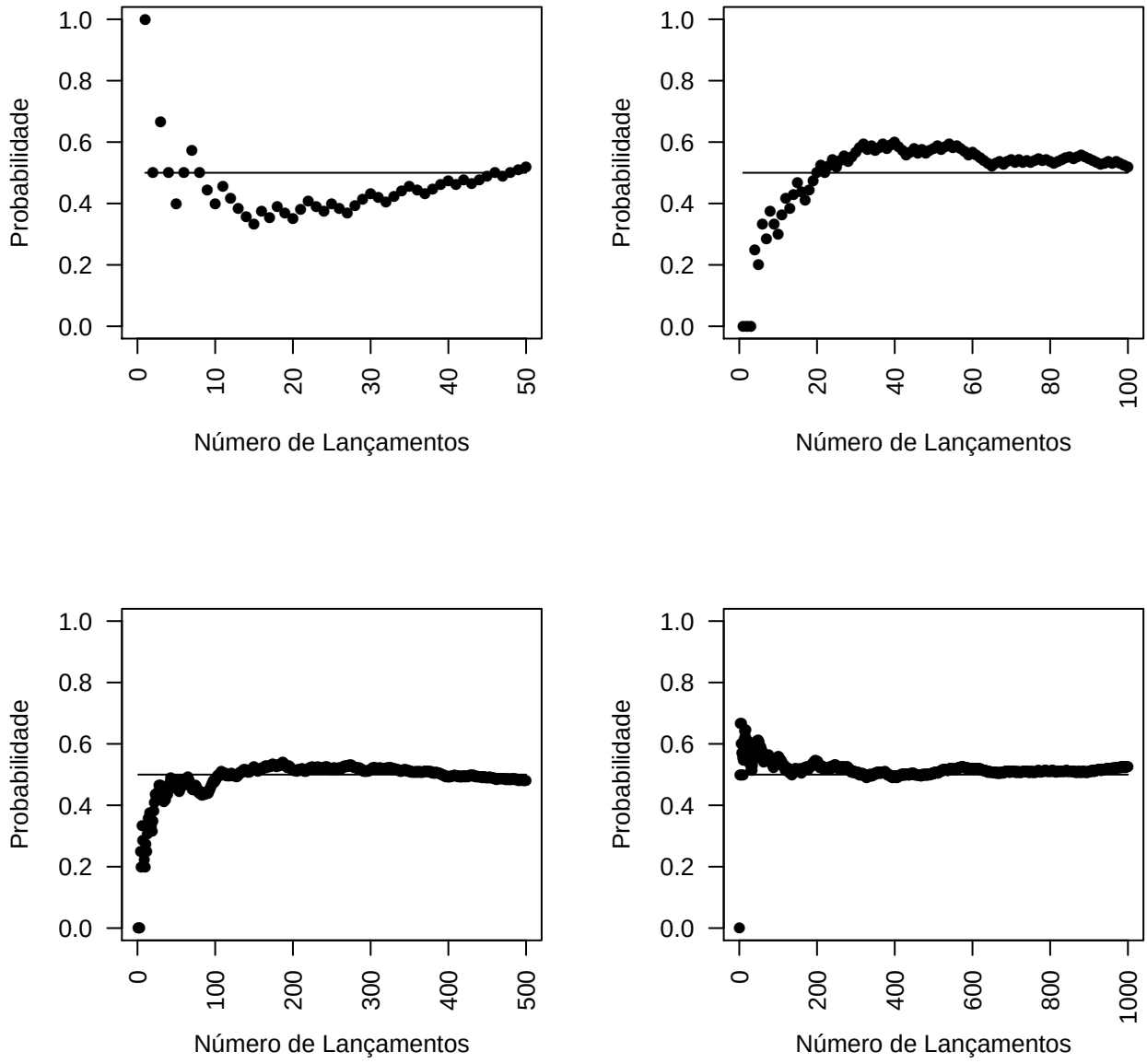


Figura 23 – Simulação do lançamento de uma moeda.

tenha uma formulação mais rigorosa para o conceito de probabilidade. Por volta do ano de 1930 o russo A. N. Kolmogorov apresentou esses axiomas matemáticos para definir probabilidade.

Axioma 1 *Para qualquer evento, A,* $P(A) \geq 0$*;*

Axioma 2 *Seja* Ω *o espaço amostral associado ao experimento aleatório, então P(*Ω*) = 1;*

Axioma 3 *Se* $A_1, A_2, \ldots, A_k$ *for um conjunto finito de eventos mutuamente exclusivos, isto*

é, $A_i \cap A_j = \emptyset$*, então* $P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i)$

Para ilustrar os axiomas anteriores considere os exemplos a seguir:

Exemplo 56 *No experimento aleatório lançar uma moeda e observar a face superior, o espaço amostral é* Ω *= {H,T}. Pelos Axiomas anteriores tem-se que P(*Ω*) = 1, isto é,*

$$1 = P(\Omega) = P(H \cup T) = P(H) + P(T) = 0,5 + 0,5.$$

5.2.6 Propriedades de Probabilidade

Seja Ω o espaço amostral associado a um experimento aleatório E, com eventos A, B e C. Então valem as seguintes propriedades:

Propriedade 1 *Para qualquer evento A tem-se* $P(A) = 1 - P(\overline{A})$.

Demonstração 1

Para ilustrar a propriedade acima, considere o exemplo a seguir.

Exemplo 57 *Considere um sistema de cinco componentes idênticos ligados em série. Qual a probabilidade de que o sistema não funcione?*

Propriedade 2 *Se A e B forem mutuamente exclusivos, então* $P(A \cap B) = 0$

Demonstração 2

Propriedade 3 *Para quaisquer dois eventos A e B vale*

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

Demonstração 3

Exemplo 58 *Depois de um longo período de testes, verificou-se que o procedimento A de recuperação de informação corre um risco de 2% de não oferecer resposta satisfatória. No procedimento B, o risco cai para 1%. O risco de ambos os procedimentos apresentarem resposta insatisfatória é de 0,5%. Qual é a probabilidade de pelo menos um dos procedimentos apresentar resposta insatisfatória?*

Solução: *Pelo menos um dos procedimentos apresentar resposta insatisfatória implica em saber a probabilidade de A ou B apresentar resposta insatisfatória. Considerando, que de acordo com o enunciado do problema os eventos A e B não são mutuamente excludentes (P(A $\cap$ B) $\neq$ 0), deve-se utilizar a expressão a seguir:*

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

Por hipótese, temos que P(A) = 0,02, P(B) = 0,01 e $P(A \cap B)$ = 0,005, então,

$$P(A \cup B) = 0,02 + 0,01 - 0,005 = 0,025$$

Portanto, a probabilidade de pelo menos 1 dos procedimentos apresentar resposta insatisfatória é aproximadamente 2,5%.

Exemplo 59 *No lançamento de dois dados, são determinados dois eventos, são eles: A = {O valor observado na face superior no primeiro dado é maior do que o valor observado na face superior do segundo dado} e B = {A soma dos valores das faces superiores dos dois dados é igual a 7}. Nesse sentido, sabendo que foi observado o evento A, qual a probabilidade do evento B ocorrer?*

Exemplo 60 *Qual é a probabilidade de sair face 4 ou 5 em um laçamento de um dado?*

Exemplo 61 *Qual é a probabilidade de **NÃO** sair face 4 ou 5 em um laçamento de um dado?*

Exemplo 62 *Uma empresa de eletricidade oferece uma taxa vitalícia de energia a qualquer lar cuja utilização de energia esteja abaixo de 240 kWh durante um determinado mês. Represente por A o evento de um lar selecionado aleatoriamente em uma comunidade que não excede a utilização da taxa vitalícia em janeiro e por B o evento análogo para o mês de julho. Supondo que P(A) = 0,8, P(B) = 0,7 e* $P(A \cup B)$ *= 0,9. Calcule* $P(A \cap B)$.

Propriedade 4 *Para quaisquer eventos A, B e C vale*

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

(Fazer de tarefa)

Demonstração 4

5.3 Probabilidade Condicional

Em certas ocasiões, o conhecimento prévio da ocorrência de um evento A pode afetar ou alterar o cálculo da probabilidade de ocorrência de B.

Uma empresa de bens de consumo veiculou um anúncio de televisão de um de seus produtos de limpeza. A partir desse anúncio ela desejou saber quais os efeitos do anúncio na venda dos produtos. Essa indagação motiva a definição de probabilidade condicional.

Definição 4 *Sejam A e B dois eventos, com $P(A) > 0$. Denotemos por $P(B|A)$ a probabilidade de ocorrência de B, a hipótese de A ter ocorrido. Ora, como A ocorreu, A passa a ser o novo espaço amostra, que vem substituir o espaço original S. Isso nos conduz à definição:*

$$P(B|A) = \frac{P(A \bigcap B)}{P(A)} \tag{5.1}$$

Chamamos de $P(B|A)$ a probabilidade condicional de B, dado A, isto é, a probabilidade de ocorrência de B, dado que A ocorreu.

Exemplo 63 *Considerando a pergunta inicial, suponha que com base em uma pesquisa, foram atribuídas probabilidades aos seguintes eventos: A = Pessoas que compraram o produto, B = Pessoas que se lembraram de ter visto o anúncio e $A \cap B$ = Pessoas que compraram o produto e que se lembram de ter visto o anúncio. As probabilidades atribuídas aos eventos foram 0,20 ; 0,40 e 0,12 respectivamente.*

a) *Qual é a probabilidade de uma pessoa comprar o produto por se lembrar de ter visto o anúncio? Ver o anúncio aumenta a probabilidade de a pessoa comprar o produto?*

b) *A empresa experimentou também outro anúncio e atribuiu a ele os valores de P(B) = 0,30, $P(A \cap B)$ = 0,10. Qual é a probabilidade de uma pessoa comprar o produto por se lembrar de ter visto o anúncio? Qual anúncio parece ter maior efeito sobre as compras efetuadas pelos clientes?*

Exemplo 64 *Os estudantes de um colégio, presentes em uma reunião, foram classificados por sexo e por opção de área de formação de acordo com as informações a seguir:*

- *Economia, sendo 10 do sexo masculino e 8 do sexo feminino;*
- *Análise e Desenvolvimento de Sistemas, sendo 6 do sexo masculino e 5 do sexo feminino;*
- *Engenharia da Computação, sendo 8 do sexo masculino e 4 do sexo feminino;*

Calcular as probabilidades de que:

a) *Alunas optem por Economia.*

b) *Alunos optem por Análise e Desenvolvimento de Sistemas.*

c) *Seja aluno sabendo-se que optou por Engenharia da Computação.*

d) *Aluno opte por Engenharia da Computação.*

A partir da definição 4, é possível determinar a probabilidade de ocorrência simultânea de dois eventos não mutuamente exclusivos. Isso nos induz a ideia de probabilidade da multiplicação dado o conhecimento prévio de uma informação. Nesse sentido, é possível reescrever a expressão 5.1 da seguinte maneira:

$$P(A \cap B) = P(A) \times P(B|A) \tag{5.2}$$

Exemplo 65 *Uma rifa composta por 15 números irá definir o ganhador de dois prêmios sorteados um de cada vez. Se você adquiriu três números, qual é a probabilidade de ganhar os dois prêmios?*

Solução: *Considere os eventos ganhar o primeiro prêmio* (G_1)*, e o evento ganhar o segundo prêmio* (G_2)*. Então, usando a definição anterior, temos a seguinte probabilidade de ganhar ambos os prêmios.*

$$P(G_1 \cap G_2) = P(G_1) \times P(G_2|G_1) = \frac{3}{15} \times \frac{2}{14} = \frac{1}{35} = 2,86\%.$$

5.3.1 Propriedades da Probabilidade Condicional

Seja Ω um espaço amostral associado a um experimento aleatório, P uma probabilidade em Ω com A $\subset \Omega$ um evento tal que P(A) $>$0. Então as probabilidades condicionais satisfazem:

1) $0 \leq P(B|A) \leq 1$;

2) $P(A|A) = 1$;

3) $P(B_1 \cup B_2|A) = P(B_1|A) + P(B_2|A)$ se B_1, B_2 são mutuamente excludentes. Analogamente pode-se estender essa propriedade para k eventos mutuamente exclusivos.

Comentário: As demonstrações destas propriedades podem ser obtidas em (LOESCH, 2012).

5.3.2 Independência entre Eventos

Diversas são as propriedades de grande importância no estudo de probabilidade. Entre elas, é possível destacar a irrelevância da ocorrência do evento A no cálculo da probabilidade do evento B. Isso motiva a próxima definição.

Definição 5 *Sejam A e B eventos do espaço amostral S. Esses eventos são chamados de independentes se a ocorrência do evento A não afeta o cálculo da probabilidade do evento B. Então tem-se que*

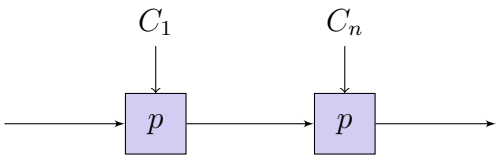
$$P(B|A) = P(B) \Rightarrow P(A \cap B) = P(A) \times P(B) \tag{5.3}$$

Teorema 1 ***Regra do Produto de Probabilidades*** *Sejam* $A_1, A_2, \ldots, A_n$ *eventos do espaço amostral S,* $P\left(\bigcap_{i=1}^{n} A_i\right) > 0$*, então o produto é dado por:*

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \times P(A_2|A_1) \times \ldots \times P\left(A_n|\bigcap_{i=1}^{n-1} A_i\right) \tag{5.4}$$

***Observação**: Para verificar a demonstração desse teorema, consultar Magalhães (2011) página 29.*

Exemplo 66 *Considere um sistema composto de n componentes ligados em série, de tal forma que, se um componente falhar, os sistema todo falha. Esquematicamente é possível representar pela Figura abaixo.*



*Se os componentes operam independentemente e cada um tem probabilidade **p** de falhar, qual é a probabilidade de o sistema funcionar?*

Solução: Considere o evento que caracteriza o componente C_i, para $i = 1, \ldots, n$ apresentar falha. Nesse caso, então temos que:

$$P(C_i) = p,$$

Como esses componentes operação de forma independente, a probabilidade de falha para cada um independe do outro. Assim,

$$P(C_1 \cap C_2 \cap \ldots \cap C_n) = p \times p \times \ldots \times p = p^n$$

Portanto, a probabilidade de que todos os componentes falhem, simultaneamente, é de aproximadamente p^n.

O próximo Teorema está associado a necessidade de calcular a probabilidade da ocorrência de parte de um evento que seja comum a diversos outros eventos, chamado de **Teorema da Probabilidade Total (TPT)**. Considere o exemplo a seguir:

Exemplo 67 *Imagine que você utiliza peças de quatro fornecedores, que têm diferentes desempenhos quanto a sua qualidade. As peças são classificadas como conformes ou não conformes e você conhece a proporção de peças não conformes de cada fornecedor (p_1, p_2, p_3, p_4). Considere a formação de um lote com peças dos quatro fornecedores,*

conforme ilustra a Figura 24. Se você selecionar, ao acaso, uma peça do lote, qual é a probabilidade de ela ser não conforme?

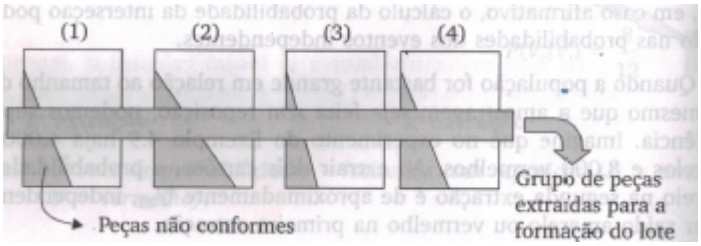


Figura 24 – Distribuição das peças por fornecedor (BARBETTA; REIS; BORNIA, 2010)

Para responder a indagação anterior é necessário conhecer alguns conceitos que compõem o TPT. As definições a seguir são baseadas em Barbetta, Reis e Bornia (2010).

Considere o espaço amostral Ω, particionado em k eventos, $E_1, E_2, \ldots, E_k$, satisfazendo ás seguintes condições:

1. $E_i \cap E_j = \emptyset, \forall i \neq j$;
2. $\bigcup_{i=1}^{k} E_k = \Omega$ (eventos exaustivos);
3. $P(E_i) > 0, \forall \quad i = 1, 2, \ldots k$; Para melhor ilustrar os itens anteriores, observe a Figura 25.

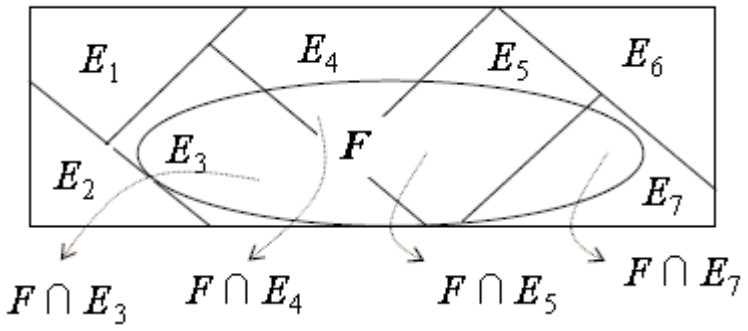


Figura 25 – Partição do espaço amostral em eventos mutuamente excludentes. (BARBETTA; REIS; BORNIA, 2010)

Teorema 2 ***(Teorema da Probabilidade Total)*** *Sejam $E_1, E_2, \ldots, E_k$ eventos que formam uma partição do espaço amostral S. Seja F um evento qualquer, então sua probabilidade de ocorrência será dada por:*

$$P(F) = \sum_{i=1}^{k} P(E_i)P(F|E_i) \tag{5.5}$$

Demonstração 5 *Seja F um evento qualquer do espaço amostral S, então tem-se que*

$$F = \bigcup_{i=1}^{k} F \cap E_i \tag{5.6}$$

em que os eventos $F \cap E_i$ são mutuamente excludentes. Logo

$$P(F) = P\left[\bigcup_{i=1}^{k} F \cap E_i\right] = \sum_{i=1}^{k} P(F \cap E_i]). \tag{5.7}$$

Utilizando o Teorema 1, temos a seguinte expressão:

$$\begin{aligned} P(F) &= P(E_1) \times P(F|E_1) + \ldots + P(E_k) \times P(F|E_k) && (5.8)\\ &= \sum_{i=1}^{k} P(E_i)P(F|E_i). && (5.9) \end{aligned}$$

Comentário: Naturalmente algumas probabilidades condicionais podem assumir valor zero, isto é, pode não existir elementos na interseção entre F e E_i. O TPT pode ser interpretado fisicamente como uma medida do peso de cada um dos eventos de E_i, na contribuição para formar o evento F.

Retomando o exemplo 67, supondo que os fornecedores 1, 2, 3 e 4 são responsáveis por 30%, 20%, 40% e 10% na composição do lote de peças, e que suas produções são 1%, 5%, 10% e 2% são não conformes, qual a probabilidade desejada?

Solução: Considere o evento de interesse peça sorteada não conforme (N). Nesse caso, ao avaliar o lote, uma peça não conforme pode ser de qualquer um dos três fornecedores (F_1, F_2, F_3, F_4). Assim, é possível construir o evento N como segue:

$$N = (N \cap F_1) \cup (N \cap F_2) \cup (N \cap F_3) \cup (N \cap F_4). \tag{5.10}$$

De modo equivalente, que aplicando probabilidade em ambos os membros de 5.10, temos:

$$\begin{aligned} P(N) &= P(N \cap F_1) + P(N \cap F_2) + P(N \cap F_3) + P(N \cap F_4)\\ &= P(F_1) \times P(N|F_1) + P(F_2) \times P(N|F_2) + P(F_3) \times P(N|F_3) + P(F_4) \times P(N|F_4)\\ &= 0,3 \times 0,01 + 0,20 \times 0,05 + 0,4 \times 0,1 + 0,1 \times 0,02\\ &= 0,055 \end{aligned}$$

Portanto, a partir das condições elencadas, a probabilidade de uma peça selecionada ao acaso desse lote, ser não conforme é de aproximadamente 5,5%.

Exemplo 68 *Uma rede local de computadores é composta por um servidor e cinco clientes (A,B,C,D,E). Registros anteriores indicam que dos pedidos de determinado tipo de processamento, realizados por meio de uma consulta, cerca de 10% vêm do cliente A, 15% do cliente B, 15% do cliente C, 40% do D e 20% do E. Se o pedido não for fito de forma adequada, o processamento apresentará erro. Usualmente, ocorrem os seguintes percentuais de pedidos inadequados: 1% do cliente A, 2% do cliente B, 0,5% do cliente C, 2% do cliente D e 8% do cliente E.*

a) *Qual é a probabilidade de o sistema apresentar erro?*

b) *Qual é a probabilidade de que o processo tenha sido pedido pelo cliente E, sabendo-se que apresentou erro?*

O próximo Teorema é um dos mais importantes encontrados na literatura. Assim como os Teoremas 1 e 2 ele relaciona probabilidades condicionais. Foi proposto por Thomas Bayes, e em sua época não foi divulgado. Ele leva seu nome e é conhecido com **Teorema de Bayes** ou também como **Teorema das Probabilidades a Posteriori**. Possui diversas aplicações que permeiam diversas áreas do conhecimento e é a base da Estatística Bayesiana.

Teorema 3 ***(Teorema de Bayes)*** *Sejam $E_1, E_2, \ldots, E_k$ eventos que formam uma partição do espaço amostral S. Seja F um evento de interesse com $P(F) > 0$. Então $\forall j = 1, \ldots, k$*

temos que:

$$P(E_j|F) = \frac{P(E_j) \times P(F|E_j)}{\sum_{i=1}^{k} P(E_i)P(F|E_i)} \tag{5.11}$$

Demonstração 6 *Sejam $E_1, E_2, \ldots, E_k$ eventos que formam uma partição do espaço amostral S, e considere o espaço equiprovável com evento F de interesse, então $\forall j = 1, \ldots, k$*

$$P(E_j|F) = \frac{P(E_j \cap F)}{P(F)} = \frac{P(E_j) \times P(F|E_j)}{\sum_{i=1}^{k} P(E_i)P(F|E_i)} \tag{5.12}$$

Exemplo 69 *A urna A contém 3 fichas vermelhas e 2 azuis, e a urna B contém 2 vermelhas e 8 azuis. Joga-se uma moeda "honesta". Se aparecer cara, extrai-se uma ficha da urna A; se aparecer coroa, extrai-se uma ficha da urna B. Uma ficha vermelha é extraída. Qual a probabilidade de ter saído cara no lançamento?*

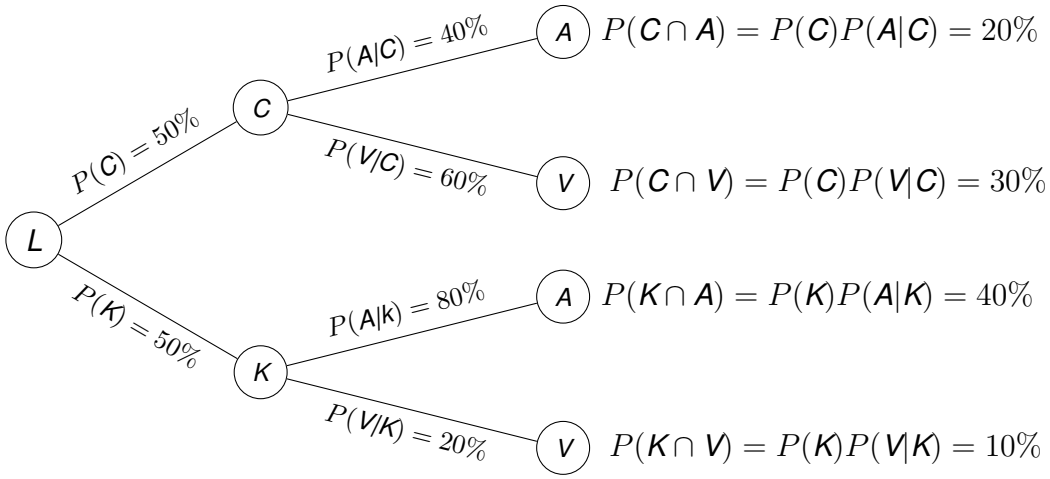
Solução: *Para resolver esse problema, devemos fazer uso do Teorema de Bayes, enunciado acima. A partir do problema enunciado, podem ocorrer as seguintes situações:*

- *Aparecer Cara - C;*
- *Aparecer Coroa - K;*
- *Sair ficha vermelha, quando apareceu cara $(V|C)$ ou sair ficha azul, quando apareceu cara $(A|C)$;*
- *Sair ficha vermelha, quando apareceu coroa $(V|K)$ ou sair ficha azul, quando apareceu coroa $(A|K)$;*

Com auxílio do Teorema de Bayes, desejamos obter o valor da probabilidade abaixo.

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

Vamos retornar ao esboço da árvore e selecionar as probabilidades necessárias para calcular a probabilidade de interesse.



Assim

$$P(C|V) = \frac{0,50 \times 0,60}{0,50 \times 0,60 + 0,50 \times 0,20} = 0,75$$

Portanto, a probabilidade de ter sido cara, quando retirou-se uma ficha vermelha, é aproximadamente 75%.

Exemplo 70 *Em um determinado posto de gasolina, 40% dos clientes usam gasolina comum (A), 35% usam gasolina aditivada (B) e 25% usam gasolina Premium. (C). Dos clientes que usam gasolina comum, apenas 30% enchem o tanque. Dos que usam gasolina aditivada e Premium, respectivamente 60% e 50% enchem o tanque.*

a) *Qual é a probabilidade de o próximo cliente pedir gasolina aditivada e encher o tanque?*

b) *Qual é a probabilidade de o próximo cliente encher o tanque?*

c) *Se o próximo cliente encher o tanque, qual é a probabilidade de pedir gasolina comum? E gasolina aditivada? E gasolina Premium?*

Exemplo 71 *As máquinas A e B são responsáveis por 60% e 40%, respectivamente, da produção de uma empresa. Os índices de peças defeituosas na produção dessas máquinas valem 3% e 7% respectivamente. Se uma peça defeituosa foi selecionada da produção desta empresa, qual é a probabilidade de que tenha sido produzida pela máquina B?*

Exemplo 72 *Setenta por cento das aeronaves leves que desaparecem em vôo em certo país são localizadas posteriormente. Das aeronaves localizadas, 60% possuem localizador de emergência, enquanto 90% das aeronaves não localizadas não possuem esse dispositivo. Suponha que uma aeronave tenha desaparecido.*

a) *Se ela tiver localizador de emergência, qual é a probabilidade de não ser localizada?*

b) *Se ela não tiver localizador de emergência, qual é a probabilidade de ser localizada?*

5.4 Exercícios de Aplicação

1. Construir o espaço amostral do lançamento de dois dados.
2. Tirando-se, uma carta de um baralho comum, de 52 cartas, qual é a probabilidade de sair um rei?
3. Uma empresa possui 105 funcionários, sendo 77 homens e 28 mulheres. Escolhendo-se aleatoriamente um desses funcionários para ser homenageado, determinar a probabilidade de que seja sorteado um homem.
4. Uma rifa tem números de 1 a 1000. Uma pessoa compra todos os números que têm exatamente três algarismos. Determinar a probabilidade dessa pessoa ganhar o prêmio.
5. Um levantamento realizado no mês passado por certa companhia aérea, mostrou que, dos 300 voos selecionados aleatoriamente, 247 chegaram no horário previsto. Qual é a probabilidade de que um voo dessa companhia não chegar no horário ? E qual a probabilidade de um voo desta companhia chegar no horário?
6. Uma bola é retirada ao acaso de uma urna que contém 6 bolas vermelhas, 8 pretas e 4 verdes. Determinar a probabilidade dessa bola:
   a) Não ser preta.
   b) Não ser verde.
   c) Ser vermelha.
7. Um experimento consiste no lançamento de dois dados um de cor vermelha(v) e outro de cor branca(b).
   a) Determine o espaço amostral, sabendo que é dado pelo lançamento do dado vermelho primeiro e depois o dado branco.
   b) A partir do espaço amostral, determine a probabilidade de que a soma dos pontos do par de dados seja 7 ou 11.
   c) Qual é a probabilidade de obter no dado branco um número que no mínimo seja 3 pontos maior que o número de pontos no dado vermelho.
   d) Qual é a probabilidade que $v > 2$ ou $b > 3$?
   e) Qual é a probabilidade que $v > 2$ e $b > 3$?
8. Dois dados são lançados , e X = Resultado do primeiro dado e Y = resultado do segundo dado. Podem-se registrar ambos os resultados na forma de pares ordenados (X,Y). Assim, o espaço amostral possui 36 elementos, o qual pode ser obtido por meio de $X \times Y$. Sejam A o evento X+Y=7 e B o evento $X > Y$.

a) Encontre os pares ordenados dos eventos A e B;

b) Encontre os valores de P(A) e P(B);

c) Encontre os pares ordenados doe evento $A \cap B$ bem como $P(A \cap B)$;

d) Calcule a probabilidade de A ou B ocorrer;

9. De uma sacola contendo 15 bolas numeradas de 1 a 15 retira-se um bola. Qual é a probabilidade desta bolsa ser divisível por 3 ou por 4?

10. Uma fábrica tem 3 linhas de produção para o mesmo produto com os seguintes resultados:

    - B1: Linha 1 produz 60% do total com um percentual de defeito de 1%
    - B2: Linha 2 produz 30% do total com um percentual de defeito de 2%
    - B3: Linha 3 produz 10% do total com um percentual de defeito de 3%

    Supondo evento A igual ao produto defeituoso, calcular as probabilidades de produto com defeito dado que produzido pela:

    a) Linha 1;

    b) Linha 2;

    c) Linha 3.

11. Joga-se dois dados. Desde que as faces mostre números diferentes, qual é a probabilidade que uma face seja 4?

12. Calcular P(B|A) se:

    a) A é subconjunto de B;

    b) A e B são mutuamente excludentes.

13. Em uma escala de idiomas com 2000 alunos, 500 fazem curso de inglês 300 fazem curso de espanhol e 200 cursam ambos os cursos. Selecionando-se um estudante do curso de inglês, qual a probabilidade dele também estar cursando o curso de espanhol?

14. Se lança uma moeda não equilibrada de modo que P(C): 2/3 e a probabilidade de P(K) = 1/3. Se obtemos cara (C), escolhe-se um número de 1 a 9, se obtemos coroa (K), escolhe-se um número de 1 a 5. Determinar a probabilidade de obter um número par.

15. Um jogador se dá 5 cartas uma de cada vez, de um baralho de 52 cartas. Qual é a probabilidade de que todas sejam espadas?

16. Uma urna contém 7 bolas vermelhas e 3 bolas brancas. Selecionam-se 3 bolas uma de cada vez. Calcular a probabilidade de que as duas primeiras sejam vermelhas e a terceira seja branca?

17. Um sistema tem dois componentes que operam independentemente. Suponha que as probabilidades de falha dos componentes 1 e 2 sejam 0,1 e 0,2, respectivamente. Determinar a probabilidade de o sistema funcionar nos dois casos seguintes:

    a) Os componentes são ligados em série;

    b) Os componentes são ligados em paralelo;

18. Um piloto de fórmula 1 tem 50% de probabilidade de vencer determinada corrida, quando esta se realiza sob chuva. Caso não chova durante a corrida, sua probabilidade de vitória é de 25%. Se o serviço de Meteorologia estimar em 30% a probabilidade de que chova durante a corrida, qual é a probabilidade deste piloto ganhar esta corrida?

19. Uma imobiliária tem três corretores que atuam no setor de vendas e locação. O histórico de vendas e locações indica que 25% dos negócios são fechados pelo corretor 1, 40% pelo corretor 2 e o restante pelo corretor 3. Dos negócios fechados pelo corretor 1, 70% são vendas, enquanto para o corretor 2 essa porcentagem é de 50% e para o corretor 3, 60%.

    a) Um contrato escolhido ao acaso nos arquivos pertence ao corretor 1. Qual a probabilidade de que este contrato seja de venda? E de locação?

    b) Um contrato é escolhido ao acaso nos arquivos. Qual a probabilidade deste contrato ser de venda? E de locação?

    c) Um contrato escolhido ao acaso nos arquivos é um contrato de venda. Qual a probabilidade deste contrato ser do corretor 2?

5.4.1 Gabarito

2. 0,0769

3. R = 0,7334

4. 0,9

5. 0,8233 e 0,1767

6a. 0,5556

6b. 0,7778

6c. 0,3334

7b. 0,2223

7c. 0,16667

7d. 0,8334

7e. 0,3334

8b. 0,1667 e 0,4167

8c. 0,0834

8d. 0,5

9. 0,4667

10a. 0,0167

10b. 0,0667

10c. 0,3000

11. 0,2778

12a. 1

12b. 0

13. 0,4

14. 0,4667

15. 0,000495

16. 0,175; 0,147

17a. 0,72

17b. 0,98

18. 0,325

19a. 0,300

19b. 0,4150

19c. 0,3419

Capítulo 6

Variáveis aleatórias

6.1 Introdução

Ao resolvermos problemas diários com o auxílio da teoria de probabilidades, é necessário conhecer uma distribuição que melhor descreve o conjunto de dados analisados. Para tal, é preciso classificar a variável aleatória de interesse, variável que possui incerteza associada, em dois tipos: Variáveis Aleatórias Discretas e Variáveis Aleatórias Contínuas.

Uma **variável aleatória** pode ser compreendida como uma variável quantitativa cujo resultado (valor) depende de fatores aleatórios (BARBETTA; REIS; BORNIA, 2010).

De um modo geral, ao realizar um experimento estamos interessados em uma ou mais quantidades. Assim, é necessário estudar a estrutura probabilística dos dados, e nesse contexto atribuir probabilidades aos eventos de interesse. Isso nos conduz ao conceito de variável aleatória.

6.2 Conceitos de Variáveis Aleatórias

Definição 6 *Considere um experimento aleatório e a ele uma espaço amostral S associado. Uma função $f : S \longrightarrow R$ é chamada de variável aleatória, se a cada elemento de S existe um número real associado.*

Por meio da definição anterior é possível compreender uma variável aleatória como uma função do espaço amostral nos reais, para qual é possível atribuir uma probabilidade de ocorrência (MAGALHÃES, 2011). Da mesma forma, de acordo com o número de resultados de um experimento aleatório, uma v.a pode ser unidimensional, bidimensional ou

multidimensional. Todavia, nos limitaremos a v.a unidimensionais. Mais detalhes podem ser obtidos em Loesch (2012).

Uma variável aleatória (v.a) pode ser classificada em **discreta** e **contínua**. Os resultados de uma **v.a discreta** são possíveis valores contidos em um conjunto finito e enumerável. Já os resultados de uma **v.a contínua** são valores que abrangem todo um intervalo de números reais.

6.3 Variáveis Aleatórias Discretas

Exemplo 73 *Ao analisar uma caixa com 3 resistores, os quais serão testados, podemos estar interessados no número de resistores defeituosos. Essa quantidade pode ser chamada de* ***variável aleatória****. Isso ocorre pois existe incerteza associada ao resultado desejado.*

O espaço amostral, para essa situação, pode ser expresso por

S = {BBB,BBD,BDB,DBB,DDB,DBD,BDD,DDD}

em que X representa a variável aleatória número de resistores com defeito. Ela pode assumir os valores 0,1,2 ou 3, isto é, nenhum, um, dois ou três resistores com defeito. É possível expressar essas informações por meio da Tabela 36.

Tabela 36 – Número de resistores com defeito

X	Evento	Probabilidade
0	A_0 = {BBB}	$\frac{1}{8}$
1	A_1 = {BBD,BDB,DBB}	$\frac{3}{8}$
2	A_2 = {DDB,DBD,BDD}	$\frac{3}{8}$
3	A_3 = {DDD}	$\frac{1}{8}$

Para os dados da Tabela 36 é possível construir o Gráfico 26.

A partir do exemplo apresentado anteriormente é possível verificar que a v.a X assume valores finitos em um conjunto enumerável finito, o que motiva a definição a seguir.

Definição 7 ***Variável Aleatória Discreta:*** *Uma v.a discreta assume somente um número enumerável de valores (finito ou infinito).*

Por meio da Tabela 36 observa-se que X assume os valores 0,1,2 ou 3.

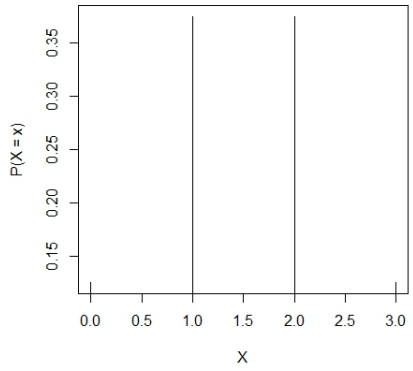


Figura 26 – Representação gráfica para variável aleatória X.

Definição 8 *Uma função de probabilidade de uma v.a discreta é uma função que atribui probabilidade a cada um dos possíveis valores assumidos pela variável.*

Em síntese, uma função de probabilidade é uma função que para cada possível valor de X, digamos x_i, a sua probabilidade de ocorrência é $p(x_i)$, isto é

$$p(x_i) = P(X = x_i), \forall i = 1, 2, \ldots$$

Para exemplificar a definição 8 basta nos reportar a terceira coluna da Tabela 36.

Exemplo 74 *Apresente a função de probabilidade para as seguintes variáveis aleatórias e construa o gráfico para cada uma:*

a) *Número de caras obtido com o lançamento de uma moeda honesta;*

b) *Número de caras obtido com o lançamento de duas moedas honestas;*

c) *Número de peças com defeito em uma amostra de duas peças, sorteadas aleatoriamente de um grande lote, em que 40% das peças são defeituosas;*

Exemplo 75 *Lançam-se dois dados. Seja X a soma das faces, determinar a distribuição de probabilidade de X bem como seu gráfico.*

6.3.1 Propriedades da Função de Probabilidade

A função de probabilidade da v.a X deve satisfazer:

1. $0 \leq p(x_i) \leq 1$;
2. $\sum_{i=1} p(x_i) = 1$.

Exemplo 76 *Dada a Tabela a seguir*

Tabela 37 – Dados do Exemplo 75

X	0	1	2	3	4	5
P(X)	0	P	P^2	P^2	P	P^2

a) *Ache p;*

$$0 + P + P^2 + P^2 + P + P^2 = 1$$
$$3P^2 + 2P - 1 = 0$$
$$P = 1/3$$

b) *P(X $\geq$ 4) e P($X < 3$);*

$$P(X \geq 4) = P(4) + P(5) = 1/3 + 1/9 = 4/9$$
$$P(X < 3) = P(2) + P(1) + P(0) = 1/9 + 1/3 = 4/9$$

c) *P($|X - 3| < 2$);*

$$P(|X - 3| < 2) = P(2) + P(3) + P(1) = 1/9 + 1/9 + 1/3 = 5/9$$

6.3.2 Função de distribuição acumulada

Definição 9 *Seja X uma v.a discreta. Uma função de distribuição ou função de distribuição acumulada (fda) é definida por*

$$F(X) = P(x \in (-\infty; x]) = P(X \leq x)$$

em que x percorre todos os números reais.

Conhecer a função de probabilidade nos permite obter qualquer informação sobre a variável, mesmo que ela assuma apenas alguns valores, visto que percorre toda a reta dos números reais.

Para uma variável aleatória discreta, a fda tem forma de escada, sendo descontínua nos valores que a variável assume. Considerando o exemplo inicial dessa seção, a fda pode ser escrita como

$$\text{F(X)} = \begin{cases} 0 & se \quad x<0 \\ \frac{1}{8} & se \quad 0 \leq x<1 \\ \frac{4}{8} & se \quad 1 \leq x<2 \\ \frac{7}{8} & se \quad 2 \leq x<3 \\ 1 & se \quad 3 \geq x \end{cases}$$

com gráfico dado por

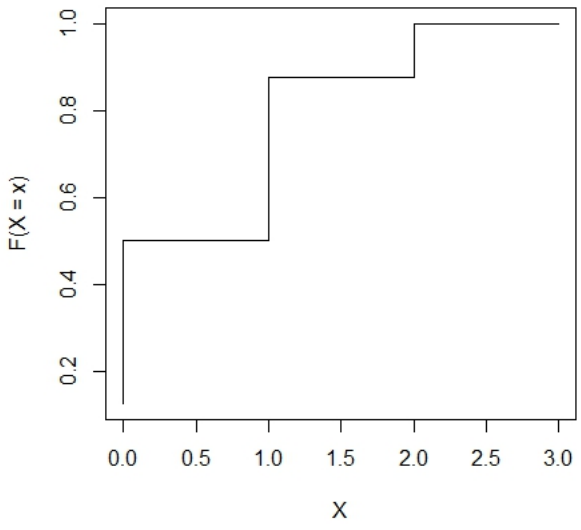


Figura 27 – Função distribuição acumulada.

6.4 Variável aleatória contínua

Para introduzir o conceito de v.a contínua considere o exemplo a seguir:

Exemplo 77 *Considere a situação em que é avaliado o tempo de duração (em anos) de uma certa lâmpada especial. A partir da situação é possível obter a Tabela 38.*

Tabela 38 – Tempo de duração em anos de lâmpadas especiais.

Duração	Probabilidade
1	0,043
2	0,171
3	0,571
4	0,171
5	0,043

O interesse está em estudar a v.a X, tempo de duração em anos de uma certa lâmpada especial. Para distribuição da variável X, é possível obter o histograma a seguir,

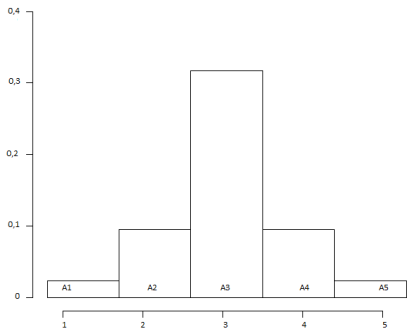
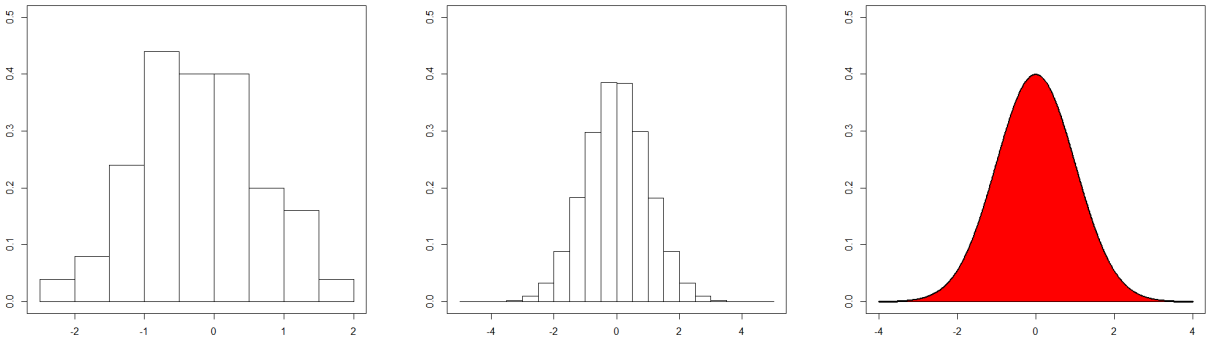


Figura 28 – Histograma para distribuição da variável X.

sendo que a base de cada retângulo mede 1 com altura $p(x = x_0)$.

Suponha que os tempos de duração de duração fossem aproximados por meses, dias, horas e assim por diante. Nesse caso, as informações fornecidas tornam-se cada vez mais precisas, a medida que são exploradas outras unidades de medida, o número de retângulos no histograma cresce, até que o polígono de frequência de origem a uma curva como observa-se na Figura 29c. Essa generalização, pode ser observada na Figura 29k.



(a) (b) (c)

Figura 29 – Histogramas de probabilidade aproximadas.

Se desejarmos obter a probabilidade de duração de uma lâmpada, em um ponto aleatório, selecionado em um intervalo [a,b] é igual a área sob a curva ajustada entre dois pontos a e b. Isso nos motiva a próxima definição.

Definição 10 *Uma variável aleatória X, com função de distribuição F, será classificada como contínua se existir uma função não-negativa f tal que*

$$F(X) = \int_{\infty}^{x} f(w)dw, \forall \quad x \in R,$$

em que f é chamada de ***função densidade de probabilidade*** *(fdp).*

6.4.1 Propriedades da função densidade de probabilidade.

Propriedade 5 *A função densidade de probabilidade de f, deve satisfazer*

1. $f(x) \geq 0, \forall \quad x \in R,$
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$

Observação: Para obter a probabilidade de uma v.a contínua X em um intervalo [a,b] basta calcular a área abaixo da curva de f(x), isto é,

$$P(a \leq x \leq b) = \int_{a}^{b} f(x)d(x) = F(b) - F(a).$$

sendo que a última igualdade decorre do Teorema Fundamental do Cálculo.

Exemplo 78 *Seja a variável aleatória T definida como o tempo de resposta na consulta a um banco de dados, em minutos. Suponha que essa variável aleatória tenha a seguinte fdp,*

$$f(x) = \begin{cases} 2e^{-2t} \quad para \quad t \geq 0 \\ 0 \quad para \quad t < 0 \end{cases}$$

determine:

a) *Que f satisfaz as propriedades de uma fdp;*

b) *P(x>3).*

Exemplo 79 *A direção de uma imperfeição em relação a uma linha de referência em um objeto circular como um pneu, um rotor de freio ou um volante normalmente apresenta alguma incerteza. Considere a linha de referência que conecta a válvula do pneu até o ponto central e seja X o ângulo medido no sentido horário até o local da imperfeição. Uma fdp possível de X é*

$$f(x) = \begin{cases} \frac{1}{360} \quad para \quad 0 \leq x < 360 \\ 0 \quad caso \quad contrário. \end{cases}$$

qual a probabilidade de um ângulo estar entre 90º e 180º?

6.4.2 Função de distribuição acumulada

De acordo com Devore (2006), a função de distribuição acumulada (fda) F(x) de uma v.a discreta X fornece, para qualquer número específico x, a probabilidade $P(X \leq x)$. Já a fda de uma v.a contínua fornece as mesmas probabilidades $P(X \leq x)$ e é obtida pela integração da fdp f(y) entre os limites $-\infty$ e x.

Definição 11 *Seja X uma v.a contínua com fdp f. A função de distribuição acumulada F(x) de uma v.a contínua X é definida para cada número x por*

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

Pela Figura 30 observa-se que para cada x, F(X) é a área abaixo da curva de densidade à esquerda de x.

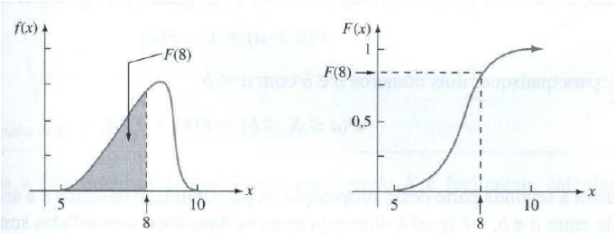


Figura 30 – Fdp e fda correspondentes.

Ainda, observe que se X é uma v.a contínua com fdp f(x) e fda F(x), para qualquer número a vale

$$P(X > a) = 1 - F(a)$$

e para quaisquer dois números a e b com $a < b$,

$$P(a \leq X \leq b) = F(b) - F(a).$$

Para compreender a expressão 6.1, podemos observar a Figura 31

Vale observar que a partir da fda de uma fdp é possível obter qualquer probabilidade. Assim para a < b, temos que:

$$\begin{aligned} P(X < a) = P(\leq a) &= F(a) \\ P(X > b) &= 1 - F(b) \\ P(a < X < b) &= F(b) - F(a) \end{aligned}$$

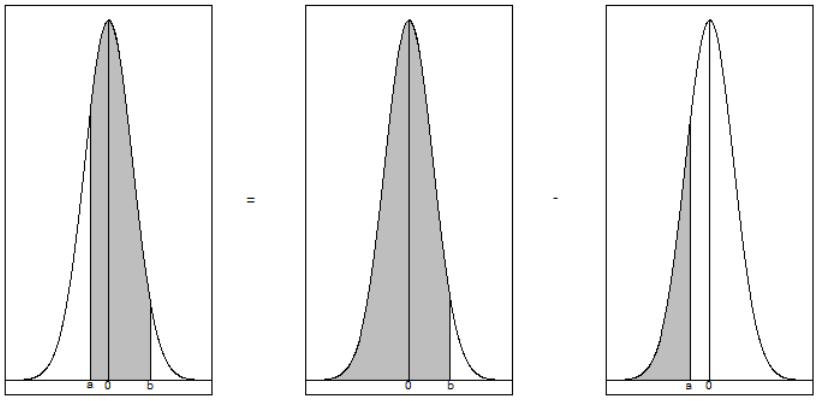


Figura 31 – Cálculo de $P(a \leq X \leq b)$ pelas probabilidades acumuladas

Observação: Dada a função de distribuição acumulada F, podemos obter a função densidade de probabilidade f por:

$$f(x) = \frac{d}{dx}F(x)$$

para todo ponto x em que F é derivável. Assim, a função F também caracteriza a distribuição de probabilidades de um variável aleatória.

Exemplo 80 *Suponha que o erro envolvido ao se fazer uma certa medida seja uma v.a contínua com fdp*

$$f(x) = \begin{cases} 0,09375(4-x^2) \quad para \quad -2 \leq x \leq 2 \\ 0 \quad caso \quad contrário. \end{cases}$$

a) *Cálculo* $P(X > 0)$.

$$F(x) = 0,09375(4X - \frac{x^3}{3} + \frac{16}{3})$$

$$P(X > 0) = 1 - F(0) = 1 - 0,09375(16/3) = 0,5$$

b) *Calcule* $P(-1 < x < 1)$

$$P(-1 < x < 1) = \int_{-1}^{1} 0,09375(4-x^2)dx = 0,6875$$

c) *Calcule* $P(X < -0,5$ *ou* $X > 5)$.

Exemplo 81 *Suponha que a fdp da grandeza X de uma carga dinâmica em uma ponte (em newtons) seja dada por:*

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & para \quad 0 \leq x \leq 2 \\ 0 & caso \quad contrário. \end{cases}$$

determine:

a) *F(x);*

$$F(x) = \frac{3}{16}x^2 + \frac{1}{8}x$$

b) $P(1 \leq X \leq 1,5)$

$$P(1 \leq X \leq 1,5) = F(1,5) - F(1) = \frac{3}{16}(1,5)^2 + \frac{1}{8}(1,5) - \frac{3}{16} - \frac{1}{8} = 0,3906$$

c) $P(X > 1)$

$$P(X > 1) = 1 - F(1) = 1 - \frac{3}{16} - \frac{1}{8}$$

Exemplo 82 *Seja X a v.a contínua tempo de avanço entre dois carros consecutivos selecionados aleatoriamente (em segundos). Suponha que, em um ambiente de tráfego diferente, a distribuição do tempo de avanço tenha forma*

$$f(x) = \begin{cases} \frac{k}{x^4} & para \quad x > 1 \\ 0 & para \quad x \leq 1. \end{cases}$$

a) *Determine o valor de k para qual f(x) é uma fdp.*

b) *Obtenha a função de distribuição acumulada.*

c) *Use a fdc do item anterior para determinar a probabilidade de o tempo de avanço exceder 2 segundos e também a probabilidade de ele estar entre 2 e 3 segundos.*

6.5 Valore Esperado (Esperança/Média) e Variância de uma variável aleatória

6.5.1 Valor Esperado ou Esperança ou Média

O valor esperado ou esperança matemática ou média de uma variável aleatória é muito utilizada(o) como medida resumo de distribuições de probabilidade.

Acredita-se, que os primeiros registros do uso do conceito de valor esperado esteja associado a ganhos em apostas com dinheiro. Posteriormente, com o formalismo matemático, teorias foram formalizadas para casos em que as variáveis aleatórias sejam discretas ou contínuas.

De modo geral, existem características numéricas que são muito importantes em uma distribuição de probabilidade, seja ela discreta ou contínua. São os *parâmetros* das distribuições.

Um primeiro parâmetro é a *esperança matemática* (ou simplesmente média) de uma variável aleatória.

6.5.1.1 Média para v.a discretas

Para introduzirmos o conceito de média, para variáveis discretas, considere o exemplo a seguir.

Exemplo 83 *Uma empreiteira paga R$ 30.000,00 em caso de atraso na entrega de obras residenciais e cobra uma taxa extra de R$ 1.000,00 de seus clientes por encargos. Sabe-se que a probabilidade de atraso na entrega das obras é de 3%. Quanto a empreiteira espera ganhar na realização de 100 obras ao longo de um período determinado?*

Observe que a expressão para média do exercício anterior pode ser escrita por meio da expressão

$$E(X) = x_1 p(x_1) + x_2 p(x_2)$$

Isso nos motiva a definição a seguir:

Definição 12 *Seja X uma v.a discreta com função distribuição de probabilidade p_x e o valor x_i para i em algum conjunto de índices I. Então, o valor esperado de X é definido por*

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

Exemplo 84 *Num jogo de dados, A paga R$ 20,00 e B lança 3 dados. Se sair face 1 em um dos dados apenas, A ganha R$ 20,00. Se sair face 1 em dois dados apenas, A ganha R$ 50,00 e se sair 1 nos três dados, A ganha R$ 80,00. Calcular o lucro líquido médio de A em uma jogada.*

Exemplo 85 *Suponhamos que um número seja sorteado de 1 a 10, inteiros positivos. Seja X o número de divisores do número sorteado. Calcular o número médio de divisores do número sorteado.*

6.5.1.2 Média para v.a contínuas

A construção feita para encontrar a expressão para cálculo do valor esperado de uma v.a discreta pode ser estendida para variáveis contínuas, porém fica a cargo do leitor.

Definição 13 *Seja X uma v.a contínua com f sendo a sua fdp. Então a sua média é dada pela expressão:*

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

desde que a integral acima esteja bem definida.

A integral definida estar bem definida significa que quando separarmos a expressão 6.1 em

$$E(X) = \int_{-\infty}^{0} x f(x) dx + \int_{0}^{+\infty} x f(x) dx$$

E(X) estará bem definida se a integral em pelo menos um dos intervalos determinados, for finita.

Segundo Magalhães (2011), nomear o valor esperado de uma variável aleatória tem origem histórica e também pode ser visto como uma referência a um resultado importante como *Lei dos Grandes Números*.

Exemplo 86 *Considerando os dados do exemplo 78, qual o tempo médio de duração das lâmpadas especiais?*

Exemplo 87 *A variável contínua X tem função de distribuição dada pela expressão a seguir:*

$$F(x) = \begin{cases} 0, & \text{para} \quad x < 0; \\ \dfrac{x^2}{4} & \text{para} \quad 0 \leq x < 1; \\ \dfrac{2x-1}{4} & \text{para} \quad 1 \leq x < 2; \\ -\dfrac{x^2-6x+5}{4} & \text{para} \quad 2 \leq x < 3; \\ 1 & \text{para} \quad x > 3. \end{cases}$$

Obter o valor esperado para essa variável aleatória.

6.5.2 Propriedades da Média

Seja k uma constante e X e Y variáveis aleatórias, as seguintes relações podem ser comprovadas:

a) E(k) = k;

b) E(X+k) = E(X)+k;

c) E(kX) = kE(X);

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

Observação: As demonstrações seguem diretamente da definição, tanto para o caso contínuo quanto para o caso discreto. Para obtê-las, consulte Morettin (2010) e Barbetta, Reis e Bornia (2010).

6.6 Variância de uma variável aleatória.

Assim como a média de uma v.a, a variância de uma v.a é um parâmetro de uma distribuição de probabilidade que caracteriza a dispersão dos dados em torno da média.

Definição 14 *Seja X uma variável aleatória, com E(X)* $< \infty$*, define-se variância de X como*

$$\begin{aligned} \sigma^2 = Var(X) &= E\left\{(X-\mu_x)^2\right\} = E(X^2) - \mu_x^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

É fácil verificar que quanto menor a variância, menor o grau de dispersão de probabilidades em torno da média e vice-versa e quanto maior a variância, maior o grau de dispersão da probabilidade em torno da média.

6.6.1 Propriedades da Variância

Seja k uma constante e X e Y variáveis aleatórias, as seguintes relações podem ser comprovadas:

a) Var(k) = 0;

b) Var(kX) = k^2Var(X);

c) Var(X $\pm$ Y) = Var(X) $+$ Var(Y) $\pm$ 2 cov(X,Y)

Observação: As demonstrações seguem diretamente da definição, tanto para o caso contínuo quanto para o caso discreto. Para obtê-las, consulte Morettin (2010) e Barbetta, Reis e Bornia (2010).

Em determinadas situações é necessário trabalhar com a unidade da variável que está sendo avaliada, e a variância é um quadrado. Nesse sentido, para resolver essa situação pode-se utilizar o **desvio padrão** de uma variável, o qual é definido a seguir.

Definição 15 *Seja X uma variável aleatória X com variância dada por Var(X). O desvio padrão de X é dado por*

$$\sigma_x = \sqrt{Var(X)}$$

Fazendo uso da tabela da distribuição normal (distribuição de probabilidade que será estudada na sequência), é possível verificar que os intervalos a seguir contemplam porcentagens específicas do volume de dados, isto é:

a) $P(\mu - \sigma_x \leq X \leq \mu + \sigma_x) \approx 68\%$;

b) $P(\mu - 2\sigma_x \leq X \leq \mu + 2\sigma_x) \approx 95\%$;

c) $P(\mu - 3\sigma_x \leq X \leq \mu + 3\sigma_x) \approx 99,7\%$;

Podemos observar essa situação por meio de um diagrama, como segue:

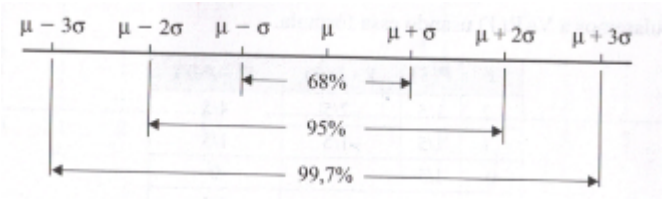


Figura 32 – Porcentagem dos dados em torno da média e desvio padrão

Para ilustrar as definições e conceitos acima, considere o exemplo a seguir:

Exemplo 88 *Os empregados A, B, C e D ganham 1, 2, 2 e 4 salários mínimos, respectivamente. Retiram-se amostras com reposição de 2 indivíduos e mede-se o salário médio da amostra retirada. Qual a média e desvio padrão do salário médio amostral?*

6.7 Modelos Discretos e Contínuos de Probabilidade

Atualmente existe um grande número de distribuições de probabilidade presentes na Literatura, os quais crescem constantemente. Em sua grande maioria, elas surgem com o intuito de modelar fenômenos cotidianos. Pesquisadores de diversas áreas da estatística

as desenvolvem com intuito de resolver problemas práticos, cuja base são distribuições já existentes.

Nas seções subsequentes serão apresentados os principais modelos discretos e contínuos de probabilidade.

6.7.1 Modelos Discretos

Considerando o grande número de distribuições que podemos encontrar na literatura, existem algumas clássicas, isto é, consagradas teoricamente e computacionalmente e que são utilizadas para modelar fenômenos discretos, como as que seguem:

- Distribuição de *Bernoulli*;
- Distribuição Binomial;
- Distribuição Geométrica;
- Distribuição Hipergeométrica;
- Distribuição de Poisson.

A seguir, faremos um estudo detalhado de cada uma delas.

6.7.1.1 Distribuição de Bernoulli ou Modelo Bernoulli

Definição 16 *Uma variável aleatória segue modelo de Bernoulli, se assume apenas valores 0 ou 1. Sua função de probabilidade é dada por:*

$$p(1) = P(X = 1) = p;$$
$$p(0) = P(X = 0) = 1 - p;$$

No modelo de Bernoulli, a probabilidade *p* é denominada de parâmetro do modelo. É prática comum considerar como *sucesso* a ocorrência de 1 e *fracasso* a ocorrência de 0. Logo, denominamos por ensaio de *Bernoulli*, o experimento que tem resposta dicotômica do tipo *sucesso-fracasso*. Alguns exemplos de ensaios de Bernoulli são:

a) Lançar uma moeda e observar se ocorre cara ou não;

b) Rodar um algoritmo e o resultado apresentar falha ou não;

c) Entregar uma obra no prazo determinado ou não;

d) Um carro apresentar defeito mecânico ou não;

e) Uma lâmpada funcionar ou não;

Em geral, a distribuição fica inteiramente caracterizado ao atribuir um valor para o parâmetro *p*.

Os dois primeiros momentos da distribuição de *Bernoulli* podem ser escritos como:

$$E(X) = p \quad e \quad Var(X) = p(1-p)$$

A notação utilizada será $X \sim B(p)$.

6.7.1.2 Distribuição Binomial ou Modelo Binomial

Em geral, em um experimento são realizados n ensaios de *Bernoulli*. O objetivo é identificar o número de ocorrências de X classificadas como sucesso, como nos exemplos a seguir:

a) Lançar uma moeda 100 vezes e observar se ocorre cara ou não em cada lançamento;

b) Rodar um algoritmo n vezes e verificar se o resultado apresentar falha ou não em cada tentativa;

Nesses casos tem-se *experimentos binomiais*.

Para determinarmos a função de probabilidade de uma variável X binomial, considere o exemplo a seguir:

Exemplo 89 *Uma indústria fabricante de componentes eletrônicos classifica os carregamentos de peças que chegam a suas instalações em A, B ou C. Suponha independência entre as chegadas de carregamentos. Suponha que a probabilidade p de classificação na classe A é a mesma para todos os carregamentos. Para os próximos carregamentos, seja X a v.a que representa o* ***número de carregamentos classificados na classe A****. Calcular a probabilidade de que X assume o valor x, isto é, a probabilidade de que x carregamentos sejam classificados na classe A (x=0,1,2,3,4), supondo que seja S quando este for classificado na classe A e seja F quando este for classificado em outra classe.*

A partir dos resultados acima, podemos generalizar a situação, caracterizando a **distribuição Binomial** de probabilidade.

Definição 17 *Seja X uma v.a com distribuição binomial de parâmetros n e p (sendo* $0 < p < 1$*). A probabilidade de X assumir um certo valor k pertencente ao conjunto* $\{0, 1, 2, \ldots, n\}$*, é dada por:*

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

A notação utilizada será $X \sim B(n, p)$*.*

Os dois primeiros momentos da distribuição de *Bernoulli* podem ser escritos como:

$$E(X) = np \quad e \quad Var(X) = np(1-p). \tag{6.1}$$

O termo $\binom{n}{k}$ representa o número de combinações que podem ser feitas com k elementos numa sequencia de n elementos (sendo $k \leq n$). Esse número de combinações pode ser calculado pela seguinte expressão:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{6.2}$$

Exemplo 90 *Considerando os dados do exemplo que motivou o uso da distribuição binomial, historicamente 30% dos carregamentos são classificados na classe A, em que podemos supor que a probabilidade p de um carregamento ser classificado na classe A é 0,3. Entre os quatro próximos carregamentos, qual é a probabilidade de que exatamente 2 sejam pertencentes a classe A?*

***Solução**. Seja X a variável aleatória número de carregamentos classificados na classe A. Nesse sentido, temos que n=4 e k=2, com probabilidade de sucesso 0,7 e fracasso 0,3. Nesse caso temos que a probabilidade de ocorrer dois sucessos em 4 tentativas é*

$$P(X = 2) \quad = \quad \binom{4}{2} 0,7^2 (1-0,7)^{4-2} = 0,2646$$

Para resolver no R, basta seguir os comandos abaixo.

```
p <- 0.7 #probabilidade de sucesso
n <- 4   #quantidade de eventos indpendentes
k <- 2   #numero de sucessos
dbinom(k, n, p)
```

Exemplo 91 *Uma equipe de programadores lança uma versão de um determinado software a cada mês. Sabendo que em 10% é a chance do software lançado apresenta algum bug, durante um ano, qual a probabilidade de exatamente 4 das versões saírem da fabrica de software com defeito?* ***Solução:***

Seja X a variável aleatória número de programas que apresentam algum bug em um ano. Nesse sentido, temos que n=12 e k=4, com probabilidade de sucesso 0,10 e fracasso 0,90. Nesse caso temos que a probabilidade de ocorrer quatro sucessos em doze tentativas é

$$P(X=4) \;=\; \binom{12}{4} 0,1^4 (1-0,1)^{12-4} = 0,02130$$

Para resolver no R, basta seguir os comandos abaixo.

```
p <- 0.1   #probabilidade de sucesso
n <- 12    #quantidade de eventos indpendentes
k <- 4     #numero de sucessos
dbinom(k,n,p)
```

Exemplo 92 *Em uma rede de computadores, em 50% dos dias ocorre alguma falha. Considere a variável aleatória X = número de dias com falha na rede. Considere o período de observação de 30 dias e suponha que os eventos são independentes. Qual a probabilidade de ocorrer exatamente 12 dias de falha na rede, considerando os mesmos 30 dias de observação?*

Exemplo 93 *O setor de teste de uma determinada empresa realiza 150 testes por dia com uma ferramenta de teste, sendo que 1/4 desses testes não são bem sucedidos. Qual a probabilidade de exatamente 30 desses testes não obtenham sucesso?*

Exemplo 94 *Um inspetor de qualidade extrai uma amostra de 10 tubos aleatoriamente de uma carga muito grande de tubos que se sabe que contém 20% de tubos defeituosos. Qual é a probabilidade de que não mais do que 2 tubos extraídos sejam defeituosos?*

Exemplo 95 *A probabilidade de um arqueiro acertar um alvo com uma única flecha é de 0,20. Lançando 30 flechas no alvo. Qual a probabilidade de que:*

a) *Acerte exatamente 4 flechas no alvo?*

b) *Acerte exatamente 6 flechas no alvo?*

Exemplo 96 *Um técnico visita os clientes que compraram assinatura de um canal de TV para verificar o decodificador. Sabe-se, que 90% desses aparelhos não apresentam defeitos. Determinar a probabilidade de que em 20 aparelhos pelo menos 17 não apresentem defeitos.*

6.7.1.3 Distribuição Hipergeométrica ou Modelo Hipergeométrico

Por motivar o estudo da distribuição Hipergeométrica, considere um problema básico de inspeção por amostragem, em que observamos uma amostra de n itens de um lote com N itens, sendo r defeituosos. Avaliamos o número X de itens defeituosos na amostra. A variável aleatória pode ser confundida com uma variável binomial. Todavia, será realmente binomial se:

a) A seleção da amostra for aleatória;

b) Com reposição.

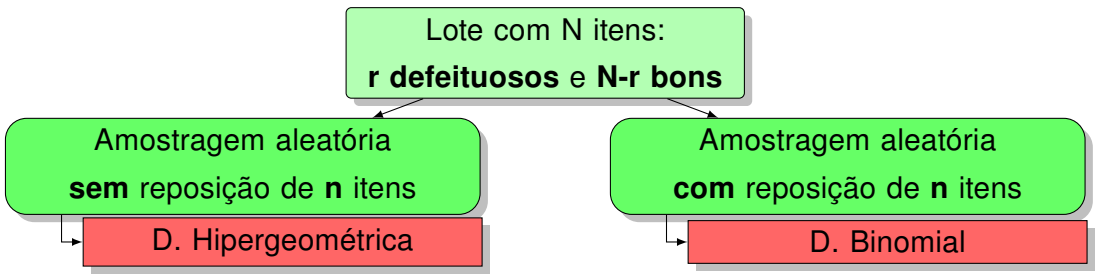
Em contrapartida, as hipóteses que levam à distribuição hipergeométrica são as seguintes:

a) A população ou conjunto de onde é retirada a amostra consiste de N indivíduos, objetos ou elementos (*população finita*).

b) Cada indivíduo é classificado como sucesso (S) ou falha (F) e há r sucessos na população.

c) É selecionada uma amostra sem reposição de n indivíduos de forma que cada subconjunto de tamanho n seja igualmente provável de ser escolhido.

Em geral a segunda condição não costuma ser satisfeita, uma vez que avaliações costuma ser **sem** reposição. Nesse caso, é necessário utilizar uma distribuição de probabilidade alternativa, que no caso é chamada de **Distribuição Hipergeométrica** com parâmetros N, n e r. Podemos representar sua lógica pelo diagrama abaixo.



A função de probabilidade de X é expressa por:

$$P(X=k)=\frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}} \quad se \quad k=0,1,\ldots,min(r,n) \tag{6.3}$$

A notação utilizada será $X \sim Hgeo(N,n,k)$. Seu valor esperado e variância podem ser calculados por meio da definição, os quais são dados por:

$$E(X)=np \quad e \quad Var(X)=np(1-p)\frac{N-n}{N-1}. \tag{6.4}$$

Exemplo 97 *Placas de vídeo são expedidas em lotes de 30 unidades. Antes que a remessa seja aprovada, um inspetor escolhe aleatoriamente cinco placas do lote e as inspeciona. Se nenhuma das placas inspecionadas for defeituosa, o lote é aprovado. Se uma ou mais forem defeituosas, todo o lote é inspecionado. Supondo que haja três placas defeituosas no lote, qual é a probabilidade de que o controle da qualidade aponte para a inspeção total?*

Solução: *Seja X a variável aleatória número de peças com defeito. Nesse caso, temos uma população com **N=30** elementos, de onde extrai-se uma amostra de **n=5** elementos. Na população há **r=3** elementos defeituosos. Para que o controle de qualidade aponte para inspeção total do lote, devemos obter a probabilidade*

$$P(X>0)=1-P(X\leq 0)=1-P(X=0).$$

Assim, a probabilidade desejada é dada por:

$$\begin{aligned} P(X>0) &= 1-\frac{\binom{3}{0}\binom{30-3}{5-0}}{\binom{30}{5}} \\ &= 1-0,5665 \\ &= 0,4335 \end{aligned}$$

Portanto, a probabilidade de que o controle da qualidade aponte para a inspeção total é aproximadamente 43,35%.

Para resolver no R, basta utilizar os comandos a seguir:

```
N<-30;n1<-5;r<-3;k1<-0
m<-r;x<-k1;n2<-N-m;k2<-n1
prob<-dhyper(x, m, n2, k2, log = FALSE)
1-prob
```

Para resolver demais exemplos no R, basta alterar a primeira linha do código acima.

Exemplo 98 *Cinco indivíduos de uma população animal supostamente ameaçada de extinção em certa região foram capturados, marcados e liberados para se misturarem à população. Após terem uma oportunidade de cruzarem, foi selecionado uma amostra aleatória de 10 desses animais. Seja X= número de animais marcados na segunda amostra. Se, na verdade, houver 25 animais desse tipo na região, qual será a probabilidade de*

a) *X = 2*

b) $X \leq 2$

Observação: É importante ressaltar que quando N é muito maior do que n, a distribuição hipergeométrica pode ser aproximada pela binomial. Muitos autores prescrevem uma relação $\frac{n}{N} \leq 0,05$ para que seja possível fazer a aproximação. Nesse caso, a binomial tem parâmetros n = tamanho da amostra e $p = \frac{r}{N}$.

6.7.1.4 Distribuição Geométrica ou Modelo Geométrico

Considere uma sequência de ensaios de Bernoulli independentes. Defina X como o número de fracassos anteriores ao primeiro sucesso ou, em outras palavras, o tempo de espera (em termos de ensaios anteriores) para o primeiro sucesso.

Definição 18 *Seja X uma variável aleatória discreta, então X segue o modelo Geométrico com parâmetro p,* $0 < p < 1$*, e tem função de probabilidade dada por:*

$$P(X = k) = p(1-p)^k, \quad para \quad k = 0, 1, \ldots \tag{6.5}$$

A notação utilizada será $X \sim Geo(p)$. Seu valor esperado e variância podem ser calculados por meio da definição, os quais são dados por:

$$E(X) = \frac{1}{p} \quad e \quad Var(X) = \frac{1-p}{p^2}. \tag{6.6}$$

A restrição dos valores de p para aqueles estritamente entre 0 e 1 evita os casos triviais mas, com o devido cuidado, os extremos poderiam ser considerados. Alguns autores preferem definir o modelo Geométrico como o número de fracassos até o primeiro sucesso, ou, ainda, o número de ensaios até o primeiro sucesso. Nesse último caso, a função de probabilidade sobre uma modificação, pois a variável inicia sues valores em 1 ao invés de 0.

Exemplo 99 *A probabilidade de se encontrar aberto o sinal de trânsito numa esquina é 0,35. Qual a probabilidade de que seja necessário passar pelo local 5 vezes para encontrar o sinal aberto pela primeira vez?*

Solução: *Nesse problema, desejamos saber qual a chance do sucesso ocorrer na 5ª tentativa, isto é, devem ocorrer 4 fracassos até o primeiro sucesso. Nesse caso, a probabilidade calculada é dada por*

$$P(X = 4) = 0,35(1 - 0,35)^4 = 0,0625.$$

Portanto, a probabilidade de precisar passar 5 vezes no local, para que na quinta tentativa o sinal esteja aberto, é aproximadamente 6,25%.

Para resolver no R, basta utilizar os comandos a seguir:

```
dgeom(4, 0.35, log = FALSE)
```

Exemplo 100 *Qual a probabilidade de que um dado deva ser lançado 10 vezes para que na 10ª ocorra a face 6 pela primeira vez?*

Exemplo 101 *Qual a probabilidade de que uma moeda não viciada deva ser lançada 3 para que na 3ª ocorra cara pela primeira vez?*

6.7.1.5 Distribuição de Poisson ou Modelo de Poisson

Historicamente, a distribuição de Poisson recebe esse nome em homenagem ao seu descobridor, Siméon-Denis Poisson. Em primeiro momento ela foi aplicada em um problema da área das ciências jurídicas (1837 - 1838), porém não chamou atenção dos teóricos da época. No entanto, ganhou destaque ao ser empregada em problemas relacionados a ciência da artilharia. A distribuição foi mais tarde e independentemente descoberta por Von Bortkiewicz, Rutherford, e Gosset. Esta distribuição expressa a probabilidade de um determinado número de eventos que ocorrem em intervalo fixo de tempo. Surge com muita frequência, associada à contagem de acontecimentos aleatórios (quando se pode admitir que não há acontecimentos simultâneos).

Para introduzirmos o conceito de modelo de Poisson, considere as situações em que se avalia o número de ocorrências de um tipo de evento por unidade de tempo, de rompimento, de área, ou de volume.

a) Número de consultas a uma base de dados em um minuto;

b) Número de fissuras em uma parede;

c) Número de defeitos em um m^2 de piso cerâmico;

d) Número de componentes eletrônicos defeituosos de um lote;

Para um processo de Poisson algumas suposições básicas são necessárias, entre as quais é possível destacar:

a) Independência entre as ocorrências do evento avaliado;

b) Os eventos ocorrem de forma aleatória, de tal forma que não haja tendência de aumentar ou reduzir as ocorrências do evento, no intervalo considerado.

Muitas vezes, no uso da distribuição binomial, acontece que n é muito grande, isto é n tende ao infinito, e p é muito pequeno, ou seja, tende a zero. Nesses casos, não encontramos valores pré-determinados ou tabelados, ou então o cálculo torna-se inviável, sendo necessário o uso de calculadoras sofisticadas.

Nesse contexto, é possível fazer uma aproximação da distribuição binomial pela distribuição de Poisson, sempre que:

a) $n \longrightarrow \infty$ (convencionalmente $n > 30$);

b) $p \longleftrightarrow 0$ (convencionalmente $p < 0,1$);

c) $0 < \mu \leq 10$;

Quando isso acontece, tomamos a média da distribuição binomial $\mu = np$ como $np = \lambda$.

Para essa situação, se $X \sim B(n,p)$, calcular $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Isso nos motiva a definição a seguir:

Definição 19 *Uma variável X segue o modelo de Poisson de parâmetro $\lambda > 0$, se sua função de probabilidade for a seguinte:*

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad para \quad k = 0, 1, \ldots \tag{6.7}$$

Usamos a notação $X \sim Poisson(\lambda)$. O parâmetro λ indica a taxa de ocorrência por unidade de medida. Seu valor esperado e variância podem ser calculados por meio da definição, os quais são dados por:

$$E(X) = Var(X) = \lambda. \tag{6.8}$$

Em geral, de acordo com o valor assumido por λ o gráfico da variável Y pode apresentar vários formatos, os quais podem ser visualizados na Figura 33 (clique em cima do gráfico).

Percebe-se que a medida que λ cresce, X se aproximada da distribuição Normal. Devido sua versatilidade, é possível estabelecer relações com as distribuições Binomial (como visto anteriormente) e Normal.

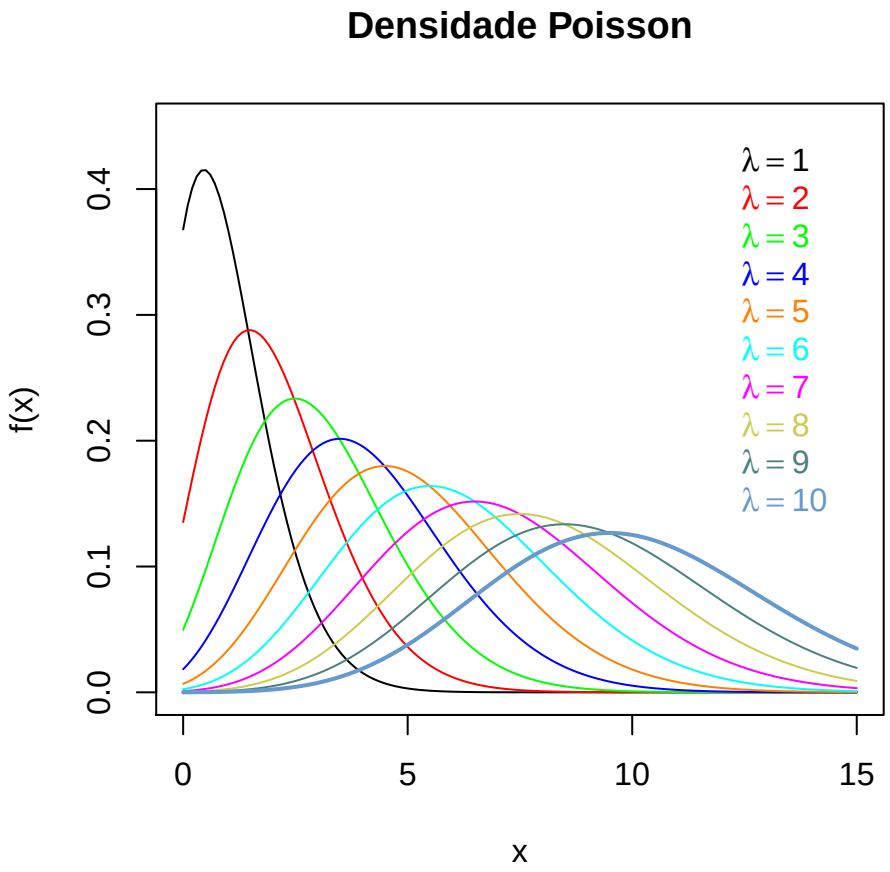


Figura 33 – Densidade de Poisson

Exemplo 102 *Supondo que as consultas num banco de dados ocorrem de forma independente e aleatória, com uma taxa média de três consultas por minuto, calcule a probabilidade de que no próximo minuto ocorram menos que três consultas.*

***Solução**: Seja X a variável aleatória número de consultas em um intervalo de tempo. Assim, temos que*

$$E(X) = V(X) = \lambda = 3,$$
$$1 - P(x = 3) = 1 - \frac{e^{-3}31^3}{3!} = 0,7760.$$

Para resolver no R, basta seguir os comandos abaixo.

```
x= 3
lambda= 3
w = dpois(x,lambda)
[1] 0.2240418

P = 1-w
```

```
P
[1] 0.7759582
```

Exemplo 103 *O artigo "Reliability-Based Service-Life Assessment ofAging Concrete Structures"(J. Structural Engr., 1993, p. 1600-1621) sugere que um processo de Poisson pode ser usado para representar a ocorrência de cargas estruturais no tempo. Suponha que o tempo médio entre as ocorrências de cargas seja 0,5 ano.*

a) *Quantas cargas podem ser esperadas durante um período de dois anos?*

b) *Qual é a probabilidade de mais de cinco cargas ocorrerem durante um período de dois anos?*

Exemplo 104 *Uma banca de jornal faz o pedido de cinco cópias de uma edição de certa revista de fotografia. Seja X = número de indivíduos que chegam para comprar a revista. Se X tiver distribuição de Poisson com parâmetro λ = 4, qual será o número esperado de cópias vendidas?*

6.7.2 Modelos Contínuos de Probabilidade.

Distribuições contínuas de probabilidade são utilizadas em diversas áreas do conhecimento. Em engenharia é comum que os fenômenos aleatórios tenham características contínua. Assim, faz-se necessário o uso de modelos probabilísticos para modelagem de dados assim caracterizados.

Entre as distribuições presentes na literatura, é comum encontrarmos as que seguem:

- Distribuição Uniforme;
- Distribuição Exponencial;
- Distribuição Normal;
- Distribuição Gama;
- Distribuição Beta;

entre outras.

Nas seções subsequentes fazermos um estudo superficial das distribuições mencionadas acima.

6.7.2.1 Distribuição Uniforme Contínua

Exemplo 105 *Considere um círculo, com medidas de ângulos, em graus, a partir de determinada origem, como mostra a Figura a seguir. Nesse caso, temos um círculo e um ponteiro que é colocado a girar. A variável aleatória de interesse é X = ângulo formado entre a posição que o ponteiro para e a linha horizontal do lado direito. Supõe-se, também, não existir região de preferência para o ponteiro parar. Nessas condições, podemos considerar que todo intervalo de mesma amplitude, contido em [00, 3600), tem a mesma probabilidade de ocorrência. É um experimento típico em que a chamada* ***distribuição uniforme*** *é apropriada.*

Isso nos motiva a definição de distribuição uniforme, como segue:

Definição 20 *Seja X uma variável aleatória contínua. Dizemos que X segue o modelo uniforme com parâmetros a e b, com $b > a$ se se sua função densidade de probabilidade (fdp) é especificada por:*

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{se} \quad x \in [a,b]; \\ 0 & \text{caso contrário} \end{cases}$$

Usualmente, se X segue distribuição uniforme contínua, então representamos por $X \sim U_c[a,b]$. Graficamente temos:

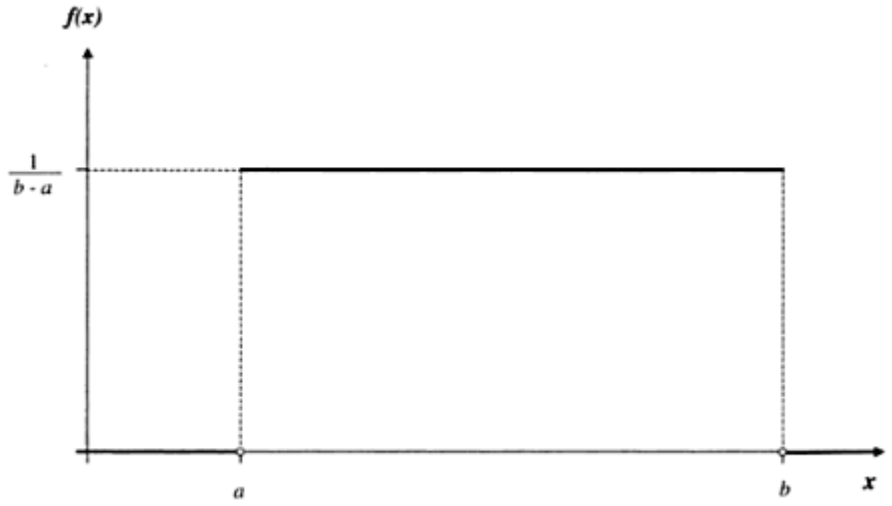


Figura 34 – Densidade da Uniforme Contínua [a,b].

Exemplo 106 *Seja X uma v.a contínua tal que segue modelo uniforme.*

a) *Se f(x) satisfaz as propriedades de um fdp;*

b) *Determine sua função densidade acumulada e seu gráfico.*

Exemplo 107 *Suponha que a temperatura de reação X (em graus celsius), em certo processo químico, tenha função de densidade uniforme a = -5 e b = 5.*

a) *Calcule* $P(X < 0)$;

b) *Calcule* $P(-2,5 < X < 2,5)$;

c) *Calcule* $P(-2 \leq X \leq 3)$.

6.7.2.2 Distribuição Exponencial

Um modelo com diversas aplicações, em diversas áreas do conhecimento com engenharia e matemática é o modelo *Exponencial*. Tempo de vida de equipamentos, intervalos entre chegadas de mensagens eletrônicas ou de chamadas telefônicas a uma central, são exemplos de problemas que podem ser modelados por meio desta distribuição.

Na distribuição de Poisson, modelamos variáveis aleatórias cujo objetivo é estudar o número de ocorrências em um certo período. Já na distribuição exponencial, a variável aleatória é definida como tempo entre duas ocorrências. A Figura 35 nos mostra as características de cada uma.

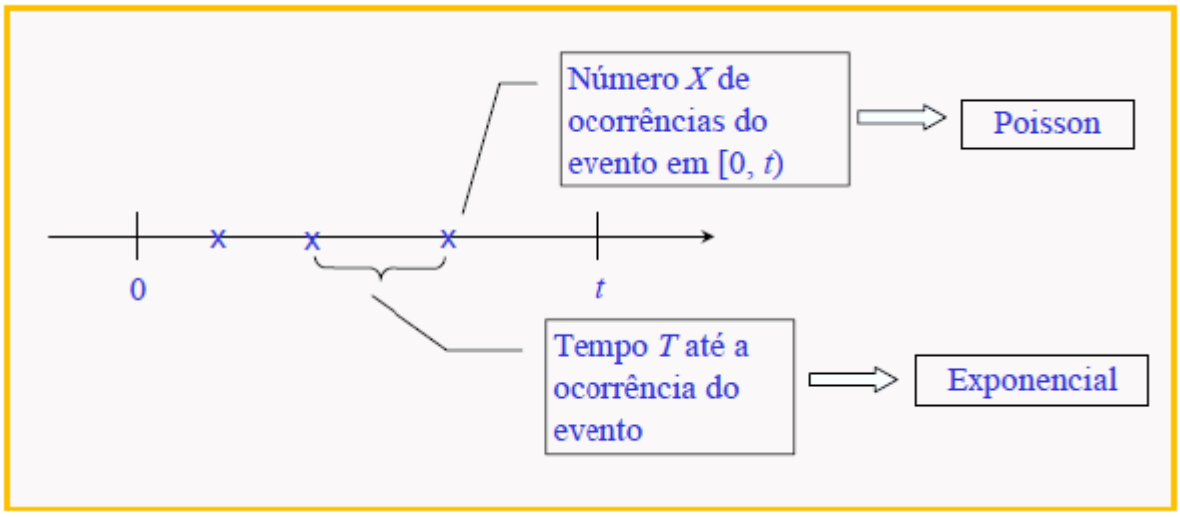


Figura 35 – Distribuição Poisson e Distribuição Exponencial.

Considere a seguinte equivalência entre eventos:

A primeira ocorrência de um evento ser depois de um tempo t $\Leftrightarrow$ Nenhuma ocorrência no intervalo [0,t)

Diante disso, é possível considerarmos as seguintes variáveis aleatórias:

- X_t = Número de ocorrências no intervalo [0,t);

- T = Tempo entre as ocorrências.

Sendo que λ representa a taxa média de ocorrências por unidade de tempo, então, considerando variáveis aleatórias independentes na realização de X_t, então essa v.a tem distribuição de Poisson com parâmetro λt. Observe que

$$T > t \Leftrightarrow X_t = 0$$

Estabelecendo uma relação entre a distribuição de probabilidade de Poisson temos:

$$P(T > t) = P(X_t = 0) = \frac{(\lambda t)e^{-\lambda t}}{0!} = e^{-\lambda t} \tag{6.9}$$

Considerando que acabou de ocorrer algum evento de interesse, o evento complementar indica o tempo até a próxima ocorrência, e assim, temos que a fda de distribuição acumulada de uma variável aleatória T com distribuição exponencial:

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t} \tag{6.10}$$

Consequentemente, a fdp para t ≥ 0 temos a função densidade de probabilidade dada por

$$f(t) = \lambda e^{-\lambda t} \tag{6.11}$$

Essa construção nos motiva a definição a seguir:

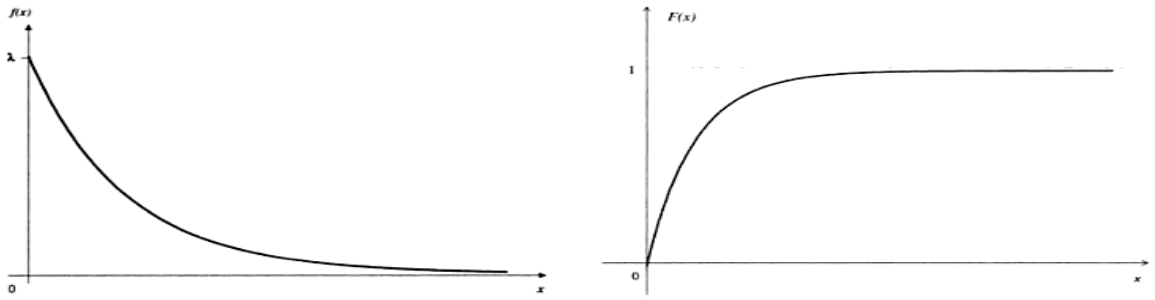
Definição 21 *Seja T uma variável aleatória contínua. Diz-se que T segue o modelo Exponencial de parâmetro $\lambda > 0$, se tiver densidade dada por:*

$$\text{f(t)} = \begin{cases} \lambda e^{-\lambda t}, & \text{se} \quad t \geq 0; \\ 0 & \text{se} \quad t < 0; \end{cases}$$

Quando T segue distribuição exponencial de parâmetro λ, escreve-se $T \sim Exp(\lambda)$. Seu valor esperado e variância podem ser calculados por meio da definição, os quais são dados por:

$$E(T) = \frac{1}{\lambda} \quad e \quad Var(T) = \frac{1}{\lambda^2}. \tag{6.12}$$

Graficamente temos:



(a) Densidade da Exponencial (λ). (b) Densidade Acumulada da Exponencial (λ).

Figura 36 – Gráfico de Densidade da Distribuição Exponencial

Exemplo 108 *O tempo de vida (em horas) de um transistor é uma variável aleatória T com distribuição exponencial. O tempo médio de vida de transistor é de 500.*

a) *Calcule a probabilidade de o transistor durar mais do que 500 horas.*

b) *Calcule a probabilidade de o transistor durar entre 300 e 1000 horas.*

c) *Sabendo que o transistor já durou 500 horas, calcule a probabilidade de ele durar mais 500 horas.*

Exemplo 109 *Suponha que o tempo de resposta X em um terminal de computador online específico (o tempo entre o final de uma consulta de um usuário e o começo da resposta do sistema para essa consulta) tenha distribuição exponencial com tempo de resposta esperado igual a 5 segundos. Qual a probabilidade de o tempo de resposta ser no máximo 10 segundos?*

6.7.2.3 Distribuição Normal

Suponha que desejamos conhecer qual é a probabilidade de que uma determinada construtora entregue uma obra no prazo estabelecido. Nesse sentido deveríamos conhecer o número de obras que foram feitas por essa construtora, assim como o número de vezes que ela foi eficaz.

Qual a disponibilidade do pesquisador para obter os dados necessários para se efetuar o cálculo da probabilidade desejada?

Em um contexto geral, essa disponibilidade é ínfima, ou seja, não conseguimos todas as informações desejadas. O que acontece na prática, é utilizar uma distribuição de probabilidade a qual baseia-se em apenas dois valores, o desvio padrão e a média do conjunto de dados. Tal distribuição é conhecida como **Distribuição Normal**.

A distribuição normal ou gaussiana é a mais importante distribuição de probabilidade no contexto estatístico. É por meio dela que modelam-se uma diversidade de fenômenos aleatórios. Os exemplos incluem alturas, pesos e outras características físicas, erros de medidas. Nesse contexto, temos a definição a seguir:

Definição 22 *Uma variável aleatória contínua X tem **distribuição normal** com parâmetros μ e σ, em que $-\infty < x < \infty, \sigma > 0, -\infty < \mu < \infty$, se a sua função densidade de probabilidade (fdp) é dada pela expressão:*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

Quando X segue distribuição normal denota-se por $X \sim N(\mu, \sigma^2)$. O gráfico de $f(x)$ pode ser representado por

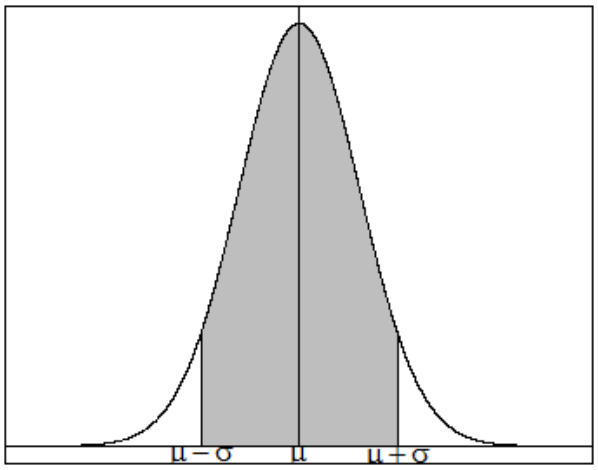


Figura 37 – Gráfico da fdp normal.

As principais características dessa função são:

- O ponto de máximo de $f(x)$ é o ponto X = μ.
- Os pontos de inflexão da função são: $X = \mu - \sigma$ e $X = \mu + \sigma$.
- A curva é simétrica com relação a μ.
- A média é $E(X) = \mu$ e a variância $VAR(X) = \sigma^2$.

Como uma variável aleatória contínua possui resposta em um intervalo, podemos ter interesse em calcular a probabilidade de que algo ocorra no intervalo [a,b]. Observe a Figura 38.

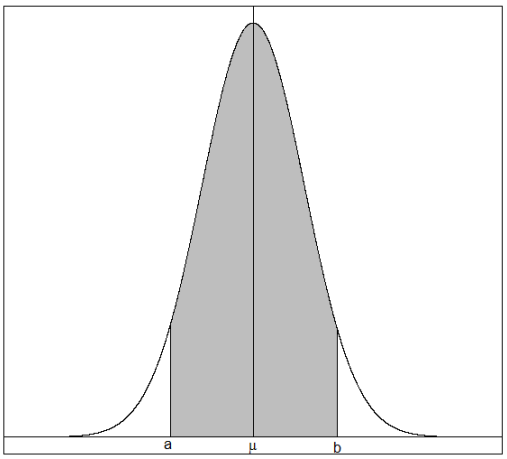


Figura 38 – Curva da Distribuição Normal

Esta probabilidade é obtida, calculando-se a área compreendida no intervalo dado, a qual pode ser obtida através da expressão a seguir,

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} dx$$

É fato que resolver a expressão anterior requer um certo conhecimento de cálculo diferencial e integral, assim como conhecimentos prévios, o que pode gerar grandes problemas. Assim, quando a média é igual a 0 e a variância é 1 diversos valores foram tabulados para os limites de integração. Essa tabela pode ser usada para calcular probabilidades de quaisquer valores da média e variância.

Assim, para "facilitar a nossa vida", se $X \sim N(\mu, \sigma^2)$, define-se que:

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

em que Z é conhecida como **distribuição Normal padronizada** tal que $Z \sim N(0, 1)$. Em problemas de pesquisa é muito importante a utilização desta distribuição. Muitas vezes, trabalhar com dados que possuem distribuição Normal, é o mesmo que viver em um mundo ideal.

A Figura 39 é a representação gráfica de $X \sim N(0, 1)$.

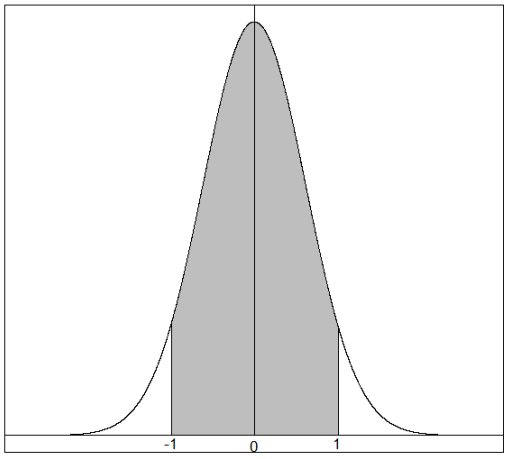
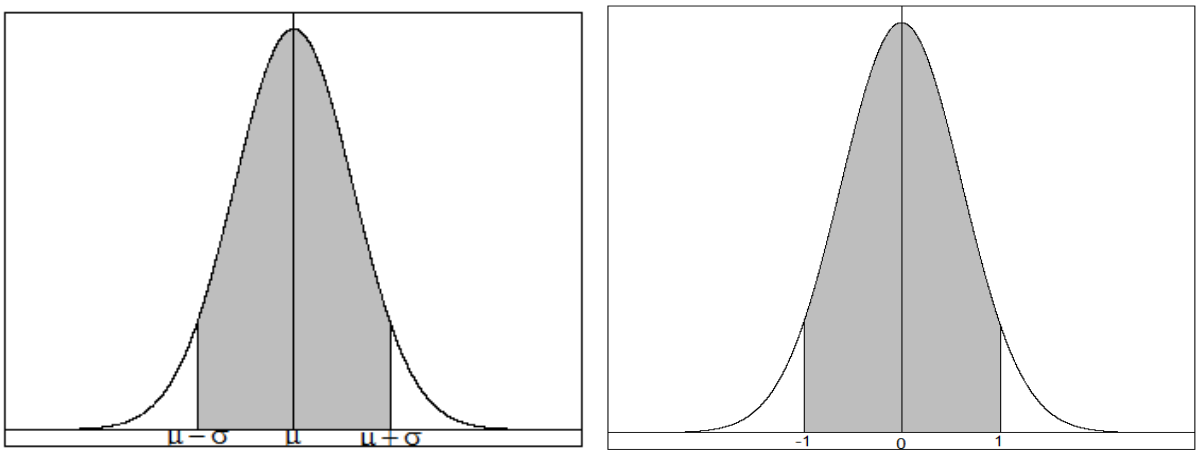


Figura 39 – Curva da Distribuição Normal Padronizada

As propriedades da distribuição normal são preservadas, ou seja esta curva continua sendo simétrica em relação a média. A variável Z indica quantos desvios padrões a variável X está afastada da média. Como as curvas são simétricas em relação às médias as Figuras 40a e 40b nos mostram como ficam os cálculos de qualquer probabilidade desejada.



(a) $P(\mu - \sigma \leq X \leq \mu) = P(\mu \leq X \leq \mu + \sigma)$ (b) $P(-1 \leq Z \leq 0) = P(0 \leq Z \leq 1)$

Figura 40 – Distribuição Normal Padronizada

Para adotar uma notação padrão, se desejarmos obter $P(Z \leq z) = \int_{-\infty}^{z} f(y; 0, 1)$, denotaremos por $\Phi(z)$.

Graficamente podemos ter a seguinte representação.

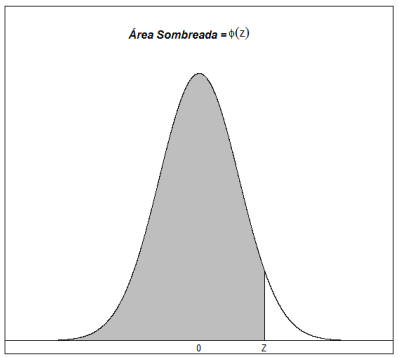


Figura 41 – Área normal padrão acumulada tabelada.

Se desejar calcular a probabilidade de um ponto aleatório do intervalo [a,b] ocorrer, tal que X segue distribuição normal padrão, temos que

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned} \tag{6.13}$$

Ainda

$$P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \quad e \quad P(X \geq a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{6.14}$$

Considerando que a distribuição normal é uma das mais importantes distribuições e pode ser utilizada em qualquer área do conhecimento, existem formulações teóricas que facilitam a análise e avaliação de conjuntos de dados. Ente as diversas formulações e regras existentes acerca dessa distribuição existe uma que por meio de evidências empíricas mostram que histogramas de dados reais frequentemente podem ser aproximados por curvas normais. Tal regra chamamos de *Regra Empírica*. Por meio dela constata-se que: se a distribuição de determinada variável resposta for (aproximadamente) normal, cerca de

- Cerca de 68% dos valores estão a 1 DP da média, isto é, $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68,26\%$;
- Cerca de 95,4% dos valores estão a 2 DP da média, isto é, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95,44\%$.
- Cerca de 99,73% dos valores estão a 3 DP da média, isto é, $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99,73\%$.

Exemplo 110 *O tempo que um motorista leva para reagir às luzes de freio em um veículo em desaceleração é crucial para evitar colisões traseiras. O artigo "Fast-Rise Brake Lamp as a Collision-Prevention Device" (Ergonomics, 1993, p.391-395) sugere que o tempo de reação de uma resposta no trânsito a um sinal de frenagem com luzes de freio convencionais pode ser modelado com uma distribuição normal de média 1,25 segundo e desvio padrão 0,46 segundos. Qual é a probabilidade de que o tempo de reação esteja entre 1,00 e 1,75 segundo?*

***Solução**: Desejamos obter P(1 < X < 1,75), sendo X a v.a tempo de reação a um sinal de frenagem, com $X \sim N(1,25; 0,2116)$. Utilizando a transformação da variável X para Z, temos:*

$$z_1 = \frac{1-1,25}{0,46} = -0,54 \quad e \quad z_2 = \frac{1,75-1,25}{0,46} = 1,09.$$

Nesse sentido, obter P(1 < X < 1,75) equivale a obter P(-0,54 < Z < 1,09). Graficamente, considerando a padronização da variável X, desejamos obter a área da região hachurada, ou a probabilidade mencionada.

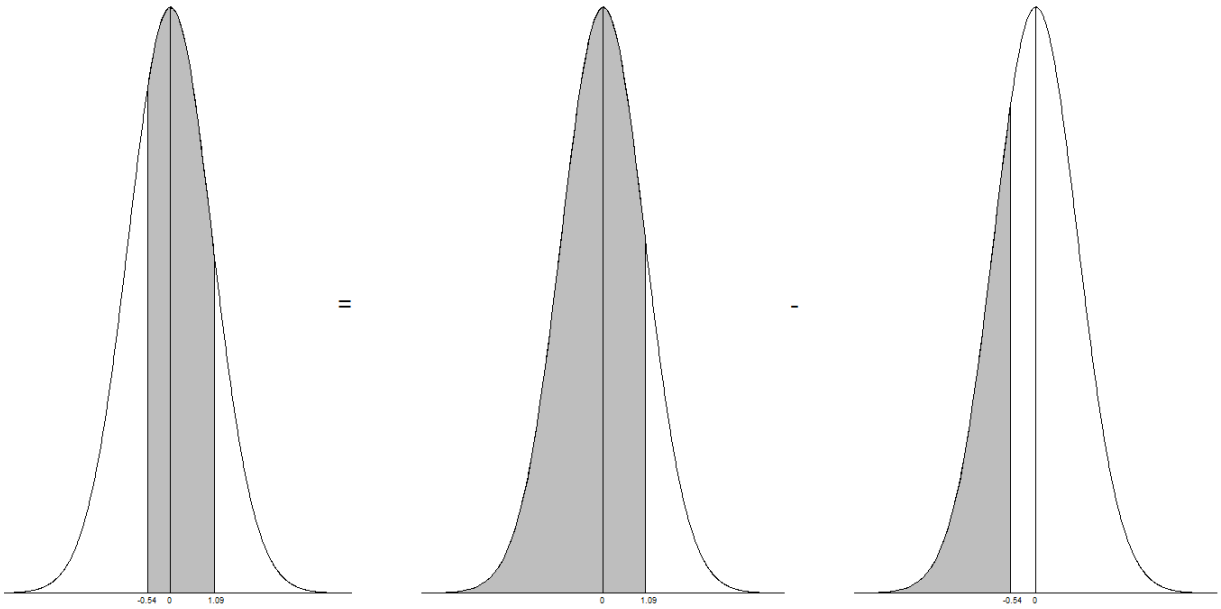


Figura 42 – Gráfico de distribuição normal do Exemplo 110

Assim,

$$P(-0,54 < Z < 1,09) = \Phi(1,09)^{1} - \Phi(-0,54) = 0,8621 - 0,2946 = 0,5675 \approx 56,75\%.$$

Portanto[2], a probabilidade de que o tempo de reação esteja entre 1,00 e 1,75 segundo é de aproximadamente 56,75%.

Para resolver no R, basta utilizar os comandos a seguir:

```
pnorm(1.75,1.25,0.46)  - pnorm(1,1.25,0.46) #pnorm(x,mu,sd)
0.5680717
```

Exemplo 111 *Suponha que o tempo de resposta na execução de um algoritmo é uma variável aleatória com distribuição normal de média 23 segundos e desvio padrão de 4 segundos. Calcule:*

a) *A probabilidade de o tempo de resposta ficar entre 20 e 30 segundos.*

***Solução**: Desejamos obter P(20 < X < 30), sendo X a v.a tempo de execução de um algoritmo, com $X \sim N(23; 16)$. Utilizando a transformação da variável X para Z, temos:*

$$z_1 = \frac{20 - 23}{4} = -0,75 \quad e \quad z_2 = \frac{30 - 23}{4} = 1,75.$$

[2] A diferença entre o valor calculado manualmente e o valor calculado computacionalmente ocorre devido aos critérios de arredondamento utilizados nos cálculos manuais.

Nesse sentido, obter P(20 < X < 30) equivale a obter P(-0,75 < Z < 1,75). Graficamente, considerando a padronização da variável X, desejamos obter a área da região hachurada, ou a probabilidade mencionada.

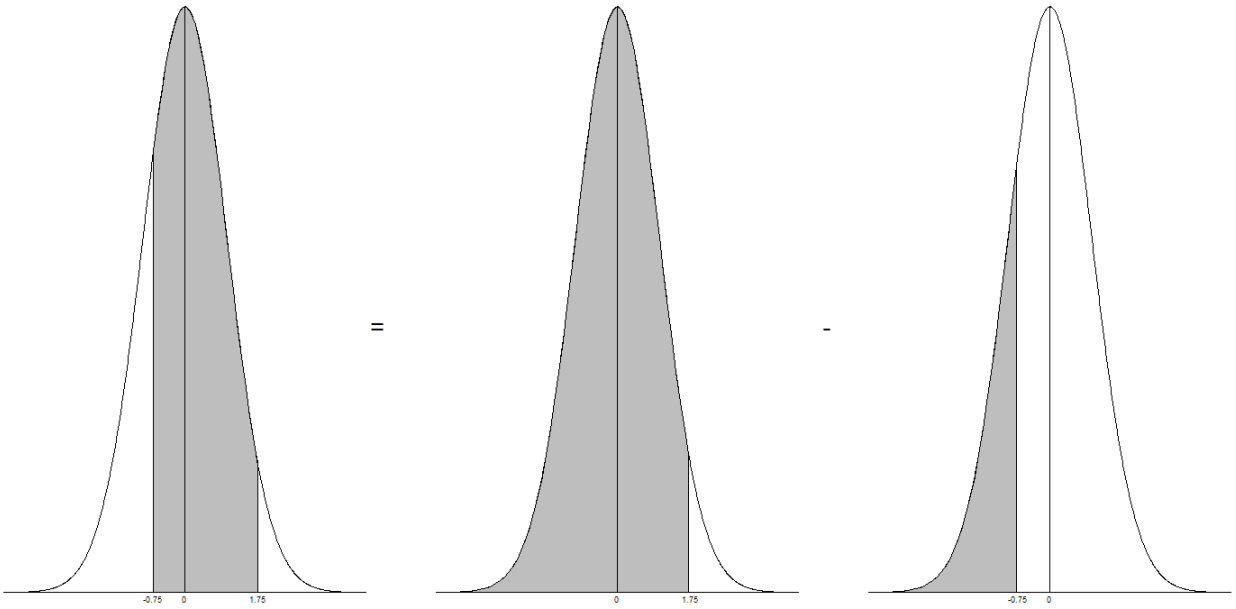


Figura 43 – Gráfico de distribuição normal do item a) do Exemplo 111

Assim,

$$\begin{aligned} P(-0,75 < Z < 1,75) &= \Phi(1,75) - [1 - \Phi(0,75)] \\ &= 0,9599 - [1 - 0,7734] \\ &= 0,9599 - 0,2266 \\ &= 0,7333 \approx 73,33\%. \end{aligned}$$

Portanto, a probabilidade de que o tempo de execução do algoritmo avaliado esteja entre 20 e 30 segundos é aproximadamente 73,33%.

Para resolver no R, basta utilizar os comandos a seguir:

```
pnorm(30,23,4)   - pnorm(20,23,4) #pnorm(x,mu,sd)
0.7333135
```

b) *A probabilidade de o tempo de resposta ser menor do que 25 segundos.* ***Solução****: Desejamos obter P(X < 25), sendo X a v.a tempo de execução de um algoritmo, com* $X \sim N(23; 16)$*. Utilizando a transformação da variável X para Z, temos:*

$$z_1 = \frac{25 - 23}{4} = 0,5$$

Nesse sentido, obter P(X < 25) equivale a obter P(Z < 0,5). Graficamente, considerando a padronização da variável X, desejamos obter a área da região hachurada, ou a probabilidade mencionada.

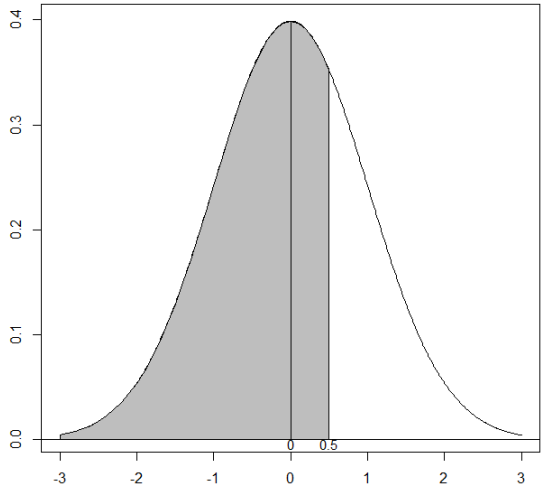


Figura 44 – Gráfico de distribuição normal do item b) do Exemplo 111

Assim,

$$P(Z < 0.5) = \Phi(0,5) = 0,6915 \approx 69,15\%.$$

Portanto, a probabilidade de que o tempo de execução do algoritmo avaliado seja inferior a 25 segundos é aproximadamente 69,15%.

Para resolver no R, basta utilizar o comando a seguir.

```
pnorm(25,23,4) #pnorm(x,mu,sd)
0.6914625
```

Exemplo 112 *A distribuição de pesos para a população masculina americana é aproximadamente normal com média 78,11 Kg e desvio padrão 13,52 Kg. Pede-se a probabilidade de que um americano selecionado ao acaso pese:*

a) *Menos do que 59 kg;*

b) *Mais do que 95,5 kg;*

c) *Entre 60 e 100 kg.*

Exemplo 113 *A resistência à compressão de amostras de cimento pode ser modelada por uma distribuição normal, com uma média de 6000* Kg/cm^2 *e um desvio padrão de 100* Kg/cm^2*.*

a) *Qual é a probabilidade de a resistência da amostra ser menor do que 6250 Kg/cm2?*

b) *Qual é a probabilidade de a resistência da amostra estar entre 5800 e 5900 Kg/cm2?*

c) *Que resistência é excedida por 95% das amostras?*

Exemplo 114 *O comprimento de uma capa de plástico, moldada por injeção, que reveste uma fita magnética é normalmente distribuído, com um comprimento médio de 90,2 milímetros e um desvio padrão de 0,1 milímetro.*

a) *Qual é a probabilidade de uma peça ser maior que 90,3 milímetros ou menor que 89,7 milímetros?*

b) *Qual deveria ser a média do processo para se usar de modo a se obter o maior número de peças entre 89,7 e 90,3 milímetros?*

Exemplo 115 *A largura do cabo de uma ferramenta usada para a fabricação de semicondutores é suposta estar distribuição normalmente, com uma média de 0,5 micrômetro e um desvio padrão de 0,05 micrômetro.*

a) *Qual é a probabilidade de a largura do cabo ser maior que 0,62 micrômetro?*

b) *Qual é a probabilidade de a largura do cabo estar entre 0,47 e 0,63 micrômetro?*

Exemplo 116 *Em um aviário foram instaladas 1000 lâmpadas novas. Sabe-se que a duração média delas é de 800 horas com desvio padrão de 100 horas, com distribuição normal. Determinar a quantidade de lâmpadas que durarão:*

a) *menos de 500 horas;*

b) *mais de 700 horas;*

c) *entre 516 e 684 horas.*

Exemplo 117 *Foi feito um estudo sobre a altura dos alunos de uma faculdade, observando-se que ela se distribuída normalmente com média 1,72 m e desvio padrão de 5 cm. Qual a porcentagem dos alunos com altura:*

a) *entre 1,57 m e 1,87 m?*

b) *acima de 1,90 m?*

Exemplo 118 *Suponha que a força que age sobre uma coluna que ajuda a suportar um edifício tenha distribuição normal com média 15,0 kips e desvio padrão 1,25 kips. Qual a probabilidade de a força ser no máximo 18 kips?*

6.7.2.4 Distribuição Normal como limite de outras distribuições

Em determinadas situações práticas em que a variável aleatória de interesse é discreta, é possível fazer a aproximação dela por uma distribuição normal. Esse é o caso das distribuições binomial e Poisson quando n e λ são grandes respectivamente (BARBETTA; REIS; BORNIA, 2010).

6.7.2.5 Relação entre distribuição Binomial e Normal

Ao considerarmos uma v.a X que segue distribuição binomial, para n muito grande, é possível que existam problemas para desenvolver os cálculos necessários, uma vez que para obter a probabilidade desejada depende-se do cálculo dos coeficientes binomiais. Nesse contexto é possível fazer uma aproximação da distribuição binomial à normal. Observe a figura a seguir:

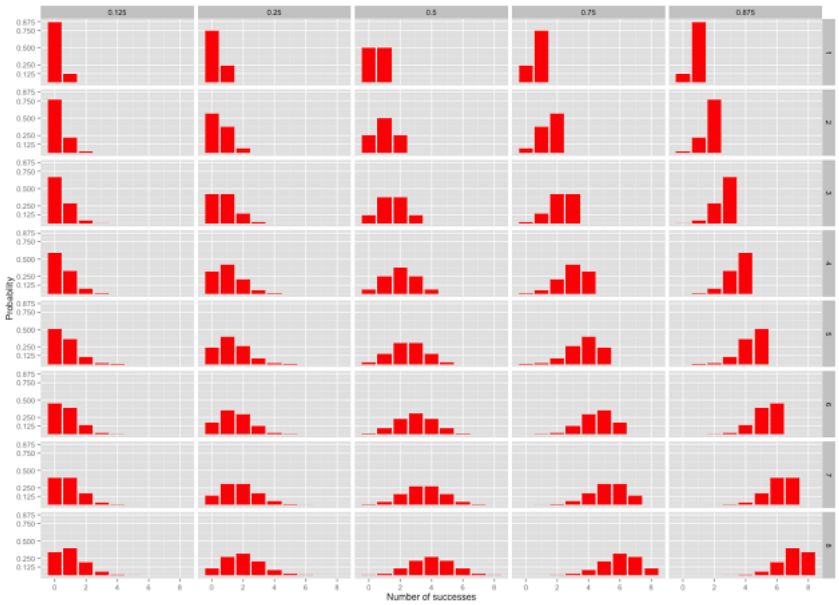


Figura 45 – Distribuição Binomial para diferentes tamanhos amostrais e probabilidade de sucesso.

De maneira geral, as condições para fazer uma aproximação da distribuição binomial pra a normal são:

- n **grande**;
- p **não** muito próximo de 0 ou de 1.

Segundo Barbetta, Reis e Bornia (2010), uma regra prática sugerida por vários autores, é razoável que as duas seguintes condições sejam satisfeitas:

$$np \geq 5 \quad e \quad n(1-p) \geq 5 \tag{6.15}$$

Os parâmetros da distribuição normal devem identificar-se ao valor esperado e ao desvio padrão da distribuição binomial, isto é:

a) $\mu = np$;

b) $\sigma = \sqrt{np(1-p)}$.

Considerando que essa situação é uma aproximação de uma variável aleatória discreta por uma contínua, em determinadas situações, para uma melhor aproximação, sem que exista perda de informação, é necessário utilizar um recurso para corrigir essa situação, chamada de *correção de continuidade*. Nesse caso, se desejarmos obter $P(a \leq X \leq b)$, uma aproximação adequada é dada por

$$P(a \leq X \leq b) \approx P(a - 0,5 \leq Y \leq b + 0,5) \tag{6.16}$$

Exemplo 119 *Historicamente, 10% dos pisos cerâmicos, que saem de uma linha de produção, têm algum defeito leve. Se a produção diária é de 1000 unidades, qual a probabilidade de ocorrer mais de 120 itens defeituosos?*

6.7.2.6 Relação entre distribuição Poisson e Normal

Variáveis aleatórias as quais podem ser modeladas por meio da distribuição de Poisson são comum na área de engenharia. Todavia, se o valor esperado da v.a for suficientemente grande é possível aproximá-la pela distribuição normal, visto que existem propriedades que garantem a "eficiência"desta distribuição.

Formalmente, para fazer essa aproximação é necessário que se $X \sim Pois(\lambda)$, é necessário que λ seja suficientemente grande e assim os parâmetros da distribuição normal devem identificar-se ao valor esperado e ao desvio padrão da distribuição de Poisson, isto é:

a) $\mu = \lambda$;

b) $\sigma = \sqrt{\lambda}$.

Graficamente é possível identificar a aproximação da distribuição de Poisson a medida que o valor da média aumenta.

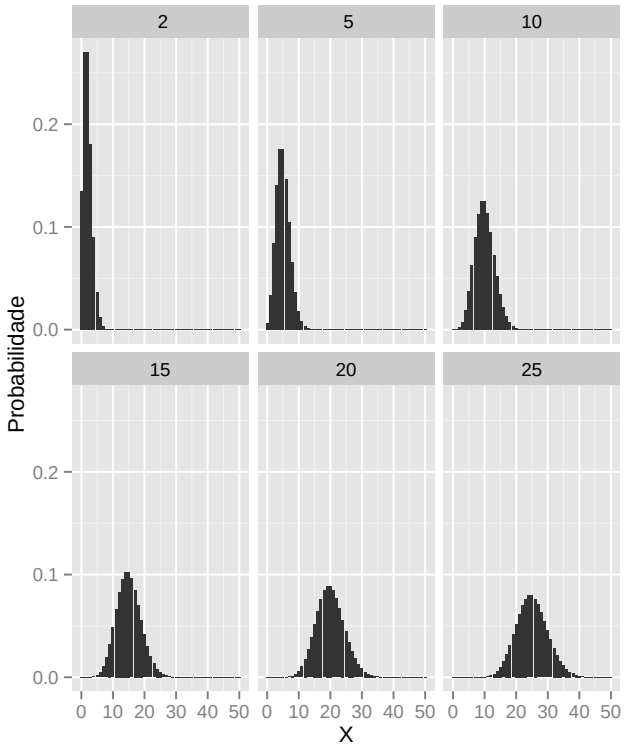


Figura 46 – Distribuição de Poisson para diferentes médias (λ = 2, 5, 10, 15, 20, 25).

Exemplo 120 *Uma empresa de auxílio à lista telefônica recebe, em média, sete solicitações por minuto, segundo uma distribuição de Poisson. Qual é a probabilidade de ocorrer mais de 80 solicitações nos próximos 10 minutos?*

6.8 Exercícios de Aplicação

1. Uma viga de concreto pode apresentar falha por cisalhamento (C) ou flexão (F). Suponha que três vigas com defeito sejam selecionadas aleatoriamente e o tipo de falha seja determinado para cada uma delas. Seja X = número de vigas entre as três selecionadas que falharam por cisalhamento. Construir a função distribuição de probabilidade para variável aleatória X.

2. Em uma sala temos 4 rapazes e 3 moças. São escolhidos aleatoriamente 2 pessoas. Construir a tabela para variável aleatória X: número de moças.

   a) Construir a tabela de distribuição de probabilidade da variável x.

   b) Determinar a probabilidade do grupo escolhido ter duas moças.

3. Lançar um par de dados equilibrados. A V.a. X corresponde ao Máximo do par (x,y).

   a) Determinar os valores da V.a.X

   b) Construir a distribuição de probabilidade e graficar.

   c) Calcular: P(X = 5); P(X = 4) e P(X = 2)

4. Suponha que o número de certo tipo de planta encontrada em uma região retangular (denominada *quadrat* pelos ecologistas) em uma determinada área geográfica seja uma v.a X com função distribuição de probabilidade dada por

$$\text{f(x)} = \begin{cases} \dfrac{c}{x^3} & \text{se} \qquad x = 1, 2, 3, \ldots \\ 0 & \text{caso contrário} \end{cases}$$

   E(X) é finito? Justifique sua resposta.

5. Sejam X = Número de defeitos graves em uma obra selecionada aleatoriamente em uma cidade. Calcule os dados a seguir:

Tabela 39 – Dados do Exercício 5

X	0	1	2	3	4
p(X)	0,08	0,15	0,45	0,27	0,05

   a) E(X);

   b) V(X).

6. Obtenha o valor (ou valores) de c, para que a expressão abaixo seja função densidade de probabilidade seguir:

$$\text{f(x)} = \begin{cases} 0, & \text{para} \quad x \leq -1; \\ -cx & \text{para} \quad -1 < x \leq 0; \\ ce^{-6x} & \text{para} \quad x > 0. \end{cases}$$

7. A dureza H de uma peça de aço pode ser pensada como uma variável aleatória com distribuição de probabilidade contínua chamada de uniforme no intervalo [50,70] da escala de Rockwel. Calcular a probabilidade de que uma peça tenha dureza entre 55 e 60, considerando a fdp dada por:

$$\text{f(x)} = \begin{cases} \dfrac{1}{20}, & \text{para} \quad 50 \leq x \leq 70; \\ 0 & \quad caso \quad contrário; \end{cases}$$

8. Considerando a distribuição uniforme, isto é, $X \sim U_c[a, b]$, mostre que $E(X) = \frac{b+a}{2}$ e $Var(X) = \frac{(b-a)^2}{12}$.

9. Seja X uma v.a contínua com fda dada por

$$\text{F(x)} = \begin{cases} 0, & \text{se} \quad x \leq 0; \\ \dfrac{x}{4}\left[1 + \ln(\dfrac{4}{x})\right] & \text{se} \quad 0 < x \leq 4; \\ 1 & \text{se} \quad x > 4; \end{cases}$$

   [Esse tipo de fda é sugerido no artigo *"Variability in Measured Bedload-Transport Rates"(Water Resources Buli.*, 1985, p. 39-48) como modelo para determinada variável hidrológica]. Determine:

   a) $P(X \leq 1)$;

   b) $P(1 \leq X \leq 3)$.

   c) A função densidade de probabilidade de X?

10. Determine as constantes *a* e *b*, para que a função *G(x)* seja a função de distribuição de alguma variável aleatória contínua.

$$\text{G(x)} = \begin{cases} a - 2b, & \text{se} \quad x < 0; \\ ax & \text{se} \quad 0 \leq x < 1; \\ a + b(x-1) & \text{se} \quad 1 \leq x < 2; \\ 1 & \text{se} \quad x \geq 2. \end{cases}$$

11. Uma empresa de cristais finos sabe por experiência que 10% de suas taças possuem defeitos cosméticos e devem ser classificadas como de "segunda linha".

    a) Entre seis taças selecionadas aleatoriamente, qual a probabilidade de uma ser de segunda linha?

b) Entre seis taças selecionadas aleatoriamente, qual é a probabilidade de no mínimo duas serem de segunda linha?

c) Se as taças forem examinadas uma a uma, qual será a probabilidade de no máximo cinco terem de ser selecionadas para encontrar quatro que não sejam de segunda linha?

12. Um lote de aparelhos de TV é recebido por uma firma. Vinte aparelhos são inspecionados. O lote é rejeitado se pelo menos 4 forem defeituosos. Sabendo-se que 1% dos aparelhos é defeituoso, determinar a probabilidade de a firma rejeitar todo o lote?

13. Sabe-se que 20% dos animais submetidos a um certo tratamento não sobrevivem. Se esse tratamento foi aplicado a 20 animais e se X é o número de não-sobreviventes:

a) Qual a distribuição de X?

b) Calcular E(X) e Var(X).

c) Calcular $P(2 < x \leq 4)$.

d) Calcular $P(X \geq 2)$.

14. Uma experiência mostra que de cada 400 lâmpadas, 2 se queimam ao serem ligadas. Qual a probabilidade de que numa instalação de:

a) 600 lâmpadas, no mínimo 3 queimem?

b) 900 lâmpadas, exatamente 8 se queimem?

15. Numa linha adutora de água, de 60 km de extensão, ocorrem 30 vazamentos no período de um mês. Qual a probabilidade de ocorrer, durante o mês, pelo menos 3 vazamentos num certo setor de 3 km de extensão?

16. Uma firma recebe 720 mensagens em seu fax em 8 horas de funcionamento. Qual a probabilidade de que:

a) em 6 minutos receba pelo menos 4 mensagens?

b) em 4 minutos não receba nenhuma mensagem?

17. Numa urna há 40 bolas brancas e 60 pretas. Retiram-se 20 bolas. Qual a probabilidade de que ocorram no mínimo 2 bolas brancas, considerando as extrações:

a) sem reposição?

b) com reposição?

18. Uma fábrica de motores para máquinas de lavar roupas separa sua linha de produção diária de 350 peças em uma amostra de 30 itens para inspeção. O número de peças defeituosas é de 14 por dia. Qual a probabilidade de que a amostra contenha pelo menos 3 motores defeituosos?

19. Uma urna tem 10 bolas brancas e 40 pretas.

    a) Qual a probabilidade de que a 6ª bola retirada com reposição seja a 1ª branca?

    b) Qual a probabilidade de que de 16 bolas retiradas sem reposição ocorram 3 brancas?

    c) Qual a probabilidade de que em 30 bolas retiradas com reposição ocorram no máximo 2 brancas?

    d) Se o número de bolas na urna fosse 50 brancas e 950 pretas, qual a probabilidade de que, retirando-se 200 bolas, com reposição, ocorressem pelo menos 3 brancas?

20. Um determinado artigo é vendido em caixa a preço de R$ 20,00 cada. É característica de produção que 20% destes artigos sejam defeituosos. Um comprador fez a seguinte proposta: de cada caixa escolhe 25 artigos, ao acaso, e paga por caixa:

    - R$ 25,00 se nenhum artigo, dos selecionados, for defeituoso;
    - R$ 17,00 se um ou dois artigos forem defeituosos;
    - R$ 10,00 se três ou mais forem defeituosos. O que é melhor para o fabricante: manter o seu preço de R$ 20,00 por caixa ou aceitar a proposta do consumidor? **Dica: Obter o valor esperado por caixa.**

21. Uma variável aleatória X tem a função de distribuição dada por

$$\text{F(X)} = \begin{cases} 0, & \text{se} \quad x \leq 0; \\ x^5, & \text{se} \quad 0 < x < 1; \\ 1, & \text{se} \quad x \geq 1 \end{cases}$$

    Calcular E(X) e Var(X).

22. O setor de manutenção de uma empresa fez um levantamento das falhas de um importante equipamento, constatando que há, em média, 0,75 falha por ano e que o tempo entre falhas segue uma distribuição exponencial. Qual é a probabilidade de o equipamento não falhar no próximo ano?

23. A vida útil de certo componente eletrônico é, em média, 10.000 horas e apresenta distribuição exponencial. Qual é a percentagem esperada de componentes que apresentarão falhas em menos de 10.000 horas?

24. A vida útil de certo componente eletrônico é, em média, 10.000 horas e apresenta distribuição exponencial. Após quantas horas se espera que 25% dos componentes tenham falhado?

25. Em um laticínio, a temperatura do pasteurizador deve ser de 75° C. Se a temperatura ficar inferior a 70° C, o leite poderá ficar com bactérias maléficas ao organismo humano. Observações do processo mostram que valores da temperatura seguem uma distribuição normal com média $75,4°C$ e desvio padrão $2,2°C$. Nesse contexto, pede-se:

    a) Qual é a probabilidade da temperatura ficar inferior a $70°C$?

    b) Qual é a probabilidade de que, em 1.000 utilizações do pasteurizador, em mais do que cinco vezes a temperatura não atinja $70°C$?

26. O tempo para que um sistema computacional execute determinada tarefa é uma variável aleatória com distribuição normal, com média 320 segundos e desvio padrão de 7 segundos. Nesse contexto, pede-se:

    a) Qual é a probabilidade de a tarefa ser executada entre 310 e 330 segundos?

    b) Se a tarefa é colocada para execução 200 vezes. Qual é a probabilidade de ela demorar mais do que 325 segundos em pelo menos 50 vezes?

27. O padrão de qualidade recomenda que os pontos impressos por uma impressora estejam entre 3,7 e 4,3 mm. Uma impressora imprime pontos, cujo diâmetro médio é igual a 4 mm e o desvio padrão é 0,19 mm. Suponha que o diâmetro dos pontos tenha distribuição normal. Nesse contexto, pede-se:

    a) Qual é a probabilidade do diâmetro de um ponto dessa impressora estar dentro do padrão?

    b) Qual deveria ser o desvio padrão para que a probabilidade do item (a) atingisse 95%?

28. Certo tipo de cimento tem resistência à compressão com média de 5.800 kg/cm2, e desvio padrão de 180 kg/cm2, segundo uma distribuição normal. Dada uma amostra desse cimento, calcule as seguintes probabilidades:

    a) resistência inferior a 5.600 kg/cm2.

    b) resistência entre 5.600 kg/cm2 e 5.950 kg/cm2.

    c) resistência superior a 6.000 kg/cm2, sabendo-se que ele já resistiu a 5.600 kg/cm2.

    d) se quer a garantia de que haja 95% de probabilidade de o cimento resistir a determinada pressão, qual deve ser o valor máximo dessa pressão?

29. O número de distribuições de probabilidade cresce constantemente, devido a necessidade de se encontrar soluções ótimas para problemas cotidianos. Uma das distribuições mais utilizadas em análise de sobrevivência é a distribuição Weibull. A

farru1ia de distribuições de Weibull foi apresentada pelo físico sueco Waloddi Weibull, em 1939. Seu artigo de 1951 ***"A Statistical Distribution Function of Wide Applicability"*** (J. Applied Mechanics, vol. 18, p. 293-297) apresenta diversas aplicações. Sua fdp é dada por:

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, & \text{se} \quad x \geq 0; \\ 0 & \text{se} \quad x < 0; \end{cases}$$

A função distribuição acumulada de uma v.a de Weibull de parâmetros α e β é

$$f(x;\alpha,\beta) = \begin{cases} 1 - e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, & \text{se} \quad x \geq 0; \\ 0 & \text{se} \quad x < 0; \end{cases}$$

Considere o problema a seguir:

Os autores do artigo *"A Probabilistic Insulation Life Model for Combined Thermal-Electrical Stresses"* (IEEE Trans. on Elect. Insulation, 1985, p. 519-522) declararam que "a distribuição de Weibull é largamente usada em problemas estatísticos relativos ao desgaste de materiais isolantes sólidos sujeitos ao desgaste e à tensão". Eles propõem o uso da distribuição como um modelo para tempo (em horas) de falha de amostras de isolantes sólidos sujeitos a voltagem de AC. Os valores dos parâmetros dependem da voltagem e da temperatura. Suponha que α = 2,5 e β = 200 (valores sugeridos pelos dados no artigo).

a) Qual é a probabilidade de o tempo de vida de um espécime ser no máximo 250? Inferior a 250? Maior de 300?

b) Qual é a probabilidade de o tempo de vida de uma amostra estar entre 100 e 250?

6.8.1 Gabarito

5. E(x) = 2,08 V(x) = 0,94

6. c = 3/2

7. 0,25.

9.

- **a)** 0,597;
- **b)** 0,369;
- **c)** f(x)= 0,3466 - 0,25 ln(x) se $0 < x < 4$.

10. $a = \frac{2}{3}$, $b = \frac{1}{3}$.

11.

- **a)** 0,354;
- **b)** 0,115;
- **c)** 0,918.

12. 0,00004.

13.

- **b)** 4 e 3,2;
- **c)** 0,4235;
- **d)** 0,93082.

14.

- **a)** 0,57681;
- **b)** 0,046330.

15. 0,191154.

16.

- **a)** 0,978774;
- **b)** 0,002479.

17.

- **a)** 0,999839;
- **b)** 0,99948.

18. 0,108453.

19.

a) 0,065536;

b) 0,293273;

c) 0,04419;

d) 0,997231.

21. $E(X) = \frac{5}{6}$ e $Var(X) = \frac{5}{252}$.

22. 0,4724.

23. 0,6321.

24. 2.877 horas.

25.

a) 0,0071;

b) 0,73.

26.

a) 0,8472;

b) 0,3859.

27.

a) 0,8858;

b) 0,153 mm.

28.

a) 0,1335;

b) 0,6632;

c) 0,1541;

d) 5.504 kg/cm^2.

29.

a) 0,826; 0,826; 0,0636;

b) 0,664.

Capítulo 7

Inferência Estatística

7.1 Introdução

Inferência Estatística ou Estatística indutiva é a parte da estatística que utiliza métodos científicos para fazer afirmações e tirar conclusões sobre características ou parâmetros de uma população, baseando-se em resultados de uma amostra . O próprio termo "indutiva"decorre da existência de um processo de indução, isto é, um processo de raciocínio em que, partindo-se do conhecimento de uma parte, procura-se tirar conclusões sobre a realidade no todo. Essas são decisões baseadas em procedimentos amostrais.

Nosso objetivo é procurar a conceituação formal desses princípios intuitivos do dia-a-dia para que possam ser utilizados cientificamente em situações mais complexas.

É fácil perceber que um processo de indução (em estatística) não pode ser exato. Ao induzir, portanto, estamos sempre sujeitos a erro. A Inferência Estatística, entretanto, irá nos dizer até que ponto poderemos estar errando em nossas induções, e com que probabilidade. Esse fato é fundamental para que uma indução (ou inferência) possa ser considerada estatística, e faz parte dos objetivos da Inferência Estatística.

Em suma, a Inferência Estatística busca obter resultados sobre as populações a partir das amostras, dizendo também, qual a precisão desses resultados e com que probabilidade se pode confiar nas conclusões obtidas. Evidentemente, a forma como as induções serão realizadas irá depender de cada tipo de problema, conforme será estudado posteriormente.

Neste capítulo trabalharemos com os principais conceitos relacionados a inferência, em específicos conceitos de Estimação (Pontual e Intervalos de Confiança) e Testes de

Hipóteses.

7.2 Parâmetros e Estatística

Em geral, ao conduzirmos uma pesquisa podem existir uma ou mais variáveis respostas associada aos elementos da população ou da amostra. Considerando isso, surgem os conceitos a seguir.

Definição 23 ***Parâmetro****: Medida descritiva obtida por meio dos valores da população. Genericamente, um parâmetro qualquer pode ser representado pela letra grega* θ*. Os parâmetros populacionais mais mencionados em qualquer literatura estatística são a média* (μ) *e variância* (σ^2)*.*

Definição 24 ***Estatística*** *ou* ***Estimador****: Medida descritiva obtida por meio dos valores da amostra. Muitas vezes, podemos dizer que o resultado numérico de uma estatística é uma estimativa de um parâmetro populacional. Como existe incerteza associada ao valor de uma estatística, ela é uma variável aleatória. A média amostral* $(\overline{x})$ *e a variância amostral* (s^2) *são exemplos de estatística.*

Os parâmetros Proporção, Média e variância, seus respectivos estimadores podem ser observados na Tabela 40.

Tabela 40 – Parâmetros e Estimadores

Estimadores	**Parâmetros**	**Estatísticas**
Proporção	$p = \frac{n_i}{N}$	$\hat{p} = \frac{n_i}{n}$
Média	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variância	$(\sigma)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$

sendo n_i o número de elementos com a característica de análise.

Exemplo 121 *Os efeitos de descargas parciais na degradação de materiais de cavidades isolante têm importantes implicações na vida útil de componentes de alta voltagem. Consideremos a seguinte amostra de n = 25 larguras de pulso de descargas lentas em uma cavidade cilíndrica de polietileno. (Esses dados são consistentes com um histograma de 250 observações no artigo "Assessment of Dielectric Degradation by Ultrawide-band PD*

5,3	8,2	13,8	74,1	85,3	88,0	90,2	91,5	92,4	92,9	93,6	94,3	94,8
94,9	95,5	95,8	95,9	96,6	96,7	98,1	99,0	101,4	103,7	106,0	113,5	

Detection,"IEEE Trans. on Dielectrics and Elec. Insul., 1995, p. 744-760.) Os dados amostrais são:

Solução: *Obtendo os valores desejados, temos:*

$$\overline{X} = \frac{5,3+8,2+13,8+74,1+\ldots+101,4+103,7+106,0+113,5}{25} = 99,76$$

$$\sigma^2 = \frac{(5,3-99,76)^2+(8,2-99,76)^2+\ldots+(106-99,76)^2+(113,5-99,76)^2}{24} = 30,9972$$

Para resolver no R, basta utilizar os comandos a seguir.

```
x <- c(5.3, 8.2, 13.8, 74.1, 85.3, 88, 90.2, 91.5, 92.4,
   92.9, 93.6, 94.3, 94.8, 94.9, 95.5, 95.8, 95.9, 96.6,
   96.7, 98.1, 99.0, 101.4, 103.7, 106.0, 113.5)
mean(x)
var(x)
```

7.3 Distribuições Amostrais

Considerando que uma estatística é uma variável aleatória, sua distribuição de probabilidade é chamada de **distribuição amostral**. Na sequência, estudaremos as distribuições amostrais da Média e da proporção.

7.3.1 Distribuição Amostral da Média

De acordo com Devore (2006), a importância da média amostral surge de seu uso para tirar conclusões sobre a média da população. Nesse caso, alguns procedimentos inferenciais baseiam-se nas propriedades da distribuição de $\overline{X}$. Considere a situação abaixo.

Exemplo 122 *Em um estudo sobre o consumo de combustível, definiu-se uma população composta por quatro ônibus de uma pequena companhia de transporte urbano. Os consumos dos ônibus (km/l), em condições padrões de teste eram 3,9, 3,8, 4,0 e 4,1.*

a) *Obter todas as amostras de tamanho dois com reposição;*

b) *Obter a média de cada uma das amostras.*

c) *Obter a variância das médias.*

d) *Construir a função distribuição de probabilidade para as médias e graficar.*

Considerando as possíveis amostras de tamanho n, de uma população de tamanho N, isto é $X = \{X_1, X_2, \ldots, X_n\}$, a distribuição amostral da média apresenta as seguintes propriedades:

Propriedade 6 *O valor esperado da média amostral é igual a média da população, isto é,*

$$E(\overline{X}) = E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu. \tag{7.1}$$

Propriedade 7 *A variância da média amostral é igual a variância populacional proporcional ao tamanho amostral (quando a amostragem for com reposição), isto é,*

$$Var(\overline{X}) = V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\sigma^2}{n} = \frac{\sigma^2}{n}. \tag{7.2}$$

Ao extrairmos a raiz quadrada da expressão 7.4 obtemos o que é chamado de **erro-padrão da estimativa da média**. Ele serve para expressar a precisão da estimativa da média amostral. Observe que a medida que n cresce a estimativa da média torna-se mais precisa.

7.3.2 Distribuição Amostral da Proporção

Consideremos a distribuição amostral da *proporção p* de sucessos em uma amostra. Seja $X = \{X_1, X_2, \ldots, X_n\}$, uma amostra aleatória de tamanho n, a distribuição amostral da proporção de p sucessos apresenta as seguintes propriedades:

Propriedade 8 *O valor esperado da proporção amostral é igual a proporção populacional, isto é,*

$$E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}np = p \tag{7.3}$$

Propriedade 9 *A variância da proporção amostral (quando a amostragem for com reposição) é dada por*

$$Var(\hat{p}) = \frac{p(1-p)}{n}. \tag{7.4}$$

se n é suficientemente grande, isto é, se $np \geq 5$ *e* $n(1-p) \geq 5$.

7.3.3 Teorema Limite Central - Algumas considerações

O teorema do limite central é um dos resultados mais importantes na área de Probabilidade e Estatística. A palavra central que aparece no nome deste teorema que se baseia em um processo limite foi dado pelo matemático George Polya. O adjetivo central se refere ao teorema pela sua importância e não ao limite calculado.

Estabelece a distribuição normal como base para a Estatística Inferencial. Este teorema afirma que a soma de muitas variáveis aleatórias independentes e com a mesma distribuição de probabilidade converge em distribuição para uma variável aleatória com distribuição normal.

Assim se tivermos uma amostra suficientemente grande, de tamanho n, a distribuição de probabilidade da média amostral pode ser aproximada por uma distribuição normal com média igual à da população e com variância igual a variância da população divida por n. O mesmo pode ser avaliado para proporção amostral, sendo que nesses casos temos que as distribuições são assintoticamente normais, isto é,

$$Z_1 = \frac{\overline{X} - \mu}{\sigma_x} \sim N(0,1) \quad e \quad Z_2 = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \tag{7.5}$$

7.4 Estimação

Em uma pesquisa, muitas vezes somos obrigados a analisar partes da população, chamadas de amostra, e a partir dos dados obtidos, **descreve-se** e **estima-se** algumas quantidades (*parâmetros*). É possível obtermos dois tipos de estimativas para um parâmetro, são elas:

- Estimativa pontual, baseada em um único valor;
- Estimativa intervalar (intervalos de confiança), baseada em um conjunto de valores.

Nas próximas seções estaremos interessados nos tipos de estimativas bem como em compreender suas características.

Definição 25 *Uma estimativa pontual de um parâmetro θ é um único número que pode ser considerado um valor sensato para θ. Obtém-se uma estimativa pontual selecionando uma estatística adequada e calculando seu valor pelos dados da amostra. A estatística selecionada é chamada estimador pontual de θ.*

Os resultados obtidos no exemplo anterior são **estimativas pontuais** para média e variância populacional.

Para qualquer situação é possível encontrar um estimador que coincide com o valor populacional do parâmetro desejado. Todavia, como os estimadores é uma função dos elementos da amostra, o mesmo é uma variável aleatória e assim, em determinadas situações a estimativa obtida se afastará do verdadeiro valor do parâmetro, isto é, existirá uma taxa de erro envolvida. Um bom estimador deve ter algumas propriedades garantidas, entre as quais destacamos a propriedade de não tendenciosidade ou ausência de vício.

Definição 26 *Dizemos que um estimador $\hat{\theta}$ é não viciado, ou não tendencioso, se $E(\hat{\theta}) = \theta$.*

Quando n tende ao infinito e a condição anterior ainda é satisfeita, dizemos que o estimador é assintoticamente não tendencioso.

Observação: Na propriedade 6 da seção de distribuição amostral dos estimadores, observamos que a condição de não tendenciosidade é satisfeita para o estimador da média populacional e o mesmo pode ser verificado para variância populacional, o qual é descrito com detalhes em Morettin (2010) página 222.

7.4.1 Estimação Pontual

Segundo Devore (2006) a inferência estatística é quase sempre direcionada à obtenção de algum tipo de conclusão sobre um ou mais parâmetros (características da população). O processo requer que o pesquisador obtenha dados de amostras de cada população em estudo. As conclusões baseiam-se, então, nos valores calculados das várias quantidades da amostra.

Para compreendermos a ideia de estimação pontual, considere o exemplo a seguir:

Exemplo 123 *No futuro próximo haverá um interesse crescente no desenvolvimento de ligas à base de Mg de baixo custo para diversos processos de fundição. Por isso, é importante ter maneiras práticas de determinar as várias propriedades mecânicas de tais ligas. O artigo "On the Development of a New Approach for the Determination of Yield Strength in Mg-based Alloys" (Light Metal Age, out. 1998, p. 50-53) propôs um método ultra-sônico para esse fim. Considere a seguinte amostra de observações no módulo de elasticidade (GPa) de espécimes da liga AZ91D de um processo de fundição:*

$$X : 44,2; 43,9; 44,7; 44,2; 44,0; 43,8; 44,6; 43,1 \tag{7.6}$$

Considerando que esses dados foram obtidos por meio de uma amostragem aleatória da população de módulo elástico sob tais circunstâncias. Devemos obter uma estimativa para média e variância da amostra.

$$\begin{aligned} \overline{x} &= \frac{44,2 + 43,9 + \ldots + 44,6 + 43,1}{8} = 44,1 \\ s^2 &= \frac{(44,2 - 44,1)^2 + (43.9 - 44,1)^2 + \ldots + (44,6 - 44,1)^2 + (43,1 - 44,1)^2}{7} = 0,251 \end{aligned}$$

Para resolver no R, basta utilizar os comandos a seguir.

```
x <- c(44.2, 43.9, 44.7, 44.2, 44.0, 43.8, 44.6, 43.1)
mean(x)
var(x)
```

7.4.2 Estimação Intervalar - Intervalos de Confiança

Quando fazemos uma estimação pontual obtemos apenas um valor, o que não nos fornece muitas informações sobre a amostra e consequentemente sobre a população

que está sendo pesquisada. Já a estimação por intervalo, é com frequência preferido. Essa técnica fornece um intervalo de valores razoável no qual se presume que esteja o parâmetro de interesse, por exemplo a média da população, com certo grau de confiança. Esse intervalo de valores é chamado de **INTERVALO DE CONFIANÇA**.

Um intervalo de confiança sempre é calculado selecionando-se primeiro o nível de confiança (1 - α), que é uma medida do grau de confiabilidade do intervalo (95%, 99% e 90%). Quanto maior o nível de confiança, mais fortemente acredita-se que o verdadeiro valor do parâmetro estimado está no intervalo construído.

Considerando um intervalo de 95% de confiança, isto é, com nível de significância $\alpha = 5\%$, podemos esperar que em 100 intervalos, 95 dele contenham o verdadeiro valor do parâmetro avaliado, e os 5 demais não contenham o valor deste parâmetro. Isso podemos observar na Figura 47.

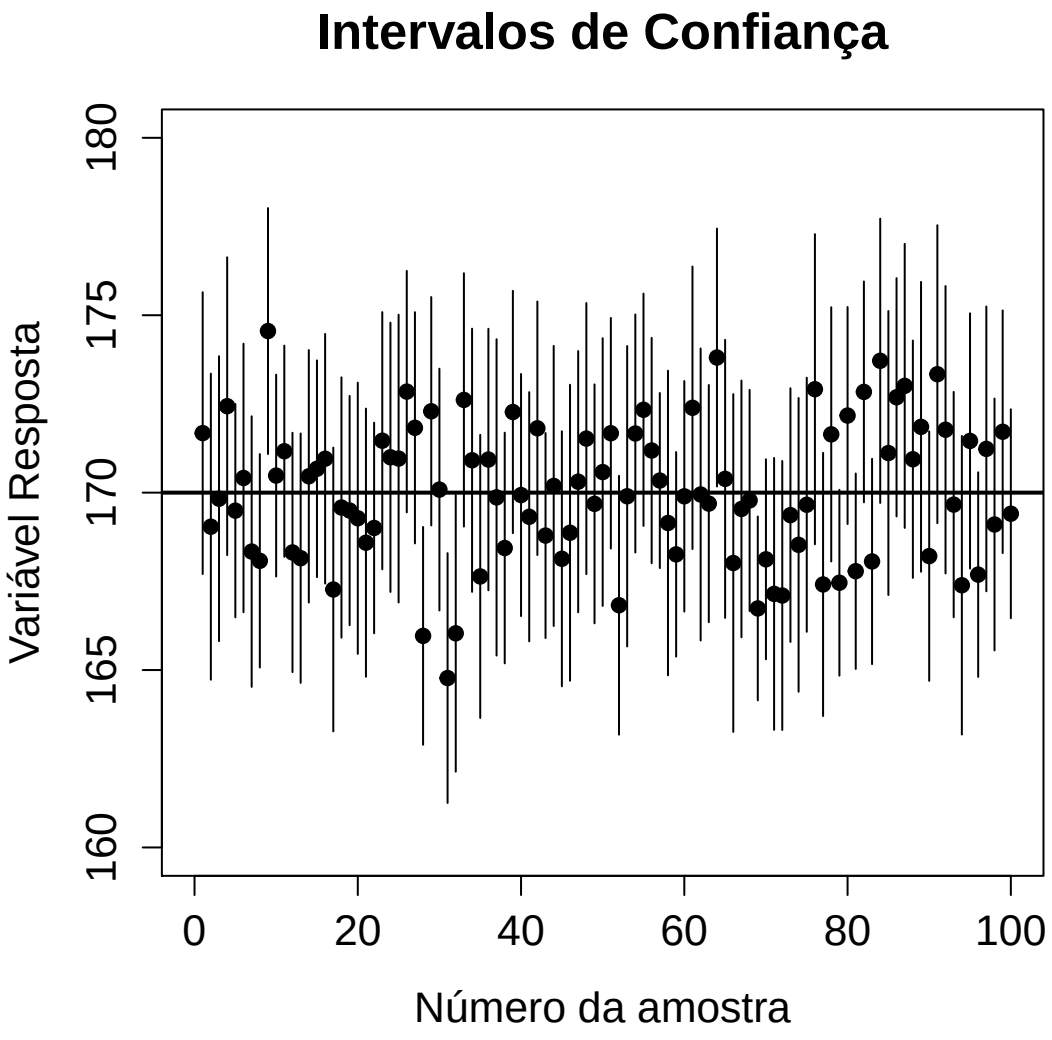


Figura 47 – Intervalo de 95% para média de 100 amostras.

As informações sobre a precisão de uma estimativa de intervalo são transmitidas pela sua extensão. Se o nível de significância for alto e intervalo resultante, bastante restrito, nosso conhecimento do valor do parâmetro será razoavelmente preciso. Um intervalo de confiança muito amplo, entretanto, passa a ideia de que há muita incerteza com relação ao valor do que estamos estimando.

Se [A,B] é um intervalo de confiança para o parâmetro θ com nível de confiança $1-\alpha$, se deve ter

$$P(A \leq \theta \leq B) = 1 - \alpha.$$

Observação: Os extremos dos intervalos se determinam em base a valores amostrais, portanto, não se tem seguridade total que o parâmetro desconhecido esteja no intervalo formado.

7.4.2.1 Intervalo de Confiança para Média com variância populacional conhecida.

Para motivar a construção do intervalo de confiança para média, considere o exemplo a seguir:

Exemplo 124 *Os engenheiros industriais que se especializam em ergonomia estão preocupados em projetar espaços e dispositivos operados por trabalhadores, de modo a obter maior produtividade e conforto. O artigo "Studies Ergonomically Designed Alphanumeric Keyboards" (Human Factors, 1985, p. 175-187) relata o estudo de altura preferida de um teclado experimental com grande apoio para o pulso e o antebraço. Uma amostra de n = 31 digitadores treinados foi selecionada, e a altura preferida do teclado foi determinada para cada digitador. A altura preferida média resultante da amostra foi* $\overline{x} = 80$ *cm. Assumindo que a altura preferida seja normalmente distribuída com* σ *= 2 cm (valor sugerido pelos dados no artigo), qual é um intervalo plausível de valores para a média populacional?*

Seja X uma variável aleatória que tem distribuição Normal, com média populacional μ e desvio padrão populacional σ, então pelo teorema central do limite, a distribuição amostral da média $\overline{X}$ tem média $\overline{x}$ e desvio padrão $\dfrac{\sigma^2}{\sqrt{n}}$ segue que:

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1). \tag{7.7}$$

Assim, utilizando a expressão 7.7

$$P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right) = P\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) \tag{7.8}$$

$$= P\left(-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \overline{x}-\mu \leq z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \tag{7.9}$$

$$= P\left(\overline{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha \tag{7.10}$$

portanto

$$I_c[\mu, 1-\alpha] = \left[\overline{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \overline{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{7.11}$$

com $z_{\frac{\alpha}{2}}$ sendo o percentil da distribuição normal que depende do nível de significância adotado. Este intervalo que acabamos de construir é chamado de **intervalo bilateral.** A figura a seguir nos mostra uma ideia do intervalo de confiança.

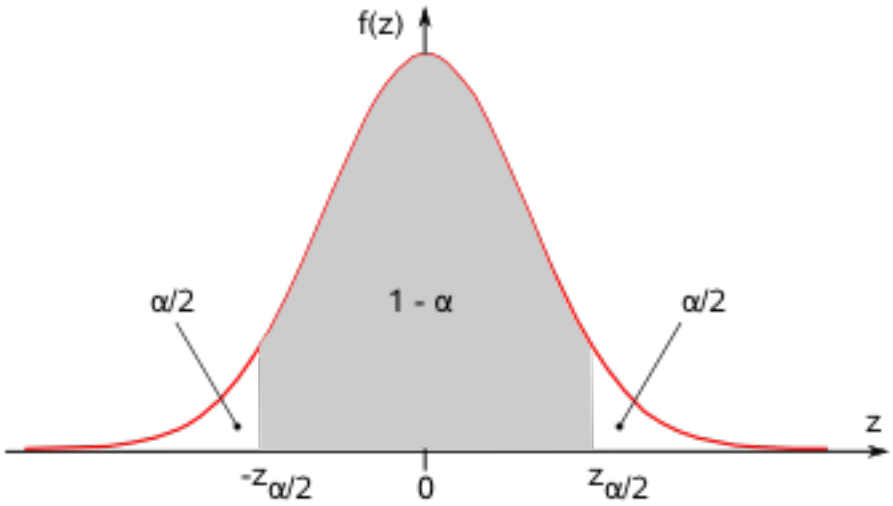


Figura 48 – Intervalo de Confiança Bilateral

Estamos confiantes a cerca de $1-\alpha$ de que o intervalo

$$\left[\overline{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \overline{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{7.12}$$

conterá a média μ. Essa afirmação não implica que a média populacional seja uma variável aleatória que assume um valor no intervalo $(1-\alpha)$ % das vezes, nem que $(1-\alpha)$

% dos valores da população se encontrem nesses imites; ao contrário, ela significa que, se selecionássemos 100 amostras aleatórias da população e as usássemos para calcular 100 intervalos de confiança diferentes para média, aproximadamente $(1-\alpha)$ dos intervalos conteriam a média verdadeira da população e α não.

Em síntese, para determinar um intervalo de confiança a certo nível, quando o desvio padrão populacional é conhecido, basta utilizarmos a expressão 7.12.

Os níveis de confiança mais utilizados são 99%, 95% ou 90%, obtendo $\alpha = 1\%, \alpha = 5\%$ e $\alpha = 10\%$. Os valores de $z_{\frac{\alpha}{2}}$ podem ser obtidos por meio da tabela de distribuição normal padrão, os quais podem ser conferidos abaixo:

Tabela 41 – Níveis de Confiança e de significância.

1 - α	α	$z_{\frac{\alpha}{2}}$
90%	10%	1,645
95%	5%	1,96
99%	1%	2,576

Assumindo que a altura preferida seja distribuída com 2 cm, então o intervalo plausível de valores para a média populacional para o exemplo 124 com um nível de significância de 5% e média conhecida igual a 80 cm:

$$\begin{aligned} I_c[\mu = 80, 1-\alpha = 95\%] &= [80 - 1,96\frac{2}{\sqrt{31}}; 80 + 1,96\frac{2}{\sqrt{31}}] \\ &= [80 - 0,704; 80 + 0,704]. \end{aligned}$$

Ou seja, a altura do teclado para os digitadores ,tolerando, com 95% de confiança uma margem de erro até 0,704 cm.

Exemplo 125 *A experiência com trabalhadores de uma certa indústria indica que o tempo necessário para que um trabalhador, aleatoriamente selecionado, realize uma tarefa é distribuído de maneira aproximadamente normal, com desvio padrão de 12 minutos. Uma amostra de 25 trabalhadores forneceu $\overline{x}$= 140 min. Determinar os limites de confiança de 95% para a média μ da população de todos os trabalhadores que fazem aquele determinado serviço.*

Solução:

$$\begin{aligned} I_c[\mu = 140, 1-\alpha = 95\%] &= [140 - 1,96\frac{12}{\sqrt{25}}; 140 + 1,96\frac{12}{\sqrt{25}}] \\ &= [140 - 4,704; 140 + 4,704] \end{aligned}$$

Os limites de confiança com uma média conhecida igual a 140 min e um intervalo de significância de 95$\%$ de todos os trabalhadores que fazem determinado serviço:

$$I_c[140, 95\%] \quad = \quad [135, 296; 144, 704]$$

Ou seja, o tempo para um trabalhador realizar uma tarefa ,tolerando, com 95% de confiança uma margem de erro até 4,704 cm.

Exemplo 126 *Um fabricante sabe que a vida útil das lâmpadas que fabrica tem distribuição aproximadamente normal com desvio padrão de 200 horas. Para estimar a vida média das lâmpadas, tomou uma amostra de 400 delas, obtendo vida média de 1.000 horas. Construir um IC para μ ao nível de 1%.*

Exemplo 127 *O artigo "Evaluating Tunnel Kiln Performance" (Amer.Ceramic Soe. Buli., ago. 1997, p. 59-63) forneceu as seguintes informações resumidas das resistências a fraturas (Mpa) de n = 169 barras cerâmicas cozidas em um determinado forno: $\overline{x} = 89, 10$, σ = 3,73. Calcule o intervalo de confiança para a resistência à fratura média real, usando um nível de confiança de 95%.*

7.4.2.2 Intervalo de Confiança para Média com variância populacional desconhecida

Nas explicações anteriores, assumimos que o desvio padrão populacional era conhecido, quando calculamos intervalos de confiança para uma média da população. Na realidade é improvável que seja o caso. Se a média populacional for desconhecida, certamente o desvio padrão também será. Porém o cálculo dos intervalos de confiança são os mesmos quando desconhecemos o desvio padrão populacional. Ao invés de usarmos a **distribuição normal padrão**, a análise depende da distribuição de probabilidade conhecida como **distribuição t de Student**.

Graficamente, a podemos representar pela Figura 49.

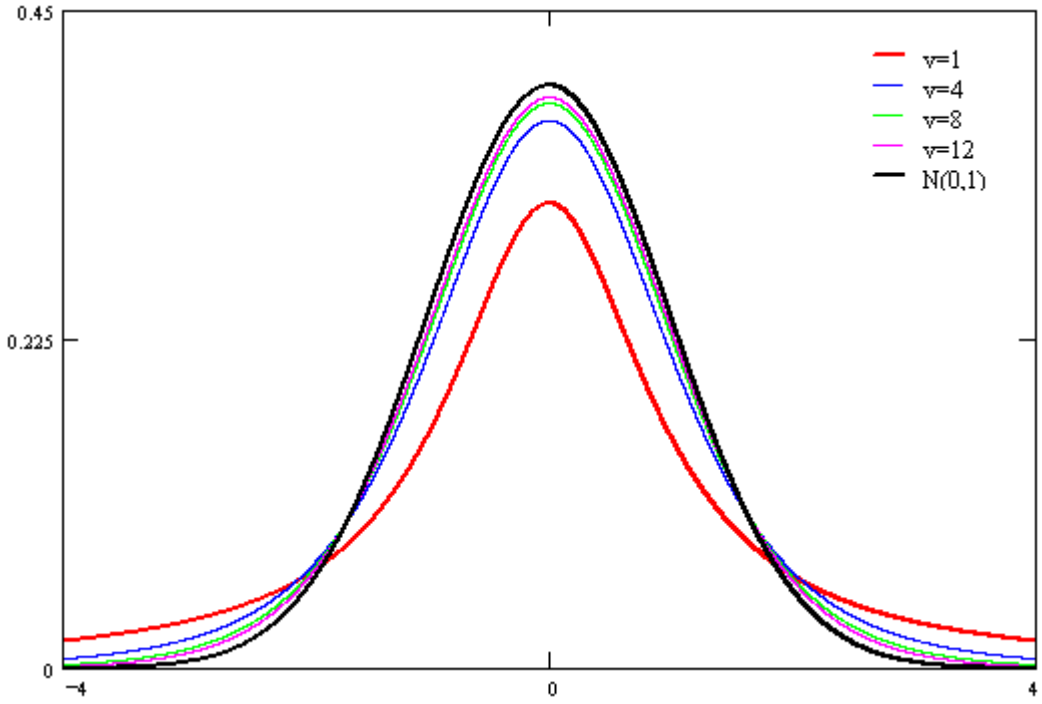


Figura 49 – Distribuição t-Student com n-1 graus de liberdade

Para cada possível valor dos graus de liberdade, há um formato "diferente"para **distribuição t**. As distribuições com um número reduzido de graus de liberdade são mais dispersas. Conforme **gl** aumenta, a **distribuição t** se aproxima da **distribuição normal padrão**. Isso ocorre porque, conforme o tamanho da amostra aumenta, *s* torna-se uma estimativa mais confiável do desvio padrão populacional; se n é muito grande, ou seja o número de elementos da amostra é grande, conhecer o valor de s é quase equivalente a conhecer o valor do desvio padrão populacional.

Portanto, o intervalo de confiança para média quando a variância populacional é desconhecida, e o número de elementos da amostra é pequeno, em geral $n < 30$, será dado por:

$$I_c[\mu, 1-\alpha] \quad = \quad \left[\overline{x} - t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}; \overline{x} + t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}\right] \tag{7.13}$$

Por outro lado, se o número de elementos da amostra é grande, e a variância populacional é desconhecida, como relatado anteriormente, a **distribuição t** se aproxima da distribuição normal padrão. Assim, para determinar o intervalo de confiança utilizaremos o coeficiente $z_{\frac{\alpha}{2}}$ e s como uma estimativa de σ, então o intervalo de confiança será dado por:

$$I_c[\mu, 1-\alpha] \quad = \quad \left[\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{7.14}$$

Exemplo 128 *Um engenheiro agrônomo está estudando a resistência do solo a penetração mecânica. A resistência a penetração é normalmente distribuída com variância de 100 psi^2. De uma amostra aleatória de 12 solos observou-se que a resistência média foi de aproximadamente 325 psi. Construir um intervalo de confiança bilateral com nível de significância de 5%.*

Solução: *Usando um nível de confiança de 95$\%$, temos gl=11 e $t_{0,025,11}$ = 2,201, temos:*

$$\begin{aligned} I_c[\mu, 95\%] &= \overline{x} \pm t_{0,025;11}\frac{s}{\sqrt{n}} \\ &= 325 \pm 2,201\frac{\sqrt{100}}{12} \\ &= [325 \pm 6,354] \end{aligned}$$

Dessa forma, um intervalo plausível:

$$I_c[\mu, 95\%] = [318,646; 331,354]$$

Ou seja, a resistência à penetração mecânica dos solos, tolerando, com 95% de confiança uma margem de erro até 6,354 psi.

Em determinadas ocasiões, é dada a amostra, e assim deve-se obter $\overline{x}$ e S, para posteriormente construir um I_C desejado.

Exemplo 129 *Deseja-se avaliar a dureza espera μ do aço produzido sob um novo processo de têmpera. Uma amostra de dez corpos de prova de aço produziu os seguintes resultados de dureza, em HRc:*

36,4 35,7 37,2 36,5 34,9 35,2 36,3 35,8 36,6 36,9

Construir um intervalo de confiança para μ, com nível de confiança de 95$\%$.

Solução*: Calculando as estatísticas para a amostra observada, temos:*

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 36,15 \qquad s^2 = \frac{\sum_{i=1}^{10}(x_i - \overline{x})^2}{10-1} = 0,5405$$

Usando um nível de confiança de 95$\%$, temos, com gl=9, o valor $t_{0,025;9} = 2,262$, resultado em:

$$IC[\mu, 95\%] = \overline{x} \pm t_{0,025;9}\frac{s}{\sqrt{n}} = 36,15 \pm 0,53 = [35,62; 36,68],$$

Ou seja, a resistência mecânica esperada do aço produzido pelo novo processo de têmpera é 36,15 HRc, tolerando, com 95% de confiança, uma margem de erro até 0,53 HRc.

Para resolver no R, basta utilizar os comandos a seguir:

```
data<-c(36.4, 35.7, 37.2, 36.5, 34.9, 35.2, 36.3, 35.8,
   36.6, 36.9)
t.test(data)
```

O resultado fornecido pelo R é apresentado abaixo, onde os limites do intervalo desejado vem logo na sequência de "95 percent confidence interval".

```
One Sample t-test

data:  data
t = 155.48, df = 9, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
35.62405 36.67595
sample estimates:
mean of x
36.15
```

Se o interesse for em intervalos com outros níveis de confiança, basta na função t.test, usar conf=0.xx, sendo xx o nível desejado.

Exemplo 130 *Como parte de um projeto maior para estudar o comportamento de painéis de revestimento tencionado, componente estrutural que está sendo usado extensivamente nos Estados Unidos, o artigo "Time-Dependent Bending Properties of Lumber" (1. of Testing and Eva/., 1996, p. 187 -193) relatou diversas propriedades mecânicas de espécimes de madeira serrada de pinho da Escócia. Considere as seguintes observações sobre o módulo de elasticidade (MPa) obtido 1 minuto depois da aplicação de carga em uma determinada configuração:*

10,490	16,620	17,300	15,480	12,970	17,260	13,400	13,900
13,630	13,260	14,370	11,700	15,470	17,840	14,070	14,760

Obter a média e a variância para variável resposta, e um intervalo de 99% de confi-

ança para média.

$$\begin{aligned} \overline{x} &= \frac{10,490 + 16,620 + 17,300 + \ldots + 17,840 + 14,070 + 14,760}{16} = 14,5325 \\ s^2 &= \frac{(10,490 - 14,5325)^2 + \ldots + (14,760 - 14,5325)^2}{15} = 4,23 \end{aligned}$$

Usando um nível de confiança de 99%, temos, com gl=15, o valor $t_{0,05;15} = 2,92$*, resultado em:*

$$IC[\mu, 99\%] = \overline{x} \pm t_{0,05;15}\frac{s}{\sqrt{n}} = 14,53 \pm 1,50 = [13,03; 16,03]$$

Para resolver no R, basta utilizar os comandos a seguir:

```
x <- c(10.490, 16.620, 17.300, 15.480, 12.970, 17.260,
   13.400, 13.900, 13.630, 13.260, 14.370, 11.700, 15.470,
    17.840, 14.070, 14.760)
mean(x)
var(x)
t.test(x,conf=0.99)
```

Portanto, a média para o módulo de elasticidade (MPa) obtido 1 minuto depois da aplicação de carga em uma determinada configuração 14,53, tolerando, com 99% de confiança, uma margem de erro até 1,50.

7.4.2.3 Erro padrão da estimativa da média

Se o valor de n for suficientemente grande, $n > 30$, e a população for infinita (ou significativamente grande para não ser preciso aplicar o fator de correção para população finita), tem-se que o erro-padrão para a média é dado por

$$e = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Sendo:

- $z_{\frac{\alpha}{2}}$: Percentil para o nível de confiança adotado;
- σ: Desvio Padrão;
- n: Tamanho da amostra.

Exemplo 131 *Um grupo de técnicos em eficiência pretende utilizar a média de uma amostra aleatória de tamanho n = 150 para estimar a aptidão mecânica média (avaliada por*

certo teste padronizado) dos operários da linha de montagem em uma grande indústria. Se, com base na experiência, os técnicos admitem que σ = 6,2 para tais dados, o que podem eles afirmar, com 0,99 de probabilidade, sobre o erro máximo de sua estimativa? E se o intervalo construído tiver 95% de confiança, qual será o erro máximo sobre sua estimativa?

Solução: Obtendo as quantidades desejadas, temos que para 99%

$$e = 2{,}576\frac{6{,}2}{\sqrt{150}} = 1{,}3040$$

Já para 95%

$$e = 1{,}96\frac{6{,}2}{\sqrt{150}} = 0{,}9922$$

7.4.2.4 Tamanho amostral

Isolando n na expressão que fornece o erro máximo de sua estimativa, podemos obter o tamanho da amostra para que o nível de confiança necessário seja atingido. Nesse sentido, temos que:e

$$n = \left(\frac{z_{\frac{\alpha}{2}}\sigma}{e}\right)^2$$

Sendo

- $z_{\frac{\alpha}{2}}$: Percentil para o nível de confiança adotado;
- σ: Desvio padrão;
- e: Erro-padrão que pode ser adotado.

Exemplo 132 *O administrador de uma agência de turismo pretende utilizar a média de uma amostra aleatória para estimar o tempo médio que os clientes gastam para serem atendidos em informações telefônicas. Ele deseja ter 95% de confiabilidade e que seu erro máximo seja de 0,25 minutos. Se, por estudos anteriores, ele sabe que é razoável supor $\sigma = 1,50$ minutos, qual o tamanho da amostra necessário?*

Solução: *Utilizando as informações fornecidas, temos que*

$$n = \left(\frac{1,96 * 1,5}{0,25}\right)^2 = 11,76^2 \approx 137$$

7.4.2.5 Intervalo de Confiança para Proporção Amostral

Em determinadas situações podemos estar interessados em obter um intervalo plausível de valores para o verdadeiro valor da proporção amostral de uma característica determinada de um conjunto de dados.

Para responder esse questionamento devemos recorrer a construção do intervalo de confiança para proporção amostral.

Seja p a proporção de *"sucessos"* de uma população em que sucesso identifica um indivíduo ou objeto que tenha uma propriedade especificada. Uma amostra aleatória de n indivíduos será selecionada e X é o número de sucesso na amostra. Se n é suficientemente grande, pelo teorema central do limite, a distribuição amostral das proporções segue uma distribuição normal padrão, isto é,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

sendo $\hat{p} = \frac{X}{n}$, a fração de sucessos da amostra e $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ o erro-padrão da estimativa das proporções.

Assim,

$$\begin{aligned} P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right) &= P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sigma_{\hat{p}}} \leq z_{\frac{\alpha}{2}}\right) && (7.15)\\ &= P\left(-z_{\frac{\alpha}{2}}\sigma_{\hat{p}} \leq \hat{p} - p \leq z_{\frac{\alpha}{2}}\sigma_{\hat{p}}\right) && (7.16)\\ &= P\left(\hat{p} - z_{\frac{\alpha}{2}}\sigma_{\hat{p}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sigma_{\hat{p}}\right) \approx 1 - \alpha && (7.17) \end{aligned}$$

portanto

$$I_c[p, 1-\alpha] = [\hat{p} - z_{\frac{\alpha}{2}}\sigma_{\hat{p}}; \hat{p} + z_{\frac{\alpha}{2}}\sigma_{\hat{p}}] \quad (7.18)$$

com $z_{\frac{\alpha}{2}}$ sendo o percentil da distribuição normal que depende do nível de significância adotado. Este intervalo que acabamos de construir é chamado de **intervalo bilateral.**

Esquematicamente um intervalo de, por exemplo, 95% de confiança para proporção amostral pode ser representado por

Exemplo 133 *Retiramos de uma população uma amostra de 100 elementos e encontramos 20 sucessos. Ao nível de 5%, construir um IC para a proporção real de sucessos na população.*

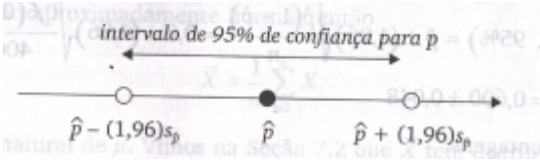


Figura 50 – Intervalo de confiança para proporção amostral. Retirado de Barbetta, Reis e Bornia (2010)

Solução: *Dessa forma as informações que podem ser retiradas do problema:*

$$n = 100$$
$$X = 20$$
$$\alpha = 1\%$$

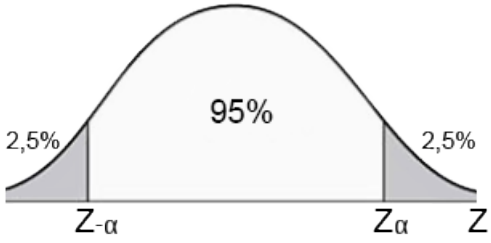
Calculamos os parâmetros iniciais para resolução do problema:

$$\hat{p} = \frac{X}{n} = \frac{20}{100} = 0,2$$

$$\hat{q} = 1 - \hat{p} = 0,8$$

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}.\hat{q}}{n}} = \sqrt{\frac{0,2.0,8}{100}} = 0,04$$

$$Z_\alpha = Z_{5\%} = 1,96$$



Dessa forma é obtido o intervalo de confiança para um nível de significância de 95%:

$$\begin{aligned} I_c[p, 1-\alpha] &= [\hat{p} - z_{\frac{\alpha}{2}}\sigma_{\hat{p}}; \hat{p} + z_{\frac{\alpha}{2}}\sigma_{\hat{p}}] \\ &= [0,2 \pm 1,96.0,04] \\ &= [0,2 \pm 0,0784] \end{aligned}$$

$$I_c[p, 95\%] = [0,1216; 0,2784]$$

Portanto, corremos um risco de 5% de que a verdadeira proporção populacional não pertença a IC dado anteriormente, ou então nossa confiança de que p pertença ao IC determinado é 95%

Exemplo 134 *Uma unidade fabril da Intel produziu 500.000 chips Pentium IV em certo período. São selecionados, aleatoriamente, 400 chips para testes.*

a) *Supondo que 20 chips não tenham velocidade de processamento adequada, construir o intervalo de confiança para proporção de chips adequados, usando 95% de confiança.*

b) *Verificar se essa amostra é suficiente para obter um intervalo de 99% de confiança, com erro amostral máximo de 0,5%, para proporção de chips adequados. Caso contrário, qual deveria ser o tamanho da amostra?*

Exemplo 135 *Em uma linha de produção de certa peça mecânica, colheu-se uma amostra de 100 itens, constatando-se que 4 peças eram defeituosas. Construir um IC para a proporção "p"das peças defeituosas ao nível de significância de 10%.*

Exemplo 136 *Ao longo de um período uma empreiteira realizou 150 construções. Ao todo, em 35 construções foram observados defeitos estruturais. Construir um IC para a proporção "p"das construções com defeitos estruturais ao nível de significância de 1%.*

7.5 Teste de Hipóteses

7.5.1 Introdução

Em diversas situações temos alguma noção, ou ideia, sobre o comportamento de uma variável resposta, ou da possível relação entre duas ou variáveis. Todavia, sem resultados concretos não podemos sair fazendo afirmação sobre o que achamos. Com base nos dados e técnicas estatísticas devemos testar essas informações para que elas possam ser generalizadas para população como um todo. Nesse sentido, adota-se que a população de pesquisa seja o "mundo real"e as ideias ou o que devemos testar sejam as hipóteses de pesquisa, que poderão (deverão) ser testadas por meio de ***Testes de Hipóteses***.

Suponha que X é uma variável aleatória tempo de duração de uma lâmpada de LED. Com base em registros acredita-se que elas durem em média aproximadamente 50.000 horas, com desvio padrão de 15.0000 horas. De uma população retiramos uma amostra de n lâmpadas e a partir das observações verifica-se que o tempo médio de reação não é exatamente o observado. Os responsáveis pelo controle de qualidade não satisfeitos com a informação, gostariam de saber se esse tempo vale para todos as lâmpadas produzidas. Podem ocorrer três tipos de situações:

- Um dos responsáveis supõe que o tempo de duração é diferente 50.000 horas, enquanto outro diz que acredita que tempo seja diferente de 50.000 horas;
- Em uma segunda opinião é feita a afirmação que tempo de duração é menor que 50.000 horas;
- Por fim pode ocorrer que alguém diga que o tempo de médio de duração é maior que 50.000 horas.

Nessas três situações podemos conduzir um teste de hipóteses para verificar a veracidade das informações. Em cada caso a primeira afirmação feita por cada pesquisador é conhecida como **Hipótese Nula**, aquela que colocamos em prova, enquanto a segunda afirmação é chamada de **Hipótese Alternativa**, que é a hipótese que contradiz a hipótese nula. Para conduzirmos um teste de hipótese fixamos um nível de significância. Os mais usuais são 1%, 5% ou 10%.

É fácil perceber que em cada uma das três situações observadas anteriormente teríamos um formato para o teste a ser realizado. No primeiro caso estaríamos colocando em prova que a o tempo médio de duração para população é igual ao parâmetro dado, enquanto na outra hipótese teríamos que o tempo médio de duração para população é

diferente do observado, ou seja, não sabemos se é maior ou menor. Nossas hipóteses seriam, sendo $\mu_0 = 50.000$ horas:

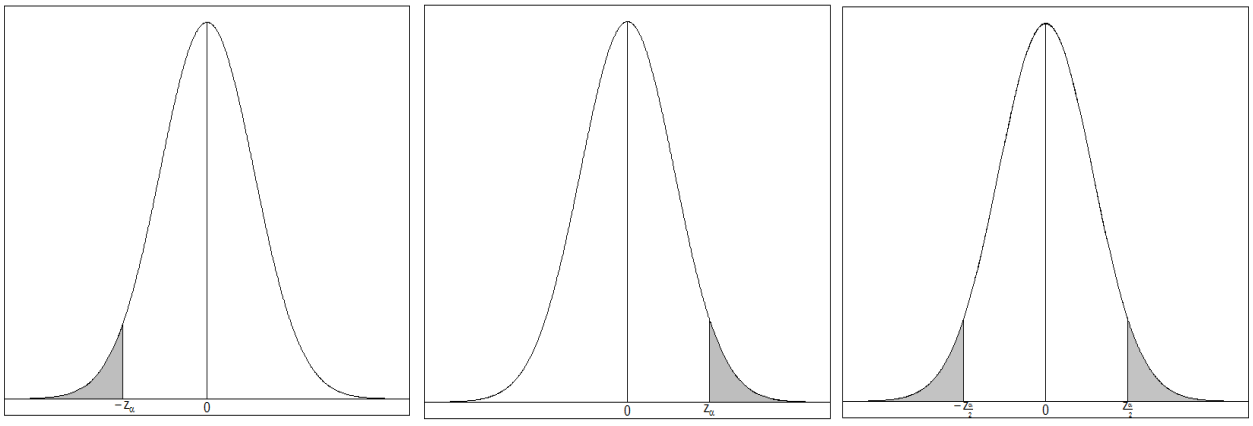
$$T : \begin{cases} H_0 : \mu = \mu_0 \\ H_a : \mu \neq \mu_0. \end{cases}$$

Nessa situação o nosso teste é chamado de **Teste de Hipóteses Bilateral**. Para os outros dois casos teríamos um **Teste de Hipóteses Unilateral**, pois nossa **hipótese alternativa** diz que o tempo médio de duração ou é apenas maior ou é menor que 50.000 horas. Assim, as hipóteses podem ser formuladas da seguinte forma:

$$T : \begin{cases} H_0 : \mu = \mu_0 \\ H_a : \mu < \mu_0 \end{cases} \quad T : \begin{cases} H_0 : \mu = \mu_0 \\ H_a : \mu > \mu_0 \end{cases}$$

Como vimos anteriormente é possível fazermos um teste bilateral ou um teste unilateral. A escolha entre um ou outro pode ser considerada controversa. Não é raro que um teste unilateral tenha significância enquanto um teste bilateral não. Usarmos um ou outro, depende do objetivo da análise que está sendo feita.

Todo teste tem uma região que é chamada de ***Região Crítica***, a qual é formada por um conjunto de valores assumidos pela variável aleatória ou estatística para os quais a hipótese nula é rejeitada.



(a) Gráfico para Teste Unilateral à Esquerda

(b) Gráfico para Teste Unilateral à Direita

(c) Gráfico para Teste Bilateral

Figura 51 – Regiões críticas para testes unilaterais e bilateral

Na Figura 51 temos a ilustração da região crítica para os testes unilaterais (Figuras 55a e 55b) e bilateral (Figura 51c) respectivamente. Quando o teste é unilateral a direita ou

a esquerda, temos apenas uma região de rejeição, enquanto um teste bilateral tem duas regiões de rejeição, como pode ser observado.

Geralmente ao conduzir um teste de hipóteses estamos expostos a dois tipos de erro: **Erro tipo I** e **Erro tipo II**. Se por um lado o **erro tipo I**, representado por α, ocorre quando rejeita-se a hipótese nula, dada que ela é verdadeira. Por outro lado, se não rejeitarmos a hipótese nula dada que ela seja falsa, estaremos cometendo um **erro tipo II**.

De um modo geral, esses dois tipos de erros podem ser comparados a um julgamento. Podemos ter duas situações: O indivíduo julgado, que é inocente é condenado ou um indivíduo culpado é inocentado. Esses são exemplos dos erros tipo I e tipo II respectivamente. A Tabela 52 nos dá uma ideia de como podemos compreender essa situação.

H_0 / Decisão	Verdadeira	Falsa
Não rejeitar	Não há erro	Erro do tipo II
Rejeitar	Erro do tipo I	Não há erro

Figura 52 – Erro tipo I e erro tipo II.

Matematicamente temos as seguintes expressões:

$$\text{P(Rejeitar } H_0 \text{| } H_0 \text{ é verdadeira)} = \alpha \text{ e P(Não Rejeitar } H_0 \text{| } H_0 \text{ é falsa)} = \beta$$

Em geral, é sempre desejável que α e β sejam próximos de zero. Todavia, é fácil ver que a medida que diminuímos α, β aumenta. A Figura 53 a seguir apresenta esta relação.

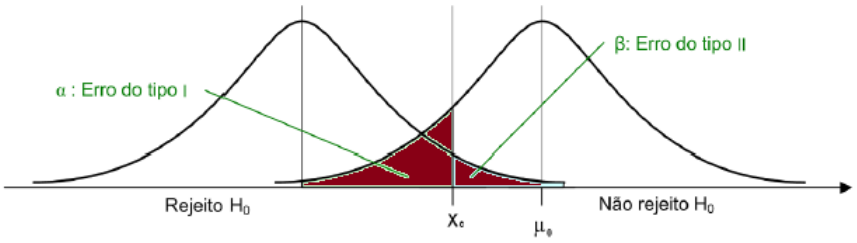


Figura 53 – Controle dos erros.

Podemos estabelecer um paralelo dessa situação com o Controle Estatístico de Processos. Quando temos limites do tipo "3-σ", aumentamos as chances de cometer um erro tipo II, ou seja, podemos estar dizendo que o processo está sob controle enquanto isso não ocorre.

Em determinadas situações deseja-se saber qual é a probabilidade de controle sobre o erro tipo II, isto é, quando a hipótese alternativa será rejeitada quando realmente

for falsa. Essa quantidade é chamada de **Poder do teste** e representamos por $1 - \beta$. Ela poderá ser calculada com base no teste utilizado.

A decisão de rejeitar ou não a hipótese nula pode ser feita com base em dois conceitos. O primeiro está relacionado com a probabilidade de obtermos um valor para média, tão ou mais extrema que a média ou proporção da amostra observada, dado que a hipótese nula é verdadeira, chamamos de **p-valor** ou **valor p**. Manualmente é muito difícil calcularmos esse valor. No entanto, computacionalmente, as funções implementadas em diversos *softwares* fornecem o. Logo para rejeitarmos ou não rejeitarmos a hipótese nula, podemos nos basear na comparação do p-valor com o nível de significância (α) que o teste de hipóteses foi conduzido. De um modo geral ocorre, em um teste bilateral.

- Se $p < \alpha$, rejeitamos H_0. Isso ocorre pelo fato da probabilidade de erro ser inferior ao nível de significância adotado.
- Se $p > \alpha$, não rejeitamos H_0. Isso ocorre pelo fato da probabilidade de erro ser superior ao nível de significância adotado.

Atualmente existem diversos artigos na literatura, os quais condenam o uso do **p-valor**. Isso ocorre por ele ser um resultado pontual e não fornecer muitas informações acerca do problema. É necessário observar o problema como um todo para que os resultados obtidos sejam confiáveis.

O segundo conceito está relacionado aos limites das regiões críticas e até mesmo comparar as estatísticas de testes com os valores fixados para os percentis das distribuições de probabilidade utilizadas, os quais serão apresentados na sequência.

O procedimento básico de teste de hipóteses relativo ao parâmetro θ de uma população, será decomposto em 4 passos:

- a) Definição das hipóteses:

$$H_0 : \theta = \theta_0 H_1 : \theta \leq \theta_0, \theta \geq \theta_0 ou \theta \neq \theta_0$$

- b) Identificação da estatística do teste e caracterização da sua distribuição;
- c) Definição da regra de decisão, com a especificação do nível de significância do teste;
- d) Cálculo da estatística de teste e tomada de decisão.

7.5.2 Teste de Hipótese para média

Seja X uma variável aleatória contínua com média μ_0 e desvio padrão σ conhecido. Pelo Teorema Limite Central, o cálculo da estatística do teste é realizado por meio da expressão

$$z_{calc} = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{7.19}$$

em que,

- $\overline{x}$: Média Amostral;
- μ_0: Média Testada;
- $\frac{\sigma}{\sqrt{n}}$: Erro-padrão.

Nossas hipóteses estatísticas são:

$$T: \begin{cases} H_0 : \mu = \mu_0 \\ H_a : \mu \neq \mu_0 \end{cases}$$

A Figura 54 nos fornece a representação gráfica para o teste bilateral para média.

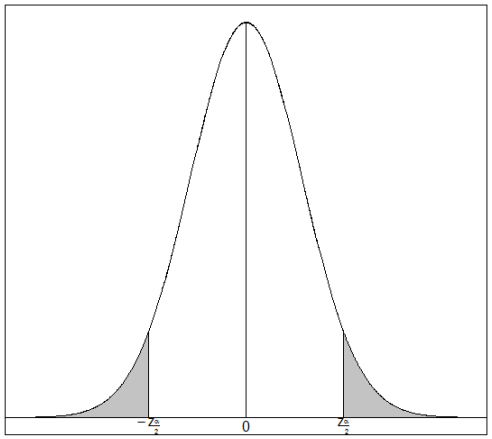


Figura 54 – Teste de Hipóteses Bilateral.

De acordo com Figura 54 a hipótese nula será rejeitada quando uma das condições abaixo for satisfeita:

a) $z_{calc} < -z_{\frac{\alpha}{2}}$;

b) $z_{calc} > z_{\frac{\alpha}{2}}$.

Vale ressaltar que nessa situação conhecemos a variância populacional e por consequência o desvio padrão populacional. Caso esse fosse desconhecido, deveríamos estimá-lo utilizando o desvio padrão amostral que é representado por **s**. Assim, ao invés de utilizar a estatística z, usamos a estatística t com n-1 graus de liberdade, a qual segue distribuição t de Student e os cálculos são análogos.

Exemplo 137 [1] *Foram coletados dados na Universidade Tecnológica Federal do Paraná, em que buscou constatar se a tensão fornecida pela rede corresponderia ao previsto, ou seja, se a tensão à qual os equipamentos eletrônicos estão sujeitos seria de 220 V ou 127 V. Nesse caso, a análise se baseou em constatar a tensão de 127 V em tomadas da Universidade. A análise de dados ocorreu no bloco I, durante o período da tarde, do dia 11 de novembro de 2015. Foram analisadas no total 15 tomadas, as quais compuseram a amostra a seguir:*

125;124;125;125;125;125;124;123; 122;123;123;123;123;124;124

O objetivo foi constatar se a tensão fornecida pela rede corresponderia ao previsto, ou seja, se a tensão à qual os equipamentos eletrônicos estão sujeitos seria de 220 V ou 127 V. Nesse caso, a análise se baseou em constatar a tensão de 127 V em tomadas da Universidade, com $\alpha = 5\%$.

Solução: Inicialmente definem-se as hipóteses a seguir.

$$T : \begin{cases} H_0 : \mu = 127 \\ H_a : \mu \neq 127 \end{cases}$$

Como n<30, utilizaremos a estatística para distribuição t-Student. Nesse caso temos:

$$t_{calc} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{123,87 - 127}{\frac{0,99}{\sqrt{15}}} \approx -12,25$$

Nesse caso, temos um teste bilateral, e assim $t_{\frac{\alpha}{2},14} = \pm 2,15$ e $t_{calc} < -t_{\frac{\alpha}{2}}$, rejeita-se a hipótese nula e conclui-se que a tensão fornecida na tomadas do bloco avaliado é diferente de 127 V.

Para resolver no R, basta utilizar os comandos a seguir.

[1] Dados coletados pelos alunos Eduardo A. K. Fonseca, Gabriel B. Polli, Joel F. Carvalho, do curso de Engenharia Elétrica, para apresentação de seminário da disciplina de Probabilidade e Estatística no segundo semestre de 2015

```
x<-c(125,124,125,125,125,125,124,123,122,123,123,123,123,124,124)
t.test(x,conf=0.95,mu=127)
```

A seguir são apresentados os resultados obtidos no R.

```
One Sample t-test

data:  x
t = -12.253, df = 14, p-value = 7.155e-09
alternative hypothesis: true mean is not equal to 127
95 percent confidence interval:
123.3182 124.4151
sample estimates:
mean of x
123.8667
```

Se a análise for baseada no p-valor, é possível identificar que seu valor correspondente é 7.155e-09, o qual é inferior ao nível de 5%, indicando que H_0 deve ser rejeitada, confirmando os cálculos manuais realizados inicialmente.

Exemplo 138 *A vida média de uma amostra de 100 lâmpadas fluorescentes fabricadas por determinada companhia é de 1750 horas, com desvio padrão de 120 horas. Teste a hipótese de que a vida média de todas as lâmpadas fabricadas por essa companhia é diferente de 1600 horas, nos níveis de 5% e 1% de significância.*

Exemplo 139 *Uma empresa construtora de aviários, utilizando seus empregados edificou 12 aviários de certo padrão, gastando 45,8, 51,4 , 46,1 , 50,9 , 48,7 , 53,2 , 47,9 , 50,1 , 49,3 , 52,6 , 44,9 , 54,4 horas por metro quadrado. Testar a hipótese de ser diferente de 50 horas o tempo médio necessário de mão-de-obra para a construção daquele padrão de residência, adotando um nível de significância de 5%.*

***Solução**: Inicialmente definem-se as hipóteses a seguir.*

$$T: \begin{cases} H_0: \mu = 50 \\ H_a: \mu \neq 50 \end{cases}$$

Como n<30, utilizaremos a estatística para distribuição t-Student. Nesse caso temos:

$$t_{calc} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{49,61 - 50}{\frac{0,95}{\sqrt{12}}} \approx -0,44$$

Para resolver no R, basta utilizar os comandos a seguir.

```
X= c(45.8,51.4,46.1,50.9,48.7,53.2,47.9,50.1,49.3,52.6,44.9,54.4)
t.test(X, conf=0.95, mu=50)
```

A seguir são apresentados os resultados obtidos no R.

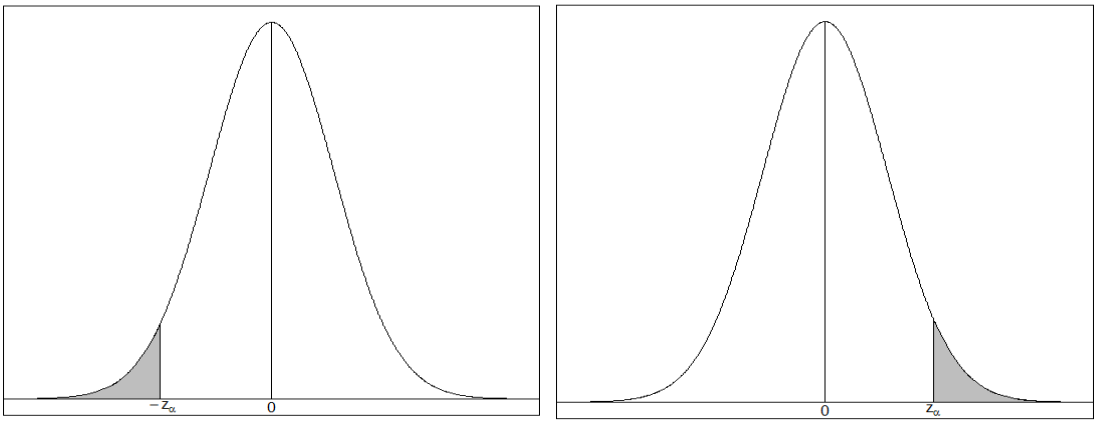
```
    One Sample t-test

    data:  X
    t = -0.44393, df = 11, p-value = 0.6657
    alternative hypothesis: true mean is not equal to 50
    95 percent confidence interval:
    47.66647 51.55020
    sample estimates:
    mean of X
    49.60833
```

Se a análise for baseada no p-valor, é possível identificar que seu valor correspondente é 0,6657, o qual é superior ao nível de 5%, indicando que H_0 não deve ser rejeitada, confirmando os cálculos manuais realizados inicialmente.

Se o interesse é conduzir um teste unilateral, tem-se as hipóteses abaixo,

$$T: \begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu < \mu_0 \end{cases} \quad T: \begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu > \mu_0 \end{cases}$$

O teste conduzido na primeira sentença é um teste unilateral a esquerda e o teste conduzido na segunda sentença é um teste unilateral a direita. Suas representações gráficas são respectivamente,



(a) Gráfico para Teste Unilateral à Esquerda (b) Gráfico para Teste Unilateral à Direita

Figura 55 – Regiões críticas para testes unilaterais

O processo para obter a estatística calculada com base nos valores amostrais é análogo ao processo feito para os testes bilaterais. Porém devemos tomar cuidado, pois existe apenas uma região de rejeição, então utiliza-se z_α e não $z_{\frac{\alpha}{2}}$.

Para o teste unilateral a esquerda a hipótese nula será rejeitada se:

$$z_{calc} < -z_\alpha$$

E para o teste unilateral a direita a hipótese nula será rejeitada se:

$$z_{calc} > z_\alpha$$

Em resumo ocorre que em um:

- Teste unilateral a esquerda: Rejeita H_0 se $z_{calc} < -z_\alpha$, caso contrário não rejeita-se H_0.
- Teste unilateral a direita: Rejeita H_0 se $z_{calc} > z_\alpha$, caso contrário não rejeita-se H_0.

Da mesma forma que em um teste bilateral, vale ressaltar que nessa situação conhecemos a variância populacional e por consequência o desvio padrão populacional. Caso esse fosse desconhecido, deveríamos estimá-lo utilizando o desvio padrão amostral que é representado por **s**. Assim, ao invés de utilizar a estatística z, usamos a estatística t com n-1 graus de liberdade, a qual segue distribuição t de Student e os cálculos são análogos.

Exemplo 140 *Para os dados do exemplo 137, suponha que o objetivo agora seja verificar se a tensão é inferior a 127 V. Nesse caso, realizar um teste de hipóteses para média, com nível de significância de 5%.*

***Solução**: Inicialmente definem-se as hipóteses a seguir.*

$$T: \begin{cases} H_0: \mu = 127 \\ H_a: \mu < 127 \end{cases}$$

Como n<30, utilizaremos a estatística para distribuição t-Student. Nesse caso temos:

$$t_{calc} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{123,87 - 127}{\frac{0,99}{\sqrt{15}}} \approx -12,25$$

Nesse caso, temos um teste unilateral a esquerda, e assim $-t_{\alpha,14} = -1,76$ *e* $t_{calc} < -t_{\alpha}$*, rejeita-se a hipótese nula e conclui-se que a tensão fornecida na tomadas do bloco avaliado é inferior a 127 V.*

Para resolver no R, fazendo uma pequena alteração, resolve-se da seguinte maneira.

```
t.test(x,conf=0.95,mu=127,alternative="less")

One Sample t-test

data:  x
t = -12.253, df = 14, p-value = 3.578e-09
alternative hypothesis: true mean is less than 127
95 percent confidence interval:
-Inf 124.3171
sample estimates:
mean of x
123.8667
```

É possível concluir que a verdadeira média é inferior a 127 V, ao nível de 5% de significância. Se a análise for baseada no p-valor, é possível identificar que seu valor correspondente é 3.578e-09, o qual é inferior ao nível de 5%*, indicando que* H_0 *deve ser rejeitada, confirmando os cálculos manuais realizados inicialmente.*

Exemplo 141 *Uma fábrica anuncia que o índice de nicotina dos cigarros da marca X apresenta-se abaixo de 26 mg por cigarro. Um laboratório realiza 10 análises do índice obtendo: 26, 24, 23, 22, 28, 25, 27, 26, 28, 24. Sabe-se que o índice de nicotina dos cigarros da marca X se distribui normalmente com variância 5,36 mg^2. Pode-se não-rejeitar a afirmação do fabricante, ao nível de 5$\%$.*

Exemplo 142 *Afirma-se que os trabalhadores de certa cidade industrial gastam em média no mínimo R$ 65,00 mensais no consumo de bebidas alcoólicas. Procurando questionar essa afirmação foi selecionada uma amostra aleatória de 25 trabalhadores daquela cidade e obteve-se média R$ 62,72 com desvio padrão de R$ 14,86. Podemos aceitar a afirmação, com base nesses dados, a um nível de significância de 5%?*

Exemplo 143 *Um fabricante de lajotas de cerâmica introduz um novo material em sua fabricação que aumentará a resistência média, que é de 206 kg. A resistência das lajotas tem distribuição normal com desvio padrão de 12 kg. Retira-se uma amostra de 30 lajotas, obtendo $\overline{x}$ = 210 kg. Ao nível de 10$\%$, pode o fabricante aceitar que a resistência média de suas lajotas tenha aumentado?*

7.5.3 Teste de Hipótese para proporção

O teste para proporção é aplicado em situações nas quais deseja-se verificar se a proporção de algum atributo na população pode ser igual a certo valor p_0. Assim, considere uma amostra aleatória X, tal que $X \sim Bernoulli(p)$, cuja proporção de indivíduos pertencem a uma classe de interesse. Nesse caso, E(X) = p e Var(p) = np(1-p). Nesse caso as hipóteses podem ser:

$$T: \begin{cases} H_0: p = p_0 \\ H_a: p \neq p_0 \end{cases} \quad T: \begin{cases} H_0: p \leq p_0 \\ H_a: p > p_0 \end{cases} \quad T: \begin{cases} H_0: p \geq p_0 \\ H_a: p < p_0 \end{cases}$$

Para testar as hipóteses mencionadas, considerando a distribuição amostral das proporções, a estatística do teste é dada por:

$$z_{calc} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{7.20}$$

As regras de decisão para rejeição ou não da hipótese nula, pata o teste sobre a proporção amostral são a mesma enunciada para o teste de hipóteses para média.

Exemplo 144 *Sabe-se que por experiência que 5% da produção de um determinado artigo é defeituosa. Um novo empregado é contratado. Ele produz 600 peças do artigo com 82 defeituosas. Ao nível de 15%, verificar se o novo empregado produz peças com maior índice de defeitos que o existente.*

***Solução**: Inicialmente definem-se as hipóteses a seguir:*

$$T: \begin{cases} H_0: p = 0,05 \\ H_a: p > 0,05 \end{cases}$$

Retirando os parâmetros necessários para os cálculos:

$$n = 600, x = 82, \hat{p_0} = \frac{82}{600} = 0,137$$

Calculando o desvio padrão:

$$\sigma_{\hat{p}} = \sqrt{\frac{0,05.0,95}{600}} = 0,0089$$

$$Z_{calc} = \frac{0,137 - 0,05}{0,0089} = 9,775$$
$$Z_\alpha = Z_{15\%} = 1,03$$

Como $Z_{calc} > Z_\alpha, Z_{calc}$ pertence a RC, rejeita-se H_0, isto é, com 15% de risco, podemos levantar sérias dúvidas quanto à habilidade do novo empregado na fabricação do artigo, sendo sua proporção de defeitos superior à dos demais.

Exemplo 145 *Um candidato a deputado estadual afirma que terá 60% dos votos dos eleitores de uma cidade. Um instituto de pesquisa colhe uma amostra de 300 eleitores dessa cidade, encontrando 160 que votarão no candidato. Esse resultado mostra que a afirmação do candidato é verdadeira, ao nível de 5%?*

***Solução**: Inicialmente definem-se as hipóteses a seguir.*

$$T: \begin{cases} H_0 : p = 0,60 \\ H_a : p \neq 0,60 \end{cases}$$

Retirando os parâmetros necessários para os cálculos:

$$n = 300, x = 160, \hat{p_0} = \frac{160}{300} = 0,53$$

Calculando o desvio padrão:

$$\sigma_{\hat{p}} = \sqrt{\frac{0,60.0,40}{300}} = 0,0283$$

$$Z_{calc} = \frac{0,53 - 0,60}{0,0283} = -2,474$$
$$Z_\alpha = Z_{2,5\%} = 1,96$$

Como Z_{calc} pertence a RC, rejeita-se H_0, isto é, podemos aceitar que a afirmação do candidato é falsa, a 5% de risco.

7.6 Teste para comparação de duas variâncias

Em geral, precisamos comparar as variâncias de duas populações ou amostras, para posteriormente, poder comparar, no caso de dados pareados (próxima seção), as médias. Ao realizar um teste de hipóteses, para esse fim, estaremos "cientes"que as variâncias **não são homogêneas** ou **não são "semelhantes"**. Caso contrário, admite-se que elas **são homogêneas** ou **"semelhantes"**.

Para um teste bilateral, as hipóteses são:

$$T: \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_a : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Uma vez que tem-se em mãos σ_1^2 ou s_1^2 e σ_2^2 ou s_2^2, a estatística do teste é calculada por meio da expressão:

$$F_{calc} = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1}, \tag{7.21}$$

com $s_1^2 \geq s_2^2$ e $F_{\alpha/2,n_1-1,n_2-1} = F_{Tab}$ um valor crítico teórico ou tabelado, que limita as regiões de rejeição e não-rejeição da hipótese nula, para distribuição F-Snedecor, na cauda superior, com nível de significância α para $n_1 - 1$ e $n_2 - 1$ graus de liberdade do numerador e denominador, respectivamente. Se $F_{calc} > F_{Tab}$, rejeita-se H_0 ao nível α de significância.

Exemplo 146 [2] *No seguinte estudo, foram coletados dados das tensões máximas suportadas (MPa) em três tipos mais usuais de madeiras da região do Sudoeste do Paraná: Itaúba, Cedrinho, Pinheiro. Aqui serão apresentados apenas os dados coletados para as madeiras Itaúba e Cedrinho, com objetivo de verificar se as variâncias dos dois grupos são homogêneas ou não, ao nível 5% de significância. Os resultados são apresentados na Tabela 42.*

Tabela 42 – Dados Coletados

	Madeiras	Resistência (MPa)
1	I1	54,07
2	I2	45,92
3	I3	44,10
4	I4	39,36
5	I5	38,46
6	I6	40,20
7	I7	40,93
8	I8	45,24
9	C1	40,42
10	C2	32,64
11	C3	45,67
12	C4	41,62
13	C5	45,08
14	C6	34,73
15	C7	32,58
16	C8	38,96

***Solução**: Nossas hipóteses são dadas por:*

$$T : \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_a : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

[2] Dados coletados pelos alunos Eduardo A. K. Fonseca, Denisson Centenaro Fortes, Geovane de Campos Soares e Paulo Eduardo B. Sabadin, do curso de Engenharia Civil, para apresentação de seminário da disciplina de Probabilidade e Estatística no segundo semestre de 2015

Inicialmente precisamos obter a média aritmética para cada um dos grupos. Nesse sentido, segue que:

$$M_I = \overline{x}_1 = \frac{54,07 + \ldots + 45,24}{8} = 43,53 \quad e \quad M_C = \overline{x}_2 = \frac{40,42 + \ldots + 38,96}{8} = 35,34.$$

Fazendo os cálculos para obter as estimativas das variâncias, temos:

$$V_I = s_1^2 = \frac{(54,07 - 43,53)^2 + \ldots + (45,24 - 43,53)^2}{8-1} = 25,86$$
$$V_C = s_2^2 = \frac{(40,42 - 35,34)^2 + \ldots + (38,96 - 35,34)^2}{8-1} = 27,17.$$

Obtendo F_{Tab}, temos:

$$F_{calc} = \frac{s^2_{maior}}{s^2_{menor}} = \frac{27,17}{25,86} = 1,051$$

Utilizando os valores tabelados para distribuição F, ano nível de $\alpha/2 = 2,5\%$ de significância , temos $F_{2,5\%,7,7} = 4,99$. Então como, $F_{calc} < F_{Tab}$, não rejeita-se a hipótese de variâncias homogêneas, ao nível de 5% de significância.

Para resolver no R, basta utilizarmos os comandos a seguir:

```
CEDRINHO<-c(40.42,32.64,45.67,41.62,45.08,34.73,32.58,38.96)
ITAUBA<-c(54.07,45.92,44.10,39.36,38.46,40.20,40.93,45.24)
var.test(CEDRINHO,ITAUBA)
```

A saída será a que segue:

```
data:  CEDRINHO and ITAUBA
F = 1.051, num df = 7, denom df = 7, p-value = 0.9494
alternative hypothesis: true ratio of variances is not
   equal to 1
95 percent confidence interval:
0.2104136 5.2496329
sample estimates:
ratio of variances
1.050997
```

Nesse caso, nossa análise poderá se basear no p-valor. E assim, observamos que $p = 0,9494$*, o que implica em* $p > \alpha$*, o que indica a não rejeição de* H_0

7.6.1 Teste de Hipótese para comparação entre médias.

Em várias situações científicas e práticas há interesse em comparar o desempenho de dois ou mais tratamentos, como por exemplo, dois processos de temperatura na produção de aço, dois tipos de cimento-e-cola para fixar azulejos, velocidade de processamento de dois sistemas operacionais, entre outros. Na comparação de tratamentos, é natural o interesse em verificar se há evidência de diferenças entre os efeitos dos tratamentos, os quais podem ser feitos por meio de testes estatísticos.

Ao compararmos as médias de grupos de dados, podemos recair em duas situações distintas. A primeira é que as medições podem ser feitas em, especificamente, dois momentos distintos, implicando uma dependência entre os dados, ou uma paridade entre eles. Nesse caso, um teste estatístico para comparação entre esses dados é o **teste t pareado**. No segundo caso, pode ser necessário comparar médias de dois grupos distintos, isto é, quando as observações são independentes. Isso implica que o teste **t para amostras independentes** deve ser utilizado, sendo que deve-se observar se as variâncias dos conjuntos são homogêneas ou não, como apresentado na seção 7.6.

Nesta seção estaremos interessados em testes estatísticos para comparação entre duas médias, em especial quando os dados possuem as características supracitadas, isto é, paridade ou independência.

7.6.1.1 Teste t para dados pareados

Cientificamente, para validação de uma teoria, é necessário utilizar técnicas estatísticas para avaliar um conjunto de dados e a partir disso fazer inferências. Em diferentes áreas do conhecimento, na grande maioria das vezes, são comparadas amostras de dois dados, em que as variáveis respostas são mensuradas em dois momentos distintos, **antes** e **depois**. Quando os dados possuem essa característica, dizemos que estes são pareados. Essa paridade gera uma dependência entre as observações, isto é, uma correlação entre elas. Quando há mais de duas observações podemos dizer que existem medidas repetidas nos dados.

Exemplos de dados pareados na área da engenharia podemos dizer que são realizações em que um mesmo produto é avaliado em dois períodos distintos de acordo com a aplicação de dois tratamentos distintos. Em síntese, dados emparelhados ocorrem quando os elementos de duas amostras são relacionados dois a dois, de acordo com algum critério

que fornece uma influência entre os vários pares e sobre os valores de cada par. Considere a seguinte situação problema:

Exemplo 147 *Seja o problema de verificar se um novo algoritmo de busca em um banco de dados é mais rápido que o algoritmo atualmente usado. Para fazer a comparação dos dois algoritmos, planeja-se realizar uma amostra aleatória de dez buscas experimentais. Em cada realização, uma dada busca é realizada pelos dois algoritmos e o tempo de resposta é registrado para ambos os processos. Considerando dez realizações, existe diferença entre as velocidades de busca para os dois algoritmos?*

Para responder essa pergunta, considere que devemos calcular as diferenças para cada par de valores, produzindo dados de uma amostra de *n* diferenças. As hipóteses a serem formuladas podem ser escritas da seguinte forma

$$T : \begin{cases} H_0 : \mu_1 - \mu_2 = \mu_{\overline{d}} \\ H_a : \mu_{\overline{d}} \neq 0 \end{cases}$$

Nesse caso, a estatística do teste é dada por

$$t = \frac{\overline{d} - \mu_{\overline{d}}}{\frac{s_{\overline{d}}}{\sqrt{n}}} \sim t_{n-1;\alpha/2} \tag{7.22}$$

em que

- $\overline{d}$ é a média da amostra das diferenças;
- $\mu_{\overline{d}}$ é o valor das diferenças entre médias das populações a ser testado;
- $s_{\overline{d}}$ desvio padrão da amostra das diferenças;
- n é o tamanho da amostra das diferenças.

Para o exemplo 147, considere os dados amostrais antes e depois contidos na Tabela 43.

Verificar, a um nível de 5% de significância, se a velocidade de busca dos dados difere para os dois algoritmos.

Solução: Inicialmente definem-se as hipóteses a seguir.

Tabela 43 – Tempo de busca dos dados

Busca	A_1	A_2
1	22	25
2	21	28
3	28	26
4	30	36
5	33	32
6	33	39
7	26	28
8	24	33
9	31	30
10	22	27

$$T: \begin{cases} H_0: \mu_1 - \mu_2 = \mu_{\overline{d}} = 0 \\ H_a: \mu_{\overline{d}} \neq 0 \end{cases}$$

$$t = \frac{\overline{d} - \mu_{\overline{d}}}{\frac{s_{\overline{d}}}{\sqrt{n}}} \sim t_{n-1;\alpha} = -2,86 \tag{7.23}$$

Para resolver no R, fazendo uma pequena alteração, resolve-se da seguinte maneira.

```
A<-c(22,21,28,30,33,33,26,24,31,22)
B<-c(25,28,26,36,32,39,28,33,30,27)



t.test(A,B,mu=0,paired=TRUE,conf.level=0.95)
  Paired t-test

  data:  A and B
  t = -2.8246, df = 9, p-value = 0.0199
  alternative hypothesis: true difference in means is not equal
     to 0
  95 percent confidence interval:
  -6.1229541 -0.6770459
  sample estimates:
  mean of the differences
  -3.4
```

Conforme o resultado do teste, percebemos que $p-value$ é menor que α. Portanto, há evidências amostrais suficientes para rejeitar a hipótese nula. Concluímos então que há diferença na velocidade de busca dos dados dos dois algoritmos.

Exemplo 148 *Um vestígio de metais em água potável afeta o sabor e concentrações extraordinariamente altas podem apresentar risco à saúde. O artigo "Trace Metals of South Indian River"relata um estudo em que foram selecionados seis locais do rio e a concentração de zinco (mg/L) determinada para a água da superfície e para a água mais profunda em cada local. Os seis pares de observações são exibidos na figura abaixo. Os dados sugerem que a concentração média real na água profunda difere da água superficial?*

	Local					
	1	2	3	4	5	6
Água profunda (x)	0,430	0,266	0,567	0,531	0,707	0,716
Água superfície (y)	0,415	0,238	0,390	0,410	0,605	0,609
Diferença	0,015	0,028	0,177	0,121	0,102	0,107

Tabela 44 – Dados do Exemplo 148

Exemplo 149 *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 5% de significância, podemos concluir que o uso da ração contribuiu para o aumento do peso médio dos animais?*

Tabela 45 – Dados do Exemplo 149

	Peso (gramas)									
Cobaia	1	2	3	4	5	6	7	8	9	10
Antes (x)	635	704	662	560	603	745	698	575	633	669
Depois (y)	640	712	681	558	610	740	707	585	635	682

7.6.1.2 Teste t para amostras independentes

Considere a seguinte situação problema:

Exemplo 150 *Um engenheiro estuda a formulação de uma argamassa de cimento Portland. Ele acredita que a adição de uma emulsão de polímero de látex na mistura tem efeito no tempo de cura e na resistência de tensão da argamassa. Deseja-se verificar se a resistência, a um nível de significância de 5%, média argamassa modificada difere da resistência média da argamassa original.*

Podemos observar que diferentemente do exemplo 147, os dados amostrais são independentes, isto é, os dados da amostra 1 não dependem da amostra 2. Quando isso ocorre, deve-se utilizar o **teste t para amostras independentes**. Nesse caso, devemos ficar atentos a homogeneidade das variâncias, isto é, se elas são semelhantes ou não.

A Tabela 46 contém as expressões que devem ser utilizadas nos casos de teste t para amostras independentes para dados com homogeneidade ou não entre as variâncias.

Tabela 46 – Teste t para amostras independentes

	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$
Estatística	t = $\dfrac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	t = $\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Graus de Liberdade	$v = n_1$+n_2-2	$v = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$
Variância Ponderada	$s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	

Os dados coletados para o exemplo inicial são dados na Tabela 47.

Tabela 47 – Argamassa experimental

	$Leste$	Oeste
n	16,85	16,62
$\overline{x}$	16,40	16,75
s^2	17,21	17,37

Com base nas expressões anteriores, verificar se a resistência média de argamassa modificada difere da resistência média da argamassa original.

Solução: Nossas hipóteses, para análise para comparação das variâncias das duas amostras são dadas por

$$T: \begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_a: \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Inicialmente precisamos obter a média aritmética para cada um dos grupos. Nesse sentido, segue que:

$$M_{A_1} = \overline{x}_1 = \frac{16,85 + \ldots + 16,57}{10} = 16,76 \quad e \quad M_{A_2} = \overline{x}_2 = \frac{16,62 + \ldots + 17,27}{10} = 17,04$$

Fazendo os cálculos para obter as estimativas das variâncias, temos:

$$V_I = s_1^2 = \frac{(16,85 - 16,76)^2 + \ldots + (16,7 - 16,76)^2}{10 - 1} = 0,1001$$
$$V_C = s_2^2 = \frac{(16.62 - 17.042)^2 + \ldots + (17,27 - 17,042)^2}{10 - 1} = 0,0615.$$

Calculando a variância ponderada (s_p), temos:

$$s_p = \sqrt{\frac{(10-1)0,1001 + (10-1)0,0615}{10 + 10 - 2}} \approx 0,2781$$

Calculando os graus de liberdade (v), temos:

$$v = 10 + 10 - 2 = 18$$

Calculando t, temos:

$$t_{calc} = \frac{16,764 - 17,042}{0,2781\sqrt{\frac{1}{10} + \frac{1}{10}}} \approx -2,23$$

Utilizando os valores tabelados para distribuição *t-Student*, ano nível de $\alpha/2 = 2,5\%$ de significância , temos $t_{2,5\%,18} = 2,101$. Então como, $t_{calc} < -t_{Tab}$, rejeita-se a hipótese de igualdade de médias, ao nível de 5% de significância.

Para resolver no R, basta utilizarmos os comandos a seguir:

```
A1=c(16.85,16.40,17.21,16.35,16.52,17.04,16.96,17.15,16.59,16.57)
A2=c(16.62,16.75,17.37,17.12,16.98,16.87,17.34,17.02,17.08,17.27)

ARG<-data.frame(A1,A2)
var.test(ARG$A1,ARG$A2)      #Teste de homogeneidade de variancias
```

A saída será a que segue para teste de homogeneidade das variâncias,

```
F test to compare two variances

data:  ARG$Y1 and ARG$Y2
F = 1.6293, num df = 9, denom df = 9, p-value = 0.4785
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.4046845 6.5593806
sample estimates:
ratio of variances
1.629257
```

Observe que p = 0,4785 é maior que o valor adotado para o nível de significância, indicando assim, que não deve-se rejeitar a hipótese de homogeneidade das variâncias.

Para a comparação entre as médias, o código considerando o resultado acima, o argumento **var.equal = TRUE**, indica homogeneidade para as variâncias, e assim temos:

```
    # Teste de comparacao de medias
    t.test(A1,A2, alternative="two.sided",mu=0, var.equal=TRUE)
```

cujo resultado apresentado é o que segue.

```
    data:  A1 and A2
    t = -2.1869, df = 18, p-value = 0.0422
    alternative hypothesis: true difference in means is not equal
        to 0
    95 percent confidence interval:
    -0.54507339 -0.01092661
    sample estimates:
    mean of x mean of y
    16.764    17.042
```

Nesse caso, nossa análise poderá se basear no p-valor. Dessa forma, observamos que p=0,0422, o que implica em p < α, o que indica a rejeição de H_0.

Exemplo 151 *Um supermercado não sabe se deve comprar lâmpadas da marca A ou B, de mesmo preço. Testa uma amostra de 100 lâmpadas de cada uma das marcas, obtendo:*

- $\overline{x}_A = 1.160h$ *e* $s_A = 90h$*;*

- $\overline{x}_B = 1.140h$ **e** $s_A = 80h$*;*

Ao nível de 2,5% de significância, testar a hipótese de que as marcas são igualmente boas quanto contra a hipótese de que as da marca A são melhores que as da marca B.

Exemplo 152 *Espécimes de concreto com razões de altura/diâmetro variáveis cortados em várias posições no cilindro original foram obtidos de uma mistura de concreto de resistência normal e de uma mistura de alta resistência. A tensão de pico (MPa) foi determinada para cada mistura, resultando nos dados a seguir ("Effect of Length On compressive Strain Softening of Concrete,"J. of Engr. Mechanics, 1997, p.25-35) . Verificar ao nível de significância de 5% se existe diferença entre as misturas de concreto de resistência normal e alta.*

Tabela 48 – Resistências das misturas de Concreto

Cond. Teste	1	2	3	4	5	6	7	8
Normal	42,8	55,6	49,0	48,7	44,1	55,4	50,1	45,7
Alta	90,9	93,1	86,3	90,3	88,5	88,1	93,2	90,8
Cond. Teste	9	10	11	12	13	14	15	
Normal	51,4	43,1	46,8	46,7	47,7	45,8	45,4	
Alta	90,1	92,6	88,2	88,6	91,0	90,0	90,1	

7.7 Exercícios de Aplicação

1. A experiência com trabalhadores de uma certa indústria indica que o tempo necessário para que um trabalhador, aleatoriamente selecionado, realize uma tarefa é distribuído de maneira aproximadamente normal, com desvio padrão de 12 minutos. Uma amostra de 25 trabalhadores forneceu média de 140 minutos. Determinar os limites de confiança de 95% para a média μ da população de todos os trabalhadores que fazem aquele serviço.

2. Uma linha de produção de certa peça mecânica, colheu-se uma amostra de 100 itens, constatando-se que 4 peças eram defeituosas. Construir o IC para a proporção "p"das peças defeituosas ao nível de 10%.

3. Um fabricante sabe que a vida útil das lâmpadas que fabrica tem distribuição aproximadamente normal com desvio padrão de 200 horas. Para estimar a vida média das lâmpadas, tomou-se uma amostra de 400 delas, obtendo vida média de 1.000 horas.

   a) Construir um IC para μ ao nível de 1%.

   b) Qual o valor do erro de estimação cometido em "a"?

   c) Qual o tamanho da amostra necessária para se obter um erro de 5 horas, com 99% de probabilidade de acerto?

4. Querendo estimar a proporção de defeitos de uma certa produção, examinou-se uma amostra de 100 itens, encontrando-se 30 defeitos. Determinar o IC para a proporção p da população ao nível de 5%.

5. Determine com 95% de confiança o intervalo que representa a média populacional referente as amostras abaixo:

   a) Tempo de execução da atividade RR (em minutos). (tempo/ unidade/ funcionários).

3,0 3,6 3,8 3,9 4,1 2,9 3,5 4,2 3,2 3,8 3,5 3,3 4,2 4,0 3,0

   b) Percurso (m) de frenagem do veículo Gol – 1000 com pneus Pirelli 165/70/13 a uma velocidade de 100km/h.

38 40 39 41 42 44 40 48 38 39 39 44 41 35

   c) Durabilidade (1000km) dos amortecedores Coffap ref. 2 x 45 vc 87.

78 84 66 84 90 84 85 69 71 70 66 60 85 80 74 66 63 59

6. Uma empresa fabricante de pastilhas para freios efetua um teste para controle de qualidade de seus produtos. Selecionou-se uma amostra de 600 pastilhas, das quais 18 apresentavam níveis de desgaste acima do tolerado. Construir um intervalo de confiança para proporção de pastilhas com desgaste acima do tolerado, do atual processo industrial, com nível de confiança de 95%. Interpretar o resultado.

7. Com o objetivo de avaliar a confiabilidade de um novo sistema de transmissão de dados, torna-se necessário verificar a proporção de bits transmitidos com erro em cada lote de 100 Mb. Considere que seja tolerável um erro amostral máximo de 2% e que em sistemas similares a taxa de erro na transmissão é de 10%. Qual deve ser o tamanho da amostra?

   a) Use α = 5%;

   b) Use α = 1%.

8. O processo de produção das unidades de caixa de controle de um tipo específico de motor foi modificado recentemente. Antes dessa modificação, os dados históricos sugeriam que a distribuição dos diâmetros do orifício dos mancais nas caixas eram normais, com um desvio padrão de O, 100 mm. Acredita-se que a modificação não tenha afetado o formato da distribuição ou o desvio padrão, mas que o valor do diâmetro médio possa ter mudado. Uma amostra de 40 unidades da caixa é selecionada e o diâmetro do orifício é determinado para cada uma, resultando em um diâmetro médio da amostra de 5,426 mm. Calcular um intervalo de confiança para o diâmetro médio real do orifício usando um nível de confiança de 90%.

9. Uma fábrica de automóveis anuncia que seus veículos consomem, em média, 11 litros por 100 km rodados. Uma revista decide testar essa afirmação e analisa 35 veículos dessa marca, obtendo 11,4 litros por 100 km como consumo médio. Admitindo que o consumo médio tenha distribuição normal, ao nível de 10% o que a revista concluirá sobre o anuncio da fábrica?

10. Em uma experiência sobre percepção extrassensorial (PES), um indivíduo A, em uma sala isolada, é solicitado a declarar a cor vermelha ou preta de cartas tiradas ao acaso de um baralho com 50 cartas, por outro indivíduo B, posicionado em outra sala. Se A identifica corretamente 32 cartas esse resultado é significativo ao nível de 5% para identificar que A tem PES?

11. A vida média de uma amostra de 100 lâmpadas produzidas por uma firma foi calculada em 1.750 horas, com desvio padrão de 120 horas. Sabe-se que a duração das lâmpadas dessa firma tem distribuição normal com média 1.600 horas . Ao nível de 1% testar se houve alteração na duração média das lâmpadas.

12. Um cliente de uma torrefação de café suspeita que os pesos dos pacotes que deveriam ser de 500 g, não estão corretos. Resolveu, então, retirar uma amostra dos pesos de 16 pacotes (supondo que provenham de uma distribuição normal):

510 495 498 500 501 499 503 500 495 492 499 499 497 495 499 501

    a) Calcule o peso médio e o desvio padrão dos elementos da amostras.

    b) O cliente tem razão na suspeita?

13. O controle estatístico de certo processo estabeleceu que pelo menos 94% dos produtos têm que estar sem defeitos. Para verificar a validade desta afirmação, foi coletada uma amostra de 150 produtos, obtendo uma proporção sem defeito igual a 92%.

    a) Com 1% de significância, há evidência de que o processo está em desacordo com o esperado?

    b) Se o percentual real sem defeito fosse 91%, qual é a probabilidade de se tomar uma decisão errada no item a)?

14. Padrões técnicos exigem que o nível de ruído em CPDs seja de, no máximo 70 dB. Foram analisados 16 CPDs de várias organizações, obtendo-se os seguintes valores máximos de ruído:

78 73 68 65 72 64 77 80 82 78 65 72 61 79 58 65

    a) Calcule a intensidade de ruído média e o desvio padrão para esses 16 CPDs.

    b) Suponha que os 16 CPDs analisados são uma amostra aleatória de CPDs do país. Para verificar se na média os CPDs atendem aos padrões técnicos, como você construiria as hipóteses?

    c) Você pode concluir que a intensidade de ruído média dos CPDs nos horários críticos é superior ao especificado? Faça o teste adequado ao nível de significância de 5%.

15. Um produto fabricado por injeção de plástico é analisado em dois níveis de percentual de talco. Os dados seguintes apresentam os resultados da dureza (HRc), segundo o percentual de talco utilizado: Os dados mostram evidência suficiente para afirmar

Tabela 49 – Dados do Exercício 15

Baixo	51,7 49,4 65,9 60,0 71,1 72,9 71,9 75,1
Alto	75,2 76,0 63,7 69,6 67,1 69,1 52,8 57,6

que a dureza média do produto é diferente nos dois níveis de percentual de talco? Use $\alpha = 0,05$.

16. Na comparação de duas topologias de rede de computadores, C1 e C2, avaliou-se o tempo de transmissão de pacotes de dados entre duas máquinas. Foram realizados 32 ensaios em C1 e 24 ensaios em C2, cujos resultados são apresentados a seguir>

Tabela 50 – Dados do Exercício 16

Topologia	Tempo (Em décimo de segundos)
C1	09 12 10 12 11 09 08 12 13 09 13 08 17 09 09 08 14 08 08 08 08 13 10 10 15 13 13 12 14 08 09 08
C2	14 15 08 13 16 12 14 17 14 10 13 12 12 17 16 12 15 13 14 14 13 14 10 15

Existe diferença significativa entre o tempo médio de transmissão nas duas topologias? Use $\alpha = 0,05$.

17. Para comparar dois algoritmos de otimização, foi realizado um experimento com seis ensaios. Em cada ensaio, foram usados separadamente os dois algoritmos em estudo, mas sob as mesmas condições (dados pareados). Os tempos de resposta ao usuário foram:

Tabela 51 – Dados do Exercício 17

Algoritmo 1	8,1 8,9 9,3 9,6 8,1 11,2
Algoritmo 2	9,2 9,8 9,9 10,3 8,9 13,1

Os tempos de resposta dos dois algoritmos são, em média, diferentes? Use $\alpha = 0,05$.

7.7.1 Gabarito

1. $IC(\mu, 95\%) = (135.3; 144, 7)$.
2. $IC(p, 90\%) = (0, 0078; 0, 07214)$.
3. **a)** $IC(\mu, 99\%) = (974, 2h; 1.025, 76h)$;

   b) 25,8h;

   c) n $\approx$ 10.651 lâmpadas.
4. $IC(p, 95\%) = (21, 02\%; 38, 98\%)$.
5. **a)** $IC(\mu, 95\%) = (3, 35; 3, 84)$;

   b) $IC(\mu, 95\%) = (38, 72; 42, 42)$;

   c) $IC(\mu, 95\%) = (69, 23; 78, 98)$;
6. $IC(\mu, 95\%) = (1, 64; 4, 36)$.
7. **a)** 865;

   b) 1.494.
8. $IC(\mu, 95\%) = (5, 400; 5, 452)$.
9. Rejeita-se H_0.
10. Rejeita-se H_0.
11. Não rejeita-se H_0.
12. **a)** $\overline{x}$= 498,94 e s^2= 4,07;

    b) Não.
13. **a)** Não;

    b) 0,74.
14. **a)** $\overline{x}$= 71,06 e s^2= 7,49;

    b) H_0= 70 e $H_0 > 70$;

    c) t= 0,57.
15. t= -0,36.
16. Sim, teste t para amostras independentes t= -4,40.
17. Sim, t= 5,175.

Capítulo 8

Análise de Variância

8.1 Introdução

A realização de experimentos no meio científico, tornou-se frequente nos dias atuais. A busca desenfreada por novas drogas, tratamentos e bens de consumo, dependem de análises estatísticas para que exista a validação das respostas inerentes ao problema.

Entre as análises podemos destacar a técnica de Análise de Variância e sua equivalente não-paramétrica, o Teste de Kruskall-Waillis. Ambos nos permitem comparar os efeitos dos níveis de um tratamento sobre a variável resposta.

8.2 Situação Problema

O artigo "Compression of Single-Wall Corrugated Shipping Containers Using Fixed and Floating Test Platens"(J. Testing and Evaluation, 1992, p.318-320) relata um experimento em que se comparou a resistência à compressão (lb) de vários tipos diferentes de caixas.

- Nesse caso, são usadas quatro níveis para o Tipo de Caixa e seis réplicas. Caixas 1, 2, 3 e 4.
- Assim, temos um experimento com **um** fator, **quatro** níveis e **seis** réplicas.

O objetivo nesse caso é avaliar se a resistência média difere para as quatro caixas avaliadas, isto é, deseja-se verificar se a variável resposta apresenta diferença na presença dos diferentes níveis.

Uma das ferramentas, usadas para comparação entre médias, é o teste t-Student (Capítulo 3). Todavia, existe o problema de que deveríamos avaliar todas as possíveis combinações entre médias, o que acaba por superestimar a probabilidade de erro tipo I, tornando o teste T ineficiente.

Os dados do experimento podem ser observados na tabela a seguir.

Tabela 52 – Dados do Exemplo Motivacional

Caixa	**Resistência a Compressão**						**Média**
	1	2	3	4	5	6	$\overline{y}_{i.}$
1	655,5	788,3	734,3	721,4	679,1	699,4	713,00
2	789,2	772,5	786,9	686,1	732,1	774,8	756,93
3	737,1	639,0	696,3	671,7	717,2	727,1	698,07
4	535,1	628,7	542,4	559,0	586,9	520,0	562,02

8.3 Análise de Variância

É evidente que em diversas situações desejamos proceder na comparação de uma ou mais médias dos níveis dos tratamentos. Entretanto, além das causas atribuídas aos níveis do tratamento, é possível existir efeitos desconhecidos, ou não controlados.

A técnica de Análise de Variância, foi proposta em 1920 por Fisher, na aplicação de experimentos agrícolas. Ela particiona a variabilidade total dos dados em duas ou mais componentes, considerando causas conhecidas ou controláveis, assim como causas desconhecidas, como por exemplo erro ou resíduo.

8.3.1 ANOVA para um fator - Dados Balanceados

Considere um experimento completamente casualizado ou de um fator com níveis ou tratamentos em *n* réplicas. A disposição dos dados pode ser observada na Tabela 53 em que o experimento é balanceado, isto é, todos os tratamentos tem o mesmo número de réplicas.

Tabela 53 – Aleatorização dos Dados

Tratamento	**Réplicas**				**Total do Tratamento**	**Média**
	1	2	...	n	$y_{i.}$	$\overline{y}_{i.}$
1	y_{11}	y_{11}	...	y_{1n}	$y_{1.}$	$\overline{y}_{1.}$
2	y_{21}	y_{22}		y_{2n}	$y_{2.}$	$\overline{y}_{2.}$
⋮	⋮	⋮	y_{ij}	⋮	⋮	⋮
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a.}$	$\overline{y}_{a.}$
					$y_{..}$	$\overline{y}_{..}$

Após a coleta dos dados, é necessário escrever um modelo estatístico, o qual supõe-se que descreva bem os dados. Assim temos o modelo de efeitos

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{8.1}$$

$$\text{com } i = 1, \ldots, a \text{ e } j = 1, \ldots, n.$$

Sendo: y_{ij} é a variável resposta para o tratamento i no indivíduo j;

μ é a constante geral (média);

τ_i é o efeito de tratamento *i*;

ε_{ij} é o erro aleatório.

8.3.1.1 Pressupostos sobre o Modelo

São necessárias as suposições:

- O erro aleatório é independente e normalmente distribuído com média 0 e variância σ^2;
- A variância é constante nos níveis do fator;
- As observações são adequadamente descritas pelo modelo.

De um modo geral, temos que:

$$\varepsilon \sim N(0, \sigma^2)$$
$$y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

É possível encontrarmos dois tipos de modelos, de acordo com o tipo de efeitos, são eles:

- Modelos de efeitos fixos, onde os fatores são determinados pelo pesquisador;
- Modelos de efeitos aleatórios, onde os fatores são determinados por meio de amostragem aleatória.

De um modo geral, em situações em que deseja-se verificar se há diferença entre os tratamentos, o objetivo é testar a igualdade entre as médias dos tratamentos. Para isso, podemos testar as hipóteses:

$$\text{T:} \begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_a \\ \\ H_a : \mu_i \neq \mu_j \qquad\qquad \text{para algum} \quad i \neq j. \end{cases}$$

8.3.1.2 Decomposição da Soma de Quadrados

A técnica ANOVA decompõe a variabilidade total dos dados em componentes. Assim, temos que a soma de quadrados total é dada por:

$$\begin{aligned} SQ_T &= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})^2 && (8.2)\\ &= \sum_{i=1}^{a}\sum_{j=1}^{n}((y_{ij} - \overline{y}_{i.}) + (y_{i.} - \overline{y}_{..}))^2 && (8.3)\\ &= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2 + 2(y_{ij} - \overline{y}_{.j}(\overline{y}_{i.} - \overline{y}_{..}) + (\overline{y}_{i.} - \overline{y}_{..})^2 && (8.4)\\ &= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2 + 2(y_{ij} - \overline{y}_{.j}(\overline{y}_{i.} - \overline{y}_{..}) + \sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{y}_{i.} - \overline{y}_{..})^2 && (8.5)\\ &= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{y}_{i.} - \overline{y}_{..})^2 && (8.6)\\ &= SQ_{TRAT} + SQ_R && (8.7) \end{aligned}$$

A SQ_R é responsável pela variabilidade nos níveis do tratamento, e SQ_{TRAT}, pela variabilidade entre os níveis do tratamento

A partir dessa soma de quadrados temos:

- *N*=$a \times n$ observações totais, então SQ_{TRAT} tem *N-1* graus de liberdade;
- *a* níveis para o fator, logo SQ_{Trat} tem *a-1* graus de liberdade;
- Dentre tratamentos há *n* repetições fornecendo *n-1* graus de liberdade para estimar o erro. Mas, como se tem a tratamentos tem-se a(n-1) = an-a = *N-a* graus de liberdade para o erro.

8.3.1.3 Quadrados Médios

Dividindo-se as Somas de quadrados para os tratamentos e para o erro por seus respectivos graus de liberdade obtemos o que chamamos de **Quadrados Médios**.

$$QM_{Trat} = \frac{SQ_{Trat}}{a-1} \quad e \quad QM_{Erro} = \frac{SQ_{Erro}}{N-a} \tag{8.8}$$

O QM_{Erro} é um estimador de σ^2 e QM_{Trat} é estimador de σ^2 caso não haja diferença entre as médias dos tratamentos.

O **Teorema de Cochran** garante a independência das Somas de Quadrados, isto é, SQ_{Trat}/σ^2 e SQ_{Erro}/σ^2 independentemente distribuídos com distribuição qui-quadrado. Nesse sentido, se H_0 é verdadeira,

$$F_{calc} = \frac{\frac{SQ_{Trat}}{a-1}}{\frac{SQ_{Erro}}{N-a}} = \frac{QM_{Trat}}{QM_{Erro}} \sim F_{a-1;N-a} \tag{8.9}$$

Sob a hipótese alternativa o valor esperado do numerador do teste estatístico 8.9 é maior que o valor esperado do denominador, e a hipótese nula será rejeitada quando o valor do teste estatístico for muito maior. Consequentemente, rejeita-se H_0, quando

$$F_{calc} > F_{\alpha,a-1,N-a} \tag{8.10}$$

Sendo $F_{\alpha,a-1,N-a}$ o valor calculado para distribuição *F-Snedecor*.Para o cálculo manual da ANOVA, podemos utilizar as expressões do quadro a seguir:

Tabela 54 – Quadro ANOVA: Dados balanceados

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	$\frac{1}{n}\sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{N}$	a-1	$\frac{SQ_{Trat}}{a-1}$	$\frac{QM_{Trat}}{QM_{Erro}}$
Erro	$SQ_T - SQ_{Trat}$	N - a	$\frac{SQ_{Erro}}{N-a}$	
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N}$	N-1		

Dos termos das expressões da Tabela 54 temos que $y_{i.}^2$ representa a soma dos elementos da linha ao quadrado, $y_{..}^2$ representa o quadrado da soma de todos os dados; y_{ij}^2 representa o quadrado de cada observação do tratamento *i* e nível *j*.

Retomando o exemplo motivacional temos as hipóteses e modelo:

- Hipóteses

T: $\begin{cases} H_0 : \mu_1 = \ldots = \mu_4 \\ \\ H_a : \mu_k \neq \mu_m \qquad \text{com} \quad k \neq m. \end{cases}$

- Modelo Estatístico

$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
com $i = 1, 2, 3, 4$ e $j = 1, \ldots, 6$

Pelas expressões para cálculo das Somas de Quadrados, segue que:

$$SQ_{Trat} = \frac{4278^2 + 4541,6^2 + 4188,4^2 + 3372,1^2}{6} - \frac{16380,1^2}{24} \approx 127,375$$

$$SQ_{Total} = (655,5^2 + \ldots + 520^2) - \frac{16380,1^2}{24} \approx 161,214$$

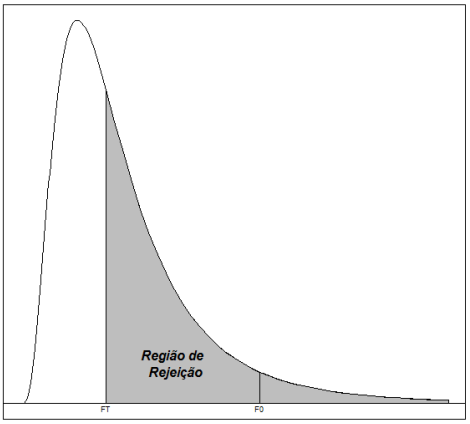
$$SQ_{Erro} = SQ_{Total} - SQ_{Trat} = 33,839$$

O que implica na Tabela a seguir.

Tabela 55 – Quadro ANOVA

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	127,38	3	42,46	25,09
Erro	33,84	20	1,69	
Total	161,21	23		

A partir dos resultados obtidos, com auxílio da Figura ao lado, observa-se que $F_0 > F_{Tabelado}$, sendo $25,09 > 3,10$, então H_0 é rejeitada ao nível de 5% de significância. A resistência de compressão média real parece depender do tipo de caixa.



Vale lembrar que em todos os exemplos serão utilizados os comandos a seguir.

```
#Instalar apenas uma vez o pacote.
install.packages("agricolae")
#Ativar o pacote toda vez que for utilizar.
library(agricolae)
```

Para resolver no R, basta seguir os comandos a seguir.

```
rm(list=ls())
dat<-read.table(header=TRUE, text="
Caixa Resistencia
1       655.5
1       788.3
1       734.3
1       721.4
1       679.1
1       699.4
2       789.2
2       772.5
2       786.9
2       686.1
2       732.1
2       774.8
3       737.1
3       639.0
3       696.3
3       671.7
3       717.2
3       727.1
4       535.1
4       628.7
4       542.4
4       559.0
4       586.9
4       520.0")

fm<-aov(Resistencia~factor(Caixa),data=dat)
anova(fm)
```

Ao extrair o resultado pelo R, a tomada da decisão pode ser feita baseando-se na comparação entre α e o p-valor. O mesmo vale para os exemplos abaixo.

Exemplo 153 *Os dados a seguir provêm de um experimento de comparação do grau de resíduos em tecidos copolimerizados com três diferentes misturas de ácido metacrílico. Ao nível de 5% de significância, é possível afirmar que o grau de resíduo médio real não é o mesmo para as três misturas?*

Tabela 56 – Dados do Exemplo 153

Mistura							
	1	2	3	4	5	$y_{i.}$	$\overline{y}_{i.}$
1	0,56	1,12	0,80	1,07	0,94	4,59	0,918
2	0,72	0,69	0,87	0,78	0,91	3,97	0,794
3	0,62	1,08	1,07	0,99	0,93	4,69	0,938

- *Hipóteses*

T: $\begin{cases} H_0 : \mu_1 = \ldots = \mu_3 \\ H_a : \mu_k \neq \mu_m \qquad \textit{com} \quad k \neq m. \end{cases}$

- *Modelo Estatístico*

$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
com i $= 1, 2, 3$ *e j* $= 1, \ldots, 5$

Pelas expressões para cálculo das Somas de Quadrados, segue que:

$$SQ_{Trat} = \frac{4,59^2 + 3,97^2 + 4,69^2}{6} - \frac{13,25^2}{15} \approx 0,061$$

$$SQ_{Total} = (0.56^2 + \ldots + 0.93^2) - \frac{13.25^2}{15} \approx 0,261$$

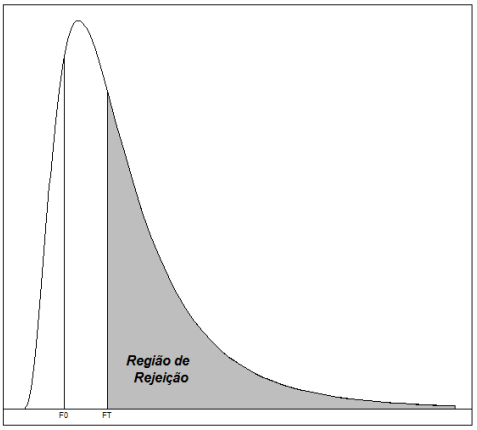
$$SQ_{Erro} = SQ_{Total} - SQ_{Trat} = 0,20$$

A partir dos cálculos anteriores é possível construir a tabela abaixo.

Tabela 57 – Quadro ANOVA

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	10,882	2	0,031	1,824
Erro	0,20	12	0,017	
Total	0,261	14		

A partir dos resultados obtidos, com auxílio da Figura ao lado, observa-se que $F_0 < F_{Tabelado}$*, sendo* $3,89 > 1,82$*,* H_0 *não é rejeitada ao nível de 5% de significância. É possível afirmar que o grau de resíduo médio real não é o mesmo para as três misturas*



Para resolver no R, basta seguir os comandos a seguir.

```
rm(list=ls())
dat<-read.table(header=TRUE, text="
Mistura Tecido
1 0.56
1 1.12
1 0.80
1 1.07
1 0.94
2 0.72
2 0.69
2 0.87
2 0.78
2 0.91
3 0.62
3 1.08
3 1.07
3 0.99
3 0.93")

fm<-aov(tecido~factor(mistura),data=dat)
anova(fm)
```

Exemplo 154 *Um engenheiro civil está interessado em determinar se quatro diferentes métodos de estimar a frequência do fluxo de inundação produzem equivalentes estimativas de pico de descarga, quando aplicado à mesma bacia hidrográfica. Cada procedimento é usado 6 vezes na bacia hidrográfica, e o resultado dos dados de descarga, em pés ao cubo por segundo são apresentados na Tabela abaixo. O que podemos concluir com* $\alpha = 5\%$*?*

Tabela 58 – Dados do Exemplo 154

Método	**Repetições**						
	1	2	3	4	5	6	$\overline{y}_{i.}$
1	0,34	0,12	1,23	0,70	1,75	0,12	4,26
2	0,91	2,94	2,14	2,36	2,86	4,55	15,76
3	6,31	8,37	9,75	6,09	9,82	7,24	47,58
4	17,15	11,82	10,95	17,20	14,35	16,82	88,29

- *Hipóteses*

T: $\begin{cases} H_0 : \mu_1 = \ldots = \mu_4 \\ H_a : \mu_k \neq \mu_m \qquad \text{com} \quad k \neq m. \end{cases}$

- *Modelo Estatístico*

$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
com $i = 1, \ldots, 4$ *e* $j = 1, \ldots, 6$

Pelas expressões para cálculo das Somas de Quadrados, segue que:

$$SQ_{Trat} = \frac{4,26^2 + 15,76^2 + 47,58^2 + 88,29^2}{6} - \frac{155,89^2}{24} \approx 708,347$$

$$SQ_{Total} = (0,34^2 + \ldots + 16,82^2) - \frac{155,8^2}{24} \approx 770,43$$

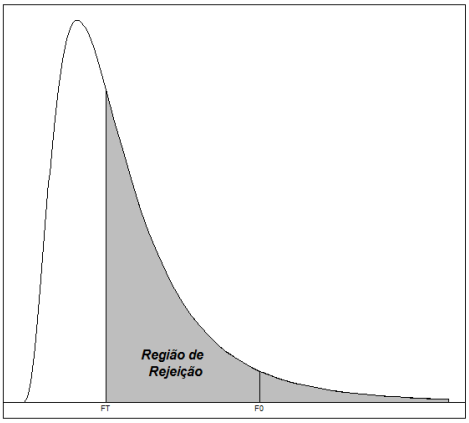
$$SQ_{Erro} = SQ_{Total} - SQ_{Trat} = 62,08$$

A partir dos cálculos anteriores é possível construir a tabela abaixo.

Tabela 59 – Quadro ANOVA

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	708,35	3	236,12	76,07
Erro	62,08	20	3,10	
Total	770,43	23		

A partir dos resultados obtidos, com auxílio da Figura ao lado, observa-se que $F_0 > F_{Tabelado}$*, sendo* $76,07 > 3,10$*, então* H_0 *é rejeitada ao nível de 5% de significância. A resistência de compressão média real parece depender do tipo de caixa.*



Para resolver no R, basta seguir os comandos a seguir.

```
rm(list=ls())
dat<-read.table(header=TRUE, text="
Metodo Bacias
1      0.34
1      0.12
1      1.23
1      0.70
1      1.75
1      0.12
```

```
2        0.91
2        2.94
2        2.14
2        2.36
2        2.86
2        4.55
3        6.31
3        8.37
3        9.75
3        6.09
3        9.82
3        7.24
4        17.15
4        11.82
4        10.95
4        17.20
4        14.35
4        16.82")

fm<-aov(Bacia~factor(Metodo),data=dat)
anova(fm)
```

8.3.2 ANOVA para um fator - Dados desbalanceados

Experimentos desbalanceados, em um planejamento inteiramente casualizado, são aqueles em que o número de repetições dentro de cada tratamento é diferente. Nesse caso, na ANOVA, é feito uma ligeira modificação na Soma de Quadrados.

Seja n_i observações em cada tratamento sendo i$= 1, 2, \ldots, a$ e $N = \sum_{i=1}^{a} n_i$. Então as equações para o cálculo das SQ_T e SQ_{TRAT} são dadas por:

$$SQ_T = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N} \text{ e } SQ_{Trat} = \sum_{i=1}^{a}\sum_{j=1}^{n} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

Exemplo 155 [1] *Segundo a revista Inovação, até 2020 os números de formandos em engenharia civil no Brasil devem chegar em torno de 80 à 107 mil por ano, enquanto o número de vagas que o mercado tende a oferecer deverá chegar a cerca de 60 mil. Esses números levam a termos em que 2,46 vezes mais engenheiros se formarão do que o mercado pode*

[1] Dados retirados dos seminários apresentados na UTFPR-PB

absorver. Mas por que essa enorme procura por engenheiros? O salário é altamente atrativo em algumas de suas ramificações, e é justamente esse o assunto de nosso trabalho. Quanto ganha, em média, os engenheiros civis por estados e regiões do Brasil Foi pesquisado o salário médio de admissão de engenheiros civis em todos os estados brasileiros. Separados em tabelas e calculados por suas respectivas regiões geográficas pesquisamos se as médias entre elas diferem ou não. Em uma tabela, seguindo a ordem dos estados antes apresentadas por região, temos os seguintes valores:

Tabela 60 – Dados do Exemplo 155

Região	**Estados**									
	Est.1	Est.2	Est.3	Est.4	Est.5	Est.6	Est.7	Est.8	Est.9	$\overline{y}_{i.}$
1	7.228	6.938	5.228	6.801	5.627	5.119	3.797			40.738
2	6.748	6.501	6.348	6.594	6.221	5.660	5.575	5.273	4.849	53.769
3	7.288	5.639	5.852	7.099						25.878
4	7.218	8.807	6.164	6.339						28.528
5	6.327	5.784	5.423							17.534

- *Hipóteses*

T: $\begin{cases} H_0 : \mu_1 = \ldots = \mu_5 \\ \\ H_a : \mu_k \neq \mu_m \qquad \textit{com} \quad k \neq m. \end{cases}$

- *Modelo Estatístico*

$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
com $i = 1, \ldots, 5$ *e* $j = 1, \ldots, 9$

Pelas expressões para cálculo das Somas de Quadrados, segue que:

$$SQ_{Trat} = \frac{40738^2}{7} + \frac{53769^2}{9} + \frac{25878^2}{4} + \frac{17534^2}{3} + \frac{28528^2}{4} - \frac{166447^2}{27} \approx 5580814.90$$

$$SQ_{Total} = (7228^2 + \ldots + 6339^2) - \frac{166447^2}{27} \approx 25154390$$

$$SQ_{Erro} = SQ_{Total} - SQ_{Trat} = 19573575.10$$

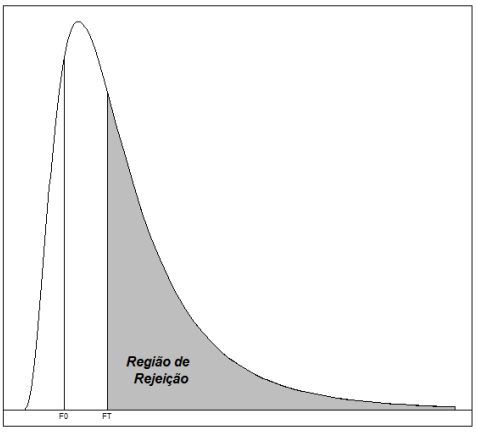
O que implica na tabela a seguir:

Tabela 61 – Quadro ANOVA

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	5.580.814,90	4	1.395.203,735	1,568
Erro	19.973.575,10	22	889707,9591	
Total	25.154.390	26		

Na tabela de F, procuramos $F_{(\alpha, a-1, N-a)}$, sendo a o número de níveis, N - a Número total de observações menos o número de tratamentos. Assim, temos que $F(0.05, 4.22, 2.8167)$

Comparando F_0 calculado e F tabelado, temos que $F_0 < F_{tabelado}$B $(1,568 < 2,8167)$. Como F_0 fica dentro do espaço crítico de F, não rejeitamos H_0 e concluímos que as médias não diferem entre si no nível de significância de 5%.



8.3.3 Comparações Múltiplas

Quando na ANOVA não rejeita-se H_0, a análise é finalizada porque não se identificou diferença estatística entre os tratamentos. Todavia, quando H_0 for rejeitada, deve-se conhecer quais das médias dos tratamentos diferem entre si. Nesse caso, deve-se efetuar um teste para comparações múltiplas. Um teste comumente utilizado é o Teste de Tukey.

8.3.3.1 Teste de Tukey

Dados balanceados

Duas médias são estatisticamente diferentes se a diferença das médias amostrais (em valor absoluto) for superior a DMS (Diferença Mínima Significativa):

$$DMS = \frac{q_{(\alpha,a,N-a)}}{\sqrt{2}}\sqrt{2 * \frac{QM_{Erro}}{n}} = q_{(\alpha,a,N-a)}\sqrt{\frac{QM_{Erro}}{n}} \tag{8.11}$$

Sendo:

$q_{(\alpha,a,N-a)}$ é um valor tabelado;

N-a são os graus de liberdade associados a estimativa $s^2(QM_{Erro})$ e a o número de tramamentos

Caso $|\overline{y}_{i.} - \overline{y}_{j.}| > DMS$, as médias comparadas diferem entre si, ao nível de significância

Dados desbalanceados

No caso de dados desbalanceados:

$$DMS = \frac{q_{(\alpha,a,N-a)}}{\sqrt{2}}\sqrt{QM_{Erro}(\frac{1}{n_i} + \frac{1}{n_j})} \tag{8.12}$$

sendon_i é o número de observações do nível *i* e n_j é o número de observações do nível *j*. Como para i $\neq$ j, $n_i \neq n_j$, tem-se um DMS para cada comparação. A conclusão para as comparações, são análogas no caso para dados balanceados.

Para fazer as comparações múltiplas, basta seguir os passos:

- Calcular as médias dos tratamentos;
- Ordenar as médias calculadas;
- Calcular a estatística de Tukey;
- Comparar as diferenças das médias com o valor DMS;

Exemplo 156 *Como no exemplo motivacional rejeitou-se H_0, realiza-se o Teste de Tukey, Seguindo os passos descritos anteriormente:*

- *Calcular as médias dos tratamentos;*

  $\overline{y}_{1.}$ *= 713,* $\overline{y}_{2.}$ *= 756,93,* $\overline{y}_{3.}$ *= 698,07 e* $\overline{y}_{4.}$ *= 562,02 .*

- *Ordenar as médias calculadas;*

  – $\overline{y}_{2.}$ *= 756,93,* $\overline{y}_{1.}$ *= 713,* $\overline{y}_{3.}$ *= 698,07 e* $\overline{y}_{4.}$ *= 562,02 .*

- *Calcular a estatística de Tukey;*

$$DMS = \frac{3,96}{\sqrt{2}}\sqrt{2 * \frac{1692}{6}} = 3,96\sqrt{\frac{1692}{6}} = 66,49$$

- *Comparar as diferenças das médias com o valor DMS;*

  – $|756,93 - 713| < 66,49$*;*

  – $|756,93 - 698,07| < 66,49$*;*

  – $|756,93 - 562,07| > 66,49$*;*

  – $|713 - 698,08| < 66,49$*;*

  – $|713 - 562,07| > 66,49$*;*

  – $|698,08 - 562,07| > 66,49$

- *Observando os resultados obtidos, pelo teste de Tukey, ao nível de 5% de significância, verifica-se que há evidência amostral que a caixa 4 apresenta menor resistência quando comparadas com as demais caixas. Para as demais caixas, as resistências não diferem significativamente.*

Para o exemplo motivacional, resolvendo o Teste de Tukey no R, temos os comandos a seguir:

```
out1 <-HSD.test(fm,'factor(Grupo)',alpha=0.05)
out1

$statistics
Mean        CV   MSerror       HSD
68.33333 33.72563 531.1111 34.56071

$parameters
Df ntr StudentizedRange alpha  test        name.t
15   3          3.673378  0.05 Tukey factor(Grupo)

$means
Baterias       std r Min Max
34 78.33333 17.22401 6  60 100
35 56.66667 31.41125 6   0  90
65 70.00000 17.60682 6  45  90

$comparison
NULL

$groups
trt     means M
1  34 78.33333 a
2  65 70.00000 a
3  35 56.66667 a
```

Para interpretar os resultados, verifica-se que no item $groups ao lado das médias há letras. Nesse sentido, letras iguais indicam que então há diferença entre os tratamentos. Letras diferentes, indicam que há diferença ao nível alfa, entre os tratamentos.

Observação: No que tange os cálculos para realizar o teste de Tukey no R, ele já identifica automaticamente se os dados são balanceados ou não e realiza os cálculos

8.4 Exercícios de Aplicação

1. [2]A amostra de 15 pregos de cabeça foram escolhidas de forma aleatória e de três fabricantes diferentes. Todas as peças foram medidas uma a uma. Este processo de medição foi repetido três vezes para melhor exatidão dos valores obtidos. O valor anotado do comprimento do prego era feito com base na média das três medições. Verificar se há diferença no comprimento dos pregos para as 3 marcas ao nível de 5%. Se existir realizar o teste de Tukey.

Tabela 62 – Dados Exercício 1

Pregos	**Repetições**					
	1	2	3	4	5	$y_{i.}$
1	8,531	8,529	8,533	8,532	8,534	42,659
2	8,523	8,528	8,526	8,525	8,529	42,631
3	8,536	8,535	8,537	8,538	8,534	42,68

2. Na tabela abaixo tem-se dados coletados a respeito da produção de energia provenientes de termelétricas no ano de 2014. O estudo consiste em analisar se houve ou não diferença estatisticamente significativas entre as médias das unidades de geração. A partir dessas informações, verificar, ao nível de 5% se há diferença na produção de energia entre elas

Tabela 63 – Dados Exercício 2

Usina	**Geração Bruta por Mês (GW.h)**					
	01	02	03	04	05	06
W. ARJONA	109,05	95,02	108,76	107,05	112,16	108.81
CHARQUEADAS	28,43	27,22	35,79	33,06	32,16	27,83
J. LACERDA	110,74	102,41	128,51	118,56	110,49	90.48
	07	08	09	010	11	12
W. ARJONA	113,00	109,60	98,23	89,65	81,62	92,97
CHARQUEADAS	31,96	25,79	25,81	37,68	15,61	31,22
J. LACERDA	102,03	84,71	81,89	104,97	10,46	106,16

3. No seguinte estudo de caso foram utilizados três instrumentos de medição para coletar medidas de dois blocos padrões diferentes, estes que são utilizados para fazer a calibração dos instrumentos de medida, pois possuem um erro de medida inferior a resolução do instrumento.

4. Três empresas de aviação oferecem voos entre Corydon e Lincolnville. Vários tempos de voo selecionados aleatoriamente (em minutos) entre as cidades para cada

[2] Dados coletados pelas alunas Isabela e Denise, para apresentação de seminário da disciplina de Probabilidade e Estatística, no segundo semestre de 2015.

Tabela 64 – Dados Exercício 3

Instrumento	**Iteração**									
	01	02	03	04	05	06	07	08	09	010
Paq. Anal.	3,01	3,01	3,10	2,90	2,99	3,00	2,99	3,05	3,03	3,02
Micro Anal.	2,99	3,09	3,09	2,995	3,01	3,00	3,001	3,005	3,00	2,995
Micro Dig.	3,001	3,002	3,00	3,01	3,001	3,004	3,002	3,001	2,983	3,00

empresa podem ser observados à direita. Suponha que as populações de tempo de voo sejam normalmente distribuídas, as amostras sejam independentes e as variâncias populacionais sejam iguais. Com $\alpha = 0.05$, você pode concluir que há uma diferença nas médias de tempos dos voos?

Tabela 65 – Dados Exercício 4

Empresa	**Tempos de voo**										
	01	02	03	04	05	06	07	08	09	010	$\overline{y}_{i.}$
1	122	135	126	131	125	116	120	108	142	113	1238
2	119	133	143	159	144	124	126	131	140	136	1355
3	120	158	155	126	147	164	134	151	131	141	1427

5. Os preços (em dólares) de 17 baterias de automóveis aleatoriamente selecionadas são exibidos na tabela. Os preços são classificados de acordo com o tipo e bateria. Com $\alpha = 0.05$, há evidência suficiente para concluir que no mínimo uma das médias dos preços de baterias é diferente das outras?

Tabela 66 – Dados Exercício 5

Grupo	**Baterias**						
	01	02	03	04	05	06	$\overline{y}_{i.}$
tam 35	60	60	90	50	80		340
tam 65	80	60	60	90	45	85	420
tam 34/78	60	100	90	70	90	60	470

8.4.1 Gabarito

1. Quadro ANOVA

Tabela 67 – Quadro ANOVA - Exercício 1

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	0,00024173	2	0,0001208	30,471
Erro	0,00004760	12	0,000003967	
Total	0,0002893	14		

Nível de significância 5% para a ANOVA, rejeita-se a hipótese de que as médias são iguais. Pelo teste de Tukey, com 5% de significância, as três marcas apresentam diferenças entre si, de acordo com a variável resposta.

Para resolver no R, basta seguir os comandos a seguir.

```
rm(list=ls())
dat<-read.table(header=TRUE, text="
Pregos Repeticoes
1        8.531
1        8.529
1        8.533
1        8.532
1        8.534
2        8.523
2        8.528
2        8.526
2        8.525
2        8.529
3        8.536
3        8.535
3        8.537
3        8.538
3        8.534")

fm<-aov(Repeticoes~factor(Pregos),data=dat)
anova(fm)

#Comparacoes multiplas
#Teste de Tukey
out1 <-HSD.test(fm,'factor(Pregos)',alpha=0.05)
out1
```

2. Quadro Anova

Tabela 68 – Quadro ANOVA - Exercício 2

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	43.588,28	2	21.794,14	205,32
Erro	3.502,94	33	106,15	
Total	47.091,22	35		

Nível de significância 5% para a ANOVA, rejeit-se a a hipótese de que as médias são iguais. Pelo teste de Tukey, com 5% de significância, existem evidencias que a ute charqueadas possui uma geração media que difere das demais. A ute William Arjona e Jorge Lacerda não possuem uma diferença significativa.

Para resolver no R, basta seguir os comandos a seguir.

```
rm(list=ls())
dat<-read.table(header=TRUE, text="
Usina Geracao
1     109.05
1     95.02
1     108.76
1     107.05
1     112.16
1     108.81
1     113.00
1     109.60
1     98.23
1     89.65
1     81.62
1     92.97
2     28.43
2     27.22
2     35.79
2     33.06
2     32.16
2     27.83
2     31.96
2     25.79
2     25.81
2     37.68
2     15.61
2     31.22
```

```
3      110.74
3      102.41
3      128.51
3      118.56
3      110.49
3      90.48
3      102.03
3      84.71
3      81.89
3      104.97
3      108.46
3      106.16")

fm<-aov(Geracao~factor(Usina),data=dat)
anova(fm)

#Comparacoes multiplas
#Teste de Tukey
out1 <-HSD.test(fm,'factor(Usina)',alpha=0.05)
out1
```

3. Quadro Anova

Tabela 69 – Quadro ANOVA - Exercício 3

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	0,00148	2	0,00074293	0,5422
Erro	0,03699	27	0,0013701	
Total	0,03847	29		

Nível de significância 5% para a ANOVA, não rejeita-se a hipótese de que as médias são iguais. Para resolver no R, basta seguir os comandos a seguir.

```
rm(list=ls())
dat<-read.table(header=TRUE, text="
Instrumento Iteracao
1           3.01
1           3.01
1           3.10
1           2.90
1           2.99
```

```
1               3.00
1               2.99
1               3.05
1               3.03
2               2.99
2               3.09
2               3.09
2               2.995
2               3.01
2               3.00
2               3.001
2               3.005
2               3.00
3               3.001
3               3.002
3               3.00
3               3.01
3               3.001
3               3.004
3               3.002
3               3.001
3               2.983")

fm<-aov(Iteracao~factor(Instrumento),data=dat)\\
anova(fm)\\

#Comparacoes multiplas
#Teste de Tukey
out1 <-HSD.test(fm,'factor(Instrumento)',alpha
   =0.05)
out1
```

4. Quado Anova

Tabela 70 – Quadro ANOVA - Exercício 4

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	1.819,8	2	909,90	5,9655
Erro	4.118,2	27	152,53	
Total	5.938,0	29		

Ao nível de significância 5% para a ANOVA, rejeita-se a hipótese de que as médias

são iguais. Pelo teste de Tukey, com 5% de significância, existem evidencias que as empresas 1 e 2, bem como 2 e 3 não apresentam diferença na variável resposta. Contudo, há diferença para as empresas 1 e 3.

Para resolver no R, basta seguir os comandos a seguir.

```
rm(list=ls())
dat<-read.table(header=TRUE, text="
Empresa Voos
1       122
1       135
1       126
1       131
1       125
1       116
1       120
1       108
1       142
1       113
2       119
2       133
2       143
2       159
2       144
2       124
2       126
2       131
2       140
2       136
3       120
3       158
3       155
3       126
3       147
3       164
3       134
3       151
3       131
3       141")
```

```
fm<-aov(Voos~factor(Empresa),data=dat)
anova(fm)

#Comparacoes multiplas
#Teste de Tukey
out1 <-HSD.test(fm,'factor(Empresa)',alpha=0.05)
out1
```

5. Quadro ANOVA

Tabela 71 – Quadro ANOVA - Exercício 5

C_v	**SQ**	**GL**	**QM**	F_0
Tratamento	1.433,3	5	716,67	1,3494
Erro	7.966,7	12	531,11	
Total	9.400,0	17		

Ao nível de significância 5% para a ANOVA, NÃO rejeita-se a hipótese de que as médias são iguais.

Para resolver no R, basta seguir os comandos a seguir.

```
rm(list=ls())
dat<-read.table(header=TRUE, text="
Grupo Baterias
35       60
35       60
35       90
35       50
35       80
35       0
65       80
65       60
65       60
65       90
65       45
65       85
34       60
34       100
34       90
34       70
34       90
34       60 ")
```

```
fm<-aov(Baterias~factor(Grupo),data=dat)
anova(fm)

out1 <-HSD.test(fm,'factor(Grupo)',alpha=0.05)
out1
```

Capítulo 9

Regressão Linear

9.1 Introdução

A modelagem de dados através de modelos de Regressão, tornou-se objetivo de estudo de diversos pesquisadores, sejam eles da área teórica ou aplicada. Todavia, esse trabalho pode tornar-se complexo e estender-se por um longo período.

Em análise de Regressão, é possível encontrar diversos modelos. É possível destacar os modelos de Regressão Linear Simples e múltipla, Regressão Logística e os modelos Lineares generalizados. O estudo da Análise de Regressão pode ser realizado com o auxílio de diversos softwares, entre os quais podemos destacar o R e SAS.

A análise de Regressão Linear Simples será objeto de estudo das próximas seções.

9.2 Situação Problema

Considere o experimento onde foi avaliado a taxa de germinação (%) de sementes de certo cultivar de Soja, em relação a umidade Relativa do Ar. Os dados podem ser observados na 72.

Tabela 72 – Dados do problema

UR(%)	**GER(%)**
20	94
30	96
40	95
50	97

Em geral, na maioria das vezes não é possível identificar a relação entre as variá-

veis do problema sem uma análise gráfica. Assim, de modo a visualizar a relação desejada é necessário a construção de um gráfico de dispersão. A figura 56 apresenta a dispersão dos dados da 72

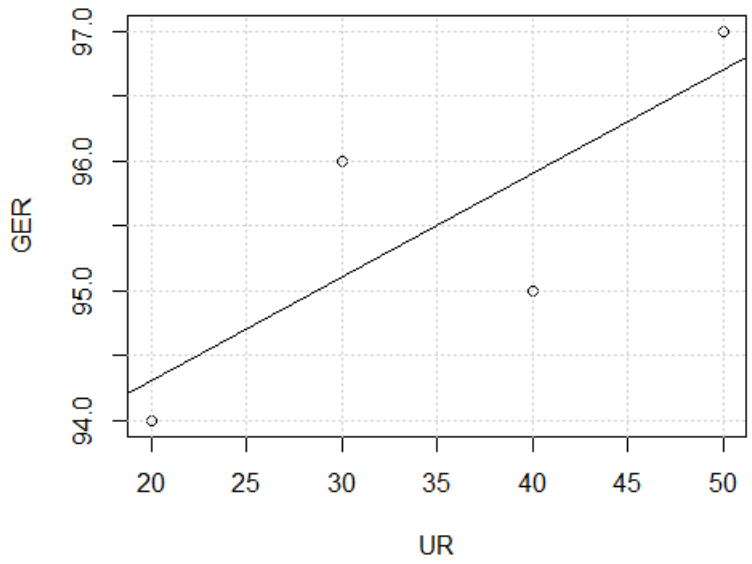


Figura 56 – Gráfico da dispersão de dados

9.3 Correlação

Um modelo de Regressão é usado para expressar a relação entre duas ou mais variáveis. Assim, uma das maneiras de quantificar esta relação, é através do uso do **coeficiente de correlação (r ou ρ)**, o qual pode ser usada para verificar se existe uma relação entre as variáveis.

A Figura 57 descreve/mostra o sentido e a força da correlação

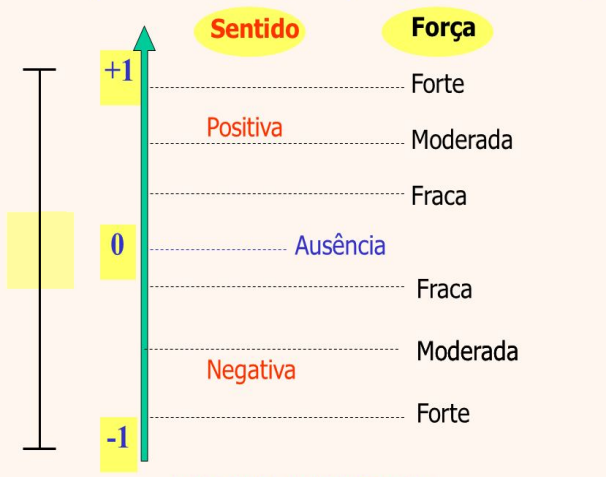


Figura 57 – Força de correlação

O coeficiente de correlação pode ser obtido por meio da expressão

$$\rho = \frac{cov(x,y)}{\sqrt{var(x).var(y)}} = \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\sum_{i=1}^{n} y_i}{n}\right)}} \quad (9.1)$$

sendo cov(x,y) a covariância[1] entre x e y, var(x) e var(y) as variâncias de x e y.

Exemplo 157 *Calcular a correlação entre as variáveis UR e GER, usando a expressão acima.*

Para resolver no R, basta utilizar os comandos a seguir.

```
dados<-read.table(header=TRUE, text="
UR GER
20 94
30 96
40 95
50 97")
cor<-cor(dados$UR,dados$GER)
cor
```

[1] A covariância ou variância conjunta é um momento conjunto de primeira ordem das variáveis aleatórias X e Y, centrados nas respectivas médias. É a média do grau de interdependência ou inter-relação numérica linear entre elas.

9.4 Regressão Linear Simples

O modelo de regressão mais simples é aquele que relaciona apenas duas variáveis e pode ser chamado de regressão linear simples. Considerando a situação problema apresentado na Seção 2 relacionou as variáveis x e y de modo linear. Nesse sentido, para descrever essa relação podemos determinar um **Modelo de Regressão Linear**.

Definição 1: Sejam $(x_1, y_1), \ldots, (x_n, y_n)$ pares de dados observados para as variáveis X e Y. Se for admitido que Y é uma função linear de X, pode-se estabelecer um modelo de regressão linear, o qual é apresentado na expressão

$$y_i \quad = \quad \beta_0 + \beta_1 x_i + \varepsilon_i \tag{9.2}$$

sendo x_i é a variável regressora e Y é a variável resposta. Nesse modelo, temos que y_i é a variável resposta; x_i é a i-ésima observação da variável regressora; ε é o erro aleatório;β_0 e β_1 são os parâmetros do modelo.

É possível destacar que $E(y_i) = \beta_0 + \beta_1 x_i$ é uma função da média e $Var(y_i) = \sigma_\varepsilon^2$ é uma função da variância, a qual é uma parte aleatória.

9.4.1 Pressupostos sobre o modelo

Por meio da definição 1, segue que:

- X é a variável que pode ser controlada;
- A relação entre X e Y é linear;
- A média do erro é nula, ou seja, $E(y_i|x_i) = 0$;
- Os erros tem variabilidade constante em torno de x, $Var(\varepsilon_i|x_i) = \sigma_\varepsilon^2$, $\forall i = 1, \ldots, n$;
- Os erros são observações independentes $\varepsilon_i \sim N(0, \sigma^2)$;
- $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ e independentes.

A suposição de normalidade é fundamental para o processo de inferência sobre o modelo.

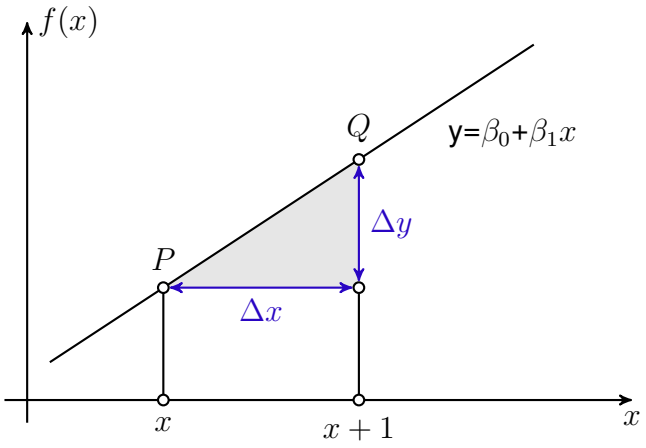
9.4.2 Estimação dos Parâmetros do modelo

Considere o modelo descrito na definição 1.

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i, \forall i = 1, \ldots, n \tag{9.3}$$

Se X=0, então β_0 representa o ponto onde a reta intercepta o eixo y e é chamado de Intercepto do modelo ou Coeficiente linear da reta. O parâmetro β_1 é chamado de coeficiente de regressão ou coeficiente angular da reta.

A Figura abaixo mostra a representação gráfica do significado dos parâmetros



Utilizando os conceitos de CDI é fácil ver que $\beta_1 = \frac{\Delta x}{\Delta y}$. Já o valor de β_0 corresponde ao ponto de interseção da reta de regressão com o eixo y. Se usarmos os valores estimados da variável resposta para uma observação X_i, temos que

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{E}(y_i|x_i), \tag{9.4}$$

sendo $\hat{\beta}_0$ e $\hat{\beta}_1$ são as estimativas de β_0 e β_1, respectivamente.

Os erros são dados por:

$$\hat{\varepsilon_i} = y_i - \hat{y_i}, \forall i = 1, 2, \ldots, n \tag{9.5}$$

os quais quantificam a diferença entre os valores observados na pesquisa (y_i) e os valores preditos pelo modelo ($\hat{y_i}$).

9.4.3 Interpretação dos Parâmetros do modelo

Diversos são os métodos de estimação dos parâmetros de um modelo de regressão encontrados na literatura, os quais fornecem uma estimativa para os parâmetros do modelo. Podemos destacar o método da Máxima Verosimilhança e o Método dos Mínimos Quadrados com suas variações.

A escolha de um método ou outro depende do problema analisado. Para estimar β_0 e β_1 usaremos o Método dos Mínimos Quadrados Ordinários (MQO).

9.4.4 Método dos MQO para estimação de β_0 e β_1

Considere o modelo

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \forall i = 1, \ldots, n. \tag{9.6}$$

O método MQO estima β_0 e β_1 de modo que os desvios entre os valores observados e preditos seja mínimo.

Sejam $X^T = (x_1, \ldots, x_n)$ e $Y^T = (y_1, \ldots, y_n)$ os vetores correspondentes as variáveis explicativas e resposta respectivamente. É necessário minimizar a soma dos desvios L, sendo:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2 \tag{9.7}$$

Considere o sistema das derivadas parciais abaixo.

$$U(X|\beta_0, \beta_1) \begin{cases} \dfrac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = -2 \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1) \\ \dfrac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = -2 \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1) x_i \end{cases} \tag{9.8}$$

Para obter $\hat{\beta}_0$ e $\hat{\beta}_1$ é necessário resolver o sistema de equações. De fato temos.

$$-2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1) = 0$$

$$\sum_{i=1}^{n}\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \tag{9.9}$$

$$n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i \tag{9.10}$$

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i - \hat{\beta}_1 \frac{1}{n}\sum x_i \tag{9.11}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{9.12}$$

Onde $\overline{x}$ e $\overline{y}$ são a média das variâncias x e y. Para obter $\hat{\beta}_1$, basta substituir (1.13) em (1.8).

$$\sum_{i=1}^{n}\hat{\beta}_1 {x_i}^2 = \sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i - (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i \tag{9.13}$$

$$\sum_{i=1}^{n}\hat{\beta}_1 {x_i}^2 = \sum_{i=1}^{n} y_i x_i - \overline{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i \tag{9.14}$$

$$\hat{\beta}_1 (\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2) = \sum_{i=1}^{n} y_i x_i - n\overline{yx} \tag{9.15}$$

Portanto segue que os valores estimados de β_0 e β_1 são dados por:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\overline{xy}}{\sum {x_i}^2 - n\overline{x}^2} \tag{9.16}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{9.17}$$

$$\tag{9.18}$$

É válido lembrar que após a estimação dos parâmetros do modelo, é necessário, verificar por meio de uma análise de Resíduos se o modelo está bem ajustado. Todavia não abordaremos nesse material.

Retomando o Exemplo Motivacional com uso dos termos que fornecem os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$, constrói-se a Tabela 73.

Tabela 73 – Tabela auxiliar para cálculo dos coeficientes do modelo linear

X	Y	X^2	XY
20	94	400	1.880
30	96	900	2.880
40	95	1.600	3.800
50	97	2.500	4.850
140	382	5.400	13.410

Assim, temos $\hat{\beta}_0$ e $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{13.410 - \frac{(140)(382)}{4}}{5.400 - \frac{140^2}{4}} = \frac{13.410 - 13.370}{5.400 - 4900} \approx 0,08 \tag{9.19}$$

$$\hat{\beta}_0 = \frac{382}{4} - 0,08\frac{140}{4} \approx 92,7 \tag{9.20}$$

O modelo ajustado é dado por:

$$\hat{y} = 92,7 + 0,08x_i, \qquad x_i = 20, 30, 40, 50.$$

Uma vez ajustado o modelo de regressão é possível obter os valores preditos pelo modelo, basta substituir x_i por 20, 30, 40 e 50. Nesse caso temos:

$$\hat{y} = 92,7 + 0,08 \times 20 = 94,30$$
$$\hat{y} = 92,7 + 0,08 \times 30 = 95,10$$
$$\hat{y} = 92,7 + 0,08 \times 40 = 95,90$$
$$\hat{y} = 92,7 + 0,08 \times 50 = 96,70$$

sendo os valores obtidos os valores preditos pelo modelo ajustado. O gráfico dos valores preditos pelos valores observados é apresentado a seguir.

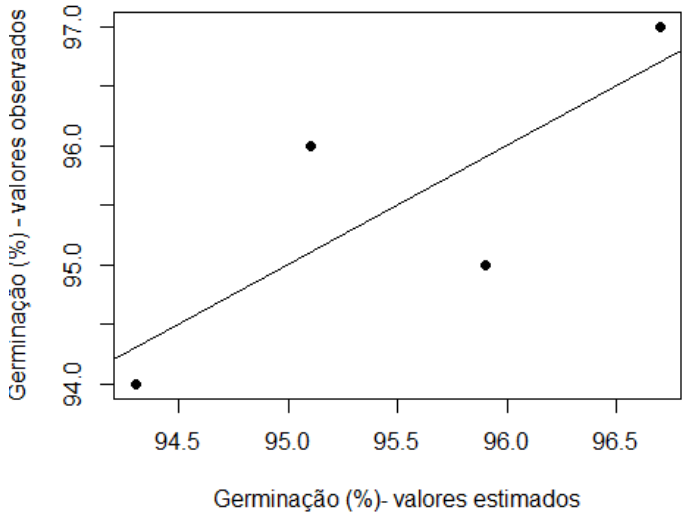


Figura 58 – Gráfico dos valores preditos X valores observados

A partir dos valores preditos, é possível obter os valores para ϵ_i, considerando a expressão 9.5. Quando observamos a Figura 58, a distância entre cada ponto e a reta ajustada, é o valor para seu respectivo ϵ_i.

Para obter os valores estimados dos coeficientes do modelo de regressão, bem como valores preditos pelo modelo, basta utilizar os comandos a seguir.

```
dados <-read.table(header=TRUE , text ="
UR GER
20 94
30 96
40 95
50 97")

reg <-lm(GER~UR,data=dados)
summary(reg)

             Estimate    Std. Error t value Pr(>|t|)
(Intercept) 92.70000     1.55885   59.467 0.000283 ***
UR           0.08000     0.04243    1.886 0.200000
---
predict(reg)
  1    2    3    4
  94.3 95.1 95.9 96.7
```

Nos resultados acima, o valor relativo a *Intercept*, representa $\hat{\beta}_0$ e UR representa o $\hat{\beta}_1$, ambos apresentados na coluna *Estimate*.

Exemplo 158 *O concreto sem finos, fabricado com um agregado rústico nivelado de maneira uniforme e uma pasta de cimento-água, é benéfico em áreas propensas a muita chuva por causa de suas excelentes propriedades de drenagem. O artigo "Pavement Thickness Design for No-Fines Concrete Parking Lots"empregou uma análise de mínimos quadrados ao estudar como y= porosidade (%) está relacionada com x= peso unitário (pcf) em amostras de concreto. Considere os dados representativos a seguir, exibidos em um formato tabular conveniente para calcular os valores das estatísticas:*

Com base nos cálculos acima, é possível obter $\overline{x}$= 109,34, $\overline{y}$=19,97. Assim, temos:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{32.308,59 - (1.640,1)(299,8)/15}{179.849,73 - (1.640,1)^2/15} \approx -0,905$$
$$\hat{\beta}_0 = 19,97 - (-0,905)(109,34) = 118,91 \approx 118,91$$

Logo o modelo ajustado, ou equação para regressão linear simples é dada por:

Tabela 74 – Dados do exemplo

Obs i	x	y	x^2	xy	y^2
1	99,0	28,8	9.801,00	2.851,20	829,44
2	101,1	27,9	10.221,21	2.820,69	778,41
3	102,7	27,0	10.547,29	2.772,90	729,00
4	103,0	25,2	10.609,00	2.595,60	635,04
5	105,4	22,8	11.109,16	2.403,12	519,84
6	107,0	21,5	11.449,00	2.300,50	462,25
7	108,7	20,9	11.815,69	2.271,83	436,81
8	110,8	19,6	12.276,64	2.171,68	384,16
9	112,1	17,1	12.566,41	1.916,91	292,41
10	112,4	18,9	12.633,76	2.124,36	357,21
11	113,6	16,0	12.904,96	1.817,60	256,00
12	113,8	16,7	12.950,44	1.900,46	278,89
13	115,1	13,0	13.248,01	1.496,30	169,00
14	115,4	13,6	13.317,16	1.569,44	184,96
15	120,0	10,8	14.400,00	1,296,00	116,64
Soma	1.640	299,8	179.849,73	32.308,59	6.430,06

$$\hat{y} = 118,91 - 0,905x_i \quad com \quad i = 1, \ldots, 15$$

O gráfico dos valores preditos pelos valores observados é apresentado a seguir

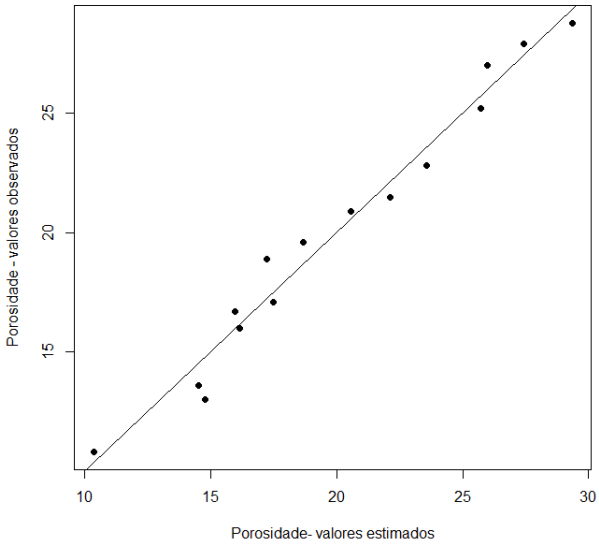


Figura 59 – Gráfico dos valores preditos X valores observados

A partir dos valores preditos, é possível obter os valores para ϵ_i, considerando a

expressão 9.5. Quando observamos a Figura 58, a distância entre cada ponto e a reta ajustada, é o valor para seu respectivo ϵ_i.

Para obter os valores estimados dos coeficientes do modelo de regressão, bem como valores preditos pelo modelo, e o gráfico dos valores preditos pelos observados basta utilizar os comandos a seguir.

```
x<- c(99,101.1,102.7,103,105.4,107,108.7,110.8,112.1,112.4,113.6,
    113.8,115.1,115.4,120)
y<- c(28.8,27.9,27,25.2,22.8,21.5,20.9,19.6,17.1,18.9,16,16.7,13,
    13.6,10.8)

cor(x,y)
dados<-data.frame(x,y)

fm <- lm(y ~ x, data = dados)
summary(fm)

predict(fm)

#Predito pelo observado
x11()
Predito1<-predict(fm)
Observado<-dados$y
plot(Predito1,Observado,lty=1,cex=1.5, pch=20,
ylab='Porosidade - valores observados',
xlab="Porosidade- valores estimados")
abline(a=0,b=1)
```

9.5 Exercícios de Aplicação

1. O dados a seguir sobre x (densidade da corrente [mA/$(cm)^2$]) e y (taxa de deposição [μm/min]) foram tirados do artigo "Plating of 60/40 Tin/Lead Solder for Head termination Metallurgy".

Tabela 75 – Dados do exercício 1

x	20	40	60	80
y	0,24	1,20	1,71	2,22

a) Determinar a correlação entre as variáveis;

b) Determinar o modelo ajustado;

c) Obter os valores preditos a partir do modelo ajustado;

d) Construir o gráfico dos valores estimados pelos valores observados;

2. O artigo "Characterization of Highway Runoff in Austin, Texas, Area"apresentou um gráfico de dispersão com a reta dos mínimos quadrados de x (volume de chuva [$(m)^3$]) y (volume de runoff [$(m)^3$]). Os valores foram reproduzidos do gráfico.

Tabela 76 – Dados do exercício 2

x	55	67	72	81	96	112	127
y	38	46	53	70	82	99	100

a) Determinar a correlação entre as variáveis;

b) Determinar o modelo ajustado;

c) Obter os valores preditos a partir do modelo ajustado;

d) Construir o gráfico dos valores estimados pelos valores observados;

3. Inúmeros estudos já mostraram que os líquens são excelentes bioindicadores de poluição do ar. O artigo "The Epiphytic Lichen Hypogymnia Physodes as a Biomonitor of Atmospheric Nitrogen and Sulphur Deposition in Norway"apresenta os dados sobre x = deposição de umidade de NO_3 (g N/m^2) e y = líquens N ($\%$ peso seco).

Tabela 77 – Dados do exercício 3

x	0,05	0,10	0,11	0,12	0,31	0,37	0,42
y	0,48	0,55	0,48	0,50	0,58	0,52	1,02

x	0,58	0,68	0,68	0,73	0,85	0,92
y	0,86	0,86	1,00	0,88	1,04	1,70

a) Determinar a correlação entre as variáveis;

b) Determinar o modelo ajustado;

c) Obter os valores preditos a partir do modelo ajustado;

d) Construir o gráfico dos valores estimados pelos valores observados;

4. O artigo "Effects of Bike Lanes on Driver and Bicyclist Behavior"relata os resultados de uma análise de regressão com x = espaço percorrido disponível em pés e a distância de separação y entre uma bicicleta e um carro em ultrapassagem.

Tabela 78 – Dados do exercício 4

x	12,8	12,9	12,9	13,6	14,5
y	5,5	6,2	6,3	7,0	7,8

x	14,6	15,1	17,5	19,5	20,8
y	8,3	7,1	10,0	10,8	11,0

a) Determinar a correlação entre as variáveis;

b) Determinar o modelo ajustado;

c) Obter os valores preditos a partir do modelo ajustado;

d) Construir o gráfico dos valores estimados pelos valores observados;

5. Os dados a seguir foram reproduzidos de um gráfico apresentado no artigo "Reactions on Painted Steel Under the Influence of Sodium Chloride, and Combinations Thereof". A variável independente é a taxa de deposição (mg/m^2/d) de SO_2 e a variável dependente é a perda de peso do aço (g/m^2).

Tabela 79 – Dados do exercício 5

x	14	18	40	41	45	112
y	280	300	470	500	560	1200

a) Determinar a correlação entre as variáveis;

b) Determinar o modelo ajustado;

c) Obter os valores preditos a partir do modelo ajustado;

d) Construir o gráfico dos valores estimados pelos valores observados;

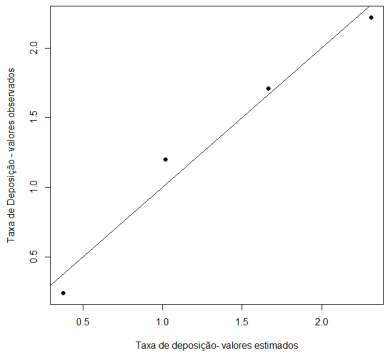
9.5.1 Gabarito

1. **a)** cor(x,y) [2] = 0,99;

   b) $\hat{\beta}_0$= -0,27, $\hat{\beta}_1$= 0,032 $\hat{y} = -0,27 + 0,032x_i$;

   c) $\hat{y_1} = 0,38$, $\hat{y_2} = 1,02$, $\hat{y_3} = 1,67$, $\hat{y_4} = 2,31$;

   d)

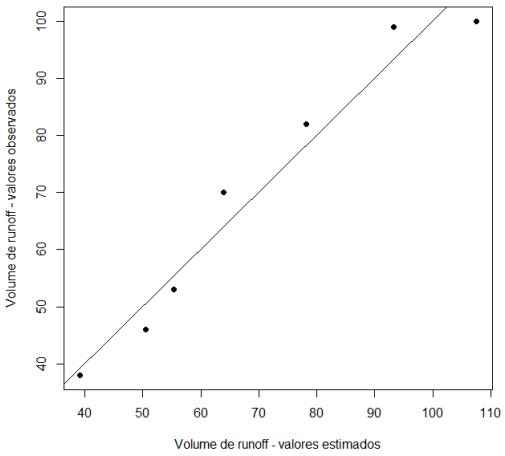


2. **a)** cor(x,y) = 0,98;

   b) $\hat{\beta}_0$= -13,09, $\hat{\beta}_1$= 0,95 $\hat{y} = -13,09 + 0,95x_i$;

   c) $\hat{y_1} = 39,17$, $\hat{y_2} = 50,57$, $\hat{y_3} = 55,33$, $\hat{y_4} = 63,88$, $\hat{y_5} = 78,13$, $\hat{y_6} = 93,34$, $\hat{y_7} = 107,59$;

   d)



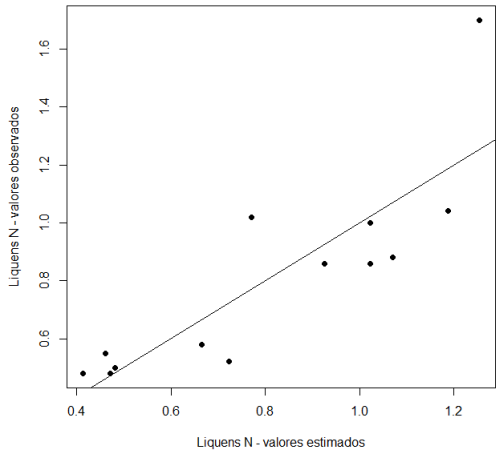
[2] No geral, os resultados são apresentados com base em valores com dois dígitos, baseados em critérios de arredondamento.

3. **a)** cor(x,y)= 0,85;

   b) $\hat{\beta}_0$= 0,37, $\hat{\beta}_1$= 0,97 $\hat{y} = 037 + 0,97x_i$;

   c) $\hat{y_1} = 0,41$, $\hat{y_2} = 0,46$, $\hat{y_3} = 0,47$, $\hat{y_4} = 0,48$, $\hat{y_5} = 0,67$, $\hat{y_6} = 0,72$, $\hat{y_7} = 0,77$, $\hat{y_8} = 0,93$, $\hat{y_9} = 1,02$, $\hat{y_{10}} = 1,02$, $\hat{y_{11}} = 1,07$, $\hat{y_{12}} = 1,19$, $\hat{y_{13}} = 1,25$;

   d)

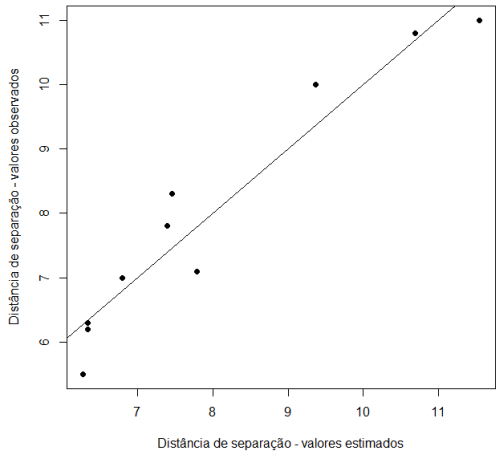


4. **a)** cor(x,y) = 0,96;

   b) $\hat{\beta}_0$= -2,18, $\hat{\beta}_1$= 0,66 $\hat{y} = -2,18 + 0,66x_i$;

   c) $\hat{y_1} = 6,27$, $\hat{y_2} = 6,34$, $\hat{y_3} = 6,34$, $\hat{y_4} = 6,80$, $\hat{y_5} = 7,39$, $\hat{y_6} = 7,46$, $\hat{y_7} = 7,79$, $\hat{y_8} = 9,34$, $\hat{y_9} = 10,69$, $\hat{y_{10}} = 11,55$,

   d)

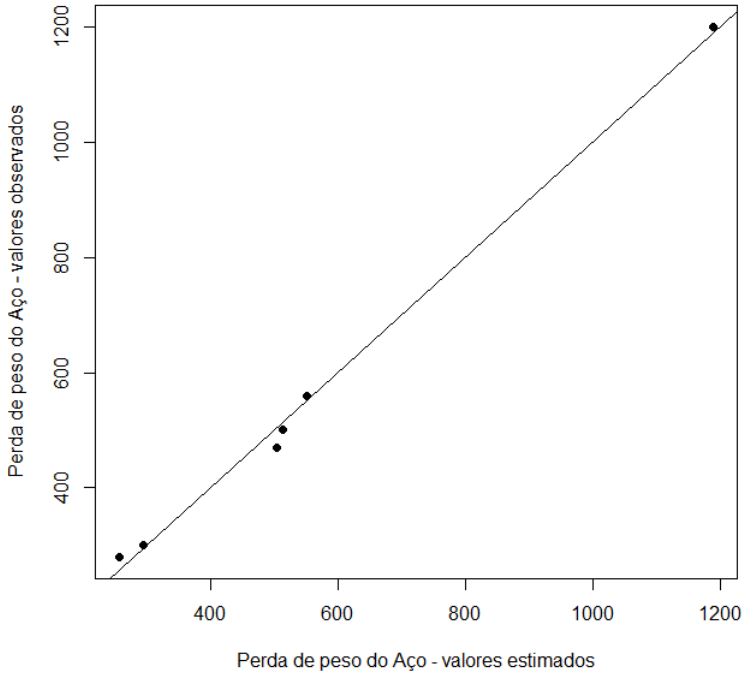


5. **a)** cor(x,y) = 0,998;

   b) $\hat{\beta}_0$= 122,86, $\hat{\beta}_1$= 9,53 $\hat{y} = 122,86 + 9,53x_i$;

   c) $\hat{y_1} = 256,27$, $\hat{y_2} = 294,39$, $\hat{y_3} = 504,02$, $\hat{y_4} = 513,55$, $\hat{y_5} = 551,67$, $\hat{y_6} = 1190,11$,

   d)



Anexos

Tabela 80 – Valores Críticos para Distribuições de Amplitude Studentizada

		a								
GL_{Den}	α	2	3	4	5	6	7	8	9	10
5	1%	5,70	6,98	7,80	8,42	8,91	9,32	9,67	9,97	10,24
	5%	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99
6	1%	5,24	6,33	7,03	7,56	7,97	8,32	8,61	8,87	9,10
	5%	3,46	4,34	4,90	5,30	5,63	5,90	6,12	6,32	6,49
7	1%	4,95	5,92	6,54	7,00	7,37	7,68	7,94	8,17	8,37
	5%	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16
8	1%	4,75	5,64	6,20	6,62	6,96	7,24	7,47	7,68	7,86
	5%	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92
9	1%	4,60	5,43	5,96	6,35	6,66	6,91	7,13	7,33	7,49
	5%	3,20	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74
10	1%	4,48	5,27	5,77	6,14	6,43	6,67	6,87	7,05	7,21
	5%	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60
11	1%	4,39	5,15	5,62	5,97	6,25	6,48	6,67	6,84	6,99
	5%	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49
12	1%	4,32	5,05	5,50	5,84	6,10	6,32	6,51	6,67	6,81
	5%	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,39
13	1%	4,26	4,96	5,40	5,73	5,98	6,19	6,37	6,53	6,67
	5%	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32
14	1%	4,21	4,89	5,32	5,63	5,88	6,08	6,26	6,41	6,54
	5%	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25
15	1%	4,17	4,84	5,25	5,56	5,80	5,99	6,16	6,31	6,44
	5%	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20
16	1%	4,13	4,79	5,19	5,49	5,72	5,92	6,08	6,22	6,35
	5%	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15
17	1%	4,10	4,74	5,14	5,43	5,66	5,85	6,01	6,15	6,27
	5%	2,98	3,63	4,02	4,30	4,52	4,70	4,86	4,99	5,11
18	1%	4,07	4,70	5,09	5,38	5,60	5,79	5,94	6,08	6,20
	5%	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07
19	1%	4,05	4,67	5,05	5,33	5,55	5,73	5,89	6,02	6,14
	5%	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04
20	1%	4,02	4,64	5,02	5,29	5,51	5,69	5,84	5,97	6,09
	5%	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01
21	1%	4,00	4,61	4,99	5,26	5,47	5,65	5,79	5,92	6,04
	5%	2,94	3,56	3,94	4,21	4,42	4,60	4,74	4,87	4,98
22	1%	3,99	4,59	4,96	5,22	5,43	5,61	5,75	5,88	5,99
	5%	2,93	3,55	3,93	4,20	4,41	4,58	4,72	4,85	4,96
23	1%	3,97	4,57	4,93	5,20	5,40	5,57	5,72	5,84	5,95
	5%	2,93	3,54	3,91	4,18	4,39	4,56	4,70	4,83	4,94
24	1%	3,96	4,55	4,91	5,17	5,37	5,54	5,69	5,81	5,92
	5%	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92
25	1%	3,94	4,53	4,89	5,14	5,35	5,51	5,65	5,78	5,89

Tabela 81 – Valores críticos para Distribuição F-Snedecor $\alpha = 5\%$

	GL_{Num}									
GL_{Den}	1	2	3	4	5	6	7	8	9	10
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,33	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,13	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
45	4,06	3,20	2,81	2,58	2,42	2,31	2,22	2,15	2,10	2,05
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93

Tabela 82 – Valores para distribuição normal padrão acumulada- P(z < Z)

	Segunda Decimal de Z									
	0	1	2	3	4	5	6	7	8	9
0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,648	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,67	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,719	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,791	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,834	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,877	0,879	0,881	0,883
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,898	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,937	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,975	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,983	0,9834	0,9838	0,9842	0,9846	0,985	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,989
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,992	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,994	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,996	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,997	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,998	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,999	0,999
3,1	0,999	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1	1	1	1	1	1	1	1	1	1

Referências

ALLENDE, H.; VALLE, C. Ensemble methods for time series forecasting. In: ____. *Claudio Moraga: A Passion for Multi-Valued Logic and Soft Computing*. Cham: Springer International Publishing, 2017. p. 217–232.

BARBETTA, P.; REIS, M.; BORNIA, A. ***Estatística: para cursos de engenharia e informática***. Atlas, 2010. ISBN 9788522437658. Disponível em: <https://books.google.com.br/books?id=ianUkQEACAAJ>.

DEVORE, J. ***Probabilidade e estatística: para engenharia e ciências***. Pioneira Thomson Learning, 2006. ISBN 9788522104598. Disponível em: <https://books.google.com.br/books?id=1JOlPgAACAAJ>.

LOESCH, C. ***Probabilidade e Estatística***. LTC, 2012. ISBN 9788521621003. Disponível em: <https://books.google.com.br/books?id=22ytNAEACAAJ>.

MAGALHÃES, M. ***Probabilidade e Variáveis Aleatórias***. Edusp, 2011. ISBN 9788531409455. Disponível em: <https://books.google.com.br/books?id=Pel8ATx9QDQC>.

MELLO, M. P.; PETERNELLI, L. A. ***Conhecendo o R - Uma Visão mais que Estatísticaa***. [S.l.], 2013. Disponível em: <https://www.editoraufv.com.br/detalhes.asp?idproduto=1593809>.

MORETTIN, L. G. ***Estatística Básica***. [S.l.]: Pearson, 2010. ISBN 9788576053705.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.