

ESTATÍSTICA BÁSICA

AUTORA

VALÉRIA FERREIRA



ESTATÍSTICA BÁSICA

AUTOR

FERNANDO DE FIGUEIREDO BALIEIRO

1^a EDIÇÃO

SESES

RIO DE JANEIRO 2015



Estácio

Conselho editorial SERGIO AUGUSTO CABRAL; ROBERTO PAES; GLADIS LINHARES

Autora do original VALÉRIA APARECIDA FERREIRA

Projeto editorial ROBERTO PAES

Coordenação de produção GLADIS LINHARES

Projeto gráfico PAULO VITOR BASTOS

Diagramação BFS MEDIA

Revisão linguística AMANDA DUARTE AGUIAR

Revisão de conteúdo PAULA TAVARES DA CUNHA MELO

Imagem de capa PAVALACHE STELIAN | DREAMSTIME.COM

Todos os direitos reservados. Nenhuma parte desta obra pode ser reproduzida ou transmitida por quaisquer meios (eletrônico ou mecânico, incluindo fotocópia e gravação) ou arquivada em qualquer sistema ou banco de dados sem permissão escrita da Editora. Copyright SESES, 2015.

Dados Internacionais de Catalogação na Publicação (CIP)

F383E FERREIRA, VALÉRIA

Estatística básica / Valéria Ferreira

Rio de Janeiro: SESES, 2015.

184 p.: il.

ISBN: 978-85-5548-129-1

1. Probabilidade. 2. Funções de variáveis. 3. Regressão Linear.

I. SESES. II. Estácio.

CDD 519.2

Diretoria de Ensino — Fábrica de Conhecimento
Rua do Bispo, 83, bloco F, Campus João Uchôa
Rio Comprido — Rio de Janeiro — RJ — CEP 20261-063

Sumário

1. Conceitos Iniciais e Apresentação dos Dados por meio de Distribuições de Frequências e Gráficos	7
Objetivos	8
1.1 Definição de Estatística	9
1.2 Conceitos básicos da Estatística	10
1.3 Coleta de dados	16
1.3.1 Técnicas de amostragem	17
1.3.1.1 Técnicas de amostragem probabilística (ou aleatória)	18
1.3.1.1.1 Amostragem aleatória simples	18
1.3.1.1.2 Amostragem estratificada	19
1.3.1.1.3 Amostragem sistemática	20
1.3.1.1.4 Amostragem por conglomerado	21
1.3.1.2 Técnicas de amostragem não probabilística (ou não aleatória)	22
1.3.1.2.1 Amostragem por conveniência	22
1.3.1.2.2 Amostragem por quota	23
1.4 Distribuição de frequências	25
1.5 Gráficos	32
1.5.1 Tipos de gráficos	33
1.5.1.1 Gráfico de linhas	33
1.5.1.2 Gráfico de barras	35
1.5.1.3 Gráfico de setores	37
1.5.1.4 Histograma	38
1.5.1.5 Polígono de frequências	39
1.5.1.6 Diagrama de Pareto	39
1.5.1.7 Diagrama de dispersão	41
1.6 Utilização do Microsoft Excel na Construção de Gráficos	43
Reflexão	46
Referências bibliográficas	47

2. Medidas Resumo 49

Objetivos	50
2.1 Medidas de tendência central	51
2.1.1 Média aritmética	51
2.1.1.1 Propriedades da média	54
2.1.2 Moda	54
2.1.3 Mediana	55
2.1.4 Cálculos das medidas de tendência central para dados agrupados em intervalos de classes	59
2.2 Medidas de dispersão	63
2.2.1 Mínimo, máximo e amplitude	64
2.2.2 Desvio médio, variância e desvio padrão amostrais	65
2.2.2.1 Uma regra prática para interpretar o desvio-padrão	67
2.2.2.2 Propriedades do desvio padrão	68
2.2.3 Coeficiente de variação	72
2.2.4 Cálculos da variância e do desvio padrão para dados agrupados em intervalos de classes	73
2.3 Medidas separatrizes ou de ordenamento	75
2.3.1 Quartis	75
2.3.2 Decis e Percentis	77
2.3.3 Cálculo das medidas separatrizes para dados agrupados em intervalos de classes	78
2.4 Medidas de assimetria e curtose	86
2.5 Utilização do Microsoft Excel na Análise de Dados	92
Reflexão	95
Referências bibliográficas	95

3. Distribuição de Probabilidade Normal 97

Objetivos	98
3.1 Variável aleatória	99
3.2 Distribuição Normal	99
3.3 Utilização do Microsoft Excel no cálculo de probabilidades normais	118
Reflexão	124
Referências bibliográficas	125

4. Teste de Hipóteses	127
Objetivos	128
4.1 Fundamentos do teste de hipóteses	129
4.2 Teste de hipóteses para a média populacional	131
4.2.1 Tipos de erros, nível de significância e estatística de teste	131
4.2.2 Decisão e interpretação	135
4.3 Teste de hipóteses para duas amostras	139
4.3.1 Testes para diferenças entre médias	140
4.3.1.1 Amostras independentes com desvios padrões desconhecidos e diferentes	141
4.3.1.2 Amostras independentes com desvios padrões desconhecidos e iguais	144
4.3.1.3 Amostras independentes com desvios padrões conhecidos	148
4.3.1.4 Amostras dependentes	150
4.4 Utilização do Microsoft Excel para testes de duas amostras	154
4.4.1 Comparação de duas médias com desvios padrões desconhecidos e diferentes	154
4.4.2 Comparação de duas médias (amostras dependentes)	157
Reflexão	162
Referências bibliográficas	163
5. Correlação e Regressão Linear Simples	165
Objetivos	166
5.1 Diagrama de dispersão	167
5.2 Coeficiente de correlação linear	168
5.3 Teste de hipóteses para correlação	173
5.4 Regressão linear simples	175
5.5 Coeficiente de determinação	181
5.6 Utilização do Microsoft Excel na análise de regressão e correlação	185
Reflexão	195
Referências bibliográficas	196

1

Conceitos Iniciais e Apresentação dos Dados por meio de Distribuições de Frequências e Gráficos

Nesse primeiro capítulo, estudaremos conceitos básicos da Estatística e como organizamos e apresentamos um conjunto de dados por meio de distribuições de frequências e gráficos apropriados.

Os conceitos abordados neste capítulo são muito importantes, pois qualquer estudo ou pesquisa deve ser conduzido a partir dos conhecimentos adquiridos neste primeiro momento, para que os resultados obtidos na análise sejam um instrumento confiável para tomadas de decisões.



OBJETIVOS

Após o estudo dos conceitos e técnicas que serão apresentados, esperamos que você consiga:

- Descrever a população e a amostra em um estudo;
 - Identificar e classificar os diferentes tipos de variáveis presentes em um estudo;
 - Compreender a que se destina cada uma das áreas da Estatística;
 - Entender as características dos vários tipos de amostragens probabilísticas utilizados para coleta de dados;
 - Construir distribuições de frequências e gráficos apropriados.
-

1.1 Definição de Estatística

É muito comum nos meios de comunicação, como jornais, revistas, televisão e internet, nos depararmos com informações estatísticas. Por exemplo:

- Os institutos de pesquisas divulgam com frequência resultados obtidos em pesquisas que têm por objetivo avaliar o governo do presidente em exercício.
- As taxas de cesárias, no Brasil, no sistema privado e no SUS.
- O percentual de aumento, ou redução, no preço da cesta básica.
- Incidência estimada de câncer de mama nos estados do Brasil.

Para que estas informações sejam obtidas, precisamos coletar dados para transformá-los em informações. Portanto, podemos definir a Estatística da seguinte maneira:

Estatística é um conjunto de técnicas utilizadas para a coleta, organização, resumo, análise e interpretação de dados.

Quando o foco está nas ciências biológicas e da saúde, usamos o termo bioestatística.

A Estatística tem um papel fundamental em diversas áreas do conhecimento, pois o uso de técnicas estatísticas apropriadas fornece informações que auxiliam no processo de tomada de decisões. Por exemplo, a eficácia de um novo medicamento para reduzir o LDL colesterol é feito por meio de um teste clínico com pacientes. A análise dos dados obtidos informará se a redução é estatisticamente significante.

Métodos estatísticos são essenciais no estudo de situações em que as variáveis de interesse estão sujeitas, inherentemente, a flutuações aleatórias. Isto acontece muito na área da saúde. Por exemplo, mesmo que o estudo seja feito com pacientes homogêneos, observamos uma grande variabilidade, por exemplo, na resposta a algum tipo de tratamento. Então, para estudar problemas clínicos, precisamos de uma metodologia capaz de tratar a variabilidade de forma adequada.

O avanço da informática e a popularização dos computadores contribuíram para o uso de métodos estatísticos. Antigamente, era muito demorado fazer análises de muitas informações, e agora, com o auxílio do computador, as análises são feitas rapidamente. Além disto, com o avanço da informática, novas técnicas de análise de dados foram introduzidas, principalmente métodos gráficos. Muitos pacotes estatísticos foram desenvolvidos e são usados tanto no meio acadêmico

como em indústrias, como, por exemplo, Minitab, SPSS e SAS. Utilizamos também o Microsoft Office Excel, que possui opções para certas técnicas estatísticas. Apesar do grande auxílio fornecido pelos pacotes estatísticos e pelo Excel, precisamos ter um conhecimento teórico sólido para saber qual técnica estatística utilizar para resolver um problema, além de saber analisar e interpretar os resultados obtidos.

A Estatística pode ser dividida em duas grandes áreas: a estatística descritiva e a inferência estatística.

Na estatística descritiva, utilizamos técnicas destinadas a organizar, descrever e resumir os dados. Os dados são tabulados e apresentados por meio de gráficos e resumidos através de medidas numéricas. Desta maneira, as informações estatísticas são apresentadas de maneira clara e de fácil entendimento.

Na inferência estatística (ou inferência indutiva), utilizamos dados amostrais para fazer estimativas, testar hipóteses e fazer previsões sobre características de uma população. Veremos, a seguir, alguns conceitos que facilitarão o entendimento da importância da inferência estatística.

1.2 Conceitos básicos da Estatística

CONCEITOS	
POPULAÇÃO	Conjunto formado por todos os elementos (pessoas, objetos, medidas, respostas e outros) que têm a característica que se deseja estudar.
AMOSTRA	Subconjunto representativo da população de interesse.
PARÂMETRO	Medida numérica que descreve alguma característica de uma população.
ESTATÍSTICA	Medida numérica que descreve alguma característica de uma amostra.

CONCEITOS

VARIÁVEL	Característica de interesse no estudo.
DADOS	Respostas coletadas da variável em estudo.
CENSO	Conjunto de dados obtidos através de todos os elementos da população.

Vale ressaltar que o termo população refere-se não somente a um conjunto de pessoas. Podemos citar alguns exemplos de populações: todos os habitantes da cidade de São Paulo; todos os carros produzidos, em determinado ano, por uma montadora; todos os acidentes ocorridos em determinada extensão de uma rodovia durante um feriado prolongado; todo o sangue no corpo de uma pessoa ou todos os pacientes traumatizados atendidos na Unidade de Emergência do Hospital das Clínicas de Ribeirão Preto da Universidade de São Paulo, no ano de 2014.

Em muitos estudos, é muito difícil podermos trabalhar com todos os elementos da população. Quando isto ocorre, retiramos um conjunto menor de elementos da população, que é denominado amostra.

A amostra é um subconjunto representativo da população de interesse e é por meio dela que o estudo estatístico é feito, de maneira a obtermos informações importantes sobre a população da qual a amostra foi extraída.

De acordo com Vieira (2008, p. 4).

As razões que levam os pesquisadores a trabalhar com amostras – e não com toda a população – são poucas, mas absolutamente relevantes.

- Custo e demora dos censos.
- Populações muito grandes.
- Impossibilidade física de examinar toda a população.
- Comprovado valor científico das informações coletadas por meio de amostras.

Podemos justificar a primeira razão, custo e demora dos censos, analisando as pesquisas eleitorais. As prévias eleitorais são feitas regularmente e publicadas. Analisar todos os milhões de eleitores do Brasil em um curto espaço de tempo torna-se impossível para o pesquisador. Vamos lembrar que nosso país possui uma vasta extensão territorial, fazendo com que a pesquisa leve muito tempo e gere um custo muito alto.

No caso de populações muito grandes, é impossível estudá-las por inteiro. Por exemplo, se temos interesse de estudar determinada planta em uma mata. O número de plantas é matematicamente finito, mas tão grande, que pode ser considerado infinito para qualquer estudo prático.

Em algumas situações, é impossível examinar toda a população. Por exemplo, na análise de sangue de uma pessoa, não podemos observar toda a população de interesse.

E, por fim, a coleta de dados por meio de uma amostra tem maior valor científico do que se estivéssemos estudando brevemente toda a população. Por exemplo, um pesquisador social tem interesse em estudar hábitos e comportamentos relacionados à saúde da criança e do adolescente de uma grande cidade brasileira. É melhor fazer uma avaliação criteriosa e cuidadosa de dados amostrais do que uma avaliação rápida e resumida de toda a população de crianças e adolescentes da cidade.

A Figura 1.1 ilustra os conceitos de população e amostra e as áreas da estatística descritiva e inferencial, com seus respectivos objetivos.

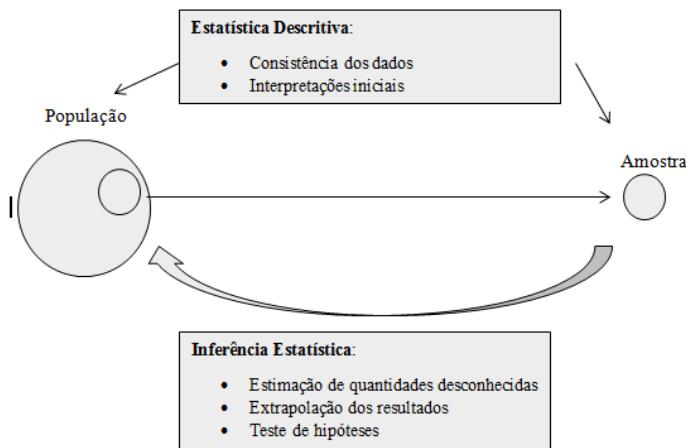


Figura 1.1 – População e amostra. Fonte: MAGALHÃES e LIMA (2004, p. 3).

Quando temos acesso a todos os elementos que desejamos estudar, ou seja, a população, não é necessário o uso de técnicas da inferência estatística.

Um levantamento de dados obtidos por meio de toda a população é chamado censo. Esta palavra é familiar, pois no nosso país, a cada 10 anos, o Instituto Brasileiro de Geografia e Estatística (IBGE) faz o Censo Demográfico do Brasil. Com as informações obtidas pelos censos, podemos conhecer a distribuição territorial e as principais características das pessoas e dos domicílios. Estas informações são imprescindíveis para a definição de políticas públicas e a tomada de decisões de investimentos.



CONEXÃO

Em épocas de recenseamento, uma declaração muito comum é: o recenseador não passou em minha residência. Para entender a metodologia adotada pelo IBGE, leia as informações disponíveis em: <<http://saladeimprensa.ibge.gov.br/noticias?view=noticia&id=1&busca=1&idnoticia=1866>> Acesso em: 30 de Abr. 2015.

Os dados obtidos por meio de uma população ou amostra, são provenientes da(s) variável(eis) em estudo. Variável é uma característica de interesse no estudo. Por exemplo, podemos ter interesse nas variáveis idade, gênero, renda e escolaridade dos clientes de determinada Unidade Básica de Saúde. As respostas obtidas em cada uma destas variáveis formarão o conjunto de dados a ser estudado.

Para uma melhor compreensão dos conceitos expostos acima, vamos analisar o exemplo a seguir.



EXEMPLO

1.1: Um hospital e maternidade possui 3 200 funcionários. O departamento de recursos humanos fez uma pesquisa de clima organizacional com 620 funcionários selecionados nos diversos setores do hospital e um dos tópicos abordados foi o grau de satisfação com os benefícios oferecidos pela empresa. A análise dos dados mostrou que 55% dos funcionários estão satisfeitos com os benefícios oferecidos. De acordo com as informações contidas no enunciado, vamos identificar:

- a) A população em estudo.
- b) A variável em estudo.
- c) O tamanho da amostra.
- d) A informação numérica 55% é um parâmetro ou uma estatística?

Resolução

- a) População em estudo: 3 200 funcionários do hospital e maternidade.
 - b) Variável em estudo: nível de satisfação com os benefícios oferecidos.
 - c) Tamanho da amostra: 620 funcionários.
 - d) A informação numérica 55% é uma estatística, pois esta informação foi obtida através de dados amostrais.
-

Quando coletamos dados referentes à variável ou às variáveis em estudo, podemos obter respostas numéricas ou não numéricas. É intuitivo pensar que quando as respostas são numéricas, estamos trabalhando com dados quantitativos e, quando as respostas não são numéricas, os dados são qualitativos.

No caso do Exemplo 1.1, os dados coletados são qualitativos, pois duas das possíveis respostas dos funcionários são: insatisfeito ou satisfeito.

Como os dados são provenientes das variáveis em estudo, podemos classificar as variáveis da mesma forma: variáveis qualitativas (ou categóricas) ou quantitativas.

As variáveis qualitativas podem ser classificadas como qualitativas nominais ou ordinais. Se existir uma ordenação natural, elas são classificadas como qualitativas ordinais. Caso contrário, elas são classificadas como variáveis qualitativas nominais. Por exemplo, variáveis como gênero (masculino e feminino) e estado civil (solteiro, casado, viúvo, etc.) são classificadas como qualitativas nominais. Agora, variáveis como desempenho de um profissional (péssimo, regular ou bom) e grau de instrução (ensino fundamental, ensino médio, superior) são classificadas como qualitativas ordinais.

No caso das variáveis quantitativas, elas podem ser classificadas como quantitativas discretas ou contínuas. As variáveis quantitativas discretas são resultantes de uma operação de contagem, assumindo respostas cujos números são inteiros. Já as variáveis quantitativas contínuas são resultantes de mensurações, assumindo valores que pertencem a um intervalo de números reais, ou seja, números decimais. Por exemplo, número de faltas ao trabalho

por motivo de saúde (0, 1, 2,...) e número de peças defeituosas em um lote (0, 1, 2, 3,...) são classificadas como quantitativas discretas, enquanto que peso, altura, renda familiar (salários mínimos) são classificadas como quantitativas contínuas.

Podemos atribuir códigos numéricos às categorias de respostas de algumas variáveis qualitativas. Por exemplo, para a variável gênero, podemos associar o código 1 para o sexo feminino e 2 para o sexo masculino. Mas isto não a torna uma variável quantitativa, ou seja, não podemos, por exemplo, calcular uma média destas respostas, pois não conseguiríamos interpretar o resultado obtido.

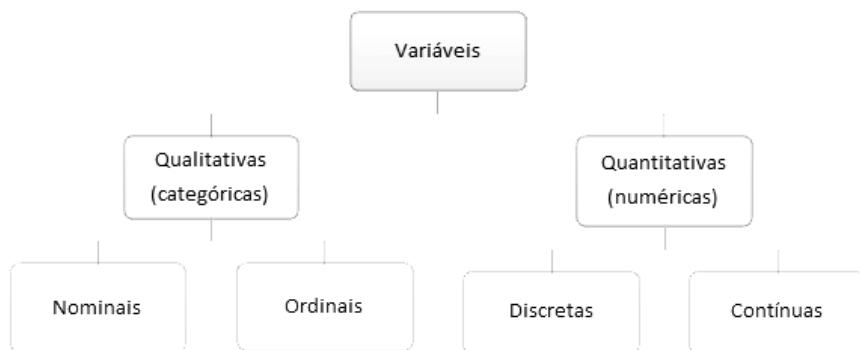


Figura 1.2 – Classificação das variáveis.



EXEMPLO

1.2: Vamos classificar as seguintes variáveis:

- Número de peças defeituosas produzidas em uma linha de montagem.
- Peso de pacientes.
- Fumante.
- Tipo sanguíneo.
- Grau de satisfação do consumidor com determinado produto.

Resolução

- Variável quantitativa discreta, pois as possíveis respostas são 0, 1, 2, 3, etc. (as respostas assumem somente valores inteiros).
- Variável quantitativa contínua, pois as possíveis respostas são 58,7; 89,8; etc. (as respostas podem assumir valores decimais).

- c) Variável qualitativa nominal, pois as possíveis respostas são sim ou não. (as possíveis respostas são categóricas).
 - d) Variável qualitativa nominal, pois as possíveis respostas são A, AB, B ou O (as possíveis respostas são categóricas).
 - e) Variável qualitativa ordinal, pois as possíveis respostas são nada satisfeito, pouco satisfeito, satisfeito, muito satisfeito (as possíveis respostas são categóricas e possuem uma ordenação natural, do menor grau de satisfação para o maior).
-

Outra maneira comum de classificar os dados é através do uso dos níveis de mensuração intervalar e de razão. No nível intervalar, as diferenças são significativas, mas não existe ponto inicial zero natural e as razões não têm sentido e, no nível de mensuração de razão, há um ponto inicial zero natural e as razões são significativas.

1.3 Coleta de dados

Já sabemos que para, fazer qualquer estudo estatístico, precisamos coletar dados. Esta coleta pode ser feita através de estudos observacionais ou experimentos.

Em estudos observacionais, não há qualquer tentativa de controlar ou modificar os elementos que farão parte do estudo. Por exemplo, uma pesquisa feita por institutos de pesquisa é um estudo observacional, pois os dados são geralmente coletados através de uma entrevista ou preenchimento de um questionário. Neste estudo, as respostas das pessoas são simplesmente coletadas e registradas, sem qualquer tipo de controle ou modificação.

Em um experimento, aplicamos algum tratamento e observamos o seu efeito sobre os elementos que estão participando do estudo. Por exemplo, uma indústria farmacêutica está interessada em testar uma nova medicação no tratamento de pessoas com colesterol alto. Um grupo de pacientes com altos níveis colesterol recebe o tratamento e passa a ser observado.

Sabemos, também, que um estudo estatístico pode ser feito com todos os elementos da população ou com uma parte desta população (amostra). Quando o estudo for feito com dados amostrais, deveremos ter muito cuidado

na maneira de coletar estes dados. De acordo com TRIOLA (2008, p. 17), “se os dados amostrais não forem coletados de maneira apropriada, eles podem ser de tal modo inúteis que nenhuma manipulação estatística poderá salvá-los”.

Para que possamos usar os resultados obtidos na amostra para fazer inferências sobre a população de interesse, precisamos garantir que a amostra seja representativa desta população. Por exemplo, no Exemplo 1.1, se os 620 funcionários forem selecionados somente em um dos setores da empresa, não podemos garantir que esta amostra seja representativa de todos os funcionários, pois parece pouco provável que os outros funcionários dos diversos setores tenham a mesma avaliação sobre o grau de satisfação com os benefícios oferecidos.

Veremos agora quais técnicas de amostragem podemos utilizar para garantir a representatividade da população.

1.3.1 Técnicas de amostragem

Temos dois tipos de amostragem, a que chamamos de probabilística (ou aleatória) e a não probabilística (ou não aleatória).

A amostragem será probabilística se todos os elementos da população tiverem probabilidade conhecida, e diferente de zero, de pertencer à amostra. Caso contrário, a amostragem será não probabilística.

Quando selecionamos os elementos que farão parte da amostra, podemos permitir que eles sejam selecionadas mais de uma vez. Neste caso, estamos trabalhando com amostragem com reposição. Na amostragem sem reposição, o elemento sorteado é removido da população. Se pensarmos na quantidade de informação que a amostra conterá, a amostragem sem reposição é mais adequada. Mas, amostragem com repetição implica independência entre os elementos selecionados. Isto facilita o desenvolvimento de propriedades de estimadores que são estudos em técnicas da inferência estatística.

Na prática podemos considerar a seleção dos elementos como independentes quando pequenas amostras são retiradas de grandes populações, pois é raro selecionar o mesmo elemento duas vezes.

Segundo TRIOLA (2008, p. 132), “Eis uma diretriz comum: se o tamanho da amostra não é maior que 5% do tamanho da população, tratamos a seleção das unidades experimentais como sendo independentes (mesmo que as seleções sejam feitas sem reposição, pois tecnicamente elas são dependentes)”.

Estudaremos agora algumas técnicas muito utilizadas de amostragem probabilística.

1.3.1.1 Técnicas de amostragem probabilística (ou aleatória)

“A grande vantagem das amostras probabilísticas é medir a precisão da amostra obtida, baseando-se no resultado contido na própria amostra” (BUSSAB; MORETTIN, 2002, p. 261).

Nas técnicas descritas a seguir, usaremos N para denotar o tamanho da população e n para indicar o tamanho da amostra.

Utilizaremos um mesmo exemplo para explicar as diferentes técnicas de amostragem, com o objetivo de evidenciar as características de cada uma delas.

1.3.1.1.1 Amostragem aleatória simples

Neste tipo de amostragem, a seleção dos elementos que farão parte da amostra é feita de maneira bem simples: quando estamos trabalhando com uma população finita, temos como obter uma listagem de todos os N elementos que compõem a população. Para fazer a seleção, escrevemos cada elemento da população em um cartão, colocamos em uma urna e sorteamos a quantidade de cartões de acordo com o tamanho da amostra. Neste procedimento, todo elemento da população tem a mesma probabilidade de pertencer à amostra. Quando a população for muito grande, o procedimento descrito torna-se inviável. Nestes casos, contamos com o auxílio do Excel, que gera números aleatórios através da função ALEATORIOENTRE. Para utilizarmos este tipo de amostragem, é desejável que a população seja homogênea, ou seja, que os elementos sejam similares sob o ponto de vista da variável em estudo. Caso a população seja heterogênea, há o risco de se obter uma amostra pouco representativa da população em estudo. Por exemplo, a população de funcionários de uma empresa difere quanto ao gênero, faixa de idade, grau de escolaridade e faixa salarial, e quando selecionarmos uma amostra aleatória de funcionários pode acontecer de não serem sorteados elementos com algumas destas características. E, os funcionários que se enquadram em cada uma destas características podem ter avaliações diferentes quanto à variável em estudo.



EXEMPLO

1.3: Uma universidade está elaborando uma pesquisa com objetivo de avaliar seu espaço físico, biblioteca, laboratórios, secretaria acadêmica, entre outros, visando aperfeiçoamento e fortalecimento das atividades de ensino. Para isto, deseja obter uma amostra de 8% dos seus 4 500 estudantes, para entrevistá-los. Qual deve ser o procedimento para a obtenção de uma amostra aleatória simples?

Resolução

Para obtermos uma amostra aleatória simples de 8% dos 4 500 estudantes, precisamos sortear 360. Como poderemos fazer este sorteio? Temos como obter o nome ou registro acadêmico de cada um dos alunos facilmente. Estas informações estão disponíveis na secretaria acadêmica da universidade. Os nomes ou registros acadêmicos são escritos em pedaços de papel. Após colocar, separadamente, as informações em 4 500 papéis, eles são colocados em uma urna. Misturamos bem e sorteamos um papel. Repetimos o procedimento até que 360 papéis sejam sorteados. Os nomes (ou registros acadêmicos) selecionados correspondem aos alunos que comporão a amostra. A descrição do sorteio foi feita desta maneira para facilitar o entendimento deste tipo de amostragem. Nos dias atuais, colocamos todos os nomes em uma planilha do Excel e utilizamos a função ALEATÓRIOENTRE.

1.3.1.1.2 Amostragem estratificada

Utilizamos esta técnica quando identificamos que a população é heterogênea para a variável de interesse no estudo. Neste caso, dividimos a população em grupos mais homogêneos (subgrupos), que são os estratos. Após a identificação dos estratos, selecionamos os elementos que farão parte da amostra através de uma amostragem aleatória simples de cada estrato ou através de uma seleção proporcional ao número de elementos existentes em cada estrato. Voltando ao exemplo da seleção de uma amostra de funcionários de um hospital e maternidade, podemos dividir a população de funcionários nos seguintes estratos: gênero, faixa de idade, grau de escolaridade e faixa salarial. Dentro de cada estrato, os elementos são similares.



EXEMPLO

1.4: Uma universidade está elaborando uma pesquisa com objetivo de avaliar seu espaço físico, biblioteca, laboratórios, secretaria acadêmica, entre outros, visando aperfeiçoamento e fortalecimento das atividades de ensino. Para isto, deseja obter uma amostra de 8% dos seus 4 500 estudantes, para entrevistá-los. Há uma suspeita de que mulheres são mais criteriosas na avaliação institucional. De acordo com informações acadêmicas, aproximadamente 60% dos estudantes são do sexo feminino. Qual deve ser o procedimento para a obtenção de uma amostra estratificada?

Resolução

De acordo com as informações, vamos separar os estudantes em dois estratos: estudantes do sexo masculino e estudantes do sexo feminino. Depois, obtemos uma amostra aleatória simples de cada estrato (gênero) e reunimos os dados selecionados dos dois estratos em uma só amostra estratificada.

Como sabemos que 60% dos estudantes são do sexo feminino e, consequentemente, 40% do sexo masculino, podemos fazer uma seleção proporcional ao número de estudantes em cada estrato. Neste caso, selecionaríamos aleatoriamente 216 estudantes no estrato do sexo feminino ($360 \times 0,6$) e 144 estudantes no estrato do sexo masculino ($360 \times 0,4$).

1.3.1.1.3 Amostragem sistemática

A seleção dos elementos, quando utilizamos a amostragem sistemática, é feita segundo um sistema preestabelecido (sistematicamente). Para estabelecermos o sistema de seleção, ordenamos os elementos da população (formando uma lista) de forma a identificá-los pela posição e, após o número inicial ser selecionado aleatoriamente, os elementos que farão parte da amostra serão selecionados segundo intervalos regulares que ocorrem a partir do número inicial. Precisamos tomar cuidado ao estabelecer o sistema de seleção dos elementos, pois tendências podem surgir se houver algum tipo de sequência periódica ou cíclica nos elementos da população que foram ordenados.



EXEMPLO

1.5: Uma universidade está elaborando uma pesquisa com o objetivo de avaliar seu espaço físico, biblioteca, laboratórios, secretaria acadêmica, entre outros, visando ao aperfeiçoamento e fortalecimento das atividades de ensino. Para isto, deseja obter uma amostra de 8% dos seus 4 500 estudantes, para entrevistá-los. Qual deve ser o procedimento para a obtenção de uma amostra sistemática?

Resolução

Na amostragem sistemática, precisamos de uma lista dos elementos que compõem a população. Após conseguir uma listagem com todos os estudantes, precisamos encontrarmos a fração amostral $k = \frac{N}{n}$. No nosso exemplo, $k = \frac{4\ 500}{360} = 12,5$. Como k não é um número inteiro, devemos arredondar para o inteiro mais próximo, ou seja, vamos considerar k = 13.

O próximo passo é escolher aleatoriamente um número entre 1 e 13 (por meio de um sorteio). Por exemplo, vamos supor que o número sorteado seja 4. Então, o primeiro estudante selecionado será o que está na quarta posição da listagem. Depois, a partir do número 4, contamos 13 e selecionamos o próximo estudante, e assim por diante, até completar a amostra de 360 estudantes.

1.3.1.1.4 Amostragem por conglomerado

Neste tipo de amostragem, dividimos a população em subgrupos (conglomerados) de elementos heterogêneos, em seguida selecionamos aleatoriamente alguns conglomerados e escolhemos todos os elementos desses conglomerados selecionados para compor a amostra.

A diferença entre a amostragem estratificada e por conglomerado é que na amostragem estratificada os elementos dentro de cada subgrupo são homogêneos e, no caso dos conglomerados, os elementos dentro de cada subgrupo são heterogêneos. A amostragem estratificada usa uma amostra de elementos de todos os estratos, enquanto que a amostragem por conglomerado usa todos os elementos dos conglomerados selecionados.



EXEMPLO

1.6: Uma universidade está elaborando uma pesquisa com objetivo de avaliar seu espaço físico, biblioteca, laboratórios, secretaria acadêmica, entre outros, visando aperfeiçoamento e fortalecimento das atividades de ensino. Para isto, deseja obter uma amostra de 8% dos seus 4 500 estudantes, para entrevistá-los. Qual deve ser o procedimento para a obtenção de uma amostra por conglomerados?

Resolução

Nesta situação, podemos formar conglomerados com os alunos matriculados em cada um dos cursos da universidade. Por exemplo, conglomerado formado com todos os estudantes matriculados no curso de Administração, ou conglomerado formado com todos os estudantes matriculados no curso de Enfermagem e assim por diante. Após a identificação dos conglomerados, sorteamos alguns deles e entrevistamos todos os estudantes dentro de cada conglomerado sorteado.

Os estudantes dentro de cada conglomerado são heterogêneos, ou seja, há diversidades de informações quanto a idade, estado civil, renda, gênero, etc.

1.3.1.2 Técnicas de amostragem não probabilística (ou não aleatória)

De acordo com BRUNI (2010, p. 173)

A amostragem não probabilística consiste em uma amostragem subjetiva, em que a variabilidade dos resultados da amostra não pode ser obtida com precisão, ao contrário da amostragem probabilística. Impede a mensuração do erro da inferência – que é indesejado -, porém, resulta em custos ainda mais baixos em uma coleta de dados ainda mais rápida.

1.3.1.2.1 Amostragem por conveniência

Na amostragem por conveniência, os elementos amostrais são escolhidos por serem mais acessíveis, gerando informações de forma rápida e barata. Por exemplo, no nosso exemplo da avaliação da universidade, um professor de

Cálculo pode escolher todos os alunos que cursam sua disciplina, nos diversos cursos em que leciona, para compor a amostra que será utilizada na pesquisa. Neste tipo de seleção, o professor restringe a escolha dos alunos que farão parte da amostra, pois estudantes que não têm aula com ele estarão excluídos de participar da amostra.

1.3.1.2.2 Amostragem por quota

Neste tipo de amostragem, os elementos que fazem parte da amostra são retirados da população segundo quotas estabelecidas de acordo com a distribuição desses elementos na população. A descrição deste tipo de amostragem nos faz lembrar da amostragem estratificada. A diferença é que, aqui, os elementos são selecionados por julgamento, e não de maneira aleatória, e depois confirmamos as características dos elementos amostrados. Por ser relativamente barato, este tipo de amostragem é muito utilizado em levantamentos de opinião e pesquisa de mercado.



CONEXÃO

Uma leitura interessante sobre a amostragem não probabilística é encontrada no seguinte trabalho: Amostragem não Probabilística: Adequação de Situações para uso e Limitações de amostras por Conveniência, Julgamento e Quotas. Disponível em: < http://www.fecap.br/adm_online/art23/tania2.htm>. Acesso em: 30 de Abr. 2015.

Após a definição de qual tipo de amostragem será utilizada em uma pesquisa, a pergunta que naturalmente surge é: Qual o tamanho da amostra que devemos utilizar? Há fórmulas estatísticas bem conhecidas para determinação do tamanho amostral, mas a aplicação dessas fórmulas exige conhecimentos que não abordaremos neste livro.

De acordo com VIEIRA (2008, p. 13)

Mais importante é saber que não basta ter em mãos uma fórmula, ou um programa de computador para estimar o tamanho da amostra. É preciso algum conhecimento prévio (estimativas preliminares de um ou mais parâmetros, obtidas de amostras pilotos ou da literatura) e uma boa dose de bom senso.



CONEXÃO

Para uma leitura introdutória sobre cálculo do tamanho de amostras, o artigo: Análise a respeito do tamanho de amostras aleatórias simples: uma aplicação na área de Ciência da Informação aborda diferentes procedimentos estatísticos para a determinação do tamanho de uma amostra aleatória simples. Disponível em: <http://dgz.org.br/ago05/Art_01.htm>. Acesso em: 30 de Abr. 2015.

Mesmo planejando e executando bem o processo de coleta da amostra, provavelmente haverá algum erro nos resultados. Por exemplo, voltando ao Exemplo 1.1, se selecionarmos uma amostra com outros 620 funcionários provavelmente encontraremos uma estimativa diferente para a proporção de funcionários satisfeitos com os benefícios oferecidos. Ou, ainda, poderíamos tirar uma amostra que forneça um resultado muito diferente daquele que seria obtido se trabalhássemos com a toda a população. Então, de acordo com o raciocínio exposto, podemos definir dois tipos de erros:

ERRO AMOSTRAL	É a diferença entre o resultado amostral e o verdadeiro resultado da população; tais erros resultam das flutuações amostrais devidas ao acaso.
ERRO NÃO AMOSTRAL	Ocorre quando os dados amostrais são coletados, registrados ou analisados incorretamente (tal como a seleção de uma amostra tendenciosa, o registro incorreto dos dados ou o uso de um instrumento de medida defeituoso).

Se os dados amostrais são coletados por meio de um processo probabilístico, esperamos que eles sejam representativos da população e, assim, podemos analisar o erro amostral, mas devemos ter o cuidado de minimizar o erro não amostral.

Agora que já sabemos que os dados são obtidos por meio de elementos provenientes de uma população ou de uma amostra e que, caso sejam dados amostrais, devemos tomar o cuidado de selecionar elementos que sejam os mais parecidos possíveis com a população do qual foram extraídos, vamos aprender a organizar os dados.

Após a obtenção dos dados, por exemplo, através de experimentos, cadastros, entrevistas ou preenchimento de questionários, obtemos o conjunto de dados brutos, ou seja, dados que ainda não foram organizados. Neste momento começamos com a apuração, isto é, organização dos dados brutos. Isto é feito por meio da construção da distribuição de frequências, que estudaremos a seguir.

1.4 Distribuição de frequências

Em um estudo estatístico, temos como maior interesse conhecer o comportamento da(s) variável(eis) presentes no estudo. Isto se torna fácil quando organizamos as respostas da variável em uma distribuição de frequências. Mas, o que é uma distribuição de frequências?

Distribuição de frequências é uma tabela em que se resumem grandes quantidades de dados, determinando o número de vezes, que cada dado ocorre (frequência) e a porcentagem com que aparece (frequência relativa).

O processo de contagem do número de vezes, que cada dado ocorre fica facilitado se ordenarmos os dados. A uma sequência ordenada (crescente ou decrescente) de dados brutos damos o nome de Rol.

Vamos formalizar os conceitos das frequências que utilizaremos na construção da distribuição de frequências:

Frequência absoluta ou simplesmente frequência (f): é o nº de vezes, que cada dado aparece na pesquisa.

Frequência relativa ou percentual (fr): é o quociente da frequência absoluta pelo número total de dados. Esta frequência pode ser expressa em porcentagem. O valor de ($fr \times 100$) é definido como fr (%).

Veremos mais adiante que, em algumas análises, precisaremos das informações das frequências acumuladas:

Frequência acumulada (fa): é a soma de cada frequência com as que lhe são anteriores na distribuição.

Frequência relativa acumulada (fra): é o quociente da frequência acumulada pelo número total de dados. Esta frequência também pode ser expressa em porcentagem. O valor de (fra x100) é definido como fra (%).

A seguir apresentamos a estrutura de uma distribuição de frequências.

NOME DA VARIÁVEL	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)
Respostas da variável		
Total	número total de elementos em estudo	100,00

Segundo VIEIRA (2003, p. 47)

1. As tabelas devem ser delimitadas, no alto e embaixo, por traços horizontais. Esses traços podem ser mais fortes do que os traços feitos no interior da tabela; as tabelas não devem ser delimitadas, à direita e à esquerda, por traços verticais;
2. O cabeçalho deve ser delimitado por traços horizontais;
3. Podem ser feitos traços verticais no interior da tabela, separando as colunas;
4. As tabelas devem ter significado próprio, isto é, devem ser entendidas mesmo quando não se lê o texto em que estão apresentadas;
5. As tabelas devem ser numeradas com algarismos arábicos. Pode ser adotada a numeração progressiva por seções.
6. Quando dois ou mais tipos de informação tiverem sido agrupados em um só conjunto, esse conjunto entra na tabela sob a denominação “outros”.



EXEMPLO

1.7: Um questionário foi aplicado aos dez candidatos a uma vaga no departamento financeiro de uma loja de departamentos e alguns dos resultados obtidos estão apresentados no quadro a seguir. Vamos organizar os dados das variáveis grau de escolaridade e idade em distribuição de frequências.

CANDIDATO DA VAGA	GRAU DE ESCOLARIDADE	IDADE	TEMPO DE EXPERIÊNCIA NA ÁREA
1	Ensino Médio	30	7
2	Ensino Superior	35	12
3	Ensino Superior	26	4
4	Ensino Médio	22	1
5	Ensino Médio	28	8
6	Pós Graduação	30	10
7	Ensino Médio	26	3
8	Ensino Superior	33	8
9	Pós Graduação	35	6
10	Ensino Médio	23	2

As variáveis, ou seja, as características de interesse nos candidatos são: grau de escolaridade, idade e tempo de experiência na área da vaga. Candidato não é variável! Esta coluna simplesmente informa que são 10 candidatos, com suas respectivas características. Os números poderiam ser substituídos pelos nomes dos candidatos.

Resolução

A distribuição de frequências contém 3 colunas: a variável em estudo, a frequência e a frequência relativa (%). Toda tabela deve conter um título que explique o conteúdo da tabela.

Também podemos utilizar como cabeçalho para a segunda coluna a palavra Frequência.

GRAU DE ESCOLARIDADE	NÚMERO DE CANDIDATOS	FREQUÊNCIA RELATIVA (%)
Ensino Médio	5	50
Ensino Superior	3	30
Pós Graduação	2	20
Total	10	100

Tabela 1.1 – Distribuição dos candidatos, segundo grau de escolaridade

O número de candidatos é 5 para o grau de escolaridade Ensino Médio, pois, analisando o Quadro 1 verificamos que 5 candidatos possuem esta escolaridade (candidatos 1, 4, 5, 7 e 10). A frequência relativa (%) para este grau de escolaridade é obtida fazendo $\frac{5}{10} \times 100 = 50\%$.

O mesmo procedimento é feito para encontrar os valores referentes ao grau de escolaridade Ensino Superior e Pós Graduação.

Analisando as informações, observamos que, dos 10 candidatos à vaga, 50% deles possuem Ensino Médio, seguidos por 30% com Ensino Superior e 20% com Pós-Graduação.

IDADE	NÚMERO DE CANDIDATOS	FREQUÊNCIA RELATIVA (%)
22	1	10
23	1	10
26	2	20
28	1	10
30	2	20
33	1	10
35	2	20
Total	10	100

Tabela 1.2 – Distribuição dos candidatos, segundo a idade.

Pelo Quadro 1, verificamos que há repetição das idades 26 (candidatos 3 e 7), 30 (candidatos 1 e 6) e 35 (candidatos 2 e 9).

Por meio das informações contidas na Tabela 1.2, observamos que a idade mínima dos candidatos é 22 anos e a máxima é 35 anos. Podemos concluir, também, que 70% dos candidatos têm no máximo 30 anos (30 anos de idade ou menos).

Podemos observar que a estrutura da distribuição de frequências é a mesma tanto para variáveis qualitativas quanto para variáveis quantitativas. No caso de variáveis quantitativas, colocamos os valores numéricos em ordem crescente.

Como dissemos anteriormente, em algumas análises precisamos da frequência acumulada, como na construção de um gráfico denominado Ogiva e no cálculo de medidas separatriizes para dados organizados em intervalos de classes. Construiremos, agora, uma distribuição de frequências com a frequência acumulada absoluta e a frequência acumulada relativa (%).

IDADE	NÚMERO DE CANDIDATOS	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA ACUMULADA	FREQUÊNCIA ACUMULADA RELATIVA (%)
22	1	10	1	10
23	1	10	2	20
26	2	20	4	40
28	1	10	5	50
30	2	20	7	70

IDADE	NÚMERO DE CANDIDATOS	FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA ACUMULADA	FREQUÊNCIA ACUMULADA RELATIVA (%)
33	1	10	8	90
35	2	20	10	100
Total	10	100		

A coluna da frequência acumulada é obtida somando cada frequência com as que lhe são anteriores e a frequência acumulada relativa (%) é obtida dividindo a frequência acumulada pelo número total de dados ($\times 100$).

Por exemplo, a frequência acumulada associada à idade 30 é obtida somando a frequência desta resposta com as frequências anteriores ($1 + 1 + 2 + 1 + 2 = 7$) e a frequência acumulada relativa (%) é obtida fazendo $\frac{7}{10} \times 100 = 70\%$.

Quando estamos analisando uma variável quantitativa contínua, é comum os valores não se repetirem. Se construirmos uma distribuição de frequências como na Tabela 1.2, ela ficará muito extensa e não atingiremos o objetivo de resumir o conjunto de dados. Nestes casos, é conveniente agrupar os dados em intervalos de classes. O mesmo procedimento pode ser feito quando a variável for quantitativa discreta e apresentar um número grande de dados, mas com valores com pouca repetição.

Identificamos os seguintes valores em um intervalo de classe:

LIMITE INFERIOR (LI)	é o menor valor que a variável pode assumir em uma classe de frequência;
LIMITE SUPERIOR (LS)	serve de limite para estabelecer qual o maior valor que a variável pode assumir em uma classe de frequência, mas, geralmente, os valores iguais ao limite superior não são computados naquela classe e sim na seguinte;
PONTO MÉDIO (PM)	é a média aritmética entre o L_i e o L_s da mesma classe, ou seja $P_m = \frac{L_i + L_s}{2}$
AMPLITUDE (H)	é a diferença entre o L_s e o L_i da classe, ou seja, $h = L_s - L_i$;

Na construção de uma distribuição de frequências com intervalos de classes devemos ter respostas para estes dois questionamentos:

- Qual o número de classes que a tabela deve ter?
- Qual o tamanho (ou a amplitude) das classes?

Podemos usar o bom senso e escolher arbitrariamente quantas classes e qual a amplitude que estas classes devem ter.

Quando não tivermos nenhuma referência sobre qual deve ser o número de classes a se trabalhar, podemos utilizar o critério que é sugerido por vários autores. Chama-se regra da raiz:

$$k \approx \sqrt{n}$$

onde k indica o número de classes que vamos construir e n é o número de observações do conjunto de dados. É muito comum o valor obtido para k não ser inteiro, então, vamos aproximar para o inteiro próximo de k .

Para encontrar a amplitude e o número de observações em cada classe, vamos seguir os seguintes passos:

- Achar o mínimo e o máximo dos dados.
- Para determinar a amplitude de cada classe calcularemos $h \approx \frac{R}{k}$, onde $R = \text{valor máximo} - \text{valor mínimo}$. O valor de h será a amplitude da classe. Normalmente o valor encontrado para h não é inteiro. Nestes casos, podemos aproximar para o inteiro próximo para facilitar a construção das classes.
- Contar o número de observações que pertencem a cada intervalo de classe. Esses números são as frequências absolutas das classes.
- Calcular as frequências relativas de cada classe.

De modo geral, a quantidade de classes não deve ser inferior a 5 e nem superior a 20. Se o número de classes for muito pequeno, perderemos informação, e com um número grande de classes, o objetivo de resumir os dados fica prejudicado.

Construiremos intervalos de classe fechados à esquerda. A representação deste tipo de intervalo é:

$$L_i | - L_s$$

Por exemplo:

$$5 | - 10$$

Pertencem a este intervalo valores iguais ou superiores ao limite inferior do intervalo (neste exemplo, 5) e inferiores ao limite superior (neste exemplo, 10). Se houver o número 10 no conjunto de dados, ele entra no próximo intervalo de classe.

1.8: Os dados abaixo referem-se à fração de colesterol de muito baixa densidade, em miligramas por decilitro (mg/dl), em indivíduos do sexo feminino. Vamos organizar este conjunto de dados numa distribuição de frequências.

22	22	24	24	25	26	26	26
26	26	26	26	27	27	27	28
28	28	28	28	28	28	28	28
28	29	29	29	29	29	29	30
30	30	30	30	30	30	30	30
30	30	32	34	34	34	34	34
35	35	35	35	35	35	35	36
36	37	39	39	40	40	45	48

Resolução:

Apesar da variável em estudo (fração de colesterol de muito baixa densidade) estar apresentada na forma discreta, há uma variação grande de números. Se construirmos uma distribuição de frequências colocando os números do menor para o maior, a tabela ficará extensa. Então, nesta situação, é conveniente agrupar os dados em intervalos de classes.

Primeiro, precisamos saber quantas classes vamos construir. Para isto, utilizaremos a fórmula:

$$\sqrt{64} = 8$$

Então, construiremos 8 classes. Agora, vamos encontrar o tamanho (amplitude) de cada uma das classes:

$$h \cong \frac{\text{valor\! m\'aximo} - \text{valor\! m\'inimo}}{8} \cong \frac{48 - 22}{8} \cong 3,3$$

Portanto, vamos construir classes de amplitude 4 cada uma (arredondamos o valor de h para facilitar a construção das classes). Quando consideramos 4 como amplitude, o número de classes passa a ser 7 (esta alteração não gera problema algum!).

FRAÇÃO DE COLESTEROL DE BAIXA DENSIDADE	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)
22 -26	5	7,81
26 -30	26	40,63
30 -34	12	18,75
34 -38	15	23,44
38 -42	4	6,25
42 -46	1	1,56
46 -50	1	1,56
Total	64	100,00

Tabela 1.4 – Distribuição de frequências do número de horas extras dos funcionários.

Neste exemplo construímos classes de mesma amplitude, mas isto não é obrigatório. Quando construímos classes de amplitudes diferentes, devemos tomar cuidado na construção de um gráfico denominado histograma, que veremos mais adiante.

Agora que já aprendemos como apresentar os dados coletados em distribuições de frequências, vamos estudar como estes mesmos dados são utilizados na construção de gráficos.

1.5 Gráficos

Os gráficos estatísticos são utilizados frequentemente nos meios de comunicação. Em geral, as pessoas tem mais facilidade de compreender as informações que estão contidas numa tabela por meio de gráficos. Há uma quantidade muito grande de gráficos disponíveis em softwares estatísticos e no Excel, mas devemos ter em mente que a construção de gráficos deve ser feita cuidadosamente! Por exemplo, a construção de um gráfico desproporcional em suas medidas pode nos levar a conclusões equivocadas.



CONEXÃO

Um texto interessante que chama à reflexão sobre a necessidade de abordagens pedagógicas mais efetivas para o ensino e a aprendizagem de gráficos está disponível em: <http://www.ufrj.br/emanped/paginas/conteudo_producoes/docs_22/carlos.pdf>. Acesso em: 30 de Abr. 2015.

1.5.1 Tipos de gráficos

Os gráficos mais utilizados são: gráfico de linhas, diagramas de área (como por exemplo: gráfico de barras e gráfico de setores) e gráficos para representar as distribuições de frequências construídas com intervalos de classes (como por exemplo: polígono de frequências e histograma).

De acordo com VIEIRA (2013, p. 17):

Cada tipo de gráfico tem indicação específica, mas, de acordo com as normas brasileiras:

- Todo gráfico deve apresentar título e escala;
- O título deve ser colocado abaixo da ilustração.
- As escalas devem crescer da esquerda para a direita e de baixo para cima.
- As legendas explicativas devem ser colocadas, de preferência, à direita da figura.
- Os gráficos devem ser numerados, na ordem em que são citados no texto.
- Os dois eixos devem apresentar legenda.

Nos itens a seguir abordaremos os gráficos de linhas, barras, setores, histograma, polígono de frequências, Pareto e dispersão. Sempre que possível utilizaremos as distribuições de frequências que construímos nos exemplos anteriores, para mostrar que as informações contidas em ambos são as mesmas.

1.5.1.1 Gráfico de linhas

O gráfico de linhas (gráfico de séries temporais) é utilizado quando os dados estiverem distribuídos segundo uma variável no tempo (meses, anos, etc.). Esse tipo de gráfico retrata as mudanças nas quantidades com respeito ao tempo através de uma série de segmentos de reta. É muito eficiente para mostrar possíveis tendências no conjunto de dados.



EXEMPLO

1.9: A Tabela 1.5 fornece o número de casos de dengue no Brasil, no período 2000 – 2013.

ANO	NÚMERO DE CASOS
2000	135.228
2001	385.783
2002	696.472
2003	274.975
2004	70.174
2005	147.039
2006	258.680
2007	496.923
2008	632.680
2009	406.269
2010	1.011.548
2011	764.032
2012	589.591
2013	1.452.489

Tabela 1.5 – Número de casos de dengue - Brasil. Fonte: Disponível em : < <http://portalsaude.saude.gov.br/images/pdf/2014/julho/31/Dengue-classica-at---2013.pdf> >. Acesso em: 17 jun. 2015.

O gráfico que melhor representa este conjunto de dados é o gráfico em linhas, já que os dados se reportam a uma série no tempo (série temporal). O gráfico está ilustrado na Figura 1.3.

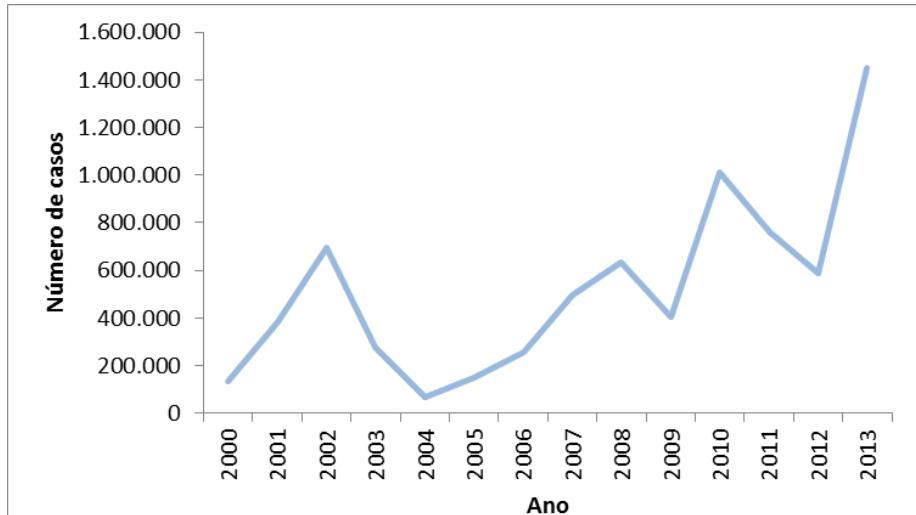


Figura 1.3 – Gráfico de linha para o número de casos de dengue no Brasil.

Analizando a Figura 1.3 observamos uma oscilação no número de casos de dengue, no Brasil, no período em estudo. O número de casos em 2013, comparado à 2012, aumentou, aproximadamente 146%!

1.5.1.2 Gráfico de barras

O gráfico de barras é bastante utilizado quando a variável em estudo for qualitativa (dados categóricos). No eixo horizontal especificamos os nomes das categorias e no eixo vertical construímos uma escala com a frequência ou a frequência relativa. As barras terão bases de mesma largura e alturas iguais à frequência ou à frequência relativa.

As barras podem estar na posição horizontal ou vertical. O Excel denomina um gráfico de barras na posição vertical como gráfico de colunas.



EXEMPLO

1.10: A Tabela 1.6 apresenta a distribuição, por tipo sanguíneo, de 120 recém-nascidos em uma maternidade. Vamos apresentar as informações por meio de um gráfico de barras (na posição horizontal e vertical).

TIPO SANGUÍNEO	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)
A	33	27,50
AB	5	4,17
B	21	17,50
O	61	50,83
Total	120	100,00

Tabela 1.6: Distribuição, por tipo sanguíneo, de recém-nascidos em uma maternidade.

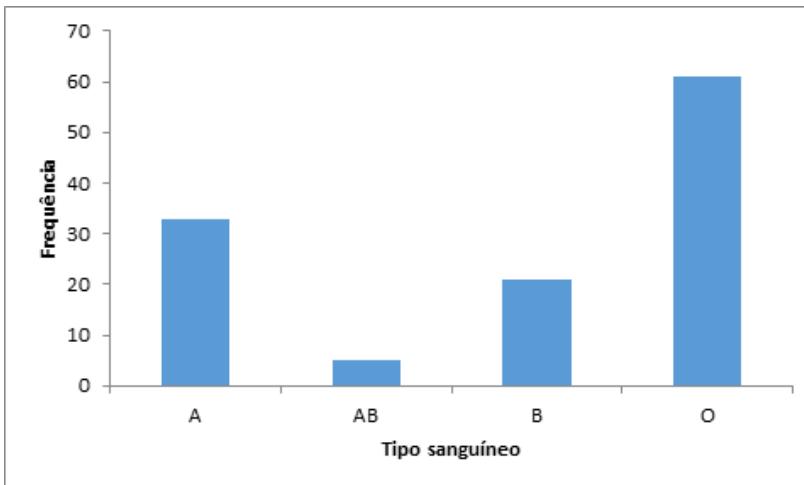


Figura 1.4 – Gráfico de barras para a variável tipo sanguíneo dos recém-nascidos.

A Figura 1.5 apresenta as barras na posição horizontal, e elas são construídas com base na frequência relativa (%). Nesta situação, as categorias são apresentadas no eixo vertical e no eixo horizontal construímos a escala, utilizando a frequência absoluta ou a frequência relativa (geralmente em porcentagem). Há a opção de colocarmos tais frequências acima das barras.

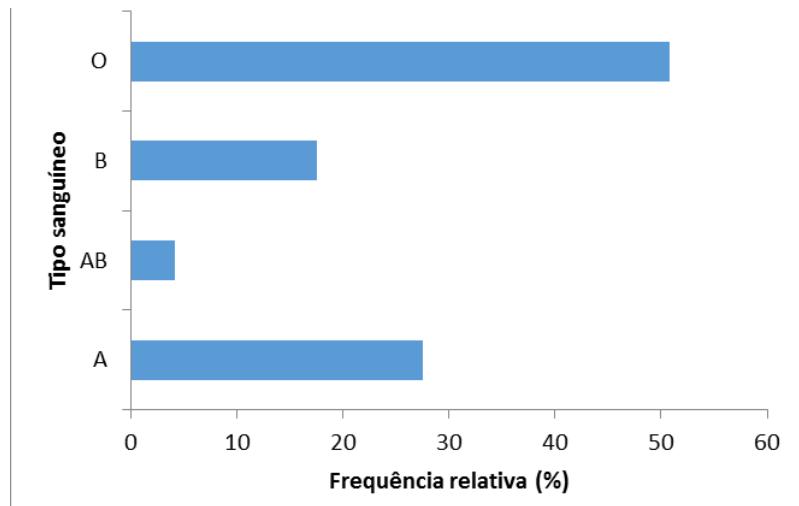


Figura 1.5 – Gráfico de barras para a variável tipo sanguíneo dos recém-nascidos.

Vale a pena ressaltar que as informações contidas nestes dois gráficos são as mesmas que estão apresentadas na Tabela 1.6.

1.5.1.3 Gráfico de setores

O gráfico de setores, também conhecido como gráfico de pizza, é um dos gráficos mais utilizados para representar variáveis qualitativas nominais (desde que o número de categorias seja pequeno) e é bastante apropriado quando se deseja visualizar a proporção que cada categoria representa do total.



EXEMPLO

1.11: Em uma universidade há 4 500 estudantes, dos quais 60% são do sexo feminino e 40% do sexo masculino. Os dados estão apresentados na Tabela 1.7.

GÊNERO	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)
Feminino	2 700	60,00
Masculino	1 800	40,00
Total	4 500	100,00

Tabela 1.7 – Distribuição dos alunos, segundo o gênero.

Vamos apresentar as informações em um gráfico de setores.

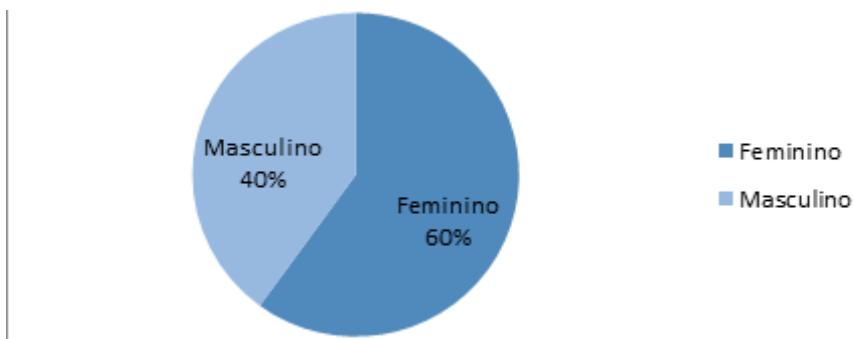


Figura 1.6 –Gráfico de setores para a variável gênero dos estudantes.

Os gráficos que serão apresentados a seguir são gráficos construídos segundo uma distribuição de frequências com intervalos de classes. São eles: o histograma e o polígono de frequências.

1.5.1.4 Histograma

Um histograma é semelhante ao diagrama de barras, porém refere-se a uma distribuição de frequências construída com intervalos de classes. Por isso, apresenta uma diferença: não há espaços entre as barras. Os intervalos de classes são colocados no eixo horizontal enquanto as frequências são colocadas no eixo vertical. As frequências podem ser absolutas ou relativas.



EXEMPLO

1.12: Vamos construir um histograma para os dados da Tabela 1.4.

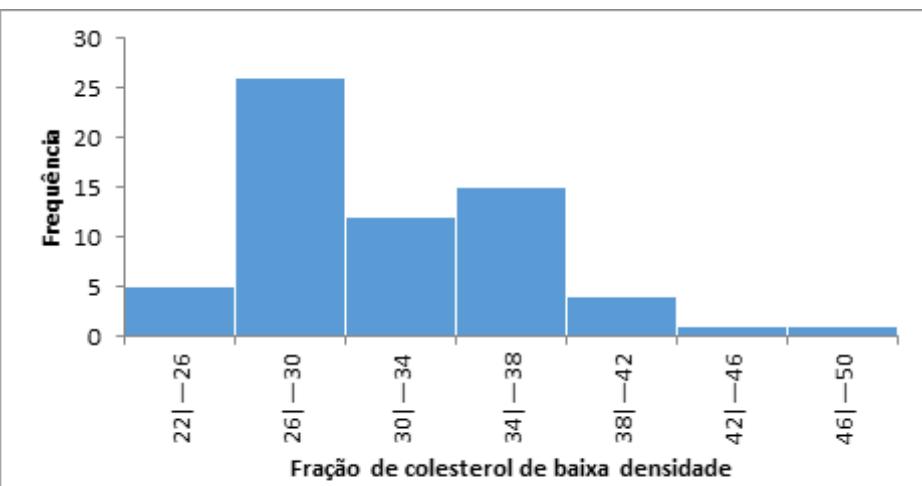


Figura 1.7 – Histograma para a fração de colesterol de baixa densidade.

O histograma é muito utilizado para visualizarmos a natureza da distribuição dos dados. Estudaremos as formas de distribuições (simétricas ou assimétricas) no próximo capítulo.

Utilizamos a frequência ou a frequência relativa para construir o histograma, desde que os intervalos de classes tenham mesma amplitude. Caso contrário, temos que encontrar a densidade de frequência, que é obtida pelo quociente da frequência absoluta pela amplitude do intervalo de classe.

1.5.1.5 Polígono de frequências

Podemos dizer que o polígono de frequências é um gráfico de linha de uma distribuição de frequências. No eixo horizontal são colocados os pontos médios de cada intervalo de classe e, no eixo vertical, são colocadas as frequências absolutas ou relativas (como no histograma). Para se obter as intersecções do polígono com o eixo das abscissas, devemos encontrar o ponto médio da classe anterior à primeira e o ponto médio da classe posterior à ultima.

O histograma e o polígono de frequências são gráficos alternativos e contêm a mesma informação. Fica a critério de quem está conduzindo o estudo a escolha de qual deles utilizar.

Considerando os dados da Tabela 1.4, temos o polígono de frequências representado pela Figura 1.8.

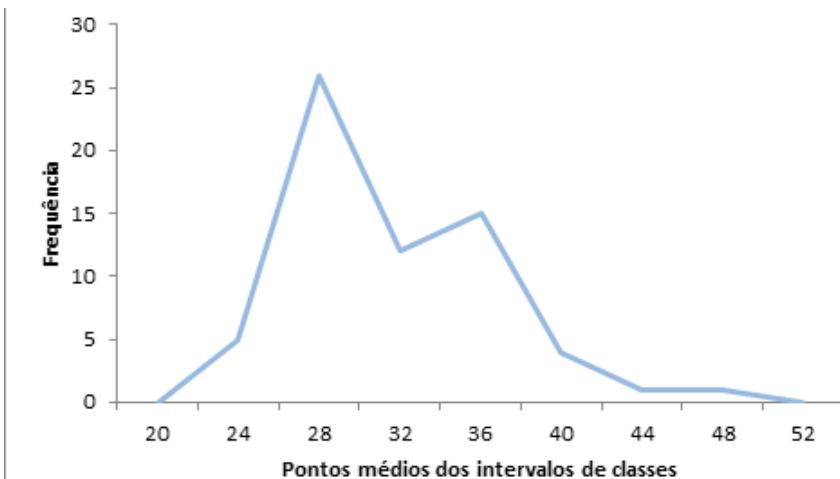


Figura 1.8 – Polígono de frequências para a fração de colesterol de baixa densidade.

1.5.1.6 Diagrama de Pareto

O Diagrama de Pareto é um gráfico de barras que é utilizado para representar as ocorrências das categorias de uma variável qualitativa. Neste tipo de gráfico, as barras são arranjadas em ordem decrescente de altura, a partir da esquerda para a direita, com a categoria que ocorre com maior frequência aparecendo em primeiro lugar.

A grande utilidade deste diagrama é a de permitir uma fácil visualização e identificação das causas ou problemas mais importantes, possibilitando a concentração de esforços sobre os mesmos. O diagrama de Pareto é uma das sete ferramentas da qualidade.



EXEMPLO

1.13: A distribuição de frequências a seguir apresenta as reclamações fundamentadas de 2013, por área, na Fundação Procon – SP.

ÁREA	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)
Produtos	9 683	31,15
Assuntos Financeiros	8 464	27,23
Serviços Essenciais	5 298	17,04
Serviços Privados	4 838	15,56
Saúde	1 408	4,53
Habitação	1 327	4,27
Alimentos	67	0,22
Total	31 085	100,00

Tabela 1.8 – Reclamações no Procon – SP por área, em 2013. Fonte: Disponível em: < http://www.procon.sp.gov.br/pdf/ranking_2013_coment.pdf >. Acesso em: 03 de Maio. 2015.

Vamos representar as informações contidas na Tabela 1.8 em um diagrama de Pareto.

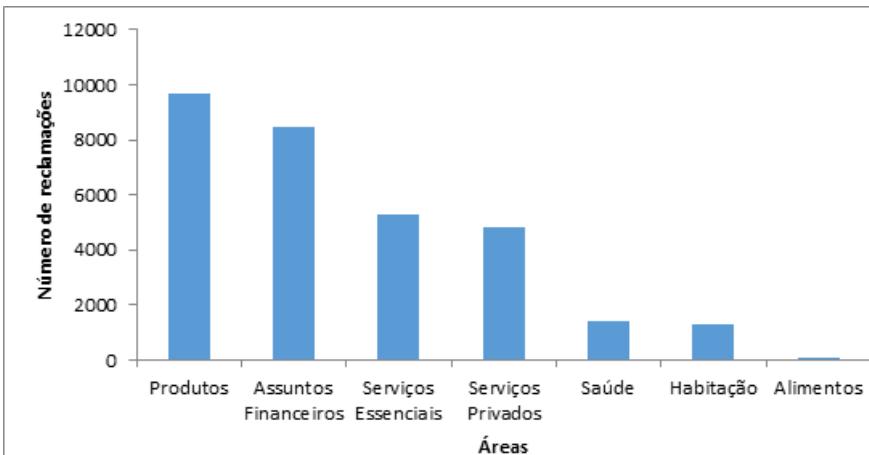


Figura 1.9 – Reclamações no Procon – SP, por área.

Analisando o gráfico, observamos que, em 2013, o maior número de reclamações fundamentadas foi na área de produtos, seguido por assuntos financeiros.

1.5.1.7 Diagrama de dispersão

O diagrama de dispersão é um gráfico muito utilizado quando temos interesse em visualizar a relação entre duas variáveis quantitativas, denominadas X e Y. Para construí-lo, cada par ordenado é colocado em suas determinadas coordenadas (x,y).



EXEMPLO

1.14: Uma construtora quer verificar a eficácia de seus anúncios em determinado programa de televisão. O objetivo é verificar se há relação entre a quantidade de anúncios e o número de apartamentos vendidos. A tabela abaixo mostra o número de anúncios que foram ao ar, durante seis meses, e o correspondente número de apartamentos vendidos de um edifício em lançamento.

NÚMERO DE ANÚNCIOS (X)	NÚMERO DE APARTAMENTOS VENDIDOS (Y)
10	4
15	7
18	6
22	12
25	15
30	19

Tabela 1.9 – Número de anúncios publicados e número de apartamentos vendidos.

Para verificarmos, visualmente, se há relação entre o número de anúncios que foram ao ar e o número de apartamentos vendidos, construímos o diagrama de dispersão.

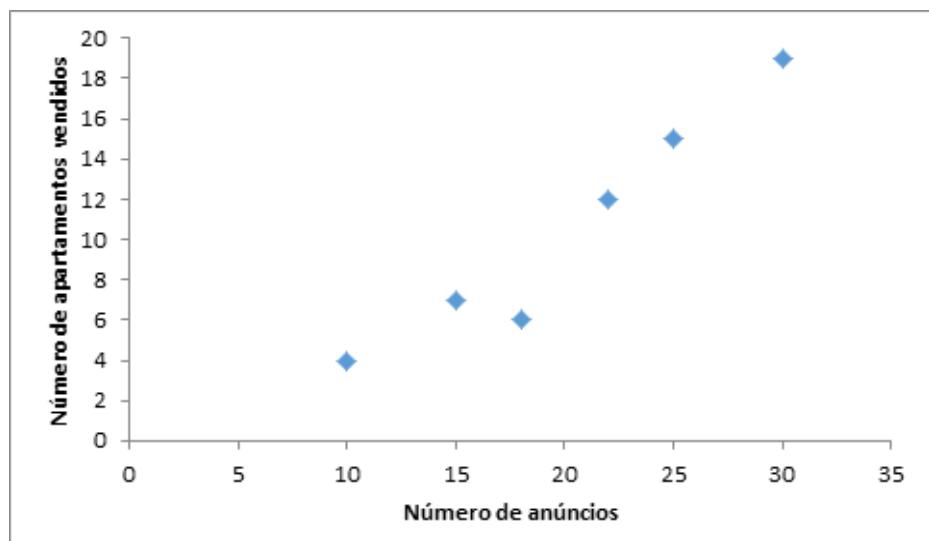


Figura 1.10 – Diagrama de dispersão do número de anúncios e número de apartamentos vendidos.

Pela análise gráfica observamos que à medida que o número de anúncios que foram ao ar aumenta, ocorre um aumento no número de apartamentos vendidos. Como identificamos uma relação entre as duas variáveis, podemos medir a intensidade da relação e fazer previsões do número de apartamentos vendidos a partir de um valor específico de anúncios. Estudaremos estes conceitos no Capítulo 5.

Vimos que os gráficos nos transmitem informações contidas no conjunto de dados, de maneira simples e de fácil compreensão. Apesar de ser uma ferramenta eficaz, precisamos tomar cuidado na construção dos gráficos para não obtermos conclusões enganosas. Os principais erros na elaboração de um gráfico são:

GRÁFICO SUCATA	<p>neste tipo de gráfico, há um uso excessivo de figuras que podem ocultar a informação que se deseja transmitir.</p>
AUSÊNCIA DE BASE RELATIVA	<p>quando utilizamos informações de mais de um conjunto de dados de tamanhos diferentes em um mesmo gráfico, com o objetivo de fazer comparações, devemos utilizar a frequência relativa em vez da frequência absoluta.</p>

EIXO VERTICAL COMPRIMIDO

as escalas empregadas devem ser coerentes com o tamanho da figura exibida. Se o eixo vertical estiver comprimido, as diferenças reais entre as categorias de respostas da variável podem ficar distorcidas.

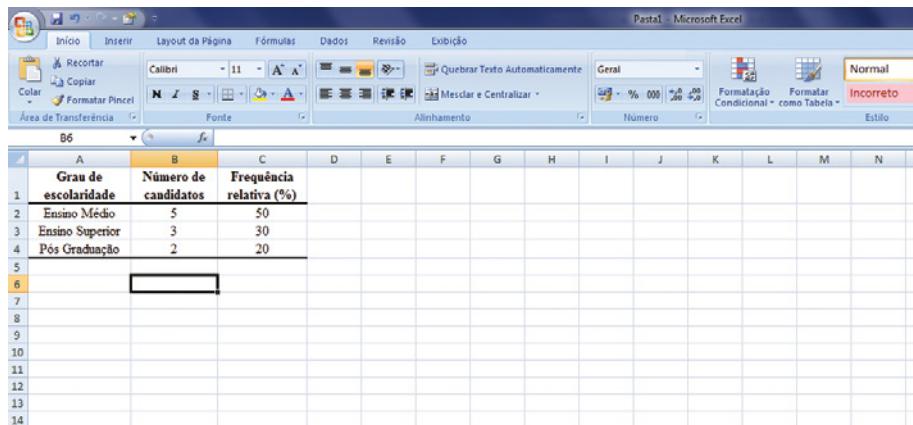
AUSÊNCIA DO PONTO ZERO

a ausência do ponto zero no eixo vertical tende a produzir uma impressão enganosa do comportamento dos dados, exagerando eventuais variações.

1.6 Utilização do Microsoft Excel na Construção de Gráficos

Os gráficos apresentados no item 1.5 foram construídos utilizando o Microsoft Excel. Estudaremos, agora, quais os procedimentos que devemos seguir para elaborar o gráfico de barras. Os procedimentos para construção de outros gráficos são semelhantes ao que vamos apresentar. Utilizaremos a versão 2010.

1º passo: Digitar em uma planilha as respostas da variável (numérica ou categórica) e suas respectivas frequências ou frequências relativas. Para exemplificar, utilizaremos os dados do Exemplo 1.7.



The screenshot shows a Microsoft Excel 2010 window with the ribbon menu at the top. The table below contains four rows of data: 'Ensino Médio' with 5 candidates and 50% relative frequency; 'Ensino Superior' with 3 candidates and 30% relative frequency; and 'Pós Graduação' with 2 candidates and 20% relative frequency. Row 5 is empty, and row 6 contains a single cell with a red border.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Grau de escolaridade	Número de candidatos	Frequência relativa (%)	D	E	F	G	H	I	J	K	L	M	N
2	Ensino Médio	5	50											
3	Ensino Superior	3	30											
4	Pós Graduação	2	20											
5														
6														
7														
8														
9														
10														
11														
12														
13														
14														

Figura 1.11 – Entrada dos dados

2º passo: Neste passo, selecionamos os dados. Podemos escolher a frequência absoluta ou relativa. Neste caso, o gráfico será construído com a frequência absoluta (colunas selecionadas: A e B, sem os títulos!). Após a seleção, escolher a aba Inserir e depois selecionar o tipo de gráfico a ser elaborado. Vamos escolher a primeira opção para o gráfico de Coluna (lembre-se, que já vimos que o Excel denomina o gráfico de barras verticais como coluna). Clicar em OK.

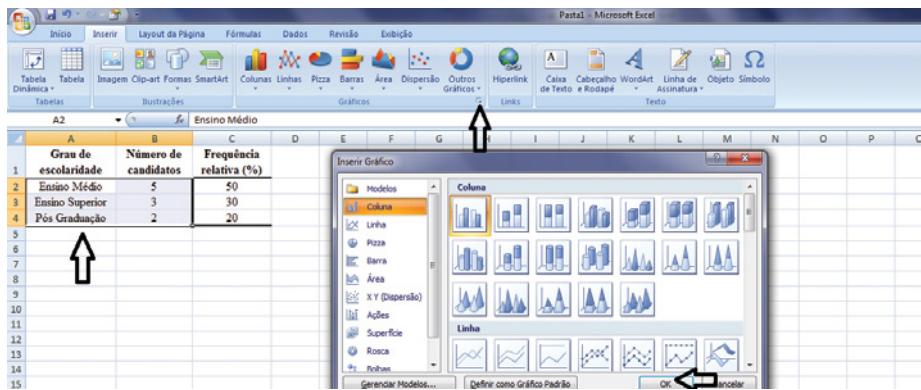


Figura 1.12 – Escolha do tipo de gráfico.

3º passo: O gráfico elaborado está na Figura 1.13. Observando as informações, percebemos que temos que formatá-lo, pois não há necessidade de legenda e os eixos estão sem título. Como opção, também podemos remover as linhas horizontais que aparecem no corpo do gráfico (linhas de grade).

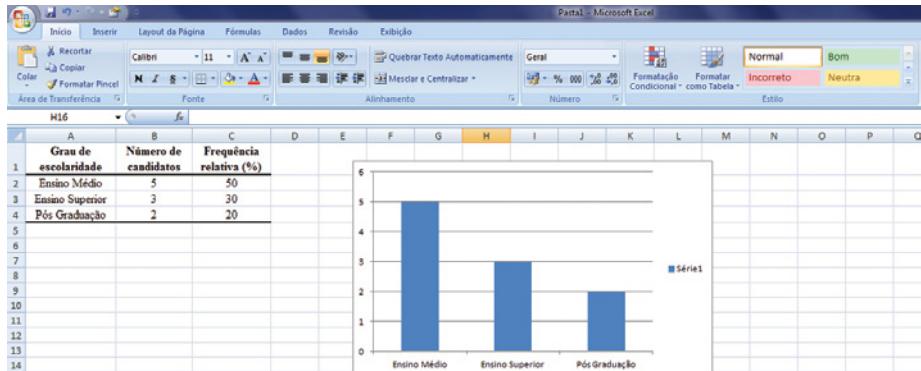


Figura 1.13 – Gráfico de barras verticais elaborado.

4º passo: Para iniciar a formatação, clicamos sobre o gráfico e aparecerá Ferramentas de Gráfico com algumas opções de escolha. Clicar em Layout e logo em seguida Títulos dos Eixos. Utilizamos as duas opções: uma para colocar título no eixo horizontal e a outra para colocar o título no eixo vertical. A Figura 1.14 ilustra a escolha para o Título do Eixo Horizontal Principal, com a opção Título Abaixo do Eixo. Após a inserção do título horizontal, seguimos o mesmo procedimento para o eixo vertical.

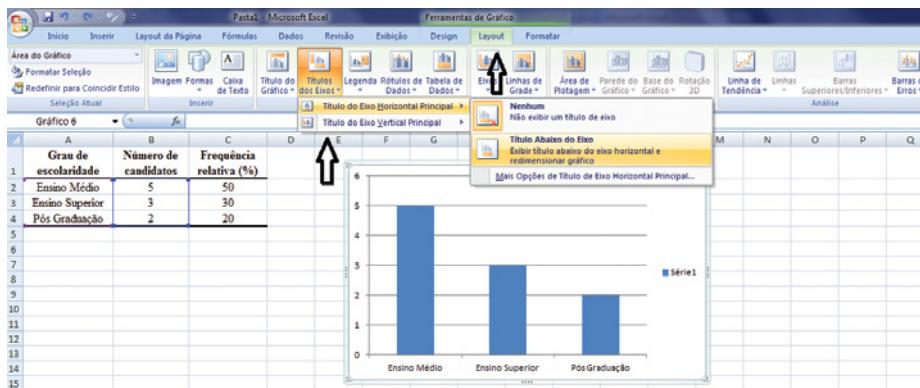


Figura 1.14 – Procedimento para inserir títulos nos eixos.

5º passo: A Figura 1.15 apresenta o gráfico com títulos nos eixos horizontal e vertical. Para finalizar, vamos excluir a legenda e as linhas de grade.

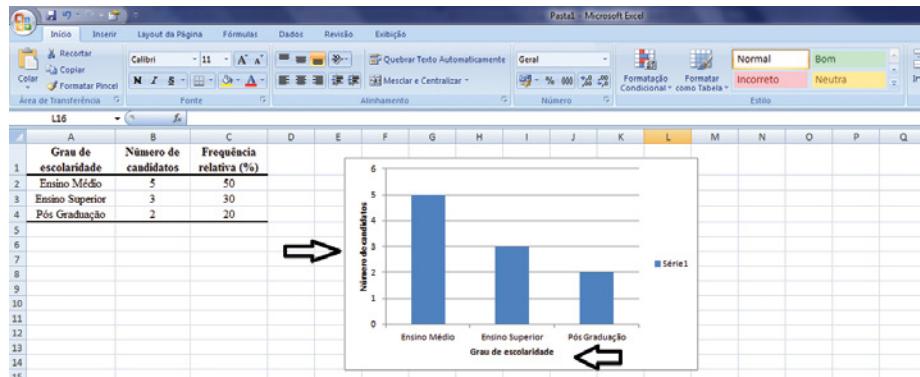


Figura 1.15 – Gráfico com título nos eixos.

6º passo: As exclusões da legenda e das linhas de grade também podem ser feitas por meio de Ferramentas de Gráfico. Clicar em Layout e logo em seguida em Legenda. Escolher a opção Nenhuma (Desativar legenda). Depois, clicar

em Linhas de grade, escolher a opção Linhas de Grade Horizontais Principais e clicar em Nenhuma. A Figura 1.16 apresenta o gráfico finalizado.

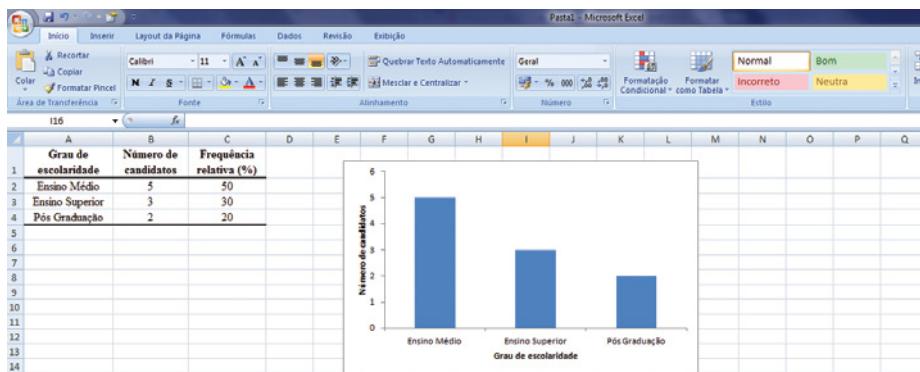


Figura 1.16 – Gráfico de barras horizontais para a variável Grau de escolaridade dos candidatos.

Agora, basta copiar e colar, por exemplo, em um arquivo formato DOC e interpretar as informações obtidas por meio da análise gráfica.

Para montar os outros gráficos com o auxílio do Excel, basta seguirmos os mesmos passos descritos acima. Há pequenas diferenças entre a montagem de um tipo de gráfico e outro, mas é fácil verificar quais procedimentos devem ser seguidos.

REFLEXÃO

Estamos encerrando nosso primeiro capítulo. Vimos, aqui, alguns conceitos que são fundamentais na compreensão do restante do conteúdo apresentado neste livro.

Com os conceitos adquiridos, você será capaz de coletar dados de maneira apropriada, saber identificá-los como qualitativos ou quantitativos e apresentá-los por meio de tabelas e gráficos.

Estamos apenas no começo. Muitas técnicas (muito interessantes!) ainda serão abordadas. E lembre-se que o conhecimento e o domínio da Estatística certamente ajudarão você a tomar às decisões mais acertadas.



LEITURA

No endereço <http://m3.ime.unicamp.br/recursos/1338> você encontrará dois áudios interessantes, primeiro módulo e segundo módulo, que introduz o conceito de Estatística e análise de dados por meio de informações sobre gravidez na adolescência.



REFERÊNCIAS BIBLIOGRÁFICAS

- BRUNI, Adriano L. **Estatística Aplicada à Gestão Empresarial**. 2. ed. São Paulo: Atlas, 2010.
- BUSSAB, Wilton de O. ; MORETTIN, Pedro A. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2002.
- MAGALHÃES, Marcos N.; LIMA, Antonio C. P de. **Noções de Probabilidade e Estatística**. 6. ed. São Paulo: Editora da Universidade de São Paulo, 2004.
- TRIOLA, Mário F. **Introdução à Estatística**. 10. ed. Rio de Janeiro: LTC, 2008.
- VIEIRA, Sonia. **Estatística básica**. São Paulo: Cengage Learning, 2013.
- VIEIRA, Sonia. **Introdução à Bioestatística**. 4 ed. Rio de Janeiro: Elsevier, 2008. Disponível em:<<http://saladeimprensa.ibge.gov.br/noticias?view=noticia&id=1&busca=1&idnoticia=1866>> Acesso em: 30 abr. 2015.
- OLIVEIRA, Tania M. Veludo. Disponível em: <http://www.fecap.br/adm_online/art23/tania2.htm>. Acesso em: 30 abr. 2015.
- OLIVEIRA, Ely F. Tannuri; GRÁCIO, Maria C. Cabrini. Disponível em: <http://dgz.org.br/ago05/Art_01.htm>. Acesso em: 30 abr. 2015.
- MONTEIRO, Carlos E. Ferreira. Disponível em: <http://www.ufrj.br/emanped/paginas/conteudo_producoes/docs_22/carlos.pdf> Acesso em: 30 abr. 2015.
- Disponível em: <http://www.procon.sp.gov.br/pdf/ranking_2013_coment.pdf>. Acesso em: 03 maio 2015.
- Disponível em : <<http://portalsaude.saude.gov.br/images/pdf/2014/julho/31/Dengue-classica-at---2013.pdf>>. Acesso em: 17 jun. 2015.
- FUSHIGIRA, Vanessa; OLIVEIRA, Samuel R.; SARTI, Luis R. Disponível em: <<http://m3.ime.unicamp.br/recursos/1338>>. Acesso em: 03 maio 2015.
-

2

Medidas Resumo

No primeiro capítulo vimos que, após a coleta dos dados brutos, é fundamental a organização e apresentação dos dados em distribuições de frequências e gráficos apropriados. Através deles, conseguimos ter uma visão geral do comportamento da variável em estudo. Além das distribuições de frequências, podemos resumir ainda mais um conjunto de dados quantitativos encontrando valores que sejam representativos de todo o conjunto.

Temos interesse em encontrar valores que descrevam duas características do conjunto de dados:

- A tendência central dos dados, ou seja, o centro em torno do qual os dados se distribuem.
- A variabilidade do conjunto de dados, ou seja, a maneira como os dados estão dispersos.

Estudaremos, primeiramente, as medidas de posição ou tendência central e, em seguida, as medidas de dispersão e separatrizes.



OBJETIVOS

Este capítulo aborda como podemos resumir um conjunto de dados quantitativos por meio de medidas resumo. Esperamos que, através dos conhecimentos aprendidos, você seja capaz de:

- Calcular e interpretar as medidas de tendência central e as medidas de dispersão;
- Compreender a importância das medidas separatrizes e utilizá-las para identificar a forma da distribuição dos dados.

2.1 Medidas de tendência central

2.1.1 Média aritmética

A média aritmética, ou simplesmente média, é a medida de tendência central mais conhecida.

Em muitas situações nos deparamos com informações referentes à média: o tempo médio de espera em um consultório médico é de 20 minutos, a média aritmética final de um estudante na disciplina de Matemática é 7,2, a taxa média de juros das operações de crédito para financiamento imobiliário está em 9,23%, etc.

Como fazemos para encontrar estas estatísticas que resumem todo o conjunto de dados em um único valor?

Para calcularmos a média precisamos somar os valores que aparecem no conjunto de dados e dividir pelo total de valores contidos neste conjunto. Vamos formalizar esta definição apresentando uma fórmula matemática:

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

em que:

\bar{x} média (lemos como 'x barra').

$\sum_{i=1}^n x_i$: somatório de n observações ($X_1, X_2, X_3, \dots, X_n$); X_1 representa o primeiro valor observado, X_2 representa o segundo valor observado e assim por diante, X_n representa o n-ésimo valor observado.

n: número de observações no conjunto de dados.

A fórmula apresentada para o cálculo da média é utilizada para dados amostrais. Quando estivermos trabalhando com dados de toda a população, usamos uma notação diferente. O número de observações é denotado por N e utilizamos a letra grega μ (Mi) para indicar a média, ou seja, $\mu = \frac{\sum_{i=1}^n x_i}{N}$.



EXEMPLO

2.1: Um questionário foi aplicado aos dez candidatos a uma vaga no setor financeiro de uma clínica de cirurgia plástica e uma das variáveis em estudo era a idade dos candidatos. Os dados obtidos foram:

30	35	26	22	28	30	26	33	35	23
----	----	----	----	----	----	----	----	----	----

Vamos encontrar a idade média dos candidatos à vaga.

Resolução

Sabemos que para encontrar a média, somamos todos os valores e dividimos pela quantidade de valores no conjunto de dados. Para nos familiarizarmos, vamos utilizar a fórmula:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_{10}}{10} = \frac{30 + 35 + 26 + \dots + 23}{10} = \frac{288}{10} = 28,8 \text{ anos}$$

Portanto, a idade média dos candidatos é 28,8 anos.

Quando os dados estiverem organizados em uma distribuição de frequências, podemos utilizar a seguinte fórmula:

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

Para utilizarmos esta fórmula, acrescentamos uma coluna na distribuição de frequências:

FREQUÊNCIA RELATIVA (%)	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)	$x_i \cdot f_i$
x_1	f_1		$x_1 \cdot f_1$
x_2	f_2		$x_2 \cdot f_2$
...
x_n	f_n		$x_n \cdot f_n$
Total	número total de observações no conjunto de dados	100,00	$\sum_{i=1}^k x_i \cdot f_i$

Tabela 2.1 – Estrutura da distribuição de frequências para o cálculo da média por meio dos dados tabelados.

Só faz sentido acrescentarmos a coluna $(x_i \cdot f_i)$ se quisermos encontrar a média, ou seja, ela é uma coluna auxiliar do cálculo.

2.2: Construindo uma distribuição de frequências para os dados do Exemplo 2.1, obtemos:

IDADE	NÚMERO DE CANDIDATOS	FREQUÊNCIA RELATIVA (%)
22	1	10
23	1	10
26	2	20
28	1	10
30	2	20
33	1	10
35	2	20
Total	10	100

Tabela 2.2 – Distribuição dos candidatos, segundo a idade.

Vamos encontrar a idade média dos candidatos à vaga por meio da distribuição de frequências.

Resolução

Como os dados já estão organizados em uma distribuição de frequências, basta acrescentarmos uma coluna na tabela:

IDADE (x_i)	NÚMERO DE CANDIDATOS (f_i)	FREQUÊNCIA RELATIVA (%)	$(x_i \cdot f_i)$
22	1	10	22
23	1	10	23
26	2	20	52
28	1	10	28
30	2	20	60
33	1	10	33
35	2	20	70
Total	10	100	288

Tabela 2.3 – Cálculo da coluna auxiliar para encontrar a média.

Então:

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot f_i}{n} = \frac{288}{10} = 28,8 \text{ anos}$$

A média aritmética possui algumas propriedades importantes, que estudaremos no próximo item.

2.1.1.1 Propriedades da média

1. A soma dos desvios é zero, ou seja:

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

Em palavras: os desvios são encontrados fazendo a diferença entre cada valor do conjunto de dados e a média aritmética do conjunto. A soma dos desvios encontrados é zero, para qualquer conjunto de dados.

2. Quando somamos (ou subtraímos) uma constante de todos os valores de um conjunto de dados, a média fica somada (ou subtraída) por esta constante.
3. Quando multiplicamos (ou dividimos) uma constante de todos os valores de um conjunto de dados, a média fica multiplicada (ou dividida) por esta constante.

Outro tipo de média muito utilizada, por exemplo, no cálculo da média final de um estudante em uma disciplina ou na nota final do candidato em um concurso, é a média ponderada. Na média ponderada são atribuídos aos valores importâncias diferentes. Por exemplo, um estudante pode fazer 4 provas durante o semestre e para cada prova é atribuído um peso. O cálculo da média ponderada é feito por meio do somatório das multiplicações entre valores e pesos, divididos pelo somatório dos pesos, ou seja,

$$\bar{x}_p = \frac{\sum x_i p_i}{\sum p_i}$$
, em que Pi são os pesos atribuídos.

2.1.2 Moda

A moda de um conjunto de dados é a resposta (ou respostas) que aparece(m) com maior frequência. A moda, diferentemente das outras medidas de posição, também pode ser encontrada quando a variável em estudo for qualitativa.

Portanto, a resposta para a moda pode ser o valor ou a categoria que aparece com a maior frequência. Existem conjuntos de dados em que nenhuma resposta aparece mais vezes que outras. Neste caso, dizemos que o conjunto de dados não apresenta moda.

Em outros casos, podem aparecer duas ou mais respostas de maior frequência no conjunto de dados. Nestes casos, dizemos que o conjunto de dados é bimodal e multimodal, respectivamente.

No conjunto de dados apresentados no Exemplo 2.1, temos que as respostas que aparecem com maior frequência (frequência 2) são: 26, 30 e 35. Portanto:

$$M_o = 26, 30 \text{ e } 35 \text{ anos}$$

Neste caso, a distribuição é multimodal.

2.1.3 Mediana

A mediana é uma medida que divide o conjunto de dados ordenados ao meio, deixando a mesma quantidade de valores abaixo dela e acima. Por isto, ela também é uma medida separatriz, pois separa o conjunto de dados em dois grupos: pelo menos 50% dos valores ordenados são maiores ou iguais ao valor da mediana e pelo menos 50% dos valores ordenados são menores ou iguais ao valor da mediana.

O cálculo para se encontrar a mediana difere no caso do número de elementos (n) do conjunto de dados ser par ou ímpar.

Se o número de elementos do conjunto de dados for ímpar, então a mediana será exatamente o valor “do meio”, ou seja:

$$Md = \frac{x_{\frac{n+1}{2}}}{2}$$

Se o número de elementos do conjunto de dados for par, então a mediana será exatamente a média “dos dois valores do meio”, isto é:

$$Md = \frac{\frac{x_n}{2} + \frac{x_{n+1}}{2}}{2}$$

em que x_n , $x_{\frac{n}{2}+1}$ e $x_{\frac{n+1}{2}}$ indicam as observações que ocupam as posições “do meio” do conjunto de dados.



EXEMPLO

2.3: Os dados abaixo se referem aos batimentos cardíacos para 15 pacientes que chegaram ao hospital em estado de parada respiratória e inconscientes. Vamos encontrar a mediana.

167	150	125	120	150	150	140	136	120	150	125	140	148	120	125
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Resolução

Para encontrarmos a mediana, os dados precisam estar ordenados:

120	120	120	125	125	125	136	140	140	148	150	150	150	150	167
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Temos $n = 15$ observações, então:

$$\begin{aligned}Md &= x_{\frac{n+1}{2}} \\Md &= x_{\frac{15+1}{2}} = x_8\end{aligned}$$

ou seja, a mediana é o valor que ocupa a oitava posição do conjunto de dados ordenados,

$$Md = 140$$

Repare que a observação 140 divide o conjunto de dados ao meio, com 7 observações abaixo dela e 7 observações acima dela.

Então, concluímos que pelo menos 50% dos valores são maiores ou iguais a 140 batidas por minuto.

Também podemos encontrar a mediana quando os dados estão apresentados em uma distribuição de frequências. Para isto, seguimos o seguinte procedimento:

1º Passo: identificaremos a frequência acumulada imediatamente superior à metade do somatório do número de observações do conjunto de dados:

$$\frac{n}{2}$$

2º Passo: a mediana será o valor da variável associada à frequência acumulada imediatamente superior ao valor encontrado no 1º Passo.

Quando $\frac{n}{2}$ for ser exatamente igual a uma das frequências acumuladas f_a , o cálculo da mediana será a média aritmética entre dois valores da variável: x_i e x_{i+1} . O valor da variável x_i será aquele associado à $\frac{n}{2} = f_a$ e o valor da variável x_{i+1} será aquele que está imediatamente após x_i na distribuição de frequências.

Para facilitar a compreensão, vamos aplicar no próximo exemplo o passo a passo descrito acima.

2.4: O número de faltas ao trabalho, no último semestre, dos 30 funcionários de uma clínica, são:

NÚMERO DE FALTAS	0	1	2	3
FREQUÊNCIA DE FUNCIONÁRIOS	9	10	5	6

Resolução

Vamos organizar uma distribuição de frequências incluindo a frequência acumulada.

Valor da variável associado à frequência acumulada igual a 19	NÚMERO DE FALTAS	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)	f_a	Frequência acumulada imediatamente superior a 15
	0	9	30,00	9	
	1	10	33,33	19	
	2	5	16,67	24	
	3	6	20,00	30	
	Total	30	100,00		

Seguindo o roteiro:

1º Passo:

$$\frac{n}{2} = \frac{30}{2} = 15$$

A frequência acumulada imediatamente superior a 15 é $f_a = 19$.

2º Passo: a mediana será o valor da variável associado à frequência acumulada imediatamente superior ao valor encontrado no 1º Passo. Portanto:

$$Md = 1 \text{ falta}$$

Lembre que o valor da variável está na primeira coluna da tabela!

Em algumas situações, a mediana pode ser a medida de tendência central mais representativa para o conjunto de dados em estudo. Vamos entender quando isto ocorre analisando o próximo exemplo.

2.5: Trinta residências de um bairro foram selecionadas para participar de uma pesquisa e uma das variáveis em estudo era a renda familiar (salários mínimos). Os dados obtidos foram:

4,3	5,1	5,7	6,4	6,8	7,1	7,4	7,6	8,2	8,7
8,9	9,2	9,5	9,7	10,0	10,4	10,6	11,2	11,4	11,6
11,7	11,9	12,1	12,3	12,4	12,4	12,7	13,2	13,5	91,3

Vamos calcular a média e a mediana para este conjunto de dados.

Resolução

Para encontrar a média, somamos todos os valores e dividimos por 30, ou seja:

$$\bar{x} = \frac{4,3 + 5,1 + 5,7 + \dots + 91,3}{30} = \frac{373,3}{30} = 12,44 \text{ s.m.}$$

Ou seja, concluímos que a renda familiar média dos moradores das 30 residências selecionadas é 12,44 salários mínimos.

Analizando o conjunto de dados, observamos que o valor encontrado para a média está acima dos valores de 26 observações do conjunto! Por que isto ocorreu? Temos uma observação discrepante, ou seja, muito maior que as outras, que é 91,3. Esta observação ‘puxa’ a média para cima, fazendo com que tenhamos uma interpretação enganosa sobre o centro em torno do qual os dados se distribuem.

A média aritmética é muito sensível a valores extremos, então, dizemos que a média não é uma medida de tendência central resistente.

Agora, vamos analisar o que acontece no cálculo da mediana.

Temos $n = 30$ observações, então:

$$Md = \frac{\frac{x_{30}}{2} + \frac{x_{30+1}}{2}}{2}$$
$$Md = \frac{x_{15} + x_{16}}{2}$$

ou seja, a mediana é a média entre os valores que ocupam a décima quinta e décima sexta posição do conjunto de dados ordenados.

$$Md = \frac{10 + 10,4}{2} = 10,2 \text{ s.m.}$$

Com o resultado obtido para a mediana, observamos que ela não é afetada pela observação discrepante, sendo, portanto, a medida de tendência central mais representativa para este conjunto de dados.

Agora que já sabemos calcular e interpretar a média, moda e mediana, podemos utilizá-las para detectar assimetria em um conjunto de dados:

- Se a distribuição dos dados for exatamente simétrica, a média, a moda e a mediana são exatamente iguais. Para distribuições aproximadamente simétricas, as três medidas são próximas.
- Se a distribuição dos dados apresentar assimetria à esquerda, em geral, a média é menor que a mediana; e se apresentar assimetria à direita, em geral, a mediana é menor que a média.

A distribuição dos dados é assimétrica quando se estende mais para um lado do que para o outro e é simétrica se a metade esquerda do seu histograma se comporta de maneira praticamente igual da sua metade direita. No Capítulo 1 vimos que o histograma é um gráfico muito utilizado para identificar a forma da distribuição dos dados.

2.1.4 Cálculos das medidas de tendência central para dados agrupados em intervalos de classes

Aprendemos, até agora, a calcular as medidas de posição central pelo conjunto de dados brutos ou pela distribuição de frequências sem intervalos de classes. E quando os dados estiverem apresentados em intervalos de classes, como vamos calcular tais medidas? Quando agrupamos as observações em classes, perdemos a informação dos valores que estão dentro de cada classe. Neste caso, vamos supor que todos os valores dentro de uma classe tenham seus valores iguais ao ponto médio desta classe.

Por exemplo, vamos supor que o intervalo de 10| –15 tenha frequência 5. Não sabemos quais são os valores destas 5 observações, só podemos afirmar que são maiores ou iguais a 10 e menores que 15. Então, assumiremos que as 5 observações são iguais a 12,5, que é o ponto médio deste intervalo.

Vamos aprender a calcular as medidas de tendência central para dados agrupados através do exemplo a seguir.



EXEMPLO

2.6: Uma professora de Ciências, interessada em fazer uma aula prática com seus alunos, fez um levantamento dos pesos, em quilogramas, de cada um deles. Os dados estão apresentados na Tabela 2.4.

PESO (KG)	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)
40 –45	8	5,59
45 –50	25	17,48
50 –55	50	34,97
55 –60	40	27,97
60 –65	20	13,99
Total	143	100,00

Tabela 2.5 – Distribuição de frequências dos pesos dos alunos.

Resolução

- Média

Para encontrarmos a média, precisamos acrescentar duas colunas na distribuição de frequências: x_i (ponto médio da classe) e $x_i \cdot f_i$.

Para o cálculo da mediana, precisaremos da frequência acumulada. Então, vamos acrescentar mais uma coluna contendo tais frequências.

PESO (KG)	FREQUÊNCIA	FREQUÊNCIA RELATIVA(%)	x_i	$x_i \cdot f_i$	FREQUÊNCIA ACUMULADA
40 –45	8	5,59	42,5	340	8
45 –50	25	17,48	47,5	1.187,5	33
50 –55	50	34,97	52,5	2.625	83
55 –60	40	27,97	57,5	2.300	123

PESO (KG)	FREQUÊNCIA	FREQUÊNCIA RELATIVA(%)	x_i	$x_i \cdot f_i$	FREQUÊNCIA ACUMULADA
60 – 65	20	13,99	62,5	1.250	143
Total	143	100,00		7.702,50	

Tabela 2.6 – Cálculos das colunas auxiliares para encontrar a média e a mediana.

Para encontrar o ponto médio, basta fazer $Pm = \frac{Li + Ls}{2}$. Então, para o primeiro intervalo, $Pm = \frac{40 + 45}{2} = 42,5$.

Substituindo os valores encontrados na fórmula, temos:

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot f_i}{n} = \frac{7.702,50}{143} = 53,86\text{kg}$$

- Moda

Existem várias definições para localizar a posição da moda em uma classe modal, mas a mais simples é definir a moda como o ponto médio da classe modal.

Portanto, neste exemplo, a classe modal é 50 | – 55 (pois, apresenta a maior frequência = 50) e, vamos considerar a moda o ponto médio desta classe, ou seja:

$$M_o = 52,5\text{kg}$$

- Mediana

Para o cálculo da mediana utilizaremos uma fórmula que, a princípio, pode parecer um pouco complexa ou trabalhosa, mas veremos que as quantidades que precisamos para substituir na fórmula são fáceis de serem obtidas. Utilizaremos a seguinte fórmula para o cálculo da mediana para dados agrupados em intervalos de classes:

$$Md = l_{inf_{md}} + \frac{h_{md}}{f_{md}} \cdot \left(\frac{n}{2} - F_{a_{ant}} \right)$$

em que:

$l_{inf_{md}}$:: limite inferior do intervalo que contém a mediana;

h_{md} :: amplitude do intervalo de classe que contém a mediana;

f_{md} :: número de observações do intervalo que contém a mediana;

n: número total de observações da distribuição de frequências;

$F_{a_{ant}}$:: frequência acumulada do intervalo anterior àquele que contém a mediana.

A primeira informação que precisamos é saber qual intervalo contém a mediana. Este intervalo está associado à frequência acumulada imediatamente superior à $\frac{n}{2}$.

Pela Tabela 2.5, como $\frac{n}{2} = \frac{143}{2} = 71,5$, o intervalo que contém a mediana é 50 | – 55 (pois $f_a = 83$).

Após a identificação do intervalo, conseguimos identificar todos os valores exigidos na fórmula:

$$l_{inf_{md}} :: 50$$

$$h_{md} :: 55 - 50 = 5$$

$$f_{md} :: 50$$

$$n: 143$$

$$F_{a_{ant}} :: 33$$

	PESO (KG)	FREQUÊNCIA	FREQUÊNCIA ACUMULADA	
Intervalo que contém a mediana	40 – 45	8	8	p_a do intervalo anterior àquele que contém a mediana
	45 – 50	25	33	
	50 – 55	50	83	Número de observações do intervalo que contém a mediana
	55 – 60	40	123	
	60 – 65	20	143	
	Total	143		

Tabela 2.7 – Identificação dos valores que serão utilizados no cálculo da mediana.

Substituindo os valores encontrados na fórmula, temos:

$$Md = l_{inf_{md}} + \frac{h_{md}}{f_{md}} \cdot \left(\frac{n}{2} - F_{a_{ant}} \right)$$

$$Md = 50 + \frac{5}{50} \cdot \left(\frac{143}{2} - 33 \right)$$

$$Md = 50 + 3,85 = 53,85\text{kg}$$

Pelo menos 50% das observações são maiores ou iguais a 53,85 kg.

As medidas resumo calculadas quando os dados estiverem agrupados em intervalos de classes são apenas aproximações dos verdadeiros valores, pois substituímos os valores das observações pelo ponto do médio do intervalo de classe.

As medidas de posição que estudamos não bastam para descrever um conjunto de dados. Tais medidas têm como objetivo indicar o centro em torno do qual os dados estão dispersos, mas não informam o quanto os dados se dispersam. Por exemplo, uma pergunta natural que surge após o cálculo da média é: será que as observações do conjunto de dados estão próximas ou distantes (dispersas) do valor médio encontrado?

Veremos, no próximo item, algumas medidas que nos auxiliam na resposta a este questionamento.

2.2 Medidas de dispersão

Antes de aprendermos a calcular algumas medidas de dispersão, vamos entender o conceito de variabilidade com o exemplo a seguir.



EXEMPLO

2.7: Os dados abaixo se referem aos salários de 10 funcionários que possuem o cargo de enfermeiro chefe nas cidades e região metropolitana de São Paulo e Belo Horizonte.

S.P.	3 250	4 125	5 270	6 029	9 840	5 127	6 350	4 250	7 125	3 850
B.H.	5 250	5 025	5 270	5 550	5 870	5 625	5 120	5 840	5 720	5 946

$$\bar{x}_{SP} = \frac{\sum_{i=1}^n x_i}{n} = \frac{3250 + 4125 + \dots + 7125 + 3850}{10} = 5521,60 \text{ reais}$$

e

$$\bar{x}_{BH} = \frac{\sum_{i=1}^n x_i}{n} = \frac{5250 + 5025 + \dots + 5720 + 5946}{10} = 5521,60 \text{ reais}$$

Embora as médias sejam iguais, observamos claramente que a variabilidade dos salários na cidade de São Paulo e região metropolitana é maior que em Belo Horizonte. Portanto, a média descreve bem a situação em Belo Horizonte, mas não em São Paulo.

Agora que ficou claro o conceito de dispersão ou variabilidade, vamos aprender a calcular as medidas de dispersão.

2.2.1 Mínimo, máximo e amplitude

O mínimo e o máximo de um conjunto de dados são, respectivamente, o menor e o maior valor do conjunto.

A amplitude de um conjunto de dados é a diferença entre o valor máximo e o valor mínimo dos dados, ou seja:

$$\text{Amplitude} = x_{(\text{máximo})} - x_{(\text{mínimo})}$$



EXEMPLO

2.8: Considerando os dados do Exemplo 2.7, vamos encontrar o mínimo, o máximo e a amplitude do conjunto de dados na cidade de São Paulo e região metropolitana.

Resolução

MÍNIMO	MÁXIMO	AMPLITUDE
3 250	9 840	6 590

Tabela 2.8 – Valores mínimo, máximo e amplitude dos salários em São Paulo e região metropolitana.

Pela amplitude, observamos que a diferença entre o salário mais alto e o mais baixo é de R\$ 6 590,00. Apesar de ser uma medida fácil de calcular e interpretar, a amplitude não é muito utilizada, pois leva em conta somente dois valores de todo o conjunto de dados. Este cálculo permite que dois conjuntos de dados com variabilidades muito diferentes tenham mesma amplitude e, permite, também, que valores extremos aumentem muito a amplitude.

O conveniente é utilizarmos uma medida que utilize todas as observações do conjunto de dados. Estudaremos nos próximos itens medidas que têm como princípio básico analisar a dispersão de cada observação em relação à média dessas observações.

2.2.2 Desvio médio, variância e desvio padrão amostrais

Antes de apresentarmos as fórmulas para o cálculo do desvio médio e da variância, vamos entender qual o conceito de desvio em estatística. Desvio nada mais é do que a distância entre qualquer observação do conjunto de dados em relação à média aritmética deste mesmo conjunto:

$$\begin{aligned}\text{desvio} &= \text{observação} - \text{média} \\ \text{desvio} &= x - \bar{x}\end{aligned}$$

É intuitivo pensar que se os desvios em relação à média são pequenos, as observações estão concentradas em torno da média e, portanto, a variabilidade dos dados é pequena. Agora, se os desvios são grandes, é porque as observações estão dispersas e, portanto, a variabilidade dos dados é grande.

Já vimos, na propriedade 1 da média que, para qualquer conjunto de dados, a soma dos desvios é igual a zero. Então, $\sum_{i=1}^n (x_i - \bar{x})$ não é uma boa medida de dispersão, pois ela não seria nada informativa sobre a dispersão das observações. Para contornar o resultado desta propriedade, podemos considerar o total dos desvios em valor absoluto, ou seja, $\sum_{i=1}^n |x_i - \bar{x}|$. Mas, somente o uso deste total pode causar dificuldades de interpretação quando estivermos comparando conjunto de dados com números diferentes de observações. Então, o conveniente é definir a medida como média, obtendo o desvio médio:

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

O desvio médio é uma média dos valores absolutos dos desvios em relação à média. Esta medida utiliza o módulo que, por suas características matemáticas, torna difícil o estudo de suas propriedades. Então, vamos definir uma medida que utiliza o quadrado dos desvios em relação à média.

A variância amostral é uma medida de dispersão que pode ser interpretada como uma média dos quadrados dos desvios, ou seja:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

O denominador, $n - 1$, são os graus de liberdade associados à variância. Uma explicação detalhada da utilização de $n - 1$ no denominador é encontrada em TRIOLA (2008, p.83).

Uma fórmula alternativa para o cálculo da variância é:

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

em que:

- x_i^2 : soma de cada valor observado ao quadrado;
- $(\sum x_i)^2$: quadrado da soma de todos os valores observados;
- n : número total de observações no conjunto de dados.

Apesar, de à primeira vista, a fórmula alternativa parecer mais complicada, os cálculos exigidos são feitos com menor número de operações aritméticas.

Quando os dados estiverem organizados em uma distribuição de frequências, podemos utilizar a seguinte fórmula:

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot f_i}{n-1} = \frac{(x_1 - \bar{x})^2 \cdot f_1 + (x_2 - \bar{x})^2 \cdot f_2 + \dots + (x_k - \bar{x})^2 \cdot f_k}{n-1}$$

Ou, pela fórmula alternativa:

$$s^2 = \frac{\sum_{i=1}^n x_i^2 \cdot f_i - \frac{(\sum_{i=1}^n x_i \cdot f_i)^2}{n}}{n-1}$$

Como a variância envolve os quadrados dos desvios, a unidade de medida da variância é igual ao quadrado da medida das observações (por exemplo, mim^2 , kg^2 , m^2 etc). As unidades elevadas ao quadrado associadas à variância tornam difícil a interpretação do valor numérico.

Para obtermos uma medida de variabilidade cuja unidade de medida seja a mesma do conjunto de dados, extraímos a raiz quadrada da variância. Esta medida é denominada desvio padrão amostral:

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

em que:

s: desvio padrão;

s^2 : variância.

Da maneira que o desvio padrão é definido, podemos concluir que:

- O desvio padrão é uma medida de variação de todos os valores a partir da média.
- O valor do desvio padrão é maior ou igual a zero. Será zero apenas quando todos os valores do conjunto de dados forem iguais.
- Valores muito próximos resultarão em desvios padrões pequenos, enquanto que valores mais espalhados resultarão em desvios padrões maiores.
- O valor do desvio padrão pode aumentar drasticamente com a inclusão de um ou mais valores discrepantes.
- A unidade de medida do desvio padrão é a mesma do conjunto de dados.
- O desvio padrão é utilizado para comparar a variabilidade de dois conjuntos de dados diferentes quando as médias forem aproximadamente iguais e quando as unidades de medidas para os dois conjuntos forem as mesmas.

As fórmulas apresentadas para o cálculo da variância e do desvio padrão são aplicadas quando estamos trabalhando com dados amostrais. No caso do conjunto de dados ser a própria população, o denominador da variância é N e substituímos s^2 por σ^2 (σ : letra grega sigma). Então, a fórmula da variância populacional é $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{N}$ e o desvio padrão populacional é $\sigma = \sqrt{\sigma^2}$

2.2.2.1 Uma regra prática para interpretar o desvio-padrão

Depois que calculamos o desvio-padrão, surge uma pergunta: como interpretá-lo?

Para conjuntos de dados que tenham distribuição com forma aproximadamente de sino, valem as seguintes considerações:

- Cerca de 68% das observações do conjunto de dados ficam a 1 desvio padrão da média, ou seja, $(\bar{x} - s)$ e $(\bar{x} + s)$.

- Cerca de 95% das observações do conjunto de dados ficam a 2 desvios padrões da média, ou seja, $(\bar{x} - 2s) \text{ e } (\bar{x} + 2s)$.
- Cerca de 99,7% das observações do conjunto de dados ficam a 3 desvios padrões da média, ou seja, $(\bar{x} - 3s) \text{ e } (\bar{x} + 3s)$.

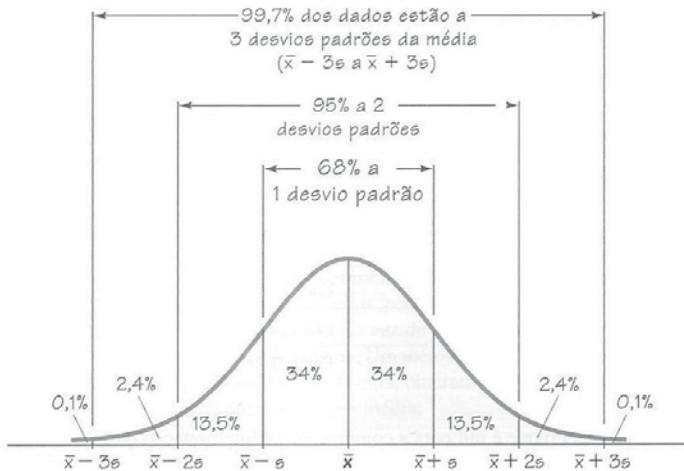


Figura 2.1: Regra prática para interpretação do desvio-padrão. Fonte: TRIOLA (2008, p. 81).

Em uma distribuição em forma de sino, as frequências começam baixas, crescem até uma frequência máxima e depois decrescem para uma frequência baixa. Além disto, a distribuição deve ser aproximadamente simétrica, com frequências igualmente distribuídas em ambos os lados da frequência máxima.

Como a média aritmética, o desvio padrão também possui algumas propriedades importantes, que apresentaremos a seguir.

2.2.2.2 Propriedades do desvio padrão

1. Quando somamos (ou subtraímos) uma constante de todos os valores de um conjunto de dados, o desvio padrão não se altera.
2. Quando multiplicamos (ou dividimos) uma constante de todos os valores de um conjunto de dados, o desvio padrão fica multiplicado (ou dividido) por esta constante.



EXEMPLO

2.9: Os dados abaixo referem-se às notas finais de dois alunos, um deles está na turma da manhã e o outro na turma da noite, na disciplina Bioestatística.

MANHÃ	9,5	7,5	3,5	6,0	6,5	2,0	7,0	1,0
NOITE	5,0	5,5	5,0	6,5	6,0	4,5	5,5	5,0

Vamos calcular as medidas de dispersão. De acordo com as informações, qual aluno apresenta maior variabilidade nas notas?

Resolução

Analisando as notas dos alunos, conseguimos identificar que as notas do aluno da manhã estão variando mais, enquanto que as notas do aluno da noite estão mais próximas uma das outras. Os dois alunos apresentam o mesmo desempenho médio na disciplina, pois:

$$\bar{x}_{manhã} = \frac{\sum_{i=1}^n x_i}{n} = \frac{9,5 + 7,5 + 3,5 + \dots + 1}{8} = 5,375$$

e

$$\bar{x}_{noite} = \frac{\sum_{i=1}^n x_i}{n} = \frac{5 + 5,5 + 5 + \dots + 5}{8} = 5,375$$

Para exercitar as fórmulas, vamos resolver este exercício de duas maneiras: da maneira como os dados estão apresentados no enunciado e depois organizando-os em uma distribuição de frequências.

Primeira maneira – aluno manhã

O valor mínimo é 1 e o máximo é 9,5. Portanto, a amplitude é 8,5 pontos, ou seja a diferença entre a menor nota e a maior é 8,5 pontos.

Para encontrarmos a variância, vamos utilizar a fórmula alternativa:

$$\mu^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{N}}{N}$$

Neste exemplo, utilizaremos a fórmula da variância populacional, pois estamos trabalhando com todas as notas dos alunos na disciplina Bioestatística.

Precisamos encontrar a soma de cada valor observado ao quadrado:

$$(9,5)^2 + (7,5)^2 + (3,5)^2 + (6,0)^2 + (6,5)^2 + (2,0)^2 + (7,0)^2 + (1,0)^2 = 291$$

Agora, precisamos encontrar o quadrado da soma de todos os valores observados:

$$(9,5 + 7,5 + 3,5 + 6,0 + 6,5 + 2,0 + 7,0 + 1,0)^2 = (43)^2 = 1849$$

Temos todos os valores necessários para substituir na fórmula:

$$\sigma^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{N}}{N} = \frac{291 - \frac{(43)^2}{8}}{8} = \frac{291 - 231,125}{8} = \frac{59,875}{8} = 7,48 \text{ pontos}^2$$

O desvio padrão é:

$$\hat{\sigma} = \sqrt{7,48} = 2,74 \text{ pontos}$$

Vamos seguir o mesmo procedimento para encontrar as medidas de dispersão para as notas do aluno da noite.

O valor mínimo é 4,5 e o máximo é 6,5. Então, a amplitude é 2,0 pontos, ou seja, a diferença entre a menor nota e a maior é 2,0 pontos.

Para o cálculo da variância precisamos das seguintes quantidades:

$$(5,0)^2 + (5,5)^2 + (5,0)^2 + (6,5)^2 + (6,0)^2 + (4,5)^2 + (5,5)^2 + (5,0)^2 = 234$$

e

$$(5,0 + 5,5 + 5,0 + 6,5 + 6,0 + 4,5 + 5,5 + 5,0)^2 = (43)^2 = 1849$$

Substituindo os valores encontrados na fórmula da variância, temos:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{N}}{N} = \frac{234 - \frac{1849}{8}}{8} = \frac{234 - 231,125}{8} = \frac{2,875}{8} = 0,36 \text{ ponto}^2$$

O desvio padrão é:

$$\hat{\sigma} = \sqrt{0,36} = 0,60 \text{ ponto}$$

Vamos colocar as informações em um quadro para facilitar a interpretação dos resultados obtidos.

MEDIDAS DE DISPERSÃO	MÍNIMO	MÁXIMO	AMPLITUDE	VARIÂNCIA	DESVIO PADRÃO
ALUNO - MANHÃ	1,0	9,5	8,5	7,48	2,74
ALUNO - NOITE	4,5	6,5	2,0	0,36	0,60

Tabela 2.9 – Medidas de dispersão para as notas dos dois alunos.

Apesar de já estar claro analisando as notas dos dois alunos, confirmamos através das medidas de dispersão que as notas do aluno da manhã apresentam maior variabilidade. Apesar das médias das notas dos dois alunos serem iguais, todas as medidas de dispersão indicam maior variabilidade nas notas do aluno da manhã. Vale ressaltar que a variância tem a unidade de medida elevada ao quadrado, portanto, utilizamos o desvio padrão para interpretar o resultado obtido.

Segunda maneira – aluno noite

Agora, vamos calcular as medidas por meio dos dados apresentados em uma distribuição de frequências. Na distribuição de frequências, acrescentamos duas colunas ($x_i \cdot f_i$ e $x_i^2 \cdot f_i$) cujos somatórios são exigidos na fórmula da variância. Utilizaremos as notas do aluno da noite para aprender a fazer os cálculos por meio dos dados tabulados. Após o aprendizado, faça o mesmo procedimento com o aluno da manhã e compare com os resultados obtidos através da primeira maneira. Você encontrará os mesmos resultados!

NOTAS (x_i)	FREQUÊNCIA (f_i)	FREQUÊNCIA RELATIVA (%)	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
4,5	1	12,50	4,5	20,25
5	3	37,50	15	75
5,5	2	25,00	11	60,5
6	1	12,50	6	36
6,5	1	12,50	6,5	42,25
Total	8	100,00	43	234

Tabela 2.10 – Cálculos das colunas auxiliares para encontrar a variância – aluno noite.

Pela distribuição de frequências também identificamos o mínimo (primeira nota) como 4,5, o máximo (última nota) como 6,5 e amplitude 2.

Utilizamos os somatórios das duas últimas colunas da Tabela 2.6 para encontrar a variância:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n x_i^2 \cdot f_i - (\sum_{i=1}^n x_i \cdot f_i)^2}{N} = \frac{234 - \frac{(43)^2}{8}}{8} = \frac{234 - 231,125}{8} = \frac{2,875}{8} = 0,36 \text{ ponto}^2$$

O desvio padrão é:

$$\sigma = \sqrt{0,36} = 0,60 \text{ ponto}$$

No Exemplo 2.9 utilizamos o desvio padrão para comparar as notas dos dois alunos, pois as médias são iguais e as variáveis em estudo são as mesmas (as notas). Agora, quando queremos comparar as variabilidades de dois conjuntos que apresentam médias bem diferentes e cujas variáveis em estudo são diferentes também, utilizamos uma medida de variabilidade denominada coeficiente de variação. Veremos, no próximo item, como calcular esta medida.

2.2.3 Coeficiente de variação

O coeficiente de variação (cv) é definido como o quociente entre o desvio-padrão e a média, e é frequentemente expresso em porcentagem. Ele mede o grau de variabilidade do conjunto de dados. Quando calculamos o desvio-padrão, obtemos um valor que pode ser grande ou pequeno, dependendo da variável em estudo. O fato de ele ser um valor considerado alto é relativo, pois dependendo da variável que está sendo estudada e da média, esta variação dos dados pode ser relativamente pequena. Então, o coeficiente de variação serve para calcular o grau de variação dos dados em relação à média aritmética. Além disto, serve também para comparar a variabilidade de conjuntos de dados cujas variáveis em estudo são diferentes, pois ele é adimensional. Obtemos esta medida por meio do seguinte cálculo:

$$cv = \frac{s}{\bar{x}} \times 100$$

onde s é o desvio-padrão e \bar{x} é a média aritmética.

Alguns autores consideram a seguinte regra empírica para a interpretação do coeficiente de variação:

- Baixa dispersão: C.V. " 15%
- Média: C.V." 15%-30%
- Alta: C.V. \geq 30%

2.2.4 Cálculos da variância e do desvio padrão para dados agrupados em intervalos de classes

O cálculo da variância e do desvio padrão para dados apresentados em tabelas com intervalos de classes é feito de maneira semelhante ao cálculo da média. Utilizamos o ponto médio do intervalo de classe para representar os valores dentro de cada classe. Então, acrescentamos três colunas na tabela x_i , $x_i \cdot f_i$ e $x_i^2 \cdot f_i$, que são necessários para o cálculo da variância.



EXEMPLO

2.10: A Tabela 2.7 apresenta as frequências de níveis séricos de colesterol para homens, de determinada cidade, entre 25 e 35 anos.

NÍVEL DE COLESTEROL (MG/100 ML)	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)
80 – 120	13	1,21
120 – 160	150	14,02
160 – 200	442	41,31
200 – 240	299	27,94
240 – 280	115	10,75
280 – 320	34	3,18
320 – 360	11	1,03
360 – 400	6	0,56
Total	1.070	100,00

Tabela 2.11 – Distribuição de frequências de níveis séricos de colesterol para homens, entre 25 e 35 anos.

Vamos encontrar a variância e o desvio padrão para os dados apresentados na Tabela 2.7.

Resolução

Acrescentando as três colunas adicionais para os cálculos, temos:

NÍVEL DE COLESTEROL (MG/100 ML)	FREQUÊNCIA	FREQUÊNCIA RELATIVA (%)	x_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
80 - 120	13	1,21	100	1.300	130.000
120 - 160	150	14,02	140	21.000	2.940.000
160 - 200	442	41,31	180	79.560	14.320.800
200 - 240	299	27,94	220	65.780	14.471.600
240 - 280	115	10,75	260	29.900	7.774.000
280 - 320	34	3,18	300	10.200	3.060.000
320 - 360	11	1,03	340	3.740	1.271.600
360 - 400	6	0,56	380	2.280	866.400
Total	1.070	100,00		213.760	44.834.400

Tabela 2.12 – Cálculos das colunas auxiliares para encontrar a variância e o desvio padrão.

Substituindo os valores na fórmula da variância, temos:

$$s^2 = \frac{\sum_{i=1}^n x_i^2 \cdot f_i - \frac{(\sum_{i=1}^n x_i \cdot f_i)^2}{n}}{n-1} = \frac{44.834.400 - \frac{(213.760)^2}{1.070}}{1.070-1}$$

$$= \frac{44.834.400 - 42.704.053,83}{1.069} = \frac{2.130.346,17}{1.069} = 1.992,84 \left(\frac{\text{mg}}{100\text{ml}} \right)^2$$

O desvio padrão é:

$$s = \sqrt{1.992,84} = 44,64 \frac{\text{mg}}{100\text{ml}}$$

Agora que já aprendemos os conceitos das medidas de tendência central e de dispersão, bem como efetuar os cálculos para encontrá-las, sabemos que a média e o desvio-padrão são influenciados pela presença de valores extremos no conjunto de dados, portanto, podem não ser adequados para representar o conjunto. Para contornarmos situações em que isto ocorre, podemos calcular outras medidas descritivas, que veremos a seguir. Estudaremos conceitos referentes às medidas separatrizes ou de ordenamento e à forma da distribuição dos dados.

2.3 Medidas separatrizes ou de ordenamento

As medidas separatrizes ou de ordenamento são: quartis, decis e percentis.

Os quartis (Q_1 , Q_2 e Q_3), como o próprio nome sugere, divide a distribuição dos dados ordenados em quatro partes, sendo, Q_1 o quartil que separa os 25% valores inferiores dos 75% superiores, Q_2 o que divide o conjunto ao meio (é a mediana) e Q_3 o que separa os 75% valores inferiores dos 25% superiores.

Os decis, por sua vez, dividem a distribuição dos dados em 10 partes ($D_i, i=1,2,\dots,9$) e os percentis dividem a distribuição em 100 partes ($P_i = 1,2,\dots,99$).

Não há um consenso universal sobre um procedimento único para o cálculo das medidas separatrizes, e diferentes calculadoras e softwares estatísticos podem produzir resultados ligeiramente diferentes.

2.3.1 Quartis

Como os quartis são medidas separatrizes precisamos, primeiramente, ordenar o conjunto de dados.

O primeiro quartil (Q_1) será a observação que ocupar a posição $\frac{n}{4}$. O segundo quartil (Q_2) será a observação que ocupar a posição $\frac{2n}{4}$ e o terceiro quartil (Q_3) será a observação que ocupar a posição $\frac{3n}{4}$. Quando fazemos estas divisões para encontrar as posições dos quartis, pode acontecer do resultado ser um número inteiro ou um número fracionário. Então, adotaremos a seguinte convenção:

- Se a divisão resultar num número fracionário, arredonde-o para cima e o valor do quartil será a observação encontrada nesta posição.
- Se a divisão for um número inteiro, o quartil será a média aritmética da observação que ocupar a posição encontrada com a observação que ocupar a posição imediatamente seguinte.



EXEMPLO

2.11 Abaixo estão listadas as medidas de entrada calórica diária, registradas em quilocalorias por quilograma, para uma amostra de adolescentes que sofrem de bulimia:

15,9	18,9	25,1	16,0	19,6	25,2	16,5	21,5	25,6	17,0
21,6	28,0	17,6	22,9	28,7	18,1	23,6	29,2	18,4	24,1
30,9	18,9	24,5	30,6						

Vamos encontrar primeiro, segundo e terceiro quartil.

Resolução

Para encontrarmos os quartis, precisamos ordenar o conjunto de dados. Então:

15,9	16,0	16,5	17,0	17,6	18,1	18,4	18,9	18,9	19,6
21,5	21,6	22,9	23,6	24,1	24,5	25,1	25,2	25,6	28,0
28,7	29,2	30,6	30,9						

- Posição do primeiro quartil (Q_1): $\frac{24}{4} = \frac{24}{4} = 6$.

Como a divisão resultou em um valor inteiro, o primeiro quartil será o resultado da média aritmética entre o valor que está na sexta posição e o valor que está sétima posição.

$$Q_1 = \frac{18,1 + 18,4}{2} = 18,25$$

Então, pelo menos 25% das observações são menores ou iguais a 18,25 quilocalorias por quilograma e, pelo menos, 75% das observações são maiores ou iguais a 18,25 quilocalorias por quilograma.

- Posição do segundo quartil (Q_2): $\frac{2 \times 4}{4} = \frac{2 \times 24}{4} = 12$

Como a divisão resultou em um valor inteiro, o segundo quartil será o resultado da média aritmética entre o valor que está na décima segunda posição e o valor que está na décima terceira posição.

$$Q_2 = \frac{21,6 + 22,9}{2} = 22,25$$

Temos que pelo menos 50% das observações são menores ou iguais a 22,25 quilocalorias por quilograma e pelo menos 50% das observações são maiores ou iguais a 22,25 quilocalorias por quilograma.

- Posição do terceiro quartil (Q_3): $\frac{3 \times n}{4} = \frac{3 \times 24}{4} = 18$

Como a divisão resultou em um valor inteiro, o terceiro quartil será o resultado da média aritmética entre o valor que está na décima oitava posição e o valor que está na décima nona posição.

$$Q_3 = \frac{25,2 + 25,6}{2} = 25,4$$

Neste conjunto de dados, pelo menos 25% das observações são maiores ou iguais a 25,4 quilocalorias por quilograma e pelo menos 75% das observações são menores ou iguais a 25,4 quilocalorias por quilograma.

Assim como a média, o desvio padrão não é uma medida de dispersão resistente. Para conjunto de dados com valores discrepantes, uma medida de dispersão alternativa ao desvio padrão é uma medida denominada amplitude interquartil, ou distância interquartil, definida como a diferença entre o terceiro e o primeiro quartil, ou seja, $D_q = Q_3 - Q_1$. No Exemplo 2.11, temos que a distância interquartil é $D_q = Q_3 - Q_1 = 25,4 - 18,25 = 7,15$. Note que a amplitude interquartil contém, aproximadamente, 50% das observações centrais.

2.3.2 Decis e Percentis

Para o cálculo dos decis e dos percentis seguiremos o mesmo procedimento que foi adotado para o cálculo dos quartis. O decil D_k será a observação que ocupar a posição P_k ; e o percentil $\frac{(k \times n)}{10}$, $k = 1, 2, \dots, 9$ será a observação que ocupar a posição $\frac{(k \times n)}{10}$, $k = 1, 2, \dots, 99$.



EXEMPLO

2.12: Uma pesquisa feita com 40 brasileiros com 16 anos e mais, durante 15 dias, teve como objetivo saber quantas horas por dia eles usavam a internet, de segunda a sexta-feira. Os dados obtidos foram:

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

Resolução

O decil D_6 será a observação que ocupar a posição $\frac{(6 \times 40)}{10} = 24$ no conjunto de dados ordenados.

Como a divisão resultou em um valor inteiro, o sexto decil será o resultado da média aritmética entre o valor que está na vigésima quarta posição e o valor que está na vigésima quinta posição.

$$D_6 = \frac{5,3 + 5,4}{2} = 5,35$$

Temos que pelo menos 60% das observações são menores ou iguais a 5,35 horas.

O percentil P_{87} será a observação que ocupar a posição $\frac{(87 \times 40)}{100} = 34,8$ no conjunto de dados ordenados.

Como a divisão resultou em um valor fracionário, vamos arredondar para 35. Portanto, o percentil P_{87} é o valor que está na trigésima quinta posição.

$$P_{87} = 6,7$$

Neste conjunto de dados, pelo menos 13% das observações são maiores ou iguais a 6,7 horas.

As medidas de ordenamento também podem ser calculadas para dados agrupados em intervalos de classes. Os cálculos são parecidos com aquele que utilizamos para calcular a mediana. Vamos estudá-los no próximo item.

2.3.3 Cálculo das medidas separatrizes para dados agrupados em intervalos de classes

Para calcularmos os quartis, decis e percentis para dados agrupados em intervalos de classes utilizamos uma única fórmula, que segue a ideia que foi descrita para o cálculo da mediana:

$$P_k = l_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right)$$

em que:

n: número total de observações da distribuição de frequências;

k: 1, 2, ..., 99;

l_{inf_k} : limite inferior da classe encontrada;

h_k : amplitude do intervalo;

Fa_{ant} : frequência acumulada anterior à da classe P_k ;

f_k : frequência absoluta da classe encontrada P_k .

Podemos utilizar esta fórmula geral, pois $Q_1 = P_{25}$, $Q_2 = P_{50}$ e $Q_3 = P_{75}$ e $D_1 = P_{10}$, $D_2 = P_{20}$, ..., $D_9 = P_{90}$.



EXEMPLO

2.13: Vamos utilizar os dados do Exemplo 2.6 para encontrar o Q_1 , D_3 e P_{85}

PESO (KG)	FREQUÊNCIA	FREQUÊNCIA ACUMULADA
40 - 45	8	8
45 - 50	25	33
50 - 55	50	83
55 - 60	40	123
60 - 65	20	143
Total	143	

- Primeiro Quartil (Q_1)

A primeira informação que precisamos é saber qual intervalo contém o primeiro quartil. Este intervalo está associado à frequência acumulada imediatamente superior à $\frac{k \cdot n}{100} = \frac{25 \times 143}{100} = 35,75$. O valor de K é igual a 25, pois $Q_1 = P_{25}$, ou seja, estamos calculando o vigésimo quinto percentil. Pelo Quadro 2.4, temos que o intervalo que contém o primeiro quartil é de 50 | - 55 (pois $f_a = 83$).

Após a identificação do intervalo, conseguimos identificar todos os valores exigidos na fórmula:

$$l_{inf_k} : 50$$

$$h_k : 55 - 50 = 5$$

$$f_k : 50$$

$$n : 143$$

$$Fa_{ant} : 33$$

k: 25 (o primeiro quartil é o vigésimo quinto percentil)

$$P_k = l_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right)$$

$$P_{25} = 50 + \frac{5}{50} \cdot \left(\frac{25 \cdot 143}{100} - 33 \right)$$

$$P_{25} = 50 + \frac{5}{50} \cdot (35,75 - 33)$$

$$P_{25} = 50 + \frac{5}{50} \cdot (2,75) = 50,275$$

Então, pelo menos 25% das observações são menores ou iguais a 50,275 kg.

- Terceiro Decil (D_3)

O intervalo que contém o terceiro decil está associado à frequência acumulada imediatamente superior à $\frac{k \cdot n}{100} = \frac{30 \cdot 143}{100} = 42,9$. Pelo Quadro 2.4, temos que o intervalo que contém o terceiro decil é de 50 | – 55 (pois $f_a = 83$). Então:

$$l_{inf_k} : 50$$

$$h_k : 55 - 50 = 5$$

$$f_k : 50$$

$$n : 143$$

$$Fa_{ant} : 33$$

k: 30 (o primeiro quartil é o vigésimo quinto percentil)

$$P_k = l_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right)$$

$$P_{30} = 50 + \frac{5}{50} \cdot \left(\frac{30 \cdot 143}{100} - 33 \right)$$

$$P_{30} = 50 + \frac{5}{50} \cdot (9,9) = 50,99$$

Pelo menos 30% das observações são menores ou iguais a 50,99 kg.

- Octogésimo quinto percentil (P_{85})

Como $\frac{k \cdot n}{100} = \frac{85 \times 143}{100} = 121,55$, temos que o intervalo que contém o octogésimo quinto percentil é de 55 | - 60 (pois $f_a = 123$). Então:

$$I_{inf_k} : 55$$

$$h_k : 60 - 55 = 5$$

$$f_k : 40$$

$$n : 143$$

$$Fa_{ant} : 83$$

$$k : 85$$

$$P_k = I_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right)$$

$$P_{85} = 55 + \frac{5}{40} \cdot \left(\frac{85 \cdot 143}{100} - 83 \right)$$

$$P_{85} = 55 + \frac{5}{40} \cdot (38,55) = 59,82$$

Por meio do P_{85} , observamos que pelo menos 15% das observações são maiores ou iguais a 59,82 kg.

Perceba que o 2º quartil, o 5º decil e o 50º percentil representam a própria mediana, ou seja, todas estas medidas separatrizes (Q_{2B} , D_{5B} e P_{50}), dividem a distribuição dos dados ao meio, deixando o mesmo número de dados em cada uma das partes.

Agora que já sabemos calcular e interpretar as medidas de dispersão e separatrizes, podemos utilizar estas informações para construir um gráfico denominado boxplot (diagrama de caixa). Este gráfico é construído utilizando os valores mínimo, máximo e os quartis. Estes valores são conhecidos como resumo dos cinco números. O boxplot informa, entre outras coisas, a posição, variabilidade e simetria dos dados. A posição central é dada pela mediana (Q_2) e a dispersão pela amplitude interquartil (d_q). Com as posições relativas de Q_2 , Q_1 , Q_2 e Q_3 , temos ideia da assimetria da distribuição. A Figura 2.3 ilustra um boxplot.

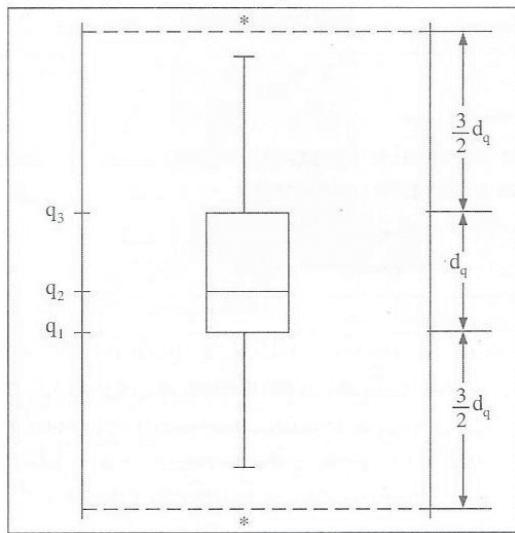


Figura 2.2 – Boxplot.Fonte: BUSSAB e MORETTIN (2002, p. 48).

De acordo com Bussab e Morettin (2002, p. 48)

Para construir este diagrama, consideremos um retângulo onde estão representados a mediana e os quartis. A partir do retângulo, para cima, segue uma linha até o ponto mais remoto que não excede $LS = q_3 + (1,5) d_q$, chamado limite superior. De modo similar, da parte inferior do retângulo, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que $LI = q_1 - (1,5) d_q$, chamado limite inferior. Os valores compreendidos entre esses dois limites são chamados valores adjacentes. As observações que estiverem acima do limite superior ou abaixo do limite inferior estabelecidos serão chamadas pontos exteioreiros e representadas por asteriscos. Essas são observações destoantes das demais e podem ou não ser o que chamamos de outliers ou valores atípicos.

Os boxplots são particularmente úteis quando temos interesse em comparar dois ou mais conjuntos de dados, especialmente quando são construídos na mesma escala. Vamos verificar sua importância através do exemplo a seguir.



EXEMPLO

2.14: Vamos utilizar os dados do Exemplo 2.9 para construir os boxplots associados a cada um dos alunos.

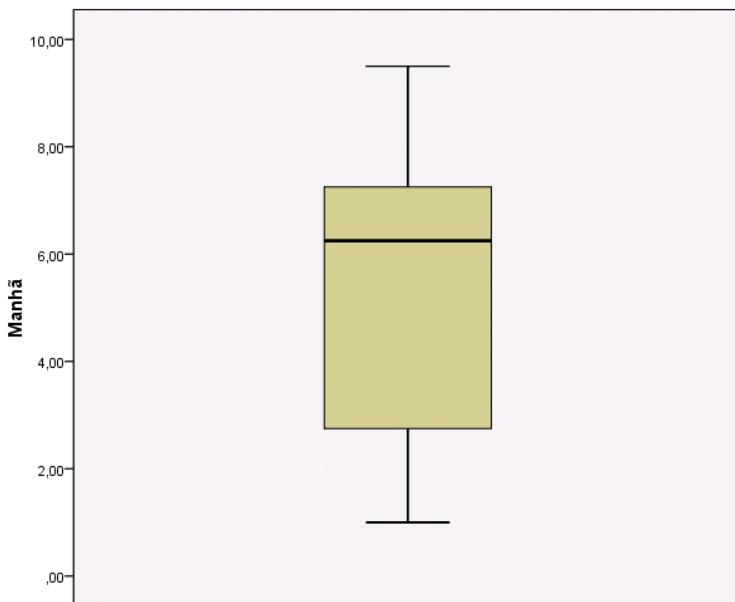


Figura 2.3 – Boxplot das notas dos dois alunos.

Pela análise gráfica, observamos que as duas distribuições são assimétricas (a distância da mediana para os quartis não é a mesma), o mesmo podendo ser observado a respeito da distância dos pontos mínimo e máximo em relação à mediana. Observamos, também, que as notas do aluno da manhã apresentam maior variabilidade (observando os valores utilizados na escala).

Sabemos que podemos identificar distribuições simétricas e assimétricas utilizando as medidas de posição e por meio da interpretação gráfica, analisando o histograma e o boxplot. Além disto, podemos calcular o grau de assimetria e o grau de achatamento ou alongamento de uma distribuição. Vamos aprender a fazer estes cálculos.

2.15: Em um estudo que investiga as causas de morte entre pessoas com asma severa, os dados foram registrados para dez pacientes que chegaram ao hospital em estado de parada respiratória e inconscientes. A Tabela 2.9 lista os batimentos cardíacos para os dez pacientes na internação do hospital. Vamos construir o boxplot para este conjunto de dados.

PACIENTE	BATIMENTO CARDÍACO
1	167
2	150
3	125
4	120
5	150
6	150
7	40
8	136
9	120
10	150

Tabela 2.13 – Batimentos cardíacos para dez pacientes asmáticos em estado de parada respiratória. Fonte: PAGANO; GAUVREAU (2004, p. 49).

Para a construção do boxplot, vamos seguir a descrição que está logo após a Figura 2.2. Precisaremos dos quartis, então vamos ordenar os dados:

40	120	120	125	136	150	150	150	150	167
----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Como $n = 10$ pacientes, e utilizando os conceitos adquiridos neste capítulo, temos:

$$Q_1 = 120$$

$$Q_2 = \frac{136 + 150}{2} = 143$$

$$Q_3 = 150$$

A distância interquartil é obtida por:

$$d_q = Q_3 - Q_1$$

$$d_q = 150 - 120 = 30$$

Agora, temos as informações necessárias para encontrar o limite superior (LS) e limite inferior (LI):

$$LS = Q_3 + (1,5) \cdot d_q$$

$$LS = 150 + (1,5) \cdot 30 = 195$$

e

$$LI = Q_1 - (1,5) \cdot d_q$$

$$LI = 120 - (1,5) \cdot 30 = 75$$

Então, com estas informações, obtemos o boxplot apresentado na Figura 2.4.

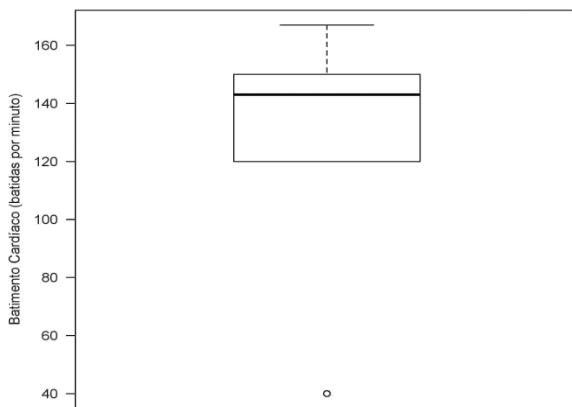


Figura 2.4 – Boxplot para os dados da Tabela 2.9.

Vamos interpretar os valores contidos na Figura 2.4:

- O retângulo é construído utilizando os quartis, ou seja, $Q_1 = 120$, $Q_2 = 143$ e $Q_3 = 150$.
- A partir do retângulo, para cima, segue uma linha até o ponto mais remoto que não excede $LS = Q_3 + (1,5) \cdot d_q$, ou seja, que não excede 195. O valor mais remoto que não excede 195 é 167.
- A partir do retângulo, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que $LI = Q_1 - (1,5) \cdot d_q$, ou seja, que não seja menor que 75. O valor mais remoto que não é menor que 75 é 120.
- As observações que estiverem acima do limite superior ou abaixo do limite inferior estabelecidos são chamados pontos exteriores e representados por asteriscos. Essas são observações destoantes das demais e podem ou não ser o que chamamos de outliers ou valores atípicos. Neste conjunto de dados temos uma observação destoante das demais, que é a resposta 40, representada no boxplot pelo asterisco.

2.4 Medidas de assimetria e curtose

Uma distribuição de frequência será simétrica se a metade esquerda de seu histograma é praticamente uma imagem espelhada de sua metade direita. Uma distribuição de frequência será assimétrica se “a cauda” do gráfico se prolongar mais de um lado do que do outro. Uma distribuição será assimétrica à esquerda (negativamente assimétrica) se a sua cauda se prolongar para a esquerda. Uma distribuição será assimétrica à direita (positivamente assimétrica) se a sua “cauda” se prolongar para a direita.

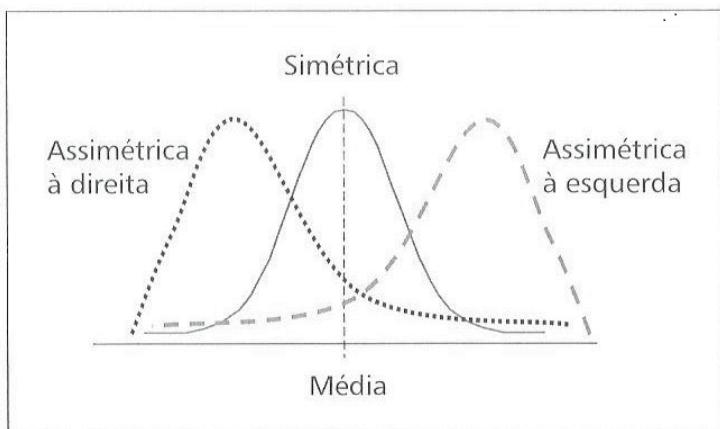


Figura 2.5 – Curvas simétricas e assimétricas. Fonte: BRUNI (2010, p. 85).

Uma das maneiras de se calcular o grau de assimetria de uma distribuição é por meio do segundo coeficiente de Pearson:

$$AS = \frac{Q_1 + Q_3 - 2 \cdot Q_2}{Q_3 - Q_1}$$

em que:

Q_1 : primeiro quartil

Q_2 : segundo quartil

Q_3 : terceiro quartil

Se:

1. $AS = 0$: distribuição simétrica (a média, a moda e a mediana são iguais)

2. $AS > 0$: distribuição assimétrica positiva ou assimétrica à direita (em geral, a média é maior que a mediana, que é maior que a moda).

3. $AS < 0$: distribuição assimétrica negativa ou assimétrica à esquerda (em geral, a média é menor que a mediana, que é menor que a moda).

Além do estudo da simetria da distribuição, podemos ter interesse em estudar o grau de achatamento ou alongamento da distribuição. De acordo com a análise das formas das distribuições, podemos classificá-las das seguintes maneiras:

- Platicúrtica
- Mesocúrtica
- Leptocúrtica

O grau de curtose pode ser medido por meio da seguinte fórmula:

$$k = \frac{Q_3 - Q_1}{2 \cdot (P_{90} - P_{10})}$$

em que:

Q_1 : primeiro quartil

Q_3 : terceiro quartil

P_{10} : décimo percentil

P_{90} : nonagésimo percentil

Dependendo do valor encontrado para o coeficiente de curtose, a distribuição será classificada da seguinte maneira:

1. $k = 0,263$: distribuição mesocúrtica, ou seja, nem chata nem delgada.
2. $k > 0,263$: distribuição leptocúrtica, ou seja, delgada.
3. $k < 0,263$: distribuição platicúrtica, ou seja, achatada.



EXEMPLO

2.16 Os dados abaixo representam as vendas ($\times 1\,000$ reais) de uma amostra de vendedores de produtos hospitalares de uma determinada empresa.

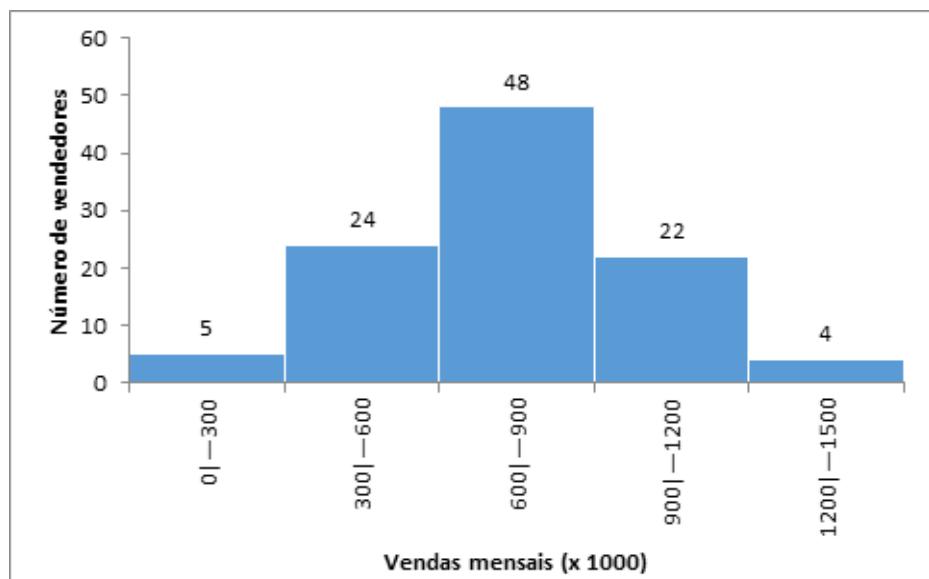


Figura 2.6 – Histograma para a variável vendas mensais de uma amostra de vendedores de produtos hospitalares de uma empresa.

Vamos calcular os coeficientes de assimetria e de curtose.

Resolução

Pela análise do histograma visualizamos uma distribuição aproximadamente simétrica, pois a metade à esquerda do histograma se comporta de maneira praticamente igual a metade à direita. Para calcular os coeficientes de assimetria e curtose, precisamos obter algumas medidas separatrizes. Vamos colocar as informações contidas no histograma em um quadro, para facilitar os cálculos, e aproveitamos para exercitar o cálculo das medidas separatrizes para dados agrupados em classes.

VENDAS MENSais ($\times 1\,000$)	NÚMERO DE VENDEDORES	FREQUÊNCIA ACUMULADA
0 —300	5	5
300 —600	24	29
600 —900	48	77
900 —1200	22	99
1 200 —1 500	4	103
Total	103	

Tabela 2.14 – Vendas mensais de vendedores do gênero alimentício.

- Primeiro Quartil

A primeira informação que precisamos é saber qual intervalo contém o primeiro quartil. Este intervalo está associado à frequência acumulada imediatamente superior à $\frac{k \cdot n}{100} = \frac{25 \cdot 103}{100} = 25,75$. O valor de k é igual a 25, pois $Q_1 = P_{25}$, ou seja, estamos calculando o vigésimo quinto percentil. Pelo Quadro 2.5, temos que o intervalo que contém o primeiro quartil é de 300 | -600 (pois $f_a = 29$).

Após a identificação do intervalo, conseguimos identificar todos os valores exigidos na fórmula:

$$I_{inf_k} : 300$$

$$h_k : 600 - 300 = 300$$

$$f_k : 24$$

$$n : 103$$

$$Fa_{ant} : 5$$

$$k : 25$$

$$\begin{aligned} P_k &= I_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right) \\ P_{25} &= 300 + \frac{300}{24} \cdot \left(\frac{25 \cdot 103}{100} - 5 \right) \\ P_{25} &= 300 + \frac{300}{24} \cdot (25,75 - 6) \\ P_{25} &= 300 + \frac{300}{24} \cdot (19,75) = 546,875 \end{aligned}$$

- Segundo Quartil

Como $\frac{k \cdot n}{100} = \frac{50 \cdot 103}{100} = 51,5$, temos que o intervalo que contém o segundo quartil é de 600 | -900 (pois $f_a = 77$). Então:

$$I_{inf_k} : 600$$

$$h_k : 900 - 600 = 300$$

$$f_k : 48$$

$$n : 103$$

$$Fa_{ant} : 29$$

$$k : 50$$

$$P_k = I_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right)$$

$$P_{50} = 600 + \frac{300}{48} \cdot \left(\frac{50 \cdot 103}{100} - 29 \right)$$

$$P_{50} = 600 + \frac{300}{48} \cdot (22,5) = 740,625$$

- Terceiro Quartil

Seguindo o mesmo procedimento utilizado para encontrar o intervalo que contém o primeiro quartil, temos que o intervalo que contém o terceiro quartil está associado à frequência acumulada imediatamente superior à $\frac{k \cdot n}{100} = \frac{75 \cdot 103}{100} = 77,25$. O valor de k é igual a 75, pois $O_3 = P_{75}$. Pelo Quadro 2.5, temos que o intervalo que contém o terceiro quartil é de 900 | -1 200 (pois $f_a = 99$).

Após a identificação do intervalo, conseguimos identificar todos os valores exigidos na fórmula:

$$I_{inf_k} : 900$$

$$h_k : 1200 - 900 = 300$$

$$f_k : 22$$

$$n : 103$$

$$Fa_{ant} : 77$$

$$k : 75$$

$$P_k = I_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right)$$

$$P_{75} = 900 + \frac{300}{22} \cdot \left(\frac{75 \cdot 103}{100} - 77 \right)$$

$$P_{75} = 900 + \frac{300}{22} \cdot (77,25 - 77)$$

$$P_{75} = 900 + \frac{300}{22} \cdot (0,25) = 900 + 3,409 = 903,409$$

- Décimo percentil (P_{10})

Como $\frac{k \cdot n}{100} = \frac{10 \cdot 103}{100} = 10,3$, temos que o intervalo que contém o décimo percentil é de 300 | - 600 (pois $f_a = 29$). Então:

$$I_{inf_k} : 600$$

$$h_k : 600 - 300 = 300$$

$$f_k : 24$$

$$n : 110$$

Fa_{ant} : 5

k: 10

$$P_k = l_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right)$$

$$P_{10} = 300 + \frac{300}{24} \cdot \left(\frac{10 \cdot 103}{100} - 5 \right)$$

$$P_{10} = 300 + \frac{300}{24} \cdot (5,3) = 366,25$$

- Nonagésimo percentil (P_{90})

Como $\frac{k \cdot n}{100} = \frac{90 \cdot 103}{100} = 92,7$, temos que o intervalo que contém o nonagésimo percentil é de 900 | - 1 200 (pois $f_a = 99$). Então:

l_{inf_k} : 900

h_k : 1200 - 900 = 300

f_k : 22

n: 103

Fa_{ant} : 77

k: 90

$$P_k = l_{inf_k} + \frac{h_k}{f_k} \cdot \left(\frac{k \cdot n}{100} - Fa_{ant} \right)$$

$$P_{90} = 900 + \frac{300}{22} \cdot \left(\frac{90 \cdot 103}{100} - 77 \right)$$

$$P_{90} = 900 + \frac{300}{22} \cdot (15,7) = 1114,09$$

Agora, substituímos os valores encontrados na fórmula do segundo coeficiente de Pearson:

$$AS = \frac{Q_1 + Q_3 - 2 \cdot Q_2}{Q_3 - Q_1}$$

$$AS = \frac{546,875 + 903,409 - 2 \cdot (740,625)}{903,409 - 546,875}$$

$$AS = \frac{-30,966}{356,534} = -0,0869$$

Apesar do AS > 0, o valor encontrado está bem próximo do zero, então, podemos considerar a distribuição aproximadamente simétrica, comprovando o que havíamos interpretado por meio do histograma.

Calculando o coeficiente de curtose:

$$k = \frac{Q_3 - Q_1}{2 \cdot (P_{90} - P_{10})}$$
$$k = \frac{903,409 - 546,875}{2 \cdot (1114,09 - 366,25)}$$
$$k = \frac{356,534}{2 \cdot (747,84)} = \frac{290,42}{1495,68} = 0,194$$

Como K = 0,263, temos que a distribuição é denominada platicúrtica (achatada).

2.5 Utilização do Microsoft Excel na Análise de Dados

A maioria das medidas apresentadas neste capítulo podem ser obtidas utilizando o Excel. Para isto, o suplemento Ferramenta de Análise deve estar ativo. Caso ele esteja ativo, deve aparecer o ícone Análise de Dados após clicar na aba Dados.

É muito comum este suplemento não aparecer ativo. Caso isto aconteça, devemos seguir o seguinte procedimento:

- Clicar no Botão Office e em seguida Opção do Excel. Escolher Suplementos e clicar;
- Escolher na lista Suplementos de Aplicativos Inativos a opção Ferramenta de Análise e clicar em Ir...
- Selecionar o seguinte suplemento disponível: Ferramenta de análise e clicar em OK.

Com o suplemento ativo, podemos fazer várias análises estatísticas!

Vamos utilizar os dados do Exemplo 2.7 para exemplificar como os cálculos são obtidos utilizando o Excel. Utilizaremos a versão Excel 2010.

1º passo: Digitar em uma planilha as respostas da(s) variável(eis).

	A	B	C	D	E	F	G
1	S.P.	B.H.					
2	3250	5250					
3	4125	5025					
4	5270	5270					
5	6029	5550					
6	9840	5870					
7	5127	5625					
8	6350	5120					
9	4250	5840					
10	7125	5720					
11	3850	5946					
12							

Figura 2.7 – Entrada dos dados.

2º passo: Neste passo, clicar em Dados e, em seguida, Análise de Dados. Aparecerá uma caixa de diálogo com uma lista de Ferramentas de análise. Clicar em Estatística descritiva e OK.

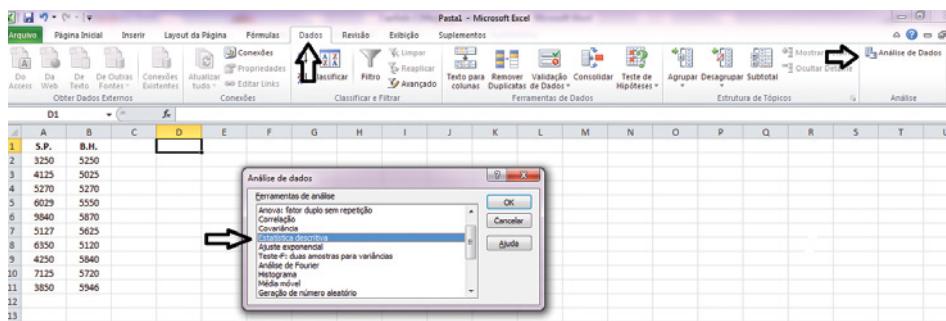


Figura 2.8 – Escolha da ferramenta de análise.

3º passo: Após clicar em Ok aparecerá uma nova caixa de diálogo. No campo Intervalo de entrada, selecionar os dados arrastando com o mouse desde A1 até B11. Marcar Rótulos na primeira linha (desde que os nomes das colunas tenham sido selecionados). Em Opções de saída, escolher Nova planilha (as estatísticas calculadas sairão em uma planilha diferente daquela que utilizamos para digitar a entrada dos dados, basta identificá-la no rodapé) e, por fim, escolher Resumo Estatístico e Ok.

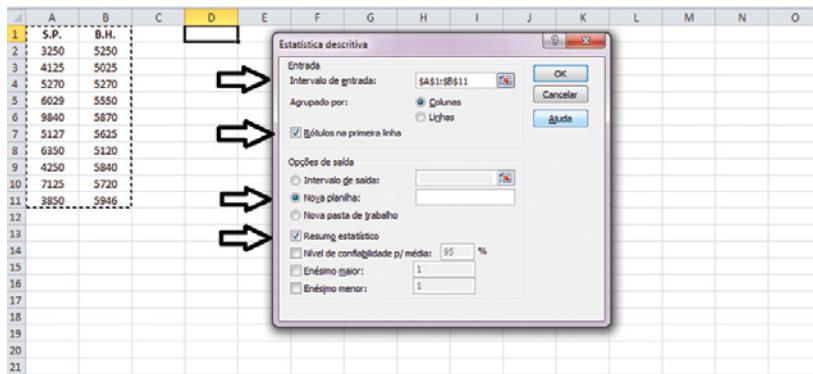


Figura 2.9 – Entrada das informações em Estatística descritiva.

4º passo: As informações obtidas estarão em uma nova planilha (rodapé da planilha). Todas as medidas que estão destacadas foram abordadas ao longo do capítulo, exceto Erro padrão.

The screenshot shows a Microsoft Excel spreadsheet titled 'Pasta1 - Microsoft Excel'. The ribbon tabs are visible at the top. The table below contains 15 rows of statistical measures. The first two rows are headers ('S.P.' and 'B.H.' in columns A and B respectively). Rows 3 to 15 list various statistical measures with their corresponding values. The table is formatted with alternating row colors.

	A	B	C	D	E	F	G	H
1	S.P.		B.H.					
2								
3	Média	5521,6	Média	5521,6				
4	Erro padrão	613,7684544	Erro padrão	105,3407381				
5	Mediana	5198,5	Mediana	5587,5				
6	Modo	#N/D	Modo	#N/D				
7	Desvio padrão	1940,906272	Desvio padrão	333,1166629				
8	Variância da amostra	3767117,156	Variância da amostra	110966,7111				
9	Curtose	1,747921869	Curtose	-1,584964468				
10	Assimetria	1,217678143	Assimetria	-0,237661059				
11	Intervalo	6590	Intervalo	921				
12	Mínimo	3250	Mínimo	5025				
13	Máximo	9840	Máximo	5946				
14	Soma	55216	Soma	55216				
15	Contagem	10	Contagem	10				

Figura 2.10 – Resumo estatístico dos salários de enfermeiros padrão nas cidades de São Paulo e Belo Horizonte.

Devemos observar que a palavra Amplitude é substituída por Intervalo e que Moda é escrita como Modo. Quando um conjunto de dados não apresenta moda, o resultado que aparece é #N/D. Vale ressaltar que o resumo estatístico do Excel não apresenta todas as respostas para a moda, caso o conjunto seja bimodal ou multimodal. No Exemplo 2.7 não calculamos as medidas de posição, a variância, o desvio padrão e os coeficientes de assimetria e curtose. Mas, sabemos como calculá-las. Encontre as medidas e compare com os resultados apresentados na Figura 2.9. Lembre-se, que há várias maneiras de se calcular as medidas separatrizes, então, os valores encontrados podem divergir daqueles encontrados pelo Excel!



REFLEXÃO

Neste capítulo aprendemos diversas medidas que são utilizadas para gerar informações estatísticas de conjuntos de dados quantitativos. Além de saber calculá-las, o mais importante é conseguir interpretar os resultados obtidos e identificar em quais situações uma medida pode ser mais representativa que outra. E, não podemos esquecer que o cálculo de uma medida resumo isoladamente pode não ser útil na comparação de dois ou mais conjuntos de dados, pois eles podem ter, por exemplo, mesma média, mas variabilidades completamente diferentes. Em situações como esta, uma análise mais completa necessita do cálculo do desvio padrão e do coeficiente de variação.



LEITURA

No endereço <http://m3.ime.unicamp.br/recursos/1315> você terá a oportunidade de ouvir dois módulos que exploram um problema envolvendo médias ponderadas e que ressalta o cuidado que devemos ter quando utilizamos a média como única informação.



REFERÊNCIAS BIBLIOGRÁFICAS

ARANGO, Héctor G. **Bioestatística Teórica e Computacional**. Rio de Janeiro: Editora Guanabara Koogan S.A., 2001.

BRUNI, Adriano L. **Estatística Aplicada à Gestão Empresarial**. 2. ed. São Paulo: Atlas, 2010.

BUSSAB, Wilton de O. ; MORETTIN, Pedro A. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2002.

PAGANO, Marcello.; GAUVREAU, Kimberlee. **Princípios de Bioestatística**. São Paulo: Pioneira Thomson Learning, 2004.

TRIOLA, Mário F. **Introdução à Estatística**. 10. ed. Rio de Janeiro: LTC, 2008.

VIEIRA, Sonia. **Estatística básica**. São Paulo: Cengage Learning, 2013.

VIEIRA, Sonia. **Introdução à Bioestatística**. 4 ed. Rio de Janeiro: Elsevier, 2008.

OLIVEIRA, Samuel R.; TEIXEIRA, Thiago; SANTOS, Joaé P. de Oliveira.

Disponível em: < <http://m3.ime.unicamp.br/recursos/1315> >. Acesso em: 20 jun. 2015.

3

Distribuição de Probabilidade Normal

Nos capítulos anteriores, tivemos como objetivo mostrar como organizamos e resumimos um conjunto de dados. Estudamos como construir distribuições de frequências e gráficos e como calcular e interpretar medidas de tendência central e variabilidade. Neste capítulo, estudaremos a distribuição mais importante na Estatística, que é a distribuição normal. Esta distribuição ocorre frequentemente em situações reais e desempenham papel importante nos métodos de inferência estatística, pois, muitos deles, exigem que os dados amostrais sejam provenientes de uma população que tenha distribuição que não se afaste drasticamente de uma distribuição normal.



OBJETIVOS

Esperamos que, através dos conhecimentos aprendidos neste capítulo, você seja capaz de:

- Compreender o conceito de variável aleatória contínua;
 - Compreender as características da curva normal, fazer a transformação de uma variável aleatória que tem distribuição normal em uma variável aleatória Z e encontrar probabilidades por meio da tabela da distribuição normal padrão.
-

3.1 Variável aleatória

Antes de começarmos a estudar a distribuição normal, precisamos esclarecer o conceito de variável aleatória.

Uma variável aleatória X representa um valor numérico associado a cada um dos resultados de um experimento aleatório.

Há dois tipos de variáveis aleatórias: as discretas e as contínuas.

As variáveis aleatórias discretas assumem valores em um conjunto enumerável e as variáveis aleatórias contínuas assumem valores em qualquer intervalo dos números reais.

Estas definições são similares àquelas apresentadas no Capítulo 1, com a diferença que agora aparece a palavra aleatória, para indicar que a cada possível valor da variável atribuímos uma probabilidade de ocorrência. Estudamos, também, que podemos representar graficamente dados contínuos, agrupados em intervalos de classes, por meio de histogramas. A análise deste gráfico nos auxilia na identificação da forma da distribuição dos dados, por exemplo, conseguimos identificar se a distribuição é simétrica e se apresenta forma de sino.

Como dissemos anteriormente, neste capítulo estudaremos a distribuição normal. Nesta distribuição, a variável em estudo é contínua, ou seja, pode assumir qualquer valor em um intervalo dos números reais e seu gráfico é simétrico e em forma de sino.

3.2 Distribuição Normal

A distribuição normal é uma distribuição contínua de probabilidade de uma variável aleatória X . Seu gráfico é chamado de curva normal.

Segundo LARSON (2004, p. 160)

A distribuição normal tem as seguintes propriedades:

1. A média, a mediana e a moda são iguais.
2. A curva normal tem formato de sino e é simétrica em torno da média.

3. A área total sob a curva normal é igual a 1.
4. A curva normal aproxima-se mais do eixo x à medida que se afasta da média em ambos os lados, mas nunca toca o eixo.

Dois parâmetros, μ e σ , determinam completamente o aspecto de uma curva normal. A média (μ) informa a localização do eixo de simetria e o desvio padrão (σ) descreve quanto os dados se espalham em torno da média.

A curva normal tem dois parâmetros, μ e σ . Eles determinam a posição e a forma da distribuição.

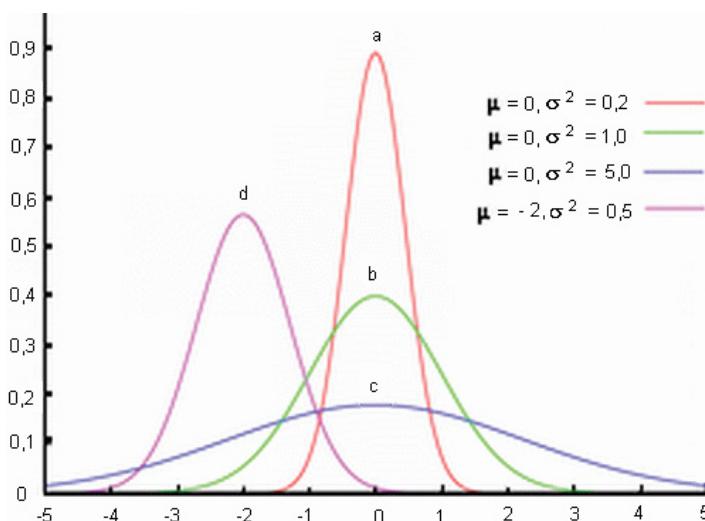


Figura 3.1 – Distribuições Normais N (μ, σ^2). Fonte: <http://www.cultura.ufpa.br/dicas/biome/bionor.htm>

As curvas normais a, b e c apresentam médias iguais (por isto estão localizadas na mesma posição no eixo x), mas apresentam desvios padrão diferentes (por isto a curva c, que apresenta maior desvio padrão, é mais achatada e a curva a, que apresenta menor desvio padrão, é mais fechada em torno da média).

A curva d apresenta média diferente das outras curvas, por isto está localizada numa posição diferente no eixo x.

A Figura 3.1 nos mostra que temos uma família de distribuições normais, diferenciadas por suas médias e desvios padrões.

Para obtermos a curva da distribuição normal, utilizamos a seguinte função densidade de probabilidade:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

em que $-\infty < x < \infty$. Valores específicos para μ e σ geram diferentes curvas, como as apresentadas na Figura 3.1. A maneira de fazer o gráfico é a mesma que utilizamos para qualquer função que relaciona x e y ou x e $f(x)$.

Como a área total sob a curva de densidade é igual a 1, existe uma correspondência entre área e probabilidade (TRIOLA, 2008, p. 196).

Quando utilizamos a função densidade de probabilidade da distribuição normal para fazer cálculos, percebemos que valores mais fáceis para μ e σ são $\mu = 0$ e $\sigma = 1$. Considerando estes valores para os parâmetros, matemáticos calcularam diferentes áreas sob a curva, que são apresentadas em uma tabela. Como existe uma correspondência entre área e probabilidade, utilizamos a tabela para encontrar probabilidades.

A distribuição normal cuja média é zero e variância 1 é chamada distribuição normal reduzida ou distribuição normal padronizada e é indicada pela letra Z.

De acordo com VIEIRA (2008, p. 213).

A distribuição normal reduzida tem grande importância:

1. As probabilidades associadas à distribuição normal reduzida são dadas em tabelas, o que torna fácil saber as probabilidades associadas a essa distribuição. Basta procurar na tabela.
2. Podemos transformar qualquer variável aleatória X com distribuição normal de média e desvio padrão conhecidos numa distribuição normal reduzida.
3. Dos itens 1 e 2 segue-se que qualquer probabilidade associada a X pode ser obtida transformando X (distribuição normal) em Z (distribuição normal reduzida).

A Figura 3.2 apresenta a curva de uma distribuição normal reduzida.

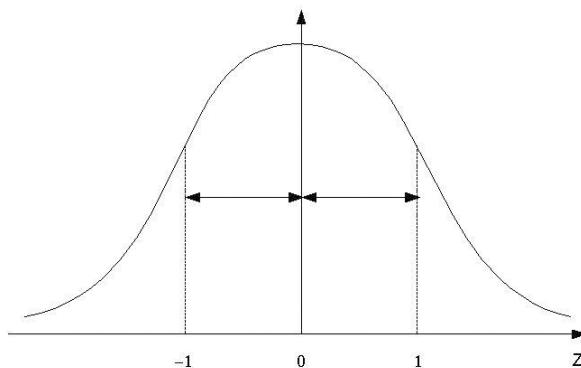


Figura 3.2 – Distribuição normal reduzida $Z \sim N(0,1)$.

Podemos transformar qualquer variável aleatória X com distribuição normal em Z (distribuição normal reduzida). Mas, como fazemos esta transformação?

Se $X \sim N(\mu, \sigma^2)$ então a variável aleatória definida por:

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

terá média zero e variância 1, ou seja, $Z \sim N(0,1)$.

A tabela fornecida no final do livro, utilizada nos cálculos das probabilidades, nos fornece $P(0 \leq Z \leq z_c) = P$, isto é,

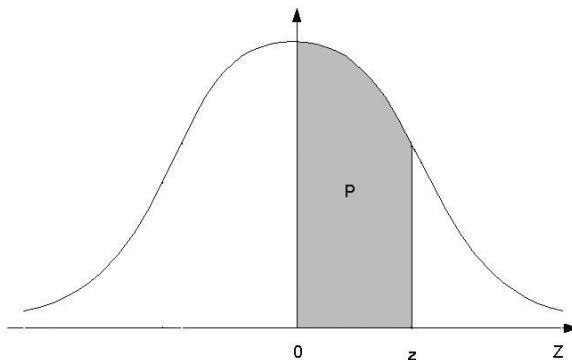


Figura 3.3 – Área correspondente à $P(0 \leq Z \leq z_c) = P$ fornecida pela tabela

A característica de simetria da distribuição normal implica em: $P(Z \geq 0) = 0,5 = P(Z \leq 0)$.

Vamos aprender a fazer a transformação e utilizar a tabela com o exemplo a seguir.



EXEMPLO

3.1: Seja $X \sim N(50, 25)$. Calcular:

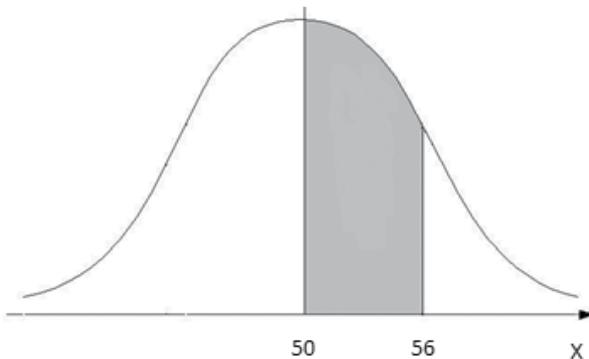
- f) $P(50 \leq x \leq 56)$
- g) $P(39 \leq x \leq 57)$
- h) $P(62 \leq x \leq 64)$
- i) $P(X \geq 58)$

Resolução

Primeiro, precisamos saber interpretar $X \sim N(50, 25)$. Lemos da seguinte maneira: a variável aleatória X tem distribuição normal com média 50 e variância 25. Como, precisamos do desvio padrão para utilizar na transformação, $\sigma = \sqrt{25} = \sqrt{25} = 5$.

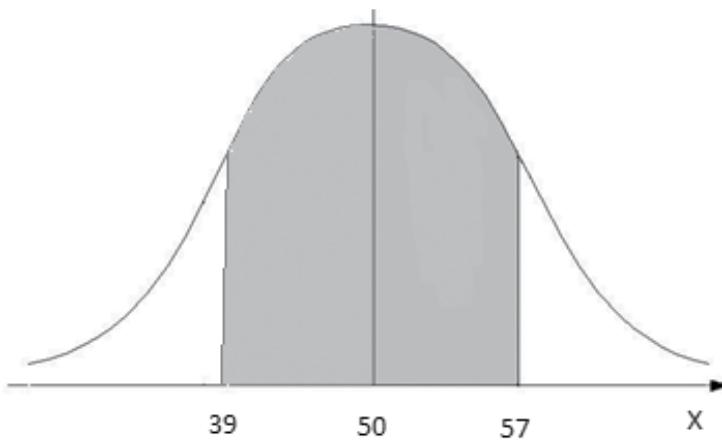
- a) $P(50 \leq x \leq 56)$

Agora, hachuramos a área do intervalo que queremos encontrar ($50 \leq x \leq 56$) na curva normal.



$$Z_1 = \frac{50 - 50}{5} = 0$$

$$Z_2 = \frac{56 - 50}{5} = \frac{6}{5} = 1,2$$



Apresentaremos, a seguir, uma parte da tabela que está no final do livro.



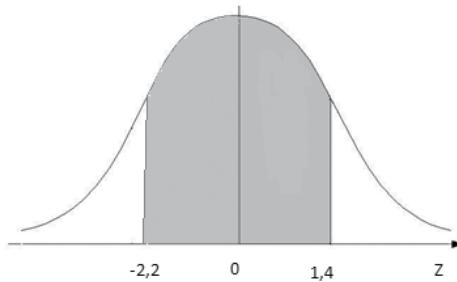
Curva Normal (p = área entre 0 e z)

z	segunda casa decimal									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4052	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319

Vamos aprender a encontrar a área (probabilidade) por meio da tabela. Na primeira coluna da esquerda (z) identificamos o número que obtemos na transformação com uma casa decimal e, a segunda casa decimal do número, está nas colunas (0 a 9). O número 1,2 é igual a 1,20, ou seja, a segunda casa decimal é 0. Vamos à linha 1,2 e na coluna 0. O número encontrado é 0,3849. Então:

$$P(50 \leq X \leq 56) = P(0 \leq Z \leq 1,2) = 0,3849$$

b) $P(39 \leq X \leq 57)$



Transformando para encontrar o novo intervalo correspondente à variável aleatória Z:

$$Z_1 = \frac{39 - 50}{5} = -2,2$$

$$Z_2 = \frac{57 - 50}{5} = 1,4$$

Observação: Devido à simetria, $P(-2,2 \leq Z \leq 0) = P(0 \leq Z \leq 2,2)$

Precisamos encontrar as áreas (probabilidades) hachuradas na tabela e somá-las.

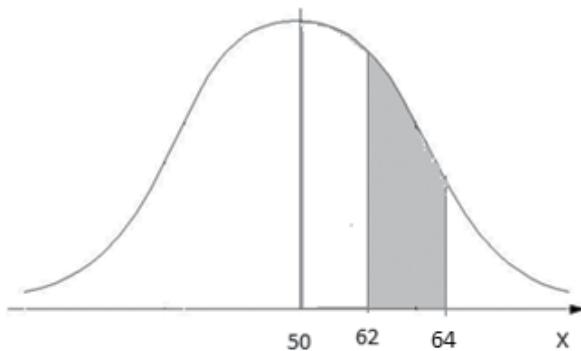
z	segunda casa decimal									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09

1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936

Então:

$$\begin{aligned} P(39 \leq X \leq 57) &= P(-2,2 \leq Z \leq 0) + P(0 \leq Z \leq 1,4) = 0,4861 + 0,4192 \\ &= 0,9053 \end{aligned}$$

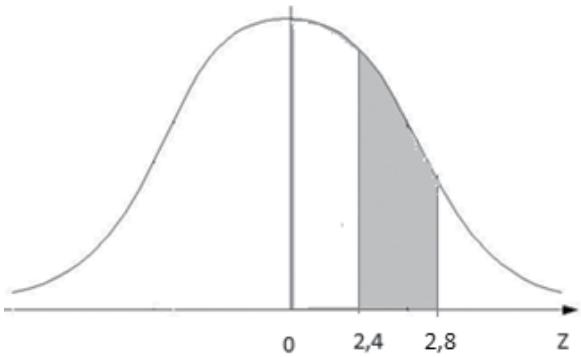
c) $P(62 \leq x \leq 64)$



Transformando:

$$Z_1 = \frac{62 - 50}{5} = -2,4$$

$$Z_2 = \frac{64 - 50}{5} = 2,8$$



O objetivo deste item é alertar para o fato que a tabela fornece a área do zero ao valor tabelado. A área hachurada neste item não corresponde à área fornecida diretamente na tabela. Então, como encontramos a área procurada? Se encontrarmos a área $0 \leq Z \leq 2,8$ e a área $0 \leq Z \leq 2,4$ (que são obtidas na tabela) e subtrairmos as duas áreas, encontraremos justamente a área hachurada!

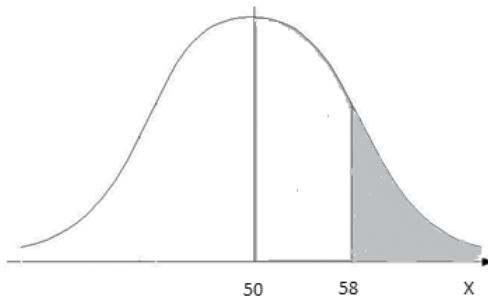
z	segunda casa decimal									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986

Portanto:

$$P(62 \leq X \leq 64) = P(2,4 \leq Z \leq 2,8) = P(0 \leq Z \leq 2,8) - P(0 \leq Z \leq 2,4)$$

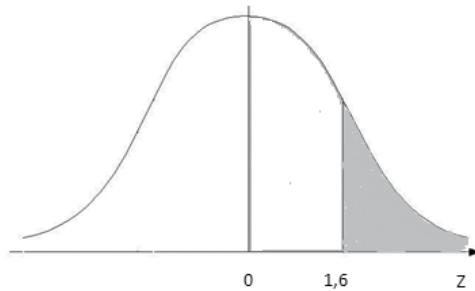
$$P(62 \leq X \leq 66) = 0,4974 - 0,4918 = 0,0056$$

d) $P(X \geq 58)$



Transformando:

$$Z = \frac{58 - 50}{5} = 1,6$$



Neste item, também temos que encontrar uma área que não é fornecida diretamente pela tabela. Como a área total sob a curva é 1 e a distribuição é simétrica, temos que $P(Z \leq 0) = P(Z \geq 0)$. Então:

z	segunda casa decimal									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09

1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706

$$P(X \geq 58) = P(Z \geq 1,6) = P(Z \geq 0) - P(0 \leq Z \leq 1,6) = 0,5 - 0,4452 = 0,0548$$

As probabilidades podem ser expressas das seguintes maneiras: frações, decimais ou percentuais. Neste livro, apresentaremos os resultados na forma decimal. Para expressarmos na forma percentual, basta multiplicar o valor decimal por 100.

3.2: A taxa de glicose no sangue humano é uma variável aleatória com distribuição normal de média $\mu = 100$ mg por 100 ml de sangue e desvio padrão $\sigma = 6$ mg por 100 ml de sangue. Calcule a probabilidade de um indivíduo apresentar taxa:

- a) Superior a 110 mg por 100 ml de sangue;
- b) Entre 90 e 100 mg por 100 ml de sangue.

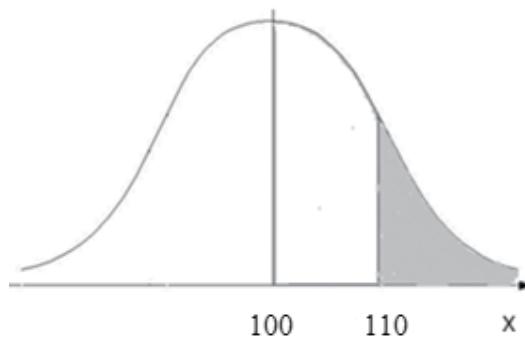
Fonte: VIEIRA (2008, p. 225).

Resolução

X: taxa de glicose no sangue humano

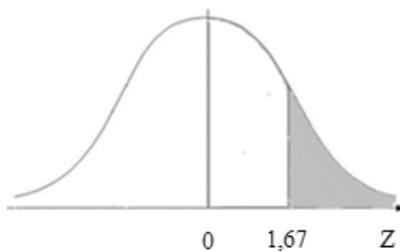
$X \sim N(100, (6^2))$

a) $P(X > 100)$



Para encontrar esta probabilidade, devemos transformar a variável X na variável normal reduzida Z:

$$Z = \frac{X - \mu}{\sigma} = \frac{110 - 100}{6} = \frac{10}{6} = 1,67$$



$$P(90 \leq X \leq 100) = P(-1,67 \leq Z \leq 0) = 0,4525$$

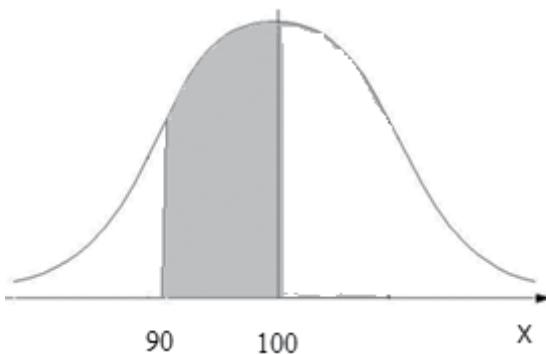
z	segunda casa decimal									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.3	0.4052	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

1.3	0.4052	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

Observação: Se quisermos apresentar o resultado obtido em forma de porcentagem, basta multiplicarmos o resultado obtido por 100, ou seja:

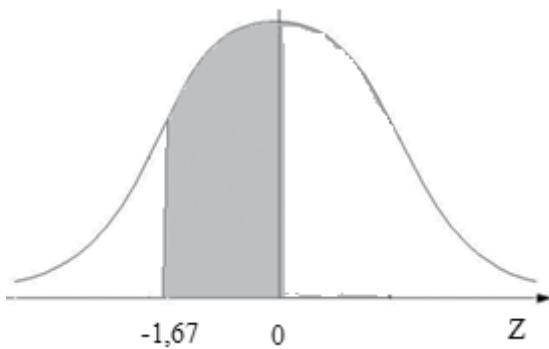
$$P(X > 110) = 0,0475 \times 100 = 4,75\%$$

b) $P(90 \leq X \leq 100)$



Transformando:

$$Z = \frac{X - \mu}{\sigma} \Leftrightarrow \frac{90 - 100}{6} = -\frac{10}{6} = -1,67$$



$$P(90 \leq X \leq 100) = P(-1,67 \leq Z \leq 0) = 0,4525$$

3.3: Uma fábrica de chocolate comercializa barras que pesam em média 200g. Os pesos são normalmente distribuídos. Sabe-se que o desvio-padrão é igual a 40g. Calcule a probabilidade de uma barra de chocolate escolhida ao acaso:

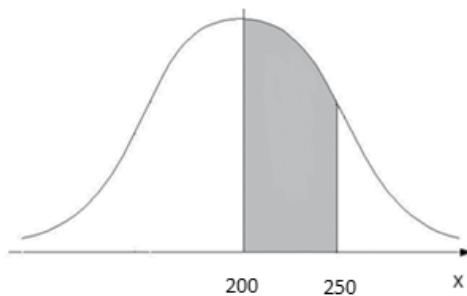
- a) pesar entre 200 e 250 g.
- b) pesar entre 170 e 200 g.
- c) pesar mais que 230 g.
- d) pesar menos que 150 g.

Resolução:

X: peso das barras de chocolate

$$X \sim N(200, 40^2)$$

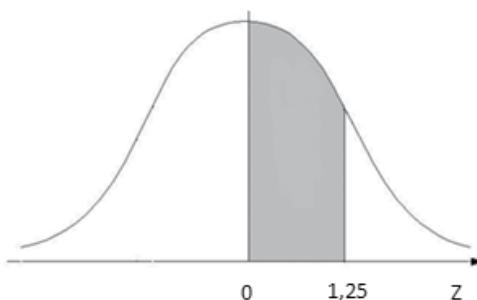
a) $P(200 \leq X \leq 250)$



Para encontrar esta probabilidade, devemos transformar a variável X na variável normal reduzida Z:

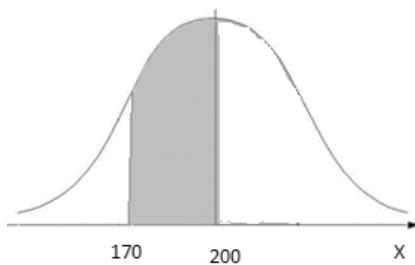
$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{200 - 200}{40} = 0$$

$$Z_2 = \frac{X_2 - \mu}{\sigma} = \frac{250 - 200}{40} = \frac{50}{40} = 1,25$$



$$P(170 \leq X \leq 200) = P(-0,75 \leq Z \leq 0) = P(0 \leq Z \leq 0,75) = 0,2734$$

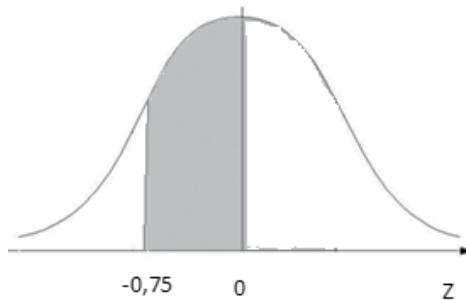
b) $P(170 \leq X \leq 200)$



Transformando:

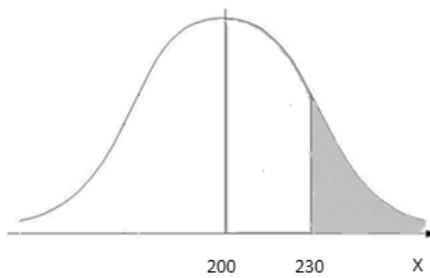
$$Z_1 = \frac{X_1 - \mu}{\hat{\sigma}} = \frac{170 - 200}{40} = \frac{-30}{40} = -0,75$$

$$Z_2 = \frac{X_2 - \mu}{\hat{\sigma}} = \frac{200 - 200}{40} = \frac{0}{40} = 0$$



$$P(170 \leq X \leq 200) = P(-0,75 \leq Z \leq 0) = P(0 \leq Z \leq 0,75) = 0,2734$$

c) $P(X \geq 200)$

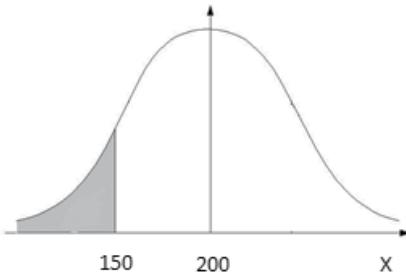


$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{230 - 200}{40} = \frac{30}{40} = 0,75$$



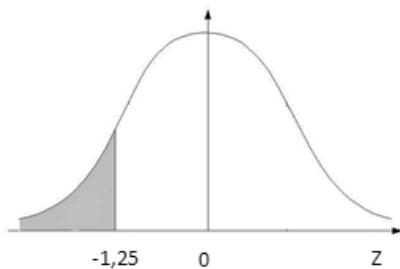
$$P(X \geq 230) = P(Z \geq 0,75) = 0,5 - P(0 \leq Z \leq 0,75) = 0,5 - 0,2734 = 0,2266$$

d) $P(X \leq 150)$



Transformando:

$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{150 - 200}{40} = -\frac{50}{40} = -1,25$$



$$P(X \leq 150) = P(Z \leq -1,25) = 0,5 - P(-1,25 \leq Z \leq 0) = 0,5 - 0,3944 = 0,1056$$

3.4: Uma clínica de emagrecimento recebe pacientes adultos com peso seguindo uma distribuição Normal com média 130 kg e desvio padrão 20 kg. Para efeito de determinar o tratamento mais adequado, os 25% pacientes de menor peso são classificado de “magros”, enquanto os 25% de maior peso de “obesos”. Determine os valores que delimitam cada uma dessas classificações.

Fonte: MAGALHÃES;LIMA (2004, p. 203).

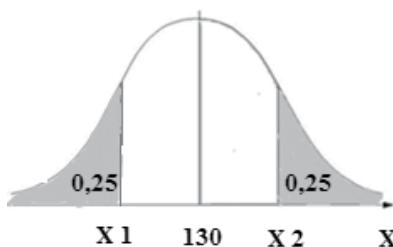
Resolução

Este exercício, diferentemente dos anteriores, fornece a área (probabilidade) e precisamos encontrar os valores críticos.

Do enunciado:

X : peso de paciente adultos

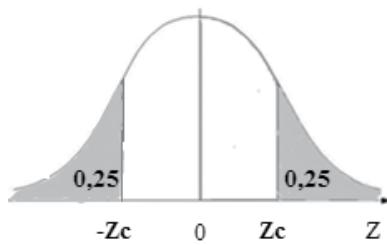
$$X \sim N(130, (20^2))$$



Construímos a curva normal desta maneira, pois o enunciado informa: os 25% pacientes de menor peso são classificado de “magros”, enquanto os 25% de maior peso, de “obesos”.

Encontrar os valores que delimitam cada uma destas classificações significa ter que encontrar X_1 e X_2 .

De acordo com a distribuição normal reduzida:



Sabemos que $P(Z \geq 0) = P(Z \leq 0) = 0,5$, então, $P(z_c \leq Z \leq 0) = 0,25$ e $P(0 \leq Z \leq z_c) = 0,25$

Portanto, temos que responder as seguintes perguntas:

- Qual o valor crítico ($-z_c$), tal que $P(z_c \leq Z \leq 0) = 0,25$?
- Qual o valor crítico (z_c), tal que $P(0 \leq Z \leq z_c) = 0,25$?

Para encontrarmos estes valores, precisamos encontrar a área = 0,25 dentro da tabela e verificar qual o valor crítico associado a esta área.



Curva Normal ($p = \text{área entre } 0 \text{ e } z$)



z	segunda casa decimal									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852

No corpo da tabela não há a área = 0,25, exatamente, então, selecionamos os dois valores mais próximos (0,2486 e 0,2517). O valor crítico associado a estas duas áreas é 0,678.

Utilizando a transformação da variável X na variável Z:

$$Z_1 = \frac{X_1 - \mu}{\sigma}$$

$$-0,678 = \frac{X_1 - 130}{20}$$

$$X_1 - 130 = -13,56$$

$$X_1 = -13,56 + 130$$

$$X_1 = 116,44 \text{ kg}$$

em que: $-z_c = Z_1$

$$Z_2 = \frac{X_2 - \mu}{\sigma}$$

$$0,678 = \frac{X_2 - 130}{20}$$

$$X_2 - 130 = 13,56$$

$$X_2 = 13,56 + 130$$

$$X_2 = 143,56 \text{ kg}$$

em que: $z_c = Z_2$

Os pacientes são classificados como “magros” se pesam até 116,44 kg e são classificados como “obesos” se pesam pelo menos 143,56 kg.

3.5: Nos dias atuais, dor crônica nas costas tornou-se frequente em crianças que carregam mochilas muito cheias e pesadas. As crianças têm o hábito de carregar livros escolares, notebooks, estojos, calculadoras, entre outros, tudo amontoado dentro da mochila, fazendo com que a chance de ocorrer algum espasmo muscular nos ombros e no pescoço e dor na coluna aumente. Uma pesquisa mostrou que o peso total carregado é diretamente proporcional ao volume da mochila. O volume de uma mochila vendida comercialmente segue uma distribuição normal com média 10 litros e desvio padrão 1,8 litros. Encontre um intervalo simétrico em torno da média, tal que 80% de todos os volumes de mochilas fiquem neste intervalo.

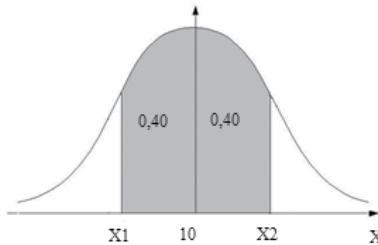
Resolução

Do enunciado:

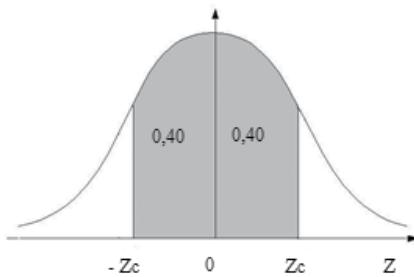
X : volume das mochilas

$X \sim N(10, (1,8^2))$

Encontrar um intervalo simétrico em torno da média tal que 80% de todos os volumes fiquem neste intervalo, significa encontrar X_1 e X_2 tal que:



De acordo com a distribuição normal reduzida:



Portanto, temos que responder as seguintes perguntas:

- Qual o valor crítico ($-z_c$), tal que $P(-z_c \leq Z \leq 0) = 0,40$?
- Qual o valor crítico (z_c), tal que $P(0 \leq Z \leq z_c) = 0,40$?

Para encontrarmos estes valores, precisamos encontrar a área = 0,40 dentro da tabela e verificar qual o valor crítico associado a esta área.



Curva Normal (p = área entre 0 e z)

z	segunda casa decimal									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015

No corpo da tabela não há a área = 0,40, exatamente, então, selecionamos os dois valores mais próximos (0,3997 e 0,4015). O valor crítico associado a estas duas áreas é 1,285.

Utilizando a transformação da variável X na variável Z:

$$Z_1 = \frac{X_1 - \mu}{\sigma}$$

$$-1,285 = \frac{X_1 - 10}{1,8}$$

$$X_1 - 10 = -2,313$$

$$X_1 = -2,313 + 10$$

$$X_1 = 7,687 \text{ litros}$$

em que: $-z_c = Z_1$

$$Z_2 = \frac{X_2 - \mu}{\sigma}$$

$$1,285 = \frac{X_2 - 10}{1,8}$$

$$X_2 - 10 = 2,313$$

$$X_2 = 2,313 + 10$$

$$X_2 = 12,313 \text{ litros}$$

em que: $z_c = Z_2$

Portanto, 80% das mochilas têm volume entre 7,687 e 12,313 litros.

3.3 Utilização do Microsoft Excel no cálculo de probabilidades normais

Para realizar os cálculos, seguimos estes procedimentos:

1º Passo: Clicar na aba Fórmulas e, em seguida, em Mais Funções. Selecionar Estatística e depois DIST.NORM.N.

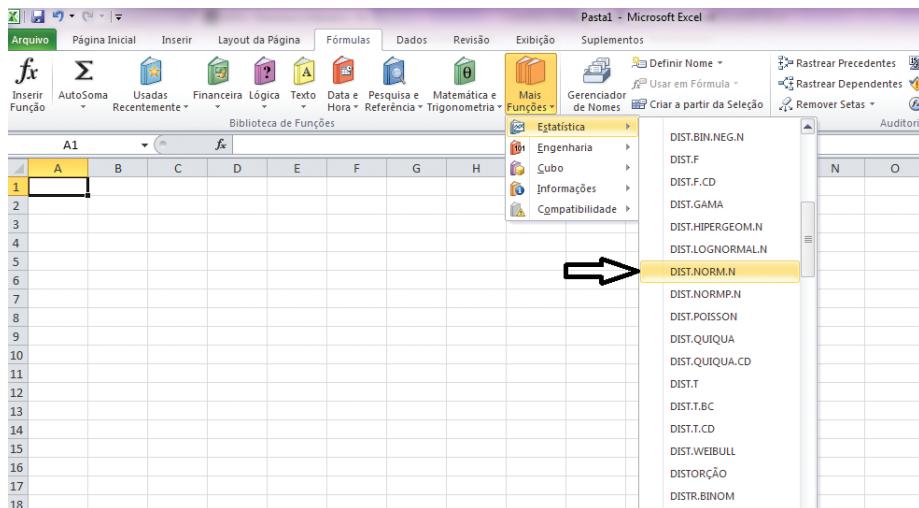


Figura 3.4 – Seleção da função Distribuição Normal.

2º Passo: Após clicar em DIST.NORM.N aparecerá uma janela onde teremos que colocar os argumentos da função. Vamos calcular as probabilidades no Excel utilizando os dados do Exemplo 3.3.

$$X \sim N(200, (402))$$

d) $P(200 \leq x \leq 250)$

X	250	= 250
Média	200	= 200
Desv_padrão	40	= 40
Cumulativo	VERDADEIRO	= VERDADEIRO
= 0,894350226		

Retorna a distribuição normal da média e do desvio padrão especificados.

Cumulativo é um valor lógico: para a função de distribuição cumulativa, use VERDADEIRO; para a função de densidade da probabilidade, use FALSO.

Resultado da fórmula = 0,894350226

Figura 3.5 – Preenchimento dos argumentos da função.

Da mesma maneira que fizemos no cálculo da distribuição binomial, vamos entender como devemos preencher cada uma das informações exigidas:

- X: é o valor cuja distribuição desejamos obter. No item a) queremos encontrar $P(200 \leq x \leq 250)$, portanto um dos valores é $X = 250$.
- Média: é a média aritmética da função. Neste exemplo, $\mu = 200$.
- Desv_padrão: é o desvio padrão da distribuição. No exemplo, $\sigma = 40$.
- Cumulativo: é um valor lógico: para a função de distribuição cumulativa, use VERDADEIRO. Para a função de densidade de probabilidade, use FALSO. Quando cumulativo = VERDADEIRO, a área calculada começa na cauda esquerda da curva normal até o x da fórmula indicada, ou seja, $P(X \leq x)$. Sempre utilizaremos VERDADEIRO.

Após o preenchimento, clicar em OK e aparecerá na planilha o resultado da probabilidade:

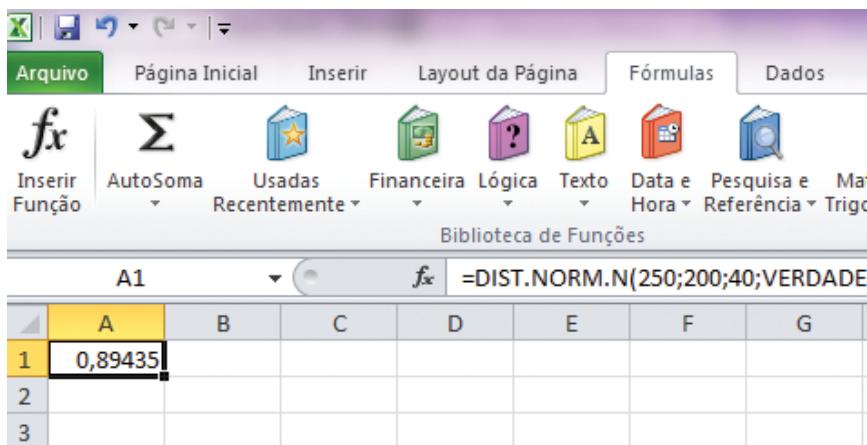


Figura 3.6 – Valor da probabilidade $P(X \leq 250)$.

Vamos interpretar o valor encontrado para a probabilidade: pela definição do argumento Cumulativo, a probabilidade encontrada é área correspondente com início na cauda esquerda da curva normal até 250. Ou seja:

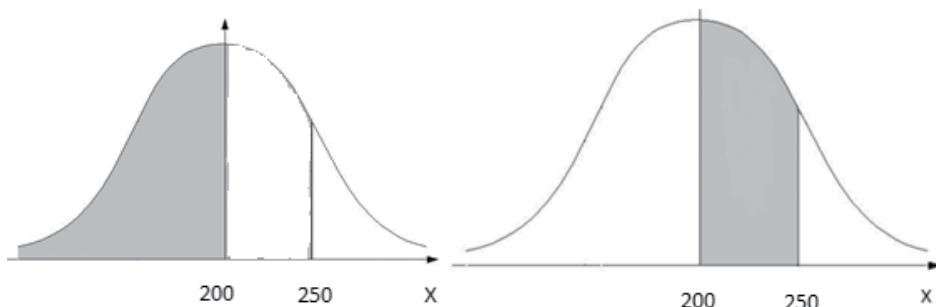
$$P(X \leq 250) = 0,89435$$

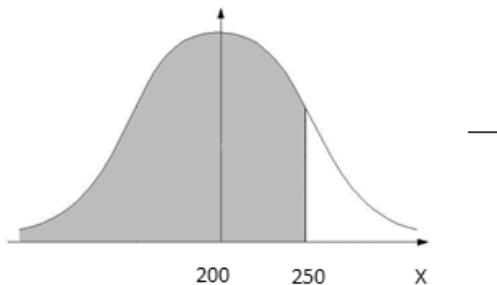
Podemos reescrever esta probabilidade da seguinte maneira:

$$P(X \leq 250) = P(X \leq 200) + P(200 \leq X \leq 250)$$

Queremos encontrar $P(200 \leq X \leq 250)$. Então:

$$P(200 \leq X \leq 250) = P(X \leq 250) - P(X \leq 200)$$





$$P(200 \leq X \leq 250) = 0,89435 - 0,5 = 0,39435$$

Obs.: Como a área total sob a curva é 1 e a distribuição é simétrica, temos que $P(X \leq 200) = P(X \geq 200) = 0,5$

e) $P(170 \leq x \leq 200)$

Para usar a probabilidade calculada pelo argumento Cumulativo, vamos reescrever a probabilidade pedida como:

$$P(170 \leq X \leq 200) = P(X \leq 200) - P(X \leq 170)$$

Já sabemos que $P(x \leq 200) = 0,5$ então, vamos calcular por meio do Excel $P(x \leq 170)$.

Argumentos da função	
DIST.NORM.N	
X	170
Média	200
Desv_padrão	40
Cumulativo	VERDADEIRO
= 0,226627352	
Retorna a distribuição normal da média e do desvio padrão especificados.	
Cumulativo é um valor lógico: para a função de distribuição cumulativa, use VERDADEIRO; para a função de densidade da probabilidade, use FALSO.	
Resultado da fórmula = 0,226627352	
Ajuda sobre esta função	

Figura 3.7 – Preenchimento dos argumentos da função.

Agora, vamos clicar em OK para encontrar o valor de P ($x \leq 170$).

The screenshot shows the Microsoft Excel interface. The ribbon at the top has tabs for Arquivo, Página Inicial, Inserir, Layout da Página, Fórmulas, and Dados. The Fórmulas tab is selected. Below the ribbon is the 'Biblioteca de Funções' (Function Library) with categories: AutoSoma (Sum), Usadas Recentemente (Recently Used), Financeira (Financial), Lógica (Logical), Texto (Text), Data e Hora (Date & Time), Pesquisa e Referência (Search & Reference), and Trig (Trigonometric). The formula bar shows the cell reference A1 and the formula =DIST.NORM.N(170;200;40;VERDADE). The main area shows a table with columns labeled A through G. Row 1 contains the formula result 0,226627 in cell A1. Row 2 is blank. Row 3 is also blank.

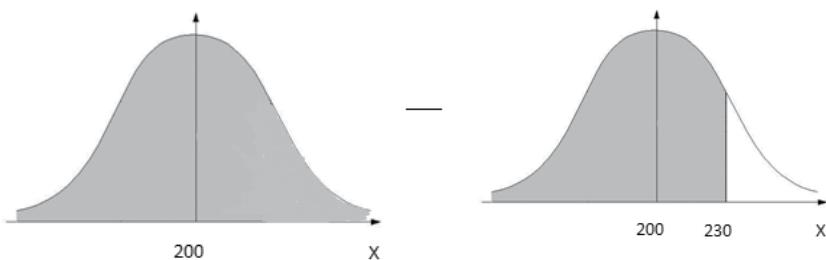
Figura 3.8 – Valor da probabilidade P ($x \leq 170$).

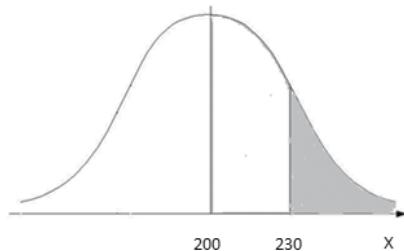
Então:

$$P(170 \leq X \leq 200) = P(X \leq 200) - P(X \leq 170)$$
$$P(170 \leq X \leq 200) = 0,5 - 0,226627 = 0,273373$$

f) $P(x \geq 230)$

Novamente, o argumento Cumulativo não fornece diretamente a probabilidade pedida. Então, reescrevendo:





Screenshot of Microsoft Excel showing the formula input process for the NORM.DIST function.

The formula bar shows: `=DIST.NORM.N(230;200;40;VERDADEIRO)`

The "Arguments da função" (Function Arguments) dialog box is open, displaying:

- DIST.NORM.N**
- X**: 230
- Média**: 200
- Desv._padrão**: 40
- Cumulativo**: VERDADEIRO

Below the dialog box, the formula `=DIST.NORM.N(230;200;40;VERDADEIRO)` is shown with its result: `= 0,773372648`.

Figura 3.9 – Preenchimento dos argumentos da função.

Screenshot of Microsoft Excel showing the final result of the formula execution.

The formula bar shows: `A1 =DIST.NORM.N(230;200;40;VERDADEIRO)`

The cell A1 contains the value: `0,773373`

Figura 3.10 – Valor de P ($x \geq 230$)

Então:

$$P(X \geq 230) = \text{area total} - P(X \leq 230)$$

$$P(X \geq 230) = 1 - 0,773373 = 0,226627$$

Obs.: O valor da probabilidade igual a 1 aparece, pois a área total sob a curva normal é 1.

g) $P(x \geq 150)$

Esta probabilidade é fornecida diretamente pelo argumento Cumulativo.

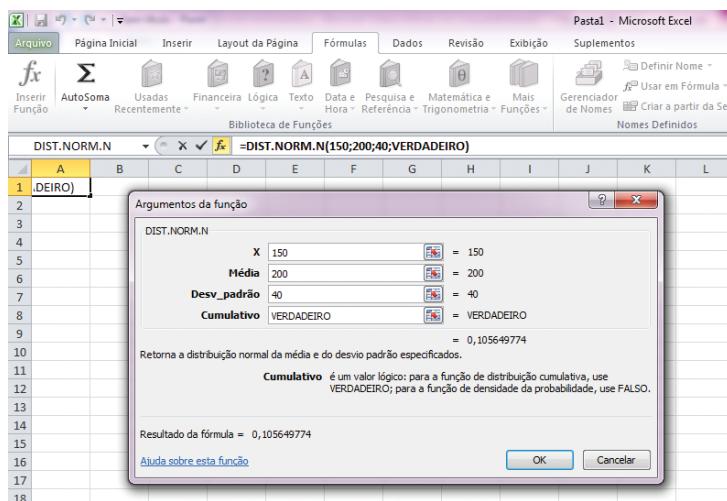


Figura 3.12 – Valor de $P(x \geq 150)$.

Portanto:

$$P(X \leq 150) = 0,10565$$

 **REFLEXÃO**

Durante todo este capítulo nos dedicamos a estudar a distribuição normal. Aprendemos que, para encontrar a probabilidade de uma variável aleatória que segue o modelo normal assumir determinados valores, precisamos utilizar a distribuição normal padrão (para encontrar probabilidades diretamente da tabela). Além de encontrar probabilidades, não podemos esquecer

quais as características da curva normal, pois esta distribuição é exigida em muitas técnicas da inferência estatística e, com isso, precisaremos saber identificar se os dados amostrais são provenientes de uma população normal.



LEITURA

No endereço <http://m3.ime.unicamp.br/recursos/1332> você encontrará comentários em dois áudios, primeiro módulo e segundo módulo, sobre a curva gaussiana (curva em forma de sino) e uma discussão envolvendo conceitos de média e mediana.



REFERÊNCIAS BIBLIOGRÁFICAS

- LARSON, Ron; FARBER, Betsy. **Estatística Aplicada**. 2. ed. São Paulo: Prentice Hall, 2004.
- LEVINE, David M.; BERENSON, Mark L.; STEPHAN, David. Estatística: Teoria e Aplicações Usando Microsoft Excel em Português. Rio de Janeiro: LTC, 2000.
- MAGALHÃES, Marcos N.; LIMA, Antonio C. P de. **Noções de Probabilidade e Estatística**. 6. ed. São Paulo: Editora da Universidade de São Paulo, 2004.
- TRIOLA, Mário F. **Introdução à Estatística**. 10. ed. Rio de Janeiro: LTC, 2008.
- VIEIRA, Sonia. **Estatística básica**. São Paulo: Cengage Learning, 2013.
- VIEIRA, Sonia. **Introdução à Bioestatística**. 4 ed. Rio de Janeiro: Elsevier, 2008.
- Disponível em: < <http://www.cultura.ufpa.br/dicas/biome/bionor.htm> >. Acesso em: 01 jun. 2015.
- NOIMAN, Caroline; OLIVEIRA, Samuel R.; SARTI, Luis R. Disponível em:
< <http://m3.ime.unicamp.br/recursos/1332> >. Acesso em: 01 jun. 2015.
-

4

Teste de Hipóteses

Estudamos, no Capítulo 1, que a Estatística pode ser dividida em duas grandes áreas: a estatística descritiva e a inferência estatística. Na inferência estatística (ou inferência indutiva), utilizamos dados amostrais para fazer estimativas, testar hipóteses e fazer previsões sobre características de uma população. Muitos pesquisadores sociais e da área da saúde trabalham com amostras, com o objetivo de generalizar os resultados obtidos para as populações de onde estas amostras foram retiradas. Por exemplo, pesquisadores da área médica utilizam testes de hipóteses para a tomada de decisões sobre novos medicamentos.

Ao longo deste capítulo, concentraremos nossos estudos em uma das técnicas da inferência estatística mais utilizada, que é o teste de hipóteses.



OBJETIVOS

Com a técnica estudada neste capítulo, esperamos que você seja capaz de:

- Compreender os fundamentos do teste de hipóteses;
 - Distinguir os erros do tipo I e do tipo II;
 - Realizar um teste de hipóteses para verificar a diferença entre duas médias populacionais, no caso de amostras dependentes;
 - Realizar um teste de hipóteses para verificar a diferença entre duas médias populacionais, no caso de amostras independentes.
-

4.1 Fundamentos do teste de hipóteses

Um teste de hipóteses é um procedimento padrão para se testar uma afirmativa sobre uma propriedade da população (TRIOLA, 2008, p. 306).

Por exemplo, com base em estudos anteriores, sabe-se que o efeito imunológico de determinada vacina se prolonga por mais de um ano em apenas 20% das pessoas que a tomam. Uma nova vacina foi desenvolvida para a mesma finalidade. É necessário testar se a nova vacina é melhor que a atual, ou seja, se a proporção de pessoas imunizadas após um ano é maior que 20%. Como a eficácia da vacina varia de pessoa para pessoa, precisamos utilizar algum método estatístico para chegarmos a uma conclusão sobre a eficácia desta nova vacina. Por meio de um teste de hipóteses, tomamos decisões em presença da variabilidade, ou seja, verificamos se estamos diante de uma diferença real ou de uma diferença devida simplesmente à flutuação aleatória ao processo.

A afirmativa sobre a propriedade da população (normalmente um parâmetro populacional) é chamada de hipótese estatística. Para testarmos uma hipótese estatística, devemos estabelecer um par de hipóteses, tal que uma delas representa uma afirmativa e a outra, o seu complemento. A hipótese que contém a afirmativa de igualdade é a hipótese nula (representada por H_0) e o complemento da hipótese nula é a hipótese alternativa (representada por H_1 ou H_a). Representamos a hipótese alternativa usando um destes símbolos: $< . >$ ou \neq .

Por exemplo, se uma afirmativa para a média populacional é que ela assume o valor k , alguns pares possíveis de hipótese nula e alternativa são:

$$\begin{cases} H_0 : \mu = k \\ H_1 : \mu > k \end{cases}$$

$$\begin{cases} H_0 : \mu = k \\ H_1 : \mu < k \end{cases}$$

$$\begin{cases} H_0 : \mu = k \\ H_1 : \mu \neq k \end{cases}$$

Segundo TRIOLA (2008, p. 309), “se você está fazendo um estudo e deseja usar um teste de hipóteses para apoiar sua afirmativa, esta deve ser escrita de modo a se tornar a hipótese alternativa (e deve ser expressa usando apenas os símbolos $< . >$ ou \neq). Ou seja, você não deve apoiar uma afirmativa de que um parâmetro seja igual a algum valor específico.



EXEMPLO

4.1: Identifique as hipóteses que estão sendo testadas em cada caso.

- Um fabricante afirma que sua vacina previne 85% dos casos de certa doença. Um grupo de médicos desconfia que a vacina não seja tão eficiente assim.
- Um fabricante de bateria para automóveis alega que a vida média de um determinado modelo é de 40 meses. Um proprietário de automóvel deseja testar essa afirmação.
- Uma empresa instalou um equipamento antipoluição sonora com o objetivo de manter o ruído médio abaixo de 65 decibéis. O sindicato decide testar se o equipamento está ou não cumprindo sua função.

Resolução

a)
$$\begin{cases} H_0 : p = 0,85 \\ H_1 : p < 0,85 \end{cases}$$

Indicamos a proporção populacional por p . O fabricante faz uma afirmação sobre o parâmetro populacional, ou seja, que a proporção de casos prevenidos pela vacina é de 85%. Como o grupo de médicos desconfia que a vacina não é tão eficiente assim (ou seja, que a proporção é menor que 85%), definimos a hipótese alternativa como $p > 0,85$.

b)
$$\begin{cases} H_0 : \mu = 40 \\ H_1 : \mu \neq 40 \end{cases}$$

A média populacional é representada por μ . Neste item, o proprietário deseja testar a afirmação do fabricante (que a vida média da bateria é de 40 meses), portanto, utilizamos o símbolo \neq na hipótese alternativa.

c)
$$\begin{cases} H_0 : \mu = 65 \\ H_1 : \mu < 65 \end{cases}$$

A empresa afirma que o equipamento instalado mantém o ruído médio abaixo de 65 decibéis. O sindicato deseja testar se o ruído médio está abaixo de 65 decibéis após a instalação do equipamento, portanto, utilizamos na hipótese alternativa o símbolo $<$.

Podemos realizar testes de hipóteses para a média, desvio padrão e proporção populacionais, mas, neste capítulo, focaremos nosso estudo em teste de hipóteses para a média.

Podemos realizar testes de hipóteses para a média, desvio padrão e proporção populacionais, mas, neste capítulo, focaremos nosso estudo em teste de hipóteses para a média.

4.2 Teste de hipóteses para a média populacional

Para a realização de um teste de hipóteses, além de estabelecermos as hipóteses nula e alternativa, precisamos seguir algumas etapas e, para isto, a compreensão dos seguintes conceitos são imprescindíveis: erros do tipo I e II, nível de significância, estatística de teste, região crítica, valor crítico e conclusão do teste baseado no método tradicional ou do valor P.

4.2.1 Tipos de erros, nível de significância e estatística de teste

Não podemos esquecer que, quando realizamos um teste de hipóteses, estamos utilizando dados amostrais e, por isto, devemos aceitar o fato de que a decisão de rejeitar ou não H_0 pode estar incorreta. A única maneira de se ter certeza de que H_0 é verdadeira ou falsa é testar toda a população e sabemos que isto é, muitas vezes, impossível. Então, quando realizamos um teste de hipóteses, dois erros podem ser cometidos:

1. Rejeitar a hipótese H_0 , quando tal hipótese é verdadeira, e
2. Não rejeitar a hipótese H_0 , quando ela deveria ser rejeitada.

Ao erro cometido em 1., denominamos erro do tipo I, enquanto que ao erro cometido em 2., denominamos erro do tipo II.

A Figura 4.1 resume os resultados possíveis na realização de um teste de hipóteses.

		SITUAÇÃO	
Decisão		H_0 é verdadeira	H_0 é falsa
	Rejeitar H_0	Erro do tipo I	Decisão correta
	Não rejeitar H_0	Decisão correta	Erro do tipo II

Figura 4.1 – Resultados possíveis na realização de um teste de hipóteses.

A probabilidade de cometermos o erro do tipo I é denotada por α e a probabilidade de cometermos o erro do tipo II é denotada por β . Desejamos que as probabilidades α e β sejam próximas de zero, mas a teoria nos mostra que,

à medida que diminuímos o erro do tipo I, a probabilidade de erro do tipo II tende a aumentar. Então, ao definir as hipóteses, o erro mais importante a ser evitado é o erro do tipo I. A probabilidade máxima permitida de ocorrer um erro do tipo I é denominada nível de significância. As escolhas comuns para α são 0,05; 0,01 e 0,10.

Após a identificação das hipóteses nula e alternativa e da especificação do nível de significância, utilizamos dados de uma amostra aleatória para calcular o valor da estatística de teste.

Segundo TRIOLA (2008, p. 310)

A estatística de teste é um valor usado para se tomar a decisão sobre a hipótese nula e é encontrada pela conversão da estatística amostral (como a proporção amostral \hat{p} ou a média amostral \bar{x} ou o desvio padrão s) em um escore (como z , t e x^2) com a suposição de que a hipótese nula seja verdadeira.

Utilizamos as seguintes estatísticas de teste para a média:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{ou} \quad t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Podemos observar que esta estatística de teste pode se basear na distribuição normal ou na distribuição t de Student. A utilização de uma estatística ou de outra depende de algumas condições que devem ser satisfeitas, que estudaremos a seguir.

A forma da distribuição t de Student é parecida com a da distribuição normal: tem média $t = 0$, como a distribuição normal padronizada, com média ; é simétrica, mas apresenta caudas mais alongadas, ou seja, maior variabilidade do que a normal. Quando aumentamos o tamanho da amostra, a distribuição t de Student tende para a distribuição normal.

A escolha da estatística de teste para a realização de um teste de hipóteses para a média populacional depende do conhecimento, ou não, do valor do desvio padrão populacional.

ESTATÍSTICA DE TESTE	CONDIÇÕES
$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	<ul style="list-style-type: none"> - A amostra é uma amostra aleatória simples. - O valor do desvio padrão populacional σ é conhecido. – Pelo menos uma das condições seguintes é verdadeira: a população é normalmente distribuída ou $n > 30$.
$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ O número de graus de liberdade (g.l.) é $n - 1$	<ul style="list-style-type: none"> - A amostra é uma amostra aleatória simples. - O valor do desvio padrão populacional σ não é conhecido. – Pelo menos uma das condições seguintes é verdadeira: a população é normalmente distribuída ou $n > 30$.¹
<p>Nota: Critérios para decidir se a população é ou não normalmente distribuída: A população não precisa ser exatamente normal, mas deve parecer simétrica de alguma forma e sem outliers. O teste t é robusto contra um afastamento da normalidade, ou seja, o teste funciona razoavelmente bem se o afastamento não for extremo.</p>	

Tabela 4.1 – Escolha da estatística de teste.

Depois que encontramos o valor da estatística de teste, passamos à etapa de decidir pela rejeição ou não da hipótese nula. Esta decisão é feita utilizando o método do valor p ou o método tradicional.

De acordo com VIEIRA (2008, p. 250), “o valor p diz quão provável seria obter uma amostra tal qual a que foi obtida, quando a hipótese nula é verdadeira”.

O valor p é o menor nível no qual H_0 pode ser rejeitado, ou seja, quando utilizamos o método do valor p a hipótese nula é rejeitada se $p \leq \alpha$. Quando utilizamos softwares estatísticos e o Excel para realizar um teste de hipóteses, os resultados obtidos informam o valor p.

Além do método do valor p podemos utilizar o método tradicional para decidir por rejeitar ou não a hipótese nula. Para utilizá-lo, precisamos das seguintes informações:

REGIÃO CRÍTICA (OU REGIÃO DE REJEIÇÃO)	conjunto de todos os valores da estatística de teste que nos fazem rejeitar a hipótese nula.
VALOR CRÍTICO	qualquer valor que separa a região crítica dos valores da estatística de teste que não levam à rejeição da hipótese nula. Para encontrarmos este valor, precisamos analisar a natureza da hipótese nula, a distribuição amostral (normal ou t de Student) e o nível de significância.

O que significa o valor crítico depender da natureza da hipótese nula?

Um teste de hipóteses pode ser bicaudal (ou bilateral), unilateral à esquerda (monocaudal esquerdo) ou unilateral à direita (monocaudal direito). A identificação de cada um destes tipos é feita por meio da hipótese alternativa.

Temos que:

- Se a hipótese alternativa H_1 contiver o símbolo $<$, o teste de hipóteses será um teste unilateral à esquerda, ou seja, a região crítica (e o valor crítico) estão na cauda esquerda sob a curva.
- Se a hipótese alternativa H_1 contiver o símbolo $>$, o teste de hipóteses será um teste unilateral à direita, ou seja, a região crítica (e o valor crítico) estão na cauda direita sob a curva.
- Se a hipótese alternativa H_1 contiver o símbolo \neq , o teste de hipóteses será um teste bilateral, ou seja, a região crítica (e os valores críticos) estão nas duas caudas sob a curva.

Estas informações são mais fáceis de serem visualizadas através da Figura 4.2.

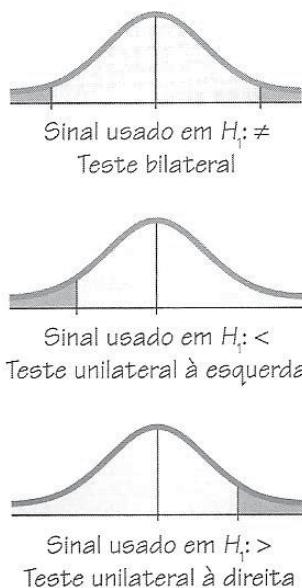


Figura 4.2 – Testes bilateral, unilateral à esquerda e unilateral à direita. Fonte: TRIOLA (2008, p. 313).

Quando estamos realizando testes bilaterais, devemos dividir igualmente o nível de significância α entre as duas caudas que constituem a região crítica. Por exemplo, em um teste bilateral com nível de significância $\alpha = 0,05$, há uma área de 0,025 em cada uma das caudas. Para testes que são unilaterais à esquerda ou à direita, a área da região crítica na cauda respectiva é α .

4.2.2 Decisão e interpretação

Para concluir um teste de hipóteses, precisamos tomar uma das seguintes decisões:

- Rejeitar a hipótese nula.
- Deixar de rejeitar a hipótese nula.

A decisão é feita usando um dos seguintes métodos estudados: método do valor P ou método tradicional. O método do valor p é muito utilizado quando a análise estatística está sendo feito através de algum software estatístico ou pelo Excel.

MÉTODOS	
Tradisional	Valor P
Rejeitar H_0 se a estatística de teste ficar dentro da região crítica.	Rejeitar H_0 se o valor $P \geq \alpha$.
Deixar de rejeitar H_0 se a estatística de teste não ficar dentro da região crítica.	Deixar de rejeitar H_0 se o valor $P > \alpha$.

Agora que já sabemos quais as etapas que devemos seguir para a realização de um teste de hipóteses, vamos apresentar, a seguir, um sumário.

1. Estabelecer as hipóteses nula e alternativa.
2. Especificar o nível de significância.
3. Calcular a estatística de teste, utilizando os dados amostrais.
4. Definir a forma da região crítica, com base na hipótese alternativa.
5. Concluir o teste com base no método tradicional ou no valor p.



EXEMPLO

4.2: Um laboratório farmacêutico lançou no mercado um novo medicamento contra dor de cabeça, retirando de circulação o antigo, com a justificativa que este novo medicamento tem ação mais rápida. O antigo medicamento tinha um tempo médio de 30 minutos para o início do efeito. Em uma amostra aleatória de 35 pessoas que tomaram o novo medicamento, obteve-se um tempo médio de 27 minutos, com desvio padrão de 4 minutos. Testar a eficácia do novo medicamento, ao nível de 5%.

Resolução

Neste estudo, temos uma amostra aleatória de 35 pessoas. Não conhecemos o desvio padrão populacional e o tamanho amostral é $n > 30$. Portanto, os requisitos necessários para a realização do teste de hipóteses para a média populacional com σ desconhecido estão satisfeitos.

Agora, seguiremos os passos necessários para a realização do teste:

1. Hipóteses:

$$\begin{cases} H_0: \mu = 30 \text{ (o novo medicamento não é mais eficaz que o antigo)} \\ H_1: \mu < 30 \text{ (o novo medicamento é mais eficaz que o antigo)} \end{cases}$$

2. O nível de significância é $\alpha = 5\%$

3. A estatística de teste é:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{27 - 30}{\frac{4}{\sqrt{35}}} = \frac{-3}{0,676123} = -4,4371$$

4. O número de grau de liberdade é $n - 1 = 35 - 1 = 34$

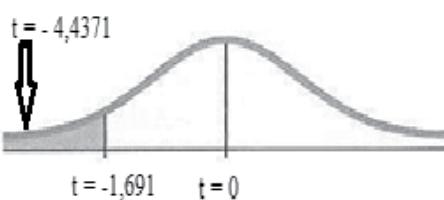
5. O valor crítico é:

$P(t \geq \text{VALOR TABELADO}) = \alpha \leftrightarrow \text{VALORES BILATERAIS}$									
G. L.	0.50	0.20	0.10	0.05	0.04	0.02	0.01	0.005	0.001
31	0.682	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.633
32	0.682	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.622
33	0.682	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.611
34	0.682	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.601

$P(t \text{ DE STUDENT} \geq \text{VALOR TABELADO}) = \alpha \leftrightarrow \text{VALORES BILATERAIS}$									
35	0.682	1.306	1.690	2.030	2.133	2.438	2.724	2.996	3.591
36	0.681	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.582
120	0.677	1.289	1.658	1.980	2.076	2.358	2.617	2.860	3.373
¥	0.674	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.290
	0,25	0,10	0,05	0,025	0,02	0,01	0,005	0,0025	0,0005

O valor crítico é $t_c = -1,691$.

6. Conclusão:



Como o teste é unilateral à esquerda (pois, H_1 contém o sinal $<$), o valor crítico é encontrado levando em conta o nível de significância que está na última linha da tabela. Por isto escolhemos a terceira coluna ($\alpha = 0,05$).

Rejeitamos H_0 se $t = t_c$. Como $-4,4371 < -1,691$, a estatística de teste está na área de rejeição. Portanto, rejeitamos H_0 , ou seja, os dados amostrais fornecem evidências suficientes para se concluir que o tempo médio de ação do novo medicamento é inferior ao tempo médio de ação do antigo medicamento.

4.3: Um experimento foi conduzido para estudar o nível médio de colesterol no sangue. Em uma amostra aleatória de 50 pacientes, a média amostral encontrada foi 268 mg/100 ml. Estudos anteriores nos informam que o desvio padrão populacional é $\sigma = 60$ mg/100ml. Teste a hipótese de que $\mu = 260$, contra a alternativa de que $\mu > 260$. Utilize um nível de 5%.

Resolução

Neste estudo, temos uma amostra aleatória de 50 pacientes. Conhecemos o desvio padrão populacional e o tamanho amostral é $n > 30$. Portanto, os requisitos necessários para a realização do teste de hipóteses para a média populacional com σ conhecido estão satisfeitos.

Agora, seguiremos os passos necessários para a realização do teste:

1. Hipóteses:

$$\begin{cases} H_0: \mu = 260 \\ H_1: \mu > 260 \end{cases}$$

2. O nível de significância é $\alpha = 0,05$.

3. A estatística de teste é:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{268 - 260}{\frac{60}{\sqrt{50}}} = \frac{8}{\frac{60}{7,0711}} = \frac{8}{8,485243} = 0,9428$$

4. O valor crítico é:

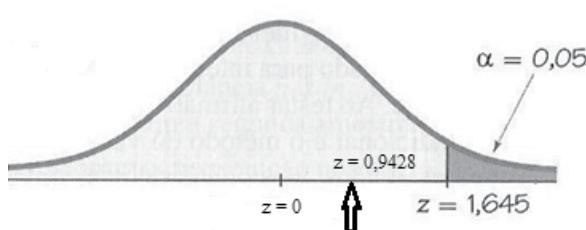


Curva Normal (p = área entre 0 e z)

		segunda casa decimal								
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

O valor crítico é $z_c = 1,645$.

5. Conclusão:



Como o teste é unilateral à direita (pois, H_1 contém o sinal $>$) e a área de $z = 0$ até o final da cauda direita é 0,5, temos que $0,5 - 0,05 = 0,45$. Devemos encontrar o valor 0,45 (ou valores próximos a este) no corpo da tabela. Portanto, o valor crítico é $z = 1,645$.

Rejeitamos H_0 se $z > z_c$. Como $0,9428 > 1,645$, a estatística de teste não está na área de rejeição. Portanto, não rejeitamos H_0 , ou seja, os dados amostrais não fornecem evidências suficientes para se concluir que o nível médio de colesterol é maior que 260 mg/100 ml.

Neste primeiro momento, apresentamos os procedimentos necessários para a realização de um teste de hipóteses para um parâmetro populacional (no caso, a média populacional). A seguir, utilizaremos o teste de hipóteses para comparar parâmetros de duas populações. Boa parte da teoria necessária ao desenvolvimento das técnicas que serão apresentadas já foi discutida no item 4.2.

4.3 Teste de hipóteses para duas amostras

A realização de um teste de hipóteses para duas amostras tem por objetivo testar uma afirmação comparando parâmetros de duas populações.

Nas mais diversas áreas de atuação profissional e de pesquisa, há uma busca contínua pelo desenvolvimento de novos métodos ou procedimentos que superem, ou melhorem, os já existentes. Por exemplo, a eficácia de um novo medicamento é testada por meio de dados amostrais, em que uma amostra utiliza o medicamento padrão e outra utiliza o novo medicamento. Por meio de um teste de hipóteses, verificamos a eficácia, ou não, do novo medicamento. Mas, por que a necessidade de dois conjuntos amostrais e da realização do teste de hipóteses? Se todos os pacientes se comportassem de maneira idêntica em relação ao tratamento utilizado, poderíamos examinar poucos deles com o novo medicamento e o medicamento padrão e a decisão seria obtida de maneira rápida e fácil, sem a necessidade de análise estatística. Porém, a reação de um tratamento varia de indivíduo para indivíduo e, na maioria dos casos, não há um tratamento ótimo para todos os pacientes. Então, para identificar o tratamento mais eficiente, o estudo é feito por meio de uma seleção de duas amostras e, por meio do teste de hipóteses, é feita a comparação dos resultados obtidos. Já estudamos, no item 4.1, que, por meio de um teste de hipóteses, tomamos decisões em presença da variabilidade.

Para a realização de um teste de hipóteses para duas amostras, extraímos uma amostra aleatória de cada uma das populações, usamos uma estatística de teste e estabelecemos uma conclusão (mesmo procedimento que utilizamos no caso de uma única amostra).

Realizaremos testes para comparação de duas médias. Sendo μ_1 e μ_2 os parâmetros populacionais, temos as possíveis hipóteses nula e alternativa:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

$$\begin{cases} H_0 : \mu_1 \leq \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$$

$$\begin{cases} H_0 : \mu_1 \geq \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

Também podemos escrever as hipóteses nula e alternativa da seguinte maneira:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

$$\begin{cases} H_0 : \mu_1 - \mu_2 \geq 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{cases}$$

4.3.1 Testes para diferenças entre médias

Quando utilizamos duas amostras, podemos nos deparar com as seguintes situações:

- Duas amostras independentes, com desvios padrões populacionais desconhecidos e diferentes;
- Duas amostras independentes, com desvios padrões populacionais desconhecidos e iguais;
- Duas amostras independentes, com desvios padrões populacionais conhecidos;
- Duas amostras dependentes.

As situações descritas acima nos informam que 2 amostras podem ser dependentes ou independentes. Qual a diferença entre estas classificações?

Duas amostras são independentes se os valores amostrais selecionados de uma população não estão relacionados com os valores amostrais selecionados da outra população. E, duas amostras são dependentes (ou emparelhadas) se os membros de uma amostra podem ser usados para se determinarem os membros da outra amostra.

Podemos identificar o uso de amostras independentes quando um grupo de pacientes é tratado com determinada droga para redução de colesterol, enquanto que outro grupo de pacientes é tratado com placebo. A independência ocorre, pois os pacientes tratados com a droga não estão de forma alguma relacionados com os pacientes tratados com placebo.

No caso de amostras dependentes, por exemplo, o peso de um grupo de pessoas é medido antes e após uma dieta. Cada par de medidas antes/depois se refere à mesma pessoa.

4.3.1.1 Amostras independentes com desvios padrões desconhecidos e diferentes

Como já verificamos na resolução de exemplos anteriores, antes da realização de qualquer teste de hipóteses, precisamos verificar se algumas condições referentes aos dados estão satisfeitas.

Segundo TRIOLA (2008, p. 372), os requisitos necessários para a realização deste teste são:

1. σ_1 e σ_2 são desconhecidos e não se faz qualquer suposição sobre igualdade de σ_1 e σ_2 .
2. As duas amostras são independentes.
3. Ambas as amostras são amostras aleatórias simples.
4. Uma, ou ambas, das seguintes condições é satisfeita: Os dois tamanhos amostrais são ambos grandes (com $n_1 > 30$ e $n_2 > 30$) ou ambas as amostras provêm de populações com distribuições normais (Para amostras pequenas, a exigência de normalidade é relaxada, no sentido de que os procedimentos funcionam bem, desde que não haja outliers e o afastamento da normalidade não seja extremo).

Seguimos os seguintes passos para a realização do teste de hipóteses:

1. Identificaremos H_0 e H_1 .
2. Especificaremos o nível de significância (α).
3. Determinaremos a estatística de teste:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

4. Determinaremos o número de graus de liberdade: menor de $n_1 - 1$ e $n_2 - 1$.
5. Determinaremos os valores críticos na Tabela 2 – Apêndice.
6. Conclusão:
 - Se t estiver na região de rejeição, rejeitamos H_0 . Caso contrário, não rejeitamos H_0 .

Neste livro, utilizaremos uma estimativa simples e conservadora para o número de graus de liberdade: o menor de $n_1 - 1$ e $n_2 - 1$. Os pacotes estatísticos, em geral, utilizam uma estimativa mais precisa, porém mais difícil de ser calculada, dada por:

$$g.l. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(s_1^2 \right)^2}{n_1} + \frac{\left(s_2^2 \right)^2}{n_2}}$$

$$\frac{n_1 - 1}{n_1 - 1} + \frac{n_2 - 1}{n_2 - 1}$$

Apesar dos dois métodos resultarem, geralmente, em números diferentes de graus de liberdade, a conclusão do teste raramente é afetada pela escolha.



EXEMPLO

4.4: Dois grupos de indivíduos participaram de um experimento planejado para testar o efeito da frustração sobre a agressividade. O grupo experimental de 35 indivíduos, escolhidos aleatoriamente, recebeu um quebra-cabeça frustrante para resolver, enquanto o grupo de controle de 35 indivíduos, escolhidos aleatoriamente, recebeu uma versão não frustrante do mesmo quebra-cabeça. Mediu-se, então, o nível de agressividade para ambos os grupos. Enquanto o grupo experimental (frustração) acusou um escore médio de agressividade $\bar{x}_1 = 5$ e um desvio padrão $S_1 = 2,4$, o grupo de controle (não frustração) teve um escore médio de agressividade $\bar{x}_2 = 3$ e um desvio padrão $S_2 = 1,5$ (escores médios mais altos indicam maior agressividade). Com esses resultados, teste a hipótese nula de que não há diferença quanto à agressividade entre as condições de frustração e não frustração. O que o resultado desse teste indica? Utilizar $\alpha = 0,05$.

Fonte: Adaptado (LEVIN, 2004, p. 259).

Resolução

Neste estudo, temos duas amostras independentes, que foram selecionadas aleatoriamente. Não conhecemos os desvios padrões das duas populações e não há suposições sobre a igualdade destes desvios. Os tamanhos amostrais são grandes (com $n_1 > 30$ e $n_2 > 30$). Portanto, os requisitos necessários para a realização do teste de hipóteses para amostras independentes com desvios padrões desconhecidos e diferentes estão satisfeitos.

GRUPO EXPERIMENTAL	GRUPO DE CONTROLE
$\bar{x}_1 = 5$	$\bar{x}_2 = 3$
$S_1^2 = (2,4)^2 = 5,76$	$S_2^2 = (1,5)^2 = 2,25$
$n_1 = 36$	$n_2 = 35$

Agora, seguiremos os passos necessários para a realização do teste:

- Hipóteses:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

- O nível de significância é $\alpha = 0,05$.

- A estatística de teste é:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5 - 3 - 0}{\sqrt{\frac{5,76}{35} + \frac{2,25}{35}}} = \frac{2}{\sqrt{0,228857143}} = \frac{2}{0,478390} = 4,1807$$

- O número de graus de liberdade é o menor entre n_1 e n_2 . Como os dois tamanhos amostrais são iguais, g.l. é $35 - 1 = 34$.

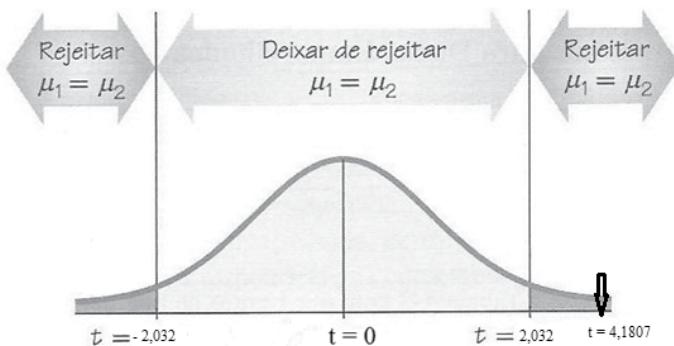
- Os valores críticos são:

P(t DE STUDENT ≥ VALOR TABELADO) = α ↔ VALORES BILATERAIS									
G. L.	0.50	0.20	0.10	0.05	0.04	0.02	0.01	0.005	0.001
29	0.683	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.660
30	0.683	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.646
31	0.682	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.633

P(t DE STUDENT ≥ VALOR TABELADO) = α ↔ VALORES BILATERAIS									
32	0.682	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.622
33	0.682	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.611
34	0.682	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.601
35	0.682	1.306	1.690	2.030	2.133	2.438	2.724	2.996	3.591
36	0.681	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.582

Os valores críticos são $t_c = -2,032$ e $t_c = 2,032$

6. Conclusão:



Como o teste é bilateral, rejeitamos H_0 se $t > t_c$ ou $t < -t_c$. Como $t = 4,1807$, a estatística de teste está na área de rejeição. Portanto, rejeitamos H_0 , ou seja, os dados amostrais fornecem evidências suficientes para apoiar a afirmativa de que há diferença quanto à agressividade entre as condições de frustração e não frustração.

Neste exemplo, consideramos desvios padrões desconhecidos e diferentes, que é o mais comum de acontecer. Caso os desvios padrões possam ser considerados iguais, o procedimento para a realização do teste muda. Abordaremos esta situação no próximo item.

4.3.1.2 Amostras independentes com desvios padrões desconhecidos e iguais

Quando as variâncias populacionais não forem conhecidas, mas for razoável supor que tenham o mesmo valor, ambas são utilizadas para se estimar σ^2 . A melhor maneira para combinar essas duas estimativas é formar uma média ponderada. O estimador resultante de σ^2 é:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Este valor é chamado estimador combinado de Image, pois combina as informações de ambas as amostras.

De acordo com TRIOLA (2008, p. 378), os requisitos necessários para a realização deste teste são:

1. Os dois desvios padrões populacionais não são conhecidos, mas supõe-se que sejam iguais, isto é, $\sigma^1 = \sigma^2$.
2. As duas amostras são independentes.
3. Ambas as amostras são amostras aleatórias simples.
4. Uma ou as duas condições seguintes são satisfeitas: Os dois tamanhos amostrais são ambos grandes (com $n_1 > 30$ e $n_2 > 30$) ou ambas as amostras provêm de populações com distribuições normais (Para pequenas amostras, a exigência de normalidade é relaxada, no sentido de que os procedimentos funcionam bem, desde que não haja outliers e os desvios da normalidade não sejam acentuados).

Os passos para a realização do teste de hipótese são:

1. Identificaremos H_0 e H_1 .
2. Especificaremos o nível de significância (α).
3. Determinaremos a estatística de teste:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

4. Determinaremos o número de graus de liberdade: $n_1 + n_2 - 1$.
5. Determinaremos os valores críticos na Tabela 2 – Apêndice.
6. Conclusão:
 - Se t estiver na região de rejeição, rejeitamos H_0 . Caso contrário, não rejeitamos H_0 .

Para usar este teste, precisamos verificar que os desvios padrões das duas amostras são iguais. Uma maneira é usar um teste preliminar de $\sigma^1 = \sigma^2$. De acordo com TRIOLA (2003), alguns autores ressaltam que dificilmente sabemos que $\sigma^1 = \sigma^2$. Eles analisam o desempenho de diferentes testes, considerando

tamanhos amostrais e poderes dos testes e concluem que o esforço deve ser empregado em aprender o método descrito no item 4.3.1.1 (desvios padrões desconhecidos e diferentes).

A menos que algum problema e/ou exercício já forneça alguma informação sobre desvios padrões desconhecidos e iguais, vamos tratá-los como diferentes e usar o método descrito no item 4.3.1.1.



EXEMPLO

4.5: Um estudo foi conduzido para determinar se a fumaça de cigarro de uma gestante tem algum efeito no conteúdo mineral ósseo da criança por ela gerada, sob outros aspectos saudáveis. Uma amostra aleatória de 77 recém-nascidos cujas mães fumaram durante a gravidez tem um conteúdo mineral médio ósseo de $\bar{x}_1 = 0,098$ g/cm e desvio padrão $S_1 = 0,026$ g/cm; uma amostra aleatória de 161 bebês cujas mães não fumavam tem média $\bar{x}_2 = 0,095$ g/cm e desvio padrão $S_2 = 0,025$ g/cm. Assuma que as variâncias das populações originais sejam iguais. Estabeleça as hipóteses nula e alternativa para o teste bilateral e conduza o teste ao nível de significância 0,05. O que podemos concluir?

Fonte: PAGANO (2004, p. 250).

Resolução

Temos duas amostras independentes, que foram selecionadas aleatoriamente. Os tamanhos amostrais são grandes (com $n_1 > 30$ e $n_2 > 30$) e os desvios padrões são desconhecidos, mas, supostamente iguais (o enunciado informa que devemos assumir que as variâncias das populações são iguais). Portanto, os requisitos necessários para a realização do teste de hipótese para amostras independentes com desvios padrões desconhecidos e iguais estão satisfeitos.

AMOSTRAS	N	\bar{x}	S
Mães que fumaram durante a gravidez	77	0,098	0,026
Mães que não fumaram durante a gravidez	161	0,095	0,025

Vamos à realização do teste:

1. Hipóteses:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

De acordo com o enunciado, o teste é bilateral.

2. O nível de significância é $\alpha = 0,05$.
 3. A estatística de teste é:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

em que:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(77 - 1)(0,026)^2 + (161 - 1)(0,025)^2}{77 + 161 - 2} = \frac{0,051376 + 0,10}{236}$$

$$= \frac{0,151376}{236} = 0,0006414$$

e

$$s_p = \sqrt{0,0006414} = 0,02533$$

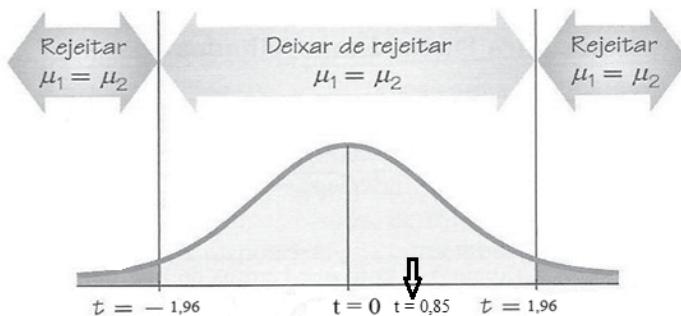
Então:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0,098 - 0,095}{0,02533 \sqrt{\frac{1}{77} + \frac{1}{161}}} = \frac{0,003}{0,02533 \cdot 0,138558} = \frac{0,003}{0,0035096} = 0,8548$$

4. O número de graus de liberdade é $n_1 + n_2 - 2 = 236$.
 5. Os valores críticos são:

P(T DE STUDENT ≥ VALOR TABELADO) = $\alpha \leftrightarrow$ VALORES BILATERAIS									
110	0,677	1,289	1,659	1,982	2,078	2,361	2,621	2,865	3,381
120	0,677	1,289	1,658	1,980	2,076	2,358	2,617	2,860	3,373
∞	0,674	1,282	1,645	1,960	2,054	2,326	2,576	2,807	3,290
	0,25	0,10	0,05	0,025	0,02	0,01	0,005	0,0025	0,0005

6. Conclusão



Como o teste é bilateral, rejeitamos H_0 se $t < t_c$ ou $t > t_c$. Como $0,85 > 1,96$, a estatística de teste não está na área de rejeição. Portanto, não rejeitamos H_0 , ou seja, os dados amostrais não fornecem evidências suficientes para apoiar a afirmativa de que a fumaça de cigarro de uma gestante tem algum efeito no conteúdo mineral ósseo da criança gerada.

4.3.1.3 Amostras independentes com desvios padrões conhecidos

Como dito anteriormente, os desvios padrões populacionais σ_1 e σ_2 raramente são conhecidos, mas, se forem, a estatística de teste baseia-se na distribuição normal. Como nos casos anteriores, para a realização do teste, temos que verificar alguns requisitos.

De acordo com TRIOLA (2008, p. 378)

1. Os dois desvios padrões populacionais são ambos conhecidos.
2. As duas amostras são independentes.
3. Ambas as amostras são amostras aleatórias simples.
4. Uma ou as duas condições seguintes são satisfeitas: Os dois tamanhos amostrais são ambos grandes (com $n_1 > 30$ e $n_2 > 30$) ou ambas as amostras provêm de popula

ções com distribuições normais (Para pequenas amostras, a exigência de normalidade é relaxada, no sentido de que os procedimentos funcionam bem, desde que não haja outliers e os desvios da normalidade não sejam acentuados).

Novamente, utilizaremos os seguintes passos para a realização do teste:

1. Identificaremos H_0 e H_1 .
2. Especificaremos o nível de significância (α).
3. Determinaremos a estatística de teste:

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

4. Determinaremos os valores críticos na Tabela 1 – Apêndice.
5. Conclusão:
 - Se z estiver na região de rejeição, rejeitamos H_0 . Caso contrário, não rejeitamos H_0 .

As situações descritas para amostras independentes podem ser visualizadas na Figura 4.3.

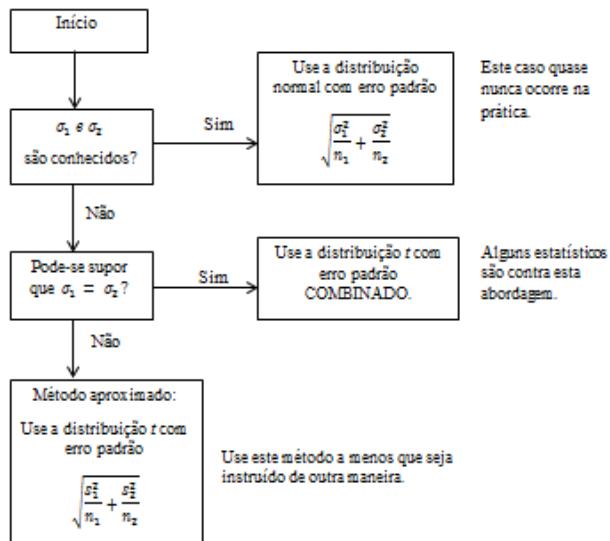


Figura 4.3 – Métodos para inferência sobre duas médias independentes. Fonte: TRIOLA (2003).

Agora, estudaremos o caso em que as amostras são dependentes.

4.3.1.4 Amostras dependentes

Neste caso, desejamos comparar duas médias populacionais sendo que, para cada unidade amostral, realizamos duas medições da característica de interesse. No geral, estas medições são tomadas antes e após uma dada intervenção. Voltando ao exemplo já citado sobre o peso de um grupo de pessoas. A medição é feita antes e após uma dieta e cada par de medidas antes/depois se refere à mesma pessoa.

No caso de amostras dependentes, também precisamos verificar alguns requisitos para a realização do teste.

Segundo TRIOLA (2008, p. 384)

1. Os dados amostrais consistem em dados emparelhados.
2. As amostras são amostras aleatórias simples.
3. Uma, ou ambas, das seguintes condições são satisfeitas: O número de pares de dados é grande ($n > 30$) ou os pares têm diferenças que são provenientes de uma população com distribuição aproximadamente normal. (Se houver um afastamento radical de uma distribuição normal, não devemos usar os métodos deste item, mas devemos usar métodos não paramétricos).

Passos para a realização do teste:

1. Identificaremos H_0 e H_1 .
2. Especificaremos o nível de significância (α).
3. Determinaremos a estatística de teste:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

4. Determinaremos o número de graus de liberdade: $n - 1$.
5. Determinaremos os valores críticos na Tabela 2 – Apêndice.
6. Conclusão:
 - Se t estiver na região de rejeição, rejeitamos H_0 . Caso contrário, não rejeitamos H_0 .

Vamos compreender a notação utilizada na estatística de teste:

d : diferença individual entre os dois valores em um único par.

μ_d : valor médio das diferenças d para a população de todos os pares.

\bar{d} : valor médio das diferenças d para dados amostrais emparelhados.

s_d : desvio padrão das diferenças d para os dados amostrais emparelhados.

n : número de pares de dados.



EXEMPLO

4.6: Um estudo foi realizado com o objetivo de investigar a eficácia de uma dieta de emagrecimento. O quadro a seguir apresenta os pesos, em kg, de 10 pessoas selecionadas aleatoriamente. Os pesos foram registrados antes e depois da dieta. Vamos supor que os dados são provenientes de uma população normalmente distribuída. Use um nível de significância de 5% para testar a afirmativa que a dieta de emagrecimento é eficaz na redução do peso.

Antes	77	61	60	80	90	75	85	58	89	67
Depois	80	57	60	74	87	68	90	50	82	63

Resolução

Temos um estudo com amostras dependentes (ou emparelhadas), pois cada par de medidas antes/depois se refere à mesma pessoa.

Avaliando os requisitos necessários para a realização do teste, temos: os dados amostrais são emparelhados, a amostra é aleatória simples e é proveniente de uma distribuição normal. Então, podemos realizar o teste de acordo a avaliação das informações do enunciado.

Realizando os passos do teste, temos:

1. Hipóteses:

$$\begin{cases} H_0: \mu_d = 0 \\ H_a: \mu_d < 0 \end{cases}$$

2. O nível de significância é $\alpha = 0,05$.
3. A estatística de teste é:

Para encontrar o valor da estatística, precisamos encontrar o valor médio das diferenças e o desvio padrão das diferenças. Vamos acrescentar algumas colunas no Quadro 4.5, para facilitar os cálculos.

Pesos (kg)				
Indivíduo	Depois	Antes	Diferença (d) Depois - Antes	d^2
1	80	77	$80 - 77 = 3$	$(3)^2 = 9$
2	57	61	$57 - 61 = -4$	$(-4)^2 = 16$
3	60	60	$60 - 60 = 0$	$(0)^2 = 0$
4	74	80	$74 - 80 = -6$	$(-6)^2 = 36$
5	87	90	$87 - 90 = -3$	$(-3)^2 = 9$
6	68	75	$68 - 75 = -7$	$(-7)^2 = 49$
7	90	85	$90 - 85 = 5$	$(5)^2 = 25$
8	50	58	$50 - 58 = -8$	$(-8)^2 = 64$
9	82	89	$82 - 89 = -7$	$(-7)^2 = 49$
10	63	67	$63 - 67 = -4$	$(-4)^2 = 16$
Total			-31	273

Tabela 4.2 – Cálculos auxiliares no cálculo da média e do desvio padrão das diferenças.

Então, a média amostral é:

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{-31}{10} = -3,1$$

e a variância amostral é:

$$s^2 = \frac{\sum_{i=1}^n d_i^2 - \frac{(\sum_{i=1}^n d_i)^2}{n}}{n-1} = \frac{273 - \frac{(-31)^2}{10}}{10-1} = \frac{273 - 96,1}{9} = \frac{176,9}{9} = 19,66$$

Portanto, o desvio padrão amostral é:

$$s = \sqrt{19,66} = 4,43$$

Substituindo os valores encontrados, temos:

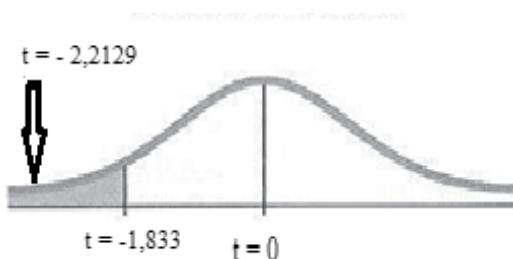
$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{-3,1 - 0}{\frac{4,43}{\sqrt{10}}} = \frac{-3,1}{1,4009} = -2,2129$$

- O número de graus de liberdade (g.l.) é $n - 1 = 10 - 1 = 9$.
- O valor crítico é:

P(t DE STUDENT ≥ VALOR TABELADO) = $\alpha \leftrightarrow$ VALORES BILATERAIS									
G. L.	0.50	0.20	0.10	0.05	0.04	0.02	0.01	0.005	0.001
1	1.000	3.078	6.314	12.706	15.894	31.821	63.656	127.321	636.578
2	0.816	1.886	2.920	4.303	4.849	6.965	9.925	14.089	31.600
3	0.765	1.638	2.353	3.182	3.482	4.541	5.841	7.453	12.924
4	0.741	1.533	2.132	2.776	2.999	3.747	4.604	5.598	8.610
5	0.727	1.476	2.015	2.571	2.757	3.365	4.032	4.773	6.869
6	0.718	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.959
7	0.711	1.415	1.895	2.365	2.517	2.998	3.499	4.029	5.408
8	0.706	1.397	1.860	2.306	2.449	2.896	3.355	3.833	5.041
9	0.703	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.781
10	0.700	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.587
11	0.697	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.437
∞	0.674	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.290
	0,25	0,10	0,05	0,025	0,02	0,01	0,005	0,0025	0,0005

Tabela 4.3 – Valores críticos da distribuição t de Student

6. Conclusão



Como o teste é unilateral à esquerda (pois, H_1 contém o sinal $<$), o valor crítico é encontrado levando em conta o nível de significância que está na última linha da tabela. Por isto que escolhemos a terceira coluna ($\alpha = 0,05$).

Rejeitamos H_0 se $t < t_0$. Como $-2,2129 < -1,833$, a estatística de teste está na área de rejeição. Portanto, rejeitamos H_0 , ou seja, os dados amostrais fornecem evidências suficientes para se concluir que a dieta é eficaz na redução do peso.

4.4 Utilização do Microsoft Excel para testes de duas amostras

Vamos utilizar duas ferramentas disponíveis no Excel para a realização de testes de hipótese para a comparação de duas médias: Teste – T: duas amostras presumindo variâncias diferentes e Teste – T: duas amostras em par para médias. Estas escolhas se devem ao fato delas serem as mais utilizadas na área profissional e de pesquisas. A versão utilizada é o Excel 2010.

Como mencionado no Capítulo 2, o suplemento Ferramenta de Análise deve estar ativo. Caso seja necessário, seguir os procedimentos descritos no Capítulo 2 para ativar este suplemento.

4.4.1 Comparação de duas médias com desvios padrões desconhecidos e diferentes



EXEMPLO

4.7: Uma empresa de computadores desenvolveu um novo curso que, comparado com o usual, apresenta novas técnicas para reparar computadores pessoais. Trinta e um estagiários foram selecionados aleatoriamente em dois grupos: 31 deles fizeram o curso usual e os outros 31 frequentaram o novo curso. Após 8 semanas, todos os estagiários foram submetidos ao mesmo exame final. De acordo com os resultados apresentados a seguir, há evidências de que os dois cursos apresentam resultados diferentes em termos de habilidade nos reparos? As pontuações mais altas indicam maior habilidade nos reparos. Use $\alpha = 0,05$.

Usual	3	5	7	9	8	9	7	4	9	9	8	7	5	4	8	8	9	7	6	5	5	4	8	9	7	6	6	4	4	8	7
Novo	8	5	9	9	5	6	4	3	2	5	8	4	8	4	9	5	7	9	6	7	7	8	5	6	4	8	7	5	6	5	4

Resolução:

Vamos seguir os seguintes passos para a realização do teste:

1º Passo: Digitar os dados das duas amostras na planilha:

Figura 4.4 – Valores das pontuações obtidas pelos estagiários, nos dois tipos de cursos.

2º Passo: Para a análise do nosso exemplo, clicamos na janela Dados e a seguir em Análise de dados. Escolhemos a Ferramenta de Análise Teste – T: duas amostras presumindo variâncias diferentes e, em seguida, OK.

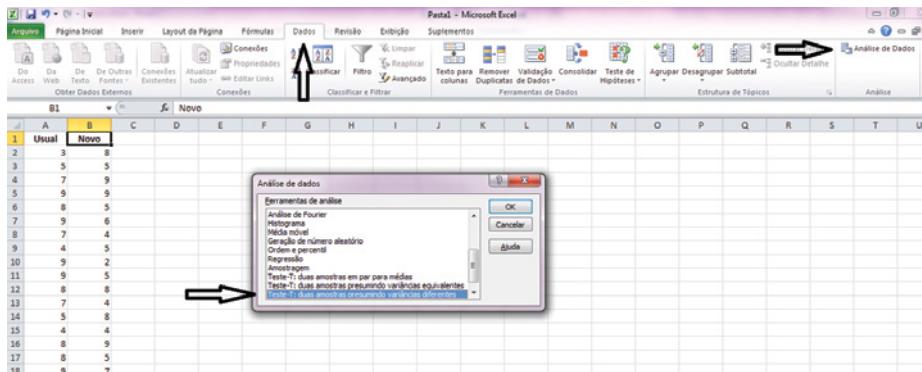


Figura 4.5 – Escolha da Análise de dados Teste – T: duas amostras presumindo variâncias diferentes.

3º Passo: Após clicar em Ok aparecerá uma nova caixa de diálogo. No campo Intervalo da variável 1, selecionar os dados arrastando com o mouse desde A2 até A32. No campo Intervalo da variável 2, selecionar os dados arrastando com o mouse desde B2 até B32. Em Hipótese da diferença de média, digitamos 0 (a hipótese $\mu_1 = \mu_2$ pode ser escrita como $\mu_1 - \mu_2 = 0$). O nível de significância é $\alpha = 0,05$. Em Opções de saída, escolher Nova planilha (as estatísticas calculadas sairão em uma planilha diferente daquela que utilizamos para digitar a entrada dos dados, basta identificá-la no rodapé) e, por fim, clicar em Ok.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Usual	Novo												
2	3	8												
3	5	5												
4	7	9												
5	9	9												
6	8	5												
7	9	6												
8	7	4												
9	4	5												
10	9	2												
11	9	5												
12	8	8												
13	7	4												
14	5	8												
15	4	4												
16	8	9												
17	8	5												
18	9	7												

Figura 4.6 – Entrada dos dados para a realização do teste.

4º Passo: Os resultados abaixo foram apresentados em uma nova planilha. Vamos entender as informações que estão grifadas:

1. Média: média de cada amostra.
2. Variância: variância de cada amostra.
3. Observações: número de observações em cada amostra
4. Hipótese da diferença de médias: $\mu_1 - \mu_2 = 0$.
5. g.l.: graus de liberdade (calculada por meio da fórmula descrita no box explicativo).
6. Stat t: valor da estatística de teste.
7. P($T \leq t$): valor p para o teste bicaudal (bilateral).
8. t crítico bicaudal: valores críticos para um teste bicaudal (bilateral).

Pasta1 - Microsoft Excel

The screenshot shows a Microsoft Excel spreadsheet titled "Pasta1 - Microsoft Excel". The ribbon menu is visible at the top, and the formula bar displays the title "Teste-t: duas amostras presumindo variâncias diferentes". The table is located in the range A1:G16. It has two columns: "Variável 1" and "Variável 2". The data rows are as follows:

	Variável 1	Variável 2
4 Média	6,612903226	6,129032258
5 Variância	3,511827957	3,449462366
6 Observações	31	31
7 Hipótese da diferença de média	0	
8 gl	60	
9 Stat t	1,021093564	
10 P(T<=t) uni-caudal	0,155654836	
11 t crítico uni-caudal	1,670648865	
12 P(T<=t) bi-caudal	0,311309671	
13 t crítico bi-caudal	2,000297822	

Figura 4.7 – Resultados obtidos a partir do Teste t – duas amostras presumindo variâncias diferentes, para os dados do Exemplo 4.7.

Sabemos que podemos concluir um teste de hipóteses utilizando o método do valor P e o método tradicional. Ao longo do capítulo, utilizamos o valor da estatística de teste e dos valores críticos para tomar uma decisão sobre rejeitar ou deixar de rejeitar a hipótese nula. Agora, por meio dos resultados obtidos pelo Excel, também podemos concluir pelo valor p. Como $0,313 > 0,05$, deixamos de rejeitar a hipótese nula. Pelo método tradicional, como o teste é bilateral, rejeitamos H_0 se $t < -t_c$ ou $t > t_c$. Como $t = 1,02$, a estatística de teste não está na área de rejeição, pois $1,02 < 2,00029$. Portanto, deixamos de rejeitar H_0 , ou seja, os dados amostrais não fornecem evidências suficientes para apoiar a afirmativa de que os cursos apresentam resultados diferentes em termos de habilidade nos reparos de computadores pessoais.

4.4.2 Comparação de duas médias (amostras dependentes)

Para este caso, utilizaremos os dados do Exemplo 4.6.

Vamos seguir os seguintes passos para a realização do teste:

The screenshot shows a Microsoft Excel spreadsheet with the following data:

	A	B	C	D	E	F	G
1	Depois	Antes					
2	80	77					
3	57	61					
4	60	60					
5	74	80					
6	87	90					
7	68	75					
8	90	85					
9	50	58					
10	82	89					
11	63	67					
12							

Figura 4.8 – Pesos, kg, de 10 pessoas, antes e depois de uma dieta.

2º Passo: Para a análise do exemplo, clicamos na janela Dados e a seguir em Análise de dados. Escolhemos a Ferramenta de Análise Teste – T: duas amostras em par para médias e, em seguida, OK.

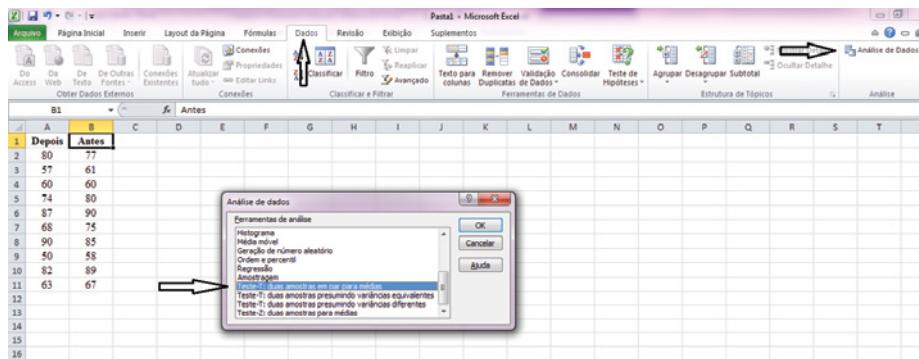


Figura 4.9 – Escolha da Análise de dados Teste – T: duas amostras em par para médias.

3º Passo: Após clicar em Ok aparecerá uma nova caixa de diálogo. No campo Intervalo da variável 1, selecionar os dados arrastando com o mouse desde A2 até A11. No campo Intervalo da variável 2, selecionar os dados arrastando com o mouse desde B2 até B11. Em Hipótese da diferença de média, digitamos 0 (a hipótese $\mu_1 = \mu_2$ pode ser escrita como $\mu_1 - \mu_2 = 0$). O nível de significância é $\alpha = 0,05$. Em Opções de saída, escolher Nova planilha (as estatísticas calculadas sairão em uma planilha diferente daquela que utilizamos para digitar a entrada dos dados, basta identificá-la no rodapé) e, por fim, clicar em Ok.

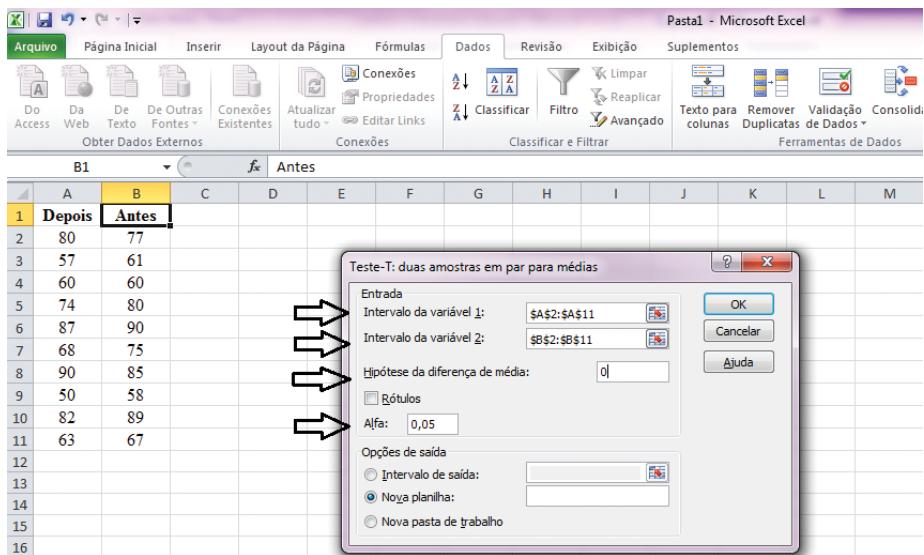


Figura 4.10 – Entrada dos dados para a realização do teste.

4º Passo: Os resultados abaixo foram apresentados em uma nova planilha. Vamos entender as informações que estão grifadas:

1. Observações: número de observações.
2. Hipótese da diferença de médias: $\mu_0 = 0$.
3. gl: graus de liberdade.
4. Stat t: valor da estatística de teste.
5. P(T \leq t): valor p para o teste unicaudal (unilateral).
6. t crítico unicaudal: valores críticos para um teste unicaudal (unilateral).

Pastas1 - Microsoft

The screenshot shows a Microsoft Excel spreadsheet titled "Pastas1 - Microsoft". The ribbon menu is visible at the top, with "Arquivo" selected. The main content is a table in the range A1:G15. The table has 15 rows and 7 columns. Rows 1 through 14 contain numerical data, while row 15 is empty. The first column contains row numbers from 1 to 14. The second column contains labels for the statistical results. The third and fourth columns provide values for "Variável 1" and "Variável 2" respectively. The fifth column contains the value "10". The sixth column contains the value "9". The seventh column contains the value "1,833112933". The font used in the table is Calibri, size 11.

	A	B	C	D	E	F	G
1	Teste-t: duas amostras em par para médias						
2							
3		Variável 1	Variável 2				
4	Média	71,1	74,2				
5	Variância	184,3222222	146,4				
6	Observações	10	10				
7	Correlação de Pearson	0,946812739					
8	Hipótese da diferença de média	0					
9	gl	9					
10	Stat t	-2,21115421					
11	P(T<=t) uni-caudal	0,027171407					
12	t critico uni-caudal	1,833112933					
13	P(T<=t) bi-caudal	0,054342815					
14	t critico bi-caudal	2,262157163					
15							

Figura 4.11 – Resultados obtidos a partir do Teste t – duas amostras em par para médias, para os dados do Exemplo 4.6.

Nesta análise, também temos a informação do valor p. Como $0,027 < 0,05$, rejeitamos a hipótese nula, mesma conclusão que aquela obtida pelo método descrito ao longo do capítulo ($-2,2111 < -1,833$). Portanto, os dados amostrais fornecem evidências suficientes para se concluir que a dieta é eficaz na redução do peso.

O valor p também pode ser obtido através da função TESTE.T. Para explicar o procedimento, vamos utilizar os dados do Exemplo 4.6. Após digitar os dois conjuntos de dados, como na Figura 4.8, seguimos os seguintes passos: na aba Fórmulas, clicar em Mais Funções , Estatística e escolher TESTE.T. Esta sequência é apresentada na Figura 4.12.

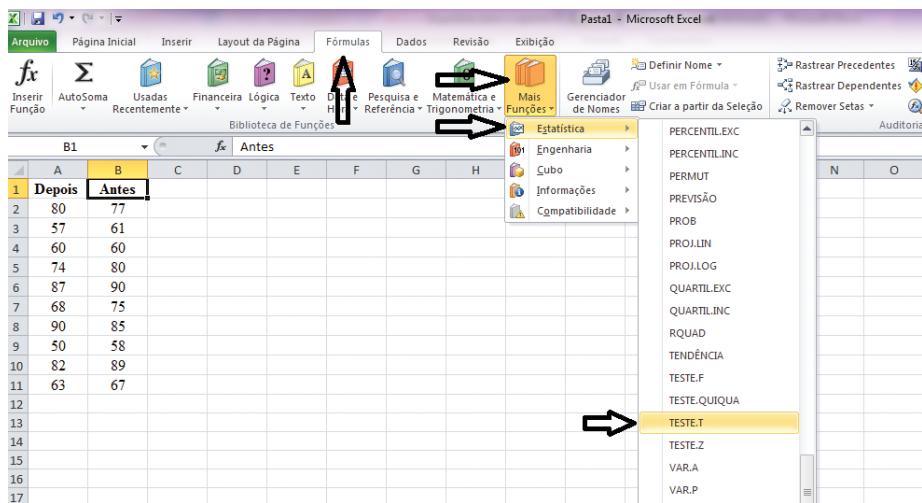


Figura 4.12 – Escolha da função estatística TESTE.T.

Após clicar em TESTE.T aparecerá uma janela em que temos que preencher os argumentos da função:

1. Matriz1: é o primeiro conjunto de dados, ou seja: A2:A11;
 2. Matriz2: é o segundo conjunto de dados, ou seja: B2:B11;
 3. Caudas: especifica o número de caudas da distribuição a ser retornado:
Para distribuição unicaudal, digitamos 1 e para distribuição bicaudal, digitamos 2. No nosso exemplo, o teste é unicaudal, portanto, digitamos 1.

4. Tipo: é o tipo de teste t. Para testes pareados, digitamos 1. Para testes com variação igual das duas amostras, digitamos 2 e para variação desigual, digitamos 3. No nosso exemplo, o teste t é para amostras dependentes (ou seja, pareadas). Então, digitamos 1.

Após o preenchimento de todos os argumentos, clicamos em OK e aparecerá o valor p. As informações estão apresentadas na Figura 4.13.

A screenshot of Microsoft Excel showing the 'TESTE.T' function dialog box. The formula bar at the top shows the formula =TESTE.T(A2:A11;B2:B11;1;1). The dialog box is titled 'Argumentos da função' (Function Arguments) and contains the following arguments:

- TESTE.T**
- Matriz1**: A2:A11 (Range: {80;57;60;74;87;68;90;50;82;63})
- Matriz2**: B2:B11 (Range: {77;61;60;80;90;75;85;58;89;67})
- Caudas**: 1
- Tipo**: 1

The result is displayed as = 0,027171407, with a black arrow pointing to the result field.

Figura 4.13 – Probabilidade associada ao teste t de Student.

Observamos que o valor p obtido é o mesmo daquele grifado na Figura 4.11. Como a conclusão de um teste pode ser feita pelo método tradicional ou do valor p, caso a escolha seja pelo valor p, a função estatística TESTE.T torna a análise mais rápida, sem a necessidade de fazer o procedimento pela Análise de Dados.

REFLEXÃO

Ao longo deste capítulo, estudamos uma das ferramentas mais importantes da inferência estatística, que são os testes de hipóteses. Aprendemos a realizar testes para a média populacional, nos casos de uma ou duas amostras.

Com os exemplos apresentados, pudemos observar a grande aplicabilidade dos testes de hipóteses na área da saúde.

Quando queremos fazer comparações sobre os parâmetros de duas populações, não basta selecionarmos duas amostras e analisarmos somente as estatísticas amostrais obtidas. Precisamos testar a afirmativa sobre estes parâmetros analisando os dados amostrais, por meio da realização de um teste apropriado e, a partir da conclusão do teste, teremos evidências para apoiar ou não a afirmativa sobre os parâmetros.

Não podemos esquecer que os testes não podem ser utilizados indiscriminadamente. Há requisitos que devem ser verificados! Com um planejamento correto para a obtenção dos dados amostrais, podemos fazer uso de mais uma ferramenta imprescindível na tomada de decisões!



LEITURA

Sugerimos que você assista ao vídeo que está no seguinte endereço: <http://m3.ime.unicamp.br/recursos/1098>. Você aprenderá algumas técnicas de planejamento de experimento, bem como verificará a importância da formulação correta de uma hipótese na análise estatística.



REFERÊNCIAS BIBLIOGRÁFICAS

FARIAS, Alfredo A.; SOARES, José F.; CÉSAR, Cibel C. **Introdução à Estatística**. 2 ed. Rio de Janeiro: LTC, 2003.

LARSON, Ron; FARBER, Betsy. **Estatística Aplicada**. 2. ed. São Paulo: Prentice Hall, 2004.

LEVIN, Jack; FOX, James A. **Estatística para Ciências Humanas**. 9 ed. São Paulo: Prentice Hall, 2004.

LEVINE, David M.; BERENSON, Mark L.; STEPHAN, David. **Estatística: Teoria e Aplicações Usando Microsoft Excel em Português**. Rio de Janeiro: LTC, 2000.

MAGALHÃES, Marcos N. ; LIMA, Antonio C. P de. **Noções de Probabilidade e Estatística**. 6. ed. São Paulo: Editora da Universidade de São Paulo, 2004.

PAGANO, Marcello.; GAUVREAU, Kimberlee. **Princípios de Bioestatística**. São Paulo: Pioneira Thomson Learning, 2004.

TRIOLA, Mário F. **Introdução à Estatística**. 10. ed. Rio de Janeiro: LTC, 2008.

VIEIRA, Sonia. **Introdução à Bioestatística**. 4 ed. Rio de Janeiro: Elsevier, 2008.

RIFO, Laura L. Ramos; CAMARNEIRO, Fábio; SANTOS, José P. de Oliveira.

Disponível em: < <http://m3.ime.unicamp.br/recursos/1098> >. Acesso em: 03 maio 2015.

5

Correlação e Regressão Linear Simples

No Capítulo 2, estudamos como podemos descrever os dados provenientes de uma variável quantitativa por meio de medidas resumo. Agora, estudaremos uma técnica estatística denominada correlação. Por meio dela, verificamos se existe relação entre duas variáveis quantitativas: uma, chamada variável Y (dependente ou resposta), e a outra, chamada variável X (independente ou explanatória). Direcionaremos nosso estudo no relacionamento linear entre as variáveis X e Y.

Se identificarmos uma relação linear entre as variáveis X e Y, podemos determinar a equação da reta que melhor modela os dados. Esta reta é chamada reta de regressão, e sua equação é chamada equação de regressão. Com esta equação, podemos prever o valor da variável resposta associada com um valor fixo da variável explicativa. Para encontrarmos a equação de regressão, utilizaremos a técnica de regressão linear simples.

Um exemplo do estudo de correlação e regressão linear simples pode ser feito para verificar a relação entre o comprimento e a idade gestacional de bebês nascidos com até 1500 gramas. Havendo uma relação, podemos encontrar a equação de regressão e utilizá-la para estimar o comprimento do bebê para determinado valor atribuído à idade gestacional.



OBJETIVOS

Com as técnicas estudadas neste capítulo, esperamos que você seja capaz de:

- Construir e interpretar o diagrama de dispersão;
- Calcular e interpretar o coeficiente de correlação linear;
- Compreender os conceitos básicos da regressão linear simples;
- Estimar a equação de regressão e utilizá-la para fazer previsões.

5.1 Diagrama de dispersão

Quando estudamos duas variáveis quantitativas, temos interesse em responder as seguintes questões:

- Há algum tipo de relação entre as variáveis X e Y?
- Qual o tipo de relacionamento entre elas?
- Qual a intensidade da relação?

Neste tipo de estudo, temos um par de resultados (x, y) para cada elemento da amostra, ou seja, a análise dos dados envolve a resposta de duas variáveis para cada elemento da amostra. Antes de conduzirmos qualquer tipo de análise, devemos construir um gráfico denominado diagrama de dispersão, com o objetivo de verificar se existe uma relação entre as variáveis X e Y. Neste diagrama, os pares ordenados (x, y) representam pontos em um plano coordenado. A variável X é representada no eixo das abscissas (horizontal) e a variável Y no eixo das ordenadas (vertical).

A Figura 5.1 apresenta alguns tipos de correlação.

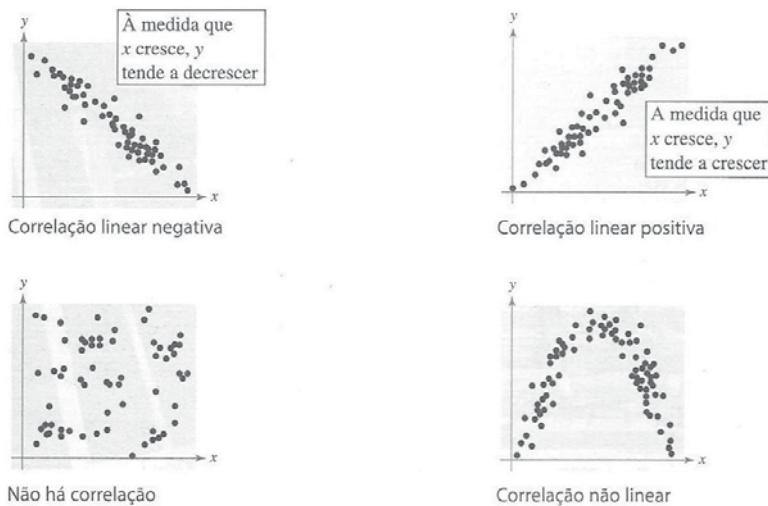


Figura 5.1 – Diagramas de dispersão com alguns tipos de correlação. Fonte: LARSON (2004, p. 334).

Com o auxílio do diagrama de dispersão, podemos identificar a forma, a direção e a intensidade da relação particular existente entre duas variáveis quantitativas. Na Figura 5.1, identificamos nos dois primeiros gráficos superiores, uma forma bem definida: os dados dispõem-se aproximadamente ao longo de uma linha reta, portanto, verificamos um padrão linear. Também, identificamos nestes dois gráficos, uma direção bem clara: No gráfico à esquerda, à medida que x cresce, y tende a decrescer e, no da direita, à medida que x cresce, y tende a crescer. A intensidade de uma relação é determinada por quanto próximo os pontos se aproximam mais de uma reta. Analisando os dois gráficos inferiores, verificamos que o da esquerda não mostra qualquer forma, sugerindo que não há relação entre x e y . O gráfico à direita mostra uma forma bem distinta, sugerindo uma relação entre x e y , cuja forma não é de uma reta.

5.2 Coeficiente de correlação linear

A análise do diagrama de dispersão nos auxilia na verificação de uma possível relação linear entre as variáveis X e Y , mas a intensidade da correlação entre as variáveis são determinadas utilizando o coeficiente de correlação linear (r).

O coeficiente de correlação é um número adimensional, ou seja, não tem unidade de medida. Os valores mínimo e máximo que o coeficiente pode assumir são, respectivamente, -1 e 1. Quando isto ocorre, dizemos que há uma relação linear perfeita entre as variáveis X e Y , ou seja, no diagrama de dispersão, todos os pares (x, y) se encontrariam sobre uma linha reta. Valores próximos de zero para o coeficiente de correlação linear indicam que x e y não estão linearmente relacionadas, ressaltando que pode haver outro tipo de relacionamento entre x e y , mas não o linear. Se os valores da variável y tendem a aumentar conforme os valores da variável x aumentam, teremos r positivo, e dizemos que x e y são positivamente correlacionadas. Agora, se os valores da variável y tendem a diminuir conforme os valores da variável x aumentam, teremos r negativo, e dizemos que x e y são negativamente correlacionadas. O coeficiente de correlação linear não é resistente, ou seja, a presença de outliers pode afetar bastante o valor de r .

O coeficiente de correlação linear de Pearson é definido pela seguinte fórmula:

$$r = \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right) \cdot \left(\sum_{i=1}^n y_i\right)}{\sqrt{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} \cdot \sqrt{n\left(\sum_{i=1}^n y_i^2\right) - \left(\sum_{i=1}^n y_i\right)^2}}, \quad -1 \leq r \leq 1$$

A correlação não faz distinção entre a variável explicativa e a variável resposta, ou seja, no cálculo do coeficiente de correlação linear, não importa qual variável é chamada de x e qual é chamada de y . O cálculo de r é feito com dados amostrais. Se tivéssemos todos os pares de valores populacionais x e y , substituiríamos r por ρ (letra grega rô).

Podemos calcular o coeficiente de correlação linear para qualquer conjunto de dados amostrais em pares. Mas, não podemos esquecer que estamos usando dados amostrais para tomar uma decisão sobre dados populacionais. Então, para determinarmos se o coeficiente de correlação populacional ρ é significante, precisamos realizar um teste de hipótese.

Para se testar hipóteses ou fazer inferências sobre r , precisamos verificar alguns requisitos.

Segundo TRIOLA (2008, p. 413)

1. A amostra de dados emparelhados (x,y) é uma amostra aleatória de dados quantitativos independentes. (É importante que os dados amostrais não tenham sido coletados com o uso de método não apropriado, por exemplo, amostra de resposta voluntária).
2. O exame visual do diagrama de dispersão deve confirmar que os pontos se aproximam do padrão de uma reta.
3. Quaisquer outliers devem ser removidos caso se saiba que são erros. Os efeitos de quaisquer outros outliers devem ser considerados pelo cálculo de r com e sem o outlier incluído.

Utilizaremos os dados do Exemplo 5.1 para construir o diagrama de dispersão e para calcular o coeficiente de correlação linear.



EXEMPLO

5.1: A Tabela 5.1 fornece o peso e a estatura de 10 pessoas adultas, do sexo feminino.

ALTURA (X)	PESO (Y)
156	53,5
158	58,4
163	59,4
162	56,4
165	61,2
172	57,5
173	67,3
174	69,7
179	77,2
183	81,6

Tabela 5.1 – Peso, em kg, e altura, em cm, de 10 pessoas adultas, do sexo feminino.

Vamos construir o diagrama de dispersão e calcular o coeficiente de correlação linear.

Resolução

Primeiro, vamos construir o diagrama de dispersão colocando cada par (x,y) no plano e depois verificamos, visualmente, o comportamento conjunto das variáveis.

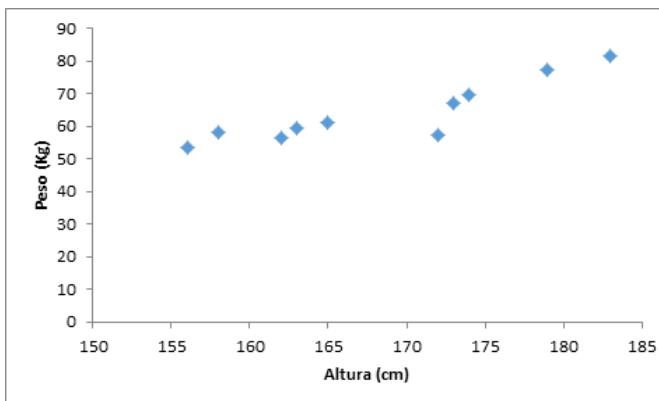


Figura 5.2 – Diagrama de dispersão para os dados sobre a altura e o peso de 10 mulheres adultas.

Analizando o diagrama de dispersão, observamos que, à medida que a altura aumenta, o peso tende a aumentar. Portanto, o diagrama nos sugere que as variáveis x e y são positivamente correlacionadas. Para medir a intensidade da correlação, vamos calcular o coeficiente de correlação linear. Para efetuar este cálculo, vamos acrescentar três colunas na tabela original dos dados, obtendo o seguinte quadro:

ALTURA (X)	PESO (Y)	X · Y	X ²	Y ²
156	53,5	8.346	24.336	2.862,25
158	58,4	9.227,2	24.964	3.410,56
163	59,4	9.682,2	26.569	3.528,36
162	56,4	9.136,8	26.244	3.180,96
165	61,2	10.098	27.225	3.745,44
172	57,5	9.890	29.584	3.306,25
173	67,3	11.642,9	29.929	4.529,29
174	69,7	12.127,8	30.276	4.858,09
179	77,2	13.818,8	32.041	5.959,84
183	81,6	14.932,8	33.489	6.658,56
$\sum = 1.685$	$\sum = 642,2$	$\sum = 108.902,5$	$\sum = 284.657$	$\sum = 42.039,6$

Para obtermos os valores da coluna ($x \cdot y$), multiplicamos cada par (x, y), ou seja, $156 \cdot 53,5$, $158 \cdot 58,4$ e assim por diante. Os valores x^2 são obtidos elevando ao quadrado cada valor da primeira coluna, ou seja, $156 \cdot 156 = 24.336$; $158 \cdot 158 = 24.964$, e assim por diante. Finalmente, obtemos y^2 fazendo cada valor da segunda coluna ao quadrado, isto é, $53,5 \cdot 53,5 = 2.862,25$; $58,4 \cdot 58,4 = 3.410,56$, e assim por diante.

Com os totais de cada uma das colunas, temos todos os valores necessários para substituir na fórmula do coeficiente de correlação linear:

$$r = \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right) \cdot \left(\sum_{i=1}^n y_i\right)}{\sqrt{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} \cdot \sqrt{n\left(\sum_{i=1}^n y_i^2\right) - \left(\sum_{i=1}^n y_i\right)^2}}$$

$$r = \frac{10(108.902,5) - (1.685) \cdot (642,2)}{\sqrt{10(284.657) - (1.685)^2} \cdot \sqrt{10(42.039,6) - (642,2)^2}}$$

$$r = \frac{1.089.025 - 1.082.107}{\sqrt{2.846.570 - 2.839.225} \cdot \sqrt{420.396 - 412.420,84}}$$

$$r = \frac{6.918}{\sqrt{7.345} \cdot \sqrt{7.975,16}} = \frac{6.918}{85,7030 \cdot 89,3038} = \frac{6.918}{7.653,60} = 0,9039$$

Como $r = 0,9039$, concluímos que as variáveis peso e altura são fortemente correlacionadas.

Após o cálculo do coeficiente de correlação linear, é comum utilizá-lo para fazer inferências sobre a natureza da relação entre x e y . Quando fazemos isto, precisamos tomar os seguintes cuidados:

- Uma alta correlação não implica necessariamente que haja uma relação de causa e efeito entre x e y .
- Uma baixa correlação não implica, necessariamente, que x e y não estejam correlacionadas. Apenas podemos afirmar que não estão fortemente e linearmente relacionadas. O diagrama de dispersão pode retratar um padrão que sugere uma forte relação não linear.

Devemos, também, ter o cuidado em interpretar correlações baseadas em médias de valores.

De acordo com TRIOLA (2008, p. 417),

As médias suprimem a variação individual e podem aumentar o coeficiente de correlação. Um estudo produziu um coeficiente de correlação 0,4 para dados emparelhados que relacionavam renda e educação entre indivíduos, mas o coeficiente de correlação linear se tornou 0,7 quando foram usadas médias regionais.

Um exemplo antigo, mas muito interessante, foi dado por um estatístico que mostrou que havia correlação positiva entre o número de recém-nascidos e o número

de cegonhas em uma pequena cidade da Dinamarca, no decorrer dos anos 30. A correlação entre essas duas variáveis é espúria: não indica relação de causa e efeito. Existe uma terceira variável, o crescimento da cidade, que implicava tanto no número de recém-nascidos (quanto maior a cidade, mais crianças nascem) quanto no número de casas com chaminés, perto das quais as cegonhas faziam seus ninhos (VIEIRA, 2008, p. 120).

Como já dissemos anteriormente, o coeficiente de correlação é obtido por meio de dados amostrais. Para fazermos inferências sobre o coeficiente de correlação populacional ρ , realizaremos um teste de hipóteses utilizando o coeficiente de correlação amostral.

5.3 Teste de hipóteses para correlação

Estudamos, no Capítulo 4, que o teste de hipóteses é um método da inferência estatística, em que utilizamos dados amostrais de uma população para testar uma afirmativa sobre uma propriedade desta população.

Para a realização de um teste de hipóteses para correlação, devemos seguir os seguintes passos:

1. Estabelecer as hipóteses nula e alternativa:

$$\begin{cases} H_0 : \rho = 0 \text{ (não há correlação linear significante)} \\ H_1 : \rho \neq 0 \text{ (há correlação linear significante)} \end{cases}$$

em que ρ é o coeficiente de correlação populacional.

2. Determinar a estatística de teste:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

em que n é o número de pares ordenados e r é o coeficiente de correlação amostral de Pearson.

3. Especificar o nível de significância. Normalmente, utilizamos $\alpha = 0,01$, $\alpha = 0,05$ ou $\alpha = 0,10$.

4. Determinar o número de graus de liberdade: g.l. = $n - 2$.

5. Conclusão:

• Se $|t| >$ valores críticos, encontrado na Tabela 2 - Apêndice, rejeitamos H_0 e concluímos que há uma correlação linear significante.

• Se $|t| \leq$ valores críticos, encontrado na Tabela 2 - Apêndice, deixamos de rejeitar H_0 . Ou seja, não há evidência suficiente para se concluir que haja uma correlação linear.



EXEMPLO

5.2: Utilizando os dados do Exemplo 5.1, vamos testar a hipótese de que há uma correlação entre o peso e a altura de pessoas adultas, do sexo feminino. Considerar $\alpha = 0,05$.

Resolução

As hipóteses são:

$$\begin{cases} H_0 : \rho = 0 \text{ (não há correlação linear significante)} \\ H_1 : \rho \neq 0 \text{ (há correlação linear significante)} \end{cases}$$

A estatística de teste é:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,9039}{\sqrt{\frac{1-(0,9039)^2}{10-2}}} = \frac{0,9039}{\sqrt{\frac{1-0,81703521}{8}}} = \frac{0,9039}{\sqrt{0,022870599}} = \frac{0,9039}{0,151230284} = 5,977$$

Como $\alpha = 0,05$ e o número de graus de liberdade é $n - 2 = 10 - 2 = 8$, os valores críticos são $t = \pm 2,306$. O teste é bicaudal devido à maneira que as hipóteses foram estabelecidas ($=$ e \neq).

9.

P(T DE STUDENT ≥ VALOR TABELADO) = α ↔ VALORES BILATERAIS									
G. L.	0.50	0.20	0.10	0.05	0.04	0.02	0.01	0.005	0.001
1	1.000	3.078	6.314	12.706	15.894	31.821	63.656	127.321	636.578
G. L.	0.50	0.20	0.10	0.05	0.04	0.02	0.01	0.005	0.001
1	1.000	3.078	6.314	12.706	15.894	31.821	63.656	127.321	636.578

$P(T \geq \text{VALOR TABELADO}) = \alpha \leftrightarrow \text{VALORES BILATERAIS}$									
2	0,816	1,886	2,920	4,303	4,849	6,965	9,925	14,089	31,600
3	0,765	1,638	2,353	3,182	3,482	4,541	5,841	7,453	12,924
4	0,741	1,533	2,132	2,776	2,999	3,747	4,604	5,598	8,610
5	0,727	1,476	2,015	2,571	2,757	3,365	4,032	4,773	6,869
6	0,718	1,440	1,943	2,447	2,612	3,143	3,707	4,317	5,959
7	0,711	1,415	1,895	2,365	2,517	2,998	3,499	4,029	5,408
8	0,706	1,397	1,860	2,306	2,449	2,896	3,355	3,833	5,041
9	0,703	1,383	1,833	2,262	2,398	2,821	3,250	3,690	4,781
10	0,700	1,372	1,812	2,228	2,359	2,764	3,169	3,581	4,587
11	0,697	1,363	1,796	2,201	2,328	2,718	3,106	3,497	4,437
12	0,695	1,356	1,782	2,179	2,303	2,681	3,055	3,428	4,318
110	0,677	1,289	1,659	1,982	2,078	2,361	2,621	2,865	3,381
120	0,677	1,289	1,658	1,980	2,076	2,358	2,617	2,860	3,373
∞	0,674	1,282	1,645	1,960	2,054	2,326	2,576	2,807	3,290
	0,25	0,10	0,05	0,025	0,02	0,01	0,005	0,0025	0,0005

Tabela 5.2 – Valores críticos da distribuição t de Student

De acordo com a estatística de teste e os valores críticos, temos que $5,977 > 2,306$. Portanto, rejeitamos H_0 , ou seja, há uma correlação linear significante entre o peso e a altura das mulheres.

Quando determinamos, por meio do teste de hipóteses, que a correlação linear é significante, podemos encontrar a reta que melhor descreve os dados observados. Esta reta é obtida por meio da equação de regressão, que é utilizada para prever o valor da variável y para determinado valor da variável x.

Aprenderemos, a seguir, como encontrar e equação de regressão.

5.4 Regressão linear simples

De acordo com Moore et al. (2006, p.95)

Uma reta de regressão é uma linha reta que descreve como uma variável de resposta y muda à medida que uma variável explicativa x também varia. Frequentemente utilizamos uma reta de regressão para predizer o valor de y a partir de um determinado valor de x.

Para obtermos a reta de regressão, precisamos da equação de regressão. Esta equação é estimada utilizando a técnica de regressão linear simples. A equação de regressão expressa a relação entre a variável independente (x) e a variável dependente (y). Voltando ao nosso exemplo do início do capítulo, a idade gestacional do bebê é a variável independente e, a partir de determinado valor atribuído a ela, podemos prever o comprimento do bebê (que é variável dependente) utilizando a equação de regressão.

A regressão linear simples envolve uma variável independente e uma variável dependente. A análise de regressão envolvendo duas ou mais variáveis independentes é chamada de análise de regressão múltipla.

Antes de começarmos o estudo para encontrar a equação de regressão, vamos relembrar qual é a equação de uma reta.

A equação típica de uma reta é $y = mx + b$, em que m é o coeficiente angular e b é o intercepto. O coeficiente angular informa a inclinação da reta em relação ao eixo das abscissas (x).

Se m for um número:

- positivo, a reta é crescente;
- negativo, a reta é decrescente;
- zero, a reta é paralela ao eixo das abscissas.

O coeficiente linear é a ordenada do ponto em que a reta corta o eixo das ordenadas (y).

Em Estatística, a equação de regressão é expressa na forma:

$$y = b_0 + b_1x$$

Os coeficientes b_0 e b_1 são estatísticas amostrais usadas para estimarem os parâmetros populacionais β_0 e β_1 . Portanto, utilizaremos dados amostrais em pares para estimar a equação de regressão. A notação \hat{y} (y “chapéu”) serve para distinguir entre um valor observado y e o valor correspondente \hat{y} , que é encontrado utilizando a reta de regressão.

Utilizaremos as seguintes fórmulas para encontrar os coeficientes b_0 e b_1 , respectivamente:

$$b_1 = \frac{n \cdot \left(\sum_{i=1}^n x_i \cdot y_i \right) - \left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n y_i \right)}{n \cdot \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2}$$

$$\text{e } b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Podemos observar que o numerador do cálculo do estimador b_1 é exatamente o numerador do coeficiente de correlação linear e o denominador é o valor obtido dentro da primeira raiz do denominador do coeficiente de correlação linear.

Vamos lembrar que:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (\text{média da variável } x)$$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad (\text{média da variável } y)$$

Analizando a fórmula para calcular b_0 , observamos que este coeficiente só pode ser encontrado após o cálculo de b_1 .



CONEXÃO

A reta de regressão é a que melhor se ajusta aos dados amostrais. O critério específico usado para se determinar qual reta se ajusta “melhor” é a propriedade dos mínimos quadrados. Uma leitura interessante sobre a propriedade dos mínimos quadrados pode ser encontrada em TRIOLA (2008, p. 435).

Uma observação importante: diferentemente do cálculo do coeficiente de correlação linear r , a distinção entre a variável independente e a variável dependente é essencial. Se invertermos os papéis das duas variáveis, obteremos uma reta de regressão diferente.

5.3: Vamos utilizar os dados do Exemplo 5.1 para encontrar a equação de regressão.

Resolução

Precisaremos das informações contidas no Quadro 5.1.

ALTURA (X)	PESO (Y)	X · Y	X ²	Y ²
156	53,5	8.346	24.336	2.862,25
158	58,4	9.227,2	24.964	3.410,56
163	59,4	9.682,2	26.569	3.528,36
162	56,4	9.136,8	26.244	3.180,96
165	61,2	10.098	27.225	3.745,44
172	57,5	9.890	29.584	3.306,25
173	67,3	11.642,9	29.929	4.529,29
174	69,7	12.127,8	30.276	4.858,09
179	77,2	13.818,8	32.041	5.959,84
183	81,6	14.932,8	33.489	6.658,56
$\sum = 1.685$	$\sum = 642,2$	$\sum = 108.902,5$	$\sum = 284.657$	$\sum = 42.039,6$

Os valores de b_1 e b_0 são, respectivamente,

$$b_1 = \frac{n \cdot (\sum_{i=1}^n x_i \cdot y_i) - (\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{n \cdot (\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

$$b_1 = \frac{10 \cdot (108.902,5) - (1.685) \cdot (642,2)}{10 \cdot (284.657) - (1.685)^2}$$

$$b_1 = \frac{6.918}{7.345} = 0,941865$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

$$b_0 = 64,22 - 0,941865 \cdot (168,5)$$

$$b_0 = 64,22 - 158,704253 = -94,4843$$

$$\text{pois, } \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1685}{10} = 168,5 \text{ e } \bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{642,2}{10} = 64,22$$

Portanto, a equação de regressão é:

$$y = -94,4843 + 0,9419x$$

Agora que já conhecemos a equação de regressão, a pergunta que surge é: como podemos interpretá-la?

Segundo TRIOLA (2008, p. 434)

Ao se trabalhar com duas variáveis relacionadas por uma equação de regressão, a mudança marginal em uma variável é a quantidade que ela varia quando a outra variável varia de exatamente uma unidade. A inclinação b_1 na equação de regressão representa a mudança marginal em y quando x varia de uma unidade.

Então, para os dados da Tabela 5.1, referentes ao peso e altura das 10 mulheres, a equação de regressão tem uma inclinação 0,9419, que mostra que, se aumentarmos x (altura) em 1 unidade, o peso aumenta em 0,9419 unidades, aproximadamente. Esta interpretação fica fácil de ser verificada se substituirmos valores para x . Por exemplo, se $x = 155$, $y = -94,4843 + 0,9419(155) = -94,4843 + 155,9945 = 51,5102$ e, se $x = 156$, $y = -94,4843 + 0,9419(156) = -94,4843 + 146,9364 = 52,4521$. A diferença entre os valores de y encontrados, $y = -94,4843 + 0,9419x = -94,4843 + 0,9419(171) = 66,58 \text{ kg}$, é exatamente o valor de b_1 , ou seja, para cada acréscimo de 1 unidade em x , y cresce de 0,9419 unidades.

A Figura 5.3 apresenta, no diagrama de dispersão, a reta de regressão.

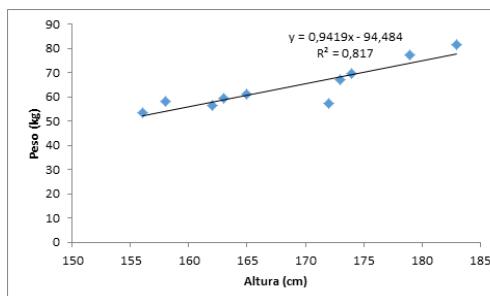


Figura 5.3 – Reta de regressão: peso (kg) em função da altura (cm).

Estudaremos, mais adiante, o que significa a informação $R_2 = 0,817$.

Podemos utilizar a equação de regressão para prever valores de Y para quaisquer valores de X dentro do intervalo estudado, mesmo que tais valores não estejam na amostra. Por exemplo, para $x = 171$ (valor que não está na Tabela 5.1), podemos estimar o valor de Y. Basta substituir este valor na equação da reta, ou seja:

$$y = -94,4843 + 0,9419x = -94,4843 + 0,9419(171) = 66,58 \text{ kg}$$

Interpretamos o valor $y = 66,58 \text{ kg}$ como uma previsão para o peso, quando a altura da mulher adulta for 171 cm.

Se atribuirmos à variável x um valor observado no conjunto de dados, por exemplo, $x = 165$, vamos encontrar o seguinte valor previsto para y:

$$y = -94,4843 + 0,9419x = -94,4843 + 0,9419(165) = 60,93 \text{ kg}$$

Analizando a Tabela 5.1, observamos que para a altura $x = 165$ cm, o peso correspondente é $y = 61,2$ kg. Esta diferença entre o valor amostral observado e o valor previsto pela equação de regressão é denominada resíduo. Então, temos a seguinte definição:

$$\text{resíduo} = y_{\text{observado}} - y_{\text{previsto}} = y - \bar{y}$$

Um gráfico de resíduos é outro instrumento útil para a análise dos resultados da correlação e regressão e para a verificação dos requisitos necessários para se fazerem inferências sobre correlação e regressão. Este gráfico é construído usando o mesmo eixo x do diagrama de dispersão, mas no eixo y (vertical) utilizamos os valores dos resíduos. Se o gráfico de resíduos não revelar qualquer padrão, a equação de regressão é uma boa representação da associação entre as duas variáveis.

A equação de regressão deve ser utilizada para fazer previsões apenas se ela for um bom modelo para os dados, ou seja, se for verificado por meio de um teste de hipóteses que a relação entre as duas variáveis é significante. Caso a relação não seja significante, o melhor valor previsto de y é \bar{y} .

Devemos tomar o cuidado de não fazer extrapolações, ou seja, utilizar a equação de regressão para fazer previsões para a variável Y utilizando valores para X muito distantes dos limites dos dados amostrais disponíveis.

De acordo com Anderson et al. (2003, p. 447), “usar a equação de regressão estimada fora do intervalo dos valores da variável independente deve ser feito com cuidado porque fora deste intervalo nós não podemos assegurar que a mesma relação seja válida”.

Agora que já aprendemos a utilizar as técnicas de correlação e regressão linear simples, vamos listar alguns conceitos importantes que foram estudadas e que não podemos esquecer:

1. O diagrama de dispersão nos dá uma ideia da relação, ou não, entre duas variáveis quantitativas.
 2. O coeficiente de correlação linear de Pearson mede a intensidade da relação linear, ou seja, só tem sentido calculá-lo se o diagrama de dispersão indicar uma relação linear.
 3. Correlação não indica causa. Uma forte relação entre duas variáveis não é suficiente para que se tirem conclusões de causa e efeito.
 4. Caso haja relação entre duas variáveis quantitativas, podemos descrevê-la através da equação de regressão que melhor representa a relação.
 5. Devemos usar a equação de regressão para previsões somente se houver uma correlação linear, confirmada pelo teste de hipóteses. Caso contrário, a melhor estimativa para a variável y é sua média amostral \bar{y} .
-

5.5 Coeficiente de determinação

Em geral, há uma variação em torno da reta de regressão, ou seja, nem todos os pontos ficam sobre a reta (pode acontecer de nenhum estar exatamente sobre a reta). Para medir a precisão da reta de regressão ajustada, isto é, a proporção da variação de Y que é explicada pela reta de regressão (variação de X), utilizamos o coeficiente de determinação. O coeficiente de determinação, R^2 , é dado pelo quadrado do coeficiente de correlação. Este coeficiente é particularmente importante se vamos usar a equação de regressão para fazer previsões. Nesse caso, queremos um R^2 tão próximo de 1 quanto possível.

Para os dados do Exemplo 5.1, o coeficiente de determinação é:

$$R^2 = (0,9039)^2 = 0,8170$$

Isto significa que 81,70% da variação do peso das mulheres se explica pela variação da altura. Este valor aparece na Figura 5.3.

Com a definição do coeficiente de determinação, podemos perceber que, se o coeficiente de correlação for $r = \pm 0,7$, teremos um coeficiente de determinação $r^2 = 0,49$, significando que a reta de regressão ajustada não consegue explicar nem a metade da variação de y . Por isso, para $-0,7 \leq x \leq 0,7$, não se deve, em geral, ajustar a reta de regressão. Para $|r| = 0,9$, a reta de regressão explica mais de 80% da variação total de y .

Vamos estudar mais um exemplo para exercitar todos os conteúdos abordados ao longo do capítulo.



EXEMPLO

5.4: Muitos acidentes de carro são causados por motoristas cansados. Vários estudos de pesquisa mostram que mudanças nas pupilas dos olhos estão relacionadas com a fadiga. Obteve-se uma amostra aleatória de 25 motoristas, e mediram-se as oscilações no tamanho da pupila (x , em milímetros por segundo) usando-se um pupilógrafo. O cansaço de cada pessoa (y) também foi registrado, usando-se o índice de pupila sem descanso (IPSD). As estatísticas resumo são:

$$\sum x_i = 7,1 ; \sum y_i = 192 ; \sum x_i \cdot y_i = 49,22 ; \sum x_i^2 = 2,1064 ; \sum y_i^2 = 2,094$$

Fonte: KOKOSKA (2013, p. 509).

De acordo com as estatísticas resumo:

- Calcule o coeficiente de correlação linear.
- Teste a hipótese de que há correlação linear significante, com um nível de significância 0,05.
- Encontre a equação de regressão por mínimos quadrados.
- Faça a previsão para o IPSD, considerando $x = 0,3$ milímetro por segundo.
- Calcule o coeficiente de determinação e interprete.

Resolução

- Neste exercício, já temos as informações necessárias para substituirmos na fórmula do coeficiente de correlação linear:

$$r = \frac{n(\sum_{i=1}^n x_i y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{\sqrt{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2} \cdot \sqrt{n(\sum_{i=1}^n y_i^2) - (\sum_{i=1}^n y_i)^2}}$$

$$r = \frac{25(49,22) - (7,1)(192)}{\sqrt{25(2,1064) - (7,1)^2} \cdot \sqrt{25(2,094) - (192)^2}}$$

$$r = \frac{1.230,5 - 1363,2}{\sqrt{52,66 - 50,41} \cdot \sqrt{52,350 - 36,864}}$$

$$r = \frac{-132,7}{\sqrt{2,25} \cdot \sqrt{15,486}} = \frac{-132,7}{1,5 \cdot 124,44} = \frac{-132,7}{186,66} = -0,7109$$

Como o coeficiente de correlação é -0,7109, concluímos que as variáveis oscilação no tamanho da pupila e IPSD são negativamente correlacionadas.

b) Estabelecendo as hipóteses:

$$\begin{cases} H_0: \rho = 0 \text{ (não há correlação linear significante)} \\ H_1: \rho \neq 0 \text{ (há correlação linear significante)} \end{cases}$$

A estatística de teste é:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0,7109}{\sqrt{\frac{1-(-0,7109)^2}{25-2}}} = \frac{-0,7109}{\sqrt{\frac{1-0,50537881}{23}}} = \frac{-0,7109}{\sqrt{0,021505269}} = \frac{-0,7109}{0,1466467} = -4,848$$

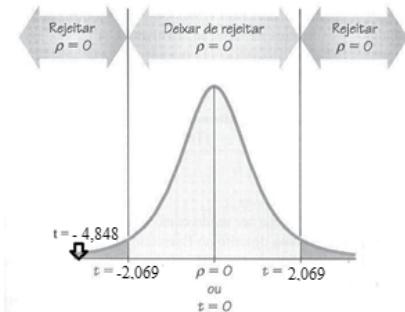
Como $\alpha = 0,05$ e o número de graus de liberdade é $n - 2 = 25 - 2 = 23$, os valores críticos são $t = \pm 2,069$. Novamente, o teste é bicaudal devido à maneira que as hipóteses foram estabelecidas ($=$ e \neq).

Tabela - Valores críticos da distribuição t de Student

P(T DE STUDENT ≥ VALOR TABELADO) = α ↔ VALORES BILATERAIS									
G. L.	0.50	0.20	0.10	0.05	0.04	0.02	0.01	0.005	0.001
20	0.687	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.850
21	0.686	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.819
22	0.686	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.792
23	0.685	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.768
24	0.685	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.745
25	0.684	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.725
26	0.684	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.707

Tabela 5.3 – Valores críticos da distribuição t de Student

De acordo com a estatística de teste e os valores críticos, temos que $| -4,848 | > \pm 2,069$. Portanto, rejeitamos H_0 , ou seja, há uma correlação linear significante entre a oscilação no tamanho da pupila e IPSD.



Quando a correlação linear é significante, podemos encontrar a reta de regressão, que melhor descreve os dados em estudo.

- c) Para encontrar a equação de regressão, precisamos encontrar os valores estimados dos parâmetros.

Os valores de b_1 e b_0 são, respectivamente,

$$b_1 = \frac{n \cdot \left(\sum_{i=1}^n x_i \cdot y_i \right) - \left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n y_i \right)}{n \cdot \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2}$$

$$b_1 = \frac{25(49,22) - (7,1) \cdot (192)}{25(2,1064) - (7,1)^2}$$

$$b_1 = \frac{-132,7}{2,25} = -58,9778$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

$$b_0 = 7,68 - (-58,9778) \cdot (0,284)$$

$$b_0 = 7,68 + 16,7497 = 24,4297$$

pois, $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{7,1}{25} = 0,284$ e $\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{192}{25} = 7,68$.

Portanto, a equação de regressão é:

$$y = 24,4297 - 58,9778x$$

- d) Para encontrar o valor previsto do IPSD, basta substituirmos $x = 0,3$ na equação de regressão:

$$y = 24,4297 - 58,9778(0,3)$$

$$y = 24,4297 - 17,69334 = 6,7364$$

- e) O coeficiente de determinação é dado pelo quadrado do coeficiente de correlação, ou seja:

$$R^2 = (-0,7109)^2 = 0,5054$$

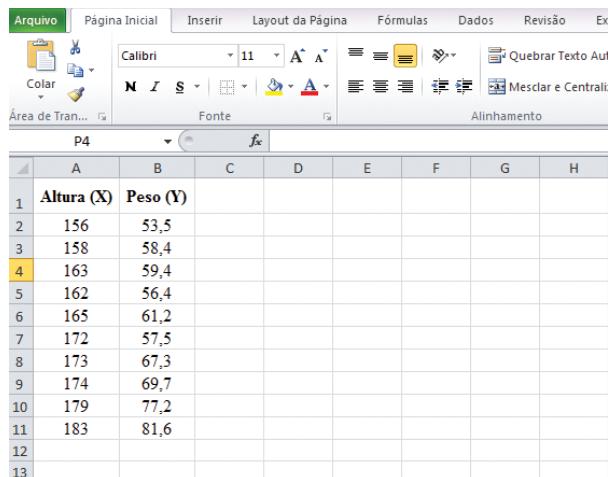
Isto significa que 50,54% da variação do IPSD se explica pela oscilação no tamanho da pupila.

5.6 Utilização do Microsoft Excel na análise de regressão e correlação

Podemos utilizar o Excel para construir o diagrama de dispersão, obter o coeficiente de correlação linear de Pearson e o coeficiente de determinação e determinar a equação de regressão. Para aprendermos o processo, vamos utilizar os dados do Exemplo 5.1. Utilizaremos a versão Excel 2010.

Para fazer as análises, seguiremos os seguintes passos:

1º Passo: Vamos digitar os pares ordenados das variáveis X e Y em uma planilha do Excel.



	A	B	C	D	E	F	G	H
1	Altura (X)	Peso (Y)						
2	156	53,5						
3	158	58,4						
4	163	59,4						
5	162	56,4						
6	165	61,2						
7	172	57,5						
8	173	67,3						
9	174	69,7						
10	179	77,2						
11	183	81,6						
12								
13								

Figura 5.4 – Valores da altura e peso de pessoas adultas, do sexo feminino

2º Passo: Neste passo, selecionamos os dados (podemos selecionar com os títulos das colunas). Após a seleção, clicar na aba Inserir e depois selecionar o tipo de gráfico a ser elaborado. Vamos escolher a primeira opção para o gráfico de Dispersão. Clicar sobre a figura.

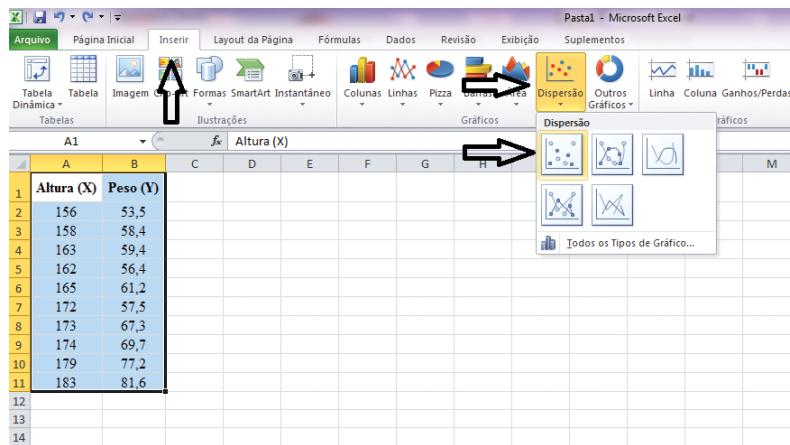


Figura 5.5 – Seleção dos dados e escolha do gráfico Dispersão.

3º Passo: Após clicar na primeira opção do gráfico Dispersão, o gráfico construído está apresentado na Figura 5.6.

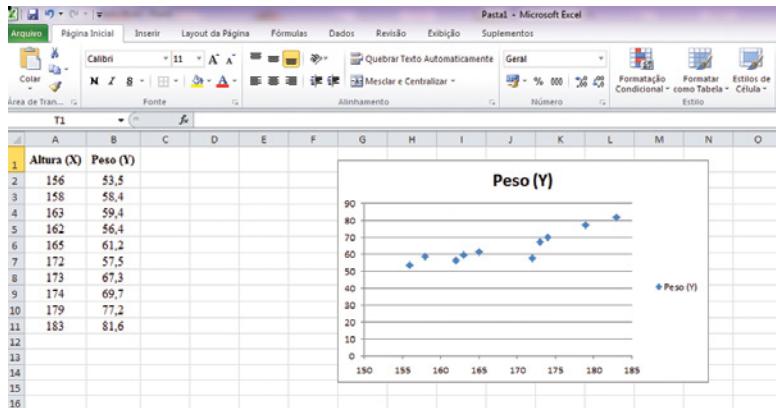


Figura 5.6 – Diagrama de dispersão.

4º Passo: Nesta etapa, vamos formatar o gráfico: deletar a legenda, o título e as linhas de grade e vamos colocar nome nos eixos. Para deletar, basta clicarmos sobre a legenda e o título e usar o botão direito do mouse ou o próprio

teclado do computador para excluir. Para as linhas de grade, basta clicar sobre qualquer uma delas e utilizar o botão direito do mouse para excluir.

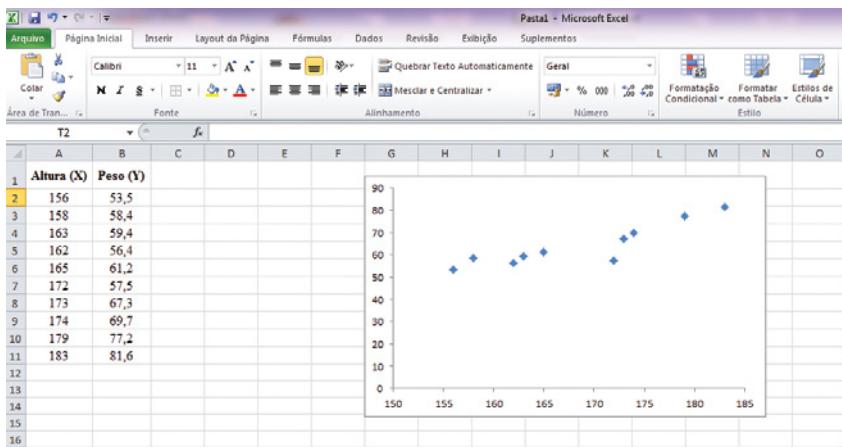


Figura 5.7 – Diagrama de dispersão (sem a legenda e sem o título).

5º Passo: Agora, vamos colocar nome nos eixos: clicamos sobre o gráfico e aparecerá Ferramentas de Gráfico com algumas opções de escolha. Clicar em Layout e logo em seguida Títulos dos Eixos. Utilizamos as duas opções: uma para colocar título no eixo horizontal e a outra para colocar o título no eixo vertical. A Figura 5.8 ilustra a escolha para o Título do Eixo Horizontal Principal, com a opção Título Abaixo do Eixo. Após a inserção do título horizontal, seguimos o mesmo procedimento para o eixo vertical.

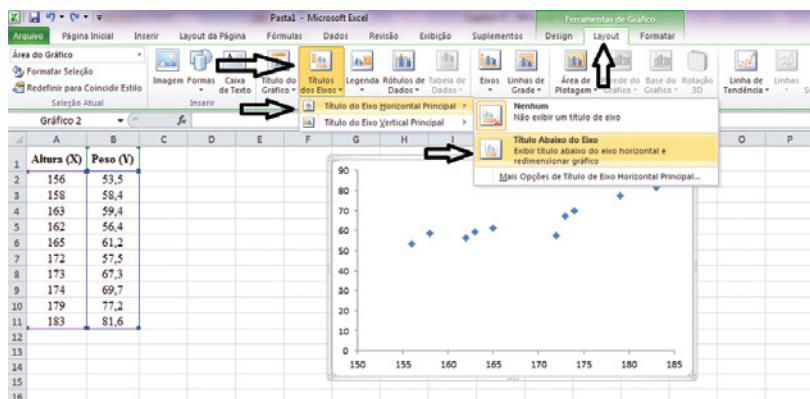


Figura 5.8 – Procedimentos para inserir títulos nos eixos.

6º Passo: A Figura 5.9 apresenta o diagrama de dispersão finalizado.

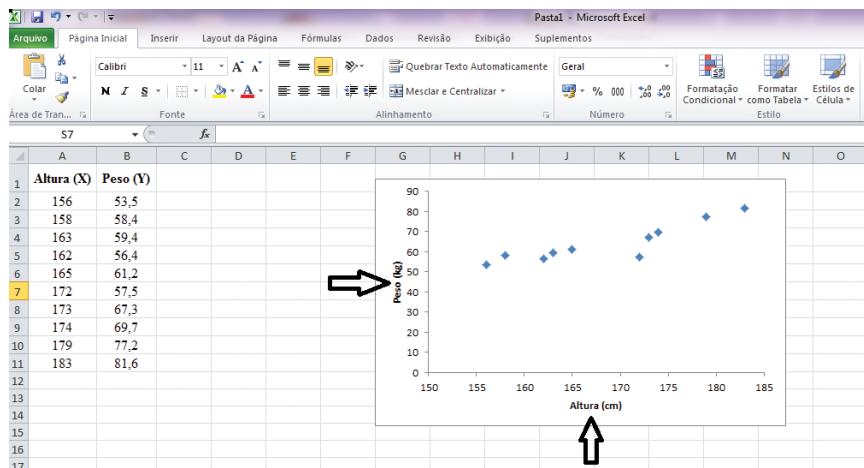


Figura 5.9 – Diagrama de dispersão da altura e peso de pessoas adultas, do sexo feminino.

Agora, vamos obter o coeficiente de correlação linear de Pearson.

1º Passo: Vamos digitar os pares ordenados das variáveis X e Y em uma planilha do Excel.

The screenshot shows a Microsoft Excel spreadsheet with the data table from Figure 5.9. The data is organized into two columns: "Altura (X)" and "Peso (Y)". The first row contains the column headers. The data points are listed in rows 2 through 11. Row 13 is empty. The table is selected, and the formula bar shows "P4".

	A	B	C	D	E	F	G	H
1	Altura (X)	Peso (Y)						
2	156	53,5						
3	158	58,4						
4	163	59,4						
5	162	56,4						
6	165	61,2						
7	172	57,5						
8	173	67,3						
9	174	69,7						
10	179	77,2						
11	183	81,6						
12								
13								

Figura 5.10 – Valores da altura e peso de pessoas adultas, do sexo feminino.

2º Passo: Para obtermos o coeficiente de correlação, clicamos na aba Fórmulas e, em seguida, clicamos em Mais Funções. Selecionando a primeira opção, Estatística, aparecerá uma lista de funções. Escolher a opção CORREL.

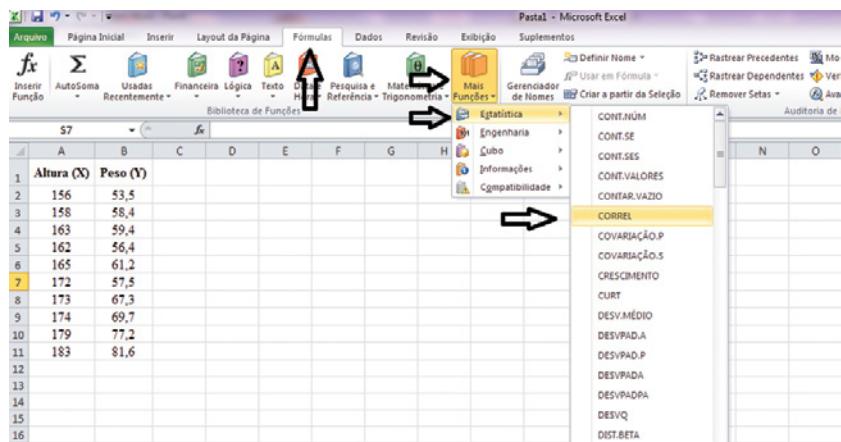


Figura 5.11 – Escolha da função CORREL para obtenção do coeficiente de correlação.

3º Passo: Após a escolha da função CORREL, aparecerá uma janela Argumentos da função. No campo Matriz 1, selecionamos os dados da variável altura (sem o título) que estão na planilha e, no campo Matriz 2, selecionamos os dados da variável peso (sem o título) que estão na planilha. Para selecionar os dados, basta clicar no primeiro valor e arrastar (com o mouse) até o último valor.

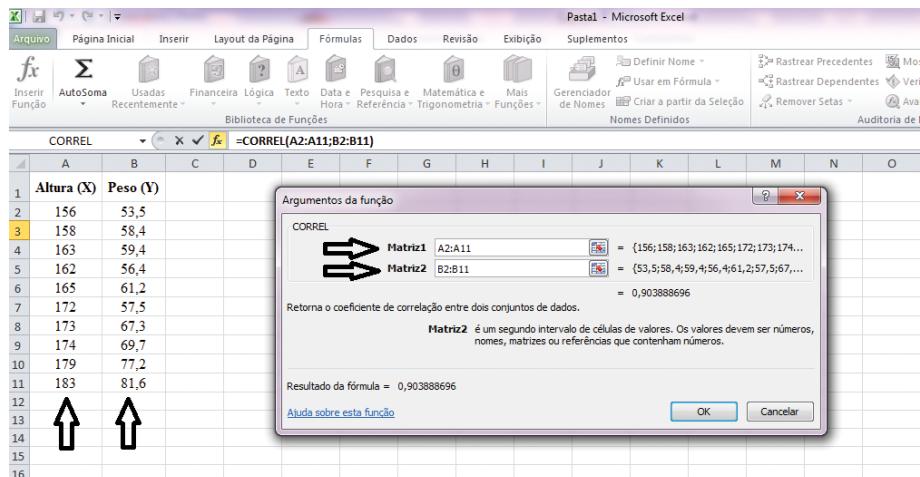


Figura 5.12 – Preenchimento dos argumentos da função.

4º Passo: Agora, clicamos em OK e obtemos o coeficiente de correlação.

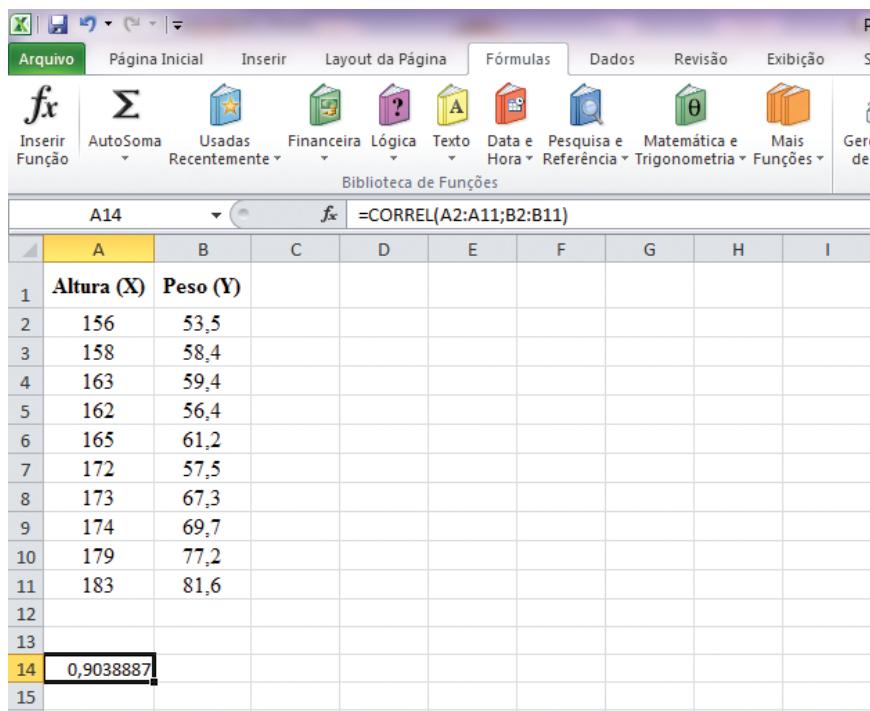


Figura 5.13 – Coeficiente de correlação linear.

O coeficiente de correlação é $r = 0,9038887$. Já sabímos que seria positivo, pois o diagrama de dispersão indica uma relação linear positiva (crescente) entre as variáveis em estudo. Por definição, o coeficiente de determinação é dado pelo quadrado do coeficiente de correlação. Portanto:

$$R^2 = (0,9038887)^2 = 0,8170$$

Isto significa que 81,70% da variação do peso se explica pela variação na altura das pessoas adultas, do sexo feminino.

E, para finalizar, vamos obter a equação de regressão.

1º Passo: Vamos digitar os pares ordenados das variáveis X e Y em uma planilha do Excel.

The screenshot shows a Microsoft Excel spreadsheet with the ribbon menu at the top. The tabs visible are Arquivo, Página Inicial, Inserir, Layout da Página, Fórmulas, Dados, Revisão, and Ex. Below the ribbon, there are several toolbars: Colar (Cut/Copy/Paste), Fonte (Font), and Alinhamento (Text Alignment). The active cell is P4. The table below contains 12 rows of data, with the first two columns labeled 'Altura (X)' and 'Peso (Y)'. The data is as follows:

	A	B	C	D	E	F	G	H
1	Altura (X)	Peso (Y)						
2	156	53,5						
3	158	58,4						
4	163	59,4						
5	162	56,4						
6	165	61,2						
7	172	57,5						
8	173	67,3						
9	174	69,7						
10	179	77,2						
11	183	81,6						
12								
13								

Figura 5.14 – Valores da altura e peso de pessoas adultas, do sexo feminino.

2º Passo: As medidas apresentadas neste capítulo podem ser obtidas utilizando o Excel. Para isto, o suplemento Análise de Dados deve estar ativo. Caso ele esteja ativo, deve aparecer o ícone Análise de Dados após clicar na janela Dados.

É muito comum este suplemento não aparecer ativo. Caso isto aconteça, devemos seguir o seguinte procedimento:

- Clicar no Botão Office e em seguida Opção do Excel. Escolher Suplementos e clicar;
- Escolher na lista Suplementos de Aplicativos Inativos a opção Ferramenta de Análise e clicar em Ir...
- Selecionar o seguinte suplemento disponível: Ferramenta de análise e clicar em OK.

Com o suplemento ativo, podemos fazer várias análises estatísticas!

Para a análise do nosso exemplo, clicamos na janela Dados e a seguir em Análise de dados. Escolhemos a Ferramenta de Análise Regressão e clicamos em OK.

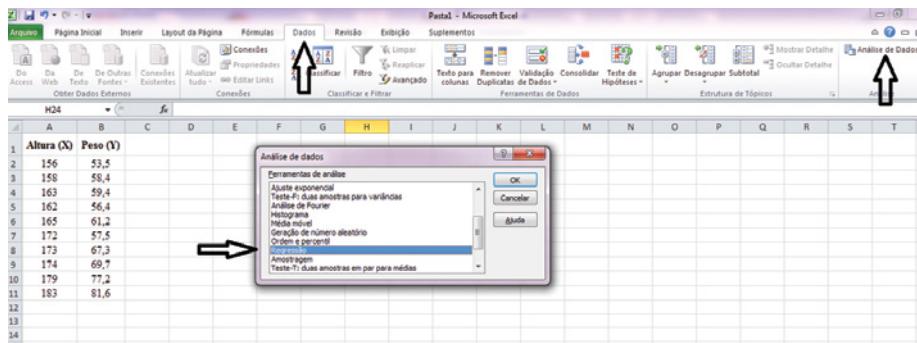


Figura 5.15 – Escolha da Análise de dados Regressão.

3º Passo: Após clicar em Ok aparecerá uma nova caixa de diálogo. No campo Intervalo Y de entrada, selecionar os dados arrastando com o mouse desde B2 até B11. No campo Intervalo X de entrada, selecionar os dados arrastando com o mouse desde A2 até A11. Devemos ficar atentos para selecionar corretamente os valores de Y e X! Em Opções de saída, escolher Nova planilha (as estatísticas calculadas sairão em uma planilha diferente daquela que utilizamos para digitar a entrada dos dados, basta identificá-la no rodapé) e, por fim, clicar em Ok.

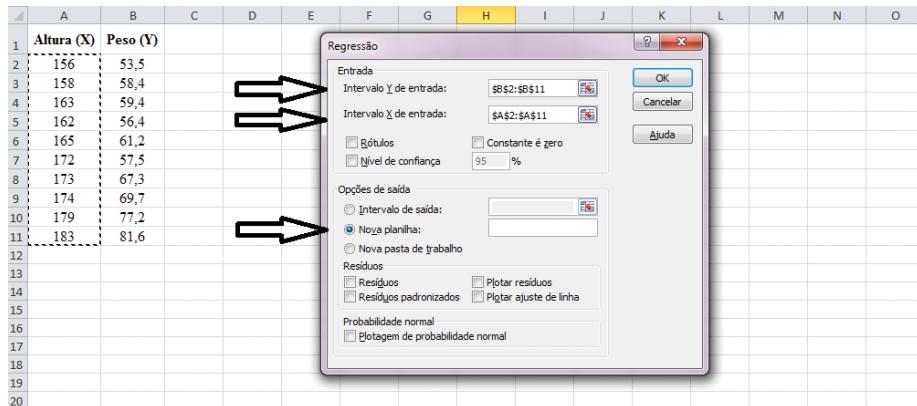


Figura 5.16 – Entrada dos dados para análise de regressão.

4º Passo: Os resultados abaixo foram apresentados em uma nova planilha. Vamos entender as informações que estão grifadas:

1. R múltiplo: é o coeficiente de correlação.
2. R – Quadrado: é o coeficiente de determinação.
3. Intersecção: é o coeficiente b_0 .
4. Variável X1: é o coeficiente b_1 .
5. Observações: número de pares ordenados (x,y).

	A	B	C	D	E	F	G	H	I	J	K
1	RESUMO DOS RESULTADOS										
2											
3	Estatística de regressão										
4	R múltiplo	0,903888696									
5	R-Quadrado	0,817014775									
6	R-quadrado ajustado	0,794141622									
7	Erro padrão	4,271030974									
8	Observações	10									
9											
10	ANOVA										
11		gl	SQ	MQ	F	F de significância					
12	Regressão	1	651,5823553	651,5823553	35,71937681	0,00033196					
13	Réstduo	8	145,9336447	18,24170558							
14	Total	9	797,516								
15											
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%		
17	Interseção	-94,48428863	26,58873986	-3,55354519	0,007470704	-155,7980327	-33,17054457	-155,7980327	-33,17054457		
18	Variável X 1	0,941865214	0,157592963	5,976568983	0,00033196	0,57845519	1,305275239	0,57845519	1,305275239		

Figura 5.17 – Resumo dos resultados.

Utilizando os coeficientes obtidos, temos que a equação de regressão é:

$$y = b_0 + b_1 x$$

$$y = -94,4843 + 0,9419 x$$

Agora que já sabemos qual é a equação de regressão, temos a opção de traçar a reta e mostrar a equação e o coeficiente de determinação no diagrama de dispersão. Para isto, quando estamos construindo o gráfico, clicamos sobre qualquer um dos pontos. Aparecerá:

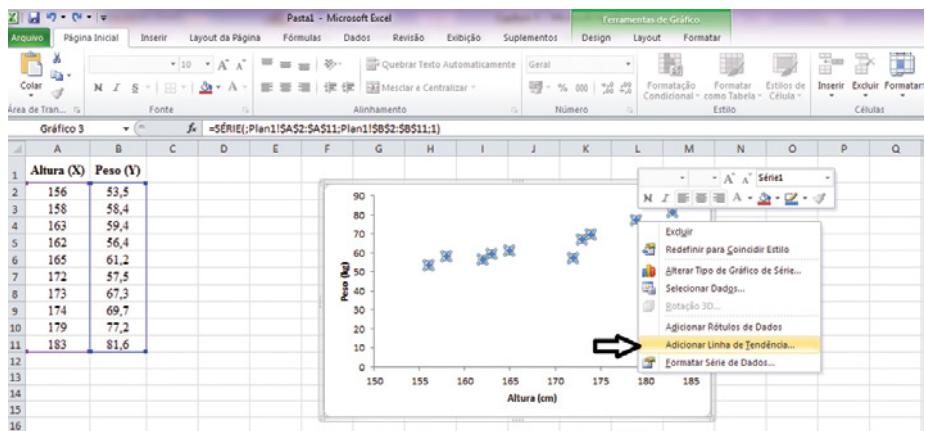


Figura 5.18 – Adicionar linha de tendência no diagrama de dispersão.

Quando clicarmos em Adicionar Linha de Tendência aparecerá a janela Formatar Linha de Tendência. Nela, escolhemos a opção Linear, Exibir Equação no gráfico e Exibir valor de R-quadrado no gráfico.

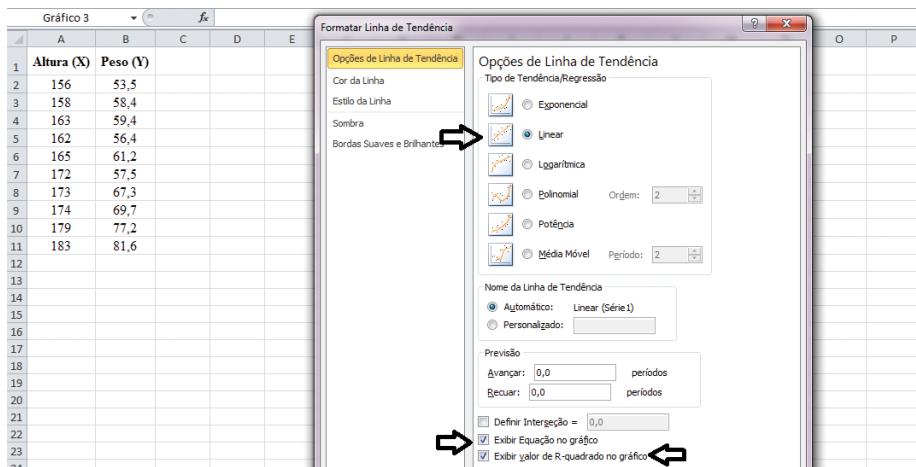


Figura 5.19 – Formatação da linha de tendência.

Finalmente, quando clicamos em fechar, aparecerá, no diagrama de dispersão, a reta ajustada, a equação da reta e o coeficiente de determinação. Podemos deslocar, com o mouse, as informações da reta e do coeficiente, colo- cando-os em uma posição mais conveniente no diagrama. Basta clicar em cima das informações e arrastar.

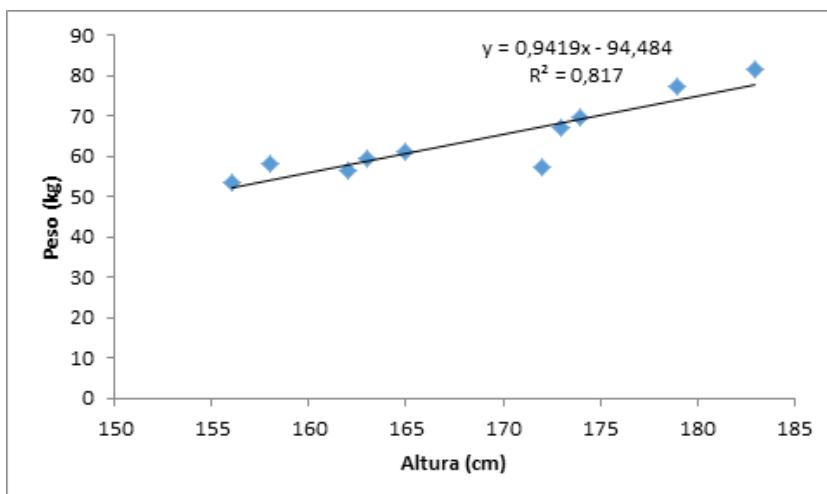


Figura 5.20 – Reta de regressão: peso (kg) em função da altura (cm).



REFLEXÃO

Chegamos ao final deste capítulo. Nele, exploramos as técnicas de correlação e regressão linear simples. Com larga aplicação, o conteúdo apresentado nos permite obter, por exemplo, funções matemáticas relacionando o preço com a demanda, a número de anos que um paciente fumou e a capacidade pulmonar, níveis de colesterol e triglicerídeos antes de uma dieta, peso da mãe e do bebê ao nascer, taxa de mortalidade infantil e expectativa de vida em uma amostra de países, entre tantas outras aplicações. Aprendemos que, quando temos informações, em pares, de duas variáveis quantitativas, podemos estudar mais profundamente um possível relacionamento entre essas variáveis, em particular, o relacionamento linear. Havendo um relacionamento linear, estimamos os coeficientes da equação de regressão pelo método de mínimos quadrados. Um dos maiores interesses é conseguir fazer previsões da variável dependente a partir, valores atribuídos para a variável independente. Mas, sabemos que para fazermos previsões, precisamos testar a adequabilidade de modelo! Além disto, temos que, tomar cuidado em não fazer extrapolações, pois não podemos garantir que a mesma relação seja válida para valores da variável independente muito distantes daqueles utilizados para encontrar a equação de regressão.

Com o uso cuidadoso destas técnicas, obtemos informações estatísticas importantes no auxílio à tomada de decisões, em várias áreas do conhecimento.



LEITURA

Um vídeo muito interessante, que aborda o conceito de correlação e correlações espúrias, é encontrado no endereço <http://m3.ime.unicamp.br/recursos/1084>. Vale a pena assistir!



REFERÊNCIAS BIBLIOGRÁFICAS

ARANGO, Héctor G. **Bioestatística Teórica e Computacional**. Rio de Janeiro: Editora Guanabara Koogan S.A., 2001.

LARSON, Ron; FARBER, Betsy. **Estatística Aplicada**. 2. ed. São Paulo: Prentice Hall, 2004.

KOKOSKA, Stephen. **Introdução à Estatística – Uma Abordagem por Resolução de Problemas**. Rio de Janeiro: LTC, 2013.

LEVINE, David M.; BERENSON, Mark L.; STEPHAN, David. **Estatística: Teoria e Aplicações Usando Microsoft Excel em Português**. Rio de Janeiro: LTC, 2000.

MOORE, David S.; McCABE, George P.; DUCKWORTH, William M.; SCLOVE, Stanley L. **A Prática da Estatística Empresarial** – Como Usar Dados para Tomar Decisões. Rio de Janeiro: LTC, 2006.

TRIOLA, Mário F. **Introdução à Estatística**. 10. ed. Rio de Janeiro: LTC, 2008.

VIEIRA, Sonia. **Estatística básica**. São Paulo: Cengage Learning, 2013.

VIEIRA, Sonia. **Introdução à Bioestatística**. 4 ed. Rio de Janeiro: Elsevier, 2008.

RIFO, Laura R. Ramos; ANNUNCIATO, Angela; SANTOS, José P. de Oliveira. Disponível em: <<http://m3.ime.unicamp.br/recursos/1084>>. Acesso em: 03 maio 2015.



ANOTAÇÕES



ANOTAÇÕES



ANOTAÇÕES



ANOTAÇÕES



P152020029

