Métodos Quantitativos Estatísticos

Paulo Ricardo Bittencourt Guimarães

1.ª edição

XXX Guimarães, Paulo Ricardo Bittencourt.

Métodos Quantitativos Estatísticos./Guimarães, Paulo Ricardo Bittencourt. — Curitiba: IESDE Brasil S.A., 2008.

245 p.

ISBN: XXX-XX-XXXX-XXX-X

1. Métodos Estatísticos 2. Probabilidade e Estatística 3. Inferência Estatística 4. Análise de Regressão 5. Análise de Dados I. Título

CDD XXX.XXXX

Inteligência Educacional e Sistemas de Ensino

IESDE Brasil S.A
Al. Dr. Carlos de Carvalho, 1 482. CEP: 80730-200
Batel – Curitiba – PR
0800 708 88 88 – www.iesde.com.br

Paulo Ricardo Bittencourt Guimarães

Doutorando em Engenharia Florestal com concentração em Economia e Política Florestal pela Universidade Federal do Paraná (UFPR). Mestre em Estatística pela Universidade Estadual de Campinas (Unicamp). Bacharel em Estatística pela Universidade Federal do Paraná (UFPR). Professor do Departamento de Estatística da Universidade Federal do Paraná (UFPR). Especialista em avaliação do Programa Nacional de Inclusão de Jovens (Projovem) da Secretaria Geral da Presidência da República. Consultor em Bioestatística e Pesquisa de Mercado.

Sumário

Sumário

Apresentação

Como se sabe, as portas do mercado de trabalho estão muito mais abertas aos profissionais que, por exemplo, tem habilidades em línguas estrangeiras. Da mesma forma, profissionais que tem uma cultura básica em Estatística estão cada vez mais valorizados, exatamente pelo seu preparo para auxiliar o processo de tomada de decisão. Mas o que significa isso? Desenvolver uma cultura estatística significa desenvolver a habilidade de planejar um estudo, controlando todos os aspectos que possam causar variações na resposta de interesse e, com base em metodologias científicas, analisar as informações coletadas para subsidiar com mais segurança a difícil tarefa de tomada de decisão.

A ciência Estatística é aplicável a qualquer ramo do conhecimento em que se manipulem dados experimentais. Assim, a Engenharia, a Economia, a Administração, a Medicina, a Biologia, as Ciências Agronômicas etc, tendem cada vez mais a servir-se dos métodos estatísticos como ferramenta de trabalho, daí sua grande e crescente importância.

O objetivo deste livro é apresentar os principais e mais freqüentes conceitos utilizados em Estatística e as técnicas básicas de análise de dados. O aluno deve estar, ao final da disciplina, apto a realizar um bom planejamento de um estudo estatístico e realizar análises estatísticas básicas dos dados resultantes desse estudo. Deve estar preparado, também, a realizar interpretações de resultados estatísticos de relatórios analíticos.

Para habilitar o estudante no uso de aplicativos de Estatística em suas análises de dados, alguns exercícios serão resolvidos fazendo uso da planilha eletrônica *Excel*.

Conceitos e Aplicações

Introdução

Geralmente, as pessoas imaginam que Estatística é uma simples coleção de números, ou tem a ver com gráficos e Censo Demográfico. Pretendemos mostrar que, na verdade, é muito mais do que isso e o seu uso surge com bastante freqüência em nossas vidas.

Estatística é um conjunto de técnicas de análise de dados, cientificamente formuladas, aplicáveis a quase todas as áreas do conhecimento que nos auxiliam no processo de tomada de decisão. É a Ciência que estuda os processos de coleta, organização, análise e interpretação de dados relevantes e referentes a uma área particular de investigação.

A origem da palavra Estatística tem a ver com uma coleção de informações populacionais e econômicas de interesse do Estado. O termo *estatística* surge da expressão em latim *statisticum collegium* palestra sobre os assuntos do Estado, da qual surgiu a palavra em língua italiana *statista*, que significa "homem de estado", ou político, e a palavra alemã *Statistik*, designando a análise de dados sobre o Estado. A palavra foi proposta pela primeira vez no século XVII, em latim, por Schmeitzel na Universidade de Lena e adotada pelo acadêmico alemão Godofredo Achenwall. Aparece como vocabulário na Enciclopédia Britânica em 1797, e adquiriu um significado de coleta e classificação de dados, no início do século 19.

Alguns exemplos de aplicação de técnicas estatísticas são: pesquisa eleitoral, pesquisa de mercado, controle de qualidade, índices econômicos, desenvolvimento de novos medicamentos, novas técnicas cirúrgicas e de tratamento médico, sementes mais eficientes, previsões meteorológicas, previsões de comportamento do mercado de ações etc., ou seja, tudo que se diz "comprovado cientificamente", em algum momento, passa por procedimentos estatísticos.

Curiosamente, apesar de a Estatística estar enquadrada entre as "ciências exatas", seus resultados estão sempre associados a uma pequena incerteza, exatamente por estarem baseados em uma amostra. O profissional de esta-

tística deve ter a habilidade de controlar esta incerteza por meio de procedimentos de Amostragem. A incerteza é conseqüência da variabilidade de um fenômeno e dificulta a tomada de decisões.

Considere um simples exemplo da vida cotidiana: a ida de uma pessoa a uma agência bancária. Em torno desse fenômeno há uma série de incertezas, por exemplo: a quantidade de pessoas na fila, o número de atendentes, o tempo de atendimento, as condições do tempo, a cotação da moeda etc.

Mesmo que um indivíduo procure informações prévias sobre todos esses elementos, sob os quais paira a incerteza, ainda assim não será possível predizer o desfecho. Podemos, por exemplo, analisar as condições do tempo, obter informações sobre o tráfego, ligar para a agência bancária e, ainda assim, não conseguiremos precisar o horário em que se receberá o desejado atendimento bancário.

Conceitos básicos

Em seguida são apresentados os principais conceitos estatísticos, os quais são diversas vezes citados ao longo do livro. É importante, nesse momento, o leitor se familiarizar com esses novos termos, o que facilita a compreensão das técnicas estatísticas apresentadas na seqüência.

Estatística Descritiva

O objetivo da Estatística Descritiva é resumir as principais características de um conjunto de dados por meio de tabelas, gráficos e resumos numéricos. Descrever os dados pode ser comparado ao ato de tirar uma fotografia da realidade. Caso a câmera fotográfica não seja adequada ou esteja sem foco, o resultado pode sair distorcido. Portanto, a análise estatística deve ser extremamente cuidadosa ao escolher a forma adequada de resumir os dados.

Inferência Estatística

Usualmente, é impraticável observar toda uma população, seja pelo custo alto, seja por dificuldades operacionais. Examina-se então uma amostra, de preferência bastante representativa, para que os resultados obtidos

possam ser generalizados para toda a população. Toda conclusão tirada por amostragem, quando generalizada para a população, apresenta um grau de incerteza. Ao conjunto de técnicas e procedimentos que permitem dar ao pesquisador um grau de confiabilidade nas afirmações que faz para a população, baseadas nos resultados das amostras, damos o nome de *Inferência Estatística*.

Dessa forma, poderíamos resumir os passos necessários para se atingir bons resultados ao realizar um experimento:

- Planejar o processo amostral e experimental.
- Obter inferências sobre a população.
- Estabelecer níveis de incerteza envolvidos nessas inferências.

População

É a totalidade de elementos que estão sob discussão e das quais se deseja informação, se deseja investigar uma ou mais características. A população pode ser formada por pessoas, domicílios, peças de produção, cobaias, ou qualquer outro elemento a ser investigado.

Para que haja uma clara definição das unidades que formam a população, é necessária a especificação de três elementos: uma característica em comum, localização temporal e localização geográfica.

Exemplos:

- Estudo da inadimplência dos clientes do banco X no Brasil

Característica comum	Clientes do banco X
Tempo	Cadastro atualizado em agosto de 2007
Localização geográfica	Agências de todo o Brasil

- Estudo de salários dos profissionais da área de seguros no estado de São Paulo

Característica comum	Profissionais da área de seguros
Tempo	Salários pagos em julho de 2007
Localização geográfica	Seguradoras de todo o estado de São Paulo

Amostra aleatória

Quando queremos obter informações a respeito de uma população, observamos alguns elementos, os quais são obtidos de forma aleatória o que chamaremos de *amostra aleatória*.

Uma amostra é uma parcela da população utilizada para uma posterior análise de dados. Em vez de utilizar toda a população, que resulta em maior custo, tempo e por muitas vezes ser inviável, o processo de amostragem utiliza uma pequena porção representativa da população. A amostra fornece informações que podem ser utilizadas para estimar características de toda a população.

É preciso garantir que a amostra ou as amostras usadas sejam obtidas por processos adequados. Se erros forem cometidos no momento de selecionar os elementos da amostra, o trabalho todo fica comprometido e os resultados finais serão provavelmente bastante viesados. Devemos, portanto, tomar especial cuidado quanto aos critérios que usados na seleção da amostra.

O que é necessário garantir, em suma, é que a amostra seja representativa da população. Isso significa que, com exceção de pequenas discrepâncias inerentes à aleatoriedade sempre presente, em maior ou menor grau, no processo de amostragem, a amostra deve possuir as mesmas características básicas da população, no que diz respeito à(s) variável(is) que desejamos pesquisar.

Os problemas de amostragem podem ser mais ou menos complexos, dependendo das populações e das variáveis que se deseja estudar. Na indústria, para efeito de controle de qualidade, as amostras são freqüentemente retiradas dos produtos e materiais. Nela os problemas de amostragem são mais simples de resolver. Por outro lado, em pesquisas sociais, econômicas ou de opinião, a complexidade dos problemas de amostragem é normalmente bastante grande. Em tais casos, deve-se ter extremo cuidado quanto à caracterização da população e ao processo usado para selecionar a amostra, a fim de evitar que os elementos constituam um conjunto com características fundamentalmente distintas das da população.

Em resumo, a obtenção de soluções adequadas para o problema de amostragem exige, em geral, muito bom senso e experiência. Além disso, é muitas vezes conveniente que o trabalho de elaboração do plano de amostragem seja baseado em informações de um especialista do assunto em questão.

Cuidado especial deve ser tomado nas conclusões em situações em que a amostra coletada não seja extraída exatamente da população de interesse (população alvo) e sim de uma população mais acessível, conveniente, nesse caso chamada de *população amostrada*.

Veja os exemplos:

1) Suponha que um sociólogo deseja entender os hábitos religiosos dos homens com 20 anos de idade em certo país. Ele extrai uma amostra de homens com 20 anos de uma grande cidade para estudar. Neste caso, tem-se:

   - População alvo – homens com 20 anos do país;
   - População amostrada – homens com 20 anos da cidade grande amostrada.

Então, ele pode fazer conclusões válidas apenas para os elementos da grande cidade (população amostrada), mas pode usar o seu julgamento pessoal para extrapolar os resultados obtidos para a população alvo, com muita cautela e certas reservas.

2) Um pesquisador agrícola está estudando a produção de certa variedade de trigo em determinado estado. Ele tem a sua disposição 5 fazendas espalhadas pelo estado, nas quais ele pode plantar trigo e observar a produção. A população amostrada, neste caso, consiste das produções de trigo nas 5 fazendas, enquanto a população alvo consiste das produções de trigo em todas as fazendas do estado.

Técnicas de Amostragem

Existem dois tipos de amostragem: *probabilística* e *não-probabilística*.

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. Uma amostragem não-probabilística é obtida quando o acesso a informações não é tão simples ou os recursos forem limitados, assim o pesquisador faz uso de dados que estão mais a seu alcance, é a chamada amostragem por conveniência.

Por exemplo, podemos realizar um estudo para avaliar a qualidade do serviço prestado por uma operadora de telefonia celular. Caso tenhamos re-

cursos suficientes, podemos realizar um plano amostral bastante abrangente de toda a população de usuários do serviço. Isso caracteriza uma amostra probabilística. Mas se por restrições orçamentárias ou de outra ordem não for possível obter uma amostra tão numerosa ou ela seja de difícil acesso, podemos restringir nossa amostra a uma pequena região delimitada de fácil acesso e de custo reduzido, usuários de uma cidade, por exemplo. Essa é uma amostragem não-probabilística.

Segundo essa definição, a amostragem probabilística implica sorteio com regras bem determinadas, cuja realização só será possível se a população for finita e totalmente acessível.

A utilização de uma amostragem probabilística é a melhor recomendação que se deve fazer no sentido de garantir a representatividade da amostra, pois o acaso é o único responsável por eventuais discrepâncias entre população e amostra. No caso em que a única possibilidade é o uso de uma amostragem não-probabilística, deve-se ter a consciência de que as conclusões apresentam alguma limitação.

A seguir, apresentamos algumas das principais técnicas de amostragem probabilística.

Amostragem aleatória simples

Esse tipo de amostragem, também chamada *simples ao acaso*, *casual*, *elementar*, *randômica* etc., é equivalente a um sorteio lotérico. Nela, todos os elementos da população têm igual probabilidade de pertencer à amostra e todas as possíveis amostras têm igual probabilidade de ocorrer.

Sendo **N** o número de elementos da população e **n** o número de elementos da amostra, cada elemento da população tem probabilidade $\mathbf{n/N}$ de pertencer à amostra. A essa relação $\mathbf{n/N}$ denomina-se *fração de amostragem*. Por outro lado, sendo a amostragem feita sem reposição, supomos, em geral, que existem $\binom{\mathbf{N}}{\mathbf{n}}$ possíveis amostras, todas igualmente prováveis.

Na prática, a amostragem simples ao acaso pode ser realizada numerando-se a população de 1 a **N**, sorteando-se, a seguir, por meio de um dispositivo aleatório qualquer, **n** números dessa seqüência, os quais correspondem aos elementos sorteados para a amostra.

Amostragem sistemática

Quando os elementos da população se apresentam ordenados e a retirada dos elementos da amostra é feita periodicamente, temos uma *amostragem sistemática*.

Assim, por exemplo, em uma linha de produção, podemos, a cada dez itens produzidos, retirar um para pertencer a uma amostra da produção diária. Assim, teremos uma produção total de **N** itens e extraíremos uma amostra de tamanho **n**, selecionando as unidades a cada dez itens. Para seleção do primeiro item, um número entre 1 e 10 é sorteado aleatoriamente e os demais subseqüentes são obtidos sistematicamente. Por exemplo, as unidades sorteadas poderão ser 8, 18, 28, 38, 48, e assim por diante, repetindo-se o procedimento até o **N-ésimo** item. Denomina-se **k = N/n** como a razão de amostragem. No exemplo, portanto, k = 10.

A principal vantagem da amostragem sistemática está na grande facilidade na determinação dos elementos da amostra. O perigo em adotá-la está na possibilidade da existência de ciclos de variação da variável de interesse, especialmente se o período desses ciclos coincidir com o período de retirada dos elementos da amostra. Por outro lado, se a ordem dos elementos na população não tiver qualquer relacionamento com a variável de interesse, então a amostragem sistemática tem efeitos equivalentes à amostragem casual simples, podendo ser utilizada sem restrições.

Amostragem estratificada

Muitas vezes, a população se divide em subpopulações ou estratos, sendo razoável supor que, de estrato para estrato, a variável de interesse apresente um comportamento substancialmente diverso, tendo, entretanto, comportamento razoavelmente homogêneo dentro de cada estrato. Em tais casos, se o sorteio dos elementos da amostra for realizado sem se levar em consideração a existência dos estratos, pode acontecer que os diversos estratos não sejam convenientemente representados na amostra, a qual seria mais influenciada pelas características da variável nos estratos mais favorecidos pelo sorteio. Evidentemente, a tendência à ocorrência de tal fato será tanto maior quanto menor o tamanho da amostra. Para evitar isso, pode-se adotar uma *amostragem estratificada*.

Constituem exemplos em que uma amostragem estratificada parece ser recomendável, a estratificação de uma cidade em bairros, quando se deseja investigar alguma variável relacionada à renda familiar; a estratificação de uma população humana em homens e mulheres, ou por faixas etárias; a estratificação de uma população de estudantes conforme suas especificações etc.

Amostragem por conglomerados

Neste método, em vez da seleção de unidades da população, são selecionados conglomerados dessas unidades. Essa é uma alternativa para quando não existe o cadastro das unidades amostrais. Se a unidade de interesse, por exemplo, for um aluno, pode ser que não exista um cadastro de alunos, mas sim de escolas. Portanto, podem ser selecionadas escolas e nelas investigar todos os alunos. Esse tipo de amostragem induz indiretamente aleatoriedade na seleção das unidades que formam a amostra e tem a grande vantagem de facilitar a coleta de dados.

Amostragem de conveniência (não-probabilística)

A *amostra de conveniência* é formada por elementos que o pesquisador reuniu simplesmente porque dispunha deles. Então, se o professor tomar os alunos de sua classe como amostra de toda a escola, está usando uma amostra de conveniência.

Os estatísticos têm muitas restrições ao uso de amostras de conveniência. Mesmo assim, as amostras de conveniência são comuns na área de saúde, em que se fazem pesquisas com pacientes de uma só clínica ou de um só hospital. Mais ainda, as amostras de conveniência constituem, muitas vezes, a única maneira de estudar determinado problema.

De qualquer forma, o pesquisador que utiliza amostras de conveniência precisa de muito senso crítico. Os dados podem ser tendenciosos. Por exemplo, para estimar a probabilidade de morte por desidratação não se deve recorrer aos dados de um hospital. Como só são internados os casos graves, é possível que a mortalidade entre pacientes internados seja maior do que entre pacientes não-internados. Conseqüentemente, a amostra de conveniência constituída, nesse exemplo, por pacientes internados no hospital, seria tendenciosa.

Finalmente, o pesquisador que trabalha com amostras sempre pretende fazer inferência, isto é, estender os resultados da amostra para toda a população. Então é muito importante caracterizar bem a amostra e estender os resultados obtidos na amostra apenas para a população da qual a amostra proveio.

Exemplos de planos amostrais:

Exemplo 1: Uma agência de seguros tem N = 100 clientes comerciantes. Seu proprietário pretende entrevistar uma amostra de 10 clientes para levantar possibilidades de melhora no atendimento. Escolha uma amostra aleatória simples de tamanho n = 10.

- Primeiro passo – atribuir a cada cliente um número entre 1 e 100.
- Segundo passo – recorrer a um gerador de números aleatórios de uma planilha eletrônica para selecionar aleatoriamente 10 números de 1 a 100. Os clientes identificados pelos números selecionados compõem a amostra.

Exemplo 2: Uma operadora de celular tem um arquivo com N = 5 000 fichas de usuários de um serviço e é selecionada, sistematicamente, uma amostra de n = 1 000 usuários. Nesse caso, a fração de amostragem é igual a n/N = 1 000/5 000 e assim podemos definir k = 5 (N/n = 5 000/1 000 = 5), ou seja, teremos 5 elementos na população para cada elemento selecionado na amostra. Na amostragem sistemática, somente o ponto de partida é sorteado dentre as 5 primeiras fichas do arquivo. Admitamos que foi sorteado o número 3, então a amostra será formada pelas fichas 3 , 8, 13 , 18, . . . , 4993 , 4998.

Tipos de variáveis

A característica de interesse de estudo (variável) pode ser dividida em duas categorias: *qualitativas* e *quantitativas*.

As *variáveis qualitativas* apresentam como possíveis realizações uma qualidade (ou atributo) do indivíduo pesquisado. Dentre as variáveis qualitativas, ainda podemos fazer uma distinção entre dois tipos: *variável qualitativa categórica* ou *nominal*, para a qual não existe nenhuma ordenação nas possíveis realizações, e *variável qualitativa ordinal*, para a qual existe certa ordem nos possíveis resultados.

Exemplo 1: (variável qualitativa nominal)

População: moradores de uma cidade.

Variável: cor dos olhos (pretos, castanhos, azuis e verdes).

Exemplo 2: (variável qualitativa ordinal)

População: moradores de um condomínio.

Variável: grau de instrução (fundamental, médio e superior).

As *variáveis quantitativas* apresentam, como possíveis realizações, números resultantes de uma contagem ou mensuração. Dentre as variáveis quantitativas, ainda podemos fazer uma distinção entre dois tipos: *variáveis quantitativas discretas*, cujos possíveis valores formam um conjunto finito ou enumerável de números e que resultam, freqüentemente, de uma contagem; e *variáveis quantitativas contínuas*, cujos possíveis valores formam um intervalo de números reais e que resultam, normalmente, de uma mensuração.

Exemplo 3: (variável quantitativa discreta)

População: hospitais de uma determinada cidade.

Variável: número de leitos (0, 1, 2, ...).

Exemplo 4: (variável quantitativa contínua)

População: moradores de uma determinada cidade.

Variável: estatura dos indivíduos.

Ampliando seus conhecimentos

(MATTAR, 2001)

Pesquisa de mercado

Em qualquer pesquisa, principalmente naquelas em que o número investigado é muito grande, torna-se quase impossível ou inviável pesquisar todos

os elementos da população. É necessário retirar uma amostra representativa para ser analisada.

A amostra em pesquisa de mercado é um fator básico para validar ou não um procedimento adotado. Vale dizer que esse item é bastante complexo porque, dependendo do universo a ser analisado e dos objetivos do estudo, teremos que usar um critério amostral.

Uma vez definida a população a ser investigada, precisamos fazer a seleção do método de escolha da amostra e definição do tamanho da amostra. Esse método vai depender do conhecimento da delimitação do universo a ser pesquisado, de suas características e ordenamento, pois nem toda amostra permite que os resultados sejam inferidos para o universo como um todo.

Etapas de uma pesquisa

Abaixo é apresentado um esquema contendo as etapas para realização de uma pesquisa.

Etapas	Fases
1. Reconhecimento e formulação do problema de pesquisa	Formulação, determinação ou constatação de um problema de pesquisa
2. Planejamento da pesquisa	a) Definição dos objetivos
	b) Estabelecimento das questões de pesquisa.
	c) Estabelecimento das necessidades de dados e definição das variáveis e de seus indicadores
	d) Determinação das fontes de dados
	e) Determinação da metodologia
	f) Planejamento da organização, cronograma e orçamento
	g) Redação do projeto de pesquisa e/ou de proposta de pesquisa
3. Execução da pesquisa	a) Preparação de campo
	b) Campo
	c) Processamento e análise
4. Comunicação dos resultados	a) Elaboração e entrega dos relatórios de pesquisa
	b) Preparação e apresentação oral dos resultados

Reconhecimento e formulação do problema de pesquisa: consiste na correta identificação do problema de pesquisa que se pretenda resolver e que possa efetivamente receber contribuições valiosas da pesquisa de *marketing* em sua solução.

Planejamento da pesquisa: compreende a definição dos objetivos da pesquisa e de toda sua operacionalização. Fontes de dados, método de pesquisa, forma de coleta, construção e teste do instrumento de coleta, plano amostral, procedimentos de campo, plano de processamento e análise, definição dos recursos necessários, definição de cronograma das etapas.

Execução da pesquisa: coleta de dados e processamento, análise e interpretação.

Comunicação dos resultados: compreende a apresentação escrita e oral das principais descobertas da pesquisa, com sugestões e recomendações.

Atividades de aplicação

Abaixo seguem alguns exemplos de aplicação da estatística. Em cada um deles são definidas algumas estratégias. Verifique se cada uma das estratégias é adequada para se atingir maior confiabilidade nos resultados atingidos. Em seguida, justifique sua resposta, apontando os motivos que levarão ou não a uma confiabilidade nos resultados.

1. Uma firma que está se preparando para lançar um novo produto precisa conhecer as preferências dos consumidores no mercado de interesse. Para isso, o que se deve fazer:

a) Uma pesquisa de mercado realizando entrevistas a domicílio com uma amostra de pessoas escolhidas aleatoriamente que se adaptam ao perfil da população de interesse.

b) Realizar entrevistas com todos os potenciais consumidores do referido produto nos estabelecimentos comerciais em que este será vendido.

c) Promover uma discussão em grupo sobre o novo produto, moderada por um especialista, com cerca de 20 donas de casa em que será feita uma degustação e posteriormente uma avaliação.

2. Antes de lançar um novo remédio no mercado, é necessário fazer várias experiências para garantir que o produto é seguro e eficiente. Para isso, o que se deve fazer:

a) Tomar dois grupos de pacientes tão semelhantes quanto possível, e dar o remédio a um grupo, mas não ao outro, e verificar se os resultados no grupo tratado são melhores.

b) Deve-se realizar um período de testes do novo medicamento, disponibilizando algumas amostras grátis em farmácias para serem avaliadas pela população durante certo período de tempo.

c) Tomar um grupo de pacientes de determinado hospital e sem que sejam informados, administrar a nova droga, comparando-se os resultados obtidos com os resultados anteriores, obtidos com a droga antiga.

3. Se estamos recebendo um grande lote de mercadorias de um fornecedor, teremos de certificar-nos de que o produto realmente satisfaz os requisitos de qualidade acordados. Para isso devemos:

a) Fazer avaliações da qualidade de todo o lote mediante inspeção de alguns itens escolhidos aleatoriamente, em quantidade que seja representativa da população.

b) Liberar uma parte do lote para comércio. Caso exista algum problema constatado pelos consumidores, deve-se devolver o lote inteiro ao fornecedor.

c) Avaliar a qualidade de aproximadamente 10% dos itens do lote. Caso não sejam encontrados itens defeituosos, liberar o lote todo ao comércio.

Análise Exploratória de Dados

Introdução

As técnicas estatísticas clássicas foram concebidas para serem as melhores possíveis, desde que se assuma um conjunto de pressupostos rígidos. Sabe-se que essas técnicas se comportam deficientemente à medida que este conjunto de pressupostos não é satisfeito.

As técnicas de Análise Exploratória de Dados contribuem para aumentar a eficácia da análise estatística, de forma fácil e rápida. Geralmente, devem ser aplicadas antes da formulação das hipóteses estatísticas para identificar padrões e características dos dados.

Uma *amostra* é um subconjunto de uma população, necessariamente finito, pois todos os seus elementos são examinados para efeito da realização do estudo estatístico desejado.

É intuitivo que, quanto maior a amostra, mais precisas e confiáveis devem ser as induções realizadas sobre a população. Levando esse raciocínio ao extremo, concluiríamos que os resultados mais perfeitos seriam obtidos pelo exame completo de toda a população, ao qual costuma-se denominar *Censo* ou *Recenseamento*. Mas essa conclusão, na prática, muitas vezes não se verifica. O emprego de amostras pode ser feito de tal modo que se obtenham resultados confiáveis.

Ocorre, em realidade, que diversas razões levam, em geral, à necessidade de recorrer-se apenas aos elementos de uma amostra. Entre ela, podemos citar o custo do levantamento de dados e o tempo necessário para realizá-lo, especialmente se a população for muito grande.

O objetivo da *Estatística Descritiva* é resumir as principais características de um conjunto de dados por meio de tabelas, gráficos e resumos numéricos. A análise estatística deve ser extremamente cuidadosa ao escolher a forma adequada de resumir os dados. Apresentamos na tabela a seguir um resumo dos procedimentos da Estatística Descritiva.

Tabela 1: Principais técnicas de estatística descritiva

Tabelas de Freqüência	Apropriada para resumir um grande conjunto de dados, agrupando informações em categorias. As classes que compõem a tabela podem ser categorias pontuais ou por intervalos.
Gráficos	Possibilita uma visualização das principais características da amostra. Alguns exemplos de gráficos são: diagrama de barras, diagrama em setores, histograma, box-plot, ramo-e-folhas, diagrama de dispersão.
Medidas Descritivas	Por meio de medidas ou resumos numéricos podemos levantar importantes informações sobre o conjunto de dados, tais como: a tendência central, variabilidade, simetria, valores extremos, valores discrepantes, etc.

Um dos objetivos da Estatística é sintetizar os valores que uma ou mais variáveis podem assumir, para que tenhamos uma visão global da variação dessa ou dessas variáveis. Isso se consegue, inicialmente, apresentando esses valores em tabelas e gráficos, que fornecem rápidas e seguras informações a respeito das variáveis.

Tabelas

Uma tabela resume os dados por meio do uso de linhas e colunas, nas quais são inseridos os números. Uma tabela compõe-se de:

- **Corpo** – conjunto de linhas e colunas que contém informações sobre a variável em estudo.
- **Cabeçalho** – parte superior da tabela que especifica o conteúdo das colunas.
- **Coluna Indicadora** – parte da tabela que especifica o conteúdo das linhas.
- **Linhas** – retas imaginárias que facilitam a leitura, no sentido horizontal, de dados que se inscrevem nos seus cruzamentos com as colunas.
- **Casas ou Células** – espaço destinado a um só número.
- **Título** – conjunto de informações (as mais completas possíveis) localizado no topo da tabela.

Existem ainda, elementos complementares que são: a *fonte*, as *notas* e as *chamadas*, os quais devem ser colocados no rodapé da tabela.

As *notas* devem esclarecer aspectos relevantes do levantamento dos dados ou da apuração. As *chamadas* dão esclarecimentos sobre os dados. Devem ser feitas de algarismos arábicos escritos entre parênteses, e colocados à direita da coluna.

Exemplo:

Tabela 2: População brasileira residente, com 15 anos e mais, segundo o estado conjugal, de acordo com o censo demográfico de 1980.

Estado conjugal	Freqüência	Percentual
solteiros[1]	25 146 484	34,18
casados[2]	41 974 865	57,06
separados	1 816 046	2,47
viúvos	3 616 046	4,92
sem declaração	1 005 234	1,37

Fonte: IBGE, 1988.

Estão computados, como separados, os desquitados e os divorciados.

[1] Exclusive as pessoas solteiras, vivendo em união consensual estável.

[2] Inclusive 4 939 528 pessoas vivendo em união consensual estável.

Observação:

Nas casas ou células devemos colocar:

- um traço horizontal (__) quando o valor é zero, não só quanto a natureza das coisas, como quanto ao resultado do inquérito;
- três pontos (...) quando não temos dados;
- ponto de interrogação (?) quando temos dúvida quanto a exatidão de um valor;
- zero (0) quando o valor é muito pequeno para ser expresso pela unidade utilizada.

Tabelas de contingência

Muitas vezes, os elementos da amostra ou da população são classificados de acordo com dois fatores. Os dados devem ser apresentados em *tabelas de contingência*, isto é, em tabelas de dupla entrada, cada entrada relativa a um dos fatores.

Vejamos um exemplo de uma tabela que apresenta o número de nascidos vivos registrados. Note que eles estão classificados segundo dois fatores: o ano do registro e o sexo.

Tabela 3: Nascidos vivos registrados segundo o ano de registro e o sexo

Ano de registro	Sexo		Total
	Masculino	Feminino	
1984	1 307 758	1 251 280	2 559 038
1985	1 339 059	1 280 545	2 619 604
1986	1 418 050	1 361 203	2 779 253

Fonte: IBGE, 1988.

Tabelas de distribuição de freqüências

As tabelas com grande número de dados são cansativas e não dão ao pesquisador visão rápida e global do fenômeno. Para isso, é preciso que os dados estejam organizados em uma *tabela de distribuição de freqüências*. As distribuições de freqüências são representações nas quais os valores da variável se apresentam em correspondência com suas repetições, evitando assim, que eles apareçam mais de uma vez na tabela, poupando, deste modo, espaço, tempo e, muitas vezes, dinheiro.

Como exemplo, considere os dados da tabela abaixo:

Tabela 4: Rendimento mensal de fundos de investimento

2,522	3,200	1,900	4,100	4,600	3,400
2,720	3,720	3,600	2,400	1,720	3,400
3,125	2,800	3,200	2,700	2,750	1,570
2,250	2,900	3,300	2,450	4,200	3,800
3,220	2,950	2,900	3,400	2,100	2,700
3,000	2,480	2,500	2,400	4,450	2,900
3,725	3,800	3,600	3,120	2,900	3,700
2,890	2,500	2,500	3,400	2,920	2,120
3,110	3,550	2,300	3,200	2,720	3,150
3,520	3,000	2,950	2,700	2,900	2,400
3,100	4,100	3,000	3,150	2,000	3,450
3,200	3,200	3,750	2,800	2,720	3,120
2,780	3,450	3,150	2,700	2,480	2,120
3,155	3,100	3,200	3,300	3,900	2,450
2,150	3,150	2,500	3,200	2,500	2,700
3,300	2,800	2,900	3,200	2,480	-
3,250	2,900	3,200	2,800	2,450	-

A partir desses dados desorganizados, chamados de *dados brutos* (dados tal como foram coletados, sem nenhum tipo de organização), é difícil chegar a alguma conclusão a respeito da variável em estudo (rendimento mensal de fundos de investimento). Obteríamos alguma informação a mais se arranjássemos os dados segundo uma certa organização como na sua ordem de magnitude, ou seja, se arrumássemos os dados na forma de um *rol* (lista em que os valores são dispostos em uma determinada ordem, crescente ou decrescente). Mas isso somente indicaria a amplitude de variação dos dados (isto é, o menor e o maior valor observado) e a ordem que os itens individuais ocupariam na ordenação.

Para se ter uma idéia geral sobre o rendimento mensal dos fundos de investimento, o pesquisador não apresenta os rendimentos observados, mas o número de observações por faixas de rendimento. O procedimento mais satisfatório é arranjar os dados em uma *distribuição de freqüências*, de modo a mostrar a freqüência com que ocorrem certas faixas de rendimento especificados.

O primeiro passo é definir o número de faixas de rendimento que recebem, tecnicamente, o nome de *classes*. Embora existam fórmulas apropriadas para esse fim, em geral, não se conhecem regras precisas que levem a uma decisão final, a qual depende, em parte, de um julgamento pessoal. Se o número de classes for muito pequeno, é comum acontecer que características importantes da variável fiquem ocultas. Por outro lado, um número elevado de classes fornece maior número de detalhes, mas resume de forma menos precisa os dados. Em geral, convém estabelecer de 5 a 20 classes. Uma das fórmulas usadas é a seguinte:

$$k = 1 + 3{,}3.\log(n),$$

em que n é o número total de dados. O número de classes é um inteiro próximo de k.

É importante deixar claro, aqui, que o resultado obtido por essa fórmula pode ser usado como referência, mas cabe ao pesquisador determinar o número de classes que pretende organizar.

Para entender como se aplica a fórmula, considere os dados da tabela de dados anterior. Como $n = 100$, tem-se que

$$k = 1 + 3{,}3.\log(100) \rightarrow k = 1+3{,}3.2 \rightarrow k = 7{,}6$$

ou seja, para aqueles dados, deve-se construir 7 ou 8 classes.

Definido o número de classes a ser utilizado, deve-se determinar o *intervalo de classe* (h_i), ou seja, a amplitude de cada classe. Um caminho para isso é dado por:

$$h_i = \frac{AT}{k},$$

em que *AT* é a amplitude total dos dados, isto é, a diferença entre o maior e o menor valor observado.

É importante deixar claro que o resultado obtido por essa fórmula será usado como referência, mas cabe ao pesquisador determinar o intervalo de classe exato.

Nos dados da tabela anterior, pode-se observar que o menor valor é 1,570 e o maior é 4,600, tem-se assim, $AT = 3{,}03$. Considerando $k = 7$, tem-se que $h_i = 0{,}43$. Dessa forma, podem então ser definidas classes de 1,5 a 2,0, de 2,0 a 2,5, e assim por diante. Logo, cada classe cobre um intervalo de 0,5, ou seja, cada intervalo de classe é de 0,5. É mais fácil trabalhar com intervalos de classe iguais.

A distribuição de freqüências para os dados da tabela apresenta-se dessa forma:

classe	freqüência
1,5 |— 2,0	3
2,0 |— 2,5	16
2,5 |— 3,0	31
3,0 |— 3,5	34
3,5 |— 4,0	11
4,0 |— 4,5	4
4,5 |— 5,0	1

Denomina-se *limites de classe* os extremos dos intervalos de cada classe. O menor número é o *limite inferior* (l_i) e o maior é o *limite superior* (L_i).

Em uma distribuição de freqüência também podem ser apresentados os *pontos médios de classe* (Pm_i). O ponto médio é dado pela soma dos limites de classe, dividida por 2. Desse modo, uma tabela típica de distribuição de freqüências tem três colunas, dadas por:

Classe (i)	Ponto Médio (Pm_i)	Freqüência (f_i)	Freqüência relativa (fr_i)	Freqüência acumulada (F_i)
1,5 |— 2,0	1,75	3	0,03	3
2,0 |— 2,5	2,25	16	0,16	19
2,5 |— 3,0	2,75	31	0,31	50

Classe (i)	Ponto Médio (Pm_i)	Freqüência (f_i)	Freqüência relativa (fr_i)	Freqüência acumulada (F_i)
3,0 |— 3,5	3,25	34	0,34	84
3,5 |— 4,0	3,75	11	0,11	95
4,0 |— 4,5	4,25	4	0,04	99
4,5 |— 5,0	4,75	1	0,01	100

As tabelas de distribuição de freqüências mostram a distribuição da variável, mas perdem em exatidão. Isso porque todos os dados passam a ser representados pelo ponto médio da classe a que pertencem. Por exemplo, a tabela acima mostra que 16 fundos de investimento apresentam rendimento com ponto médio igual a 2,25, mas não dá informação exata sobre o rendimento de cada um deles.

Em uma tabela de distribuição de freqüências, pode-se ter, ainda, outros dois tipos de freqüências: *freqüência relativa* e *freqüência acumulada*. A freqüência relativa é obtida dividindo-se a freqüência simples pelo número total de observações e a freqüência acumulada é obtida somando-se as freqüências simples das classes anteriores.

Gráficos

A representação gráfica dos dados tem por finalidade representar os resultados obtidos, permitindo chegar-se a conclusões sobre a evolução do fenômeno ou sobre como se relacionam seus valores. A escolha do gráfico mais apropriado fica a critério do analista. Contudo, os elementos simplicidade, clareza e veracidade devem ser considerados quando da elaboração de um gráfico.

Os principais tipos de gráficos usados na representação estatística são:

- **Histograma e gráfico de barras** – apresentam os resultados por meio do desenho de diversas barras, em que cada categoria da variável em estudo é associada à uma barra e o comprimento da barra diz respeito ao resultado indicado para a categoria. Pode ser usada também em representações envolvendo diversas variáveis, acompanhadas em diversos momentos de tempo.

- **Gráficos de linha** – útil quando se deseja representar a evolução de diversas variáveis ao longo de vários momentos de tempo. É um grá-

fico de duas dimensões formado por dois eixos perpendiculares, em que o tempo é representado no eixo horizontal X e os resultados das variáveis no eixo vertical Y.

- **Gráfico em setores (*pizza*)** – composto de um círculo repartido em **n** fatias, com tamanhos proporcionais à ocorrência da variável nos resultados da pesquisa, representando um certo instante no tempo. Sugere-se que seja aplicado em variáveis com no máximo 8 categorias.

Descrição gráfica das variáveis qualitativas

No caso das variáveis qualitativas, a representação gráfica é bem simples, basta computar as freqüências ou freqüências relativas das diversas classificações existentes e elaborar a seguir um gráfico conveniente. Esse gráfico pode ser um gráfico de barras, um gráfico de setores, ou outro qualquer tipo de gráfico equivalente.

Exemplo: Este exemplo foi extraído do Anuário da Bolsa de Valores de São Paulo, edição 1970. Nessa publicação, na parte "Fundos – Decreto Lei 157", existe uma tabela que fornece a distribuição dos fundos relativos a cada região econômica do Brasil. Essa tabela é reproduzida aqui.

Tabela 5: Distribuição de fundos relativos às regiões do Brasil

Estado	Número de estabelecimentos	
	Unidades	%
São Paulo	38	28,1
Rio de Janeiro	30	22,2
Rio Grande do Sul	35	25,9
Minas Gerais	15	11,1
Demais Estados	17	12,7
Total	**135**	**100**

As duas colunas referentes ao número de estabelecimentos contêm, respectivamente, as freqüências e as freqüências relativas, dadas em porcentagem, com que os fundos existem nos estados considerados. A variável qualitativa considerada no presente exemplo é dada pelas regiões consideradas.

Esses dados podem ser representados de diversas formas, conforme podemos notar a partir das figuras a seguir:

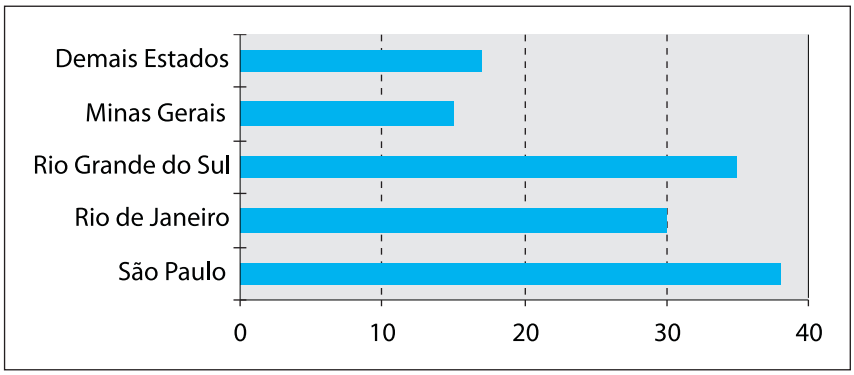

Figura 1: Gráfico de barras

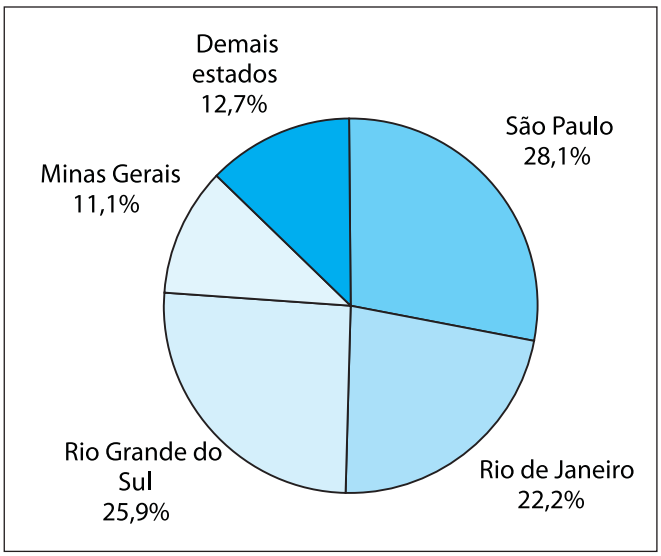

Figura 2: Gráfico de setores

Descrição gráfica das variáveis quantitativas discretas

No caso das variáveis quantitativas discretas, a representação gráfica é, normalmente, feita por meio de um gráfico de barras. A diferença para com o caso anterior está na variável quantitativa e seus valores numéricos podem ser representados num eixo de abscissas, o que facilita a representação. Note que, aqui, existe uma enumeração natural dos valores da variável, o que não havia no caso das variáveis qualitativas.

Exemplo: Vamos representar graficamente o conjunto dado a seguir, constituído hipoteticamente por vinte valores da variável "número de defeitos por unidade", obtidos a partir de aparelhos retirados de uma linha de montagem.

Sejam os seguintes valores obtidos:

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

Usando a letra x para designar os diferentes valores da variável, podemos construir a distribuição de freqüências dada a seguir, a partir da qual elaboramos o gráfico de barras correspondentes.

Distribuição de freqüências		
x_i	f_i	fr_i
0	4	0,20
1	7	0,35
2	5	0,25
3	2	0,10
4	1	0,05
5	1	0,05
	20	**1**

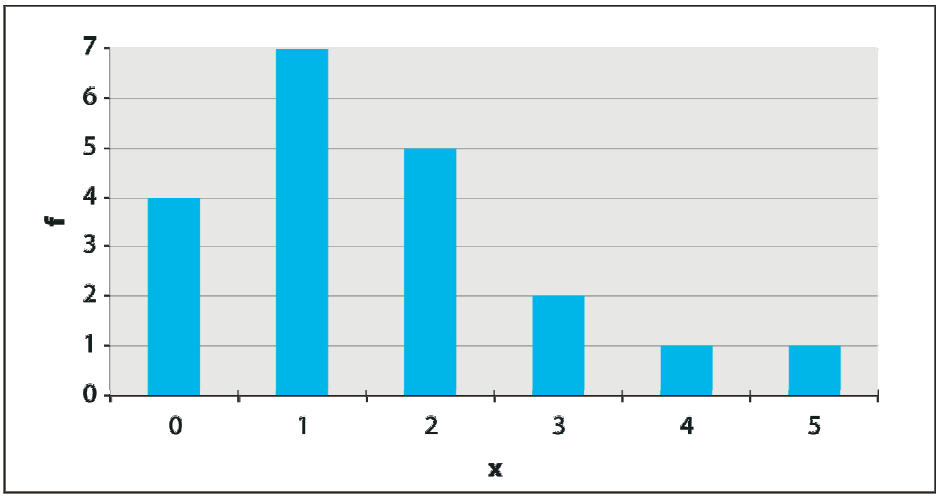

Figura 3: Gráfico de barras

Descrição gráfica das variáveis quantitativas contínuas – classes de freqüências

No caso das variáveis quantitativas contínuas, o procedimento até a obtenção da tabela de freqüências pode ser análogo ao visto no caso anterior.

Entretanto o diagrama de barras não mais se presta à correta representação da distribuição de freqüências, devido à natureza contínua da variável.

Os gráficos apropriados para representar esse tipo de variável são: *o histograma, o polígono de freqüências* e *a Ogiva de Galton*.

- **Histograma** – Para construir um histograma, primeiro se traça o sistema de eixos cartesianos. Depois, se os intervalos de classe são iguais, traçam-se barras retangulares com bases iguais, correspondentes aos intervalos de classe, e com alturas determinadas pelas respectivas freqüências.

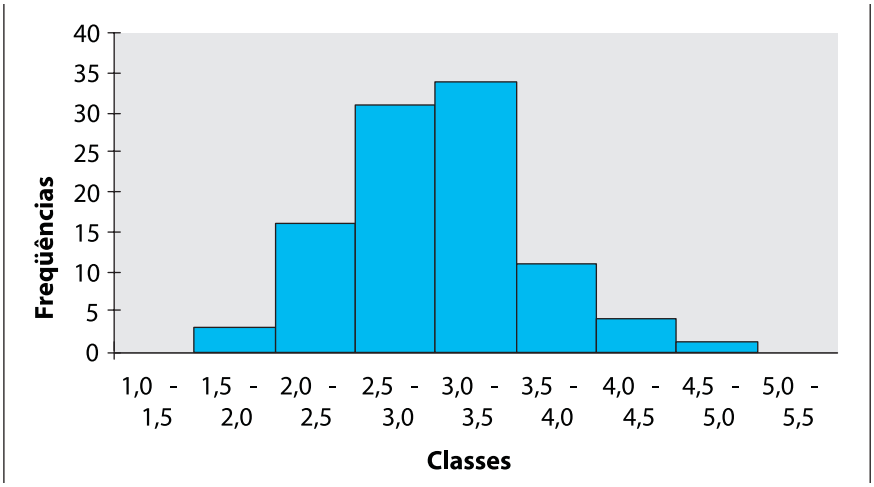

Figura 4: Histograma

- **Polígono de freqüências** – Para se construir um polígono de freqüências, primeiro se traça o sistema de eixos cartesianos. Depois, se os intervalos de classes são iguais, marcam-se pontos com abscissas iguais aos pontos médios de classe e ordenadas iguais às respectivas freqüências. Se os intervalos de classe são diferentes, marcam-se pontos com abscissas iguais aos pontos médios de classe e ordenadas iguais às respectivas densidades de freqüência relativa. Para fechar o polígono, unem-se os extremos da figura com o eixo horizontal, nos pontos de abscissas iguais aos pontos médios de uma classe imediatamente inferior à primeira, e de uma classe imediatamente superior à última.

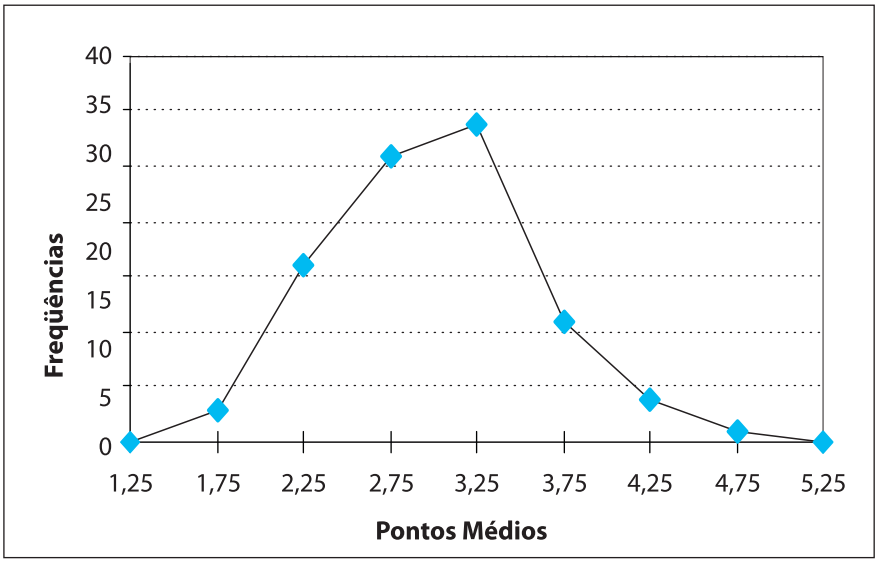

Figura 5: Polígono de freqütências

- **Ogiva de Galton** – Esse é um gráfico representativo de uma distribuição de freqüências acumuladas, seja ela crescente ou decrescente. Consta de uma poligonal ascendente. No eixo horizontal, colocam-se as extremidades de cada classe e no eixo vertical as freqüências acumuladas. Ao contrário do polígono de freqüências, a ogiva utiliza os pontos extremos das classes, e não os pontos médios.

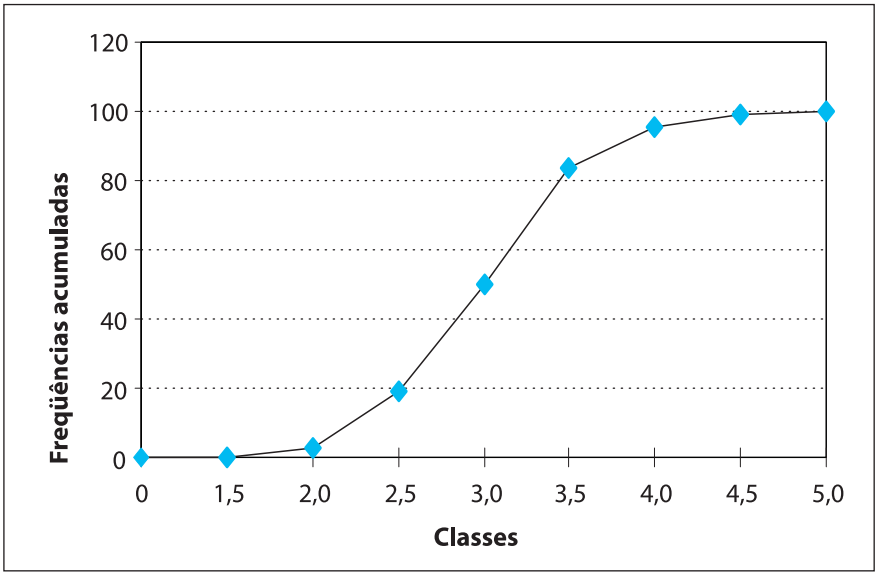

Figura 6: Ogiva de Galton Crescente

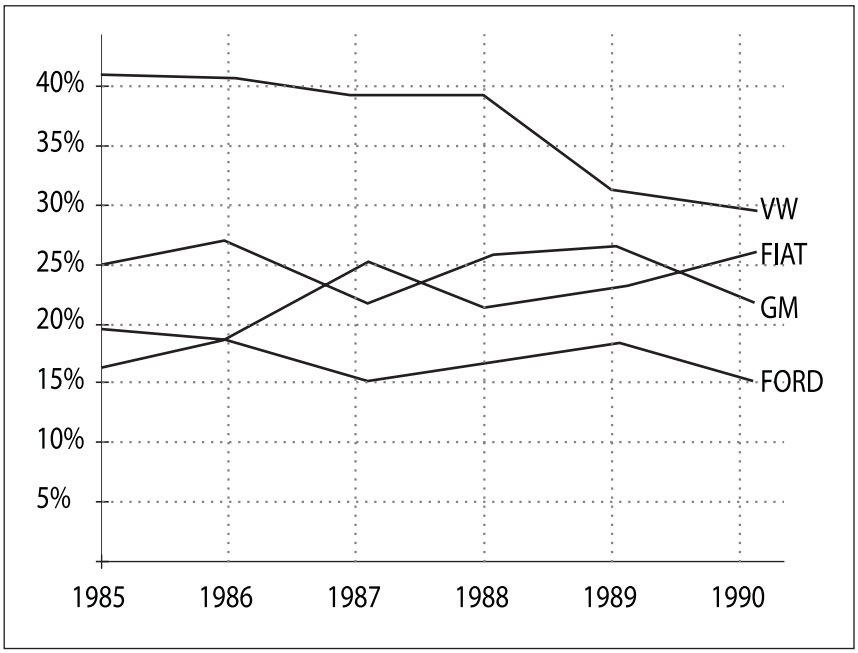

Figura 7: Gráfico de linhas

Ramo-e-folhas

Este tipo de gráfico é um modo simples de organizar os dados e que pode facilitar a construção de tabelas de freqüências. Podem ser usados para dados quantitativos (numéricos), mas não qualitativos (por exemplo, dados nominais ou por categorias).

Veja o seguinte exemplo: considere que se tenha anotado 20 valores relativos ao tempo de uma atividade, e que se deseja organizá-los em um diagrama de ramos e folhas. Os valores são os seguintes:

23 - 31 - 42 - 45 - 51 - 52 - 57 - 61 - 61 - 64 - 68 - 69 - 73 - 75 - 75 - 82 - 89 - 94 - 118 - 120

1º passo: determina-se o menor e o maior valor; neste exemplo, 23 minutos o menor valor e 120 minutos o maior.

2º passo: constroem-se categorias nas quais se deseja agrupar os dados a partir da menor dezena até a maior. Nas colunas, o 2 representa a dezena dos "20" minutos e o 12 representa a dezena dos "120 minutos".

Figura 8. Passo inicial da construção de um gráfico de ramos e folhas

Dezenas de minutos
2|
3|
4|
5|
6|
7|
8|
9|
10|
11|
12|

3º passo: retorna-se aos dados originais e simplesmente coloca-se as unidades referentes às dezenas em cada uma das linhas, ordenadamente. Por exemplo, o número 23 é representado por um 3 colocado na linha 2, e 118 pode ser representado na linha 11 por um 8. Uma vez feito para todos os valores, o diagrama fica com o aspecto da Figura 9.

Figura 9. Diagrama de ramos e folhas

Dezenas de minutos	Minutos
2|	3
3|	1
4|	2 5
5|	1 2 7
6|	1 1 4 8 9
7|	3 5 5
8|	2 9
9|	4
10|	
11|	8
12|	0

Analisando a figura acima podemos observar que o tempo de atividade mais freqüente está na faixa dos 60 minutos, apresentando-se em seguida, as faixas de 50 e 70 minutos. Se analisássemos a figura acima como se fosse um histograma poderíamos considerar que a figura apresenta certa simetria, observa-se as maiores freqüências ao redor da média.

Ampliando seus conhecimentos

(HOAGLIN. D. C.; MOSTELLER. F. & TUKEY.J. W., 1983)

Uma técnica de análise exploratória de dados: o *box-plot*

O *Box-Whisker-Plot*, mais conhecido por *Box-Plot*, é uma representação gráfica de valores, conhecidos como resumo de 5 números. Essa técnica nos revela uma boa parte da estrutura dos dados, por meio da visualização de características como:

- tendência central;
- variabilidade;
- assimetria;
- outliers (valores discrepantes).

O chamado resumo de cinco números é constituído pelo: mínimo (menor valor), primeiro quartil (Q1), a Mediana (Md), o terceiro quartil (Q3) e o máximo (maior valor).

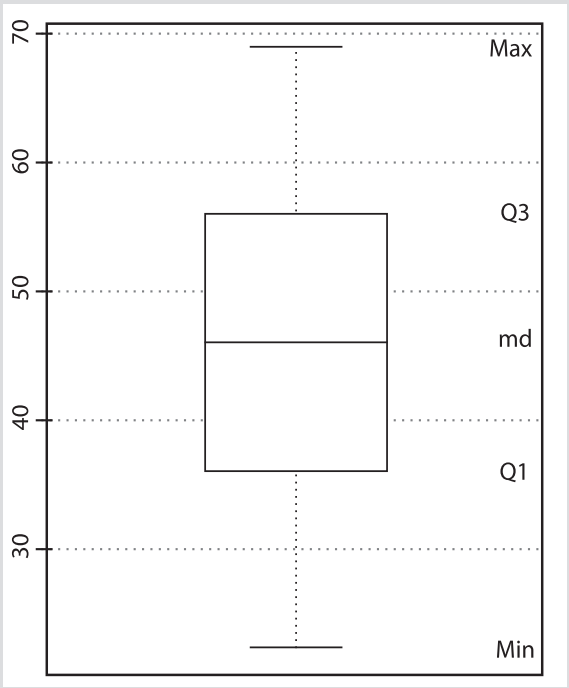

Figura 1: *Box-plot*

A parte central do gráfico é composta de uma "caixa" com o nível superior dado por Q3 e o nível inferior por Q1. O tamanho da caixa é uma medida de dispersão chamada amplitude interquartíllica (AIQ = Q3 - Q1).

A mediana, medida de tendência central, é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até os valores máximo e mínimo.

Detalharemos agora o procedimento para construção de um *Box-plot* para um conjunto de dados, por meio de um exemplo relacionado com o Censo dos EUA de 1960:

Tabela 6: Censo dos EUA (1960) – População das principais capitais

Cidade	População (1 000 hab)	Cidade	População (1 000 hab)
New York	778	Washington	76
Chicago	355	St. Louis	75
Los Angeles	248	Milwaukee	74
Filadélfia	184	San Francisco	74
Detroit	167	Boston	70
Baltimore	94	Dallas	68
Houston	94	New Orleans	63
Cleveland	88		

Para a construção do *box-plot* é necessário que sejam calculadas as medidas que compõem o resumo de 5 números:

- **A Mediana** (88) – neste exemplo, a variável em estudo tem *n* ímpar; a mediana será o valor da variável que ocupa o posto de ordem $\frac{n+1}{2}$, ou seja, o oitavo valor.
- **Os Quartis Q_1 e Q_3** (74 e 184) – devemos contar $\frac{n}{4}$ valores para se achar Q_1 e $\frac{3n}{4}$ para determinar Q_3.
- **Os valores Mínimo e o Máximo** (63 e 778)

  as barreiras de outliers[1] são obtidas por meio do cálculo:

$$Q_1 - \frac{3}{2} \cdot d_F \quad (1)^2 \quad \text{e} \quad Q_3 + \frac{3}{2} \cdot d_F \quad (2)^2$$

  em que $d_F = Q_3 - Q_1$

[1] *Outliers* são elementos ou valores que distorcem a média da distribuição pois encontram-se distantes dos demais valores da distribuição.

[2] O *outlier* mínimo é 74 – 1,5.110 = -91. O *outlier* máximo é 184 +1,5.110 = 349

Isso significa que os valores inferiores a (1) ou superiores a (2) são considerados *outliers* ou valores discrepantes. O *Box-plot* nos apresenta a localização (mediana), a dispersão (comprimento da caixa), a assimetria (pela distância dos quartis à mediana) e os *outliers* (Chicago e Nova Iorque):

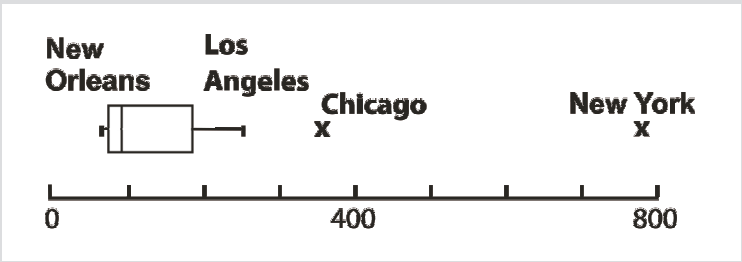

Figura 2: *Box-plot* – População das principais capitais (1960)

Observe que a barreira inferior de *outliers* é –91. Entretanto, na representação gráfica, substituiremos esse valor pelo mínimo observado (63). As expressões utilizadas para as barreiras de *outliers* são de certo modo arbitrárias, mas a experiência dos autores dessa técnica indicou que esta definição serve perfeitamente para a identificação de valores que requerem uma atenção especial.

Atividades de aplicação

Resolva as questões abaixo utilizando as definições vistas neste capítulo.

1. Uma firma de consultoria investiga as instituições financeiras que mais lucraram durante a gestão do governo atual. Do cadastro de instituições selecionou-se uma amostra aleatória de 20 para realização de uma auditoria completa. Coletou-se então o lucro de cada uma no período especificado. Os dados seguem abaixo (em US$ milhões):

58	62	55	80	74
51	60	79	50	65
68	72	54	81	65
119	82	75	86	61

Você como analista da empresa de consultoria deve elaborar um relatório sucinto, realizando uma descrição do conjunto de dados acima.

2. A tabela de dados brutos abaixo apresenta os pesos (kg) relativos de uma turma de alunos:

96	72	56	59	57	52	50	
75	85	64	68	51	66	64	
56	59	76	49	54	64	58	
80	61	74	55	72	78	78	
69	52	63	50	75	53	52	
70	53	80	67	48	90	76	
94	52	51	82	61	64	78	76

Utilizando os dados complete a tabela de distribuição de freqüência abaixo:

i	Pesos (kg)	Tabulação	f_i	Pm_i	fr_i	%
1	48 |— 53					
2	53 |— 58					
3	58 |— 63					
4	63 |— 68					
5	68 |— 73					
6	73 |— 78					
7	78 |— 83					
8	83 |— 88					
9	88 |— 93					
10	93 |— 98					
–	TOTAL					

De posse da tabela de distribuição de freqüência completa, determine:

a) O limite superior da 2ª classe.

b) O limite inferior da 5ª classe.

c) A amplitude do intervalo da 3ª classe.

d) A amplitude total.

e) O ponto médio da 4ª classe.

f) A freqüência da 1ª classe.

g) O número de alunos com peso abaixo de 68kg.

h) O número de alunos com peso igual ou acima de 73kg.

i) O número de alunos com peso maior ou igual a 58 e menor que 78.

j) A freqüência percentual da última classe.

k) A percentagem de alunos com peso inferior a 58kg.

l) A percentagem de alunos com peso superior ou igual a 78kg.

3. Faça no mesmo gráfico um esboço das três distribuições descritas abaixo:

a) Distribuição das alturas dos brasileiros adultos.

b) Distribuição das alturas dos suecos adultos.

c) Distribuição das alturas dos japoneses adultos.

4. Para estudar o desempenho de duas companhias corretoras de ações, selecionou-se de cada uma delas amostras aleatórias das ações negociadas. Para cada ação selecionada, computou-se a porcentagem de lucro apresentada durante um período fixado de tempo. Os dados estão a seguir, representados pelos diagramas de ramos-e-folhas:

Corretora A	**Corretora B**
3 | 8	5 | 0012234
4 | 588	5 | 5556677788999
5 | 44555569	6 | 1
6 | 00245	
7 | 0	

Que tipo de informação revelam esses dados ?

Medidas de Posição e Variabilidade

Introdução

Para melhor compreender o comportamento do conjunto de dados, é importante que conceituemos o que chamamos de *medidas descritivas*. Existem duas categorias de medidas descritivas:

- **Medidas de posição ou tendência central** – servem para dar uma idéia acerca dos valores médios da variável em estudo.
- **Medidas de dispersão** – servem para dar uma idéia acerca da maior ou menor concentração dos valores da variável em estudo.

Observação: Quando as medidas de tendência central e as de dispersão são calculadas sobre a população, elas são chamadas de *parâmetros*. Por outro lado, quando essas medidas são obtidas considerando-se uma amostra retirada de uma população, elas são chamadas de *estatísticas*.

Medidas de Posição ou de Tendência Central

Como o próprio nome indica, a medida de tendência central visa determinar o centro da distribuição dos dados observados. Essa determinação depende, portanto, da definição de *centro* da distribuição. Todavia, o centro de um conjunto de valores não está definido e pode ser interpretado de várias maneiras, cada uma das quais descreve uma propriedade da distribuição, que pode ser razoavelmente chamada de tendência central.

As principais medidas de tendência central são:

- média aritmética;
- mediana;
- moda.

Média Aritmética ($\overline{X}$)

Dada uma distribuição de freqüências, chama-se de média aritmética desta distribuição, e representa-se por $\overline{\mathbf{X}}$, a soma de todos os valores da variável, dividida pela freqüência total (número total de observações).

Por exemplo, considerando-se os dados da tabela abaixo, tem-se:

Tabela 1: Pacientes com hipertensão, segundo a idade em anos completos.

Idade em anos completos	Número de indivíduos (freqüência - f_i)	$x_i . f_i$	Idade em anos completos	Número de indivíduos (freqüência - f_i)	$x_i . f_i$
22	1	22	47	1	47
27	1	27	48	1	48
30	1	30	50	2	100
31	1	31	53	3	159
34	1	34	56	1	56
35	3	105	58	1	58
36	5	180	59	2	118
40	1	40	60	1	60
42	1	42	61	1	61
43	1	43	63	1	63
44	2	88	65	3	195
45	1	45	67	2	134
46	2	92			
			Total	**40**	**1 878**

$$\overline{X} = \frac{22+27+30+31+\cdots+65+65+65+67+67}{40}$$

$$\overline{X} = \frac{22.1+27.1+30.1+31.1+\cdots+65.3+67.2}{40} = \frac{1878}{40} = 46,95 \text{ anos} = 46 \text{ anos}$$

e 11 meses, ou seja, a idade média dos hipertensos é igual a 46 anos e 11 meses.

De maneira geral, ao se ter a seguinte distribuição de freqüências:

Valores x_i da variável X	Freqüência (f_i)	Produto ($x_i . f_i$)
x_1	f_1	$x_1 . f_1$
x_2	f_2	$x_2 . f_2$
.	.	.
.	.	.
.	.	.
x_k	f_k	$x_k . f_k$
Total	$\sum_{i=1}^{k} f_i$	$\sum_{i=1}^{k} x_i . f_i$

a média aritmética será:

$$\bar{X} = \frac{\sum_{i=1}^{k} x_i . f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i . f_i}{n}$$

Se os dados da tabela anterior estivessem agrupados em classes, como mostra a tabela a seguir, seria preciso, antes de calcular $\bar{X}$, determinar os pontos médios das classes.

Tabela 2. Pacientes com hipertensão, segundo a idade em anos completos.

Classes	Ponto Médio (Pm_i)	Número de pacientes (f_i)	Produto $Pm_i . f_i$
20 |— 30	25	2	50
30 |— 40	35	11	385
40 |— 50	45	10	450
50 |— 60	55	9	495
60 |— 70	65	8	520
Total		**40**	**1 900**

$$\bar{X} = \frac{1\,900}{40} = 47{,}5 \text{ anos} = 47 \text{ anos e } 6 \text{ meses ou } 47 \text{ anos (completos).}$$

De maneira geral, ao se ter uma distribuição de freqüências por classes, a média aritmética será:

$$\bar{X} = \frac{\sum_{i=1}^{k} PM_i . f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} PM_i . f_i}{n}$$

Observação: a idade média calculada a partir dos dados da tabela 2 não coincide com a idade média verdadeira dos 40 hipertensos, calculada a partir dos dados da Tabela 1. Isso se deve ao fato de ter sido suposto, para o cálculo da média aritmética com os dados da Tabela 2, que todos os indivíduos de uma determinada classe tinham a idade dada pelo ponto médio da classe, o que, em geral, não corresponde à realidade.

Da própria definição segue que a média aritmética de uma distribuição de freqüências:

- é da natureza da variável considerada;
- sempre existe, e quando calculada admite um único valor;
- não pode ser calculada quando os dados estiverem agrupados em classes e a primeira ou última classe tiverem extremos indefinidos;
- sofre muito a influência de valores aberrantes.

Mediana (Md)

A mediana é uma quantidade que, como a média, também procura caracterizar o centro da distribuição de freqüências, porém, de acordo com um critério diferente. Ela é calculada com base na ordem dos valores que formam o conjunto de dados.

A mediana é a realização que ocupa a posição central da série de observações quando estas estão ordenadas segundo suas grandezas (crescente ou decrescente).

Dada uma distribuição de freqüências e supondo-se os valores da variável dispostos em ordem crescente ou decrescente de magnitude, há dois casos a considerar:

1º A variável em estudo tem *n* ímpar. Neste caso a mediana será o valor da variável que ocupa o posto de ordem $\frac{n+1}{2}$.

Exemplo: Admita-se que o número de demissões em certa empresa nos meses de janeiro dos últimos 7 anos, ordenando, fosse:

24, 37, 41, 52, 65, 68 e 82.

A mediana neste caso vale: *Md* = 52 demissões, valor que ocupa o posto $\frac{7+1}{2} = 4º$.

2º A variável tem n par. Neste caso, não existe na graduatória um valor que ocupe o seu centro, isto é, a mediana é indeterminada, pois qualquer valor compreendido entre os valores que ocupam os postos $\frac{n}{2}$ e $\frac{n+2}{2}$ pode ser considerado o centro da graduatória.

O problema é resolvido por uma convenção que consiste em tomar como mediana da graduatória a média aritmética dos valores que ocupam os postos $\frac{n}{2}$ e $\frac{n+2}{2}$.

Exemplo: Considerando o número de demissões de certa empresa nos meses de janeiro dos 6 últimos anos e ordenando-se os valores, tem-se:

24, 37, 41, 65, 68 e 82

A mediana será, por convenção:

$$\frac{41+65}{2} = 53 \text{ demissões,}$$

ou seja, a média aritmética dos valores que ocupam os postos $\frac{6}{2} = 3º$ e $\frac{6+2}{2} = 4º$.

A mediana tem interpretação muito simples quando as observações são diferentes umas das outras, pois ela é tal que o número de observações com valores maiores a ela é igual ao número de observações com valores menores do que ela. Todavia, quando há valores repetidos, a sua interpretação não é tão simples. Assim, admitindo, como resultado da aplicação de um teste a um conjunto de alunos, as seguintes notas:

2, 2, 5, 5, 5, 5, 7, 7, 8, 8,

a mediana seria a nota 5 e, no entanto, só existem 2 notas menores e 4 maiores do que 5. Essa desvantagem, unida ao fato da inadequacidade da sua expressão para o manejo matemático, faz com que, em análises estatísticas, a mediana seja menos utilizada do que a média aritmética. No entanto, existem casos nos quais o emprego da mediana faz-se necessário; assim:

- Nos casos em que existem valores aberrantes, pois têm influência muito menor sobre a mediana do que sobre a média aritmética.

Exemplo: Se na graduatória

24, 37, 41, 52, 65, 68, 82

em lugar de 82 houvesse 1000 casos, isto é,

24, 37, 41, 52, 65, 68, 1000,

o valor da mediana manter-se-ia o mesmo 52 demissões, ao contrário do que acontece com a média aritmética, que passaria de 52,7 demissões a 183,85 demissões.

- Nos casos em que na distribuição em estudo a primeira ou última classe (ou ambas) tenham, respectivamente, o extremo inferior e o extremo superior indefinidos e o centro da distribuição não esteja contido em nenhuma delas. Nessas condições é possível determinar a mediana, o que não acontece com a média aritmética.

Observação: Além da mediana que, por definição, divide um conjunto ordenado de valores em duas partes iguais, existem outras medidas que dividem o conjunto de valores em 4, 10 e 100 partes iguais. Conquanto essas medidas não sejam de tendência central, elas podem ser consideradas medidas de posição, uma vez que fornecem pontos à esquerda ou à direita, dos quais

são encontradas frações da freqüência total. Estas medidas são os *quartis*, os *decis* e os *percentis*.

Os três *quartis* são definidos como os valores que dividem o conjunto ordenado de valores em 4 partes iguais; 25% dos valores são menores do que o primeiro quartil, que é denotado por Q_1; 50% dos valores caem abaixo do segundo quartil, Q_2 (mediana), e 75% dos valores são menores que o terceiro quartil, Q_3. O cálculo de um quartil se faz de maneira análoga ao cálculo de uma mediana, com a diferença de que é necessário contar $\frac{n}{4}$ valores para se achar Q_1 e $\frac{3n}{4}$ para determinar Q_3.

Os *decis* são valores que dividem o conjunto ordenado de valores em 10 partes iguais, isto é, 10% das observações caem abaixo do primeiro decil, denotado por D_1 etc.

Os *percentis* são valores que dividem o conjunto ordenado de valores em 100 partes iguais, isto é, 1% das observações caem abaixo do primeiro percentil, denotado por C_1 etc.

Moda (Mo)

Dada uma distribuição de freqüências, a *moda* é o valor da variável que corresponde à freqüência máxima, isto é, é o valor mais freqüente.

Conquanto o seu resultado seja o mais simples possível, a moda nem sempre existe e nem sempre é única. Quando numa distribuição existem poucos valores da variável, muito freqüentemente não há valores repetidos, com o que nenhum deles satisfaz à condição de moda.

Exemplo: Se os pesos (em quilos) correspondentes a 8 adultos são:

82, 65, 59, 74, 60, 67, 71 e 73,

essas 8 medidas não definem uma moda.

Por outro lado, a distribuição dos pesos de 13 adultos:

63, 67, 70, 69, 81, 57, 63, 73, 68, 71, 71, 71, 83,

possui duas modas, a saber: *Mo* = 63 quilos e Mo = 71 quilos. Nesse caso, a distribuição é chamada de *bimodal*. Será *unimodal* no caso de apresentar uma só moda e *multimodal* se apresentar várias modas.

Observação: É interessante notar que a moda pode ser usada como uma medida de tendência central também no caso de a variável considerada ser de natureza qualitativa. De fato, quando se diz que as faltas ao trabalho constituíram a causa principal de demissão em certo ano, isso quer dizer que na distribuição das demissões, segundo a *causa*, a falta ao trabalho correspondeu a um maior número de demissões, isto é, a rubrica "falta ao trabalho" é a moda da distribuição.

Em se tratando de distribuições de classes de valores, a moda pertence à classe de maior freqüência. Resta, todavia, saber qual o valor da classe deve ser escolhido para representar a moda. Relativamente simples, o cálculo da moda, neste caso, é dado por:

$$Mo = L + t.\frac{f_1}{f_1 + f_2}$$

onde **L** é o extremo inferior da classe em que está a moda, **t** é a amplitude desta classe, $\mathbf{f_1}$ e $\mathbf{f_2}$ são, respectivamente, as freqüências das classes adjacentes à classe da moda.

Exemplo: Na tabela 2, a moda está na classe 30 |– 40, logo,

$L = 30$

$t = 10$

$f_1 = 2$

$f_2 = 10$

e, portanto,

$$Mo = 30 + 10.\frac{2}{2+10} = 30 + \frac{10}{6} = 31{,}667$$

= 31 anos e 8 meses = 31 anos completos.

Observação: o valor da moda, em se tratando de classes, é fortemente afetado pela maneira como as classes são construídas.

Medidas de Dispersão

Sejam A e B duas localidades com mesma renda média por habitante. Esse simples fato de igualdade das duas médias permite concluir que a situação econômica das duas localidades é a mesma? Evidentemente que não, pois essa igualdade poderia existir mesmo que A fosse perfeitamente esta-

bilizada no sentido de que todos os seus habitantes tivessem praticamente a mesma renda (igual à renda média por habitante) e B tivesse uns poucos indivíduos com rendas extraordinariamente altas e a maioria com rendas baixas. Esse simples exemplo basta para mostrar que o conhecimento da intensidade dos valores assumidos por uma grandeza, isto é, da posição de uma distribuição, não é suficiente para a sua completa caracterização.

O fato de em A todos os indivíduos terem a mesma renda pode ser traduzido dizendo que em A as rendas não variam de indivíduo para indivíduo, ou ainda que a distribuição das rendas não apresenta *variabilidade*. Analogamente, o fato de em B alguns indivíduos terem rendas muito elevadas em detrimento da grande maioria, que tem rendas muito baixas, pode ser expresso dizendo-se que em B as rendas variam ou que a distribuição das rendas apresentam variabilidade.

Nesse sentido, várias medidas foram propostas para indicar o quanto os dados se apresentam dispersos em torno da região central. Caracterizam, portanto, o grau de variação (variabilidade) existente no conjunto de dados.

Amplitude de Variação (R)

Uma das medidas mais elementares é a *amplitude*, a qual é definida como sendo a diferença entre o maior e o menor valor do conjunto de dados:

$$R = x_{max} - x_{min}$$

Evidentemente que essa medida é muito precária, pois a amplitude não dá informe algum a respeito da maneira pela qual os valores se distribuem entre os valores extremos.

Por exemplo, nos dois conjuntos de valores:

4, 6, 6, 6, 8

4, 5, 6, 7, 8

a amplitude de variação é a mesma e igual a 4 (8 – 4 = 4) e, no entanto, as dispersões desses dois conjuntos são diferentes. Além disso, os valores mínimo e máximo, estando muito sujeitos às flutuações de amostras, fazem com que a amplitude da distribuição fique igualmente sujeita a tais flutuações. Assim, por exemplo, se existir uma série de indivíduos cujos pesos oscilam entre 50

e 80 quilos, o aparecimento de um único indivíduo que pese 110 quilos fará a amplitude passar de 30 a 60.

Amplitude Semiquartil ou Desvio Quartil

Esta medida, que se baseia na posição ocupada pelos 50% centrais da distribuição, é definida por:

$$Q = \frac{Q_3 - Q_1}{2},$$

onde Q_1 e Q_3 são o primeiro e o terceiro quartis.

Essa medida, conquanto se baseia também em apenas dois valores, apresenta sobre a anterior a vantagem de não estar tão sujeita às flutuações amostrais quanto os valores extremos.

A dispersão poderia ser medida pela *amplitude quartil*, ou seja, $Q_3 - Q_1$; todavia, a divisão por 2 dá a distância média pela qual os quartis se desviam da mediana.

Desvio Padrão e Variância

Para medir a dispersão de uma distribuição faz-se uso da diferença entre cada valor e a média aritmética da distribuição.

As medidas que se baseiam na diferença entre cada valor e a média aritmética da distribuição partem do fato de que a média aritmética é o valor que todas as observações teriam se fossem iguais entre si. Uma vez introduzida a noção de variabilidade, essa propriedade poderia ser expressa dizendo-se que a média aritmética é o valor que todas as observações teriam se não houvesse variabilidade. Daí resulta que o desvio (diferença) de cada observação para a média aritmética representa o quanto as observações variam com relação à média. Nada mais natural, portanto, que definir uma medida de variabilidade baseada nesses desvios. A primeira idéia foi calcular a média aritmética desses desvios.

Se, por exemplo, as observações tivessem os valores:

$$1, 2, 3, 4, 5$$

cuja média é $\overline{X} = 3$, calcular-se-iam as diferenças, como mostrado na tabela 3,

Tabela 3: Diferenças entre as observações e a respectiva média

x_i	$(x_i - \overline{X})$
1	1 – 3 = –2
2	2 – 3 = –1
3	3 – 3 = 0
4	4 – 3 = 1
5	5 – 3 = 2
Total	$\Sigma (x_i - \overline{X}) = 0$

obtendo-se para a medida de variabilidade $\frac{0}{5} = 0$, a qual indica que na distribuição acima não existe variabilidade.

É fácil ver que esta medida, que se apóia num argumento lógico, leva a uma informação errônea sobre a variabilidade. A explicação deste fato reside na propriedade da média aritmética, que diz que a soma de todos os desvios das observações para a média aritmética é nula. Por esta razão, a simples média aritmética dos desvios não pode ser usada como medida de variabilidade.

Ao se atentar para o fato de que a soma dos desvios é sempre igual a zero, porque a cada desvio positivo corresponde um desvio igual, mas de sinal contrário, compreende-se que a situação pode ser contornada calculando-se a média dos módulos dos desvios ou apenas dos quadrados dos desvios.

No primeiro caso ter-se-ia:

x_i	$(x_i - \overline{X})$	$\lvert x_i - \overline{X} \rvert$
1	1 – 3 = –2	2
2	2 – 3 = –1	1
3	3 – 3 = 0	0
4	4 – 3 = 1	1
5	5 – 3 = 2	2
Total	$\Sigma (x_i - \overline{X}) = 0$	**6**

e a medida de variabilidade seria

$$\frac{\Sigma \left| x_i - \overline{X} \right|}{n} = \frac{6}{5} = 1{,}2$$

a qual recebe o nome de *desvio médio* (*DM*), que por motivos de ordem teórica, quase não é usado.

No segundo caso, ter-se-ia:

x_i	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$
1	$1 - 3 = -2$	4
2	$2 - 3 = -1$	1
3	$3 - 3 = 0$	0
4	$4 - 3 = 1$	1
5	$5 - 3 = 2$	4
Total	$\Sigma (x_i - \bar{X}) = 0$	**10**

e a medida de variabilidade seria

$$\frac{\Sigma\left(x_i - \bar{X}\right)^2}{n} = \frac{10}{5} = 2$$

a qual recebe o nome de *variância* (*Var* ou σ^2).

Entretanto, quando calculamos a variância de um grupo de observações, este grupo provém de um outro ainda maior, que inclui todos os possíveis valores da variável X. Em geral, desejamos que a variância do nosso grupo seja uma estimativa da variância de todas as observações de onde os nossos dados particulares foram retirados. Pode ser mostrado que, quando a variância do grupo maior é definida como feito acima, a variância do grupo derivado deveria ser definida como

$$S^2 = \text{Var}(X) = \frac{\Sigma\left(x_i - \bar{X}\right)^2}{n-1}$$

com o objetivo de obter uma boa estimativa da variância do grupo mais amplo. Por isso usaremos *n* - 1 em lugar de *n* como divisor.

A unidade em que a variância é expressa será a unidade original ao quadrado e, para comparar a unidade da nossa medida de variabilidade com a dos dados originais, extraímos a raiz quadrada,

$$S = \sqrt{\frac{\Sigma\left(x_i - \bar{X}\right)^2}{n-1}}$$

a qual recebe o nome de *desvio-padrão*. O desvio-padrão é expresso nas

mesmas unidades dos dados originais. Tanto o desvio-padrão (S) quanto a variância (S^2 ou Var(X)), são usados como medidas de variabilidade. Conforme a finalidade, é conveniente o uso de uma ou de outra.

De maneira geral, ao se ter uma distribuição de freqüências, utiliza-se para o cálculo da variância a seguinte expressão:

$$\frac{\sum\left(x_i - \overline{X}\right)^2 . f_i}{n-1}$$

onde, os *xi's* podem ser os valores individuais da variável X ou os pontos médios das classes.

Como exemplo, tome a Tabela 2, lembrando-se que a média aritmética foi igual a 47,5 anos:

Valores x_i de X (anos)	Ponto médio da classe	f_i	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2 . f_i$
20 |– 30	25	2	-22,5	506,25	1 012,50
30 |– 40	35	11	-12,5	156,25	1 718,75
40 |– 50	45	10	-2,5	6,25	62,50
50 |– 60	55	9	7,5	56,25	506,25
60 |– 70	65	8	17,5	306,25	2 450,00
Total		**40**			**5 750,00**

$$S^2 = \frac{\sum\left(x_i - \overline{X}\right)^2}{n-1} . f_i = \frac{5\,750}{39} = 147{,}44 \text{ anos}$$

$$S = \sqrt{S^2} = \sqrt{147{,}44} = 12{,}14 \text{ anos.}$$

Considerações finais sobre o desvio-padrão:

- O desvio-padrão é uma quantidade essencialmente positiva.
- O desvio-padrão só é *nulo* se todos os valores da distribuição forem iguais entre si, isto é, se não houver variabilidade.
- O desvio-padrão é da mesma natureza da variável X e depende também de sua magnitude.

Coeficiente de Variação

Para comparar duas distribuições quanto à variabilidade, deve-se usar *medidas de variabilidade relativa*, tais como o *coeficiente de variação de*

Pearson (*CV*), o qual é dado por: $CV = \frac{S}{\overline{X}}$ o qual independe da natureza e magnitude da variável X.

Esse resultado é multiplicado por 100, para que o coeficiente de variação seja dado em porcentagem.

Exemplo: Para duas emissões de ações ordinárias da indústria eletrônica, o preço médio diário, no fechamento dos negócios, durante um período de um mês, para as ações A, foi de R$ 150,00 com um desvio padrão de R$ 5,00. Para as ações B, o preço médio foi de R$ 50,00 com um desvio padrão de R$ 3,00. Em termos de comparação absoluta, a variabilidade do preço das ações A foi maior, devido ao desvio padrão maior. Mas em relação ao nível de preço, devem ser comparados os respectivos coeficientes de variação:

$$CV(A) = \frac{S_A}{\overline{X}_A} = \frac{5}{150} = 0,033 \text{ ou } 3,3\%$$

$$CV(B) = \frac{S_B}{\overline{X}_B} = \frac{3}{50} = 0,060 \text{ ou } 6\%$$

Portanto, relativamente ao nível médio de preços das ações, podemos concluir que o preço da ação B é quase duas vezes mais variável que o preço da ação A.

Ampliando seus conhecimentos

(MATTAR, 1996)

É importante que um pesquisador que vá realizar uma coleta de informações tenha noções básicas sobre os diferentes tipos e aplicações de metodologias de pesquisa. Veremos aqui algumas definições que irão facilitar a diferenciação entre os diferentes tipos de pesquisa:

Projeto de Pesquisa: Cada planejamento de pesquisa realizado cientificamente tem um padrão específico para controlar a coleta de dados. Este padrão chama-se *projeto de pesquisa*. Sua função é assegurar que os dados exigidos sejam coletados de maneira precisa e econômica.

Os projetos de pesquisa podem ser agrupados nas seguintes categorias: exploratória, descritiva e experimental.

a) **Pesquisa Exploratória** – Visa fornecer ao pesquisador um maior conhecimento do tema ou problema de interesse. É apropriada para os primeiros estágios da investigação quando a familiaridade, o conhecimento e a compreensão do fenômeno por parte do pesquisador são insuficientes.

O projeto formal está quase ausente nos estudos exploratórios. A imaginação do explorador é o fator principal. Entretanto, há 4 linhas de ataque que podem ajudar na descoberta de hipóteses valiosas:

- **Levantamentos em fontes secundárias** – Levantamentos bibliográficos, levantamentos documentais, levantamentos de estatísticas e levantamentos de pesquisas realizadas.
- **Levantamentos de experiências** – Muitas pessoas, em função da posição estratégica que ocupam numa empresa ou instituição, acumulam experiências e conhecimentos sobre um tema ou problema em estudo. Informações são levantadas a partir de entrevistas individuais ou em grupo, realizadas com especialistas ou conhecedores do assunto.
- **Estudo de casos selecionados** – Exame de registros existentes, observação da ocorrência do fato, entrevistas etc. (cases). Casos que reflitam mudanças, comportamentos ou desempenhos extremados, dificuldades superadas etc.
- **Observação informal** – A utilização do processo de observação do dia-a-dia em pesquisa exploratória deve ser informal e dirigida, ou seja, centrada unicamente em observar objetos, comportamentos e fatos de interesse para o problema em estudo.

b) **Pesquisa Descritiva** – Destinam-se a descrever as características de determinada situação. Ao contrário do que o ocorre nas pesquisas exploratórias, a elaboração das questões de pesquisa pressupõe profundo conhecimento do problema a ser estudado. Os estudos descritivos não devem ser encarados como simples coletas de dados, embora infelizmente, muitos deles não são mais do que isso. Para ser valioso, o estudo descritivo precisa coletar dados com um objetivo definido e deve incluir uma interpretação por um investigador. Pode ser dividido nos seguintes tipos:

- **Levantamentos de campo (método estatístico)** – Procuram-se dados representativos da população de interesse, a amostra é ge-

rada a partir de métodos estatísticos, tem-se total controle sobre a representatividade dos dados obtidos em relação à população. Permite a geração de tabelas sumarizadas por categorias e a generalização dos resultados para toda a população. No entanto não permite aprofundar os tópicos da pesquisa pela própria característica de gerar sumários estatísticos. É dispendioso em termos de tempo e isto requer grandes conhecimentos técnicos.

- **Estudos de campo** – É o método de estudo intensivo de um número relativamente pequeno de casos. Por exemplo, um investigador pode fazer um estudo detalhado entre alguns consumidores, alguns varejistas, alguns sistemas de controle de vendas, ou alguns mercados de cidades pequenas. Deve ser considerado como um estágio diferente no desenvolvimento de um método científico comum. Servem para geração de hipóteses em vez de teste de hipóteses, recomendados quando há grande homogeneidade entre os elementos da população. Entretanto somente investigam após a ocorrência do fato e geralmente não podem ser generalizados.

c) **Pesquisa Experimental** – Este método pode ser resumido na expressão: "Se ocorrer isto, provavelmente ocorrerá aquilo". Neste caso, ocorre uma observação da relação de causalidade entre várias possíveis causas e o efeito pressuposto.

$$y = f(x, z, t, v, s,...)$$

onde y, é a variável dependente e as demais são independentes. Ganha-se maior confiabilidade nos resultados, à medida que repetidas experimentações com as mesmas variáveis independentes e dependente indicam sempre as mesmas conclusões.

Atividades de aplicação

1. Em uma determinada empresa X, a média dos salários é 10 000 unidades monetárias e o 3º quartil é 5 000. Pergunta-se:

a) Se você se apresentasse como candidato a esta empresa e se o seu salário fosse escolhido ao acaso entre todos os possíveis salários, o que seria mais provável: ganhar mais ou menos que 5 000 unidades monetárias? Justifique!

b) Suponha que na empresa Y a média dos salários é 7 000 unidades monetárias e a variância é praticamente zero, e lá o seu salário também seria escolhido ao acaso. Em qual empresa você se apresentaria para procurar emprego X ou Y? Justifique!

2. A média aritmética é a razão entre:

a) o número de valores e o somatório deles.

b) o somatório dos valores e o número deles.

c) os valores extremos.

d) os dois valores centrais.

e) nenhuma das alternativas anteriores.

3. Na série 60, 90, 80, 60, 50 a moda é:

a) 50

b) 60

c) 66

d) 90

e) nenhuma das anteriores.

4. A estatística que possui o mesmo número de valores abaixo e acima dela é:

a) a moda.

b) a média.

c) a mediana.

d) o elemento mediano.

e) nenhuma das anteriores.

5. A soma dos desvios entre cada valor e a média sempre será:

a) positiva.

b) negativa.

c) zero.

d) diferente de zero.

e) nenhuma das alternativas anteriores.

6. Considere a série 6, 5, 7, 8, 9 o valor 7 será:

a) a média e a moda.

b) a média e a mediana.

c) a mediana e a moda.

d) a média, a mediana e a moda.

e) nenhuma das alternativas anteriores.

7. Quando desejamos verificar a questão de uma prova que apresentou maior número de erros, utilizamos:

a) moda.

b) média.

c) mediana.

d) qualquer das anteriores.

e) nenhuma das anteriores.

8. O coeficiente de variação é uma estatística denotada pela razão entre:

a) desvio padrão e média.

b) média e desvio padrão.

c) mediana e amplitude interquartílica.

d) desvio padrão e moda.

e) nenhuma das alternativas anteriores.

9. Uma prova de estatística foi aplicada para duas turmas. Os resultados seguem abaixo

Turma 1: média = 5 e desvio padrão = 2,5

Turma 2: média = 4 e desvio padrão = 2,0

Com esses resultados podemos afirmar:

a) a turma 2 apresentou maior dispersão absoluta.

b) a dispersão relativa é igual à dispersão absoluta.

c) tanto a dispersão absoluta quanto a relativa são maiores para a turma 2.

d) a dispersão absoluta da turma 1 é maior que a turma 2, mas em termos relativos as duas turmas não diferem quanto ao grau de dispersão das notas.

e) nenhuma das alternativas anteriores.

10. Uma empresa possui dois serventes recebendo salários de R$ 250,00 cada um, quatro auxiliares recebendo R$ 600,00 cada um, um chefe com salário de R$1.000,00 e três técnicos recebendo R$ 2.200,00 cada um. O salário médio será:

a) R$ 1.050,00

b) R$ 1.012,50

c) R$ 405,00

d) R$ 245,00

e) nenhuma das alternativas anteriores.

11. O cálculo da variância supõe o conhecimento da:

a) média.

b) mediana.

c) moda.

d) ponto médio.

e) desvio padrão.

12. Em uma determinada distribuição de valores iguais, o desvio padrão é:

a) negativo.

b) positivo.

c) a unidade.

d) zero.

e) nenhuma das alternativas anteriores.

13. Dados os conjuntos de números X = {–2, –1, 0, 1, 2} e Y = {220, 225, 230, 235, 240}, podemos afirmar, de acordo com as propriedades do desvio padrão, que o desvio padrão de Y será igual:

a) ao desvio padrão de X.

b) ao desvio padrão de X, multiplicado pela constante 5.

c) ao desvio padrão de X, multiplicado pela constante 5, e esse resultado somado a 230.

d) ao desvio padrão de A mais a constante 230.

e) nenhuma das alternativas anteriores.