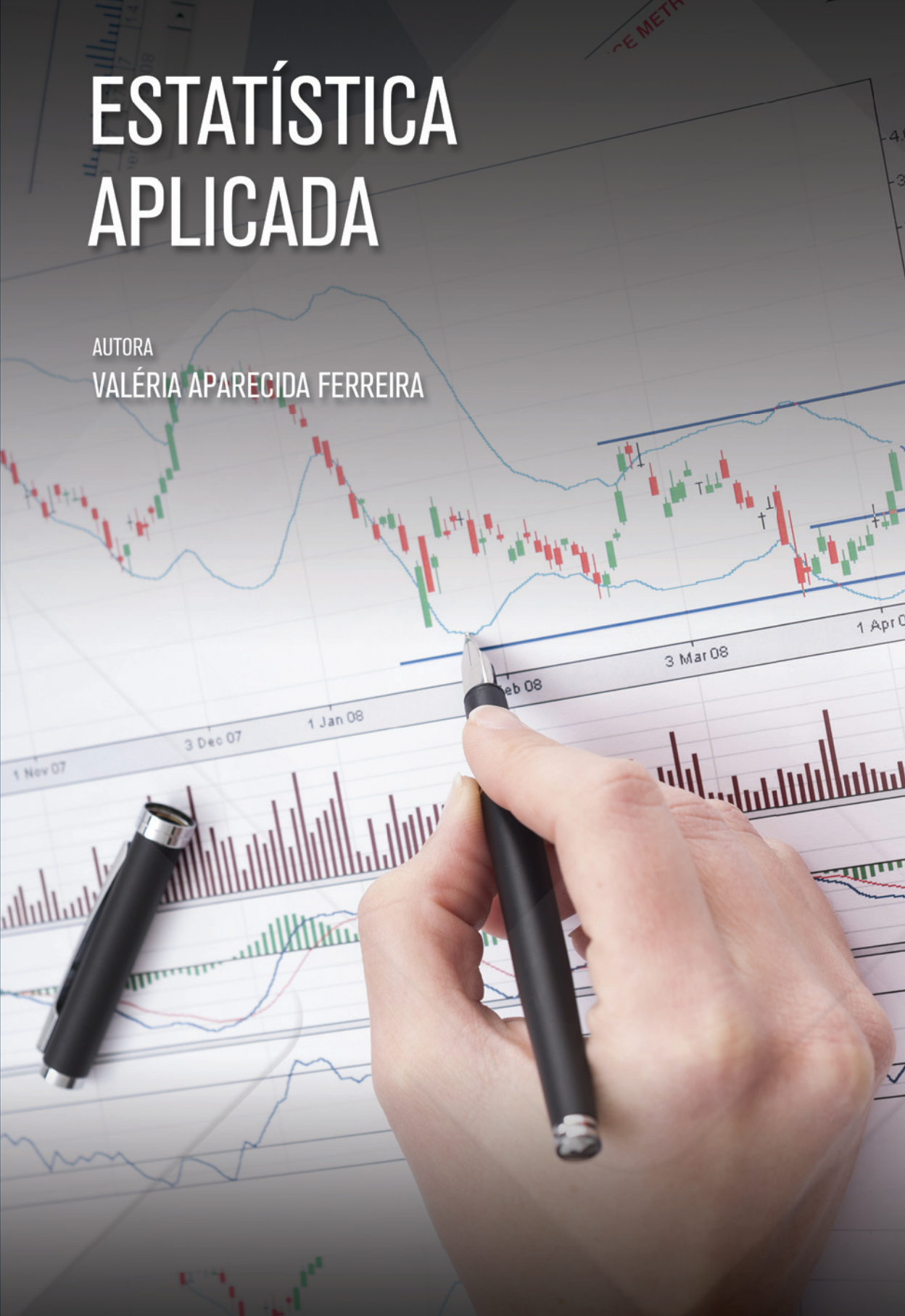
ESTATÍSTICA APLICADA

AUTORA
VALÉRIA APARECIDA FERREIRA



ESTATÍSTICA APLICADA

AUTOR DO ORIGINAL
VALÉRIA APARECIDA FERREIRA

1ª EDIÇÃO
SESES
RIO DE JANEIRO 2015

Estácio

Conselho editorial DURVAL CORRÊA MEIRELLES, RONALD CASTRO PASCHOAL, OTÁVIA TRAVENÇOLO MUNIZ SALA

Autor do original VALÉRIA APARECIDA FERREIRA

Projeto editorial ROBERTO PAES

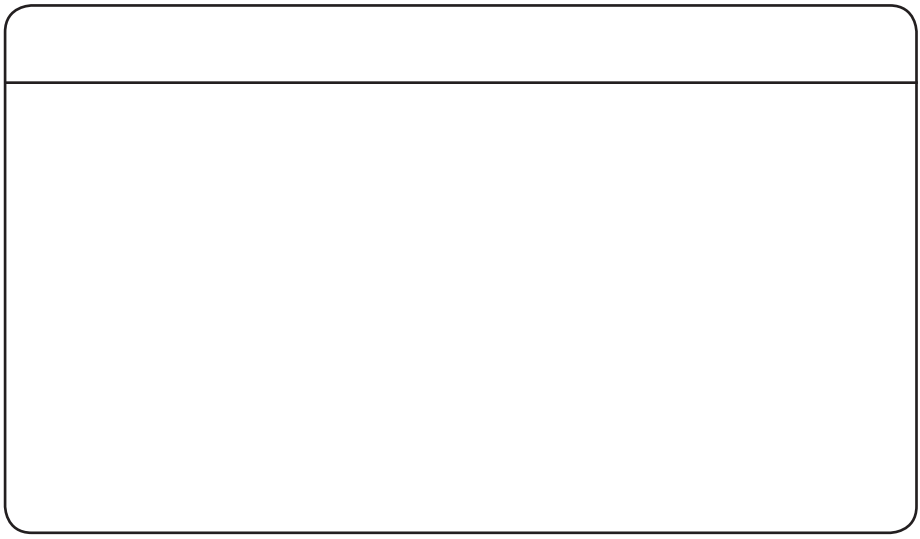
Coordenação de produção RODRIGO AZEVEDO DE OLIVEIRA

Projeto gráfico PAULO VITOR BASTOS

Diagramação FABRICO

Revisão linguística ADERBAL TORRES BEZERRA

Imagem de capa NOME DO AUTOR — SHUTTERSTOCK



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

Prefácio 5

1. Conceitos Introdutórios e Técnicas de Amostragem 7

Breve histórico 8
Definição de Estatística 9
Amostragem 12
Exemplo Prático Envolvendo Técnicas de Amostragem Probabilística 16

2. Distribuição de Frequências e Medidas de Posição Central 27

Distribuição de Frequências 29
Medidas de posição 34

3. Medidas de Ordenamento e Forma, Medidas de Dispersão e Gráficos 51

Medidas de ordenamento 53
Medidas de Dispersão 63
Gráficos 79

4. Distribuições Amostrais e Estimação 99

Conceitos Básicos 100
Estimador de uma Média Populacional 102
Estimador de uma Proporção Populacional 104
Propriedades da Distribuição Normal 106
Distribuições Amostrais 107
Erro Padrão de um Estimador 109
Intervalos de Confiança para a Média Populacional 112
Intervalos de Confiança para a Proporção Populacional 114

5. Distribuição Normal e Teste de Hipótese 121

Variável Aleatória 122
Função Densidade de Probabilidade 123
Modelo Probabilístico para Variáveis Aleatórias Contínuas 123
Teste de Hipótese 130

Prefácio

Prezados(as) alunos (as)

Estatística é uma palavra de origem latina, que significou por muito tempo *ciência dos negócios do Estado*. Ela pode ser vista como uma *Matemática Aplicada*, uma disciplina da área das ciências exatas que tem aplicação em praticamente todas as áreas de estudo. Esse fato serve para desmistificar o temor vivido pelos alunos com relação ao ensino da matemática em si (aquela que nós aprendemos até o ensino médio). As dificuldades enfrentadas e a falta de conexão com a prática são talvez os fatores que mais contribuem para que este temor ocorra.

No entanto, o ensino da Estatística, mesmo provocando sentimentos semelhantes aos estudantes, proporciona a esses uma visão prática do conteúdo que está sendo abordado. Mais que isso, ele possibilita, a quem o está aplicando, a obtenção de importantes informações do fato que está sendo estudado. O conhecimento mínimo em Estatística se tornou pré-requisito para ler um jornal ou uma revista conceituada, pois muitas informações se encontram resumidas em tabelas ou gráficos que grande parte da população não tem condições de interpretar e por isso ignoram (ou não entendem) reportagens importantes para a formação de uma pessoa esclarecida social, econômica e politicamente.

Procuramos, aqui, apresentar a Estatística de forma clara e prática. Não com o intuito de formar especialistas nessa área, mas sim de proporcionar a você, futuro gestor, uma compreensão dos elementos básicos que compõem essa ciência, visando a aplicação na sua área de atuação. Não tivemos a intenção de esgotar o assunto, mas sim de apresentar os elementos necessários para que você realize uma leitura satisfatória da realidade que o cerca e das informações que têm a sua volta.

Muitos dos exemplos aqui apresentados são hipotéticos. São exemplos de situações que ocorrem de forma semelhante na realidade, mas os dados apresentados não são reais, foram criados apenas para ilustrar a aplicação do conteúdo apresentado.

1

Conceitos Introdutórios e Técnicas de Amostragem

1 Conceitos Introdutórios e Técnicas de Amostragem

Nesse primeiro capítulo, apresentaremos conceitos básicos da Estatística e as principais técnicas de amostragens probabilísticas e não probabilísticas.
Em qualquer estudo e/ou pesquisa que envolve a coleta e análise de dados, é imprescindível o conhecimento dos conceitos que serão abordados neste capítulo, para que os resultados obtidos na análise sejam um instrumento confiável para tomadas de decisão.

OBJETIVOS

▪ Identificar os diferentes tipos de variáveis que podem estar presentes em uma pesquisa, compreender a que se destina cada uma das áreas da Estatística e entender as características dos vários tipos de amostragens utilizados para coleta de dados.

REFLEXÃO

Você se lembra de já ter visto nos meios de comunicação informações de pesquisas, por exemplo, eleitorais ou de avaliação de um governo, obtidas através de amostras? Neste capítulo, apresentaremos o conceito de população e amostra e estudaremos como (e por que), na maioria das vezes, fazemos levantamento de dados através de amostras.

1.1 Breve histórico

O interesse por levantamento de dados não é algo que surgiu somente nos dias atuais. Há indícios de que 3000 anos a.C. já se faziam censos na Babilônia, China e Egito. Havia interesse dos governantes das grandes civilizações antigas por informações sobre suas populações e riquezas. Usualmente estas informações eram utilizadas para taxação de impostos e alistamento militar.

A palavra Estatística surgiu, pela primeira vez, no séc. XVIII. Alguns autores atribuem esta origem ao alemão Gottfried Achemmel (1719-1772), que teria utilizado pela primeira vez o vocábulo Estatística, em 1746.

Na sua origem, a Estatística estava ligada ao Estado. Na atualidade, a Estatística não se limita apenas ao estudo de dados demográficos e econômicos. Ela é empregada em praticamente todas as áreas do conhecimento, sempre que estiverem envolvidas coleta e análise de dados.

CONEXÃO

Para saber um pouco sobre a evolução histórica da Estatística, assista o vídeo " História da Estatística" produzido pela Fundação Universidade de Tocantins, disponível em:<http://www.youtube.com/watch?v=jCzMPL7Ub2k&feature=related>.

1.2 Definição de Estatística

A Estatística é uma ciência que trata de métodos científicos para a coleta, organização, descrição, análise e interpretação (conclusão) de um conjunto de dados, visando a tomada de decisões.

Podemos dividir a aplicação da Estatística basicamente em três etapas, que são descritas resumidamente a seguir:

1. Refere-se à coleta de dados, na qual devemos utilizar técnicas estatísticas que garantirão uma amostra representativa da população.
2. Depois dos dados coletados, devemos resumi-los em tabelas de frequências e/ou gráficos e, posteriormente, encontrar as medidas de posição e variabilidade (quantidades). Esta etapa também é conhecida como *Estatística Descritiva* ou *Dedutiva*.
3. Esta etapa envolve a escolha de um possível modelo que explique o comportamento dos dados para posteriormente se fazer a inferência dos dados para a população de interesse. Esta etapa também é chamada de *Estatística Inferencial* ou *Indutiva*. Nesta etapa, se faz necessário um conhecimento mais aprofundado, principalmente no que se refere aos tópicos de probabilidades. A probabilidade fornece métodos para quantificar a incerteza existente em determinada situação, usando ora um número ora uma função matemática.

Podemos citar inúmeros exemplos da Estatística em várias áreas do conhecimento, mas só para convencê-lo da importância das técnicas estatísticas, vamos enumerar alguns deles:

1. Se estamos interessados em abrir um supermercado em um determinado local precisamos saber se fatores como sexo, grau de escolaridade, idade, estado civil, renda familiar, entre outros, interferem na abertura deste supermercado e os tipos de produtos que devem ser priorizados nesse estabelecimento, além de definir as estratégias de *marketing* mais eficientes.
2. Uma empresa, quando está interessada em lançar um novo produto no mercado, precisa saber as preferências dos consumidores. Para isso, é necessário realizar uma pesquisa de mercado.
3. O gestor precisa saber escolher uma amostra representativa de uma população de interesse para não perder muito tempo e, consequentemente, dinheiro da empresa em que trabalha.
4. Para se lançar um novo medicamento no mercado farmacêutico, é necessário a realização de várias experiências. O medicamento deve ser testado estatisticamente quanto à sua eficiência no tratamento a que se destina e quanto aos efeitos colaterais que pode causar, antes de ser lançado no mercado.
5. Para uma empresa, é muito importante fazer previsões de demanda de seus produtos. Para isto existem várias técnicas estatísticas como regressão linear, regressão logística, análise de séries temporais, etc.
6. Controles estatísticos de qualidade (ou controles estatísticos do processo) são indispensáveis em todos os tipos de empresas. Eles são realizados através de um conjunto de técnicas estatísticas, geralmente aplicadas por engenheiros de produção e administradores, para garantir o nível de qualidade exigido para a produção (ou serviço) dentro de uma indústria.

São inúmeras e diversificadas as aplicações de técnicas estatísticas que um gestor pode utilizar. Não conseguiremos falar sobre todas elas, mas apresentaremos os principais conceitos e técnicas que quando utilizados podem auxiliar na tomada de decisões.

Começaremos por apresentar alguns conceitos elementares bastante utilizados no processo estatístico.

Unidade experimental ou de observação é um objeto (isto é, pessoa, objeto, transação ou evento) a partir do qual coletamos os dados. Vamos analisar os exemplos a seguir para entender o conceito dos possíveis tipos de objeto:

- Os eleitores da cidade de São Paulo são pessoas;
- Os carros produzidos em determinado ano por uma montadora são objetos;
- As vendas realizadas durante um mês numa loja de departamento são transações;
- Os acidentes ocorridos em determinada extensão de uma rodovia durante um feriado são eventos.

População é o conjunto total de unidades experimentais que têm determinada característica que se deseja estudar. Uma população pode ser finita ou infinita.

Populações finitas permitem que seus elementos sejam contados. Por exemplo: todas as lojas existentes em determinado *shopping*, todos os alunos matriculados em determinada universidade, todos os veículos licenciados pelo departamento de trânsito em um ano. Indicamos o tamanho de uma população finita por N.

Na prática, uma população que está sendo estudada é usualmente considerada infinita se ela envolve um processo contínuo que torna impossível a listagem ou contagem de cada elemento na população. Por exemplo: quantidades de porções que se pode extrair da água do mar para análise.

Vale ressaltar que, em alguns estudos, a população de interesse é tão grande que só pode ser estudada por meio de amostras. Por exemplo: quantos peixes existem no mar? Esse número é matematicamente finito, mas tão grande que pode ser considerado infinito para qualquer finalidade prática.

Amostra é uma parte da população de interesse que se tem acesso para desenvolver o estudo estatístico. Se a amostra não for fornecida no estudo, devemos retirá-la da população através de técnicas de amostragem adequadas, para que os resultados fornecidos sejam confiáveis.

Estatística Descritiva é a parte da estatística que trata da organização e do resumo do conjunto de dados por meio de gráficos, tabelas e medidas descritivas (quantidades).

Estatística Indutiva é a parte que se destina a encontrar métodos para tirar conclusões (ou tomar decisões) sobre a população de interesse, geralmente, baseado em informações retiradas de uma amostra desta população.

Variável é uma característica da unidade experimental. Vamos estudar dois tipos de variáveis: quantitativas e qualitativas.

Variáveis quantitativas são aquelas cujas respostas da variável são expressas por números (quantidades). Podemos distinguir dois tipos de variáveis quantitativas: quantitativa contínua e discreta.

Variáveis quantitativas contínuas são aquelas que podem assumir, teoricamente, infinitos valores entre dois limites (num intervalo), ou seja, podem assumir valores não inteiros. Por exemplo: altura (em metros) de alunos de uma determinada faixa etária, peso (em kg), salário, etc.

Variáveis quantitativas discretas são aquelas que só podem assumir valores inteiros. Por exemplo: número de filhos por casal, número de livros em uma biblioteca, número de carros vendidos, etc.

Variáveis qualitativas são as variáveis cujas respostas são expressas por um atributo. Podemos distinguir dois tipos de variáveis qualitativas: nominal e ordinal.

Variáveis qualitativas nominais definem-se como aquelas em que as respostas são expressas por um atributo (nome) e esse atributo não pode ser ordenado. Por exemplo: tipo sanguíneo, religião, estado civil, etc.

Variáveis qualitativas ordinais têm suas respostas expressas por um atributo (nome) e esse atributo pode ser ordenado. Por exemplo: grau de instrução, classe social, etc.

1.3 Amostragem

Em qualquer estudo que envolva coleta de dados, sabemos que dificilmente podemos estudar a população de interesse como um todo. Para isso, utilizamos técnicas de amostragem para selecionar amostras que sejam representativas da população de interesse. No próximo item estudaremos algumas técnicas de amostragem muito utilizadas em estudos e/ou pesquisas.

1.3.1 Técnicas de Amostragem

Quando selecionamos uma amostra devemos garantir que esta amostra seja representativa da população, ou seja, no processo de amostragem, a amostra selecionada deverá possuir as mesmas características básicas da população.

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.

Indenpendente do tipo de amostragem, podemos trabalhar com amostragem com reposição ou sem reposição. Na amostragem *com reposição* é permitido que uma unidade experimental seja sorteada mais de uma vez, e na amostragem *sem reposição*, a unidade experimental sorteada é removida da população. Quando pensamos na quantidade de informação contida na amostra, amostrar sem reposição é mais adequado. Mas, amostragem com reposição implica que tenhamos independência entre as unidades experimentais selecionadas, o que facilita o desenvolvimento de propriedades de estimadores que serão abordados mais adiante. Na prática, é comum considerarmos a selação das unidades experimentais como independentes quando pequenas amostras são retiradas de grandes populações. De acordo com (TRIOLA, 2008), uma diretriz comum a ser seguida é:

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).

Agora, vamos estudar alguns tipos de técnicas de amostragens probabilísticas e não probabilísticas.

1.3.1.1 Definições das Técnicas de Amostragem Probabilística (ou Aleatória)

Sempre que possível, devemos escolher trabalhar com amostragem probabilística. Este tipo de amostragem nos garante, com alto grau de confiança, a representatividade da amostra com relação à população que se tem interesse em estudar.

Usaremos N para denotar o tamanho da população e n indicando o tamanho da amostra.

1.3.1.1.1 Amostragem Aleatória Simples

É utilizada quando todos os elementos da população têm a mesma chance (ou probabilidade igual) de pertencer à amostra.

Para trabalhar com a amostragem casual simples devemos conseguir listar a população de 1 a N. Os elementos da população que irão pertencer a amostra serão sorteados de forma aleatória. Sortearemos n números dessa sequência, os quais corresponderão aos elementos sorteados para a amostra.

Exemplo 1.1: Se desejamos, por exemplo, selecionar 50 elementos de uma população de 500 elementos, então "numeramos" a população de 1 a 500 e sorteamos, dessa forma, cada um dos 50 que irão compor a amostra.

1.3.1.1.2 Amostragem Sistemática

Utilizamos amostragem sistemática quando os elementos da população se apresentam ordenados (ou em filas) e a retirada dos elementos da amostra é feita periodicamente.

! ATENÇÃO

Cuidado com ciclos de variação. Às vezes, podem ocorrer ciclos de variação e os elementos sorteados para a amostra terão sempre a mesma característica. Se isto for detectado, o salto poderá ser diversificado, podendo então selecionar, por exemplo, o 3º, o 5º e o 9º elementos, depois novamente conto 3, 5 e 9 e assim por diante até obter a amostra desejada.

Exemplo 1.2: Usando o exemplo anterior, onde a população é composta de 500 elementos ordenados, poderíamos utilizar a amostragem sistemática primeiramente determinando qual o "*salto*" que deverá ser dado. Para isto, fazemos a divisão do tamanho da população pelo tamanho da amostra desejada:

$$\frac{N}{n} = \frac{500}{50} = 10$$

Em seguida, podemos iniciar a amostragem com qualquer indivíduo escolhido (de forma aleatória) entre os 10 primeiros. A partir desse elemento, selecionamos os demais sempre "saltando" de 10 em 10.

1.3.1.1.3 Amostragem por Conglomerados (Clusters)

Em algumas vezes a população se apresenta numa subdivisão em pequenos grupos, chamados *conglomerados*. Neste caso é possível, e até conveniente, fazermos uma amostragem por meio desses conglomerados. Este tipo de amostragem consiste em sortear um número suficiente de conglomerados, cujos elementos constituirão a amostra. Quando um conglomerado é sorteado, todos os elementos dentro dele são selecionados para a amostra. Este tipo de amostragem é muitas vezes utilizado por motivos de ordem prática e econômica.

Exemplo 1.3: Suponhamos que desejamos estudar alguma característica dos indivíduos que moram num determinado bairro de sua cidade. A população de interesse é constituída, portanto, por todos os indivíduos que moram nesse bairro e cada residência constitui um conglomerado. Podemos sortear alguns conglomerados (residências) e cada morador da unidade sorteada fará parte da nossa amostra.

1.3.1.1.4 Amostragem Estratificada

Esta técnica é muito utilizada quando a população é heterogênea ou quando se consegue dividi-la em subpopulações ou estratos. A amostragem estratificada consiste em especificar quantos elementos da amostra serão retirados em cada estrato. O número de elementos sorteados em cada estrato deve ser proporcional ao número de elementos existente no estrato.

Exemplo 1.4: Vamos supor que uma pesquisa tem como objetivo estudar uma determinada característica do povo brasileiro, como por exemplo, a renda familiar. Nesse caso, a população de interesse é constituída por todo cidadão que mora no Brasil. Podemos considerar cada estrato como sendo cada um dos estados brasileiros. Em cada um deles será selecionado um número x de elementos, proporcional à população de cada estado.

1.3.1.2 Técnicas de Amostragem Não Probabilística (ou Não Aleatória)

Somente recomendamos o uso de métodos de amostragem não probabilística nos casos em que é impossível ou inviável a utilização de métodos probabilísticos.

1.3.1.2.1 Amostragem a Esmo ou Sem Norma

É a amostragem em que o pesquisador, para simplificar o processo, procura ser aleatório sem, no entanto, usar algum dispositivo aleatório confiável.

Exemplo 1.5: Suponha que desejamos retirar uma amostra de 50 parafusos de uma caixa contendo 5.000. Nesse caso, poderíamos, ao invés de sortear os parafusos, escolher a esmo aqueles que fariam parte da amostra. Não é um procedimento totalmente aleatório porque, mesmo sem percebermos, poderíamos estar privilegiando alguma parte da caixa, não dando, dessa forma, a mesma chance de participação a qualquer um dos parafusos.

1.3.1.2.2 Amostragem Intencional

Neste caso, o amostrador escolhe deliberadamente os elementos que irão compor a amostra, muitas vezes, por julgar tais elementos bem representativos da população.

Exemplo 1.6: Um diretor de uma instituição de ensino deseja avaliar o quanto determinada disciplina está sendo bem ministrada por seu professor. Para isso, seleciona, para uma entrevista, alguns dos alunos com melhor desempenho nessa disciplina.

1.4 Exemplo Prático Envolvendo Técnicas de Amostragem Probabilística

Exemplo 1.7: O quadro 1.1 lista a idade e a opinião de 50 profissionais de empresas públicas e privadas que estão sendo entrevistados para responder se são "contra" ou "a favor" da inclusão de deficientes visuais e auditivos em suas empresas e em que tipo de empresa trabalha: pública ou privada.

PROFISSIONAIS	IDADE	OPINIÃO	TIPO DE EMPRESA
1	52	contra	Pública
2	22	a favor	pública
3	36	a favor	privada
4	35	a favor	privada
5	35	a favor	privada
6	50	contra	pública
7	44	contra	pública
8	42	contra	pública
9	40	contra	pública
10	45	contra	pública

PROFISSIONAIS	IDADE	OPINIÃO	TIPO DE EMPRESA
11	36	a favor	privada
12	34	a favor	privada
13	23	contra	pública
14	26	a favor	pública
15	28	a favor	pública
16	28	a favor	pública
17	29	a favor	privada
18	30	a favor	privada
19	30	a favor	privada
20	34	a favor	privada
21	38	a favor	privada
22	41	contra	pública
23	42	contra	pública
24	50	contra	pública
25	49	contra	pública
26	38	contra	privada
27	26	a favor	privada
28	29	a favor	privada
29	26	a favor	privada
30	36	a favor	privada
31	27	a favor	privada

PROFISSIONAIS	IDADE	OPINIÃO	TIPO DE EMPRESA
32	32	a favor	privada
33	31	a favor	privada
34	33	a favor	privada
35	33	a favor	privada
36	36	contra	pública
37	34	a favor	pública
38	46	contra	privada
39	44	contra	pública
40	65	contra	pública
41	56	contra	pública
42	52	contra	pública
43	35	a favor	pública
44	24	a favor	privada
45	23	a favor	privada
46	28	a favor	privada
47	30	a favor	privada
48	34	a favor	privada
49	46	contra	pública
50	26	a favor	privada

Quadro 1.1 – Idade e a opinião de profissionais de empresas públicas e privadas com relação a inclusão de deficientes visuais e auditivos

a) Utilizando o quadro acima, retire uma **amostra sistemática** de 10 profissionais, iniciando pelo 3º profissional, e liste o nº do profissional sorteado, a idade, a opinião e o tipo de empresa que ele trabalha.

Resolução:

Dividindo 50 por 10, temos grupos com 5 elementos cada. Se sortearmos o terceiro elemento do primeiro grupo, por exemplo, os participantes da amostra serão os listados abaixo:

PROFISSIONAIS	IDADE	OPINIÃO	TIPO DE EMPRESA
3	36	a favor	privada
8	42	contra	pública
13	23	contra	pública
18	30	a favor	privada
23	42	contra	pública
28	29	a favor	privada
33	31	a favor	privada
38	46	contra	privada
43	35	a favor	pública
48	34	a favor	privada

b) Com a amostra selecionada no item a), calcule:

- a idade média dos profissionais;

Resolução:

$$\bar{x} = \frac{348}{10} = 34,8 \text{ anos}$$

- a porcentagem de profissionais contra a inclusão;

Resolução:

$$\frac{\text{nº de profissionais contra a inclusão na amostra}}{\text{nº total de profissionais na amostra}} = \frac{4}{10} = 0,4$$

ou seja, 40% dos profissionais são contra a inclusão de deficientes visuais ou auditivos nas empresas em que trabalham.

- a porcentagem de profissionais que são da rede pública.

Resolução:

$$\frac{\text{nº de profissionais de empresas na amostra}}{\text{nº total de profissionais na amostra}} = \frac{4}{10} = 0,4$$

ou seja, 40% dos profissionais trabalham em empresas públicas.

c) É possível retirar uma amostra estratificada dos profissionais considerando a variável tipo de empresa? Diga, em poucas palavras, como você procederia neste caso?

Resolução:

Sim, pois é possível identificar dois estratos: empresa pública e privada. O procedimento deve ser: retirar uma amostra proporcional de profissionais de empresas públicas e privadas e depois fazer as análises devidas.

ATIVIDADE

1. Classifique as variáveis abaixo em quantitativas (discretas ou contínuas) ou qualitativas (nominal ou ordinal).
   a) cor dos olhos
   b) número de peças produzidas por hora
   c) diâmetro externo
   d) número de pontos em uma partida de futebol

e) produção de algodão
f) salários dos executivos de uma empresa
g) número de ações negociadas na bolsa de valores
h) sexo dos filhos
i) tamanho de pregos produzidos por uma máquina
j) quantidade de água consumida por uma família em um mês
k) grau de escolaridade
l) nível social
m) tipo sanguíneo
n) estado civil

2. A guerra das ´Colas´ é o termo popular para a intensa competição entre Coca-Cola e Pepsi mostrada em suas campanhas de marketing. As campanhas têm estrelas de cinema e televisão, vídeos de rock, apoio de atletas e afirmações das preferências dos consumidores com base em testes de sabor. Como uma parte de uma campanha de marketing da Pepsi, suponha que 1 000 consumidores de refrigerante sabor cola submetam-se a um teste cego de sabor (isto é, as marcas estão encobertas). Cada consumidor é questionado quanto à sua preferência em relação à marca A ou B.
   a) Descreva a população.
   b) Descreva a variável de interesse.
   c) Descreva a amostra
   d) Descreva a inferência.

3. Suponha que você tenha determinado conjunto de dados e classifique cada unidade da amostra em quatro categorias: A, B, C ou D. Você planeja criar um banco de dados no computador com esses dados e decide codificá-los como A = 1, B = 2, C = 3 e D = 4. Os dados A, B, C e D são qualitativos ou quantitativos? Depois de introduzidos no banco de dados como 1, 2, 3 e 4, os dados são qualitativos ou quantitativos? Explique sua resposta.

4. Os institutos de pesquisa de opinião regularmente fazem pesquisas para determinar o índice de popularidade do presidente em exercício. Suponha que uma pesquisa será conduzida com 2.500 indivíduos, que serão questionados se o presidente está fazendo um bom ou um mau governo. Os 2.500 indivíduos serão selecionados por números de telefone aleatórios e serão entrevistados por telefone.
   a) Qual a população relevante?
   b) Qual a variável de interesse? É qualitativa ou quantitativa?

c) Qual é a amostra?
d) Qual é o interesse da inferência para o pesquisador?
e) Qual é o método de coleta de dados que foi empregado?
f) A amostra em estudo é representativa?

5. Uma população se encontra dividida em quatro estratos. O tamanho de cada estrato é, N1 = 80, N2 = 120, N3 = 60 e N4 = 60 Sabe-se que uma amostragem proporcional foi realizada e dezoito elementos da amostra foram retirados do segundo estrato. Qual o número total de elementos da amostra?

6. Uma pesquisa precisa ser realizada em uma determinada cidade. A amostragem proposta para este problema é a seguinte: dividir a cidade em bairros (pelo próprio mapa da cidade): em cada bairro, sorteia-se certo número de quarteirões proporcional à área do bairro; de cada quarteirão, são sorteadas quatro residências, destas quatro residências, todos os moradores são entrevistados.
   a) Essa amostra será representativa da população ou poderá apresentar algum vício (não confiável)?
   b) Quais tipos de amostragem foram utilizados no procedimento?

7. Uma empresa de seguros mostra que, entre 4000 sinistros reportados à empresa durante um mês, 2700 são sinistros pequenos (inferiores a R$400,00), enquanto os outros 1300 são sinistros grandes (R$400,00 ou mais). Foi extraída uma amostra proporcional de 1% para estimar o valor médio desses sinistros, Os dados estão a seguir, separados por tipo de sinistro.

SINISTROS PEQUENOS	SINISTROS GRANDES
84 330 126 156 90 296 390 132 36 73 55 178 340 82 184 206 44 276 98 124 176 226 58 144 58 166 228	492 710 1744 1298 506 676 982 1720 1510 976 1004 2600 2420

   a) Determine a média de cada uma das amostras (sinistros pequenos e sinistros grandes),
   b) Determine sua média ponderada, tomando como pesos os dois tamanhos de estratos N1 = 2700 e N2 = 1300.

8. Consideremos um estudo realizado em empresas de pequeno e médio porte de uma determinada região composto por 1000 empresas, distribuídas, quanto ao número de funcionários, como mostra a tabela abaixo, e que nesta região sejam amostrados 50 empresas.

Distribuição do nº de empresas de uma região qualquer, quanto ao nº de funcionários.

Nº DE FUNCIONÁRIOS	Nº DE PROPRIEDADES	AMOSTRA ESTRATIFICADA (N = 50)	
		UNIFORME	PROPORCIONAL
0 ├ 20	500		
20 ├ 50	320		
50 ├ 100	100		
100 ├ 200	50		
200 ├ 400	30		
Total	**1000**	**50**	**50**

a) Qual deverá ser o tamanho da amostra dentro de cada estrato no caso uniforme e no proporcional?
b) Determine a média amostral obtida para a amostragem estratificada uniforme e para a amostragem estratificada proporcional. Comente os resultados.

Observação: Amostragem estratificada uniforme é quando retiramos o mesmo número de elementos de cada estrato, independente do tamanho do estrato.

REFLEXÃO

Estamos encerrando nosso primeiro capítulo. Vimos, aqui, alguns conceitos que serão fundamentais na compreensão dos outros conteúdos abordados no livro. Já deve ter dado para perceber que, mesmo estando no início da disciplina, as aplicações práticas que você poderá fazer na sua área de atuação serão muitas. A compreensão e interpretação das mais variadas informações, com as quais nos deparamos em nosso cotidiano, dependem, em parte, do conhecimento de certos elementos estatísticos.

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

LEITURA

Sugerimos que você ouça os áudios que estão no seguinte endereço: <http://m3.ime.unicamp.br/recursos/1252>. No primeiro módulo você conhecerá um pouco da história da Estatística e, no segundo módulo, a história da Probabilidade.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. Estatística aplicada à administração e economia. São Paulo: Pioneira Thomson Learning, 2003.

BUSSAB, Wilton de O.; MORETTIN, Pedro A.. Estatística básica. São Paulo: Saraiva, 2003.

COSTA NETO, Pedro Luiz de Oliveira. Estatística, São Paulo: Edgard Blucher, 2002.

DOWNING, Douglas; CLARK, Jeffrey. Estatística aplicada. São Paulo: Saraiva, 2002.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. Introdução à estatística. Rio de Janeiro: LTC, 2003.

FONSECA, Jairo Simon; MARTINS, Gilberto de Andrade; TOLEDO, Geraldo Luciano. Estatística Aplicada. São Paulo: Atlas, 1985.

LEVIN, Jack; FOX, James Alan. Estatística para ciências humanas. São Paulo: Prentice Hall, 2004.

LEVINE, David M.; BERENSON, Mark L.; STEPHAN David. Estatística: teoria e aplicações. Rio de Janeiro: LTC, 2000.

MARTINS, Gilberto de Andrade. Estatística geral e aplicada. São Paulo: Atlas, 2002.

McClave, James T.; BENSON P. George.; SINCICH, Terry. Estatística para adminstração e economia. São Paulo: Pearson Prentice Hall, 2009.

MEMÓRIA, José M. P. Breve História da Estatística. Disponível em: <http://www.im.ufrj.br/~lpbraga/prob1/historia_estatistica.pdf>. Acesso em: 25 setembro 2014.

MILONE, Giuseppe. Estatística geral e aplicada. São Paulo: Pioneira Thomson Learning, 2004.

TRIOLA, Mario F.. Introdução à estatística. Rio de Janeiro: LTC, 1999.

VIEIRA, Sonia. Elementos de estatística. São Paulo: Atlas, 2003. Disponível em: <http://www.ufrgs.br/mat/graduacao/estatistica/historia-da-estatistica>. Acesso em: 25 set. 2014

NO PRÓXIMO CAPÍTULO

Agora que já sabemos como coletar dados, aprenderemos no próximo capítulo como organizar, resumir e apresentar os dados coletados em distribuições de frequências. Estudaremos, também, como resumir informações de conjuntos de dados numéricos em alguns valores que sejam representativos de todo o conjunto.

2

Distribuição de Frequências e Medidas de Posição Central

2 Distribuição de Frequências e Medidas de Posição Central

Nesse capítulo, estudaremos como organizar os dados numa distribuição de frequências e aprenderemos a resumir conjuntos de dados numéricos em alguns valores representativos de todo conjunto.

Quando realizamos uma coleta de dados, geralmente estamos lidando com uma quantidade muito grande de informações. Portanto, torna-se imprescindível a utilização de certas técnicas visando simplificar a leitura de tais informações. Para que se tenha uma visão do todo (sobre o fenômeno que está sendo estudado) precisamos, por exemplo, dispor as informações em tabelas ou apresentá-las em gráficos. É o que estaremos abordando num primeiro momento. Logicamente, há mais técnicas que podem ser aplicadas, mas elas serão vistas nos próximos capítulos.

OBJETIVOS

- Organizar, resumir e apresentar, através de distribuição de frequências, as informações contidas em grandes conjuntos de dados. Calcular e interpretar as medidas de posição central.

REFLEXÃO

Você se lembra de já ter visto tabelas em jornais, livros ou revistas, em que eram utilizados percentuais para indicar as frequências de ocorrências de respostas em uma pesquisa? Ou com os percentuais referentes à avaliação de um governo? E informações como média salarial de determinada categoria de profissionais, ou ainda, idade média dos estudantes do primeiro ciclo de determinada universidade? Neste capítulo, veremos como (e para quê) construimos tabelas dessa natureza e como calculamos medidas descritivas como a média aritmética.

2.1 Distribuição de Frequências

Para entendermos a ideia de distribuição de frequências, vamos analisar a seguinte situação: quando um pesquisador termina de coletar os dados para sua pesquisa, geralmente fica com muitos questionários em mãos (respondidos pelas pessoas que foram sorteadas para pertencer a sua amostra) ou com os dados digitados em alguma planilha eletrônica. O fato é que os dados "brutos" (sem tratamento) não trazem as informações de forma clara, por isso devemos tabular esses dados. Quando tabulamos os dados estamos resumindo as informações para melhor compreensão da variável em estudo. A esta tabulação damos o nome de *distribuição de frequências* (ou *tabela 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).

Para facilitar a contagem do número de vezes que cada dado ocorre, podemos ordenar os dados. A uma sequência ordenada (crescente ou decrescente) de dados brutos damos o nome de *Rol*.

Os tipos de frequências com os quais iremos trabalhar são:

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

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

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

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

Exemplo 2.1: Com as informações fornecidas na tabela 2.1, vamos indicar e classificar a variável em estudo. Depois, completaremos a distribuição de frequências encontrando a frequência relativa (%).

FAIXA DE RENDA (EM SALÁRIOS MÍNIMOS)	NÚMERO DE OPERÁRIOS (f)	fr (%)
0 ├ 2	43	39,09
2 ├ 4	39	35,45
4 ├ 6	16	14,55
6 ├ 8	8	7,27
8 ├ 10	4	3,64
Total	**110**	**100**

Tabela 2.1 – Distribuição de renda de operários de uma determinada empresa.

Resolução:

A variável em estudo é a renda dos operários de uma determinada empresa. Esta variável é classificada como quantitativa contínua, pois pode assumir qualquer valor dentro de um intervalo numérico.

! ATENÇÃO

Em todos os nossos exemplos, na distribuição de frequências construída com intervalos de classes, vamos considerar que o intervalo de classe é fechado à esquerda e aberto à direita. Por exemplo, no caso dessa tabela, considerando a terceira classe de frequência, podemos dizer que os 16 operários que estão nesta classe recebem de 4 a menos que 6 salários mínimos por mês.

As frequências absolutas (f) são fonecidas no problema. As frequências relativas (f_r(%)) são encontradas dividindo cada frequência absoluta (de cada classe de frequência) pelo total de operários (110) e multiplicando por 100.

Uma distribuição de frequências apresenta, basicamente, as 3 colunas apresentadas na tabela 2.1. Desta maneira, conseguimos organizar de forma resumida um conjunto de dados.

Em alguns estudos podemos ter interesse em outras quantidades relacionadas à tabela, como, por exemplo, a frequência acumulada ou a frequência acumulada (%). Veremos mais adiante que a frequência acumulada é utilizada na construção de um gráfico denominado Ogiva. A tabela 2.2 apresenta a frequência acumulada e a frequência relativa acumulada (%).

FAIXA DE RENDA (EM SALÁRIOS MÍNIMOS)	NÚMERO DE OPERÁRIOS (f)	fr (%)	FREQUÊNCIA ACUMULADA (f_a)	f_{ra} (%)
0 ├ 2	43	39,09	43	39,09
2 ├ 4	39	35,45	82	75,55
4 ├ 6	16	14,55	98	89,09
6 ├ 8	8	7,27	106	96,36
8 ├ 10	4	3,64	110	100,00
Total	**110**	**100**		

Tabela 2.2 – Distribuição das frequências acumuladas da variável faixa de renda.

A coluna frequência acumulada (f_a) decada classe é obtida somando a frequência da respectiva classe com as que lhe são anteriores e a f_{ra}(%) é obtida dividindo a f_a pelo número total de dados e multiplicando por 100.

Para organizar dados de variáveis qualitativas ou quantitativas discretas (cujos valores não estão agrupados em classes) seguimos o mesmo procedimento que foi utilizado na construção da tabela 2.1.

Exemplo 2.2: Uma determinada empresa resolveu traçar o perfil socioeconômico de seus empregados. Uma das variáveis estudadas foi o número de filhos, com idade inferior a 18 anos, de cada um dos empregados. A tabela 2.3 fornece a frequência e a frequência relativa (%) para cada valor obtido.

NÚMERO DE FILHOS	NÚMERO DE OPERÁRIOS (f)	fr (%)
0	0	13,33
1	1	24,44
2	2	28,89
3	3	15,56
4	4	11,11
5	5	2,22
6	6	4,44
Total	**45**	**100,00**

Tabela 2.3 – Distribuição de frequências dos empregrados, segundo o número de filhos.

Para encontrarmos a f_a e a f_{ra} (%) seguimos o mesmo procedimento que foi utilizado na tabela 2.2.

2.1.1 Agrupamento em Classes

Como vimos no exemplo 2.1, para representar a variável contínua "renda", organizamos os dados em classes. Portanto, podemos dizer que a variável renda foi dividida em "5 classes de frequências".

Quando agrupamos em classes de frequências perdemos informações, pois não sabemos exatamente quais são os valores que estão contidos em cada uma das classes (a não ser que seja possível pesquisar esta informação no conjunto de dados brutos). Na análise das distribuições de frequências com intervalos de classes podemos identificar os seguintes valores:

Limite inferior (L_i)**:** é o menor valor que a variável pode assumir em uma classe de frequência;

Limite superior (L_s)**:** 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 (P_m): é a média aritmética entre o L_i e o L_s da mesma classe, ou seja, $Pm = \frac{Li + Ls}{2}$

Amplitude (h): é a diferença entre o L_s e o L_i da classe, ou seja, $h = L_s - L_i$;

Amplitude total (h_t): é a diferença entre o L_s da última classe de frequência e o L_i da primeira classe, ou seja: $h_t = L_s - L_i$.

Na construção de uma distribuição de frequências com intervalos de classes devemos determinar o número de classes que uma tabela deve ter e qual o tamanho (ou a amplitude) destas 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* e será apresentado a seguir.

Considere:

$$k \approx \sqrt{n} \qquad \text{e} \qquad h = \frac{k}{k'} \tag{2.1}$$

onde k é o número de classes que vamos construir na distribuição de frequências; n é o tamanho da amostra que estamos trabalhando; h é a amplitude de cada uma das classes e R é a amplitude total dos dados.

Os valores de k e h devem ser arredondados sempre para o maior valor. Por exemplo, para uma amostra de tamanho n = 50 cujo menor valor é 4 e o maior valor é 445 temos que R = 441 (maior valor – menor valor). O número de classes seria dado por $k \cong \sqrt{n} = \sqrt{50} = 7,07106... \gg 8$ (maior inteiro depois de 7) e a amplitude (tamanho) de cada uma das 8 classes acima deverá ser $h = \frac{R}{k} = \frac{441}{8} = 55,125 \gg 56$ (maior inteiro depois de 55). Ou seja, deveríamos, para este exemplo, montar uma tabela com 8 classes e de amplitude 56. A tabela pode ser iniciada pelo menor valor do conjunto de dados.

Resumindo, para montar uma distribuição de frequências com intervalos de classes devemos:

- Achar o mínimo e o máximo dos dados.
- Determinar as classes de frequências que na verdade nada mais é do que escolher intervalos de mesmo comprimento que cubra a amplitude entre o mínimo e o máximo. Para determinar o número de classes, usaremos $k \approx \sqrt{n}$ e para determinar o "tamanho" das classes usaremos $h = \frac{k}{k'}$

- Contar o número de observações que pertencem a cada intervalo de classe. Esses números são as frequências observadas da classe.
- Calcular as frequências relativas e acumuladas de cada classe.
- De modo geral, a quantidade de classes não deve ser inferior a 5 e nem superior a 25.

Agora, aprenderemos como caracterizar um conjunto de dados através de medidas numéricas que sejam representatativas de todo o conjunto.

2.2 Medidas de posição

As *medidas de posição*, também chamadas de *medidas de tendência central*, têm o objetivo de apresentar um ponto central em torno do qual os dados se distribuem. As mais conhecidas são: a média, a mediana e a moda. Vamos estudar cada uma dessas medidas de posição (estatísticas).

Primeiramente, vamos fazer um estudo para os dados não tabulados, ou seja, quando os dados não estiverem na forma de distribuição de frequência. Em seguida, as mesmas medidas serão calculadas com base em dados tabulados.

2.2.1 Média aritmética

A média aritmética $(\overline{x})$ é a mais comum e mais simples de ser calculada dentre todas as medidas de posição mencionadas.

Para calculá-la, basta fazer a divisão da soma de todos os valores $(x_1, x_2, ..., x_n)$ da variável pelo número total de elementos do conjunto de dados (n):

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

onde:

$\bar{x}$ = a média aritmética;

x_i = os valores da variável;

n = o número de valores no conjunto de dados.

2.2.2 Mediana

A *mediana* é outra medida de posição, dita mais robusta que a média, pois, da forma como ela é determinada, não permite que alguns valores muito altos ou muito baixos interfiram de maneira significativa em seu valor. Desta forma, se o conjunto de dados apresentar alguns poucos valores discrepantes em relação à maioria dos valores do conjunto de dados, em geral, é aconselhável usar a mediana ao invés da média.

! ATENÇÃO

A mediana é a medida de posição mais frequentemente usada quando a variável em estudo for renda (R$), pois algumas rendas extremamente elevadas podem inflacionar a média. Neste caso, a mediana é uma melhor medida de posição central.

A mediana é encontrada *ordenando* os dados do menor para o maior valor e em seguida identificando o valor central destes dados ordenados. É uma medida que divide o conjunto de dados ao meio, deixando a mesma quantidade de valores abaixo dela e acima.

A determinação da mediana difere no caso do tamanho (n) do conjunto de dados ser par ou ímpar. Vejamos a seguir.

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

$$\text{Md} = x_{\left(\frac{n+1}{2}\right)} \tag{2.3}$$

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:

$$\text{Md} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}\right)+1}}{2} \tag{2.4}$$

onde $x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n+1}{2}\right)}$ e $x_{\left(\frac{n}{2}\right)+1}$ indicam as posições onde os dados se encontram.

2.2.3 Moda

A moda de um conjunto de dados é o valor (ou valores) que ocorre com maior frequência. A moda, diferentemente das outras medidas de posição, também pode ser encontrada quando a variável em estudo for qualitativa. Existem conjuntos de dados em que nenhum valor aparece mais vezes que os outros. Neste caso, dizemos que o conjunto de dados ***não apresenta moda.***

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

Por conta das definições diferentes, a ***média***, a ***mediana*** e a ***moda*** fornecem, muitas vezes, informações diferentes sobre o centro de um conjunto de dados, embora sejam todas ***medidas de tendência central***.

No exemplo 2.3 apresentaremos os cálculos das medidas de posição para dados não tabelados (dados brutos).

Exemplo 2.3: Um gerente de banco deseja estudar a movimentação de pessoas em sua agência na segunda semana de um mês qualquer. Ele constata que no primeiro dia entraram 1.348 pessoas, no segundo dia, 1.260 pessoas, no terceiro, 1.095, no quarto, 832 e no último dia do levantamento, 850 pessoas. Encontre a média aritmética, a mediana e a moda para este conjunto de dados e interprete os resultados.

Resolução:

A *média aritmética* é dada por:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1.348 + 1.260 + 1.095 + 832 + 850}{5} = \frac{5.385}{5} = 1.077$$

O número médio de pessoas que entram na agência bancária na segunda semana do mês é de 1.077 pessoas. Isto quer dizer que, alguns dias entram menos que 1.077 e outros dias entram mais, ou seja, 1.077 é um valor em torno do qual o número de pessoas que entram na agência, durante a segunda semana de cada mês, se concentra.

Para encontrar a *mediana*, devemos, primeiramente, ordenar os dados em ordem crescente (pode ser decrescente também):

$$832,\ 850,\ 1095,\ 1.260, 1.348$$

Como a quantidade de dados (n) é um número ímpar, a mediana é exatamente o valor que se encontra no meio do conjunto de dados:

$$Md = x_{\left(\frac{n+1}{2}\right)} = x_{\frac{6}{2}} = x_3 = 1.095 \text{ pessoas}$$

Isto significa que temos o mesmo número de observações menores ou iguais ao valor da mediana e o mesmo número de observações maiores ou iguais ao valor da mediana.

Este conjunto de dados não possui ***moda***, pois não existe nenhum valor que "aparece" com mais frequência que os outros.

Agora, vamos fazer um estudo para os ***dados tabulados***, ou seja, quando os dados **estiverem** na forma de uma distribuição de frequências.

Neste caso, a maneira de se calcular a média aritmética muda um pouco. Como as frequências são números que indicam quantas vezes aparece determinado valor ou quantos valores têm em cada classe de frequência, elas funcio-

narão como “fatores de ponderação”. Estas situações serão apresentadas nos exemplos 2.4 e 2.5, respectivamente.

Média Aritmética

No caso de dados tabulados, o cálculo da média aritmética é:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i} \qquad (2.5)$$

onde:

x_i é o valor da variável (ou o ponto médio de uma classe de frequência);

f_i é a frequência referente a cada valor (ou classe);

$\sum_{i=1}^{k} f_i$ é a soma dos valores das frequências (n).

A expressão (2.5) apresentada acima também é conhecida como fórmula da ***média ponderada.***

No caso de distribuições de frequências que não apresentam intervalos de classes, a *mediana* e a *moda* são encontradas utilizando os conceitos apresentados nos itens 2.2.2 e 2.2.3, respectivamente.

Exemplo 2.4: Em um determinado mês, foi computado o número de faltas ao trabalho, por motivos de saúde, que cada funcionário de uma determinada empresa teve. Os dados estão apresentados na tabela a seguir.

NÚMERO DE FALTAS	*f*
0	31
1	20
2	8
3	2
4	0
5	1
6	1
Total	**63**

Tabela 2.4 – Número de faltas ao trabalho, por motivos de saúde.

Encontre a média aritmética, a mediana e a moda para este conjunto de dados e interprete os resultados.

Resolução:

Média Aritmética

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i \times f_i}{\sum_{i=1}^{k} f_i} = \frac{(0\times 31)+(1\times 20)+(2\times 8)+(3\times 2)+(4\times 0)+(5\times 1)+(6\times 1)}{63} = \frac{53}{63} \cong 0,84$$

ou seja, nesta empresa ocorreram, em média, 0,84 faltas por funcionário, por motivo de saúde.

Mediana

Como os dados estão tabelados, eles já se encontram ordenados. Para ficar mais fácil encontrar o valor da mediana, vamos incluir na distribuição de frequências uma coluna com as frequências acumuladas.

! ATENÇÃO

Quando os dados estiverem dispostos numa distribuição de frequências, o cálculo da média aritmética pode ser feito acrescentando uma coluna na distribuição de frequências. Esta coluna é denominada $x_i \times f_i$ e é obtida multiplicando cada valor da variável (x_i) pela sua respectiva frequência (f_i). A média aritmética é obtida dividindo a soma dos valores desta coluna pela soma dos valores da coluna da frequência.

NÚMERO DE FALTAS	f	f_a
0	31	31
1	20	51
2	8	59
3	2	61
4	0	61
5	1	62
6	1	63
Total	**63**	

Para encontrar o valor da mediana, seguimos os seguintes passos:

1º Passo: identificaremos a frequência acumulada imediatamente superior à metade do somatório das frequências absolutas:

2º Passo: a mediana será o valor da variável associado à frequência acumulada imediatamente superior ao valor encontrado no 1º Passo. Então, a frequência acumulada imediatamente superior a 31,5 é $f_a = 51$. . Portanto, o valor da mediana é o valor da variável associado à $f_a = 51$, ou seja,

$$Md = 1\ falta$$

Neste conjunto de dados, pelo menos 50% das observações são maiores ou iguais a 1 falta.

No caso do valor $\frac{\sum_{i=1}^{k} f_i}{2}$ 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 cujo $\frac{\sum_{i=1}^{k} f_i}{2} = f_a$ e o valor da variável $x_i + 1$ será aquele que está imediatamente após xi na distribuição de frequência.

Moda

A resposta que aparece com maior frequência neste conjunto de dados é o 0 (com frequência 31), ou seja, é mais frequente encontrar funcionários que não faltam.

! ATENÇÃO

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.

No caso do exemplo 2.5 veremos que os dados estão agrupados em intervalos de classes. Quando o conjunto de dados for apresentado sob a forma agrupada perdemos a informação dos valores das observações. Neste caso, vamos supor que todos os valores dentro de uma classe tenham seus valores iguais ao ponto médio desta classe.

Os cálculos da média, da moda e da mediana para distribuição de frequências agrupadas em classes estão apresentados a seguir.

Vale ressaltar que, sempre que possível, as medidas de posição e dispersão devem ser calculadas antes dos dados serem agrupados.

Exemplo 2.5: A tabela abaixo apresenta a distribuição de frequências do tempo de vida de 60 componentes eletrônicos (medido em dias) submetidos à experimentação num laboratório especializado. Calcular as medidas de posição.

TEMPO DE VIDA (DIAS)	f	PONTO MÉDIO (x_i)
3 ├ 18	3	10,5
18 ├ 33	4	25,5
33 ├ 48	4	40,5
48 ├ 63	8	55,5
63 ├ 78	10	70,5
78 ├ 93	28	85,5
93 ├ 108	2	100,5
108 ├ 123	1	115,5
Total	**60**	

Tabela 2.5 – Tempo de vida de componentes eletrônicos.

Resolução:

Neste tipo de tabela, como temos classes de frequências, devemos encontrar um valor que represente cada classe, para que possamos efetuar os cálculos. Por exemplo, considerando a primeira classe de frequência,

TEMPO DE VIDA (DIAS)	f	PONTO MÉDIO (x_i)
3 ├ 18	3	10,5

sabemos que 3 componentes eletrônicos tiveram tempo de vida entre 3 e 18 dias, porém, não sabemos exatamente qual foi o tempo de vida de cada um. Se considerarmos o limite inferior da classe (3) para efetuarmos os cálculos estaremos subestimando as estimativas. Por outro lado, se considerarmos o limite superior da classe (18) estaremos superestimando as estimativas. Portanto, vamos utilizar o ponto médio de cada classe para podermos fazer os cálculos sem grandes prejuízos. A terceira coluna da tabela acima contém os **pontos médios** calculados para cada intervalo de classe. O valor do ponto médio passa a ser o nosso valor x_i a ser utilizado nos cálculos. Vamos aprender como se faz:

Média Aritmética

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

$$= \frac{(10,5 \cdot 3) + (25,5 \cdot 4) + (40,5 \cdot 4) + (55,5 \cdot 8) + (70,5 \cdot 10) + (85,5 \cdot 28) + (100,5 \cdot 2) + (115,5 \cdot 1)}{60}$$

$$= \frac{4155}{60} = 69,25$$

Podemos dizer, através da média aritmética, que os componentes eletrônicos têm uma duração média de 69 dias e 6 horas (69,25 dias).

Mediana

Como os dados estão tabelados em classes de frequências, calculamos a mediana através da seguinte fórmula:

$$M_e = X_e + \frac{h \times (X_m - F_{iaa})}{F} \qquad (2.6)$$

onde

X_e o limite inferior da classe que contém a mediana;

X_m: metade do valor da frequência total;

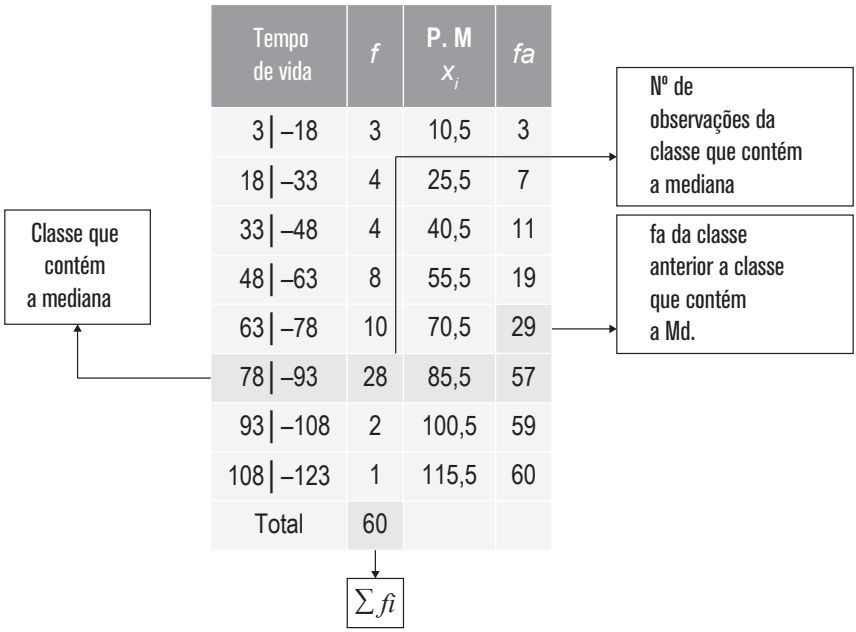
F_{iaa}: frequência acumulada da classe anterior à classe que contém a mediana;

F_i: número de observações na classe que contém a mediana;

h: amplitude da classe que contém a mediana.

No cálculo da mediana para os dados da tabela 2.5 temos que primeiramente encontrar a classe que contém a mediana. Esta classe corresponde à classe associada à frequência acumulada imediatamente superior à $\frac{\sum f_i}{2}$.

Como $\frac{\sum f_i}{2} = \frac{60}{2} = 30$, temos que a classe que contém a mediana é de $78 \vdash 93$ (pois $f_a = 57$).



Além disso, temos:

X_e 78;

X_m: 30;

F_{iaa}: 29;

F_i: 28;

h: 15

Agora, basta substituirmos todos os valores encontrados na fórmula 2.6 e encontramos o valor da mediana:

$$M_e = 78 + \frac{15 \times (30 - 29)}{28} = 28 + 0.54 \text{ @ } 78.5$$

Através da mediana podemos dizer que pelo menos 50% dos componentes eletrônicos avaliados têm duração igual ou inferior a 78 dias e 12 horas.

Para calcularmos a moda para distribuições de frequências com intervalos de classes, utilizaremos a seguinte fórmula:

$$Mo = X_o + \frac{h \times (F_m - F_a)}{2F_m - (F_a + F_p)}$$

onde (2.7)

X_o o limite inferior da classe que contém a moda;
F_m: frequência máxima;
F_a: frequência anterior à frequência máxima;
F_p: frequência posterior à frequência máxima;
h: amplitude da classe que contém a moda.

No cálculo da moda para os dados da Tabela 2.5 temos que primeiramente encontrar a classe que contém a moda. Esta classe corresponde à classe que possui a frequência máxima. Então, a classe que contém a moda é de 78 ⊣ 93 (pois $f = 28$).

Além disso, temos:
X_o 78;
F_m: 28;
F_a: 10;
F_p: 2;
h: 15.

Agora, basta substituirmos todos os valores encontrados na fórmula 2.7 e encontramos o valor da moda:

$$Mo = 78 + \frac{15 \times (28 - 10)}{2\,(28) - (10 + 2)} = 78 + \frac{270}{44} \approx 84{,}1$$

Portanto, é comum encontrar componentes eletrônicos que durem, aproximadamente, 84 dias e 2 horas.

CONEXÃO

Sugerimos os vídeos: "Novo Telecurso – E. Fundamental – Matemática – Aula 34 (parte 1)" e "Novo Telecurso – E. Fundamental – Matemática – Aula 34 (parte 2)" disponíveis, respectimante em <http://www.youtube.com/watch?v=SyWbYOtAIYc&NR=1> e <http://www.youtube.com/watch?v=ejMyWfuSO5k>. que apresenta de modo bem prático a utilização das medidas de posição.

ATIVIDADE

1. Abaixo temos as idades dos funcionários de uma determinada empresa. Construir uma distribuição de frequências, agrupando os dados em classes.

   OBS.: A tabela de distribuição de frequências deve ser completa com f, f_r e f_a.

   Idades (dados brutos)

48	28	37	26	29	59	27	28	30	40	42	35	23	22	31
21	51	19	27	28	36	25	40	36	49	28	26	27	41	29

   Baseado na distribuição de frequências construída, responda:
   a) Quantos são os funcionários com idade inferior a 33 anos?
   b) Que porcentagem de funcionários tem idade igual ou superior a 47 anos?
   c) Quantos são os funcionários com idade maior ou igual a 26 anos e não tenham mais que 40 anos?
   d) Qual a porcentagem de funcionários com idade abaixo de 40 anos?
   e) Qual a porcentagem de funcionários que têm no mínimo 40 anos?

2. Um consultor estava interessado em saber quanto, geralmente, cada pessoa gastava em um determinado supermercado no primeiro sábado após receberem seus pagamentos (salários). Para isso ele entrevistou 50 clientes que passaram pelos caixas entre 13h e 18h, e anotou os valores gastos por cada um deles. Estes valores estão listados abaixo:

4,89	11,00	5,60	73,85	24,83	98,00	186,00	234,87	58,00	198,65
223,86	341,42	94,76	445,76	82,80	35,00	455,00	371,00	398,60	234,00
64,90	54,98	48,80	68,90	120,32	126,98	76,43	6,35	9,98	12,68
243,00	18,65	134,90	11,10	321,09	290,76	74,00	48,80	74,52	138,65
26,00	210,13	15,78	197,45	75,00	76,55	32,78	166,09	105,34	99,10

Analisando o conjunto de dados, responda os seguintes itens:

a) Qual é a variável em estudo? Classifique-a.

b) Construa uma tabela de frequências a partir do conjunto de dados brutos.

3. Os dados abaixo referem-se ao número de horas extras de trabalho que uma amostra de 64 funcionários de uma determinada empresa localizada na capital paulista.

10	10	12	14	14	14	15	16
18	18	18	18	18	19	20	20
20	20	20	21	22	22	22	22
22	22	22	22	22	22	22	22
23	23	24	24	24	24	24	24
24	25	25	25	25	26	26	26
26	26	26	27	27	27	28	28
29	30	30	32	35	36	40	41

Pede-se:

a) Calcule e interprete as seguintes medidas descritivas calculadas para os dados brutos (dados não tabulados): média aritmética; mediana; moda.

b) Construir uma distribuição de frequências completa (com freq. absoluta, freq. relativa, freq. acumulada e ponto médio).

c) Com a tabela construída no item b), encontre as seguintes medidas: média aritmética, mediana, moda. Interprete os resultados.

4. Os dados abaixo representam as vendas mensais (em milhões de reais) de vendedores de gênero alimentícios de uma determinada empresa.

VENDAS MENSAIS (EM MILHÕES DE REAIS)	NÚMERO DE VENDEDORES
0 |– 1	6
1 |– 2	12
2 |– 3	20
3 |– 4	48
4 |– 5	14
5 |– 6	10
Total	**110**

a) Qual a variável em estudo? Que tipo de variável é esta?
b) Encontre a média, mediana e moda e interprete os resultados.
c) Qual a porcentagem de vendedores com vendas mensais inferior a 2 milhões de reais?
d) Qual a porcentagem de vendedores com vendas mensais superior a 4 milhões de reais?
e) Qual a porcentagem de vendedores com vendas mensais entre 3 (inclusive) e 5 (exclusive) milhões de reais?
f) Qual a porcentagem de vendedores que vendem, pelo menos, 3 milhões de reais mensais?

5. Numa pesquisa realizada com 91 famílias, levantaram-se as seguintes informações com relação ao número de filhos por família:

NÚMERO DE FILHOS	0	1	2	3	4	5
FREQUÊNCIA DE FAMÍLIAS	19	22	28	16	2	4

Calcule e interprete os resultados da:
a) média aritmética
b) mediana
c) moda

6. Define-se a média aritmética de n números dados como o resultado da divisão por n da soma dos n números dados. Sabe-se que 4,2 é a média aritmética de 2.7; 3.6; 6.2; e x. Determine o valor de x.

LEITURA

Recomendamos a leitura do texto "Como analisar de forma simples um grande número de dados?", disponível no endereço: <http://www.klick.com.br/conteudo/pagina/0,6313,POR-1453--1453,00.html>, que aborda de maneira clara alguns procedimentos que podem ser utilizados quando nos deparamos com situações em que precisamos resumir as informações de grandes conjuntos de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. Estatística aplicada à administração e economia. São Paulo: Pioneira Thomson Learning, 2003.

BUSSAB, Wilton de O.; MORETTIN, Pedro A.. Estatística básica. São Paulo: Saraiva, 2003.

COSTA NETO, Pedro Luiz de Oliveira. Estatística, São Paulo: Edgard Blucher, 2002.

DOWNING, Douglas; CLARK, Jeffrey. Estatística aplicada. São Paulo: Saraiva, 2002.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. Introdução à estatística. Rio de Janeiro: LTC, 2003.

TRIOLA, Mario F.. Introdução à estatística. Rio de Janeiro: LTC, 1999.

VIEIRA, Sonia. Elementos de estatística. São Paulo: Atlas, 2003.

NO PRÓXIMO CAPÍTULO

Se até agora vimos como organizar os dados (informações) em distribuições de frequências e como resumir as informações numéricas em medidas de posição central, iremos incrementar esse processo através da inserção de medidas de ordenamento e dispersão. As medidas de ordenamento nos fornecem uma ideia sobre a distribuição dos dados ordenados e apresentam a vantagem de não serem afetadas pela forma da distribuição dos dados ou por valores discrepantes. Para que tenhamos informações mais completas do conjunto de dados, é necessário estudar a sua variabilidade. As estatísticas que têm essa função são denominadas medidas de variabilidade ou de dispersão. Finalizaremos o capítulo apresentando vários tipos de gráficos que representam de maneira adequada as informações contidas em um conjunto de dados.

3

Medidas de Ordenamento e Forma, Medidas de Dispersão e Gráficos

3 Medidas de Ordenamento e Forma, Medidas de Dispersão e Gráficos

Nesse terceiro capítulo estudaremos, primeiramente, as medidas de ordenamento, que são usadas para comparação de valores dentro do conjunto de dados. Estas medidas apresentam a vantagem de não serem afetadas pela presença de valores extremos no conjunto de dados. Estudaremos, também, as medidas de dispersão, que servem para indicar o quanto os dados se apresentam dispersos em torno da região central. Fornecem, portanto, o grau de variação existente no conjunto de dados. Dois ou mais conjuntos de dados podem, por exemplo, ter a mesma média, porém, os valores poderão estar muito mais dispersos num conjunto do que no outro. Ou seja, podem ter maior ou menor grau de homogeneidade.
E, para finalizar, apresentaremos os principais tipos de gráficos utilizados para representar a distribuição de uma variável.

OBJETIVOS

- Calcular e interpretar as medidas de ordenamento e de dispersão;
- Saber escolher representações gráficas mais apropriadas para variáveis qualitativas e quantitativas.

REFLEXÃO

Você se lembra de, após ter feito uma prova bimestral, algum professor ter informado que o desempenho médio da turma ficou em torno de 7,2? Para efeito de comparação, ele também pode ter calculado a média de outra turma, e verificado que o desempenho médio também ficou em torno de 7,2. Surgem os seguintes questionamentos:

- será que as notas das duas turmas foram iguais?
- será que as notas das turmas estão próximas da média ou dispersas?

Utilizaremos os conceitos deste capítulo para responder a estas perguntas.

3.1 Medidas de ordenamento

Os quartis, decis e percentis são muito similares à mediana, uma vez que também subdividem a distribuição de dados de acordo com a proporção das frequências observadas.

Já vimos que a mediana divide a distribuição em duas partes iguais, então, 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 igual à mediana) e Q_3 o que separa os 75% valores inferiores dos 25% superiores. Não há um consenso universal sobre um procedimento único para o cálculo dos quartis, e diferentes programas de computador muitas vezes produzem resultados diferentes.

! ATENÇÃO

Perceba que o 2º quartil, o 5º decil e o 50º percentil representam a própria mediana, ou seja, todas estas medidas separatrizes ($\boldsymbol{Q}_2$, $\boldsymbol{D}_5$, e $\boldsymbol{P}_{50}$), dividem a distribuição dos dados ao meio, deixando aproximadamente 50% dos dados abaixo delas e 50% acima.

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

As medidas separatrizes, geralmente, só são calculadas para grandes quantidades de dados.

No Excel, por exemplo, temos a opção de pedir o cálculo de tais medidas.

Com os cálculos dos quartis, juntamente com os valores mínimo e máximo do conjunto de dados, podemos construir um gráfico chamado ***desenho esquemático ou boxplot.*** A análise deste gráfico é bastante útil no sentido de informar, entre outras coisas, a variabilidade e a simetria dos dados.

CONEXÃO

Para se entender quais são os procedimentos utilizados na construção de um boxplot, bem como sua interpretação, leia o texto: "Diagramas de Caixa (Boxplots)" em: TRIOLA, Mario F. Introdução à estatística. 10.ed. Rio de Janeiro: LTC, 2008. pp. 98 a 102.

3.1.1.1 Quartis

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

Podemos obter os quartis utilizando a seguinte fórmula:

$$Q_{nq} = x_{\left(\frac{nq \times n}{4} + \frac{1}{2}\right)} \tag{3.1}$$

Onde

Q: quartil que se deseja obter;

nq: número do quartil que se deseja obter (1, 2 ou 3);

x: elemento do conjunto de dados ordenados;

n: tamanho da amostra.

Exemplo 3.1: Um escritório que presta consultoria em administração levantou os tempos de espera de pacientes que chegam a uma clínica de ortopedia para atendimento de emergência. Foram coletados os seguintes tempos, em minutos, durante uma semana. Encontre os quartis.

2 5 10 11 3 14 8 8 7 12 3 4 7 3 4 2 6 7

Resolução:

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

2 2 3 3 3 4 4 5 6 7 7 7 8 8 10 11 12 14

- **Primeiro quartil (Q_1)**

$$Q_1 = x_{\left(\frac{1\times 18}{4}+\frac{1}{2}\right)}$$
$$Q_1 = x_5$$

Portanto, o primeiro quartil é o elemento que está na quinta posição, no conjunto de dados ordenados. Então:

$$Q_1 = 3$$

Pelo menos, 25% das observações são menores ou iguais a 3 minutos.

- **Segundo quartil (Q_2)**

$$Q_2 = x_{\left(\frac{2\times 18}{4}+\frac{1}{2}\right)}$$
$$Q_2 = x_{9,5}$$

Portanto, o segundo quartil está entre os elementos que ocupam a nona e a décima posição do conjunto de dados orenados. Para encontrá-lo, fazemos a média entre os valores que estão nestas posições, ou seja:

$$Q_2 = \left(\frac{6+7}{2}\right) = 6,5$$

Pelo menos, 50% das observações são menores ou iguais a 6,5 minutos.

- **Terceiro quartil (Q_3)**

$$Q_3 = x_{\left(\frac{3\times 18}{4}+\frac{1}{2}\right)}$$
$$Q_3 = x_{14}$$

Portanto, o terceiro quartil é o elemento que ocupa a décima quarta posição do conjunto de dados ordenados. Então:

$$Q_3 = 8$$

Pelo menos, 25% das observações são maiores ou iguais a 8 minutos.

3.1.1.2 Decis

Os decis dividem o conjunto de dados ordenados em 10 partes. Eles podem ser obtidos através da seguinte fórmula:

$$D_{nq} = x_{\left(\frac{nd \times n}{10} + \frac{1}{2}\right)} \tag{3.2}$$

Onde
D: decil que se deseja obter;
nd: número do decil que se deseja obter (1, 2 ou 3);
x: elemento do conjunto de dados ordenados;
n: tamanho da amostra.

Exemplo 3.2: Vamos utilizar os dados do Exemplo 3.1 para encontrar o oitavo decil:

2 5 10 11 3 14 8 8 7 12 3 4 7 3 4 2 6 7

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

2 2 3 3 3 4 4 5 6 7 7 7 8 8 10 11 12 14

- **Oitavo decil (D_8)**

$$D_8 = x_{\left(\frac{8 \times 18}{10} + \frac{1}{2}\right)}$$

$$D_8 = x_{14,9}$$

Portanto, o oitavo decil corresponde ao elemento 14,9 no conjunto de dados ordenados, situados entre as posições 14 e 15. Como o oitavo decil não está exatamente no meio de dois elementos ($x_{14,9}$), obtemos este elemento seguindo o seguinte raciocínio:

$$\frac{x - 8}{0,9} = \frac{2}{1}$$

Usando a propriedade fundamental da proporção ("multiplicação em cruz"), obtemos:

$$x - 8 = 1,8$$
$$x = 1,8 + 8$$
$$x = 9,8$$

Aproximadamente 20% das observações são maiores ou iguais a 9,8 minutos.

! ATENÇÃO

A proporção montada neste exemplo foi obtida através do seguinte raciocínio: a diferença entre o elemento x que queremos encontrar e o elemento que está na décima quarta posição (8) **está para** a diferença entre as posições 14,9 e 14 (14,9 – 14) **assim como** a diferença entre os elementos 10 e 8 **está para** a diferença entre as posições 15 e 14 (15-14).

3.1.1.3 Percentis

Os percentis dividem o conjunto de dados ordenados em 100 partes. Eles são obtidos de maneira similiar aos quartis e decis, ou seja:

$$P_{nq} = x_{\left(\frac{np \times n}{100} + \frac{1}{2}\right)}$$

Onde

P: percentil que se deseja obter;

np: número do percentil que se deseja obter (1, 2, ..., 99);

x: elemento do conjunto de dados ordenados;

n: tamanho da amostra.

Exemplo 3.3: Vamos utilizar os dados do Exemplo 3.1 para encontrar o trigésimo quinto percentil:

2 5 10 11 3 14 8 8 7 12 3 4 7 3 4 2 6 7

Resolução:

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

2 2 3 3 3 4 4 5 6 7 7 7 8 8 10 11 12 14

- **Trigésimo quinto percentil (P_{35})**

$$P_{35} = x_{\left(\frac{35\times 18}{10}+\frac{1}{2}\right)} \qquad (3.3)$$

$$P_{35} = x_{6,8}$$

Portanto, o trigésimo quinto percentil corresponde ao elemento 6,8 no conjunto de dados ordenados, situados entre as posições 6 e 7. A sexta e a sétima posição são ocupadas pela observação 4, portanto:

$$P_{35} = 4$$

Então, aproximadamente 35% das observações são menores ou iguais a 4 minutos.

3.1.2 Cálculo dos quartis e percentis para dados agrupados em classes

Podemos encontrar os quartis e decis para dados agrupados em classes, utilizando a fórmula dos percentis para dados agrupados:

$$P_{nx} = L_i + \left(\frac{\frac{nx \times x}{100} - Fa_{ant}}{F_i}\right) \times h \qquad (3.4)$$

onde

n: número de elementos da amostra.

nx: 1, 2, ..., 99;

L_i: limite inferior da classe encontrada;

h: amplitude do intervalo;

Fa_{ant}: frequência acumulada anterior à classe P_i.

F_i: frequência absoluta da classe encontrada P_i.

CONEXÃO

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

Exemplo 3.4: Vamos utilizar os dados do Exemplo 2.5 para encontrar o terceiro quartil, o quinto decil e o décimo quinto percentil.

TEMPO DE VIDA (DIAS)	f	P. M. x_1	f_a
3 ├ 18	3	10,5	3
18 ├ 33	4	25,5	7
33 ├ 48	4	40,5	11
48 ├ 63	8	55,5	19
63 ├ 78	10	70,5	29
78 ├ 93	28	85,5	57
93 ├ 108	2	100,5	59
108 ├ 123	1	115,5	60
Total	**60**		

Resolução

- **Terceiro Quartil**

Primeiramente, temos que encontrar a classe que contém o terceiro quartil. Esta classe corresponde à classe associada à frequência acumulada imediatamente superior à $\frac{75 \cdot \sum f_i}{100}$.

Como $\frac{75 \cdot \sum f_i}{100} = \frac{75 \cdot 60}{100} = 45$, temos que a classe que contém o primeiro quartil é de 78 ├ 93 (pois $f_a = 57$).

Além disso, temos:

60: número de elementos da amostra.

nx: 75;

L_i: 78;

h: 15;

Fa_{ant}: 29.

F_i: 28.

Agora, basta substituirmos todos os valores encontrados na fórmula 3.4 e encontrar o valor do terceiro quartil:

$$P_{nx} = L_i + \left(\frac{\frac{nx \cdot n}{100} - Fa_{ant}}{F_i} \right) \cdot h$$

$$P_{75} = 78 + \left(\frac{\frac{75 \cdot 60}{100} - 29}{28} \right) \cdot 15$$

$$P_{75} = 78 + \left(\frac{45 - 29}{28} \right) \cdot 15 = 78 + 8,57 \approx 86,6$$

De acordo com o resultado obtido podemos esperar que aproximadamente 25% dos dados são maiores ou iguais a 86,6, ou seja, aproximadamente 25% dos componentes eletrônicos têm duração superior a 86 dias e 14 horas.

- **Quinto decil**

Primeiramente, temos que encontrar a classe que contém o quinto decil. Esta classe corresponde à classe associada à frequência acumulada imediatamente superior à $\frac{50 \cdot \sum f_i}{100}$.

Como $\frac{50 \cdot \sum f_i}{100} = \frac{50 \text{ x } 60}{100} = 30$, temos que a classe que contém o quinto decil é de 78 ├ 93 (pois $f_a = 57$).

Além disso, temos:

60: número de elementos da amostra.

nx: 50;

L_i: 78;

h: 15;

Fa_{ant}: 29.

F_i: 28.

Agora, basta substituirmos todos os valores encontrados na fórmula 3.4 e encontrar o valor do quinto percentil:

$$P_{nx} = L_i + \left(\frac{\frac{nx \cdot n}{100} - Fa_{ant}}{F_i} \right) \cdot h$$

$$P_{50} = 78 + \left(\frac{\frac{50 \cdot 60}{100} - 29}{28} \right) \cdot 15$$

$$P_{50} = 78 + \left(\frac{30 - 29}{28} \right) \cdot 15 = 78 + 0,54 \approx 78,5$$

De acordo com o resultado obtido podemos esperar que aproximadamente 50% dos dados são menores ou iguais a 78,5, ou seja, aproximadamente 50% dos componentes eletrônicos têm duração inferior a 78 dias e 12 horas.

Vale ressaltar que o P_{50} = Md e este cálculo já havia sido feito no exemplo 2.5.

- **Décimo quinto percentil**

Primeiramente, temos que encontrar a classe que contém o décimo quinto percentil. Esta classe corresponde à classe associada à frequência acumulada imediatamente superior à $\frac{15 \cdot \sum f_i}{100}$.

Como $\frac{50 \cdot \sum f_i}{100} = \frac{15 \times 60}{100} = 9$, temos que a classe que contém o quinto decil é de 33 ├ 48 (pois $f_a = 11$).

Além disso, temos:
60: número de elementos da amostra.
nx: 15;
L_i: 33;
h: 15;
Fa_{ant}: 7.
F_i: 4.

Agora, basta substituirmos todos os valores encontrados na fórmula 3.4 e encontrar o valor do décimo quinto percentil:

$$P_{nx} = L_i + \left(\frac{\frac{nx \cdot n}{100} - Fa_{ant}}{F_i} \right) \cdot h$$

$$P_{15} = 33 + \left(\frac{\frac{15 \cdot 60}{100} - 7}{4} \right) \cdot 15$$

$$P_{15} = 33 + \left(\frac{9 - 7}{4} \right) \cdot 15 = 33 + 7{,}5 = 40{,}5$$

De acordo com o resultado obtido podemos esperar que aproximadamente 15% dos dados são menores ou iguais a 40,5, ou seja, aproximadamente 15% dos componentes eletrônicos têm duração inferior a 40 dias e 12 horas.

3.2 Medidas de Dispersão

Para termos uma ideia da importância de se conhecer as medidas de dispersão para a tomada de decisões, vamos analisar o exemplo a seguir.

Exemplo 3.5: Imagine que estamos interessados em fazer uma viagem para Honolulu (Havaí) ou Houston (Texas) e para arrumar as malas necessitamos saber se a localidade a ser visitada faz calor, faz frio ou ambos. Se tivéssemos apenas a informação de que a temperatura média diária (medida durante um ano) das duas localizações fosse igual a 25º C, poderíamos colocar na mala apenas roupas de verão? A resposta é não. Por exemplo, se estivéssemos interessados em viajar para o Havaí (em Honolulu), poderíamos levar apenas roupas de verão, pois a temperatura mínima observada durante um ano foi de 21ºC e a máxima foi de 29ºC. Porém, se resolvermos ir ao Texas (Houston), devemos tomar cuidado com a época, pois as temperaturas, durante um ano, variaram de 4ºC (mínima) a 38ºC (máxima). Com estas informações concluímos que as temperaturas em Honolulu variam pouco em torno da média diária, ou seja, podemos levar uma mala apenas com roupas leves. Porém, em Houston, as temperaturas variam muito, com períodos de muito frio ou muito calor. Portanto, para ir à Houston sem perigo de sofrer com a temperatura, devemos analisar o período do ano para saber se a temperatura estará alta ou baixa.

Percebemos, através deste exemplo bem simples, que uma simples medida de dispersão (a amplitude, por exemplo) já ajudaria muito a tomar certos cuidados com a arrumação das bagagens.

Veremos, nos próximos itens, como calcular e interpretar as seguintes *medidas de dispersão*: amplitude, amplitude interquartil, desvio-padrão, variância e coeficiente de variação.

Primeiramente, vamos apresentar os cálculos das medidas de dispersão para dados ***não-tabulados***, ou seja, quando os dados ***não estiverem*** na forma de distribuição de frequências.

! ATENÇÃO

As medidas de dispersão indicam o grau de variabilidade das observações. Estas medidas possibilitam que façamos distinção entre conjuntos de observações quanto à sua homogeneidade. Quanto menor as medidas de dispersão, mais homogêneo é o conjunto de dados.

3.2.1 Amplitude Total

A amplitude total é a diferença entre o maior e o menor valor observado no conjunto de dados, ou seja:

$$R = x_{(máximo)} - x_{(mínimo)} \tag{3.5}$$

A amplitude não é uma medida muito utilizada, pois só leva em conta dois valores de todo o conjunto de dados e é muito influenciada por valores extremos. No próximo item estudaremos uma medida de dispersão mais resistente a valores extremos.

3.2.2 Amplitude Interquartil

A amplitude interquartil, ou distância interquartil, é uma medida de variabilidade que não é facilmente influenciada por valores discrepantes no conjunto de dados. Ela engloba 50% das observações centrais do conjunto de dados e seu cálculo é definido como:

$$\textit{Amplitude do interquartil} = Q_3 - Q_1 \tag{3.6}$$

Agora, vamos estudar uma medida de dispersão muito utilizada e que leva em conta todos os valores do conjunto de dados: o desvio-padrão.

3.2.3 Desvio-Padrão

Primeiramente, vamos entender qual é a definição da palavra desvio em estatística. Desvio nada mais é do que a distância entre qualquer valor do conjunto de dados em relação à média aritmética deste mesmo conjunto de dados.

Existem várias medidas de dispersão que envolvem os desvios. São elas: o desvio-padrão (mais utilizada), a variância e o coeficiente de variação.

O desvio-padrão é a medida mais utilizada na comparação de diferenças entre grupos, por ser mais precisa e estar na mesma medida do conjunto de dados. Matematicamente, sua fórmula é dada pela raiz quadrada da média aritmética aproximada dos quadrados dos desvios, ou seja:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} \quad (3.7)$$

onde x_i é cada uma das observações do conjunto de dados, $\bar{x}$ é a média do conjunto de dados e n é o número total de observações do conjunto de dados. Desenvolvendo a fórmula (3.7) chegamos a fórmula (3.8) que, para alguns casos, tornam os cálculos mais simples e rápidos.

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} \quad (3.8)$$

onde:

$\sum x_i^2$ é a soma de cada valor da variável ao quadrado;

$\left(\sum x_i\right)^2$ é o quadrado da soma de todos os valores da variável;

n é o número total de valores do conjunto de dados.

Como o desvio-padrão é uma medida de dispersão e mede a variabilidade entre os valores temos que valores muito próximos resultarão em desvios-padrões pequenos, enquanto que valores mais espalhados resultarão em desvios -padrões maiores.

! ATENÇÃO

O valor do desvio-padrão nunca é negativo. É zero apenas quando todos os valores do conjunto de dados são os mesmos. A unidade do desvio-padrão é a mesma unidade dos dados originais.

3.2.3.1 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 em forma de sino, valem as seguintes considerações:

- Cerca de 68% das observações do conjunto de dados ficam a 1 desvio-padrão da média, ou seja, $(\overline{x} - s)$ e $(\overline{x} + s)$
- Cerca de 95% das observações do conjunto de dados ficam a 2 desvios-padrões da média, ou seja, $(\overline{x} - 2s)$ e $(\overline{x} + 2s)$
- Cerca de 99,7% das observações do conjunto de dados ficam a 3 desvios -padrões da média, ou seja, $(\overline{x} - 3s)$ e $(\overline{x} + 3s)$

3.2.3.2 Propriedades do desvio-padrão

1. Somando-se (ou subtraindo-se) uma constante c de todos os valores de uma variável, o desvio padrão não se altera:

$$y_i = x_i \pm c \Rightarrow S_y = S_x$$

2. Multiplicando-se (ou dividindo-se) todos os valores de uma variável por uma constante (diferente de zero), o desvio padrão fica multiplicado (ou dividido) por essa constante:

$$y_i = x_i \times c \Rightarrow S_y = c \times S_x \quad \text{ou} \quad y_i = \frac{x_i}{c} \Rightarrow S_y = \frac{S_x}{c}$$

3.2.4 Variância (s^2)

A variância de um conjunto de dados nada mais é do que o valor do desvio-padrão elevado ao quadrado, ou seja,

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{3.9}$$

ou

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} \tag{3.10}$$

A variância não é uma medida muito utilizada para mostrar a dispersão de um conjunto de dados, pois, expressa o seu resultado numa medida ao quadrado, não sendo possível interpretar o seu valor. Portanto, na análise descritiva dos dados, não vamos trabalhar com esta medida constantemente. Se um determinado problema fornecer a variância do conjunto de dados, basta calcularmos a raiz quadrada deste valor (variância) e obteremos o desvio-padrão, que é facilmente interpretado por estar na mesma medida do conjunto de dados.

3.2.5 Coeficiente de Variação (cv)

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 cal-

cular o grau de variação dos dados em relação à média aritmética. E é obtido através do seguinte cálculo:

$$cv = \frac{s}{\overline{x}} \times 100 \qquad (3.11)$$

onde *s* é o desvio-padrão e $\overline{x}$ 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.* ³ 30%

Em geral, o coeficiente de variação é uma estatística útil para comparar a variação para valores originados de diferentes variáveis (por exemplo: peso, em Kg e altura, em cm), pois ele é adimensional.

3.2.6 Exemplo de Aplicação das Medidas de Dispersão para Dados não Tabulados

Vamos exemplificar o cálculo da **amplitude, da amplitude interquartil**, do **desvio-padrão**, da **variância** e do **coeficiente de variação** utilizando o exemplo 2.3, que apresenta o conjunto de dados brutos.

Exemplo 3.6: Um gerente de banco deseja estudar a movimentação de pessoas em sua agência na segunda semana de um mês qualquer. Ele constata que no primeiro dia entraram 1348 pessoas, no segundo dia, 1260 pessoas, no terceiro, 1095, no quarto, 832 e no último dia do levantamento, 850 pessoas. Encontre a amplitude, o desvio-padrão, a variância e o coeficiente de variação para este conjunto de dados e interprete os resultados.

Resolução:

A amplitude é dada por:

$$R = x_{(máximo)} - x_{(mínimo)} = 1348 - 832 = 516 \text{ pessoas}$$

A diferença, no número de pessoas que entram na agência, entre o dia de maior movimento e o dia de menor movimento é de 516 pessoas.

Para encontrarmos a amplitude interquartil precisamos calcular o primeiro e o terceiro quartil. Para isto, vamos seguir os procedimentos descritos no item 3.1.1.1.

- **Primeiro quartil** (Q_1)

$$Q_1 = x_{\left(\frac{1 \times 5}{4} + \frac{1}{2}\right)}$$

$$Q_1 = x_{1,75}$$

Portanto, o primeiro quartil está entre os elementos que ocupam a primeira e a segunda posição do conjunto de dados ordenados. Então:

$$\frac{x - 832}{0,75} = \frac{850 - 832}{1}$$

$$x - 832 = 13,5$$

$$x = 832 + 13,5 = 845,5$$

- **Terceiro quartil** (Q_3)

$$Q_3 = x_{\left(\frac{3 \times 5}{4} + \frac{1}{2}\right)}$$

$$Q_3 = x_{4,25}$$

Portanto, o terceiro quartil está entre os elementos que ocupam a quarta e a quinta posição do conjunto de dados ordenados. Então:

$$\frac{x - 1260}{0,25} = \frac{1348 - 1260}{1}$$

$$x - 1260 = 22$$

$$x = 1260 + 22 = 1282$$

$$Amplitude\ interquartil = Q_3 - Q_1$$

$$= 1282 - 845,5$$

$$= 436,5\ pessoas$$

Então, a amplitude do intervalo que contém 50% das observações centrais é 436,5 pessoas.

O desvio-padrão é obtido através das fórmulas (3.7) ou (3.8). Como a média aritmética é um número inteiro e existem poucos dados, a fórmula (3.7) é mais rápida de ser calculada. Porém, fica a critério de cada um a utilização de uma ou de outra. Lembrando que a média aritmética encontrada anteriormente é igual a 1077 e utilizando a fórmula (3.7), temos:

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

$$= \sqrt{\frac{(1348-1077)^2 + (1260-1077)^2 + (1095-1077)^2 + (832-1077)^2 + (850-1077)^2}{5-1}} =$$

$$= \sqrt{\frac{(271)^2 + (183)^2 + (18)^2 + (-245)^2 + (-227)^2}{4}} =$$

$$= \sqrt{\frac{(73441) + (33489) + (324) + (60025) + (51529)}{4}} =$$

$$= \sqrt{\frac{218808}{4}} = \sqrt{54702} \ @\ 233,88 \quad \text{pessoas}$$

Neste exemplo, entram na agência, em média, 1077 pessoas por dia. O número de pessoas que entram na agência varia, mas, tipicamente, a diferença em relação à média foi de aproximadamente 234 pessoas.

A **variância**, como vimos, é obtida através das fórmulas (3.9) ou (3.10), ou simplestemente, como já temos o desvio-padrão, a variância é o valor que está dentro da raiz quadrada, ou seja:

$$s^2 = 54.702 \text{ pessoas}^2$$

Não há como interpretar a expressão *pessoas*². Por esse motivo, utilizamos o **desvio-padrão** no lugar da variância.

O **coeficiente de variação**, dado pela fórmula (3.11), é muito fácil de ser obtido desde que já conheçamos os valores da média aritmética e do desvio-padrão. Pela fórmula podemos observar que basta fazermos uma simples divisão. Para este exemplo temos que:

$$cv = \frac{s}{\bar{x}} = \frac{233,88}{1077} \approx 0,2172 \text{ ou } 21,72\%$$

Utilizando a regra empírica, podemos dizer que o conjunto de dados apresenta uma média dispersão.

Agora, vamos aprender a calcular as medidas de dispersão através de dados tabulados.

Quando os dados estiverem na forma tabulada, haverá uma pequena diferença no cálculo das medidas de dispersão, pois agora será necessário considerar as frequências, que funcionarão como "fatores de ponderação", referentes a cada valor da variável.

3.2.7 Desvio-Padrão para Dados Tabulados

Se os dados estiverem tabulados, o desvio-padrão pode ser encontrado da seguinte forma:

$$s = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot f_i}{n-1}} \qquad (3.12)$$

Desenvolvendo a fórmula (3.12) chegamos a fórmula (3.13) que também é utilizada para o cálculo do desvio-padrão:

$$s = \sqrt{\frac{\sum x_i^2 \cdot f_i - \frac{\left(\sum x_i \cdot f_i\right)^2}{n}}{n-1}} \qquad (3.13)$$

onde, para ambas as fórmulas (3.12) e (3.13), x_i representa cada uma das observações do conjunto de dados ou, se os dados estiverem agrupados em classes de frequências, x_i representa o ponto médio da classe, $\bar{x}$ é a média do conjunto de dados, f_i é a frequência associada a cada observação (ou classe de observações) do conjunto de dados e n é o número de total de observações no conjunto de dados.

3.2.8 Variância para Dados Tabulados

A variância de um conjunto de dados agrupados é dada por:

$$s^2 = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot f_i}{n-1}} \qquad (3.14)$$

ou

$$s^2 = \sqrt{\frac{\sum x_i^2 \cdot f_i - \frac{\left(\sum x_i \cdot f_i\right)^2}{n}}{n-1}} \qquad (3.15)$$

A **amplitude, a amplitude interquartil** e o **coeficiente de variação** não sofrem modificações significativas. A amplitude continua sendo a diferença entre o maior e o menor valor (se os dados estiverem em classes de frequências, *R* será a diferença entre o limite superior da última classe e o limite inferior da primeira classe). A amplitude interquartil continua sendo a diferença entre o terceiro e o

primeiro quartil e o cálculo do coeficiente de variação é feito utilizando a fórmula (3.11), porém, se os dados estiverem em classes de frequências, o desvio-padrão e a média aritmética são obtidos utilizando x_i como o ponto médio da classe.

3.2.9 Exemplos de Aplicações das Medidas de Dispersão para Dados Tabulados

Para demonstração dos cálculos para dados tabelados, vamos continuar utilizando os exemplos desenvolvidos no item 2.2 (Exemplos 2.4 e 2.5).

NÚMERO DE FALTA	*f*
0	31
1	20
2	8
3	2
4	0
5	1
6	1
Total	**63**

Exemplo 3.7: Em um determinado mês, foi computado o número x de faltas ao trabalho, por motivos de saúde, que cada funcionário de uma determinada empresa teve. Os dados estão apresentados na tabela abaixo:

Encontre a amplitude, o desvio-padrão, a variância e o coeficiente de variação para este conjunto de dados e interprete os resultados.

Resolução:

A **amplitude** para este conjunto de dados é dada por:

$$R = x_{(máximo)} - x_{(mínimo)} = 6 - 0 = 6 \text{ pessoas}$$

A maior diferença entre os números de faltas ao trabalho, por motivo de saúde, que funcionários de uma determinada empresa tiveram no período de um mês, é 6 faltas.

O **desvio-padrão** é obtido através das fórmulas (3.12) ou (3.13). Para exemplificar, vamos trabalhar com a fórmula (3.13). Para facilitar, vamos montar um quadro com os resultados que nos interessa para aplicar tal expressão.

NÚMERO DE FALTAS (x_i)	f	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
0	31	0	0
1	20	20	20
2	8	16	32
3	2	6	18
4	0	0	0
5	1	5	25
6	1	6	36
Total $(\bullet)$	**63**	**53**	**131**

Substituindo os valores encontrados no quadro acima na fórmula 3.13, obtemos:

$$s = \sqrt{\frac{\sum x_i^2 \cdot f_i - \frac{\left(\sum x_i \cdot f_i\right)^2}{n}}{n-1}} = \sqrt{\frac{131 - \frac{(53)^2}{63}}{63-1}} =$$

$$= \sqrt{\frac{131 - \frac{2809}{63}}{62}} \approx \sqrt{\frac{131 - 44{,}59}{62}} \approx$$

$$\approx \sqrt{1{,}3938} \approx 1{,}18 \text{ faltas}$$

Podemos dizer que, em média, ocorre aproximadamente 1 falta por funcionário, por mês. Na verdade, sabemos que esse número de faltas por funcionário varia em torno da média, mas, tipicamente, a diferença em relação à média é de, aproximadamente, 1 falta.

A **variância** é obtida através das fórmulas (3.14) ou (3.15), porém, como já temos o desvio-padrão, a variância é o valor que está dentro da raiz quadrada. Portanto, temos:

$$s^2 = 1{,}3938 \text{ faltas}^2$$

! ATENÇÃO

O valor da média, calculado anteriormente para este conjunto de dados, é igual a 0,84 falta. Se arredondarmos esse valor para um valor inteiro, podemos dizer que a média é aproximadamente igual a 1 falta.

Como 1,3938 faltas^2 não tem interpretação, utilizamos o desvio-padrão para interpretar o comportamento dos dados.

O **coeficiente de variação** para este exemplo é dado por:

$$cv = \frac{s}{\bar{x}} = \frac{1{,}18}{0{,}84} \approx 1{,}4048 \text{ ou } 140.48\%$$

O coeficiente de variação nos diz que este conjunto de dados apresenta uma alta dispersão.

Para finalizarmos, vamos fazer os cálculos para os dados agrupados em classes de frequências. Para isto vamos utilizar o exemplo 2.5 que se encontra no item 2.2.

Exemplo 3.8: A tabela abaixo apresenta a distribuição de frequências do tempo de vida de 60 componentes eletrônicos (medido em dias) submetidos à experimentação num laboratório especializado.

TEMPO DE VIDA (DIAS)	f	PONTO MÉDIO (x_i)
3 ├ 18	3	10,5
18 ├ 33	4	25,5
33 ├ 48	4	40,5
48 ├ 63	8	55,5
63 ├ 78	10	70,5
78 ├ 93	28	85,5
93 ├ 108	2	100,5
108 ├ 123	1	115,5
Total	**60**	

Calcule a amplitude, o desvio-padrão, a variância e o coeficiente de variação para este conjunto de dados e interprete os resultados.

Resolução:

A **amplitude** para este conjunto de dados é dada por:

$$R = x_{(máximo)} - x_{(mínimo)} = 123 - 3 = 120 \text{ dias}$$

A maior diferença entre os tempos de vida (em dias) dos componentes eletrônicos foi de 120 dias, ou seja, o componente com maior sobrevivência durou 120 dias a mais do que o componente que durou menos tempo.

Para o cálculo do **desvio-padrão**, podemos utilizar as fórmulas (3.12) ou (3.13), onde o termo x_i é o ponto médio de cada classe de frequên-cia. Como a média aritmética envolve valores decimais, é mais simples efetuar os cálculos através da fórmula (3.13). Como no exemplo anterior, vamos construir um quadro acrescentando as colunas que fornecerão os valores que precisamos para substituir na fórmula 3.13.

CLASSES DE FREQUÊNCIAS	f	PONTO MÉDIO (x_i)	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
3 ├ 18	3	10,5	31,5	330,75
18 ├ 33	4	25,5	102	2601
33 ├ 48	4	40,5	162	6561
48 ├ 63	8	55,5	444	24642
63 ├ 78	10	70,5	705	49702,5
78 ├ 93	28	85,5	2394	204687
93 ├ 108	2	100,5	201	20200,5
108 ├ 123	1	115,5	115,5	13340,25
Total	**60**		**4155**	**322065**

Com os valores obtidos, temos:

$$s = \sqrt{\frac{\sum x_i^2 \cdot f_i - \frac{\left(\sum x_i \cdot f_i\right)^2}{n}}{n-1}} = \sqrt{\frac{322065 - \frac{(4155)^2}{60}}{60-1}} =$$

$$= \sqrt{\frac{322065 - \frac{17264025}{60}}{59}} \approx \sqrt{\frac{322065 - 287733,75}{59}} \approx$$

$$\approx \sqrt{581,89} \approx 24,12 \text{ dias}$$

Em média, os componentes eletrônicos têm duração de 69 dias e 6 horas com uma variação de, aproximadamente, 24 dias e 3 horas para mais ou para menos com relação à média.

A **variância**, como já sabemos, é o desvio-padrão ao quadrado. Assim, temos:

$$s^2 = 581,89 \text{ dias}^2$$

Como 581,89 dias^2 não tem interpretação, utilizamos o desvio-padrão para interpretar o comportamento dos dados.

O **coeficiente de variação** para este exemplo é:

$$cv = \frac{s}{\bar{x}} = \frac{24,12}{69,25} \approx 0,3483 \text{ ou } 34,83\%$$

o que indica uma variabilidade alta no conjunto de dados.

3.3 Gráficos

O objetivo da utilização de gráficos em análise de dados é o de facilitar a compreensão do fenômeno estatístico por meio do efeito visual imediato que os gráficos proporcionam.

CONEXÃO

Vamos refletir um pouco sobre a necessidade de abordagens pedagógicas para o ensino e a aprendizagem de gráficos acessando o endereço <http://www.ufrrj.br/emanped/paginas/conteudo_producoes/docs_22/carlos.pdf>.

3.3.1 Tipos de Gráficos

Existem vários tipos de gráficos. Os mais usados são: gráfico em linhas, diagramas de área (como por exemplo: gráfico em colunas, gráfico em barras e gráfico em setores) e gráficos para representar as distribuições de frequências construídas com intervalos de classes (como por exemplo: polígono de frequências, histograma e ogiva).

Segundo 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.

Vamos saber um pouco quando usar e como construir cada um destes gráficos.

3.3.1.1 Gráfico em Linhas

Sempre que os dados estiverem distribuídos segundo uma variável no tempo (meses, anos, etc.), assim como sucede com os dados do exemplo 3.9 –figura 1, os dados podem, também, ser descritos através de um gráfico em linhas. 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 3.9: A tabela 3.1 fornece uma lista do número de assinantes de telefones celulares, em milhões, de 1997 a 2007, do país X. Construa um gráfico para resumir os dados da tabela abaixo.

ANO	ASSINANTES (EM MILHÕES)
1997	1,1
1998	1,3
1999	1,5
2000	1,9
2001	2,4
2002	2,6
2003	3,1
2004	7,4
2005	18,6
2006	21,5
2007	29

Tabela 3.1 – Assinantes de telefones celulares, em milhões, de 1997 a 2007.

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..

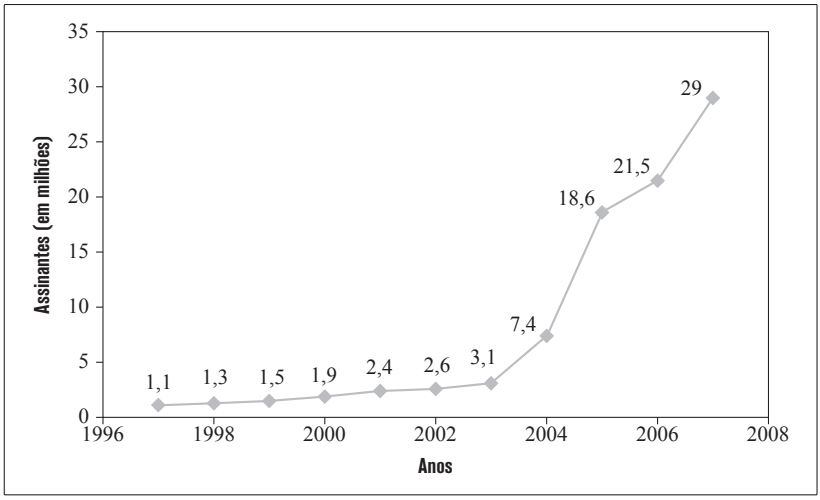


Figura 1 – Gráfico em linha para os dados de assinantes de telefones celulares.

3.3.1.2 Gráfico (ou Diagrama) em Barras (ou Colunas)

Os diagramas em barras (ou colunas) são bastante utilizados quando trabalhamos com variáveis qualitativas (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. O gráfico em barras, quando as barras estão dispostas no sentido vertical, também é chamado de *gráfico em colunas*.

! ATENÇÃO

Quando construímos o gráfico de barras para variáveis qualitativas e as barras são arranjadas em ordem descendente de altura, a partir da esquerda para a direita, com o atributo que ocorre com maior frequência aparecendo em primeiro lugar, denominamos este gráfico de barras de *Diagrama de Pareto*.

Exemplo 3.10: Uma grande indústria de materiais de construção, com diversas lojas espalhadas pelo país, fez um levantamento das principais causas de perda de ativos durante o ano de 2007 e as informações estão dispostas na tabela seguinte.

CAUSAS	VALOR PERDIDO (MILHÕES DE REAIS)
Má administração	5,2
Roubos de funcionários	3,9
Fraudes nas vendas	5,5
Assaltos às lojas	1,8
Perda do estoque	1,6
Atendimento ruim	0,8

Tabela 3.2: Causas de perda de ativos durante o ano de 2007.

Graficamente, podemos representar este conjunto de dados de três formas diferentes: gráfico em colunas, gráfico em barras e o gráfico em setores (ou pizza ou circular), que será apresentado no próximo item.

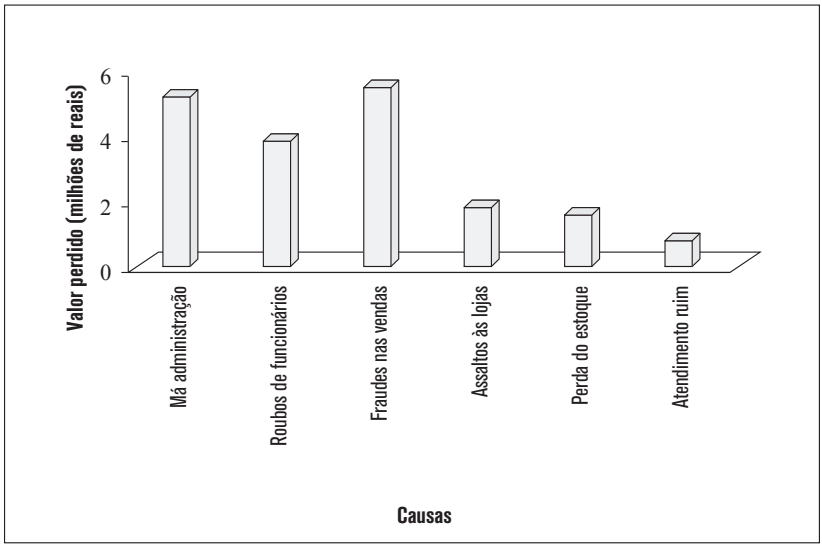


Figura 2a – Gráfico em colunas para a variável causas de perdas de ativos.

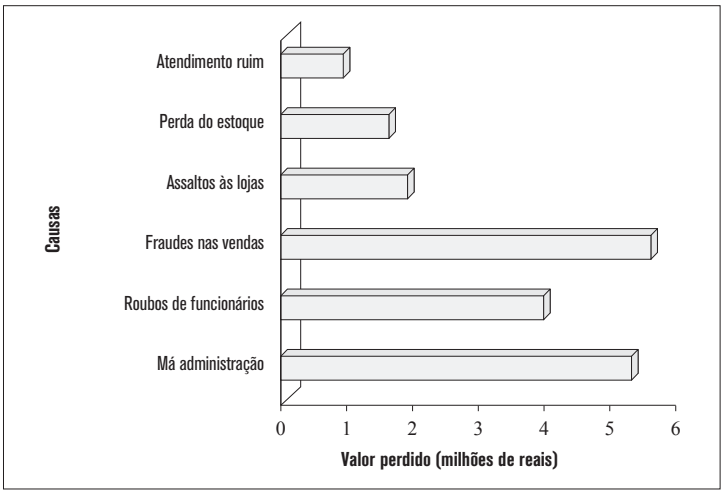


Figura 2b – Gráfico em barras para a variável causas de perdas de ativos.

3.3.1.3 Gráfico (ou Diagrama) em Setores

O diagrama em setores, também conhecido como gráfico de pizza, é um dos gráficos mais utilizados para representar variáveis qualitativas (ou categóricas) e é bastante apropriado quando se deseja visualizar a proporção que cada categoria representa do total.

Vamos utilizar os dados do exemplo 3.10 para mostrar um gráfico em setores.

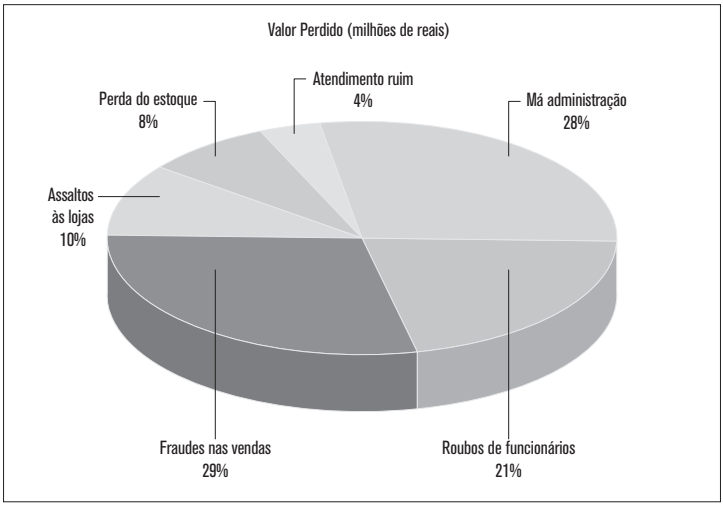


Figura 3 – Gráfico em setores para a variável causas de perdas de ativos.

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

3.3.1.4 Histograma

Um histograma é semelhante ao diagrama de barras, porém refere-se a uma distribuição de frequências para dados quantitativos contínuos. 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 3.11: A tabela abaixo apresenta o salário de funcionários de uma empresa no interior de Minas Gerais.

SALÁRIO (R$)	FREQ. ABSOLU TA (f)	FREQ. ACUMULADA (f_a)
400,00 ├ 800,00	38	38
800,00 ├ 1200,00	18	56
1200,00 ├ 1600,00	12	68
1600,00 ├ 2000,00	8	76
2000,00 ├ 2400,00	8	84
2400,00 ├ 2800,00	5	89
2800,00 ├ 3200,00	3	92
3200,00 ├ 3600,00	0	92
3600,00 ├ 4000,00	2	94
4000,00 ├ 4400,00	0	94
4400,00 ├ 4800,00	1	95
Total	**95**	

Tabela 3.3 – Distribuição de frequências dos salários dos funcionários de uma empresa no interior de Minas Gerais.

Como os dados da tabela 3.3 estão apresentados em intervalos de classes podemos representá-los graficamente através de um histograma ou do polígono de frequências, como mostram as figuras 4 e 5, respectivamente.

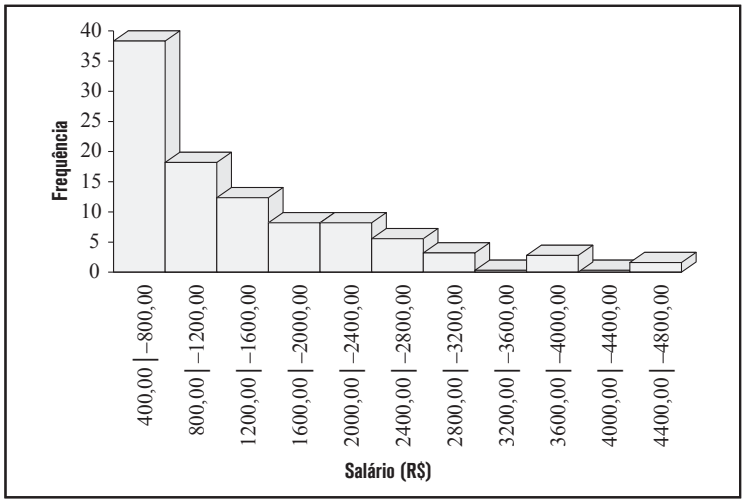


Figura 4 – Histograma dos salários dos funcionários de uma empresa no interior de Minas Gerais.

3.3.1.5 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 do Exemplo 3.11, temos o polígono de frequências representado pela Figura 5.

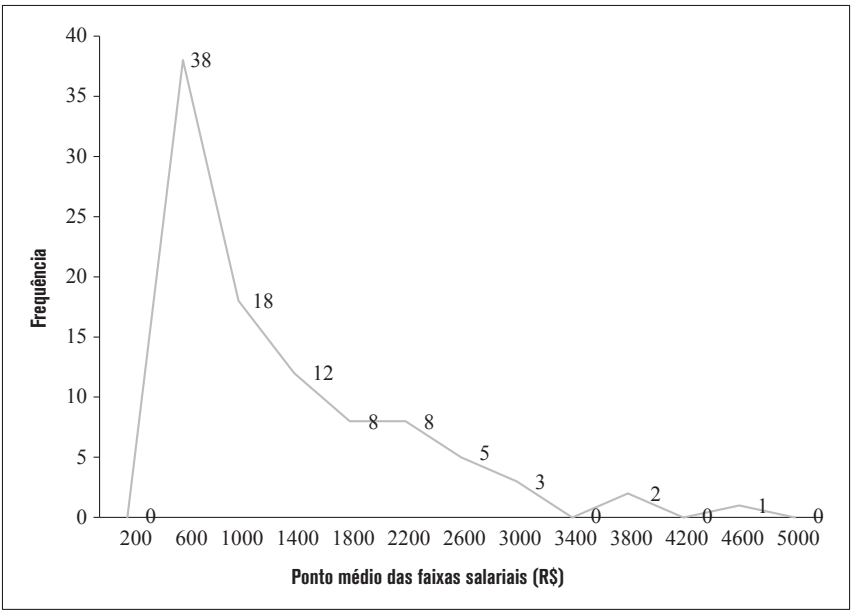


Figura 5 – Polígono de frequências dos salários dos funcionários de uma empresa no interior de Minas Gerais.

3.3.1.6 Ogiva

Uma ogiva é um gráfico para uma distribuição de frequên-cias acumuladas. Utilizando o Exemplo 3.11, a terceira coluna traz a frequência acumulada dos dados e a ogiva fica representada pela Figura 6.

! ATENÇÃO

Para construir um gráfico de ogiva, devemos usar o limite superior de cada intervalo no eixo horizontal e a frequência acumulada no eixo vertical. A frequência acumulada relacionada com o limite inferior da primeira classe é sempre zero.

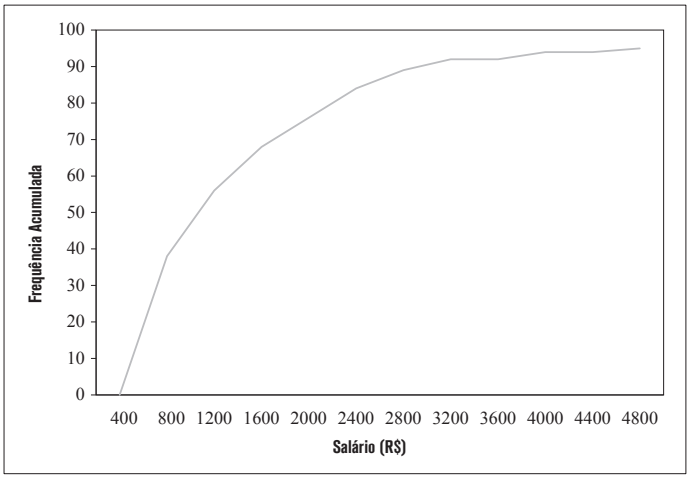


Figura 6 – Ogiva dos salários dos funcionários de uma empresa no interior de Minas Gerais.

3.3.1.7 Diagrama de Dispersão

O diagrama de dispersão é um gráfico muito utilizado quando temos interesse em identificar a associação entre duas variáveis quantitativas (X e Y). Para construí-lo, cada para ordenado é colocado em suas determinadas coordenadas (x,y). Vamos construir um diagrama de dispersão utilizando os dados do exemplo a seguir.

Exemplo 3.12: Uma concessionária de veículos quer verificar a eficácia de seus anúncios em determinado jornal na venda de carros novos. A tabela, a seguir, mostra o *número de anúncios publicados*, por mês, e o correspondente *número de carros vendidos* nos últimos seis meses.

NÚMERO DE ANÚNCIOS PUBLICADOS (X)	NÚMERO DE CARROS VENDIDOS (Y)
28	140
20	110
22	100
14	75
10	60
7	52

Tabela 3.4 – Número de anúncios publicados e número de carros vendidos

Para verificarmos, visualmente, se há relação entre o número de anúncios publicados e o número de carros vendidos, construímos o diagrama de dispersão.

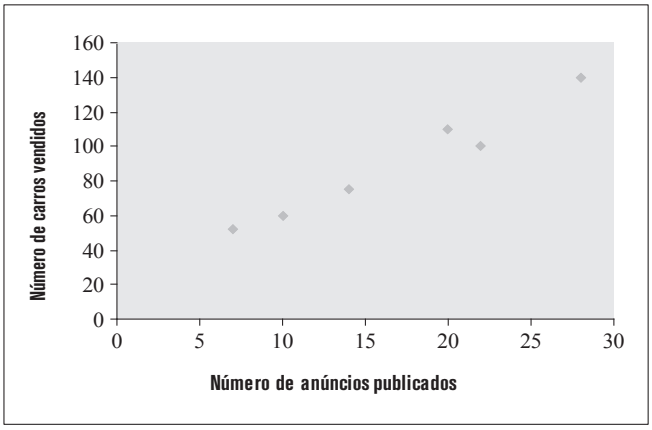


Figura 7 – Diagrama de dispersão do número de anúncios publicados e número de carros vendidos.

Pela análise gráfica observamos que à medida que o número de anúncios publicados aumenta, ocorre um aumento no número de carros vendidos.

! ATENÇÃO

Um tipo de gráfico muito utilizado em jornais e revistas é o gráfico pictórico. Estes gráficos são construídos a partir de figuras ou conjunto de figuras que são representativas do fenômeno em estudo. Como são representados por figuras, despertam a atenção do leitor.

Como vimos nos itens anteriores, os gráficos nos auxiliam no estudo do comportamento da variável em estudo no conjunto de dados. 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:

1. **Gráfico sucata**: neste tipo de gráfico, há um uso excessivo de figuras que podem ocultar a informação que se deseja transmitir.
2. **Ausência de base relativa**: 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 re-

lativa ao invés da frequência absoluta.

3. **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.
4. **Ausência do ponto zero**: a ausência do ponto zero no eixo vertical tende a produzir uma impressão enganosa do comportamento dos dados, exagerando ou reduzindo eventuais variações.

ATIVIDADE

1. Uma agência de turismo está interessada em saber o perfil dos seus clientes com relação à variável estado civil. Para isso, o gerente desta agência pediu ao funcionário do setor de vendas para fazer um gráfico que resuma estas informações. Construa o gráfico e interprete-o.

ESTADO CIVIL	NÚMERO DE CLIENTES
Solteiro	2600
Casado	900
Viúvo	345
Separado	1200
Outros	1020
Total	6065

2. Um consultor estava interessado em saber quanto, geralmente, cada pessoa gastava em um determinado supermercado no primeiro sábado após receberem seus pagamentos (salários). Para isso ele entrevistou 50 clientes que passaram pelos caixas entre 13h e 18h, e anotou os valores gastos por cada um deles. Estes valores estão listados abaixo:

4,89	11,00	5,60	73,85	24,83	98,00	186,00	234,87	58,00	198,65
223,86	341,42	94,76	445,76	82,80	35,00	455,00	371,00	398,60	234,00
64,90	54,98	48,80	68,90	120,32	126,98	76,43	6,35	9,98	12,68

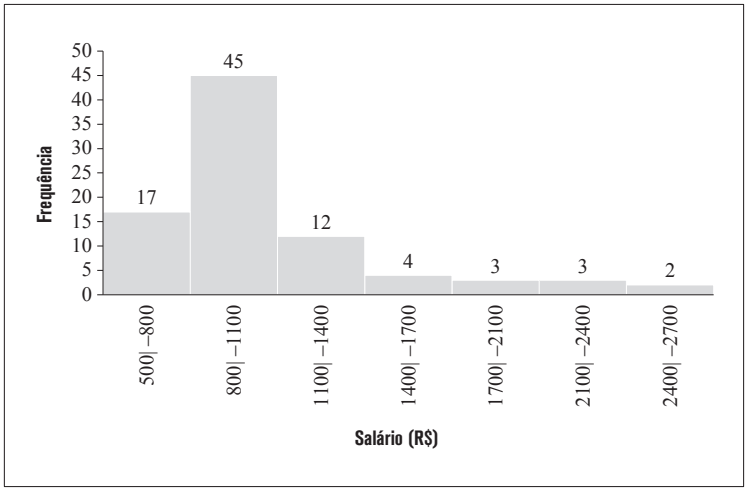
243,00	18,65	134,90	11,10	321,09	290,76	74,00	48,80	74,52	138,65
26,00	210,13	15,78	197,45	75,00	76,55	32,78	166,09	105,34	99,10

Analisando o conjunto de dados, responda os seguintes itens:

a) Construa uma distribuição de frequências a partir do conjunto de dados brutos.

b) Construa um histograma e um polígono de frequências para a tabela construída no item b).

3. Analise o gráfico abaixo e responda:



a) Qual a variável em estudo? Classifique-a.

b) Quantos funcionários ganham entre R$800,00 (inclusive) e R$1100,00 (exclusive)?

c) Qual o número de funcionários total desta empresa?

d) Qual a porcentagem de funcionários que ganham R$1700,00 ou mais?

e) Qual a porcentagem de funcionários que ganham entre R$500,00 (inclusive) e não mais que R$1100,00?

f) A partir do histograma, monte uma tabela de distribuição de frequências.

4. Os dados abaixo referem-se ao número de horas extras de trabalho que uma amostra de 64 funcionários de uma determinada empresa localizada na capital paulista.

10	10	12	14	14	14	15	16
18	18	18	18	18	19	20	20
20	20	20	21	22	22	22	22
22	22	22	22	22	22	22	22
23	23	24	24	24	24	24	24
24	25	25	25	25	26	26	26
26	26	26	27	27	27	28	28
29	30	30	32	35	36	40	41

Pede-se:

a) Calcule e interprete as seguintes medidas de dispersão, calculadas para os dados brutos (dados não tabulados): amplitude, desvio-padrão, variância e coeficiente de variação e interprete os resultados.
b) Construir uma distribuição de frequências completa (com freq. absoluta, freq. relativa, freq. acumulada e ponto médio).
c) Através da distribuição de frequências construída no item b), encontre a amplitude, o desvio-padrão, a variância e o coeficiente de variação e interprete os resultados.
d) Com a tabela construída no item b), encontre as seguintes medidas: 1° quartil, 7° decil e 99° Percentil. Interprete os resultados.
e) Construa o histograma para este conjunto de dados.

5. Os dados a seguir representam as notas de 5 disciplinas de um determinado candidato em um concurso público. São elas:

2, 5, 8, 8, 9

Calcule a amplitude, o desvio-padrão, a variância e o coeficiente de variação. Interprete os resultados.

6. Numa pesquisa realizada com 91 famílias, levantaram-se as seguintes informações com relação ao número de filhos por família:

NÚMERO DE FILHOS	0	1	2	3	4	5
FREQUÊNCIA DE FAMÍLIAS	19	22	28	16	2	4

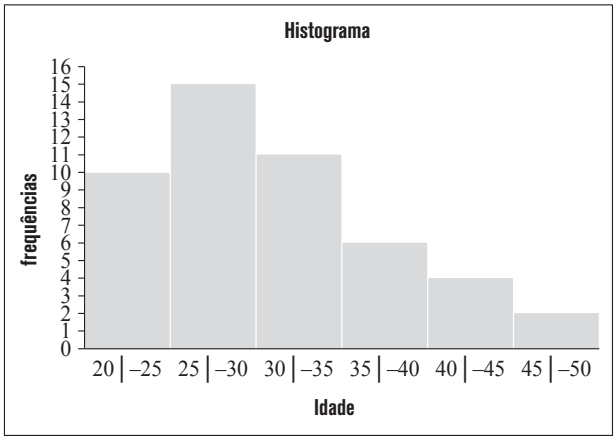
Calcule e interprete os resultados da:

a) amplitude

b) desvio-padrão

c) coeficiente de variação

7. O histograma abaixo representa a distribuição das idades dos funcionários de uma agência bancária. Com base no histograma abaixo, responda:



Qual a amplitude, o desvio-padrão, a variância e o coeficiente de variação para as idades dos funcionários? Interprete os resultados.

8. Um fabricante de caixas de cartolina fabrica três tipos de caixa. Testa-se a resistência de cada caixa, tomando-se uma amostra de 100 caixas e determinando-se a pressão necessária para romper cada caixa. São os seguintes os resultados dos testes:

TIPOS DE CAIXAS	A	B	C
Pressão média de ruptura (bária)	15	20	30
Desvio-padrão das pressões (bária)	4	5	6

a) Que tipo de caixa apresenta a menor variação absoluta na pressão de ruptura?
b) Que tipo de caixa apresenta a maior variação relativa na pressão de ruptura?

REFLEXÃO

Vimos, nesse capítulo, que tão importante quanto conhecer a média de um conjunto de dados, por exemplo, é determinar o seu grau de variabilidade (ou dispersão). Na maioria dos estudos que realizamos, nos deparamos com conjuntos que podem apresentar maior ou menor grau de homogeneidade.

Conjuntos com características de maior homogeneidade tendem a nos fornecer informações mais precisas e confiáveis. Imagine, por exemplo, um estabelecimento que diariamente presta atendimento aos seus clientes. Se a quantidade desses clientes varia muito de um dia para outro, fica mais difícil você determinar quantos funcionários disponibilizar para realizar o atendimento. No entanto, se esse número varia pouco (apresenta-se mais homogêneo) de um dia para o outro, fica muito mais fácil montar uma estrutura adequada de atendimento.

LEITURA

Sugerimos a leitura do artigo "E se todos fossem ao cinema ao mesmo tempo?" do professor Luiz Barco, disponível em:<http://super.abril.com.br/ciencia/lei-regularidade-estatistica-se-todos-fossem-ao-mesmo-cinema-ao-mesmo-tempo-439499.shtml>. Ele retrata, de forma bem interessante, a questão da regularidade dos fenômenos relacionados ao comportamento social.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. Estatística aplicada à administração e economia. São Paulo: Pioneira Thomson Learning, 2003.

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

BUSSAB, Wilton de O.; MORETTIN, Pedro A.. Estatística básica. São Paulo: Saraiva, 2003

COSTA NETO, Pedro Luiz de Oliveira. Estatística, São Paulo: Edgard Blucher, 2002.

DOWNING, Douglas; CLARK, Jeffrey. Estatística aplicada. São Paulo: Saraiva, 2002.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. Introdução à estatística. Rio de Janeiro: LTC, 2003.

TRIOLA, Mario F.. Introdução à estatística. Rio de Janeiro: LTC, 1999.

VIEIRA, Sonia. Estatística Básica. São Paulo: Cengage Learning, 2013.

NO PRÓXIMO CAPÍTULO

No próximo capítulo estudaremos uma das técnicas mais utilizadas da Inferência Estatística: Estimação Pontual e Estimação Intervalar. Veremos como estimar a média populacional a partir da média de uma amostra retirada dessa população. Nesse tipo de estudo, surgem definições conhecidas como: margem de erro, nível de confiança da pesquisa, etc. Além da estimação das médias, também estudaremos a estimação de proporções populacionais. Na construção do intervalo de confiança precisamos de conhecimentos sobre a distribuição Normal, cujo conceito também será introduzido no próximo capítulo.

4

Distribuições Amostrais e Estimação

4 Distribuições Amostrais e Estimação

A Inferência Estatística é um conjunto de técnicas muito utilizadas em problemas práticos do dia a dia. Com estas técnicas podemos tirar conclusões acerca de uma população de interesse utilizando informações de uma amostra aleatória.
A grande vantagem em se utilizar a Inferência Estatística é economizar tempo e dinheiro que seriam gastos para analisar uma população inteira, ressaltando que, algumas vezes, é impossível trabalhar com toda a população de interesse.
Uma das técnicas mais importante e utilizada da Inferência Estatística é a Estimação.
Veremos neste capítulo como estimar uma característica de interesse na população através da *estimação pontual e por intervalo*.
Alguns conceitos básicos necessários para o desenvolvimento teórico das técnicas de Inferência Estatística também serão apresentados.

OBJETIVOS

- Saber estimar tanto a média como a proporção populacional (referentes a uma variável presente na população) a partir de dados coletados em uma amostra aleatória retirada dessa população.

REFLEXÃO

Você se lembra de já ter ouvido notícias sobre divulgação de resultados de pesquisas em que foram citados termos como margem de erro e nível de confiança da pesquisa? Sempre que as pesquisas são realizadas em amostras, esses conceitos passam a fazer parte dos resultados que serão obtidos. Vamos compreender melhor o que eles significam e como são determinados.

4.1 Conceitos Básicos

Parâmetro é uma quantidade numérica, em geral desconhecida, que descreve uma característica da população. Normalmente é representado por letras gregas como θ, μ e σ, entre outras.

Estimador é uma função dos valores da amostra que utilizamos para estimar um parâmetro populacional. Os estimadores, em geral, são representados

por letras gregas com acento circunflexo: $\hat{\theta}$, $\hat{\mu}$ e $\hat{\sigma}$ etc.

Estimativa é o valor numérico obtido através do estimador.

Erro amostral é a diferença entre o resultado amostral e o verdadeiro resultado da população; tais erros resultam de flutuações amostrais devidas ao acaso.

Erro não amostral ocorre quando os dados amostrais são coletados, registrados ou analisados incorretamente. Por exemplo: os dados são selecionados através de uma amostra tendenciosa, uso de um instrumento de medida defeituoso ou o registro incorreto dos dados.

Amostra Aleatória Simples (AAS) de tamanho n de uma v.a. X, com determinada distribuição, é o conjunto de n v.a. ´s independentes $X_1, X_2, ..., X_n$ cada uma com a mesma distribuição de X.

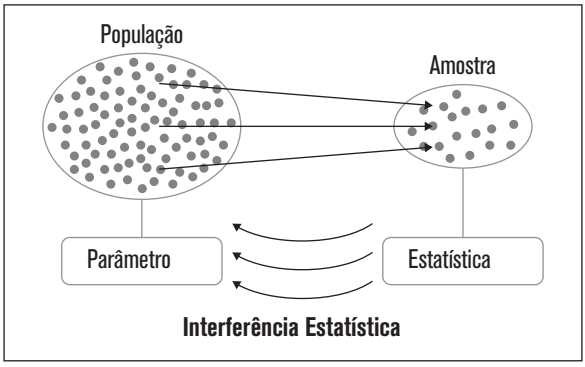


Figura 8 – Esquema de inferência sobre um parâmetro
Fonte: (MARTINS, 2006).

Como o estimador é uma função de valores da amostra aleatória, ou seja, $\hat{\theta} = f(X_1, X_2, ..., X_n)$, para diferentes amostras vamos obter valores diferentes para o estimador $\hat{\theta}$. Portanto, $\hat{\theta}$ também é uma variável aleatória. Como uma estimativa para o parâmetro populacional pode ser obtida utilizando mais de um estimador, precisamos estudar algumas propriedades dos estimadores para decidir qual utilizar. Vamos estudar agora duas propriedades dos estimadores: *vício* e *consistência*.

Vício: um estimador $\hat{\theta}$ é não viciado para θ se $E(\hat{\theta}) = \theta$, ou seja, o valor esperado do estimador é igual ao valor do parâmetro.

Consistência: um estimador $\hat{\theta}$ é consistente se:

$$\lim_{n \to \infty} E(\hat{\theta}) = \theta$$
$$\lim_{n \to \infty} Var(\hat{\theta}) = 0 \qquad (4.1)$$

Podemos observar que um estimador é consistente se, quando aumentamos o tamanho da amostra, o valor esperado do estimador é igual ao valor do parâmetro, portanto, não viciado, e a variância do estimador convergir para zero.

Dois parâmetros populacionais muito importantes e de grande interesse em se estimar são a *média* e a *proporção*. Portanto, vamos escolher agora os estimadores utilizados para estimar estes dois parâmetros de maneira que eles satisfaçam as propriedades de vício e consistência.

4.2 Estimador de uma Média Populacional

O melhor estimador da média populacional μ é a média amostral:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) \qquad (4.2)$$

Para mostrar que $\overline{X}$ é um estimador não viciado e consistente da média populacional precisamos encontrar $E(\bar{X})$ e $Var(\bar{X})$.

Seja X_1, X_2, ..., X_n uma amostra aleatória retirada de uma população, identificada pela variável X, com média μ e variância σ^2 Da definição de AAS temos que cada X_i, $i = 1, 2, ..., n$ tem a mesma distribuição de X, ou seja, $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$.

Pelas propriedades da esperança e da variância temos:

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n}\left[E(X_1) + E(X_2) + \ldots + E(X_n)\right]$$

$$= \frac{1}{n}[\mu + \mu + \ldots + \mu] = \frac{n\mu}{n} = \mu$$

e

$$Var(\bar{X}) = Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n^2}\left[Var(X_1) + Var(X_2) + \ldots + Var(X_n)\right]$$

$$= \frac{1}{n^2}[\sigma^2 + \sigma^2 + \ldots + \sigma^2] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

O primeiro resultado mostra que a média amostral é um estimador não viciado da média populacional. O segundo resultado mostra que, conforme *n* cresce, a variância da média amostral tende a zero, portanto $\overline{X}$ é um estimador consistente para μ.

Exemplo 4.1: Uma amostra aleatória de 20 famílias de determinado bairro foi selecionada e observou-se o número de pessoas em cada família com nível superior completo. Os dados obtidos foram:

1, 2, 2, 0, 3, 1, 1, 3, 0, 1, 2, 2, 1, 3, 1, 0, 2, 1, 1, 2

Encontre a estimativa para a média de pessoas com nível superior completo neste bairro.

Resolução:

Pelo que vimos na teoria, o estimador utilizado para se estimar a média populacional é a média amostral, ou seja:

$$\bar{X} = \frac{1}{20}(1+2+2+0+...+2)$$
$$= \frac{29}{30} \approx 0,97$$

Então, podemos concluir que, neste bairro, aproximadamente 1 pessoa possui nível superior completo.

4.3 Estimador de uma Proporção Populacional

O melhor estimador da proporção populacional p é a proporção amostral:

$$\hat{p} = \frac{\text{número de indivíduos na amostra com determinada característica}}{n} \quad (4.3)$$

Se definirmos uma variável aleatória X_i como:

$$X_i = \begin{cases} 1 & \text{, se o indivíduo apresenta a característica} \\ 0 & \text{, se o indivíduo não apresenta a característica} \end{cases}$$

podemos reescrever fórmula da proporção amostral como:

$$\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n} = \bar{X}$$

Portanto, o estimador da proporção populacional é uma média de variáveis aleatórias convenientemente definidas.

ATENÇÃO

A distribuição de Bernoulli é uma distribuição de probabilidade discreta com as seguintes características: o experimento é realizado somente uma vez e a v.a. X assume apenas dois valores, *P(sucesso) = P(X = 1) = p e P (fracasso) = P(X = 0) = 1 – p*, com *E(X) = p e VAR(X) = 1 – p*.

Seja $X_1, X_2, ..., X_n$ uma sequência de variáveis aleatórias independentes com distribuição de Bernoulli. Pelas propriedades da esperança e da variância temos:

$$E\left(\hat{p}\right) = E\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n}\left[E(X_1) + E(X_2) + ... + E(X_n)\right]$$
$$= \frac{1}{n}[\mathrm{p} + \mathrm{p} + ... + \mathrm{p}] = \frac{np}{n} = p$$

$$Var\left(\hat{p}\right) = Var\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n^2}\left[Var(X_1) + Var(X_2) + ... + E(X_n)\right]$$
$$= \frac{1}{n^2}[p(1-p) + p(1-p) + ... + p(1-p)] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Como no caso do estimador $\overline{X}$ vemos que $\hat{p}$ é um estimador não viciado pois $E(\hat{p}) = p$ e consistente pois, a medida que n aumenta, a variância da proporção amostral tende a zero.

Exemplo 4.2: Uma determinada academia, interessada em abrir uma filial em certo bairro, selecionou uma amostra aleatória de 30 adultos e perguntou se a pessoa fazia atividades físicas pelo menos 3 vezes por semana. As respostas foram classificadas da seguinte maneira: foi atribuído o valor 1 se a pessoa respondeu sim e 0 se a pessoa respondeu não. Os resultados são:

0 1 0 0 1 1 1 1 1 0 1 1 0 0 0
1 0 0 1 1 0 1 0 0 1 0 1 1 1 0

Obtenha (estime) a proporção de pessoas, neste bairro, que fazem atividades físicas pelo menos três vezes por semana.

Para estimarmos esta proporção utilizamos a proporção amostral dada por:

$$\hat{p} = \frac{0+1+0+0+1+\ldots+0}{30} = \frac{16}{30} = 0,5333$$

Portanto, baseado nesta amostra, aproximadamente 53,33% das pessoas deste bairro fazem atividade física pelo menos três vezes por semana.

Antes de passarmos para o conceito de distribuições amostrais e estimação intervalar vamos estudar as características de uma variável aleatória cuja distribuição é Normal. Precisaremos do conceito da distribuição Normal para construir intervalos de confiança.

4.4 Propriedades da 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.

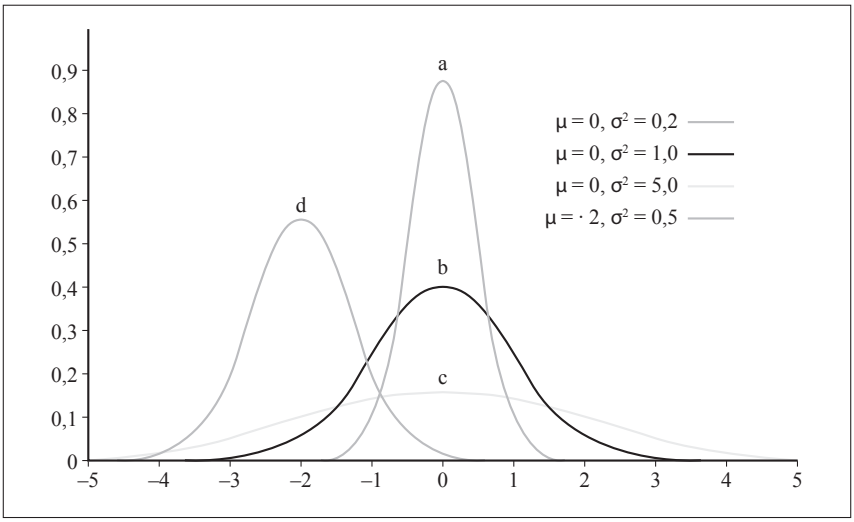


Figura 9 – Distribuições Normais, N (μ, σ^2)

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*.

4.5 Distribuições Amostrais

Estudamos como determinar os estimadores da média e da proporção populacional. Encontramos o valor esperado e a variância de cada estimador sem especificar a sua distribuição.

Agora vamos obter informação sobre a forma da distribuição dos estimadores da média e da proporção.

A distribuição amostral de um estimador é a distribuição de todos os valores do estimador quando todas as amostras possíveis de mesmo tamanho n são extraídas da mesma população. A distribuição amostral de um estimador é representada, frequentemente, através de uma tabela ou de um histograma.

Para encontrar a distribuição amostral da média utilizaremos um resultado fundamental na teoria da Inferência Estatística, conhecido como *Teorema do Limite Central*.

Teorema do Limite Central (TLC)

Para amostras aleatórias simples ($X_1, X_2, ..., X_n$), selecionadas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média $\overline{X}$ pode ser aproximada, para n grande, pela distribuição normal, com média μ e variância σ^2/n (desvio padrão, $\sigma^2/\sqrt{n}$).

Observação: para amostras com 30 elementos ou mais a aproximação é considerada boa.

Se a população é normal $N(\mu, \sigma^2)$, a distribuição amostral da média tem distribuição exata normal com média μ e variância σ^2/n para qualquer tamanho de amostra.

Na aplicação do teorema do limite central, o resultado $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ supõe que a população seja infinitamente grande. Quando trabalhamos com amostragem com reposição, a população é infinitamente grande. Agora, em populações finitas, cuja amostragem é feita sem reposição, precisamos fazer um ajuste no resultado $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Vamos utilizar a seguinte regra empírica:

Quando a amostragem for sem reposição e o tamanho amostral n for maior que 5% do tamanho finito N da população, ou seja, $n > 0{,}05 \cdot N$, devemos ajustar o desvio padrão das médias amostrais $\sigma_{\bar{x}}$ multiplicando-o pelo fator de correção para população finita:

$$\sqrt{\frac{N - n}{N - 1}}$$

4.5.2 Distribuição amostral da proporção

No item 4.3 vimos que o estimador da proporção populacional é uma média de v.a´s, ou seja,

$$\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n} = \bar{X}$$

Portanto, para n grande podemos considerar a distribuição amostral de p como aproximadamente normal:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \tag{4.5}$$

4.6 Erro Padrão de um Estimador

Seja $\hat{\theta}$ um estimador do parâmetro θ. O erro padrão de $\hat{\theta}$ é a quantidade:

$$EP(\hat{\theta}) = \sqrt{Var(\hat{\theta})} \tag{4.6}$$

Em palavras, o erro padrão avalia a precisão do cálculo do estimador populacional.

No caso da média amostral, que é estimador da média populacional, temos:

$$EP(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \tag{4.7}$$

Como σ é, em geral, desconhecido podemos obter o erro padrão estimado de $\overline{X}$, ou seja:

$$EP(\bar{X}) = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} \tag{4.8}$$

onde s^2 é a variância amostral.

No caso da proporção amostral, que é o estimador da proporção populacional, temos:

$$EP(\hat{p}) = \sqrt{\frac{p\,(1 - p)}{n}} \tag{4.9}$$

Quando não conhecemos p obtemos o erro padrão estimado de $\hat{p}$ substituindo p por $\hat{p}$:

$$\hat{EP}(\hat{p}) = \sqrt{\frac{\hat{p}\,(1 - \hat{p})}{n}} \tag{4.10}$$

Os estimadores vistos até agora são *pontuais*, ou seja, produzem um único valor como estimativa do parâmetro. Se quisermos medir a precisão da estimativa obtida podemos construir *intervalos de confiança* que são baseados na distribuição amostral do estimador pontual.

A um intervalo de confiança está associado um nível de confiança $1 - \alpha$ que fornece a probabilidade de que o intervalo incluirá o verdadeiro parâmetro populacional em várias amostras repetidas.

Devemos tomar bastante cuidado na interpretação do intervalo de confiança. Uma interpretação conveniente é a seguinte: se selecionarmos várias amostras de mesmo tamanho e calcularmos, para cada uma delas, os correspondentes intervalos de confiança com nível de confiança $1 - \alpha$, esperamos que a proporção de intervalos que contenham o valor do parâmetro populacional seja igual a $1 - \alpha$. Por exemplo, se selecionarmos 100 amostras de mesmo tamanho e construirmos seus respectivos intervalos de confiança, 95 intervalos irão conter o verdadeiro valor do parâmetro populacional.

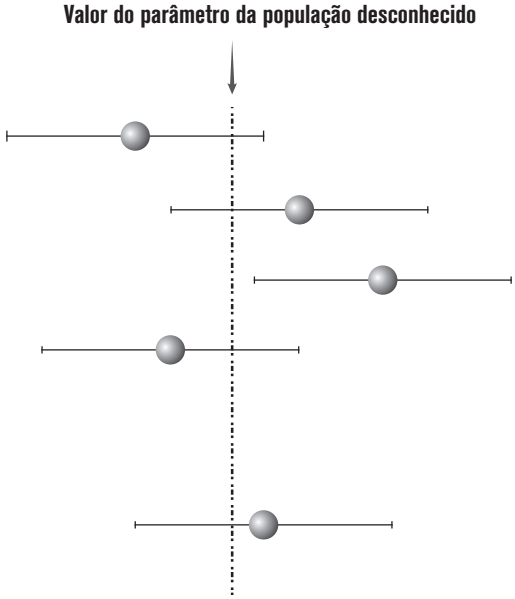


Figura 10 – intervalos de confiança para o parâmetro populacional

As escolhas mais comuns para o nível de confiança e, consequentemente, os respectivos valores críticos obtidos da distribuição normal são:

NÍVEL DE CONFIANÇA	α	VALOR CRÍTICO $Z_{\alpha/2}$
90%	0,10	1,645
95%	0,05	1,96
99%	0,01	2,575

A escolha de 95% é a mais comum porque resulta em um bom equilíbrio entre precisão (que é refletido na largura do intevalo de confiança) e confiabilidade (conforme expresso pelo nível de confiança). Dependendo da necessidade, o nível de confiança pode superar 99%. No entanto, quanto maior esse nível,

maior também será a margem de erro, que significa perda na precisão dos resultados. Portanto, é necessário dosar nível de confiança e margem de erro, para se chegar aos resultados mais apropriados.

! ATENÇÃO

De acordo com TRIOLA (2008, p. 255) , " um valor crítico é um número na fronteira que separa estatísticas amostrais que têm chance de ocorrer daquelas que não tem. O número $Z_{\alpha/2}$ é um valor crítico que é um escore z com a propriedade de separar uma área de $\alpha/2$ na cauda direita da distribuição normal padronizada".

4.7 Intervalos de Confiança para a Média Populacional

Podemos construir intervalos de confiança para a média populacional considerando 2 casos: σ conhecido ou σ desconhecido. Estudaremos cada um dos casos a seguir.

1º CASO – Com σ conhecido

Para construirmos um intervalo de confiança para a média populacional com σ conhecido temos que verificar os seguintes requisitos:

1. A amostra em estudo é uma amostra aleatória simples.
2. O valor do desvio padrão populacional, σ, é conhecido.
3. Uma ou ambas as condições são satisfeitas: a população é normalmente distribuída ou $n > 30$.

Um intervalo de confiança para a média populacional, verificado os requisitos acima, é dado por:

$$\bar{x} \pm z_{a/2} \times \frac{S}{\sqrt{n}} \qquad (4.11)$$

onde $1 - \alpha$ é o nível de confiança e $z_{a/2}$ é o valor tabelado da distribuição normal padronizada.

O valor obtido de $z_{a/2} \times \frac{S}{\sqrt{n}}$ é chamado margem de erro ou erro máximo da estimativa.

2º CASO – Com σ desconhecido

Para construirmos um intervalo de confiança para a média populacional com σ desconhecido temos que verificar os seguintes requisitos:

1. A amostra em estudo é uma amostra aleatória simples.

! ATENÇÃO

Quando coletamos um conjunto de dados amostrais para estimar um parâmetro populacional, o valor obtido pelo estimador deste parâmetro é tipicamente diferente do valor do parâmetro. A diferença entre estes dois valores é chamada margem de erro ou erro máximo de estimativa, ou seja, é a diferença máxima provável entre a estimativa obtida através do estimador e o verdadeiro valor do parâmetro populacional.

2. Uma ou ambas as condições são satisfeitas: a população é normalmente distribuída ou $n > 30$.

Um intervalo de confiança para a média populacional, verificado os requisitos acima, é dado por:

$$\bar{x} \pm t_{a/2} \times \frac{s}{\sqrt{n}} \qquad (4.12)$$

onde $1 - \alpha$ é o nível de confiança, $t_{a/2}$ é o valor tabelado da distribuição t de Student com $n - 1$ graus de liberdade e s é o desvio-padrão amostral.

O valor obtido de $t_{a/2} \times \frac{s}{\sqrt{n}}$ é chamado margem de erro ou erro máximo da estimativa.

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 $z = 0$; 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.

O quadro a seguir resume os Casos 1 e 2.

DISTRIBUIÇÃO	CONDIÇÕES
Use a distribuição normal (z)	σ conhecido e população normalmente distribuída ou σ conhecido e $n > 30$
Use a distribuição t	σ desconhecido e população normalmente distribuída ou σ desconhecido e $n > 30$
Use um método não paramétrico ou bootstrap	População não é normalmente distribuída e $n \leq 30$

Notas: 1. 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, com uma única moda e sem outliers.
2. Tamanho amostral *n > 30*: Essa é uma diretriz comumentemente usada, mas tamanhos amostrais de 15 a 30 são adequados se a população parecer ter uma distribuição que não se afasta muito da normal e se não há outliers. Para algumas distribuições populacionais que se afastam extremamente da normal, o tamanho amostral pode precisar ser maior do que 50, ou mesmo 100.

Quadro 4.1 – Escolha entre ***z*** e ***t***

Fonte: (TRIOLA, 2008, p. 280).

4.8 Intervalos de Confiança para a Proporção Populacional

Vimos na distribuição amostral do estimador $\hat{p}$ que, para n grande,

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Verificados os requisitos, temos que um intervalo de confiança para a proporção é dado por:

$$\hat{p} \pm z_{a/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

onde 1 – α é o nível de confiança e $z_{a/2}$ é o valor tabelado da distribuição normal padronizada.

O valor obtido de $z_{a/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ é chamado ***margem de erro*** ou ***erro máximo da estimativa.***

CONEXÃO

Os requisitos necessários para a costrução de intervalos de confiança para a proporção populacionam exigem conhecimento da distribuição de probabilidade Binomial. Para isto, leia a Seção 7.2 Estimação da Proporção Populacional, em (TRIOLA, 2008, p. 252).

Exemplo 4.3: De uma amostra de 40 observações de uma população normal com média desconhecida e desvio-padrão σ =5, obtemos uma média amostral $\bar{x} = 25$.

Construir um intervalo de 95% de confiança para a média populacional.

Resolução:

Neste exemplo vamos usar a fórmula do intervalo de confiança descrito no 1º CASO. Os dados que o exercício fornece são:

$$\bar{x} = 25; \quad n = 40; \quad s = 5; \quad 1 - a = 0,95;$$
$$a = 0,05; \quad a/2 = 0,025; \quad z_{a/2} = 1,96$$

Substituindo na fórmula temos:

$$\left[\bar{x} - z_{a/2} \times \frac{s}{\sqrt{n}} \; ; \; \bar{x} + z_{a/2} \times \frac{s}{\sqrt{n}}\right]$$

$$\left[25 - 1,96 \times \frac{5}{\sqrt{40}} \; ; \; 25 + 1,96 \times \frac{5}{\sqrt{40}}\right]$$

$$\left[25 - \frac{9,8}{6,3246} \; ; \; 25 + \frac{9,8}{6,3246}\right]$$

$$[25 - 1,5495 \; ; \; 25 + 1,5495]$$

$$[23,4505 \; ; \; 26,5495]$$

Podemos interpretar este intervalo da seguinte maneira: estamos 95% confiantes que o intervalo de 23,4505 a 26,5495 realmente contenha o verdadeiro valor de μ. Ou ainda, se selecionássemos muitas amostras diferentes de tamanho 40 e construíssemos os intervalos de confiança como fizemos aqui, 95% deles conteriam realmente o valor da média populacional μ.

Exemplo 4.4: Uma amostra de tamanho 15, extraída de uma população normal, fornece uma média amostral $\bar{x}$ = 23 e s = 0,5. Construir um intervalo de 90% de confiança para a média populacional.

Resolução:

Vamos usar o intervalo de confiança descrito no 2º CASO, pois temos uma população normal com σ desconhecido. Os dados são:

$$\bar{x} = 23,5; \quad n = 15; \quad S = 0,5; \quad 1 - a = 0,9; \quad a = 0,1;$$
$$a/2 = 0,05; \quad t_{a/2} = 1,761; \quad n - 1 = 14$$

Usando a fórmula:

$$\left[\bar{x} - t_{a/2} \times \frac{S}{\sqrt{n}} \; ; \; \bar{x} + t_{a/2} \times \frac{S}{\sqrt{n}}\right]$$

$$\left[23,5 - 1,761 \times \frac{0,5}{\sqrt{15}} \; ; \; 23,5 + 1,761 \times \frac{0,5}{\sqrt{15}}\right]$$

$$\left[23,5 - \frac{0,8805}{3,8730} \; ; \; 23,5 + \frac{0,8805}{3,8730}\right]$$

$$[23,5 - 0,2273 \; ; \; 23,5 + 0,2273]$$

$$[23,2727 \; ; \; 23,7273]$$

Exemplo 4.5: Em uma cidade foram entrevistadas 2.000 pessoas e constatou-se que 1.200 estão satisfeitas com o atual prefeito. Construir um intervalo de 95% de confiança para a proporção populacional que está satisfeita com o atual prefeito.

Resolução:

Os dados são:

$$\hat{p} = \frac{1.200}{2.000} = 0,6; \quad 1 - \hat{p} = 0,4; \quad n = 2.000; \quad 1 - a = 0,95;$$

$$a = 0,05; \quad a/2 = 0,025; \quad z_{a/2} = 1,96$$

Um intervalo de 95% de confiança para a proporção populacional é dado por:

$$\left[\hat{p} - z_{a/2} \times \frac{S}{\sqrt{n}} \; ; \; \hat{p} + z_{a/2} \times \frac{S}{\sqrt{n}}\right]$$

$$\left[0,6 - 1,96 \times \sqrt{\frac{0,6 \times 0,4}{2.000}} \; ; \; 0,6 + 1,96 \times \sqrt{\frac{0,6 \times 0,4}{2.000}}\right]$$

$$[0,6 - 1,96 \times 0,01095 \; ; \; 0,6 + 1,96 \times 0,01095]$$

$$[0,6 - 0,021462 \; ; \; 0,6 + 0,021462]$$

$$[0,57854 \; ; \; 0,62146]$$

Se quisermos um intervalo de 95% de confiança para a porcentagem populacional podemos expressar este resultado como [57,85% ; 62,15%].

ATENÇÃO

Podemos interpretar este intervalo da seguinte maneira: entre os moradores desta cidade, a porcentagem dos que estão satisfeitos com o atual prefeito é estimada em 60%, com uma margem de erro de ± 2,15%.

ATIVIDADE

1. Uma agência de publicidade está interessada em estimar a idade média em que os adolescentes começam a fumar. Uma amostra aleatória de 25 fumantes, extraída de uma população normal, forneceu uma média amostral de 15 anos e um desvio-padrão amostral de 1,7 ano. Construir um intervalo de 99% de confiança para estimar a idade média em que a população adolescente começa a fumar. Determine a margem de erro e o erro padrão estimado da média.

2. 02. A fim de averiguar a popularidade da gestão da nova reitoria de determinada universidade, uma amostra aleatória de 400 estudantes foi selecionada e constatou-se que 45% estavam satisfeitos com a nova gestão.
   a) Obtenha o erro padrão estimado da proporção;
   b) Determine o intervalo de 95% de confiança para a proporção populacional.

3. Os dados abaixo referem-se ao número de horas semanais que os estudantes do primeiro semestre de administração passaram se preparando para o exame de Estatística.

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

   Determine:
   a) o erro padrão estimado da média;
   b) o intervalo de 96% de confiança para a média populacional.

4. Dos 1.600 funcionários entrevistados numa empresa, 880 estão de acordo com a nova política salarial. Construir um intervalo de 98% de confiança para a proporção populacional dos funcionários desta empresa favoráveis à nova política salarial.

5. Para se avaliar a popularidade de certo candidato à próxima eleição para prefeito de determinada cidade, extraiu-se uma amostra aleatória de 1.000 eleitores e constatou-se que 400 votariam no candidato. Estimar a proporção de eleitores em toda a cidade que têm a intenção de votar no candidato. Encontre o erro padrão estimado da proporção.

REFLEXÃO

Vimos que é possível encontrar a melhor estimativa pontual da média e da proporção, mas não temos indicação de quão boa é esta nossa melhor estimativa. Para contornar isto, utilizamos a estimativa intervalar, ou intervalo de confiança, que consiste em uma faixa de valores em vez de apenas um único valor. Como dissemos anteriormente, precisamos ser cuidadosos para interpretar corretamente os intervalos de confiança. Na interpretação correta, o nível de confiança se refere à taxa de sucesso do *processo* em uso para se estimar o parâmetro populacional e não a chance de que o verdadeiro valor do parâmetro esteja entre os limites do intervalo de confiança. Por exemplo, um nível de confiança de 95% nos diz que o *processo* em uso resultará, a longo prazo, em limites de intervalo de confiança que contenham o verdadeiro valor do parâmetro populacional 95% das vezes.

LEITURA

Sugerimos que você ouça os áudios que estão no seguinte endereço: <http://m3.ime.unicamp.br/recursos/1288>. Nestes áudios você aprenderá o significado da expressão margem de erro no contexto da Matemática.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. Estatística aplicada à administração e economia. São Paulo: Pioneira Thomson Learning, 2003.

BUSSAB, Wilton de O.; MORETTIN, Pedro A.. Estatística básica. São Paulo: Saraiva, 2003.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. Introdução à estatística. Rio de Janeiro,: LTC, 2003.

LARSON, R; FARBER, Betsy. Estatística aplicada. São Paulo: Pearson Prentice Hall, 2004.

SMAILES, Joanne; McGRANE, Angela. Estatística aplicada à administração com Excel. São Paulo: Atlas, 2002.

SPIEGEL, Murray R.. Estatística. São Paulo: Makron Books, 1993.

TRIOLA, Mario F.. Introdução à estatística. Rio de Janeiro: LTC, 2008.

NO PRÓXIMO CAPÍTULO

Neste capítulo estudamos a técnica de Estimação, uma das mais importantes da inferência estatística. Usando dados amostrais, aprendemos a obter estimativas pontuais e intervalares para parâmetros populacionais importantes: média e proporção. Para a construção de intervalos de confianças vimos que precisamos encontrar valores críticos da distribuição normal padrão ou da distribuição *t-student*.

No próximo capítulo estudaremos outra técnica da inferência estatística: teste de hipótese. E, nos aprofundaremos no estudo da distribuição normal.

5

Distribuição Normal e Teste de Hipótese

5 Distribuição Normal e Teste de Hipótese

Neste capítulo estudaremos a distribuição de probabilidade mais importante em Estatística: *a distribuição normal*. As distribuições normais podem ser usadas para modelar muitos conjuntos de medidas na natureza, na indústria e no comércio. Por exemplo, a altura de uma determinada população ou a duração de aparelhos de televisão segue, em geral, uma distribuição normal.
Ao finalizarmos o conceito da distribuilçao normal aprenderemos outra técnica da inferência estatística, além da estimação, muito utilizada: teste de hipótese. Veremos as características dos testes paramétricos e não paramétricos, bem como os procedimentos necessários para a construção de um teste de hipótese.

OBJETIVOS

- Identificar situações nas quais podemos aplicar o modelo de probabilidade normal, bem como calcular probabilidades associadas a tal modelo.
- Compreender quais são as etapas necessárias para a realização de um teste de hipótese.

REFLEXÃO

Você se lembra de ter ouvido de alguma companhia de transporte que, em média, o intervalo entre sucessivos ônibus é de 20 minutos? Aprenderemos neste capítulo como podemos testar a afimartiva da companhia utilizando o conceito de teste de hipótese.
Antes de iniciarmos o estudo da distribuição normal, é importante a compreensão do conceito de variável aleatória.

5.1 Variável Aleatória

Uma variável aleatória (v.a.) é uma variável que associa um valor numérico a cada ponto do espaço amostral. Ela é denominada discreta quando pode assumir apenas um número finito ou infinito enumerável de valores e é dita contínua quando assume valores num intervalo da reta real.

É comum utilizarmos letras latinas para representarmos variáveis aleatórias.

Quando trabalhamos com uma v.a. que pode assumir valores num intervalo de números reais, como mensuração de peso, altura e temperatura, estamos lidando com uma distribuição contínua de probabilidade. Em distribuições deste tipo podemos construir uma curva contínua que é a representação gráfica da função densidade de probabilidade, usualmente designada por f(x).

5.2 Função Densidade de Probabilidade

Uma função *f(x)* é uma função densidade de probabilidade (f.d.p.) para uma v.a. contínua *X* se satisfaz as condições:

- $f(x) \geq o$ para todo $x \in (-\infty, \infty)$;
- a área definida por *f(x)* é igual a 1;
- $P(X = x_0) = 0$, ou seja, a probabilidade da v.a. assumir um valor pontual é zero.

O valor esperado e a variância de uma v.a. contínua são definidos, respectivamente, por:

$$E(X) = \int_{-\infty}^{\infty} x \times f(x)\,dx \tag{5.1}$$

e

$$Var(X) = E(X^2) - \left[E(X)\right]^2 \tag{5.2}$$

onde $E(X) = \int_{-\infty}^{\infty} x \times f(x)\,dx$. Essa expressão é uma integral que é um conceito matemático cuja compreensão não é nada elementar. Mas não se preocupe, pois não teremos que saber como calculá-la.

5.3 Modelo Probabilístico para Variáveis Aleatórias Contínuas

Estudaremos aqui a distribuição de probabilidade mais importante: A Distribuição Normal. Esta distribuição desempenha papel fundamental na Inferência Estatística. A curva da função densidade de probabilidade desta distribuição é conhecida por muitos como a "curva em forma de sino".

5.3.1 Distribuição Normal

A Distribuição Normal tem função densidade de probabilidade dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty \tag{5.3}$$

em que μ e σ^2 são os parâmetros da distribuição.

A Figura 11 ilustra uma curva normal típica:

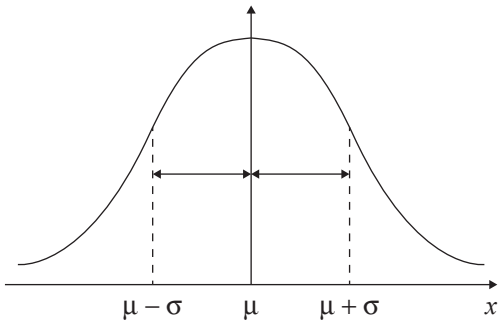


Figura 11 – f.d.p. de uma v.a. Normal com média μ e desvio-padrão σ.

Representaremos v.a. ´s com distribuição normal por $X \sim N(\mu, \sigma^2)$.
As principais características da distribuição normal são:

- o ponto de máximo de $f(x)$ é o ponto $x = \mu$;
- os pontos de inflexão são: $x = \mu + \sigma$ e $x = \mu - \sigma$;
- a curva é simétrica com relação a μ;
- $f(x) \to 0$ quando $x \to \pm\infty$.

Para se obter o cálculo de probabilidades de uma v.a. $X \sim N(\mu, \sigma^2)$ devemos resolver a integral da função densidade de probabilidade no intervalo de interesse, isto é,

$$P(a \le X \le b) = \int_a^b f(x)\,dx \tag{5.4}$$

onde a integral indica a área sob a curva da densidade entre os pontos a e b.

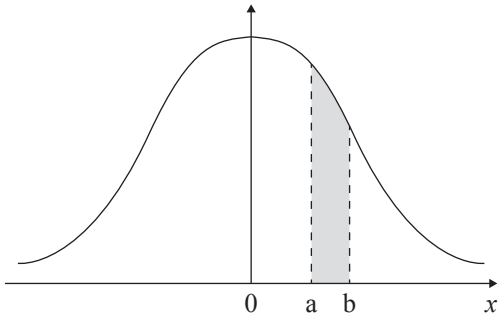


Figura 12 – área sob a curva normal, no intervalo de a a b

Esta integral só pode ser resolvida, aproximadamente, por meio de integração numérica. Para contornar esta dificuldade, as probabilidades para a distribuição normal são calculadas com o auxílio de tabelas. Para isto, utilizamos uma transformação da v.a. X em uma v.a. Z definida por:

$$Z = \frac{X - \mu}{\sigma} \tag{5.5}$$

onde μ = média e σ = desvio padrão.

Esta nova variável é denominada de variável normal padronizada com média 0 e variância 1, ou seja, $Z \sim N(0, 1)$

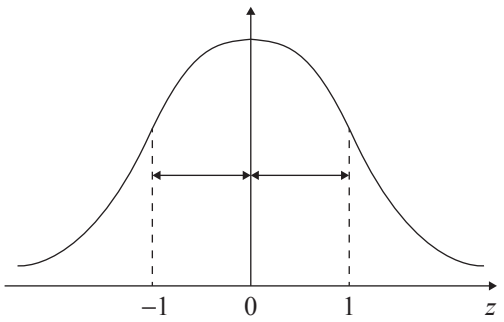


Figura 13 – f.d.p. de uma v.a. Z ~ N(0,1)

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

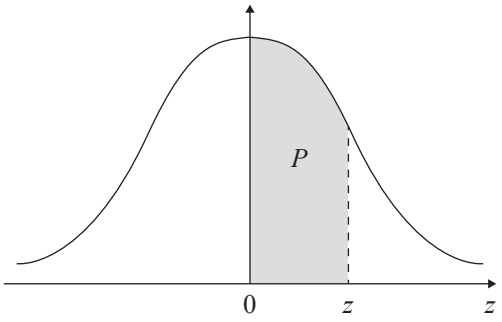


Figura 14 – área correspondente à P(0 ≤ Z ≤ zc) fornecida pela tabela

ATENÇÃO

A característica de simetria da distribuição normal implica que a probabilidade de estar acima (ou abaixo) de zero é 0,5, ou seja, $P(Z \geq 0) = 0,5 = P(Z \leq 0)$.

Exemplo 5.1: As vendas mensais de determinado produto têm distribuição aproximadamente normal, com média 500 unidades e desvio-padrão 50 unidades. Determine as probabilidades de que, em um mês, as vendas do produto sejam:

a) no máximo 530 unidades;
b) no mínimo 460 unidades;
c) entre 450 e 550 unidades;
d) no mínimo 530 unidades.

ATENÇÃO

O enunciado desse exemplo forneceu o valor do desvio-padrão que, por definição, é a raiz quadrada da variância. Portanto, $\sigma^2 = 50^2 = 2.500$ unidades.

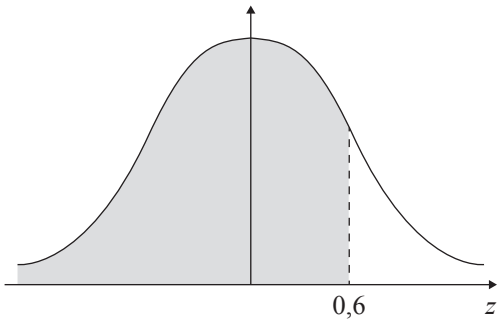
Resolução:

Vamos definir a v.a. como sendo *X*: *vendas mensais de determinado produto*. Portanto, $X \sim N(500,2500)$.

a) $P(X \leq 530)$

Para calcularmos esta probabilidade, vamos transformar a v.a. *X* na v.a. *Z* para podermos usar a tabela.

$$Z = \frac{X - \mu}{\sigma} = \frac{530 - 500}{50} = \frac{30}{50} = 0,6$$



Portanto,

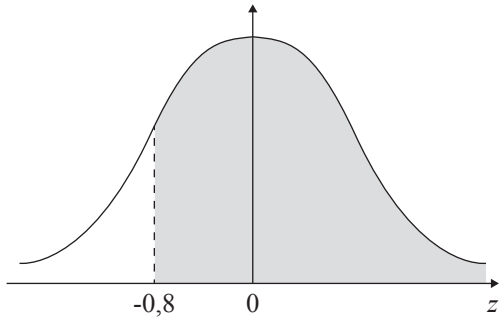
$P(X \leq 530) = P(Z \leq 0,6) = 0,5 + P(o \leq Z \leq o,6) = 0,5 + 0,2257 = 0,7257$

O valor 0,5 vem do fato que $P(Z \leq 0) = 0,5$.

b) $P(X \geq 460)$

Usando a transformação:

$$Z = \frac{X - \mu}{\sigma} = \frac{460 - 500}{50} = 40/50 = -0,8$$



Portanto,

$P(X \geq 460) = P(Z \geq -0{,}8) = P(-0{,}8 \leq Z \leq 0) + 0{,}5 = 0{,}2881 + 0{,}5 = 0{,}7881$

Observação.: Devido à simetria da distribuição normal temos que

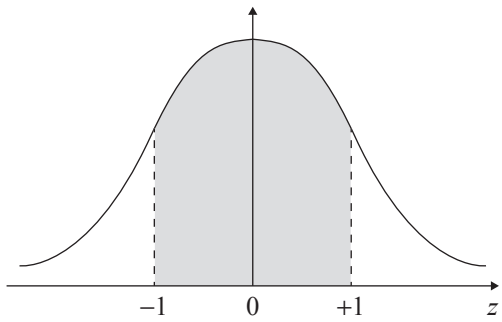
$P(-0{,}8 \leq Z \leq 0) = P(0 \leq Z \leq 0{,}8)$ e, como no item anterior, o valor 0,5 vem do fato que $P(Z \geq 0) = 0{,}5$.

c) $P(450 \leq X \leq 550)$

Transformando:

$$Z_1 = \frac{X - \mu}{\sigma} = \frac{450 - 500}{50} = \frac{50}{50} = -1$$

$$Z_2 = \frac{X - \mu}{\sigma} = \frac{550 - 500}{50} = \frac{50}{50} = 1$$



$P(450 \leq X \leq 550) = P(-1 \leq Z \leq 1) = P(-1 \leq Z \leq 0) + P(0 \leq Z \leq 1) =$

$0{,}3413 + 0{,}3413 = 0{,}3826$

Observação: Devido à simetria $P(-1 \leq Z \leq 0) + P(0 \leq Z \leq 1)$.

d) $P(X \geq 530)$

Para calcularmos esta probabilidade, vamos transformar a v.a. X na v.a. Z para podermos usar a tabela.

$$Z = \frac{X - \mu}{\sigma} = \frac{530 - 500}{50} = \frac{30}{50} = 0,6$$

Portanto,

$P(X \geq 530) = P(Z \geq 0,6) = 0,5 - P(0 \leq Z \leq 0,6) = 0,5 - 0,2257 = 0,2743$

O valor 0,5 vem do fato que $P(Z \geq 0) = 0,5$

Exemplo 5.2: O tempo de vida médio de certo aparelho é de dez anos, com desvio-padrão de 1,5 ano. O fabricante substitui os aparelhos que acusam defeito dentro do prazo de garantia. Qual deve ser o prazo de garantia para que a porcentagem de aparelhos substituídos seja no máximo 5%?

Resolução

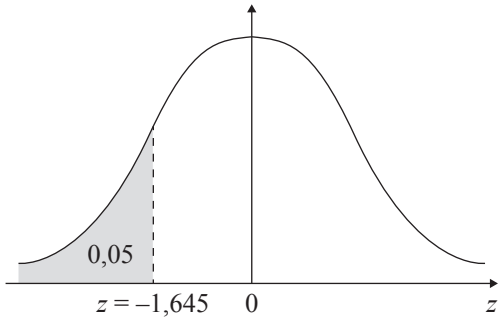
X: tempo de vida do aparelho

$X \sim N(10, 2,25)$

Neste exercício queremos encontrar X_c: *prazo de garantia para que no máximo 5% dos aparelhos produzidos sejam substituídos dentro desse prazo*. Observe que o exercício já forneceu a probabilidade e queremos encontrar qual o valor correspondente a esta probabilidade, isto é,

$$P(X \geq X_c) = 0,05$$

Transformando na v.a. Z temos $P(Z \geq Z_c) = 0,5$ e procurando no corpo da tabela 0,45 (0,5-0,05) encontramos $Z_c = -1,645$.



Portanto:

$$Z_c = \frac{X_c - \mu}{\sigma}$$

$$-1,645 = \frac{X_c - \mu}{1,5}$$

$$X_c - 10 = -2,4675$$

$$X_c = -2,4675 + 10$$

$$X_c = 7,5325$$

Então, o prazo de garantia deve ser de 7,5 anos.

5.4 Teste de Hipótese

A inferência estatística utiliza os dados amostrais principalmente para: estimar um parâmetro populacional (como vimos no Capítulo 4) e para testar uma hipótese ou uma afirmativa sobre um parâmetro populacional (que veremos neste capítulo).

Um teste de hipótese usa estatísticas amostrais para testar uma afirmativa sobre uma propriedade da população. Por exemplo, pesquisadores da área médica e da política utilizam testes de hipóteses para a tomada de decisões sobre novos medicamentos ou resultados de uma eleição.

Os testes de hipótese podem ser paramétricos ou não paramétricos.

Testes paramétricos têm requisitos sobre a natureza ou a forma das populações envolvidas. São baseados em parâmetros da distribuição.

Testes não paramétricos não exigem que as amostras sejam provenientes de populações normais ou qualquer outra distribuição específica. Com isso, em geral, são chamados ***testes livres de distribuição.***

5.4.1 Passos para a construção de um teste de hipótese

1. Dada uma afirmativa, identificar a hipótese nula e a hipótese alternativa e expressar ambas em forma simbólica;
2. Através de uma afirmativa e dos dados amostrais, calcular o valor da estatística de teste.
3. Fixar α e identificar o(s) valor(es) crítico(s).
4. Concluir o teste com base na estatística de teste e na região crítica.

5.4.2 Hipótese nula e hipótese alternativa

- **A hipótese nula** (representada por H_0) é uma afirmativa de que o valor de um parâmetro populacional é ***igual*** a algum valor especificado.
- **A hipótese alternativa** (representada por H_1 ***ou*** H_a) é a afirmativa de que o parâmetro tem um valor que, de alguma forma, difere da hipótese nula. 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:

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

Segundo (TRIOLA, 2008, p. 309), "se você está fazendo um estudo e deseja usar um teste de hipótese 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 $\neq$. Você não pode usar um teste de hipótese para apoiar uma afirmativa de que um parâmetro seja *igual* a algum valor específico".

! ATENÇÃO

Alguns livros texto usam os símbolos ≤ ou ≥ na hipótese nula H_0, mas seguiremos a notação da maioria dos periódicos profissionais que usam apenas o símbolo de igualdade.

Exemplo 5.3: Identifique as hipóteses nulas e alternativa e identifique qual representa a afirmação em cada um dos itens abaixo.

a) Uma universidade alega que a proporção de seus alunos que são do sexo masculino é de 46%. O departamento de *marketing* da universidade deseja testar esta afirmação.
b) Os amortecedores de automóveis que circulam em cidades duram, em média, 35.000 quilômetros, segundo informação de algumas oficinas especializadas. Um proprietário de automóvel deseja testar essa afirmação.
c) Um veterinário conseguiu ganho médio diário de 2,5 litros de leite por vaca com uma nova composição de ração. Um pecuarista acredita que o ganho não é tão grande assim.

Resolução:

a) $H_0 : p = 0{,}46$ afirmação
$H_1 : p \neq 0{,}46$

b) $H_0 : \mu = 35.000$ afirmação
$H_1 : \mu \neq 35.000$

c) $H_0 : \mu = 2{,}5$ afirmação
$H_1 : \mu < 2{,}5$

5.4.3 Estatística de teste paramétrico para a média

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* ou χ^2) com a suposição de que a hipótese nula seja verdadeira.

Estudaremos, neste capítulo, a seguinte estatística de teste para a média:

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad \text{ou} \qquad t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Podemos observar que esta estatística de teste pode se basear na distribuição normal ou na distribuição t de *Student*, dependendo das condições que sejam satisfeitas. Utilizaremos, aqui, os mesmos requisitos descritos no item 4.7 (ver quadro 4.1).

5.4.4 Região crítica, nível de significância e valor crítico

A ***região crítica*** é composta por todos os valores da estatística de teste que nos fazem rejeitar a hipótese nula.

O nível de significância (α) é a probabilidade da estatística de teste cair na região crítica quando a hipótese nula for realmente verdadeira. Portanto, α é a probabilidade de cometermos o erro de rejeitar a hipótese nula quando ela é verdadeira. Esse erro é conhecido como **Erro do tipo I**. Este nível de significância α é o mesmo que aquele foi definido para a construção do intervalo de confiança no Capítulo 4, cujas escolhas comuns para α são 0,05; 0,01 e 0,10.

5.4.5 Teste bilateral, unilateral à esquerda e unilateral à direita

Um teste de hipótese pode ser bilateral, unilateral à esquerda ou unilateral à direita. O tipo de teste depende da região da distribuição amostral que favorece uma rejeição de $\boldsymbol{H_0}$

Temos que:

- Se a hipótese alternativa $\boldsymbol{H_1}$ contiver o símbolo $<$, o teste de hipótese será um teste unilateral à esquerda, ou seja, a região crítica está na cauda esquerda sob a curva;
- Se a hipótese alternativa $\boldsymbol{H_1}$ contiver o símbolo $>$, o teste de hipótese será um teste unilateral à direita, ou seja, a região crítica está na cauda direita sob a curva;
- Se a hipótese alternativa $\boldsymbol{H_1}$ contiver o símbolo $\neq$, o teste de hipótese será um teste bilateral, ou seja, a região crítica está nas duas caudas sob a curva;

Nos testes bilaterais, o nível de significância α é dividido igualmente entre as duas caudas que constituem a região crítica. Em testes unilaterais à esquerda ou à direita, a área da região crítica na cauda respectiva é α.

5.4.6 Conclusão do teste de hipótese

O objetivo de um teste de hipótese é testar a hipótese nula, de modo que nossa conclusão será uma das seguintes:

1. Rejeitar a hipótese nula.
2. Deixar de rejeitar a hipótese nula.

CONEXÃO

O uso do método do valor *P* está sendo utilizado com bastante frequência, pois tal valor aparece nos resultados de pacotes estatísticos. Para a compreensão de tal método, leia o **Procedimento para a Determinação de Valores *P***, que se encontra em (TRIOLA, 2008, p. 314).

A decisão de rejeitar ou deixar de rejeitar uma hipótese nula pode ser feita utilizando o método tradicional (método clássico), o método do valor P, ou baseando-se em intervalos de confiança.

Utilizaremos o método clássico para concluir um teste de hipótese.

Quando concluímos um teste de hipótese e a estatística de teste não cair na região crítica, vamos usar a terminologia deixar de rejeitar a hipótese nula. Alguns textos escrevem *aceitar a hipótese nula*. Mas, *não estamos provando a hipótese nula*. Apenas estamos nos baseando em evidências amostrais que não garantiram a rejeição da hipótese nula e, por isso, o termo *deixar de rejeitar* parece o mais correto.

Exemplo 5.4: Uma grande revista de negócios brasileira afirmou que o faturamento médio das indústrias de uma determinada região do sul do país seria igual a R$ 820.000,00. Sabe-se que o desvio padrão do faturamento de todas as empresas da região é igual a R$ 120.000,00? Um pesquisador independente analisou os dados de uma amostra formada por 35 empresas, encontrando um faturamento médio igual a R$ 780.000,00. Assumindo nível de significância igual a 8%, seria possível concordar com a alegação?

Resolução:

1. Identificar a hipótese nula e alternativa:

$$H_0 : \mu = 820.000$$
$$H_1 : \mu \neq 820.000$$

2. Através de uma afirmativa e dos dados amostrais, calcular o valor da estatística de teste.

Como o desvio padrão é conhecido e n > 30, utilizaremos a seguinte estatística de teste:

$$Z = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}}$$

$$Z = \frac{780.000 - 820.000}{\frac{120.000}{\sqrt{35}}} = \frac{40.00}{20283,702} = -1,97$$

3. Fixar α e identificar o(s) valor(es) crítico(s).

Do enunciado, temos $\alpha = 0,08$. Como o teste é bilateral, o nível de significância $\alpha = 0,08$ é dividido igualmente entre as dus caudas que constituem a região crítica. Portanto:

$$Z_{\alpha/2} = -1,755 \quad \text{e} \quad Z_{\alpha/2} = 1,755$$

4. Concluir o teste com base na estatística de teste e na região crítica.

Temos que a estatística de teste caiu na região de rejeição, do lado esquerdo da cauda, pois – 1,97 < –1,755. Portanto, rejeitamos a hipótese nula.

Há evidência suficiente para garantir a rejeição da afirmativa de que o faturamento médio das indústrias de uma determinada região do sul do país é de R$ 820.000,00.

Exemplo 5.5: Uma grande construtora nacional afirma que seus funcionários recebem um salário médio igual a, no mínimo, R$ 1.450,00, com desvio padrão igual a R$ 700,00 e a distribuição supostamente normal. Uma amostra com 500 funcionários apresentou uma média de R$ 1 000,00. A alegação da empresa poderia ser aceita? Justifique. Considere $\alpha = 2\%$.

Resolução:

1. Identificar a hipótese nula e alternativa:

$$H_0 : \mu = 1.450$$
$$H_1 : \mu < 1.450$$

2. Através de uma afirmativa e dos dados amostrais, calcular o valor da estatística de teste.

Como a distribuição é supostamente normal e o desvio padrão é conhecido, utilizaremos a seguinte estatística de teste:

$$Z = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}}$$

$$Z = \frac{1.000 - 1.450}{\frac{700}{\sqrt{500}}} = \frac{450}{31,204952} = -14,37$$

3. Fixar α e identificar o(s) valor(es) crítico(s).

Do enunciado, temos $\alpha = 0,02$. Como o teste é unilateral à esquerda, a área da região crítica na cauda respectiva é α:

$$Z_{\alpha} = -2,055$$

4. Concluir o teste com base na estatística de teste e na região crítica.

Temos que a estatística de teste caiu na região de rejeição, pois $-14{,}37 < -2{,}055$.

Portanto, rejeitamos a hipótese nula.

Há evidência suficiente para garantir a rejeição da afirmativa de que salário médio dos funcionários da construtora seja de, no mínimo, R$ 1.450,00.

5.4.7 Testes não paramétricos

Como vimos no item 5.4, os testes paramétricos têm requisitos sobre a natureza ou a forma das populações envolvidas. São baseados em parâmetros da distribuição. Quando não for possível supor ou assumir características sobre parâmetros da população de onde os dados foram extraídos, torna-se necessário aplicar testes não paramétricos de hipótese.

Listaremos os principais testes na paramétricos e em que situação devem ser utilizados.

De acordo com (BRUNI, 2010, p. 256), "dentre os principais modelos de testes não parmétricos, podem ser destacados os relacionados a seguir:

a) **Teste do qui-quadrado**: empregado na análise de frequências, quando uma característica da amostra é analisada;
b) **Teste do qui-quadrado para independência ou associação**: também empregado na análise de frequências, porém quando duas características da amostra são analisadas;
c) **Teste dos sinais**: empregado no estudo de dados emparelhados, quando um mesmo elemento é submetido a duas medidas;
d) **Teste de Wilcoxon**: também analisa dados emparelhados, permitindo, porém, uma consideração das magnitudes encontradas;
e) **Teste de Mann-Whitney**: analisa se dois grupos originam-se de populações com médias diferentes;
f) **Teste da mediana**: analisa se dois grupos originam-se de populações com medianas diferentes;
g) **Teste de Kruskal-Wallis**: analisa se mais de dois grupos originam-se de populações com médias diferentes".

Os testes não paramétricos não são tão eficientes quanto os testes paramétricos. Então, precisamos, em geral, de evidência mais forte (amostra maior ou diferenças maiores) para rejeitar a hipótese nula.

ATIVIDADE

1. A durabilidade de um tipo de pneu de determinada marca é descrita por uma v.a.Normal de média 70.000 km e desvio-padrão de 9.000 km.
   a) Se o fabricante desta marca garante os pneus pelos primeiros 50.000 km,qual a proporção de pneus que deverão ser trocados pela garantia?
   b) Qual deve ser a garantia (em km) para assegurar que o fabricante troque sob garantia no máximo 2% dos pneus?

2. As vendas de determinado produto têm distribuição aproximadamente normal, com média 700 unidades e desvio-padrão 80 unidades. Se a empresa decide fabricar 800 unidades no mês em estudo, qual é a probabilidade de que não possa atender a todos os pedidos desse mês, por estar com a produção esgotada?

3. As velocidades dos carros numa rodovia têm distribuição normal, com média de 90km/h. Determinar:
   a) o desvio-padrão das velocidades, se 5% dos carros ultrapassar 100 km/h;
   b) a porcentagem dos carros que trafegam a menos de 80 km/h.

4. Uma fábrica de embalagens de papelão afirma que suas caixas modelo padrão têm uma resistência média não inferior a 14 kg. Uma amostra de cinco caixas revelou uma resistência média igual a 12,6 kg. Assumindo um nível de significância igual a 2%, é possível confiar na palavra da fábrica? Sabe-se que o desvio padrão populacional das resistências das caixas é igual a 2 kg e que esta variável encontra-se normalmente distribuída.

5. 05. A campanha WZA fabrica um determinado analgésico que alega ter duração não inferior a quatro horas. Uma análise de 30 medicamentos escolhidos aleatoriamente acusou uma média de 3,8 horas de duração. Teste a alegação da campanhia, contra a alternativa de que a duração seja inferior a quatro horas ao nível de 0,05, se o desvio populacional for de 0,5 hora.

REFLEXÃO

Agora, acreditamos que você poderá interpretar de maneira mais profunda muitas das informações que recebe. O conhecimento dos conceitos abordados neste livro é de fundamental importância nas análises que qualquer profissional necessita fazer em seu cotidiano. Comprovadamente, o uso da Estatística em qualquer área leva a tomada de decisões com maiores chances de acerto. Agora, aplicar ou não o que você aprendeu, depende exclusivamente de você. Boa sorte e muito sucesso!

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

ANDERSON, David R.; SWEENEY, Denis J.; WILLIAMS, Thomas A. Estatística aplicada à administração e economia. São Paulo: Pioneira Thomson Learning, 2003.

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

BUSSAB, Wilton de O.; MORETTIN, Pedro A.. Estatística básica. São Paulo: Saraiva, 2003.

FARIAS, Alfredo Alves de; SOARES, José Francisco; CÉSAR, Cibele Comini. Introdução à estatística. Rio de Janeiro,: LTC, 2003.

LARSON, R; FARBER, Betsy. Estatística aplicada. São Paulo: Pearson Prentice Hall, 2004.

MAGALHÃES, Marcos Nascimento; LIMA, Antônio Carlos Pedroso de. Noções de probabilidade e estatística. São Paulo: Editora da Universidade de São Paulo, 2004.

TRIOLA, Mario F.. Introdução à estatística. Rio de Janeiro: LTC,.

EXERCÍCIO RESOLVIDO

Capítulo 1

1. Resposta
   a) qualitativa nominal
   b) quantitativa discreta
   c) quantitativa contínua
   d) quantitativa discreta
   e) quantitativa contínua
   f) quantitativa contínua
   g) quantitativa discreta
   h) qualitativa nominal
   i) quantitativa contínua
   j) quantitativa contínua
   k) qualitativa ordinal
   l) qualitativa ordinal
   m) qualitativa nominal
   n) qualitativa nominal

2. Resposta
   a) Como estamos interessados nas respostas dos consumidores de refrigerantes sabor cola no teste de sabor, um consumidor desse tipo de refrigerante é uma unidade experimental. Assim, a população de interesse é a coleção ou conjunto de todos esses consumidores.
   b) A característica que a Pepsi deseja medir é a preferência do consumidor de refrigerante sabor cola revelada sob a aplicabilidade de um teste cego, logo, a preferência pelo tipo de refrigerante é a variável de interesse.
   c) A amostra é de 1.000 consumidores de refrigerante sabor cola selecionados da população de todos os consumidores desse tipo de refrigerante.
   d) A inferência de interesse é a generalização da preferência de refrigerantes sabor cola dos 1.000 consumidores da amostra para a população de todos os consumidores desse tipo de refrigerante. Em particular, as preferências dos consumidores da amostra podem ser usadas para estimar o percentual de todos os consumidores que preferem cada marca.

3. Qualitativos.

   Qualitativos, pois as categorias foram simplesmente codificadas. Mas, isto não torna a variável quantitativa. Não há sentido, por exemplo, calcular a média para estes dados codificados.

4. Resposta
   a) Todos os cidadãos brasileiros.
   b) Avaliação do trabalho do presidente (bom ou mau); qualitativa.
   c) 2.500 indivíduos sorteados.
   d) Estimar a proporção de todos os cidadãos que acreditam que o presidente está fazendo um bom trabalho.
   e) Pesquisa.
   f) A amostra em estudo não é representativa, pois foram entrevistadas somente pessoas que possuem telefone.

5. Aproximadamente 48.

6. Resposta
   a) Sim, será representativa.
   b) Foram utilizados pelo menos 3 tipos de técnicas de amostragem: Amostragem Estratificada no primeiro momento, Amostragem casual simples no segundo momento e Amostragem por meio de conglomerados para finalizar.

7. Resposta
   a) $\bar{x}_p = 161.33;\ \bar{x}_g = 1279,85$
   b) $\bar{x}_{\text{ponderada}} = 524,85$

8. Resposta

a)

Nº DE FUNCIONÁRIOS	Nº DE PROPRIEDADES	AMOSTRA ESTRATIFICADA (N = 50)	
		UNIFORME	PROPORCIONAL
0 ├ 20	500	10	25
20 ├ 50	320	10	15
50 ├ 100	100	10	6
100 ├ 200	50	10	2
200 ├ 400	30	10	2
Total	**1.000**	**50**	**50**

b) $\bar{x}_u = 114;\ \bar{x}_p = 42,5$ A média obtida através da amostragem estratificada uniforme não mostra a realidade das empresas com relação ao número de funcionários, já que a grande maioria tem no máximo 50 funcionários (como mostra a média obtida através da amostragem estratificada proporcional).

Capítulo 2

Antes das respostas gostaríamos de deixar claro que as interpretações das questões ficam a cargo do estudante. Se ocorrer dúvidas, entrar em contato com o tutor.

1. k ≈ 6 classes e amplitude da classe h ≈7

IDADES	f	f_r	f_a
19 ├ 26	5	0,1667	5
26 ├ 33	13	0,4333	18
33 ├ 40	4	0,1333	22
40 ├ 47	4	0,1333	26
47 ├ 54	3	0,1000	29
54 ├ 61	1	0,0333	30
Total	**30**	**1**	

Tabela 1: Distribuição de frequências das idades dos funcionários.

a) 18
b) 13,33%
c) 17
d) 73,33%
e) 26,67%

2. Resposta
   a) Valores gastos com supermercado. Variável quantitativa contínua.
   b)

CLASSES(GASTOS EM R$)	f	f_r	f_a
4,89 ├ 61,89	17	0,34	17
61,89 ├ 118,89	13	0,26	30
118,89 ├ 175,89	5	0,10	35
175,89 ├ 232,89	5	0,10	40
232,89 ├ 289,89	3	0,06	43
289,89 ├ 346,89	3	0,06	46
346,89 ├ 403,89	2	0,04	48
403,89 ├ 460,89	2	0,04	50
Total	**30**	**1**	

Tabela 1: Distribuição de frequências para a variável Valores gastos com supermercado.

3. Resposta

a) $\bar{x} = 23{,}1,\ Md = 22{,}5,\ Mo = 22$

b)

CLASSES	f	f_r	f_a	Pm
10 ├ 14	3	0,0469	3	12
14 ├ 18	5	0,0781	8	16
18 ├ 22	12	0,1875	20	20
22 ├ 26	25	0,3906	45	24
26 ├ 30	12	0,1875	57	28
30 ├ 34	3	0,0469	60	32
34 ├ 38	2	0,0313	62	36
38 ├ 42	2	0,0313	64	40
Total	**64**	**1,0000**		

c) $\bar{x} = 24{,}06,\ Md = 23{,}92,\ Mo = 24$

4. Resposta

a) Vendas mensais. Variável quantitativa contínua.

b) $\bar{x} \approx 3{,}2,\ Md = 3{,}4,\ Mo = 3{,}5$

c) 16,36%

d) 21,82%

e) 56,36%

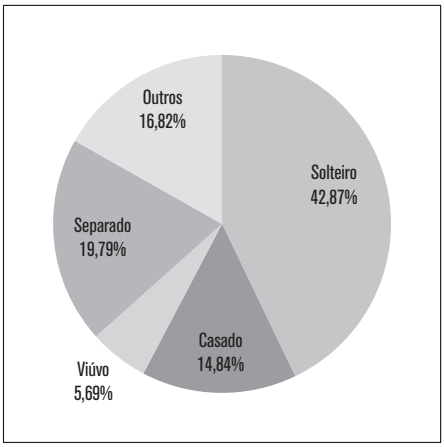
f) 65,45%

5. $\bar{x} \approx 1{,}7,\ Md = 2,\ Mo = 2$

6. 4,3

Capítulo 3

1. Resposta



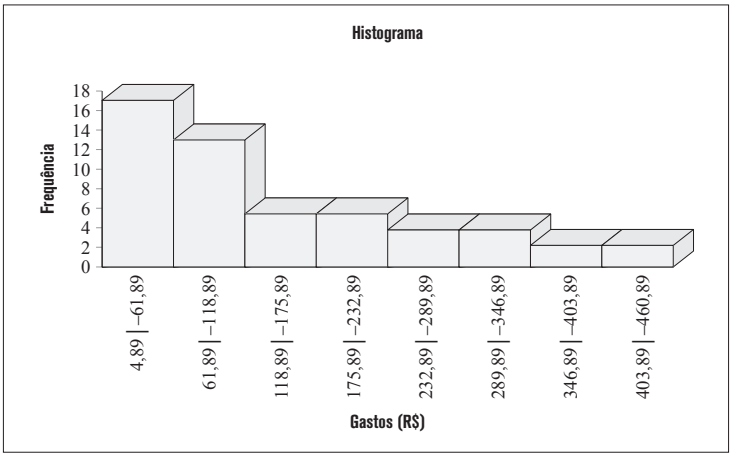
Através do gráfico, podemos dizer que aproximadamente 43% dos clientes desta agência de turismo são solteiros, 20% são separados, 17% têm outro tipo de estado civil, 15% são casados e apenas 5% são viúvos. Esta informação é importante na hora de lançar pacotes de viagens. A agência deve se lembrar que grande parte de seus clientes são solteiros. Também pode criar estratégias para trazer mais clientes casados ou viúvos, que provavelmente devem ter outro tipo de perfil.

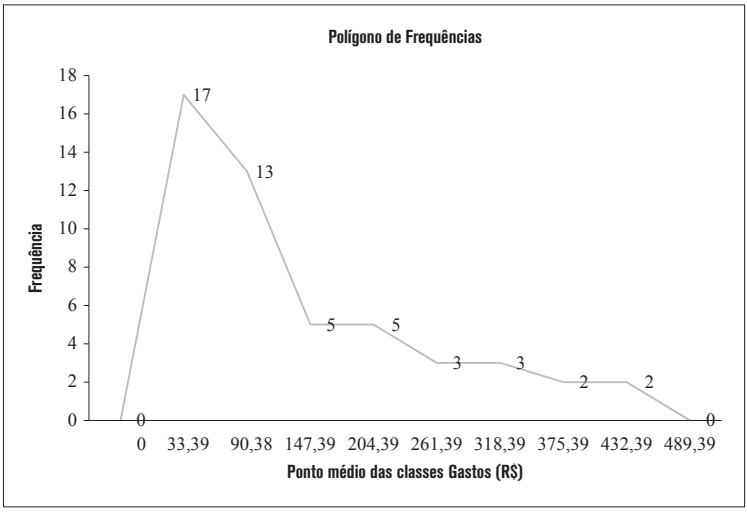
2. Resposta

a)

IDADES	f	f_r	f_a
4,89 ├ 61,89	17	0,34	17
61,89 ├ 118,89	13	0,26	30
118,89 ├ 175,89	5	0,10	35
175,89 ├ 232,89	5	0,10	40
232,89 ├ 289,89	3	0,06	43
289,89 ├ 346,89	3	0,06	46
346,89 ├ 403,89	2	0,04	48
403,89 ├ 460,89	2	0,04	50
Total	**50**	**1**	

b)





3. Resposta
   a) Salário de funcionários de uma empresa. Esta variável é classificada como quantitativa contínua.
   b) 45 funcionários
   c) 86 funcionários
   d) 9,30%
   e) 72,09%
   f) Tabela 1: Distribuição de frequências dos salários dos funcionários de uma empresa.

IDADES	f	f_r	f_a
500,00 ├ 800,00	17	19,77	17
800,00 ├ 1100,00	45	52,33	62
1100,00 ├ 1400,00	12	13,95	74
1400,00 ├ 1700,00	4	4,65	78
1700,00 ├ 2100,00	3	3,49	81
2100,00 ├ 2400,00	3	3,49	84
2400,00 ├ 2700,00	2	2,33	86
Total	**86**	**100**	

4. Resposta

a) $R = 31,\ s \approx 6,1,\ s^2 \approx 37,2,\ cv \approx 0,2633$ ou $26,33\%$

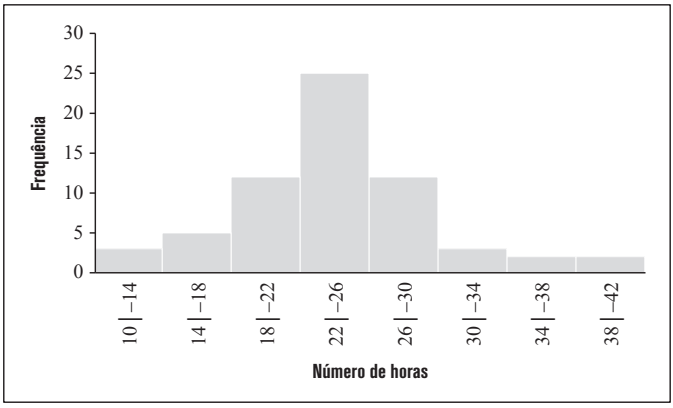
b)

CLASSES	f	f_r	f_a	Pm
10 ├ 14	3	0,0469	3	12
14 ├ 18	5	0,0781	8	16
18 ├ 22	12	0,1875	20	20
22 ├ 26	25	0,3906	45	24
26 ├ 30	12	0,1875	57	28
30 ├ 34	3	0,0469	60	32
34 ├ 38	2	0,0313	62	36
38 ├ 42	2	0,0313	64	40
Total	**64**	**1,0000**		

c) $R = 32,\ s \approx 5,8,\ s^2 \approx 33,6,\ cv \approx 0,2358$ ou $23,58\%$

d) $Q_1 \approx 20,7,\ D_7 \approx 26,\ P_{99} = 40,7$

e)



5. $R = 7,\ s \approx 2,88,\ s^2 \approx 8,29,\ cv \approx 0,45$ ou 45%

6. $R = 5,\ s \approx 1,29,\ s^2 \approx 8,29,\ cv \approx 0,7588$ ou $75,88\%$

7. $R = 30,\ s \approx 6,9,\ s^2 \approx 47,6,\ cv \approx 0,2233$ ou $22,33\%$

8. Resposta
   a) Caixa A (menor variação absoluta (*s*))
   b) Caixa A (maior variação relativa (*cv*))

Capítulo 4

1. [14,04902 ; 15,95098] Margem de erro: 0,95098 $\hat{EP}(\bar{X}) = \frac{S}{\sqrt{n}} = 0,34$

2. a) $\hat{EP}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0,02487$ b) [0,40125 ; 0,49875]

3. a) $\hat{EP}(\bar{X}) = \frac{S}{\sqrt{n}} = 0,3863$ b) [4,3749 ; 6,0251]

4. [0,52110 ; 0,57890]

5. a) $\hat{p} = 0.4$ b) $\hat{EP}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0,01549$

Capítulo 5

6. a) 0,0132 b) 51.496 km

7. 0,1056

8. a) $\sigma = 6.08$ b) 5,05%

9. $z_\alpha = -2,055$ e $z = -1,5652$ (estatística de teste)
Não há evidência suficiente para garantir a rejeição da afirmativa de que as caixas modelo padrão da fábrica de embalagens têm resistência média não inferior a 14 kg.

10. $z_\alpha = -1.645$ e $z = -2.1909$ (estatística de teste)
Há evidência suficiente para garantir a rejeição da afirmativa de que a duração do efeito do analgésico fabricado por esta companhia não seja inferior a 4 horas.

Curva Normal (p = área entre 0 à 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
0.1	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.2	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.3	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.4	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.5	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879

0.6	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.7	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.8	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	2764	0.2794	0.2823	0.2852
0.9	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
1.0	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.2	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.3	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.4	0.4032	0.4049	0.4066	0.4082	0.4099	0.1415	0.4131	0.4147	0.4162	0.4177
1.5	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.6	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.7	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.2545
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	04732	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	04916
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
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Tabela – Valores críticos da distribuição *t* de *Student*

P(|T DE STUDENT| $\geq$ VALOR TABELADO) = α $\leftrightarrow$ 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

P(|T DE STUDENT| ≥ VALOR TABELADO) = α ↔ VALORES BILATERAIS									
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
13	0.694	1.350	1.771	2.160	2.282	2.650	3.012	3.372	4.221
14	0.692	1.345	1.761	2.145	2.264	2.624	2.977	3.326	4.140
15	0.691	1.341	1.753	2.131	2.249	2.602	2.947	3.286	4.073
16	0.690	1.337	1.746	2.120	2.235	2.583	2.921	3.252	4.015
17	0.689	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.965
18	0.688	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.922
19	0.688	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.883
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
27	0.684	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.689
28	0.683	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.674
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| $\geq$ VALOR TABELADO) = α $\leftrightarrow$ 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
37	0.681	1.305	1.687	2.026	2.129	2.431	2.715	2.985	3.574
38	0.681	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.566
39	0.681	1.304	1.685	2.023	2.125	2.426	2.708	2.976	3.558
40	0.681	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.551
41	0.681	1.303	1.683	2.020	2.121	2.421	2.701	2.967	3.544
42	0.680	1.302	1.682	2.018	2.120	2.418	2.698	2.963	3.538
43	0.680	1.302	1.681	2.017	2.118	2.416	2.695	2.959	3.532
44	0.680	1.301	1.680	2.015	2.116	2.414	2.692	2.956	3.526
45	0.680	1.301	1.679	2.014	2.115	2.412	2.690	2.952	3.520
46	0.680	1.300	1.679	2.013	2.114	2.410	2.687	2.949	3.515
47	0.680	1.300	1.678	2.012	2.112	2.408	2.685	2.946	3.510
48	0.680	1.299	1.677	2.011	2.111	2.407	2.682	2.943	3.505
49	0.680	1.299	1.677	2.010	2.110	2.405	2.680	2.940	3.500
50	0.679	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.496
60	0.679	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.460
70	0.678	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.435
80	0.678	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.416
90	0.677	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.402
100	0.677	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.390
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