ESTATÍSTICA

Professora Me. Ivnna Gurniski de Oliveira
Professora Me. Renata Cristina de Souza Chatalov

GRADUAÇÃO
Unicesumar

UNICESUMAR
Av. Guedner, 1610 - Jardim Aclimação
Cep 87050-900 - MARINGÁ - PARANÁ
unicesumar.edu.br
44 3027.6360

UNICESUMAR EDUCAÇÃO A DISTÂNCIA
NEAD - Núcleo de Educação a Distância
Bloco 4 - MARINGÁ - PARANÁ
unicesumar.edu.br
0800 600 6360

FICHA CATALOGRÁFICA

C397 **CENTRO UNIVERSITÁRIO DE MARINGÁ.** Núcleo de Educação a Distância; **OLIVEIRA**, Ivnna Gurniski de; **CHATALOV**, Renata Cristina de Souza.

Estatística. Ivnna Gurniski de Oliveira; Renata Cristina de Souza Chatalov.
Maringá-Pr.: UniCesumar, 2017.
182 p.
"Graduação - EaD".

1. Estatística. 2. Probabilidade. 3. EaD. I. Título.

ISBN 978-85-459-0719-0

CDD - 22 ed. 515.5
CIP - NBR 12899 - AACR/2

Ficha catalográfica elaborada pelo bibliotecário
João Vivaldo de Souza - CRB-8 - 6828

Impresso por:

Professor
Wilson de Matos Silva
Reitor

Em um mundo global e dinâmico, nós trabalhamos com princípios éticos e profissionalismo, não somente para oferecer uma educação de qualidade, mas, acima de tudo, para gerar uma conversão integral das pessoas ao conhecimento. Baseamo-nos em 4 pilares: intelectual, profissional, emocional e espiritual.

Iniciamos a Unicesumar em 1990, com dois cursos de graduação e 180 alunos. Hoje, temos mais de 100 mil estudantes espalhados em todo o Brasil: nos quatro campi presenciais (Maringá, Curitiba, Ponta Grossa e Londrina) e em mais de 300 polos EAD no país, com dezenas de cursos de graduação e pós-graduação. Produzimos e revisamos 500 livros e distribuímos mais de 500 mil exemplares por ano. Somos reconhecidos pelo MEC como uma instituição de excelência, com IGC 4 em 7 anos consecutivos. Estamos entre os 10 maiores grupos educacionais do Brasil.

A rapidez do mundo moderno exige dos educadores soluções inteligentes para as necessidades de todos. Para continuar relevante, a instituição de educação precisa ter pelo menos três virtudes: inovação, coragem e compromisso com a qualidade. Por isso, desenvolvemos, para os cursos de Engenharia, metodologias ativas, as quais visam reunir o melhor do ensino presencial e a distância.

Tudo isso para honrarmos a nossa missão que é promover a educação de qualidade nas diferentes áreas do conhecimento, formando profissionais cidadãos que contribuam para o desenvolvimento de uma sociedade justa e solidária.

Vamos juntos!

Janes Fidélis Tomelin
Pró-Reitor de Ensino de EaD

Kátia Solange Coelho
Diretoria de Graduação e Pós

Débora do Nascimento Leite
Diretoria de Design Educacional

Leonardo Spaine
Diretoria de Permanência

Seja bem-vindo(a), caro(a) acadêmico(a)! Você está iniciando um processo de transformação, pois quando investimos em nossa formação, seja ela pessoal ou profissional, nos transformamos e, consequentemente, transformamos também a sociedade na qual estamos inseridos. De que forma o fazemos? Criando oportunidades e/ou estabelecendo mudanças capazes de alcançar um nível de desenvolvimento compatível com os desafios que surgem no mundo contemporâneo.

O Centro Universitário Cesumar mediante o Núcleo de Educação a Distância, o(a) acompanhará durante todo este processo, pois conforme Freire (1996): "Os homens se educam juntos, na transformação do mundo".

Os materiais produzidos oferecem linguagem dialógica e encontram-se integrados à proposta pedagógica, contribuindo no processo educacional, complementando sua formação profissional, desenvolvendo competências e habilidades, e aplicando conceitos teóricos em situação de realidade, de maneira a inseri-lo no mercado de trabalho. Ou seja, estes materiais têm como principal objetivo "provocar uma aproximação entre você e o conteúdo", desta forma possibilita o desenvolvimento da autonomia em busca dos conhecimentos necessários para a sua formação pessoal e profissional.

Portanto, nossa distância nesse processo de crescimento e construção do conhecimento deve ser apenas geográfica. Utilize os diversos recursos pedagógicos que o Centro Universitário Cesumar lhe possibilita. Ou seja, acesse regularmente o Studeo, que é o seu Ambiente Virtual de Aprendizagem, interaja nos fóruns e enquetes, assista às aulas ao vivo e participe das discussões. Além disso, lembre-se que existe uma equipe de professores e tutores que se encontra disponível para sanar suas dúvidas e auxiliá-lo(a) em seu processo de aprendizagem, possibilitando-lhe trilhar com tranquilidade e segurança sua trajetória acadêmica.

AUTORAS

Professora Me. Ivnna Gurniski de Oliveira

Mestre em Ensino de Ciências e Educação Matemática da Universidade Estadual de Londrina, especialista em Docência no Ensino Superior pela Unicesumar - Centro Universitário Cesumar, graduada em Licenciatura em Matemática pela Universidade Estadual de Maringá.

Professora Me. Renata Cristina de Souza Chatalov

Possui graduação em Tecnologia Ambiental pelo Centro Federal de Educação Tecnológica do Paraná. Especialista em Gestão Ambiental pela Faculdade Estadual de Ciências e Letras de Campo Mourão - FECILCAM. Mestre em Engenharia Urbana pela Universidade Estadual de Maringá - UEM. Tem experiência em pesquisa na área de Sistema de Gestão de Qualidade, na Área Ambiental, com ênfase em Tecnologias Avançadas de Tratamento de Efluentes, Gestão e Tratamento de Resíduos Sólidos. Trabalha como Professora Formadora no curso de Gestão Ambiental, Gestão de Recursos Humanos, Gestão de Negócios Imobiliários, Segurança do Trabalho no EAD no Centro Universitário Cesumar - UniCesumar. Professora no curso de graduação em Administração na Faculdade Metropolitana de Maringá. Coordenadora do Curso de Tecnologia em Gestão Ambiental na Faculdade Metropolitana de Maringá. Professora da disciplina de Indústria e Meio Ambiente na Pós-graduação em Gestão Ambiental na Faculdade Metropolitana de Maringá. Professora da pós-graduação EAD UniCesumar.

SEJA BEM-VINDO(A)!

Caro(a) estudante, é com muito prazer que apresentamos a você o livro que fará parte da disciplina de Estatística. A Estatística é uma ciência que se dedica ao desenvolvimento e ao uso de métodos para a coleta, resumo, organização, apresentação e análise de dados. Um exemplo do uso da estatística está na previsão do tempo em uma região, em tendências numa eleição, a posição dos bancos dos trens em certa linha e, até, o hábito de lavar as mãos após usar o banheiro.

Fazendo uma pequena viagem pelo tempo, em 3000 a.C., registrava-se os primeiros indícios de censos na Babilônia, na China e no Egito. No Velho Testamento, Livro 4° (Números), registra-se uma instrução de Moisés: "Fazer levantamento dos homens de Israel aptos a guerrear" (TOREZANI, 2004, p. 2).

A palavra "Censo" deriva do verbo latino "censere", que significa taxar. O objetivo inicial da realização dos censos era buscar informações sobre as populações para orientar a taxação de impostos. Era, portanto, uma atividade que interessava, particularmente, aos governos, ao Estado. Daí deriva a palavra ESTATÍSTICA (de STATUS). Trata-se da ferramenta de trabalho dos estadistas.

Em 1805, Guilherme, o Conquistador, determinou que se fizesse, na Inglaterra, um levantamento, visando obter informações sobre posse de terras, sua utilização, seus proprietários, número de empregados, posse de animais etc., para taxação de impostos.

No século XVII, John Graint publica "Aritmética Política", uma análise sobre nascimentos e óbitos, a partir das chamadas Tábuas de Mortalidade.

Já, no século XVIII (1797), surge, na Enciclopédia Britânica, o verbete "STATISTICS" pela primeira vez.

O termo "Estatística" é usado, hoje, com alguns significados diferentes. Ele pode se referir a:

- meros registros de eventos que interessem ao administrador em geral;
- uma simples medida estatística que seja obtida de uma amostra;
- métodos estatísticos padronizados utilizados em pesquisa por amostragem;
- ciência estatística em geral, hoje, grandemente desenvolvida e com aplicação disseminada como auxiliar as mais diferentes áreas de conhecimento.

De forma simplificada, podemos admitir que a Ciência Estatística tem como objetivo obter informações confiáveis sobre determinado fenômeno de interesse.

A Estatística está de forma muito presente na mídia, seja em jornais, revistas ou meios de comunicação. Além disso, uma vez que está diretamente envolvida com pesquisa, é a partir dela que as decisões são tomadas. Podemos dizer que a Estatística é uma ferramenta para qualquer pesquisador na busca pelas respostas aos vários problemas relacionados ao meio em que trabalha. Entretanto, para que ela seja bem utilizada, é necessário conhecer os seus fundamentos, seus princípios e suas ferramentas para que possamos utilizá-la de forma adequada.

APRESENTAÇÃO

Este material foi separado em cinco unidades, sendo iniciado com a importância da Estatística básica, passando por probabilidades e finalizando com medidas de associação.

A unidade I vai do início de sua utilização até a importância dos gráficos e das tabelas na apresentação dos dados. Essa unidade trata, basicamente, dos conceitos que você precisará saber para entender a Estatística nas unidades posteriores.

Na unidade II, nos aprofundaremos no estudo de tabelas e de gráficos, mais especificamente, leitura e construção de tabelas, aplicação e utilização de alguns tipos de gráficos.

A unidade III mostra as medidas de posição e de dispersão. Essas medidas são, amplamente, empregadas dentro de pesquisas em nível científico e, também, nos problemas mais simples do cotidiano.

A unidade IV trata sobre probabilidades. As probabilidades podem tratar de eventos simples a extremamente complexos. De forma abrangente, elas tratam das chances de determinados fenômenos ocorrerem. A importância de se estudar probabilidades está na verificação de que alguns eventos ocorrem com uma facilidade maior que outros e, assim, podemos prever situações futuras sobre esses eventos.

A unidade aborda as probabilidades de forma geral, mostrando desde os cálculos mais simples, passando por suas propriedades, e indo até as probabilidades condicionais e distribuições de probabilidades. As principais distribuições são aquelas que utilizamos com maior frequência, uma vez que existem inúmeros tipos. Essas distribuições do comportamento da variável com a qual estamos trabalhando é importante, pois, por meio delas é que determinamos como calcular probabilidades de forma correta.

Finalizando o material, a unidade V trata das medidas de associação, mais especificamente a correlação e a análise de regressão. Essas medidas nos mostram o grau de relação entre duas variáveis. A correlação informa a intensidade da relação e a análise de regressão mostra a quantidade de variação em uma por meio da variação em outra.

Este material está bastante sintetizado, focando os pontos principais da Estatística de modo a proporcionar encaminhamentos que possibilitem a compreensão dos conceitos, ao contrário do que muitas vezes é posto em se tratando de estudar Matemática e, especificamente, Estatística.

A resolução de tarefas é importante desde que o(a) estudante procure fazê-la à luz da teoria que ela contempla. Com isso, afirmo: será necessário, também, muito empenho de sua parte para a realização desse intenso trabalho. No decorrer de suas leituras, procure interagir com os textos, fazer anotações, responder as atividades de estudo, anotar suas dúvidas, ver as indicações de leitura e realizar novas pesquisas sobre os assuntos tratados, pois com certeza não será possível esgotá-los em apenas um livro.

Prof.ª Ivnna Gurniski de Oliveira

Prof.ª Renata C. de Souza Chatalov

SUMÁRIO

UNIDADE I

CONCEITOS E IMPORTÂNCIA DA ESTATÍSTICA

UNIDADE II

TABELAS E GRÁFICOS

SUMÁRIO

UNIDADE III

MEDIDAS DESCRITIVAS ASSOCIADAS A VARIÁVEIS QUANTITATIVAS

UNIDADE IV

PROBABILIDADES

SUMÁRIO

UNIDADE V

CORRELAÇÃO LINEAR E REGRESSÃO

Professora Me. Ivnna Gurniski de Oliveira
Professora Me. Renata Cristina de Souza Chatalov

UNIDADE I

CONCEITOS E IMPORTÂNCIA DA ESTATÍSTICA

Objetivos de Aprendizagem

- Entender o que significa Estatística.
- Compreender a importância da Estatística.
- Assimilar os principais conceitos dentro da Estatística.
- Compreender as principais formas de apresentação de dados estatísticos.

Plano de Estudo

A seguir, apresentam-se os tópicos que você estudará nesta unidade:

- A Importância da Disciplina de Estatística
- População e Amostra
- Amostragem
- Tipos de Variáveis
- Fases do Método Estatístico

Unicesumar

INTRODUÇÃO

Normalmente, as pessoas imaginam que a Estatística é simplesmente uma coleção de números, ou que tem a ver apenas com censo demográfico, com a construção de tabelas ou de gráficos. Podemos afirmar que a Estatística vai muito além de disso e que, na verdade, ela é muito frequente na nossa vida.

Como exemplos de aplicações de técnicas estatísticas, temos: a pesquisa eleitoral, a pesquisa de mercado, o controle de qualidade, os índices econômicos, o desenvolvimento de novos medicamentos, as novas técnicas cirúrgicas e de tratamento médico, as sementes mais eficientes, as previsões meteorológicas, as previsões de comportamento do mercado de ações, dentre outros, isto é, tudo que se diz cientificamente comprovado, por algum momento, passa por procedimentos estatísticos.

Portanto, podemos definir Estatística como um conjunto de técnicas de análise de dados, que são aplicáveis a quase todas as áreas do conhecimento, e que nos auxiliam no processo de tomada de decisão. Estatística é a ciência que estuda os processos de coleta, de organização, de análise e de interpretação de dados relevantes e referentes a uma área particular de investigação.

Você verá que a Estatística é uma ciência multidisciplinar que permite a análise de dados em todas as áreas e que fornece ferramentas para que sejamos capazes de transformar dados brutos em informações acessíveis e de fácil compreensão, de modo que possamos compará-los com outros resultados ou, ainda, verificar sua adequação com alguma teoria pronta.

Abordaremos que a Estatística tem uma base na formação do acadêmico, pois é de extrema importância para o desenvolvimento dos alunos saber observar as tabelas e os gráficos e usar essa ferramenta para a tomada de decisões.

Nesta unidade, serão apresentados conceitos básicos em Estatística, que são subsídios para o desenvolvimento de todo o estudo proposto neste livro. Então, aproveite bem essa unidade, e lembre-se que ela será um subsídio para toda nossa disciplina.

A IMPORTÂNCIA DA DISCIPLINA DE ESTATÍSTICA

A palavra Estatística deriva do latim "*status*", que significa ESTADO. Os primeiros usos da Estatística se deram com base no conhecimento da população em relação às suas riquezas e na coleta de impostos. Posteriormente, foi empregada no manuseio de dados que descreviam aspectos de um Estado ou país, daí a origem da palavra estar relacionada a Estado.

A Estatística está presente no dia a dia de qualquer indivíduo, seja na abordagem governamental em que somos questionados, como o censo demográfico, ou então quando somos abordados sobre qual candidato que iremos votar na próxima eleição, quando nasce um indivíduo, ou quando estamos consumindo algum tipo de produto; em todas essas situações, dentre tantas outras, estamos fazendo parte da Estatística.

A análise estatística tem, assim, por objetivo, a resolução de problemas, bem como a produção de conhecimentos que geram novos problemas e, portanto, podemos dizer que está envolvida diretamente em um processo iterativo, sendo seu principal objetivo auxiliar na tomada de decisão.

Segundo Rao (1999), a Estatística é uma ciência que estuda e pesquisa tanto o levantamento de dados, quanto o processamento destes para a quantificação da incerteza existente na resposta para um determinado problema; é a tomada de decisões sob condições de incerteza, sob o menor risco possível.

A importância da Estatística está presente em todos os segmentos ligados à pesquisa, de forma geral e abrangente. A maioria desses órgãos possui departamentos oficiais destinados à realização de estudos estatísticos. A Estatística tornou-se responsável, nos últimos tempos, pelo desenvolvimento científico e tecnológico, sendo que é a partir dela que analisamos dados e tomamos as decisões.

Ainda, podemos dizer que ela fornece meios precisos e rigorosos na verificação e na análise dos dados, transformando-os em informações claras e a partir das quais tomamos nossas decisões baseados em comprovações científicas, e não em "achismos".

Dentre outros atributos, podemos dizer ainda que o estudo da Estatística justifica-se pela necessidade de desenvolver pesquisas e pela utilização dos resultados visando à comprovação de alguma hipótese e à solução de algum problema.

Ademais, atualmente as empresas têm procurado admitir profissionais que tenham certo nível de conhecimento em Estatística, pois este tem resultado em diferença significativa nos processos decisórios. Torna-se fundamental para qualquer indivíduo ter conhecimentos básicos e saber aplicá-los de maneira coerente, utilizando técnicas estatísticas nos diferentes casos que podem surgir.

CONCEITOS BÁSICOS EM ESTATÍSTICA

A Estatística tem por objetivo fornecer métodos e técnicas para que se possa lidar com situações de incerteza, e pode ser subdividida em três áreas: descritiva, probabilística e inferencial.

A estatística descritiva, também chamada de estatística dedutiva, tem como objetivo organizar, resumir e simplificar as informações, a fim de torná-las mais fáceis de serem entendidas, transmitidas e discutidas. Como o nome indica, ela descreve os fenômenos de forma prática e acessível, ou seja, por meio de tabelas, gráficos e medidas resumo, que veremos nas próximas unidades. Assim, podemos captar rapidamente, por exemplo, o significado de uma "taxa de desemprego", de um "consumo médio de combustível por quilômetro", ou de uma "nota média de estudantes".

A Estatística Inferencial objetiva "inferir" conclusões sobre a população, interpretando os dados colhidos de uma amostra. Para isso, utiliza amplamente a

"Teoria das Probabilidades", que é fundamental para avaliar situações que envolvam o acaso. A aplicação de métodos probabilísticos nos permite "quantificar" a importância do acaso.

Assim, resultados obtidos por amostragem são "testados", utilizando-se conhecimentos probabilísticos, a fim de se determinar até que ponto eles são significativos, isto é, não são obra do acaso.

Quando tratamos de dados estatísticos, podemos optar por dois processos: o Censo e as Estatísticas, que podem ser assim definidos:

- Censo: processo que consiste no exame de todos os elementos da população. Exemplo: censo demográfico, censo industrial, etc.
- Estatísticas: utilizadas para avaliar os elementos de uma amostra.

A partir do censo, são encontradas medidas que descrevem toda a população, os chamados Parâmetros e, ao se trabalhar com amostras, são obtidas as estimações e, a partir delas, os Estimadores, como definidos a seguir:

- Parâmetros: medidas descritivas de uma população. Exemplo: a contagem do número total de habitantes de uma região.
- Estimadores: medidas descritivas de uma amostra e que indiretamente estimam um parâmetro pelo cálculo de probabilidades. Exemplo: proporção de votantes em certo candidato obtido por amostragem.

POPULAÇÃO E AMOSTRA

A Estatística fornece vários métodos para organizar e para resumir um conjunto de dados e, com base nestas informações, tirar conclusões. Quando se fala em conjunto de dados ou fatos coletados, este se refere ao material tomado a partir de um conjunto de elementos. Deve-se, então, definir de onde esses dados serão tomados, e assim surge o conceito de População (BARBETTA, 2014).

População pode ser definida como sendo uma coleção de elementos que possuem alguma característica em comum, podendo estes ser animados ou inanimados.

Quando as informações desejadas estiverem disponíveis para todos os objetos da população, temos o chamado **censo**. Normalmente, é impraticável ou inviável trabalhar com a população quando se faz Estatística (CRESPO, 2009). Isto é devido a alguns fatores:

- Restrição de tempo ou de recursos.
- População "infinita", entre outros.

Assim, o procedimento comum é coletar desta população um subconjunto de elementos, as chamadas Amostras.

Amostra pode ser definida como uma parte da população. Entretanto, este conceito deve ir um pouco mais além. Uma amostra deve ser representativa da população, ou seja, deve ter todas as características da população de onde foi extraída.

A partir do estudo do conjunto de dados obtido na amostra, faz-se uma extrapolação dos seus resultados para a população toda. Essa extrapolação é chamada Inferência.

Um exemplo pode ser dado em estudos de opinião pública sobre a aceitação de um candidato às eleições, ou então sobre a durabilidade de aparelhos, resistência de materiais, etc.

A escolha das unidades que comporão a amostra é feita por um processo chamado Amostragem, e este pode ser feito de várias maneiras, dependendo do que se tem em mãos, por exemplo, do tamanho da população e do conhecimento que se tem dela.

AMOSTRAGEM

A amostragem é utilizada no nosso cotidiano. Por exemplo, para verificar o tempero de um alimento em preparação, provamos uma pequena porção (amostra), dessa maneira estamos fazendo uma amostragem, isto é, extraindo do todo (população) uma parte (amostra) com o objetivo de se ter uma ideia (inferência) sobre a qualidade do tempero em todo o alimento preparado (BARBETTA, 2014).

Em pesquisas científicas, na qual desejamos, muitas vezes, conhecer algumas características (parâmetros) de uma população, podemos utilizar as técnicas de amostragens, a fim de obtermos os valores aproximados (estimativas) para os parâmetros em que estamos interessados.

Esse tipo de pesquisa é chamado de levantamento por amostragem (BARBETTA, 2014). No entanto, a seleção dos elementos que serão utilizados para fazerem parte da amostra, deve ser feita sob uma metodologia adequada, de tal forma que os resultados da amostra sejam suficientemente informativos para se inferir sobre os parâmetros populacionais.

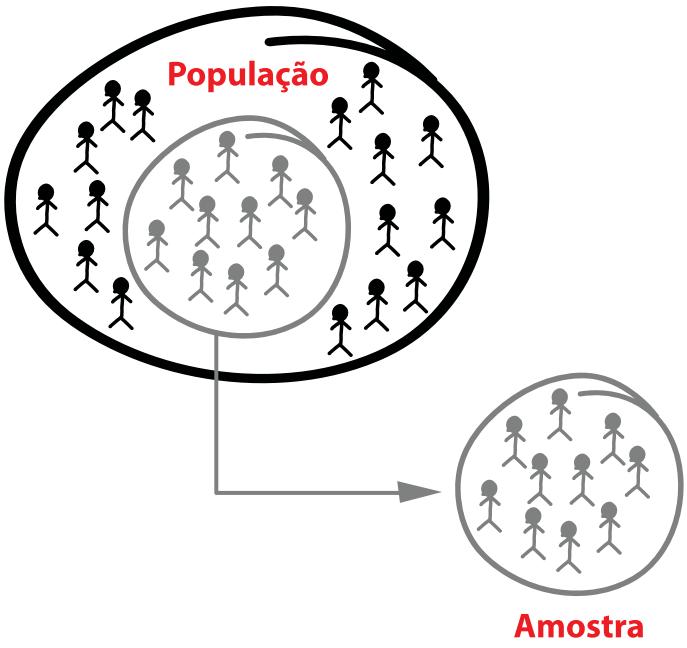


Figura 1 - Exemplo de amostra

E por que utilizamos as técnicas de amostragem?

01. **Economia**: normalmente é mais econômico fazermos o levantamento somente em uma parcela da população, não em um todo.

02. **Tempo**: Muitas vezes não temos tempo suficiente para analisar toda população, por exemplo: queremos fazer uma pesquisa eleitoral, a cinco dias antes das eleições. Imaginem o tempo para isso.

03. **Confiabilidade e operacionalidade**: quando pesquisamos em um número reduzido de elementos, podemos dar mais atenção aos casos individuais.

E quando a amostragem não se torna interessante?

01. **População pequena**: quando temos uma população pequena é inviável utilizarmos uma técnica de amostragem, imaginem a situação: vamos fazer uma entrevista com alunos de uma turma com dez alunos; nesse caso, é mais interessante entrevistamos os dez alunos do que aplicarmos uma técnica de amostragem a fim de obtermos a amostra.

02. **Característica de fácil mensuração**: a população pode não ser tão pequena, mas a variável que se quer observar é de tão fácil mensuração que não compensa investir em um plano de amostragem. Como por exemplo, queremos fazer uma pesquisa sobre o local de uma festa de confraternização em uma empresa, assim podemos entrevistar todos os colaboradores no próprio local de trabalho.

03. Necessidade de alta precisão: a cada dez anos o Instituto Brasileiro de Geografia e Estatística (IBGE) realiza um censo demográfico a fim de estudar várias características da população brasileira. Para Barbetta (2014), dentre essas características, tem-se o parâmetro número de habitantes residentes no país. É um parâmetro que precisa ser avaliado com grande precisão; por isso, é pesquisada toda população.

Para que se obtenha uma amostra representativa da população, o processo de coleta deve ser feito de forma adequada, no qual cada situação exige uma maneira apropriada.

Existem dois grandes grupos de técnicas amostrais:

- Probabilísticas: quando todos os elementos da população têm probabilidade conhecida e diferente de zero de pertencer à amostra.
- Não probabilísticas: quando nem todos os elementos da população têm probabilidade conhecida de pertencer à amostra.

A amostragem probabilística é a mais recomendada para garantir a representatividade da amostra, pois implica em um sorteio dos elementos com regras bem determinadas, sendo possível, apenas, quando a população é finita.

As principais técnicas de amostragem probabilísticas são:

AMOSTRAGEM CASUAL SIMPLES

Para se ter uma amostra casual simples, precisa-se de uma listagem com todos os elementos da população de origem. Os elementos que farão parte da amostra devem ser obtidos de forma totalmente aleatória, ou seja, por sorteio e sem restrição. É escrito cada elemento em um cartão e assim sorteado os participantes da amostra. Todos os elementos da população têm igual probabilidade $\left(\frac{n}{N}\right)$ de pertencer à amostra.

Essa técnica de sorteio se torna inviável quando a população é significativamente grande. Neste caso, é necessário o uso de tabelas de números aleatórios ou algoritmos que geram números aleatoriamente.

SAIBA MAIS

Tendo em vista que a amostragem aleatória é vital para a inferência estatística, existem tabelas que são elaboradas e são denominadas Tabelas de Números Aleatórios, construídas de modo que os dez algarismos (0 a 9), são distribuídos ao acaso nas linhas e nas colunas.

Nesta tabela de números aleatórios, os dez algarismos (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) podem ser lidos: isoladamente ou em grupos; em qualquer ordem, como por colunas, em um sentido ou outro, por linhas, diagonalmente, dentre outras formas e podem ser considerados aleatórios. A opção de leitura, porém, deve ser feita antes de iniciado o processo.

Fonte: Bussab e Morettin (2003).

De forma geral, na amostragem casual simples, sorteia-se um elemento da população, sendo que todos os elementos têm a mesma probabilidade de serem selecionados. Repete-se o procedimento até que sejam sorteadas todas as unidades da amostra.

Exemplo:

Para obter uma amostra representativa de 10% de uma população de 100 elementos:

- Numerar os apartamentos de 1 a 100.
- Escrever os números de 1 a 100 em pedaços de papel e colocá-los em uma urna.
- Retirar 10 pedaços de papel um a um da urna, formando a amostra.

Observe que cada elemento tem a mesma probabilidade $\left(\frac{1}{100}\right)$ de ser selecionado.

AMOSTRAGEM SISTEMÁTICA

A amostragem sistemática é utilizada quando os elementos da população se apresentam ordenados, sendo a retirada dos elementos feita, periodicamente, para compor a amostra. O sorteio é feito de forma sistematizada.

De posse de uma listagem de todos os elementos da população, estabelece-se o intervalo de seleção: I = N / n.

Em seguida, sorteia-se um número dentro desse intervalo. Esse será o número de ordem do primeiro sorteado da lista. Os demais sujeitos da amostra serão selecionados utilizando o intervalo I, a partir do primeiro número sorteado.

Exemplo:

Em uma turma de 32 alunos, desejamos obter uma amostra de 5 alunos. Utilizando a técnica de amostragem sistemática, temos que $\boldsymbol{N} = \mathbf{32}$, $\boldsymbol{n} = \mathbf{5}$ e o intervalo de seleção $\boldsymbol{I} = \left(\frac{32}{5}\right) = \mathbf{6,4}$. Para o valor do intervalo, deve ser considerado apenas o valor inteiro, logo, $\boldsymbol{I} = \mathbf{6}$.

O primeiro elemento da amostra deve ser retirado entre os seis primeiros da lista. Para obter os demais elementos da amostra, somamos o intervalo I ao elemento anterior. Se o sorteado for, por exemplo, o número **4**, a amostra será formada pelos sujeitos de números **4, 10, 16, 22 e 28**.

AMOSTRAGEM ESTRATIFICADA

Neste tipo de amostragem, a população deve ser dividida em subgrupos (estratos). Dentro de cada subgrupo, os indivíduos devem ser semelhantes entre si. Assim, pode-se obter uma amostra aleatória de pessoas em cada grupo. Esse processo pode gerar amostras bastante precisas, mas só é viável quando a população pode ser dividida em grupos homogêneos, devendo, na composição da amostra, serem sorteados elementos de todos os estratos.

Quando os estratos possuem, aproximadamente, o mesmo tamanho, sorteia-se igual número de elementos em cada estrato e a amostragem é chamada estratificada uniforme. Caso contrário, sorteia-se, em cada estrato, um número de elementos proporcional ao número de elementos do estrato, chamada amostragem estratificada proporcional.

Exemplo:

Um corretor possui 200 imóveis à disposição, há 120 à venda e 80 para locação. Para extrair uma amostra representativa de 10% dessa população, devem-se identificar seus subconjuntos, sendo neste caso os tipos de investimentos (à venda ou locação).

Tabela 1 - Imóveis de uma imobiliária

INVESTIMENTO	POPULAÇÃO	AMOSTRA (10%)
À venda	120	12
Locação	80	8
Total	**200**	**20**

Fonte: elaborada pelas autoras.

Portanto, a amostra com 20 elementos deve conter 12 imóveis à venda e 8 para locação.

- Deve-se "sortear" 12 elementos entre os 120 imóveis à venda e 8 entre os 80 imóveis para locação, formando a amostra da população.

AMOSTRAGEM POR CONGLOMERADO

Nesta amostragem, a população é dividida em diferentes grupos (conglomerados), extraindo-se uma amostra apenas dos conglomerados selecionados e não de toda a população. O ideal seria que cada conglomerado representasse tanto quanto possível o total da população.

Exemplo:

Para estudar a população de uma cidade, dispondo apenas do mapa dos bairros, deve-se:

- Numerar os bairros e colocar os pedaços de papéis numa urna.
- Retirar um pedaço de papel da urna e realizar o estudo sobre o bairro (conglomerado) selecionado.

É importante saber que a amostra não pode conter vícios, ou seja, não ser tendenciosa. Deve ser selecionada com cuidado, aplicando a técnica de amostragem adequada com tamanho amostral (n) que seja informativo ao que consta na população. O tamanho da população pode ser obtido por fórmulas encontradas facilmente na literatura ou pode ser dado pelo bom senso do pesquisador. O importante é que ele seja representativo da população.

No caso da amostra não ser representativa da população, devemos ter cuidado com o conjunto de dados para que não haja grandes erros de inferência ou então não devemos fazer a inferência.

REFLITA

É um erro básico teorizar antes de ter os dados.

(Sir Arthur Conan Doyle – 1859-1930).

AMOSTRAGEM NÃO ALEATÓRIA

As amostragens não aleatórias procuram gerar amostras que, de alguma forma, representem razoavelmente bem a população de onde foram extraídas. As principais podem ser a amostragem por cotas e por julgamento.

Amostragem por cotas

A amostragem por cotas é semelhante à amostragem estratificada proporcional. A população é separada, dividida em diversos subgrupos. É selecionada uma cota de cada subgrupo proporcional ao seu tamanho. Para compensar a falta de aleatoriedade da seleção, para Barbetta (2014), costuma-se dividir a população em um grande número de subgrupos, como em uma pesquisa socioeconômica na qual a população pode ser dividida por localidade, por nível de instrução, por faixas de renda.

Amostragem por julgamento

Os elementos escolhidos são aqueles julgados como típicos da população que se deseja estudar. Por exemplo em um estudo sobre a produção científica de um departamento de uma instituição de ensino superior, um pesquisador sobre esse assunto pode escolher os departamentos que ele considera serem aqueles que melhor representam a instituição como um todo.

DIMENSIONAMENTO DA AMOSTRA

Para Barbetta (2014), o tamanho de uma amostra é um problema complexo, pois a heterogeneidade e os tipos de parâmetros que queremos estimar são pontos muito importante para determinarmos esse tamanho.

O tamanho da amostra irá depender do nível de confiança estabelecido para pesquisa em função direta do erro amostral, previamente fixado pelo pesquisador. O erro amostral mais utilizado é de 5%. No máximo, é recomendado que esse erro seja de 10%, para que se possa manter uma precisão razoável nos parâmetros estimados pela pesquisa. Quanto menor o erro amostral, maior o tamanho da amostra, implicando em maior precisão nas estimativas populacionais e consequentemente maior custo para execução da pesquisa (FONSECA, 2008).

O erro amostral é a diferença entre uma estatística e o parâmetro que se quer estimar (BARBETTA, 2014).

Para que possamos determinar o tamanho de uma amostra, o pesquisador precisa especificar o erro amostral tolerável, isto é, o quanto ele admite errar na avaliação dos parâmetros de interesse; por exemplo, em uma pesquisa eleitoral, é comum encontrarmos algo que diz que a pesquisa tem um erro de 2%. Isso significa que, quando a pesquisa aponta que determinado candidato tem 45% de intenções de votos, o pesquisador afirma na verdade que a preferência por este candidato, em toda a população de eleitores, é um valor no intervalo de 43% a 47%, isto é, 45% mais ou menos 2%.

Para dimensionamento da amostra, utilizaremos a equação a seguir:

$$\frac{Z^2 \cdot p\hat{} \cdot q\hat{} \cdot N}{d^2 \cdot (N-1) + Z^2 \cdot p\hat{} \cdot q\hat{}}$$

Em que:

N = Tamanho da população.

$p\hat{}$ = estimativa da verdadeira proporção de um dos níveis da variável escolhida.

$q\hat{} = 1 - p$

$Z\ \alpha\ /2$ = abcissa da curva norma padrão, fixado em um nível de confiança

d = erro amostral.

Quando não sabemos o valor de $p\hat{}$ e $q\hat{}$ podemos adotar 0,5 para cada. O valor de z α /2, observamos no Quadro 01.

Quadro 01 – Intervalo de confiança e valores de Z

INTERVALO DE CONFIANÇA	VALOR DE Z
90%	1,645
95%	1,96
98%	2,33
99%	2,575

Fonte: elaborado pelas autoras adaptado de Crespo (2009); Barbetta (2014).

Os valores que devem ser utilizados na equação são os valores de z, por exemplo se tivermos o intervalo de confiança de 90%, utilizaremos 1,645 e assim por diante.

Exemplo: Temos uma população com 43.373 pessoas e desejamos fazer uma pesquisa utilizando os níveis de confiança e os erros amostrais a seguir:

a. Determine o tamanho da amostra com o nível de confiança de 90% e erro amostral de 10%

Temos:

$$\frac{\boldsymbol{Z^2 \,.\, p\hat{} \,.\, q\hat{} \,.\, N}}{\boldsymbol{d^2 \,.\, (N-1) + Z^2 \,.\, p\hat{} \,.\, q\hat{}}}$$

N (número total) = 43.373

p = 0,5

q = 0,5

Z α /2 (para 90% de grau de confiança, encontrado na tabela Z de distribuição normal) = 1,645

d = 0,10

$$\frac{\mathbf{1{,}645^2 \,.\, 0{,}5 \,.\, 0{,}5 \,.\, 43.373}}{\mathbf{0{,}10^2 \,.\, (43.373-1) + 1{,}645^2 \,.\, 0{,}5 \,.\, 0{,}5}}$$

$$\frac{\mathbf{29.342{,}10558}}{\mathbf{434{,}40}} = 67{,}54 \text{ arredondando} = 68 \text{ entrevistados}$$

b. Determine o tamanho da amostra com o nível de confiança de 95% e erro amostral de 5%

Temos:

$$\frac{Z^2 \,.\, p\hat{} \,.\, q\hat{} \,.\, N}{d^2 \,.\, (N - 1) + Z^2 \,.\, p\hat{} \,.\, q\hat{}}$$

N (número total) = 43.373

p = 0,5

q = 0,5

Z α /2 (para 95% de grau de confiança, encontrado na tabela Z de distribuição normal) = 1,96

d = 0,05

$$\frac{1{,}96^2 \,.\, 0{,}5 \,.\, 0{,}5 \,.\, 43.373}{0{,}05^2 \,.\, (43.373 - 1) + 1{,}96^2 \,.\, 0{,}5 \,.\, 0{,}5}$$

$$\frac{41.655{,}43}{109{,}39} = 380{,}79 \text{ arredondando} = 381 \text{ entrevistados}$$

c. Determine o tamanho da amostra com o nível de confiança de 95,5% e erro amostral de 4,5%

N (número total) = 43.373

p = 0,5

q = 0,5

Z α /2 (para 95,5% de grau de confiança, encontrado na tabela Z de distribuição normal) = 2

d = 0,045

$$\frac{2^2 \,.\, 0{,}5 \,.\, 0{,}5 \,.\, 43.373}{0{,}045^2 \,.\, (43.373 - 1) + 2^2 \,.\, 0{,}5 \,.\, 0{,}5}$$

$$\frac{43.373}{88{,}83} = 488{,}26 \text{ arredondando} = 489 \text{ entrevistados}$$

d. Determine o tamanho da amostra com o nível de confiança de 98% e erro amostral de 2%

N (número total) = 43.373

p = 0,5

q = 0,5

Z α /2 (para 98% de grau de confiança, encontrado na tabela Z de distribuição normal) = 2,33

d = 0,02

$$\frac{\mathbf{2{,}33^2 \,.\, 0{,}5 \,.\, 0{,}5 \,.\, 43.373}}{\mathbf{0{,}02^2 \,.\, (43.373 - 1) + 2{,}33^2 \,.\, 0{,}5 \,.\, 0{,}5}}$$

$$\frac{\mathbf{58.869{,}63}}{\mathbf{18{,}71}} = 3.146{,}43 \text{ arredondando} = 3147 \text{ entrevistados}$$

TIPOS DE VARIÁVEIS

VARIÁVEL – é uma característica que possa ser avaliada (ou medida) em cada elemento da população, sob as mesmas condições. Uma variável observada (ou medida) em um elemento da população deve gerar um e apenas um resultado.

Exemplo:

Seja uma população formada pelos funcionários de determinada empresa.

Podemos considerar variáveis como: tempo de serviço, salário, estado civil, idade, sexo, escolaridade, inteligência, peso, estatura, autoestima, grau de satisfação com o emprego, autoritarismo, religiosidade, etc.

Como medir essas características? Devemos fixar uma unidade de medida (kg, cm, anos completos,...) ou definir atributos (casado, solteiro, masculino, feminino, forte, fraco...).

Para descrever o grupo ou a amostra, há a necessidade de identificar o tipo dessa variável para definir a melhor metodologia de trabalho. As variáveis podem ser: Quantitativas ou Qualitativas.

Variáveis Qualitativas ou Categóricas - são variáveis que assumem como possíveis valores atributos ou qualidades. Se tais variáveis assumem uma ordenação natural, são chamadas de **qualitativas ordinais** (ex.: grau de escolaridade, classe social); caso contrário, são chamadas **qualitativas nominais** (ex.: cor dos olhos, campo de estudo).

Variáveis Quantitativas - são variáveis que assumem como possíveis valores os números. Quando estas variáveis são resultantes de contagens, são chamadas de **quantitativas discretas** (ex.: quantidade de irmãos, de defeitos num carro novo); caso assumam qualquer valor em intervalos dos números reais, são chamadas **quantitativas contínuas** (ex.: altura, peso, velocidade).

Exercício:

Classificar as seguintes variáveis:

a. tempo de vida de uma placa-mãe.

b. tipo sanguíneo.

c. raça.

d. produção de amortecedores de uma indústria em um período de dois minutos.

e. produção de mel das caixas de um apiário.

f. religião.

g. estado civil.

h. número de pessoas na fila de um banco.

i. número de produtores associados a uma cooperativa.

Respostas:

a. Quantitativa Contínua.

b. Qualitativa Nominal.

c. Qualitativa Nominal.

d. Quantitativa Discreta.

e. Quantitativa Contínua.

f. Qualitativa Nominal.

g. Qualitativa Nominal.

h. Quantitativa Discreta.

i. Quantitativa Discreta.

FASES DO MÉTODO ESTATÍSTICO

Após a definição do problema a ser estudado, a marcha natural do processo de pesquisa é a seguinte:

- Planejamento.
- Coleta de dados.
- Crítica, organização e sumarização dos dados.
- Apresentação dos dados.
- Análise e interpretação.

PLANEJAMENTO – devem-se estabelecer com clareza os objetivos e os procedimentos a serem adotados. Nesta fase, define-se a maneira de coletar os dados (entrevista, questionário ou simples medição), determinando, também, o tamanho necessário para a amostra e a maneira mais indicada para selecioná-la.

COLETA DE DADOS – de acordo com a finalidade da pesquisa, a coleta pode ser:

- contínua - obtendo-se registros de fenômenos de interesse do administrador.
- periódica - quando se necessita de avaliações sistemáticas. Um exemplo bem característico é o censo realizado pelos governos em períodos pré-estabelecidos.
- ocasional - quando existe um interesse momentâneo em determinado fenômeno.

O conjunto de dados coletados dá origem às SÉRIES ESTATÍSTICAS. Didaticamente, podemos caracterizá-las como:

- HISTÓRICAS ou CRONOLÓGICAS: quando o fenômeno é estudado ao longo do tempo, em determinado local.
- GEOGRÁFICAS ou TERRITORIAIS: quando se observam valores da variável em determinado momento, segundo sua localização.
- ESPECÍFICAS ou CATEGÓRICAS: quando a variável é observada em determinado tempo e local, discriminada por especificações ou categorias.

CRÍTICA, ORGANIZAÇÃO E SUMARIZAÇÃO - têm a finalidade de eliminar erros. Neste processo, procede-se a uma revisão crítica dos dados, retirando os valores estranhos que podem ocorrer tanto por erro de quem coletou os dados ou de quem foi abordado na pesquisa. Para um melhor entendimento, diante de grande quantidade de dados, é adequado que se faça uma compilação deles para sua apresentação.

APRESENTAÇÃO DOS DADOS – será feita por meio de TABELAS e GRÁFICOS. As tabelas são mais ricas em detalhes e em precisão. Os gráficos proporcionam maior rapidez de interpretação, embora percam exatidão em detalhes.

ANÁLISE E INTERPRETAÇÃO - têm como objetivo a determinação de medidas estatísticas que, como já vimos, têm a finalidade de descrever de forma prática e objetiva as características gerais de uma população. Determinadas as medidas estatísticas básicas, a análise desejada poderá ter sequência já no campo da Estatística Inferencial, baseada sempre em raciocínios probabilísticos.

TABELAS E GRÁFICOS

O objetivo da utilização de tabelas e gráficos é transformar dados em informações que permitam a fácil visualização e interpretação da nossa pesquisa. Também servem para verificar a existência de algum padrão para comparar esse padrão com outros resultados ou, ainda, para julgar sua adequação a alguma teoria.

As tabelas são quadros em que serão dispostas as informações por alguma categoria pelo cálculo de alguma frequência. Devem ter as laterais abertas, ou seja, sem bordas, e devem ainda ter um título explicativo e localizado acima delas, precedido da palavra Tabela e seguido de sua numeração. Veja o exemplo abaixo:

Tabela 2 - Número de computadores pessoais instalados no Brasil e percentual sobre total de habitantes de 2000 a 2004

ANO	NÚMERO DE COMPUTADORES (MILHÕES)	PERCENTUAL DE COMPUTADORES POR TOTAL DE HABITANTES
2000	5,0	3,0
2001	8,5	4,0
2002	13,0	7,3
2003	24,0	13,1
2004	30,0	16,1

Fonte: elaborada pelas autoras.

Um gráfico é uma figura utilizada na Estatística para representar um fenômeno. Ele deve refletir padrões gerais e específicos do conjunto de dados. Apesar de comum, a utilização dos gráficos fornece menos detalhes do conjunto de dados em relação às tabelas, entretanto, é um meio rápido e prático para visualização dos dados.

Um gráfico dispõe tendências, os valores mínimos e máximos, as variações dos dados e, também, as ordens de grandezas dos fenômenos que estão sendo observados. Todo gráfico deve visar clareza e objetividade, além de ser fiel às informações pertinentes ao conjunto original de dados.

Existem diversos tipos de gráficos e de tabelas disponíveis e que podemos utilizar na apresentação dos dados de nossas pesquisas. Na próxima unidade, você verá esses exemplos associados a vários tipos e formatos. Veja o exemplo a seguir:

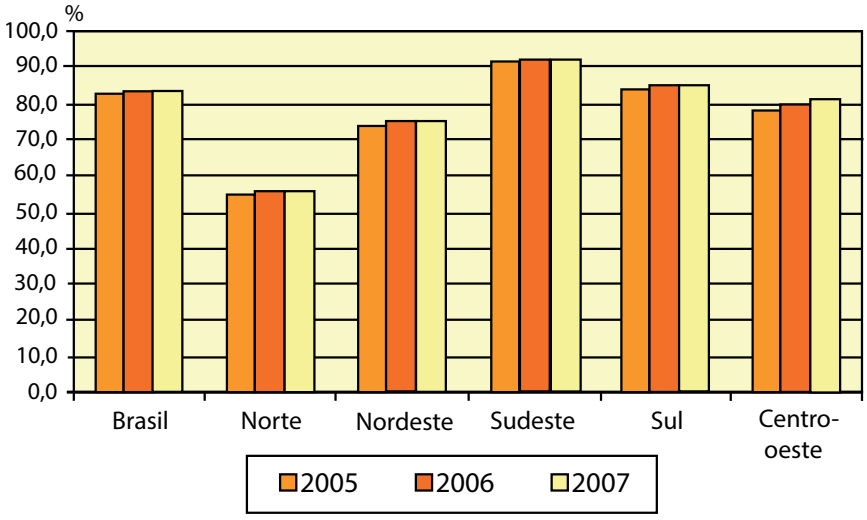


Gráfico 1 - Percentual de domicílios atendidos por rede geral de abastecimento de água no total de domicílios particulares permanentes (%)

Fonte: IBGE - Pesquisa Nacional por Amostra de Domicílios (2006-2007, online).

Algumas perguntas podem ser feitas ao se optar pela utilização de um gráfico, como:

- O gráfico é uma opção que realmente demonstra o que quero mostrar na pesquisa?
- Qual é o tipo de gráfico adequado para os dados da pesquisa?
- Como este gráfico deve ser mostrado ao público?
- Devo usar somente um gráfico para visualização dos dados da pesquisa?

CONSIDERAÇÕES FINAIS

A Estatística é utilizada para coleta, organização, descrição e análise de informações obtidas em uma pesquisa, sendo que a estatística descritiva é utilizada para a descrição dos dados. O objetivo principal é transformar os dados brutos em informações.

Nesta unidade, você viu os principais conceitos utilizados dentro da estatística descritiva, tipos de amostras e a importância da utilização de gráficos e de tabelas como forma de apresentação dos dados. Dos conceitos abordados, podemos destacar o censo, que é o processo que consiste no exame de todos os elementos da população e cujas medidas são chamadas de parâmetros. Entretanto, se utilizamos uma parte dessa população, não temos um parâmetro, e sim uma estatística ou um estimador; portanto, um estimador é uma medida tomada em uma parte dessa população, mas não nela toda, embora este estimador represente o parâmetro.

Também vimos os conceitos de População e Amostra. População, no sentido estatístico, pode ser definida como um conjunto de elementos que possuem alguma característica em comum. Como na maioria das vezes é difícil ou custoso trabalharmos com população, utilizamos uma parte dela. A esta parte chamamos de amostra. Essa amostra deve, porém, representar a população, ou seja, deve ter as mesmas características da população que se irá amostrar.

Para que isso seja conseguido, o processo de coleta de uma amostra, também conhecido como amostragem, deve ser feito de forma casual ou aleatória. Existem algumas maneiras de se fazer uma amostragem e, para cada situação, existe uma maneira ideal. Nesta unidade, foram discutidos os principais tipos de amostras utilizadas nas pesquisas, sendo que a escolha deve ser feita de modo que as amostras representem de fato a população e de forma que sejam não tendenciosas.

Após a coleta da amostra, é necessário descrever os dados. Para isso, primeiramente, devemos saber com quais tipos de variáveis estamos trabalhando, para, assim, escolhermos qual é a melhor maneira de apresentar a pesquisa.

Finalizando, é importante que a apresentação dos dados seja feita de forma precisa. As duas formas vistas nesta unidade foram as tabelas e os gráficos, e o uso correto das formas de apresentação dos dados é fundamental para o sucesso da pesquisa.

ATIVIDADES

1. Defina Estatística, Estatística Descritiva e Estatística Inferencial.

2. Apresente os conceitos para os termos abaixo relacionados e dê um exemplo para cada um deles:

- População.
- Amostra.
- Censo.
- Estimação.
- Variáveis.

3. Explique os principais tipos de amostras.

4. Comente as vantagens de apresentar resultados de pesquisa por meio de tabelas gráficos.

5. Identifique a população em estudo e o tipo de amostragem a ser utilizado em cada alternativa:

a. Uma empresa tem 3.414 empregados repartidos nos seguintes departamentos: Administração (914), Transporte (348), Produção (1401) e Outros (751). Deseja-se extrair uma amostra entre os empregados para verificar o grau de satisfação em relação à qualidade da refeição servida no refeitório.

b. Um cabo eleitoral escreve o nome de cada senador do Brasil em cartões separados, mistura e extrai 10 nomes.

c. Um administrador hospitalar faz uma pesquisa com as pessoas que estão na fila de espera para serem atendidas pelo sistema SUS, entrevistando uma a cada 10 pessoas da fila.

d. Para dar a porcentagem de defeitos das 3000 peças fabricadas por dia, a cada 6 peças, uma é retirada para teste.

SOBRE O IBGE - Censo

Para entender como funciona o IBGE, que tal conhecer um pouco sobre sua organização? Você vai ver que o IBGE está presente no Brasil inteiro!

O IBGE é composto por quatro diretorias (Executiva, de Pesquisas, de Geociências e de Informática), um Centro de Documentação e Disseminação de Informações e pela Escola Nacional de Ciências Estatísticas (ENCE). Essas diretorias possuem funções específicas e estão localizadas no município do Rio de Janeiro.

Para que nossas atividades possam cobrir todo o território nacional, possuímos uma rede nacional de pesquisa e disseminação, composta por:

- 27 Unidades Estaduais (26 nas capitais dos estados e 1 no Distrito Federal)
- 27 Supervisões de Documentação e Disseminação de Informações (26 nas capitais e 1 no Distrito Federal)
- 584 Agências de Coleta de dados nos principais municípios brasileiros.

O IBGE mantém, ainda, a Reserva Ecológica do Roncador, situada a 35 quilômetros ao sul de Brasília. O Censo Demográfico é uma pesquisa realizada pelo IBGE a cada dez anos. Através dele, reunimos informações sobre toda a população brasileira.

Nosso primeiro Censo aconteceu em 1872 e recebeu o nome de Recenseamento da População do Império do Brasil. O mais recente foi o Censo 2010, cujos resultados você pode buscar no site do IBGE. Antes dele, o IBGE realizou o Censo 2000.

No Censo, os pesquisadores do IBGE visitam todos os domicílios do país para aplicar um questionário. Depois de percorrer todos os cantos do Brasil, indo de casa em casa, os pesquisadores organizam e analisam as informações coletadas nos questionários. Em seguida, divulgam os resultados em uma série de publicações sobre os temas estudados.

Os resultados do Censo Demográfico são importantes para a sociedade ter informações atualizadas sobre a população e para o governo planejar suas ações de forma mais adequada.

No IBGE, muitas pessoas trabalham coletando, analisando e armazenando informações sobre a população e as atividades econômicas de nosso país. Em nosso instituto existem duas grandes áreas de pesquisa: a de informações sociais, demográficas e econômicas: para produzir estas informações, são realizadas pesquisas como o Censo Demográfico, o Censo Agropecuário e os Índices de Preços; a de informações geográficas: onde são feitos os mapas, os estudos de recursos naturais e de meio ambiente.

Com estas informações, o governo pode saber quem está estudando, onde são necessárias mais escolas, onde o número de lojas e fábricas é maior, onde há mais empregos, o que é produzido em uma determinada região e uma série de outras coisas.

Fonte: adaptado de IBGE (2017, on-line).

MATERIAL COMPLEMENTAR

LIVRO

A Estatística Fácil

Antônio Arnot Crespo

Editora: Saraiva

Sinopse: Este livro apresenta todos os tópicos exigidos pelo programa estabelecido para os cursos profissionalizantes da rede de ensino particular e oficial, de forma acessível ao aluno, dentro de um esquema de ensino objetivo e prático. O estudo é complementado por exercícios em abundância com situações práticas.
Trabalha com estatística descritiva, probabilidades, distribuições de probabilidades, correlação e regressão linear, de forma prática e fácil linguagem.

REFERÊNCIAS

BARBETTA, P. A. **Estatística aplicada às Ciências Sociais**. 9. ed. Florianópolis – SC: UFSC, 2014.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. São Paulo: Saraiva, 2003.

CRESPO, A. A. **Estatística Fácil**. 19. ed. São Paulo: Saraiva, 2009.

FONSECA, J. S. da; MARTINS, G. de A. **Curso de Estatística**. 6. ed. 11. Reimp. São Paulo: Atlas, 2008.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Pesquisa Nacional por Amostra de Domicílios**. 2006-2007. Disponível em: <http://www.ibge.gov.br/home/estatistica/pesquisas/pesquisa_resultados.php?id_pesquisa=40>. Acesso em: 18 abr. 2017.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Pesquisa Nacional por Amostra de Domicílios**. 2007-2009. Disponível em: <http://www.ibge.gov.br/home/estatistica/pesquisas/pesquisa_resultados.php?id_pesquisa=40>. Acesso em: 18 abr. 2017.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Sobre o IBGE - Censo**. Disponível em: <http://7a12.ibge.gov.br/sobre-o-ibge/nossas-pesquisas.html> Acesso em: 18 abr. 2017.

RAO, C. R. Statistics: a technology for the millennium Internal. **J. Math. & Statist. Sci**, v. 8, n. 1, p. 5-25, jun. 1999.

TOREZANI, W. **Apostila de Estatística I**. Vila Velha: Faculdade Univila, 2004.

GABARITO

1. A Estatística pode ser definida como uma parte da matemática que se preocupa em coletar, organizar, descrever, analisar e interpretar um conjunto de dados.

   A estatística descritiva se preocupa em descrever os dados.

   A estatística inferencial se preocupa com a análise dos dados e sua interpretação. Ela analisa os dados com base na amostra e, então, estende as conclusões desta amostra à população.

2. População – conjunto de elementos que possuem alguma característica em comum.

   Amostra – parte da população, devendo ser representativa dela.

   Censo – levantamento de dados de toda uma população.

   Estimação – obtenção de valores de uma amostra.

   Variáveis – características tomadas em uma população ou amostra, por exemplo: sexo, idade, região de procedência, peso, etc.

3. Amostra Casual Simples - é aquela em que todos os elementos da população têm igual probabilidade de pertencer à amostra. Pode ser obtida sorteando os elementos a partir da população de estudo.

   Amostra Sistemática - é uma forma simplificada da amostragem casual simples, podendo ser utilizada quando os elementos da população se apresentam ordenados, sendo a retirada dos elementos para compor a amostra feita com certa periodicidade.

   Amostra Estratificada – é uma amostra em que a população é separada em grupos ou estratos e, dentro de cada estrato, os indivíduos são sorteados, devendo ser semelhantes entre si dentro de cada estrato.

   Amostra por Conglomerado – é uma amostra em que a população é dividida em diferentes conglomerados, extraindo-se uma amostra apenas dos conglomerados selecionados, e não de toda a população.

4. Representar os dados por meio de gráficos e tabelas: os dados são apresentados de forma resumida, em que há uma visualização rápida e fácil deles para o público. Há um entendimento melhor dos dados, ficando fácil sabermos o que está ocorrendo com os dados coletados.

5.
   a) Funcionário da empresa, amostragem estratificada.
   b) Senadores do Brasil, amostragem aleatória simples.
   c) Pessoas na fila de atendimento, amostragem sistemática.
   d) Peças fabricadas, amostragem sistemática.

Professora Me. Ivnna Gurniski de Oliveira
Professora Me. Renata Cristina de Souza Chatalov

UNIDADE II

TABELAS E GRÁFICOS

Objetivos de Aprendizagem

- Entender a importância dos gráficos e das tabelas.
- Aprender a construir gráficos e tabelas para variáveis qualitativas.
- Aprender a construir gráficos e tabelas para variáveis quantitativas.

Plano de Estudo

A seguir, apresentam-se os tópicos que você estudará nesta unidade:

- Tabelas
- Gráficos

Unicesumar

INTRODUÇÃO

Em uma pesquisa, geralmente, os dados são descritos e analisados com auxílio de técnicas estatísticas. As pesquisas precisam da Estatística para alcançar seus objetivos, principalmente, quando envolvem grande quantidade de informações que precisam ser resumidas.

A organização dos dados em tabelas de frequências nos proporciona um meio eficaz de estudo do comportamento de características de interesse. Muitas vezes, a informação contida nas tabelas pode ser mais facilmente visualizada por meio de gráficos. Como exemplo, podemos citar os meios de comunicação que nos apresentam, diariamente, gráficos das mais variadas formas para auxiliar na apresentação das informações. Também os órgãos públicos e empresas se municiam de gráficos e de tabelas em seus documentos internos e relatórios de atividades e de desempenho. Graças ao aumento dos recursos gráficos, sua construção tem sido cada vez mais simplificada por meio do uso de programas computacionais, existindo hoje uma infinidade de tipos de gráficos que podem ser utilizados. É importante salientar que existem diversas formas de gráficos e de tabelas e a escolha de uma ou outra forma depende da característica com a qual estamos trabalhando.

Diante disso, nesta unidade, temos o objetivo de ensiná-lo(a) a construir as tabelas e os principais tipos de gráficos. Para essa construção, há necessidade de separação das variáveis. As variáveis em estudos estatísticos são valores que assumem determinadas características dentro de uma pesquisa e podem ser classificadas em: qualitativas ou quantitativas.

Para alguns tipos de gráficos, podem ser utilizados tanto para uma, quanto para outra variável, entretanto, existem alguns tipos que são específicos para variáveis qualitativas ou quantitativas, portanto, é interessante conhecer o tipo adequado para cada caso.

É importante desenvolver tanto a habilidade de construir tabelas e gráficos, quanto a de fazer uma leitura adequada deles.

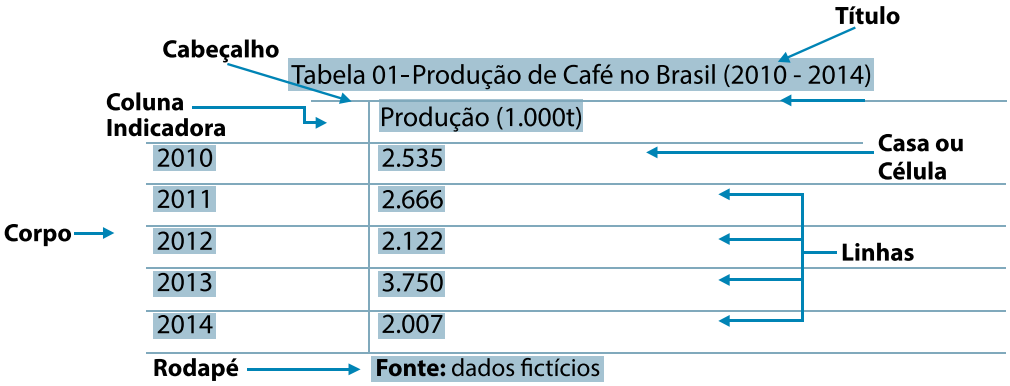
TABELAS

Quando retiramos as informações da pesquisa, temos em mãos os dados brutos. A ideia é transformar os dados brutos em informações para que seu entendimento e visualização se tornem mais simples e rápidos.

Existem normas nacionais para a organização de tabelas, ditadas pela Associação Brasileira de Normas e Técnicas (ABNT), que não serão abordadas aqui, mas convém saber que as tabelas são formadas por título, cabeçalho, corpo e fonte:

- **Título**: precede a tabela e resume o dado em estudo (O quê? Onde? Quando?). Deve vir precedido da palavra tabela e de sua numeração. As tabelas devem ser numeradas em ordem crescente à maneira que aparecem no texto, ex.: Tabela 1; Tabela 2 e assim por diante.
- **Cabeçalho**: especifica o conteúdo de cada coluna.
- **Corpo**: formado por linhas e colunas contendo os dados.
- **Fonte**: na parte inferior, informa-se a fonte da coleta de dados ou o autor. A fonte cita o informante, caracterizando a confiabilidade dos dados.

As tabelas deverão ser fechadas com traços horizontais nas bordas superior e inferior, enquanto que nas bordas esquerda e direita não. Dentro das tabelas pode haver traços verticais na separação das colunas no corpo da tabela ou entre as linhas. É conveniente, também, que o número de casas decimais seja padronizado. Vejamos um exemplo a seguir:



Fonte: elaborada pelas autoras.

Uma tabela contém as categorias da variável estudada e suas respectivas frequências. Essas frequências podem ser:

- Absoluta (*Fi*), dada pela contagem do número de ocorrências de cada categoria.
- Relativa (*Fr*), dada pela frequência absoluta em relação ao total de elementos (n) a serem estudados, ou seja, $Fr = \frac{Fi}{n}$.
- Percentual (*Fr*%), dada pela frequência relativa multiplicada por cem, ou seja, $Fr\% = \left(\frac{Fi}{n}\right) \text{X } 100$.
- Acumulada (*Fac*), dada pela soma das frequências de todas as linhas anteriores até a classe atual. Pode ser: frequência acumulada absoluta, frequência acumulada relativa ou frequência acumulada percentual.

TABELAS PARA VARIÁVEIS QUALITATIVAS

Variáveis qualitativas apresentam-se em categorias e, portanto, a representação tabular deve ser feita por meio das frequências referentes a cada uma das categorias. Podem se apresentar de forma simples (com apenas uma variável) ou conjunta (com duas ou mais variáveis).

Exemplo:

Tabela 02 - Distribuição de frequências de indivíduos que acessam o site quanto ao sexo

SEXO	NÚMERO DE CLIENTES (*Fi*)	*Fr*	*Fr%*	*Fac*
Masculino	7	7/11 = 0,636	63,6	7
Feminino	4	4/11 = 0,364	36,4	11
Total	11	1,0	100	-

Fonte: dados fictícios elaborados pelas autoras.

Outro exemplo:

Tabela 03 - Grupo de atributos que mais valorizam os imóveis

GRUPO DE ATRIBUTOS	PORCENTAGEM (%)
Localização	27,47
Conforto	22,71
Segurança	20,51
Incorporação	17,58
Lazer	11,73

Fonte: dados fictícios elaborados pelas autoras.

REFLITA

Existem diversos tipos de variáveis demonstradas em tabelas. O formato das tabelas é sempre o mesmo, podendo apresentar uma única frequência como a frequência absoluta ou a porcentagem ou, ainda, várias frequências combinadas.

As tabelas também podem se apresentar mostrando a combinação de algumas variáveis conjuntas. Observe que, na tabela a seguir, foram tomadas 2 variáveis: Região e Ano.

Tabela 4 - Custo médio (R$/m²) das áreas geográficas de um dado país

REGIÃO	2005	2006	2007	2008	2009
Norte	870,2	893,4	921,0	923,1	925,7
Nordeste	574,4	573,6	573,8	571,1	582,0
Sudeste	659,2	670,4	671,5	680,9	681,4
Sul	1094,3	1112,0	114,6	1240,3	1500,4
Centro-Oeste	897,5	902,4	909,5	1002,1	1004,9

Fonte: dados fictícios elaborados pelas autoras.

TABELAS PARA VARIÁVEIS QUANTITATIVAS

Para variáveis quantitativas contínuas ou discretas, com elevado número de valores diferentes, a **distribuição de frequências** apropriada é apresentar os dados em classes de valores.

Para esse procedimento, primeiramente, precisamos determinar o número de classes. Uma classe é uma linha da distribuição de frequências.

Quando temos um conjunto de dados com muitos valores, é recomendado que seja colocado em uma tabela de frequências. A seguir, aprenderemos como fazer uma distribuição de frequências de forma adequada.

Tabela Primitiva

Vamos considerar a forma pela qual podemos descrever os dados estatísticos resultantes de variáveis quantitativas, como por exemplo, o peso dos colaboradores de um setor, a estatura de um grupo de pessoas, as notas obtidas por alunos de uma turma, dentre outros.

Supondo que resolvemos fazer uma coleta de dados referentes às estaturas de 40 colaboradores de uma empresa, que resultou nos dados a seguir:

ESTATURAS DE 40 COLABORADORES DE UMA EMPRESA									
166	160	161	150	162	160	165	167	164	160
162	168	161	163	156	173	160	155	164	168
155	152	163	160	155	155	169	151	170	164
154	161	156	172	153	157	156	158	158	161

Fonte: dados fictícios elaborados pelas autoras.

A esse tipo de tabela, na qual os valores e/ou elementos não foram trabalhados, nem numericamente organizados, denominamos de tabela primitiva.

Rol

A partir dos resultados anteriores, composta pelos 40 colaboradores de uma empresa (tabela primitiva), é difícil nós averiguarmos em torno de que valor tende a se concentrar as estaturas, qual a menor ou qual a maior estatura ou, ainda, quantos alunos se acham abaixo ou acima de uma dada estatura (CRESPO, 2009).

Depois que nós conhecemos os valores de certa variável, fica difícil termos uma ideia exata do comportamento do grupo como um todo, por meio dos dados não ordenados. A maneira mais simples de organizar os dados é por meio de certa ordenação (crescente ou decrescente). A tabela obtida mediante a ordenação dos dados recebe o nome de rol.

O Rol é definido como uma organização dos dados, que pode ser em ordem crescente ou decrescente. Para construirmos as tabelas de distribuição de frequências, fica mais fácil organizarmos os dados de forma crescente. Vamos ver como ficou os dados organizados em Rol a seguir.

ESTATURAS DE 40 COLABORADORES DE UMA ORGANIZAÇÃO									
150	154	155	157	160	161	162	164	166	169
151	155	156	158	160	161	162	164	167	170
152	155	156	158	160	161	163	164	168	172
153	155	156	160	160	161	163	165	168	173

Fonte: dados fictícios elaborados pelas autoras.

Observem que é bem melhor para trabalharmos com os dados organizados, não acham? A partir disso, já conseguimos visualizar algumas coisas quanto à estatura dos colaboradores, por exemplo, que o colaborador com a menor estatura tem 150 cm e a maior estatura 173 cm.

Construindo uma Distribuição de Frequência

a) Distribuição de Frequência sem intervalo de classes

Uma distribuição de frequências sem intervalo de classes consiste na simples condensação dos dados de acordo as repetições de seus valores. Vejamos um exemplo a seguir, referente às idades de 20 colaboradores de uma organização:

20	18	18	19	21	25	28	28	28	28
28	21	25	29	29	29	18	18	25	21

Para agrupá-los em uma tabela de frequências sem intervalo de classes, vamos organizar os dados em Rol (crescente).

18	18	18	18	19	20	21	21	21	25
25	25	28	28	28	28	28	29	29	29

Agora vamos construir a tabela de frequências; basta colocarmos na coluna Idades as referências às idades, e nas colunas Frequências quantas vezes as idades se repetiram, vejam a seguir:

Tabela 5 - Idades de colaboradores de uma empresa

IDADES	FREQUÊNCIA (Fi)
18	4
19	1
20	1
21	3
25	3
28	5
29	3
Total	**20**

Fonte: dados fictícios elaborados pelas autoras.

Vejam como é simples, construímos a tabela e distribuímos suas frequências, ou seja, quantas vezes a idade se repetiu.

b) Distribuição de Frequência com intervalo de classes

Vamos construir uma tabela de frequências com intervalo de classes com os dados referentes à Estatura de 40 alunos colaboradores; a variável em questão, estatura, para Crespo (2009), será analisada e estudada mais facilmente quando colocarmos valores ordenados em uma coluna e colocarmos, ao lado de cada valor, o número de vezes que aparece repetido. Para isso, vamos construir uma distribuição de frequências.

Segue os dados em rol seguir:

ESTATURAS DE 40 COLABORADORES DE UMA ORGANIZAÇÃO									
150	154	155	157	160	161	162	164	166	169
151	155	156	158	160	161	162	164	167	170
152	155	156	158	160	161	163	164	168	172
153	155	156	160	160	161	163	165	168	173

Fonte: dados fictícios elaborados pelas autoras.

Podemos chamar de frequência o número de colaboradores que fica relacionado a um determinado valor da variável, isto é, quantas vezes determinado valor se repetiu (CRESPO, 2009). Podemos obter, então, uma tabela que receberá o nome de distribuição de frequência. Primeiro apresentamos a vocês esses dados em uma distribuição de frequências sem intervalos de classes, analisem:

Tabela 06 - Estaturas de 40 colaboradores de uma organização

ESTATURAS (CM)	FREQUÊNCIA
150	1
151	1
152	1
153	1
154	1
155	4
156	3

ESTATURAS (CM)	FREQUÊNCIA
157	1
158	2
160	5
161	4
162	2
163	2
164	3
165	1
166	1
167	1
168	2
169	1
170	1
172	1
173	1
Total	**40**

Fonte: dados fictícios elaborados pelas autoras.

O que podemos concluir por meio dessa tabela?

- Vejam que nesse caso, quando temos muitos valores e resolvemos distribuir em uma tabela de frequências sem intervalo de classes, a tabela fica muito grande.
- As frequências (Fi) são as repetições dos valores no rol, por exemplo, a altura 150 cm se repetiu uma vez, portanto sua frequência é 1; enquanto a altura 151, também se repetiu uma vez, e tem sua frequência igual a 1; e assim sucessivamente, até completarmos a tabela.

Entretanto, essa tabela, feita essa distribuição de frequências (sem o intervalo de classes), está de fácil leitura? Poderia ser melhorada?

- Essa tabela tem 22 classes, ou seja, 22 linhas (fora o cabeçalho e o total); poderíamos condensar esses valores, construindo uma tabela de distribuição de frequências com intervalo de classes.

Vamos aprender a construir essa tabela? Vejamos os passos para sua elaboração.

1º Passo: Amplitude Total (AT): consiste na diferença entre o maior valor do conjunto de dados e o menor valor do conjunto de dados, portanto:

$AT = X_{max} - X_{mín}$

No nosso exemplo:

$AT = 173 - 150 = 23$

(Vamos guardar esse valor, que também será usado no 3º passo).

2º passo: determinar o número de classes (k): não há regras absolutas para a escolha do número de classes (***k***), geralmente entre 5 e 20 classes serão satisfatórias para a maior parte dos conjuntos de dados. Uma regra prática razoável é:

$$K \approx \sqrt{\textit{número de observações}}$$

$$\mathbf{K} = \sqrt{n}$$

Usar um número pequeno de classes poderia concentrar a maioria das observações em uma ou duas classes. Se for usado um número grande de classes, muitas delas terão frequências iguais à zero.

No nosso exemplo, $K = \sqrt{40} = 6{,}32...$ = arredondando = 6.

Portanto, nossa tabela terá 6 classes.

3º passo: determinar a amplitude total: consiste na diferença entre o limite inferior e o limite superior, sendo dado pela equação:

$$AC\ (h) = \frac{AT}{k}$$

Em que:

AT = Amplitude Total

K = número de classes

Portanto, AC ou h é a divisão entre a amplitude total (AT) pelo número das classes (K).

No nosso exemplo:

$h = \frac{At}{K} = \frac{23}{6} = 3{,}83$ = arredondando 4.

Portanto, a diferença entre o limite inferior (li) e o limite superior (ls) da classe será igual a 4.

4º passo: construir a Tabela de frequências

O menor valor da classe é denominado limite inferior (L_i) e o maior valor da classe, limite superior (L_s).

a. Para obtenção da primeira classe, tomar como L_i o menor valor. Ao L_i, somar o valor da AC (ou h) e assim se obtém o L_s.

b. Para construção da segunda classe, repetir o L_s da primeira classe, sendo que este na segunda classe passa a ser o L_i. A este valor adicionar o valor de AC (ou h) e se obtém o L_s.

c. Para a terceira classe, repetir o procedimento. O L_s da segunda classe é repetido na terceira classe e se torna o L_i. A esse L_i, adicionar o valor de AC e assim se obtém o L_s. Esse definido procedimento deve ser repetido até que se obtenha o número de classes. O L_s da última classe deve, obrigatoriamente, ser igual ou ultrapassar o maior valor do conjunto de dados.

Obs.: Não se esqueça de obedecer à simbologia do limite entre as classes.

São os valores extremos de cada intervalo de classe representado por:

L_i = limite inferior e L_s = limite superior, em que o limite inferior é o menor número da classe e o limite superior é o maior número da classe. No nosso exemplo, o L_i =150 e o L_s = 173.

Notações entre os limites de classes:

- L_i |--- L_s: o limite inferior está incluído na contagem da frequência absoluta da classe e o limite superior não.
- L_i ---| L_s: o limite superior está incluído na contagem da frequência absoluta da classe e o limite inferior não.
- L_i |---| L_s: os limites inferiores e superiores estão incluídos na contagem da frequência absoluta da classe.
- L_i --- L_s: os limites inferiores e superiores não estão incluídos na contagem da frequência absoluta da classe.

Agora que já sabemos todas as regras, vamos construir a tabela de frequências seguindo os passos anteriores.

Com os dados organizados em rol (apresentados novamente para facilitar na elaboração da tabela).

ESTATURAS DE 40 COLABORADORES DE UMA ORGANIZAÇÃO									
150	154	155	157	160	161	162	164	166	169
151	155	156	158	160	161	162	164	167	170
152	155	156	158	160	161	163	164	168	172
153	155	156	160	160	161	163	165	168	173

Fonte: dados fictícios elaborados pelas autoras.

Tabela 07 - Estaturas de 40 colaboradores de uma organização

ESTATURAS	FREQUÊNCIA (Fi)
150 |----- 154	4
154 |----- 158	9
158 |----- 162	11
162 |----- 166	8
166 |----- 170	5
170 |----- 174	3
Total	**40**

Fonte: dados fictícios elaborados pelas autoras.

Após, contamos as frequências (fi), quantos valores foram determinados para cada classe da tabela.

Além disso, temos as colunas complementares da tabela de frequência, que são:

a. Ponto médio das Classes

O ponto médio de uma classe é dado por:

$$x_i = \frac{L_i + L_s}{2}$$

Em que:

x_i = Ponto médio

L_i = limite inferior

L_s = limite superior

Veremos, na próxima unidade, que o ponto médio de uma classe é utilizado para calcular a média aritmética ponderada para um conjunto de dados agrupados.

Obs.: Só existe limite de classes em tabelas de frequências com intervalo de classes.

b. Frequência relativa (fr ou fr %): é a proporção dos dados que aparece em cada classe, dada pela equação a seguir:

$$fr(\%) = \frac{fi}{n} \cdot 100$$

Em que:

$fr(\%)$ = Frequência relativa em percentual

fi = Frequência da classe

n = número total de elementos (ou somatória da frequência)

c. Frequência acumulada (*Fac*): é a representação da frequência absoluta (*fi*) de forma acumulada.

d. Frequência Relativa Acumulada (*FRac*): é dada pela divisão da frequência acumulada pelo número total de elementos da série em porcentagem (%).

$$FRac(\%) = \frac{Fac}{n} \cdot 100$$

Em que:

$FRac(\%)$ = Frequência relativa acumulada em percentual

Fac = Frequência acumulada

n = número total de elementos (ou somatória da frequência)

Vamos resolver as colunas complementares na tabela a seguir:

Tabela 08 - Estaturas de 40 colaboradores de uma organização

IDADES	(Fi)	FR (%)	Fac	FRac(%)	Xi
150 |----- 154	4	(4/40)*100 = 10	4	(4/40)*100 = 10	(150+154)/2= 152
154 |----- 158	9	(9/40)*100 = 22,5	13	(13/40)*100 = 32,5	(154+158)/2= 156
158 |----- 162	11	(11/40)*100 = 27,5	24	(24/40)*100 = 60	(158+162)/2= 160
162 |----- 166	8	(8/40)*100 = 20	32	(32/40)*100 = 80	(162+166)/2= 164
166 |----- 170	5	(5/40)*100 = 12,5	37	(37/40)*100 = 92,5	(166+170)/2= 168
170 |----- 174	3	(3/40)*100 = 7,5	40	(40/40)*100 = 100	(170+174)/2= 172
Total	**40**	**100**			**972**

Fonte: dados fictícios elaborados pelas autoras.

Para melhor apresentar nossa tabela, vamos apresentá-la somente com os resultados, sem os cálculos dentro dela, observem:

Tabela 09 - Estaturas de 40 colaboradores de uma organização

IDADES	(Fi)	FR (%)	Fac	FRac(%)	Xi
150 |----- 154	4	10	4	10	152
154 |----- 158	9	22,5	13	32,5	156
158 |----- 162	11	27,5	24	60	160
162 |----- 166	8	20	32	80	164
166 |----- 170	5	12,5	37	92,5	168
170 |----- 174	3	7,5	40	100	172
Total	**40**	**100**			**972**

Fonte: dados fictícios elaborados pelas autoras.

Observe que conseguimos ir "tirando provas" do nosso cálculo em uma tabela de frequências, a saber:

a. Na coluna Fac (frequência acumulada), o último valor é igual ao total da somatória das frequências, que também é igual ao total de número de elementos do conjunto de dados (no nosso exemplo é igual a 40).

b. Na coluna FR(%), a somatória de todos os números é igual a 100.

c. Na coluna FRac(%), o último elemento é igual a 100.

Analise que, para conseguirmos fazer uma tabela de frequências, temos que fazer um bom ROL, pois, se você errar na distribuição da coluna frequências (fi), você poderá errar os dados de toda tabela.

Exemplo 02: Suponhamos que perguntamos a idade a um grupo de 100 pessoas e, ainda, que desejamos representar os resultados obtidos distribuídos em intervalos. As idades desse grupo estão apresentadas a seguir (em rol):

17	17	18	18	18	18	19	19	19	20
20	21	21	21	22	23	24	24	25	25
25	26	26	26	27	27	27	28	28	28
29	29	29	30	30	30	30	31	32	32
33	34	34	35	35	36	36	36	36	37
38	39	39	40	40	40	40	40	41	42
43	43	43	43	43	44	44	44	44	44
44	44	45	45	45	46	46	46	46	46
47	47	47	47	47	48	48	48	48	48
48	48	48	49	50	52	52	53	55	55

Fonte: elaborados pelas autoras.

1º passo: AT = X máx. – X mín.: 55 – 17 = 38

2º passo: $\sqrt{n} = 100 = 10$

3º passo: $h = \frac{AT}{K} = \frac{38}{10} = 3{,}8$ = arredondando 4.

Tabela 10 - Idades de 100 colaboradores de uma organização

IDADES	Fi	FR (%)	Fac	FRac (%)	Xi
17 |----- 21	11	11	11	11	19
21 |----- 25	7	7	18	18	23
25 |----- 29	12	12	30	30	27
29 |----- 33	10	10	40	40	31
33 |----- 37	9	9	49	49	35
37 |----- 41	9	9	58	58	39
41 |----- 45	14	14	72	72	43
45 |----- 49	21	21	93	93	47
49 |----- 53	4	4	97	97	51
53 |----- 57	3	3	100	100	55
Total	**100**	**100**			**370**

Fonte: elaborada pelas autoras.

GRÁFICOS

Gráficos são ferramentas de representação dos dados que servem para facilitar a visualização deles. Devem ter simplicidade e clareza para permitir se chegar a conclusões sobre a evolução do fenômeno ou como se relacionam os valores da série.

Cotidianamente, observa-se que meios de comunicação utilizam essa ferramenta para mostrar as pesquisas. Isso se deve ao fato da facilidade de interpretação demonstrada nos gráficos e da eficiência com que resume informações, embora apresente menor grau de detalhes em relação às tabelas, dando uma ideia mais global dos dados.

Ao optar pela utilização de um gráfico em uma pesquisa, devemos levar em conta que sua construção exige cuidados, como escolher o tipo que melhor se encaixa na representação dos dados.

ELEMENTOS FUNDAMENTAIS DOS GRÁFICOS

Os elementos fundamentais de um gráfico para que ele cumpra sua função de racionalização das informações são:

- Título: para indicar o que ele representa.
- Legenda: para facilitar a leitura do gráfico.
- Fonte: para informar a origem dos dados.

A identificação ou o título de um gráfico deve aparecer na parte inferior dele, precedido pela palavra Gráfico, seguido de seu número de ordem de ocorrência no texto. Se necessário, uma legenda explicativa pode ser utilizada. Se os dados observados no gráfico forem extraídos de terceiros, como dados obtidos de uma revista, de uma fundação, prefeitura, etc., então, é obrigatório o uso de sua fonte.

GRÁFICOS PARA VARIÁVEIS QUALITATIVAS

Gráficos em colunas

Esse tipo de gráfico é formado por retângulos verticais, em que cada um dos retângulos representa a intensidade de um atributo. É o gráfico mais utilizado para representar variáveis qualitativas. Indicado quando as categorias são breves.

Exemplo:

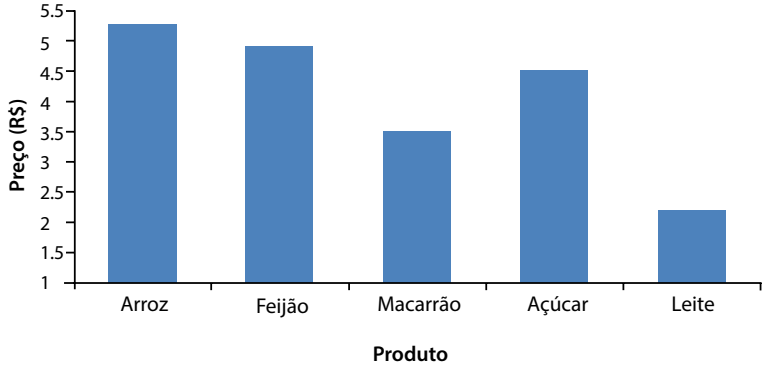


Gráfico 1 - Valores de Produtos em um Supermercado
Fonte: elaborados pelas autoras.

No caso de estarmos trabalhando com duas variáveis, podemos utilizar os gráficos comparativos. Por exemplo, em uma escola temos 120 alunos e o gráfico a seguir nos indica o número de alunos inscritos em cada modalidade esportiva praticada na escola. Cada aluno só pratica um esporte.

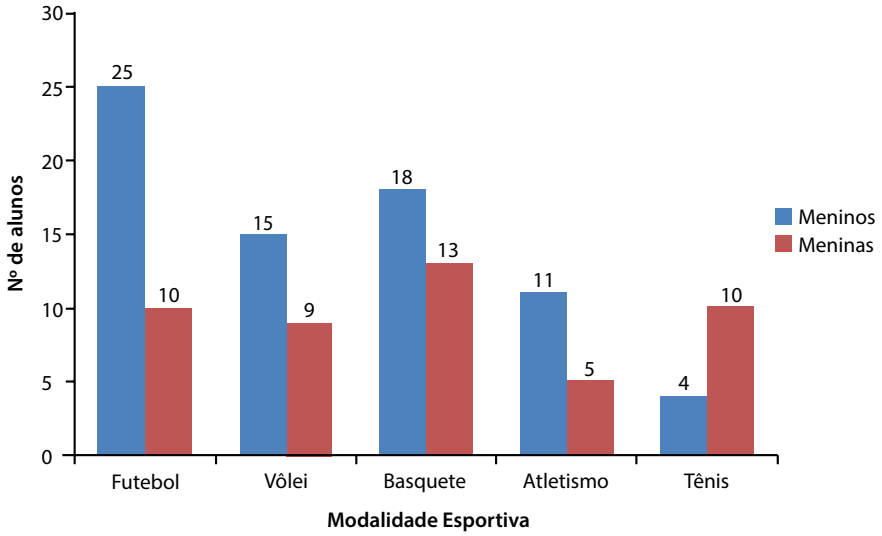


Gráfico 2 - Modalidades Esportivas de uma escola
Fonte: elaborados pelas autoras.

Para os gráficos comparativos, podemos utilizar as barras empilhadas, uma acima da outra, como visto abaixo:

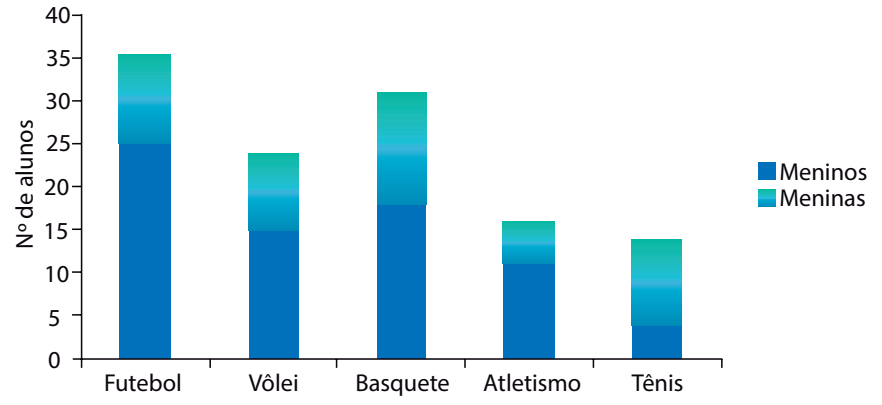


Gráfico 3 - Modalidades Esportivas de uma escola
Fonte: elaborados pelas autoras.

É importante observar, neste tipo de gráfico, que cada uma das modalidades de esporte refere-se a uma das barras e é dividida por sexo.

Gráfico em barras

É um gráfico formado por retângulos horizontais, em que cada um deles representa a intensidade de um atributo. O objetivo deste gráfico é de comparar grandezas, e é recomendável para variáveis cujas categorias tenham designações extensas.

Exemplo:

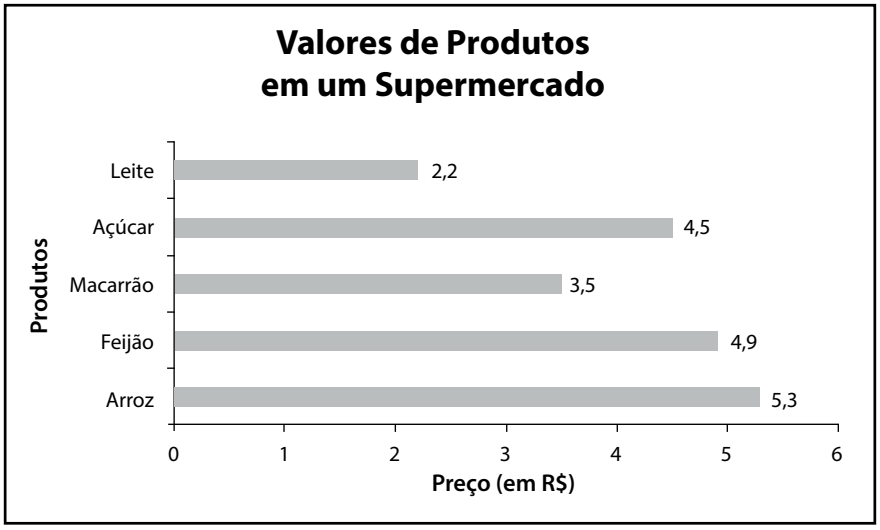


Gráfico 4 – Valores de Produtos em um Supermercado
Fonte: elaborado pelas autoras.

Gráfico de setores

Também conhecido como gráfico de "pizza". Neste tipo de gráfico, a variável em estudo é projetada num círculo dividido em setores com áreas proporcionais às frequências das suas categorias. É recomendado para o caso em que o número de categorias não é grande e não obedece a alguma ordem específica.

Exemplo:

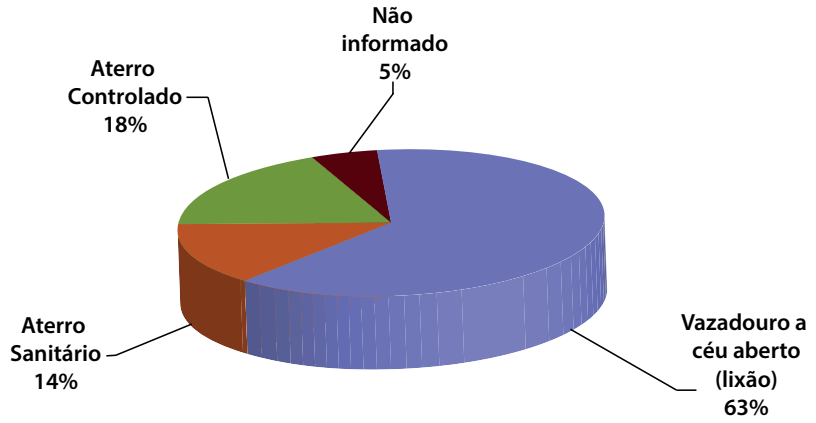


Gráfico 5 - Destinação final de resíduos sólidos por número de município (2000)
Fonte: IBGE (2000, on-line).

Gráfico de linhas

Este tipo de gráfico é utilizado para representar dados relacionados ao tempo. É feito colocando-se no eixo vertical (y) a mensuração da variável em estudo e no eixo horizontal (x) as unidades da variável numa ordem crescente. Esse tipo permite mostrar as flutuações da variável ao longo do tempo além de analisar as tendências.

Exemplo:

Suponha uma empresa que esteja analisando o número de vendas de notebooks de certa marca nos primeiros 4 meses do ano.

Tabela 11 - Número de vendas de notebooks por mês

MÊS	NÚMERO DE VENDAS
Janeiro	10
Fevereiro	16
Março	9
Abril	12

Fonte: elaborada pelas autoras.

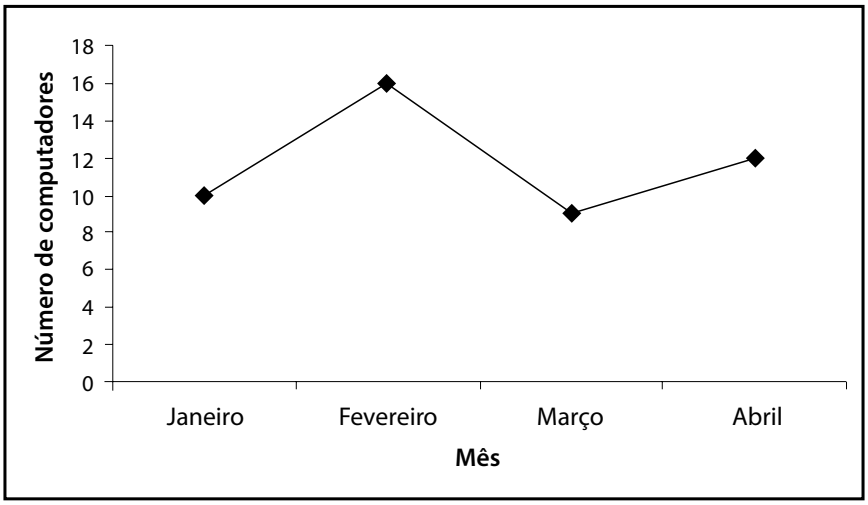


Gráfico 6 - Número de vendas de notebooks por mês
Fonte: elaborado pelas autoras.

GRÁFICOS PARA VARIÁVEIS QUANTITATIVAS

Se o conjunto de dados consiste de muitas observações, seria trabalhoso construir gráficos como os já mencionados. Assim, para variáveis quantitativas, são usados outros dois gráficos importantes: Histograma e Polígono de Frequência.

Histograma

É um gráfico de colunas, sendo dispostos, no eixo horizontal, os limites das classes da variável em questão, segundo as quais os dados foram agrupados e, no eixo vertical, as frequências para cada agrupamento. Um detalhe importante é que, no histograma, as colunas são retângulos justapostos.

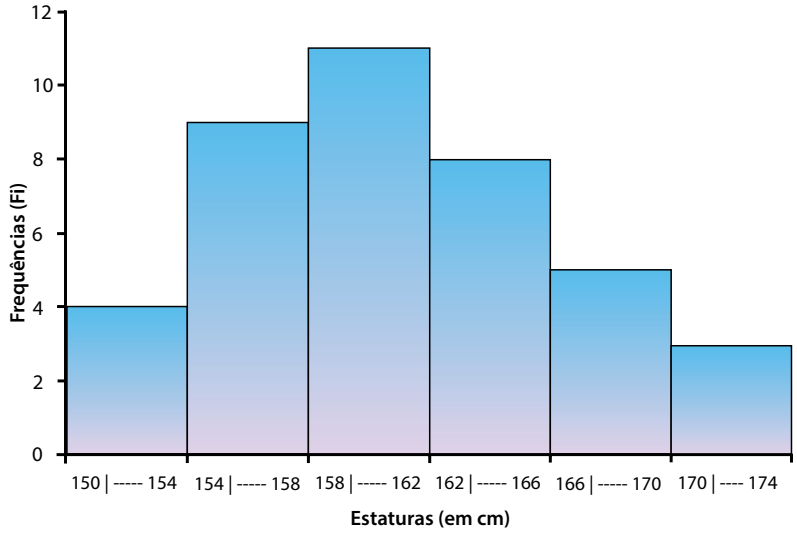


Gráfico 7 - Estatura (em cm) de 40 colaboradores de uma organização
Fonte: dados fictícios elaborados pelas autoras.

Polígono de frequência

Segundo Crespo (2009), o polígono de frequência é um gráfico de linhas, sendo as frequências marcadas sobre perpendiculares ao eixo horizontal, levantadas pelos pontos médios dos intervalos de classes. Vejamos um exemplo desse gráfico a seguir:

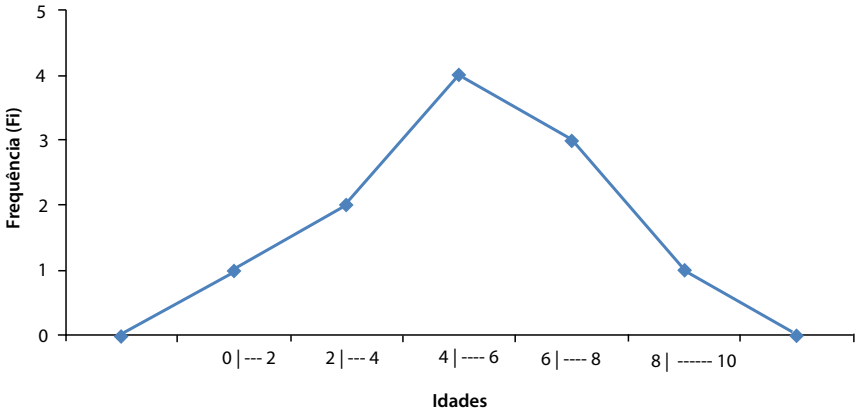


Gráfico 8 - Idades de crianças em uma creche
Fonte: dados fictícios – elaborados pelas autoras.

REFLITA

Os diversos tipos de gráficos sempre têm o mesmo objetivo: mostrar os dados de forma resumida. O tipo de gráfico a ser utilizado depende da escolha e do objetivo do pesquisador.

A interpretação adequada de um gráfico ou tabela é fundamental para o entendimento da pesquisa. Ler o título de forma minuciosa e observar valores máximos, mínimos e suas variações são pontos fundamentais para uma interpretação adequada.

SAIBA MAIS

O Microsoft Excel não fornece mais o assistente de gráfico. Agora, para criar um gráfico básico, selecione uma parte do conteúdo que pretende representar e clique no tipo de gráfico desejado na guia Inserir, no grupo Gráficos.

O Microsoft Excel possui um assistente para facilitar a geração de gráficos, no qual ele divide este processo em quatro etapas subsequentes, apresentando a cada etapa apenas as opções, diretamente, relacionadas e necessárias para a conclusão do gráfico.

Fonte: Support Office (2017)

CONSIDERAÇÕES FINAIS

Nesta unidade, tratamos da necessidade de que a apresentação dos dados seja feita de forma precisa. As duas formas vistas, nesta unidade, foram tabelas e gráficos. Enfatizamos que o uso correto das formas de apresentação dos dados é fundamental para o sucesso da pesquisa.

Os gráficos são formas de sintetizar as informações coletadas. São importantes para dispormos as informações de forma clara e para que consigamos enxergar o que aconteceu na nossa pesquisa. Existem diversos tipos de gráficos. Nesta unidade, vimos os tipos mais comuns como os de barras e colunas, os de linha, de setores ou pizza, histograma e polígono de frequência.

De forma geral, os gráficos demonstram dados quantitativos associados a alguma variável qualitativa. Todos os gráficos têm o mesmo objetivo, que é o de demonstrar de forma clara e rápida os dados da pesquisa. A escolha do tipo adequado fica ou a critério do pesquisador ou a critério do objetivo da pesquisa estudada.

Nessa etapa, o interesse maior consiste em tirar conclusões que auxiliem o pesquisador a resolver seu problema. Portanto, além da organização e da tabulação dos dados, as tabelas e os gráficos nos apresentam de uma forma clara, sucinta e objetiva os resultados de uma pesquisa, para tirarmos conclusões e nos ajudarem na tomada de decisões.

Também observamos que podemos construir gráficos e tabelas por meio de programas computacionais, por exemplo, o Microsoft Excel, que é uma planilha de dados que dispõe de ferramentas para construção de gráficos, a partir de tabelas. Esse programa é fácil de usar, tem inúmeras ferramentas que podem ser úteis ao gestor.

Esperamos que você tenha compreendido essa unidade, porque ela é de extrema importância aos futuros profissionais, pois tabelas e gráficos estão presentes no nosso cotidiano, e cabe a nós entender, interpretar e avaliar os dados apresentados por meio de tabelas e de gráficos.

ATIVIDADES

Considere a seguinte planilha de dados quanto às topologias de rede de computadores na resposta do tempo ao usuário:

INFORMAÇÃO	TOPOLOGIA	TEMPO DE RESPOSTA
1	C1	6,0
2	C2	7,0
3	C3	5,0
4	C1	6,3
5	C2	6,8
6	C2	7,2
7	C1	6,0
8	C2	6,7
9	C1	5,7
10	C2	6,5
11	C3	6,4
12	C1	5,7
13	C3	7,2
14	C3	6,8
15	C3	6,5
16	C2	7,5

1. Construa uma tabela de distribuição de frequências para Topologia.

2. Construa um gráfico de setores para Topologia.

3. Construa uma tabela de distribuição de frequências com intervalo de classes para a variável tempo de resposta em quatro classes.

4. Demonstre um histograma para a variável tempo de resposta.

5. Demonstre um polígono de frequências para a variável tempo de resposta.

REGRAS DE ARREDONDAMENTO

Arredondamentos são de fundamental importância para nossos estudos, principalmente ao calcular valores que têm muitas casas decimais. Muitas vezes, é conveniente suprimir unidades inferiores às de determinada ordem. Esta técnica é denominada arredondamento de dados ou valores.

Muitas vezes é muito mais fácil e mais compreensível usarmos valores arredondados para melhor entendimento do público que terá acesso à informação.

De acordo com a Resolução nº 886/66 do IBGE:

I) < 5 (menor que 5). Quando o primeiro algarismo a ser abandonado é 0,1,2,3 ou 4, ficará inalterado o último algarismo que permanece.

Exemplo:

43,24 passa para 43,2.

54,13 passa para 54,1.

II) > 5 (maior que 5). Quando o primeiro algarismo a ser abandonado é o 6,7,8, ou 9, aumenta-se em uma unidade o algarismo que permanece.

Exemplos:

23,87 passa para 23,9.

34,08 passa para 34,1.

74,99 passa para 75,0.

III) = 5 (igual a 5). Quando o primeiro algarismo a ser abandonado é 5, há duas soluções:

A) Se após o 5 seguir, em qualquer casa, um algarismo diferente de zero, aumenta-se uma unidade ao algarismo que permanece.

Exemplos:

6,352 passa para 6,4.

55,6501 passa para 55,7.

96,250002 passa para 96,3.

B) Se o 5 for o último algarismo ou após o 5 só se seguirem zeros, o último algarismo a ser conservado só será aumentando de uma unidade se for ímpar.

Exemplos:

14,75 passa para 14,8

24,65 passa para 24,6

34,75000 passa para 34,8

44,8500 passa para 44,8

Observação: Nunca devemos fazer arredondamentos de sucessivos.

Para melhor entendimento didático quando o último primeiro algarismo a ser abandonado for 5 o último a permanecer aumenta em uma unidade.

Exemplo:

72,5 passa para 73 inteiros.

72,45 passa para 72,5 (setenta e dois inteiros e cinco décimos) - uma casa após a vírgula.

72,445 passa para 72,45 (setenta e dois inteiros e quarenta e cinco centésimos) duas casas após a vírgula.

Fonte: Portal da Educação (2013, on-line).

MATERIAL COMPLEMENTAR

LIVRO

Estatística Básica

Geraldo Luciano Toledo, Ivo Izidoro Ovalle

Editora: Atlas

Sinopse: Este livro contém a matéria fundamental para estudos subsequentes no campo da estatística inferencial. Além disso, aborda os tópicos mais importantes da estatística básica.

REFERÊNCIAS

CRESPO, A. A. **Estatística Fácil**. São Paulo: Saraiva, 19. ed., 2009.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Comentários**. Disponível em: <https://www.ibge.gov.br/home/estatistica/economia/ppm/2006/comentarios.pdf >. Acesso em: 20 abr. 2017.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Pesquisa Nacional por Amostra de Domicílios**. 2006-2007. Disponível em: <http://www.ibge.gov.br/home/estatistica/pesquisas/pesquisa_resultados.php?id_pesquisa=40>. Acesso em: 20 abr. 2017.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Pesquisa Nacional por Amostra de Domicílios**. 2007-2009. Disponível em: <http://www.ibge.gov.br/home/estatistica/pesquisas/pesquisa_resultados.php?id_pesquisa=40>. Acesso em: 20 abr. 2017.

PORTAL DA EDUCAÇÃO. **Regras de Arredondamento**. Disponível em:<https://www.portaleducacao.com.br/conteudo/artigos/administracao/regras-de-arredondamento/30568> Acesso em: 20 abr. 2017.

SUPPORT OFFICE. **Criar um gráfico do início ao fim**. Disponível em: <https://support.office.com/pt-br/article/Criar-um-gr%C3%A1fico-do-in%C3%ADcio-ao-fim-0baf399e-dd61-4e18-8a73-b3fd5d5680c2> Acesso em: 20 abr. 2017.

GABARITO

1. Tabela 01 - Distribuição de frequências para a variável topologia

TOPOLOGIA	*Fi*	*Fr%*	*Fac*
C1	5	31,25	5
C2	6	37,50	11
C3	5	31,25	16
Total	**16**	**100,00**	**-**

Fonte: as autoras.

2.

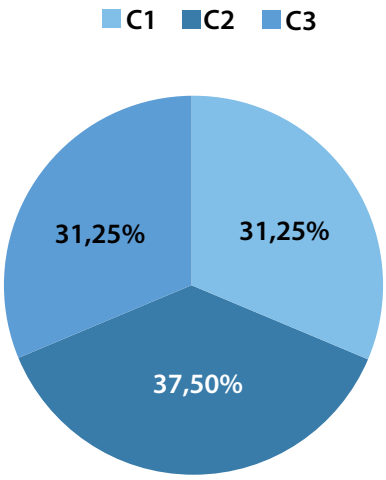


Gráfico 01 - Porcentagem de clientes para a variável topologia
Fonte: as autoras.

3. $AC = \frac{7,5 - 5,0}{4} = 0,625 =$ arredondando para 0,63

Tabela 01 - Distribuição de frequências para a variável tempo de resposta

TEMPO	*Fi*	*Fr%*	*Fac*	Xi
5,00 |--5,63	1	6,25	1	5,32
5,63 |--6,26	4	25,00	5	5,95
6,26 |--6,89	7	43,75	12	6,58
6,89|--7,52	4	25,00	16	7,21
Total	**16**	**100,00**	**-**	**-**

Fonte: as autoras.

4.

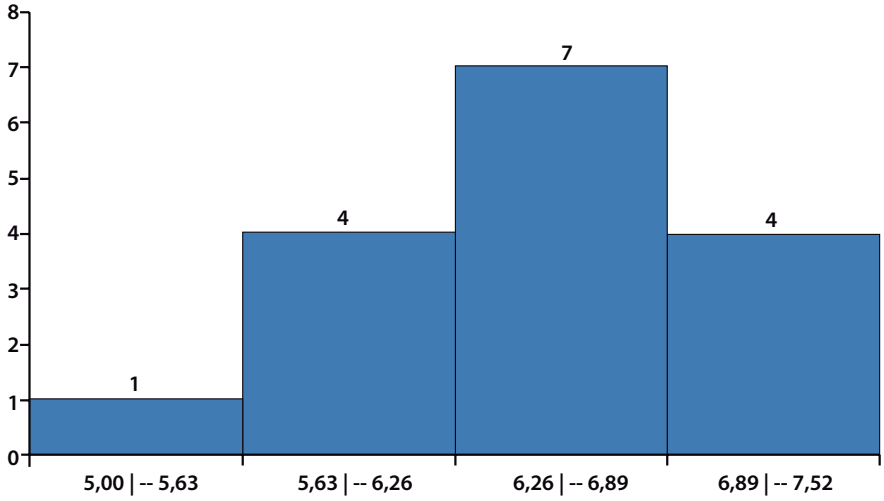


Gráfico 02 - Porcentagem de clientes para a variável tempo de resposta ao usuário
Fonte: as autoras.

5.

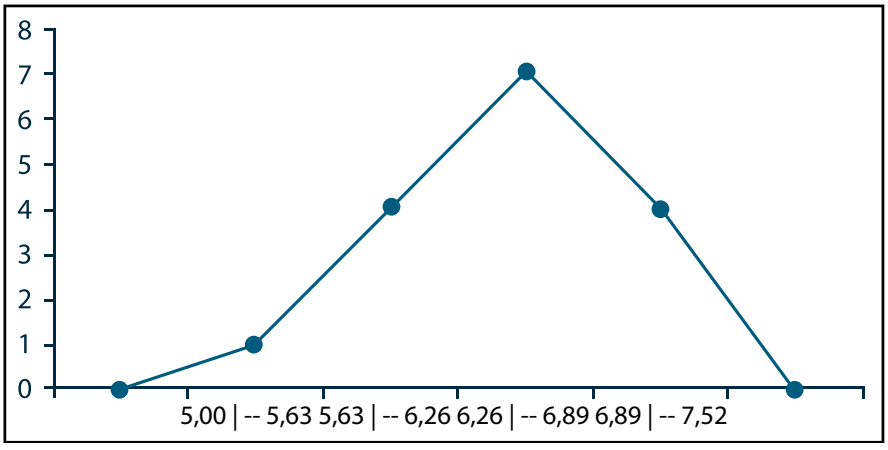


Gráfico 3 - Porcentagem de clientes para a variável tempo de resposta
Fonte: as autoras.

Professora Me. Ivnna Gurniski de Oliveira
Professora Me. Renata Cristina de Souza Chatalov

MEDIDAS DESCRITIVAS ASSOCIADAS A VARIÁVEIS QUANTITATIVAS

UNIDADE III

Objetivos de Aprendizagem

- Compreender as principais medidas estatísticas de posição, dispersão e separatrizes.
- Entender a aplicação das medidas estatísticas de posição, dispersão e separatrizes.

Plano de Estudo

A seguir, apresentam-se os tópicos que você estudará nesta unidade:

- Medidas de Posição
- Medidas Separatrizes
- Medidas de Dispersão

Unicesumar

INTRODUÇÃO

Quando estamos realizando uma pesquisa, podemos fazer a apresentação dos dados por meio de gráficos, tabelas, ou fazendo o uso de medidas que resumem as informações obtidas na coleta dos dados, chamadas medidas descritivas.

Nesta unidade, estudaremos as medidas de posição e de dispersão utilizadas para descrever dados quantitativos. Essas medidas são demasiadas importantes na representação dos dados.

As medidas de posição ou de tendência central mostram o centro de uma distribuição de dados, dando-nos uma noção do que está ocorrendo com eles. Por meio dessas medidas, podemos localizar a maior concentração de valores em uma distribuição, ou seja, se ela localiza-se no início, no meio ou no centro, ou, ainda, se há uma distribuição por igual. As medidas de tendência central mais importantes são a média aritmética, a mediana e a moda. Vale salientar que temos outras medidas de posição que são as separatrizes, que englobam: a própria mediana, os quartis e os percentis.

As medidas de dispersão, porém, são utilizadas para avaliar o grau de variabilidade do conjunto de dados, mostrando se ele é homogêneo ou heterogêneo. Essas medidas servem para analisar o quanto os dados são semelhantes e descrevem o quanto os dados distanciam do valor central, portanto as medidas de dispersão servem também para avaliar o grau de representação da média. As medidas de dispersão mais utilizadas são: a amplitude total, a variância, o desvio padrão e o coeficiente de variação.

Assim, para descrevermos um conjunto de dados, é de bom grado sempre termos uma medida de posição e uma de dispersão para representá-lo. A de posição, para dizer o que está ocorrendo com a pesquisa, e a de dispersão, para dizer se há alta ou baixa variabilidade.

Nesta unidade, vamos estudar as principais medidas de posição e medidas de dispersão utilizadas nas pesquisas para descrever e representar o conjunto de dados.

MEDIDAS DE POSIÇÃO

Para sumarizar as informações de um conjunto de observações, muitas vezes é necessário utilizar medidas que resumem em um só número certas características. Assim, temos as medidas de posição, de dispersão, de assimetria e de curtose. Se as medidas são calculadas para dados a partir de uma amostra, são chamadas de estatísticas da amostra; se são calculadas a partir de uma população, são chamadas de parâmetros da população.

As principais medidas de posição e as principais medidas separatrizes são:

- *Medidas de Posição*
  - *Medidas de tendência central*
    - *Média Aritmética*
    - *Mediana*
    - *Moda*
  - *Separatrizes*
    - *Mediana*
    - *Quartis*
    - *Decis*
    - *Percentis*

As principais medidas de dispersão são:

- *Medidas de Dispersão*
  - *Amplitude Total*
  - *Variância*
  - *Desvio – Padrão*
  - *Coeficiente de Variação*

As medidas de posição servem para representar o ponto central de equilíbrio de um conjunto de observações ordenadas segundo suas grandezas. Dentre as medidas de posição, destacamos: média, mediana e moda, sendo que a medida a ser escolhida para representar coerentemente os dados depende das características deles.

MÉDIA ARITMÉTICA

A média de uma variável é a medida mais importante e mais simples de ser calculada. Esta fornece uma medida de posição central. Se os dados são de uma amostra, a média é denotada por $\overline{x}$; se os dados são de uma população, a média é denotada pela letra grega μ.

A média de um conjunto de dados é encontrada somando seus valores e dividindo pelo número de observações. Seja x_1, x_2, ..., x_n um conjunto de dados, a média será dada por:

População

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Amostra

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

Exemplo:

Suponha que estamos estudando a idade de cinco indivíduos de uma família. As idades observadas foram: 5, 10, 12, 35, 38. Logo, a idade média dessa família é:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$$

$$\overline{x} = \frac{5 + 10 + 12 + 35 + 38}{5} = 20 \text{ anos}$$

Exercício

Calcule a média para a quantidade de atendimentos realizados em um mês por um consultor financeiro.

18, 19, 20, 21, 21, 22, 24, 34, 35, 37

R: 25,1

MÉDIA ARITMÉTICA PONDERADA

Existem situações em que não temos todos os dados disponíveis ou então temos "pesos" diferentes para os dados considerados. Nestes casos, utilizamos o que chamamos de média aritmética ponderada para obtermos a média, cujas fórmulas para População e para Amostra são dadas da seguinte maneira:

População

$$\mu = \frac{\sum F_i x_i}{N}$$

Amostra

$$\bar{x} = \frac{\sum F_i x_i}{n}$$

Se a situação for de dados agrupados, a média é obtida a partir de uma ponderação em que os pesos são as frequências absolutas (Fi) de cada classe e xi é o ponto médio da classe i. Observe o exemplo abaixo:

Tabela 1 - Idades de colaboradores de uma empresa

IDADES	FREQUÊNCIA (Fi)
18	4
19	1
20	1
21	3
25	3
28	5
29	3
Total	**20**

Fonte: dados fictícios elaborados pelas autoras.

A média ponderada será dada por:

$$\bar{x} = \frac{(18*4)+(19*1)+(20*1)+(21*3)+(25*3)+(28*5)+(29*3)}{20} = \frac{476}{20} = 23{,}8$$

Tabela 2 - Idades de colaboradores de uma empresa

IDADES	FREQUÊNCIA	Xi. Fi
18	4	72
19	1	19
20	1	20
21	3	63
25	3	75
28	5	140
29	3	87
Total	**20**	**476**

Fonte: dados fictícios elaborados pelas autoras.

$$\frac{476}{20} = 23{,}8.$$

Quando desejamos obter a média de dados que estão distribuídos em uma tabela com intervalo de classes, a equação para determinar a média continua sendo a mesma, no entanto, temos que tirar o ponto médio antes. Observem um exemplo a seguir.

Tabela 03 - Idades de 100 colaboradores de uma organização

IDADES	Fi	Xi	Xi. Fi
17 |----- 21	11	19	(11*19) = 209
21 |----- 25	7	23	(7*23) = 161
25 |----- 29	12	27	(12*27) = 324
29 |----- 33	10	31	(10*31) = 310
33 |----- 37	9	35	(9*35) = 315
37 |----- 41	9	39	(9*39) = 351
41 |----- 45	14	43	(14*43) = 602
45 |----- 49	21	47	(21*47) = 987
49 |----- 53	4	51	(4*51) = 204

53 |----- 57	3	55	(3*55) = 165
Total	**100**	**370**	**3.628**

Fonte: dados fictícios elaborados pelas autoras.

$$\frac{3.628}{100} = 36{,}28 = 36.$$

Exercício:

Tabela 4 - Distribuição de frequências para a idade dos clientes de uma imobiliária para efetuar uma compra

CLASSES	Fi	Fr(%)	Fac	Xi
17 |---- 29	4	36,4	4	23
29 |---- 41	4	36,4	8	35
41 |----| 53	3	27,3	11	47
Total	**11**	**100**	**-**	**-**

Fonte: dados fictícios elaborados pelas autoras.

Resposta: 33,91

Existem situações em que os dados não estão agrupados, mas existem "pesos" diferentes para cada um deles. Vejamos um exemplo:

A média da nota bimestral dos alunos da Unicesumar é composta pela nota de uma prova (com peso 8) e pela nota dos trabalhos (com peso 2). Calcule a média bimestral do aluno que tirou as seguintes notas:

Prova: 7 (peso 8) Trabalho: 9 (peso 2)

A média será dada por:

$$\bar{x} = \frac{(8 \times 7) + (2 \times 9)}{8 + 2} = 7{,}4$$

Exercício: Calcule as médias ponderadas das notas bimestrais dos alunos abaixo:

Tabela 5 - Médias bimestrais dos alunos da Escola X

ALUNO	PROVA	TRABALHO
João	5,0	3,0
Antônio	7,0	4,0

*Considere que o peso da prova seja igual a 9,0 e o peso do trabalho seja igual a 1,0.
Fonte: dados fictícios elaborados pelas autoras.

Resposta: 4,8 e 6,7

A média é a medida mais importante dentro de um conjunto de dados e possui algumas propriedades importantes. São elas:

1. A média é única em um conjunto de dados.
2. A média é afetada por valores extremamente pequenos ou grandes.
3. A média depende de todos os valores observados, assim, qualquer modificação nos dados fará que a média fique alterada.
4. A soma das diferenças dos valores observados em relação à média é zero:

$$\sum (x_i - \bar{x}) = 0$$

A propriedade 2 é importante, pois, em um conjunto de dados muito heterogêneo, a média torna-se uma medida não apropriada para representar os dados, devendo o pesquisador optar por uma outra medida.

A propriedade 4 é importante na definição de variância, uma medida de dispersão que veremos ainda nesta unidade.

SAIBA MAIS

O conceito e a ideia de média estão sempre relacionados com a soma dos valores de um determinado conjunto de medidas, dividindo-se o resultado dessa soma pela quantidade dos valores que foram somados. Esse procedimento é o que definimos como média aritmética simples, e que estamos acostumados a aplicar nas estimativas que fazemos diariamente.

Não faltam brincadeiras em relação a esse tipo de cálculo quando, ironicamente, calculamos a média salarial de, por exemplo, determinada empresa, somando o maior salário com o menor e dividindo por dois. A média aritmética simples produz a média ponderada em função da repetição das medidas. Geralmente, a média ponderada é apresentada com regras pré-estabelecidas para os seus pesos, dando a aparência de que se trata de outra fórmula, muito diferente da média aritmética.

Fonte: Neto (2009).

MODA

Chamamos de moda o valor que aparece com maior frequência em um conjunto de dados. Para o caso de valores individuais, a moda pode ser determinada observando-se o rol dos dados.

Exemplos:

Observe as notas da prova de estatística de uma turma do curso de administração:

4; 5; 6; 6; 6; 6; 7; 7; 7; 8.

A moda é 6, pois esse é o valor que ocorreu com maior frequência.

Essa sequência é unimodal, pois tem apenas uma moda.

Veja essa outra sequência:

4; 5; 5; 5; 6; 7; 7; 7; 8; 9.

Nesta, existem duas modas (5 e 7), ela é bimodal.

Essa outra:

1; 2; 3; 4; 5; 6; 7; 8; 9; 10.

Não existe moda, nenhum valor aparece com maior frequência, é amodal ou antimodal.

Quando os dados estão agrupados em classes, primeiramente é necessário identificar a classe modal que apresenta a maior frequência e calcular então a moda da seguinte maneira:

$$Mo = l_i + \frac{h(F_i - F_{i-1})}{(F_i - F_{i-1}) + (F_i - F_{i+1})}$$

Em que:

i é a ordem da classe modal.

l_i é o limite inferior da classe modal.

h é a amplitude da classe modal.

F_i é a frequência absoluta da classe modal.

F_{i-1} é a frequência absoluta da classe anterior à classe modal.

F_{i+1} é a frequência absoluta da classe posterior à classe modal.

Se o conjunto de dados apresentar todos seus elementos com a mesma frequência absoluta, não existirá a moda. Se ocorrer várias frequências iguais, então teremos uma distribuição com mais de uma moda.

A moda tem o atributo de não ser afetada pelos valores extremos no conjunto de dados.

Exemplo:

Tabela 06 - Idades de 100 colaboradores de uma organização

IDADES	Fi	Xi
17 |----- 21	11	19
21 |----- 25	7	23
25 |----- 29	12	27
29 |----- 33	10	31
33 |----- 37	9	35
37 |----- 41	9	39
41 |----- 45	14	43
45 |----- 49	21	47
49 |----- 53	4	51
53 |----- 57	3	55
Total	**100**	**370**

Fonte: dados fictícios elaborados pelas autoras.

Para isso, devemos determinar a classe modal. A classe modal é a classe com a maior frequência absoluta e, neste caso, é a oitava classe, pois essa possui o maior valor de F_i. Determinada a classe modal, vamos calcular a moda por meio da fórmula para dados agrupados.

Nesse exemplo, temos que a classe modal é a 8º classe, pois é nela que temos a maior frequência, agora que localizamos a classe modal, vamos aplicar a equação dada:

$$Mo = l_i + \frac{h(F_i - F_{i-1})}{(F_i - F_{i-1}) + (F_i - F_{i+1})} = 45 + \frac{4(21 - 14)}{(21 - 14) + (21 - 4)} = 46{,}16$$

Portanto, a moda para o conjunto de dados da Tabela 46,16 ou arredondando é 46.

MEDIANA

Corresponde ao valor central ou à média aritmética dos dois valores centrais de um conjunto de observações organizadas em ordem crescente. Ou seja, 50% das observações são inferiores à mediana e 50% superiores.

Exemplo:

Uma pesquisa em uma empresa apresentou os seguintes dados relacionados ao tempo de trabalho de seus funcionários:

5, 13, 12, 3, 15, 17, 8, 15, 6, 16, 9.

Para encontrarmos a mediana, primeiramente devemos ordenar os dados brutos transformando-os em um rol, ou seja, organizando os dados:

3, 5, 6, 8, 9, 12,13, 15, 15, 16, 17

Identificamos a posição da mediana, após verificar que o conjunto de dados é ímpar, pois n = 11 elementos. Utilizamos a fórmula:
Se n for ímpar: Md = $\frac{n+1}{2}$, portanto: $\frac{11+1}{2} = \frac{12}{2} = 6$. Nesse caso, a mediana é o 6º elemento do conjunto de dados. Depois localizamos o elemento central, no caso 12, pois à esquerda dele temos 5 elementos e à direita também. Assim temos:

$$\text{Md} = 12.$$

$$\underbrace{3, 5, 6, 8, 9,}\ 12,\ \underbrace{13, 15, 15, 16, 17}$$

Quando o rol tiver número par de elementos, a mediana será a média aritmética entre os dois elementos centrais. Vejamos, por exemplo, um rol com 10 elementos (número par de elementos):

3, 5, 6, 8, 9, 13, 14, 15, 15, 16.

$$\text{Md} = \frac{9+13}{2} = 11$$

Assim, considerando n o número de elementos da série, o valor mediano será dado pelo termo de ordem dado pelas seguintes fórmulas:

Se n for ímpar: $\text{Md} = \frac{n+1}{2}$

Se n for par: $\text{Md} = \left[\left(\frac{n}{2}\right)+\left(\frac{n}{2}+1\right)\right].\frac{1}{2}$ (média entre dois números)

Exercício:

Calcule a mediana para as notas dos alunos nas duas situações seguintes:

- 6.0, 4.5, 5.0, 7.0, 6.5;
- 4.8, 6.3, 8.9, 9.5, 6.0, 7,8;

Resposta: 6.0 e 7.05

Para os dados agrupados em distribuição de frequências em classes, tem-se:

$$\boldsymbol{Md = l_i + \frac{h(p - F_{ac\,-1})}{F_i}}$$

Em que:

l_i é o limite inferior da classe da mediana.

h é a amplitude da classe da mediana.

p indica a posição da mediana, onde n , sendo o número total de elementos.
$F_{ac\ -\ 1}$ é a frequência acumulada da classe anterior à da mediana.
F_i é a frequência absoluta da classe da mediana.
Exemplo:
Vamos encontrar a mediana para o seguinte conjunto de dados:

Tabela 08 - Idades de 100 colaboradores de uma organização

IDADES	Fi	Fac	Xi
17 |----- 21	11	11	19
21 |----- 25	7	18	23
25 |----- 29	12	30	27
29 |----- 33	10	40	31
33 |----- 37	9	49	35
37 |----- 41	9	58	39
41 |----- 45	14	72	43
45 |----- 49	21	93	47
49 |----- 53	4	97	51
53 |----- 57	3	100	55
Total	**100**		**370**

Fonte: dados fictícios elaborados pelas autoras.

Primeiramente devemos determinar em qual classe a mediana está, para isso calculamos o valor de p.

$$p = \frac{n}{2} \quad p = \frac{100}{2} = 50$$

Quando o valor de p for decimal, sempre aproximamos seu valor para "cima". Para saber qual é a classe da mediana, devemos olhar na coluna da frequência acumulada, de modo que $p \leq F_{ac}$. Logo, a mediana está na 6ª classe. Em seguida, aplicamos a equação da mediana:

$$\boldsymbol{Md = l_i + \frac{h(p - F_{ac\ -\ 1})}{F_i} = 37 + \frac{4(50 - 49)}{9} = 37{,}44}$$

REFLITA

Para qualquer assunto que trate de dados numéricos, sempre trabalhamos com uma medida de posição. Normalmente usamos a média, que é a medida mais conhecida. Observe também como essas medidas são importantes no seu cotidiano.

(As autoras)

MEDIDAS SEPARATRIZES

As separatrizes são os valores que dividem as séries em partes iguais. As principais medidas separatrizes são: a mediana (já estudada) e os quartis, os decis e os percentis.

QUARTIS

Chamamos de quartis os valores que dividem a distribuição em 4 partes iguais, e podem ser obtidos da seguinte maneira:

Temos três quartis:

Primeiro quartil (Q1) – é o valor que tem 25% dos dados à sua esquerda e o restante (75%) à direita.

Segundo quartil (Q2) – tem 50% dos dados de cada lado, coincide com a mediana.

Terceiro quartil (Q3) – tem 75% dos dados à sua esquerda e 25% à direita.

Fórmulas:

1º Quartil (Q_1)	P=0,25(n +1)
2º Quartil (Q_2)	P=0,50(n +1)
3º Quartil (Q_3)	P=0,75(n +1)

DECIS

Chamamos de decis os valores que dividem uma série em dez partes iguais. Portanto, temos nove decis, o primeiro tem 10% dos dados à sua esquerda e 90% à sua direita, o segundo tem 20% dos dados à sua esquerda e 80% à sua direita e assim por diante até o nono decil, que tem 90% dos dados à sua esquerda e 10% à sua direita.

1º Decil (D_1)	P=0,10(n +1)
2º Decil (D_2)	P=0,20(n +1)
3º Decil (D_3)	P=0,30(n +1)
4º Decil (D_4)	P=0,40(n +1)
5º Decil (D_5)	P=0,50(n +1)
6º Decil (D_6)	P=0,60(n +1)
7º Decil (D_7)	P=0,70(n +1)
8º Decil (D_8)	P=0,80(n +1)
9º Decil (D_9)	P=0,90(n +1)

PERCENTIS

Chamamos de percentis os noventa e nove valores que separam uma série em 100 partes iguais. O cálculo dos percentis está relacionado com percentagem.

No quadro seguinte, são mostrados alguns percentis:

5º Percentil (P_5)	P=0,05(n +1)
25º Percentil (P_{25})	P=0,25(n +1)
50º Percentil (P_{50})	P=0,50(n +1)
75º Percentil (P_{75})	P=0,75(n +1)
90º Percentil (P_{90})	P=0,90(n +1)

Em que a letra *n* nas fórmulas de calcular a posição dos quartis, dos decis e dos percentis representa o número total de elementos da amostra.

Quando o valor de p for inteiro, temos que a medida separatriz está na posição de número p, caso contrário, o cálculo das medidas separatrizes, para os dados em rol, é dado por:

$$S_k = X_i + (p - i)\,(X_{i+1} - X_i)$$

Em que:

S_k é a medida separatriz a ser utilizada, podendo ser os quartis, os decis ou os percentis.

X_i e X_{i+1} são as posições dos dados no rol.

p é a posição da medida separatriz adotada.

i é a parte inteira de p.

Calcule o 3º quartil (Q_3) e o 90º percentil (P_{90}) para a idade média de um grupo de indivíduos que têm as seguintes idades: 18, 19, 20, 21, 21, 22, 24, 24, 25, 27, 30, 33, 38.

Primeiramente, verifique se os dados do rol estão ordenados; caso não estejam, coloque-os em ordem crescente. Em seguida, calcule a posição do dado e, por fim, substitua os valores numéricos na fórmula: $S_k = X_i + (p - i)\,(X_{i+1} - X_i)$.

- Para o (Q_3) tem-se:

  $p = 0{,}75(13 + 1) = 10{,}5 \rightarrow k = 10$

  Assim,

  $Q_3 = X_{10} + (p - k)(X_{11} - X_{10})$

  $Q_3 = 27 + (10{,}5 - 10)(30 - 27)$

  $Q_3 = 27 + (0{,}5 \,.\, 3)$

  $Q_3 = 28{,}5$ anos.

  Portanto, pode-se afirmar que 75% dos indivíduos têm idade inferior a 28,5 anos.

- Para o ($P90$) tem-se:

  $p = 0{,}90(13 + 1) = 12{,}6 \rightarrow k = 12$

  Assim,

  $P_{90} = X_{12} + (p - k)(X_{13} - X_{12})$

  $P_{90} = 33 + (12{,}6 - 12)(38 - 33)$

  $P_{90} = 33 + (0{,}6 \cdot 5)$

  $P_{90} = 33 + (0{,}6 \cdot 5)$

  $P_{90} = 36$ anos.

  Portanto, 90% dos indivíduos têm idade inferior a 36 anos.

Para os dados agrupados, o cálculo das medidas separatrizes é dado por

$$S_k = l_i + \frac{h(p - F_{ac-1})}{F_i}$$

Em que:

S_k é a medida separatriz a ser utilizada, podendo ser os quartis, os decis ou os percentis.

l_i é o limite inferior da classe da separatriz.

h é a amplitude da classe da separatriz.

p é a posição da medida separatriz adotada. Sendo que para os quartis $p = \frac{nk}{4}$, $k = 1, 2, 3$; para os decis $p = \frac{nk}{10}$, $k = 1, 2, ..., 9$; para os percentis $p = \frac{nk}{100}$, $k = 1, 2, ..., 99$; onde n é a quantidade de elementos da amostra.

F_{ac-1} é a frequência acumulada da classe anterior a da separatriz.

F_i é a frequência absoluta da classe da separatriz.

Exemplo:

Vamos determinar o Q_3 e o D_7 para o seguinte conjunto de dados:

Tabela 9 - Teor de oxigênio (mg/L) em vários rios da região Norte do Brasil

CLASSES	Fi	Fr	Fr%	Fac	Xi
0,5|---0,8	4	0,2500	25,00	4	0,65
0,8|---1,1	4	0,2500	25,00	8	0,95
1,1|---1,4	7	0,4375	43,75	15	1,25
1,4|---1,7	1	0,0625	6,25	16	1,55
Total	**16**	**1,0000**	**100,00**	-	-

Fonte: dados fictícios elaborados pelas autoras.

- Para o Q_3:

  Primeiro vamos determinar a posição da medida e, em seguida, determinar qual é a sua classe.

$$p = \frac{nk}{4} = \frac{16 \cdot 3}{4} = 12$$

  Para saber qual é a classe do Q_3, devemos olhar na coluna da frequência acumulada, de modo que $p \leq F_{ac}$. Logo, o Q_3 está na 3ª classe, pois $12 \leq 15$ e corresponde à

$$Q_3 = 1{,}1 + \frac{0{,}3(12 - 8)}{7} = 1{,}27$$

  Portanto, pode-se afirmar que 75% dos rios da região norte do Brasil têm teor de oxigênio inferior a 1,27 mg/L.

  Para o D_7:

  Primeiro vamos determinar a posição da medida e, em seguida, determinar qual é a sua classe.

$$p = \frac{nk}{10} = \frac{16 \cdot 7}{10} = 11{,}2$$

Para saber qual é a classe do D_7, devemos olhar na coluna da frequência acumulada, de modo que $p \leq F_{ac}$. Logo, o D_7 está na 3ª classe, pois 11,2 ≤ 15 e corresponde à

$$D_7 = 1{,}1 + \frac{0{,}3(11{,}2 - 8)}{7} = 1{,}24$$

Portanto, pode-se afirmar que 70% dos rios da região norte do Brasil têm teor de oxigênio inferior a 1,24 mg/L.

MEDIDAS DE DISPERSÃO

As medidas de dispersão mostram a variabilidade de um conjunto de observações em relação à região central. Essas medidas indicam se um conjunto de dados é homogêneo ou heterogêneo. Além disso, mostram se a medida de tendência central escolhida representa bem o conjunto de dados que está sendo trabalhado pelo pesquisador. Vejamos um exemplo:

Considere as idades de três grupos de pessoas A, B e C:

A: 15; 15; 15; 15; 15

B: 13; 14; 15; 16; 17

C: 5; 10; 15; 20; 25

A média aritmética do conjunto A é 15, do B é 15 e do C também é 15.

A média aritmética é a mesma para os três conjuntos acima, porém o grau de homogeneidade entre eles é muito diferente, ou seja, a variação dos seus elementos em relação à média é bem distinta. O conjunto A não tem dispersão, o B tem certo grau de variabilidade e o conjunto C tem grande variabilidade. Por isso, devemos estudar as medidas de dispersão, pois conjuntos de dados diferentes podem ter médias iguais, porém isso não indica que são iguais, uma vez que a variabilidade entre eles pode ser diferente.

AMPLITUDE TOTAL

A amplitude total de um conjunto de dados é a diferença entre o maior e o menor valor. Essa medida nos diz pouco, pois embora fácil de ser calculada, é baseada em somente duas observações, sendo altamente influenciada pelos valores extremos; quanto maior a amplitude, maior será a variabilidade. Veja sua fórmula abaixo:

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

Em que:

x_{max} é o maior valor no conjunto de dados.

x_{min} é o menor valor no conjunto de dados.

Exemplo:

Suponha que estamos estudando a idade de cinco indivíduos de uma família. As idades observadas foram: 5, 10, 12, 35, 38. Logo, a amplitude das idades nessa família:

$$AT = 38 - 5 = 33 \text{ anos}$$

Esta medida de dispersão não leva em consideração os valores intermediários, perdendo a informação de como os dados estão distribuídos.

Exercício

Calcule a Amplitude total dos seguintes conjuntos de dados:

A: 15; 15; 15; 15; 15

B: 13; 14; 15; 16; 17

C: 5; 10; 15; 20; 25

R: 0; 4; 20

VARIÂNCIA

A variância é uma medida de variabilidade que utiliza todos os dados. É calculada considerando o quadrado dos desvios em relação à média aritmética dos dados em estudo.

Se os dados são para uma população, a variância é denotada pelo símbolo grego σ^2 e sua definição é dada como segue:

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

No qual μ é a média da população e N o número de observações.

Se os dados são para uma amostra, a variância, denotada por s^2, é definida como:

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

No qual $\bar{x}$ é a média da amostra e n o número de observações. O uso de (n – 1) neste denominador é necessário para que a variância da amostra resultante forneça uma estimativa não induzida da variância da população.

Na maioria das vezes, trabalhamos nas pesquisas com dados amostrais. Portanto, iremos nos basear sempre na variância amostral.

Exemplo:

Vamos calcular a variância do conjunto de dados do exemplo anterior, ou seja, vamos calcular a variância das idades observadas de uma família, sendo elas: 5, 10, 12, 35, 38.

Primeiramente, devemos calcular a média $\bar{x}$ para as idades:

$$\bar{x} = \frac{5 + 10 + 12 + 35 + 38}{5} = 20$$

Agora vamos calcular a variância das idades:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1} = \frac{\sum_{i=1}^{5} (x_i - 20)^2}{5 - 1}$$

$$s^2 = \frac{(5 - 20)^2 + (10 - 20)^2 + (12 - 20)^2 + (35 - 20)^2 + (38 - 20)^2}{4} =$$

$$\frac{938}{4} = 234{,}5 \text{ anos}^2$$

A unidade da variância é a mesma unidade da característica, entretanto, por simbologia apenas, devemos colocar o símbolo do quadrado junto à unidade. Assim, dizemos que a variância é dada em unidades quadráticas, o que dificulta a sua interpretação. O problema é resolvido extraindo-se a raiz quadrada da variância, definindo-se, assim, o desvio padrão.

DESVIO PADRÃO

O desvio padrão dá a ideia de distribuição dos desvios ao redor do valor da média. Para obtermos o desvio padrão, basta que se extraia a raiz quadrada da variância e, seguindo a notação adotada para as variâncias de população e amostra, *s* denotará o desvio padrão da amostra, enquanto σ, o desvio padrão da população. Assim:

População

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

Amostra

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

De forma mais simplificada:

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

Considerando o exemplo, em que a variância foi $s^2 = 234{,}5$ anos^2, o cálculo do desvio padrão (s) fica bastante simples, ou seja:

$$s = \sqrt{234{,}5} = 15{,}31 \text{ anos}$$

Esta medida é interpretável e dizemos que a dispersão média entre os indivíduos desta família é de 15,31 anos.

Para saber se o desvio padrão está alto ou baixo, vamos compará-lo com o valor da média. Quanto maior o valor do desvio padrão em relação à média, maior então será a variação dos dados e mais heterogêneo é o nosso conjunto de observações.

COEFICIENTE DE VARIAÇÃO

O Coeficiente de Variação (CV) envolve cálculos percentuais, por isso é uma medida relativa, e não absoluta. Assim, observe as fórmulas a seguir:

População	Amostra
$CV = \frac{\sigma}{\mu} \cdot 100$	$CV = \frac{s}{\bar{x}} \cdot 100$

A partir do valor do coeficiente de variação, podemos verificar se o conjunto de dados é homogêneo e também conseguimos saber se a média é uma boa medida para representar o conjunto de dados. Outra utilização para esta medida é comparar conjuntos com unidades de medidas distintas, uma vez que o CV é dado em porcentagem (%).

O CV tem o problema de deixar de ser explicativo da variação quando a média está perto de zero, pois esta situação pode deixá-lo alto demais. Um coeficiente de variação alto sugere alta variabilidade ou heterogeneidade do conjunto de observações. Quanto maior for este valor, menos representativa será a média.

Se isto acontecer, deve-se optar para representar os dados por outra medida, podendo ser essa a mediana ou moda, não existindo uma regra prática para a escolha de uma dessas. Fica, então, essa escolha a critério do pesquisador. Ao mesmo tempo, quanto mais baixo for o valor do CV, mais homogêneo é o conjunto de dados e mais representativa será sua média.

Quanto à representatividade em relação à média, podemos dizer que quando o coeficiente de variação (CV) é ou está (CRESPO, 2009):

- Menor que 10%: significa que é um ótimo representante da média, pois existe uma pequena dispersão (desvio padrão) dos dados em torno da média.
- Entre 10% e 20%: é um bom representante da média, pois existe uma boa dispersão dos dados em torno da média.
- Entre 20% e 35%: é um razoável representante da média, pois existe uma razoável dispersão dos dados em torno da média.
- Entre 35% e 50%: representa fracamente a média, pois existe uma grande dispersão dos dados em torno da média.
- Acima de 50%: não representa a média, pois existe uma grandíssima dispersão dos dados em torno da média.

Exemplo: vamos determinar o coeficiente de variação para o exemplo das idades dos indivíduos de uma família, sendo elas: 5, 10, 12, 35, 38.

Já efetuamos anteriormente os cálculos da média e do desvio-padrão: $\overline{x} = 20$ e s = 15,31. Logo, o coeficiente de variação para esse conjunto de dados é

$$CV = \frac{s}{\overline{x}} = \frac{15,31}{20} \times 100 = 76,55\%$$

Verificamos que há uma grande variação, ou seja, uma alta dispersão dos dados. Portanto, concluímos que a média não é uma boa representante desse conjunto de dados.

Exercício:

Calcule as medidas de dispersão para um grupo de indivíduos que têm as seguintes idades: 18, 19, 20, 21, 21, 22, 24, 24, 25, 27, 30, 33 e verifique se a média é uma medida que representa bem este conjunto de dados.

R: Média: 23,67; Variância: 20,42; desvio padrão amostral: 4,51; C.V.(%): 19%.

Observe que, para dados agrupados, há uma pequena diferença nas fórmulas de variância da população e da amostra:

Amostra

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

População

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

Em que cada x_i, é o ponto médio de cada classe estudada e F_i a frequência respectiva a cada classe, sendo μ e $\bar{x}$ as médias populacional e amostral, respectivamente.

Observe que a única diferença é que, com dados agrupados, os desvios ao quadrado devem ser multiplicados por suas respectivas frequências.

Exemplo: vamos determinar a variância para o seguinte conjunto de dados agrupados em tabelas de frequências.

Tabela 10 - Idades de 100 colaboradores de uma organização

IDADES	Fi	Xi	(Xi - MÉDIA)²* Fi
17 |----- 21	11	19	(19 – 36)² * 11 = 3.179
21 |----- 25	7	23	(23 – 36)² * 7 = 1.183
25 |----- 29	12	27	(27 – 36)² * 12 = 972
29 |----- 33	10	31	(31 – 36)² * 10 = 250
33 |----- 37	9	35	(35 – 36)² * 9 = 9
37 |----- 41	9	39	(39 – 36)² * 9 = 81
41 |----- 45	14	43	(43– 36)² * 14 = 686

45 |----- 49	21	47	$(47 - 36)^2 * 21 = 2.541$
49 |----- 53	4	51	$(51 - 36)^2 * 4 = 900$
53 |----- 57	3	55	$(55 - 36)^2 * 3 = 1.083$
Total	**100**	**370**	**10.884**

Fonte: dados fictícios elaborados pelas autoras.
Obs.: a média foi arredondada para 36.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 F_i}{n-1} = \frac{10884}{100 - 1} = 109{,}93$$

DESVIO PADRÃO

Para calcular o desvio padrão, o procedimento continua sendo o mesmo, ou seja, basta extrairmos a raiz quadrada da variância. Assim, observe as fórmulas:

População

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

Amostra

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

De forma mais simplificada:

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

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

REFLITA

Observe como é importante diferenciar população de amostra.

Exemplo

Considerando a situação exposta acima, em que a variância foi igual a s^2 = 109,93, o desvio padrão será:

$$S = \sqrt{109,93} = 10,48$$

E, consequentemente, o Coeficiente de variação será:

$$C.V.(\%) = \frac{10,48}{36} = 0,2911 * 100 = 29,11\%$$

Exercício:

Calcule as medidas de dispersão para dados agrupados considerando a tabela abaixo:

Tabela 11 - Distribuição de frequências para a idade dos clientes de uma loja em Maringá

CLASSES	Fi	Fr(%)	Fac	Xi
17 |---- 29	4	36,4	4	23
29 |---- 41	4	36,4	8	35
41 |----| 53	3	27,3	11	47
Total	**11**	**100**	-	-

Fonte: as autoras.

R: variância 99,49; desvio padrão 9,97; coeficiente de variação 29,40%

CONSIDERAÇÕES FINAIS

Nas pesquisas, após a coleta e organização dos dados, convém verificar o que ocorre com eles. Nos dados quantitativos, a principal forma de análise é calcular as medidas de posição e de dispersão.

Nesta unidade, você aprendeu a calcular as principais medidas de Posição e Dispersão, além das medidas Separatrizes. Vimos que as principais medidas de posição dentro da estatística são média aritmética, moda, mediana e separatrizes. Pelo menos uma dessas medidas sempre deve estar presente na descrição das informações coletadas.

As principais medidas de dispersão são variância, desvio padrão e coeficiente de variação. Analisamos que as medidas de dispersão são utilizadas para um estudo descritivo de um conjunto de dados numéricos qualquer, que têm por objetivo determinar a variabilidade ou a dispersão dos dados em relação à medida de localização do centro da amostra em análise.

Aprendemos o passo a passo de como calcular essa dispersão, diferenciando os cálculos de população e amostra. Vimos que, para calcular essa dispersão, precisamos da média; após calcularmos a média, calculamos a variância em relação à média, sendo que, para se calcular a variância, soma-se os quadrados dos desvios da amostra observada, em relação à média, e divide-se pelo número de observações da amostra menos um; o que a diferencia da população é que a divisão é feita somente pelo número de observações. Logo após, calculamos o desvio padrão, que é simplesmente a raiz quadrada da variância. O desvio padrão é uma medida de extrema importância, porque quanto maior for a variabilidade dos dados, maior será o valor do desvio padrão.

É importante salientar que, de todas essas medidas vistas, as mais utilizadas nas pesquisas são a média e o desvio padrão e que essas são representativas da população e da amostra também. As medidas representarão sempre os dados, portanto é fundamental que saibamos qual ou quais são as medidas mais adequadas para o tipo de informação que temos em mãos.

ATIVIDADES

1. Das medidas de posição vistas na unidade, explique:

a. Qual é a mais utilizada e por quê.

b. Quais são os problemas que a média pode ter em sua utilização como medida representativa de um conjunto de dados.

2. Considere os seguintes diâmetros (mm) de eixos produzidos em certa fábrica de autopeças:

93	94	96	100	96	102	89	87	105

Calcule:

a. A média aritmética, a moda e a mediana.

b. A variância, o desvio padrão.

c. O coeficiente de variação (interprete).

d. O 3º quartil e o 6º decil.

3. Considere a seguinte tabela de distribuição de frequências com os tempos (em dias) que um corretor demora a concluir um negócio, observado em 40 operações:

TEMPO (DIAS)	Fi	Fac	Xi
0 |– 2,5	2	2	1,25
2,5 |– 5,0	3	5	3,75
5,0 |– 7,5	25	30	6,25
7,5 |– 10,0	10	40	8,75
Total	**40**	-	-

Fonte: as autoras.

Calcule:

a. A média aritmética, a moda e a mediana.

b. A variância, o desvio padrão.

c. O coeficiente de variação (interprete).

d. O 3° quartil e o 4° percentil.

Anuário Estatístico de Acidentes do Trabalho – AEAT

A partir da experiência com a publicação do Anuário Estatístico de Previdência Social – AEPS, da crescente necessidade de informações sobre os acidentes de trabalho no país e considerando que a única fonte de dados disponível sobre esses acidentes era o documento Comunicação de Acidente do Trabalho – CAT, recebido pelo Instituto Nacional do Seguro Social – INSS a Secretaria de Políticas de Previdência Social – SPS decidiu produzir uma publicação que sistematizasse as informações coletadas pela CAT e agregasse dados oriundos dos sistemas de concessão de benefícios do INSS. Essa publicação, o Anuário Estatístico de Acidentes do Trabalho – AEAT, começou a ser editada em 2000 e também introduziu indicadores que permitiam a mensuração da evolução relativa da incidência de acidentes do trabalho e de outras características desses acidentes. Inicialmente os níveis de análise eram Brasil e UF, mas atendendo demandas de pesquisadores e do Ministério do Trabalho e Emprego foi introduzido um detalhamento dos acidentes em nível municipal.

Objetivo:

Tornar públicos dados estatísticos consistentes e confiáveis sobre os acidentes do trabalho ocorridos no Brasil, criando um registro histórico dessas estatísticas e subsidiando a formulação de políticas públicas e estudos relativos ao tema.

População Alvo: -

Abrangência Geográfica:

Nacional.

Metodologia:

O AEAT é produzido basicamente a partir de dados armazenados na Empresa de Tecnologia e Informações da Previdência Social – DATAPREV. Os dados são provenientes das Comunicações de Acidentes do Trabalho encaminhadas ao INSS e de dados de benefícios por acidente do trabalho concedidos pelo INSS. Para o cálculo dos indicadores, são utilizados dados da base de contribuintes do INSS. Uma vez ao ano são ativados processos automáticos que fazem a extração dos dados das bases transacionais. Esses dados são avaliados e com eles preparadas as tabelas publicadas no AEAT.

Principais Variáveis:

Acidentes do Trabalho:

Registrado: Para Brasil, UF e Regiões: Número de Acidentes do Trabalho por Motivo, Tipo de Registro, Mês do Ano, Grupos de Idade, Sexo, Classe de Atividade Econômica, Para Brasil e UF: 200 códigos CID mais incidentes.

Para Municípios: Quantidade de Acidentes do Trabalho por Motivo, Tipo de Registro e Número de Óbitos.

Liquidado: Para Brasil, UF e Regiões: Número de Acidentes do Trabalho Liquidados, Consequência do Acidente, Classe de Atividade Econômica, Mês do Ano.

Indicadores: Para Brasil e UF: Indicadores de Incidência, Incidência de Doenças do Trabalho, Incidência de Acidentes Típicos, Incidência de Incapacidade Temporária, Taxa de Mortalidade, Taxa de Letalidade, Taxa de Acidentalidade Proporcional Específica para a Faixa Etária de 16 a 34 anos, Classe de Atividade Econômica.

Observação: Nem todas as variáveis estão disponíveis em todos os anos de publicação do AEAT.

Documentação Operacional: -

Época da Coleta:

Meses de Maio a Julho do ano seguinte ao de competência.

Tempo Previsto entre o Início da Coleta e a Liberação dos Dados:

Sete meses.

Nível de Divulgação:

Brasil, Grandes Regiões, UF e Municípios.

Formas de Disseminação:

Há 4 formas de disseminação: volumes impressos, CD-ROM, tabelas na Internet e tabulador de dados na Internet.

Fonte: adaptado de IBGE (2017, on-line).

MATERIAL COMPLEMENTAR

LIVRO

Estatística Aplicada

Douglas Downing; Jeffrey Clarck

Editora: Saraiva

Sinopse: este livro aborda assuntos, técnicas estatísticas e suas aplicações, estatística descritiva, probabilidades, teste de hipóteses, pesquisa e amostragem, regressão linear simples e múltipla, métodos não-paramétricos, indicadores econômicos e teoria da decisão. Os capítulos começam com os "Termos-chave", trazendo um resumo dos conceitos fundamentais de cada capítulo. A seção "Lembre-se" retoma, ao longo do estudo, tópicos essenciais a serem fixados, e no "Conheça os conceitos", encontram-se exercícios para aplicação do aprendizado.

REFERÊNCIAS

CRESPO, A. A. **Estatística Fácil**. São Paulo: Saraiva, 19. ed., 2009.

IBGE (2017). **Anuário Estatístico de Acidentes do Trabalho – AEAT**. Disponível em: <http://ces.ibge.gov.br/base-de-dados/metadados/mps/anuario-estatistico-de-acidentes-do-trabalho-aeat.html>. Acesso em: 24 abr. 2017.

NETO, A. R. **Conceito de média**: A média ponderada é também uma média aritmética. 2009. Disponível em: <https://educacao.uol.com.br/disciplinas/matematica/conceito-de-media-a-media-ponderada-e-tambem-uma-media-aritmetica.htm>. Acesso em: 24 abr. 2017.

GABARITO

1.

a) A média é mais utilizada, pois é a medida mais precisa, é única em um conjunto de dados e sempre existe.

b) Os problemas da média ocorrem porque ela é afetada por medidas extremas, ou seja, valores muito altos ou muito baixos, destoando da maioria dos outros valores, podem comprometer o valor da média. Além disso, em conjuntos de dados muito heterogêneos, ela não é uma medida que representa bem o conjunto de dados.

2.

a) A média aritmética: $\bar{x}$ =95,8 mm

A moda: Mo = 96 mm

A mediana: Md = 96 mm

b) Variância: $s^2 = 34,5\ mm^2$

Desvio padrão: s = 5,9 mm

c) CV = 6,2%. Temos uma baixa dispersão dos dados em torno da média, logo esta é uma ótima representante do conjunto de dados.

d) Q_3 = 101 mm, portanto, 75% dos diâmetros dos eixos estão abaixo de 101 mm.

D_6 = 96 mm, portanto, 60% dos diâmetros dos eixos estão abaixo de 96 mm.

3.

a) $\bar{x}$ = 6,44 dias

Mo = 6,48 dias

Md = 6,5 dias

b) $s^2 = 3,33\ dias^2$

s = 1,82 dias

c) CV = 28,26%. Temos uma dispersão razoável dos dados em torno da média, logo esta é uma representante aceitável do conjunto de dados.

d) Q_3 = 7,5 dias, portanto, 75% da demora de concluir um negócio está abaixo de 7,5 dias.

P_4 = 2 dias, portanto, 4% da demora de concluir um negócio está abaixo de 2 dias.

Professora Me. Ivnna Gurniski de Oliveira
Professora Me. Renata Cristina de Souza Chatalov

PROBABILIDADES

UNIDADE IV

Objetivos de Aprendizagem

- Entender os conceitos relacionados a probabilidades.
- Saber aplicar as probabilidades nas diversas situações.
- Compreender probabilidade condicional.
- Conhecer as principais distribuições de probabilidades.

Plano de Estudo

A seguir, apresentam-se os tópicos que você estudará nesta unidade:

- Probabilidade
- Distribuição de Probabilidades Discreta
- Distribuição de Probabilidades Contínua

Unicesumar

INTRODUÇÃO

Nesta unidade, vamos tratar das probabilidades. Quando estamos falando de probabilidade, queremos identificar a chance de ocorrência de um determinado resultado de interesse em situações nas quais não é possível calcular com exatidão o valor real do evento. Então, desta forma, trabalhamos com chances ou com probabilidades.

A palavra probabilidade deriva do Latim *probare* (provar ou testar), e designa eventos incertos, ou mesmo "sorte", "risco", "azar", "incerteza" ou "duvidoso". A probabilidade como ramo da matemática data de mais de 300 anos e se aplicava a jogos de azar, em que jogadores que tinham mais conhecimento sobre suas teorias planejavam estratégias para levar vantagem nos jogos. Hoje, essa prática ainda é utilizada, porém também passou a ser empregada por governos, empresas e organizações profissionais nas suas tomadas de decisões ou ainda na escolha de produtos, sendo úteis também para o desenvolvimento de estratégias.

As decisões nos negócios são frequentemente baseadas na análise de incertezas, tais como: chances de um investimento ser lucrativo, chances das vendas decrescerem se o preço for aumentado, probabilidade de projetos terminarem no prazo, etc. As probabilidades medem o grau de incerteza, assim, não podemos antecipar o evento, mas lidar com as chances maiores ou menores dele ocorrer.

Nesta unidade, serão apresentados conceitos básicos de probabilidade, como a probabilidade pode ser interpretada e como suas regras podem ser utilizadas para calcular as possibilidades de ocorrência de eventos futuros, além de trabalharmos com as principais distribuições de probabilidades discretas e contínuas. Veremos a importância de estudarmos as probabilidades, pois é necessário que os futuros gestores saibam que muitas das decisões a serem tomadas são baseadas na incerteza.

PROBABILIDADE

As probabilidades são utilizadas para delinear a chance de ocorrência de determinado evento. Seus valores são sempre atribuídos numa escala de 0 a 1. A probabilidade próxima de 1 indica um evento quase certo, enquanto que a probabilidade próxima de zero indica um evento improvável de acontecer.

Ao discutirmos probabilidade, definimos experimentos como qualquer ação ou processo que gera resultados bem definidos. Os experimentos aleatórios são aqueles que, repetidos várias vezes, apresentam resultados imprevisíveis. Ao descrever um experimento aleatório, deve-se sempre especificar o que deverá ser observado.

ANÁLISE DE RISCO E A PROBABILIDADE

A análise qualitativa de risco é definida como o processo de avaliação do impacto e probabilidade de riscos identificados. Este processo prioriza riscos de acordo com os seus efeitos potenciais nos objetivos do projeto. Análise qualitativa de risco é um modo de determinar a importância de se endereçar riscos específicos e guiar respostas de risco. A questão crítica do tempo e as ações relacionadas ao risco podem ampliar a importância de um risco (SILVEIRA; ORTH; PRINKLANDICKI, 2006).

Essa análise qualitativa de risco requer que a probabilidade e as consequências dos riscos sejam avaliadas, usando métodos e ferramentas de análise qualitativa estabelecidos. Tendências nos resultados, quando a análise qualitativa é repetida, pode indicar a necessidade de mais ou menos ação da gerência de risco. O uso dessas ferramentas ajuda a corrigir influências que estão frequentemente presentes em um plano de projeto. (SILVEIRA; ORTH; PRINKLANDICKI, 2006).

EXEMPLOS PRÁTICOS DE PROBABILIDADES

Queremos estudar a ocorrência das faces de um dado. Esse seria o experimento aleatório. A partir do conhecimento de que o dado tem 6 faces, sendo o dado equilibrado, de modo a não favorecer nenhuma das faces, podemos construir o modelo probabilístico da seguinte maneira:

Tabela 1 - Modelo probabilístico do lançamento de um dado

Face	1	2	3	4	5	6
Frequência	1/6	1/6	1/6	1/6	1/6	1/6

Fonte: as autoras.

Se o experimento aleatório for o lançamento de uma moeda. Sabendo que só podem ocorrer duas situações ao lançamento dela: cara ou coroa, o modelo probabilístico para esta situação seria:

Tabela 2 - Modelo probabilístico do lançamento de uma moeda

Face	Cara	Coroa
Frequência	1/2	½

Fonte: as autoras.

Se um grupo for composto por 20 homens e 30 mulheres e um deles for sorteado ao acaso para ganhar um determinado prêmio, o modelo probabilístico será:

Tabela 3 - Modelo probabilístico do sorteio de um prêmio

Indivíduo	Homem	Mulher
Frequência	20/50	30/50

Fonte: as autoras.

Verificamos que em todos os exemplos mostrados, precisamos ter um modelo probabilístico. Um modelo probabilístico envolve os conceitos de espaço amostral e eventos. Vejamos a seguir suas características.

Espaço amostral

Chamamos de espaço amostral o conjunto de todos os resultados possíveis de um experimento. Os elementos do espaço amostral são chamados de *pontos amostrais*. Representamos o espaço amostral por **Ω**.

Exemplo: considere o lançamento de uma moeda. Os possíveis resultados (n) são dois: cara (c) ou coroa (k). Então, o espaço amostral é dado por **Ω** = {***c***,***k***}. Se quisermos lançar a moeda duas vezes, os possíveis resultados são quatro: cara e cara; cara e coroa; coroa e cara; coroa e coroa. Logo, o espaço amostral é **Ω** = {***cc***,***ck***,***kc***,***kk***}.

Eventos

Chamamos de evento um subconjunto do espaço amostral Ω de um experimento aleatório. O evento é dito simples se consistir em um único resultado, ou composto se consistir em mais de um resultado.

Exemplo

No lançamento de uma moeda Ω = {cara, coroa}. Um evento de interesse A pode ser "obter cara no lançamento de uma moeda" e então A = {cara} e o n para este evento será 1, sendo n o número de resultados para o evento.

No lançamento de um dado, o evento de interesse A pode ser obter face par e então A será igual a:

$$A = \{2; 4, 6\} \text{ e } n = 3.$$

Probabilidade de um evento

Podemos fazer cálculos de probabilidades utilizando três formas distintas:

- Método clássico – quando o espaço amostral tem resultados equiprováveis.
- Método empírico – baseado na frequência relativa de um grande número de experimentos repetidos.
- Método subjetivo – baseia-se em estimativas pessoais de probabilidade com certo grau de crença.

Utilizaremos aqui o método clássico.

Considerando um experimento aleatório em que se queira um determinado evento A, a probabilidade deste evento ocorrer é dada por P(A).

Assim: a probabilidade de A ocorrer será dada por:

$P(A) = \frac{n(A)}{\Omega}$, para qualquer evento discreto. Ou seja,

Considere um experimento aleatório em que se queira determinar um evento E. A probabilidade de este evento ocorrer, denotada por P(E), é dada pela razão do número de resultados do evento E, ($\boldsymbol{n(E)}$), pelo número total de resultados no espaço amostral, ($\boldsymbol{\Omega}$). Isto é,

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

Por exemplo, considere o lançamento de um dado. Queremos calcular a probabilidade de obtermos uma face ímpar (evento A) e a probabilidade de sair as faces 2 e 5 (evento B).

Primeiro vamos determinar o espaço amostral, que é composto por todos os resultados possíveis:

$\boldsymbol{\Omega = \{1, 2, 3, 4, 5, 6\}, n = 6}$

Em seguida, determinamos os resultados possíveis para os eventos A e B:

$\boldsymbol{A = \{1, 3, 5\}, n = 3}$
$\boldsymbol{B = \{2, 5\}, n = 2}$

Assim,

$\mathbf{P(A) = \frac{n(A)}{\Omega} = \frac{3}{6} = 0,5}$ **ou, em porcentagem,** $\mathbf{P(A) = 0,5 \times 100 = 50\%.}$

$\mathbf{P(B) = \frac{n(B)}{\Omega} = \frac{2}{6} = 0,33}$ **ou, em porcentagem,** $\mathbf{P(B) = 0,33 \times 100 = 33\%.}$

REGRAS BÁSICAS

Tendo um modelo probabilístico e conhecendo suas frequências relativas, podemos estabelecer no cálculo das probabilidades algumas regras:

- A probabilidade deverá ser um valor que varie entre 0 e 1, sendo representado por:

$$0 < P(A) < 1$$

- Um evento impossível é um conjunto vazio (φ) e atribui-se probabilidade 0, enquanto que um evento certo tem probabilidade 1, assim:

$$P(\Omega) = 1 \qquad P(\varphi) = 0$$

- A soma das probabilidades para todos os resultados experimentais tem de ser igual a 1.

OPERAÇÕES COM EVENTOS

Nos cálculos de probabilidades, algumas vezes, o interesse do pesquisador está na determinação da probabilidade de combinação dos eventos relacionados ao experimento aleatório. Podemos ter dois tipos de combinações, dados dois eventos A e B:

- O evento intersecção de A e B, denotado $A \cap B$, é o evento em que A e B ocorrem simultaneamente.
- O evento reunião de A e B, denotado $A \cup B$, é o evento em que A ocorre ou B ocorre (ou ambos).
- O evento complementar de A, denotado A^c, é o evento em que A não ocorre.

Assim:

- A probabilidade de um ou outro evento ocorrer é dada por $P(A \cup B)$.
- A probabilidade de ambos os eventos ocorrerem simultaneamente é dada por $P(A \cap B)$.

Exemplo: considere um baralho completo de 52 cartas. Desejamos saber qual é a probabilidade de sair um rei de copas. Para este evento, vamos calcular a probabilidade **P(A∩B)**, na qual A é a probabilidade da carta ser um rei e B é a probabilidade da carta ser de copas. Se desejarmos saber a probabilidade de sair uma carta de valor 2 ou uma carta de valor 5, vamos calcular a probabilidade **P(C∪D)**, na qual C é a probabilidade de sair uma carta de valor 2 e D é a probabilidade de sair uma carta de valor 5.

Regra da adição

Essa regra leva em consideração a ocorrência do evento A ou a ocorrência do evento B ou ainda de ambos os eventos. É denotada matematicamente por P(A∪B) e dizemos união de A e B que é a probabilidade de ocorrência de pelo menos um dos dois eventos.

No cálculo dessa probabilidade, surgem duas situações:

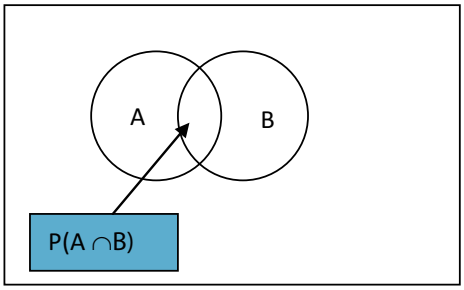
A primeira quando os eventos A e B são mutuamente excludentes (não têm elementos em comum). Nesta situação, a fórmula é dada por:

$$P(A \cup B) = P(A) + P(B)$$

A segunda, quando os eventos A e B não são mutuamente excludentes (têm elementos em comum). Nesta situação, a fórmula é dada por:

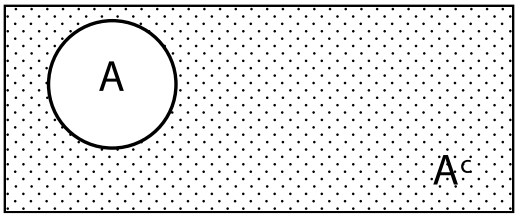
$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \text{ em que:}$$

$P(A \cap B)$ – é a probabilidade de A e B ocorrerem simultaneamente; a intersecção entre os eventos A e B.



COMPLEMENTO DE UM EVENTO

Dado um evento A, o complemento de A (A^c) é um evento que consiste de todos os pontos amostrais que não estão em A (TOLEDO; OVALLE, 1997). O diagrama abaixo ilustra o conceito. A área retangular representa o espaço amostral; o círculo representa o evento A e a região com preenchimento, os pontos do complemento de A.



O cálculo da probabilidade usando o complemento é feito por meio da relação:

$$P(A^C) = 1 - P(A) \text{ para todo evento A.}$$

REFLITA

A teoria da probabilidade é no fundo nada mais do que o senso comum reduzido ao cálculo.

(Pierre Simon de Laplace)

Exemplo:

Considere o lançamento de um dado e os seguintes eventos: sair faces pares (A), sair faces ímpares (B), sair faces cujo valor é maior do que 3 (C). Ou seja,

$$\Omega = \{1, 2, 3, 4, 5, 6\} \qquad n(\Omega) = 6$$
$$A = \{2, 4, 6\} \qquad n(A) = 3$$
$$B = \{1, 3, 5\} \qquad n(B) = 3$$
$$C = \{4, 5, 6\} \qquad n(C) = 3$$

Vamos calcular as seguintes probabilidades: $P(A \cap B)$, $P(A \cap C)$, $P(A \cup B)$ e $P(A^c)$. Temos que:

$$A \cap B = \{\} \qquad n(A \cap B) = 0$$
$$A \cap C = \{4, 6\} \qquad n(A \cap C) = 2$$
$$A \cup B = \{1, 2, 3, 4, 5, 6\} = \Omega \qquad n(A \cup B) = 6$$
$$A^c = \{1, 3, 5\} \qquad n(A^c) = 3$$

Logo,

$$P(A \cap B) = \frac{0}{6} = 0$$

$$P(A \cap C) = \frac{2}{6} = 0{,}33$$

$$P(A \cup B) = \frac{6}{6} = 1$$

$$P(A^c) = \frac{3}{6} = 0{,}5$$

Exercícios

1. Determine o espaço amostral no lançamento de um dado.

2. Considere o experimento aleatório do lançamento de dois dados e:
   a. Encontre Ω.
   b. Demonstre o n(A) e P(A) nos casos abaixo.

      A1: apareçam faces iguais.

      A2: a segunda face é o dobro da primeira.

      A3: apareçam somente números ímpares.

      A4: apareçam faces iguais ou a segunda face é o quadrado da primeira.

      A5: a soma das faces é igual a 7.

3. Um consultor está estudando dois diferentes tipos de imóveis quanto a quantidades disponíveis à venda, por região, em uma determinada cidade. Os dados são mostrados abaixo:

Tabela 4 - Tipos de imóveis e apartamentos por região

REGIÃO	TIPO DE IMÓVEL		TOTAL
	APARTAMENTO	CASA	
Norte	30	28	58
Sul	40	56	96
Leste	38	34	72
Oeste	52	22	74
Total	**160**	**140**	**300**

Fonte: as autoras.

Considere Norte por N; Sul por S; Leste por L; Oeste por O; Apartamento por A e Casa por C.

Calcule as seguintes probabilidades:

a. $P(N) =$

b. $P(S) =$

c. $P(L) =$

d. $P(O) =$

e. $P(A) =$

f. $P(C) =$

g. $P(N \cap A) =$

h. $P(S \cap C) =$

i. $P(L \cap A) =$

j. $P(O \cap C) =$

k. $P(N \cup A) =$

l. $P(S \cup C) =$

m. $P(L \cup A) =$

n. $P(O \cup C) =$

Resposta:

1. $\Omega = \{1, 2, 3, 4, 5, 6\}$

2.
   a. $\Omega = \{(1,1)\ (1,2)\ (1,3)\ (1,4)\ (1,5)\ (1,6)\ (2,1)\ (2,2)\ (2,3)\ (2,4)\ (2,5)\ (2,6)\ (3,1)\ (3,2)\ (3,3)\ (3,4)\ (3,5)\ (3,6)\ (4,1)\ (4,2)\ (4,3)\ (4,4)\ (4,5)\ (4,6)\ (5,1)\ (5,2)\ (5,3)\ (5,4)\ (5,5)\ (5,6)\ (6,1)\ (6,2)\ (6,3)\ (6,4)\ (6,5)\ (6,6)\}$

   b. $A1 = \{(1,1)\ (2,2)\ (3,3)\ (4,4)\ (5,5)\ (6,6)\}$

   $P(A1) = 6/36 = 0,17$

   $A2 = \{(1,2)\ (2,4)\ (3,6)\}$

   $P(A2) = 3/36 = 0,08$

   $A3 = \{(1,1)\ (1,3)\ (1,5)\ (3,1)\ (3,3)\ (3,5)\ (5,1)\ (5,3)\ (5,5)\}$

   $P(A3) = 9/36 = 0,25$

   $A4 = \{(1,1)\ (2,2)\ (2,4)\ (3,3)\ (4,4)\ (5,5)\ (6,6)\}$

   $P(A4) = 7/36 = 0,19$

   $A5 = \{(1,6)\ (2,5)\ (3,4)\ (4,3)\ (5,2)\ (6,1)\}$

   $P(A5) = 6/36 = 0,17$

3.

a. $P(N) = 58/300 = 0,1933$
b. $P(S) = 96/300 = 0,320$
c. $P(L) = 72/300 = 0,240$
d. $P(O) = 74/300 = 0,247$
e. $P(A) = 160/300 = 0,533$
f. $P(C) = 140/300 = 0,467$
g. $P(N \cap A) = 30/300 = 0,100$
h. $P(S \cap C) = 56/300 = 0,187$
i. $P(L \cap A) = 38/300 = 0,127$
j. $P(O \cap C) = 22/300 = 0,073$
k. $P(N \cup A) = (58/300)+(160/300)-(30/300)= 188/300 = 0,627$
l. $P(S \cup C) = 0,32+0,467-0,187 = 0,60$
m. $P(L \cup A) = 0,240+0,533-0,127 = 0,647$
n. $P(O \cup C) = 0,247+0,467-0,073 = 0,641$

PROBABILIDADE CONDICIONAL

Frequentemente, a probabilidade de um evento é influenciada pela ocorrência de um evento paralelo. Seja A um evento com probabilidade P(A). Se obtivermos a informação extra que o evento B ocorreu paralelamente, iremos tirar vantagem dela no cálculo de uma nova probabilidade para o evento A. Esta será escrita como $P(A \mid B)$ e lida como "probabilidade de A dado B".

Neste caso, podemos utilizar esta informação extra para realocar probabilidades aos outros eventos. Vamos utilizar o exemplo da tabela do exercício anterior.

Tabela 5 - Tipos de imóveis e apartamentos por região

REGIÃO	TIPO DE IMÓVEL		TOTAL
	APARTAMENTO	CASA	
Norte	30	28	58
Sul	40	56	96
Leste	38	34	72
Oeste	52	22	74
Total	**160**	**140**	**300**

Fonte: as autoras.

Se soubermos que o imóvel é um apartamento, qual é a chance de ser da região norte? Reformulando a pergunta, poderíamos ter o interesse de saber: dado que o imóvel é um apartamento, qual a probabilidade de pertencer à região norte? Observe que estamos impondo uma condição ao evento. Sabemos que o imóvel é um apartamento, essa é a condição imposta. Quando impomos alguma condição em probabilidade, dizemos então que a probabilidade é condicional e, assim, reduzimos então o espaço amostra à condição imposta.

Assim, escrevemos:

$P(N \mid A)$ e lê-se probabilidade de N dado A, sendo a condição A, ou seja, ser apartamento, sendo que:

$$P(N \mid A) = \frac{30}{160}$$

De forma geral, para dois eventos quaisquer A e B, sendo $P(B) > 0$, definimos a probabilidade condicional de $A \mid B$ como sendo $P(A \mid B)$ dado pela seguinte fórmula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Caso a condição seja A:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Para o exemplo acima mencionado, se N e A indicam, respectivamente, norte para região e apartamento para tipo, então:

$$P(N \mid A) = \frac{P(N \cap A)}{P(A)} = \frac{30/300}{160/300} = \frac{30}{160}$$ como mostrado acima.

Observe que, se trocarmos a condição para ser do tipo A, dado que a região é Norte, a condição agora é ser da região Norte e o problema ficaria da seguinte maneira:

$$P(N \mid A) = \frac{P(N \cap A)}{P(N)} = \frac{30/300}{58/300} = \frac{30}{58}$$

Exercícios

Baseado na tabela acima, calcular as seguintes probabilidades:

a. $P(S \mid C) =$

b. $P(C \mid S) =$

c. $P(L \mid A) =$

d. $P(A \mid L) =$

e. $P(O \mid C) =$

f. $P(C \mid O) =$

Resposta: a. 56/140 = 0,4 b. 56/96 = 0,5833 c. 38/160 = 0,2375 d. 38/72 = 0,5278 e. 22/140 = 0,1571 f. 22/74 = 0,2973

EVENTOS INDEPENDENTES

Dois eventos A e B são independentes se $P(A \mid B) = P(A)$ ou $P(A \mid B) = P(B)$. Caso contrário, os eventos são dependentes.

Regra da multiplicação

A relação geral mostrada acima foi:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Desta relação, obtemos a regra do produto das probabilidades, em que:

$$P(A \cap B) = P(B) \cdot P(A \mid B).$$

Observe que a probabilidade de A e B ocorrerem conjuntamente está sob uma condição, pois a probabilidade de A está sob a condição de B, mostrando que há uma dependência de uma probabilidade em relação ao evento ocorrido anteriormente.

Em caso de A e B serem eventos independentes, ou seja, a probabilidade de um evento não depender da ocorrência do outro evento, nesta condição, a probabilidade de A e B ocorrer é dada pela probabilidade de A vezes a probabilidade de B.

$$P(A \cap B) = P(A) \cdot P(B)$$

Exemplo:

Uma urna contém duas bolas brancas e três bolas pretas. Sorteamos duas bolas ao acaso sem reposição. Isto quer dizer que sorteamos a primeira bola, verificamos sua cor e não a devolvemos à urna. As bolas são novamente misturadas e sorteamos então a segunda bola. Para resolver as probabilidades nesta situação, ilustraremos a situação por um diagrama de árvore em que em cada "galho da árvore" estão indicadas as probabilidades.

Diagrama de árvore

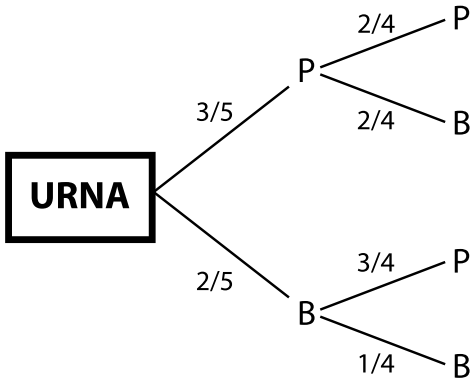


Figura 1 - Diagrama de árvores para o sorteio de duas bolas sem reposição
Fonte: as autoras.

Observe que o cálculo das probabilidades, na segunda retirada, fica condicionado aos resultados da primeira retirada. Na Tabela 24, estão os resultados possíveis do sorteio com suas respectivas probabilidades.

Indicando bola branca por B e bola preta por P, vejamos o cálculo das probabilidades para as seguintes situações:

Tabela 6 - Resultados e probabilidades do diagrama de árvore

RESULTADOS	PROBABILIDADES
BB	2/5 x 1/4 = 2/20
BP	2/5 x 3/4 = 6/20
PB	3/5 x 2/4 = 6/20
PP	3/5 x 2/4 = 6/20
Total	**1,0**

Fonte: as autoras.

Considere agora que vamos fazer o mesmo sorteio, mas repondo a primeira bola sorteada novamente na urna. Assim, as probabilidades são:

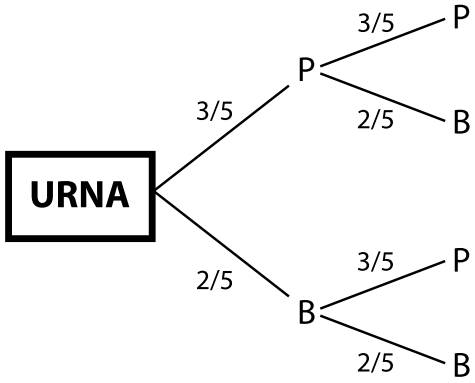


Figura 2 - Diagrama de árvores para o sorteio de duas bolas com reposição
Fonte: as autoras.

a. Qual é a probabilidade de sair bola branca na primeira retirada?

$$P(B) = 2/5$$

b. Qual é a probabilidade de sair bola branca na primeira retirada e bola preta na segunda retirada?

$$P(B \text{ na } 1^a \cap P \text{ na } 2^a) = 6/20$$

c. Qual é a probabilidade de sair bola preta na segunda retirada, dado que saiu branca na primeira retirada?

$$P(P \text{ na } 2^a \mid B \text{ na } 1^a) = 3/4$$

d. Qual é a probabilidade de sair bola branca na segunda retirada, dado que saiu preta na primeira retirada?

$$P(B \text{ na } 2^a \mid P \text{ na } 1^a) = 2/4$$

e. Qual é a probabilidade de sair bola preta na segunda retirada?

$$P(P \text{ na } 2^a) = 6/20 + 6/20 = 12/20$$

Observe que os cálculos das probabilidades na segunda retirada não ficam condicionados aos resultados da primeira retirada.

Tabela 7 - Resultados e probabilidades do diagrama de árvore

RESULTADOS	PROBABILIDADES
BB	2/5 x 2/5 = 4/25
BP	2/5 x 3/5 = 6/25
PB	3/5 x 2/5 = 6/25
PP	3/5 x 3/5 = 9/25
Total	**1,0**

Fonte: as autoras.

Observe que os cálculos das probabilidades na segunda retirada não ficariam condicionados aos resultados da primeira retirada. Assim, indicando B por "branca" e P por "preta", vejamos o cálculo das probabilidades.

a. Qual é a probabilidade de sair bola branca na primeira retirada?

$$P(B) = 2/5$$

b. Qual é a probabilidade de sair bola branca na primeira retirada e bola preta na segunda retirada?

$$P(B \text{ na } 1^a \cap P \text{ na } 2^a) = 6/25$$

c. Qual é a probabilidade de sair bola preta na segunda retirada, dado que saiu branca na primeira retirada?

$$P(P \text{ na } 2^a \mid B \text{ na } 1^a) = 3/5$$

d. Qual é a probabilidade de sair bola branca na segunda retirada, dado que saiu preta na primeira retirada?

$$P(B \text{ na } 2^a \mid P \text{ na } 1^a) = 2/5$$

e. Qual é a probabilidade de sair bola preta na segunda retirada?

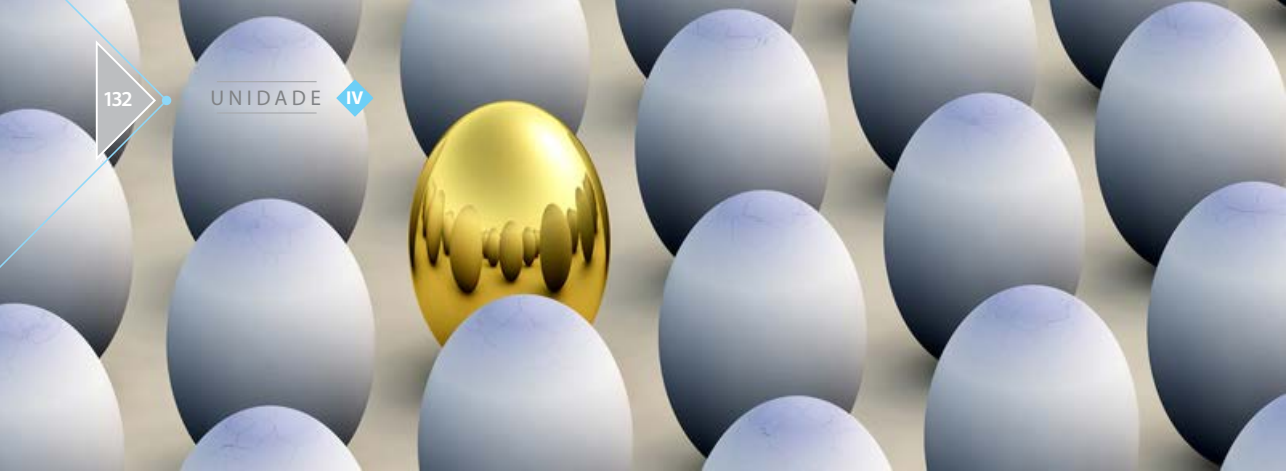
$$P(P \text{ na } 2^a) = 6/25 + 9/25 = 15/25 = 3/5$$

Observe que as probabilidades da segunda retirada não são alteradas pela extração da primeira bola.

$$\text{Assim, } P(P \text{ na } 2^a \mid B \text{ na } 1^a) = 3/5 = P(P \text{ na } 2^a).$$

Nesse caso, dizemos que o evento A independe do evento B e:

$$P(A \cap B) = P(A) \cdot P(B)$$



REGRAS BÁSICAS DA PROBABILIDADE

De acordo com o evento estudado, existem algumas regras para o cálculo de probabilidades. São elas:

P(A ou B), para eventos não mutuamente excludentes:

$$P(A \text{ ou } B \text{ ou ambos}) = P(A) + P(B) - P(A \text{ e } B)$$

Para eventos mutuamente excludentes:

$$P(A \text{ ou } B) = P(A) + P(B)$$

Para eventos independentes:

$$P(A \text{ e } B) = P(A) \cdot P(B)$$

Para eventos dependentes:

$$P(A \text{ e } B) = P(B).P(A \mid B) \text{ ou } P(A).P(B \mid A)$$

Exercícios:

1. Uma urna contém 5 bolas pretas e 4 azuis. Em duas extrações consecutivas, sem reposição, determine os resultados esperados e calcule as seguintes probabilidades:

   a. De retirar a primeira azul e a segunda preta.

   b. De retirar a primeira azul e a segunda azul.

   c. De retirar a segunda azul, dado que a primeira foi preta.

2. Em um lote de 15 peças, sendo 5 defeituosas, retira-se uma peça e inspeciona-se. Qual a probabilidade:

   a. Da peça ser defeituosa?

   b. Da peça não ser defeituosa?

3. Uma loja dispõe de cartuchos de tintas novas e recondicionadas. Entre 30 cartuchos, sabe-se que 10 são recondicionados.

   a. Se um cliente levar um cartucho, qual a probabilidade de que ele seja recondicionado?

   b. Se um cliente levar dois cartuchos, qual a probabilidade de que ambos sejam recondicionados?

   c. Se um cliente levar 4 cartuchos, qual a probabilidade de que todos sejam recondicionados?

Resposta:

1. a. 0,278

   b. 0,167

   c. 0,5

2. a. 0,333

   b. 0,667

3. a. 0,333

   b. 0,103

   c. 0,008

Distribuições De Probabilidade

Os métodos de análise estatística requerem sempre que sejam enfocados certos aspectos numéricos dos dados (média, desvio padrão, etc.), independentemente de o experimento originar resultados qualitativos ou quantitativos.

Um meio para descrever, por valores numéricos, os resultados experimentais é o conceito de Variável Aleatória.

Uma variável aleatória permite passar cada um dos resultados do experimento para uma função numérica dos resultados. Para ilustrar, em uma amostra de componentes, ao invés de manter o registro de falhas individuais, o pesquisador pode registrar apenas quantos apresentaram falhas dentro de mil horas. Em geral, cada resultado é associado por um número, especificando-se uma regra de associação (TRIOLA, 1999).

Uma variável aleatória pode ser classificada como discreta ou contínua, dependendo dos valores numéricos que ela assume. Uma variável aleatória é:

- Discreta: quando pode assumir tanto um número finito de valores como uma infinita sequência de valores, tais como 0, 1, 2, ..., n.
- Contínua: quando pode assumir qualquer valor numérico em um intervalo ou associação de intervalos.

Observe o exemplo do lançamento de uma moeda duas vezes. A variável aleatória é o "número de caras" em duas jogadas. Considerando C como sair cara e K como sair coroa, os possíveis resultados são:

Tabela 8 - Distribuição de probabilidades Cara ou Coroa

RESULTADOS	VALOR DA VARIÁVEL ALEATÓRIA (SAIR CARA)	PROBABILIDADE DO RESULTADO
C C	2	½ x ½ = ¼
C K	1	½ x ½ = ¼
K C	1	½ x ½ = ¼
K K	0	½ x ½ = ¼

Fonte: as autoras.

A distribuição de probabilidades ficará:

Tabela 9 - Distribuição de probabilidades

VALOR DA VARIÁVEL ALEATÓRIA (SAIR CARA)	PROBABILIDADE DO RESULTADO
0	¼
1	¼ + ¼ = 2/4
2	¼
Total	**1,0**

Fonte: as autoras.

Para cada possível evento, associamos um número e em seguida montamos o modelo probabilístico. Assim, conhecemos a distribuição de probabilidades que essa variável aleatória (v.a.) segue.

DISTRIBUIÇÃO DE PROBABILIDADES DISCRETA

Existem experimentos cujos resultados, refletidos em uma variável aleatória, seguem um comportamento previsível em relação às suas probabilidades de ocorrência e, portanto, podem ser modelados por uma equação específica.

Dentre as principais distribuições discretas, destacam-se a Distribuição de Bernoulli, Distribuição Binomial e Distribuição de Poisson.

DISTRIBUIÇÃO DE BERNOULLI

A distribuição de Bernoulli consiste em uma distribuição em que a variável aleatória assume apenas dois possíveis resultados: sucesso (o evento se realiza) ou fracasso (o evento não se realiza).

Exemplos:

- Lançamento de uma moeda: o resultado é cara ou não.
- Uma peça é escolhida ao acaso: o resultado é defeituosa ou não.
- Uma cidade tem esgotamento sanitário: sim ou não.

Deve ficar claro que nem sempre o que é "bom" é o sucesso, mas sim o que se está estudando. Assim, o fato da peça ser defeituosa, por exemplo, seria o sucesso da pesquisa em si.

Em todos os casos, definimos uma variável aleatória X que só assuma dois valores possíveis:

$$X = \begin{cases} 0 \rightarrow \textbf{fracasso} \\ 1 \rightarrow \textbf{sucesso} \end{cases}$$

onde $\mathbf{P\ (X = 0) = q}$ e $\mathbf{P(X = 1) = p}$.

A função probabilidade de Bernoulli é dada por:

$$\mathbf{P(X = k) = p^k \cdot q^{1-k}}$$

O cálculo da média (chamada de Esperança e denotada por E(X)) e da variância (Var(X)) e do desvio padrão (σ) para a distribuição de Bernoulli são:

$$\mathbf{E(X) = p}$$
$$\mathbf{Var(X) = pq}$$
$$\mathbf{\sigma(X) = \sqrt{pq}}$$

Exemplo:
Supondo que a probabilidade de venda amanhã seja de 0,8.
Seja a variável aleatória "vender", temos que:

- A probabilidade de não vender este produto é:

  $\mathbf{P\ (X = 0) = q}$

  $\mathbf{P(X = 0) = 1 - p}$

  $\mathbf{P(X = 0) = 1 - 0{,}8 = 0{,}2}$

  Ou seja, **20%** de chances de não vender.

- A probabilidade de vender este produto é:

  $\mathbf{P(X = 1) = p}$

  $\mathbf{P(X = 1) = 0{,}8}$

  Ou seja, **80%** de chances de vender.

- A média, a variância e o desvio padrão da venda são:

  $\mathbf{E(X) = p = 0,8}$

  $\mathbf{Var(X) = pq = 0,8 \cdot 0,2 = 0,16}$

  $\mathbf{\sigma(X) = \sqrt{pq} = \sqrt{0,16} = 0,4}$

DISTRIBUIÇÃO BINOMIAL

Um experimento binomial é aquele que consiste em uma sequência de n ensaios idênticos e independentes. Cada tentativa pode resultar em apenas dois resultados possíveis: sucesso e fracasso, e a probabilidade de sucesso é constante de uma tentativa para outra.

Exemplos:

- Lançar uma moeda 5 vezes e observar o número de caras.
- 10 peças são escolhidas ao acaso e observamos as falhas.
- 5 cidades são observadas quanto ao acesso à rede de internet.

Designando por X o número total de sucessos em n tentativas, com probabilidade p de sucesso, sendo $0 < p < 1$, os possíveis valores de X são 0, 1, 2, ... , n. Os pares (x, p(x)), em que $p(x) = P(X=x)$, constituem a distribuição binomial, de modo que:

$$P(X = k) = \binom{n}{k} p^k \cdot q^{n-k}$$

$$\binom{n}{k} = \frac{n!}{K!\,(n-k)!}$$

$$P(X = k) = \binom{n}{k} = \frac{n!}{K!\,(n-k)!}\, p^k \cdot q^{n-k}$$

Em que:

k = número de sucessos.

n = número de elementos da amostra.

p = probabilidade de sucesso.

q = probabilidade de fracasso.

A média, a variância e o desvio padrão de uma distribuição binomial são dadas por:

$$\mathbf{E(x)....}$$

$$\mathbf{Var(x).....}$$

$$\mathbf{\sigma(X) = \sqrt{npq}}$$

Exemplos:

Um processo industrial na fabricação de monitores opera com média de 5% de defeituosos. Baseado em amostras de 10 unidades, calcule as probabilidades de uma amostra apresentar:

a. Nenhum monitor com defeito:

$$P(x = 0) = \frac{10!}{0!\,(10 - 0)!}\, 0{,}05^{0} \,.\, 0{,}95^{10} = 0{,}598 \text{ ou } 59{,}8\%$$

Observe que:

n = 10

k = 0

p = 5% ou 0,05

q = 1 – 0,05 = 0,95

Após a retirada dos dados, basta substituir os valores na fórmula.

Vejamos outro exemplo:

b. 3 monitores com defeito:

$$P(x = 3) = \frac{10!}{3!\,(10 - 3)!}\, 0{,}05^{3} \,.\, 0{,}95^{7} = 0{,}010 \text{ ou } 1\%$$

c. Pelo menos 9 monitores com defeito:

$P(x \geq 9) = P(x = 9) + P(x = 10)$

$P(x = 9) = \frac{10!}{9!\,(10 - 9)!}\, 0{,}05^{9} \,.\, 0{,}95^{1} = 1{,}85 \times 10^{-11}$

$P(x = 10) = \frac{10!}{10!\,(10 - 10)!}\, 0{,}05^{10} \,.\, 0{,}95^{0} = 9{,}76 \times 10^{-14}$

$P(x \geq 9) = 1{,}85$ x 1011 + 9,76 x 10-14 = 1,86 x 10-11 ou 0,0000000000186 ou 0,00000000186%

d. No máximo 2 monitores com defeito:

$P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$

$P(x = 0) = \frac{10!}{0!\,(10 - 0)!}\, 0{,}05^{0} \,.\, 0{,}95^{10} = 0{,}598$ ou 59,8%

$P(x = 1) = \frac{10!}{1!\,(10 - 1)!}\, 0{,}05^{1} \,.\, 0{,}95^{9} = 0{,}315$ ou 31,5%

$P(x = 2) = \frac{10!}{2!\,(10 - 2)!}\, 0{,}05^{2} \,.\, 0{,}95^{8} = 0{,}074$

$P(x \leq 2) = 0{,}598 + 0{,}315 + 0{,}074 = 0{,}987$ ou 98,7%

A média e a variância de monitores defeituosos serão:

$E(X) = 10 \times 0{,}05 = 0{,}5$

$Var\,(X) = 10 \times 0{,}05 \times 0{,}95 = 0{,}475$

Desvio padrão = 0,689

DISTRIBUIÇÃO DE POISSON

A distribuição de Poisson é frequentemente útil para estimar o número de ocorrências sobre um intervalo de tempo ou de espaços específicos. A probabilidade de uma ocorrência é a mesma para qualquer dois intervalos de igual comprimento e a ocorrência ou não em um intervalo é independente da ocorrência ou não em qualquer outro intervalo.

Exemplos:

- Número de chamadas telefônicas durante 10 minutos.
- Número de falhas de uma máquina durante um dia de operação.
- Número de acidentes ocorridos numa semana.
- Número de mensagens que chegam a um servidor por segundo.
- Defeitos por m^2, etc.

A distribuição de Poisson é dada por

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Em que:

- **λ** é a taxa de ocorrência do evento em um intervalo.
- **k** é o número de ocorrências do evento.
- **e** é uma constante matemática e ≈ 2,71828.

A média (E(X)), a variância (Var(X)) e o desvio padrão (σ) para a distribuição de Poisson são dadas por:

$$\mathbf{E(X) = \lambda}$$
$$\mathbf{Var(X) = \lambda}$$
$$\mathbf{\sigma(X) = \sqrt{\lambda}}$$

Vale ressaltar que a Distribuição de Poisson não tem um limite superior, ou seja, o número de ocorrências X pode assumir uma infinita sequência de valores.

Exemplos:

1. Um departamento de polícia recebe 5 solicitações por hora, em média, relacionadas a crimes cometidos. Qual a probabilidade de receber:

   a. 2 solicitações numa hora selecionada aleatoriamente?

   $$P\ (X=2) = \frac{2{,}71828^{-5} \cdot 5^2}{2!} = 0{,}0842 \text{ ou } 8{,}42\%$$

b. No máximo 2 solicitações numa hora selecionada aleatoriamente?

$$P(X \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$$

$$P(x = 0) = \frac{2{,}71828^{-5} \cdot 5^{0}}{0!} = 0{,}0067 \text{ ou } 0{,}67\%$$

$$P(x = 1) = \frac{2{,}71828^{-5} \cdot 5^{1}}{1!} = 0{,}0337 \text{ ou } 3{,}37\%$$

$$P(X=2) = \frac{2{,}71828^{-5} \cdot 5^{2}}{2!} = 0{,}0842 \text{ ou } 8{,}42\%$$

$$P(X \leq 2) = 0{,}0067 + 0{,}0337 + 0{,}0842 = 0{,}1246 \text{ ou } 12{,}46\%$$

2. Em um posto de gasolina, sabe-se que, em média, 10 clientes por hora param para colocar gasolina numa bomba. Pergunta-se:

a. Qual é a probabilidade de 3 clientes pararem qualquer hora para abastecer?

$$P(X) = \frac{2{,}71828^{-10} \cdot 10^{3}}{3!} = 0{,}0076 \text{ ou } 0{,}76\%$$

b. Qual é a média, a variância e o desvio padrão para essa distribuição?

Valor médio: $E(X) = 10$

Variância: $Var(X) = 10$

Desvio padrão: $\sigma(X) = \sqrt{10} = 3{,}16$

DISTRIBUIÇÃO DE PROBABILIDADES CONTÍNUA

As variáveis aleatórias contínuas são aquelas que assumem qualquer valor numérico em um intervalo de números reais. Como este tipo de variável pode assumir infinitos valores dentro de um intervalo e, por consequência, infinitos valores de probabilidade, não faz sentido tratar as variáveis contínuas da mesma forma que são tratadas as variáveis discretas.

Por exemplo, suponhamos que quiséssemos calcular a probabilidade de, num grupo, uma pessoa ter 170 cm de altura. Observe que a variável aleatória agora é a altura e X pode assumir qualquer valor entre 0 e infinito. Assim, se cada ponto fosse uma probabilidade, iríamos obter probabilidades com valores tendendo a zero. O valor para probabilidade citada no exemplo seria $1/\infty$. Assim, para calcular a probabilidade X, usamos o artifício de que X esteja compreendido entre dois pontos quaisquer. Exemplo: podemos calcular a probabilidade de um indivíduo medir entre 160 cm e 180 cm. Podemos fazer isso por meio da construção de um histograma, como pode ser visto abaixo:

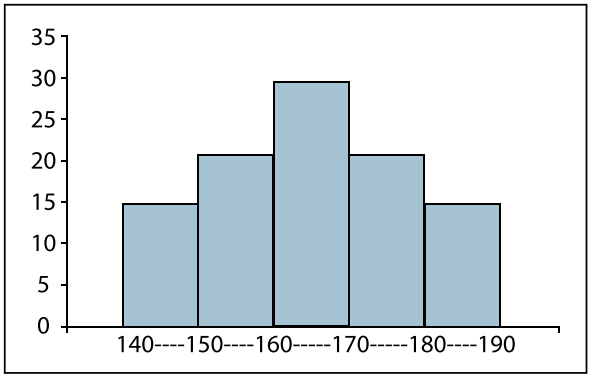


Gráfico 1 - Alturas de indivíduos
Fonte: as autoras.

Com o conhecimento da área na qual o intervalo 160 – 180 está compreendido, sabemos a probabilidade correspondente de um indivíduo ter entre 160 cm e 180 cm. Para o cálculo da área, usamos o artifício matemático chamado de integral. Assim, definidos dois pontos [a, b], a probabilidade da variável estar entre a e b é dado por:

$$P(a \leq X \bullet b) = \int_a^b f(x)dx$$

A função $f(x)$ é chamada de densidade de probabilidade (f.d.p) da variável aleatória X. Assim, podemos construir modelos teóricos para variáveis aleatórias contínuas, escolhendo adequadamente as funções densidade de probabilidade.

Dentre as principais distribuições contínuas, destacam-se a Distribuição Uniforme, Distribuição Exponencial e Distribuição Normal.

DISTRIBUIÇÃO UNIFORME

A Distribuição Uniforme é uma das mais simples de se conceituar. É usada em situações em que a função densidade de probabilidade é constante dentro de um intervalo de valores da variável aleatória X.

Usualmente, associamos uma distribuição uniforme a uma determinada variável aleatória, simplesmente por falta de informação mais precisa, além do conhecimento do seu intervalo de valores. O gráfico, a seguir, representa a função dada:

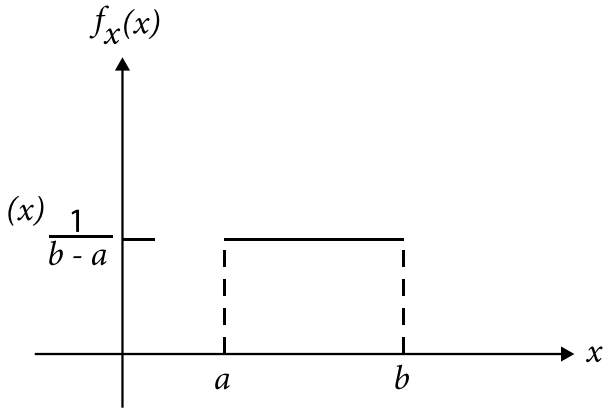


Gráfico 2 - Função
Fonte: as autoras.

Sendo que:

$$f(x) = \frac{1}{b-a} \text{ se } \quad a \leq x \leq b$$

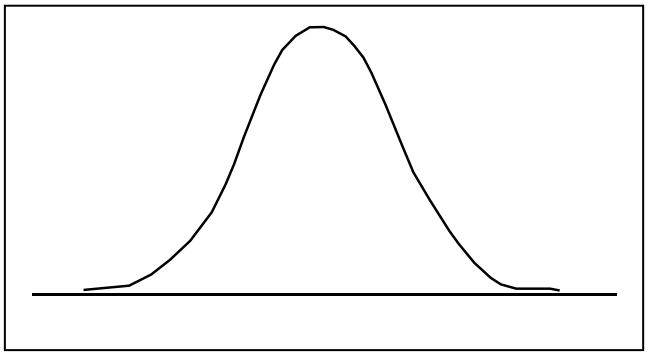
As fórmulas para o valor esperado e para a variância são:

$$E(x) = \frac{a+b}{2}$$

$$\text{Var}(x) = \frac{(b-a)^2}{12}$$

DISTRIBUIÇÃO NORMAL DE PROBABILIDADE

A distribuição de probabilidade contínua mais importante e mais utilizada na prática é a Distribuição Normal. A forma desta distribuição é ilustrada por uma curva em forma de sino, cujo ponto mais alto está na média, que também é a mediana e a moda da distribuição. Seu formato é simétrico em relação à média e seus extremos se estendem ao infinito em ambas as direções e, teoricamente, nunca tocam o eixo horizontal.



Fonte: as autoras.

O desvio-padrão determina a curva. Curvas mais largas e planas resultam de valores maiores de desvio-padrão, mostrando maior variabilidade dos dados. A área total sob a curva para a Distribuição Normal é 1. Como ela é simétrica, o valor da área sob a curva à esquerda e à direita é equivalente a 0,5 de cada lado.

Para simplificar a notação de uma variável aleatória com distribuição normal, com média μ e variância, utiliza-se:

$$X \sim N(\mu, \sigma^2)$$

Dizemos que a variável aleatória X tem distribuição normal com parâmetros μ e σ^2 se sua densidade é dada por:

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

Em que:

e = constante matemática (aproximada por 2,71828).

π = constante matemática (aproximada por 3,14159).

μ = média aritmética da população.

σ = desvio padrão da população.

x = qualquer valor da variável aleatória contínua onde $-\infty < X < \infty$.

DISTRIBUIÇÃO NORMAL PADRÃO

Para calcular $P(a \leq X \leq b)$ quando X é uma variável aleatória normal com parâmetros μ e σ, devemos calcular:

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx$$

Quando uma variável aleatória tem uma distribuição normal com média zero e desvio padrão 1, tem uma distribuição normal padrão de probabilidade.

Nenhuma das técnicas de integração padrão pode ser usada para calcular a integral acima. Assim, quando $\mu = 0$ e $\sigma = 1$, essa expressão foi calculada e tabulada para valores determinados de a e b. Nesta tabela, entra-se com a variável reduzida ou a variável padronizada Z e se encontra f(Z) ou vice-versa.

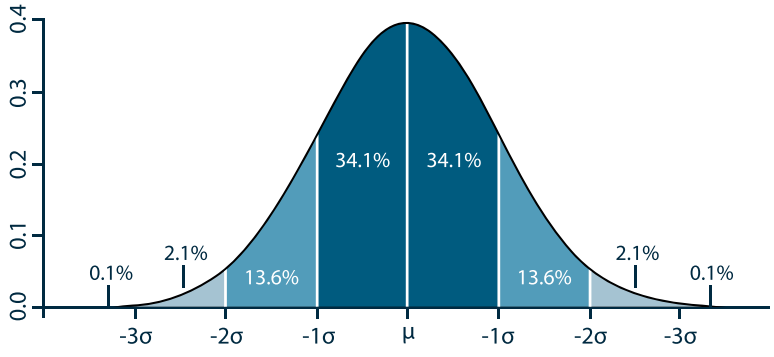


Gráfico 3 - Distribuição Normal Padrão
Fonte: as autoras.

A partir dessas integrais obtidas numericamente e utilizando a curva normal padronizada, podemos obter as probabilidades por meio de tabelas prontas que mostram a área sob a curva normal correspondente. A tabela para utilização das probabilidades é mostrada a seguir:

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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 10 - Tabela de distribuição normal reduzida
Fonte: EEL-USP ([2017], on-line)[1].

Vale ressaltar que tabelas com diferentes integrais calculadas podem ser encontradas. A tabela acima fornece sempre a seguinte área sob a curva:

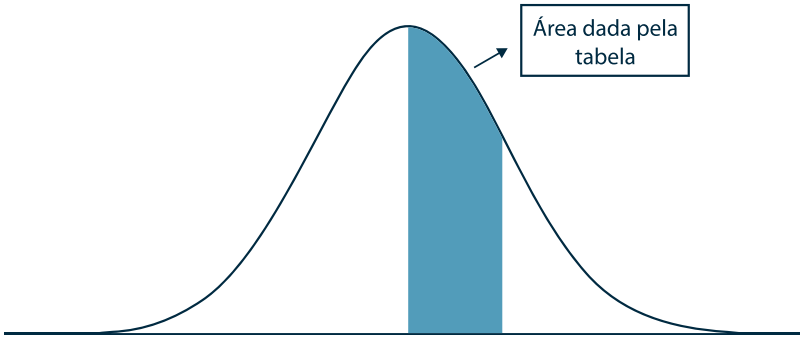


Gráfico 4 - Área sobre a curva
Fonte: as autoras.

A tabela anterior retorna a probabilidade de ocorrência de um evento entre 0 e z. Na margem esquerda, há o valor de z com uma decimal e, se for necessário considerar a segunda decimal, deve-se procurá-la na margem superior.

Exemplos:

- Para calcular a probabilidade de z entre 0 e 1, procuramos na margem esquerda a linha que tem z = 1,0 e a coluna 0,00, e encontramos o valor 0,3413. Isto significa que a probabilidade de encontrar um valor de x entre a média zero e z = 1,0 é 0,3413 ou 34,13%.

- Por outro lado, para se obter a probabilidade de z maior que 1, procuramos na tabela a probabilidade igual a 1,00 que é igual a 0,3413, e a seguir fazemos 0,5 - 0,3413 = 0,1587 ou 15,87%.

- Para se obter a probabilidade de z entre 0 e 1,87, procuramos a célula cuja linha é 1,8 e coluna 0,07. O resultado é o valor 0,4693 ou 46,93%.
- Valores procurados abaixo da média, ou seja, abaixo de 0 irão aparecer como negativos, porém observe que na tabela não há valores negativos. Como a curva é simétrica, valores negativos são equivalentes aos valores positivos, ou seja, a área procurada é a mesma equivalente aos valores positivos.

Para utilizar a tabela, as variáveis aleatórias x precisam ser padronizadas. A fórmula usada para esta conversão é:

$$z = \frac{x_i - \mu}{\sigma}$$

Em que:

- x_i = ponto que se deseja converter em z.
- μ = média da normal original.
- σ = desvio padrão da normal original.

Vejamos o exemplo:

Suponha que a média da taxa de falhas de dados é transmitida em lotes. Sabe-se que essa característica segue uma distribuição normal com média de 2,0 e desvio padrão igual a 0,5. Calcule, então, as seguintes probabilidades:

- De tomarmos um lote ao acaso e este ter uma taxa de falhas entre 2,0 e 2,5. Traduzindo para linguagem probabilística, queremos:

$$P(2,0 < x < 2,5) = ?$$

Primeiramente, vamos padronizar os dados.

Lembre-se que a fórmula da padronização é:

$$Z = \frac{x_i - \mu}{\sigma} \text{ e que } \quad \mu = 2,0 \quad \sigma = 0,5$$

Assim:

$$Z = \frac{2,5 - 2,0}{0,5} = 1$$

$$Z = \frac{2,0 - 2,0}{0,5} = 0$$

Novamente traduzimos para a linguagem probabilística, mas agora usando os dados padronizados:

$P(2,0 < x < 2,5) = P(0 < z < 1) = 0,3413$ ou 34,13%

Queremos uma área que esteja entre 0 e 1 desvios padrão.

Essa área é exatamente o que a tabela nos dá. Basta olhar, como explicado acima, o valor da linha 1,0 e na linha 0,0 para obtermos o valor 0,3413.

Assim, dizemos que a chance de tomarmos um lote que tenha uma taxa de falhas de dados entre 2,0 e 2,5 é de 34,13%.

Vamos ver outra probabilidade:

- De tomarmos um lote ao acaso e ter mais que uma taxa de falhas de 2,5.

  $P(x < 2,5) = P(z < 1) = 0,5 - 0,3413 = 0,1587$ ou 15,87%

- De tomarmos um lote ao acaso e ter menos que uma taxa de falhas de 2,5.

  $P(x > 2,5) = P(z > 1) = 0,5 + 0,3413 = 0,8413$ ou 84,13%

CONSIDERAÇÕES FINAIS

Vimos nesta unidade a importância das probabilidades no nosso cotidiano.

A teoria das probabilidades tenta quantificar a noção de provável, sendo uma ferramenta estatística de grande utilidade quando se trabalha com inúmeros eventos relacionados a pesquisas em empresas, órgãos governamentais e instituições de ensino. Essa ferramenta lida com as chances de ocorrências de algo que vai acontecer, então dizemos que ela lida com fenômenos aleatórios. Portanto, é necessário conhecer o material de estudo para poder calcular essas chances ou probabilidades de maneira correta e então tomarmos nossas decisões com base em nossas estimativas.

Um efeito importante da teoria da probabilidade no cotidiano está na avaliação de riscos. Normalmente, governos, por exemplo, utilizam processos envolvidos em probabilidades para suas tomadas de decisões. Uma aplicação importante das probabilidades é a questão da confiabilidade, por exemplo, no lançamento de algum produto, nas chances deles falharem.

Para inferir sobre probabilidades, é necessário saber que tipo de variável aleatória está sendo trabalhada. Cada variável aleatória possui um tipo de comportamento chamado de distribuição de probabilidades. Isso é importante, pois cada distribuição de probabilidade possui algumas características e elas devem ser respeitadas para que se possa chegar a resultados precisos e então conclusões válidas possam ser tomadas sobre aquilo que estamos estudando. Vimos, nesta unidade, os conceitos básicos de probabilidade, a forma clássica de calculá-la e também vimos as principais distribuições de probabilidades utilizadas.

Deve-se entender que é razoável pensar ser de extrema importância compreender como estimativas de chance e probabilidades são feitas e como elas contribuem para reputações e decisões em nossa sociedade.

ATIVIDADES

1. Explique espaço amostral e eventos.

2. Uma máquina de fabricação de computadores tem probabilidade de produzir um item defeituoso de 10%. **Em uma amostra de 6 itens, calcule a probabilidade de:**

   a. Haver no máximo um item defeituoso.

   b. Haver 3 itens defeituosos.

   c. Não haver itens defeituosos.

   d. Determine a média e a variância do experimento.

3. A qualidade de alguns CDs foi avaliada sobre a resistência a arranhões e sobre a adequação de trilhas. Os resultados foram:

RESISTÊNCIA A ARRANHÕES	ADEQUAÇÃO DE TRILHAS		TOTAL
	APROVADO	REPROVADO	
Alta	700	140	840
Baixa	100	60	160
Total	**800**	**200**	**1000**

Fonte: adaptado de Barbetta et al. (2010)

Se um CD for selecionado aleatoriamente deste lote, qual é a probabilidade de:

   a. Ter resistência alta a arranhões.

   b. Ter resistência baixa a arranhões.

   c. Ser aprovado na avaliação das trilhas.

   d. Ser reprovado na avaliação das trilhas.

   e. Ter resistência alta ou ser aprovado.

   f. Ter resistência baixa ou ser reprovado.

   g. Ter resistência alta dado que seja reprovado.

   h. Ter resistência baixa dado que seja aprovado.

ATIVIDADES

4. Um sistema de banco de dados recebe em média 80 requisições por minuto, segundo uma distribuição de Poisson. **Qual a probabilidade de que no próximo minuto ocorram 100 requisições? Determine a média e a variância para essa variável aleatória.**

5. A distribuição da duração de monitores pode ser aproximada por uma distribuição normal de média $\mu = 6$ anos e desvio padrão $\sigma = 2$ anos. **Determine a probabilidade de um monitor durar:**

   a. Entre 6 e 9 anos.

   b. Acima de 9 anos.

   c. Entre 4 e 9 anos.

   d. Acima de 4 anos.

Qual a probabilidade de ganhar na Mega Sena?

Acho que todos os brasileiros gostariam de ganhar na Mega Sena, não acham? Por isso as casas lotéricas estão sempre lotadas. Muitos pensam em jogar suas datas de nascimento, de casamento, do aniversário de alguém importante, mas nem sempre esses números são os sorteados. E esse é o momento esperado, o sorteio dos números que irão decidir se existiram ganhadores.

A cartela da Mega Sena é composta por 60 números, enumerados de 01 a 60. A aposta mínima é feita por seis números e a máxima de 15, mas os valores das apostas podem variar de acordo com o aumento dos números apostados. Quanto mais números marcados, maior a chance de ganhar. Os seis números sorteados (dentre os sessenta) e os prêmios em dinheiro são pagos para os ganhadores da quadra (que é o acerto de quatro números), da quina (acerto de cinco números) e a tão esperada sena (o acerto dos seis números). E aqueles sonhados prêmios milionários são pagos somente a quem acertar os seis números sorteados; caso mais de uma pessoa acerte os seis números, o prêmio é dividido em partes iguais dentre os acertadores. E agora, pergunta-se: qual é a chance de uma pessoa ganhar com apenas seis números?

Essas chances de acerto de seis números são por meio de combinação simples de sessenta elementos tomados seis a seis $C_{60,6}$, ou seja, existem 50.063.860 (cinquenta milhões, sessenta e três mil e oitocentos e sessenta) probabilidade de se acertar os seis números de 1 a 60.

Essa probabilidade corresponde a 1/50.063.860 = 0,00000002 que corresponde a 0,000002%.

A pergunta é: quantos jogos de 6 números são possíveis utilizando os 60 números disponíveis?

Para a escolha do 1º número, temos 60 possibilidades.

Para a escolha do 2º número, temos 59 possibilidades

Para a escolha do 3º número, temos 58 possibilidades.

Para a escolha do 4º número, temos 57 possibilidades.

Para a escolha do 5º número, temos 56 possibilidades.

Para a escolha do 6º número, temos 55 possibilidades.

Pelo princípio multiplicativo, temos então **60 x 59 x 58 x 57 x 56 x 55** jogos, ou seja, um número bem grande.

Contudo, devemos verificar que, por exemplo, existe o jogo 1,2,3,4,5,6 que é a mesma coisa que 2,3,4,1,6,5, ou seja, existe a permutação dos 6 números que foram sorteados.

Portanto, para sabermos quantos jogos **distintos** podemos formar, dividiremos pela permutação dos 6 algarismos da mega sena, que é **6**!

Fonte: adaptado de ALAMADA (2010, on-line)[2].

MATERIAL COMPLEMENTAR

LIVRO

Estatística Básica. Probabilidade e Inferência

Luis Gonzaga Morettin

Editora: Pearson

Sinopse: O livro Estatística básica traz o conteúdo programático de um curso de estatística. Fornece diversos exemplos para ilustrar a teoria ao longo dos capítulos e, ao final de cada um deles, apresenta exercícios resolvidos e propostos para auxiliar na aprendizagem dos estudantes.

REFERÊNCIAS

TOLEDO, G. L.; OVALLE, I. I. **Estatística Básica**. São Paulo: Atlas, 1997.

TRIOLA, M. F. **Introdução à Estatística**. Rio de Janeiro: LTC, 1999.

SILVEIRA, F. K.; ORTH, A. I.; PRINKLADNICKI, R. RiskFree – Uma Ferramenta de Gerenciamento de Riscos Baseada no PMBOK e Aderente ao CMMI. **V Simpósio Brasileiro de Qualidade de Software – SBQS'2006** Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbqs/2006/014.pdf>. Acesso em: 25 abr. 2017.

REFERÊNCIAS ON-LINE

[1]Em: <http://www.dequi.eel.usp.br/~fabricio/tabela_dist_normal.pdf>. Acesso em: 25 abr. 2017.

[2]Em: <http://cassiusalmadamatematica.blogspot.com.br/2010/08/analise-combinatoria-chance-de-ganhar.html>. Acesso em: 25 abr. 2017.

GABARITO

1. Espaço amostral é o conjunto de todos os possíveis resultados do experimento aleatório.

   Eventos: é um dos possíveis resultados do experimento aleatório e do qual se deseja saber a probabilidade de ocorrência.

2.
   a. 0,53 + 0,35 = 0,88 ou 88%
   b. 0,0145 ou 1,45%
   c. 0,53 ou 53%....
   d. Média = 0,6. Variância = 0,54

3.
   a. Ter resistência alta a arranhões 840/1000
   b. Ter resistência baixa a arranhões 160/1000
   c. Ser aprovado na avaliação das trilhas 800/1000
   d. Ser reprovado na avaliação das trilhas 200/1000
   e. Ter resistência alta ou ser aprovado $p(a \cup ap) = \frac{840 + 800 - 700}{1000} = 940/1000$
   f. Ter resistência baixa ou ser reprovado $p(b \cup r) = \frac{160 + 200 - 60}{1000} = 300/1000$
   g. Ter resistência alta dado que seja reprovado p(a/r) = 140/200
   h. Ter resistência baixa dado que seja aprovado 100/800

4.

$P(X) = \frac{2{,}71828^{-80} \cdot 80^{100}}{100!} = 0{,}0039$ ou 0,39%

E(x) = 80

Var (x) = 80

5.
   a. 0,4332 ou 43,32%
   b. 0,0668
   c. 0,3413 + 0,4332 = 0,77 ou 77%
   d. 0,8413 ou 84,13%

Professora Me. Ivnna Gurniski de Oliveira
Professora Me. Renata Cristina de Souza Chatalov

UNIDADE V

CORRELAÇÃO LINEAR E REGRESSÃO

Objetivos de Aprendizagem

- Conhecer o coeficiente de correlação linear.
- Entender associação entre duas variáveis.
- Saber interpretar correlação positiva e negativa.
- Compreender a correlação e aplicação da correlação de Pearson.
- Conhecer a utilização da regressão linear.
- Entender a predição de uma variável por meio de outra.

Plano de Estudo

A seguir, apresentam-se os tópicos que você estudará nesta unidade:

- Correlação Linear
- Regressão Linear

Unicesumar

INTRODUÇÃO

A Estatística apresenta muitas ferramentas para descrever e analisar dados de pesquisas. A escolha das ferramentas a serem utilizadas na pesquisa depende dos seus objetivos, bem como do tipo de variável com a qual se trabalha.

Como visto na unidade I, as variáveis podem ser qualitativas e quantitativas. Esta distinção é importante, uma vez que as ferramentas utilizadas para um tipo de variável nem sempre podem ser utilizadas para o outro tipo. É importante também saber que, nas pesquisas, se utiliza não só uma, mas um grupo de variáveis.

Em alguns casos, algumas variáveis podem estar relacionadas de alguma forma e a variação de uma vai depender da variação da outra. As decisões gerenciais, geralmente, são baseadas nas relações entre duas ou mais variáveis. Por exemplo, após considerar a relação entre gastos com publicidade e vendas, um gerente poderia tentar prever as vendas de acordo com o nível de gastos com a publicidade.

O fato de duas variáveis estarem ligadas permite tomar decisões se baseando em uma variável, porém esperando resposta em outra que seja de difícil mensuração ou só possa ser medida tardiamente. Existem algumas medidas estatísticas que permitem medir o grau de associação entre duas variáveis.

Nesta unidade, iremos ver duas delas: a correlação linear e a regressão linear. Entretanto, essas duas ferramentas só podem ser utilizadas quando as variáveis medidas são quantitativas. Assim, nesta unidade, você verá como podemos verificar a associação entre variáveis ou a dependência de uma variável em função da outra, e também como quantificar esta associação.

A importância da correlação e da regressão linear está associada em algumas decisões gerenciais, que são baseadas na relação entre duas ou mais variáveis, como a relação entre o consumo e vendas, matéria-prima e produto acabado, sendo uma ferramenta de fundamental conhecimento para os futuros gestores.

CORRELAÇÃO LINEAR

Em diversas situações, o objetivo é apenas estudar o comportamento conjunto de duas variáveis e verificar se elas estão relacionadas, ou seja, saber se as alterações sofridas por uma das variáveis são acompanhadas por alterações nas outras.

Em Estatística, o termo correlação é usado para indicar a força que mantém unidos dois conjuntos de valores. O estudo da correlação tem como objetivo estudar a existência ou não de uma relação; e seu grau de relação entre as variáveis.

Uma medida do grau da correlação e sua direção é dada pela covariância entre duas variáveis aleatórias, mas é mais conveniente medir o grau da correlação por meio do Coeficiente de Correlação Linear de Pearson.

COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON

O coeficiente de correlação é uma medida que dimensiona a correlação. É representado pela letra "r" e dado pela seguinte fórmula:

$$r = \frac{\sum xy - \frac{\left(\sum x_i \cdot \sum y_i\right)}{n}}{\sqrt{\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right] \cdot \left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right]}}$$

O valor de r não depende de qual das duas variáveis em estudo é chamada de "x" e de "y" e independe das unidades com as quais as variáveis são medidas.

A intensidade do coeficiente de correlação pode variar entre -1 e 1, sendo que quanto mais próximo de -1 ou de 1, mais forte será a associação entre as duas variáveis, e quanto mais próximo de 0, mais fraca será a associação. Quando r = 1, todos os pares (x, y) estarão alinhados em linha reta com coeficiente angular positivo e quando r = -1, todos os pares (x, y) estarão alinhados com o coeficiente angular negativo (BUSSAB e MORETTIN, 2003).

Quanto ao direcionamento entre as duas variáveis, o coeficiente de correlação pode ser positivo ou negativo. Se a correlação entre duas variáveis for positiva, dizemos que as duas variáveis variam para o mesmo sentido.

Exemplo: se sabemos que a correlação entre renda familiar e gastos com alimentação é positiva, podemos dizer, que, à medida que a renda familiar aumenta, também, aumentam os gastos com alimentação, ou à medida que a renda familiar diminui, também diminuem os gastos com alimentação. Entretanto, se dizemos que o conhecimento e tempo gasto para aprender a operar uma máquina têm uma correlação negativa, então, podemos pensar que à medida que o conhecimento aumenta, o tempo gasto para aprender a operar uma máquina diminui ou vice-versa.

Se o coeficiente de correlação for igual a 0, dizemos que não existe associação linear entre as duas variáveis.

Outra forma de representação das correlações é por meio do diagrama de dispersão, mostrando a variação conjunta entre as duas variáveis. Observe os diagramas de dispersão abaixo:

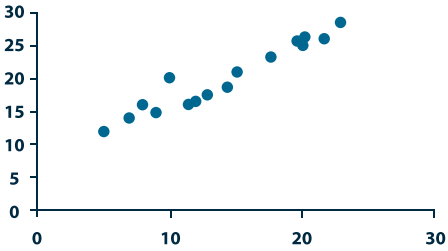


Gráfico 1 - Correlação positiva
Fonte: as autoras

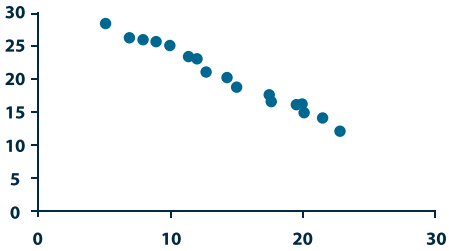


Gráfico 2 - Correlação negativa
Fonte: as autoras.

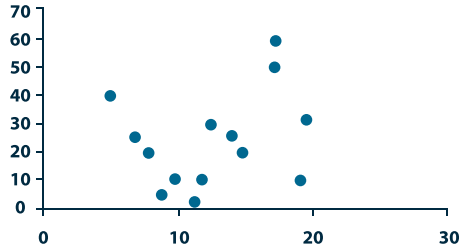


Gráfico 3 - Correlação Nula
Fonte: as autoras.

Exemplo:

Verifique se existe correlação linear entre o número de nascidos vivos e a taxa de mortalidade infantil na região de Maringá-PR.

Tabela 01 - Taxa de Mortalidade Infantil em municípios da região de Maringá

REGIONAL DE SAÚDE E MUNICÍPIOS	NASCIDOS VIVOS (X) *	TOTAL MENOR DE 01 ANO (Y)
Astorga	289	6
Colorado	246	2
Floresta	68	1
Itambé	67	1
Mandaguaçu	251	4
Mandaguari	423	6
Marialva	378	5
Nova Esperança	423	9
Paiçandu	443	4
São Jorge do Ivaí	59	0

Fonte: adaptada de SESA/ISEP/CIDS - Departamento de Sistemas de Informação em Saúde (2002, on-line).

Inicialmente, é interessante traçar um diagrama de dispersão para as duas variáveis, para nos dar uma ideia de como ocorre a variação conjunta dos dados.

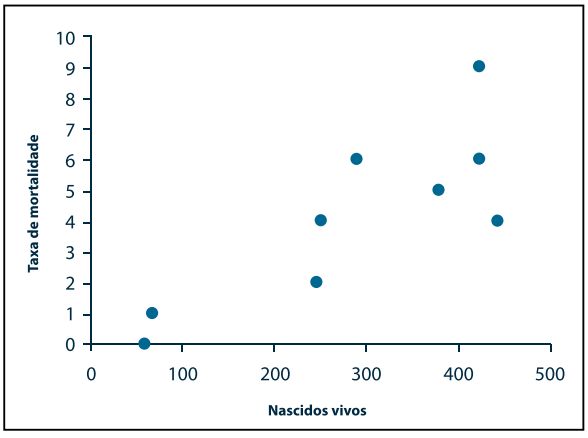


Gráfico 04 - Dispersão da taxa de mortalidade até 1 ano e o número de nascidos vivos na região de Maringá
Fonte: adaptado de SESA/ISEP/CIDS - Departamento de Sistemas de Informação em Saúde (2002, on-line).

Por meio do diagrama de dispersão, podemos notar que há uma variação crescente entre as duas variáveis.

As maiores correlações positivas e negativas são obtidas somente quando todos os pontos estão bem próximos a uma linha reta.

O próximo passo é calcular o coeficiente de correlação entre as duas variáveis. Podemos utilizar uma tabela para organizar os dados:

Tabela 02 - Nascidos Vivos das Cidades Metropolitanas de Maringá

CIDADE	NASCIDOS VIVOS (X)	TAXA DE MORTALIDADE (Y)	X.Y	X^2	Y^2
Astorga	289	6	1734	83521	36
Colorado	246	2	492	60516	4
Floresta	68	1	68	4624	1
Itambé	67	1	67	4489	1
Mandaguaçu	251	4	1004	63001	16
Mandaguari	423	6	2538	178929	36
Marialva	378	5	1890	142884	25
Nova Esperança	423	9	3807	178929	81
Paiçandu	443	4	1772	196249	16
São Jorge do Ivaí	59	0	0	3481	0
Soma	**2647**	**38**	**13372**	**916623**	**216**

Fonte: adaptada de SESA/ISEP/CIDS - Departamento de Sistemas de Informação em Saúde (2002, on-line).

Agora, vamos calcular o r:

$$r = \frac{13372 - \frac{(2647 \cdot 38)}{10}}{\sqrt{\left(916623 - \frac{2647^2}{10}\right) \cdot \left(216 - \frac{38^2}{10}\right)}} = 0{,}84$$

Observe que o valor do coeficiente de correlação foi 0,84. Assim, podemos concluir que existe uma forte correlação ou associação entre o número de nascidos vivos e a taxa de mortalidade antes de 1 ano de idade, uma vez que o valor 0,84 é um valor próximo de 1. Em seguida, podemos verificar que, como mostrado

no diagrama de dispersão, o valor dessa correlação é positivo, então, podemos pensar que, se aumenta o número de nascidos vivos, também há um aumento na taxa de mortalidade até 1 ano de idade.

REFLITA

O coeficiente de correlação de Pearson (r) é uma medida de associação linear entre variáveis.

COEFICIENTE DE DETERMINAÇÃO

O coeficiente de determinação expressa a porcentagem de variação dos valores de Y em função do valor X, ou seja, este coeficiente mostra até que ponto a variação conjunta dos dados é de fato linear.

Esse coeficiente varia de 0 a 1, sendo que, quanto mais perto de 1, **maior é a variação conjunta das duas variáveis somadas com a variável x que dará um resultado na qual a variável x pode ser explicada pela variável y**. O R^2 varia entre 0 e 1, e indica, em percentual, o quanto o modelo consegue explicar os valores observados, e quanto maior o R^2, mais explicativo é o modelo, melhor ele se ajusta à amostra.

O coeficiente de determinação é dado por R^2, ou seja, o símbolo do coeficiente de determinação é dado por "R" maiúsculo e é dado pelo valor encontrado para a correlação linear de Pearson ao quadrado.

Utilizando um exemplo em que r = 0,84, o coeficiente de determinação será dado por:

$R^2 = 0{,}84^2 = 0{,}7056$

Isso mostra que a variação conjunta dos dados é boa, ou seja, a variação da taxa de mortalidade pode ser explicada pela variação no número de nascidos vivos. Observe que os dados retirados pela estatística se encaixam.

REGRESSÃO LINEAR

A análise de regressão é uma técnica estatística cujo objetivo é investigar e descrever a relação entre variáveis por meio de um modelo matemático. Esta relação é explorada de modo que se possa obter informações sobre uma variável, por meio dos valores conhecidos das outras.

Primeiramente, é preciso estudar a dependência de uma variável em relação à outra e, assim, indicar a variável independente para o eixo "x" e a variável dependente para o eixo "y". À medida que a variável independente (ou explicativa) varia, provoca uma mudança na variável dependente (ou resposta).

Aplicações da regressão:

- Estimar valores de uma variável com base em valores conhecidos de outra variável.
- Situações em que as duas variáveis medem, aproximadamente, a mesma situação, mas uma delas é relativamente dispendiosa ou difícil de lidar, enquanto a outra não.
- Explicar valores de uma variável em termos da outra, isto é, pode-se suspeitar de uma relação de causa e efeito.
- Predizer valores de uma variável para a análise de regressão: resta saber como é o tipo dessa relação.

REGRESSÃO LINEAR SIMPLES

A regressão linear simples é assim chamada quando duas variáveis, X e Y (numéricas e contínuas) estão relacionadas linearmente. Isso quer dizer que à medida que X aumenta, Y também aumenta, ou à medida que X aumenta, y diminui. Essa relação é dada por uma equação que chamamos de equação de regressão linear:

$$\hat{y} = a + bx \quad \text{em que}$$

$\hat{y}$ = valor predito da variável resposta.

a = constante de regressão que representa o intercepto entre a linha de regressão e o eixo y.

b = coeficiente linear de regressão da variável resposta y em função da variável explicativa x; inclinação da reta; taxa de mudança na variável y por unidade de mudança na variável x.

x = valor da variável explicativa.

O coeficiente de regressão "b" fornece uma estimativa da variação esperada de y a partir da variação de uma unidade em x (BARBETTA et al., 2010).

A partir dessa equação, é possível encontrar os valores preditos para y e a reta de regressão. Além disso, a relação entre x e y pode ser mostrada por um diagrama de dispersão. Vejamos o diagrama de dispersão abaixo, mostrando a relação entre as variáveis x e y, bem como o exemplo de uma reta de regressão.

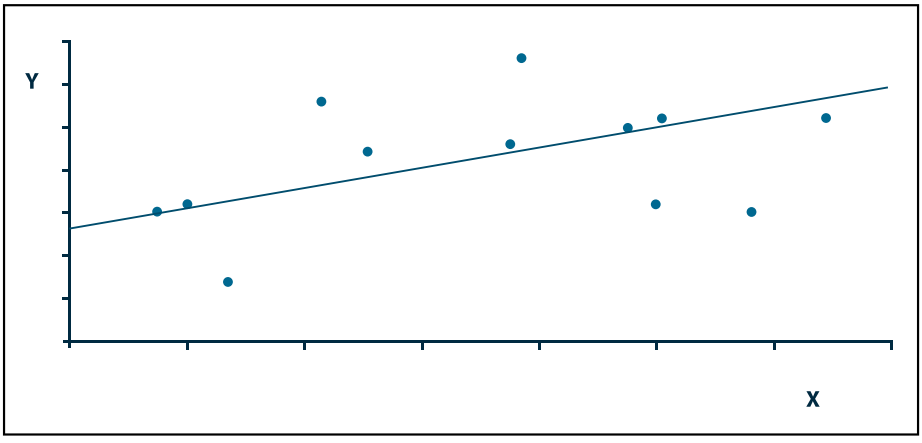


Gráfico 05 - Reta de Regressão Linear
Fonte: as autoras.

O diagrama de dispersão mostra o tipo de relação que existe entre x e y e também verifica se o modelo proposto (y = a + bx) explica bem a variação dos dados. O modelo explicará melhor quanto mais perto dos dados ou dos pontos (visto no gráfico) a reta estiver.

O método mais usado para ajustar uma linha reta a um conjunto de pontos é conhecido como método dos mínimos quadrados que nos fornece os seguintes resultados para estimarmos a e b:

$$b = \frac{\sum xy - \frac{\left(\sum x_i . \sum y_i\right)}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} \qquad e \qquad a = \bar{y} - b\bar{x}$$

Em que $\bar{y}$ e $\bar{x}$ são as médias de y e x, respectivamente.

Exemplo:

Tabela 03 - Nascidos Vivos das Cidades Metropolitanas de Maringá

REGIONAL DE SAÚDE E MUNICÍPIOS	NASCIDOS VIVOS *	TOTAL MENOR DE 01 ANO
Astorga	289	6
Colorado	246	2
Floresta	68	1
Itambé	67	1
Mandaguaçu	251	4
Mandaguari	423	6
Marialva	378	5
Nova Esperança	423	9
Paiçandu	443	4
São Jorge do Ivaí	59	0

Fonte: adaptada de SESA/ISEP/CIDS - Departamento de Sistemas de Informação em Saúde (2002, on-line).

Nesse caso, a variável x é o número de nascidos vivos, a variável y é a taxa de mortalidade, uma vez que é evidente que a taxa de mortalidade depende do número de nascidos vivos. A partir dessa definição, é necessária a estimação dos parâmetros da equação a e b.

Tabela 04 - Nascidos Vivos das Cidades Metropolitanas de Maringá

CIDADE	NASCIDOS VIVOS (X)	TAXA DE MORTALIDADE (Y)	X.Y	X^2
Astorga	289	6	1734	83521
Colorado	246	2	492	60516
Floresta	68	1	68	4624
Itambé	67	1	67	4489
Mandaguaçu	251	4	1004	63001
Mandaguari	423	6	2538	178929
Marialva	378	5	1890	142884
Nova Esperança	423	9	3807	178929
Paiçandu	443	4	1772	196249
São Jorge do Ivaí	59	0	0	3481
Soma	**2647**	**38**	**13372**	**916623**

Fonte: adaptada de SESA/ISEP/CIDS - Departamento de Sistemas de Informação em Saúde (2002, on-line).

Assim, o valor de b e de a será:

$$b = \frac{13372 - \frac{(2647 \,.\, 38)}{10}}{916623 - \frac{2647^2}{10}} = 0{,}015$$

$$\bar{x} = \frac{2647}{10} = 264{,}7 \qquad \bar{y} = \frac{38}{10} = 3{,}8$$

$$a = 3{,}8 - 0{,}015 \times 264{,}7 = -0{,}17$$

De acordo com o valor de b, dizemos que a cada 1 nascido vivo esperamos um aumento (b positivo) de 0,015% na taxa de mortalidade em crianças de até 1 ano de idade.

Assim, a equação da reta de regressão é dada por:

$$\hat{y} = -0{,}17 + 0{,}015x$$

As equações de regressão mostram as taxas de mortalidades preditas em função do número de nascidos vivos, como segue:

Tabela 05 - Nascidos Vivos das Cidades Metropolitanas de Maringá

NÚMERO DE NASCIDOS VIVOS (X)	A + BX	Ŷ
59	- 0,17 + 0,015 . 59	0,715
67	- 0,17 + 0,015 . 67	0,835
68	- 0,17 + 0,015 . 68	0,85
246	- 0,17 + 0,015 . 246	3,52
251	- 0,17 + 0,015 . 251	3,595
289	- 0,17 + 0,015 . 289	4,165
378	- 0,17 + 0,015 . 378	5,5
423	- 0,17 + 0,015 . 423	6,175
443	- 0,17 + 0,015 . 443	6,475

Fonte: adaptada de SESA/ISEP/CIDS - Departamento de Sistemas de Informação em Saúde (2002, on-line).

Para cada número de nascidos vivos, temos uma taxa de mortalidade prevista. A representação gráfica para esta situação pode ser observada no diagrama de dispersão abaixo.

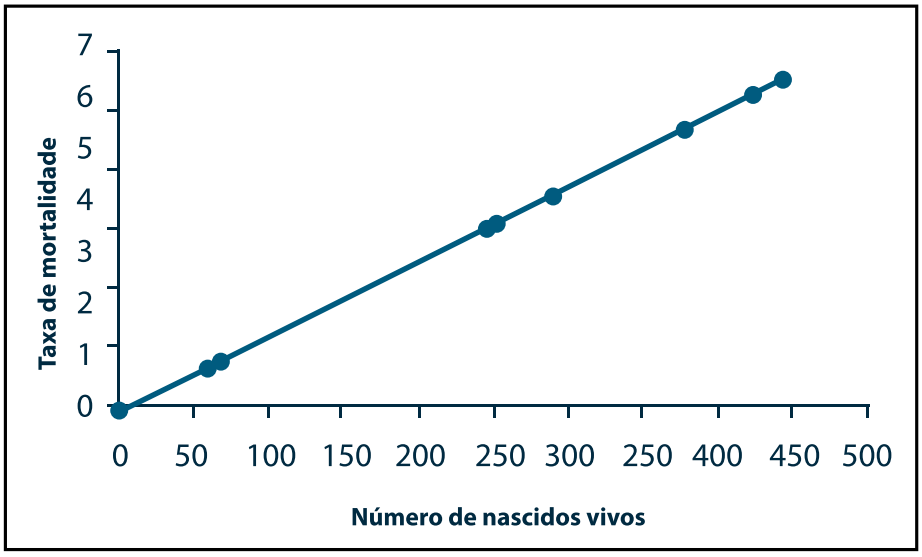


Gráfico 06 - Diagrama de dispersão para a taxa de mortalidade prevista em função do número de nascidos vivos
Fonte: as autoras.

Observe que, para cada número de nascidos vivos, foi construída uma equação e obtida uma taxa de mortalidade prevista. No gráfico 2, é mostrada a reta que descreve os dados e, a título de exemplificação, foi mostrado um ponto no gráfico em que o número de nascidos vivos (disposto no eixo x) foi 378 e a taxa de mortalidade (disposta no eixo y) prevista é de 5,41.

Vimos que a correlação linear de Pearson mostrou um valor alto para a relação entre o número de nascidos vivos e a taxa de mortalidade (0,84), e que foi verificada uma correlação populacional acima de "0". Na análise de regressão, utilizamos, também, o coeficiente de determinação para verificar a precisão da reta de regressão e dizer se ela explica bem ou não a variação dos dados.

Como vimos:

$$R^2 = 0{,}7056$$

A explicação para a análise de regressão será:

70,56% da variação observada na taxa de mortalidade é explicada pela reta de regressão. Isto mostra que a reta se aproxima bem dos pontos observados. A reta de regressão mostra as equações de regressão previstas e os pontos são os valores observados.

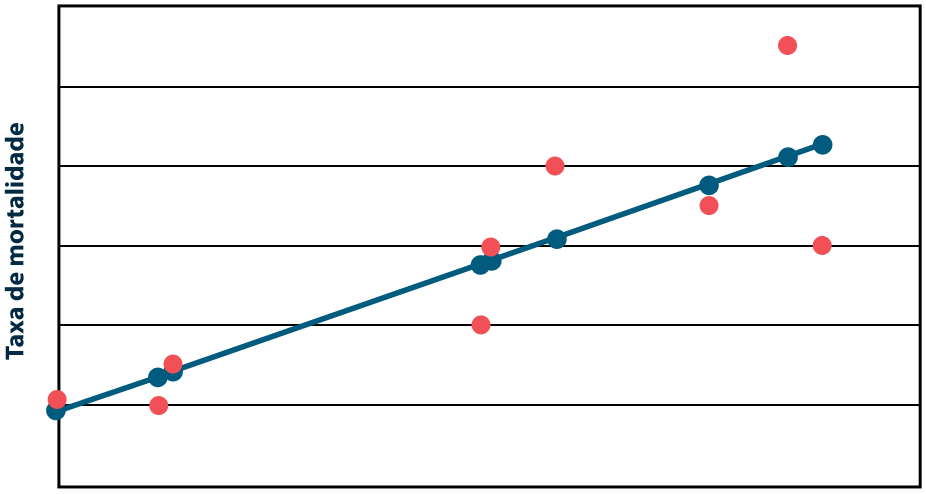


Gráfico 07 - Diagrama de dispersão para a taxa de mortalidade prevista e taxa de mortalidade observada em função do número de nascidos vivos

Fonte: as autoras.

Os pontos em vermelho nos mostram as taxas de mortalidade observadas. Observe que os pontos estão próximos da reta de regressão. Isso, associado ao coeficiente de determinação, indica boa precisão da variação dos dados.

A partir da equação dada acima e também do diagrama de dispersão, podemos fazer predições para a variável dependente em função da variável independente. Como exemplo, podemos prever a taxa de mortalidade em função de qualquer número de nascidos vivos dentro do intervalo estudado (59 a 443). Supondo que quiséssemos saber a taxa de mortalidade esperada para um número de 300 nascidos vivos, se substituirmos esse valor na equação de regressão estimada, temos que:

$$\hat{y} = - 0,17 + 0,015 \times 300 = 4,33\% \text{ de mortalidade esperada}$$

No entanto, essas predições devem seguir alguns critérios:

- Só podemos fazer predições em casos de valores dentro do intervalo trabalhado para a variável independente.
- Só devemos fazer essas predições caso, de fato, a variável independente explique a variação da variável dependente.

Exercícios:

Resultado de um teste (de 0 a 100) sobre conhecimento (X) e tempo gasto (minutos) para aprender a operar uma máquina (Y) para oito indivíduos.

Tabela 06 - Tempo gasto para operação de uma máquina

INDIVÍDUO	TESTE (X)	TEMPO (Y)
A	45	343
B	52	368
C	61	355
D	70	334
E	74	337
F	76	381
G	80	345
H	90	375

Fonte: as autoras.

Responda:

1. Construa um diagrama de dispersão entre as variáveis teste e tempo.
2. Calcule os valores do coeficiente de regressão linear b e do intercepto a.
3. Demonstre os valores das equações para o tempo esperado ou previsto para todos os valores do teste.
4. Demonstre o diagrama de dispersão para as duas variáveis utilizando os y preditos.
5. Demonstre o valor do tempo esperado quando o valor do teste for igual a 69.

R: 2)

$$b = \frac{194.850 - [(548x2.838)/8]}{39.102 - [(548)^2/8]}$$

$$b = \frac{194.850 - (1.555.224/8)}{39.102 - (300.304/8)}$$

$$b = \frac{194.850 - 194.403}{39.102 - 37.538}$$

$$b = \frac{447}{1.564} = 0,2858$$

$$x = \frac{548}{8} = 68,5$$

$$y = \frac{2.838}{8} = 354,75$$

$$y = a + bx$$
$$354,75 = a + 0,2858x68,5$$
$$354,75 = a + 19,5773$$
$$a = 354,75 - 19,5773$$
$$a = 335,17$$

SAIBA MAIS

A análise de regressão também pode ser feita com várias variáveis independentes sobre uma única variável dependente. A esse tipo de análise damos o nome de análise de regressão múltipla, que é muito utilizada em aplicações financeiras como renda, poupança e juros. Para ver um exemplo, leia o livro Estatística para os cursos de engenharia e informática de Pedro Barbetta.

Fonte: Barbetta (2010, p. 346).

CONSIDERAÇÕES FINAIS

Estudar o grau de relacionamento entre duas variáveis é de grande importância dentro das análises estatísticas. Para verificar o grau de associação entre duas variáveis, há a necessidade de conhecer os métodos estatísticos utilizados para tal procedimento.

Vimos, nesta unidade, duas ferramentas importantes para estudar o grau de associação entre duas características numéricas: a correlação e a regressão. Na Estatística, estuda-se casos com 1 variável. No estudo de Correlação e Regressão, deve-se levar em conta 2 ou mais variáveis. Dentre esse estudo, o principal objetivo é investigar a existência ou não de relação entres essas variáveis, quantificando a força dessa relação por meio da correlação, ou explicitando a forma dessa relação por meio da regressão.

As correlações podem ser Positivas, quando o aumento de uma variável corresponde ao aumento da outra; Negativas, quando o aumento de uma variável corresponde à diminuição da outra; Lineares, quando é possível ajustar uma reta, que podem ser fortes (quanto mais próximas da reta) ou fracas (quanto menos próximas da reta), e ainda Não Lineares, quando não é possível ajustar uma reta.

Após estabelecida uma relação linear e uma boa correlação entre as variáveis, deve-se, agora, determinar uma fórmula matemática para fazer predições de uma das variáveis por meio da outra, e a essa técnica damos o nome de análise de regressão.

É importante entender que nem sempre duas variáveis estão de fato associadas. Para isso, há necessidade da avaliação do coeficiente de determinação na análise de regressão. É, também, importante termos bom senso na hora de calcular algumas medidas, uma vez que estamos trabalhando com fórmulas matemáticas para explicar fenômenos. Assim, sempre algum valor será extraído numericamente, porém, nem sempre esses valores podem ser explicados biologicamente ou socialmente. Portanto, cabe ao pesquisador escolher quais variáveis devem participar das análises.

ATIVIDADES

1. Um estudo foi desenvolvido para verificar o quanto o comprimento de um cabo da porta serial de microcomputadores influencia na qualidade da transmissão de dados, medida pelo número de falhas em 10000 lotes de dados transmitidos (taxa de falha) (BARBETTA et al., 2010).

   Os resultados foram:

COMPRIMENTO DO CABO (M)	TAXA DE FALHA
8	2,2
8	2,1
9	3,0
9	2,9
10	44,1
10	4,5
11	6,2
11	5,9
12	9,8
12	8,7
13	12,5
13	13,1
14	19,3
14	17,4
15	28,2

Fonte: as autoras.

Desenvolva os exercícios abaixo:

a. Explique quem é a variável independente (x) e a dependente (y).

b. Demonstre e interprete o valor da correlação entre o comprimento do cabo e a taxa de falha.

c. Verifique a significância da correlação populacional em nível de 1% de erro.

d. Explique a significância da correlação por meio dos intervalos.

e. Demonstre os valores de b e de a na análise de regressão linear.

f. Demonstre os valores das equações de predição ($y = a + bx$) para todos os comprimentos de cabo mostrados na tabela.

ATIVIDADES

g. Demonstre o diagrama de dispersão entre os valores dos comprimentos dos cabos (x) e das taxas de falhas preditas ($\hat{y}$).

h. Calcule e interprete o coeficiente de determinação e de alienação.

i. Explique a diferença entre a análise de correlação linear e de regressão linear.

2. Uma pesquisa foi realizada para verificar o efeito da área (m^2) sobre o preço de terrenos na cidade de Mogi Mirim-SP. Considere a equação $y = 20 + 0{,}5x$ para estimar os preços em função da área. **Considerando terrenos com 200, 300 e 400 m^2, estime o preço de cada terreno.**

3. É esperado que a massa muscular de uma pessoa diminua com a idade. Para estudar essa relação, uma nutricionista selecionou 18 mulheres, com idade entre 40 e 79 anos, e observou em cada uma delas a idade (X) e a massa muscular (Y). **Determine a correlação linear entre a idade a massa muscular.**

Relação entre massa muscular e idade

MASSA MUSCULAR (Y)	IDADE (X)
82.0	71.0
91.0	64.0
100.0	43.0
68.0	67.0
87.0	56.0
73.0	73.0
78.0	68.0
80.0	56.0
65.0	76.0
84.0	65.0
116.0	45.0
76.0	58.0
97.0	45.0
100.0	53.0
105.0	49.0
77.0	78.0
73.0	73.0
78.0	68.0

Fonte: as autoras.

Hipertensão arterial e consumo de sal em população urbana

A hipertensão arterial é considerada um problema de saúde pública por sua magnitude, risco e dificuldades no seu controle. É também reconhecida como um dos mais importantes fatores de risco para o desenvolvimento do acidente vascular cerebral e infarto do miocárdio.

Vários estudos populacionais evidenciam a importância do controle da hipertensão para a redução da morbimortalidade cardiovascular. Desta forma, as elevadas taxas de morbimortalidade cardiovascular em países de industrialização recente parecem depender de modo importante da elevada prevalência de hipertensão arterial nesses países. Apesar de não se dispor de estudos com boa representatividade em nível nacional sobre a hipertensão arterial no Brasil, pesquisas localizadas mostram prevalências elevadas, situando-se no patamar de 20 a 45% da população adulta.

Na maioria dos casos, desconhece-se a causa da hipertensão arterial. Porém, vários são os fatores que podem estar associados à elevação da pressão arterial como o sedentarismo, o estresse, o tabagismo, o envelhecimento, a história familiar, a raça, o gênero, o peso e os fatores dietéticos.

Apesar de consolidada a relação entre hipertensão arterial e os fatores nutricionais, ainda não são bem esclarecidos os mecanismos de atuação destes sobre a elevação da pressão arterial. São conhecidos, no entanto, os efeitos de uma dieta saudável (rica em frutas e vegetais e pobre em gordura) sobre o comportamento dos níveis pressóricos. Dentre os fatores nutricionais estudados e que se associam à alta prevalência de hipertensão arterial estão o elevado consumo de álcool e sódio e excesso de peso. Recentemente vêm sendo, também, associados o consumo de potássio, cálcio e magnésio, os quais atenuariam o progressivo aumento dos níveis pressóricos com a idade.

A avaliação dietética de sódio é extremamente complexa, já que sua ingestão diária varia substancialmente e pode subestimar a quantidade de sódio ingerida, pois não leva em consideração as diferenças interpessoais na adição de sal. Além disso, outro problema encontrado para a realização da avaliação dietética é a tabela de composição de alimentos utilizada, que pode variar muito de um país para o outro e não contemplar preparações regionais e os produtos industrializados produzidos internamente.

Levando-se em consideração que mais de 95% do sódio ingerido é excretado na urina, e que a avaliação dietética apresenta muitos problemas operacionais, a excreção urinária de 24h vem sendo utilizada como um marcador do consumo diário de sódio, apesar da grande variabilidade intraindividual. Assim sendo, interpretações clínicas e fisiológicas baseadas numa única avaliação devem ser cautelosas. Este problema, porém, pode ser superado em estudos de base populacional, visto que a excreção urinária de sódio é considerada um bom índice de consumo de sal num dado dia.

A maior parte dos estudos que visa associar o consumo de sódio à hipertensão arterial utiliza a excreção urinária de sódio de 24h como marcador diário e em muitos há uma consistente relação. Evidências sobre a associação do consumo de sódio e hipertensão foram relatadas também pelo INTERSALT Group, principalmente quando foram avaliadas as diferenças nas prevalências de hipertensão arterial associadas ao nível de industrialização das populações estudadas. Populações ocidentais e com alto consumo de sal apareceram como tendo os maiores percentuais de hipertensão, enquanto as populações rurais ou primitivas que não faziam uso de sal de adição apresentaram menores prevalências ou nenhum caso de hipertensão arterial. Porém, o sobrepeso e o sedentarismo, presentes nessas populações, podem ser importantes variáveis de confusão. Outros estudos foram conduzidos nesta direção em várias populações com objetivo de comprovar a hipótese de que uma grande ingestão de sal na dieta aumenta os níveis pressóricos, independentemente da idade e de outros fatores, hoje já bem estabelecidos.

Fonte: Bisi Molina et al. (2003)

MATERIAL COMPLEMENTAR

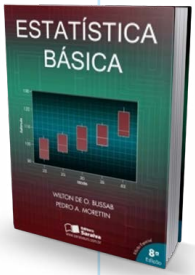
LIVRO

Estatística Básica

Wilton de O. Bussab, Pedro A. Morettin

Editora: Saraiva

Sinopse: O livro trata da estatística básica e é dividido em três partes. A primeira trata da análise de dados unidimensionais e bidimensionais, com atenção especial para métodos gráficos. A segunda parte trata dos conceitos básicos de probabilidades e variáveis aleatórias. Por fim, a terceira parte estuda os tópicos principais da interferência estatística, além de alguns temas especiais, como regressão linear simples.



REFERÊNCIAS

BARBETTA, P. A. et al. **Estatística para os cursos de engenharia e informática**. 3. ed. São Paulo: Atlas, 2010.

BISI MOLINA, M. D. C.; CUNHA, R. de S.; HERKENHOFF, L. F.; MILL, J. G. Hipertensão arterial e consumo de sal em população urbana. ***Rev. Saúde Pública*** [on-line]. 2003, v. 37, n. 6, p. 743-750. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102003000600009&lng=en&nrm=iso&tlng=pt>. Acesso em: 25 abr. 2017.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. São Paulo: Saraiva, 2003.

SESA/ISEP/CIDS - Departamento de Sistemas de Informação em Saúde. Situação da Mortalidade Infantil no Paraná – Triênio 2000-2002. Disponível em: <http://www.saude.pr.gov.br/arquivos/File/SPP_Arquivos/comite_mort_mat_infant/infantil/15SITUA__O_DA_MORTALIDADE_INFANTIL_NO_PARAN__2000_2002.pdf>. Acesso em: 26 abr. 2017.

GABARITO

1.

a. A variável independente é o comprimento do cabo x e a variável dependente é a variável taxa de falhas y, ou seja, a taxa de falhas depende do comprimento do cabo.

b. 0,47 – verifica-se uma correlação mediana e positiva entre as duas variáveis a, mostrando que, quanto maior o comprimento do cabo, maior a taxa de falhas.

c. $tc = 2,17 \quad t_{tabelado} = 3,06$

$tc < t_{tabelado}$ conclui-se que não existe significância na correlação entre as duas variáveis em nível de 1% de erro.

d. Correlação substancial entre as duas variáveis.

e. $b = 2,4 \quad a = -15,2$

f.

COMPRIMENTO DO CABO (M)	Ŷ
8	4,08
9	6,49
10	8,9
11	11,31
12	13,72
13	16,13
14	18,54
15	20,95

g. Diagrama de dispersão

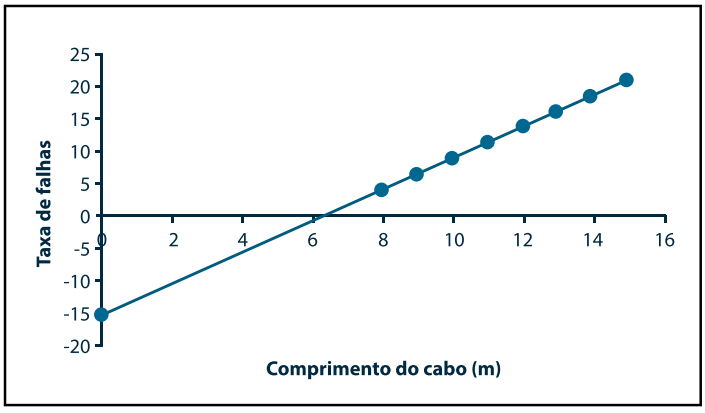


Gráfico 01 - Diagrama de dispersão entre o comprimento do cabo e as taxas de falhas preditas
Fonte: as autoras.

h. $R^2 = 0,47^2 = 0,22$ $k = 0,78$

22% dos dados são explicados pela equação linear de regressão.

Pelo coeficiente de alienação, observamos que há mais ausência que relação entre as duas variáveis.

i. Na correlação linear, medimos somente o grau de associação linear entre duas variáveis. Na análise de regressão, porém, medimos o quanto de variação de uma variável é explicada pela outra; podemos, também, fazer predições de uma variável baseada em outra.

2. 120; 170; 220.

3. -0,837.

CONCLUSÃO

Caro(a) aluno(a)!

Este material foi feito para contribuir com seu processo de formação. Atualmente, as informações chegam a nós de forma rápida e não podemos deixar de pensar o quanto a Estatística é útil para quem precisa tomar decisões.

Nesse sentido, a Estatística aparece como suporte na compreensão dos fatos, dando base para o seu entendimento e compreensão adequada para eles.

Este material tratou de alguns pontos importantes no ensino da Estatística. O primeiro ponto tratou da importância da Estatística, dos seus conceitos e da aplicação de algumas de suas ferramentas. Em qualquer pesquisa, seja ela de ordem observacional ou experimental, utilizamos a Estatística.

Na unidade II, foram discutidas formas de apresentação dos dados estatísticos, mais especificamente a estruturação e a interpretação de gráficos e tabelas.

A unidade III tratou das medidas descritivas, mostrou como devemos calculá-las e onde devemos aplicá-las. Vimos as principais medidas de posição, as separatrizes e as medidas de dispersão.

Na unidade IV, trabalhamos com parte da teoria das probabilidades e algumas de suas principais distribuições. As distribuições de probabilidades, vistas também nessa unidade, lida com probabilidades, porém associadas ao tipo de variável aleatória em questão. Para utilizarmos qualquer distribuição, é necessário saber se a variável aleatória numérica é contínua, discreta.

Finalizando o material, a unidade V tratou das medidas de associação, duas ferramentas importantes dentro da estatística. Tanto a correlação quanto a regressão envolvem associação entre variáveis, embora a função de cada uma delas seja diferente. Na correlação, tem-se o grau de associação entre as duas variáveis; na regressão, o que se obtém é estimação de uma variável por meio da outra. Entender como duas variáveis se relacionam é importante dentro da análise de dados. Alguns critérios devem ser seguidos, entretanto, ao se trabalhar com regressão ou correlação, essas medidas só podem ser utilizadas em casos de variáveis quantitativas.

Professora Me. Ivnna Gurniski de Oliveira

Professora Me. Renata Cristina de Souza Chatalov