Coleção UAB–UFSCar

Engenharia Ambiental

Luis Aparecido Milan

Estatística Aplicada

Estatística Aplicada

UAB-UFSCar
Universidade Federal de São Carlos
Rodovia Washington Luís, km 235
13565-905 - São Carlos, SP, Brasil
Telefax (16) 3351-8420
www.uab.ufscar.br
uab@ufscar.br

Luis Aparecido Milan

Estatística Aplicada

São Carlos

2014

Concepção Pedagógica
Daniel Mill

Supervisão
Douglas Henrique Perez Pino

Revisão Linguística
Clarissa Galvão Bengtson
Daniel William Ferreira de Camargo
Kamilla Vinha Carlos
Paula Sayuri Yanagiwara
Rebeca Aparecida Mega

Diagramação
Izis Cavalcanti
Juan Toro
Vagner Serikawa

Capa e Projeto Gráfico
Luís Gustavo Sousa Sguissardi

SUMÁRIO

UNIDADE 3: Introdução à probabilidade

UNIDADE 4: Introdução à Inferência Estatística

APRESENTAÇÃO

A Estatística é uma ciência que desenvolve e disponibiliza métodos e técnicas para lidarmos com números. Podemos transformá-los em gráficos e tabelas ou aplicar métodos mais complexos de análise, de forma a extrair deles informações úteis que podem ser utilizadas para guiar nossas decisões.

A rigor, a incerteza permeia tudo o que vivenciamos no dia a dia. No mundo real, nada ou quase nada é exatamente o que aparenta ser. Veja o exemplo: Quando compramos 1 kg de arroz em um supermercado, estamos, de fato, comprando algo muito próximo de 1 kg, mas que pode variar um pouquinho para cima ou para baixo. Temos nesse caso uma dose de incerteza. O mesmo ocorre com medições feitas no exercício profissional. Para lidar com situações que envolvam incerteza, apresentamos algumas Leis de Probabilidade, que permitem, de alguma forma, medir o quanto essa incerteza afeta os nossos resultados. Portanto, podemos também nos referir à Estatística como a ciência que nos possibilita tratar com a incerteza.

Um aspecto importante da Estatística é que ela nos permite analisar pequenos conjuntos de dados, que chamamos de amostra, e concluir para todos os elementos de um conjunto maior, no qual está o nosso interessse, ao qual chamamos população.

A esse procedimento chamamos inferência, quando analisamos uma pequena parte e generalizamos para o todo.

No caso do exemplo do arroz, seria como analisar uma amostra de 10 saquinhos de arroz e concluir para todo um lote de um mesmo fornecedor. Como fica a incerteza nesse caso? Se encontrarmos uma diferença para menos, poderemos afirmar que o consumidor está sendo prejudicado? Ou se trata apenas de uma variação eventual e que não se repetirá com frequência? Questões similares podem ser aplicadas à relação entre empresas que comercializam produtos entre si.

O objetivo deste texto é apresentar uma introdução à estatística direcionada especialmente aos usuários desses métodos. Uma dificuldade que todo principiante tem com métodos estatísticos está relacionada ao volume de contas que devem ser realizadas. Por esse motivo, incluímos, no início, uma introdução a uma ferramenta computacional que permite ao leitor realizar os cálculos relacionados aos temas estatísticos apresentados.

O texto está organizado de forma que, na Unidade 1, apresentamos uma breve introdução ao R, um programa que permitirá aos leitores implementar todos os métodos e técnicas tratados neste texto. O R é um programa de livre

acesso, disponível na internet, e que pode ser obtido gratuitamente. Após essa introdução, nas unidades seguintes, as instruções do R relacionadas a cada nova técnica estatística serão também apresentadas.

Estatística Descritiva é o tema da Unidade 2. Nessa Unidade, apresentamos algumas técnicas voltadas à descrição, sumarização e apresentação de dados. Essas técnicas são geralmente utilizadas no primeiro contato com um conjunto de dados e permitem traçar um perfil do fenômeno que se está estudando. Também são muito úteis em apresentações em geral.

Apresentamos, na Unidade 3, uma introdução à teoria de probabilidades. Os tópicos buscam introduzir o leitor às ideias de aleatoriedade e incerteza, características que estão associadas aos fenômenos em geral. Os conceitos de probabilidade fornecidos nessa Unidade formam a base necessária ao entendimento dos métodos apresentados na Unidade 4.

Por fim, na Unidade 4, apresentamos os métodos inferenciais. Esses métodos permitem, a partir da observação de amostras, estabelecer conclusões sobre as populações que as geram, sempre com o controle da incerteza associada a essas conclusões. Introduzimos nessa Unidade os temas estimação, intervalo de confiança, teste de hipóteses, amostragem, regressão linear simples e planejamento de experimentos.

Para encerrar, eu gostaria de agradecer a todos que, de forma direta ou indireta, contribuíram para a elaboração deste material. Também gostaria de deixar aberta a possibilidade de envio de sugestões e críticas, já que este é um projeto em andamento e este texto deverá ser alterado em breve com a inclusão de novos tópicos e melhoramento dos já existentes. Quem desejar contribuir com sugestões, pode fazê-lo enviando uma mensagem para o endereço eletrônico dlam@ufscar.br. Desde já, agradeço.

Luis Aparecido Milan

UNIDADE 1

Aprendendo a utilizar o R

1.1 Primeiras palavras

Nesta Unidade, introduzimos a ferramenta computacional necessária para implementar todos os procedimentos estatísticos descritos nas unidades que seguem. Trata-se do R, um aplicativo que está disponível para Windows, Linux e MacOS X, podendo ser obtido livremente na internet.

1.2 Problematizando o tema

Os métodos estatísticos em geral envolvem a realização de muitos cálculos e o R nos permite realizá-los de forma rápida e segura.

1.3 Texto básico para estudos

O R é um programa de uso geral, versátil e que pode servir a muitas finalidades. Ele pode, por exemplo, ser utilizado para fazer o orçamento doméstico.

Uma de suas características importantes e que nos interessa em particular é sua capacidade de realizar cálculos e gráficos estatísticos. O R será muito utilizado nesse texto. Por essa razão, é essencial que todos consigam instalá-lo em seus computadores e aprendam a utilizá-lo.

Vamos, então, instalar o programa R no computador e aprender a fazer algumas operações simples com ele.

1.3.1 Instalando o R em seu computador

O R, no Brasil, pode ser encontrado no seguinte endereço de internet: <http://brieger.esalq.usp.br/CRAN/>.[7]

Para utilizar o programa, você deve acessar o endereço e proceder da seguinte forma:

Procedimento de instalação do R:

1. Acesse a página indicada anteriormente;

2. Escolha o sistema operacional a ser utilizado: Linux, MacOS X ou Windows;

7 Caso esse endereço não esteja funcionando, o programa está disponível em muitos outros endereços pelo mundo. Para encontrá-los, basta realizar uma busca na internet com a sigla “CRAN”, o resultado da busca será o site do do projeto que desenvolve o R, “http://cran.r-project.org/”. Em seguida, selecione a opção “mirrors”, isso fornecerá uma lista de sites espelhos onde o programa está disponível. Escolha o mais próximo e siga em frente.

3. Selecione a opção “base”;

4. Baixe o instalador do programa R selecionando a opção adequada ao seu sistema operacional. Nesse ponto, o seu navegador pedirá autorização para executar um arquivo com extensão “exe”. Isso ocorrerá de forma diferente, dependendo do navegador utilizado (se Internet Explorer, ou Firefox, ou Chrome, ou outro), e você deve autorizar a execução para que a mesma ocorra, isto é feito clicando sobre a opção “Run” ou “Execute”, ou algo similar. Alguns navegadores também oferecem a opção “Salvar arquivo”, nesse caso você deve escolher a pasta em que deseja salvar o arquivo com o programa instalador. Procedendo dessa forma, você terá em seu computador um arquivo executável — com extensão “exe”;

5. Clicando sobre o nome desse arquivo executável, você instalará o R em seu computador. Durante a instalação, é oferecida a opção de idiomas, você pode escolher “Português (Brasil)”. A seguir, várias escolhas são oferecidas ao usuário que instala o programa. Sugerimos que você aceite as escolhas padrão, clicando sobre a opção “Avançar”, para que ele inicie a instalação. Uma janela com a opção “Concluir” é apresentada. Selecione-a para encerrar a instalação;

6. Um ícone com a letra R deverá aparecer em sua área de trabalho. Toda vez que você quiser executar o programa, você deverá clicar sobre esse ícone.

1.3.2 Primeiros passos

O R é um programa que permite vários níveis de utilização. O usuário inexperiente pode utilizá-lo de uma forma muito simples, como se fosse uma calculadora. Com o uso continuado, o usuário vai se familiarizando com o aplicativo. Um especialista pode realizar cálculos estatísticos extremamente sofisticados com o mesmo programa. Isso torna o R interessante para utilização nos mais variados níveis, de forma que o usuário pode utilizá-lo de maneira elementar no início, e, à medida que avança, novas instruções vão sendo adicionadas.

Para iniciar a execução do R, você deve clicar sobre o ícone “R” na área de trabalho do seu computador. Surgirá, então, uma janela com a expressão “R Console” no canto superior esquerdo. Esta é a tela de comandos.

A tela de comandos permite ao usuário apresentar instruções ao computador e obter respostas imediatas para cada instrução, individualmente, ou para um grupo delas. Por essa razão, dizemos que esta é uma linguagem interpretativa ou interpretada.

Assim sendo, iniciaremos a utilização do R com instruções bem elementares. É altamente recomendável que o leitor leia esta seção com o R aberto e que, a cada nova instrução apresentada, execute-a e procure entender o seu funcionamento.

Operações diretas

O R pode ser utilizado como se fosse uma calculadora. Experimente digitar, na tela de comandos, por exemplo, a sequência "`4 * 2`" seguida da tecla <enter>,[8] o número obtido deve ser o 8, que é o resultado da multiplicação de 4 por 2. Para evitar repetições a tecla <enter> será omitida a partir daqui.

Instruções para o R serão apresentadas em quadros como o exposto a seguir. O símbolo # indica o início de um comentário, assim o conteúdo de uma linha a partir do # é ignorado pelo interpretador da linguagem, não sendo portanto executado.

```
4 * 2 <enter>          #    multiplicação de 4 por 2
```

A partir deste ponto, apresentamos vários exemplos de instruções em linguagem R. É importante frisar que todos estes exemplos devem ser experimentados pelo leitor, o que facilita a assimilação da informação. Os exemplos em R aparecem em quadros como o exposto acima. Nos quadros, as instruções estão prontas e devem ser experimentadas como estão. Basta digitá-las na tela de comandos.

As operações aritméticas de soma, subtração, multiplicação e divisão são implementadas pelos símbolos `*`, `+`, `-` e `/` respectivamente.

```
2 + 3                  #    adição
5 - 3                  #    subtração
9 / 2                  #    divisão
```

As operações de exponenciação (potência), divisão inteira e módulo (resto da divisão inteira), por sua vez, são implementadas por meio dos símbolos ^, %/% e %%, respectivamente. O valor da constante π está disponível no R para ser usada em expressões matemáticas como pi, sendo pi = 3,141593... .

8 Em alguns teclados, a tecla <enter> pode ser representada pelo símbolo "↵".

```
2^3                 #   exponenciação (dois elevado ao cubo)
9 %/% 2             #   divisão inteira
9 %% 2              #   módulo (resto da divisão inteira)
pi                  #   mostra o valor de π (pi)
```

É importante notar que, na tela de comandos, utilizamos ponto (.) e não vírgula (,) para separar a parte inteira da parte decimal dos números, já que as instruções executáveis do R seguem a norma internacional. No texto, contudo, utilizamos vírgula, seguindo a norma brasileira.

As principais funções matemáticas estão disponíveis no R. Apresentamos uma lista dessas funções na Tabela 1.1, em que uma lista de instruções (e/ou funções) em sua forma geral e os argumentos das funções aparecem na forma de letras, sendo que, na maioria dos casos, utilizamos a letra "x", mas outras letras também são utilizadas. Para fazer uso dessas instruções no R, devemos substituir as letras por números adequados a cada função ou atribuir valores a essas letras (são as variáveis, como descrevemos mais adiante). Mais uma vez, recomendamos ao leitor que experimente cada uma das funções para entender seu funcionamento.

Tabela 1.1 Funções matemáticas disponíveis no R.

Instrução	Efeito
`abs(x)`	Valor absoluto de x
`exp(x)`	Exponencial de x
`gamma(x)`	Gama de x (matemática)
`lgamma(x)`	Log-gama de x
`log(x)`	Logaritmo de x na base e
`log10(x)`	Logaritmo de x na base 10
`sqrt(x)`	Raiz quadrada de x
`sin(x)`	Seno de x
`cos(x)`	Cosseno de x
`tan(x)`	Tangente de x
`asin(x)`	Arco seno de x
`acos(x)`	Arco cosseno de x
`atan(x)`	Arco tangente de x
`sinh(x)`	Seno hiperbólico de x
`cosh(x)`	Cosseno hiperbólico de x
`tanh(x)`	Tangente hiperbólica de x
`asinh(x)`	Arco seno hiperbólico de x
`acosh(x)`	Arco cosseno hiperbólico de x
`atanh(x)`	Arco tangente hiperbólico de x

Tabela 1.1 *Continuação...*

Instrução	Efeito
`ceiling(x)`	Menor inteiro > ou = a x
`floor(x)`	Maior inteiro < ou = a x
`round(x,digits = k)`	Arredonda x para k casas decimais

Um aspecto relevante na execução de instruções em R é que os nomes das funções devem ser digitados exatamente com estão. Pequenas mudanças, tais como passar uma letra de minúscula para maiúscula, faz com que o R não execute a instrução adequadamente. Por exemplo, para o R, "`Pi`" é diferente de "`pi`".

No quadro que segue, mostramos a aplicação de algumas das funções introduzidas na tabela Tabela 1.1.

```
sqrt(9)              #    calcula a raiz quadrada de 9

log(9)               #    calcula logaritmo de 9 na base e

ceiling(1.9)         #    calcula o maior inteiro

ceiling(-1.9)        #    calcula o maior inteiro

floor(1.9)           #    calcula o menor inteiro

floor(-1.9)          #    calcula o menor inteiro

round(pi, j)         #    arredonda o conteúdo do primeiro argumento para j
                     #    casas decimais (j deve ser um número inteiro)
```

Operações com variáveis

O R também nos permite armazenar valores numéricos ou alfanuméricos em variáveis. Variáveis são referências (nomes) que utilizamos para armazenar objetos que podem ser números, sequências, etc. Por exemplo, o número real 3.2 pode ser armazenado na variável "`X`" por meio da instrução mostrada no quadro a seguir, que deve ser lido como "`X` recebe 3.2".

```
X = 3.2              #    variável numérica
```

O conteúdo de uma variável pode ser verificado digitando-se o nome da variável seguido da tecla <enter>.

```
X <enter>
```

Todas as funções listadas na Tabela 1.1 também se aplicam a variáveis. Experimente os comandos listados no quadro a seguir.

Exemplos de instruções com variáveis.

```
y = 7.3                              #   Armazena 7.3 em y
z1 = sin(y)                          #   Calcula seno de 7.3 e guarda em z1
z1                                   #   Mostra o conteúdo de z1
z2=log(sqrt(sin(atan(y))))           #   Calcula a expressão e guarda em z2
z2                                   #   Mostra o conteúdo de z2
```

Escolha dos nomes das variáveis

Ao escolher os nomes das variáveis, devemos atentar para alguns aspectos:

i) Os nomes das variáveis devem ser formados por combinações de letras, números e pontos, não podendo começar por um número;

ii) O R considera diferenças entre letras maiúsculas e minúsculas, portanto 'x' e 'X' se referem a variáveis diferentes, da mesma forma que o conteúdo "Marte" é diferente de "marte", e ambos são distintos de "MARTE";

iii) Ao escolher nomes para as variáveis, devem ser evitados os nomes reservados para uso da própria linguagem. Entre esses nomes, estão, por exemplo, os nomes das funções matemáticas, mostrados na Tabela 1.1, e outros que poderão ou não ser citados aqui. Para saber se um nome está ou não sendo usado pela linguagem, digite o nome seguido da tecla <enter>. Caso esteja sendo usado, a tela mostrará seu conteúdo; caso não esteja em uso, a resposta será *Object* "nome" *not found*.

Faça um teste para verificar alguns nomes de variáveis de sua escolha e compare-os com alguns nomes reservados.

As variáveis previamente utilizadas pelo usuário podem ter seu conteúdo redefinido, bastando, para isso, atribuir às mesmas um novo conteúdo.

Tipos de variáveis

Variáveis no R podem ser dos seguintes tipos: numéricas, lógicas, alfanuméricas ou complexas.

As **variáveis numéricas** são as mais comuns e são usadas para armazenar conteúdo númerico. Tal conteúdo pode ser um número inteiro ou real. No exemplo mostrado acima, X armazenou o número real 3,2.

As **variáveis lógicas** assumem as condições “verdadeiro” e “falso”, que são representadas pelas palavras TRUE e FALSE (“verdadeiro” e “falso” em inglês, respectivamente). Veja no quadro a seguir a definição de variáveis lógicas. Note que em TRUE e FALSE todas as letras são maiúsculas.

```
y = TRUE
X = FALSE
```

As variáveis lógicas serão utilizadas mais adiante como argumentos de funções estatísticas e também para controle de fluxo em programas. Variáveis com conteúdo lógico podem ser utilizadas em expressões aritméticas e, nesse caso, assumem os valores 1 e 0 em lugar de TRUE e FALSE, respectivamente. Elas também têm utilização ampla em programação de novos procedimentos, mas esse tipo de programação está fora do escopo deste texto.

Experimente os comandos que seguem, você verá que dependendo do resultado da operação lógica, a operação `y * 5` assume valores diferentes. Na primeira linha y recebe o resultado da operação lógica “4 maior ou igual a 3”, cujo resultado é verdadeiro, portanto o conteúdo de y passa ser TRUE. A Tabela 1.2 mostra os operadores disponíveis no R.

```
y = 4 >= 3
y
y * 5
y = 4 < 3
y
y * 5
```

As **variáveis alfanuméricas** armazenam conteúdo alfanumérico, como mostrado no quadro a seguir.

```
x = ‘Marte’
y = ‘2000’
```

Nesse caso, o conteúdo deve estar entre aspas e não deve ser utilizado em operações numéricas.

Sempre que o resultado de uma operação não estiver definido, por exemplo, quando tentarmos obter o logaritmo de um número negativo, o conteúdo atribuído será “NA”. O conteúdo “NA” também é utilizado em situações em que temos dados com valores não conhecidos, os chamados *missing values*.

As operações com variáveis são usualmente separadas por meio de uma mudança de linha. Também é possível arranjar as instruções de forma a termos mais que uma por linha, bastando, para isso, separá-las com ponto e vírgulas (;), como mostrado a seguir.

```
X = 10; z1 = sin(x); z2 = cos(x)
```

A linguagem R permite que trabalhemos com variáveis complexas. No entanto, não tratamos desse tópico aqui, pois ele está fora do escopo deste material.

Comparadores e operadores lógicos

Comparações podem ser feitas entre variáveis, entre constantes ou entre variáveis e constantes. O resultado da operação será TRUE ou FALSE e poderá ser atribuído a outra variável. Os operadores lógicos usuais estão disponíveis na linguagem. A lista desses comparadores é mostrada na Tabela 1.2.

Tabela 1.2 Operadores para comparações.

`>`	Maior
`>=`	Maior ou igual
`<`	Menor
`<=`	Menor ou igual
`==`	Igual
`!=`	Diferente
`&`	E
`&&`	E sequencial
`|`	Ou
`||`	Ou sequencial
`xor(e1,e2)`	Ou exclusivo
`!`	Não

No quadro a seguir, apresentamos alguns exemplos de operações com variáveis lógicas.

Exemplos de operações com variáveis e operações lógicas.

```
x = 2                  #   atribui valor 2 a x
y = 3                  #   atribui valor 3 a y
x < y                  #   operador “menor”
x >= y                 #   operador “maior ou igual”
x<y & x>y              #   operador “e”
x<y | x>y              #   operador “ou”
x != y                 #   operador “diferente”
! x<y                  #   operador “não”
```

Os operadores “e” e “ou” sequenciais, representados por && e ||, respectivamente, agilizam o processamento na medida em que evitam operações desnecessárias. Se em uma sequência de operações “e” ocorrer um FALSE ou se em uma sequência de operações “ou” um TRUE for obtido, as demais operações não serão realizadas.

Sequências

Uma possibilidade no R é tratar dados na forma de sequências (também chamadas de vetores). Para efeitos práticos, do ponto de vista da linguagem R, é indiferente se considerarmos um ou outro. Dessa forma, passamos a usar o termo “sequência”. Esse instrumento aumenta o poder de processamento de informações, quando comparadas com variáveis simples, e simplifica as instruções. Uma sequência é construída concatenando-se vários elementos de mesma natureza.

Instrução `c( )`

Para concatenar elementos, usamos a instrução `c( )`. Como podemos observar nos exemplos listados no quadro a seguir, os elementos concatenados podem ser valores numéricos, conteúdos alfanuméricos, variáveis ou mesmo outras sequências previamente definidas.

Exemplos de instruções para criação de sequências.

```
s1=c(2,5,0,-1,2.5)     #   concatena os números listados e guarda em s1
s1                     #   mostra o conteúdo de s1
s2=c('Jose','Maria')   #   concatena o conteúdo listado e guarda em s2
x=10; y=20             #   inicializa x e y com 10 e 20 respectivamente
s3=c(x,y,2,3)          #   concatena o conteúdo listado e guarda em s3
s4=c(s1,s3,x,y,2,5)    #   concatena o conteúdo listado e guarda em s4
```

Instrução `seq(a,b,c)`

A função `seq(a,b,c)` gera uma sequência iniciando em a, terminando em b e com incrementos de c.

```
seq(10,25)              #   Sequência de 10 a 25
seq(1.3,2.5,0.3)        #   Sequência de 1.3 a 2.5, de 0.3 em 0.3
```

Uma forma abreviada de definir uma sequência é por meio da instrução de repetição ":". O comando a:b define uma sequência iniciando em a, terminando em b e com saltos unitários. Veja a seguir exemplos de utilização de ":".

```
3:10                    #   Cria a sequência {3, 4, 5, 6, 7, 8, 9, 10}
2.5:10                  #   Experimente!
```

Instrução `rep(a,b)`

A função `rep(a,b)` resulta numa sequência constituída de b repetições do conteúdo de a. No quadro que segue, temos exemplos de utilização da instrução `rep( )`.

```
rep(1.3,20)             #   Sequência de 20 valores 1.3
rep(1:3,4)              #   Sequência {1,2,3}, quatro vezes
```

Os operadores `seq( )` e `rep( )` podem ser combinados. Experimente os comandos mostrados no quadro que segue.

Exemplos de instruções para repetições e sequências.

```
rep(1:4,3)              #   Sequência de 1 a 4, três vezes
rep(2:4,1:3)            #   Veja o resultado no R!
rep(seq(1,5),rep(3,5))  #   Veja o resultado no R!
```

Instruções para sequências

Os operadores aritméticos e funções matemáticas descritos anteriormente também se aplicam a sequências numéricas. A aplicação de uma operação ou função matemática a uma sequência resulta na aplicação dessa operação ou função a cada um dos elementos da sequência.

Algumas instruções foram desenvolvidas para serem aplicadas especificamente a sequências. Na Tabela 1.3, mostramos uma lista dessas instruções com suas respectivas funções.

Tabela 1.3 Operadores para sequências.

Chamada	Efeito
`length(s)`	Retorna o comprimento da sequência
`sort(s)`	Retorna os elementos ordenados
`order(s)`	Retorna a ordem dos elementos
`rank(s)`	Retorna os postos dos elementos
`rev(s)`	Retorna os elementos em ordem reversa
`max(s)`	Retorna o maior dos elementos
`min(s)`	Retorna o menor dos elementos
`range(s)`	Retorna o menor e o maior dos elementos
`mean(s)`	Retorna a média aritmética dos elementos
`sd(s)`	Retorna o desvio padrão dos elementos
`var(s)`	Retorna a variância dos elementos
`median(s)`	Retorna a mediana dos elementos
`quantile(s,prob)`	Retorna os quantis estipulados em prob
`sum(s)`	Retorna a soma dos elementos
`prod(s)`	Retorna o produto dos elementos
`cumsum(s)`	Retorna a sequência de somas acumuladas
`cumprod(s)`	Retorna a sequência de produtos acumulados

Algumas das instruções mostradas na Tabela 1.3 são de especial interesse para a estatística. Essas funções serão tratadas com detalhes mais adiante, são elas: `mean(s)`, `sd(s)`, `var(s)`, `median(s)` e `quantile(s,prob)`.

No quadro a seguir, apresentamos exemplos de instruções para sequências. Experimente cada uma delas e verifique o efeito que produzem.

Exemplos de instruções para sequências.

```
length(0:50)            # fornece o número de elementos da sequência
sort(c(2.3,5.1,3.2))    # ordena valores da sequência fornecida
order(c(2.3,5.1,3.2))   # fornece a ordem dos valores da sequência
range(c(2.3,5.1,3.2))   # fornece os valores mínimo e máximo
exp(c(2.3,5.1,3.2))     # calcula a exponencial
log(1:100)              # calcula o logaritmo
median(1:4)             # calcula a mediana da sequência
sum(1:5)                # calcula a soma da sequência
mean(1:5)               # calcula a média da sequência
var(1:5)                # calcula a variância da sequência
sd(1:5)                 # calcula o desvio padrão da sequência
cumsum(1:5)             # calcula a soma acumulada da sequência
cumprod(1:5)            # calcula o produto acumulado da sequência
```

1.3.3 Ajuda on-line

Nesta primeira seção, buscamos introduzir o leitor à utilização do aplicativo R, que será muito requisitado durante todo o desenvolvimento dos métodos estatísticos que seguem. Mas, como não poderia deixar de ser, este material é extremamente conciso e fornece apenas a base para uma utilização inicial do recurso. O programa R traz consigo recursos para expandir esse conhecimento por parte dos usuários.

Um aspecto importante do R é a ajuda ao usuário. Suponha que você esteja no meio de um trabalho e esqueceu o que faz uma instrução ou como passar os argumentos para uma função do R. Nesse caso, você tem à disposição todos os manuais. Basta clicar em “Ajuda” na barra de ferramentas – barra horizontal na parte superior da janela – e, logo em seguida, deslocar o indicador do *mouse* para “Manuais (em pdf)”. Aparecerá, então, uma lista de todos os manuais disponíveis *online*.

Você também pode aprofundar o seu conhecimento sobre o R. Para isso, é recomendável a leitura do manual “An introduction to R”, que está disponível *online*. Para acessá-lo basta clicar na palavra “Ajuda” na barra de controle. O “R reference manual” pode ser muito útil para aprofundar seu conhecimento sobre instruções específicas.

Uma forma mais rápida de obter ajuda sem ter de passar pelos manuais é por meio da instrução `help( )`. Suponha que você esteja na janela "R Console" e deseje obter mais informação sobre a instrução `log( )`, que calcula logaritmos. Para obter uma explicação detalhada sobre como calcular logaritmos, basta digitar `help(log)` ou, de forma abreviada, `?log`. Em ambos os casos, uma nova janela será aberta com a informação desejada.

1.4 Considerações finais

Estamos encerrando esta breve introdução ao R. Não se preocupe em memorizar o que fazem todas as instruções apresentadas. Neste ponto, basta que você tenha executado cada uma delas pelo menos uma vez e tenha buscado entender o seu funcionamento para ganhar alguma habilidade com o R. As instruções relativas a cada tópico futuro serão apresentadas à medida que o texto for se desenvolvendo.

1.5 Atividades de aplicação, prática e avaliação

O R é um aplicativo que permite uma forma flexível de utilização. Procure identificar, em seu ambiente profissional e/ou domiciliar, situações em que se pode aplicar os conhecimentos sobre o R.

1.5.1 Atividades individuais

Identifique, no seu cotidiano, situações em que se é possível aplicar os conhecimentos adquiridos.

1.5.2 Atividades coletivas

Troque ideias sobre possibilidades de aplicação do R.

1.6 Estudos complementares

Recomendamos a leitura do manual "An introduction to R", que acompanha o programa, e também do "R reference manual", para aprofundar seu conhecimento sobre instruções específicas.

1.6.1 Saiba mais

As instruções apresentadas anteriormente estão todas em sua forma simplificada. Para conhecer em detalhes o potencial de cada uma, utilize o “R reference manual”, que também está disponível a partir da barra de ferramentas do R.

UNIDADE 2

Estatística descritiva

2.1 Primeiras palavras

Nesta Unidade, apresentamos algumas técnicas para descrição, sumarização e apresentação de dados. Para isso, definimos variáveis qualitativas e quantitativas e descrevemos as técnicas que melhor se adaptam a cada situação.

2.2 Problematizando o tema

Nas atividades cotidianas de um usuário de métodos estatísticos, surgem situações nas quais é necessário conhecer a fundo determinados fenômenos. Para atingir esse objetivo, é preciso fazer levantamentos e conduzir experimentos que gerarão conjuntos de dados que, por sua vez, devem ser analisados para produzir a informação desejada. A análise de dados deve ser executada com ferramentas adequadas a cada situação para que possamos chegar a conclusões corretas.

2.3 Texto básico para estudos

Todo fenômeno, seja ele natural ou provocado, pode ser transformado em dados. Isso é feito por meio do registro de informações na forma de variáveis.

Vamos chamar de elemento a unidade básica sobre a qual registramos as informações. Por exemplo, se estamos estudando lâmpadas, cada lâmpada é considerada um elemento. Chamamos de variáveis as informações que registramos sobre cada elemento. Por exemplo, se estamos interessados no tempo de funcionamento de lâmpadas, então o tempo é a variável de interesse.

Elementos podem ser pessoas, cidades, ruas, fábricas, rios, lagos, etc.

São exemplos de variáveis: peso, altura, gênero, volume, comprimento, número de habitantes, quantidade de CO_2 lançado no ambiente por dia, nível de oxigênio por unidade de volume, etc.

As variáveis podem ser classificadas como quantitativas ou qualitativas:

- **Variáveis quantitativas**: Como o nome diz, são variáveis que representam características numéricas de elementos. São exemplos de variáveis quantitativas: o peso, a altura, o volume, etc.;
- **Variáveis qualitativas**: São variáveis que se referem a características não numéricas dos elementos. São exemplos de variáveis qualitativas: o gênero (masculino ou feminino), grau de escolaridade (fundamental, médio ou superior), estado civil (solteiro ou casado), etc.

As variáveis quantitativas podem ser classificadas em discretas, como é o caso do número de pessoas numa sala, ou contínuas, como é o caso de peso e altura. As variáveis qualitativas também podem ser classificadas em categóricas, como no caso do estado civil, e ordinais, como é o caso de escolaridade, pois sabe-se que um indivíduo com ensino médio tem mais escolaridade que um com ensino básico, e o mesmo ocorre quando comparamos um indivíduo com ensino superior em relação a outro com ensino médio.

Geralmente, um estudo sobre um determinado fenômeno consiste na seleção de um conjunto de elementos, ao qual damos o nome de amostra, e no registro da informação relevante sobre cada um dos elementos. Essa informação relevante é representada por uma ou mais variáveis. A seguir, procedemos a análise desses dados para a obtenção de resultados que, por sua vez, levam às conclusões.

Definição: Estatística descritiva é a denominação dada ao conjunto de técnicas voltadas à descrição, sumarização e apresentação de um conjunto de dados.

Existe uma grande quantidade de técnicas voltadas à descrição dos dados. Apresentamos a seguir algumas dessas técnicas.

2.3.1 Tabela de frequências

Uma tabela de frequências é uma forma simples de organizar e sumarizar um conjunto de dados. Ela mostra como os dados se distribuem em classes. Classes podem ser definidas naturalmente, como no caso das variáveis qualitativas, em que temos as categorias, ou de uma forma arbitrária, como no caso das variáveis quantitativas, situação em que o analista fixa o início e o fim de intervalos que formam as classes.

A forma pela qual as tabelas de frequências são organizadas depende do tipo de variável que elas representam. Veja, a seguir, exemplos para variáveis quantitativas e qualitativas.

2.3.1.1 Tabela de frequências para variáveis quantitativas

Tabela de frequências é um instrumento cuja função é sumarizar um conjunto de dados. Essa sumarização permite ao observador uma melhor compreensão do comportamento desses dados se comparada à observação direta dos próprios dados.

A sumarização de dados na forma de uma tabela de frequências é feita em passos:

1. No primeiro passo, dividimos o intervalo que contém todas as observações em subintervalos, de forma que cada observação esteja contida em apenas um desses subintervalos. A esses subintervalos chamamos classes. Classes são um conjunto de intervalos que contém todos os dados observados, de forma que cada observação pertence a apenas um dos intervalos.
2. No segundo passo, contamos quantos elementos estão em cada classe e obtemos a “Frequência absoluta”, mostrada na segunda coluna da Tabela 2.1.
3. As frequências relativas são obtidas dividindo as frequências absolutas (segunda coluna) pelo número total de elementos, última linha da segunda coluna. As frequências relativas são apresentadas na terceira coluna da Tabela 2.1.
4. Outra componente da tabela de frequências é a frequência acumulada. Para obter frequência acumulada para uma certa linha da tabela, somamos a frequência absoluta desde a primeira classe até a linha em questão. As frequências acumuladas são listadas na quarta coluna da Tabela 2.1.
5. As frequências relativas acumuladas são obtidas dividindo as frequências acumuladas pelo total de elementos. As frequências relativas acumuladas são apresentadas na quinta coluna da Tabela 2.1.

No Exemplo 2.1, tratamos da quantidade de resíduos sólidos coletados por catadores em uma região geográfica. Temos portanto uma variável quantitativa.

Exemplo 2.1 Resíduos coletados.

Considere os dados apresentados no quadro que segue, a quantidade de resíduos sólidos coletada por catadores por dia (em kg).

```
94  85  79 100 87 78 94  88 80  90 85 108 88 91
88 105  98  80 93 73 86  99 84 103 88  96 86 86
86  88 114  77 79 82 85 102 89  86 95 102
```

A sumarização desses dados na forma de uma tabela de frequências é feita aplicando os passos descritos acima. A tabela de frequência resultante é representada pela Tabela 2.1.

Tabela 2.1 Tabela de Frequências para a quantidade de resíduos coletados.

Quantidade (Kg/dia)	Frequência absoluta	Frequência relativa	Frequência acumulada	Frequência relativa acumulada
(70,80][5]	7	0,175	7	0,175
(80,90]	18	0,450	25	0,625
(90,100]	9	0,225	34	0,850
(100,110]	5	0,125	39	0,975
(110,120]	1	0,025	40	1,000
Total	40	1,000		

Nota: É conveniente observar que, multiplicando as frequências relativas — terceira coluna — por 100, obtemos a porcentagem de elementos em cada classe ou categoria.

2.3.1.2 Tabela de frequências para variáveis qualitativas

As tabelas de frequências para variáveis qualitativas são construídas da mesma forma que para variáveis quantitativas, exceto no que diz respeito à divisão em classes que, neste caso, não é necessária, pois as categorias já formam as linhas da tabela. Veja o exemplo que segue.

Exemplo 2.2 O conjunto de dados simulados apresentado na Tabela 2.2 se refere aos tipos de utilização dada a pneus inservíveis (para sua finalidade original).

Tabela 2.2 Tabela de Frequências para aplicação de pneus inservíveis.

Tipo de aplicação	Frequência absoluta	Frequência relativa	Frequência acumulada	Frequência relativa acumulada
Pavimentação asfáltica	235	0,29	235	0,29
Produtos moldados	247	0,31	482	0,60
Superfície para atletismo	113	0,14	595	0,74
Pneus automotivos	90	0,11	685	0,86
Desvulcanização	29	0,04	714	0,89
Utensílios de plástico e borracha	31	0,04	745	0,93
Construção civil e outros	56	0,07	801	1,00
Total	801	1,00		

5 É interessante notar que na delimitação de intervalos, o parêntese indica a não inclusão da borda no intervalo, enquanto que o colchete indica a inclusão. Assim (70,80] indica que 70 não pertence ao intervalo e 80 pertence.

Na Tabela 2.2, temos, na coluna da esquerda, as categorias de classificação. As frequências absolutas são apresentadas na segunda coluna e correspondem à contagem do número de elementos que pertencem a cada uma das categorias listadas na primeira coluna da tabela. As demais colunas são obtidas da mesma forma que no caso da variável contínua descrita anteriormente.

Para construir a tabela do exemplo sobre resíduos sólidos no R, utilizamos o seguinte procedimento.

```
#
#    Registro dos dados
Res.Solido=c(94,85,79,100,87,78,94,88,80,90,85,108,88,91,88
,
105,98,80,93,73,86,99,84,103,88,96,86,86,86,88,114,77,79,82
, 85,102,89,86,95,102)
#
#    Classifica cada observação em uma classe
Classes=cut(Res.Solido,c(70,80,90,100,110,120))
#
#    Cálculo da frequência absoluta
Fre.Ab=table(Classes)
#
#    Cálculo da frequência relativa
Fre.Re=Fre.Ab/sum(Fre.Ab)
#
#    Cálculo da frequência acumulada absoluta
Fre.Ac=cumsum(Fre.Ab)
#
#    Cálculo da frequência acumulada relativa
Fre.Re.Ac=Fre.Ac/sum(Fre.Ab)
#
#    Concatenação das variáveis
Tabela=cbind(Fre.Ab,Fre.Re,Fre.Ac,Fre.Re.Ac)
#
#    Apresentação dos resultados
Tabela
```

No procedimento anterior, a primeira instrução armazena os dados na forma de uma sequência em "`Res.Solido`". A seguir, a instrução `cut(.)` estabelece as classes e classifica cada observação em uma classe e a instrução `table(.)` calcula (conta), a partir dos dados, as frequências absolutas. As instruções seguintes calculam as frequências relativas, acumuladas e relativas acumuladas, e as armazena em `Fre.Ab`, `Fre.Re`, `Fre.Ac` e `Fre.Re.Ac`, respectivamente. A instrução `cbind(.)` do R concatena (ou cola, uma ao lado da outra) as colunas que lhe são fornecidas, formando uma matriz. A última instrução apresenta o resultado armazenado em Tabela.

2.3.2 Medidas de posição

As medidas de posição, como o nome diz, indicam aos usuários onde os dados estão localizados, entre os possíveis valores. Um exemplo óbvio de seu uso é quando se diz que a altura de indivíduos adultos está em torno de 1,70 metros. Outro exemplo é quando se diz que um equipamento industrial produz 20 mg de uma substância poluente por hora de funcionamento.

Entre as medidas de posição, aquelas que apontam o centro da distribuição dos dados são as mais utilizadas, são indicadores do centro em torno do qual os dados se distribuem, contudo não são as únicas. Os quartis e quantis são indicadores de posição que apontam para outra posição que não o centro, como veremos mais adiante.

As medidas de posição mais utilizadas são a média e a mediana, sendo a média a mais comum entre elas.

2.3.2.1 Média

A média, ou mais especificamente a média aritmética, para um conjunto de dados, é a soma dos componentes dividida pelo número de elementos da amostra.

Para formalizar o conceito, representamos um conjunto de n observações por $x = \{x_1, x_2, \ldots, x_n\}$. A média aritmética dessas observações é representada por $\bar{x}$ e é obtida pela aplicação da expressão

$$\bar{x} = (x_1 + x_2 + \ldots + x_n)/n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Na expressão acima, Σ, que corresponde à letra grega sigma, indica a soma dos elementos que a seguem. Nós lemos $\sum_{i=1}^{n}$ como "a somatória para i variando de 1 até n".

Exemplo 2.3 Média.

Considere uma empresa que produz refugos[6] diários cujos pesos registrados foram $x = \{71, 81, 64, 61, 74\}$. A média aritmética dessas observações ou, em outras palavras, o peso médio diário desses refugos é dado por

$$\bar{x} = (71 + 81 + 64 + 61 + 74)/5 = 351/5 = 70,2.$$

A produção média diária de refugos, nesse caso, é 70,2 kg.

Para calcular a média do exemplo precedente no aplicativo R, utilizamos o procedimento que segue.

```
#
#     Exemplo de cálculo de média
X = c( 71, 81, 64, 61, 74 )
Mx = mean( X )
Mx
```

As médias são muito úteis para fazer projeções elementares, tais como a do exemplo a seguir.

Exemplo 2.4 Poluição.

Considerando que uma fábrica lança diariamente no ambiente em média 100 kg de um determinado poluente, perguntamos: quanto essa fábrica lançará em 30 dias de atividade?

Para responder a essa questão, podemos utilizar a média. Multiplicamos o lançamento médio diário (100 kg) pelo número de dias (30) e temos a resposta:

$$\text{lançamento em 30 dias} = 100 \times 30 = 3000\,\text{kg},$$

ou seja, projetamos o lançamento de 3000 kg, ou 3 toneladas, de poluentes em 30 dias.

É importante deixar claro que esta é uma projeção grosseira, métodos mais apurados, que consideram também a incerteza associada, são apresentados na Unidade 4.

6 Refugos são peças que foram produzidas e descartadas devido à não conformidade com as especificações do produto.

2.3.2.2 Mediana

Do mesmo modo que a média, a mediana também indica o centro em torno do qual os dados estão distribuídos, mas usa um critério diferente para definir esse centro. Isso é feito ordenando os dados e escolhendo o ponto central dos dados ordenados. Caso o número de observações seja ímpar, escolhemos o dado central; ou fazemos a média dos dois dados centrais caso o número de observações seja par.

Formalizando, para um conjunto de n observações, representadas por $x = \{x_1, x_2, \ldots, x_n\}$, e para dados ordenados por $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$, a mediana dessas observações é

$$\text{Mediana}\,(x) = x_{((n+1)/2)}, \text{ se n for ímpar,}$$

ou

$$\text{Mediana}\,(x) = \left(x_{(n/2)} + x_{(n/2+1)}\right)/2, \text{ se n for par.}$$

Exemplo 2.5 Mediana com n ímpar.

Considere os mesmos dados utilizados no Exemplo 2.3, em que se calcula a média, $x = \{71, 81, 64, 61, 74\}$. A mediana é obtida ordenando as observações, $\{61, 64, 71, 74, 81\}$, e, como n é ímpar ($n = 5$), a mediana assume o valor da observação central, posição $(5+1)/2 = 3$, ou seja,

$$\text{Mediana}\,(x) = x_{((n+1)/2)} = x_{(3)} = 71.$$

Nesse caso, a mediana das observações ou o peso mediano é 71 kg.

Para ilustrar o caso em que n é par, vamos acrescentar uma observação ao conjunto de dados e refazer o exemplo.

Exemplo 2.6 Mediana com n par.

Considere os mesmos dados utilizados no exemplo anterior ampliados em uma observação, $x = \{71, 81, 64, 61, 74, 70\}$. Os dados ordenados são $\{61, 64, 70, 71, 74, 81\}$. Como n é par (n = 6), a mediana assume o valor da média das duas observações centrais, ou seja,

$$\text{Mediana}\,(x) = \left(x_{(n/2)} + x_{(n/2+1)}\right)\Big/2 = \left(x_{(3)} + x_{(4)}\right)\Big/2 = (70+71)/2 = 70{,}5.$$

Nesse caso, a mediana da produção diária de refugos é 70,5 kg.

Para calcular a mediana no aplicativo R, utilizamos o seguinte procedimento.

```
#
#     Exemplo de cálculo da mediana
X = c( 71, 81, 64, 61, 74 )
Md = median( X )
Md
```

Notas:

1. A mediana de um conjunto de dados separa esses dados de forma que o número de elementos acima e abaixo da mediana sejam iguais;
2. Da mesma forma que a média, a mediana indica o centro em torno do qual os dados estão distribuídos;
3. A mediana não sofre influência de valores extremos, o que pode acontecer com a média.

2.3.2.3 Quantis

Os quantis são medidas que apontam para as posições que dividem os dados em proporções preestabelecidas.

Os quantis são medidas de posição que dividem os dados de forma que uma proporção “p” dos dados fique abaixo e uma proporção “1 – p” dos dados fique acima do referido quantil. Ao quantil calculado para uma proporção p chamamos “quantil p” e denotamos por q(p). Os valores de p devem estar entre zero e um, ou seja, $0 \leq p \leq 1$.

Para calcular o quantil p para um conjunto de dados $x = \{x_1, x_2, \ldots, x_n\}$, calculamos inicialmente p_i, que é dado por

$$p_i = (i - 0{,}5)/n,$$

para $i = 1, \ldots, n$.

A seguir, ordenamos os dados formando a sequência de dados ordenados $\left(x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right)$. O quantil p é então definido por

$$q(p) = x_{(1)}, \text{ se } p < p_1$$

$$q(p) = x_{(i)}, \text{ se } p = p_i \quad \text{e} \quad i = 1, 2, \ldots, n;$$

$$q(p) = \left(1 - f_i\right) x_{(i)} + f_i x_{(i+1)}, \text{ se } p = p_i$$

$$q(p) = x_{(n)}, \text{ se } p > p_n,$$

em que $f_i = \dfrac{p - p_i}{p_{i+1} - p_i}$.

Pelo procedimento descrito acima, podemos notar que se $p < p_1$ o quantil p corresponde à menor das observações; se $p > p_n$ o quantil p corresponde à maior das observações; se $p = p_i$ o quantil p corresponde à i-ésima observação; e se p estiver entre dois p_i's então fazemos a interpolação mostrada acima.

2.3.2.4 Quartis

Os quartis são casos particulares de quantis que dividem as observações em quatro partes de 25% cada.

O primeiro quartil corresponde à posição que separa um quarto das observações (25%) abaixo de seu valor e três quartos (75%) acima.

O segundo quartil, seguindo a definição, divide as observações em dois quartos, ou seja, metade (2/4 = ½), ou 50%, abaixo e metade acima. O segundo quartil corresponde à mediana e ao quantil 0,5, ou q(0,5).

O segundo quartil é equivalente à mediana e praticamente não é citado, já que nos referimos a essa medida como mediana.

O terceiro quartil divide as observações de forma a termos três quartos das observações abaixo (75%) e um quarto acima (25%) de seu valor.

Para obter o valor dos quartis, utilizamos os dados ordenados já utilizados na obtenção da mediana, $\left\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right\}$, e calculamos a posição na sequência de valores ordenados do primeiro e terceiro quartis por meio das expressões

$$p_1 = (n+1)/4 \quad \text{e} \quad p_3 = 3(n+1)/4,$$

respectivamente. Quando valores fracionários são obtidos, uma interpolação deve ser utilizada.

O primeiro quartil é definido como

$$Q1(x) = x(\lfloor p_1 \rfloor) + \left(x(\lfloor p_1 \rfloor + 1) - x(\lfloor p_1 \rfloor)\right)(p_1 - \lfloor p_1 \rfloor),$$

em que [.] denota a parte inteira de seu conteúdo. O segundo quartil é a própria mediana, $Q2(x) = \text{Mediana}(x)$.

O terceiro quartil é definido como

$$Q3(x) = x(\lfloor p_3 \rfloor) + \left(x(\lfloor p_3 \rfloor + 1) - x(\lfloor p_3 \rfloor)\right)\ (p_3 - \lfloor p_3 \rfloor).$$

Se p_1 ou p_3 não tiverem parte fracionária, as expressões anteriores simplificarão para $Q1 = x(p_1)$ e $Q3 = x(p_3)$, respectivamente.

Exemplo 2.7 Utilizando os dados do Exemplo 2.1, sobre coleta de resíduos sólidos, temos, neste exemplo, n = 40.

Para calcular os quartis

$$p_1 = (n+1)/4 = 41/4 = 10,25 \quad \text{e} \quad p_3 = 3(n+1)/4 = 30,75.$$

Os dados nas posições 10 e 11, após a ordenação, são ambos iguais a 85, não sendo necessária a interpolação para obter Q1(x) = 85. Os dados nas posições 30 e 31 são 95 e 96, interpolamos, portanto, para obtenção de Q3

$$Q3(x) = 95 + (96 - 95)\ (30,75 - 30) = 95,75.$$

Portanto, o primeiro e terceiro quartis são 85 e 95,75, respectivamente.

Para obter a posição que separa os dados em 5% abaixo e 95% acima, calculamos o quantil 0,05, ou q(0,05).

O primeiro passo é obter os dados ordenados.

```
73 77 78  79  79  80  80  82  84  85
85 85 86  86  86  86  86  87  88  88
88 88 88  89  90  91  93  94  94  95
96 98 99 100 102 102 103 105 108 114
```

No segundo passo, calculamos os p_i's. Como buscamos separar uma proporção de 0,05 dos menores valores, sabemos que o quantil correspondente está próximo de $x_{(2)}$ (já que 2 é 5% de 40), portanto não precisamos calcular todos os p_i's. Temos que $p_2 = (2 - 0{,}5)/40 = 0{,}0375$; e para enquadrar 0,05, calculamos $p_3 = (3 - 0{,}5)/40 = 0{,}0625$.

Calculamos então a fração $f_2 = (0{,}05 - 0{,}0375)/(0{,}0625 - 0{,}0375) = 0{,}5$ e, como $p_2 < 0{,}05 < p_3$, o quantil 0,05 é dado por

$$q(p) = (1 - f_i)x_{(i)} + f_i x_{(i+1)} = (1 - 0{,}5)77 + (0{,}5)78 = 77{,}5.$$

Portanto, temos $q(0{,}05) = 77{,}5$, que é o valor que divide os dados nas proporções 0,05 e 0,95.

Nota: Os quantis tratados aqui também são chamados de "quantis empíricos" por serem baseados nos dados. Existem também os quantis teóricos, baseados em distribuições probabilísticas, que serão tratados na Unidade 3.

Utilizamos a instrução `quantile(.)`, como apresentada a seguir, para obter no R os mesmos quantis calculados no Exemplo 2.7, incluindo o quantil q(0,05) e o primeiro, segundo e terceiro quartis.

```
#
#    Registro dos dados
Res.Solido=c(94,85,79,100,87,78,94,88,80,90,85,108,88,91,88
,
105,98,80,93,73,86,99,84,103,88,96,86,86,86,88,114,77,79,82
,
 85,102,89,86,95,102)
#
#    Exemplo de cálculo de quartis
quantile(Res.Solido,prob=c(0.25,0.5,0.75))
```

Como pode ser observado no quadro acima, utilizamos três argumentos para obter os quantis desejados no R, o primeiro é a sequência de dados "`Res.Solido`"; o segundo argumento, "`prob=c(0.25,0.5,0.75)`", estabelece as proporções dos quantis que correspondem aos quartis. Também é possível escolher o tipo de interpolação a ser utilizada, o argumento usado seria "type=k", em que k pode assumir valores inteiros de 1 a 9 na versão atual do R.

Exercício: Verifique no "R reference guide", disponível no site do R, os demais tipos de interpolação disponíveis.

2.3.3 Medidas de dispersão

Como o nome já diz, as medidas de dispersão nos informam o quanto os dados estão espalhados ou dispersos, ou, em outras palavras, elas medem a variabilidade dos dados. Medida de variabilidade é um nome que também é utilizado por alguns autores para se referir a essas medidas.

Apresentamos a seguir as medidas de dispersão mais utilizadas: a amplitude, o desvio médio, a variância, o desvio padrão e a distância entre quartis.

2.3.3.1 Amplitude

A amplitude é a medida de dispersão mais simples e consiste na distância entre o valor máximo e o mínimo entre os dados:

$$\text{Amplitude}(x) = \text{máximo}(x_i) - \text{mínimo}(x_i).$$

2.3.3.2 Desvio médio

O desvio médio é a média das distâncias entre a média dos dados e cada valor observado

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

em que o par de barras verticais, $|.|$, indica que deve ser considerado o valor absoluto do conteúdo.

2.3.3.3 Variância

A variância é a média dos quadrados das distâncias entre os valores observados e a média,

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

Note que, na expressão acima, para calcular a variância, dividimos por “n”, sendo possível encontrar situações em que a mesma é dividida por “n - 1”. Nós utilizaremos “n” quando tivermos toda a população ou quando tivermos uma amostra e estivermos apenas descrevendo-a. Utilizaremos “n - 1” quando tivermos uma amostra e estivermos fazendo inferência (assunto da Unidade 4 deste texto).

2.3.3.4 Desvio padrão

O desvio padrão é também uma medida de dispersão e como tal mede como os dados se espalham em torno da média. É definido como sendo a raiz quadrada da variância, ou seja,

$$\mathrm{Dp}(x) = \sqrt{\mathrm{Var}(x)}.$$

O diferencial do desvio padrão em relação à variância está na escala, enquanto a variância está em uma escala quadrática o desvio padrão mede a dispersão dos dados na escala natural dos mesmos.

2.3.3.5 Distância entre quartis

A distância entre quartis é obtida pela diferença entre o terceiro e o primeiro quartil,

$$\mathrm{Dq}(x) = Q3(x) - Q1(x).$$

Todas as medidas apresentadas medem de alguma forma a variabilidade dos dados. Isso quer dizer que quanto maior for o valor obtido, maior será a dispersão dos dados. Apresentamos, a seguir, um exemplo para dois conjuntos de dados, com variabilidades diferentes, em que podemos ver essas medidas em ação.

Exemplo 2.8 Utilizando os dados sobre resíduos sólidos, Exemplo 2.1, calculamos a seguir todas as medidas de dispersão.

$$\text{Amplitude}(x) = \text{máximo}(x_i) - \text{mínimo}(x_i) = 114 - 73 = 41$$

$$DM(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = (4{,}075 + 4{,}925 + \ldots + 12.075)/40 = 7{,}26$$

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left((4{,}075)^2 + (-4{,}925)^2 + \ldots + (12.075)^2\right)/40 = 81{,}47$$

$$Dp(x) = \sqrt{Var(x)} = \sqrt{81{,}46} = 9{,}0$$

$$Dq(x) = Q3(x) - Q1(x) = 95{,}25 - 85 = 10{,}25.$$

Como podemos observar no exemplo, as diversas medidas de dispersão apresentam resultados diferentes, não sendo comparáveis entre si. A comparabilidade acontece quando aplicamos a mesma medida a conjuntos de dados diversos, como no Exemplo 2.9.

As medidas de dispersão mostradas anteriormente podem ser calculadas no R por meio das instruções mostradas no quadro que segue. Note que, embora o R disponha de uma instrução própria para calcular a variância, a `var(.)`, utilizamos a seguir uma expressão. O motivo para utilizarmos a expressão e não a instrução `var(.)` é que nessa instrução a divisão da soma de quadrados se dá por (n – 1) e não por n, como definido anteriormente.

```
#
#     Cálculo da amplitude
Amplitude=Max(Res.Solido)-min(Res.Solido)
#
#     Cálculo do desvio médio
mRS=mean(Res.Solido)
DM=sum(abs(Res.Solido-mRS))/length(Res.Solido)
#
#     Cálculo da variância
Var=sum((Res.Solido-mRS)^2)/length(Res.Solido)
#
#     Cálculo do desvio padrão
Dp=sqrt(sum((Res.Solido-mRS)^2)/length(Res.Solido))
#
#     Cálculo da distância entre quatis
Dq=sum(quantile(Res.Solido,c(0.25,0.75))*c(-1,1))
```

Exemplo 2.9 Utilizando os dados da Tabela 2.3, calculamos as medidas de posição e dispersão para x e y.

Tabela 2.3 Dados para comparação de medidas de posição e dispersão.

i	1	2	3	4	5	6	7	8	9	10
x_i	52	58	36	70	62	62	59	53	55	69
y_i	109	82	67	55	75	75	49	115	46	91

As medidas de posição e dispersão dos dados, “x” e “y”, calculadas como descrito acima, são apresentadas na Tabela 2.4.

Tabela 2.4 Medidas de posição e dispersão para x e y.

	Média	Mediana	Q1	Q3	Amplitude	DM	Var	Dp	Dq
x	57,6	58,5	53,5	62,0	34	6,88	85,04	9,22	8,5
y	76,4	75	58	88,75	69	18,28	502,24	22,41	30,75

Os mesmos dados são apresentados na Figura 2.1 na forma de diagramas de linha. Nestes, a média aparece marcada por um triângulo apontando para

cima e a mediana por um triângulo apontando para baixo, ambas indicam o centro em torno do qual os dados se distribuem.

Observando x e y na Figura 2.1, notamos que os valores de x estão mais concentrados em torno da média que os valores de y. Essa diferença de dispersão entre x e y pode ser percebida de formas diferentes através das medidas apresentadas na Tabela 2.4, ou seja, Amplitude, DM, Var, Dp e Dq, mas todas elas mostram valores maiores para a dispersão de y, como constatado no gráfico.

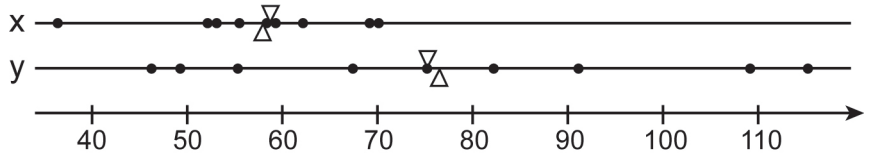


Figura 2.1 Locação e dispersão de x e y.

Notas:

1. Uma medida de posição e uma medida de dispersão, conjuntamente, são uma forma importante de sumarização do conjunto de dados, pois nos permitem identificar onde estão e como estão espalhados (ou dispersos) os dados;
2. O desvio padrão é a medida de dispersão mais utilizada, seguido pela variância;
3. Das medidas apresentadas anteriormente, apenas a variância não está na escala natural dos dados. Note o expoente na expressão.

2.3.4 Gráficos

Gráficos são ferramentas utilizadas para a visualização dos dados. A informação visual é a mais universal e com maior capacidade de comunicação ao ser humano, daí a importância dos gráficos. Eles permitem comunicar ideias estatísticas mesmo a pessoas com pouco conhecimento da área.

Apresentamos, a seguir, alguns dos principais tipos de gráficos utilizados e descrevemos o tipo de informação que eles proporcionam.

2.3.4.1 Histograma

O histograma é possivelmente o gráfico mais utilizado. Pode ser empregado tanto com variáveis qualitativas como com quantitativas.

No caso de variáveis qualitativas, traçamos barras verticais cujas alturas são proporcionais ao número de elementos em cada categoria.

O exemplo a seguir ilustra um histograma para variáveis quantitativas.

Exemplo 2.10 A Figura 2.2 mostra o histograma para os dados sobre resíduos sólidos. No caso de variáveis quantitativas, utilizamos as classes calculadas na tabela de frequências do Exemplo 2.1. As alturas das barras são proporcionais às frequências absolutas.

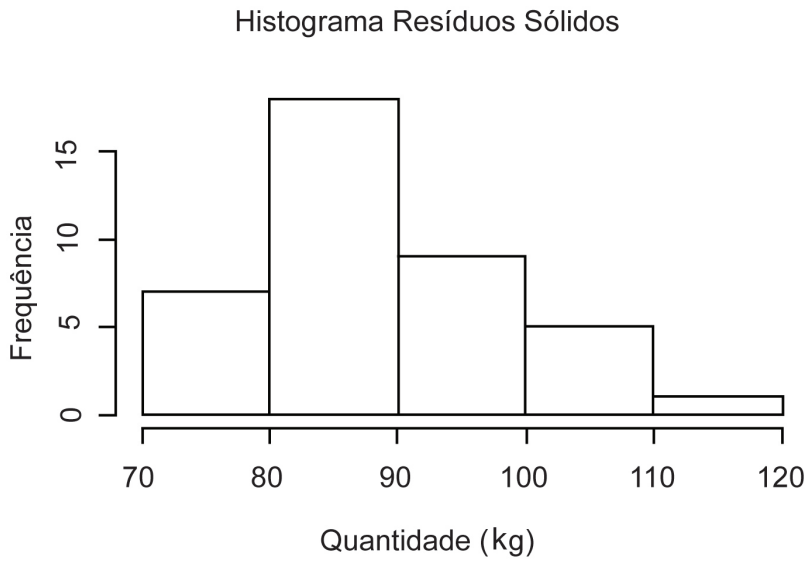


Figura 2.2 Histograma para a quantidade (em kg) de resíduos coletada.

Podemos construir o histograma no R usando a instrução a seguir.

```
#
#    Exemplo de construção de histograma
hist(Res.Solido,main='Histograma Resíduos Sólidos',
 xlab='Quantidade(Kg)',ylab='Frequência')
```

Nota: O histograma é uma ferramenta muito útil para apresentações. Nele, a informação apresentada é a forma com que os dados estão distribuídos, se concentrados ou dispersos e quais categorias são mais ou menos frequentes.

2.3.4.2 Ramo-e-folhas

O ramo-e-folhas é um instrumento de visualização que apresenta informação em forma de distribuição dos dados, tal como o histograma, porém mantém

a informação numérica. Uma característica importante do ramo-e-folhas é a facilidade de construção, podendo ser feito de forma direta, sem necessitar de computador ou instrumentos de desenho.

Exemplo 2.11 A aplicação do ramo-e-folhas ao conjunto de dados sobre resíduos, disponibilizada no Exemplo 2.1, é mostrada a seguir.

```
 7 | 3
 7 | 7899
 8 | 0024
 8 | 555666667888889
 9 | 01344
 9 | 5689
10 | 0223
10 | 58
11 | 4
```

Procedimeto para elaborar manualmente um ramo-e-folhas:

1. Identifique os valores máximo e mínimo entre os dados.
2. Estabeleça os limites superior e inferior da escala, a coluna à esquerda do traço vertical. Para isso, arredonde o maior para cima e o menor para baixo.
3. Dividindo o intervalo entre os limites superior e inferior, formamos as classes. No exemplo acima, as classes têm comprimento 5, iniciando na classe com números terminados em zero (70) até os terminados em 4 (74), seguida da classe com números terminados em 5 até 9 (75 a 79), seguida da classe incluindo números de 80 a 84, e assim por diante. As classes são marcadas pela coluna da esquerda, onde são anotados os dígitos dos números que compõem as classes, exceto o último.
4. A seguir, anotamos à direita do traço vertical o último dígito de cada dado observado na linha correspondente à classe a que pertence. Cada dígito à direita do traço vertical combinado com o(s) dígito(s) à esquerda recompõem o dado original. Do exemplo, na última linha, combinando 11 com 4, temos 114 do conjunto de dados originais. Durante essa anotação, é importante manter os dígitos anotados alinhados na vertical para garantir a eficiência do ramo-e-folhas.
5. Opcionalmente, podemos ordenar os dados de cada linha à direita do traço vertical para apresentação.

A construção do ramo-e-folhas no R é feita através da instrução `stem( )` apresentada a seguir.

```
#
#     Exemplo de ramo-e-folhas
#
stem(Res.Solido)
```

2.3.4.3 Gráfico de composição em setores (Pizza)

Os gráficos de composição em setores são utilizados principalmente para representar dados que indicam composição, ou seja, variáveis na forma de proporções cuja soma é a unidade ou porcentagens em que a soma é 100%.

Este gráfico apresenta forma arredondada e é dividido em fatias cujos tamanhos representam as proporções desejadas. Assemelham-se a uma pizza e, por essa razão, são chamados de gráficos tipo “pizza” ou “torta” (do inglês *pie chart*), conforme pode ser observado na Figura 2.3.

Em geral, esse gráfico é composto por um círculo cujos 360 graus são particionados em proporções iguais àquelas que desejamos representar.

A melhor forma de construção desse gráfico é a partir da respectiva tabela de frequência, fazendo o ângulo correspondente a cada item proporcional à sua frequência relativa (= 360 f_i).

Exemplo 2.11 O gráfico de composição para os dados do exemplo sobre destinação de pneus inservíveis é apresentado na Figura 2.3. Como podemos observar, o gráfico apresenta a composição do destino dado aos pneus inservíveis.

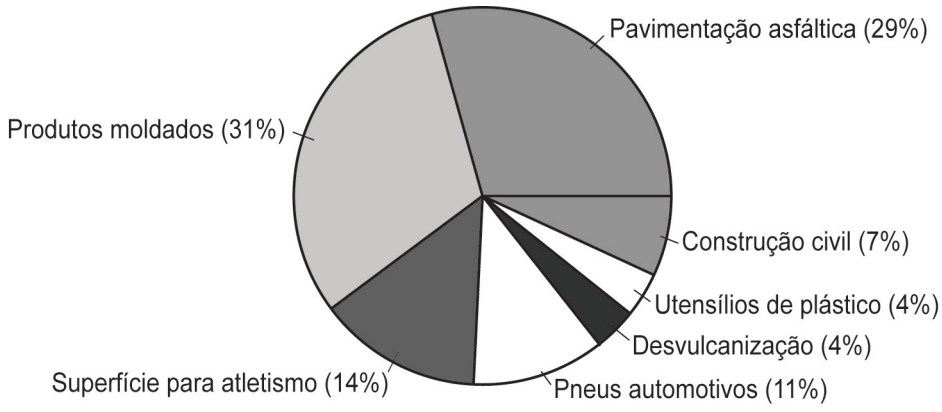


Figura 2.3 Gráfico por setores para dados do exemplo sobre pneus inservíveis.

A implementação do gráfico de composição no R pode ser feita pela instrução `pie( )` do R, como mostrado a seguir.

```
#

#    Exemplo de construção de gráfico de composição

Pneus=c(235,247,113,90,29,31,56)
names(Pneus)=c('Pavimentação asfáltica (29%)','Produtos molda-
dos (31%)', 'Superfície para atletismo (14%)','Pneus automo-
tivos (11%)', 'Desvulcanização (4%)', 'Utensílios de plástico
(4%)', 'Construção civil (7%)'
pie(Pneus, col=c("purple", "violetred1", "green3", "cornsilk",
"cyan", "white"))
```

2.3.4.4 Box-plot

O box-plot, também chamado de "diagrama de caixa", apresenta graficamente informações que sumarizam os dados. Sua forma é apresentada na Figura 2.4.

O seu corpo principal apresenta o formato de um retângulo cortado ao meio. Nesse retângulo, a linha horizontal inferior corresponde ao primeiro quartil, Q1; a linha superior, ao terceiro quartil, Q3; e a linha horizontal central corresponde à mediana.

O traço superior, chamado de limite superior e denotado por LS, é posicionado uma vez e meia a distância entre quartis acima do terceiro quartil, ou seja, $LS = Q3 + (1{,}5)Dq$.

De forma análoga, o traço inferior, chamado de limite inferior e denotado por LI, é posicionado uma vez e meia a distância entre quartis abaixo do primeiro quartil, ou seja, $LI = Q1 - (1{,}5)Dq$.

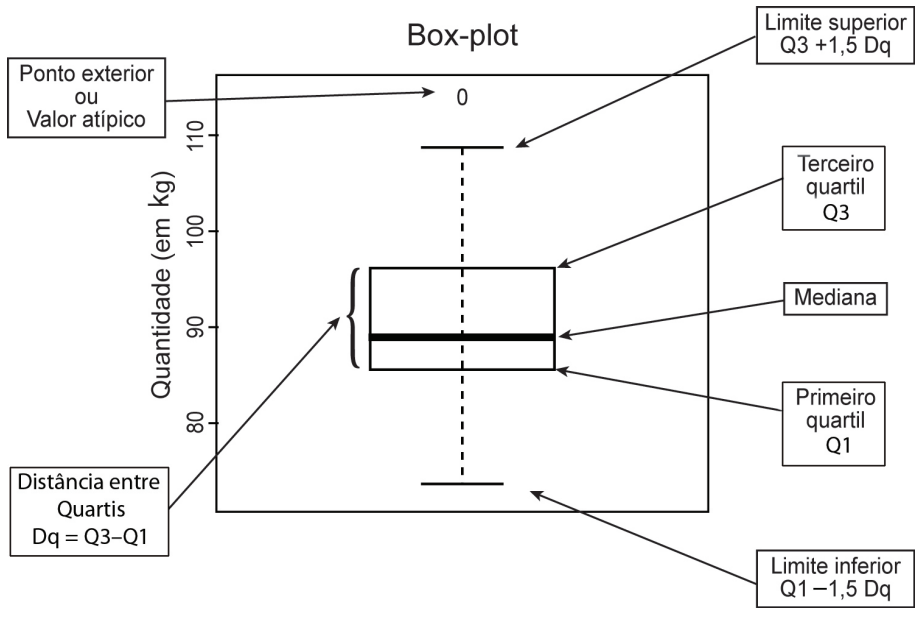


Figura 2.4 Box-plot para dados do exemplo.

Os pontos acima do limite superior ou abaixo do limite inferior são chamados de pontos exteriores e são representados individualmente no gráfico.

Pontos exteriores merecem atenção especial dos analistas. Estes podem eventualmente ser gerados por erros de coleta dos dados, caso em que devem ser corrigidos ou até descartados.

Em geral, observações com valores atípicos distorcem os valores da média, afastando-a do centro da distribuição e inflando o valor da variância – e consequentemente o valor do desvio padrão. A utilização destes valores distorcidos causam problemas quando estas medidas forem utilizadas em inferência, assunto que será tratado na Unidade 4.

O box-plot é uma poderosa forma de apresentação de informações sumarizadas sobre os dados, pois informa sobre o centro da distribuição dos mesmos por meio da mediana; sobre a dispersão por meio da distância entre quartis; e também sobre dados atípicos por meio de asteriscos.

Exemplo 2.12 O box-plot para os dados do exemplo é apresentado na Figura 2.4.

O box-plot pode ser implementado no R por meio da instrução `boxplot( )` demonstrada no quadro que segue, para os dados do Exemplo 2.1.

```
#
#             Exemplo de Box-plot
boxplot(Res.Solido,main='Box-plot',ylab='Quantidade          (em
Kg)')
```

2.3.4.5 Diagrama de dispersão

O diagrama de dispersão é um gráfico utilizado para mostrar a relação entre duas variáveis quantitativas.

Esse tipo de gráfico é construído a partir de dois eixos ortogonais, denominados ordenada e abscissa, cada um representando uma das variáveis envolvidas. Cada elemento é representado no gráfico por um ponto que aparece na posição (x_i, y_i).

Exemplo 2.13 Considere os dados x e y, cujas observações são apresentadas na Tabela 2.5.

Tabela 2.5 Dados para diagrama de dispersão.

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	17,4	13	12,9	13,1	11,9	16,3	16	19,3	11,3	14,7	15,2	18,5
y_i	36,4	28,8	26,8	30,3	29,3	34,6	38,3	44,6	26,6	33,6	35,5	40,8

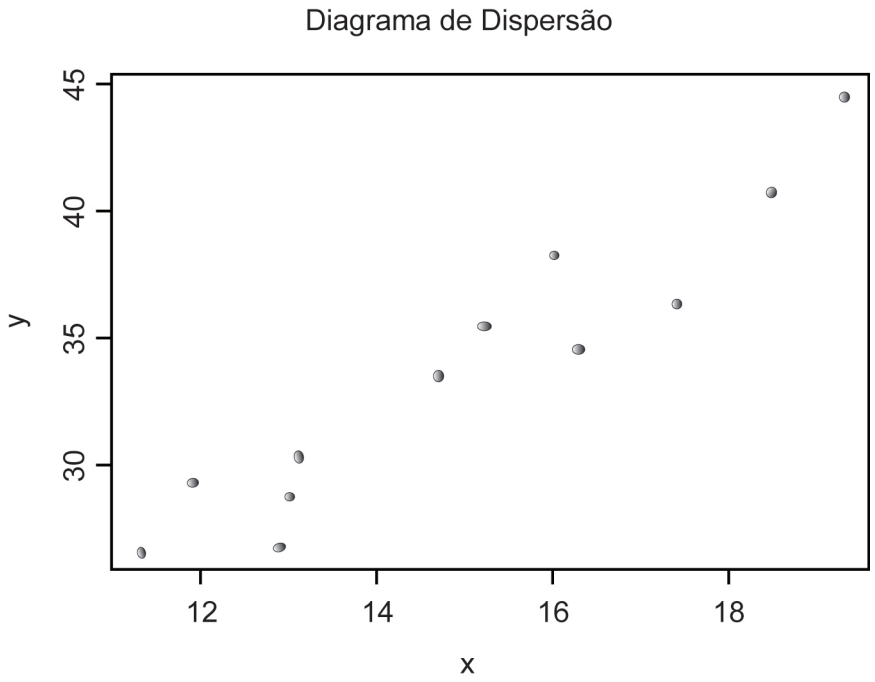


Figura 2.5 Exemplo de diagrama de dispersão.

O diagrama de dispersão é apresentado na Figura 2.5. Como podemos observar, existe uma associação positiva entre x e y, ou seja, à medida que x cresce, y também cresce.

Para produzir o gráfico do exemplo no R, utilizamos as instruções do quadro a seguir.

```
#      Exemplo de diagrama de dispersão

#      Entrada de dados

x=c(17.4,13,12.9,13.1,11.9,16.3,16,19.3,11.3,14.7,15.2,18.5)

y=c(36.4,28.8,26.8,30.3,29.3,34.6,38.3,44.6,26.6,33.6,35.5,40.
8)

#      Realização do gráfico

plot(x,y,main='Diagrama de Dispersão')
```

2.4 Considerações finais

Apresentamos, nesta seção, as principais técnicas utilizadas para descrição, sumarização e apresentação de dados. Também apresentamos as instruções em R para implementá-las.

2.5 Atividades de aplicação, prática e avaliação

É altamente recomendável que o leitor exercite a sua capacidade de experimentação com os procedimentos apresentados em R. Devido à limitação de espaço, apenas a forma mais simples das instruções é apresentada, entretanto uma rápida consulta ao manual permitirá muitas outras alternativas, especialmente nas instruções gráficas.

Busque, no seu dia a dia, fontes de dados que possam ser utilizadas para aplicar os métodos descritos anteriormente. Procure aplicá-las manualmente e utilizando o R.

2.5.1 Atividades individuais

Busque, no seu ambiente de trabalho, conjuntos de dados, aplique as técnicas descritas aqui e veja quais informações você consegue extrair desses dados.

2.5.2 Atividades coletivas

Troque ideias com seus colegas sobre os métodos que você aplicou aos seus dados e verifique se eles(as) utilizariam o mesmo método que você utilizou.

2.6 Estudos complementares

Recomendamos a leitura dos capítulos correspondentes à estatística descritiva nos textos citados na lista de referências.

2.6.1 Saiba mais

Você pode ampliar os seus conhecimentos de estatística descritiva incluindo novos métodos em seu repertório ou aprofundando seus conhecimentos sobre os métodos já tratados estudando as referências.

UNIDADE 3

Introdução à probabilidade

3.1 Primeiras palavras

Todo fenômeno, seja ele natural ou provocado, envolve algum tipo de incerteza. A teoria de probabilidades nos disponibiliza instrumentos que nos permitem mensurar e trabalhar com a incerteza. Nesta Unidade, estudaremos alguns desses instrumentos.

3.2 Problematizando o tema

Como medir a incerteza associada a fenômenos naturais e experimentos?

3.3 Texto básico para estudos

Nossa vida cotidiana é cercada de eventos aos quais está associada alguma incerteza. Geralmente, nós não atentamos para este fato. São inúmeros os exemplos, vejamos alguns. O ônibus que deveria passar no ponto às dez horas não passa exatamente às dez; as embalagens de produtos em supermercados não contêm exatamente a quantidade marcada; o número de frutas em uma árvore varia de ano para ano; a quantidade de resíduos produzida por uma fábrica varia de um dia para o outro; a quantidade de peixes em um lago ou rio; etc.

Também temos a incerteza associada a eventos do tipo do lançamento de uma moeda ou um dado. No caso da moeda, os resultados possíveis são cara e coroa, e no lançamento do dado o resultado da face que fica para cima pode ser 1, 2, 3, 4, 5 ou 6. São muitos os exemplos possíveis, detalhamos a seguir o exemplo da embalagem.

Exemplo 3.1 A incerteza sobre o peso real do pacote de arroz.

Considere a compra de um pacote de 1 kg de arroz em um mercado.

Utilizando uma balança de alta precisão, verificaríamos que o peso de um pacote escolhido ao acaso da prateleira não é de exatamente 1 kg. Mesmo verificando todos os pacotes da prateleira, não encontraríamos um pacote com exatamente 1 kg. E mais, não encontraríamos sequer dois pacotes com pesos exatamente iguais. O que temos é uma incerteza a respeito do peso real de arroz em pacotes de 1 kg.

Essa incerteza decorre de uma variação natural no processo de empacotamento. Também temos o erro de medida devido à balança. Em geral, para balanças de precisão, esse tipo de erro é muito pequeno se comparado ao erro devido à variação natural.

A incerteza apresentada no exemplo anterior pode ser transportada para a maioria dos fenômenos observáveis, daí a importância de entendermos os processos envolvidos, o que é feito por meio da teoria de probabilidades.

Apresentamos, a seguir, uma breve introdução à teoria de probabilidades. Esse conhecimento é fundamental para podermos entender os métodos analíticos que vêm a seguir.

3.3.1 Espaço amostral

Para estudar a incerteza associada a *fenômenos aleatórios*[7], o primeiro passo é fazer uma lista de todos os resultados possíveis.

Definição: Chamamos de **espaço amostral** a qualquer conjunto que contenha todos os resultados possíveis do fenômeno em estudo.

Representamos o espaço amostral discreto, isto é, um conjunto com número finito ou infinito enumerável de elementos, pela letra grega ômega maiúscula, Ω, e seus componentes por letras ômega minúsculas devidamente indexadas, ou seja, $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$. A cada elemento de Ω, ω_j, para j = 1,2,..., chamamos de ponto amostral e $\{\omega_j\}$ chamamos de evento elementar.

Veja, a seguir, alguns exemplos de espaços amostrais associados a fenômenos simples.

Exemplo 3.2 Lançamento de uma moeda.

Considere o lançamento de uma moeda. Antes de lançarmos a moeda, não sabemos o resultado do lançamento. Sabemos apenas que o resultado será cara ou coroa, representados por C e $\underline{C}$, respectivamente. Assim, temos $\Omega = \{C, \underline{C}\}$.

Exemplo 3.3 Lançamento de um dado.

Considere o lançamento de um dado. Antes de lançarmos o dado, não sabemos o resultado do lançamento, sabemos apenas que o resultado será uma das faces, podendo ser 1, 2, 3, 4, 5 ou 6. Dessa forma, temos $\Omega = \{1, 2, 3, 4, 5, 6\}$.

No caso do exemplo do pacote de um quilo de arroz, Exemplo 3.1, podemos definir o espaço amostral como sendo o conjunto dos números reais positivos, ou seja, $\Omega = R^+$. Nesse caso o Ω não é discreto e dizemos que o espaço amostral é contínuo.

Note que, nesse caso, o espaço proposto contém elementos que jamais ocorrerão — por exemplo, uma embalagem de arroz de 1 kg contendo 10 kg. Isso não constitui um problema, já que o espaço amostral, além de conter todos os

7 Fenômeno aleatório é todo fenômeno que apresente alguma incerteza associada a seus possíveis resultados.

valores possíveis, também pode conter valores com probabilidade zero, e para simplificar a sua definição, incluímos também eventos impossíveis de ocorrer.

Definição: Chamamos de evento a qualquer subconjunto do espaço amostral[8].

Denotamos os eventos por letras maiúsculas, por exemplo A, S, etc.

No lançamento da moeda, podemos estar interessados no evento "cara", $A = \{C\}$. No lançamento do dado, podemos estar interessados no evento "o resultado obtido é par", ou seja, $A = \{2, 4, 6\}$.

No exemplo da embalagem de 1 kg, ou 1000 g de arroz, estamos interessados no evento "a diferença entre o valor real e o valor nominal do peso é menor que 10 g", ou seja, o conjunto A, nesse caso, corresponde ao intervalo contendo os números reais maiores que 990 g e menores que 1010 g, $A = (990, 1010)$.

Exercícios:

1. Em um jogo de futebol do campeonato nacional, estamos interessados no número de gols que a equipe visitante marca. Defina um espaço amostral e um evento de interesse.
2. Uma empresa X está interessada em estudar o número de peças fora de especificação (Não Conformes) em um posto de trabalho. Defina um espaço amostral e um evento de interesse.
3. Um ônibus circulando pela área urbana de um município emite poluentes. Defina um espaço amostral que represente a quantidade de um certo poluente emitido. Indique um evento de interesse.
4. Identifique, no seu cotidiano, três fenômenos que envolvam incerteza. Defina espaços amostrais e fenômenos de interesse para cada um deles.

3.3.2 Probabilidade de um evento

Uma forma de medir a incerteza associada a um evento E é atribuir ao mesmo uma probabilidade. Denotamos a probabilidade de um evento E por P(E).

Definição: Probabilidade é um número entre zero e 1 que é atribuído a cada evento do espaço amostral e que deve seguir a três regras básicas, as quais chamamos de axiomas.

8 Essa definição vale sempre para o caso de espaço amostral discreto, se Ω não é discreto pode-se provar que existem exceções a essa regra, mas discutir esse nível de detalhe está fora do escopo desse texto.

Axiomas de probabilidade:

1. $P(\Omega) = 1$;

2. $0 \leq P(E) \leq 1$;

3. Para dois eventos disjuntos E_1 e E_2, temos que $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Eventos disjuntos são eventos cujos conjuntos que os representam têm a interseção vazia, ou seja, se os conjuntos E_1 e E_2 forem disjuntos, então $E_1 \cap E_2 = \varnothing$, em que $\varnothing$ denota o conjunto vazio.

Partindo dessa definição de probabilidade, podemos estabelecer alguns resultados que nos ajudam a entender melhor o que significa.

Probabilidade é um número entre zero e um. Quanto mais próximo de zero, menor a chance de o evento correspondente acontecer, e quanto mais próximo de um maior a chance do evento ocorrer.

Levando essa regra ao extremo, temos que um evento com probabilidade um acontecerá com certeza. Chamamos isso de evento certo. É o que acontece quando o evento é o próprio espaço amostral, $E = \Omega$, axioma um, apresentado acima.

Quando a probabilidade de um evento é zero, esse evento não ocorre. Chamamos esse evento de "evento impossível". Esse é o caso, por exemplo, de se obter o número 7 no lançamento de um dado de seis faces, algo que nunca vai ocorrer, portanto $P\{7\} = 0$. O conjunto vazio também tem probabilidade zero, $P(\varnothing) = 0$.

A probabilidade de um evento qualquer é a soma das probabilidades dos eventos elementares que o compõem. Essa propriedade decorre do axioma três e é muito útil para estabelecer probabilidade de eventos não elementares.

Exemplo 3.4 Considere o lançamento de uma moeda equilibrada. Nesse caso, temos que a probabilidade do evento cara é igual à probabilidade do evento coroa, $P(C) = P(\underline{C}) = 0,5$.

A probabilidade do evento E, "cara ou coroa", num lançamento, é a probabilidade da união dos eventos elementares, ou seja,

$$P(E) = P(E_1 \cup E_2) = P(C \cup \underline{C}) = P(C) + P(\underline{C}) = 0,5 + 0,5 = 1.$$

Nesse caso, utilizamos o fato de os eventos cara e coroa serem eventos elementares e, portanto, disjuntos. Cara e coroa não podem ocorrer ao mesmo tempo em um único lançamento para aplicar o axioma três e obter a probabilidade do evento E. Nesse caso, o evento E coincide com o espaço amostral.

3.3.3 Regra da interseção

Toda vez que trabalhamos com dois eventos, A e B, e desejamos determinar a probabilidade de que ambos ocorram, devemos aplicar a regra da interseção, $A \cap B$ = "ambos os eventos, A e B ocorrem".

Exemplo 3.5 Considere o lançamento de um dado e defina dois eventos tais que o evento A corresponda a {O resultado do lançamento é uma face com número par} e o evento B corresponda a {O resultado do lançamento é uma face cujo número é menor que 4}. Então, temos $A = \{2, 4, 6\}$ e $B = \{1, 2, 3\}$.

Eventos A e B ocorrerem simultaneamente significa $A \cap B$ ocorrer, ou seja, $C = A \cap B = \{2\}$.

3.3.4 Regra da união

Toda vez que trabalhamos com dois eventos, A e B, e desejamos saber a probabilidade de que pelo menos um deles ocorra, devemos aplicar a regra da união, $A \cup B =$ "pelo menos um dos eventos, A ou B ocorre",

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

Exemplo 3.6 Considere o lançamento de um dado equilibrado. Nesse caso, temos $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$. Considere os eventos A e B tais que A = "O resultado do lançamento é par" e B = "o resultado do lançamento é maior que 4".

Nesse caso, podemos obter a probabilidade de cada um dos eventos usando a propriedade da união de eventos disjuntos, ou seja,

$$P(A) = P(\{2, 4, 6\}) = P(\{2\} \cup \{4\} \cup \{6\}) = P(\{2\}) + P(\{4\}) + P(\{6\})$$

$$= 1/6 + 1/6 + 1/6 = 3/6$$

e

$$P(B) = P(\{5, 6\}) = P(\{5\} \cup \{6\}) = P(\{5\}) + P(\{6\}) = 1/6 + 1/6 = 2/6.$$

O evento "ambos A e B ocorrem" somente é atendido pelo resultado {6}, pois é o único que atende às condições "ser par e maior que quatro ao mesmo tempo". Portanto,

$P(A \cap B) = P(\{6\}) = 1/6$.

A probabilidade do evento “A ou B ocorre” pode ser obtida pela regra da união, fazendo

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 3/6 + 2/6 - 1/6 = 4/6.$$

Esse resultado pode ser verificado, já que

$$P(A \cup B) = P(\{2, 4, 5, 6\}) = 4/6.$$

3.3.5 Probabilidade condicional

Considere dois eventos, A e B, relativos a um espaço amostral.

Definição: A probabilidade condicional de que um evento B ocorra dado que A ocorreu, $P(B|A)$, é dada por

$$P(B|A) = P(A \cap B)/P(A),$$

para $P(A) > 0$.

Como consequência direta da definição de probabilidade condicional, podemos escrever que, para dois eventos A e B,

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

Exemplo 3.7 Considere o lançamento de um dado equilibrado.

Evento A: o resultado do lançamento é par;

Evento B: o resultado do lançamento é maior que 3.

Temos, então,

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

$$P(B) = P\{4, 5, 6\} = 3/6 = 1/2$$

e

$$P(A \cap B) = P\{4, 6\} = 2/6 = 1/3.$$

Portanto,

$$P(B|A) = P(A \cap B)/P(A) = \frac{1/3}{1/2} = 2/3.$$

Podemos chegar ao mesmo resultado de forma intuitiva. Quero saber a probabilidade de uma face maior que 3 ocorrer no lançamento, mas já tenho a informação de que o resultado é par. Portanto, os resultados possíveis ficam reduzidos a {2, 4, 6}. Como todas as faces têm a mesma probabilidade, temos duas possibilidades {4, 6} de que A ocorra, ou seja, uma probabilidade de 2/3.

Exemplo 3.8 Em uma região, temos 12 fábricas, das quais sete são têxteis e cinco são químicas. Entre as químicas, quatro poluem, e, entre as têxteis, apenas duas poluem. Considere a seleção de uma empresa ao acaso e os seguintes eventos,

Evento C: a empresa é têxtil;

Evento D: a empresa polui.

Por consequência, o complementar de C, denotado por C^c, equivale à indústria química, e o complementar de D, D^c, equivale a "a empresa não polui", e, portanto,

$$P(C) = 7/12$$

$$P(C^c) = 1 - P(C) = 1 - 7/12 = 5/12$$

$$P(D) = 6/12 = 1/2$$

$$P(D^c) = 6/12 = 1/2.$$

Se não tivermos nenhuma informação extra, uma empresa selecionada ao acaso terá probabilidade de 1/2 de ser uma empresa que polui. Mas se tivermos informação adicional, por exemplo, de que se trata de uma empresa química, a probabilidade de ser poluente mudará para 4/5, ou seja,

$$P(D|C^c) = P(C^c \cap D)/P(C^c) = \frac{4/12}{5/12} = 4/5.$$

Exemplo 3.9 Probabilidade Condicional.

Considere uma fábrica de autopeças feitas de plástico, com duas linhas de produção, I e II, que produzem dois produtos, A e B. Valores históricos indicam que a geração de itens não conformes com as especificações (refugos que são descartados) se distribuem como mostrado pelas proporções na Tabela 3.1, a seguir.

Tabela 3.1 Proporção de refugos.

	A	B	Total
I	0,3	0,4	0,7
II	0,1	0,2	0,3
Total	0,4	0,6	1,0

Note que os valores apresentados na Tabela poderiam ser expressos em porcentagens.

Ponderando que os valores históricos podem ser considerados como boas aproximações das probabilidades dos eventos, assim a probabilidade de que um item refugado escolhido ao acaso seja do produto A e tenha sido produzido pela linha II é 0,1, as seguintes probabilidades podem ser calculadas:

i. Selecionando um refugo do produto A ao acaso, a probabilidade de que o mesmo tenha sido produzido pela linha I é dada pela probabilidade condicional

$$\Pr(\mathrm{I}|\mathrm{A}) = \Pr(\mathrm{I}, \mathrm{A})/\Pr(\mathrm{A}) = 0,3/0,4 = 3/4.$$

ii. Selecionando ao acaso um item refugado proveniente da linha II, a probabilidade que seja do produto B é

$$\Pr(\mathrm{B}|\mathrm{II}) = \Pr(\mathrm{II}, \mathrm{B})/\Pr(\mathrm{II}) = 0,2/0,3 = 2/3.$$

Notas:

1. Todo condicionamento significa acrescentar mais informação sobre o evento de interesse;
2. Todo condicionamento impõe uma restrição ao espaço amostral. No Exemplo 3.6, o número de elementos caiu de 6 para 3 e, no exemplo das indústrias, caiu de 12 para 5.

3.3.6 Independência

Definição: Dois eventos A e B serão ditos independentes se

$$P(B|A) = P(B).$$

Caso essa condição não seja satisfeita, os eventos são ditos dependentes. Dizer que os eventos A e B são independentes equivale a dizer que $P(A \cap B) = P(A)\,P(B)$.

Em outras palavras, dois eventos serão independentes se o fato de um deles ter ocorrido – ou não ter ocorrido – não afetar a probabilidade de o outro ocorrer.

Exemplo 3.10 Retomando o Exemplo 3.5, temos que os eventos A e B são dependentes, já que

$$P(B) = 1/2 \neq P(B|A) = 2/3.$$

Definindo um evento E, de forma que

E: {o resultado do lançamento é maior que 2}.

Nesse caso, temos $P(E) = P\{3, 4, 5, 6\} = 4/6 = 2/3$, e os eventos A e E são independentes, já que

$$P(E|A) = P(A \cap E)/P(A) = \frac{1/3}{1/2} = 2/3 = P(E).$$

No caso das fábricas (Exemplo 3.8), os eventos C e D são dependentes.

3.3.7 Variáveis aleatórias discretas

Se Ω é discreto, variável aleatória (VA) é uma função que associa a cada elemento do espaço amostral um número real. As VAs podem ser discretas ou contínuas.

Definição: Variável aleatória discreta é uma função definida sobre o espaço amostral e que pode assumir um número finito de valores ou um número infinito que seja contável.

Exemplo 3.11 Retomamos o exemplo do lançamento da moeda, com resultados possíveis C e $\underline{C}$. Nesse caso, podemos definir uma VA X, associando o valor X = 0 ao evento C e X = 1 ao evento $\underline{C}$. A Figura 3.1 ilustra a relação entre espaço amostral e variável aleatória, no caso do lançamento de uma moeda.

Exemplo 3.12 No caso do lançamento do dado, podemos associar a cada face o seu valor inteiro. Nesse caso associamos X = 1 com a face 1 do dado, X = 2 com a face 2, e assim por diante, até X = 6.

A variável aleatória definida no exemplo 3.12 não é única e muitas outras poderiam ser usadas.

Apresentamos, a seguir, um exemplo de VA que assume um número infinito contável de valores.

Exemplo 3.13 Considere a situação em que o fenômeno de interesse é o tráfego em uma rodovia, mais especificamente o número de veículos que passam na rodovia em período de tempo predeterminado (por exemplo, em um minuto). Nesse caso, a VA associada pode assumir valores no conjunto dos números naturais, $N = \{0, 1, 2, \ldots\}$, que é um conjunto infinito contável – podemos, assim, enumerar os seus componentes.

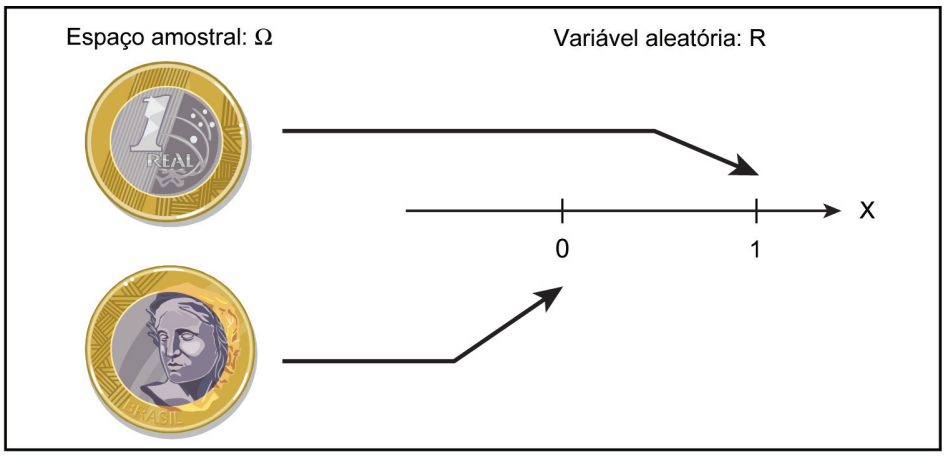


Figura 3.1 Espaço amostral e variável aleatória.

3.3.8 Valor esperado de uma VA discreta

O valor esperado de uma VA discreta X, denotado por E[X], também chamado de valor médio, é a média dos valores que a VA pode assumir ponderada pelas respectivas probabilidades.

Definição: Considere uma VA X que pode assumir os valores x_1, x_2,...,x_n, o valor esperado de X é definido como

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

Exemplo 3.13 Retomando o Exemplo 3.2, do lançamento de uma moeda, com uma VA que assume valores em $\{0, 1\}$, com probabilidades 0,5 cada um, temos

$$E[X] = 0P(X = 0) + 1P(X = 1) = 0(0,5) + 1(0,5) = 0,5.$$

Nesse caso, o valor esperado ou valor médio da VA X é 0,5.

Exemplo 3.14 No exemplo do lançamento do dado, a VA X pode assumir valores em $\{1, 2, 3, 4, 5, 6\}$, com probabilidades 1/6 cada um

$$E[X] = 1P(X = 1) + 2P(X = 2) + 3P(X = 3) + 4P(X = 4) + 5P(X = 5) + 6P(X = 6)$$

$$= (1 + 2 + 3 + 4 + 5 + 6)/6 = 21/6 = 3,5.$$

O valor esperado, nesse caso, é 3,5.

O valor esperado de VAs tem propriedades que podem ser exploradas na sua aplicação. Apresentamos a seguir duas dessas propriedades. Considere que X e Y sejam duas variáveis aleatórias e *a* e *b* duas constantes, então podemos afirmar que

1. $E[a\,X] = a\,E[X]$

e

2. $E[a\,X + b\,Y] = a\,E[X] + b\,E[Y]$.

Exercício: Aplique essas propriedades aos exemplos acima e verifique a sua validade.

Exemplo 3.15 Considere uma unidade de produção de um produto Y. O produto pode estar em conformidade com as especificações, e, então, denotamos por C (Conforme); ou pode não estar em conformidade com as especificações, nesse caso denotamos por NC (Não Conforme). O espaço amostral que representa os possíveis resultados para cada unidade produzida é Ω = {C, NC}. Historicamente, sabemos que a proporção de não conformes é 0,01, o que nos permite usar esse valor como uma aproximação para a probabilidade de que uma determinada unidade seja NC.

Considerando que cada produto em conformidade com as especificações permite um ganho à empresa de R$ 0,10, enquanto que um NC é descartado e produz uma perda de R$ 1,20, qual é o ganho esperado em cada unidade produzida?

Para responder a essa pergunta, basta definir uma VA X que associa C ao valor 10 e NC ao valor –120, que correspondem ao ganho e perda em centavos. Calculamos então o valor esperado de X,

$$E[X] = 10\,P(X=10) + (-120)\,P(X=120) = 10(0,99) + (-120)(0,01) = 8,7.$$

Logo, o ganho esperado por unidade produzida é de 8,7 centavos.

3.3.9 Variância de uma VA discreta

Definição: A variância de uma VA discreta X é definida como

$$\mathrm{Var}(X) = E\left[\left(x - E[X]\right)^2\right] = \sum_{i=1}^{n} \left(x_i - E[X]\right)^2 P(X = x_i).$$

O desvio padrão de X, dp(X), é a raiz quadrada positiva da variância de X.

Exemplo 3.16 Retomando o exemplo da moeda, a variância é dada por

$$\mathrm{Var}(X) = E\left[\left(x - E[X]\right)^2\right] = \sum_{i=1}^{n} \left(x_i - E[X]\right)^2 P(X = x_i)$$

$$= (0-0,5)^2\,(0,5) + (1-0,5)^2\,(0,5) = 0,125 + 0,125 = 0,25,$$

e o desvio padrão é $\mathrm{dp}(X) = \sqrt{0,25} = 0,5$.

Exercício: Use as informações do exemplo do dado e calcule a variância e o desvio padrão. Note que o valor médio e a variância de uma VA têm naturezas diferentes da média e variância estudadas em estatística descritiva. Aqui, estas dizem respeito à VAs, ou seja, a possíveis valores que ainda não se realizaram. Já em estatística descritiva, os valores dizem respeito a um conjunto de dados que estamos descrevendo ou sumarizando. A relação entre esses objetos de nomes semelhantes ficará mais clara na seção de inferência.

A variância de uma VA apresenta as seguintes propriedades. Considere que X seja uma variável aleatória e *a* uma constante, então podemos afirmar que

1. $\mathrm{Var}[a + X] = \mathrm{Var}[X]$

e

2. $\mathrm{Var}[a\,X] = a^2\ \mathrm{Var}[X]$.

Exercício: Considere que no lançamento de uma moeda, se o resultado for cara há um ganho de dois reais e no caso de coroa há uma perda de um real. Quais são o ganho esperado e a variância a cada lançamento? E qual é o desvio padrão? Comente os resultados.

3.3.10 Distribuições probabilísticas discretas

A distribuição de probabilidades de uma VA X é uma descrição das probabilidades associadas aos valores possíveis de X. No caso de uma VA discreta, a distribuição pode ser especificada por meio de uma lista de valores possíveis e suas respectivas probabilidades, como nos exemplos da moeda e do dado. Em alguns casos, é conveniente recorrer a uma fórmula matemática, a função de probabilidade.

Definição: A função de probabilidade de uma VA discreta X é

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

para todo $x \in R$. Esta é uma função definida em R e com valores em R.

Considerando que a VA X pode assumir os valores x_1, x_2, ..., então $P(X = x_1)$, $P(X = x_2)$, $P(X = x_3)$,... é a distribuição de probabilidade da VA X.

Exemplo 3.17 Considere um experimento cujos eventos, $\Omega = \{A, B, C, D\}$, têm probabilidades $\{0,1; 0,3; 0,2; 0,4\}$, respectivamente. Definimos uma VA X de forma que $X(A) = 1$, $X(B) = 2$, $X(C) = 3$ e $X(D) = 4$.

Na Figura 3.2, apresentamos o gráfico da função de probabilidade da VA X definida no exemplo precedente.

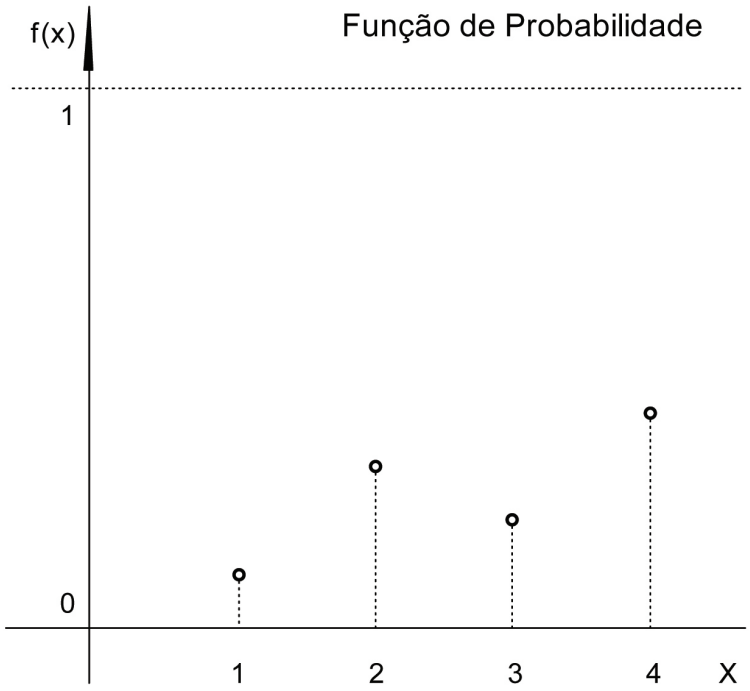


Figura 3.2 Função de probabilidade da VA X.

Como são definidos a partir de uma probabilidade, os valores que a função de probabilidade pode assumir são todos positivos ou nulos, $f(x_i) \geq 0$, e sua soma é sempre igual a 1, $\sum_{i=1}^{n} f(x_i) = 1$.

Definição: A função de distribuição acumulada (fda), F(x), de uma VA X é definida para todo $x \in R$ como a soma de todas as probabilidades para x_i menor ou igual a x, ou seja, $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$.

Uma característica importante de F(x) é que, à medida que x vai para $-\infty$ (menos infinito), F(x) vai para zero e, à medida que x vai para ∞, F(x) vai para 1, como pode ser verificado na Figura 3.3.

A Figura 3.3 apresenta o gráfico da função de distribuição acumulada para a VA X definida no exemplo do dado.

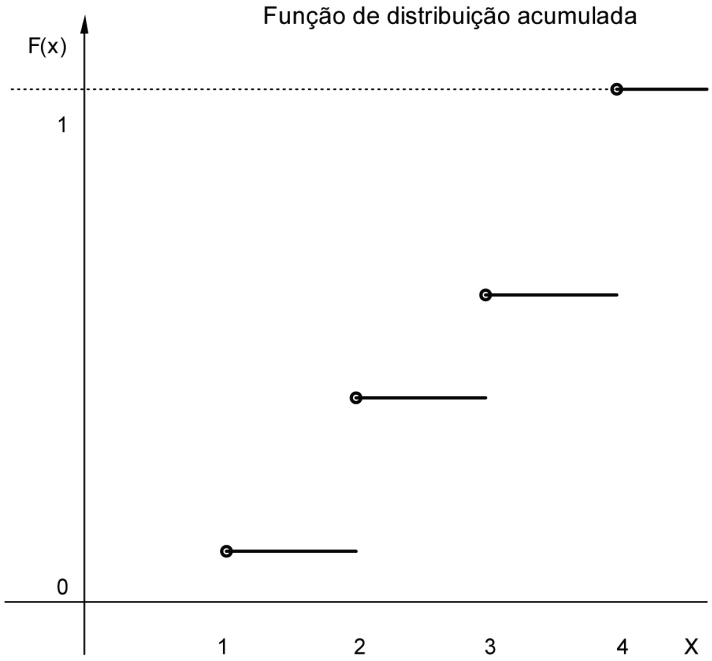


Figura 3.3 Função de distribuição acumulada para a VA X.

3.3.11 Algumas distribuições discretas mais comuns

Apresentamos, a seguir, algumas das distribuições discretas mais utilizadas.

3.3.11.1 Distribuição uniforme discreta

Definição: Uma VA X assumindo valores x_1, x_2,..., x_n terá distribuição uniforme discreta se

$$f(x_i) = P(X = x_i) = 1/n,$$

para todo i = 1, 2,..., n.

A fda é dada por

$$F(x) = P(X \leq x) = \sum_{i:x_i \leq x} f(x_i) = \frac{\text{número de } x_i\text{'s menores ou iguais a } x}{n},$$

para todo $x \in R$

O que caracteriza a distribuição uniforme discreta é que todos os valores possíveis de X têm a mesma probabilidade.

Exemplo 3.18 No lançamento de um dado equilibrado, todas as faces têm a mesma probabilidade de ocorrer, 1/6. Nesse caso, temos

$$P(X=5)=1/6$$

$$P(X<3)=P(X\leq 2)=P(X=1)+P(X=2)=1/6+1/6=2/6$$

$$P(X>3)=P(X\geq 4)=P(X=4)+P(X=5)+P(X=6)=3/6=1/2.$$

No exemplo do lançamento do dado equilibrado, as formas da função de probabilidade e função de probabilidade acumulada são apresentadas na Figura 3.4.

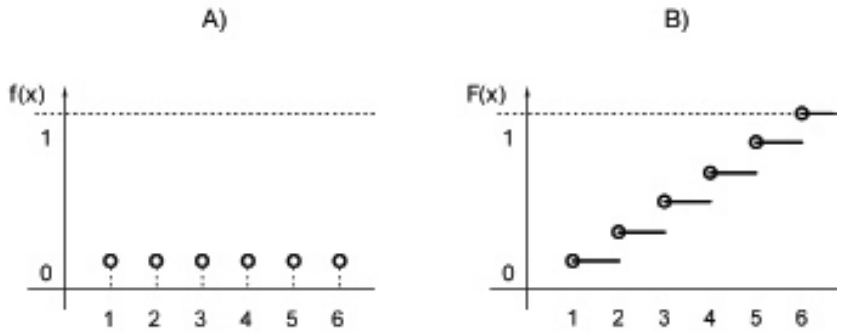


Figura 3.4 A) Função de probabilidade; e B) Função de probabilidade acumulada para a distribuição uniforme discreta.

Nota: Podemos observar, no gráfico, que a fda é contínua à direita, ou seja, limite de $F(x)=F(x_i)$, quando x tende a x_i pela direita, ou seja,

$$\lim_{x\downarrow x_i} F(x)=F(x_i).$$

Quando x tende a x_i pela esquerda, temos

$$\lim_{x\uparrow x_i} F(x)=F(x_i)-1/n.$$

3.3.11.2 Distribuição de Bernoulli

Definição: Uma VA X que assume apenas os valores zero e 1, com probabilidades

$$P(X=0)=(1-p) \text{ e } P(X=1)=p,$$

é chamada de VA de Bernoulli, e a distribuição correspondente, distribuição de Bernoulli com parâmetro p, sendo denotada por Ber(p).

Se X tem distribuição Ber(p), então

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

A função de probabilidade e a função de distribuição acumulada são apresentadas na Figura 3.5.

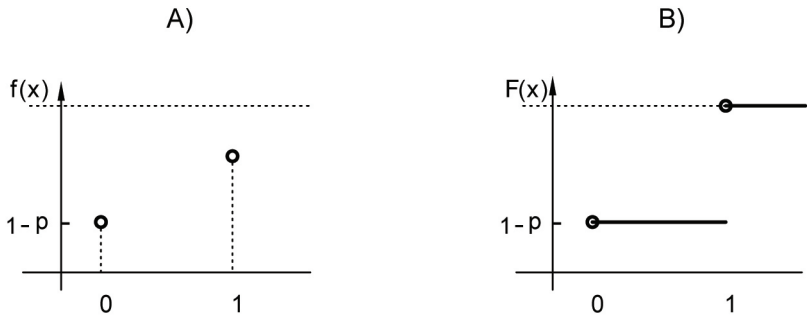


Figura 3.5 A) fdp da Bernoulli(p); e B) fda da Bernoulli(p).

Por ser uma VA com resposta binária, isto é, que admite apenas dois possíveis resultados, essa VA encontra uma vasta gama de aplicação. Em geral, todos os fenômenos cujo resultado é do tipo sucesso ou fracasso podem ser modelados por essa distribuição. Citamos, a seguir, alguns exemplos de aplicação:

- Numa linha de produção, uma peça pode ser considerada conforme (C) ou não (NC) às especificações;
- Num estudo sobre saúde pública, um indivíduo pode ser portador ou não de determinada doença;
- Uma compra com cartão de crédito pode ser classificada como fraude ou não;
- Um córrego pode ser classificado como contaminado ou não;
- Uma oferta de um novo produto a um cliente pode ser bem sucedida ou não.

Exemplo 3.19 O lançamento da moeda é o exemplo clássico de uma distribuição de Bernoulli, com parâmetro p = P(C) = P(cara). No caso de uma moeda equilibrada temos p = 0,5.

Exemplo 3.20 Considere uma unidade de produção de parafusos. Cada parafuso produzido pode estar em conformidade com as especificações de produção, denotada por C, ou não estar em conformidade, NC. Associamos C com 1 e NC com 0. Temos, então, uma VA de Bernoulli. Se p = 0,98, teremos $P(X=1)=0,98$.

Exercício: Considere que a VA X tem distribuição binomial com parâmetros n e p, X~bin(n,p). Mostre que se fixamos o valor de n em 1, n = 1, então a VA resultante tem distribuição Bernoulli com parâmetro p, X~Ber(p).

3.3.11.3 Distribuição binomial

Definição: Uma VA X que pode assumir os valores 0, 1,..., n, onde n é um número natural maior ou igual a 1, com probabilidades

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \qquad \text{para } k=0,1,\ldots,n,$$

tem distribuição binomial com parâmetros n e p, sendo denotada por bin(n,p), e $\binom{n}{k}=\frac{n!}{k!(n-k)!}$ é a combinação de n, k-a-k. Se n = 1, então X tem distribuição Ber(p), descrita na seção anterior.

Se X tiver distribuição Bin(n,p), então

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

A função de probabilidade e a função de distribuição acumulada de uma Bin(5;0,2) são apresentadas na Figura 3.6.

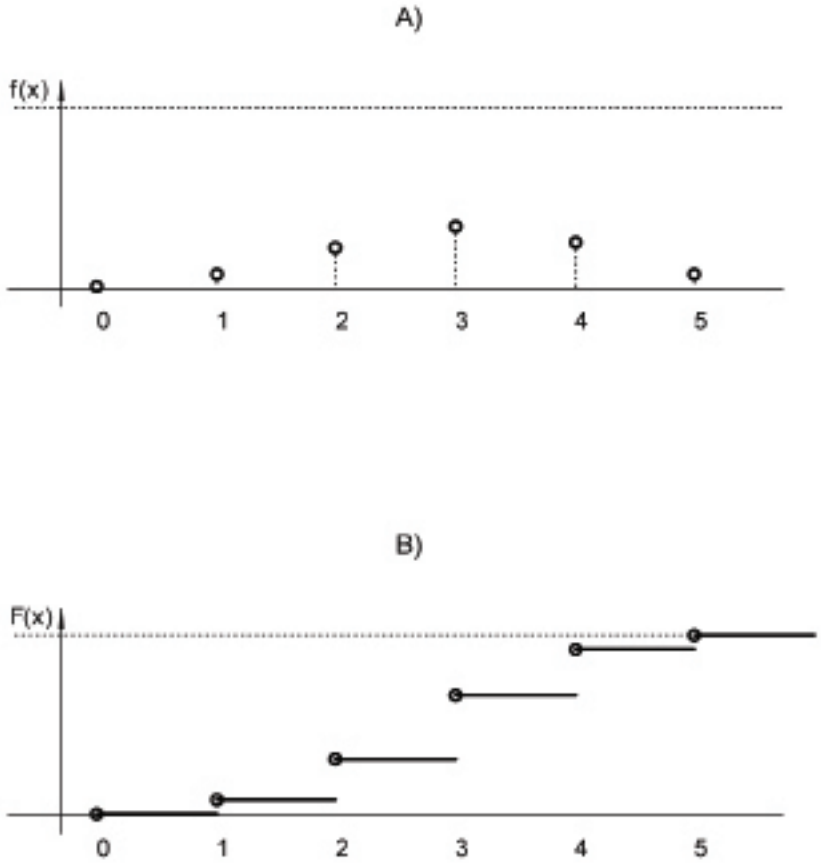


Figura 3.6 A) Função de probabilidade; e B) Função de distribuição acumulada para a distribuição Bin(5;0,6).

Exemplo 3.21 Um lote de peças tem 10 unidades, cada peça com probabilidade p = 0,002 de ser NC. A variável de interesse é o número de peças NC no lote. A distribuição, nesse caso, é a binomial com parâmetros n = 10 e p = 0,002, ou Bin(n = 10, p = 0,002).

Podemos dizer então que a probabilidade de zero NC's no lote é 0,98 já que aplicando a definição acima temos

$$P(X=0)=\binom{10}{0}(1-0{,}002)^{10}=0{,}98.$$

Podemos determinar no R a função de probabilidade e a função de probabilidade acumulada para a distribuição binomial utilizando as instruções mostradas no quadro que segue.

```
#
# Cálculo da probabilidade do exemplo acima
dbinom(0,10,.002)
#
# Cálculo das probabilidades da distribuição Bin(10,0,4)
# para valores de x entre 0 e 5
dbinom(0:5,10,0.4)
#
# Cálculo da fda da distribuição Bin(10,0,4) para x=4
pbinom(4,10,0.4)
```

Relação entre as distribuições Bernoulli e binomial

- Na distribuição binomial com parâmetros n e p, se fixarmos n = 1 obtemos a distribuição de Bernoulli com parâmetro p.
- Se realizarmos n experimentos de Bernoulli com probabilidade de sucesso p e contarmos o número de sucessos o resultado dessa contagem tem distribuição binomial com parâmetros n e p. Essa contagem corresponde a soma das VA's de Bernoulli já que estamos lidando com valores zeros e uns apenas.

3.3.11.4 Distribuição de Poisson

Definição: Uma VA X que assume os valores N = {0, 1, 2, ...}[9], com probabilidades

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!} \qquad \text{para } k = 0, 1, \ldots,$$

tem distribuição Poisson com parâmetro λ, onde λ é um número real estritamente positivo, e é denotada por Pois(λ).

Se X tiver distribuição de Poisson com parâmetro λ, então

$$E[X] = \lambda \text{ e } Var[X] = \lambda.$$

9 Neste caso, os valores que uma VA Poisson pode assumir vão ao infinito, daí a notação "...".

A distribuição de Poisson tem ampla aplicação na modelagem de fenômenos envolvendo contagem. Apresentamos, a seguir, alguns exemplos:

1. Número de veículos passando por certo ponto de uma rodovia em um intervalo de tempo predeterminado;
2. Número de chamadas telefônicas chegando a uma central em um intervalo de tempo predeterminado;
3. Número de acidentes ocorridos em uma fábrica no período de um ano.

A função de probabilidade é apresentada na Figura 3.7.

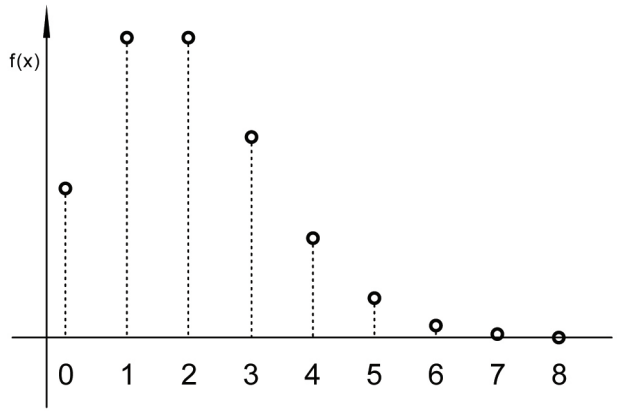


Figura 3.7 Função de probabilidade para a Poisson com $\lambda = 4$.

Exemplo 3.22 Numa central telefônica, o número de chamadas que chegam por unidade de tempo segue a distribuição de Poisson com média 7. Então, temos

$$P(X = 8) = \frac{e^{-7}7^8}{8!} = 0{,}130$$

$$P(X < 7) = P(X \leq 6) = \sum_{i=0}^{6} \frac{e^{-7}7^i}{i!} = 0{,}0009 + 0{,}0064 + 0{,}022 + \ldots + 0{,}149 = 0{,}450$$

e

$$P(X > 8) = P(X \geq 9) = 1 - P(X \leq 8) = 1 - \sum_{i=0}^{8} \frac{e^{-7}7^i}{i!} = 1 - 0{,}729 = 0{,}271$$

Podemos determinar as probabilidades da distribuição de Poisson no R usando as instruções do quadro abaixo.

```
#
#  Cálculo da probabilidade P(X=5) para uma Pois(4)
dpois(5,4)
#  Cálculo da fda P(X≤5) para uma Pois(4)
ppois(5,4)
```

3.3.12 Variáveis aleatórias contínuas

Definição: Variável aleatória absolutamente contínua é uma função definida sobre o espaço amostral e que assume valores em um ou mais intervalos no conjunto dos números reais, R.

Exemplo 3.23 No caso em que a variável de interesse é o peso do arroz em uma embalagem de 1 kg, temos uma VA contínua. Podemos definir como espaço amostral o conjunto de números reais pertencentes ao intervalo (0,2). Note que o intervalo contém infinitos elementos e os mesmos não são enumeráveis.

Definição: A função f(x) é uma função de densidade de probabilidade (fdp) de uma VA se for não negativa, ou seja, assumir somente valores positivos ou nulos, e sua integral for um,

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

A área sob a curva f(x), entre dois valores a e b, nos dá a probabilidade de a variável pertencer ao intervalo [a,b],

$$P(a \le X \le b) = \int_{a}^{b} f(x)\,dx.$$

Definição: O valor esperado de uma variável aleatória contínua é definido como

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\,dx.$$

Definição: A variância de uma variável aleatória absolutamente contínua é definida como

$$\mathrm{Var}[X] = E\left[\left(x - E[X]\right)^2\right] = \int_{-\infty}^{\infty} \left(x - E[X]\right)^2 f(x)\ dx.$$

Definição: A função de distribuição acumulada (fda) de uma variável aleatória absolutamente contínua X é

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\ dx,$$

para todo $x \in R$.

Uma utilização interessante para a fda é que ela pode ser usada para determinar a probabilidade da VA assumir valores no intervalo (a, b] sem a necessidade de integração, isto é,

$$P(a \leq X \leq b) = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = F(b) - F(a).$$

Definição: Para uma VA X, absolutamente contínua, o quantil p, denotado por Q(p), é definido com sendo o número real q(p), cuja fda é igual a p,

$$F(Q(p)) = p.$$

A fda F(x) admite função inversa, que denotamos por $F^{-1}(y)$, definida para todo $y \in (0,1)$, e $F^{-1}(y)$ é tal que $F^{-1}(p) = q(p)$. O quantil pode ser visto como a imagem de p pela função inversa da função de distribuição acumulada, $F^{-1}(y)$.

3.3.13 Distribuições probabilísticas contínuas

Apresentamos, a seguir, algumas distribuições de probabilidades absolutamente contínuas.

3.3.13.1 Distribuição uniforme

Definição: Uma VA X terá distribuição uniforme com parâmetros a e b, $-\infty < a < b < +\infty$, denotada por $X \sim U(a,b)$, se a sua função de densidade de probabilidade for

$$f(x; a, b) = \frac{1}{b-a}, \qquad \text{se } x \in [a, b],$$

e $f(x; a, b) = 0$, caso contrário. O valor esperado e a variância são dados por

$$E[X] = \frac{a+b}{2} \quad \text{e} \quad Var[X] = \frac{(b-a)^2}{12},$$

respectivamente.

A função de distribuição acumulada de uma VA uniforme é F(x) = 0 para x < a, é

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\, dx = \int_{a}^{x} \frac{1}{b-a} dx = \frac{x-a}{b-a},$$

para $x \in [a, b]$, e $F(x) = 1$, para $x \geq b$.

Exemplo 3.24 Considere a VA $X \sim U(0, 1)$. A probabilidade de $X \in [0,5; 0,8]$ é dada pela integral

$$P(0,5 \leq x \leq 0,8) = \int_{0,5}^{0,8} \frac{1}{b-a} dx = F(0,8) - F(0,5) = 0,3.$$

Nesse caso, o valor esperado e a variância são dados por

$$E[X] = \frac{a+b}{2} = \frac{0+1}{2} = 0 \quad \text{e} \quad Var[X] = \frac{(b-a)^2}{12} = \frac{(1-0)^2}{12} = \frac{1}{12},$$

respectivamente.

A Figura 3.8 mostra os gráficos da função de densidade e função de distribuição acumulada de uma VA com distribuição U(0,1).

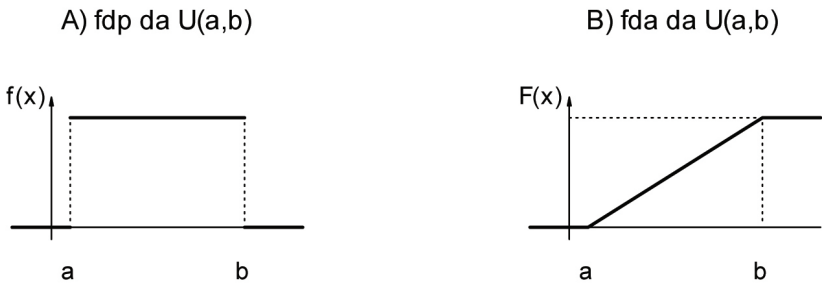


Figura 3.8 A) fdp de uma U(0,1); e B) fda da U(0,1).

Podemos determinar as probabilidades da distribuição uniforme no R utilizando as instruções mostradas no quadro que segue.

```
#
#  Cálculo da densidade de uma U(3,5) no ponto 4
dunif(4,3,5)
#
#  Cálculo da fda para uma U(-1,2) no ponto 1
punif(1,-1,2)
#
#  Inicializa p com probabilidades
p=c(0.25,0.5,0.75)
#
#  Cálculo dos quantis da distribuição U(5,9)
qunif(p,5,3)
#
```

3.3.13.2 Distribuição normal

Definição: Uma VA X terá distribuição normal com parâmetros μ e σ^2, denotada por $X \sim N(\mu, \sigma^2)$, se a sua função de densidade de probabilidade for

$$f(x; m, s^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right),$$

para $-\infty < x < \infty, -\infty < \mu < \infty$ e $0 < \sigma^2 < \infty$.

O valor esperado e a variância para uma VA X com distribuição normal são

$$E[X] = \mu \quad \text{e} \quad Var[X] = \sigma^2,$$

respectivamente.

No caso em que $\mu = 0$ e $\sigma^2 = 1$, temos a distribuição normal padrão, denotada por $Z \sim N(0,1)$, e cuja densidade é dada por

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right),$$

para $-\infty < z < \infty$.

Um resultado importante, relacionado à distribuição normal, estabelece que, se $X \sim N(\mu, \sigma^2)$, então esta pode ser reduzida à normal padrão por meio da transformação

$$Z = \frac{X - \mu}{\sigma},$$

em que $Z \sim N(0,1)$. A essa transformação damos o nome de padronização.

O caminho inverso também é possível, partindo de uma VA Z com distribuição normal padrão podemos obter uma VA X com distribuição $N(\mu, \sigma^2)$ através da transformação

$$X = \mu + \sigma Z.$$

A função de distribuição acumulada da distribuição normal padrão é

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(z)\,dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(\frac{-z^2}{2}\right) dz.$$

Essa integral não pode ser calculada analiticamente. Assim sendo, $\Phi(z)$ é obtida numericamente, e seus valores encontram-se tabelados. A Figura 3.9 mostra a fdp e a fda da distribuição normal padrão.

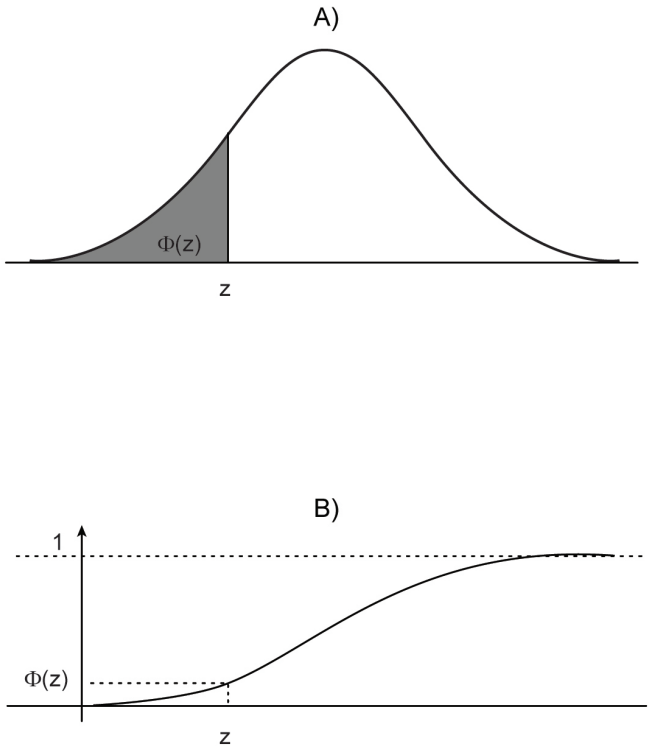


Figura 3.9 A) Função de densidade; e B) fda da $N(0,1)$.

Para determinar a probabilidade do evento $(a < X < b)$, ou $X \in (a,b)$, com $X \sim N(m, s^2)$, temos

$$P(X \in (a,b)) = P(a < X < b).$$

Como não podemos integrar analiticamente a densidade da normal e somente a normal padrão está tabelada, explicitamos o evento (a < X < b) em função da normal padrão, $Z = \frac{X-\mu}{\sigma}$. Isso é feito subtraindo μ e dividindo por σ em todos os termos da desigualdade, ou seja,

$$P(a < X < b) = P(a - m < X - m < b - m) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) =$$

$$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

$\Phi(z)$ está tabelada para valores de $z \in [-3,9; 0]$. Para valores de z maiores que zero, usamos a propriedade de simetria da normal padrão. Com base nessa propriedade, temos que $\Phi(z) = 1 - \Phi(-z)$.

Exemplo 3.25 Para $X \sim N(10, 4)$, determine as probabilidades a) $P(X < 12)$, b) $P(9 < X < 12)$ e c) $P(X > 9)$.

a) $P(X < 12) = \Phi\left(\frac{12-10}{2}\right) = \Phi(1) = 1 - \Phi(-1) = 1 - 0{,}158655 = 0{,}841345$

b) $P(9 < X < 12) = \Phi\left(\frac{12-10}{2}\right) - \Phi\left(\frac{9-10}{2}\right) = \Phi(1) - \Phi(0{,}5)$

$= 0{,}841345 - 0{,}308538 = 0{,}532807$

c) $P(X > 9) = 1 - P(X < 9) = 1 - \Phi\left(\frac{9-10}{2}\right) = 1 - \Phi(-0{,}5)$

$= 1 - 0{,}308538 = 0{,}691462$

Podemos determinar as probabilidades associadas à distribuição normal com o R utilizando as instruções mostradas no quadro que segue.

```
#
# Inicializa os valores de x
x=c(8,10,11)
#
# Cálculo da densidade de uma N(10,4) nos pontos x
dnorm(x,10,2)
#
# Cálculo da fda para uma N(10,4) nos pontos x
pnorm(x,10,2)
#
# Inicializa p com probabilidades
p=c(0.25,0.5,0.75)
#
# Cálculo dos quantis da distribuição N(5,9)
qnorm(p,5,3)
```

3.3.13.3 Distribuição t de Student

Definição: Uma VA X terá distribuição t de Student com v graus de liberdade, denotada por t(v), se a sua função de densidade de probabilidade for

$$f(x;v)=\frac{\Gamma\left((v+1)/2\right)}{\Gamma\left(v/2\right)\sqrt{v\pi}}\left(1+x^2/v\right)^{-(v+1)/2}$$

para $-\infty < x < \infty$.

Se X for uma VA com distribuição t de Student com v graus de liberdade, então

$$E[X]=0 \text{ e } Var[X]=\frac{v}{v-2}, \text{ para } v > 2.$$

No que diz respeito à forma, as distribuições normal padrão e t de Student são muito similares, como mostra o gráfico na Figura 3.10. A diferença entre elas é que a distribuição t de Student tem caudas mais pesadas, tanto mais pesadas quanto menor for o número de graus de liberdade.

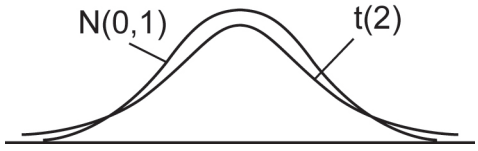


Figura 3.10 Comparação das densidades das distribuições N(0,1) e t(2).

Exemplo 3.26 Para X com distribuição t de Student, com cinco graus de liberdade, $X \sim t(5)$, obtemos as seguintes probabilidades usando a tabela.

$$P(X < -2) = 0,05$$

$$P(X < 2) = 0,949$$

$$P(X > 2) = 1 - P(X < 2) = 1 - 0,949 = 0,051$$

$$P(-1 < X < 1) = 0,818 - 0,182 = 0,636$$

Podemos determinar probabilidades associadas à distribuição t de Student utilizando as instruções do R mostradas no quadro.

```
#
# Inicializa os valores de x
x=c(-1,0.5,2)
#
# Cálculo da densidade de uma t de Student com 9
# graus de liberdade nos pontos x
dt(x,9)
#
# Cálculo da fda para uma t de Student com 9
# graus de liberdade nos pontos x
pt(x,9)
#
# Inicializa p com probabilidades
p=c(0.25,0.5,0.75)
#
# Cálculo dos quantis da t de Student com 9
qt(p,9)
```

3.3.13.4 Distribuição qui-quadrado

Definição: Uma VA X terá distribuição qui-quadrado com v graus de liberdade, $X \sim \chi^2(v)$, se a sua função de densidade de probabilidade for

$$f(x;v) = \frac{1}{\Gamma(v/2)2^{v/2}} x^{v/2-1} \, e^{-y/2},$$

para $0 < x < \infty$, onde v é um número real estritamente positivo.

Se X for uma VA com distribuição qui-quadrado com v graus de liberdade, então

$$E[X] = v \;\; e \;\; Var[X] = 2v.$$

Os valores da distribuição qui-quadrado estão tabelados para um conjunto selecionado de v. A forma da distribuição é mostrada no gráfico da Figura 3.11.

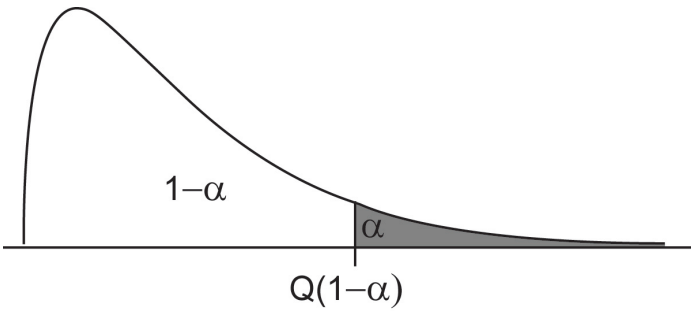


Figura 3.11 Densidade da distribuição $\chi^2(3)$.

Exemplo 3.27 Para $X \sim \chi^2(5)$, a probabilidade $P(X > 6)$ é dada por

$P(X > 6) = 1 - P(X < 6) = 1 - 0,694 = 0,306.$

As instruções listadas no quadro a seguir podem ser utilizadas para calcular valores da função de densidade, função de distribuição acumulada e quantis da distribuição $\chi^2(5)$, respectivamente.

```
#
# Cálculo da função de densidade da qui-quadrado(6) em x=2
dchisq(2,6)
# Cálculo da fda da qui-quadrado(6) em x=2
pchisq(2,6)
# Cálculo do quantil 0,95 da qui-quadrado(6)
qchisq(0.95,6)
```

3.3.13.5 Distribuição F de Snedecor

Definição: Uma VA W terá distribuição F de Snedecor com v_1 e v_2 graus de liberdade, $W \sim F(v_1, v_2)$, se a sua função de densidade de probabilidade for

$$f(w; v_1, v_2) = \left(\frac{v_1}{v_2}\right)^{v_1/2} \frac{w^{(v1-2)/2}}{(1 + v_1 w / v_2)^{(v1+v2)/2}},$$

para $0 < w < \infty$, v1>0 e v2>0.

Se W for uma VA com distribuição F de Snedecor com v_1 e v_2 graus de liberdade, então

$$E[W] = \frac{v_2}{v_2 - 2} \text{ e } Var[W] = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)},$$

para $v_2 > 4$.

Os valores da distribuição F de Snedecor estão tabelados para um par (v_1, v_2) selecionados. A forma da distribuição é mostrada no gráfico da Figura 3.12.

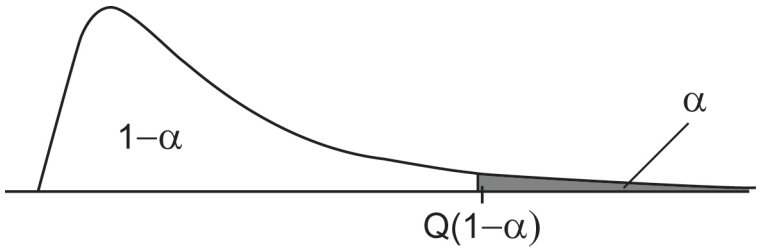


Figura 3.12 Densidade da distribuição F(6,6).

A relação $F(v_1, v_2) = 1/F(v_2, v_1)$ pode ser utilizada para encontrar valores não disponíveis na tabela.

Exemplo 3.28 Para $W \sim F(8, 9)$, a probabilidade $P(W > 4,0)$ é dada por

$$P(W > 4,0) = 1 - P(W < 4,0) = 1 - 0,973 = 0,027.$$

Instruções para uso da tabela F

As instruções listadas no quadro a seguir podem ser utilizadas para calcular valores da função de densidade, função de distribuição acumulada e quantis da distribuição F de Snedecor com v_1 e v_2 graus de liberdade, respectivamente.

```
#
# Cálculo da função de densidade da F(3,6) no ponto x=2
df(2,3,6)
# Cálculo da função de distribuição acumulada da F(3,6) no pon-
to x=2
pf(2,3,6)
# Cálculo do quantil 0,95 da F(3,6)
qf(0.95,3,6)
```

3.3.13.6 Distribuições bivariadas

Definição: Sejam X e Y duas VAs absolutamente contínuas, a função de densidade conjunta para X e Y é definida por

1. $f(x, y) \geq 0$;
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$.

A integral definida de f(x,y), numa determinada região, nos dá a probabilidade de as variáveis pertencerem à região

$$P(a \leq X \leq b,\ c \leq Y \leq d) = \int_{a}^{b} \int_{c}^{d} f(x,y)\,dx\,dy.$$

Definição: A função de distribuição acumulada conjunta é definida como

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

3.3.13.7 Independência

Definição: Duas VAs X e Y, absolutamente contínuas, serão independentes se a função de densidade conjunta de X e Y fatorar no produto das funções densidades de X e Y, ou seja,

$$f(x, y) = f_1(x) f_2(y),$$

para todo x e y reais, sendo $f_1(x)$ e $f_2(y)$ as fdp's de X e Y respectivamente

O mesmo vale para a função de distribuição acumulada conjunta. X e Y serão independentes se

$$F(x,y) = F_1(x)\,F_2(y).$$

A independência entre VAs é importante em estatística, pois nos permite escrever a distribuição conjunta de amostras quando os elementos da amostra são independentes. Isso será tratado na Unidade 4.

3.4 Considerações finais

Nesta terceira Unidade, apresentamos os tópicos necessários para compreensão dos principais conceitos em probabilidade, bem como a aplicação destes aos métodos em inferência, que serão apresentados na próxima Unidade. Todos os tópicos foram acompanhados das instruções para sua implementação no R.

3.5 Atividades de aplicação, prática e avaliação

É altamente recomendável que o leitor exercite a sua capacidade de experimentação com os métodos apresentados e busque, no seu dia a dia, fontes de dados que possam ser utilizados para aplicar os métodos descritos anteriormente. Procure aplicá-los manualmente e utilizando o R.

3.5.1 Atividades individuais

Busque, no seu ambiente de trabalho, conjuntos de dados, aplique as técnicas descritas aqui e veja quais informações você consegue extrair desses dados.

3.5.2 Atividades coletivas

Troque ideias com seus colegas sobre os métodos que você aplicou aos seus dados e verifique se eles(as) utilizariam o mesmo método que você utilizou.

3.6 Estudos complementares

Recomendamos a leitura dos capítulos correspondentes à inferência estatística nos textos citados na lista de referências.

3.6.1 Saiba mais

Você pode ampliar os seus conhecimentos de inferência estatística incluindo novos métodos em seu repertório ou aprofundando seus conhecimentos dos métodos já tratados estudando as referências.

UNIDADE 4

Introdução à Inferência Estatística

4.1 Primeiras palavras

Nesta seção, apresentamos o conceito de população e amostra e aquele que é o ponto central da inferência estatística: “observando apenas uma amostra, obtemos conclusões para a população correspondente, tendo sob controle o grau de incerteza decorrente de não dispor da população como um todo”. Para isso, utilizamos métodos adequados juntamente com as informações sobre probabilidade apresentadas na Unidade anterior.

4.2 Problematizando o tema

Como podemos obter resultados para uma população a partir de uma amostra? Qual é o grau de incerteza associado ao procedimento utilizado?

4.3 Texto básico para estudos

4.3.1 Introdução à inferência estatística

A inferência estatística contempla o estudo de um conjunto de métodos que possibilitam, a partir de uma amostra, concluir sobre toda uma população.

4.3.2 Parâmetros, estimadores, estimativas e estatísticas

Como visto anteriormente, parâmetros são características das distribuições de probabilidades que, por sua vez, caracterizam as variáveis aleatórias, VAs.

Em inferência, os parâmetros serão o nosso foco, o nosso objetivo final, algo que não conhecemos e sobre o qual procuraremos aumentar nosso conhecimento.

Exemplo 4.1 Uma operação industrial produz resíduos que, lançados ao meio ambiente, se transformam em poluição. Cada vez em que a operação é realizada, temos um evento. Em cada evento, a variável de interesse é a quantidade de resíduos gerados, variável que apresenta um comportamento aleatório, isto é, a quantidade não é constante a cada vez em que a operação é realizada. A quantidade produzida é representada por uma variável aleatória. Essa VA, por sua vez, é caracterizada por uma distribuição de probabilidades e seus parâmetros.

Neste nosso exemplo, o interesse está em determinar a quantidade média de resíduos lançada ao meio ambiente, ou seja, nosso interesse é determinar a média da distribuição da VA de interesse. Como sabemos que essa determinação

não será exata, é importante que seja acompanhada de uma medida de incerteza sobre o erro que estaremos cometendo.

Na busca por informações (estimativas) dos parâmetros, utilizamos n replicações da operação, ou seja, utilizamos uma amostra de tamanho n da VA.

Definição: Uma Amostra Aleatória de tamanho n é um conjunto de variáveis aleatórias X_1, X_2,..., X_n independentes e tal que todas têm a mesma distribuição.

Quando temos uma amostra aleatória, dizemos que as variáveis aleatórias são independentes e identicamente distribuídas e a denotamos pela sigla iid.

Definição: Uma estatística é qualquer função das observações em uma amostra aleatória e como tal também é uma VA.

Exemplo 4.2 Dois exemplos de estatísticas muito usadas são a média e a variância amostrais, que são definidas como

$$\overline{X} = \left(X_1 + X_2 + \ldots + X_n\right)/n = \sum_{i=1}^{n} X_i/n$$

e

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

respectivamente.

Definição: Se X for uma VA com uma distribuição de probabilidades definida por uma função de densidade $f(x;\theta)$, caracterizada por um parâmetro desconhecido θ e se $X_1, X_2, \ldots, X_n$ for uma amostra aleatória de X, de tamanho n, então a estatística $\hat{\theta} = g(X_1, X_2, \ldots, X_n)$, utilizada para aproximar θ, será chamada estimador de θ. Note que g(.) é uma função da amostra.

Exemplo 4.3 As estatísticas $\overline{X}$ e S^2 apresentadas anteriormente são estimadores da média e da variância populacionais.

Note que, sendo o estimador uma função de VAs, ele próprio é uma VA.

Definição: Uma estimativa de um parâmetro θ é um valor numérico assumido por um estimador.

Exemplo 4.4 Dois exemplos de estimativas muito usadas são a média e a variância amostrais, definidas após a observação da amostra. Seja x a amostra realizada, ou seja, já com seus valores numéricos, $x = \{x_1, x_2, \ldots, x_n\} = \{21, 23, 21, 16, 24, 21\}$, então as estimativas da média e da variância são determinadas fazendo

$$\bar{x} = (x_1 + x_2 + \ldots + x_n)/n = \sum_{i=1}^{n} x_i/n = 21$$

e

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = 7{,}6,$$

respectivamente.

Note que as VAs são denotadas por letras maiúsculas, enquanto que os valores obtidos para elas através das amostras (o que também chamamos de realizações) são denotados por letras minúsculas.

4.3.3 Distribuições amostrais

Definição: Distribuição amostral é a distribuição de probabilidades de uma estatística.

Exemplo 4.5 Se $X_1, X_2, \ldots, X_n$ for uma amostra aleatória de tamanho n, de uma variável aleatória X, tal que $E[X] = \mu$ e $Var[X] = \sigma^2$, então a média amostral $\bar{X}$ tem valor esperado

$$E\left[\bar{X}\right] = E\left[(X_1 + X_2 + \ldots + X_n)/n\right] = \left(E[X_1] + E[X_2] + \ldots + E[X_n]\right)/n$$

$$= (\mu + \mu + \ldots + \mu)/n = \mu$$

e tem variância

$$\sigma^2\left(\bar{X}\right) = Var[X] = Var\left[(X_1 + X_2 + \ldots + X_n)/n\right] = \left(\sigma^2 + \ldots + \sigma^2\right)/n^2 = \sigma^2/n.$$

Note que na expressão acima, a variância da soma é a soma das variâncias, isso se deve à indenpendência entre as variáveis envolvidas.

A utilização da média amostral como estimador para a média populacional dá margem a questões sobre se essa “aproximação” é adequada e sobre a incerteza associada a essa estimação. Existem vários resultados matemáticos importantes sobre esse tema. Uma abordagem rigorosa sobre ele, no entanto, está totalmente fora do escopo deste livro. Contudo, apresentamos de forma superficial dois resultados importantes: a Lei dos Grandes Números e o Teorema

Central do Limite, que nos permitirão seguir adiante.

4.3.3.1 Lei dos Grandes Números

Considere uma amostra aleatória de tamanho n, X_1, X_2,..., X_n extraída de uma população com média μ. A lei dos grandes números estabelece que a média amostral converge para a média populacional à medida que n cresce.

A demonstração matemática desse resultado está fora do escopo deste livro, mas ilustramos a convergência com o gráfico da Figura 4.1. Os valores mostrados no gráfico é resultado de simulação com valores gerados em computador com parâmetros preestabelecidos. A linha horizontal no centro do gráfico mostra o valor da média populacional, $\mu = 10$, e a linha contínua mostra a evolução da média amostral a medida que aumentamos o tamanho da amostra. É fácil notar que, à medida que n cresce, a média amostral se aproxima cada vez mais da média populacional.

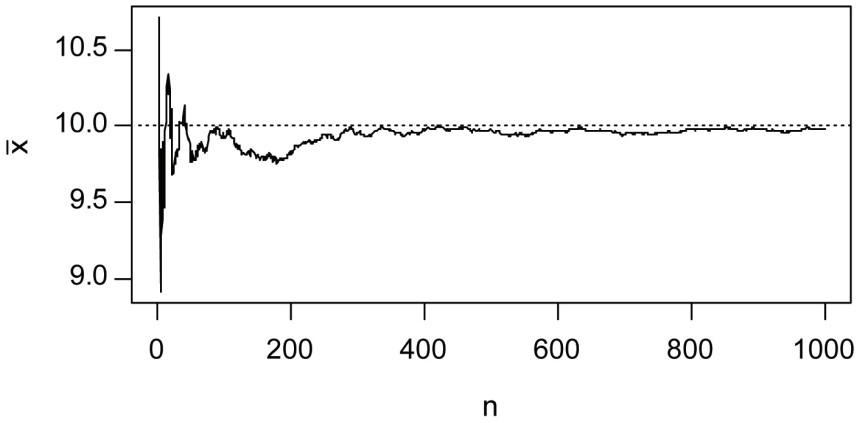


Figura 4.1 Ilustração da Lei dos Grandes Números.

O teorema central do limite, que enunciamos a seguir, de forma sucinta, é um resultado importante para conclusões estatísticas em geral, pois ele associa à estatística S_n a distribuição normal padrão, N(0,1).

4.3.3.2 Teorema Central do Limite – TCL

Sejam X_1, X_2,..., X_n variáveis aleatórias independentes, tais que $E[X_i] = \mu$ e $Var[X_i] = \sigma^2 < \infty$, então a VA $\overline{X}$ padronizada converge para a distribuição normal padrão a medida que n cresce, ou seja,

$$S_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$$

quando $n \rightarrow \infty$. Da mesma forma a fda de S_n converge para a fda da distribuição normal.

Colocando em outras palavras o resultado acima, podemos dizer que “a distribuição da estatística S_n se aproxima da distribuição normal padrão à medida que n cresce”.

O TCL é um resultado importante, pois permite utilizar a aproximação normal para a média de VAs independentes sob condições suaves, ou seja, sob condições que a grande maioria das amostras coletadas no dia a dia atende.

4.3.4 Estimação por ponto

Na estimação por ponto, buscamos aproximar o valor do parâmetro populacional por meio de um número.

Existem diversos métodos para determinar estimadores. Apresentaremos, neste texto, dois métodos de estimação: o de máxima verossimilhança, que é apresentado logo a seguir, e o de mínimos quadrados, apresentado na seção sobre Regressão Linear.

4.3.4.1 Estimador de Máxima Verossimilhança – EMV

Como o nome indica, esse método obtém as estimativas dos parâmetros maximizando a função de verossimilhança, que definimos a seguir.

Definição: Se $X_1, X_2, ..., X_n$ for uma amostra aleatória de uma função de densidade $f(x;\theta)$, em que θ é um parâmetro desconhecido, e $x_1, x_2, ..., x_n$ são os valores observados da amostra aleatória, então a função de verossimilhança é definida como

$$L(\theta) = f(x_1;\theta) f(x_2;\theta) \dots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Notas:

1. Como $x_1, x_2, ..., x_n$ são valores observados, somente θ permanece desconhecido em $L(\theta)$;

2. Na forma como está proposta, a função de verossimilhança depende da independência entre as VAs observadas, por esse motivo a utilização de uma amostra aleatória na definição da função de verossimilhança.

Definição: A estimativa de máxima verossimilhança para um parâmetro θ é o valor de θ que maximiza a função de verossimilhança.

Exemplo 4.6 Numa linha de produção, uma peça é NC (Não Conforme) com probabilidade p. Uma amostra aleatória de tamanho n = 1000 peças apresenta 18 itens NC's.

O modelo probabilístico que se ajusta a esse caso é a distribuição de Bernoulli com parâmetro p, Ber(p). A função de verossimilhança é dada por

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}.$$

Para encontrar a estimativa de máxima verossimilhança recorremos ao cálculo. Poderíamos derivar e igualar a zero a função $L(\theta)$, mas, por facilidade, aplicamos o logaritmo antes. A aplicação do logaritmo não altera o valor de $\hat{\theta}$ que leva ao máximo de $L(\theta)$ e facilita na hora de derivar e igualar a zero. Assim, a função log-verossimilhança fica

$$l(\theta) = \log(L(\theta)) = \left(\sum_{i=1}^{n} x_i\right)\log(\theta) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta).$$

Para encontrar o máximo de $L(\theta)$, primeiro derivamos a função $l(\theta)$, com relação a θ, como em

$$\frac{\partial \log(L(\theta))}{\partial\theta} = \frac{1}{\theta}\sum_{i=1}^{n} x_i \log(\theta) - \frac{1}{1-\theta}\left(n - \sum_{i=1}^{n} x_i\right),$$

e a seguir igualamos o resultado da derivação a zero e isolamos o θ, para obter o seu estimador,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

A partir da estimativa $\hat{\theta}$ podemos definir o estimador

$$\hat{\theta} = \hat{\theta}\ (X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Note que a estimativa é o valor que assume o estimador quando a amostra $(X_1, X_2, ..., X_n)$ assume o valor $(x_1, x_2, ..., x_n)$.

A estimativa para a probabilidade de NC pode ser obtida, então, substituindo na expressão acima os valores observados na amostra, o que, nesse caso, corresponde à proporção de NC's na amostra, ou seja,

$$\hat{\theta} = 18/1000 = 0{,}018.$$

Tendo por base essa estimativa, podemos dizer que cada novo item produzido na linha de produção tem probabilidade 0,018 de ser um NC (Não Conforme). De forma equivalente, também podemos dizer que, mantidas as condições que geraram essa amostra, essa linha de produção produzirá em média 1,8% dos itens fora da especificação (ou NC's).

Notas:

1. O EMV é uma ferramenta muito poderosa e pode ser facilmente estendida para o caso de múltiplos parâmetros. A diferença é que a função de verossimilhança, L(.), será uma função de múltiplas variáveis (que, nesse caso, são os parâmetros), e, após as derivadas parciais, teremos um sistema de equações a resolver para determinar os estimadores;
2. O logaritmo utilizado na expressão para $L(\theta)$ é o neperiano, com base "e", que é a constante matemática e = 2,718282..., também conhecida como "número de Euler". No restante do texto, caso outra base seja utilizada, o leitor será avisado.

Exercícios:

1. Determine o estimador EMV para λ, baseado em uma amostra aleatória de tamanho n, de uma distribuição de Poisson com parâmetro λ, $X \sim \text{Pois}(\lambda)$;
2. Determine o estimador EMV para μ, baseado em uma amostra aleatória de tamanho n, de uma distribuição $N(\mu, \sigma^2)$ quando a variância é conhecida.

4.3.4.1.1 Propriedades do EMV

As propriedades 1 a 3 do EMV, apresentadas a seguir, são assintóticas, ou seja, valem para grandes amostras:

1. O EMV é aproximadamente não viesado, $E[\hat{\theta}] \approx \theta$;
2. A variância do EMV é quase tão pequena quanto se poderia obter com qualquer outro estimador, o que leva a estimadores mais precisos;
3. A distribuição do EMV é aproximadamente normal. Essa propriedade é muito interessante, pois permite estabelecer regiões de confiança;
4. O EMV é invariante. Isso quer dizer que, se tivermos um EMV $\hat{\theta}$ para θ, mas estivermos interessados em uma função $h(\theta)$, pela propriedade da invariância o EMV de $h(\theta)$ será $h(\hat{\theta})$.

Para maiores detalhes sobre essas propriedades ver literatura específica.

4.3.4.1.2 Dificuldades no uso do EMV

Quando uma solução analítica para encontrar o máximo da função de verossimilhança não for viável, uma solução possível passa a ser encontrar o máximo da função de verossimilhança através de métodos numéricos.

O R dispõe da instrução `nlm( )`, que encontra o mínimo de funções não lineares.

A utilização da instrução é apresentada no quadro que segue. Na primeira linha, definimos a função log–verossimilhança, a qual chamamos `F1`. A inversão do sinal da função (multiplicação por `-1`) é necessária, pois `nlm( )` calcula mínimos – e desejamos um máximo. Na segunda linha, temos a instrução propriamente dita. O primeiro argumento é a função a ser minimizada, `F1`, e o segundo é o ponto de partida do método iterativo.

```
#
#   Obtenção do EMV para o Exemplo 4.6
#
# Definição da função da qual desejamos encontrar o mínimo
F1=function(tt) (-1)*(18*log(tt) + 982*log(1-tt))
#
# Minimização propriamente dita
nlm(F1,.01)
#
```

A função de verossimilhança apresentada no quadro é a mesma do Exemplo 4.6. Recomendamos que se execute as instruções do quadro e se compare o resultado obtido com o resultado analítico obtido no exemplo. Como o leitor pode notar, o resultado computacional é uma aproximação em relação ao anterior.

Nota: Sempre que estiver disponível, o resultado analítico deve ser preferido em detrimento do numérico, já que o analítico é exato.

4.3.5 Estimação por intervalo

Na estimação por intervalo, buscamos aproximar o valor de um parâmetro populacional desconhecido através de um intervalo. O diferencial em relação à estimação por pontos é que, no caso da estimação por intervalo, também obtemos a probabilidade de que o parâmetro esteja no intervalo apresentado.

Na estimação por intervalo, temos duas informações muito interessantes. Uma delas diz respeito à medida da precisão do estimador, que é observada através da amplitude do intervalo, quanto mais largo o intervalo menos preciso o estimador e quanto mais estreito o intervalo mais preciso o estimador.

A segunda informação interessante é uma medida de incerteza, a cada intervalo obtido temos a probabilidade de que ele contenha o parâmetro de interesse. Quanto mais próxima de um for a probabilidade, mais confiança temos que o verdadeiro valor do parâmetro esteja no intervalo apresentado, e quanto mais distante de um for a probabilidade, menos confiança teremos que o intervalo contenha o parâmetro que desejamos estimar.

4.3.5.1 Estimação por intervalo para a média populacional

Considere o caso em que temos uma amostra aleatória, X_1, X_2,..., X_n, e desejamos estimar a média populacional através de um intervalo com probabilidade $1-\alpha$.

Utilizando o resultado do Teorema Central do Limite, apresentado acima, aproximamos a estatística S_n pela distribuição normal,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \tag{4.1}$$

tem distribuição aproximadamente normal com média zero e variância um para todo n suficientemente grande. Utilizando a tabela normal podemos escrever

$$P\left(z(\alpha/2) < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z(1-\alpha/2) \right) = 1-\alpha,$$

em que $z(\alpha/2)$ e $z(1-\alpha/2)$ são os quantis $\alpha/2$ e $(1-\alpha/2)$ da distribuição normal, respectivamente.

Considerando que $z(1-\alpha/2) = -z(\alpha/2)$, temos

$$P\left(-z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \right) = 1-\alpha$$

ou

$$P\left(\overline{X} - z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \right) = 1-\alpha$$

Como podemos observar, temos um intervalo para a média populacional μ com probabilidade $1-\alpha$. Esse intervalo é denominado intervalo de confiança e nos referiremos a ele pelas suas iniciais, IC, podendo ser expresso na forma

$$IC(\mu; 1-\alpha) = \left(\overline{X} \pm z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \right).$$

Quando trabalhamos com ICs, o mais comum é trabalharmos com probabilidade $1 - \alpha = 0{,}95$. Uma possível justificativa para essa escolha é que a probabilidade de 0,95 significa uma alta probabilidade de o evento associado ocorrer, ou seja, 95 em 100, ou equivalentemente 19 em 20. Outras probabilidades que também são utilizadas para IC's são 0,90, 0,99 e 0,999.

Quanto maior a probabilidade $1 - \alpha$, maior a chance de que o IC venha a conter o parâmetro de interesse, o que nos levaria a escolher IC's com a maior probabilidade possível. Essa abordagem tem um problema, quanto maior a probabilidade, mais amplo será o IC, o que também não é interessante. Assim sendo, na busca por um equilíbrio entre a maior probabilidade e o menor tamanho do intervalo, o uso consagrou $1 - \alpha = 0{,}95$.

Particularizando o IC para $1 - \alpha = 0{,}95$, podemos definir os valores de z( ), o que pode ser feito através da tabela normal, então obtemos o intervalo

$$P\left(\overline{X} - \frac{(1{,}96)\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{(1{,}96)\sigma}{\sqrt{n}}\right) = 0{,}95,$$

ou na forma de um IC,

$$IC(\mu; 0{,}95) = \left(\overline{X} \pm \frac{(1{,}96)\sigma}{\sqrt{n}}\right).$$

Esse intervalo pode ser determinado para outras probabilidades. Basta ter os valores correspondentes a essas novas probabilidades da tabela normal. Para 0,90, seria $\pm$ 1,64; para 0,99 seria $\pm$ 2,58 em lugar de $\pm$ 1,96.

Exemplo 4.7 Uma amostra de tamanho 10, relativa à emissão de material particulado em uma operação industrial, gerou os seguintes dados: 17,5; 20,6; 14,1; 20,4; 18,9; 14,6; 19,8; 15,5; 23,9; e 24,1. Observações históricas sobre o fenômeno nos permitem considerar $\sigma = 3{,}2$, constante e conhecido. Da amostra, temos $\overline{x} = 18{,}94$. O IC com probabilidade 0,95, usando aproximação normal, é dado por

$$IC(\mu;\ 0{,}95) = \left(\overline{x} \pm (1{,}96)\frac{\sigma}{\sqrt{n}}\right) = \left(18{,}94 \pm (1{,}96)\frac{3{,}2}{\sqrt{10}}\right) = (18{,}94 \pm 1{,}98).$$

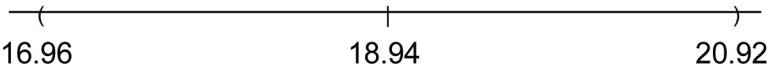


Figura 4.2 Ilustração do intervalo de confiança do exemplo.

Notas:

1. A probabilidade do IC, $1 - \alpha$, está associada à confiança ou à crença de que o intervalo contenha o verdadeiro valor do parâmetro de interesse, no caso anterior a média μ – embora essa observação seja válida para todos os tipos de ICs;

2. O tamanho do intervalo está associado à precisão do resultado. Intervalos muito grandes podem ser inúteis. Um exemplo disso está na afirmação de que a altura média de um grupo de pessoas adultas, da população em geral, está entre 1,5 e 1,9 m. Embora essa afirmação possa ser feita com probabilidade muito alta, ela não acrescenta nada ao que todos já sabemos;

3. A precisão de uma estimativa pode ser controlada pelo tamanho da amostra, o "n". Vejamos, aumentando o tamanho da amostra e mantendo a probabilidade do intervalo de confiança a ser construído, nossa expectativa é que tamanho dele seja reduzido, tornando a informação mais precisa. De forma inversa, se reduzirmos o tamanho da amostra e mantivermos a probabilidade, nossa expectativa é que o intervalo tenha seu tamanho aumentado, sendo portanto menos preciso. Note que, em ambas as situações apresentadas, a incerteza associada ao intervalo obtido permanece a mesma, pois a probabilidade não foi alterada;

4. A escolha da probabilidade associada ao IC, $1 - \alpha$, depende da área em que a pesquisa é feita. Um valor usual para $1 - \alpha$ é 0,95, valor utilizado na grande maioria das aplicações. Existem situações em que valores diferentes são recomendados. Aplicações na indústria farmacêutica, por exemplo, em que a dosagem em medicamentos tem implicações para a vida de seus usuários, utilizam probabilidades tais como 0,9999, ou maiores. Já estudos em que as variáveis de interesse apresentam variabilidade elevada, como é o caso de estudos nas áreas sociais, trabalham com valores tais como 0,90 para $1 - \alpha$ ou menores. Nesses casos, utilizar $1 - \alpha = 0{,}95$ implicaria em amostras tão grandes que poderia inviabilizar a pesquisa;

5. ICs que utilizam a aproximação normal, mostrados anteriormente, são recomendados para situações em que o desvio padrão σ é conhecido. Isso pode ocorrer em função de informações históricas, como mostrado no exemplo. Quando a amostra é grande, alguns autores recomendam $n > 100$, tambem podemos utilizar a distribuição normal. Quando essas condições não ocorrem, devemos utilizar a distribuição t de Student, como descrito a seguir.

Exemplo 4.8 Utilizando os mesmos dados da amostra do Exemplo 4.7, para construir um IC com probabilidade 0,99 usando aproximação normal, utilizamos

$$IC(\mu; 0,99) = \left(\bar{x} \pm (2,58)\frac{\sigma}{\sqrt{n}} \right) = \left(18,94 \pm (2,58)\frac{3,2}{\sqrt{10}} \right) = (18,94 \pm 2,61).$$

Exercício: Utilizando os mesmos dados da amostra do Exemplo 4.7 construa um IC com probabilidade 0,90 usando aproximação normal.

Um aspecto importante sobre σ é que este é, em geral, desconhecido. Uma solução é substituí-lo pelo seu estimador natural, o desvio padrão amostral, s. Para amostras grandes, n > 100, a aproximação é boa.

Para amostras pequenas, a distribuição normal na equação (4.1) é substituída pela distribuição t de Student com n – 1 graus de liberdade. Assim, temos

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

utilizando então a distribuição t de Student, podemos escrever

$$P\left(t(n-1)(\alpha/2) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t(n-1)(1-\alpha/2) \right) = 1-\alpha,$$

em que $t(n-1)(\alpha/2)$ e $t(n-1)(1-\alpha/2)$ são os quantis $(\alpha/2)$ e $(1-\alpha/2)$ da distribuição t de Student com n – 1 graus de liberdade, que estão tabelados.

Reorganizando a desigualdade, temos

$$P\left(\bar{X} - t(n-1)(\alpha/2)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t(n-1)(\alpha/2)\frac{S}{\sqrt{n}} \right) = 1-\alpha.$$

O intervalo para a média populacional μ, mostrado anteriormente, pode ser expresso na forma

$$IC(\mu; 1-\alpha) = \left(\bar{X} \pm t(n-1)(\alpha/2)\frac{S}{\sqrt{n}} \right).$$

Para obter o IC estimado substituimos $\bar{X}$ e S pelos seus respectivos estimadores obtidos da amostra $\bar{x}$ e

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

obtendo

$$IC(\mu; 1-\alpha) = \left(\bar{x} \pm t(n-1)(\alpha/2)\frac{s}{\sqrt{n}}\right).$$

Exemplo 4.9 Resíduos industriais.

A quantidade de resíduos produzidos por um processo de produção industrial é representada pela VA X, com média μ e variância σ^2. Suponha que historicamente esse processo tem produzido 10 kg de resíduos por lote do produto. Desejamos fazer afirmações sobre a quantidade de resíduos produzidos tendo controle sobre a incerteza associada a essas afirmações.

Considere que um estudo sobre novos filtros aplicados a equipamentos semelhantes, de outras empresas, produziu as observações $X = \{9, 10, 12, 11, 10, 11, 12, 11\}$. Dessa forma, temos $\bar{x} = 10,75$ e $s = 1,035$. O intervalo de confiança para média é

$$IC(\mu; 1-\alpha) = \left(\bar{X} \pm t(n-1)(\alpha/2)\frac{S}{\sqrt{n}}\right)$$

$$IC(\mu; 0,95) = \left(10,75 \pm 2,306\frac{1,035}{\sqrt{9}}\right) = (10,75 \pm 0,796) = (9,954;\ 11,546).$$

4.3.6 Testes de hipóteses

Uma forma interessante de fazer inferência é o teste de hipóteses. Nessa forma de inferência, estabelecemos uma hipótese inicial a respeito de um parâmetro de uma variável de interesse, que chamamos de hipótese nula e denotamos por H_0, e uma hipótese alternativa à hipótese nula, que chamamos hipótese alternativa e denotamos por H_1. A seguir, coletamos uma amostra que trará evidências em favor de uma dessas hipóteses, fazendo com que rejeitemos H_0 em favor de H_1, ou não.

Definição: Uma hipótese estatística é uma afirmação sobre um ou mais parâmetros.

Geralmente, a hipótese nula se refere ao estado atual de um fenômeno, e a hipótese alternativa a uma possível modificação desse estado atual, veja o exemplo que segue.

Aproveitando o exemplo sobre resíduos industriais, podemos formular basicamente três tipos de hipóteses sobre a média populacional:

1. H_0: $\mu = \mu_0$ *versus* H_1: $\mu \neq \mu_0$;
2. H_0: $\mu = \mu_0$ *versus* H_1: $\mu > \mu_0$;
3. H_0: $\mu = \mu_0$ *versus* H_1: $\mu < \mu_0$.

em que μ_0 é uma constante escolhida pelo analista em função do problema tratado.

No caso 1, o teste é chamado de bilateral pois os valores de μ que contrariam H_0 podem estar tanto acima como abaixo de μ_0; já nos casos 2 e 3, o teste é chamado de unilateral.

4.3.6.1 Teste de hipótese bilateral H_0: $\mu = \mu_0$ *versus* H_1: $\mu \neq \mu_0$

Exemplo 4.10 Retomando o exemplo sobre resíduos industriais, considere que um estudo sobre novos filtros aplicado a equipamentos semelhantes de outras empresas produziu as observações $X = \{9, 10, 12, 11, 10, 11, 12, 11\}$, e, portanto, $\bar{x} = 10{,}75$. O desvio padrão é conhecido, $\sigma = 1$.

Suponha que desejamos testar a hipótese H_0: $\mu = 10$ *versus* H_1: $\mu \neq 10$. Temos, então, supondo que H_0 seja verdadeira e usando o TCL, a estatística de teste

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 10}{1/\sqrt{8}} \sim N(0,1).$$

Valores aceitáveis para Z devem estar num intervalo de alta probabilidade para a VA Z, enquanto que valores improváveis de Z conduzem à rejeição de H_0.

Como Z tem distribuição normal padrão, consideramos $1 - \alpha$ uma probabilidade alta, então o intervalo $(z(\alpha/2), z(1-\alpha/2))$ contém os valores mais prováveis de Z, enquanto que valores de Z menores que $z(\alpha/2)$ e maiores que $z(1-\alpha/2)$ são os valores improváveis de Z. A esses valores improváveis chamamos de região de rejeição ou região crítica do teste.

Como a normal padrão é simétrica em torno de zero, temos $z(1-\alpha/2) = -z(\alpha/2)$, a região crítica do teste é $|Z| > z(1-\alpha/2)$ e

$$P\left(|Z| > z\left(1-\alpha/2\right)\right) = \alpha.$$

Devemos, portanto, rejeitar H_0 para $|Z| > z\left(1-\alpha/2\right)$.

Exemplo 4.11 Retomando o exemplo anterior e fazendo $\alpha = 0{,}05$, temos a região crítica determinada por $|Z| > z\left(1-\alpha/2\right) = 1{,}96$. Essa região crítica é ilustrada no gráfico da Figura 4.3.

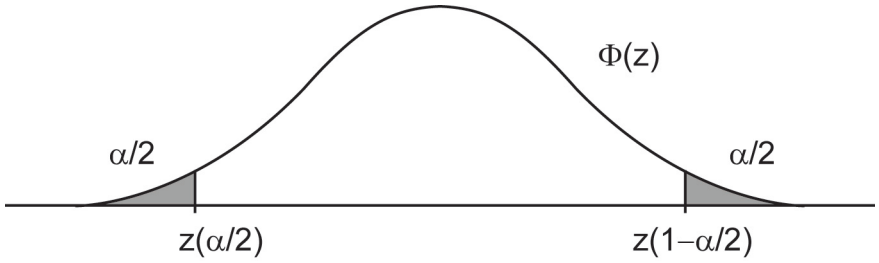


Figura 4.3 Região crítica do teste normal bilateral.

A realização da estatística de teste por meio da amostra é

$$Z_0 = \frac{10{,}75 - 10}{1/\sqrt{8}} = 2{,}12,$$

valor que se encontra dentro da região crítica, o que nos leva, portanto, a rejeitar H_0.

Exercício: Defina a região crítica do teste apresentado no exemplo anterior para valores de α em 0,10 e 0,01. Refaça os passos do exemplo para verificar seus conhecimentos.

Nota: O valor α, definido anteriormente, é chamado nível de significância do teste.

4.3.6.2 Teste de hipótese unilateral H_0: $\mu = \mu_0$ *versus* H_1: $\mu > \mu_0$

Nesse caso, o procedimento para montar o teste é análogo ao apresentado anteriormente para hipótese bilateral. A estatística de teste, Z, é a mesma, somente a região crítica do teste é diferente.

A região crítica é estabelecida em função da hipótese alternativa. Com H_1: $\mu > 10$, devemos rejeitar H_0 para valores altos de Z. Fixando o nível de significância do teste em α, a região crítica fica estabelecida como sendo os valores de Z maiores que $z\left(1-\alpha\right)$, pois

$$P\left(Z > z\left(1-\alpha\right)\right) = \alpha.$$

Assim, rejeitamos H_0 para valores de $Z > z\left(1-\alpha\right)$.

Exemplo 4.12 Aplicando o teste H_0: $\mu = 10$ *versus* H_1: $\mu > 10$, para $\alpha = 0{,}05$, aos dados do Exemplo 4.10, temos que $z(0{,}95) = 1{,}64$.

Como $Z_0 = 2{,}12$ e $Z_0 > z(0{,}95)$, rejeitamos H_0. A região crítica do teste é mostrada no gráfico da Figura 4.4.

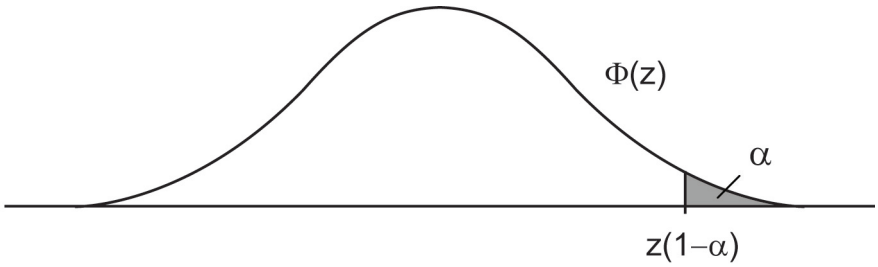


Figura 4.4 Região crítica do teste normal, $Z > z\left(1-\alpha\right)$.

4.3.6.3 Teste de hipótese unilateral H_0: $\mu = \mu_0$ versus H_1: $\mu < \mu_0$

Novamente, o procedimento para montar o teste é análogo ao apresentado anteriormente. A estatística de teste, Z, é a mesma. Somente a região crítica do teste é diferente.

A região crítica é estabelecida em função da hipótese alternativa. Como H_1: $\mu < 10$, nesse caso devemos rejeitar H_0 para valores baixos de Z. Fixando o nível de significância do teste em α, a região crítica fica estabelecida como sendo os valores de Z menores que $z\left(\alpha\right)$:

$$P\left(Z < z\left(\alpha\right)\right) = \alpha.$$

Assim, rejeitamos H_0 para valores de $Z < z\left(\alpha\right)$.

Exemplo 4.13 Aplicando o teste H_0: $\mu = 10$ *versus* H_1: $\mu < 10$, para $\alpha = 0{,}05$, aos dados do exemplo, temos que $z\left(0{,}05\right) = -1{,}64$.

Como $Z_0 = 2{,}12$ e $Z_0 > z\left(0{,}95\right)$, não rejeitamos H_0. A região crítica do teste é mostrada no gráfico da Figura 4.5.

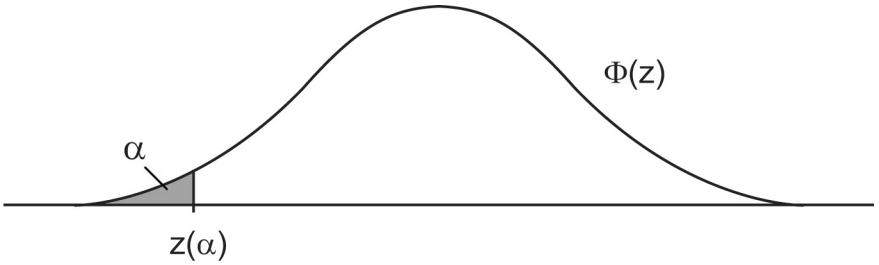


Figura 4.5 Região crítica do teste normal, $Z < z(\alpha)$.

Geralmente, podemos aplicar a seguinte lógica de raciocínio em teste de hipóteses: é mais razoável considerar que H_0 seja falsa e rejeitá-la do que aceitar que valores exóticos de Z ocorram – valores com probabilidade muito baixa de ocorrer. É claro que valores com probabilidades baixas são raros, mas ocorrem, e, nesse caso, temos um erro. Discutimos, a seguir, os erros cometidos em testes de hipóteses.

4.3.6.4 Erros tipo I e II

Quando testamos uma hipótese H_0, o resultado pode ser um acerto quando rejeitamos H_0 e ela é falsa, ou quando não a rejeitamos e ela é verdadeira. Mas temos duas situações de erro, que chamamos de erros tipo I e II:

- **Erro tipo I**: ocorre quando rejeitamos H_0, mas ela é verdadeira;
- **Erro tipo II**: ocorre quando não rejeitamos H_0, mas ela é falsa.

Controlar as probabilidades de cometer os erros tipo I e II é muito importante para sabermos o peso real das afirmações feitas decorrentes dos resultados dos testes e para ponderarmos as decisões tomadas em decorrência dos resultados. Veja o exemplo a seguir.

Exemplo 4.14 Uma empresa tem uma proposta de alteração em seu produto principal. Historicamente esse produto tem 20% do mercado. Uma pesquisa de opinião é feita para determinar a aceitação do produto modificado. Podemos comparar a aceitação atual, representada por μ com com a histórica de 20%, testando as hipóteses H_0: $\mu = 20$ *versus* H_1: $\mu > 20$. Rejeitar H_0 significa alterar o processo de produção, o que implica alto investimento. Os erros tipo I e II são importantes nesse caso. O erro tipo I, rejeitar H_0 quando ela é verdadeira, significa, para a empresa, realizar um alto investimento sem obter o retorno esperado. O erro tipo II, não rejeitar H_0 quando H_1 é verdadeira, significa a perda da oportunidade de melhorar o produto e perda do investimento feito no desenvolvimento das melhorias. Em ambos os casos, o erro significa a perda de competitividade da empresa.

As probabilidades dos erros tipo I e II são denotadas por

P(Erro tipo I) = α e P(Erro tipo II) = β.

No erro tipo II, a condição na qual H_0 deveria ter sido rejeitada pode ser um conjunto com infinitos valores do parâmetro, portanto a probabilidade de erro tipo II, β, depende desse valor verdadeiro do parâmetro para ser expressa.

Exemplo 4.15 Considere o caso do teste H_0: $\mu = \mu_0$ *versus* H_1: $\mu > \mu_0$, então

$$P(\text{Erro tipo II}) = P\left(\text{Nª o rejeitar } H_0 \mid \mu \in (\mu_0, \infty)\right),$$

ou seja, a probabilidade do erro tipo II $|\mu$ é função do verdadeiro valor do parâmetro testado cujo valor é desconhecido. Por essa razão o erro tipo II é expresso como função de μ, como não sabemos seu valor consideramos todas as possibilidades. O mesmo acontece com a função poder apresentada a seguir.

O erro tipo II está associado a uma outra função que é a função poder do teste, que é denotada por $\beta(\mu)$, e é dada por

$$\beta(\mu) = 1 - P(\text{Erro tipo II} \mid \mu).$$

A função poder é assim chamada pois ela reflete a capacidade ou o poder que o teste tem para rejeitar a hipótese H_0 quando ela é falsa. Aos leitores interessados em aprender mais sobre poder de testes de

hipóteses, recomendamos a literatura especializada em inferência estatística.

4.3.6.5 Teste t para a média

Nos testes tratados anteriormente, é comum que a variância não seja conhecida. Nesse caso, adotamos a mesma solução utilizada para intervalos de confiança, substituímos o desvio padrão σ por seu estimador s e a consequência é que passamos a trabalhar com a estatística T com distribuição t de Student com n – 1 graus de liberdade

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(n-1).$$

A estrutura dos testes é a mesma apresentada anteriormente. As únicas mudanças estão na estatística de teste – agora temos a T mostrada anteriormente – e na distribuição utilizada para determinar a região crítica.

As regiões críticas para os testes são definidas de forma análoga ao caso das normais, mostrados anteriormente. A região crítica de teste, no caso do teste H_0: $\mu = \mu_0$ *versus* H_1: $\mu \neq \mu_0$ é definida pelos quantis $t(n-1)(\alpha/2)$ e $t(n-1)(1-\alpha/2)$. Como a distribuição t de Student é simétrica em torno de zero, a região crítica para o teste bilateral é $|T| > t(n-1)(1-\alpha/2)$. No caso do teste para H_0: $\mu = \mu_0$ *versus* H_1: $\mu > \mu_0$, a região crítica para o teste é $T > t(n-1)(1-\alpha)$. E, no caso do teste da hipótese H_0: $\mu = \mu_0$ *versus* H_1: $\mu < \mu_0$, a região crítica para o teste é $T < t(n-1)(\alpha)$.

Apresentamos, a seguir, um exemplo para o teste bilateral. Os demais seguem de forma análoga ao caso da normal.

Exemplo 4.16 Retomando o exemplo sobre resíduos industriais, considere que um estudo sobre novos filtros aplicados a equipamentos semelhantes de outras empresas produziu as observações $X = \{9, 10, 12, 11, 10, 11, 12, 11\}$. Dessa forma, temos que

$$\bar{x} = 10.75 \text{ e } s = 1.035.$$

Suponha que desejamos testar a hipótese H_0: $\mu = 10$ *versus* H_1: $\mu \neq 10$. Então, sob H_0, temos que $T \sim t(7)$ e

$$T = \frac{10,75 - 10}{(1,035)/\sqrt{8}} = 2,05.$$

A região crítica de teste, para $\alpha = 0,05$, é

$$|T| > t(n-1)(1-\alpha/2) = t(7)(0,975) = 2,365.$$

Logo, não rejeitamos a hipótese H_0: $\mu = 10$.

Valores críticos para o teste t ou quantis da distribuição t de Student podem ser obtidos de tabelas da distribuição ou podem ser obtidos no R por meio da instrução `qt(.)`.

Os quantis da distribuição t de Student com v graus de liberdade podem ser obtidos como mostra o quadro a seguir.

```
#
# quantil 0,975 da distribuição t com 7 graus de liberdade
qt(0.975,7)
#
```

4.3.7 Introdução à amostragem

Vários aspectos devem ser considerados quando pretendemos analisar uma amostra e concluir para toda a população. Um aspecto importante é a forma como esses dados são coletados. Esse assunto é tratado em uma área da estatística a qual chamamos amostragem.

Amostragem é a área da estatística que reúne técnicas de coleta de dados, técnicas que buscam garantir que os resultados não apresentam viés.

Exemplo 4.17 Em pesquisas de opinião, um dos tipos que mais atrai a atenção das pessoas em geral são as pesquisas pré-eleitorais. Imagine que, em um município, as preferências pelos candidatos são regionalizadas, ou seja, os eleitores do candidato A estão mais concentrados em uma região, já os do B em outra, e assim por diante. Se um pesquisador desavisado concentrar sua amostra em um bairro que concentra mais eleitores de A, podemos dizer que o resultado apresentará um viés a favor do candidato e será diferente do resultado pretendido, que é estimar os votos em todo o município.

Existem técnicas para garantir que resultados não viesados sejam produzidos. Mostraremos algumas dessas técnicas a seguir.

Definição: População é todo o conjunto de elementos nos quais estamos interessados.

Definição: Amostra é um subconjunto de elementos selecionados a partir da população.

Exemplo 4.18 Exemplos de população e amostra.

População	**Amostra**
Eleitores de um município	200 eleitores selecionados ao acaso
Um lote com 5000 unidades de um determinado produto	50 unidades do produto selecionadas ao acaso
O conjunto de todas as fábricas de uma região	20 fábricas escolhidas entre todas

4.3.7.1 AAS (Amostra Aleatória Simples)

Definição: Uma amostra aleatória simples, de tamanho n, consiste em n elementos da população escolhidos de forma que qualquer elemento tenha a mesma chance (probabilidade) de ser escolhido.

Uma AAS pode ser obtida com ou sem reposição. Considere que a população tenha um formato que permita numerar cada um de seus elementos. Considere também que, para cada elemento, podemos ter uma bola com seu número em uma urna.

4.3.7.1.1 AAS sr (Amostra Aleatória Simples sem reposição)

Num procedimento similar a um bingo doméstico, podemos obter uma amostra de tamanho 10 retirando em sequência 10 bolas sem recolocá-las na urna. Nesse caso, teremos uma AAS sem reposição.

4.3.7.1.2 AAS cr (Amostra Aleatória Simples com reposição)

Outra forma de obter uma AAS de tamanho 10 é retirar uma bola, anotar o seu número e retorná-la à urna. Repetindo esse procedimento 10 vezes, também teremos uma amostra de tamanho 10, mas agora com reposição.

Existem várias diferenças entre os procedimentos com e sem reposição. De imediato, notamos que, com reposição, o mesmo elemento pode participar mais de uma vez na amostra. As diferenças no que se refere à análise estatística são marcantes. Neste texto, trataremos apenas da AAS sem reposição.

Considere $X = \{X_1, X_2, \ldots, X_n\}$ uma amostra coletada pelo esquema da AAS sem reposição de uma população finita com N elementos. Um estimador para a média populacional é a média amostral, descrito pela expressão

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

e o estimador para a variância é dado por

$$\hat{\text{var}}(\bar{x}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right).$$

Consequentemente, o intervalo de confiança para a média populacional, μ é

$$IC(\mu; 1-\alpha) = \bar{x} \pm z(\alpha/2)\sqrt{\hat{var}(\bar{x})}.$$

Exemplo 4.19 Diâmetro de um lote de 100 peças.

Em um lote de tamanho N = 100 peças, estamos interessados em estimar um IC para a média dos diâmetros. Uma AAS sr de tamanho n = 5 é coletada e produz $\bar{x} = 10{,}1$ e $s^2 = 0{,}04$. Uma estimativa para a variância da média amostral é

$$\hat{var}(\bar{x}) = \frac{s^2}{n}\left(1-\frac{n}{N}\right) = \frac{0{,}04}{5}\left(1-\frac{5}{100}\right) = (0{,}008)(0{,}95) = 0{,}0076.$$

O IC para a média populacional, IC(μ; 0,95), usando aproximação normal, é

$$IC(\mu; 0{,}95) = \bar{x} \pm z(\alpha/2)\sqrt{\hat{var}(\bar{x})} = 10{,}1 \pm (1{,}96)(0{,}0872) = 10{,}1 \pm 0{,}17.$$

4.3.7.1.3 Estimação de uma proporção

A estimação de uma proporção é, de fato, um caso particular da estimação da média. A proporção é a razão entre o número de casos favoráveis e o número total de casos, tanto para a população quanto para a amostra. Considere a situação em que a variável de interesse pode assumir somente os valores zero e um, sendo que a mesma assume o valor um quando o elemento amostrado apresenta a característica de interesse. Neste caso, a média populacional corresponde à proporção populacional e a média amostral corresponde à proporção amostral.

A proporção populacional é dada por

$$P = \frac{\text{Casos Favoráveis}}{\text{Casos Totais}} = \sum_{i=1}^{N}\frac{X_i}{N} = \bar{X},$$

em que $X_i = 1$, se o i-ésimo elemento da amostra apresentar a característica de interesse, e $X_i = 0$, caso contrário.

Dada a equivalência entre a média e a proporção, os estimadores utilizados para a média são também utilizados para a proporção, com as devidas simplificações.

O estimador da proporção populacional é a proporção amostral:

$$p = \frac{\text{Casos Favoráveis na amostra}}{\text{Tamanho da amostra}}.$$

Note a diferença: Na proporção populacional, o “P” é maiúsculo, enquanto que na amostral o “p” é minúsculo.

O estimador para a variância de p é

$$\hat{\text{var}}(p) = \frac{p(1-p)}{n}\frac{N-n}{N-1},$$

e o intervalo de confiança para P, utilizando aproximação normal para p, é

$$IC(P;1-\alpha) = p \pm z(\alpha/2)\sqrt{\hat{\text{var}}(p)}.$$

Exemplo 4.20 Em uma AAS sr de tamanho 100 de uma população com 1000 elementos, 15 elementos apresentam a característica de interesse.

A estimativa da proporção populacional é

$$p = 15/100 = 0{,}15.$$

A estimativa da variância do estimador é

$$\hat{\text{var}}(p) = \frac{(0{,}15)(1-0{,}15)}{100}\frac{1000-100}{1000-1} = (0{,}001275)(0{,}9009) = 0{,}001148.$$

O intervalo de confiança para a proporção populacional é

$$IC(P;0{,}95) = 0{,}15 \pm (1{,}96)(0{,}034) = 0{,}15 \pm 0{,}067.$$

4.3.7.1.4 Determinação do tamanho da amostra

Uma questão que surge no planejamento de uma AAS sr diz respeito ao tamanho da amostra. A partir do tamanho do IC considerado aceitável, da probabilidade do IC resultante, $(1-\alpha)$, e do valor aproximado dos parâmetros, podemos aproximar o tamanho da amostra.

Uma aproximação inicial do tamanho da amostra necessária para estimar uma proporção é

$$n_0 = \frac{t^2 p(1-p)}{d^2},$$

em que d é o limite superior para o erro considerado aceitável – comprimento do intervalo dividido por 2 – e t é o quantil $(1-\alpha/2)$ para um IC com probabilidade $(1-\alpha)$, ou seja, desejamos um IC tal que $P(|p-P|>d)=\alpha$. O valor de p na expressão anterior geralmente não é conhecido. Assim sendo, o seu valor é aproximado por resultados de estudos anteriores, opinião de especialistas ou mesmo amostra piloto.

Um refinamento do tamanho da amostra, utilizado quando n_0/N não é desprezível, é dado por

$$n = \frac{n_0}{1+(n_0/N)}.$$

Exemplo 4.21 Qual é o tamanho de amostra necessário para estimar uma proporção que acreditamos estar próxima de 0,2, com IC de comprimento não maior que 0,06, em uma população de tamanho 900?

Uma aproximação inicial é

$$n_0 = \frac{(1{,}96)^2\, 0{,}2(1-0{,}2)}{(0{,}03)^2} = 0{,}615/0{,}0009 = 683{,}3,$$

e o refinamento é

$$n = \frac{n_0}{1+(n_0/N)} = \frac{683{,}3}{1+(683{,}3/900)} = 388{,}4.$$

Portanto, devemos usar uma amostra de pelo menos 389 elementos para obter o intervalo desejado.

Para estimar a média de uma população, a determinação do tamanho da amostra é similar. A aproximação inicial é dada por

$$n_0 = \left(\frac{tS}{r\bar{X}}\right)^2,$$

em que S é uma aproximação do desvio padrão, $\bar{X}$ é uma aproximação da média que deseja estimar e r diz respeito ao comprimento máximo aceitável do IC, expresso como proporção da média. O IC desejado é tal que $P(|\bar{x}-\bar{X}|>r\bar{X})=\alpha$.

O refinamento na determinação do tamanho da amostra é idêntico ao caso da proporção mostrado anteriormente.

Exemplo 4.22 Uma população tem média próxima de 20 e um desvio padrão em torno de 2. Qual o tamanho da amostra necessário para estimar a média com um IC de comprimento não maior que 5% da média em uma população de tamanho 1000 e com probabilidade $1-\alpha = 0,95$

A aproximação inicial é

$$n_0 = \left(\frac{tS}{r\bar{X}}\right)^2 \left(\frac{(1,96)(2)}{(0,025)(20)}\right)^2 = 61,5,$$

e o refinamento é

$$n = \frac{n_0}{1+(n_0/N)} = \frac{61,5}{1+(61,5/1000)} = 57,9.$$

Portanto, devemos usar uma amostra com pelo menos 58 elementos para obter o intervalo desejado.

Existem, disponíveis na literatura, tabelas de números aleatórios, também chamados de randômicos, em que é possível obter uma sequência de números para coleta de uma AAS. No nosso caso, recomendamos a utilização da instrução `sample(.)` do R, que pode fornecer diretamente a lista de elementos para fazer parte da amostra. Um exemplo está disponível no quadro a seguir.

A AAS é geralmente a mais indicada quando sua aplicação é viável. Em situações em que é possível obter uma lista dos elementos da população e numerá-los, geralmente podemos aplicar a AAS. São exemplos de situações como esta quando queremos aplicar um questionário aos funcionários de uma fábrica, aos alunos de uma escola, etc. Em todos esses casos, existe uma lista dos elementos da população. Também podemos aplicar a AAS em casos em que, embora não disponhamos de uma lista, os elementos da população estão organizados de forma que cada um pode ser identificado sem ambiguidade. Este é o caso de itens em estoque organizados em prateleiras, estrados ou outra forma que, sendo sorteados, podem ser localizados por meio de uma contagem.

Em situações em que a AAS não pode ser utilizada, outros métodos de coleta de amostras estão disponíveis. Esses métodos mais sofisticados não serão tratados aqui. Aos leitores interessados, recomendamos buscar literatura técnica específica. Um exemplo de situação em que a AAS não pode ser aplicada são os estudos de pesquisa de opinião, em que não é possível, por exemplo, obter uma lista de todos os potenciais compradores de um certo produto.

Amostras aleatórias podem ser obtidas no R por meio da instrução `sample(.)`. Esta permite amostrar a partir de uma lista de elementos ou a partir de uma sequência numérica. No quadro a seguir, mostramos como obter uma amostra sem reposição, de tamanho 20, de uma sequência numérica que vai de 1 a 500 – 500, nesse caso, seria o tamanho da população. Utilize essas instruções no R e veja o resultado.

```
#
#     Exemplo de seleção de amostra AASsr
#
N=500
Amostra=sample(1:N,size=20,replace=FALSE)
Amostra
#
```

4.3.8 Regressão Linear Simples

Um problema comum em várias áreas do conhecimento é a análise da relação linear entre duas variáveis.

A relação linear ente as variáveis x e y pode ser representada pelo modelo

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (1)$$

em que y é a variável resposta, x é a variável explicativa, β_0 e β_1 são os parâmetros do modelo, o intercepto e a inclinação, respectivamente, e ε é o erro aleatório.

Podemos particularizar esse modelo para os dados observados, para cada uma das n observações, na forma

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

em que i = 1, 2,..., n, e i indexa as observações.

No modelo linear simples, o parâmetro β_0 corresponde ao intercepto ou ao ponto em que a reta corta o eixo das ordenadas, e β_1 corresponde ao coeficiente angular, o quanto y varia quando x varia uma unidade.

Outra forma de representar o modelo é por meio da curva esperada,

$$E[Y] = \beta_0 + \beta_1 x.$$

4.3.9 Estimação pelo método de mínimos quadrados

Estimar o modelo (1) significa encontrar boas aproximações para os parâmetros β_0 e β_1. Isso pode ser feito por meio de vários métodos, um dos mais utilizados é o método de mínimos quadrados.

Considere a soma de quadrados dos erros cometidos por um modelo quando dois valores quaisquer são escolhidos para β_0 e β_1,

$$Q = \sum_{i=1}^{n} (\varepsilon_i)^2 = (y_i - \beta_0 - \beta_1 x_i)^2.$$

Derivando essa expressão para Q em relação a β_0 e β_1 e igualando a zero, temos as equações

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

e

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

que correspondem ao sistema de equações normais. A solução desse sistema de equações produz as estimativas de mínimos quadrados para os parâmetros β_0 e β_1, dados por

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

e

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

Exemplo 4.23 Considere duas variáveis X e Y cujos dados são apresentados na Tabela 4.1. O ajuste do modelo linear pode ser feito como descrito a seguir.

Tabela 4.1 Dados para aplicação do modelo linear simples.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	59	57	76	68	53	77	58	72	63	57	77	67	60	72	70
Y	38	38	40	50	39	55	38	50	42	44	53	44	38	45	37

Para ajustar o modelo, determinamos inicialmente partes da expressão

$$n = 15 \qquad \bar{x} = 65,733 \qquad \bar{y} = 43,40$$

$$\sum_{i=1}^{n} x_i^2 = 65736 \qquad \sum_{i=1}^{n} y_i^2 = 28761 \qquad \sum_{i=1}^{n} x_i y_i = 43247.$$

A seguir, aplicamos esses valores na expressão das estimativas dos parâmetros e obtemos,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{43247 - (15)(65,733)(43,40)}{65736 - (15)(65,733)^2} = 0,492$$

e

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 43,40 - (0,492)(65,733) = 11,06.$$

A Figura 4.6 mostra o diagrama de dispersão desses dados já com o modelo ajustado, representado pela linha reta no gráfico.

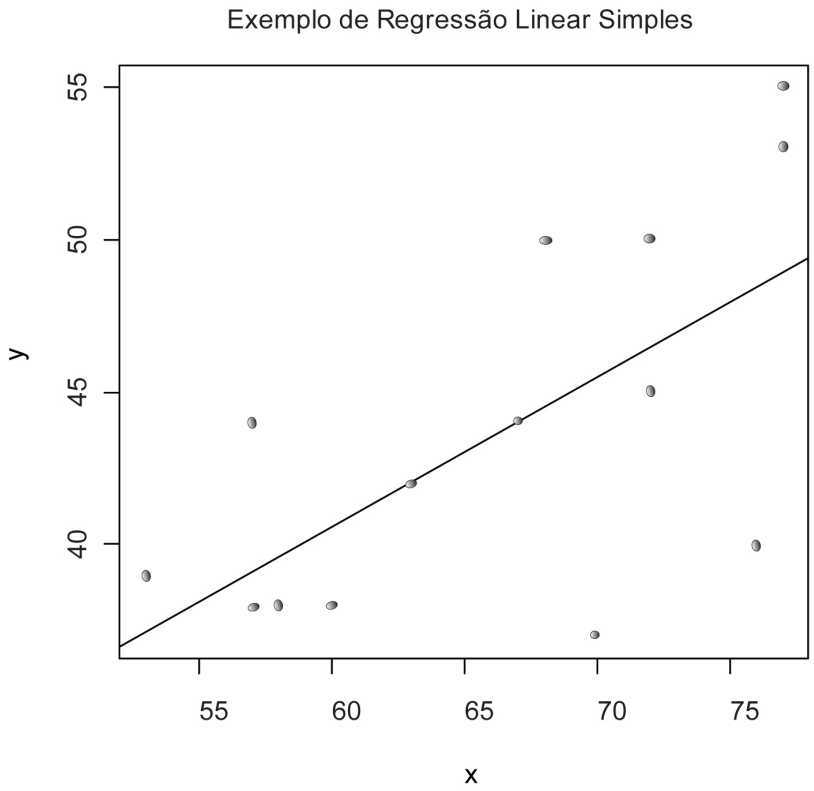


Figura 4.6 Regressão com dados do exemplo.

Tendo uma boa aproximação dos parâmetros, podemos prever a resposta por meio do modelo ajustado

$$\hat{y}_{(x)} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Assim, se desejarmos prever uma nova observação, por exemplo, para $x = 70$, substituiremos 70 na equação anterior e obteremos

$$\hat{y}(70) = 11{,}13 + 0{,}491(70) = 45{,}5.$$

4.3.10 Inferência

Para poder construir intervalos de confiança e realizar testes de hipóteses relativos aos parâmetros e previsões, precisamos estabelecer uma base probabilística para o modelo.

Essa base probabilística é estabelecida por meio das seguintes suposições básicas:

1. Os erros ε_i são variáveis aleatórias;

2. Os erros têm valor esperado igual a zero, $E[\varepsilon_i] = 0$;

3. Os erros têm variância constante, $Var[\varepsilon_i] = \sigma^2$;

4. Os erros são não correlacionados, $cor(\varepsilon_i, \varepsilon_j) = 0$, para $i \neq j$;

5. Os erros têm distribuição normal.

A partir dessa base, podemos deduzir os resultados

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right) \quad \text{e} \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right).$$

Como o parâmetro σ^2 não é conhecido e deve ser substituído por seu estimador S_e^2, a distribuição a ser utilizada é a t de Student com $(n-2)$ graus de liberdade.

Assim, os intervalos de confiança com nível de confiança $(1-\alpha)$ para β_0 e β_1 são

$$IC(\beta_0; 1-\alpha) = \hat{\beta}_0 \pm t_{(n-2)}(1-\alpha/2)\ S_e \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}}$$

e

$$IC(\beta_1; 1-\alpha) = \hat{\beta}_1 \pm t_{(n-2)}(1-\alpha/2) S_e \sqrt{\frac{1}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}},$$

respectivamente.

Para a previsão de novas observações, utilizamos

$$IC(y(x); 1-\alpha) = \hat{y}(x) \pm t_{(n-2)}(1-\alpha/2) S_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}}.$$

Exemplo 4.24 Aplicando esses ICs aos dados do Exemplo 4.23, temos que o intervalo de confiança para o intercepto β_0, para $\alpha = 0{,}05$, é

$$IC(\beta_0; 1-\alpha) = 11{,}02 \pm (2{,}16)(4{,}67)\sqrt{\frac{65736}{15\left(65736 - 15(65{,}7)^2\right)}}$$

$$IC(\beta_0; 1-\alpha) = 11{,}02 \pm (2{,}16)(9{,}83) = 11{,}02 \pm 21{,}24$$

e o intervalo de confiança para a inclinação da reta β_1 é

$$IC(\beta_1; 1-\alpha) = 0{,}49 \pm (2{,}16)(4{,}67)\sqrt{\frac{1}{65736 - 15(65{,}7)^2}}$$

$$IC(\beta_1; 1-\alpha) = 0{,}49 \pm 0{,}32.$$

O valor predito para uma nova observação com x = 60 é

$$\hat{y}(60) = 40{,}6$$

e o intervalo de confiança associado é

$$IC(y(60); 0{,}95) = 40{,}6 \pm (2{,}16)(4{,}67)\sqrt{1 + \frac{1}{15} + \frac{(60 - 65{,}7)^2}{65736 - 15(65{,}7)^2}}$$

$$IC(y(x); 1-\alpha) = 40{,}6 \pm 10{,}58.$$

Para verificar se o intercepto deve ou não ser mantido no modelo, podemos testar a hipótese H_0: $\beta_0 = 0$ *versus* H_1: $\beta_0 \neq 0$. Para isso, utilizamos a estatística de teste

$$t(\hat{\beta}_0) = \frac{\hat{\beta}_0 - \beta_0}{S_e}\sqrt{\frac{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}{\sum_{i=1}^{n} x_i^2}},$$

que segue a distribuição t de Student com $(n-2)$ graus de liberdade.

Para verificar se a variável explicativa deve permanecer no modelo, testamos a hipótese H_0: $\beta_1 = 0$ *versus* H_1: $\beta_1 \neq 0$. Para isso, a estatística utilizada é

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1 - \beta_1}{S_e}\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\ ,$$

que segue a distribuição t de Student com $(n-2)$ graus de liberdade.

Exemplo 4.25 Utilizando os dados do Exemplo 4.23, podemos testar a hipótese H_0: $\beta_0 = 0$ *versus* H_1: $\beta_0 \neq 0$ usando a estatística $t(\hat{\beta}_0)$,

$$t(\hat{\beta}_0) = \frac{11{,}06 - 0}{(4{,}67)}\sqrt{\frac{15\left(65736 - 15(65{,}7)^2\right)}{65736}} = 1{,}125,$$

que segue a distribuição t de Student com 13 graus de liberdade, tendo, portanto, como região crítica $|t(\hat{\beta}_0)| > 2{,}16$. Assim sendo, não rejeitamos H_0. Nesse caso, os dados não mostram evidência suficiente para afirmar que o intercepto é diferente de zero. Dessa forma, uma alternativa seria a exclusão do intercepto do modelo, o que resultaria em uma reta passando pela origem, $y = \beta_1 x$. Muitos analistas preferem manter o intercepto no modelo mesmo este não sendo significativamente diferente de zero, reservando a exclusão somente para os casos em que há razões estruturais para a exclusão.

Para esclarecer o que seriam razões estruturais, lançamos mão de um exemplo. Considere o caso em que modelamos preço de imóveis em função da área de terrenos. Caso o modelo linear seja adequado, é natural que a reta passe pela origem, já que um terreno de área zero deve ter custo zero.

Também podemos testar a hipótese H_0: $\beta_1 = 0$ *versus* H_1: $\beta_1 \neq 0$ usando a estatística $t(\hat{\beta}_1)$,

$$t(\hat{\beta}_1) = \frac{0{,}492 - 0}{(4{,}67)}\sqrt{65736 - 15(65{,}7)^2} = 3{,}313,$$

em que $t(\hat{\beta}_1) \sim t_{13}$. A região crítica é $|t(\hat{\beta}_1)| > 2{,}16$, e, portanto, rejeitamos H_0. Ao rejeitarmos H_0, estamos dizendo que a variável explicativa, x nesse caso, contribui significativamente para explicar a resposta y. Caso não a tivéssemos rejeitado, significaria que os dados disponíveis não teriam evidência suficiente para

afirmar que a contribuição de x é significativa na explicação de y e, portanto, esta poderia ser excluída do modelo.

4.3.11 Avaliação do modelo

Uma forma de avaliar o modelo é decompondo a soma de quadrados da variável resposta.

Considere inicialmente o modelo mais simples para explicar a variável resposta, a média. A soma dos erros ao quadrado para esse modelo é o que chamamos de Soma de Quadrados Total, ou SQTot, e é expressa por

$$SQTot = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Considere agora o modelo linear simples. O erro de previsão do modelo ajustado para a i-ésima observação, ao qual chamamos resíduo, é dado por $e_i = y_i - \bar{y}$, sendo que $\hat{y}_i$ é o valor predito para a i-ésima observação e é definido como $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Uma medida da qualidade de ajuste do modelo é a Soma dos Quadrados dos Resíduos, ou SQRes, que é expressa por

$$SQRes = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Em ambos os casos, SQTot e SQRes, quanto maior a soma de quadrados, pior o ajuste do modelo e vice-versa.

Um aspecto sobre essas somas é que SQTot é sempre maior que SQRes. A justificativa para essa diferença é que, dispondo da variável explicativa, o modelo linear sempre se ajusta melhor aos dados. Atribuímos a diferença entre essa soma de quadrados ao ajuste da reta, ou à regressão, e a chamamos de Soma de Quadrados de Regressão, ou SQReg, e calculamos como em

SQReg = SQTot – SQRes.

Essas somas de quadrados podem ser organizadas na forma de uma tabela, que chamamos tabela de Análise de Variâncias ou de forma abreviada de "Tabela ANOVA", termo que vem do inglês *ANalysis Of VAriance*. A tabela ANOVA para o modelo linear simples é mostrada na Tabela 4.2.

Tabela 4.2 Tabela ANOVA para modelo linear simples.

Fontes de variação	Graus de liberdade	Somas de quadrados	Quadrados médios	Estatística F	Pr(>F)
Regressão	1	SQReg	QMReg	F	Valor-p
Resíduos	n – 2	SQRes	QMRes		
Total	n – 1	SQTot			

A tabela ANOVA é organizada de forma que, na primeira coluna, estão listadas as fontes de variação, que são Regressão, Resíduos e Total; na segunda, os graus de liberdade relativos a cada fonte, que são 1, n – 2 e n – 1, respectivamente; na terceira coluna, estão as somas de quadrados; na quarta coluna, estão os quadrados médios, que são as somas de quadrados divididas pelos graus de liberdade correspondentes, QMReg = SQReg e QMRes = SQRes/$(n-2)$; na quinta coluna, está a estatística de teste, F = QMReg/QMRes, que segue a distribuição F com 1 e n – 2 graus de liberdade, ou $F \sim F_{1,\, n-2}$; e na última coluna, segue o "valor-p" da estatística F com relação à distribuição correspondente.

O quadrado médio de resíduos, também denotado por S_e^2, é o estimador de σ^2.

A hipótese sendo testada pela estatística F da tabela ANOVA mostrada anteriormente é H_0: $\beta_1 = 0$ *versus* H_1: $\beta_1 \neq 0$, que equivale ao teste t de Student para a estatística $t(\beta_1)$, também mostrado anteriormente. Essa coincidência somente ocorre porque temos apenas uma variável explicativa. Em casos de regressão múltipla – que estão fora do escopo deste livro, mas que estão disponíveis na bibliografia –, essa coincidência não ocorre.

Um subproduto da tabela ANOVA é o coeficiente de determinação, denotado e popularmente conhecido como R^2. Ele é definido como

$$R^2 = \frac{\text{SQReg}}{\text{SQTot}}.$$

O R^2 é um número entre zero e um, $0 \leq R^2 \leq 1$, que indica a qualidade do ajuste. Em geral, quanto maior o seu valor, melhor o modelo se ajusta aos dados. Este pode ser interpretado como sendo a proporção da variabilidade da Variável Resposta explicada pela Variável Explicativa.

No caso do modelo linear simples, $R^2 = 1$ indica que todos os pontos estão sobre a reta ajustada. Não é incomum o R^2 surgir multiplicado por 100 e expresso em porcentagens.

Exemplo 4.26 Aplicando as definições para a tabela ANOVA aos dados do Exemplo 4.23, obtemos a Tabela 4.3.

Tabela 4.3 Tabela ANOVA para Modelo Ajustado.

Fonte de variação	Graus de liberdade	Somas de quadrados	Quadrados médios	Estatística F	Pr(> F)
Regressão	1	223,9	223,9	10,3	0,007
Resíduos	13	283,7	21,8		
Total	14	507,6			

Como podemos observar na tabela ANOVA, o p-valor é 0,007. Nesse caso, rejeitamos a hipótese de que a variável explicativa seja não significativa para qualquer $\alpha > 0,007$. Portanto, para o valor usual $\alpha = 0,05$, rejeitamos H_0 e concluímos pela permanência da variável explicativa no modelo.

O coeficiente de determinação para os dados do exemplo é

$$R^2 = \frac{SQReg}{SQTot} = \frac{223,9}{507,6} = 0,443,$$

o que indica que a proporção da variabilidade de y explicada por x é 0,443, ou 44,3%.

A sequência de instruções mostradas no quadro a seguir ilustra como ajustar um modelo linear simples no R. Os dados utilizados são os mesmos do Exemplo 4.23 trabalhado anteriormente.

```
#
#     Registro dos dados
x=c(59,57,76,68,53,77,58,72,63,57,77,67,60,72,70)
y=c(38,38,40,50,39,55,38,50,42,44,53,44,38,45,37)
#
#     Ajuste do modelo
m1=lm(y~x)
#
#     Saída dos coeficientes ajustados
summary(m1)
#
#     Cálculo da tabela ANOVA
anova(m1)
#
#     Gráfico dos dados com reta ajustada
plot(x,y)
abline(m1$coef)
title("Exemplo de Regressão Linear Simples")
```

Note que as duas primeiras instruções guardam os dados nas variáveis x e y, respectivamente. A instrução seguinte, `lm(.)`, ajusta o modelo e guarda os resultados em `m1`. A instrução `summary(.)` apresenta um resumo dos parâmetros estimados e a instrução `anova(.)` calcula os dados para a tabela ANOVA. A instrução `plot(.)` constrói o diagrama de dispersão, `abline(.)` acrescenta ao diagrama a reta ajustada e `title(.)` acrescenta o título ao gráfico.

Experimente essas instruções no R e compare com os resultados apresentados acima. As pequenas diferenças que possam surgir eventualmente são devidas a diferenças de arredondamento.

4.3.12 Coeficiente de correlação linear

O coeficiente de correlação linear entre x e y ou simplesmente a correlação entre x e y mede a associação entre as variáveis x e y. Ele é definido como

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

e pode assumir valores entre –1 e 1, $-1 \leq r_{xy} \leq 1$. Valores próximos de 1 ou –1 indicam forte associação entre as variáveis; já valores próximos de zero indicam associação fraca. Valores positivos indicam associação positiva ou direta, ou seja, à medida que x aumenta y também aumenta e vice-versa; valores negativos indicam, por sua vez, uma associação negativa ou inversa, ou seja, à medida que x aumenta y diminui e vice-versa.

Exemplo 4.27 O coeficiente de correlação entre x e y para os dados do Exemplo 4.23 é

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

$$= \frac{43247 - (15)(65.73)(43.40)}{\sqrt{\left(65736 - (15)(65.73)^2\right)\left(28761 - (15)(43.40)^2\right)}} = 0,66,$$

o que indica uma associação positiva não muito forte.

No R, a instrução para calcular o coeficiente de correlação linear entre duas variáveis x e y é mostrada no quadro a seguir.

```
#
#    Cálculo do coeficiente de correlação entre x e y
cor(x,y)
#
```

4.3.13 Análise de resíduos

Toda a inferência utilizada para validar o modelo, intervalos de confiança e testes de hipóteses é baseada nas suposições básicas apresentadas anteriormente e também na suposição de validade do modelo linear.

Todas essas suposições devem ser verificadas. Isso é feito por meio da análise de resíduos, tanto via medidas de diagnóstico como via gráficos de resíduos.

4.3.13.1 Gráficos de resíduos

Apresentamos, a seguir, dois tipos de gráficos de resíduos: o de valores preditos *versus* resíduos e o probabilístico normal.

4.3.13.1.1 Gráfico de valores preditos *versus* resíduos

O gráfico de valores preditos *versus* resíduos é utilizado para verificar a suposição de variância constante e a adequabilidade do modelo linear. A Figura 4.7 ilustra algumas possibilidades para esse tipo de gráfico.

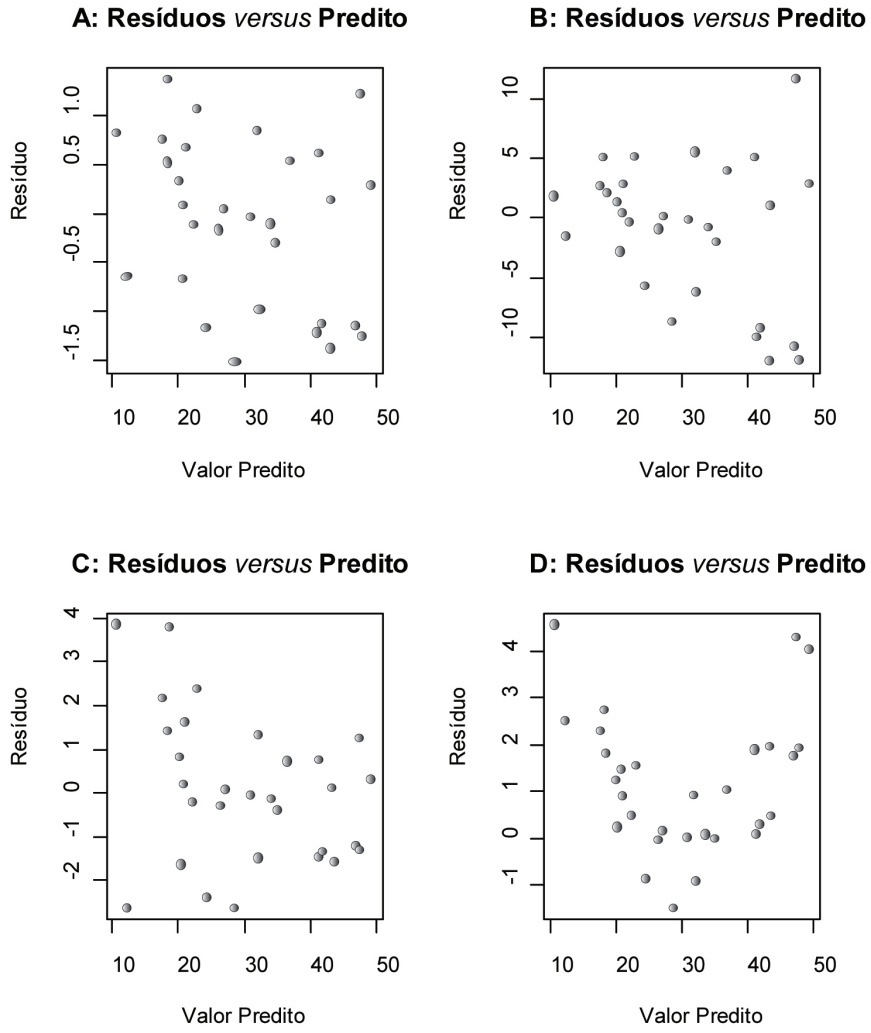


Figura 4.7 Possíveis configurações do gráfico valores preditos *versus* resíduos.

Observando a Figura 4.7, temos, no gráfico A, a situação esperada quando as suposições básicas são todas atendidas. Nesse tipo de gráfico, esperamos pontos concentrados em torno de zero e reduzindo a densidade à medida que a distância de zero aumenta. Também esperamos pontos igualmente espalhados, independente do valor predito, esperamos ainda que não haja tendências no gráfico.

No gráfico B, observamos que a variabilidade dos resíduos aumenta à medida que o valor predito aumenta, contrariando a suposição de variância constante. No gráfico C, a variância decresce à medida que o valor predito cresce, também contrariando a suposição de variância constante. No gráfico D, notamos uma tendência em forma de parábola, ou "Ü", contrariando o próprio modelo linear.

Para calcular no R os diagnósticos apresentados anteriormente, utilize as instruções mostradas no quadro a seguir. Note que, para funcionar, as instruções do quadro anterior devem ter sido executadas recentemente.

```
#
#     Gráfico de valores preditos versus residuos
plot(m1$fit,m1$res,xlab='valor predito',ylab='resíduo')
title('Gráfico de Resíduos')
#
```

4.3.13.1.2 Gráfico de probabilidade normal

Uma suposição importante é a de normalidade dos erros. Como não dispomos dos erros, já que estes dependem dos parâmetros, utilizamos os seus valores estimados, os resíduos. Para verificar a normalidade dos resíduos, usamos o gráfico de probabilidade normal, também chamado de gráfico probabilístico normal, mostrado na Figura 4.8.

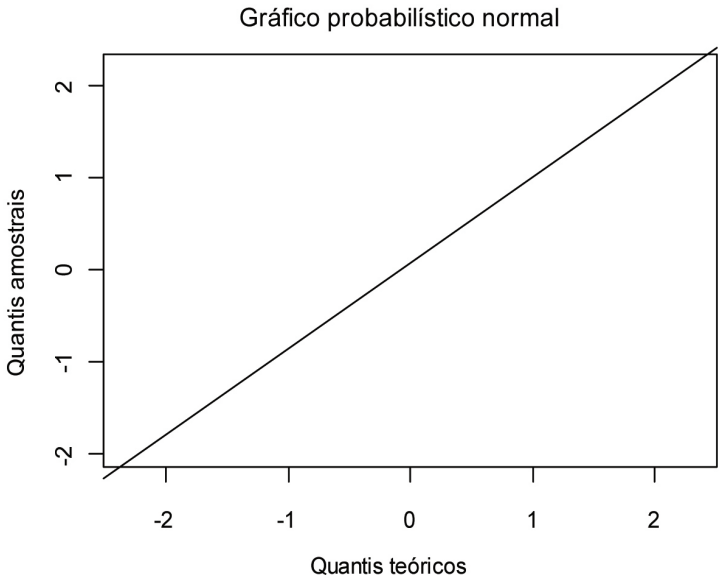


Figura 4.8 Gráfico probabilístico normal.

No caso dos resíduos terem distribuição normal, os pontos no gráfico devem aparecer próximos da linha reta. Caso isso não ocorra, a suposição de normalidade deve ser questionada, e uma análise mais detalhada deve ser conduzida.

Exemplo 4.28 Aplicando o gráfico de probabilidade normal aos resíduos do ajuste do modelo linear aos dados do Exemplo 4.23, obtemos o gráfico da Figura 4.9.

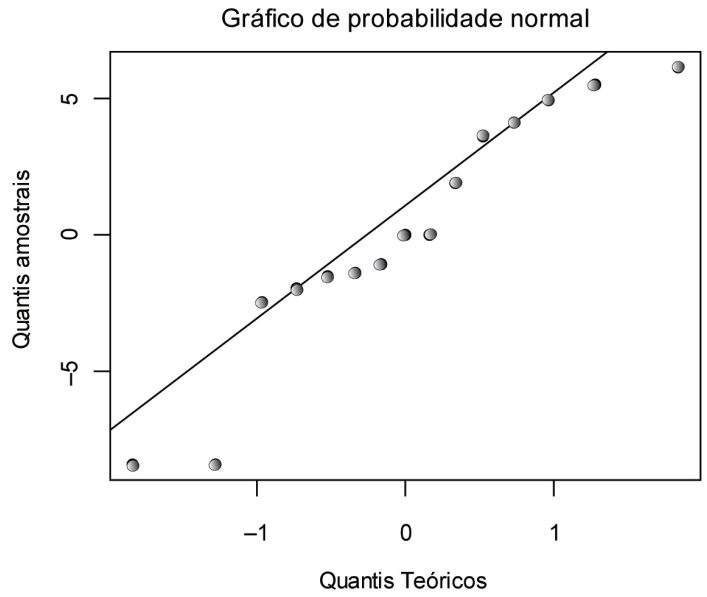


Figura 4.9 Gráfico probabilístico normal para resíduos do Exemplo 4.23.

Como podemos observar, a maioria dos pontos tem uma boa aderência à linha reta. A exceção são os pontos no canto inferior esquerdo. Esses pontos merecem uma análise mais cuidadosa.

O gráfico de probabilidade normal para os dados do Exemplo 4.23 pode ser gerado utilizando as instruções do quadro a seguir.

```
#

#     Gráfico probabilístico normal para resíduos do modelo

qqnorm(m1$res,xlab='Quantis teóricos', ylab='Quantis
amostrais',main='Gráfico probabilístico normal')

qqline(m1$res)

#
```

4.3.14 Medidas de diagnósticos

Apresentamos, a seguir, o resíduo padronizado e o resíduo estudentizado.

4.3.14.1 Resíduo padronizado

O resíduo padronizado é definido como

$$\hat{z}_i = \frac{y_i - \hat{y}_i}{S_e} = \frac{e_i}{S_e}.$$

4.3.14.2 Resíduo estudentizado

O resíduo estudentizado é definido como

$$r_i = \frac{e_i}{S_e\sqrt{1-\nu_{ii}}},$$

em que $\nu_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$.

Ambos os resíduos seguem a distribuição t de Student, portanto valores acima de 2 alertam o analista para possíveis problemas com as observações correspondentes.

A diferença entre resíduos padronizados e estudentizados é que, nos estudentizados, r_i, a i-ésima observação é excluída da estimação do i-ésimo resíduo. Isto é interessante na medida que evita que uma observação influencie no cálculo de seu próprio diagnóstico.

Exemplo 4.29 Calculando esses diagnósticos para os dados do Exemplo 4.23, obtemos o resultado mostrado na Tabela 4.4.

Tabela 4.4 Diagnósticos.

i	$\hat{z}_i$	e_i	$\hat{z}_i$	r_i
1	40.083	–2.083	–0.474	–0.460
2	39.098	–1.098	–0.255	–0.246
3	48.457	–8.457	–2.000	–2.310
4	44.516	5.484	1.219	1.244
5	37.128	1.872	0.460	0.446
6	48.950	6.050	1.452	1.524
7	39.591	–1.591	–0.365	–0.353
8	46.487	3.513	0.797	0.785
9	42.054	–0.054	–0.012	–0.011
10	39.098	4.902	1.138	1.152
11	48.950	4.050	0.972	0.970
12	44.024	–0.024	–0.005	–0.005
13	40.576	–2.576	–0.582	–0.567
14	46.487	–1.487	–0.337	–0.325
15	45.502	–8.502	–1.904	–2.154

Como podemos observar na Tabela 4.4, enquanto os resíduos padronizados não detectaram observações suspeitas, a observação 3 está exatamente sobre o limite, 2.0, mas não o ultrapassa. Os resíduos estudentizados detectaram duas observações suspeitas, a 3 e a 15, justamente as circuladas na Tabela.

Essas observações devem passar por uma investigação mais detalhada, com a verificação de possíveis erros de coleta da informação, tais como digitação ou condução da leitura desses dados. Caso sejam detectados erros, estes deverão ser corrigidos, se isso for possível, ou a observação pode até mesmo ser descartada. Caso não sejam detectados erros, a observação deve ser mantida.

Para calcular os valores da Tabela 4.4, no Exemplo 4.29, o R dispõe do procedimento mostrado no quadro que segue.

```
#
#     Cálculo de resíduos padronizados e estudentizados
Rpad=rstandard(m1)
Rstd=rstudent(m1)
D=round(cbind(m1$fit,m1$res, Rpad, Rstd),3)
colnames(D)=c('y.chap','resid','r.pad','r.est')
D
```

Nas instruções mostradas anteriormente, a primeira linha calcula os resíduos padronizados, a segunda calcula os resíduos estudentizados, a terceira linha junta as informações de interesse em uma única matriz, `D`, a quarta linha dá nomes às colunas da matriz `D`, e a quinta linha explicita o conteúdo de `D`.

4.3.15 Introdução ao planejamento de experimentos

As técnicas de planejamento de experimentos têm aplicação muito ampla. São exemplos: a agricultura, a indústria e a pesquisa em geral. Uma área de particular interesse no que se refere ao meio ambiente são os processos industriais.

As técnicas de planejamento de experimentos são muito úteis para calibrar processos produtivos e, assim, obter:

1. Melhor rendimento com reduções em matéria-prima e energia por unidade produzida;
2. Redução da variabilidade e consequentemente melhor qualidade e atendimento às especificações do produto;
3. Redução nos tempos de desenvolvimento;
4. Redução de custos;
5. Redução na produção de resíduos.

Em processos industriais em geral, são muitas as variáveis controláveis que podem ser utilizadas para atingir os objetivos citados anteriormente. São exemplos: temperatura, pressão, velocidade, tipos de materiais utilizados, etc. A essas variáveis controláveis chamamos fatores.

Num experimento planejado, temos a variável resposta, que é medida em todos os elementos analisados, e os fatores, que são objeto de verificação para saber se sua influência sobre a resposta é significativa ou não. Cada fator tem dois

ou mais níveis, os quais denominamos tratamentos. A cada combinação de níveis de diferentes fatores, temos um número de observações a que chamamos réplicas.

A aleatorização é de fundamental importância nos experimentos. Ela evita que o efeito de variáveis não controladas (perturbadoras) venha a comprometer os resultados obtidos.

4.3.15.1 Experimento completamente aleatorizado com um fator

Um experimento é dito completamente aleatorizado quando cada unidade amostral – unidade que gera uma observação da resposta para posterior análise – é atribuída sem restrições aos tratamentos.

Exemplo 4.30 Uma linha de produção de peças plásticas tem como opção dois tipos de materiais, A e B. O fabricante quer saber se há diferença entre os materiais no que se refere à resistência da peça à quebra.

Como não existe suspeita de outra variável interferindo na resistência da peça, o tipo de material é a única variável explicativa no modelo. Caso houvesse suspeitas sobre outras variáveis, estas deveriam fazer parte do modelo.

Temos, portanto, apenas um fator – o material a ser usado na fabricação – com dois níveis, A e B.

Como temos disponibilidade para 20 testes de resistência, serão produzidas 10 peças com material A e 10 com material B. Logo, temos 10 réplicas.

O próximo passo é definir a aleatorização do experimento. Nesse caso, atribuímos aleatoriamente – por sorteio – o tratamento – material A ou B – a cada peça a ser produzida, ou seja, um experimento completamente aleatorizado. A sequência do experimento pode ser observada na Tabela 4.5.

Tabela 4.5 Sequência de implementação dos experimentos.

Sequência	Material	Sequência	Material	Sequência	Material	Sequência	Material
1	B	6	A	11	A	16	B
2	A	7	A	12	B	17	A
3	B	8	A	13	B	18	A
4	A	9	B	14	A	19	B
5	B	10	B	15	A	20	B

A instrução `sample(.)` do R pode ser utilizada para fazer a alocação aleatória de tratamentos. O quadro a seguir mostra como fazê-la.

```
#
#  Selecao de amostra para Experimento Completamente
#  aleatorizado com um fator
#
#     Determina o número de níveis
niveis=2
#
#     Determina o número de réplicas
replica=10
#
#     Seleciona a amostra
x=sample(rep(1:niveis,10))
#
#     Organiza saída
cbind(1:(niveis*replica),x)
#
```

Nesse ponto, temos, no exemplo anterior, o planejamento de um experimento completamente aleatorizado, com um fator, dois níveis e 10 replicações.

O passo seguinte é a realização do experimento propriamente dito, o que dá origem aos dados. Não há muito o que discutir neste ponto, já que depende muito dos aspectos técnicos de cada área, seja ela química, mecânica, elétrica, farmacêutica, da saúde etc. O que deve ser lembrado sempre é que os procedimentos de implementação do experimento devem ser padronizados (homogeneizados), evitando a criação de variabilidade extra nos resultados.

Feito o experimento e obtidos os dados, passamos a discutir a análise dos dados produzidos. Essa análise é chamada de análise de variância ou ANOVA.

Os dados gerados por um experimento completamente aleatorizado com um fator, com “a” níveis (linhas da tabela) e “n” réplicas (colunas da tabela), pode ser apresentado como mostrado na Tabela 4.6. Como podemos observar, além de listar os dados, a Tabela apresenta, nas duas últimas colunas, o total e a média para cada nível (linha) do fator.

Tabela 4.6 Apresentação dos dados.

Tratamentos	Observações				Totais	Médias
1	y_{11}	y_{12}	...	y_{1n}	$y_{1\cdot}$	$\bar{y}_{a\cdot}$
2	y_{21}	y_{22}	...	y_{2n}	$y_{2\cdot}$	$\bar{y}_{a\cdot}$
.	.	.	...	.	.	.
.	.	.	...	.	.	.
.	.	.	...	.	.	.
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}$

O modelo linear associado ao experimento pode ser descrito por

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

em que i = 1, 2,..., a e j = 1, 2,..., n; μ é a média global; τ_i é o efeito do tratamento i; e ε_{ij} é o erro aleatório associado à observação j do tratamento i.

Nesse tipo de experimento, estamos interessados em testar hipóteses do tipo

$$H_0: \tau_1 = \tau_2 = ... = \tau_a = 0 \quad \textit{versus} \quad H_1: \tau_i \neq 0,$$

para pelo menos um nível do fator i, ou seja, queremos saber se existe efeito de tratamento para pelo menos um nível.

O teste é implementado em procedimento similar ao utilizado em regressão linear, mostrado anteriormente, sendo a componente regressão substituída por tratamento. A tabela ANOVA correspondente é mostrada a seguir.

Tabela 4.7 Tabela ANOVA para experimento com um fator.

Fontes de variação	Graus de liberdade	Soma de quadrados	Quadrados Médios	Estatística F	Pr(> F)
Tratamento	a – 1	SQTrat	QMTrat	F	Valor-p
Erro	a(n – 1)	SQE	QME		
Total	an – 1	SQTot			

Os elementos da tabela ANOVA são definidos como segue:

$$SQTot = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij}\right)^2 - an\bar{y}^2;$$

$$SQTrat = \sum_{i=1}^{a}\left(y_{i\circ}\right)^2 - an\bar{y}^2;$$

$SQE = SQTot - SQTrat;$

$QMTrat = SQTrat/(a-1);$

$QME = SQE/\left[a(n-1)\right];$ e

$F = QMTrat/QME.$

A estatística F segue a distribuição F com $(a-1)$ e $\left[a(n-1)\right]$ graus de liberdade, ou seja, $F \sim F_{(a-1),\left[a(n-1)\right]}$.

Exemplo 4.31 Adicionando os dados da Tabela 4.8 ao Exemplo 4.30, podemos realizar a análise que segue.

Tabela 4.8 Dados sobre resistência à quebra para materiais A e B.

Tratamento\Obs.	1	2	3	4	5	6	7	8	9	10	Soma	Médias
A	19	18	15	17	18	28	23	22	23	23	206	20,6
B	16	17	15	13	15	20	25	25	25	24	195	19,5

Uma abordagem inicial dos dados pode ser feita por meio de um box-plot duplo com os tipos de materiais mostrados separadamente, como mostra a Figura 4.10.

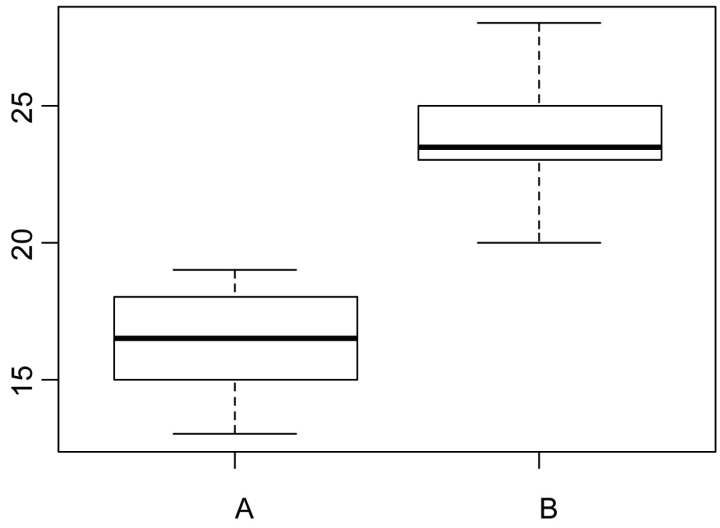


Figura 4.10 Box-plot para a resistência à quebra para materiais A e B.

Obtemos os componentes da tabela ANOVA utilizando as expressões precedentes e mostramos na Tabela 4.9.

$$SQTot = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij}\right)^2 - an\bar{y}^2 = 8393 - (2)(10)(20,05)^2 = 352,95;$$

$$SQTrat = \sum_{i=1}^{a} n\left(_{i\circ}\right)^2 - an^2 = 8321,3 - (2)(10)(20,05)^2 = 281,25;$$

$$SQE = SQTot - SQTrat = 352,95 - 281,25 = 71,7;$$

$$QMTrat = SQTrat/(a-1) = 281,25/1 = 281,25;$$

e

$$F = QMTrat/QME = 281,25/3,98 = 70,66.$$

Tabela 4.9 Tabela ANOVA para experimento com um fator.

Fontes de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	Estatística F	Valor-p
Tratamento	1	281,25	281,25	70,66	1,2E-7
Erro	18	71,7	3,98		
Total	19	352,95			

A última coluna pode ser obtida de uma tabela da distribuição F1,18 ou calculada no R pela instrução `1-pf(70.66,1,18)`. O valor-p obtido é tão pequeno que utilizamos notação científica para apresentá-lo. Nesse caso, o valor da expressão 1,2E – 7 é $1,2 \times 10^{-7}$.

O valor-p corresponde à área sob a função de densidade da distribuição $F_{1,18}$ a partir de 70,66. Sua utilidade é nos indicar que testes de hipóteses, na forma H_0: $\tau_1 = \tau_2 = ... = \tau_a = 0$ *versus* H_1: $\tau_i \neq 0$, para pelo menos um i, com nível de significância, α, maiores que seu valor seriam rejeitados. Como usamos $\alpha = 0,05$, rejeitamos H_0 e, portanto, rejeitamos a hipótese de igualdade de resistência entre os materiais.

A conclusão apresentada pelo teste concorda com a informação visual do box-plot apresentado na Figura 4.10, que mostra o limite inferior para o material B acima do limite superior para o material A.

Nota: Embora as apresentações gráficas sejam muito úteis e nos permitam visualizar o que ocorre com os dados, como é o caso neste exemplo, conclusões não devem ser baseadas em gráficos e sim em procedimentos inferenciais, ou seja, intervalos de confiança ou testes de hipóteses, entre os métodos tratados neste texto.

O quadro a seguir mostra como realizar a análise dos dados do exemplo no R. É importante que o leitor execute essas instruções no R e compare os resultados obtidos com os dados no exemplo precedente.

```
#
# Registro dos dados no R
resistencia=c(19,18,15,17,18,28,23,22,23,23,16,17,15,13,15,20,2
5,25,25,24)
material=factor(rep(c('A','B'),c(10,10)))
#
# Construção do box-plot apresentado na Figura 4.10
boxplot(resistencia~material)
#
# Cálculo das médias
mean(resistencia[1:10])
mean(resistencia[11:20])
#
# Cálculo das Somas
sum(resistencia[1:10])
sum(resistencia[11:20])
#
# Ajuste do modelo linear
m1=lm(y~material)
#
# Análise de variância do modelo ajustado
anova(m1)
```

Nota: Observando a Tabela 4.9, notamos que, em todos os níveis, o número de observações é o mesmo, n. Por esse motivo, esse plano de experimentos é chamado de balanceado. Se tivermos um número diferente de observações para pelo menos um nível do fator, diremos que o experimento está desbalanceado.

O tratamento de experimentos desbalanceados, bem como outras estratégias de planejamento de experimentos, está fora dos objetivos deste material.

Recomendamos aos leitores interessados no assunto que recorram à literatura especializada em planejamento de experimentos.

4.4 Considerações finais

Nesta Unidade, apresentamos os principais instrumentos da estatística frequentista, que são o intervalo de confiança e o teste de hipótese. Também apresentamos introduções muito breves de métodos específicos, como amostragem, regressão linear e planejamento de experimentos. Todos os tópicos foram acompanhados das instruções para sua implementação no R.

4.5 Atividades de aplicação, prática e avaliação

É altamente recomendável que o leitor exercite a sua capacidade de experimentação com os métodos apresentados. Busque, no seu dia a dia, fontes de dados que podem ser utilizadas para aplicar os métodos descritos anteriormente. Procure aplicá-las manualmente e utilizando o R.

4.5.1 Atividades individuais

Busque, no seu ambiente de trabalho, conjuntos de dados, aplique as técnicas descritas aqui e veja quais informações você consegue extrair desses dados.

4.5.2 Atividades coletivas

Troque ideias com seus colegas sobre os métodos que você aplicou aos seus dados e verifique se eles(as) utilizariam o mesmo método que você utilizou.

4.6 Estudos complementares

Recomendamos a leitura dos capítulos correspondentes à inferência estatística nos textos citados na lista de referências.

4.6.1 Saiba mais

Você pode ampliar os seus conhecimentos de inferência estatística incluindo novos métodos em seu repertório ou aprofundando seus conhecimentos dos métodos já tratados estudando as referências.

REFERÊNCIAS

BUSSAB, W. O.; MORETTIN, P. A. *Estatística básica*. São Paulo: Saraiva, 2006.

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. *Statistics for experimenters*. Nova Iorque (EUA): John Willey & Sons, 1978.

COCHRAN, W. G. *Sampling techniques*. Nova Iorque (EUA): John Wiley & Sons, 1977.

MCCABE, G. P.; DUCKWORTH, W. M.; SCLOVE, S. L. *Estatística empresarial*. Tradução de Luis Antonio Fajardo. São Paulo: LTC, 2006.

MONTGOMERY, D. C.; RUNGER, G. C. *Estatística Aplicada e probabilidade para engenheiros*. Tradução de Verônica Calado. São Paulo: LTC, 2003.

MOORE, D. S. *A estatística básica e sua prática*. Tradução de Cristiana Filizola Carneiro Pessoa. São Paulo: LTC, 2005.

MOORE, D. S.; MCCABE, G. P.; DUCKWORTH, W. M.; SCLOVE, S. L. *Estatística empresarial*. Tradução de Luis Antonio Fajardo. São Paulo: LTC, 2006.

R DEVELOPMENT CORE TEAM. *R reference manual*. Bristol (ENG): Network Theory, 2003.

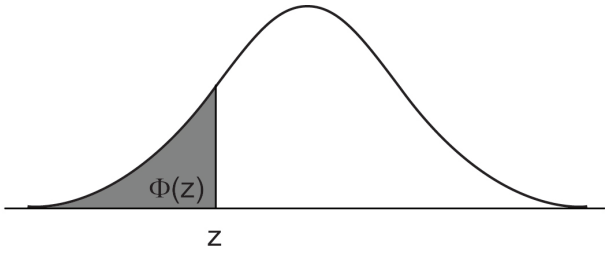
VENABLES, W. N.; SMITH, D. M. *An introduction to R*. 2008. Disponível em: <http://cran.r-project.org/doc/manuals/R-intro.pdf>. Acesso em: 29 jul. 2010.

APÊNDICE

Tabelas

APÊNDICE A: Tabela da Distribuição Normal

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \Phi(z)dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(\frac{-z^2}{\sigma^2}\right)dz$$



z	-0,09	-0,08	-0,07	-0,06	-0,05	-0,04	-0,03	-0,02	-0,01	-0,00
-3.9	0.000033	0.000034	0.000036	0.000037	0.000039	0.000041	0.000042	0.000044	0.000046	0.000048
-3.8	0.000050	0.000052	0.000054	0.000057	0.000059	0.000062	0.000064	0.000067	0.000069	0.000072
-3.7	0.000075	0.000078	0.000082	0.000085	0.000088	0.000092	0.000096	0.000100	0.000104	0.000108
-3.6	0.000112	0.000117	0.000121	0.000126	0.000131	0.000136	0.000142	0.000147	0.000153	0.000159
-3.5	0.000165	0.000172	0.000178	0.000185	0.000193	0.000200	0.000208	0.000216	0.000224	0.000233
-3.4	0.000242	0.000251	0.000260	0.000270	0.000280	0.000291	0.000302	0.000313	0.000325	0.000337
-3.3	0.000349	0.000362	0.000376	0.000390	0.000404	0.000419	0.000434	0.000450	0.000466	0.000483
-3.2	0.000501	0.000519	0.000538	0.000557	0.000577	0.000598	0.000619	0.000641	0.000664	0.000687
-3.1	0.000711	0.000736	0.000762	0.000789	0.000816	0.000845	0.000874	0.000904	0.000935	0.000968
-3.0	0.001001	0.001035	0.001070	0.001107	0.001144	0.001183	0.001223	0.001264	0.001306	0.001350
-2.9	0.001395	0.001441	0.001489	0.001538	0.001589	0.001641	0.001695	0.001750	0.001807	0.001866
-2.8	0.001926	0.001988	0.002052	0.002118	0.002186	0.002256	0.002327	0.002401	0.002477	0.002555
-2.7	0.002635	0.002718	0.002803	0.002890	0.002980	0.003072	0.003167	0.003264	0.003364	0.003467
-2.6	0.003573	0.003681	0.003793	0.003907	0.004025	0.004145	0.004269	0.004396	0.004527	0.004661
-2.5	0.004799	0.004940	0.005085	0.005234	0.005386	0.005543	0.005703	0.005868	0.006037	0.006210
-2.4	0.006387	0.006569	0.006756	0.006947	0.007143	0.007344	0.007549	0.007760	0.007976	0.008198
-2.3	0.008424	0.008656	0.008894	0.009137	0.009387	0.009642	0.009903	0.010170	0.010444	0.010724
-2.2	0.011011	0.011304	0.011604	0.011911	0.012224	0.012545	0.012874	0.013209	0.013553	0.013903
-2.1	0.014262	0.014629	0.015003	0.015386	0.015778	0.016177	0.016586	0.017003	0.017429	0.017864
-2.0	0.018309	0.018763	0.019226	0.019699	0.020182	0.020675	0.021178	0.021692	0.022216	0.022750
-1.9	0.023295	0.023852	0.024419	0.024998	0.025588	0.026190	0.026803	0.027429	0.028067	0.028717
-1.8	0.029379	0.030054	0.030742	0.031443	0.032157	0.032884	0.033625	0.034380	0.035148	0.035930
-1.7	0.036727	0.037538	0.038364	0.039204	0.040059	0.040930	0.041815	0.042716	0.043633	0.044565
-1.6	0.045514	0.046479	0.047460	0.048457	0.049471	0.050503	0.051551	0.052616	0.053699	0.054799
-1.5	0.055917	0.057053	0.058208	0.059380	0.060571	0.061780	0.063008	0.064255	0.065522	0.066807
-1.4	0.068112	0.069437	0.070781	0.072145	0.073529	0.074934	0.076359	0.077804	0.079270	0.080757
-1.3	0.082264	0.083793	0.085343	0.086915	0.088508	0.090123	0.091759	0.093418	0.095098	0.096800
-1.2	0.098525	0.100273	0.102042	0.103835	0.105650	0.107488	0.109349	0.111232	0.113139	0.115070
-1.1	0.117023	0.119000	0.121000	0.123024	0.125072	0.127143	0.129238	0.131357	0.133500	0.135666
-1.0	0.137857	0.140071	0.142310	0.144572	0.146859	0.149170	0.151505	0.153864	0.156248	0.158655
-0.9	0.161087	0.163543	0.166023	0.168528	0.171056	0.173609	0.176186	0.178786	0.181411	0.184060
-0.8	0.186733	0.189430	0.192150	0.194895	0.197663	0.200454	0.203269	0.206108	0.208970	0.211855
-0.7	0.214764	0.217695	0.220650	0.223627	0.226627	0.229650	0.232695	0.235762	0.238852	0.241964
-0.6	0.245097	0.248252	0.251429	0.254627	0.257846	0.261086	0.264347	0.267629	0.270931	0.274253
-0.5	0.277595	0.280957	0.284339	0.287740	0.291160	0.294599	0.298056	0.301532	0.305026	0.308538
-0.4	0.312067	0.315614	0.319178	0.322758	0.326355	0.329969	0.333598	0.337243	0.340903	0.344578
-0.3	0.348268	0.351973	0.355691	0.359424	0.363169	0.366928	0.370700	0.374484	0.378280	0.382089
-0.2	0.385908	0.389739	0.393580	0.397432	0.401294	0.405165	0.409046	0.412936	0.416834	0.420740
-0.1	0.424655	0.428576	0.432505	0.436441	0.440382	0.444330	0.448283	0.452242	0.456205	0.460172
0.0	0.464144	0.468119	0.472097	0.476078	0.480061	0.484047	0.488034	0.492022	0.496011	0.500000

APÊNDICE B: Tabela da Distribuição t de Student

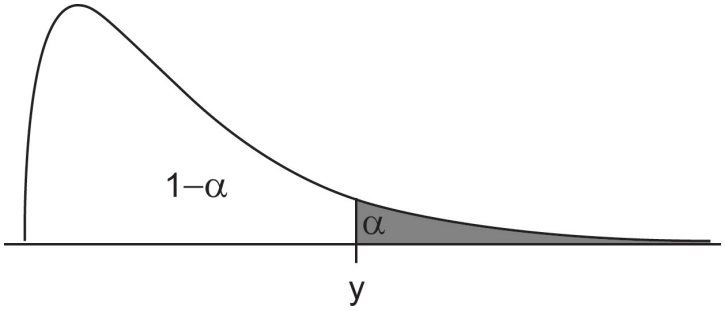
$$F(t;v) = \int_{-\infty}^{t} \frac{\Gamma\left((v+1)/2\right)}{\Gamma\left(v/2\right)\sqrt{v\pi}}\left(1+t^2/v\right)^{-(v+1)/2}$$

v\F(.)	0.9	0,95	0,975	0,99	0,995	0,9995
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460
35	1.3062	1.6896	2.0301	2.4377	2.7238	3.5911
40	1.3031	1.6839	2.0211	2.4233	2.7045	3.5510
45	1.3006	1.6794	2.0141	2.4121	2.6896	3.5203
50	1.2987	1.6759	2.0086	2.4033	2.6778	3.4960
55	1.2971	1.6730	2.0040	2.3961	2.6682	3.4764
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.4602
120	1.2886	1.6577	1.9799	2.3578	2.6174	3.3735

Nota: v é o número de graus de liberdade (primeira coluna da tabela); F(.) corresponde à probabilidade acumulada (primeira linha da tabela); e no corpo da tabela está o valor de t.

APÊNDICE C: Tabela da Distribuição Qui-Quadrado

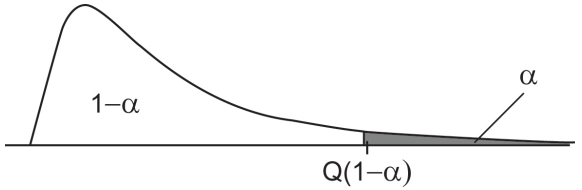
$$\alpha = \int_{y}^{\infty} \frac{1}{\Gamma(v/2)2^{v/2}} x^{v/2-1} \; e^{\;-y/2} dx$$



gl \ α	0.990	0.980	0.975	0.950	0.900	0.100	0.050	0.025	0.020	0.010
1	0.000	0.001	0.001	0.004	0.016	2.706	3.841	5.024	5.412	6.635
2	0.020	0.040	0.051	0.103	0.211	4.605	5.991	7.378	7.824	9.210
3	0.115	0.185	0.216	0.352	0.584	6.251	7.815	9.348	9.837	11.345
4	0.297	0.429	0.484	0.711	1.064	7.779	9.488	11.143	11.668	13.277
5	0.554	0.752	0.831	1.145	1.610	9.236	11.070	12.833	13.388	15.086
6	0.872	1.134	1.237	1.635	2.204	10.645	12.592	14.449	15.033	16.812
7	1.239	1.564	1.690	2.167	2.833	12.017	14.067	16.013	16.622	18.475
8	1.646	2.032	2.180	2.733	3.490	13.362	15.507	17.535	18.168	20.090
9	2.088	2.532	2.700	3.325	4.168	14.684	16.919	19.023	19.679	21.666
10	2.558	3.059	3.247	3.940	4.865	15.987	18.307	20.483	21.161	23.209
11	3.053	3.609	3.816	4.575	5.578	17.275	19.675	21.920	22.618	24.725
12	3.571	4.178	4.404	5.226	6.304	18.549	21.026	23.337	24.054	26.217
13	4.107	4.765	5.009	5.892	7.042	19.812	22.362	24.736	25.472	27.688
14	4.660	5.368	5.629	6.571	7.790	21.064	23.685	26.119	26.873	29.141
15	5.229	5.985	6.262	7.261	8.547	22.307	24.996	27.488	28.259	30.578
16	5.812	6.614	6.908	7.962	9.312	23.542	26.296	28.845	29.633	32.000
17	6.408	7.255	7.564	8.672	10.085	24.769	27.587	30.191	30.995	33.409
18	7.015	7.906	8.231	9.390	10.865	25.989	28.869	31.526	32.346	34.805
19	7.633	8.567	8.907	10.117	11.651	27.204	30.144	32.852	33.687	36.191
20	8.260	9.237	9.591	10.851	12.443	28.412	31.410	34.170	35.020	37.566
21	8.897	9.915	10.283	11.591	13.240	29.615	32.671	35.479	36.343	38.932
22	9.542	10.600	10.982	12.338	14.041	30.813	33.924	36.781	37.659	40.289
23	10.196	11.293	11.689	13.091	14.848	32.007	35.172	38.076	38.968	41.638
24	10.856	11.992	12.401	13.848	15.659	33.196	36.415	39.364	40.270	42.980
25	11.524	12.697	13.120	14.611	16.473	34.382	37.652	40.646	41.566	44.314
26	12.198	13.409	13.844	15.379	17.292	35.563	38.885	41.923	42.856	45.642
27	12.879	14.125	14.573	16.151	18.114	36.741	40.113	43.195	44.140	46.963
28	13.565	14.847	15.308	16.928	18.939	37.916	41.337	44.461	45.419	48.278
29	14.256	15.574	16.047	17.708	19.768	39.087	42.557	45.722	46.693	49.588
30	14.953	16.306	16.791	18.493	20.599	40.256	43.773	46.979	47.962	50.892

APÊNDICE D: Tabela para a distribuição F

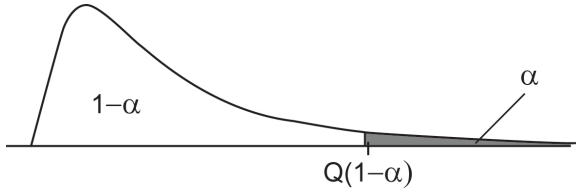
Tabela F para $\alpha = 0{,}10$



$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.47	60.71	60.90	61.07	61.22	61.35	61.46	61.57	61.66	61.74	62.05	62.26	62.53	62.79	63.06	63.32
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41	9.41	9.42	9.42	9.43	9.43	9.44	9.44	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22	5.21	5.20	5.20	5.20	5.19	5.19	5.19	5.18	5.17	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90	3.89	3.88	3.87	3.86	3.86	3.85	3.85	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27	3.26	3.25	3.24	3.23	3.22	3.22	3.21	3.21	3.19	3.17	3.16	3.14	3.12	3.11
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90	2.89	2.88	2.87	2.86	2.85	2.85	2.84	2.84	2.81	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67	2.65	2.64	2.63	2.62	2.61	2.61	2.60	2.59	2.57	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50	2.49	2.48	2.46	2.45	2.45	2.44	2.43	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38	2.36	2.35	2.34	2.33	2.32	2.31	2.30	2.30	2.27	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28	2.27	2.26	2.24	2.23	2.22	2.22	2.21	2.20	2.17	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21	2.19	2.18	2.17	2.16	2.15	2.14	2.13	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15	2.13	2.12	2.10	2.09	2.08	2.08	2.07	2.06	2.03	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10	2.08	2.07	2.05	2.04	2.03	2.02	2.01	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.07	2.05	2.04	2.02	2.01	2.00	1.99	1.98	1.97	1.96	1.93	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.96	1.95	1.94	1.93	1.92	1.89	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99	1.97	1.95	1.94	1.93	1.92	1.91	1.90	1.89	1.86	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96	1.94	1.93	1.91	1.90	1.89	1.88	1.87	1.86	1.83	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.95	1.93	1.92	1.90	1.89	1.87	1.86	1.85	1.84	1.84	1.80	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.93	1.91	1.89	1.88	1.86	1.85	1.84	1.83	1.82	1.81	1.78	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.91	1.89	1.87	1.86	1.84	1.83	1.82	1.81	1.80	1.79	1.76	1.74	1.71	1.68	1.64	1.61
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.84	1.82	1.80	1.79	1.77	1.76	1.75	1.74	1.73	1.72	1.68	1.66	1.63	1.59	1.56	1.52
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77	1.75	1.74	1.72	1.71	1.70	1.69	1.68	1.67	1.63	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.74	1.71	1.70	1.68	1.66	1.65	1.64	1.62	1.61	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66	1.64	1.62	1.60	1.59	1.58	1.56	1.55	1.54	1.50	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.63	1.60	1.58	1.56	1.55	1.53	1.52	1.50	1.49	1.48	1.44	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.95	1.85	1.77	1.72	1.67	1.63	1.60	1.57	1.55	1.52	1.51	1.49	1.47	1.46	1.44	1.43	1.42	1.38	1.34	1.30	1.24	1.17	1.03

APÊNDICE D: Tabela para a distribuição (*continuação...*)

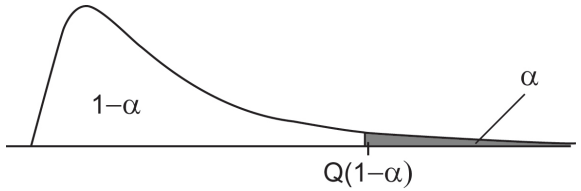
Tabela F para $\alpha = 0{,}05$



$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	60	120	∞
1	161.4	199.5	215.7	224.5	230.1	233.9	236.7	238.8	240.5	241.8	242.9	243.9	244.6	245.3	245.9	246.4	246.9	247.3	247.7	248.0	249.3	250.1	251.1	252.2	253.2	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45	19.46	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56	4.52	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15	3.11	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77	2.73	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65	2.60	2.57	2.53	2.49	2.45	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54	2.50	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46	2.41	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39	2.34	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33	2.28	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30	2.29	2.28	2.23	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23	2.18	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19	2.14	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12	2.07	2.04	1.99	1.95	1.90	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.07	2.05	2.04	2.02	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.99	1.98	1.96	1.95	1.93	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.90	1.89	1.87	1.85	1.84	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.82	1.80	1.78	1.76	1.75	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.83	1.80	1.78	1.75	1.73	1.71	1.69	1.67	1.66	1.60	1.55	1.50	1.43	1.35	1.26
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.72	1.69	1.67	1.64	1.62	1.60	1.59	1.57	1.51	1.46	1.40	1.32	1.22	1.03

APÊNDICE D: Tabela para a distribuição (*continuação...*)

Tabela F para $\alpha = 0{,}01$



$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	40	60	120	∞
1	4052	4999	5403	5624	5763	5858	5928	5981	6022	6055	6083	6106	6125	6142	6157	6170	6181	6191	6200	6208	6239	6260	6286	6313	6339	6365
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.42	99.43	99.43	99.44	99.44	99.44	99.45	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.98	26.92	26.87	26.83	26.79	26.75	26.72	26.69	26.58	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25	14.20	14.15	14.11	14.08	14.05	14.02	13.91	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77	9.72	9.68	9.64	9.61	9.58	9.55	9.45	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56	7.52	7.48	7.45	7.42	7.40	7.30	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31	6.28	6.24	6.21	6.18	6.16	6.06	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.48	5.44	5.41	5.38	5.36	5.26	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96	4.92	4.89	4.86	4.83	4.81	4.71	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.52	4.49	4.46	4.43	4.41	4.31	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.21	4.18	4.15	4.12	4.10	4.01	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.97	3.94	3.91	3.88	3.86	3.76	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.78	3.75	3.72	3.69	3.66	3.57	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.62	3.59	3.56	3.53	3.51	3.41	3.35	3.27	3.18	3.09	3.01
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.49	3.45	3.42	3.40	3.37	3.28	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.37	3.34	3.31	3.28	3.26	3.16	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.27	3.24	3.21	3.19	3.16	3.07	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.19	3.16	3.13	3.10	3.08	2.98	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15	3.12	3.08	3.05	3.03	3.00	2.91	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09	3.05	3.02	2.99	2.96	2.94	2.84	2.78	2.69	2.61	2.52	2.42
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89	2.85	2.81	2.78	2.75	2.72	2.70	2.60	2.54	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70	2.66	2.63	2.60	2.57	2.55	2.45	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.61	2.56	2.52	2.48	2.45	2.42	2.39	2.37	2.27	2.20	2.11	2.02	1.92	1.81
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.44	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.10	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.28	2.23	2.19	2.15	2.12	2.09	2.06	2.03	1.93	1.86	1.76	1.66	1.53	1.38
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.19	2.13	2.08	2.04	2.00	1.97	1.94	1.91	1.88	1.77	1.70	1.59	1.48	1.33	1.05

SOBRE O AUTOR

Luis Aparecido Milan

Graduado em Estatística pela Universidade Estadual de Campinas (Unicamp) em 1978, onde também obteve o título de mestre em 1987. Obteve o título de PhD pela Universidade de Lancaster, Inglaterra, em 1993. Atuou em diversas empresas de 1978 a 1985. É docente no Departamento de Estatística da UFSCar desde 1985.

Este livro foi impresso pelo Departamento de Produção Gráfica – UFSCar.