

Universidade Federal de Santa Catarina
Pró-Reitoria de Ensino de Graduação
Departamento de Ensino de Graduação a Distância
Centro Sócio-Econômico
Departamento de Ciências da Administração

Estatística Aplicada à Administração

Professor

Marcelo Menezes Reis

2012
2ª edição revisada e atualizada

R375e Reis, Marcelo Menezes

Estatística aplicada à administração / Marcelo Menezes Reis. –
2. ed. rev. atual. – Florianópolis: Departamento de Ciências da Administração/UFSC, 2012.

280p. : il.

Inclui bibliografia

Curso de Graduação em Administração, Modalidade a Distância
ISBN: 978-85-7988-149-7

1. Matemática na administração. 2. Administração – Métodos Estatísticos. 3. Probabilidades. 4. Educação a distância. I. Título.

CDU: 51-77: 65

Catálogo na publicação por: Onélia Silva Guimarães CRB-14/071

PRESIDENTA DA REPÚBLICA – *Dilma Vana Rousseff*
MINISTRO DA EDUCAÇÃO – *Aloizio Mercadante*
COORDENADOR DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR – *Jorge Almeida Guimarães*
DIRETOR DE EDUCAÇÃO A DISTÂNCIA – *João Carlos Teatini de Souza Clímaco*

UNIVERSIDADE FEDERAL DE SANTA CATARINA

REITOR – *Alvaro Toubes Prata*
VICE-REITOR – *Carlos Alberto Justo da Silva*
PRÓ-REITORA DE ENSINO DE GRADUAÇÃO – *Yara Maria Rauh Muller*
DIRETOR DE GESTÃO E DESENVOLVIMENTO ACADÊMICO – *Carlos José de Carvalho Pinto*
COORDENADORA UAB – *Eleonora Milano Falcão Vieira*

CENTRO SÓCIO-ECONÔMICO

DIRETOR – *Ricardo José Araújo Oliveira*
VICE-DIRETOR – *Alexandre Marino Costa*

DEPARTAMENTO DE CIÊNCIAS DA ADMINISTRAÇÃO

CHEFE DO DEPARTAMENTO – *Gilberto de Oliveira Moritz*
SUBCHEFE DO DEPARTAMENTO – *Marcos Baptista Lopez Dalmau*
COORDENADOR DE CURSO – *Rogério da Silva Nunes*
SUBCOORDENADOR DE CURSO – *Sinésio Stefano Dubiela Ostroski*

COMISSÃO DE PLANEJAMENTO, ORGANIZAÇÃO E FUNCIONAMENTO

Alexandre Marino Costa (Presidente)
Gilberto de Oliveira Moritz
Luiz Salgado Klaes
Marcos Baptista Lopez Dalmau
Maurício Fernandes Pereira
Raimundo Nonato de Oliveira Lima

COORDENADOR DE TUTORIA – *Marilda Todescat*
COORDENADOR DE POLOS – *Luiz Salgado Klaes*
SUBCOORDENADOR DE POLOS – *Allan Augusto Platt*
COORDENADOR ACADÊMICO – *Irineu Manoel de Souza*
COORDENADOR DE APOIO E ACESSORAMENTO PEDAGÓGICO – *Raimundo Nonato de Oliveira Lima*
COORDENADOR FINANCEIRO – *Alexandre Marino Costa*
COORDENADOR DE AMBIENTE VIRTUAL DE ENSINO-APRENDIZAGEM (AVEA) – *Mário de Souza Almeida*
COORDENADOR EDITORIAL – *Luís Moretto Neto*
COMISSÃO EDITORIAL E DE REVISÃO – *Alessandra de Linhares Jacobsen*
Mauricio Roque Serva de Oliveira
Paulo Otolini Garrido
Claudelino Martins Dias Junior

COORDENAÇÃO DE PRODUÇÃO DE RECURSOS DIDÁTICOS – *Denise Aparecida Bunn*

DESIGN INSTRUCIONAL – *Denise Aparecida Bunn*
Patrícia Regina da Costa

PROJETO GRÁFICO E DIAGRAMAÇÃO – *Annye Cristiny Tessaro*

REVISÃO DE PORTUGUÊS – *Mara Aparecida Andrade da Rosa Siqueira*
Sergio Meira

ORGANIZAÇÃO DO CONTEÚDO – *Marcelo Menezes Reis*

POLOS DE APOIO PRESENCIAL

CRUZEIRO DO OESTE – PR

PREFEITO – *Zeca Dirceu*

COORDENADORA DE POLO – *Maria Florinda Santos Risseto*

CIDADE GAÚCHA – PR

PREFEITO – *Vitor Manoel Alcobia Leitão*

COORDENADORA DE POLO – *Eliane da Silva Ribeiro*

PARANAGUÁ – PR

PREFEITO – *José Baka Filho*

COORDENADORA DE POLO – *Meire A. Xavier Nascimento*

HULHA NEGRA – RS

PREFEITO – *Carlos Renato Teixeira Machado*

COORDENADORA DE POLO – *Margarida de Souza Corrêa*

JACUIZINHO – RS

PREFEITO – *Diniz José Fernandes*

COORDENADORA DE POLO – *Jaqueline Konzen de Oliveira*

TIO HUGO – RS

PREFEITO – *Verno Aldair Muller*

COORDENADORA DE POLO – *Fabiane Kuhn*

SEBERI – RS

PREFEITO – *Marcelino Galvão Bueno Sobrinho*

COORDENADORA DE POLO – *Ana Lúcia Rodrigues Guterra*

TAPEJARA – RS

PREFEITO – *Seger Luiz Menegaz*

COORDENADORA DE POLO – *Loreci Maria Biasi*

SÃO FRANCISCO DE PAULA – RS

PREFEITO – *Décio Antônio Colla*

COORDENADORA DE POLO – *Maria Lúcia da Silva Teixeira*

Apresentação

Caro estudante!

Toda vez que alguém ouve a palavra “Estatística”, as reações costumam ser de aversão, medo, negação da importância, restrições ideológicas, inclusive; e, sempre, a noção de que se trata de algo muito complicado... “É matemática braba”, “são muitas fórmulas difíceis”, “pode-se obter qualquer resultado com Estatística”, “métodos quantitativos são dispensáveis”, “não se aplica à minha realidade” são algumas das expressões que ouvi nesses anos em que leciono a disciplina. Talvez você tenha ouvido tais expressões também, mas eu lhe asseguro que elas são exageradas ou mesmo falsas. É preciso acabar com alguns mitos e mostrar a importância que a Estatística tem na formação do administrador.

E não se esqueça: o essencial no ato de administrar está em tomar decisões.

É por isso que esta disciplina faz parte do currículo de Administração, afinal, por meio dela você aprende como obter dados confiáveis (conceitos de planejamento de pesquisa estatística e amostragem); como resumir e organizá-los (análise exploratória de dados) a fim de empregar técnicas apropriadas (probabilidade aplicada e inferência estatística) e, desse modo, generaliza os resultados encontrados para, então, tomar decisões. Assim, apresento exemplos concretos de aplicação, que empregam ferramentas computacionais simples (como as planilhas eletrônicas, conhecidas na disciplina de *Informática Básica*).

Você está iniciando a disciplina de *Estatística Aplicada à Administração* e, acredite, o domínio dos métodos estatísticos significará importante diferencial, pois permitirá tomar decisões acertadas, o que, em essência, é o objetivo primordial de qualquer organização.

Sucesso em sua caminhada!

Prof. Marcelo Menezes Reis

Sumário

Unidade 1 – Introdução à Estatística e ao Planejamento Estatístico

Definição e Subdivisões da Estatística.	13
Conceito de Estatística	14
Variabilidade.	14
Importância da Estatística para o Administrador	17
Planejamento Estatístico de Pesquisa.	18
Resumindo.	31
Atividades de aprendizagem.	32

Unidade 2 – Técnicas de Amostragem

Técnicas e Definições de Amostragem.	37
O que é Amostragem?	37
Aspectos Necessários para o Sucesso da Amostragem	41
Plano de Amostragem	44
Cálculo do Tamanho de uma Amostra Probabilística (Aleatória)	55
Resumindo.	59
Atividades de aprendizagem.	61

Unidade 3 – Análise Exploratória de Dados I

O que é Análise Exploratória de Dados?	67
Estrutura de um Arquivo de Dados	68
Resumindo.	92
Atividades de aprendizagem.	94

Unidade 4 – Análise Exploratória de Dados II

Medidas de Posição ou de Tendência Central	99
Média ($\bar{x}$).	100
Mediana (Md).	104
Moda (Mo).	106
Quartis.	107
Medidas de Dispersão ou de Variabilidade.	109
Cálculo de Medidas de Síntese de uma Variável em Função dos Valores de Outra.	115
Resumindo.	120
Atividades de aprendizagem.	121

Unidade 5 – Conceitos Básicos da Probabilidade

Probabilidade: conceitos gerais	125
Definições Prévias	127
Definições de Probabilidade.	132
Regra do Produto.	141
Eventos Independentes.	145
Resumindo.	147
Atividades de aprendizagem.	148

Unidade 6 – Variáveis Aleatórias

Conceito de Variável Aleatória	153
Distribuições de Probabilidades para Variáveis Aleatórias Discretas	155
Distribuições de Probabilidades para Variáveis Aleatórias Contínuas	159
Valor Esperado e Variância.	160
Resumindo.	163
Atividades de aprendizagem.	165

Unidade 7 – Modelos Probabilísticos Mais Comuns

Modelos Probabilísticos para Variáveis Aleatórias Discretas	169
Modelo Binomial	170
Modelo de Poisson.	174
Modelos para Variáveis Aleatórias Contínuas	177
Resumindo.	198
Atividades de aprendizagem.	199

Unidade 8 – Inferência Estatística e Distribuição Amostral

Conceito de Inferência Estatística	205
Parâmetros e Estatísticas	207
Distribuição Amostral	209
Resumindo.	215
Atividades de aprendizagem.	217

Unidade 9 – Estimação de Parâmetros

Estimação por Ponto de Parâmetros	221
Estimação por Ponto dos Principais Parâmetros	222
Tamanho Mínimo de Amostra para Estimação por Intervalo.	233
“Empate Técnico”.	237
Resumindo.	239
Atividades de aprendizagem.	240

Unidade 10 – Testes de Hipóteses

Lógica dos Testes de Hipóteses	245
Tipos de Hipóteses	248
Tipos de Testes Paramétricos.	249
Testes de Hipóteses sobre a Média de uma Variável em uma População	257
Testes de Hipóteses sobre a Proporção de uma Variável em uma População. . .	262
Teste de Associação de Quiquadrado	265
Resumindo.	274
Atividades de aprendizagem.	275
Referências.	277
Minicurriculo.	280

1

UNIDADE

Introdução à Estatística e ao Planejamento Estatístico



Ao finalizar esta Unidade, você deverá ser capaz de compreender o conceito de Estatística; de perceber a importância dela para o administrador; e de entender os principais aspectos do planejamento estatístico, que garantem a obtenção de dados confiáveis.

Definição e Subdivisões da Estatística

Caro estudante, seja bem-vindo!

Convidamos você a ingressar conosco neste universo amplo, desafiador e instigante da **Estatística**. A partir da leitura do material, poderemos construir e socializar saberes, articular teoria e prática. Portanto, estamos em campo fértil para debates, por isso teremos muito a discutir. Não esqueça que dúvidas e indagações são sempre pertinentes, já que delinearão o processo a que estamos nos dispondo, coletivamente, nesta disciplina.

Não é possível tomar decisões corretas sem dados confiáveis. Os governantes do Antigo Egito e da Suméria (seus administradores) já sabiam disso, por essa razão, mandavam seus escribas registrar e compilar os dados da produção agrícola e dos homens aptos para o serviço militar. Em outras palavras, eles já usavam métodos estatísticos: a raiz da palavra Estatística vem de Estado. Com o passar do tempo e diante da expansão do conhecimento, os métodos estatísticos tornaram-se mais sofisticados com a adoção de modelos probabilísticos e da inferência estatística e, nos últimos 30 anos, observa-se a utilização de computadores não apenas pelos governos, mas também por empresas, universidades e pessoas comuns.

A intensiva aplicação da informática possibilitou a automatização de muitos cálculos e a busca por informações em gigantescas bases de dados, o que vem constituindo o campo de conhecimento de mineração de dados e inteligência empresarial.

Todo administrador, atualmente, precisa usar métodos estatísticos e, para tanto, precisa conhecê-los, a começar por suas definições e subdivisões, o que veremos nesta Unidade, além de apresentarmos os conceitos de Estatística e de Planejamento Estatístico.

Conceito de Estatística

"Estatística é a Ciência que permite obter conclusões a partir de dados"

(Paul Velleman).

Estatística é a Ciência que parte de perguntas e desafios do mundo real. Veja os exemplos:

- cientistas querem verificar se uma nova droga consegue eliminar o vírus HIV;
- uma montadora de automóveis precisa saber a qualidade de um lote inteiro de peças fornecidas através de uma pequena amostra;
- a um político interessa conhecer o percentual de eleitores que votarão nele nas próximas eleições;
- um empresário deseja saber se há mercado potencial para abrir uma casa noturna em determinado bairro da cidade; e
- aos investidores saber em quais ações devem investir para obter maior rendimento.

Variabilidade

Variabilidade – diferenças encontradas por sucessivas medições realizadas em pessoas, animais ou objetos, em tempos ou situações diferentes. Fonte: Montgomery (2004).

O principal problema que surge ao tentar responder essas perguntas é que todas as medidas feitas para tal, por mais acurados que sejam os instrumentos de medição, apresentarão sempre uma variabilidade, ou seja, não há respostas perfeitas. Feliz ou infelizmente, a natureza comporta-se de forma variável: não há dois seres humanos iguais, não há dois insetos iguais, não há dois consumidores iguais. Mesmo os tão comentados “clones” ou os gêmeos idênticos (“clones” naturais) somente apresentam um código genético comum, pois, se submetidos à experiências de vida diferentes, terão um desenvolvimento distinto. Sendo assim, variabilidade é **inevitável** e **inerente** à vida.

Antes de prosseguir, faça uma reflexão sobre as seguintes questões:

Suas preferências musicais são as mesmas de dez anos atrás? Sua aparência é a mesma de há dez anos? Você votaria no mesmo candidato a deputado federal em que votou na última eleição? Você permanece com o mesmo peso de dez anos atrás? Imagine, então, as diferenças que existem de pessoa para pessoa, de cidade para cidade, de povo para povo...

A Estatística permite descrever e identificar as fontes e até indicar meios de controlar a variabilidade. Vamos apresentar as suas subdivisões para que você entenda como isso ocorre.

Subdivisões da Estatística

Os dados são coletados para responder a uma pergunta do mundo real. Para respondê-la, é preciso estudar uma ou mais características de uma População de interesse.

Como o interesse maior está na população, o ideal seria pesquisá-la na sua integralidade, em suma, realizar um **censo** (como o IBGE faz periodicamente no Brasil). Contudo, por razões econômicas ou práticas (para obter rapidamente a informação ou evitar a extinção ou exaustão da população) nem sempre é possível realizar um censo, então, torna-se necessário pesquisar apenas uma amostra da população, um subconjunto finito e representativo da população.

Para Saber sobre o censo no Brasil, acesse o site do IBGE: <<http://www.ibge.gov.br>>. Acesso em: 18 abr. 2012.

Às etapas dos parágrafos anteriores, somam-se outros tópicos que estudaremos mais adiante, para constituir o **planejamento estatístico** da pesquisa.

Lembre-se: a qualidade de uma pesquisa nunca será melhor do que a qualidade dos seus dados.

Probabilidade – medida da possibilidade relativa de ocorrência de um evento qualquer relacionado a certo fenômeno, pode ser calculada através da definição de um modelo probabilístico para o fenômeno. Fonte: Lopes (1999).

Uma das principais subdivisões da Estatística é a **Amostragem**, que reúne os métodos necessários para coletar adequadamente amostras representativas e suficientes para que os resultados obtidos possam ser generalizados para a população de interesse.

Após a coleta dos dados, por censo ou amostragem, a Análise Exploratória de Dados permite apresentá-los e resumi-los de maneira que seja possível identificar padrões e elaborar as primeiras conclusões a respeito da população. Em suma, descrever a **variabilidade** encontrada. Se a pesquisa foi feita por censo, basta realizar a análise exploratória de dados para obter as conclusões.

Posteriormente, através da Inferência Estatística, é possível generalizar as conclusões oriundas dos dados para a população quando estes forem provenientes de uma **amostra** utilizando-se a probabilidade para calcular a confiabilidade das conclusões obtidas.

A Figura 1 ilustra a subdivisão da Estatística. Veja:

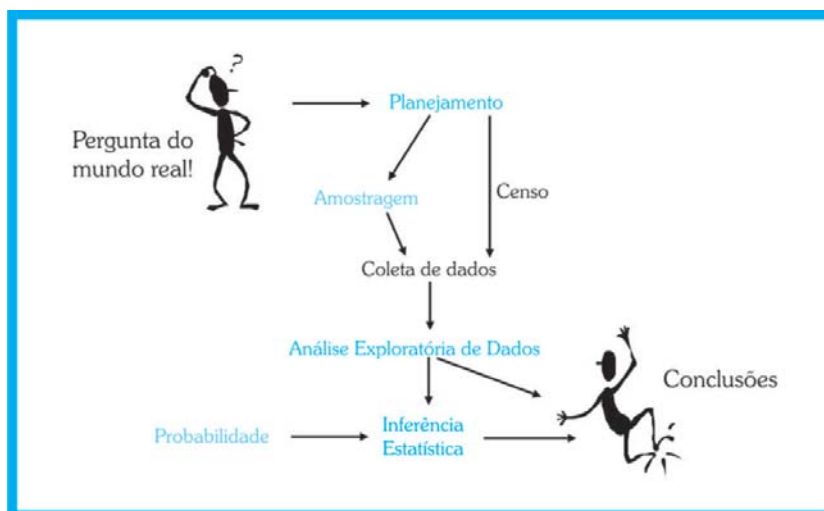


Figura 1: Subdivisões da Estatística
Fonte: Elaborada pelo autor deste livro

Importância da Estatística para o Administrador

O administrador precisa tomar decisões. Para tanto, precisa de informações confiáveis, mas já sabemos que para obtê-las é preciso coletar dados e resumi-los. Posteriormente é necessário interpretá-los, levando em conta a variabilidade inerente e inevitável

em todos os fenômenos. Como a Estatística fornece os meios para todas essas etapas, trata-se de um conhecimento indispensável para o administrador.

Não se esqueça: em qualquer profissão é preciso analisar dados (verificar a confiabilidade das fontes), relacioná-los ao contexto ao qual estão inseridos e, por várias vezes, compará-los com dados passados a fim de prever o comportamento futuro. Veja o exemplo a seguir (Figura 2) extraído de um jornal de grande circulação.

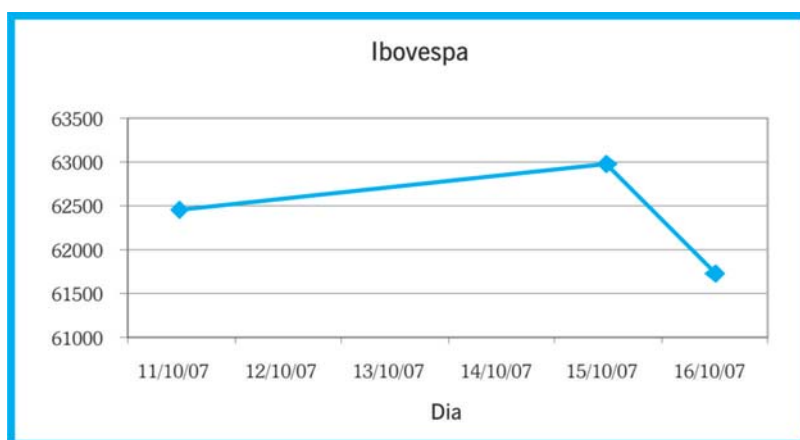


Figura 2: Variação do IBOVESPA de 11/10/2007 a 16/10/2007
Fonte: Adaptada de Diário Catarinense (2007)

Através de um simples [gráfico de linhas](#), podemos observar uma “queda” no Índice da Bolsa de Valores de São Paulo – Ibovespa, entre 15 e 16 de outubro de 2007, motivada pelo temor da depreciação nas bolsas internacionais, por sua vez causada pelo possível agravamento da crise imobiliária nos EUA. Tal constatação refere-se à situação contemporânea relacionada ao contexto internacional; e ao mesmo tempo em que se verifica a queda, sabemos que o Ibovespa nunca antes havia ultrapassado 60 mil pontos, o que, para a Estatística, representa, nesse caso, considerar os dados do passado. Os investidores em ações negociadas na Bovespa provavelmente tomarão novas decisões, sobre a compra e a venda de ações, baseadas nessas informações.

Na Unidade 3, você conhecerá o tema com mais detalhes.

O primeiro passo para qualquer análise bem-sucedida é obter dados confiáveis. Por isso, vamos iniciar o tópico de Planejamento Estatístico.

Planejamento Estatístico de Pesquisa

Planejamento estatístico de pesquisa – conjunto de métodos cuja implementação visa garantir a confiabilidade dos dados coletados. Fonte: Barbetta, Reis e Bornia (2008).

Essas e outras informações você encontra em: <http://www.crasc.org.br/index.php?pg=inicial/oque.htm>. Acesso em: 18 abr. 2012.

O planejamento estatístico de pesquisa é parte do planejamento geral da pesquisa.

Antes de se pensar em qualquer abordagem estatística, é preciso definir o que se quer pesquisar, em qualquer campo do conhecimento. “Como poderemos escolher o melhor caminho se não sabemos para onde ir”? Em outras palavras, é preciso definir corretamente a “**pergunta do mundo real**” que queremos responder, porque isso afetará profundamente as etapas do planejamento estatístico.

Para facilitar a compreensão, vamos fazer o planejamento de uma pesquisa fictícia, mas que muito auxiliará na compreensão do conteúdo. O Conselho Regional de Administração (CRA) “é um órgão consultivo, orientador, disciplinador e fiscalizador do exercício da profissão de Administrador” no qual somente bacharéis em Administração (graduados em cursos de Administração) podem registrar-se nele. O CRA preocupa-se muito com a qualidade dos cursos de Administração e, frequentemente, apresenta sugestões para aperfeiçoar currículos e disciplinas, visando à melhoria da formação dos profissionais.

Com isso em mente, imagine que o CRA de Santa Catarina está interessado em conhecer a opinião dos seus registrados sobre o curso em que se graduaram; desde que tal curso esteja situado em Santa Catarina. Esta é a “pergunta do mundo real”: qual é a opinião dos profissionais registrados no CRA-SC, e graduados no Estado, sobre o curso em que se formaram. Observe: não se falou em Estatística ainda, o CRA apenas definiu o que quer pesquisar. Agora, podemos passar ao planejamento estatístico da pesquisa.

Para realizar o planejamento estatístico, precisamos definir o objetivo geral, os objetivos específicos, a população, as variáveis, o delineamento, a forma de coleta de dados e o instrumento de pesquisa. Todos esses tópicos serão temas das próximas seções.

Objetivos da Pesquisa

Como você já sabe, há dois tipos de objetivos: o geral e o específicos. A pesquisa pode ter apenas o **objetivo geral**. Esse objetivo inclui o propósito que motivou a pesquisa, e a sua justificativa e relevância.

As características que precisam ser pesquisadas para permitir a consecução do objetivo geral são os **objetivos específicos**. Trata-se do detalhamento do objetivo geral, em que explicamos o que queremos medir (preferências, opiniões sobre fatos ou pessoas, resultados de experimentos, entre outras).

Para o nosso exemplo (pesquisa sobre os cursos de Administração de Santa Catarina), podemos enunciar os objetivos:

- **Objetivo geral:** avaliar a opinião dos registrados no CRA-SC, graduados no Estado, sobre os seus respectivos cursos.

Propósito: buscar elementos que indiquem os pontos fortes e fracos dos cursos.

Relevância: a pesquisa é relevante, pois poderá obter informações úteis para a melhoria da qualidade dos cursos de Administração. Tal melhoria certamente motivará mais os atuais e futuros acadêmicos, propiciando-lhes uma formação mais adequada e abrindo-lhes mais oportunidades. Para a sociedade, o efeito seria benéfico por contribuir para a formação de profissionais mais qualificados.

- **Objetivos específicos:**

- Avaliar a opinião dos registrados sobre o corpo docente dos seus cursos.
- Avaliar a opinião dos registrados sobre o currículo dos seus cursos.
- Avaliar a opinião dos registrados sobre a infraestrutura dos seus cursos (salas, bibliotecas, laboratórios, ventilação, limpeza, iluminação).
- Identificar as razões que levaram os registrados a escolher a instituição onde se graduaram.



Observe que é necessário “dividir” o objetivo geral em específicos para que a pesquisa possa ser executada; afinal, é através destes que vamos chegar às variáveis, tópico que estudaremos mais à frente. O próximo passo consiste em definir quem será pesquisado, ou seja, a população da pesquisa.

População

Uma parte importante do delineamento de qualquer pesquisa é a definição da população. Tal definição dependerá obviamente dos objetivos da pesquisa, das características a mensurar, dos recursos disponíveis.

“População é o conjunto de medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m)”. Se, por exemplo, estivéssemos avaliando as opiniões de eleitores sobre os candidatos à presidência, a população da pesquisa seria constituída pelas opiniões declaradas pelos eleitores em questão. A população pode se referir a seres humanos, animais e mesmo a objetos e o conjunto de medidas à altura de pessoas adultas do sexo masculino, ao peso de bois adultos ou ao diâmetro dos parafusos produzidos em uma fábrica.

É muito importante conhecer o tamanho da população. Isso ajudará a calcular os custos da pesquisa, a área de abrangência, o tempo necessário para concluí-la e os recursos necessários para fazer a tabulação e a análise dos resultados.



Com relação à pesquisa do CRA-SC, qual seria a população?

- Conjunto das opiniões dos registrados no CRA-SC, graduados no Estado, sobre os seus cursos.
- Tamanho da população: em 24/10/2007 havia 11.676 registrados no CRA-SC. Destes, por suposição, 9.000 foram graduados em faculdades catarinenses.

Com esses aspectos definidos podemos partir para a determinação das variáveis e conhecer o que efetivamente será medido.

Variáveis

Quando um fenômeno específico é estudado, determinadas características são analisadas: as **variáveis**. Por meio delas é possível descrever o fenômeno. As variáveis são características que podem ser observadas ou medidas em cada elemento pesquisado, sob as mesmas condições. Para cada variável, para cada elemento pesquisado, em dado momento, **há um e apenas um resultado possível**. Os resultados obtidos permitirão, então, a consecução dos objetivos específicos da pesquisa.

As variáveis são as medidas que precisam ser realizadas para a consecução dos objetivos específicos da pesquisa.

Tenha em mente que as variáveis precisam ser relacionadas aos objetivos específicos. Faça uma experiência com o seguinte questionamento: qual era a sua estatura aos 12 anos? Naquela época, apenas um valor possível correspondia a variável altura. No ano seguinte, **outra fase**, provavelmente a altura era diferente; como é diferente da altura que você apresenta hoje. O importante é perceber que, em cada momento, ela [altura] teve um único valor.

As variáveis podem ser classificadas de acordo com o seu **nível de mensuração** (o quanto de informação cada variável apresenta) e seu **nível de manipulação** (como uma variável relaciona-se com as outras no estudo). Observe a Figura 3 e entenda a classificação das variáveis por nível de mensuração.

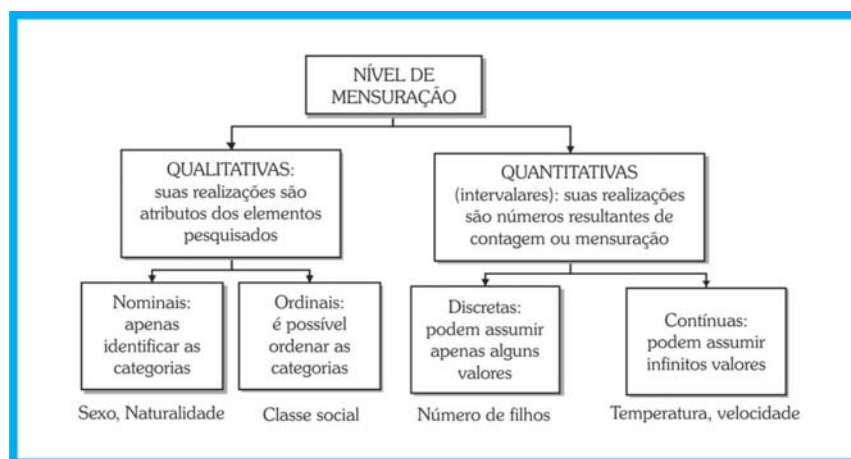


Figura 3: Classificação das variáveis por nível de mensuração

Fonte: Elaborada pelo autor deste livro

As variáveis **qualitativas** ou categóricas são aquelas cujas realizações são atributos (categorias) do elemento pesquisado, como sexo, grau de instrução e espécie. Elas podem ser nominais ou ordinais:

- As qualitativas **nominais** podem ser medidas apenas quanto aos itens (a especificação deles) pertencentes a diferentes categorias, mas não podemos quantificar nem mesmo ordenar tais categorias. Para ilustrar, podemos dizer que dois indivíduos são diferentes em relação à variável A (sexo, por exemplo), mas não podemos dizer qual deles tem mais da qualidade representada pela variável. Modelos típicos de variáveis nominais: sexo, naturalidade, cor dos olhos etc.
- As qualitativas **ordinais** permitem ordenar os itens medidos em relação à quantidade – qual tem mais ou menos – da qualidade representada pela variável; porém, não é possível precisar quanto mais. Um modelo típico de uma variável ordinal é o *status* socioeconômico das famílias residentes em uma localidade; sabemos que média-alta é mais alta do que média, mas não podemos dizer, por exemplo, que é 18% mais alta.

As variáveis **quantitativas** são aquelas cujas realizações são números resultantes de contagem ou mensuração, como número de filhos, número de clientes, velocidade em km/h, peso em kg, entre outros. Elas podem ser discretas ou contínuas:

- As quantitativas **discretas** são aquelas que podem assumir apenas alguns valores numéricos, que geralmente podem ser listados (número de filhos, número de acidentes etc.).
- As quantitativas **contínuas** são aquelas que podem assumir teoricamente qualquer valor em um intervalo (velocidade, peso etc.).

A predileção dos pesquisadores em geral por variáveis quantitativas explica-se porque elas costumam conter mais informação do que as qualitativas. Quando a variável peso de um indivíduo é descrita em termos de “magro” e “gordo”, sabemos que o gordo é mais pesado do que o magro, mas não temos ideia de quão mais pesado ele é. Se, contudo, descreve-se o peso de forma numérica, medido em quilogramas, e um indivíduo pesa 60 kg e outro pesa 90 kg, não somente sabemos que o segundo é mais pesado, mas que é 30 kg mais pesado do que o primeiro.

Você deve estar se perguntando: por que eu preciso saber disso? Deve saber porque a escolha da forma de medição da variável influenciará na qualidade dos resultados da pesquisa.

Vejamos, na Figura 4, a classificação das variáveis por nível de manipulação.

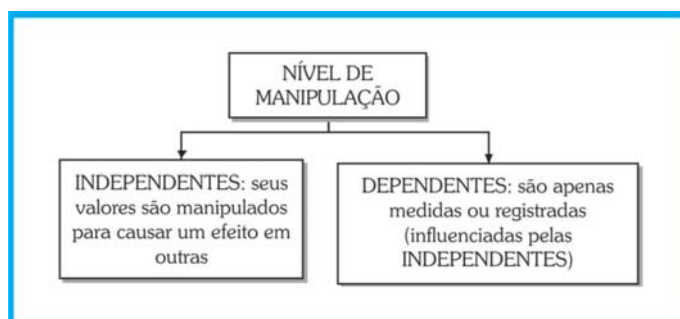


Figura 4: Classificação das variáveis por nível de manipulação

Fonte: Elaborada pelo autor deste livro

Variáveis **independentes** correspondem àquelas que são manipuladas, enquanto que as **dependentes** são apenas medidas ou

registradas como resultado da manipulação das variáveis independentes. Essa distinção confunde muitas pessoas, que dizem que “todas as variáveis dependem de alguma coisa”. Entretanto, trata-se de distinção indispensável!

Os termos variável independente e dependente aplicam-se principalmente à pesquisa experimental, em que algumas variáveis são manipuladas, e, nesse sentido, são “independentes” dos padrões de reação inicial, intenções e características das unidades experimentais. Espera-se que outras variáveis sejam “dependentes” da manipulação ou das condições experimentais. Portanto, elas dependem do que as unidades experimentais farão em resposta.

Contrariando a natureza da distinção, esses termos também são usados em estudos em que não se manipulam variáveis independentes, literalmente falando, mas apenas se designam sujeitos a “grupos experimentais” (blocos) baseados em propriedades pré-existentes dos próprios sujeitos.

Muitas vezes, fazemos a pesquisa para tentar identificar a ligação existente entre as variáveis. Em uma pesquisa eleitoral para presidente do Brasil, uma variável independente poderia ser a região do país e, a dependente, o candidato escolhido pelo eleitor pesquisado.

Vejamos um exemplo para entender esse processo de análise e observar se há relação entre as variáveis. Para isso, retomamos o exemplo da pesquisa com os registrados no CRA-SC, para o qual as variáveis a serem medidas devem definir, pelo menos, uma variável para cada objetivo específico, conforme a seguir:

Tentando identificar o primeiro objetivo específico, vamos avaliar a opinião dos registrados sobre o corpo docente dos seus cursos para definir as variáveis:

- Conhecimento sobre o conteúdo da disciplina;
- habilidade didática;
- forma de avaliação; e
- relacionamento com os estudantes.

Repare que cada um desses quatro aspectos pode ser segmentado em outros para obtenção de mais detalhes. Desse modo, como mensurá-los? Nesse caso, devemos utilizar uma escala ordinal. Observe a pergunta:

No que diz respeito ao **conhecimento teórico** sobre a disciplina X, o professor pode ser considerado:

- () ótimo
- () bom
- () satisfatório
- () insuficiente
- () horrível.

Atente que para cada acadêmico, em determinado momento, há apenas um resultado possível (já limitado no enunciado da questão) para a pergunta. Poderíamos construir perguntas semelhantes para os outros três itens e para cada objetivo específico.

Passaremos agora à definição do delineamento da pesquisa, momento em que as preocupações lógicas e teóricas das fases anteriores cedem lugar às questões mais práticas de verificação.

Delineamento da Pesquisa

Conhecendo os objetivos da pesquisa, a população e as variáveis, precisamos definir como ela será conduzida. Há basicamente dois modos de fazê-lo: **levantamento** e **experimento**.

A maioria das pesquisas socioeconômicas é encaminhada por **levantamento**, em que o pesquisador usualmente apenas registra os dados, através de um questionário ou de qualquer outro instrumento de pesquisa. Procura-se responder às perguntas da pesquisa por meio da identificação de associações entre as variáveis ou entre grupos de

elementos da população, mas o pesquisador não tem controle sobre as variáveis. Por esse motivo, para que os resultados sejam confiáveis, faz-se necessária a obtenção de grande conjunto de dados.

A nossa pesquisa com os registrados no Conselho Regional de Administração de Santa Catarina, poderia ser encaminhada por um levantamento, através da aplicação de um questionário aos acadêmicos de Administração, conforme a Figura 5:

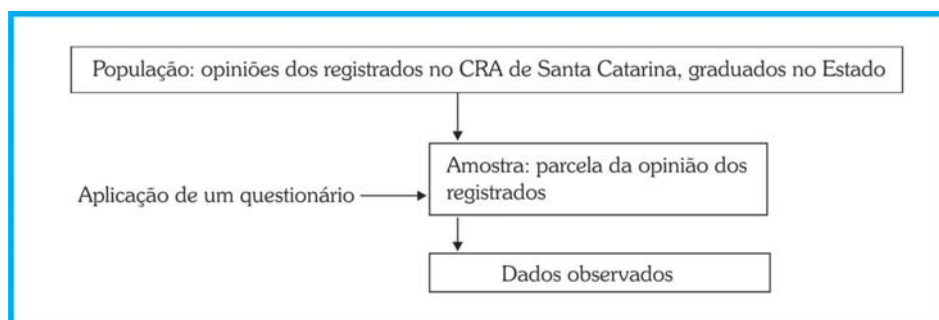


Figura 5: Pesquisa por levantamento
Fonte: Adaptada de Barbetta (2007)

Quando há absoluta necessidade (e viabilidade) de provar relações de causa e efeito, o delineamento apropriado é o **experimento**. Nesse tipo de delineamento podemos manipular algumas variáveis para observar o efeito em outras, removendo (ou tentando remover) todas as outras variáveis que poderiam influenciar no resultado final: assim, se o experimento for adequadamente conduzido, será possível provar, dentre outras coisas, que a variação nos valores de uma ou mais variáveis causou as mudanças. Como o pesquisador tem muito controle sobre o estudo não há necessidade de um grande conjunto de dados.

No seu dia a dia como administrador você encontrará os dois tipos de delineamento:

- para pesquisas de opinião, de mercado, de desemprego, de produção industrial, por exemplo, adota-se o modo por levantamento; e
- para pesquisas na indústria farmacêutica (sobre eficácia e segurança de medicamentos), na indústria química (quais

fatores irão propiciar maior rendimento nas reações químicas), na indústria siderúrgica (qual a composição necessária de uma liga de aço para obter a dureza especificada), por exemplo, adota-se o modo por experimento.

Forma de Coleta de Dados

Há duas formas básicas de coletar os dados: por **censo** ou por **amostragem**.

No censo, a pesquisa é realizada com **todos** os elementos da população, o que permite (teoricamente) precisão absoluta. É recomendável quando estamos reunindo dados para tomar decisões de longo alcance; por exemplo, um grande programa de controle de natalidade, ou incentivo à redução da desigualdade regional. Portanto, precisamos ter um quadro muito completo da situação atual. É exatamente isso que o IBGE faz a cada dez anos no Brasil, com o censo demográfico. Mas há também outros censos, como o industrial e o agropecuário.

Obviamente, o censo exige um grande volume de recursos e de tempo apreciável para a sua realização, consolidação dos dados, produção dos relatórios e análise dos resultados.

Nas pesquisas por amostragem, apenas uma pequena parte da população – parte esta considerada representativa – é pesquisada. Os resultados podem ser, então, generalizados para toda a população, usualmente através de métodos estatísticos apropriados. A economia de tempo e de dinheiro é evidente ao se utilizar amostragem, bem como se torna obrigatório o seu uso em casos em que há destruição ou exaustão dos elementos pesquisados, como em testes destrutivos: imagine o indivíduo interessado em testar todos os palitos de uma caixa de fósforos para ver se funcionam.

A partir de uma amostra de 3.000 eleitores podemos obter um retrato confiável da preferência do eleitorado brasileiro. Contudo, sempre há **risco** de que a amostra, por maiores que sejam os cuidados na sua retirada, não seja representativa da população.

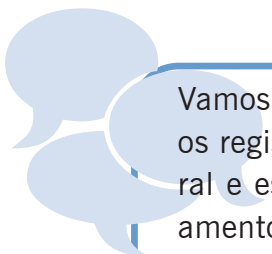
Além da decisão por censo ou amostragem devemos decidir se utilizaremos dados **primários** ou **secundários**.

Os dados secundários são dados existentes, coletados por outros pesquisadores e disponíveis em relatórios ou publicações. A sua

Na Unidade 2 você vai estudar as formas de minimizar tal risco.

utilização pode reduzir muito os custos de uma pesquisa. Se fosse necessário obter informações demográficas, poderíamos utilizar os relatórios do IBGE referentes ao último censo, ou à pesquisa nacional por amostragem de domicílios (PNAD) e não haveria a necessidade de realizar nova pesquisa.

Quando os dados não existem, ou estão ultrapassados, ou não correspondem exatamente aos objetivos de nossa pesquisa (foram coletados com outra finalidade), torna-se necessário coletar dados primários, diretamente dos elementos da população.



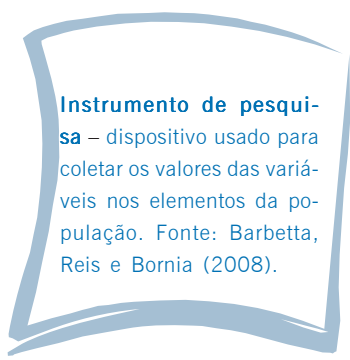
Vamos recordar o que já fizemos na pesquisa com os registrados no CRA-SC: definimos objetivos (geral e específicos), população, variáveis e o delineamento. Os dados que procuramos existem em algum lugar? Provavelmente não, ou talvez, estejam ultrapassados, o que exige que levantemos tais características diretamente dos elementos da população: precisamos obter dados primários. Como há um número muito grande de registrados, distribuídos por todo o Estado, será muito mais econômico conduzir a pesquisa por amostragem. Na Unidade 2, vamos apresentar os vários tipos de amostragem.

Quando decidimos coletar dados primários, diretamente dos elementos da população, precisamos pensar no instrumento de pesquisa: onde as variáveis serão efetivamente registradas.

Instrumento de Pesquisa

Coletamos os valores das variáveis, os dados da pesquisa, através do instrumento de pesquisa. É importante ressaltar que ele está intrinsecamente relacionado às variáveis da pesquisa. Portanto, no seu projeto, precisamos deixar claro qual é a relação existente com as variáveis, da mesma forma que as variáveis devem ser relacionadas aos objetivos específicos.

O senso comum confunde instrumento de pesquisa com questionário, o que não é verdade. O questionário é apenas um tipo de



Instrumento de pesquisa – dispositivo usado para coletar os valores das variáveis nos elementos da população. Fonte: Barbetta, Reis e Bornia (2008).

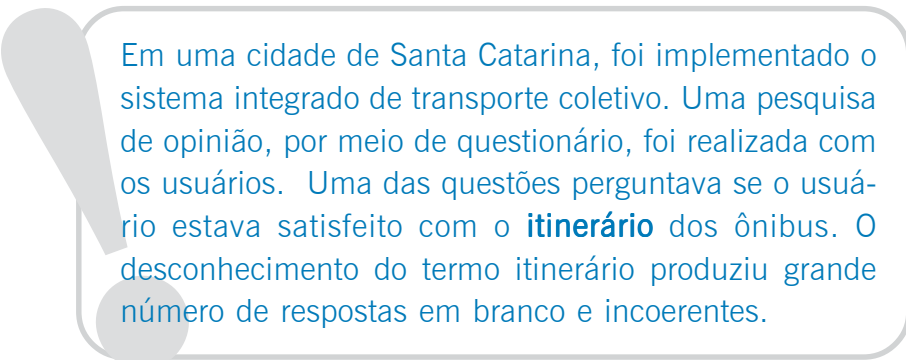
instrumento de pesquisa, e em muitas situações ele não é o mais apropriado.

Imagine a situação: queremos registrar o movimento em lojas de um Shopping Center com a finalidade de saber quais apresentam clientela suficiente para continuar a merecer a permanência. Não precisamos aplicar um questionário aos clientes, que poderão se recusar a responder; ou aos lojistas, que poderão ser “criativos demais” nas respostas basta registrar em uma **planilha** quantas pessoas entram na loja, o horário, se fizeram compras ou não, entre outros aspectos. Uma outra situação seria uma pesquisa climática, em que são registradas medidas de temperatura, de umidade relativa do ar, de velocidade do vento: obviamente, não precisamos de um questionário para isso.

O questionário torna-se quase que indispensável quando precisamos mensurar ou avaliar atitudes, preferências, crenças e comportamentos que exigem a manifestação dos pesquisados. Pesquisas de mercado, acerca da aceitação de um produto ou de uma propaganda, pesquisas de comportamento, pesquisas de opinião eleitoral, todas elas envolvem algum tipo de questionário.

O questionário pode ser enviado pelo correio, feito por telefone, feito com a presença física do entrevistador, ou mesmo via Internet. Todos eles têm suas vantagens e desvantagens.

O aspecto mais importante do questionário consiste na obtenção das informações sem induzir ou confundir o respondente. As perguntas precisam ser claras, afirmativas ou interrogativas, evitando-se negações e coerentes com o nível intelectual dos elementos da população.



Em uma cidade de Santa Catarina, foi implementado o sistema integrado de transporte coletivo. Uma pesquisa de opinião, por meio de questionário, foi realizada com os usuários. Uma das questões perguntava se o usuário estava satisfeito com o **itinerário** dos ônibus. O desconhecimento do termo itinerário produziu grande número de respostas em branco e incoerentes.

Na nossa pesquisa, precisaríamos aplicar alguma espécie de questionário. O CRA-SC dispõe de várias informações sobre os registrados, incluindo endereço postal e, talvez, o telefone e o endere-

ço eletrônico. Poderíamos enviar os questionários por um desses três meios.

Saiba mais...

Para saber mais sobre experimentos, consulte: MOORE, David S.; McCABE, George P.; DUCKWORTH, William M.; SCLOVE, Stanley L. *A prática da estatística empresarial: como usar dados para tomar decisões*. Rio de Janeiro: LTC, 2006.

Para saber mais sobre elaboração de questionários, consulte: BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Editora da UFSC, 2007, Capítulo 2.



O resumo desta Unidade está esquematizado na Figura 6.

Observe:

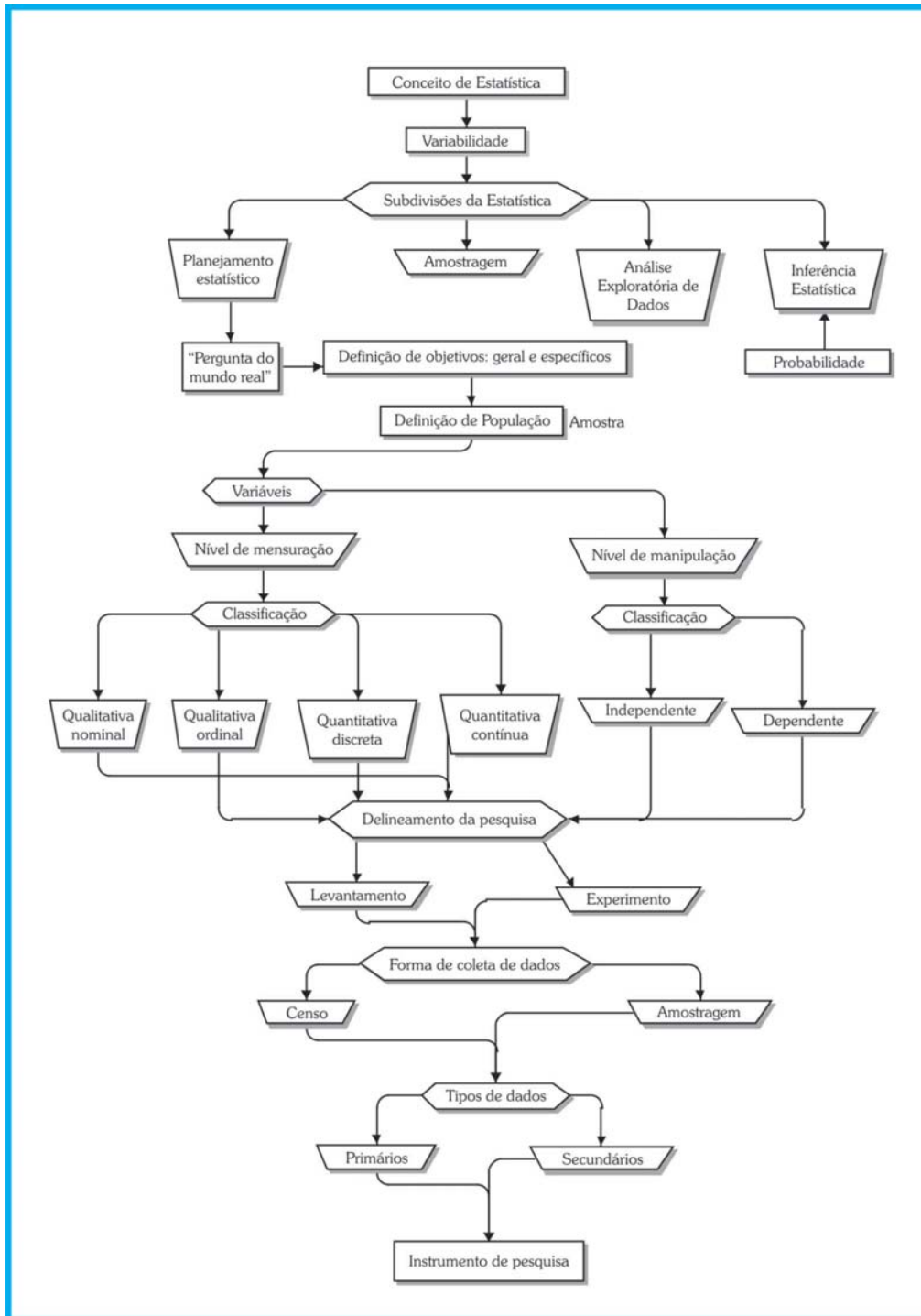


Figura 6: Resumo da Unidade 1

Fonte: Elaborada pelo autor deste livro



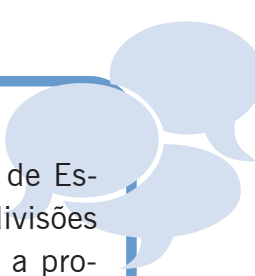
Atividades de aprendizagem



Veja se você obteve bom entendimento acerca do que tratamos nesta Unidade respondendo às questões, conforme os conceitos estudados. Ao final, encaminhe-as para seu tutor através do Ambiente Virtual de Ensino-Aprendizagem.

Bom estudos! Se precisar de auxílio, não deixe de fazer contato com seu tutor.

1. A direção do CED (Centro de Educação da UFSC) e os departamentos desse Centro têm interesse em avaliar se a biblioteca setorial está atendendo convenientemente aos alunos de graduação, de pós-graduação, aos docentes e aos técnicos-administrativos do CED. Há uma preocupação com o acervo em si (atualização, composição, número de cópias disponíveis e adequação às necessidades de cada curso) e com o atendimento aos usuários (número de atendentes, horário, cortesia). Faça o planejamento da pesquisa estabelecendo:
 - a) Objetivo geral da pesquisa.
 - b) Objetivos específicos da pesquisa.
 - c) Tipo de pesquisa (Levantamento ou Experimento).
 - d) População da pesquisa.
 - e) Quais são as variáveis da pesquisa? Por quê? Como serão medidas? Por quê?
 - f) Como serão coletados os dados: secundários ou primários, censo ou amostragem? Por quê?
 - g) Escolha do instrumento de pesquisa. Justificativa.
 - h) Elaboração do instrumento de pesquisa. Justificativa dos itens e opções escolhidas.



Caro estudante,

Fazer com que você entendesse o conceito de Estatística, as suas variabilidades e suas subdivisões na aplicação de estudos e experimentos foi a proposta desta Unidade. Com esse conhecimento, você será capaz de obter, de organizar e de analisar dados, determinando as correlações que se apresentem e tirando delas suas consequências para descrição e explicação do que se passou, além de aprender a prever e a organizar o futuro.

Leia as indicações de textos complementares, responda às atividades de aprendizagem e interaja com a equipe de tutoria. Não fique com dúvidas, questione!

Saiba que você não está sozinho nesse processo e pode contar com uma equipe de apoio para a construção do seu conhecimento.

2

UNIDADE

Técnicas de Amostragem



Ao finalizar esta Unidade, você deverá ser capaz de compreender o conceito de amostragem; de constatar suas principais técnicas; de perceber quando devem ser empregadas; de aprender a definição do plano de amostragem; e, ainda, capaz de utilizar uma fórmula simplificada para cálculo do tamanho mínimo de amostra.

Técnicas e Definições de Amostragem

Caro estudante,

Conforme vimos na Unidade 1, a amostragem é uma das formas de coleta de dados e diz respeito a uma das subdivisões da Estatística, cujo conhecimento é indispensável ao administrador. Tenha em mente que estamos interessados na obtenção de dados confiáveis para a tomada de decisões e, muitas vezes, precisaremos realizar pesquisas para coletar tais dados. Convidamos você para conhecer um pouco mais sobre essa técnica de pesquisa e seus diferentes métodos de aplicação.

Há vários argumentos para justificar a utilização da amostragem e o administrador precisa conhecê-los para que, confrontando com os recursos disponíveis e os objetivos da pesquisa, possa tomar a melhor decisão sobre a forma de coleta de dados. Todavia, há casos em que o uso da amostragem pode não ser a melhor opção; mas se o administrador decidir por ela precisará delinear o plano de amostragem, indicando como esta será implementada e qual será seu tamanho, item crucial que influenciará muito nos custos da pesquisa. Veremos isso em detalhes nesta Unidade.

O que é Amostragem?

Amostragem é uma das subdivisões da Estatística que reúne os métodos necessários para coletar adequadamente amostras representativas e suficientes para que os resultados obtidos possam ser generalizados para a população de interesse. A pressuposição básica é que todas as etapas prévias do planejamento da pesquisa já tenham sido cumpridas e que o administrador, após isso, decidirá se a coleta dos dados será por censo ou por amostragem.

Censo – forma de coleta de dados em que a pesquisa é realizada com todos os elementos da população. Fonte: Barbetta, Reis e Bornia (2008).

Amostragem – forma de coleta de dados em que apenas uma pequena parte, considerada representativa, da população é pesquisada. Os resultados podem ser então generalizados, usualmente através de métodos estatísticos apropriados, para toda a população. Fonte: Barbetta (2007).

O censo consiste simplesmente em estudar todos os elementos da **população** e a amostragem pesquisa apenas uma pequena parte dela, suposta representativa do todo. Para realizar um estudo por amostragem de maneira que seus resultados sejam válidos e possam ser generalizados para a população, algumas técnicas precisam ser empregadas. A essência desse processo é mostrada na Figura 7 a seguir:

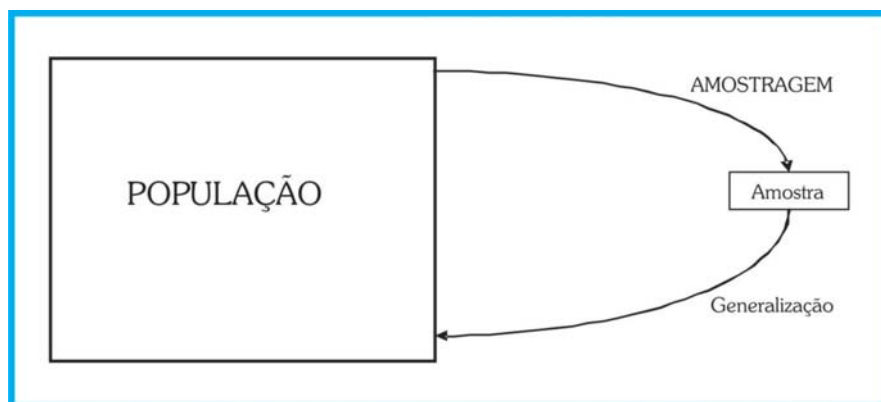


Figura 7: Processo de Amostragem e Generalização

Fonte: Elaborada pelo autor deste livro

É importante sabermos avaliar os argumentos a favor de cada forma de coleta.

Testes destrutivos – são ensaios realizados para avaliar a durabilidade, resistência ou conformidade com as especificações de determinados produtos, que causam a sua inutilização, impedindo a sua comercialização. Muitos testes destrutivos são previstos em legislação específica das mais diversas áreas. Fonte: Elaborado pelo autor deste livro.

Quando Devemos Usar Amostragem

Podemos enumerar basicamente três motivos para usar amostragem em uma pesquisa: economia, rapidez de processamento e quando há a necessidade e testes destrutivos.

- **Economia:** é muito mais barato levantar as características de uma pequena parcela da população do que de todos os seus integrantes, especialmente para grandes populações. O custo do censo demográfico do IBGE é tão colossal que somente pode ser feito a cada dez anos.

- **Rapidez de processamento:** como a quantidade de dados coletada é menor do que a produzida em um censo, especialmente para grandes populações, o seu processamento é mais rápido. Os resultados são conhecidos em pouco tempo e, por isso, as decisões são tomadas com agilidade. Tal característica é especialmente importante em pesquisas de opinião eleitoral, cujo resultado precisa ser conhecido rapidamente, para que candidatos e partidos possam reavaliar suas estratégias.
- **Testes destrutivos:** se para realizar a pesquisa precisamos realizar testes destrutivos (de resistência, tempo de vida útil, entre outros), o censo torna-se impraticável por exigir a utilização de amostragem. Em muitos casos, como no caso de produtos alimentícios e farmacêuticos, há normas legais que precisam ser cumpridas rigorosamente quando da realização dos ensaios.

A Figura 8 sintetiza os motivos:



Figura 8: Economia, Rapidez de processamento e Testes destrutivos

Fonte: Clipart Microsoft® Office (2012)

Após conhecer os motivos de se utilizar a amostragem, pense em algumas situações nas quais seria recomendável utilizar essa técnica.

Quando Não Devemos Usar Amostragem

Existem situações em que a utilização de amostragem pode não ser a melhor opção. Nesse caso, podemos enumerar basicamente quatro

Veremos isso mais adiante, ainda nesta Unidade.

razões: população pequena, característica de fácil mensuração, exigências políticas e necessidade de alta precisão.

- **População pequena:** quando é utilizada uma amostra probabilística (aleatória) com população pequena (menos de 100 elementos), o tamanho mínimo de amostra para obter bons resultados será semelhante ao próprio tamanho da população. Vale a pena, nesse caso, realizar um censo.
- **Característica de fácil mensuração:** a característica pode não precisar de mecanismos sofisticados de mensuração, simplesmente resume-se em uma opinião direta – a favor ou contra uma proposta. Nesse caso, a coleta dos dados seria bastante simples, possibilitando avaliar todos os elementos da população. Um caso frequente na indústria se refere aos sistemas automatizados de medição, por exemplo. Em uma fábrica de cubos de rodas de bicicletas, situada na zona franca de Manaus, os diâmetros de todos os cubos produzidos são medidos automaticamente por um sistema de telemetria a laser, dispensando a coleta por amostragem e um inspetor humano para realizar a medição.
- **Necessidades políticas:** muitas vezes, uma proposta irá afetar dramaticamente todos os elementos da população, como a adoção de um regime ou forma de governo, por exemplo, o que pode ensejar a realização de um censo, para que todos manifestem sua opinião.
- **Necessidade de alta precisão:** por que o IBGE conduz um censo a cada dez anos? Porque as informações demográficas requerem precisão para orientar políticas governamentais. Somente dessa maneira esse objetivo pode ser atingido. A Figura 9 sintetiza os motivos:

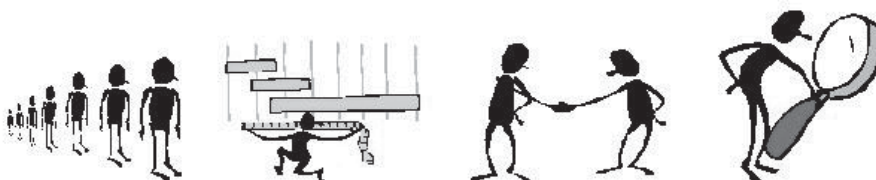


Figura 9: População pequena, fácil mensuração, necessidades políticas e alta precisão

Fonte: Clipart Microsoft® Office (2012)

Exercite a mente! Pense em algumas situações em que seja aconselhável usar um censo. Considere a pesquisa que esboçamos na Unidade 1: "o CRA-SC está interessado em conhecer a opinião dos seus registrados sobre o curso em que se graduaram, desde que tal curso esteja situado em Santa Catarina". Vimos que o número de registrados no CRA, com graduação no Estado, foi suposto igual a 9.000 e há uma listagem com os registrados para fins de cobrança de anuidade que contém informações sobre endereço, curso em que se graduou, entre outras. Para conhecer a opinião das pessoas precisamos entrevistá-las (via correio, Internet, telefone ou pessoalmente). Diante disso, você entende que a pesquisa deve ser conduzida por censo ou por amostragem? Vejamos juntos!

Representatividade ou amostra representativa

– é aquela que traz na sua composição todas as subdivisões da população, procurando retratar da melhor maneira possível a sua variabilidade. Fonte: Elaborado pelo autor deste livro.

Suficiência ou amostra suficiente

– é aquela que tem um tamanho tal que permite representar adequadamente a variabilidade da população. Fonte: Elaborado pelo autor deste livro.

Aspectos Necessários para o Sucesso da Amostragem

Há três aspectos necessários para que uma pesquisa realizada por amostragem gere resultados confiáveis: representatividade, suficiência e aleatoriedade da amostra.

A **representatividade** é o mais óbvio. A amostra precisa retratar a variabilidade existente na população e precisa ser uma “cópia reduzida” desta. Sendo assim, todas as subdivisões da população precisam ter representantes na amostra. A chave é avaliar se as subdivisões da população (por sexo, classe econômica, cidade, atividade profissional) podem influenciar nos resultados da pesquisa. Imagine uma pesquisa eleitoral para governador: deveremos entrevistar eleitores em todas as regiões do Estado (assume-se que haja diferenças de opinião de região para região), pois se escolhermos apenas uma delas, e ela for base política de um candidato, o resultado será distorcido.

A **suficiência** também é um aspecto relativamente óbvio. É necessário que a amostra tenha um tamanho suficiente para representar a variabilidade existente na população. Quanto mais homogê-

Na Unidade 9, veremos o cálculo de tamanho de amostra com detalhes.

Amostra aleatória, casual ou probabilística

– é a amostra retirada por meio de um sorteio não viciado, que garante que cada elemento da população terá uma probabilidade maior do que zero de pertencer à amostra. Fonte: Barbetta, Reis e Bornia (2008).

Algoritmos de geração de números pseudoaleatórios

– são programas computacionais que geram números aleatórios (pseudoaleatórios, pois têm uma regra de formação), com o intento de simular os sorteios manuais de números de 0 a 9. Fonte: Elaborado pelo autor deste livro.

Nesse caso, há sempre o risco de os números se repetirem caso a série seja muito longa, descaracterizando a aleatoriedade.

nea for a população (menor variabilidade), menor poderá ser o tamanho da amostra, e quanto mais heterogênea (maior variabilidade), maior terá que ser o tamanho da amostra para representá-la.

A **aleatoriedade** da amostra é o aspecto menos intuitivo, mas extremamente importante. Significa que os elementos da amostra serão selecionados da população por meio de sorteio não viciado: todos os elementos da população têm chance de pertencer à amostra. É necessária uma listagem com os elementos da população, que permita a atribuição de números a cada um deles, quando é feito o sorteio. Idealmente, escreveríamos os números dos elementos da população em pequenos papéis, depositaríamos em uma urna, misturaríamos e, de olhos vendados, escolheríamos os números, selecionando a amostra. Para grandes populações esse procedimento é inviável e contraproducente considerando-se a disponibilidade de recursos computacionais.

O sorteio pode ser realizado através de tabelas de números aleatórios ou **algoritmos de geração de números pseudoaleatórios**.

As tabelas de números aleatórios são instrumentos usados para auxiliar na seleção de amostras aleatórias. São formadas por sucessivos sorteios de algarismos do conjunto {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, fazendo com que todo número com a mesma quantidade de algarismos tenha a mesma probabilidade de ocorrência. Quando o sorteio é realizado “manualmente”, a tabela é, de fato, chamada de tabela de números aleatórios. Muitos estatísticos realizaram tais sorteios, registraram os resultados e os publicaram em livros e periódicos para uso geral. Se, porém, os números são obtidos mediante simulação computacional, passamos a ter uma tabela de números pseudoaleatórios, pois os números são provenientes da execução de um algoritmo matemático, que tem uma lógica e uma lei de **formação dos resultados**. Não obstante, tal problema somente ocorre, caso o algoritmo seja bom, após milhões ou bilhões de sorteios, quantidade muitíssimo superior àquela usada nas nossas pesquisas. Alguns estatísticos construíram tabelas de números pseudoaleatórios e as deixaram disponíveis para o público em geral.

Nos dias de hoje, com todas as facilidades da informática, é cada vez mais comum bases de dados armazenadas em meio digital, desde uma simples planilha do Microsoft Excel®, ou do Br.Office Calc® até grandes bancos de dados.

Então, perguntamos: por que não realizar também o processo de amostragem, em meio digital, com os algoritmos citados no parágrafo anterior: os **algoritmos de geração de números pseudoaleatórios**?

Trata-se de programas computacionais que procuram simular os sorteios reais de números. A grande vantagem do seu uso é a possibilidade de adaptar facilmente o sorteio ao tamanho da população envolvida e, obviamente, à velocidade de processamento. Veja um exemplo de **números aleatórios** de quatro dígitos (de 0001 a 9000) gerados pelo Br.Office Calc®:

3439	907	5369	8092	7962	8626	131	3667	7769	1248
2206	410	292	1478	1977	155	2566	3088	4983	3217
3347	3201	8193	4195	3836	2736	8781	7260	8921	2307

Na seção *Para saber mais*, vamos disponibilizar um *link* que explica como gerar números pseudoaleatórios com esse aplicativo.

No caso da nossa pesquisa para o CRA-SC, em que há 9.000 registrados graduados em Santa Catarina, e existe uma listagem da população, reflita: como seria o sorteio? No caso mais simples de amostragem aleatória, o registrado de número 3.439 seria sorteado; seguido pelo número 907; este pelo número 5.369; e assim por diante, até completar o tamanho da amostra. Em geral, cria-se automaticamente uma nova base de dados com os elementos sorteados.

Toda a **teoria de inferência estatística** pressupõe que a amostra, a partir da qual será feita a generalização (veja a Figura 10), foi retirada de forma aleatória.

Estudaremos sobre a teoria da inferência estatística nas Unidades 8, 9 e 10.

Agora que já conhecemos os aspectos principais para o sucesso da amostragem, podemos detalhar o plano de amostragem.

Plano de Amostragem

Uma vez tendo decidido realizar a pesquisa selecionando uma amostra da população, é preciso elaborar o **plano de amostragem**, que consiste em definir as unidades amostrais, o modo como a amostra será retirada (o tipo de amostragem) e o próprio tamanho da amostra.

As **unidades amostrais** são as selecionadas para se chegar aos elementos da própria população. Podem ser os próprios elementos da população, quando há acesso direto a eles, ou qualquer outra unidade que possibilite chegar até eles: selecionar os domicílios como unidades de amostragem, para chegar até as famílias (que são os elementos da população); selecionar as turmas como unidades de amostragem, para chegar até os alunos (que são os elementos da população). No caso da pesquisa do CRA-SC, as unidades amostrais são os próprios elementos da população, uma vez que temos a sua listagem. No caso da Pesquisa Nacional por Amostragem de Domicílios do IBGE, as unidades amostrais são os domicílios, através dos quais chegamos às famílias.

O modo como a amostra será retirada é outra decisão importante, que precisa constar do plano de amostragem. Na Figura 10, a seguir, observe o resumo dos diversos tipos de amostragem:

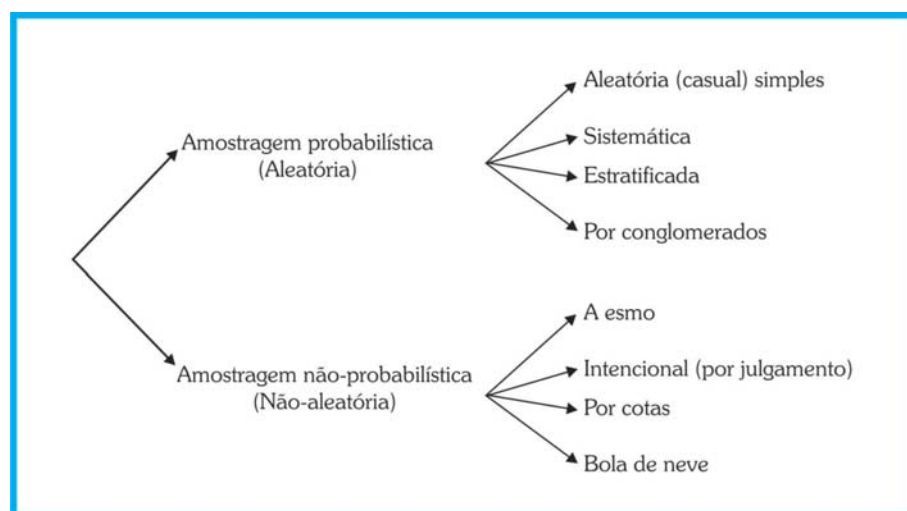
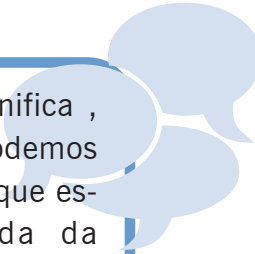


Figura 10: Tipos de amostragem
Fonte: Elaborada pelo autor deste livro

Amostragem Probabilística (aleatória)

Amostragem probabilística, aleatória ou casual, é aquela que garante que cada elemento da população tenha probabilidade de pertencer à amostra. Para que isso ocorra, é necessário que a amostra seja selecionada por sorteio não viciado, ou seja, exige-se aleatoriedade. A sua importância decorre do fato de que apenas os resultados provenientes de uma amostra probabilística podem ser generalizados, estatisticamente, para a população da pesquisa.



Você deve estar se perguntando: o que significa, afinal, estatisticamente? Significa que podemos associar aos resultados a probabilidade de que estejam corretos, ou seja, uma medida da confiabilidade das conclusões obtidas. Se a amostra não for probabilística não há como saber se há 95% ou 0% de probabilidade de que os resultados sejam corretos, e as técnicas de inferência estatística, porventura utilizadas, terão validade questionável.

A condição primordial para o uso da amostragem probabilística é a de que todos os elementos da população tenham uma probabilidade maior do que zero de pertencerem à amostra. Tal condição é materializada se:

- 1) Há acesso a toda a população; ou seja, não há teoricamente problema em selecionar nenhum dos elementos, todos poderiam ser pesquisados. Concretamente, há uma lista da população; como no caso da pesquisa do CRA-SC que dispõe de uma lista com os 9.000 registrados que se graduaram em Santa Catarina.
- 2) Os elementos da amostra são selecionados por meio de alguma forma de sorteio não viciado: tabelas de números aleatórios, números pseudoaleatórios gerados por computador. Com a utilização de sorteio, elimina-se a ingerência do pesquisador na obtenção da amostra e garante-se que todos os integrantes da população podem, por probabilidade, pertencer à amostra.

Agora vamos lhe apresentar os tipos de amostragem probabilística.

Amostragem Aleatória (casual) Simples

A amostragem aleatória simples é o tipo de amostragem probabilística recomendável somente se a população for homogênea em relação aos objetivos da pesquisa. Por exemplo: quando admitimos que todos os elementos da população têm características semelhantes em relação aos objetivos da pesquisa. Há uma listagem dos elementos da população, atribuem-se números a eles e, através de alguma espécie de sorteio não viciado, por meio de tabelas de números aleatórios ou números pseudoaleatórios gerados por computador, os integrantes da amostra são selecionados. Nesse tipo de amostragem probabilística todos os elementos da população têm a mesma probabilidade de pertencer à amostra. Foi exatamente o que fizemos no final da seção *Aspectos necessários para o sucesso da amostragem* para a nossa pesquisa do CRA-SC.

Amostragem aleatória simples – é o processo de amostragem em que todos os elementos da população têm a mesma probabilidade de pertencer à amostra, e cada elemento é sorteado. Fonte: Barbetta, Reis e Bornia (2008).

Tabelas de números aleatórios – são instrumentos usados para auxiliar na seleção de amostras aleatórias, formadas por sucessivos sorteios de algarismos do conjunto {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, fazendo com que todo número com a mesma quantidade de algarismos tenha a mesma probabilidade de ocorrência. Fonte: Barbetta (2007).

Amostragem sistemática – é a variação da amostragem aleatória simples em que os elementos da população são retirados a intervalos regulares, até compor o total da amostra, sendo o sorteio realizado apenas no ponto de partida. Fonte: Barbetta (2007).

Amostragem Sistemática

Quando a lista de respondentes for muito grande, a utilização de amostragem aleatória simples pode ser um processo moroso; ou se o tamanho de amostra for substancial, teremos que realizar um grande número de sorteios: caso estejamos utilizando números pseudoaleatórios aumenta o risco de repetição dos números. Utiliza-se então uma variação, a amostragem sistemática, que também supõe que a população é homogênea em relação à variável de interesse, mas que consiste em retirar elementos da população a intervalos regulares, até compor o total da amostra. A amostragem sistemática somente pode ser retirada se a ordenação da lista não tiver relação com a variável de interesse. Imagine que queiramos obter uma amostra de idades de uma listagem justamente ordenada dessa forma, nesse caso a amostragem sistemática não seria apropriada, a não ser que reordenássemos a lista.

Veja a seguir o procedimento para a amostragem sistemática:

- obtém-se o tamanho da população (N);
- calcula-se o tamanho da amostra (n), (veremos isso mais adiante); e

- encontra-se o intervalo de retirada $k = N/n$:
 - se k for fracionário, deve-se aumentar (n) até tornar o resultado inteiro;
 - se (N) for um número primo, excluem-se por sorteio alguns elementos da população para tornar k inteiro;
- sorteia-se o ponto de partida (um dos k , números do primeiro intervalo) usando-se uma tabela de números aleatórios, ou qualquer outro dispositivo (isso precisa ser feito para garantir que todos os elementos da população tenham chance de pertencer à amostra); e
- a cada k , elementos da população, retira-se um para fazer parte da amostra, até completar o valor de (n) .

O resumo desse processo é retratado na Figura 11, veja:

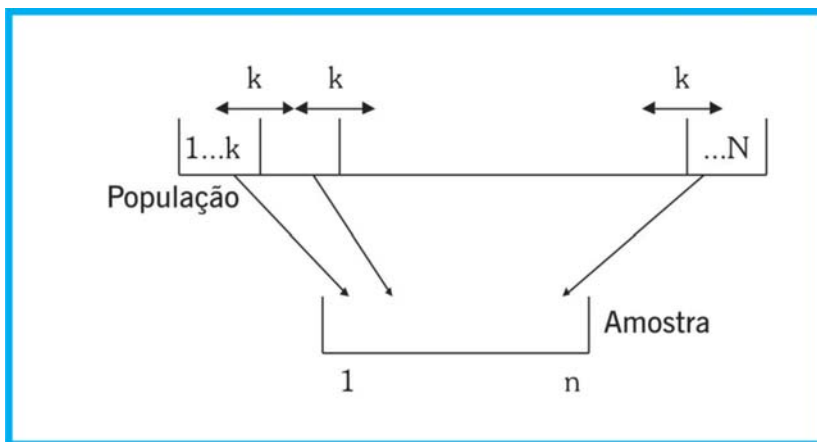


Figura 11: Processo de amostragem sistemática

Fonte: Elaborada pelo autor deste livro

O Exemplo, a seguir, ajudará você a entender melhor sobre o processo de amostragem sistemática. Leia com atenção!

Exemplo 1 – Uma operadora telefônica pretende saber a opinião de seus assinantes comerciais sobre seus serviços na cidade de Florianópolis. Supondo que há 25.037 assinantes comerciais e a amostra precisa ter no mínimo 800 elementos, mostre como seria organizada uma amostragem sistemática para selecionar os respondentes.

Amostra representativa

– aquela que representa, na sua composição, todas as subdivisões da população, procurando retratar da melhor maneira possível a sua variabilidade. Fonte: Elaborado pelo autor deste livro.

Amostragem estratificada

– é a amostragem probabilística usada quando a população for heterogênea em relação aos objetivos da pesquisa (as opiniões tendem a variar muito de subgrupo para subgrupo), e a amostra precisa conter elementos de cada subgrupo da população para representá-la adequadamente. Fonte: Barbetta (2007).

A operadora dispõe de uma lista ordenada alfabeticamente com todos os seus assinantes, o intervalo de retirada será: $k = N/n = 25037/800 = 31,2965$.

Como o valor de k é fracionário, algo precisa ser feito. Aumentar o tamanho da amostra não resolverá o problema, porque 25.037 é um número primo. Como não podemos reduzir o tamanho de amostra, devendo permanecer igual a 800, se excluirmos por sorteio 237 elementos da população e refizermos a lista, teremos: $k = N/n = 24800/800 = 31$.

A cada 31 assinantes um é retirado para fazer parte da amostra. Devemos sortear o ponto de partida: um número de 1 a 31 (do 1º ao 31º assinante). Imagine que o sorteio resultasse em 5, então, a amostra seria (número de assinantes): {5, 36, 67, 98,..., 24.774}.

Amostragem Estratificada

É bastante comum a população de uma pesquisa ser heterogênea em relação aos objetivos dela. No caso de uma investigação eleitoral para governador, por exemplo, podemos esperar que a opinião seja diferente dependendo da região onde o eleitor mora, sua classe social e, inclusive, sua profissão. Contudo, podemos supor que haja certa homogeneidade de opinião dentro de cada grupo e, apesar disso, presumir que haja heterogeneidade entre os estratos, mas homogeneidade em cada estrato; e que eles sejam mutuamente exclusivos (cada elemento da população pode pertencer a apenas um estrato). Para garantir que a amostra seja representativa da população, precisamos garantir que os diferentes estratos sejam nela representados: deve usar a amostragem estratificada, como representa a Figura 12:

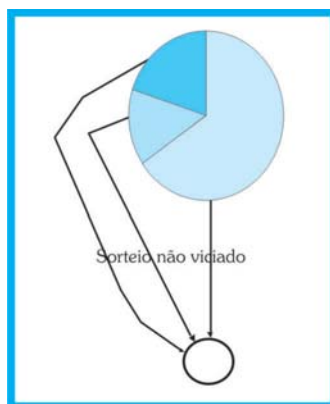


Figura 12: Amostragem estratificada
Fonte: Elaborada pelo autor deste livro

Veja que a seleção dos elementos de cada estrato pode ser feita usando amostragem aleatória simples ou sistemática.

A amostragem estratificada pode ser:

- proporcional, quando o número de elementos selecionados de cada estrato é proporcional ao seu tamanho na população (por exemplo, se o estrato representa 15% da população, 15% da amostra deverá ser retirada dele); e
- uniforme, quando os mesmos números de elementos são selecionados de cada estrato.

A amostragem estratificada proporcional possibilita resultados melhores, mas exige conhecimento da população (para saber quantos são e quais são os tamanhos dos estratos). A amostragem estratificada uniforme é mais utilizada em estudos comparativos.

No caso da pesquisa do CRA-SC, você acredita que a população é heterogênea em relação aos objetivos da pesquisa? Aspectos como: a região do Estado onde estudaram ou se oriundos de faculdade pública ou de particular, podem influenciar na opinião dos registrados sobre os cursos em que se graduaram?

Amostragem por Conglomerados

Teoricamente, a amostragem estratificada proporcional apresenta os melhores resultados possíveis. Sua grande dificuldade de uso deve-se ao grau de conhecimento necessário sobre a população, que geralmente não existe ou é impraticável de se obter. Uma alternativa consiste no uso de conglomerados.

Os conglomerados também são grupos mutuamente exclusivos de elementos da população, mas são definidos de forma mais arbitrária do que os estratos: é bastante comum definir os conglomerados

Amostragem por conglomerados – é a amostragem probabilística em que a população é subdividida em grupos definidos por conveniência (usualmente geográfica), e alguns desses grupos são selecionados por sorteio, e elementos dos grupos sorteados podem também ser sorteados para compor a amostra. Fonte: Barbetta (2007).

geograficamente. Por exemplo, os bairros de uma cidade, que constituiriam conglomerados de domicílios.

O procedimento para a amostragem por conglomerados ocorre da seguinte forma:

- divide-se a população em conglomerados;
- sorteiam-se os conglomerados (usando tabela de números aleatórios ou qualquer outro método não viciado); e
- pesquisam-se todos os elementos dos conglomerados sorteados, ou sorteiam-se elementos deles.

A utilização de amostragem por conglomerados permite uma redução substancial nos custos de obtenção da amostra, sem comprometer demasiadamente a precisão, sendo que em alguns casos é a única alternativa possível. Veja a Figura 13 e entenda como ocorre essa amostragem:

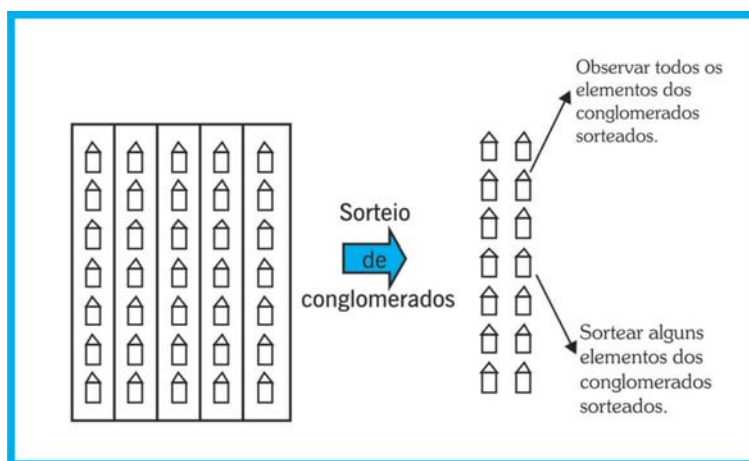


Figura 13: Amostragem por conglomerados

Fonte: Elaborada pelo autor deste livro

Mais informações em
< h t t p : / /

www.ibge.gov.br/home/
estatistica/populacao/
trabalhoerendimento/
pnad98/saude/
metodologia.shtm>.

Acesso em: 14 fev.
2012.

A **Pesquisa Nacional por Amostra de Domicílios (PNAD)** do IBGE, coleta informações demográficas e socioeconômicas sobre a população brasileira. Utiliza amostragem por conglomerados em três estágios:

- **Primeiro estágio:** amostras de municípios (conglomerados) para cada uma das regiões geográficas do Brasil;
- **Segundo estágio:** setores censitários sorteados em cada município (conglomerado sorteado); e

- **Terceiro estágio:** domicílios sorteados em cada setor censitário.

Você deve estar se perguntando: e se não for possível garantir a probabilidade de todo elemento da população pertencer à amostra? Então, esse é o momento de partirmos para a amostragem não probabilística.

Amostragem Não Probabilística

A obtenção de uma amostra probabilística exige uma listagem com os elementos da população. Em suma, exige acesso a todos os elementos da população. Nem sempre é possível obter tal listagem na prática, o que teoricamente inviabilizaria a retirada de uma amostra probabilística. Desse modo, pode-se recorrer à amostragem não probabilística.

Ao usar a amostragem não probabilística, o pesquisador não sabe qual a probabilidade que tem um elemento da população de pertencer à amostra. Portanto, os resultados da amostra não podem ser estatisticamente generalizados para a população, porque não se pode estimar o erro amostral.

Alguns dos usos habituais da amostragem não probabilística são os seguintes:

- a etapa preliminar em projetos de pesquisa;
- em projetos de pesquisa qualitativa; e
- em casos em que a população de trabalho não pode ser enumerada.

Veja que existem ainda vários tipos de amostragem não probabilística, os quais serão descritos na sequência.

Amostragem não probabilística – é o processo de amostragem em que nem todos os elementos da população têm chance de pertencer à amostra, pois a seleção não é feita por sorteio não viciado. Fonte: Barbetta (2007).

Erro amostral – é o valor máximo que o pesquisador admite errar na estimativa de uma característica da população a partir de uma amostra aleatória dessa mesma população. Fonte: Barbetta (2007).

Amostragem a Esmo

Na amostragem a esmo, o pesquisador procura ser o mais aleatório possível, mas sem fazer um sorteio formal. Imagine um lote de 10.000 parafusos, do qual queremos tirar uma amostra de 100. Se fôssemos realizar uma amostragem aleatória simples, o processo talvez fosse trabalhoso demais. Então simplesmente retiramos os elementos a esmo. Esse tipo de amostragem também pode ser utilizado quando a população for formada por material contínuo (gases, líquidos, minérios), bastando homogeneizar o material e retirar a amostra.

Amostragem por Julgamento (intencional)

Na amostragem por julgamento, o pesquisador, deliberadamente, escolhe alguns elementos para fazer parte da amostra com base no julgamento de que estes seriam representativos da população. Esse tipo de amostragem é bastante usado em estudos qualitativos. Obviamente, o risco de se obter uma amostra viciada é grande, pois se baseia totalmente nas preferências do pesquisador, que pode se enganar (involuntária ou “voluntariamente”).

Amostragem por Cotas

A Amostragem por cotas é semelhante a uma amostragem estratificada proporcional, da qual se diferencia por não empregar sorteio na seleção dos elementos. A população é dividida em vários subgrupos, na realidade é comum dividi-la em um grande número deles para compensar a falta de aleatoriedade, e seleciona-se uma cota de cada subgrupo, proporcional ao seu tamanho.

Em pesquisa de opinião eleitoral, por exemplo, poderíamos dividir a população de eleitores por sexo, nível de instrução, faixas de renda entre outros aspectos, e obter cotas proporcionais ao tamanho dos grupos (que poderia ser obtido através das informações do IBGE). Na amostragem por cotas, os elementos da amostra são escolhidos pelos entrevistadores (de acordo com os critérios), geralmente em pontos de grande movimento, o que sempre acarreta certa subjetividade (e impede a seleção de quem não esteja passando pelo local no exato momento da pesquisa).

Na prática, muitas pesquisas são realizadas utilizando-se amostragem por cotas, incluindo as polêmicas **pesquisas eleitorais**.

No exemplo apresentado no Quadro 1 a seguir, imagine que queiramos saber a opinião dos eleitores do bairro Goiaba sobre o governo municipal. Supõe-se que as principais variáveis que condicionariam as respostas seriam sexo, idade e classe social. O bairro apresenta a seguinte composição demográfica para as variáveis:

Leia um texto muito interessante sobre o tema em: <<http://www.ime.unicamp.br/~dias/falaciaPesquisaEleitoral.pdf>>. Acesso em: 14 fev. 2012.

SEXO	IDADE (FAIXA ETÁRIA)	CLASSE SOCIAL	% POPULACIONAL
Masculino	18 -- 35	A	1%
Masculino	18 -- 35	B	4%
Masculino	18 -- 35	C	10%
Feminino	18 -- 35	A	1%
Feminino	18 -- 35	B	2%
Feminino	18 -- 35	C	9%
Masculino	35 -- 60	A	5%
Masculino	35 -- 60	B	8%
Masculino	35 -- 60	C	12%
Feminino	35 -- 60	A	4%
Feminino	35 -- 60	B	8%
Feminino	35 -- 60	C	10%
Masculino	Mais de 60	A	1%
Masculino	Mais de 60	B	9%
Masculino	Mais de 60	C	3%
Feminino	Mais de 60	A	3%
Feminino	Mais de 60	B	7%
Feminino	Mais de 60	C	3%

Quadro 1: Esquema de amostragem por cotas
Fonte: Adaptado de Marconi e Lakatos (2003)

Se, por exemplo, o tamanho de nossa amostra fosse igual a 200 (200 pessoas serão entrevistadas), o número de pessoas deveria ser dividido de forma proporcional: 1% do sexo masculino, com idade entre 18 e 25 anos, da classe A, totalizando duas pessoas; 4% do sexo masculino, com idade entre 18 e 25 anos, da classe B, totalizando oito pessoas; e assim por diante. Os entrevistadores receberiam suas cotas e deveriam escolher pessoas, conforme critérios estabelecidos e em pontos de movimento do referido bairro, e entrevistá-las a fim de colher suas opiniões sobre o governo municipal. Usualmente, os resul-

tados são generalizados estatisticamente para a população, com o emprego das técnicas que serão vistas na Unidade 9 deste livro-texto; mas, rigorosamente, os resultados da amostragem por cotas **não têm validade estatística**, visto que não contemplam o princípio de aleatoriedade na seleção da amostra.

Amostragem “Bola de Neve”

A Amostragem “bola de neve” é particularmente importante quando é difícil identificar respondentes em potencial. A cada novo respondente, identificado e entrevistado, pede-se que indique outros que possam ser qualificados como respondentes. Por conta disso, poderá acontecer de as amostras serem compostas apenas por “amigos” dos primeiros entrevista-dos, o que pode causar viesamentos nos resultados finais.

Agora que você já conhece sobre o importante e interessante tema do cálculo do tamanho de amostra, passaremos para uma amostra probabilística.

Viesamentos – distorções na obtenção da amostra que podem comprometer a confiabilidade dos dados.
Fonte: Elaborado pelo autor deste livro.

Cálculo do Tamanho de uma Amostra Probabilística (aleatória)

A determinação do tamanho de amostra é um dos aspectos mais controversos da técnica de amostragem e envolve uma série de conceitos (probabilidade, inferência estatística e a própria teoria da amostragem). Nesta seção, apresentaremos uma visão simplificada para obtenção do tamanho mínimo de uma amostra aleatória simples que atenda aos seguintes requisitos:

- o interesse na proporção de ocorrência de um dos valores de uma variável qualitativa na população;
- a confiabilidade dos resultados da amostra deve ser aproximadamente igual a 95% (ou seja, há 95% de probabilidade de que a proporção populacional do valor da variável

qualitativa esteja no intervalo definido pelos resultados da amostra);

- a consideração quanto à estimativa, exagerada ou não, do tamanho de amostra; e
- a despreocupação com os aspectos financeiros relacionados ao tamanho da amostra (embora, obviamente, seja uma consideração importante).

O primeiro passo para calcular o tamanho da amostra é definir o **erro amostral** tolerável, que será chamado de E_0 . Esse erro corresponde ao valor máximo que o pesquisador admite errar na estimativa de uma característica da população.

Lembre-se das pesquisas de opinião eleitoral: “o candidato Fulano está com 18% de intenção de voto, a margem de erro da pesquisa é de 2% para mais ou para menos”. O 2% é o valor do erro amostral tolerável, então o percentual de pessoas declarando o voto no candidato Fulano é igual a $18\% \pm 2\%$. Além disso, existe a probabilidade de que esse intervalo não contenha o valor real do parâmetro, ou seja, o percentual de eleitores que declararam seu voto ao candidato; isso porque se fez uso de uma amostra, o que raramente é revelado pela mídia, especialmente a televisiva.

É razoável imaginar que, quanto menor o erro amostral tolerável escolhido, maior será o tamanho da amostra necessário para obtê-lo. Isso fica mais claro ao ver a fórmula para obtenção da primeira estimativa do tamanho de amostra: $n_0 = \frac{1}{E_0^2}$

Onde E_0 é o erro amostral tolerável, e n_0 é a primeira estimativa do tamanho de amostra. Se o tamanho da população (N) for conhecido, podemos corrigir a primeira estimativa: $n = \frac{N \times n_0}{N + n_0}$

Com esse exemplo, pense em como obter o tamanho mínimo de uma amostra aleatória simples admitindo, com alto grau de confiança, um erro amostral máximo de 2%, supondo que a população tenha:

- a) 200 elementos.
- b) 200.000 elementos.

Observe a diferença entre os tamanhos das duas populações: a da letra b é mil vezes maior do que a da letra a. Como na primeira estimativa n_0 não depende do tamanho da população e o erro amostral é de 2% para ambas, podemos calculá-lo apenas uma vez. Para isso, dividiremos 2% por 100 antes de substituí-lo na fórmula:

$$n_0 = \frac{1}{E_0^2} = \frac{1}{(0,02)^2} = 2.500$$

Então nossa primeira estimativa, para um erro amostral de 2%, é retirar uma amostra de 2.500 elementos.

- Obviamente precisamos corrigir a primeira estimativa, pois a população conta com apenas 200 elementos. Então:

$$n = \frac{N \times n_0}{N + n_0} = \frac{200 \times 2.500}{200 + 2.500} = 185,185$$

Precisamos arredondar, sempre para cima, o tamanho mínimo da amostra. Desse modo, a amostra deverá ter pelo menos 186 elementos para garantir um erro amostral de 2%. Observe que a amostra representa 93% da população. Nesse caso, um censo não seria mais aconselhável?

- Corrigindo a primeira estimativa com o tamanho da população:

$$n = \frac{N \times n_0}{N + n_0} = \frac{200.000 \times 2.500}{200.000 + 2.500} = 2.469,136$$

Arredondando, a amostra deverá ter no mínimo 2.470 elementos para garantir um erro amostral de 2%. Observe que a amostra representa 1,235% da população. Claríssimo caso em que a amostragem é a melhor opção de coleta.

Poderíamos ter usado diretamente a primeira estimativa: 2.500 elementos, pois a correção não causou mudança significativa. Esse exemplo prova que não precisamos de grandes amostras para obter precisão nos resultados.

A Figura 14 exibe um gráfico que relaciona os tamanhos de amostra a diferentes tamanhos de população, considerando-se um erro amostral tolerável igual a 2%.

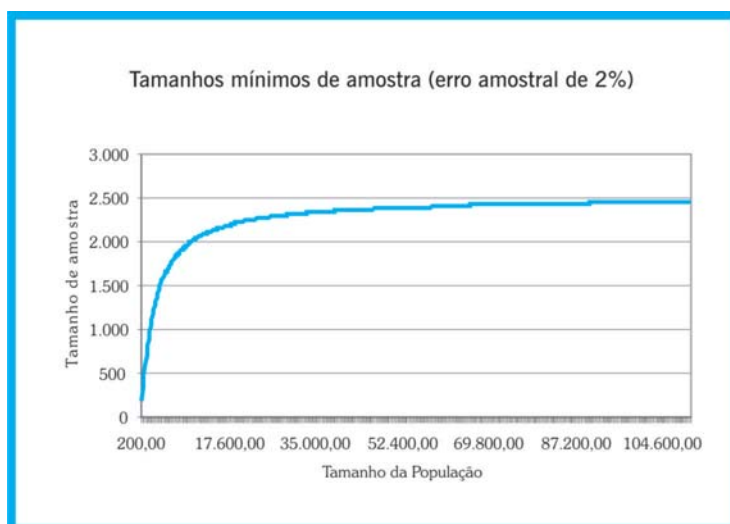


Figura 14: Tamanho de amostra x tamanho da população ($E_0 = 2\%$)

Fonte: Elaborada pelo autor deste livro

Observe que a partir de determinado tamanho de população, para o mesmo erro amostral, o ritmo de crescimento do tamanho da amostra diminui para 70.000 elementos ou mais, praticamente não há mais aumento. Isso mostra que não há necessidade de retirar, por exemplo, 50% da população para que se obtenha uma boa amostra.

É importante alertar que, ao calcular o tamanho de amostra para amostragem estratificada, devemos fazê-lo para cada estrato e o tamanho total será a soma dos valores. Se isso não for feito, não podemos garantir o erro amostral dentro de cada estrato: se calcularmos um valor geral e dividirmos o tamanho da amostra por estrato (mesmo proporcionalmente), a margem de erro dentro de cada estrato será maior do que a prevista.

Saiba mais.....

Sobre amostragem, consulte: BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Ed. da UFSC, 2008, Capítulo 3.

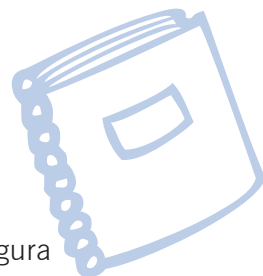
Sobre características de fácil mensuração consulte em LAGO NETO, João C. *O Efeito da Autocorrelação em Gráficos de Controle para Variável Contínua: um estudo de caso*. Florianópolis. 1999. Dissertação (Mestrado em Engenharia de Produção). Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina. 1999.

Sobre pesquisas eleitorais, consulte: SOUZA, Jorge. *Pesquisas Eleitorais: críticas e técnicas*. Brasília (DF): Centro Gráfico do Senado Federal, 1990.

Sobre como gerar números pseudoaleatórios ou obter amostras aleatórias simples no Br.Office Calc®, leia o texto: *Como gerar uma amostra aleatória simples com o Br.Office Calc®*, no Ambiente Virtual de Ensino-Aprendizagem (AVEA).

Sobre Amostragem a esmo, leia: COSTA NETO, Pedro L. da O. *Estatística*. 2. ed. São Paulo: Edgard Blücher, 2002.

Resumindo



O resumo desta Unidade está esquematizado na Figura 15. Observe:

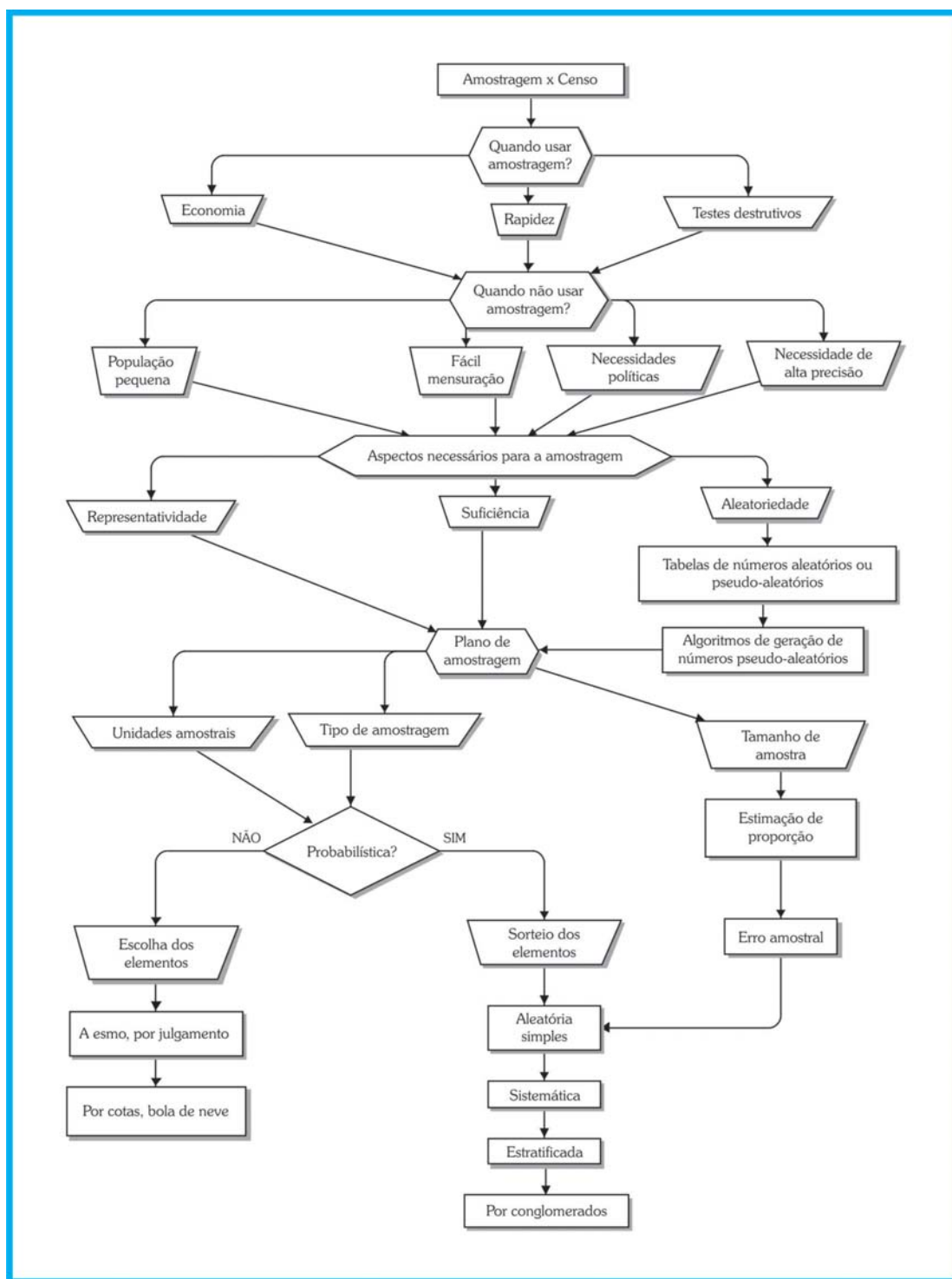


Figura 15: Resumo da Unidade 2
Fonte: Elaborada pelo autor deste livro

Caro estudante,

Chegamos ao final da Unidade 2 e por meio dela aprendemos sobre os temas amostragem e censo, conhecimentos necessários para o bom administrador. Essa Unidade foi repleta de Figuras, de Quadros, de representações e de exemplos relativos ao emprego das técnicas e das diferentes formas de utilizá-las, conforme suas especificidades. Todo esse estudo servirá de sustentação para as discussões propostas nas próximas Unidades.

Releia, caso necessário, todos os exemplos, leia as indicações do *Saiba mais* e discuta com seus colegas. Na realização da atividade de aprendizagem você colocará em prática os ensinamentos repassados. Conte sempre com o acompanhamento da tutoria e das explicações do professor. Lembre-se que você não está sozinho. Conte com a gente!



Atividades de aprendizagem

Hora de testar seus conhecimentos com relação ao estudo da Unidade 2. Para isso, faça as atividades propostas a seguir e encaminhe-as para seu tutor por meio do AVEA. Não hesite em buscar o auxílio do seu tutor se encontrar dificuldades.

1. Analise os planos de amostragens apresentados a seguir e responda: Você concorda com a maneira como foram elaborados? Justifique. Apresente as soluções que julgar necessárias.
 - a) Para ser conhecida a opinião dos estudantes da UFSC sobre o Jornal Universitário, foram colhidas as opiniões de

40 estudantes da última fase do curso de Jornalismo dessa instituição.

- b) Há interesse em medir o índice de luminosidade das salas de aula da UFSC. A coleta de dados será feita em todos os centros dessa Universidade durante os períodos diurno e noturno nas salas que estiveram desocupadas no momento da pesquisa. Cada centro será visitado apenas uma vez.
- c) As constantes reclamações dos usuários motivaram a direção da Biblioteca Central da UFSC a realizar uma pesquisa sobre o nível de ruído em suas dependências. O ruído será medido em todas as seções da Biblioteca, na primeira e na penúltima semanas do semestre, de segunda a sábado, durante todo o horário de funcionamento.
- d) No controle de qualidade de uma fábrica de peças, que trabalha 24 horas por dia, sete dias por semana, um item produzido é retirado de cada máquina a cada meia hora para avaliação. O procedimento é feito durante todo o dia, ao longo da semana.
- e) O Comando de um Batalhão da Polícia Militar de Santa Catarina quer conhecer a opinião das pessoas que residem na área em que ele atua com intuito de formular novas escalas de policiamento ostensivo. Para tanto, serão feitas entrevistas com os transeuntes que passarem pela frente do Batalhão de segunda a sexta das 8h30 às 12 h e das 14 h às 17h30 durante duas semanas.
- f) Com a finalidade de estudar o perfil dos consumidores de um supermercado, observou-se os consumidores que compareciam ao supermercado no primeiro sábado do mês.
- g) Com a finalidade de estudar o perfil dos consumidores de um supermercado, fez-se a coleta de dados durante um mês. Para esse fim, entrevistou-se, a cada dia, um consumidor em cada fila e em caixas diferentes, variando-se sistematicamente o horário de coleta dos dados.
- h) Para avaliar a qualidade dos itens que saem de uma linha de produção, observou-se todos os itens das 14 h às 14h30.

- i) Para avaliar a qualidade dos itens que saem de uma linha de produção, observou-se um item a cada meia hora, durante todo o dia.
- j) Para estimar a porcentagem de empresas que investiram em novas tecnologias no último ano, enviou-se um questionário a todas as empresas de um Estado. A amostra foi formada pelas empresas que responderam ao questionário.

2. Determinada faculdade do interior de Santa Catarina possui seis cursos. Seus alunos estão matriculados de acordo com a tabela a seguir:

CURSO	DIREITO	ADMINISTRAÇÃO	ECONOMIA	AGRONOMIA	VETERINÁRIA	COMPUTAÇÃO
Alunos	250	200	150	150	150	100

A diretoria pretende selecionar, por amostragem, alguns alunos para uma atividade extracurricular.

- a) Os cursos de Direito, de Administração e de Economia formam um estrato (socioeconômico); os de Agronomia e de Veterinária formam outro (agrários); e os de Computação outro estrato (tecnológicos). Extraia uma amostra estratificada proporcional de 20 alunos (use o Microsoft Excel® ou o Br.Office Calc®).
 - b) Através de uma amostragem de conglomerados de dois estágios, extraia uma amostra aleatória de 21 alunos. Selecione três cursos e, depois, sete alunos por curso (use o Microsoft Excel® ou o Br.Office Calc®).
 - c) Qual das duas amostras você acredita que tem resultados mais confiáveis? Justifique.
3. Será feito um levantamento por amostragem de uma população de 2.000 famílias para a realização de uma pesquisa.
- a) Calcule o tamanho mínimo de uma amostra para que se tenha um erro amostral máximo de 5%.
 - b) Supondo a população dividida em dois estratos iguais, qual o tamanho mínimo de amostra para se obter um erro amostral máximo de 5% em cada estrato?

- c) Qual seria o erro amostral em cada estrato do item b se o tamanho da amostra em cada estrato fosse simplesmente o valor definido no item a dividido por 2?

3

UNIDADE

Análise Exploratória de Dados



Ao finalizar esta Unidade, você deverá ser capaz de compreender a definição de Análise Exploratória de Dados; de entender como se realiza a descrição tabular e gráfica de conjuntos de dados pertinentes a variáveis qualitativas e quantitativas.

O Conceito de Análise Exploratória de Dados

Caro estudante,

Nas Unidades anteriores, estudamos o planejamento de uma pesquisa e as principais técnicas de amostragem. Conforme vimos, independente de os dados terem sido coletados via censo ou amostragem, eles precisam ser interpretados para atingir os objetivos propostos na pesquisa. O passo inicial para isso é usar os conceitos e as técnicas de Análise Exploratória de Dados para resumir e organizar os dados de maneira que seja possível identificar padrões e elaborar as primeiras conclusões a respeito da população, isto é, descrever a sua variabilidade.

Para entender esse contexto, é importante relembrar a definição de variável e a sua classificação por nível de mensuração e nível de manipulação, estudadas na Unidade 1.

Nesta Unidade, vamos estudar como realizar a Análise Exploratória de Dados em cinco casos por meio de tabelas e gráficos, tipos de conjuntos de dados: uma variável qualitativa, uma variável quantitativa, duas variáveis qualitativas, uma qualitativa e uma quantitativa e duas quantitativas.

É indispensável que o administrador seja capaz de realizar a Análise Exploratória de Dados, pois, sem isso, sua capacidade de tomada de decisões ficará seriamente comprometida.

No passado a Análise Exploratória de Dados era chamada de Estatística

Descritiva, por preocupar-se com a descrição dos dados tão somente.

Variáveis estatísticas – são características que podem ser observadas ou medidas em cada elemento pesquisado, em um dado momento, há um e apenas um resultado possível. Fonte: Barbetta (2007).

A **Análise Exploratória de Dados**, antigamente chamada apenas de Estatística Descritiva, diz respeito ao que a maioria das pessoas entende como Estatística e, inconscientemente, usa no dia a dia. Consiste em **resumir** e em **organizar** os dados coletados através de tabelas, de gráficos ou de medidas numéricas e, a partir dos dados resumidos, procurar alguma regularidade ou padrão nas observações (**interpretar** os dados).

A partir dessa interpretação inicial, é possível identificar se os dados seguem alguns modelos conhecidos, que permitam estudar o fenômeno sob análise, ou se é necessário sugerir um novo modelo. Usualmente, a concretização dos objetivos de uma pesquisa passa pela análise de uma ou de mais variáveis estatísticas, ou da sua relação.

Assim, o processo da análise exploratória de dados baseia-se em organizar, resumir e interpretar as medidas das variáveis da melhor maneira possível e, para isso, é necessário construir um arquivo de dados, que tem algumas características especiais.

Estrutura de um Arquivo de Dados

Uma vez disponíveis, os dados precisam ser tabulados a fim de possibilitar sua análise. Atualmente, os dados costumam ser armazenados em meio computacional, seja em grandes bases de dados, programas estatísticos ou mesmo planilhas eletrônicas, oriundos de pesquisa de campo, de registros de operações financeiras, ou de arquivos de recursos humanos, entre outros. Possuem uma estrutura fixa, que possibilita a aplicação de várias técnicas para extrair as informações de interesse.

As variáveis são registradas nas colunas, e os casos (os elementos da população) nas linhas. As variáveis são as características pesquisadas ou registradas. Imagine a base de dados do Departamento de Administração Escolar (DAE) da Universidade Federal de Santa Catarina (UFSC), que armazena as informações dos acadêmicos, contendo as variáveis: o nome do aluno, a data de nascimento, o número de matrícula, o Índice de Aproveitamento Acumulado (IAA), o Índice de Aproveitamento por Período (IAP) e outras informações; ou uma operadora de cartão de crédito, que armazena as transações efetuadas,

contendo o número do cartão, nome do titular, data e hora da transação, valor do crédito, do bem ou do serviço adquirido, estabelecimento.

Os casos constituem cada indivíduo ou registro. Para a base do DAE, por exemplo: João Ninguém; nascido em 20 de fevereiro de 1985; matrícula 02xxxxxxx-01; IAA = 3,5; IAP = 6,0. Para a operadora de cartão de crédito: cartão número xxxxxxxxxx-84; José Nenhum; 11 de dezembro de 2003; 14h28min; R\$ 200,00; supermercado.

Exemplo 1 – A Megamontadora **TOYORD** conduz, com regularidade, pesquisas de mercado com os clientes que compraram carros zero km diretamente de suas concessionárias. O objetivo é avaliar a satisfação deles em relação aos diferentes modelos, ao *design* e, consequentemente, perceber a adequação, ou não, desses itens ao perfil do cliente. A última pesquisa foi terminada em julho de 2007, quando foram entrevistados 250 clientes do total de 30.000 que compraram veículos novos entre maio de 2006 e maio de 2007. A pesquisa foi restringida aos modelos mais vendidos e aos que estão no mercado há dez anos.

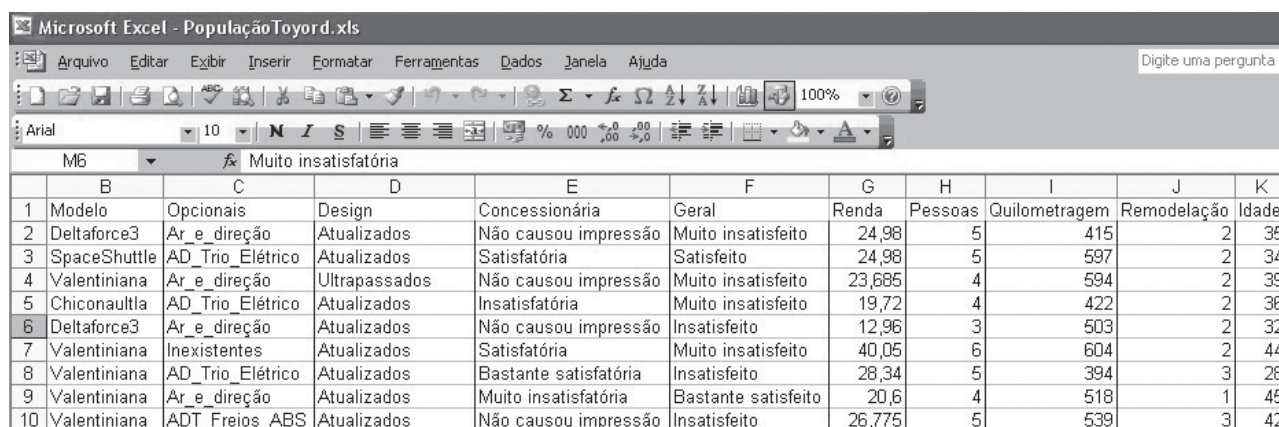
Trata-se de uma empresa e de uma pesquisa fictícias.

- **Modelo comprado:** o compacto Chiconaultla, o sedã médio DeltaForce3, a perua familiar Valentiniana, a van SpaceShuttle ou o luxuoso LuxuriousCar.
- **Opcionais:** inexistentes (apenas os itens de série); ar-condicionado e direção hidráulica; ar-condicionado, direção hidráulica e trio elétrico; ar-condicionado, direção hidráulica, trio elétrico e freios ABS.
- **Opinião sobre o *design*:** se os clientes consideram o design do veículo comprado ultrapassado, atualizado, ou adiante dos concorrentes.
- **Opinião sobre a concessionária onde comprou o veículo (incluindo atendimento na venda, manutenção programada e eventuais problemas):** muito insatisfatória, insatisfatória, não causou impressão, satisfatória, bastante satisfatória.
- **Opinião geral sobre o veículo adquirido:** muito insatisfeito, insatisfeito, satisfeito, bastante satisfeito.
- **Renda declarada pelo cliente:** em salários mínimos mensais.
- **Número de pessoas** geralmente transportadas no veículo.

- **Quilometragem** mensal média percorrida com o veículo.
- **Percepção do cliente** relativa aos anos que o veículo comprado sofreu sua última remodelação de *design*: em anos completos (se há menos de um ano o entrevistador anotou zero).
- **Idade do cliente** em anos completos.

Imagine que você é *trainee* da Toyord e tem por missão analisar os resultados da pesquisa em relatório e, dependendo do seu desempenho, poderá ser contratado ou dispensado (sem carta de recomendação). A questão é: como deve ser estruturada a base de dados para permitir a análise?

Digamos que você dispõe dos 250 questionários que foram aplicados e irá tabulá-los em uma planilha eletrônica. Há dez variáveis; logo, a base de dados deve conter então dez colunas e 250 linhas (no Calc® são 251, já que a primeira é usada para pôr o nome das variáveis). Veja o resultado com as primeiras linhas (casos) na Figura 16:



	B	C	D	E	F	G	H	I	J	K
1	Modelo	Opcionais	Design	Concessionária	Geral	Renda	Pessoas	Quilometragem	Remodelação	Idade
2	Deltaforce3	Ar_e_direção	Atualizados	Não causou impressão	Muito insatisfeito	24,98	5	415	2	35
3	SpaceShuttle	AD_Trio_Elétrico	Atualizados	Satisfatória	Satisfeito	24,98	5	597	2	34
4	Valentiniana	Ar_e_direção	Ultrapassados	Não causou impressão	Muito insatisfeito	23,685	4	594	2	39
5	Chiconaultla	AD_Trio_Elétrico	Atualizados	Insatisfatória	Muito insatisfeito	19,72	4	422	2	36
6	Deltaforce3	Ar_e_direção	Atualizados	Não causou impressão	Insatisfeito	12,96	3	503	2	32
7	Valentiniana	Inexistentes	Atualizados	Satisfatória	Muito insatisfeito	40,05	6	604	2	44
8	Valentiniana	AD_Trio_Elétrico	Atualizados	Bastante satisfatória	Insatisfeito	28,34	5	394	3	28
9	Valentiniana	Ar_e_direção	Atualizados	Muito insatisfatória	Bastante satisfeito	20,6	4	518	1	45
10	Valentiniana	ADT_Freios_ABS	Atualizados	Não causou impressão	Insatisfeito	26,775	5	539	3	42

Figura 16: Base de dados da Toyord
Fonte: Adaptada de Linux (2008)

Veja que cada uma das variáveis é registrada em uma coluna específica e, nas linhas, encontram-se os registros de cada funcionário. Por exemplo, o respondente 1 adquiriu um modelo Deltaforce3, com os opcionais: ar-condicionado e direção hidráulica e considera o *design* do veículo atualizado. Disse que o atendimento da concessionária

ria onde comprou o veículo não causou impressão e que está muito insatisfeito com seu veículo. Tem renda mensal de 24,98 salários mínimos (R\$ 9.492,00), costuma levar cinco pessoas no veículo, trafega em média 415 km por mês, afirma que a última remodelação foi feita há dois anos e tem 35 anos de idade. Esse raciocínio pode ser estendido para os outros 249 respondentes. Analisando as variáveis isoladamente ou em conjunto, podemos atingir os objetivos da pesquisa.

O arquivo de dados mostrado na Figura 16 está disponível no Ambiente Virtual de Ensino-Aprendizagem no formato do Excel® e do Calc®. Juntamente com ele está disponibilizado o [texto Como realizar análise exploratória de dados no Br.Office Calc®](#), e como realizar análise exploratória de dados no Microsoft Excel®.

A maioria dos programas estatísticos, gerenciadores de bases de dados e de planilhas eletrônicas com capacidade estatística exige que os dados sejam estruturados de acordo com o formato da Figura 16. Podemos ter tantas colunas e linhas quantas quisermos, respeitando, porém, a capacidade dos programas. O Microsoft Excel®, por exemplo, admite 65.000 linhas, o que é suficiente para muitas aplicações.

Estando os dados no formato apropriado, especialmente se em meio digital, podemos passar para a etapa de análise. Uma das ferramentas mais úteis para isso é a distribuição de frequências, como veremos a seguir.

Distribuição de Frequências

O processo de resumo e de organização dos dados busca basicamente registrar as ocorrências dos possíveis valores das variáveis que caracterizam o fenômeno. Em suma, consiste em elaborar **Distribuições de Frequências** das variáveis para que o conjunto de dados possa ser reduzido e, assim, possibilite sua análise.

A construção da distribuição de frequências exige que os possíveis valores da variável sejam discriminados e seja contado o número de vezes em que cada valor ocorreu no conjunto de dados. Para grandes arquivos de informações, tal processo é viável somente utilizando-se meios computacionais.

Uma distribuição de frequências pode ser expressa através de tabelas ou de gráficos, que terão algumas particularidades dependendo do nível de mensuração da variável e de quantas variáveis serão analisadas. Veremos cinco casos: 1- quando há apenas uma variável

Veja a seção *Saiba mais* desta Unidade. O arquivo de dados e o texto servirão para as Unidades 3 e 4.

Distribuições de Frequências – organização dos dados de acordo com as ocorrências dos diferentes resultados observados. Fonte: Barbetta, Reis e Bornia (2008).

Se alguém diz que 33,33% das mulheres de um curso se casaram com professores, você pode ter uma má impressão delas. Mas se diz que das três mulheres (dados brutos) do tal curso uma delas se casou com o professor, o efeito é diferente. Fonte: Huff (1992).

Frequência absoluta – registro dos valores da variável por meio de contagem das ocorrências no conjunto de dados. Fonte: Barbetta, Reis e Bornia (2008).

Frequência relativa ou percentual – registro dos valores da variável por meio de proporção (relativa) ou percentagem (percentual) do total das ocorrências do conjunto de dados. Fonte: Barbetta, Reis e Bornia (2008).

qualitativa; 2- com apenas uma variável quantitativa; 3- com duas variáveis qualitativas; 4- com duas variáveis quantitativas; e 5- com uma variável qualitativa e uma quantitativa.

Caso de Uma Variável Qualitativa

Usualmente, uma variável qualitativa assume apenas alguns valores: basta, então, discriminá-los e contar quantas vezes eles ocorrem no conjunto. Essa contagem pode ser registrada em números absolutos, **frequência absoluta**, ou em números relativos, **frequência relativa ou percentual**. Ambos os registros devem ser feitos e apresentados: a frequência absoluta permite avaliar se os resultados são sólidos (é temerário tomar decisões com base em pequenas quantidades de dados); já a **frequência relativa** possibilita comparar os resultados da distribuição de frequências com outros conjuntos de tamanhos diferentes. A distribuição de frequências pode ser apresentada em forma de tabela ou gráfico.

Exemplo 2 – Imagine que você está interessado em descrever a variável “opinião sobre as concessionárias” (vista no Exemplo 1), isoladamente; e representar os dados em forma de tabela. Como ficariam os resultados? Saiba que o resultado seria semelhante ao ilustrado no Quadro 2, uma apresentação tabular da variável “opinião sobre concessionárias”.

VALORES	FREQUÊNCIA	PERCENTUAL
Muito insatisfatória	29	11,60%
Insatisfatória	58	23,20%
Não causou impressão	75	30,00%
Satisfatória	50	20,00%
Bastante satisfatória	38	15,20%
Total	250	100%

Quadro 2: Opinião dos clientes sobre as concessionárias Toyord

Fonte: Elaborado pelo autor deste livro

Concluimos, com esse segundo exemplo, que as concessionárias não são exatamente bem vistas pelos clientes: apenas 35,20% dos entrevistados as consideram satisfatórias ou bastante satisfatórias. Pense que, nesse caso, o administrador terá que descobrir as causas de tal resultado e atuar para resolver os problemas.

Podemos aplicar um raciocínio semelhante para as outras variáveis qualitativas e apresentar uma descrição gráfica da distribuição de frequências. Quando a variável é qualitativa, podemos usar dois tipos de gráficos: **em barras** ou **em setores**.

No gráfico de barras (Figura 17), são colocadas, em um dos eixos, as categorias da variável e, no outro, as frequências ou percentuais de cada categoria. As barras podem ser horizontais ou verticais (use essas preferencialmente). Para os dados do segundo exemplo, usando as frequências, temos:

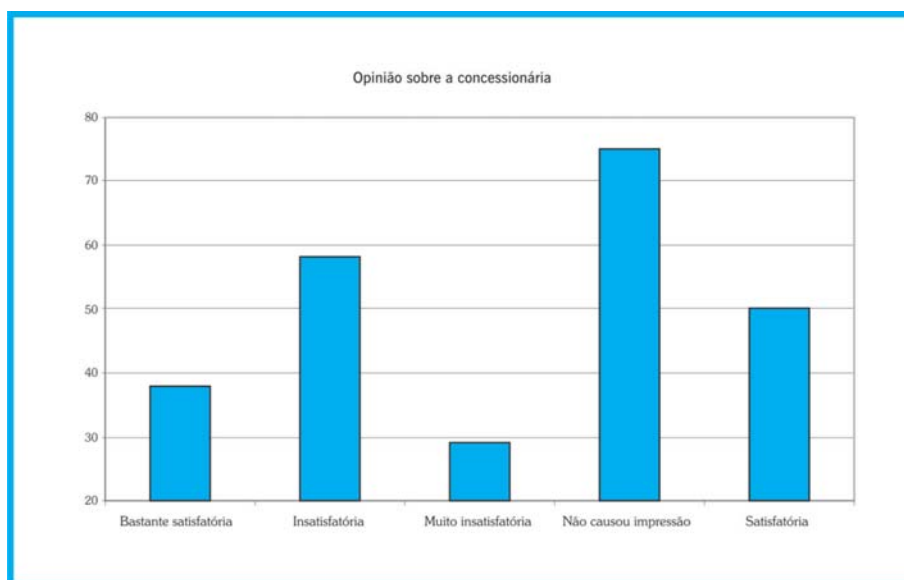


Figura 17: Opinião sobre as concessionárias
Fonte: Elaborada pelo autor deste livro

Trata-se da mesma distribuição de frequências observada no Quadro 2. A apreensão da informação, porém, é muito mais rápida através de um gráfico. Percebe-se claramente que a opção “Não causou impressão” apresenta maior frequência.

Contudo, você consegue identificar alguma particularidade no gráfico? Olhe bem!

A escala começa em 20, não em zero. Sendo assim, as diferenças relativas entre as frequências podem ser distorcidas, o que pode levar a uma interpretação diferente dos resultados: Portanto, cuidado com as escalas dos gráficos! É muito comum vermos erros grosseiros nessas escalas veiculados na mídia em geral. Provavelmente por ignorância, mas devemos estar atentos. Os administradores tomam decisões com base na interpretação de gráficos. Logo, estes devem retratar fielmente a realidade. Veja a Figura 18, com a escala correta.

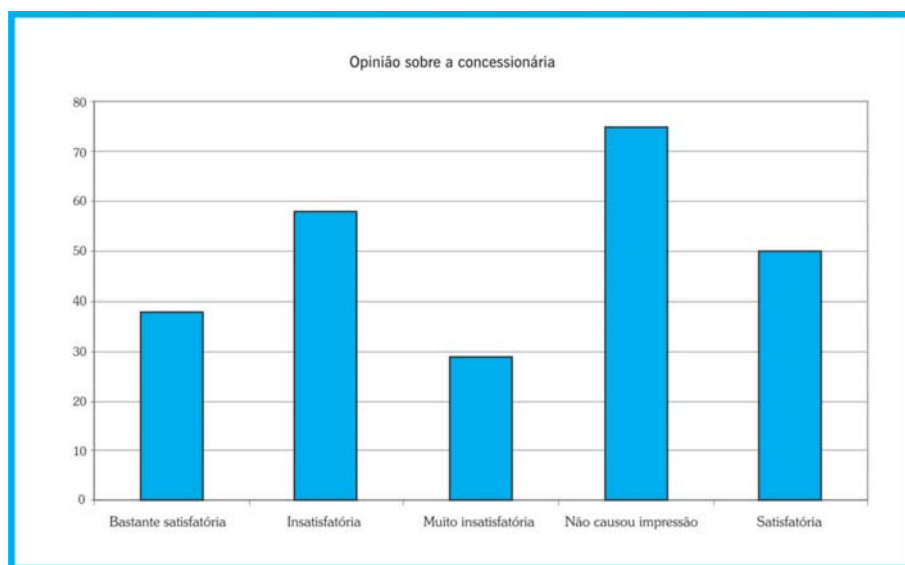


Figura 18: Opinião sobre as concessionárias
Fonte: Elaborada pelo autor deste livro

Outro tipo de gráfico bastante utilizado é o gráfico circular, em setores ou em "pizza". Ele é apropriado quando o número de valores da variável qualitativa não é muito grande, mas sua construção é um pouco mais elaborada do que o gráfico de barras. Consiste em dividir um círculo (360°) em setores proporcionais às realizações de cada categoria por meio de uma regra de três simples, na qual a frequência total (ou o percentual total: 100%) corresponderia aos 360° e a frequência, ou a proporção de cada categoria, corresponderia a um valor desconhecido em graus.

$$\text{Graus de uma categoria} = \frac{360^\circ \times \text{frequência (proporção) da categoria}}{\text{frequência (proporção) total}}$$

Observe os valores em graus correspondentes aos resultados do Quadro 3.

VALORES	FREQÜÊNCIA	PERCENTUAIS	GRAUS
Muito insatisfatória	29	11,60%	41,76
Insatisfatória	58	23,20%	83,52
Não causou impressão	75	30,00%	108
Satisfatória	50	20,00%	72
Bastante satisfatória	38	15,20%	54,72
Total	250	100%	360

Quadro 3: Opinião dos clientes sobre as concessionárias Toyord

Fonte: Elaborado pelo autor deste livro

E o gráfico em setores será conforme apresentado na Figura 19:

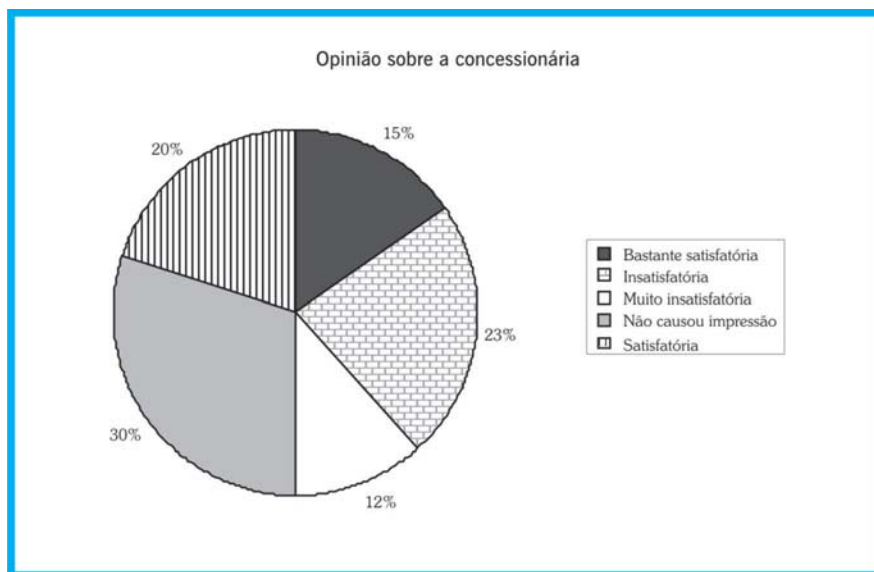


Figura 19: Opinião sobre as concessionárias

Fonte: Elaborada pelo autor deste livro

Pela observação dos percentuais, é possível reparar o predomínio da opção “Não causou impressão”, com 30% das respostas. Se a variável qualitativa contivesse muitos valores (por exemplo, bairros da região metropolitana de São Paulo), o gráfico dificilmente resumiria alguma coisa, pois teria um número excessivo de fatias. Isso também ocorre com variáveis quantitativas, especialmente as contínuas.

Tabela de contingências – tabela que permite analisar o relacionamento entre duas variáveis, nas linhas são postos os valores de uma delas, e nas colunas os da outra, e nas células contam-se as frequências de todos os cruzamentos possíveis. Fonte: Barbetta (2007).

Caso de Duas Variáveis Qualitativas

O administrador, frequentemente, precisa estudar a relação entre duas ou mais variáveis para tomar decisões. Por exemplo: existe relação entre o sexo do consumidor e a preferência por um modelo de carro? Ou entre a escolaridade do eleitor e o candidato eleito presidente?

Quando as duas variáveis são qualitativas (originalmente, ou quantitativas categorizadas), usualmente é construída uma distribuição conjunta de frequências, também chamada de **tabela de contingências** ou dupla classificação e, nela, são contadas as frequências de cada cruzamento possível entre os valores das variáveis. A expressão pode incluir o cálculo de percentuais em relação ao total das linhas, colunas ou total geral da tabela. A representação gráfica também é possível.

Vamos ver um caso para a mesma situação do Exemplo 1. Agora você está interessado em observar a relação entre a variável modelo adquirido e a opinião geral do cliente sobre o veículo e expressá-la de forma tabular e gráfica.

A variável modelo apresenta cinco resultados possíveis (cinco modelos foram considerados nesta pesquisa), e a variável opinião geral pode assumir quatro resultados (Bastante satisfeito, satisfeito, insatisfeito e muito insatisfeito). Isso significa que podemos ter até 20 cruzamentos possíveis para os quais precisamos contar as frequências. Para grandes bases de dados, inclusive para o nosso exemplo com apenas 250 casos, seria um processo tedioso e suscetível a erros. Portanto, o mais inteligente é utilizar alguma ferramenta computacional, como a planilha eletrônica Microsoft Excel® ou o Br.Office Calc®.

Empregando-se uma ferramenta computacional, chegaremos ao Quadro 4.

	OPINIÃO GERAL SOBRE O VEÍCULO				
MODELO	MUITO INSATISFEITO	INSATISFEITO	SATISFEITO	BASTANTE SATISFEITO	TOTAL
		1			1
Chiconaultla	69	11	1	0	81
DeltaForce3	29	22	5	0	56
Valentiniana	11	18	9	3	41
SpaceShuttle	1	14	17	10	42
LuxuriousCar	0	1	9	19	29
Total	110	67	41	32	250

Quadro 4: Tabela de contingências de modelo por opinião geral (apenas frequências)

Fonte: Elaborado pelo autor deste livro

Observe a última coluna e a última linha do Quadro 4: são os chamados **totais marginais**, isto é, as frequências dos valores das variáveis Modelo e Opinião geral sobre o veículo, respectivamente. Percebe-se que os modelos Chiconaultla e DeltaForce3 são os mais vendidos, e que as opiniões negativas (muito insatisfeito e insatisfeito) são mais frequentes do que as positivas.

Além disso, é fácil perceber que as opiniões negativas são as predominantes nos modelos Chiconaultla, DeltaForce3 e em menor grau no Valentiniana. Apenas os modelos SpaceShuttle e LuxuriousCar têm proprietários predominantemente satisfeitos.

Você deve ter percebido também uma linha com várias células vazias (apenas uma observação na opção insatisfeito). Trata-se de um **dado perdido**: o entrevistado esqueceu de mencionar o modelo adquirido, ou o entrevistador não o registrou durante a realização da pesquisa, ou mesmo houve um erro de digitação. Como a quantidade aqui é muito pequena (1 em 250, 0,4%), não causará problemas. Apenas quando a quantidade ultrapassa 5% da base de dados há motivo para preocupação; pois, ou houve muitos erros de digitação na tabulação dos dados ou o **instrumento de pesquisa** foi mal projetado, pois muitos elementos da população não forneceram as informações desejadas.

O Quadro 4 pode ser apresentado de forma gráfica, através de um gráfico de barras múltiplas.

Totais marginais – totais das linhas ou das colunas de uma tabela de contingência, permitem avaliar individualmente as variáveis componentes da tabela. Fonte: Bussab e Morettin (2003).

Conforme vimos na Unidade 1.

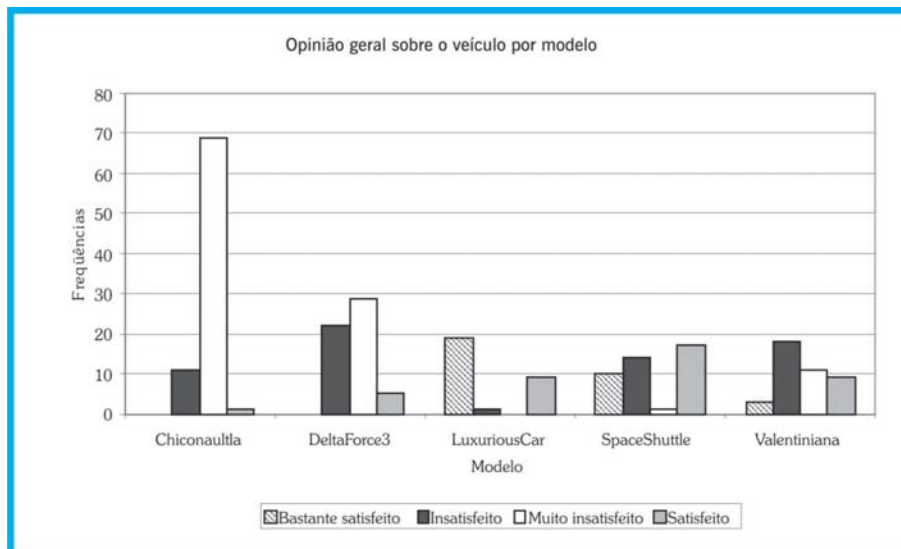


Figura 20: Opinião geral por modelo
Fonte: Elaborada pelo autor deste livro

As frequências absolutas podem ser insuficientes para a interpretação dos resultados, especialmente quando comparamos os resul-

tados com outros conjuntos de dados de tamanhos diferentes. Assim, podemos calcular percentuais em relação aos totais de cada coluna, ou em relação aos totais de cada linha ou ainda ao total geral da tabela. Vamos apresentar, no Quadro 5, apenas um dos percentuais possíveis em relação aos totais das linhas (maiores detalhes nos textos “Como realizar análise exploratória de dados no Microsoft Excel®” e “Como realizar análise exploratória de dados com o Br.Office Calc®”):

MODELO	OPINIÃO GERAL SOBRE O VEÍCULO				
	MUITO INSATISFEITO	INSATISFEITO	SATISFEITO	BASTANTE SATISFEITO	TOTAL
Chiconaultla	69 85,19%	11 13,58%	1 1,23%	0 0,00%	81 100%
DeltaForce3	29 51,79%	22 39,29%	5 8,93%	0 0,00%	56 100%
Valentiniana	11 26,83%	18 43,90%	9 21,95%	3 7,32%	41 100%
SpaceShuttle	1 2,38%	14 33,33%	17 40,48%	10 23,81%	42 100%
LuxuriousCar	0 0,00%	1 3,45%	9 31,03%	19 65,52%	29 100%
Total	110 44,18%	66 26,51%	41 16,47%	32 12,85%	249 100%

Quadro 5: Tabela de contingência de opinião geral por Modelo
(com % por linha)

Fonte: Elaborado pelo autor deste livro

Vistos os exemplos, o que você pode concluir acerca da satisfação dos clientes com relação aos modelos? Qual modelo deveria receber atenção prioritária?

Veja que o cruzamento de duas variáveis qualitativas é atividade corriqueira para o administrador que, frequentemente, precisa avaliar mais de duas variáveis, o que exige métodos matemáticos sofisticados, implementados computacionalmente. A seguir, aprofundaremos-nos nesse tema.

Caso de Uma Variável Quantitativa

A construção das distribuições de frequências para variáveis quantitativas é semelhante ao caso das variáveis qualitativas: relacionar os valores da variável com as suas ocorrências no conjunto de dados. Todavia, apresenta algumas particularidades a depender de a variável ser **discreta** ou **contínua**.

Se a variável fosse quantitativa discreta e pudesse assumir apenas alguns valores, a abordagem seria semelhante a das variáveis qualitativas. A diferença reside na substituição de atributos por números, que gera uma **distribuição de frequência para dados não agrupados**. Vamos ver um exemplo.

Exemplo 4 – para a mesma situação do Exemplo 1: imagine que você está interessado em descrever a variável número de pessoas usualmente transportadas no veículo, isoladamente, e representar os dados em forma de tabela. Como ficariam os resultados? O resultado seria semelhante ao mostrado no Quadro 6, uma apresentação tabular (em forma de tabela) da variável número de pessoas transportadas.

VALORES	FREQUÊNCIA	PERCENTUAL
1	19	7,60%
2	29	11,60%
3	43	17,20%
4	42	16,80%
5	57	22,80%
6	60	24,00%
Total	250	100%

Quadro 6: Número de pessoas usualmente transportadas no veículo

Fonte: Elaborado pelo autor deste livro

Pela observação do Quadro 6, podemos concluir que os veículos têm uso predominantemente “familiar” (várias pessoas transportadas usualmente). Sabendo disso, o administrador pode decidir por direcionar o marketing ou mesmo a produção de modelos visando ao segmento de famílias maiores. Uma abordagem semelhante poderia ser aplicada para as outras variáveis discretas: anos de remodelação e a idade dos consumidores.

A maioria dos programas (estatísticos ou não) que constroem em histogramas para variáveis quantitativas discretas costuma ignorar isso.

Como representar a distribuição de frequências para variáveis qualitativas discretas graficamente? O Quadro 6 poderia ser representado através de um **Histograma**, um gráfico de barras justapostas, e que as áreas das barras são proporcionais às frequências de cada valor. Vamos ver a Figura 21:

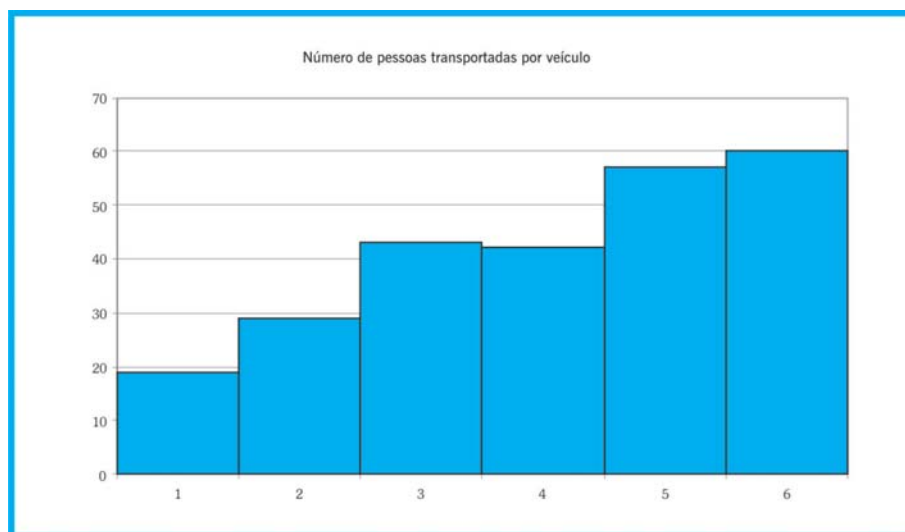


Figura 21: Histograma do número de pessoas transportadas por veículo
Fonte: Elaborada pelo autor deste livro

Nesse caso, poderíamos usar o gráfico em setores? A resposta é não, pois formalmente o gráfico em setores deve ser usado apenas para variáveis qualitativas. A interpretação é a mesma, mas a apreensão da informação é mais rápida. Observe que não há problemas com a escala vertical, pois ela começa em zero.

Se a variável quantitativa for contínua, o procedimento descrito anteriormente será inviável como instrumento de resumo do conjunto, porque todos os valores têm praticamente frequência baixa, o que resultaria em uma tabela enorme.

Se o conjunto de dados for pequeno, até 100 observações, é possível usar ferramentas gráficas como o diagrama de pontos e o ramo em folhas.

Se o conjunto for grande, é preciso representar os dados por meio de um conjunto de faixas de valores mutuamente exclusivas (para que cada valor pertença apenas a uma faixa), que contenha do menor ao maior valor do conjunto: registram-se, então, o número de valores do conjunto que se encontram em cada faixa. Há duas maneiras de fazer isso:

- através da **categorização** (recodificação) da variável. Por exemplo: todos que ganham até quatro salários mínimos (R\$ 1.680,00) pertencem à classe baixa; os que ganham entre 4,01 e 20 salários mínimos (até R\$ 8.400,00) pertencem à classe média; e, acima disso, pertencem à classe alta. Essa abordagem é largamente utilizada na mídia.
- mediante a **distribuição de frequências para dados agrupados** (ou agrupada em classes), processo mais elaborado e mais “estatístico”. Veremos o procedimento a seguir.

O processo para montagem da distribuição de frequências para dados agrupados é o seguinte:

- Determinar o intervalo do conjunto (diferença entre o maior e o menor valor do conjunto).
- Dividir o intervalo em um número conveniente de classes, onde: $N^{\circ} \text{ de classes} = \sqrt{N^{\circ} \text{ de elementos}}$. Nesse ponto há grande controvérsia entre os estatísticos e a fórmula apresentada é apenas uma das opções possíveis. Admite-se que o número mínimo de classes seja igual a 5 e o máximo 20, mas se aceita uma definição arbitrária nesse intervalo.
- Estabelecer as classes com a seguinte notação:

Li = limite inferior; Ls = limite superior;

Li | -- Ls = limite inferior incluído, superior excluído; e

Li | -- | Ls = ambos incluídos.

Determinar as frequências de cada classe.

Determinar os pontos médios de cada classe por meio da média dos dois limites (serão os representantes das classes).

Categorização – processo pelo qual se transforma uma variável quantitativa em qualitativa, associando atributos a intervalos de valores numéricos, por exemplo, classe A para uma certa faixa de renda familiar. Fonte: Elaborado pelo autor deste livro.

Distribuição de frequências para dados agrupados – distribuição de frequências na qual os valores da variável são agrupados em faixas de ocorrência, e as frequências contadas para cada faixa, para facilitar o resumo do conjunto de dados, usualmente empregado para variáveis quantitativas contínuas. Fonte: Barbetta, Reis e Bornia (2008).

Vamos ver exemplos de ambas as abordagens.

Exemplo 5 – para a mesma situação do Exemplo 1: imagine que você está interessado em descrever a variável renda dos consumidores, isoladamente, e representar os dados em forma de tabela. Como ficariam os resultados nos seguintes casos:

- Se optássemos por categorizar a variável da seguinte forma: todos que ganham até quatro salários mínimos (R\$ 1.680,00) pertencem à classe baixa; os que ganham entre 4,01 e 20 salários mínimos (até R\$ 8.400,00) pertencem à classe média; e, acima disso, pertencem à classe alta?
- Se optássemos por uma distribuição de frequências para dados agrupados?

A resolução passo a passo desse problema está na seção *Saiba mais* desta Unidade, que explica como realizar Análise Exploratória de Dados no Excel. Aqui, apresentaremos apenas os resultados finais.

No caso do item a, a categorização levará à criação de uma nova variável, agora qualitativa, permitindo uma abordagem semelhante a que vimos anteriormente. No Quadro 7 e na Figura 22 estão os resultados: tabela de frequências e gráfico em setores.

VALORES	FREQUÊNCIA	PERCENTUAL
Classe baixa (até 2 s.m.)	2	0,8%
Classe média (entre 2,01 e 20 s.m.)	104	41,6%
Classe alta (acima de 20 s.m.)	144	57,6%
Total	250	100%

Quadro 7: Renda categorizada em classe social
Fonte: Elaborado pelo autor deste livro

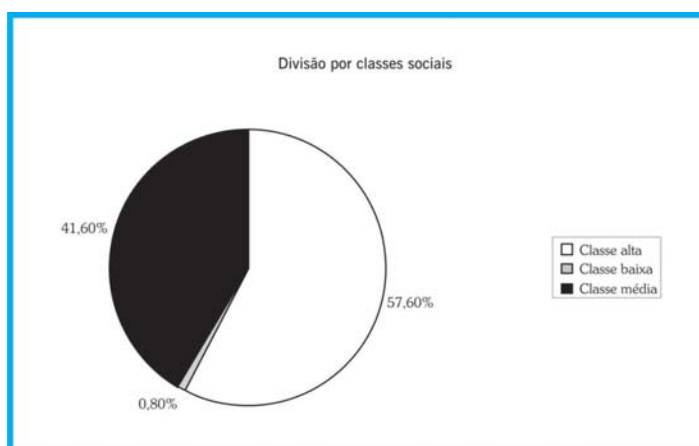


Figura 22: Gráfico em setores para a renda categorizada em classes
Fonte: Elaborada pelo autor deste livro

Observe que perdemos informação sobre os dados originais de renda ao fazer a categorização. A interpretação é relativamente simples: a maioria absoluta (mais de 50%) dos clientes da montadora pode ser considerada de classe alta (renda superior a 20 salários mínimos mensais). A grande discussão que surge nesse caso é quem define o que é classe baixa, média ou alta (ou A, B, C, D e E). Uma sugestão é utilizar a classificação do IBGE.

Passando para o item b, devemos seguir os passos:

- $\text{Intervalo} = \text{Maior} - \text{Menor} = 86,015 - 1,795 = 84,22$ (a maior renda é de 86,015 salários mínimos e a menor de 1,795, as classes devem englobar do menor ao maior valor).
- $\text{N}^\circ \text{ de classes} = \sqrt{\text{N}^\circ \text{ de elementos}} = \sqrt{250} = 15,81 \cong 16$. Por esse expediente, deveríamos usar 16 classes. Porém, conforme foi dito anteriormente, o número de classes pode ser definido de forma arbitrária. Para simplificar nosso problema vamos usar cinco classes.

$\text{Amplitude das classes} = 86,015/5 = 16,844$ (valor exato).

A amplitude das classes pode ser ligeiramente maior do que a obtida anteriormente, poderíamos, novamente tentando a simplificação do problema, usar amplitude igual a 16,85. Se a amplitude não for um valor exato, deve sempre ser arredondado para cima, garantindo que as classes contêm do menor ao maior valor. As classes podem, então, ser definidas.

- Classes:

1,795 | - 18,645 18,645 | - 35,495 35,495 | - 52,345
52,345 | - 69,195 69,195 | - 86,045

(nesse caso, o ponto inicial foi o próprio menor valor do conjunto. Poderia ser outro valor conveniente abaixo do menor valor).

- Pontos médios de cada classe: $(\text{limite inferior} + \text{limite superior})/2$
(os pontos médios calculados estão no quadro a seguir)
- Frequências de cada classe (Quadro 8):

Esses valores foram obtidos no arquivo de dados citado no início desta Unidade.

CLASSES	FREQUÊNCIA	PERCENTUAL	PONTOS MÉDIOS
1,795 - 18,645	98	39,2%	10,22
18,645 - 35,495	102	40,8%	27,07
35,495 - 52,345	38	15,2%	43,92
52,345 - 69,195	9	3,6%	60,77
69,195 - 86,045	3	1,2%	77,62
Total	250	100%	-

Quadro 8: Renda agrupada em classes

Fonte: Elaborado pelo autor deste livro

Observe que perdemos informação sobre o conjunto original: sabemos que há 98 pessoas com renda entre 1,795 e 18,645 salários mínimos, mas não quais são os seus valores exatos, ou seja, as frequências das classes passam a ser as frequências dos pontos médios. Podemos afirmar que quase 80% dos clientes têm renda até 35,495 salários mínimos.

O Quadro 8 também pode ser representado através de um histograma (Figura 23), uma vez que a variável permanece sendo formalmente quantitativa. Mas o histograma para uma tabela de dados agrupados é diferente daquele visto anteriormente. O número de barras é igual ao número de classes. Cada barra é centrada no ponto médio de cada classe, o ponto inicial de cada barra é o limite inferior da classe, e o ponto final é o limite superior.

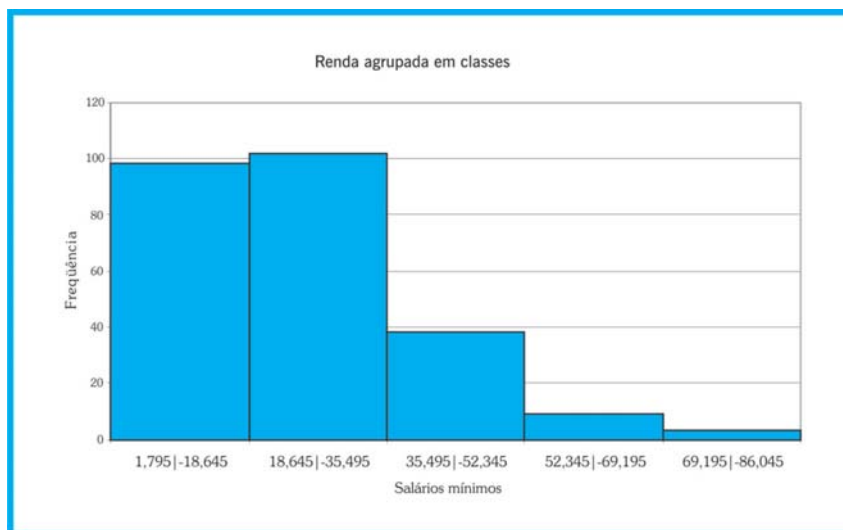


Figura 23: Histograma para renda agrupada em classes

Fonte: Elaborada pelo autor deste livro

Note que a interpretação é mais direta quando observamos o gráfico apresentado na Figura 23.

Mas aqui surge um fato interessante. Parece haver contradição com a interpretação do item a, pois concluímos que a maioria absoluta dos clientes é de classe alta. Isso ocorre devido à definição arbitrária das classes. Pense em como você resolveria essa contradição.

O agrupamento em classes apresenta algumas desvantagens, além da já citada perda de informação sobre o conjunto original.

Os pontos médios nem sempre são os representantes mais fiéis das classes. Para grande quantidade de dados, existe maior probabilidade de que essas estimativas correspondam exatamente aos verdadeiros valores. Outro problema são as **medidas estatísticas** calculadas com base na distribuição de frequências para dados agrupados: serão apenas estimativas dos valores reais devido à perda de informação referida anteriormente.

A tendência atual é **não calcular** medidas estatísticas com base em tabelas de dados grupados.

Agora, vamos ver como analisar o relacionamento entre duas variáveis quantitativas.

Caso de Duas Variáveis Quantitativas

Muitas vezes também estamos interessados em avaliar a relação entre variáveis quantitativas, sejam elas discretas ou contínuas.

Basicamente, há interesse em, a partir de dados, verificar **se e como** duas variáveis quantitativas relacionam-se entre si em uma população, ou seja, avaliar se há correlação entre elas, e avaliar a força e a direção (se elas caminham na mesma direção ou em direções opostas) dessa correlação, caso ela exista.

Uma das variáveis é chamada de independente. Essa pode ser uma variável que o pesquisador manipulou para observar o efeito em outra, ou alguma cuja medição possa ser feita de maneira mais fácil ou precisa, sendo então suposta sem erro.

Há outra variável chamada de dependente cujos valores são resultado da variação dos valores das variáveis independentes.

Correlação – medida de associação entre duas variáveis quantitativas. Fonte: Barbetta, Reis e Bornia (2008).

A denominação de correlação costuma levar a má interpretação do significado da “correlação” entre variáveis: se há correlação entre variáveis significa que os seus valores variam em uma mesma direção, ou em direções opostas, com certa “força”, ou seja, correlação não significa causalidade.

Observações emparelhadas – medidas de duas ou mais variáveis que foram realizadas com a mesma unidade experimental/amostral, no mesmo momento. Fonte: Elaborado pelo autor deste livro.

Saiba mais nos textos *Como realizar Análise Exploratória de Dados no Br. Office Calc®* e *Como realizar Análise Exploratória de Dados no Microsoft Excel®*.

Por exemplo, pode haver correlação entre a pluviosidade mensal (em mm) em Florianópolis e o número de ratos exterminados por mês na cidade de Sidney, na Austrália, mas seria um pouco forçado imaginar que uma coisa “causou” a outra. É necessário usar bom senso!

Em outro caso, ao avaliarmos o relacionamento entre renda mensal em reais e área em m^2 da residência de uma família, esperamos um relacionamento positivo entre ambas: para maior renda (independente), esperamos maior área (dependente).

Para que seja possível avaliar o relacionamento entre duas variáveis (nesse caso quaisquer delas, não apenas quantitativas), os dados devem provir de **observações emparelhadas** e em condições semelhantes. Desse modo, ao avaliarmos a correlação existente entre a altura e o peso de um grupo de crianças, por exemplo, o peso de determinada criança deve ser medido e registrado no mesmo instante em que é medida e registrada a sua altura e correlacionado à variável renda e à área da residência da mesma família, no mesmo momento.

Se estivermos analisando duas variáveis quantitativas, cujas observações constituem pares ordenados, denominadas essas variáveis de **X** (independente) e **Y** (dependente), poderemos plotar o conjunto de pares ordenados (x,y) em um diagrama cartesiano, que é chamado de **Diagrama de Dispersão**. Atualmente, isso pode ser feito com aplicativos computacionais, uma planilha eletrônica como o **Br.Office Calc®** ou **Microsoft Excel®**.

Por meio do diagrama de dispersão, podemos ter uma ideia inicial de como as variáveis estão relacionadas: a direção da correlação (isto é, quando os valores de **X** aumentam, os valores de **Y** aumentam também ou diminuem); a força da correlação (em que “taxa” os valores de **Y** aumentam ou diminuem em função de **X**); e a nature-

za da correlação (se é possível ajustar uma reta, parábola, exponencial aos pontos).

Para esclarecer, vamos ver um exemplo.

Uma empresa agroindustrial processa soja para obter óleo. A direção quer estudar a relação entre o valor da soja (em dólares por tonelada) na bolsa de cereais de Chicago e a cotação da ação da empresa (em dólares) na bolsa de Nova Iorque. Para tanto coletou um conjunto de 400 pares de observações e plotou o diagrama de dispersão exposto na Figura 24.

Observando o diagrama (Figura 24), é possível afirmar que a relação entre as variáveis é fortemente linear?

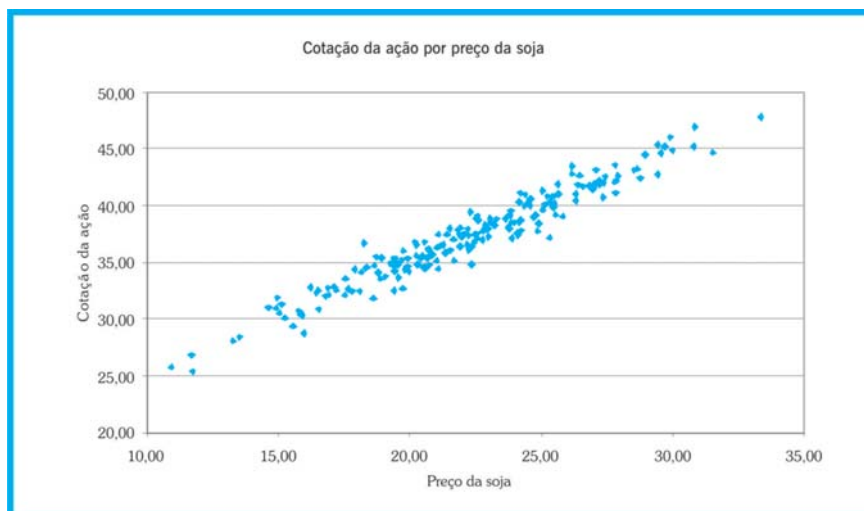


Figura 24: Diagrama de dispersão de cotação da ação por preço da soja
Fonte: Elaborada pelo autor deste livro

A correlação entre as variáveis é claramente positiva: maiores valores de preço da soja correspondem a maiores valores de cotação da ação, o que parece plausível. A correlação parece ser muito forte, pois os pontos estão muito próximos. Quanto à natureza, é possível observar que seria possível ajustar uma reta entre os pontos. Portanto, concluímos que a relação entre as variáveis é fortemente linear. Poderíamos então obter a equação da reta para, a partir dos valores da soja, prever a cotação da empresa agroindustrial.

Se uma das variáveis quantitativas for o tempo (medido em anos, meses, semanas, dias) teremos uma série temporal.

Série temporal – conjunto de observações de uma variável quantitativa, ordenado no tempo (diário, semanal, mensal, anual).
Fonte: Moore, McCabe, Duckworth e Sclove (2006).

Em *Saiba mais*, vamos apresentar algumas referências sobre o assunto, que serão extremamente úteis

caso você precise lidar com séries temporais.

Você já deve ter visto em algum lugar uma tabela ou gráfico mostrando evolução do PIB do Brasil ao longo dos anos, ou a evolução da população de um país, ou mesmo os percentuais de intenção de voto para os candidatos à presidência a cada pesquisa eleitoral. O objetivo da **análise de uma série temporal** é identificar a existência de padrões que nos auxiliem a tomar decisões.

Vamos agora ao último caso desta Unidade, que é muito importante para o administrador por ser comum a necessidade de estudar a relação entre uma variável qualitativa e outra quantitativa.

Caso de Uma Variável Qualitativa e Uma Quantitativa

Em geral, pressupõe-se que analisaremos a variável quantitativa em função dos valores da variável qualitativa, visto que esta costuma ter menos opções, o que simplificaria o processo e permitiria resumir mais os dados.

Na Unidade 1, falamos sobre classificação das variáveis por nível de manipulação, em independente e dependente. Se estivermos estudando duas variáveis, uma qualitativa e outra quantitativa, a qualitativa será considerada independente (ou de agrupamento) e a quantitativa a dependente. Vejamos dois exemplos rápidos.

Imagine que você está realizando uma pesquisa experimental. Há interesse em avaliar a resposta a um medicamento contra o diabetes, que deveria reduzir o nível de glicose no sangue dos indivíduos portadores da doença. Para testar a eficiência do medicamento, você realiza um experimento sorteando dois grupos de voluntários. Um grupo receberá o medicamento e o outro o placebo durante um período de tempo. Ao final do experimento, os níveis de glicose dos indivíduos dos dois grupos são medidos para avaliar se no grupo que recebeu o medicamento eles sofreram redução significativa. Há duas variáveis, a independente, grupo de indivíduos, com dois valores (grupo tratado e grupo placebo), qualitativa; e a dependente, nível de glicose no sangue, quantitativa. Nesse caso, a definição de variável independente como a que é manipulada para causar um efeito na dependente é aceitável.

Em outra situação, em uma pesquisa de levantamento, a variável independente seria meramente uma variável de agrupamento, para categorizar a variável dependente. Vamos ver um exemplo a respeito.

Para a mesma situação do Exemplo 1. No caso, gostaríamos de avaliar se existe alguma relação entre a renda do consumidor e o modelo adquirido. Espera-se que tal relação exista, pois os modelos Chiconaultla e DeltaForce3 são os mais baratos, e o sofisticado LuxuriousCar é o mais caro de todos.

Nesse caso, podemos obter distribuições de frequências da variável Renda para cada valor da variável Modelo. Seria uma situação semelhante a do item b do Exemplo 4, mas agora com cinco tabelas, uma para cada opção de Modelo.

Muito Importante! Se optarmos por agrupamento em classes, todas as tabelas precisam ter o mesmo número de classes e as mesmas amplitudes de classe para que possamos comparar os grupos. No nosso caso, vamos usar as classes obtidas no item b do Exemplo 4 para as cinco tabelas:

1,795 | -18,645 18,645 | -35,495 35,495 | -52,345
52,345 | -69,195 69,195 | -86,045

Basta, então, ordenar as rendas em função dos modelos e contar as frequências em cada modelo, resultando os dados ilustrados no Quadro 9:

RENDA	MODELO					
	CHICONAULTLA	DELTAFORCE3	VALENTINIANA	SPACESHUTTLE	LUXURIOUSCAR	TOTAL
1,795 -18,645	73	20	4	0	0	97
18,645 -35,495	7	35	32	24	4	102
35,495 -52,345	1	1	4	18	14	38
52,345 -69,195	0	0	1	0	8	9
69,195 -86,045	0	0	0	0	3	3
Total	81	56	41	42	29	249

Quadro 9: Distribuições de frequências de renda agrupadas em classe por modelo

Fonte: Elaborado pelo autor deste livro

Há 249 dados na tabela porque o dado perdido (descoberto no Quadro 5) foi removido do conjunto.

Observe a semelhança da tabela indicada no Quadro 9 com aquela do Quadro 8. Da mesma forma que lá fizemos, é possível calcular percentuais em relação aos totais das linhas, colunas ou total geral.

Podemos perceber que a relação esperada entre as variáveis foi confirmada: para os modelos mais baratos a renda mais alta está na classe de 35,495 a 52,345 salários mínimos; já os clientes do modelo mais caro (LuxuriousCar) estão nas classes mais altas.

A tabela do Quadro 9 poderia ser expressa por intermédio de um gráfico, um **histograma categorizado**. Infelizmente tal gráfico não pode ser feito em uma planilha eletrônica sem consideráveis manipulações. Mas através de um *software* estatístico: o Statsoft Statistica 6.0®, isso é possível:

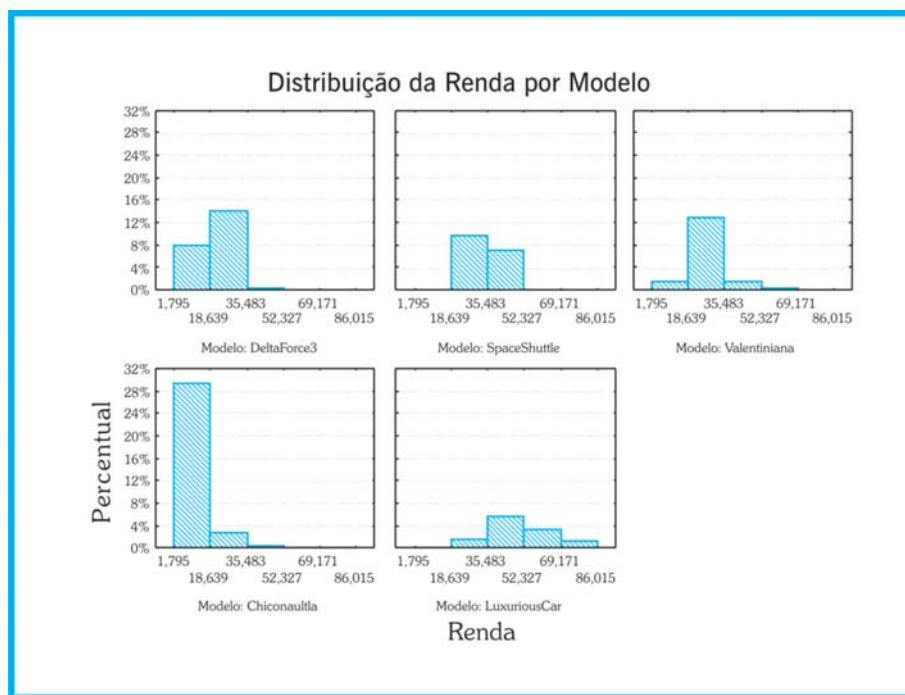


Figura 25: Histograma categorizado da renda por modelo adquirido
Fonte: Elaborada pelo autor deste livro

Observe que o *software* dividiu a variável renda em cinco classes também, mas com limites ligeiramente diferentes dos nossos. Além disso, optamos por apresentar os resultados em percentuais relativos ao total dos dados (249). A interpretação é semelhante à da tabela.

Na prática, o mais comum quando analisamos uma variável quantitativa em função de uma qualitativa é calcular medidas de sín-

tese daquela para cada grupo definido pelos valores desta. A partir dos resultados é possível verificar se existe relação entre as variáveis. Veremos na Unidade 4 as medidas de síntese.

Saiba mais...

Sobre correlação entre variáveis (quantitativas): BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Ed. da UFSC, 2007. Capítulo 13.

Sobre correlação entre variáveis (quantitativas): MOORE, David S.; McCABE, George P.; DUCKWORTH, William M.; SCLOVE, Stanley L. *A prática da estatística empresarial: como usar dados para tomar decisões*. Rio de Janeiro: LTC, 2006. Capítulo 2.

Sobre análise de séries temporais: LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística: Teoria e Aplicações – Usando Microsoft Excel em Português*. 5. ed. Rio de Janeiro: LTC, 2005. Capítulo 13.

Sobre análise de séries temporais: STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001. Capítulo 16.

Sobre como realizar as análises descritas nesta Unidade e na Unidade 4 empregando-se o Microsoft Excel® consulte *Como realizar Análise Exploratória de Dados no Microsoft Excel®*, disponível no ambiente virtual assim como o arquivo de dados usado nos exemplos apresentados.

Sobre como realizar as análises descritas nesta Unidade e na Unidade 4 através do Br.Office Calc® consulte *Como realizar Análise Exploratória de Dados com o Br.Office Calc®* disponível no ambiente virtual assim como o arquivo de dados usado nos exemplos apresentados.



A Figura 26 contém o resumo desta Unidade.

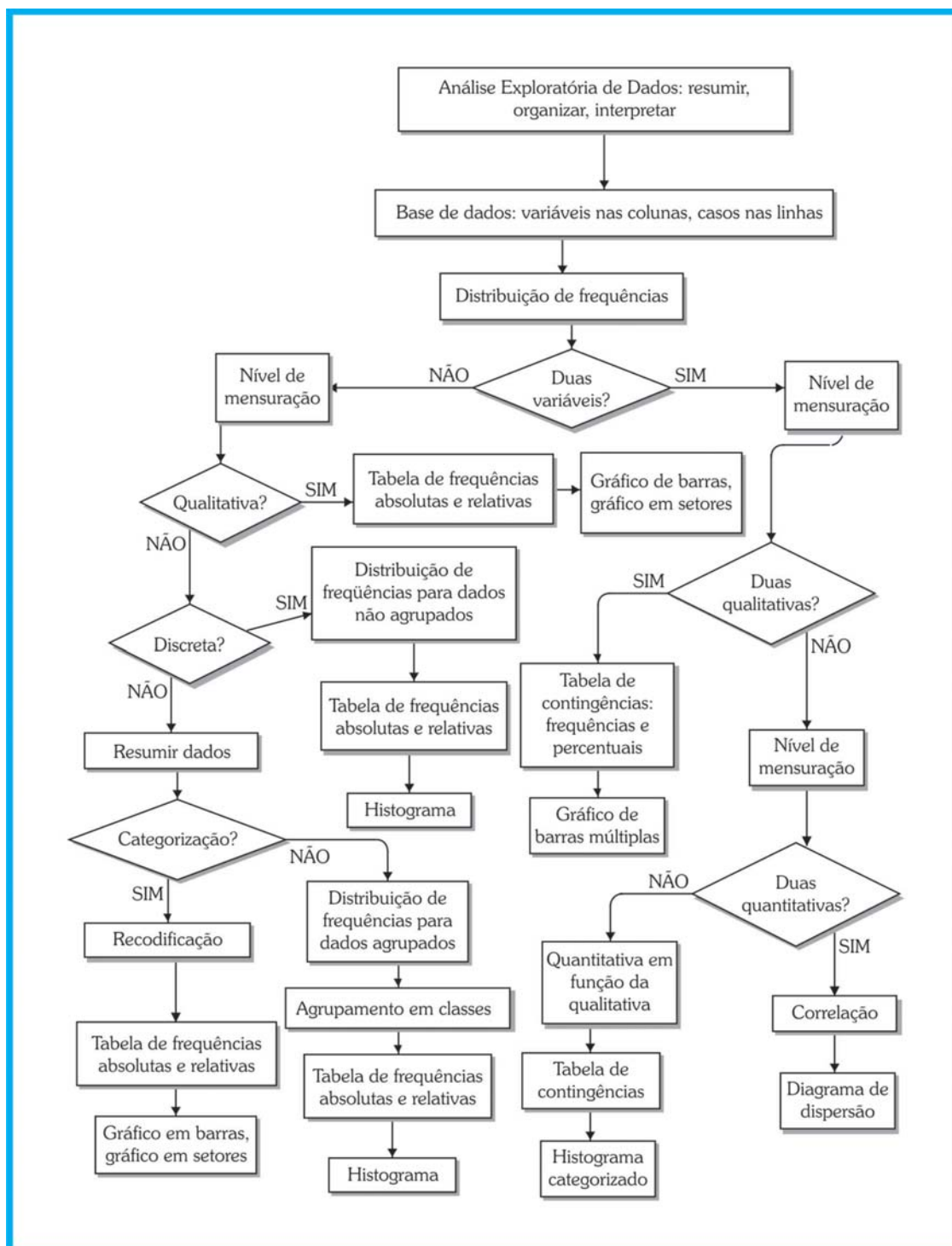


Figura 26: Resumo da Unidade 3
Fonte: Elaborada pelo autor deste livro



Caro estudante,

Esta Unidade foi importantíssima para você entender a Análise Exploratória de Dados. Vimos como organizar, interpretar e resumir as informações coletadas, níveis de mensuração e número de variáveis. Você aprendeu a elaborar tabelas, planilhas e gráficos de acordo com as especificidades das informações colhidas. Chegamos ao final da Unidade e ao começo de uma nova aprendizagem.

Esta Unidade lhe deu base para o aprendizado proposto nas Unidades seguintes. Releia, quantas vezes precisar, os variados exemplos propostos para cada categoria estudada. As Figuras, os quadros, as representações e os exemplos são importantes recursos nesse processo de aprendizagem. Interaja com sua turma e responda às atividades. A tutoria está pronta a auxiliá-lo e o professor interessado em seu desenvolvimento.

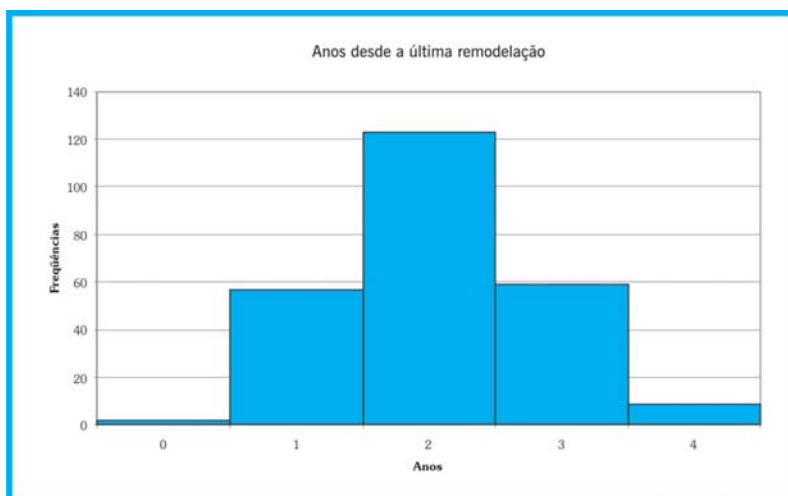
Vamos em frente!



Atividades de aprendizagem

Ao final da Unidade 3, chegou o momento de avaliar seus conhecimentos. Para isso, responda às atividades propostas a seguir e encaminhe-as para seu tutor por meio do Ambiente Virtual de Ensino-Aprendizagem. As atividades devem ser feitas usando o Microsoft Excel®, através do arquivo AmostraToyord.xls que está nesse ambiente, salvo onde indicado.

1. Construa a distribuição de frequências para a variável opinião sobre o *design* dos veículos da Toyord. De acordo com a análise dos resultados, os clientes têm, de forma geral, boa impressão sobre o *design* dos veículos da TOYORD? Justifique.
2. A variável anos de remodelação dos veículos (na percepção do cliente) está representada no histograma a seguir:



O departamento de marketing alega que precisa de mais orçamento para “convencer” os clientes de que os veículos da TOYORD têm *design* avançado, pois eles creem que a maioria dos clientes acha

- que eles foram remodelados “há vários anos”. Os dados confirmam a crença do departamento de marketing? Justifique.
3. Construa a distribuição de frequências agrupada em classes para a variável quilometragem média mensal percorrida com o veículo. Você considera que os clientes da Toyord usam bastante o veículo, ou não? Justifique.
 4. Construa a tabela de contingências para modelo por opinião sobre concessionárias. Com base nela avalie se os clientes de todos os modelos estão satisfeitos com os serviços prestados. Justifique sua resposta.
 5. Os executivos da Toyord creem que seus clientes abastados são mais críticos e tendem à insatisfação com seus veículos. Para verificar se isso é verdade, construíram a tabela a seguir. Com base nela, constatarem a crença dos executivos. Calcule os percentuais que entender necessários. Justifique.

RENDA	OPINIÃO GERAL SOBRE O VEÍCULO				
	BASTANTE SATISFEITO	INSATISFEITO	MUITO INSATISFEITO	SATISFEITO	TOTAL
1,795 - 18,645	0	16	78	4	98
18,645 - 35,495	7	47	26	22	102
35,495 - 52,345	15	4	5	14	38
52,345 - 69,195	7	0	1	1	9
69,195 - 86,045	3	0	0	0	3
Total geral	32	67	110	41	250

Fonte: Elaborado pelo autor deste livro

4

UNIDADE

Análise Exploratória de Dados II



Ao finalizar esta Unidade, você deverá ser capaz de aprender mais uma maneira de descrever e de analisar um conjunto de dados referente a uma variável quantitativa (discreta ou contínua): por meio das medidas de síntese; e de perceber as medidas de posição e de dispersão que permitem sintetizar o comportamento da variável individualmente ou em função dos valores de outra variável.

Medidas de Posição ou de Tendência Central

Caro estudante!

Na Unidade 3, estudamos como fazer a descrição tabular e gráfica das variáveis, seja isoladamente ou relacionadas a outras e como interpretar os resultados obtidos. Além das técnicas, nos casos em que a variável sob análise for **quantitativa discreta** ou **quantitativa contínua**, há uma terceira forma de descrição: as **medidas de síntese** ou estatísticas. A sua utilização pode ser feita de forma a complementar as técnicas vistas na Unidade 3, ou como alternativa a elas.

As medidas de síntese subdividem-se em **medidas de posição (ou de tendência central)** e **medidas de dispersão**. Vamos estudar as medidas de posição: média, mediana, moda e quartis; e as medidas de dispersão: intervalo, variância, desvio padrão e coeficiente de variação percentual. Cada uma delas pode ser muito útil para caracterizar um conjunto de dados referente a uma variável quantitativa.

Tenha em mente a necessidade de o administrador conhecer as medidas de síntese para que possa realizar Análise Exploratória de Dados através delas. Veremos que são ferramentas que geram resultados objetivos, o que torna mais racional o processo de tomada de decisão.

As Medidas de Posição procuram caracterizar a tendência central do conjunto, um valor numérico que o “represente”. Esse valor pode ser calculado levando em conta todos os valores do conjunto ou apenas alguns valores ordenados. As medidas mais importantes são média, mediana, moda e quartis.

Média ($\bar{x}$)

A Média aqui citada é a **média aritmética simples**, a soma dos valores observados dividida pelo número desses valores. Seja um conjunto de **n** valores de uma variável quantitativa X, a média do conjunto será:

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

Onde x_i é um valor qualquer do conjunto, $\sum_{i=1}^n x_i$ é a soma dos valores do conjunto, e n é o tamanho do conjunto.

Vamos ver um exemplo, ao qual iremos recorrer algumas vezes.

O Quadro 10 se refere às notas finais de três turmas de estudantes.

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Quadro 10: Notas finais das turmas A, B, e C

Fonte: Elaborado pelo autor deste livro

Como o objetivo é calcular a média de cada turma, ao somar os valores teremos o mesmo resultado: 48. Como cada turma tem 8 alunos, as três turmas terão a mesma média: 6.

No exemplo que acabamos de ver, as três turmas têm a mesma média (6). Então, se apenas essa medida fosse utilizada para caracterizá-las poderíamos ter a impressão de que as três turmas têm desempenhos idênticos. Será? Observe atentamente o Quadro 10.

Repare que na primeira turma há realmente os dados distribuídos regularmente em torno da média, com a mesma variação tanto

No Microsoft Excel e no Br.Office Calc a média aritmética simples é implementada através da função MÉDIA().

Média aritmética simples – medida de posição que é resultado da divisão da soma de todos os elementos do conjunto divididos pela quantidade de elementos do conjunto. Conceitualmente, é o centro de massa do conjunto de dados. Fonte: Barbetta (2007).

abaixo quanto acima. Na segunda, vemos uma distorção maior, embora a maioria das notas seja alta, algumas notas baixas "puxam" a média para um valor menor. E, na terceira turma, há apenas uma nota baixa, mas seu valor é tal que consegue diminuir a média do conjunto.

Um dos problemas da utilização da média é que, por levar em conta todos os valores do conjunto, ela pode ser distorcida por **valores discrepantes** ("outliers") que nele existam. É importante então interpretar corretamente o valor da média.

O valor da média pode ser visto como o centro de massa de cada conjunto de dados, ou seja, o ponto de equilíbrio do conjunto: "se os valores do conjunto fossem pesos sobre uma tábua, a média seria a posição em que um suporte equilibra essa tábua".

Vamos ver como os valores do exemplo distribuem-se em um diagrama apropriado:

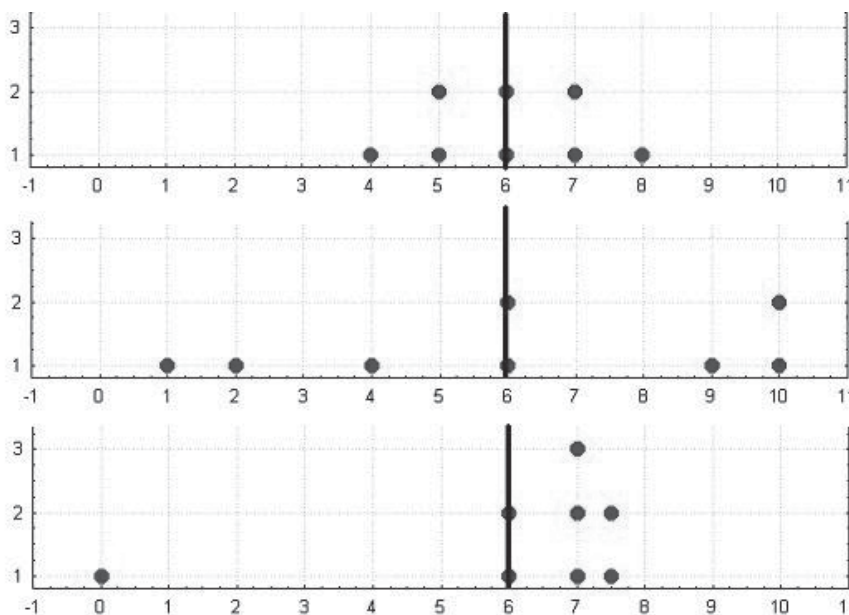


Figura 27: Interpretação do valor da média

Fonte: Elaborada pelo autor deste livro

A média dos três conjuntos é a mesma, mas observe as diferentes disposições dos dados. O primeiro grupo apresenta os dados distribuídos de forma **simétrica** em torno da média. No segundo grupo a distribuição já é mais irregular, com valores mais "distantes" na parte de baixo; e o terceiro grupo é claramente **assimétrica** em relação à média (que foi distorcida pelo valor discrepante 0). **Portanto, muito cuidado ao caracterizar um conjunto apenas por sua média.**

Valores discrepantes (outliers) – valores de uma variável quantitativa que se distanciam muito (para cima ou para baixo) da maioria das observações. Por exemplo, a renda de Bill Gates é um valor discrepante da variável renda de pessoas morando nos EUA. Fonte: Adaptado de Bussab e Morettin (2003).

Assimétrica – uma distribuição dos valores de uma variável quantitativa é dita assimétrica caso a média e a mediana sejam diferentes, indicando que os valores do conjunto se estendem mais, apresentam maior variabilidade, em uma direção do que na outra. Fonte: Barbetta (2007).

Essa foi a grande crítica nas décadas de 60 e 70 sobre as medições de nível de desenvolvimento. Era comum medir o nível de desenvolvimento de um país por sua *renda per capita* (PIB/número de habitantes), uma média que não revelava, porém, a concentração de renda do país, levando a conclusões errôneas sobre a qualidade de vida em muitos países.

Outro aspecto importante a ressaltar é que a média pode ser um valor que a variável não pode assumir. Isso é especialmente verdade para variáveis quantitativas discretas, resultantes de contagem, como número de filhos, quando a média pode assumir um valor “quebrado”: 4,3 filhos, por exemplo.

Rompemos com o mito de que “média é o valor mais provável do conjunto”, erro que é cometido quase que diariamente pela mídia, em vários países.

É extremamente comum calcular médias de variáveis quantitativas a partir de distribuições de frequências representadas em tabelas: simplesmente multiplica-se cada valor (ou o ponto médio da classe) pela frequência associada, somam-se os resultados e divide-se o somatório pelo número de observações do conjunto. Na realidade trata-se de uma média ponderada pelas frequências de ocorrência de cada valor da variável.

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

Onde k é o número de valores da variável discreta, ou o número de classes da variável agrupada; x_i é um valor qualquer da variável discreta, ou o ponto médio de uma classe qualquer; f_i é a frequência de um valor qualquer da variável discreta ou de uma classe qualquer; e n é o número total de elementos do conjunto.

Nesse segundo exemplo vamos calcular a média do número de pessoas usualmente transportadas no veículo, através da distribuição de frequências obtida no terceiro exemplo exposto na Unidade 3 (Quadro 11).

VALORES	FREQUÊNCIA	PERCENTUAL
1	19	7,60%
2	29	11,60%
3	43	17,20%
4	42	16,80%
5	57	22,80%
6	60	24,00%
Total	250	100%

Quadro 11: Número de pessoas usualmente transportadas no veículo

Fonte: Elaborado pelo autor deste livro

Precisamos multiplicar a coluna de valores x_i pela da frequência f_i , somar os resultados, e dividi-los por 250, que é o número de elementos do conjunto (n). Observe que a variável discreta pode assumir seis valores diferentes, logo $k = 6$. No Quadro 12, podemos observar o resultado:

VALORES X_i	FREQUÊNCIA f_i	$X_i \times f_i$
1	19	19
2	29	58
3	43	129
4	42	168
5	57	285
6	60	360
Total	250	1019

Quadro 12: Número de pessoas usualmente transportadas no veículo

Fonte: Elaborado pelo autor deste livro

Agora podemos calcular a média:

$$\bar{x} = \frac{\sum_{i=1}^k (x_i \times f_i)}{n} = \frac{\sum_{i=1}^6 (x_i \times f_i)}{250} = \frac{1019}{250} = 4,076 \text{ pessoas usualmente transportadas no veículo.}$$

Veja novamente a Figura 21 da Unidade 3 e observe como o va-

lor da média permite equilibrar os pesos e as frequências dos vários valores da variável.

No Exemplo 2, o resultado da média é um valor (4,076) que a variável número de pessoas usualmente transportadas não pode assumir. **Mas, trata-se do centro de massa do conjunto.**

Se quisermos calcular média aritmética simples a partir de uma distribuição de frequências para dados agrupados, devemos tomar cuidado. Os pontos médios das classes serão usados no lugar dos xi da expressão da média vista acima. Eles podem ou não ser bons representantes das classes (geralmente serão melhores representantes quanto maiores forem as frequências das classes), pois perdemos a informação sobre o conjunto original de dados ao agrupá-lo em classes. Sendo assim, as medidas calculadas a partir de uma distribuição de frequências para dados agrupados – não apenas a média aritmética simples, mas todas as outras – tornam-se meras estimativas dos valores reais.

Importante! Não calcule nenhuma medida estatística com base em uma distribuição de frequência para dados agrupados se você tiver acesso aos dados originais.

Além da média aritmética simples, outra medida de posição bastante usada é a mediana, que veremos a seguir.

Mediana (Md)

A **mediana** é o ponto que divide o conjunto em duas partes iguais: 50% dos dados têm valor menor do que a mediana e os outros 50% têm valor maior do que a mediana e é pouco afetada por eventuais **valores discrepantes** existentes no conjunto (que costumam distorcer substancialmente o valor da média).

A mediana de um conjunto de valores é o valor que ocupa a posição $(n + 1)/2$, quando os dados estão **ordenados** crescente ou decrescentemente. Se $(n + 1)/2$ for fracionário, toma-se como mediana a média dos dois valores que estão nas posições imediatamente

abaixo e acima de $(n + 1)/2$; onde n é o número de elementos do conjunto.

Nesse terceiro exemplo, vamos calcular a mediana para as notas das três turmas do Exemplo 1.

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Quadro 13: Notas finais das turmas A, B, e C

Fonte: Elaborado pelo autor deste livro

Posição mediana = $(n + 1)/2 = (8+1)/2 = 4,5^a$. Significa que o valor da mediana será calculado através da média entre os valores que estiverem na **4ª e na 5ª posição do conjunto**.

$$\text{Turma A: Md} = (6 + 6)/2 = 6$$

$$\text{Turma B: Md} = (6 + 6)/2 = 6$$

$$\text{Turma C: Md} = (7 + 7)/2 = 7$$

Observe que a mediana da Turma C é diferente, mais alta, refletindo melhor o conjunto de dados, uma vez que há apenas uma nota baixa. Perceba também que apenas os dois valores centrais foram considerados para obter a mediana, deixando o resultado “imune” aos valores discrepantes.

No exemplo 4, vamos Calcular a mediana para o grupo a seguir:

10 11 12 13 15 16 16 35 60

Posição mediana = $(n + 1)/2 = (9+1)/2 = 5^a$. Como o conjunto tem um número ímpar de valores, o valor da mediana será igual ao valor que estiver na quinta posição.

$$\text{Mediana} = 15$$

$$\text{Média} = 20,89$$

Observe que, no exemplo, média e mediana são diferentes, pois a média foi distorcida pelos valores mais altos 35 e 60, que constituem uma minoria. Nesse caso, a medida de posição que melhor representaria o conjunto seria a mediana. Se a média é diferente da mediana, a distribuição da variável quantitativa no conjunto de dados é dita **assimétrica**.

Por esse motivo, os dados precisam estar ordenados crescentemente.

No Microsoft Excel® e no Br.Office Calc ® a mediana é implementada através da função MED(), tal como explicado no texto *Como realizar análise exploratória de dados no Microsoft Excel®*.

Tal como a média, a mediana pode ser calculada a partir de uma tabela de frequências, com as mesmas ressalvas feitas para aquela medida. Os programas estatísticos e muitas planilhas eletrônicas dispõem de funções que calculam a mediana.

Moda (Mo)

A **moda** é o valor da variável que ocorre com maior frequência no conjunto. Pode, então, ser considerado o mais provável.

É a medida de posição de obtenção mais simples e também pode ser usada para variáveis qualitativas, pois apenas registra qual é o valor mais frequente, podendo esse valor ser tanto um número quanto uma categoria de uma variável nominal ou ordinal.

Um conjunto pode ter apenas uma Moda, várias Modas ou nenhuma Moda. Esse último caso geralmente ocorre com variáveis quantitativas contínuas.

A proposta, no exemplo 5, é encontrar a moda das notas das três turmas do Exemplo 1.

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Quadro 14: Notas finais das turmas A, B, e C
Fonte: Elaborado pelo autor deste livro

A turma A tem 3 modas: os valores 5, 6 e 7 ocorrem duas vezes cada. A turma B tem duas modas: os valores 6 e 10 ocorrem duas vezes cada. A turma C tem uma moda apenas: o valor 7 ocorre três vezes.

Quartis

Para alguns autores, os **quartis** não são medidas de posição, são separatrizes. Porém, como sua forma de cálculo é semelhante a da mediana, resolvemos incluí-los no tópico de Medidas de Posição. Os quartis são medidas que dividem o conjunto em quatro partes iguais.

O primeiro quartil ou **quartil inferior (Qi)** é o valor do conjunto que delimita os 25% dos menores valores: 25% dos valores são menores do que **Qi** e 75% são maiores do que **Qi**.

O segundo quartil ou **quartil do meio** é a própria mediana (**Md**), que separa os 50% menores dos 50% maiores valores.

O terceiro quartil ou **quartil superior (Qs)** é o valor que delimita os 25% dos maiores valores: 75% dos valores são menores do que **Qs** e 25% são maiores do que **Qs**.

Como são medidas baseadas na ordenação dos dados, é necessário calcular, inicialmente, as posições dos quartis.

$$\text{Posição do quartil inferior} = (n + 1)/4$$

$$\text{Posição do quartil superior} = [3 \times (n + 1)]/4$$

Onde **n** é o número total de elementos do conjunto.

Após calcular a posição, encontrar o elemento do conjunto que nela está localizado. O conjunto de dados precisa estar ordenado! Se o valor da posição for fracionário deve-se fazer a média entre os dois valores que estão nas posições imediatamente anterior, e imediatamente posterior à posição calculada. Se os dados estiverem dispostos em uma distribuição de frequências, utilizar o mesmo procedimento observando as frequências associadas a cada valor (variável discreta) ou ponto médio de classe.

No exemplo 6, iremos encontrar os quartis para a renda no conjunto de dados apresentados no Quadro 15:

Microsoft Excel® e no Br.Office Calc® os quartis são implementados através da função **QUARTIL(;1)** para quartil inferior, e **QUARTIL(;3)** para quartil superior.

VALORES									
4,695	5,750	7,575	12,960	13,805	14,000	15,820	18,275	18,985	18,985
19,595	19,720	20,600	22,855	22,990	23,685	24,400	24,400	24,685	24,980
24,980	26,775	27,085	27,240	28,340	31,480	40,050	43,150	47,075	

Quadro 15: Renda em salários mínimos

Fonte: Elaborado pelo autor deste livro

Há 29 elementos no conjunto que já estão ordenados crescentemente. Podemos calcular as posições dos quartis.

$$\text{Posição do quartil inferior} = (n + 1)/4 = (29 + 1)/4 = 7,5^{\text{a}}.$$

$$\text{Posição do quartil superior} = [3 \times (n+1)]/4 = [3 \times (29 + 1)]/4 = 22,5^{\text{a}}.$$

Para encontrar o quartil inferior precisamos calcular a média dos valores que estão na 7ª e 8ª posição do conjunto: no caso, 15,820 e 18,275, resultando:

$$Q_i = (15,820 + 18,275)/2 = 17,0475$$

Imagine que fosse um grande conjunto de dados, referente a salários de uma população: apenas 25% dos pesquisados teriam renda **abaixo** de 17,0475 salários mínimos (ou R\$ 6.478,05 pelo salário mínimo de maio de 2007). Com base nisso, poderíamos ter uma ideia do nível de renda daquela população.

Para encontrar o quartil superior, precisamos calcular a média dos valores que estão na 22ª e 23ª posição do conjunto: no caso, 15,820 e 18,275 para obtermos:

$$Q_s = (26,775 + 27,085)/2 = 26,93$$

Novamente, imagine que fosse um grande conjunto de dados referente a salários de uma população: apenas 25% dos pesquisados teriam renda acima de 26,93 salários mínimos (ou R\$ 10.233,40 pelo salário mínimo de maio de 2007).

Com todas as medidas de posição citadas, já é possível obter um retrato razoável do comportamento da variável. Mas as medidas de posição são insuficientes para caracterizar adequadamente um conjunto de dados. É preciso calcular também medidas de dispersão.

Medidas de Dispersão ou de Variabilidade

O objetivo das medidas de dispersão é medir quão próximos uns dos outros estão os valores de um grupo (algumas mensuram a dispersão dos dados em torno de uma medida de posição). Com isso, é obtido um valor numérico que sintetiza a variabilidade.

Vamos estudar o intervalo, a variância, o desvio padrão e o coeficiente de variação percentual.

Intervalo

O intervalo é a medida mais simples de dispersão. Consiste em identificar os valores extremos do conjunto (mínimo e máximo), podendo ser expresso:

- pela diferença entre o valor máximo e o mínimo; e
- pela simples identificação dos valores.

O intervalo é muito útil para nos dar uma ideia da variabilidade geral do conjunto de dados. Alguém que calculasse o intervalo da variável renda mensal familiar no Brasil provavelmente ficaria abismado pela gigantesca diferença entre o valor mais baixo e o mais alto. Se essa mesma pessoa fizesse o mesmo cálculo na Noruega, a diferença não seria tão grande.

No exemplo 7 vamos obter o Intervalo para os conjuntos de notas das duas turmas apresentadas no Quadro 16:

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	4 4 4,2 4,3 4,5 5 5 8

Quadro 16: Notas das turmas A e B
Fonte: Elaborado pelo autor deste livro

Medidas de dispersão – medidas numéricas que objetivam avaliar a variabilidade do conjunto de dados, sintetizando-a em um número. Fonte: Elaborado pelo autor deste livro.

O intervalo será o mesmo para ambas as turmas: [4,8] ou 4.

Observe que no Exemplo 7 as duas turmas apresentam o mesmo intervalo (4). Mas, observando os dados, percebemos facilmente que a dispersão dos dados tem comportamento diferente nas duas turmas, e essa é a principal desvantagem do uso do intervalo como medida de dispersão.

Se colocarmos os dados do Exemplo 7 em um diagrama apropriado:

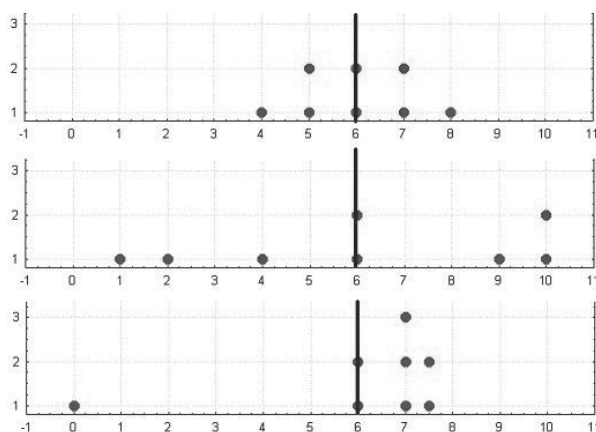


Figura 28: Desvantagem do uso do intervalo como medida de dispersão

Fonte: Elaborada pelo autor deste livro

No Microsoft Excel® e no Br.Office Calc® podemos obter o Intervalo através das funções MÁXIMO () e MÍNIMO ().

Observamos claramente que os dados da turma A apresentam uma dispersão bem mais uniforme do que os da turma B, embora ambos os conjuntos tenham o mesmo intervalo. O intervalo não permite ter ideia de como os dados estão distribuídos entre os extremos (não permite identificar que o valor 8 na turma B é um valor discrepante).

Torna-se necessário obter outras medidas de dispersão, capazes de levar em conta a variabilidade entre os extremos do conjunto, o que nos leva a estudar variância e desvio padrão.

Variância (s^2)

A variância é uma das medidas de dispersão mais importantes. É a média aritmética dos quadrados dos desvios de cada valor em relação à média: proporciona uma mensuração da dispersão dos dados em torno da média.

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

Onde x_i é um valor qualquer do conjunto; $\bar{x}$ é a média do conjunto; e n é o número de elementos do conjunto. Se os dados referem-se a uma **população**, usa-se **n** no denominador da expressão.

Você sabe por que é preciso elevar os desvios ao quadrado para avaliar a dispersão? Não podemos apenas somar os desvios dos valores em relação à média do conjunto? Deixamos como exercício para você os cálculos dos desvios (diferença entre cada valor e a média) para as notas das três turmas descritas no Quadro 10, do Exemplo 1. Após calcular os desvios, some-os e veja os resultados. Lembre-se de que a média é o centro de massa do conjunto.

A razão dessa distinção será explicada mais adiante na Unidade 7. Podemos adiantar que

a utilização de $n - 1$ no denominador é indispensável para que a variância da variável na amostra possa ser um bom estimador da variância da variável na população.

A unidade da variância é o quadrado da unidade dos dados e, portanto, o quadrado da unidade da média, causando dificuldades para avaliar a dispersão: se, por exemplo, temos a variável peso com média de 75 kg em um conjunto e ao calcular a variância obtemos 12 kg², a avaliação da dispersão torna-se difícil. Não obstante, a variância e a média são as medidas geralmente usadas para caracterizar as distribuições probabilísticas (que serão vistas adiante, na Unidade 6).

O que podemos afirmar, porém, é que quanto maior a variância mais dispersos os dados estão em torno da média (maior a dispersão do conjunto).

No Microsoft Excel® e no Br.Office Calc® a variância populacional é obtida por meio da função `VARP()`, e a variância amostral através da função `VAR()`.

Para fins de Análise Exploratória de Dados, caracterizar a dispersão por intermédio da variância não é muito adequado. Costumamos usar a raiz quadrada positiva da variância, o desvio padrão. Vamos ver mais sobre isso? Continuemos os estudos!

Desvio Padrão (s)

É a raiz quadrada positiva da variância, que apresenta a mesma unidade dos dados e da média, e permite avaliar melhor a dispersão.

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

As mesmas observações, sobre população e amostra, feitas para a variância são válidas para o desvio padrão. É prática comum, ao se resumir através de várias medidas de síntese um conjunto de dados referente a uma variável quantitativa, apresentar apenas a média e o desvio padrão desse conjunto para que seja possível uma ideia do valor típico e da distribuição dos dados em torno dele.

Deixamos como exercício para você a função de elevar ao quadrado os desvios obtidos com os dados das turmas, somá-los e dividi-los por 7 (suponha que se trate de uma amostra). Assim, você obterá os desvios padrões das notas das turmas. Use como referência o Quadro 10, Exemplo 1.

O desvio padrão pode assumir valores menores do que a média, da mesma ordem de grandeza da média, ou até mesmo maiores do que a média. Obviamente, se todos os valores forem iguais não haverá variabilidade e o desvio padrão será igual a zero.

A fórmula acima costuma levar a consideráveis erros de arredondamento, basicamente porque exige o cálculo prévio da média. Se o valor desta for uma dízima, um arredondamento terá que ser feito, causando um pequeno erro que será propagado pelas várias operações de subtração (de cada valor em relação à média) e potenciação (elevação ao quadrado da diferença entre cada valor e a média). Assim, a fórmula é modificada para reduzir o erro de arredondamento apenas ao resultado final:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i^2) - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}{n-1}} \quad (\text{amostra})$$

Dessa forma, cada valor (x_i) do conjunto é elevado ao quadrado e somam-se todos os resultados obtendo-se $\sum_{i=1}^n x_i^2$. Somam-se tam-

bém todos os valores do conjunto para obter $\sum_{i=1}^n x_i$, somatório que será elevado ao quadrado. Os somatórios e o valor de **n** (número de elementos no conjunto) são substituídos na fórmula para a obtenção dos resultados.

Tal como no caso da média, pode haver interesse em calcular o desvio padrão de variáveis quantitativas a partir de distribuições de frequências representadas em tabelas e, ainda, os valores da variável (ou os pontos médios das classes) e os quadrados desses valores serão multiplicados por suas respectivas frequências:

$$s = \sqrt{\frac{\sum_{i=1}^k (x_i^2 \times f_i) - \frac{\left(\sum_{i=1}^k x_i \times f_i\right)^2}{n}}{n-1}} \quad (\text{amostra})$$

Onde x_i é o valor da variável ou ponto médio da classe; f_i a frequência associada, k é o número de valores da variável discreta (ou o número de classes da variável agrupada), e n é o número de elementos do conjunto.

Veremos, neste oitavo exemplo, como calcular o desvio padrão da renda para os dados do Exemplo 6.

VALORES									
4,695	5,750	7,575	12,960	13,805	14,000	15,820	18,275	18,985	18,985
19,595	19,720	20,600	22,855	22,990	23,685	24,400	24,400	24,685	24,980
24,980	26,775	27,085	27,240	28,340	31,480	40,050	43,150	47,075	

Quadro 17: Renda em salários mínimos

Fonte: Elaborado pelo autor deste livro

Há 29 elementos no conjunto, $n = 29$.

Somando os valores, temos: $\sum_{i=1}^n x_i = \sum_{i=1}^{29} x_i = 654,935$

Elevando cada valor ao quadrado e somando-os, obtemos:

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^{29} x_i^2 = 17497,919125$$

Agora basta substituir os somatórios na expressão e calcular o desvio padrão, supondo que se trate de uma amostra:

É desse modo que os programas computacionais calculam o desvio padrão.

No Microsoft Excel® e no Br.Office Calc® podemos obter o desvio padrão populacional através da função DESVPADP() e amostral por meio da função DESVPAD().

$$s = \sqrt{\frac{\sum_{i=1}^{29} (x_i^2) - \left[\frac{\left(\sum_{i=1}^{29} x_i \right)^2}{29} \right]}{29-1}} = \sqrt{\frac{17497,919125 - \left[\frac{(654,935)^2}{29} \right]}{29-1}} =$$

$$s = \sqrt{\frac{17497,919125 - 14791,02946}{28}}$$

$s \cong 9,83$ salários mínimos

Se calcularmos a média, obteremos 22,584 salários mínimos. Observe que o desvio padrão é menor do que a média, não chega à metade. Com base nisso, poderíamos avaliar a variabilidade do conjunto.

Quanto menor o desvio padrão, mais os dados estão concentrados em torno da média. Pensando nisso, alguém teve a ideia de criar uma medida de dispersão que relacionasse média e desvio padrão, o coeficiente de variação percentual, que veremos a seguir.

Coeficiente de Variação Percentual (c.v.%)

O coeficiente de variação percentual é uma medida de dispersão relativa, pois permite comparar a dispersão de diferentes distribuições com diferentes médias e desvios padrões.

$$\text{c.v. \%} = \frac{s}{\bar{x}} \times 100\%$$

Onde s é o desvio padrão da variável no conjunto de dados e $\bar{x}$ é a média da variável no mesmo conjunto.

Quanto menor o coeficiente de variação percentual, mais os dados estão concentrados em torno da média, pois o desvio padrão é pequeno em relação à média.

No exemplo a seguir, vamos calcular o coeficiente de variação percentual para as notas das turmas do Exemplo 1 e indicar qual das três apresenta as notas mais homogêneas.

TURMA	VALORES
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Quadro 18: Notas finais das turmas A, B, e C

Fonte: Elaborado pelo autor deste livro

Coeficiente de variação percentual – resultado da divisão do desvio padrão pela média do conjunto multiplicado por 100, que permite avaliar quanto o desvio padrão representa da média. Fonte: Barbetta, Reis e Bornia (2008); Anderson, Sweeney e Williams (2007).

Para a turma A: $\bar{x} = 6$ $s = 1,31$ $\text{c.v.}\% = (1,31/6) \times 100 = 21,82\%$

Para a turma B: $\bar{x} = 6$ $s = 3,51$ $\text{c.v.}\% = (3,51/6) \times 100 = 58,42\%$

Para a turma C: $\bar{x} = 6$ $s = 2,49$ $\text{c.v.}\% = (2,49/6) \times 100 = 41,55\%$

A turma mais homogênea é a A, pois apresenta o menor coeficiente de variação das três. Isso era esperado, uma vez que as notas da turma A estão distribuídas mais regularmente do que as das outras.

No caso apresentado, a comparação ficou ainda mais simples, uma vez que as médias dos grupos eram iguais. Bastaria, no entanto, avaliar apenas os desvios padrões dos grupos, mas para comparar a dispersão de distribuições com médias diferentes é imprescindível a utilização do coeficiente de variação percentual.

Você deve se perguntar: por que é tão importante calcular a média e o desvio padrão dos valores de uma variável registrados em um conjunto de dados? A resposta é que talvez a mediana seja a melhor medida de posição, já que os quartis permitem uma boa ideia da dispersão. Contudo, há um teorema que permite, a partir da média e do desvio padrão, a obtenção de estimativas dos extremos do conjunto, especialmente quando se trata de uma amostra. Trata-se do **teorema de Chebyshev**, também chamado de Desigualdade de Chebyshev.

Há um pequeno texto sobre o Teorema de Chebyshev no ambiente virtual.

Cálculo de Medidas de Síntese de Uma Variável em Função dos Valores de Outra

Na Unidade 3, estudamos como analisar em conjunto uma variável quantitativa e outra qualitativa. Naquela ocasião, mostramos como os dados da variável quantitativa poderiam ser avaliados em função dos valores da variável qualitativa, uma vez que esta costuma ter menos opções, possibilitando resumir mais o conjunto.

Recomendamos a você que, novamente, sirva-se do oitavo exemplo exibido nesta Unidade. Você verá que construímos distribuições de frequências agrupadas em classes para a variável renda (quantitativa) em função dos valores da variável modelo (qualitativa). Poderíamos fazer o mesmo com as medidas de síntese? Vejamos o exemplo seguinte.

Para a mesma situação dos Exemplos 1 (Unidade 3) e 7 (Unidade 4). Gostaríamos de avaliar se existe alguma relação entre a renda do consumidor e o modelo adquirido. Esperamos que exista tal relação, pois os modelos Chiconaultla e DeltaForce3 são os mais baratos, e o sofisticado LuxuriousCar é o mais caro de todos.

Através do Microsoft Excel® e do Br.Office Calc®, podemos calcular várias medidas de síntese da variável renda em função dos modelos de veículos. O Excel® permite obter as seguintes medidas em função dos valores de outra variável: média, desvio padrão (amostral e populacional), variância (amostral e populacional), mínimo e máximo (infelizmente, não permite cálculo de mediana ou quartis). O Calc permite obter as mesmas medidas, mas para cada uma delas faz-se necessário acionar o assistente de dados, enquanto no Excel é possível agrupá-las em uma única tabela. Ao realizar esse procedimento, empregando-se os dados do arquivo AmostraToyord.xls, obtemos:

MODELO	MEDIDA	VALOR
Chiconaultla	Frequência	81
	Mínimo	1,795
	Máximo	40,160
	Média	12,704
	Desvio padrão (amostral)	6,038
DeltaForce3	Frequência	56
	Mínimo	10,820
	Máximo	48,220
	Média	22,063
	Desvio padrão (amostral)	6,956

Quadro 19: Medidas de síntese de renda por modelo

Fonte: Elaborado pelo autor deste livro

MODELO	MEDIDA	VALOR
LuxuriousCar	Frequência	29
	Mínimo	29,800
	Máximo	86,015
	Média	50,932
	Desvio padrão (amostral)	14,922
SpaceShuttle	Frequência	42
	Mínimo	18,865
	Máximo	47,300
	Média	33,050
	Desvio padrão (amostral)	7,620
Valentiniana	Frequência	41
	Mínimo	13,055
	Máximo	65,390
	Média	27,353
	Desvio padrão (amostral)	8,383
Frequência Total		249
Mínimo Total		1,795
Máximo Total		86,015
Média Total		25,105
Desvio padrão (amostral) Total		14,505

Quadro 19: Medidas de síntese de renda por modelo

Fonte: Elaborado pelo autor deste livro

Se analisarmos as medidas de renda para os cinco modelos, vamos identificar alguns aspectos interessantes:

- os mínimos de Chiconaultla e DeltaForce3 são efetivamente menores do que os dos outros modelos (o mínimo de Chiconaultla é o menor do conjunto todo);
- o mínimo de LuxuriousCar é o maior de todos, e seu máximo também (sendo o valor máximo do conjunto todo);
- quanto às médias, podemos observar um comportamento na seguinte ordem crescente: Chiconaultla, DeltaForce3, Valentiniana, SpaceShuttle e LuxuriousCar; e
- a média de renda dos clientes do LuxuriousCar é quase quatro vezes maior do que as dos compradores do Chiconaultla.

Portanto, a relação entre renda e modelo parece realmente existir.

Agora, devemos avaliar a dispersão da renda em função dos modelos. Como as médias são diferentes, é recomendável calcular os coeficientes de variação percentual, mostrados no Quadro 20.

MODELO	MEDIDA	VALOR
Chiconaultla	Coeficiente de Variação Percentual	47,526%
DeltaForce3	Coeficiente de Variação Percentual	31,528%
LuxuriousCar	Coeficiente de Variação Percentual	29,298%
SpaceShuttle	Coeficiente de Variação Percentual	23,054%
Valentiniana	Coeficiente de Variação Percentual	30,646%
Coeficiente de Variação Percentual Total		57,777%

Quadro 20: Coeficientes de variação percentual de renda por modelo

Fonte: Elaborado pelo autor deste livro

Aparentemente, a relação existente entre renda média e os modelos não se reproduz completamente no que tange à dispersão. Embora o Chiconaultla (modelo mais barato, cujos compradores têm a média mais baixa de renda) tenha o maior coeficiente de variação percentual (47,526%), o modelo mais sofisticado: o LuxuriousCar, cujos compradores têm a média mais alta, não apresenta o menor coeficiente de variação percentual. O modelo, que os compradores possuem a renda mais concentrada em torno da média, é o SpaceShuttle, cujo coeficiente de variação percentual vale 23,054%. Podemos concluir que, embora o Chiconaultla seja um modelo mais “simples”, teoricamente visando ao público de menor renda, ele também é adquirido por compradores mais abastados. Já o SpaceShuttle tem compradores de nível mais elevado (segunda maior média de renda), com pouca variação entre eles.

Utilizando-se um *software* estatístico, podemos calcular outras medidas além das mostradas nos Quadros anteriores. No nosso caso, com o Statsoft Statistica 6.0®, podemos obter:

MODELO	MEDIDAS							
	$\bar{X}$	FREQ.	S	MIN.	MAX.	QI	MD	Qs
Deltaforce3	22,064	56	6,956	10,82	48,22	16,575	21,378	26,392
SpaceShuttle	33,05	42	7,62	18,865	47,3	26,62	33,85	39,65
Valentiniana	27,353	41	8,383	13,055	65,39	23,685	25,715	30,13
Chiconaultla	12,705	81	6,038	1,795	40,16	8,88	12,245	15,4
LuxuriousCar	50,932	29	14,922	29,800	86,015	41,89	47,525	58,92
Total	25,105	249	14,505	1,795	86,015	14,095	23,545	32,17

Quadro 21: Medidas de síntese de renda por modelo

Fonte: Elaborado pelo autor deste livro

Observe que as medianas, quartis inferiores e superiores comportam-se de forma semelhante às médias. A propósito, médias e medianas são próximas, o que indicaria simetria das distribuições das rendas para todos os modelos.

Saiba mais...

Sobre medidas de síntese, assimetria, diagramas em caixa e outros aspectos, procure em: BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Ed. da UFSC, 2007, Capítulo 7.

Sobre outros tipos de médias (harmônica, geométrica), SPIEGEL, Murray R. *Estatística*. 3. ed. São Paulo: Makron Books, 1993, Capítulo 3.

Sobre outros aspectos de Análise Exploratória de Dados com medidas de síntese, teorema de Chebyshev e assimetria, ANDERSON, David R.; SWEENEY, Dennis J.; WILLIAMS, Thomas A. *Estatística Aplicada à Administração e Economia*. 2. ed. São Paulo: Thomson Learning, 2007, Capítulo 3.

Sobre Análise Exploratória de Dados utilizando o Excel, LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português*. 5. ed. Rio de Janeiro: LTC, 2005.

Para saber como realizar as análises descritas nesta Unidade mediante o emprego do Microsoft Excel®, consulte *Como realizar Análise Exploratória de Dados no Microsoft Excel®*, disponível no Ambiente Virtual de Ensino-Aprendizagem, assim como o arquivo de dados usado nos exemplos apresentados.

Sobre como realizar as análises descritas nesta Unidade através do Br.Office Calc®, consulte *Como realizar análise exploratória de dados com o Br.Office Calc®*, disponível no Ambiente Virtual assim como o arquivo de dados usado nos exemplos apresentados.



A Figura 29, a seguir, apresenta o resumo desta Unidade:

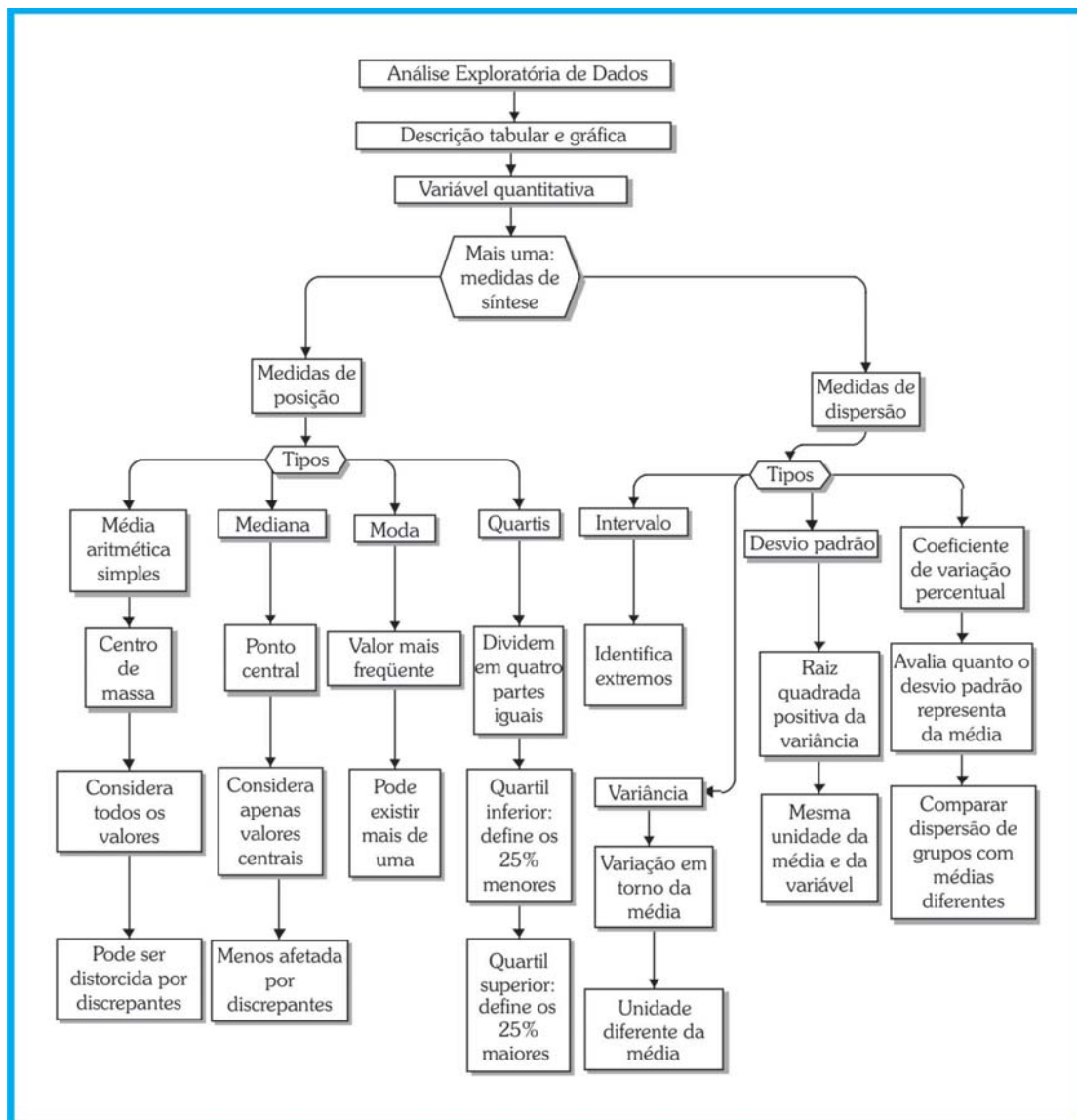


Figura 29: Resumo da Unidade 4

Fonte: Elaborada pelo autor deste livro

Finalizamos nosso estudo acerca da Análise Exploratória de Dados. É extremamente importante que você realize todos os exercícios e, em contato com a tutoria, esclareça suas dúvidas, pois não há outra forma de aprender a não ser praticando. Na Unidade 5, veremos os conceitos de Probabilidade, indispensáveis à compreensão do processo de inferência (generalização) estatística. Vamos em frente!



Atividades de aprendizagem

As atividades devem ser feitas com o emprego do Microsoft Excel® ou do Br.Office Calc®, por meio do arquivo AmostraToyord.xls que está no Ambiente Virtual de Ensino-Aprendizagem.

1. A variável anos de remodelação dos veículos (na percepção do cliente) está representada na distribuição de frequências expressa no Quadro a seguir:

ANOS DE REMODELAÇÃO	FREQUÊNCIAS
0	2
1	57
2	123
3	59
4	9
Total	250

Fonte: Elaborado pelo autor deste livro

- a) Calcule a média, mediana, moda e quartis da variável anos de remodelação.
- b) A direção da Toyord acredita que se uma parcela considerável dos clientes perceber que seus modelos são atualizados (foram remodelados há no máximo 2 anos) o

design e o marketing dos veículos estão coerentes. Com base nos resultados da letra **a**, os dados mostram isso? Justifique.

- c) Calcule o intervalo, desvio padrão e coeficiente de variação percentual da variável anos de remodelação.
- d) Com base nos resultados dos itens **a** e **c**, você considera que os dados estão fortemente concentrados em torno da média? Justifique.

2. Na questão 5 das atividades de aprendizagem da Unidade 3, foi dito: “os executivos da Toyord creem que seus clientes mais abastados são mais críticos e tendem à insatisfação com seus veículos”. Naquela questão, foi construída uma distribuição de frequências conjunta, relacionando a renda agrupada em classes com a opinião geral dos clientes sobre seus veículos, para verificar se os executivos estavam certos. Agora, analise a renda dos clientes (variável quantitativa) em função da opinião geral dos clientes (através do Microsoft Excel® ou do Br.Office Calc®), e calcule medidas de síntese de renda em função das opiniões.

- a) Com base nos resultados, os executivos estão certos? Justifique.
- b) Compare com as conclusões que você obteve na questão 5 da Unidade 3.

5

UNIDADE

Conceitos Básicos da Probabilidade



Ao finalizar esta Unidade, você deverá ser capaz de compreender os conceitos de Probabilidade; de perceber a importância do uso do raciocínio probabilístico; e de fazer uso desse raciocínio, que é fundamental para o administrador na tomada de decisões em ambiente de incerteza.

Probabilidade: conceitos gerais

Caro estudante,

Nesta Unidade, vamos estudar os conceitos básicos de Probabilidade, tais como: Experimento Aleatório; Espaço Amostral e Eventos; Axiomas e Propriedades; Probabilidade Condicional; e Independência Estatística. Nos EUA há uma anedota que diz: “as únicas coisas que são certas são a morte e os impostos”. Em outras palavras, estamos imersos na incerteza! Os administradores diariamente precisam tomar decisões. Muitas delas extremamente sérias em cenário de grande incerteza:

- lançamos ou não um novo modelo de automóvel?
- Convertemos nossos fornos de óleo combustível para gás natural?
- Qual será a reação do nosso público às mudanças na grade de programação?

Por que há incerteza? Porque a **variabilidade** inerente à natureza impede a compreensão completa dos fenômenos naturais e humanos. Mas, os seres humanos precisam tomar decisões! Assim, é necessário levar a incerteza em conta no processo. Alguns apelam para a sabedoria popular, outros para o místico. Os administradores precisam tomar decisões de forma objetiva; surge, então, a Probabilidade como uma das abordagens de tratamento da incerteza.

A utilização de métodos probabilísticos proporciona grande auxílio na tomada de decisões, pois permite avaliar riscos e otimizar recursos (sempre escassos) para as situações mais prováveis. Você está convidado a conhecer mais sobre esse tema nesta Unidade.

Variabilidade – diferenças encontradas por sucessivas medições realizadas em pessoas, animais ou objetos, em tempos ou situações diferentes. Fonte: Montgomery (2004).

Probabilidade – descrição quantitativa da certeza de ocorrência de um evento, geralmente representada por um número real entre 0 e 1 (0% e 100%). Fonte: Elaborado pelo autor deste livro.

Modelo Probabilístico – modelo matemático para descrever a certeza de ocorrência de eventos, onde são definidos os eventos possíveis e uma regra de ocorrência para calcular quão provável é cada evento ou conjunto de eventos. Fonte: Barbetta (2007).

Nas Unidades 3 e 4, utilizamos raciocínio predominantemente indutivo. Os dados foram coletados e, através da sua organização em distribuições de frequências e medidas de síntese, foi possível caracterizar a variabilidade do fenômeno observado e elaborar hipóteses ou conjecturas a respeito.

Suponha que estejamos estudando o percentual de meninos e de meninas nascidos em um estado brasileiro. Consultando dados do IBGE provenientes de censos e de levantamentos anteriores (portanto distribuições de frequências da variável qualitativa sexo dos recém-nascidos), há interesse em prever o percentual de nascimentos no ano de 2009: em suma, será usado o raciocínio dedutivo, a partir de algumas suposições sobre o problema (a definição dos resultados possíveis, os percentuais registrados em anos anteriores) na tentativa de se obter novos valores.

Se o percentual de meninos no passado foi de 49%, a pergunta é: qual será o percentual de meninos nascidos no ano de 2009? É possível que seja um valor próximo de 49%, talvez pouco acima ou pouco abaixo, mas não há como responder com certeza absoluta, pela simples razão que o fenômeno ainda não ocorreu, e que sua natureza é aleatória, ou seja, é possível identificar quais serão os resultados possíveis (menino ou menina), e há certa regularidade nos percentuais de nascimentos (verificados anteriormente), mas não é possível responder qual será o resultado exato antes do fenômeno ocorrer.

A regularidade citada (que foi observada para grande número de nascimentos) permite que seja calculado o grau de certeza ou confiabilidade da previsão feita, que recebe o nome de Probabilidade. Haverá grande probabilidade de que realmente o percentual de meninos nascidos em 2009 seja de 49%, mas nada impede que um valor diferente venha a ocorrer.

Montamos um Modelo Probabilístico para o problema em questão:

- foram definidos todos os resultados possíveis para o fenômeno (experimento); e
- definiu-se uma **regra** que permite dizer quão provável será cada resultado ou grupo de resultados.

O Modelo Probabilístico permite expressar o grau de incertezas através de probabilidades.

A regra citada foi definida a partir de observações anteriores do fenômeno, mas também poderia ser formulada com base em considerações teóricas. Por exemplo, se há interesse em estudar as proporções de ocorrências das faces de um dado, de um dado não é viciado, espera-se que cada face ocorra em $1/6$ do total de lançamentos: se o dado for lançado numerosas vezes isso provavelmente ocorrerá, mas um resultado diferente poderia ser obtido sem, no entanto, significar a viciação do dado, principalmente se forem feitos **poucos lançamentos**.

Nesse ponto, é importante ressaltar que os modelos probabilísticos não têm razão de ser para fenômenos (experimentos) não aleatórios (**determinísticos**): aqueles em que, adotando-se teorias e fórmulas apropriadas, podemos prever exatamente qual será o seu re-sultado antes do fenômeno ocorrer. Por exemplo: o lançamento de uma pedra de 5 kg de uma altura de 10 metros, se interessados em cronometrar o tempo que ela leva para atingir o chão. Conhecendo-se o peso da pedra, a altura do lançamento, a aceleração da gravidade e as leis da Física, é perfeitamente possível calcular o tempo de queda, sem a necessidade de se realizar o experimento.

Para prosseguirmos, iremos conhecer algumas definições importantes para o estudo dos modelos probabilísticos. Precisamos definir exatamente as condições em que devemos usar tais modelos, o que exige compreensão acerca de experimento aleatório, de espaço amostral e de eventos. Vamos lá?

Para construir ou utilizar modelos probabilísticos, é necessário que haja grande número de realizações do fenômeno (experimento) para que uma regularidade possa ser verificada: é a Lei dos Grandes Números. No início do Século XX, o estatístico inglês Karl Pearson lançou uma moeda não viciada 24.000 vezes (!) para verificar a validade dessa lei: obteve 12.012 caras, praticamente o valor esperado (12.000, 50%).

Definições Prévias

Experimento Aleatório

Experimento Aleatório é um processo de obtenção de um resultado ou medida que apresenta as seguintes características:

- não podemos afirmar, antes de efetivar o experimento, qual será o resultado de uma realização, mas é possível determinar o conjunto de resultados possíveis.

- quando é realizado grande número de vezes (replicado), apresentará regularidade que permitirá construir um modelo probabilístico para analisar o experimento.

São experimentos aleatórios:

- O lançamento de um dado com uma das faces voltadas para cima; não sabemos exatamente qual face vai ocorrer, apenas que será uma das seis, e que se o dado não for viciado e o lançamento imparcial, todas as faces têm a mesma chance de ocorrer.
- A observação dos diâmetros, em mm, de eixos produzidos em uma metalúrgica; sabemos que as medidas devem estar próximas de um valor nominal, mas não sabemos exatamente qual é o diâmetro de cada eixo antes de efetuar as mensurações.
- O número de mensagens que são transmitidas corretamente por dia em uma rede de computadores; sabemos que o mínimo possível é zero, mas não sabemos o número máximo de mensagens que serão transmitidas.

Todo experimento aleatório terá alguns resultados possíveis, que constituirão o Espaço Amostral.

Espaço Amostral (S ou Ω)

Espaço Amostral é o conjunto de **todos** os resultados possíveis de um experimento aleatório. “Para cada experimento aleatório haverá um espaço amostral único Ω associado a ele”. Neste primeiro exemplo veremos alguns experimentos aleatórios com os respectivos espaços amostrais:

- Lançamento de um dado e a observação da face voltada para cima: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Retirada de uma carta de baralho comum (52 cartas) e a observação do naipe: $\Omega = \{\text{copas, espadas, ouros, paus}\}$.
- O número de mensagens que são transmitidas corretamente por dia em uma rede de computadores: $\Omega = \{0, 1, 2, 3, \dots\}$.
- A observação do diâmetro, em mm, de um eixo produzido em uma metalúrgica: $\Omega = \{D, \text{ tal que } D > 0\}$.

Note que não há um limite superior conhecido, mas somente é possível a ocorrência de valores inteiros.



- e) As vendas mensais, em unidades, de determinado modelo de veículo: $\Omega = \{0, 1, \dots\}$

O espaço amostral pode ser:

- **finito**, formado por um número limitado de resultados possíveis, como nos casos a e b;
- **infinito numerável**, formado por um número infinito de resultados, mas que podem ser listados, como nos casos c ou e;
- **infinito**, formado por intervalos de números reais, como no caso d.

Um espaço amostral é dito discreto quando ele for finito ou infinito enumerável; e dito contínuo quando é infinito, formado por intervalos de números reais.

A construção do modelo probabilístico dependerá do tipo de espaço amostral, como veremos adiante.

Eventos

Eventos são quaisquer subconjuntos do espaço amostral. Um evento pode conter um ou mais resultados; se pelo menos um dos resultados ocorrer, o evento acontece. Geralmente, há interesse em calcular a probabilidade de que determinado evento venha a ocorrer e, tal evento, pode ser definido de forma verbal, **precisando ser “traduzido” para as definições da Teoria de Conjuntos, que veremos a seguir.**

Seja o Experimento Aleatório o lançamento de um dado não vi-ciado e a observação da face voltada para cima: o seu espaço amostral será $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Definindo três eventos:

$$E_1 = \{2, 4, 6\}$$

$$E_2 = \{3, 4, 5, 6\}$$

$$E_3 = \{1, 3\}$$

serão apresentadas as definições de **Evento União**, **Evento Intersecção**, **Eventos Mutuamente Exclusivos** e **Evento Complementar**.

Não há um limite superior e, teoricamente, pode haver uma infinidade de valores.

Embora nem todos os autores concordem com essa abordagem, ela auxilia bastante na compreensão dos conceitos.

Evento **União** de E_1 com E_2 ($E_1 \cup E_2$): evento que se dá quando E_1 ou E_2 ocorrem.

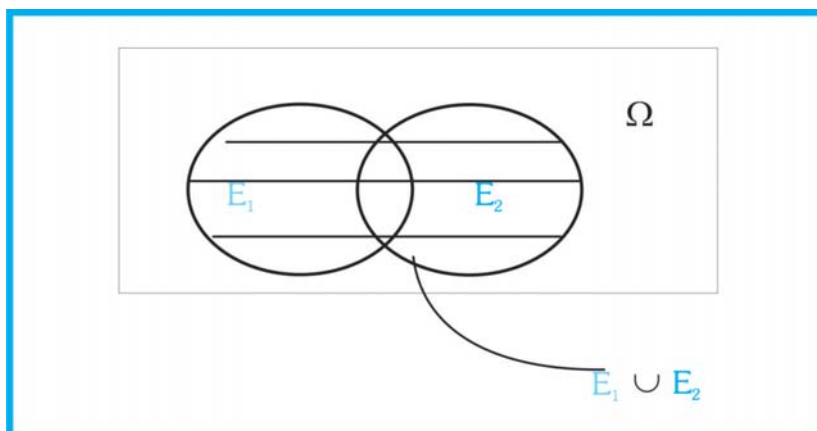


Figura 30: Evento União

Fonte: Elaborada pelo autor deste livro

$$E_1 \cup E_2 = \{2, 3, 4, 5, 6\}.$$

Composto por todos os resultados que pertencem a um **ou** ao outro, **ou** a ambos.

Evento **Intersecção** de E_1 com E_2 ($E_1 \cap E_2$): evento que ocorre se E_1 e E_2 ocorrem simultaneamente.

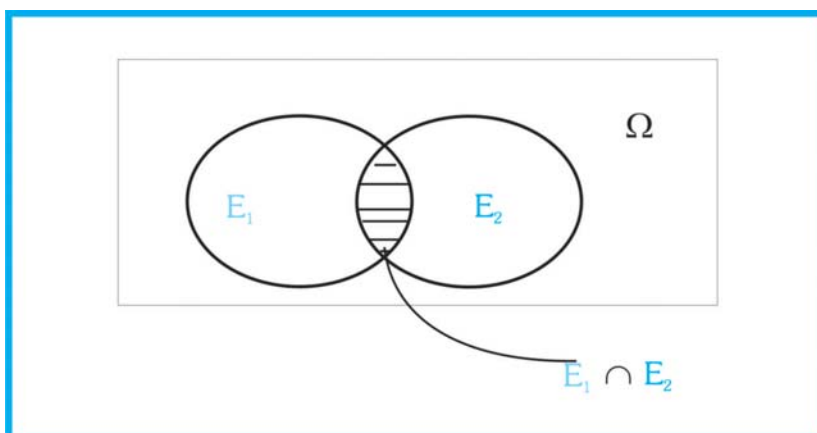


Figura 31: Evento intersecção

Fonte: Elaborada pelo autor deste livro

Composto por todos os resultados que pertencem a ambos:
 $E_1 \cap E_2 = \{4, 6\}$.

Eventos **Mutualmente Exclusivos** (M.E.): são eventos que não podem ocorrer simultaneamente; não apresentando elementos em comum (sua intersecção é o conjunto vazio).

Dentre os três eventos definidos acima, observamos que os eventos E_1 e E_3 não têm elementos em comum:

$E_1 = \{2, 4, 6\}$ $E_3 = \{1, 3\}$ $E_1 \cap E_3 = \emptyset \Rightarrow E_1$ e E_3 são mutuamente exclusivos.

Evento **Complementar** de um evento qualquer é formado por todos os resultados do espaço amostral que não pertencem ao evento. A união de um evento e seu complementar formará o próprio Espaço Amostral, e a intersecção de um evento e seu complementar é o conjunto vazio.

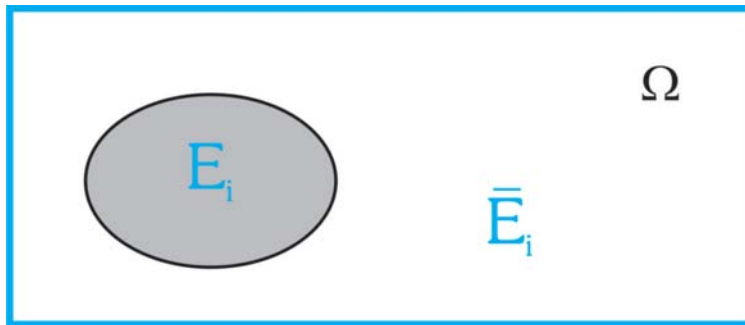


Figura 32: Evento complementar
 Fonte: Elaborada pelo autor deste livro

$$\begin{array}{ll} E_i \cup \bar{E}_i = \Omega & E_i \cap \bar{E}_i = \emptyset \\ E_1 = \{2, 4, 6\} & \bar{E}_1 = \{1, 3, 5\} \\ E_2 = \{3, 4, 5, 6\} & \bar{E}_2 = \{1, 2\} \end{array}$$

A compreensão dessas definições será extremamente útil quando calcularmos probabilidades, pois as expressões poderão ser deduzidas ou simplificadas se identificarmos de tipo de evento se trata. Podemos, agora, conhecidas as definições, passar aos conceitos de probabilidade, que são complementares.

Definições de Probabilidade

Usamos “conceitos” de probabilidade anteriormente porque, ao longo dos séculos, vários deles foram apresentados, embora se complementem.

A repetição de um experimento aleatório, mesmo sob condições semelhantes, poderá levar a resultados (eventos) diferentes. Mas se o experimento for repetido um número “suficientemente grande” de vezes haverá uma regularidade nesses resultados que permitirá calcular a sua probabilidade de ocorrência. Essa é a base para as definições que veremos a seguir.

Definição Clássica de Probabilidade

Intuitivamente, as pessoas sabem como calcular algumas probabilidades para tomar decisões. Observe os seguintes exemplos.

Exemplo 1 – Vamos supor que você fez uma aposta com um amigo. O vencedor será aquele que acertar a face que ficar para cima após o lançamento de uma **moeda honesta**. Qual a chance de você ganhar?

Você responderia, por intuição, que há 50% ($1/2$) de chances de ganhar, uma vez que há apenas duas faces (resultados) possíveis. Mesmo sem saber o que seja probabilidade, você pode calcular a chance de ocorrência de um evento de interesse, a face na qual você apostou.

Você continua apostando com o mesmo amigo. O vencedor agora será aquele que acertar o naipe de uma carta que será retirada ao acaso de um baralho comum de 52 cartas. Veremos, nesse segundo exemplo, qual a chance de você ganhar.

Novamente, de forma intuitiva, você responderia que há 25% ($1/4$) de chance, uma vez que há apenas quatro naipes (resultados) possíveis.

O que há em comum entre as situações dos exemplos 1 e 2? Refletindo um pouco, você observará que em ambos temos experimentos aleatórios e, a cada realização, apenas um dos resultados possíveis pode ocorrer. Além disso, como se supõe que a moeda e o baralho são honestos, cada um dos resultados possíveis tem a mesma probabilidade de acontecer: tanto cara quanto coroa possuem 50% de

Usaremos o termo moeda honesta para referenciar uma moeda perfeitamente equilibrada e lançamentos imparciais. De forma análoga, usaremos o adjetivo honesto para dado, baralho, entre outros.

chance de ocorrer; todos os quatro naipes (copas, espadas, ouros e paus) têm 25%. Sem que você soubesse, aplicou a **definição clássica de probabilidade** para obter as chances de ganhar.

Se um experimento aleatório puder resultar em n diferentes e igualmente prováveis resultados, e n_{E_i} desses resultados referem-se ao evento E_i , então a probabilidade do evento E_i ocorrer será:

$$P(E_i) = \frac{n_{E_i}}{n}$$

O problema reside em calcular o número total de resultados possíveis e o número de resultados associados ao evento de interesse. Isso pode ser feito usando-se técnicas de análise combinatória (que serão vistas posteriormente) ou por considerações teóricas (“bom senso”).

Seja o seguinte Experimento Aleatório: lançamento de um dado honesto e a observação da face voltada para cima. Nesse Exemplo 3, vamos calcular as probabilidades de ocorrência dos seguintes eventos:

- a) Face 1.
- b) Face par.
- c) Face menor ou igual a 2.

O Espaço Amostral deste experimento será: $\Omega = \{1, 2, 3, 4, 5, 6\}$. Sendo assim, há um total de seis resultados possíveis, com $n = 6$. Basta então definir quantos resultados estão associados a cada evento para que seja possível calcular suas probabilidades pela definição clássica.

O evento “face 1” tem apenas um resultado associado: $\{1\}$. Então $n_{E_i} = 1$, e a probabilidade de ocorrer a face 1 será: $P(E_i) = \frac{n_{E_i}}{n} = \frac{1}{6}$

O evento “face par” tem três resultados associados: $\{2, 4, 6\}$. Então $n_{E_i} = 3$, e a probabilidade de ocorrer face par será: $P(E_i) = \frac{n_{E_i}}{n} = \frac{3}{6} = \frac{1}{2}$

O evento “face menor ou igual a 2” tem dois resultados associados: $\{1, 2\}$. Então $n_{E_i} = 2$, e a probabilidade de ocorrência de face menor ou igual a 2 será: $P(E_i) = \frac{n_{E_i}}{n} = \frac{2}{6} = \frac{1}{3}$

Como visto nos exemplos, a definição clássica, que foi desenvolvida a partir do Século XVII, foi inicialmente aplicada para orientar apostas em jogos de azar. Surgiram dois problemas dessa aplicação.

O primeiro é relativamente óbvio: muitos jogos de azar não eram “honestos”. Os donos das casas inescrupulosamente “viciavam” dados e roletas, marcavam baralhos de maneira a fazer com que os

clientes perdessem sistematicamente, ou seja, o lançamento dos dados ou a retirada da carta do baralho não eram mais experimentos aleatórios.

O segundo problema decorre da pergunta: será que em todos os experimentos aleatórios todos os eventos terão a mesma probabilidade de ocorrer? Será que a probabilidade de chover no mês de novembro na cidade de Brest (cidade francesa que tem em média 225 dias nublados por ano) é a mesma na cidade de Sevilha (cidade espanhola que tem em média 240 dias de sol por ano)? Precisamos partir para a **definição experimental de probabilidade**.

Definição Experimental de Probabilidade

Seja um experimento aleatório repetido n vezes, e E_i um evento associado. A frequência relativa do evento E_i :

$$f_{REi} = \frac{n_{Ei}}{n} = \frac{\text{n}^\circ \text{ vezes que } E_i \text{ ocorreu}}{\text{total de tentativas}}$$

Quando o número de repetições tende ao infinito (ou a um número suficientemente grande), f_{REi} tende a um limite: a probabilidade de ocorrência do evento E_i . A probabilidade do evento pode ser estimada através da frequência relativa. Lembre-se da Unidade 3, a descrição de um fenômeno pode ser feita por distribuição de frequências.

Quando não há outra maneira de se obter as probabilidades dos eventos, é necessário realizar o experimento (veja novamente a Unidade 1) várias vezes para que seja possível a obtenção de um número tal de tentativas que permita que as frequências relativas estimem as probabilidades, para que seja possível construir um modelo probabilístico para o experimento. Isso pode ser feito em laboratório, em condições controladas. Por exemplo, a vida útil das lâmpadas vendidas no comércio é definida por meio de testes de sobrevivência realizados pelos fabricantes.

Em alguns casos, porém, não é possível realizar experimentos: a maioria dos fenômenos socioeconômicos e climáticos, por exemplo. Nesse caso, precisaremos estimar as probabilidades mediante as frequências relativas históricas.

Independente de como obtenhamos as probabilidades, elas obedecerão a alguns axiomas e a propriedades, que veremos a seguir.

Axiomas e Propriedades de Probabilidade

Alguns autores chamam esses axiomas e propriedades de definição axiomática da Probabilidade.

Seja um experimento aleatório e um espaço amostral associado a ele. A cada evento E_i associaremos um número real denominado $P(E_i)$ que deve satisfazer os seguintes axiomas:

a) $0 \leq P(E_i) \leq 1,0$

A probabilidade de ocorrência de um evento **sempre** é um número real entre 0 e 1 (0% e 100%);

b) $P(\Omega) = 1,0$

A probabilidade de ocorrência do Espaço Amostral é igual a 1 (100%) pois pelo menos um dos resultados do Espaço Amostral ocorrerá. Por isso o Espaço Amostral é chamado de **Evento Certo**; e

c) Se $E_1, E_2, \dots, E_n$ são eventos mutuamente exclusivos, então

$$P(E_1 \cup E_2 \cup \dots \cup E_n) = P(E_1) + P(E_2) + \dots + P(E_n)$$

Esse axioma afirma que ao unir resultados diferentes, devemos somar as probabilidades.

Além dos axiomas, há algumas propriedades básicas da Probabilidade:

a) $P(\emptyset) = 0$

A probabilidade de ocorrência do conjunto vazio é nula (igual a zero), uma vez que não há resultados no conjunto vazio. Por isso o conjunto vazio é chamado de Evento Impossível.

b) $\sum P(E_i) = 1,0$

Se a probabilidade de ocorrência do Espaço Amostral é igual a 1 (100%) ao somar as probabilidades de todos os eventos que compõem o Espaço Amostral o resultado deverá ser igual a 1 (100%);

c) $P(E_i) = 1 - P(\bar{E}_i)$:

A probabilidade de ocorrência de um evento qualquer será igual a probabilidade do Espaço Amostral (1 ou 100%) menos a probabilidade de seu evento complementar (a soma das probabilidades de todos os outros eventos do Espaço Amostral); e

Evento Impossível – evento com probabilidade de ocorrer igual a 0% é o conjunto vazio. Fonte: Barbetta, Reis e Bornia (2008).

d) Sejam E_i e E_j dois eventos quaisquer: $P(E_i \cup E_j) = P(E_i) + P(E_j) - P(E_i \cap E_j)$

A probabilidade de ocorrência do evento União de dois outros eventos será igual a soma das probabilidades de cada evento menos a probabilidade de ocorrência do evento Intersecção dos mesmos dois eventos. Essa propriedade também é chamada de **regra da adição**.

Veja, neste quarto exemplo, o que seja o Experimento Aleatório lançamento de um dado não viciado e observação da face voltada para cima, definido no Exemplo 3: o seu espaço amostral será $\Omega = \{1, 2, 3, 4, 5, 6\}$

Definindo três eventos: $E_1 = \text{face } 1 = \{1\}$, $E_2 = \text{face par} = \{2, 4, 6\}$ e $E_3 = \text{face } \leq 2 = \{1, 2\}$, cujas probabilidades já foram calculadas. Calcular a probabilidade de ocorrência dos seguintes eventos:

- Complementar de E_1 ;
- Complementar de E_2 ;
- União de E_2 e E_3 ; e
- União de E_1 e E_2 .

No Exemplo 2, obtivemos $P(E_1) = 1/6$, $P(E_2) = 3/6$ e $P(E_3) = 2/6$

Usando as propriedades:

$P(E_1) = 1 - P(\overline{E_1})$ então $P(\overline{E_1}) = 1 - P(E_1) = 1 - 1/6 = 5/6$
 $E_1 = \{2, 3, 4, 5, 6\}$

$P(E_2) = 1 - P(\overline{E_2})$ então $P(\overline{E_2}) = 1 - P(E_2) = 1 - 3/6 = 3/6$
 $E_2 = \{1, 3, 5\}$

$P(E_2 \cup E_3) = P(E_2) + P(E_3) - P(E_2 \cap E_3)$.

Observe que há apenas um elemento em comum entre os eventos E_2 e E_3 : apenas um resultado associado $\Rightarrow P(E_2 \cap E_3) = 1/6$

$P(E_2 \cup E_3) = 3/6 + 2/6 - 1/6 = 4/6$

$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

Não há elementos em comum entre os eventos E_1 e E_2 : eles são mutuamente exclusivos, sua intersecção é o conjunto vazio, e a probabilidade de ocorrência do conjunto vazio é nula.

$P(E_1 \cup E_2) = 1/6 + 3/6 - 0 = 4/6$

Agora, vamos exercitar a mente! Imagine que você trabalha em uma corretora de ações e precisa aconselhar um cliente sobre investir ou não em ações da Petrobras. Supõe-se que o preço do barril do petróleo subirá cerca de 10% nos próximos dias, há uma probabilidade estimada de acontecer. E, sabendo disso, você gostaria de saber qual é a probabilidade de que as ações da empresa subam também 10% na Bovespa. Esse caso, em que se deseja calcular a probabilidade de ocorrência de um evento condicionada à ocorrência de outro, somente poderá ser resolvido por Probabilidade Condicional, que veremos a seguir.

Probabilidade Condicional

Muitas vezes há interesse de calcular a probabilidade de ocorrência de um evento A qualquer, dada a ocorrência de outro evento B. Por exemplo, qual a probabilidade de chover amanhã em Florianópolis sabendo-se que hoje choveu? Ou, qual a probabilidade de um dispositivo eletrônico funcionar sem problemas por 200 horas consecutivas sabendo-se que ele já funcionou por 100 horas? Ou, ainda, a situação levantada anteriormente: qual a probabilidade de as ações da Petrobras aumentarem em 10% se o preço do barril de petróleo subir 10% previamente?

Veja, queremos calcular a probabilidade de ocorrência de A condicionada à ocorrência prévia de B, simbolizada por $P(A | B)$ – lê-se probabilidade de A dado B – e sua expressão será:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \text{ para } P(B) > 0$$

A probabilidade de ocorrência de A condicionada à ocorrência de B será igual à probabilidade da intersecção entre A e B, dividida pela probabilidade de ocorrência de B (o evento que já ocorreu).

Se houvesse interesse no oposto, probabilidade de ocorrência de B condicionada à ocorrência prévia de A:

$$P(B | A) = \frac{P(B \cap A)}{P(A)} \text{ para } P(A) > 0$$

No denominador da expressão é colocada **sempre** a probabilidade do evento que já ocorreu.

Comutativa – operação em que a sequência de realização não modifica o resultado, “a ordem dos fatores não altera o produto”. Fonte: Elaborado pelo autor deste livro.

Nesse caso, o valor no denominador seria a probabilidade de A uma vez que esse evento ocorreu previamente, tal como B na outra expressão. É importante ressaltar que a operação de intersecção é comutativa e implica:

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

Seja o lançamento de dois dados honestos, um após o outro, e a observação das faces voltadas para cima. Nesse quinto exemplo, iremos calcular as probabilidades:

- a) de que as faces sejam iguais supondo-se que sua soma é menor ou igual a 5.
- b) de que a soma das faces seja menor ou igual a 5, supondo-se que as faces são iguais.

Observe que há interesse em calcular a probabilidade de eventos, supondo que outro evento ocorreu previamente.

Como em todo o problema de probabilidade, é preciso montar o Espaço Amostral. Nesse caso, para os pares de faces dos dados, considerando-se que são lançados um após o outro, a ordem das faces é importante:

$$\Omega = \left\{ \begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\}$$

Figura 33: Espaço amostral do Exemplo 5

Fonte: Elaborada pelo autor deste livro

Há um total de 36 resultados possíveis: $n = 36$. Agora é preciso definir os eventos de interesse:

- a) “Faces iguais, sabendo que sua soma é menor ou igual a 5”: significa dizer probabilidade de ocorrência de faces iguais supondo que já ocorreram faces cuja soma é menor ou igual a 5; chamando o evento faces iguais de E_1 e o evento soma das faces menor ou igual a 5 de E_2 , estamos procurando $P(E_1 | E_2)$, probabilidade de ocorrência de E_1 condicionada à ocorrência prévia de E_2 .

Empregando-se a fórmula:

$$P(E_1 | E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} \text{ é preciso encontrar os valores das pro-}$$

babilidades.

Primeiramente, definimos o número de resultados do Espaço Amostral que pertencem aos eventos de interesse, para que seja possível calcular a sua probabilidade usando a definição clássica de probabilidade:

$E_1 = \{(1,1) (2,2) (3,3) (4,4) (5,5) (6,6)\}$ – faces iguais, 6 resultados, $n_{E_1} = 6$

$E_2 = \{(1,1) (1,2) (1,3) (1,4) (2,1) (2,2) (2,3) (3,1) (3,2) (4,1)\}$ – soma das faces ≤ 5 , 10 resultados, $n_{E_2} = 10$

Os elementos em comum formarão o evento intersecção:

$E_1 \cap E_2 = \{(1,1) (2,2)\}$ – faces iguais e soma das faces ≤ 5 , 2 resultados,

$$n_{E_1 \cap E_2} = 2$$

$$P(E_2) = n_{E_2} / n = 10/36 \quad P(E_1 \cap E_2) = n_{E_1 \cap E_2} / n = 2/36$$

Obtendo-se as probabilidades acima, é possível calcular a probabilidade condicional:

$$P(E_1 | E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{2/36}{10/36} = \frac{2}{10} = 0,2(20\%)$$

Então, a probabilidade de as faces serem iguais, sabendo-se que sua soma é menor ou igual a 5, é de 20%.

Esse resultado poderia ser obtido de outra forma. Se a soma das faces é menor ou igual a 5, o evento E_2 já ocorreu previamente, então o Espaço Amostral modificou-se, passando a ser o conjunto de resultados do evento E_2 :

$$\text{novo } \Omega = \{(1,1) (1,2) (1,3) (1,4) (2,1) (2,2) (2,3) (3,1) (3,2) (4,1)\}$$

O novo Espaço Amostral tem 10 resultados, novo $n = 10$

O número de resultados do evento faces iguais (E_1) no novo Espaço Amostral é igual a 2, novo $n_{E_1} = 2$ (há apenas dois pares no novo Espaço Amostral, de soma das faces menor ou igual a 5, em que as faces são iguais).

Então, a probabilidade de ocorrer o evento E_1 no novo Espaço Amostral, ou seja a probabilidade de ocorrência do evento E_1 condicionada à ocorrência prévia do evento E_2 , $P(E_1 | E_2)$, será:

Houve uma mudança no evento que ocorreu previamente.

$P(E_1|E_2) = \text{novo } n_{E1} / \text{novo } n = 2/10 = 0,2 \text{ (20\%)}$ o mesmo resultado obtido anteriormente.

b) “Soma das faces menor ou igual a 5, sabendo-se que as faces são iguais: significa dizer probabilidade de ocorrência de faces cuja soma é menor ou igual a 5, supondo-se que já ocorreram faces que são iguais; chamando o evento faces iguais de E_1 e o evento soma das faces menor ou igual a 5 de E_2 , estamos procurando $P(E_2|E_1)$, probabilidade de ocorrência de E_2 condicionada à ocorrência prévia de E_1 .

Empregando-se a fórmula: $P(E_2|E_1) = \frac{P(E_2 \cap E_1)}{P(E_1)}$ todos os valores já foram obtidos no item a.

$$P(E_2|E_1) = \frac{P(E_2 \cap E_1)}{P(E_1)} = \frac{2/36}{6/36} = \frac{2}{6} = 0,33(33\%)$$

Então, a probabilidade de que as faces possuírem soma menor ou igual a 5, sabendo que são iguais, é de 33%.

Da mesma forma que no item a o resultado poderia ser obtido de outra forma. Se as faces são iguais, o evento E_1 já ocorreu previamente, então o Espaço Amostral modificou-se, passando a ser o conjunto de resultados do evento E_1 : novo $\Omega = \{(1,1) (2,2) (3,3) (4,4) (5,5) (6,6)\}$.

O novo Espaço Amostral tem 6 resultados, novo $n = 6$.

O número de resultados do evento soma das faces menor ou igual a 5 (E_2) no novo Espaço Amostral é igual a 2, novo $n_{E2} = 2$ (há apenas dois pares no novo Espaço Amostral, de faces iguais, em que a soma das faces é menor ou igual a 5).

Então, a probabilidade de ocorrer o evento E_2 no novo Espaço Amostral, ou seja, a probabilidade de ocorrência do evento E_2 condicionada à ocorrência prévia do evento E_1 , $P(E_2|E_1)$, será: $P(E_2|E_1) = \text{novo } n_{E2} / \text{novo } n = 2/6 = 0,33 \text{ (33\%)}$ o mesmo resultado obtido anteriormente.

É extremamente importante lembrar que, conceitualmente, $P(A|B) \neq P(B|A)$, pois os eventos que ocorreram previamente são diferentes.

No quinto exemplo utilizamos a definição clássica para obter as probabilidades necessárias, mas poderíamos usar distribuições de frequências de dados históricos ou experimentais para obtê-las.

Regra do Produto

Uma das consequências da expressão da probabilidade condicional é a regra do produto, isolando a probabilidade da intersecção:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(B) \times P(A | B)$$

Nesse caso, o evento B ocorreu previamente, e o segundo valor é a probabilidade de ocorrência de A dado que B ocorreu.

$$P(A | B) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(A) \times P(B | A)$$

Nesse caso, o evento A ocorreu previamente, e o segundo valor é a probabilidade de ocorrência de B dado que A ocorreu.

Não se esqueça que a intersecção é comutativa.

É importante que seja observada com cuidado a sequência dos eventos para montar as expressões acima: analisar corretamente que evento já ocorreu.

Exemplo 6: digamos que uma urna contém duas bolas brancas e três vermelhas. Retiramos duas bolas ao acaso, uma após a outra. Veremos nos itens a seguir se a retirada foi feita **sem reposição**.

- Qual a probabilidade de as duas bolas retiradas serem da mesma cor?
- Qual a probabilidade de as duas bolas retiradas serem vermelhas, supondo-se que são da mesma cor?

Como em todos os problemas de probabilidade, primeiramente é preciso definir o Espaço Amostral. Há duas cores e duas retiradas, então poderemos ter:

- a 1ª e a 2ª bolas brancas (duas bolas da mesma cor): evento $E_1 = B_1 \cap B_2$;
- a 1ª bola branca e a 2ª bola vermelha: evento $E_2 = B_1 \cap V_2$;
- a 1ª bola vermelha e a 2ª bola branca: evento $E_3 = V_1 \cap B_2$; e
- a 1ª bola vermelha e a 2ª bola vermelha (duas bolas da mesma cor): evento $E_4 = V_1 \cap V_2$.

Então, o Espaço Amostral será:

$$\Omega = \{ B_1 \cap B_2, B_1 \cap V_2, V_1 \cap B_2, V_1 \cap V_2 \}$$

Todos os quatro eventos acima são mutuamente exclusivos: quando as bolas forem retiradas apenas um, e somente um, dos eventos acima pode ocorrer.

As retiradas são feitas sem reposição: a segunda retirada depende do resultado da primeira. Se as retiradas forem feitas sem reposição, elas serão dependentes, pois o Espaço Amostral será modificado: a cada retirada, as probabilidades de ocorrência são modificadas porque as bolas não são repostas.

- A probabilidade de retirar bola branca na 1ª retirada é de $2/5$ (2 bolas brancas no total de 5): $P(B_1) = 2/5$; e
- a probabilidade de retirar bola vermelha na 1ª retirada é de $3/5$ (3 bolas vermelhas em 5): $P(V_1) = 3/5$

Se a primeira bola retirada foi branca (o evento B_1 ocorreu previamente), restaram 4 bolas, 1 branca e 3 vermelhas:

- a probabilidade de retirar uma bola branca na 2ª retirada, se na 1ª foi extraída uma branca, é de $1/4$ (1 bola branca em 4): $P(B_2 | B_1) = 1/4$
- a probabilidade de retirar uma bola vermelha na 2ª retirada, se na 1ª foi extraída uma branca, é de $3/4$ (3 bolas vermelhas em 4): $P(V_2 | B_1) = 3/4$

Se a primeira bola retirada foi vermelha (o evento V_1 ocorreu previamente), restaram 4 bolas, 2 brancas e 2 vermelhas:

- a probabilidade de retirar uma bola branca na 2ª retirada, se na 1ª foi extraída uma vermelha, é de $2/4$ (2 bolas brancas em 4): $P(B_2 | V_1) = 2/4$

Repare que o número de bolas, número de resultados, diminuiu de 5 para 4 porque as retiradas são feitas sem reposição.

- a probabilidade de retirar uma bola vermelha na 2ª retirada, se na 1ª foi extraída uma vermelha, é de 2/4 (2 bolas vermelhas em 4): $P(V_2 | V_1) = 2/4$

a) O evento que nos interessa: “bolas da mesma cor”: brancas ou vermelhas, evento união brancas-vermelhas.

Chamando-se bolas da mesma cor de evento F: $F = [(B_1 \cap B_2) \cup (V_1 \cap V_2)]$

Empregando-se as propriedades de probabilidade:

$$P(F) = P[(B_1 \cap B_2) \cup (V_1 \cap V_2)] = P(B_1 \cap B_2) + P(V_1 \cap V_2) - P(B_1 \cap B_2) \cup (V_1 \cap V_2)$$

Os eventos $(B_1 \cap B_2)$ e $(V_1 \cap V_2)$ são mutuamente exclusivos, se as bolas são da mesma cor ou são brancas ou são vermelhas, então a intersecção entre eles é o conjunto vazio, e a probabilidade do conjunto vazio ocorrer é igual a zero, então simplesmente: $P(F) = P[(B_1 \cap B_2) \cup (V_1 \cap V_2)] = P(B_1 \cap B_2) + P(V_1 \cap V_2)$

Empregando-se a regra do produto:

$$P(B_1 \cap B_2) = P(B_1) \times P(B_2 | B_1) = (2/5) \times (1/4) = 2/20 = 1/10$$

$$P(V_1 \cap V_2) = P(V_1) \times P(V_2 | V_1) = (3/5) \times (2/4) = 6/20 = 3/10$$

Substituindo na expressão:

$$P(F) = P[(B_1 \cap B_2) \cup (V_1 \cap V_2)] = P(B_1 \cap B_2) + P(V_1 \cap V_2) = 1/10 + 3/10 = 4/10 = 0,4 \text{ (40\%)}$$

Então, se as retiradas forem feitas sem reposição, a probabilidade de que as duas bolas sejam da mesma cor será igual a 0,4 (40%).

b) Neste caso, sabemos que as duas bolas são da mesma cor (o evento F acima JÁ OCORREU) e há interesse em saber a probabilidade de as duas bolas serem vermelhas:

$$P[(V_1 \cap V_2) | F] = P\{(V_1 \cap V_2) | [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\}$$

Empregando-se a expressão de probabilidade condicional:

$$P\{(V_1 \cap V_2) | [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\} = \frac{P\{(V_1 \cap V_2) \cap [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\}}{P[(B_1 \cap B_2) \cup (V_1 \cap V_2)]}$$

A probabilidade do denominador já é conhecida do item a. E a do numerador pode ser obtida facilmente.

Repare: o que há em comum entre o evento $(V_1 \cap V_2)$ e o evento $[(B_1 \cap B_2) \cup (V_1 \cap V_2)]$? Em suma, qual será o evento intersecção? O que há em comum entre duas bolas vermelhas e duas bolas da mesma cor? O próprio evento duas bolas vermelhas $(V_1 \cap V_2)$, então:

$$(V_1 \cap V_2) \cap [(B_1 \cap B_2) \cup (V_1 \cap V_2)] = (V_1 \cap V_2);$$

$$P\{(V_1 \cap V_2) \cap [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\} = P(V_1 \cap V_2) = 3/10$$

Sabendo-se que $P\{(V_1 \cap V_2) \mid [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\} = 4/10$ (do item a.1) e substituindo-se os valores na fórmula:

$$P\{(V_1 \cap V_2) \mid [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\} =$$

$$\frac{P(V_1 \cap V_2)}{P[(B_1 \cap B_2) \cup (V_1 \cap V_2)]} = \frac{3/10}{4/10} = \frac{3}{4}$$

$$P\{(V_1 \cap V_2) \mid [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\} = 0,75 \text{ (75\%)}$$

Então, se as retiradas forem feitas sem reposição e as duas bolas forem da mesma cor, a probabilidade de serem vermelhas será igual a 0,75 (75%).

As retiradas e as probabilidades podem ser representadas através de um diagrama chamado de “Árvore de Probabilidades”:

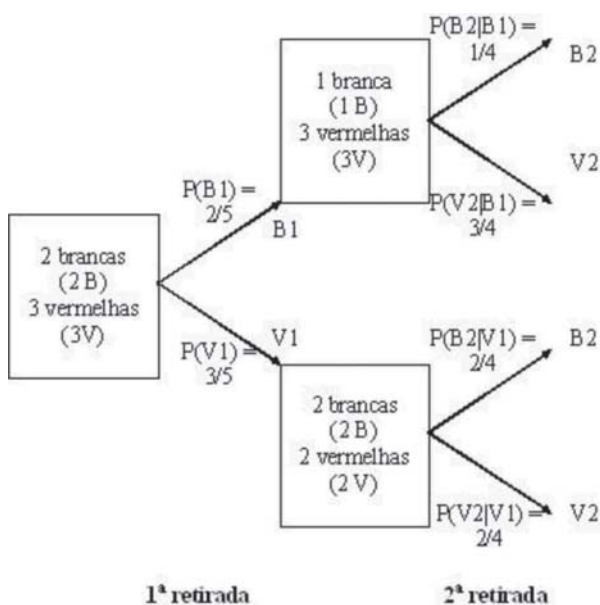


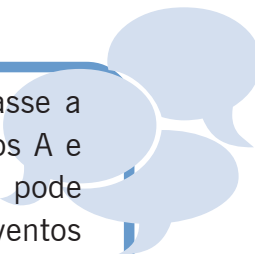
Figura 34: Árvore de probabilidades – retiradas sem reposição
Fonte: Elaborada pelo autor deste livro

Observe que através da Árvore de Probabilidades podemos chegar aos mesmos resultados obtidos anteriormente. Partindo-se do Espaço Amostral original um dos ramos significa 1ª bola branca (B_1) e o outro 1ª bola vermelha (V_1). Dependendo do resultado da primeira retirada haverá um Espaço Amostral diferente: 1 bola branca e 3 vermelhas se na 1ª retirada obteve-se uma bola branca, ou 2 bolas brancas e 2 vermelhas se na 1ª retirada obteve-se uma bola vermelha.

A partir dos novos Espaços Amostrais, é possível calcular as probabilidades condicionais para cada caso e, depois, substituí-las nas fórmulas adequadas. Contudo, a árvore será inútil se o evento para o qual se deseja calcular a probabilidade não for definido adequadamente: nesse caso, no item a: bolas da mesma cor $[(B_1 \cap B_2) \cup (V_1 \cap V_2)]$; e no item b: bolas vermelhas sabendo-se que são da mesma cor $\{(V_1 \cap V_2) \mid [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\}$.

A árvore será igualmente inútil se não forem usadas as definições de eventos dependentes (porque não há reposição) e de eventos mutuamente exclusivos (porque os eventos não podem ocorrer simultaneamente), e as expressões de probabilidade condicional e os axiomas de probabilidade.

O grande inconveniente da Árvore de Probabilidades surge quando o número de “retiradas” aumenta e/ou o número de resultados possíveis para cada retirada é considerável: torna-se impraticável desenhar a Árvore, enumerando-se todos os resultados. Nesses casos, usamos Análise Combinatória, que veremos adiante.



E se a ocorrência do evento A não modificasse a probabilidade de ocorrência de B? Os eventos A e B seriam chamados de independentes. Você pode imaginar situações práticas em que dois eventos sejam independentes?

Eventos Independentes

Dois ou mais eventos são independentes quando a ocorrência de um deles não influencia a probabilidade de ocorrência dos outros. Se dois eventos A e B são independentes, então a probabilidade de A

ocorrer dado que B ocorreu é igual à própria probabilidade de ocorrência de A, e a probabilidade de B ocorrer dado que B ocorreu é igual à própria probabilidade de ocorrência de B.

Se A e B são independentes, então:

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

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

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

As expressões acima são válidas se os eventos A e B forem independentes.

Em situações práticas, dois eventos são independentes quando a ocorrência de um deles não modifica, ou modifica muito pouco, o Espaço Amostral do Experimento Aleatório. É o que ocorria na Unidade 2 quando fazíamos amostragem aleatória simples: naquele momento não foi dito que a amostragem era com reposição, que dificilmente é feita na prática, mas admite-se que sendo o tamanho da população muito grande, a retirada de uma pequena amostra não modificará muito as proporções dos eventos.

Exemplo 7: resolva o Exemplo 6, mas agora supondo que as retiradas foram feitas **com reposição**.

- Qual a probabilidade de as duas bolas retiradas serem da mesma cor? Resposta: 0,52 (52%)
- Qual a probabilidade de as duas bolas retiradas serem vermelhas, supondo-se que são da mesma cor? Resposta: 0,69 (69%).

Saiba mais...

Sobre conceitos básicos de Probabilidade, BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Editora da UFSC, 2007, Capítulo 7.

Também sobre conceitos básico de Probabilidade: STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Editora Harbra, 2001, Capítulo 3.

LOPES, Paulo A. *Probabilidades e Estatística*. Rio de Janeiro: Reichmann e Affonso Editores, 1999, Capítulo 3.



A Figura 35 apresenta um resumo desta Unidade.

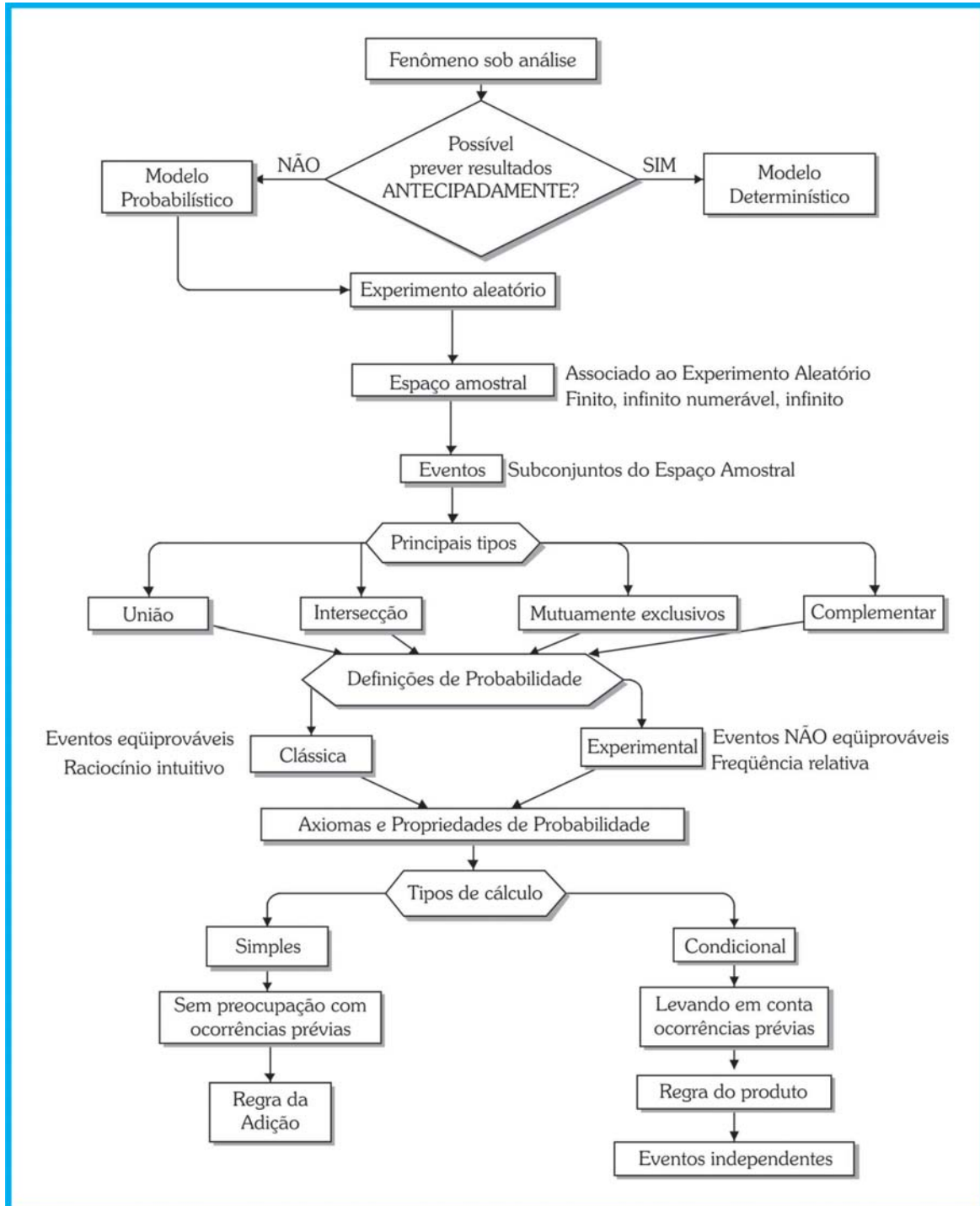


Figura 35: Resumo da Unidade 5
Fonte: Elaborada pelo autor deste livro

Chegamos ao final de Unidade 5. Esperamos que você tenha aprendido todos os conceitos trabalhados e, por meio dos exemplos, colocado em prática as informações adquiridas. Neles, propusemos que você reconhecesse os modelos probabilísticos, modelos determinísticos, principais tipos de eventos e os diferentes tipos de cálculos. Na Unidade 6, daremos continuidade ao tema explicando o conceito de variável aleatória, indispensável para as Unidades 7, 8 e 9, nas quais conheceremos também alguns dos modelos probabilísticos mais empregados, além dos conceitos de probabilidade no processo de inferência estatística.



Atividades de aprendizagem

Adaptado de
BARBETTA, Pedro A.
*Estatística Aplicada às
Ciências Sociais*. 7.
ed. Florianópolis: Ed.
da UFSC, 2007.

1. Numa eleição para a prefeitura de uma cidade, 30% dos eleitores pretendem votar no candidato A; 50% no candidato B; e 20% em branco ou nulo. Sorteia-se um eleitor na cidade e verifica-se o candidato de sua preferência.
 - a) Construa um modelo probabilístico para o problema.
 - b) Qual a probabilidade de o eleitor sorteado votar em um dos dois candidatos? Resposta: 0,8
2. Extraem-se ao acaso duas cartas de um baralho de 52 cartas; uma após a outra SEM reposição. Calcule as seguintes probabilidades:
 - a) Ambas as cartas são vermelhas. Resposta: 0,245
 - b) Ambas as cartas são de paus. Resposta: 0,058
 - c) Ambas as cartas são de “Figuras” (ás, rei, dama ou valete). Resposta: 0,0905
 - d) Uma carta de paus e outra de copas. Resposta: 0,1274

3. Repita o exercício 2 supondo que as retiradas foram feitas COM reposição.
- a) Resposta: 0,25
 - b) Resposta: 0,0625
 - c) Resposta: 0,0947
 - d) Resposta: 0,125
4. Para determinado telefone a probabilidade de se conseguir linha é de 0,75 em dias normais e de 0,25 em dias de chuva. A probabilidade de chover em um dia é 0,1. Além disso, tendo-se conseguido linha a probabilidade de um número estar ocupado é $11/21$.
- a) Qual a probabilidade de um telefone ter sua ligação completada? Resposta: 0,333
 - b) Dado que um telefonema foi completado, qual a probabilidade de estar chovendo? Resposta: 0,0357

6

UNIDADE

Variáveis Aleatórias



Ao finalizar esta Unidade, você deverá ser capaz de compreender o conceito de Variável Aleatória; de perceber a relação dela com os modelos probabilísticos; e, ainda, de constatar que os modelos probabilísticos podem ser construídos para tais variáveis.

Conceito de Variável Aleatória

Caro estudante!

Há uma pergunta que normalmente é feita àqueles que trabalham com ciências exatas: por que a obsessão por reduzir tudo a números? Vimos em Análise Exploratória de Dados que uma variável quantitativa, geralmente, apresenta mais informação que uma variável qualitativa e pode ser resumida não somente através de tabelas e de gráficos, mas também através de medidas de síntese. Dissemos "geralmente" porque, afinal, nem tudo pode ser reduzido a números; é o caso da inteligência e da criatividade.

Nos exemplos sobre probabilidade apresentados na Unidade 5, os eventos foram geralmente definidos de forma verbal: bolas da mesma cor, 2 bolas vermelhas, soma das faces menor ou igual a 5 etc. Não haveria problema em definir os eventos através de números. Bastaria associar aos resultados do Espaço Amostral números, por meio de uma função; esta chamada de Variável Aleatória. Os modelos probabilísticos podem, então, ser construídos para as variáveis aleatórias. O administrador precisa conhecer esses conceitos porque eles proporcionam maior objetividade na obtenção das probabilidades, o que torna o processo de tomada de decisões mais seguro. Vamos conhecê-los nesta Unidade!

Uma definição inicial de Variável Aleatória: trata-se de uma variável quantitativa, cujo resultado (valor) depende de fatores aleatórios.

Formalmente, **Variável Aleatória** é uma função matemática que associa números reais (contradomínio da função) aos resultados de um Espaço Amostral (domínio da função), por sua vez vinculado a

Espaço Amostral – é o conjunto de todos os resultados possíveis de um experimento aleatório. Fonte: Barbetta, Reis e Bornia (2008).

Experimento Aleatório – é o processo de obtenção de um resultado ou medida que apresenta as seguintes características: não se pode afirmar, antes de realizar o experimento, qual será o resultado de uma realização, mas é possível determinar o conjunto de resultados possíveis; quando é realizado um grande número de vezes (replicado) apresentará uma regularidade que permitirá construir um modelo probabilístico para analisar o experimento. Fonte: Adaptado de Lopes (1999).

Distribuição de probabilidades – função que relaciona os valores possíveis que uma variável aleatória pode assumir com as respectivas probabilidades, em suma é o próprio modelo probabilístico da variável aleatória. Fonte: Barbetta, Reis e Bornia (2008).

um Experimento Aleatório. Se o Espaço Amostral for finito ou infinito numerável, a variável aleatória é dita discreta. Se o Espaço Amostral for infinito, a variável aleatória é dita contínua.

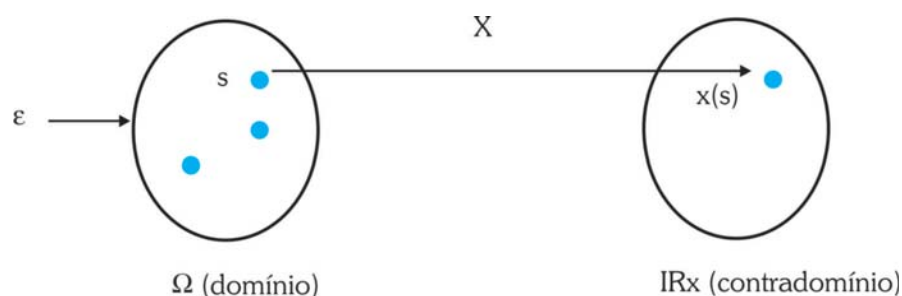


Figura 36: Variável aleatória
Fonte: Elaborada pelo autor deste livro

Por exemplo, imaginemos o seguinte Experimento Aleatório: o lançamento de uma moeda honesta duas vezes, e observar a face voltada para cima. O Espaço Amostral seria finito:

$$\Omega = \{\text{CaraCara; CaraCoroa; CoroaCara; CoroaCoroa}\}$$

Se houvesse interesse no número de caras obtidas, poderia ser definida uma variável aleatória discreta X ; onde X = Número de caras em dois lançamentos. Os valores possíveis de X seriam: $X = \{0, 1, 2\}$

O valor 0 é associado ao evento CoroaCoroa; o valor 1 é associado aos eventos CaraCoroa e CoroaCara; e o valor 2 é associado ao evento CaraCara.

Quando o Espaço Amostral é infinito, muitas vezes já está definido de forma numérica, pela própria natureza quantitativa do fenômeno analisado, facilitando a definição da variável aleatória.

Os Modelos Probabilísticos podem ser construídos para as variáveis aleatórias: assim haverá Modelos Probabilísticos Discretos e Modelos Probabilísticos Contínuos. Para construir um modelo probabilístico para uma variável aleatória é necessário definir os seus possíveis valores (contradomínio), e como a probabilidade total (do Espaço Amostral, que vale 1) distribui-se entre eles: é preciso, então, definir a distribuição de probabilidades.

Veja que, dependendo do tipo de variável aleatória, haverá diferenças na construção da distribuição.

Distribuições de Probabilidades para Variáveis Aleatórias Discretas

Podemos ver alguns exemplos de variáveis aleatórias discretas:

- a) número de coroas obtido no lançamento de 2 moedas;
- b) número de itens defeituosos em uma amostra retirada aleatoriamente de um lote;
- c) número de defeitos em um azulejo numa fábrica de revestimentos cerâmicos; e
- d) número de pessoas que visitam determinado *site* em um período de tempo.

Quando uma variável aleatória X é discreta, a obtenção da distribuição de probabilidades consiste em definir o conjunto de pares $[x_i, p(x_i)]$; onde x_i é o i -ésimo valor da variável X , e $p(x_i)$ é a probabilidade de ocorrência de x_i , como na Tabela 1:

Tabela 1: Distribuição de Probabilidades para uma Variável Aleatória Discreta

$X = x_i$	$p(X = x_i)$
x_1	$p(x_1)$
x_2	$p(x_2)$
...	...
x_n	$p(x_n)$

Fonte: Elaborada pelo autor deste livro

Onde $p(x_i) \geq 0$, n é o número de valores que X pode assumir, e

$$\sum_{i=1}^n p(x_i) = 1,0.$$

Ao obter a distribuição de probabilidades para uma variável aleatória discreta e desejar conferir os resultados, some as probabilidades. Se elas não somarem 1, algo está errado. Vamos ao primeiro exemplo.

Imagine que o jogador Ruinzinho está treinando cobranças de pênaltis. Dados históricos mostram que: a probabilidade de ele acertar uma cobrança, supondo que ele acertou a anterior é de 60%; mas

se ele errou, a probabilidade de ele acertar uma cobrança cai para 30%. Construa a distribuição de probabilidades do número de acertos em três tentativas de cobrança.

A variável aleatória X , número de acertos em três tentativas, é uma variável aleatória discreta: o seu contradomínio é finito, o jogador pode acertar 0, 1, 2 ou 3 vezes. Mas para calcular as probabilidades associadas a esses valores é preciso estabelecer todos os eventos possíveis, pois mais de um evento contribui para as probabilidades de 1 e 2 acertos. Observando a árvore de eventos a seguir (onde A é acertar a cobrança e E significa erro).

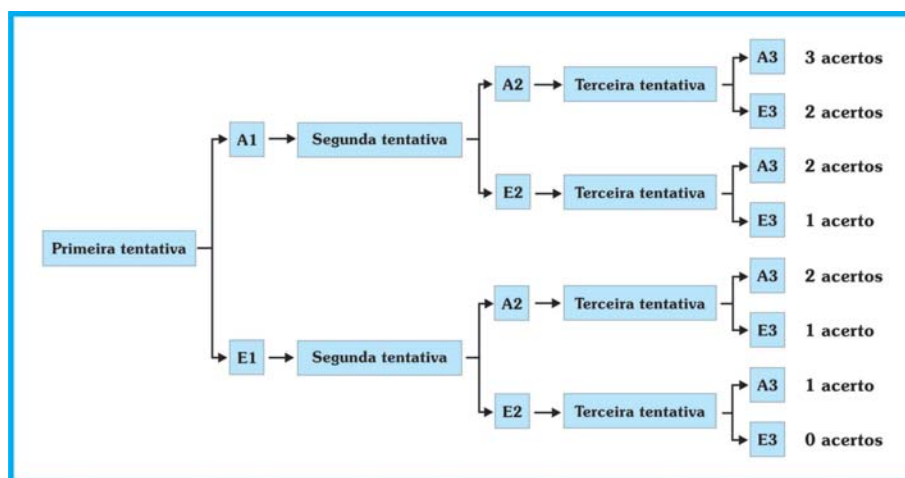


Figura 37: Árvore de eventos
Fonte: Elaborada pelo autor deste livro

Observe que todos os eventos são mutuamente exclusivos, o jogador não pode, na mesma sequência de 3 cobranças, errar e acertar a primeira. É preciso explicitar os valores da variável, e os eventos em termos de teoria dos conjuntos.

Valores possíveis = $\{0, 1, 2, 3\}$ acertos. A equivalência entre os valores da variável e os eventos é estabelecida a seguir:

$$X = 0 \Leftrightarrow [E_1 \cap E_2 \cap E_3]$$

$$X = 1 \Leftrightarrow [(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$$

$$X = 2 \Leftrightarrow [(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap A_3) \cup (A_1 \cap E_2 \cap A_3)]$$

$$X = 3 \Leftrightarrow [A_1 \cap A_2 \cap A_3]$$

Então:

$$P(X=0) = P[E_1 \cap E_2 \cap E_3]$$

$$P(X=1) = P[(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$$

$$P(X=2) = P[(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap A_3) \cup (A_1 \cap E_2 \cap A_3)]$$

$$P(X=3) = P[A_1 \cap A_2 \cap A_3]$$

Assume-se que na primeira tentativa o jogador tem 50% de chance de acertar, então: $P(A_1) = 0,5$ e $P(E_1) = 0,5$

Além disso, estabeleceu-se que quando o jogador acertou a cobrança na tentativa anterior, a probabilidade de acertar a próxima é de 0,6 e, caso tenha errado na anterior, a probabilidade de acertar na próxima é de apenas 0,3. Trata-se de duas probabilidades condicionais, estabelecidas em função de eventos já ocorridos.

Se o jogador acertou na tentativa i (qualquer uma), as probabilidades de acertar e errar na próxima tentativa serão: $P(A_{i+1} | A_i) = 0,6$

Pelo complementar, obtém-se: $P(E_{i+1} | A_i) = 0,4$

Se o jogador errou na tentativa i , as probabilidades de acertar e errar na próxima tentativa serão: $P(A_{i+1} | E_i) = 0,3$

Pelo complementar, obtém-se: $P(E_{i+1} | E_i) = 0,7$

Com essas probabilidades estabelecidas, considerando-se a regra do produto e o fato de os eventos serem mutuamente exclusivos, é possível calcular as probabilidades de ocorrência de cada valor da variável aleatória X .

$$P(X=0) = P[E_1 \cap E_2 \cap E_3] = P(E_1) \times P(E_2 | E_1) \times P(E_3 | E_1 \cap E_2)$$

Como os resultados em uma tentativa só dependem daqueles obtidos na imediatamente anterior, o terceiro termo da expressão acima pode ser simplificado para $P(E_3 | E_2)$, e a probabilidade será:

$$P(X=0) = P(E_1) \times P(E_2 | E_1) \times P(E_3 | E_2) = 0,5 \times 0,7 \times 0,7 = 0,245$$

(24,5%)

Estendendo o procedimento acima para os outros valores:

$$P(X=1) = P[(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$$

$$P(X=2) = P[(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap A_3) \cup (A_1 \cap E_2 \cap A_3)]$$

$$P(X=3) = P[A_1 \cap A_2 \cap A_3]$$

Como os eventos são mutuamente exclusivos:

$$P(X=1) = P(A_1 \cap E_2 \cap E_3) + P(E_1 \cap A_2 \cap E_3) + P(E_1 \cap E_2 \cap A_3)$$

$$P(X=1) =$$

$$P(A_1) \times P(E_2 | A_1) \times P(E_3 | E_2) + P(E_1) \times P(A_2 | E_1) \times P(E_3 | A_2) + P(E_1) \times P(E_2 | E_1) \times P(A_3 | E_2)$$

E_1 , errar a primeira cobrança, é o evento complementar de A_1 , acertar a primeira cobrança.

$$P(X=1) = 0,5 \times 0,4 \times 0,7 + 0,5 \times 0,3 \times 0,4 + 0,5 \times 0,7 \times 0,3 = 0,305$$

$$P(X=2) = P(A_1 \cap A_2 \cap E_3) + P(E_1 \cap A_2 \cap A_3) + P(A_1 \cap E_2 \cap A_3)$$

$$P(X=2)$$

$$= P(A_1) \times P(A_2 | A_1) \times P(E_3 | A_2) + P(E_1) \times P(A_2 | E_1) \times P(A_3 | A_2) + P(A_1) \times P(E_2 | A_1) \times P(A_3 | E_2)$$

$$P(X=2) = 0,5 \times 0,6 \times 0,4 + 0,5 \times 0,3 \times 0,6 + 0,5 \times 0,4 \times 0,3 = 0,27$$

(27%)

$$P(X=3) = P[A_1 \cap A_2 \cap A_3] = P(A_1) \times P(A_2 | A_1) \times P(A_3 | A_2) = 0,5 \times 0,6 \times 0,6 = 0,18 \text{ (18\%)}$$

Com os valores calculados acima é possível construir a Tabela 2 com os pares valores-probabilidades.

Tabela 2: Distribuição de probabilidades: número de acertos em 3 cobranças

X	$P(X = x_i)$
0	0,245
1	0,305
2	0,270
3	0,180
Total	1,0

Fonte: Elaborada pelo autor deste livro

Ao longo dos séculos, matemáticos e estatísticos deduziram modelos para tornar mais simples a obtenção de distribuição de probabilidades para uma variável aleatória discreta. Alguns desses modelos serão vistos na Unidade 7.

Agora, vamos passar para a Análise das Variáveis Aleatórias Discretas.

Distribuições de Probabilidades para Variáveis Aleatórias Contínuas

Podemos ver alguns exemplos de variáveis aleatórias contínuas:

- volume de água perdido em um sistema de abastecimento;
- renda familiar em salários mínimos de pessoas selecionadas por amostragem aleatória para responder a uma pesquisa;
- a demanda por um produto em um mês; e
- o tempo de vida de uma lâmpada incandescente.

Uma variável aleatória contínua está associada a um Espaço Amostral infinito. Assim, a probabilidade de a variável assumir exatamente um valor x_i é zero, não havendo mais sentido em representar a distribuição pelos pares $x_i - p(x_i)$. Igualmente sem sentido fica a distinção entre $>$ e $\geq$ existente nas variáveis aleatórias discretas. Utiliza-se então uma função não negativa: a função densidade de probabilidades, definida para todos os valores possíveis da variável aleatória.

Uma função densidade de probabilidades poderia ser apresentada graficamente da seguinte forma:

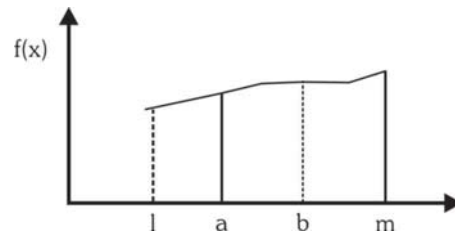


Figura 38: Função densidade de probabilidades

Fonte: Elaborada pelo autor deste livro

Para calcular a probabilidade de uma variável aleatória contínua assumir valores entre **a** e **b** (dois valores quaisquer), basta calcular a área abaixo da curva entre **a** e **b**. Se a área for calculada entre **l** e **m** (limites da função) tem que dar 1, que é a probabilidade total. Usualmente isso é feito calculando-se a integral da função no intervalo de interesse. Em muitas situações de nosso interesse tais probabilidades podem ser calculadas através de fórmulas matemáticas relativamente simples, ou foram dispostas em tabelas, que são encontradas

em praticamente todos os livros de estatística, e que serão vistas na Unidade 7.

Vejamos alguns conceitos muito importantes, como valor esperado e variância de uma variável aleatória.

Valor Esperado e Variância

Todos os modelos probabilísticos apresentam duas medidas (dois momentos) que permitem caracterizar a variável aleatória para a qual eles foram construídos: o Valor Esperado e a Variância da variável aleatória. O Valor Esperado (simbolizado por $E(X)$) nada mais é do que a média aritmética simples, vista em Análise Exploratória de Dados (Unidade 4), que utiliza probabilidades ao invés de frequências no cálculo. Analogamente, a Variância (simbolizada por $V(X)$), vista anteriormente, faz uso de probabilidades. Da mesma forma que na Análise Exploratória de Dados, é comum trabalhar-se com o Desvio Padrão, raiz quadrada positiva da Variância (que aqui será simbolizado por $\sigma(X)$, “sigma de X”). A interpretação dos resultados obtidos pode ser feita de forma semelhante à Análise Exploratória de Dados; apenas ‘e preciso ter claro de que se trata de uma variável aleatória e, por isso, probabilidades e não frequências estão sendo usadas.

Para uma variável aleatória discreta o valor esperado e a variância podem ser calculados da seguinte forma:

$$E(X) = \sum_{i=1}^n x_i \times p(x_i) \quad V(X) = E(X^2) - [E(X)]^2, \text{ onde } E(X^2) = \sum_{i=1}^n x_i^2 \times p(x_i)$$

Para uma variável aleatória contínua, a obtenção do valor esperado e da variância exige o cálculo de integrais das funções de densidade de probabilidades. Para as distribuições mais importantes as equações encontram-se disponíveis nos livros de Estatística, em função dos parâmetros da distribuição, e algumas serão vistas na Unidade 7.

Uma das principais utilidades do valor esperado é observada na comparação de propostas. Suponha que os valores de uma variável aleatória sejam lucros, ou prejuízos, advindos de decisões toma-

das. Por exemplo: decidir por uma proposta de compra do cliente A ou do cliente B. Associados aos valores, há probabilidades. Como decidir qual a mais vantajosa? O cálculo do valor esperado possibilita uma comparação objetiva: decidiríamos pela que apresentasse o lucro esperado mais elevado. Há um campo de conhecimento que se ocupa especificamente de fornecer as ferramentas necessárias para tais tomadas de decisão: a Teoria Estatística da Decisão ou Análise Estatística da Decisão.

O valor esperado (média) e a variância apresentam algumas propriedades, tanto para variáveis aleatórias discretas quanto contínuas. O seu conhecimento facilitará muito a obtenção das medidas em problemas mais sofisticados.

Para o valor esperado $E(X)$, sendo k uma constante:

- a) $E(k) = k$ – A média de uma constante é a própria constante.
- b) $E(k \pm X) = k \pm E(X)$ – A média de uma constante somada a uma variável aleatória é a própria constante somada à média da variável aleatória.
- c) $E(k \times X) = k \times E(X)$ – A média de uma constante multiplicada por uma variável aleatória é a própria constante multiplicada pela média da variável aleatória.
- d) $E(X \pm Y) = E(X) \pm E(Y)$ – A média da soma de duas variáveis aleatórias é igual à soma das médias das duas variáveis aleatórias.
- e) Sejam X e Y duas variáveis aleatórias independentes $E(X \times Y) = E(X) \times E(Y)$ – A média do produto de duas variáveis aleatórias independentes é igual ao produto das médias das duas variáveis aleatórias.

Para a variância $V(X)$, sendo k uma constante:

- a) $V(k) = 0$ – Uma constante não varia, portanto sua variância é igual a zero.
- b) $V(k \pm X) = V(X)$ – A variância de uma constante somada a uma variável aleatória é igual apenas à variância da variável aleatória.
- c) $V(k \times X) = k^2 \times V(X)$ – A variância de uma constante multiplicada a uma variável aleatória é igual ao quadrado da constante multiplicada pela variância da variável aleatória.

d) Sejam X e Y duas variáveis aleatórias **independentes**

$V(X \pm Y) = V(X) + V(Y)$ – A variância da soma ou subtração de duas variáveis aleatórias independentes será igual à soma das variâncias das duas variáveis aleatórias.

Vejamos um exemplo. Exemplo 2 – Calcular o valor esperado e a variância da distribuição do Exemplo 1.

Para uma variável aleatória discreta, é aconselhável acrescentar mais uma coluna à Tabela 3 com os valores e as probabilidades para poder calcular o valor de $E(X^2)$:

Tabela 3: Distribuição de probabilidades do Exemplo 1
(com coluna $x_i^2 \times p(X = x_i)$)

X	$p(X = x_i)$	$x_i \times p(X = x_i)$	$x_i^2 \times p(X = x_i)$
0	0,245	0	0
1	0,305	0,305	0,305
2	0,270	0,540	1,08
3	0,180	0,540	1,62
Total	1,0	1,385	3,005

Fonte: Elaborada pelo autor deste livro

Substituindo nas expressões o valor esperado e a variância:

$$E(X) = \sum_{i=1}^n x_i \times p(x_i) = \mathbf{1,385 \text{ acertos}}$$

$$V(X) = \sum_{i=1}^n x_i^2 \times p(x_i) - \left[\sum_{i=1}^n x_i \times p(x_i) \right]^2 = \mathbf{3,005 - (1,385)^2 = 1,087 \text{ acertos}^2}$$

$$\sigma(X) = \sqrt{V(X)} = \sqrt{1,087} = \mathbf{1,042 \text{ acertos}}$$

Observe que o valor esperado (1,385 acertos) é um valor que a variável aleatória não pode assumir! Não é o “valor mais provável”, é o ponto de equilíbrio do conjunto. Repare que a unidade da variância dificulta sua comparação com o valor esperado, mas ao se utilizar o desvio padrão é possível verificar que a dispersão dos resultados é quase do valor da média (valor esperado).

Saiba mais...

Sobre Variáveis Aleatórias, BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulos 5 e 6.

Sobre as propriedades de valor esperado e variância, BARBETTA, P. A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulos 5 e 6.

Também sobre variáveis aleatórias, STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Ed. Harbra, 2001, Capítulos 5 e 6.

Sobre teoria estatística da decisão: BEKMAN, Oto R.; COSTA NETO, Pedro O. *Análise Estatística da Decisão*. São Paulo: Edgard Blücher, 1980, 4. reimpressão, 2006.

Resumindo



O resumo desta Unidade está demonstrado na Figura 39:

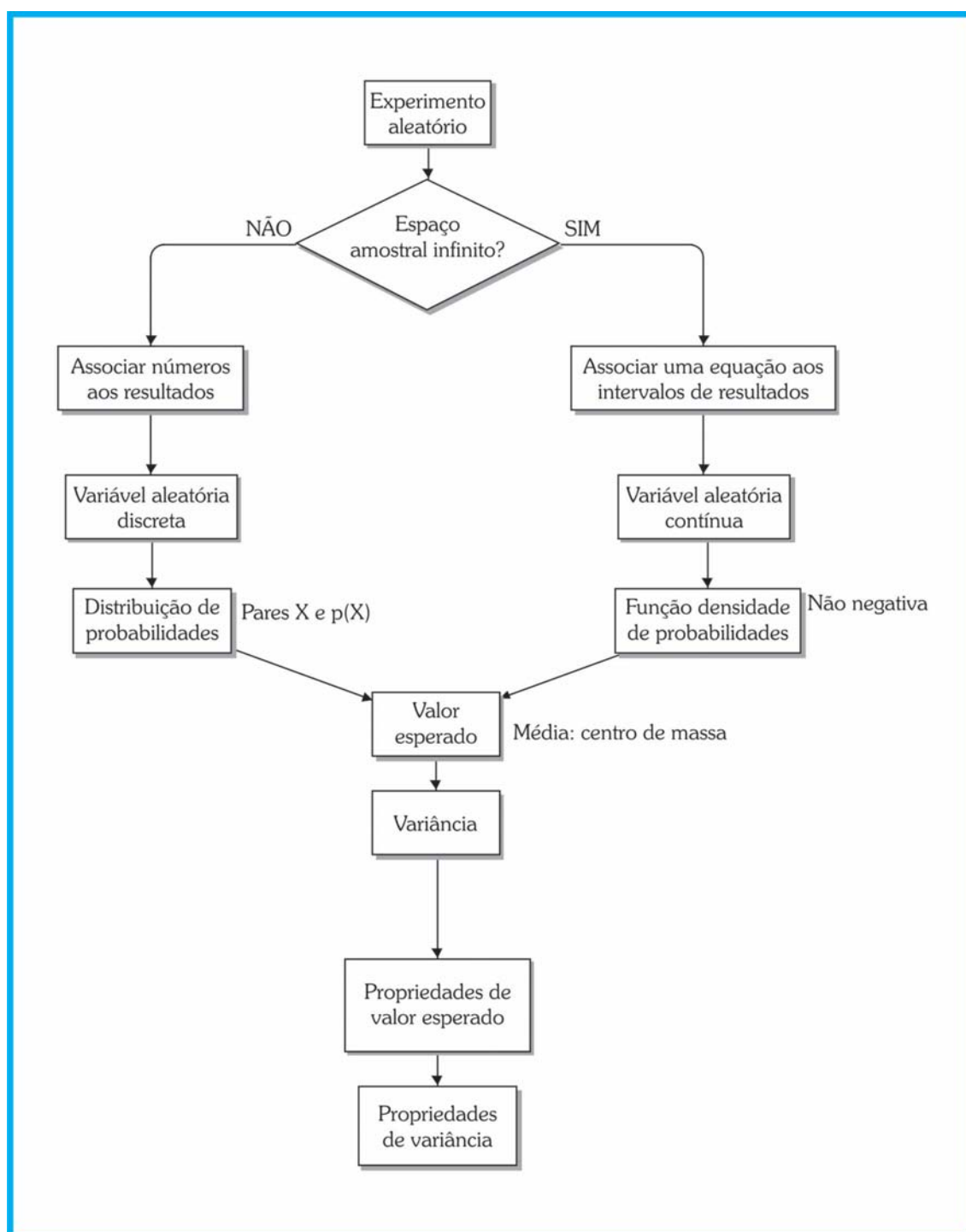


Figura 39: Resumo da Unidade 6
Fonte: Elaborada pelo autor deste livro

Chegamos ao final de mais uma Unidade. Veremos mais sobre os temas aqui abordados na Unidade 7, quando estudarmos várias distribuições de probabilidade (modelos probabilísticos) que são extremamente úteis para modelar muitas situações práticas e para auxiliar na tomada de decisões. Esses conhecimentos serão depois aplicados nas Unidades 8 e 9.



Atividades de aprendizagem

- Três alunos estão tentando, de modo independente, resolver um problema. A probabilidade de o aluno A resolver o problema é de $\frac{4}{5}$; de B resolver é de $\frac{2}{3}$; e de C resolver é de $\frac{3}{7}$. Seja X o número de soluções corretas apresentadas para esse problema.
 - Construa a distribuição de probabilidades de X . (Resposta: 0,038; 0,257; 0,476; 0,228); e
 - Calcule $E(X)$ e $V(X)$. (Resposta: 1,893; 0,630).
- Um prédio possui três vigias dispostos em vários pontos de onde têm visão do portão de entrada. Se alguém não autorizado entrar, o vigia que o vir fará soar um alarme. Suponha que os vigias trabalham independentemente entre si, e a probabilidade de cada um deles ver uma pessoa entrar é 0,8. Seja X o número de alarmes que soam ao entrar uma pessoa não autorizada. Encontre a distribuição de probabilidades de X . (Resposta: 0,008; 0,096; 0,384; 0,512).

7

UNIDADE

Modelos Probabilísticos mais Comuns



Ao finalizar esta Unidade, você deverá ser capaz de compreender os modelos probabilísticos mais importantes para variáveis aleatórias discretas e contínuas; de identificar as situações reais em que tais modelos podem ser usados para o cálculo de probabilidades; e, além disso, de perceber a relevância desses aspectos para o administrador.

Modelos Probabilísticos para Variáveis Aleatórias Discretas

Nas Unidades 5 e 6 vimos os conceitos gerais de Probabilidade e Variáveis Aleatórias: podemos construir um modelo probabilístico do zero para um problema de administração, a partir de dados históricos ou experimentais.

Embora plenamente possível, o processo de construção de um modelo probabilístico do zero pode ser bastante longo: é preciso coletar os dados (ver Unidades 1 e 2), fazer a análise exploratória deles (ver Unidades 3 e 4), obter as probabilidades e validar o modelo. Mesmo tomando todos os cuidados, muitas vezes reinventaremos a roda, com o risco de fazê-la quadrada...

Por que não usar os conhecimentos prévios desenvolvidos ao longo de centenas de anos de pesquisa e de experimentação? Vamos procurar, dentre os vários modelos probabilísticos existentes aquele mais apropriado para o fenômeno que estamos estudando, que é materializado através de variáveis aleatórias.

Por meio da análise exploratória de dados podemos avaliar qual modelo é mais apropriado para os nossos dados. Contudo, para fazer isso precisamos conhecer tais modelos.

Nesta Unidade vamos estudar os modelos mais usados para variáveis aleatórias discretas (binomial e Poisson), e para variáveis aleatórias contínuas (uniforme, normal, t e qui-quadrado).

É importante determinar com cuidado a variável aleatória discreta.

Variável aleatória – é uma função matemática que associa números reais aos resultados de um Espaço Amostral, por sua vez vinculado a um Experimento Aleatório. Fonte: Barbetta, Reis e Bornia (2008).

É preciso, também, identificar se o **Espaço Amostral é finito** ou **infinito numerável** porque alguns modelos são apropriados para um caso e não para o outro.

Vamos ver, a seguir, os dois modelos mais importantes: o binomial e o Poisson.

Espaço Amostral finito

– é aquele formado por um número limitado de resultados possíveis. Fonte: Barbetta, Reis e Bornia (2008).

Espaço Amostral infinito numerável

– é aquele formado por um número infinito de resultados, mas que podem ser listados. Fonte: Barbetta, Reis e Bornia (2008).

Experimento Aleatório

– é um processo de obtenção de um resultado ou medida que apresenta as seguintes características: não se pode afirmar, antes de se efetivar o experimento, qual será o resultado de uma realização, mas é possível determinar o conjunto de resultados possíveis; quando é realizado um grande número de vezes (replicado) apresentará uma regularidade que permitirá construir um modelo probabilístico para analisar o experimento. Fonte: Adaptado de Lopes (1999).

Variável Aleatória Discreta

– o Espaço Amostral ao qual ela está associada é finito ou infinito numerável. Fonte: Barbetta, Reis e Bornia (2008).

Modelo Binomial

Seja um Experimento Aleatório qualquer que apresente as seguintes características:

- consiste na realização de um número finito e conhecido **n** de ensaios (ou repetições);
- cada um dos ensaios tem apenas dois resultados possíveis: “sucesso” ou “fracasso” (termos entre aspas porque a definição de sucesso não remete necessariamente a algo “positivo” e, também, porque poderá incluir/significar determinado grupo de resultados); e
- os ensaios são independentes entre si, apresentando probabilidades de “sucesso” (**p**) e de “fracasso” (**1-p**) constantes.

Nesse caso estamos interessados no número de “sucessos” obtidos nos **n** ensaios: como o Espaço Amostral é finito (vai de 0 a **n**) uma variável aleatória associada seria discreta. Este tipo de experimento é chamado de Binomial.

Então, a Variável Aleatória Discreta **X**, número de “sucessos” nos **n** ensaios, apresenta uma distribuição (modelo) binomial com os seguintes parâmetros:

n = número de ensaios; e

p = probabilidade de “sucesso”.

Com esses dois parâmetros é possível calcular as probabilidades de determinado número de sucessos, bem como obter o Valor Esperado e a Variância da variável **X**:

$$E(X) = n \times p \quad V(X) = n \times p \times (1-p)$$

Exemplo 1: Experimentos binomiais:

- Observar o número de caras em três lançamentos imparciais de uma moeda honesta: $n=3$; $p=0,5$
- Observar o número de meninos nascidos em três partos de uma mesma família: $n=3$; $p = x$
- Observar o número de componentes defeituosos em uma amostra de dez componentes de um grande número de peças que apresentaram anteriormente 10% de peças defeituosas: $n = 10$; $p= 0,1$

Vamos ver com mais detalhes o caso do número de meninos (e meninas) nascidos em uma família. Chamando menino de evento H, será o “sucesso”, e menina de evento M. Sabendo-se, pela história da família, que:

$$P(H) = 0,52 \text{ e } P(M) = 0,48 \text{ (então } p = 0,52 \text{ e } 1-p = 0,48)$$

Perguntamos: quais as probabilidades obtidas para a variável aleatória número de meninos em três nascimentos? Vamos obter a distribuição de probabilidades.

Tentando a resolução por meio dos conceitos gerais de probabilidade, é preciso, inicialmente, determinar o Espaço Amostral, quais serão os sexos das três crianças:

$$\Omega = \{H \cap H \cap H, H \cap H \cap M, H \cap M \cap H, M \cap H \cap H, H \cap M \cap M, M \cap H \cap M, M \cap M \cap H, M \cap M \cap M\}.$$

Suposto que os nascimentos sejam independentes, podemos calcular as probabilidades de cada intersecção simplesmente multiplicando as probabilidades individuais de seus componentes:

$$\begin{aligned} P\{H \cap H \cap H\} &= P(H) \times P(H) \times P(H) = p \times p \times p = p^3 \\ P\{H \cap H \cap M\} &= P(H) \times P(H) \times P(M) = p \times p \times (1-p) = p^2 \times (1-p) \\ P\{H \cap M \cap H\} &= P(H) \times P(M) \times P(H) = p \times (1-p) \times p = p^2 \times (1-p) \\ P\{M \cap H \cap H\} &= P(M) \times P(H) \times P(H) = (1-p) \times p \times p = p^2 \times (1-p) \\ P\{H \cap M \cap M\} &= P(H) \times P(M) \times P(M) = p \times (1-p) \times (1-p) = p \times (1-p)^2 \\ P\{M \cap H \cap M\} &= P(M) \times P(H) \times P(M) = (1-p) \times p \times (1-p) = p \times (1-p)^2 \\ P\{M \cap M \cap H\} &= P(M) \times P(M) \times P(H) = (1-p) \times (1-p) \times p = p \times (1-p)^2 \\ P\{M \cap M \cap M\} &= P(M) \times P(M) \times P(M) = (1-p) \times (1-p) \times (1-p) = (1-p)^3 \end{aligned}$$

Em qualquer livro de Matemática do ensino médio é possível encontrar a definição e os exemplos de combinações.

Observe que:

$P\{H \cap H \cap M\} = P\{H \cap M \cap H\} = P\{M \cap H \cap H\} = p^2 \times (1 - p) =$ Probabilidade de 2 “sucessos”.

$P\{H \cap M \cap M\} = P\{M \cap H \cap M\} = P\{M \cap M \cap H\} = p \times (1 - p)^2 =$ Probabilidade de 1 “sucesso”.

Importa apenas a “natureza” dos sucessos, não a ordem em que ocorrem: com a utilização de **combinações** é possível obter o número de resultados iguais para cada número de sucessos. Suposto que o número de ensaios n é o número de “objetos” disponíveis, e que o número de “sucessos” em que estamos interessados (doravante chamado k) é o número de “espaços” onde colocar os objetos (um objeto por espaço), o número de resultados iguais será:

$$C_{n,k} = \frac{n!}{k! \times (n-k)!}$$

Para o caso acima, em que há 3 ensaios ($n = 3$):

- para 2 sucessos ($k = 2$) $C_{3,2} = \frac{3!}{2! \times (3-2)!} = 3$ (o mesmo resultado obtido por enumeração); e
- para 1 sucesso ($k = 1$) $C_{3,1} = \frac{3!}{1! \times (3-1)!} = 3$ (o mesmo resultado obtido por enumeração).

O procedimento acima poderia ser feito para quaisquer valores de n e k (desde que $n \geq k$), que permita obter uma expressão geral para calcular a probabilidade associada a um resultado qualquer.

A probabilidade de uma variável aleatória discreta X – número de sucessos em n ensaios com distribuição binomial de parâmetros n e p – assumir certo valor k ($0 \leq k \leq n$) será:

$$P(X = k) = C_{n,k} \times p^k \times (1 - p)^{n-k} \text{ onde } C_{n,k} = \frac{n!}{k! \times (n-k)!}$$

É importante lembrar que a probabilidade de ocorrer k sucessos é igual à probabilidade de ocorrer $n - k$ fracassos, e que todos os axiomas e propriedades de probabilidade continuam válidos.

Exemplo 2: admitamos que a probabilidade de a companhia não entregar seus produtos no prazo é igual a 18%. Nesse caso, quais

as probabilidades de as 3 entregas (1, 2 ou as 3 entregas) serem feitas no prazo? Calcule também o valor esperado, variância e desvio padrão do número de entregas no prazo.

Para cada entrega (“ensaio”) há apenas dois resultados: no prazo ou não. Há um número limitado de realizações, $n = 3$. Definindo “sucesso” como no prazo, e supondo as operações independentes, a variável aleatória X , número de entregas no prazo em 3, teremos distribuição binomial com parâmetros

$$n = 3 \text{ e } p = 0,82 \text{ (e } 1-p = 0,18).$$

Então:

$$P(X = 0) = C_{3,0} \times 0,82^0 \times (0,18)^3 = \frac{3!}{0! \times (3-0)!} \times 0,82^0 \times (0,18)^3 = 0,006$$

$$P(X = 1) = C_{3,1} \times 0,82^1 \times (0,18)^2 = \frac{3!}{1! \times (3-1)!} \times 0,82^1 \times (0,18)^2 = 0,080$$

$$P(X = 2) = C_{3,2} \times 0,82^2 \times (0,18)^1 = \frac{3!}{2! \times (3-2)!} \times 0,82^2 \times (0,18)^1 = 0,363$$

$$P(X = 3) = C_{3,3} \times 0,82^3 \times (0,18)^0 = \frac{3!}{3! \times (3-3)!} \times 0,82^3 \times (0,18)^0 = 0,551$$

Somando-se todas as probabilidades, o resultado é **igual a 1**, como deveria ser. O Valor Esperado, Variância e o Desvio Padrão serão:

$$E(X) = n \times p = 3 \times 0,82 = 2,46 \text{ entregas.}$$

$$V(X) = n \times p \times (1-p) = 3 \times 0,82 \times 0,18 = 0,4428 \text{ entregas}^2.$$

$$\sigma(X) = \sqrt{V(X)} = \sqrt{0,4428} = 0,665 \text{ entregas.}$$

A média é quase igual ao número de operações devido à alta probabilidade de sucesso.

Lembre-se que a soma das probabilidades de todos os eventos que compõem o Espaço Amostral é igual a 1. E que $0! = 1$, e que um número diferente de 0 elevado a zero é igual a 1.

Mas, se o Espaço Amostral fosse infinito numerável? Teríamos que usar o modelo de Poisson. Você conhece esse modelo? Sabe como tirar proveito de suas facilidades? Juntos, vamos aprender mais!

Modelo de Poisson

Vamos supor um experimento binomial, com apenas dois resultados possíveis, mas com a seguinte característica: apesar da probabilidade **p** ser constante o valor de **n** teoricamente é infinito.

Para tal situação o modelo binomial não poderá ser utilizado, o modelo de Poisson é o indicado.

Como seria a solução para o caso acima?

Como **n** é “infinito”, devemos fazer a análise das ocorrências em um período contínuo (de tempo, de espaço, entre outros) subdividido em certo número de subintervalos (número tal que a probabilidade de existir mais de uma ocorrência em uma subdivisão é desprezível, e supondo ainda que as ocorrências em subdivisões diferentes são independentes); novamente é preciso trabalhar com uma quantidade constante que será chamada de **m** também:

$$m = \lambda \times t$$

onde **λ** é uma taxa de ocorrência do evento em um período contínuo (igual ou diferente do período sob análise), e **t** é justamente o período contínuo sob análise.

Apesar do símbolo **t**, o período contínuo não é necessariamente um intervalo de tempo.

Como obter a taxa **λ**? Há duas opções: realizar um número suficiente de testes de laboratório para a obtenção da taxa de ocorrência do evento a partir dos resultados, ou observar dados históricos e calcular a taxa.

Se uma variável aleatória discreta **X**, número de ocorrências de um evento, segue a distribuição de Poisson, a probabilidade de **X** assumir um valor **k** será:

$$P(X = k) = \frac{e^{-m} \times m^k}{k!}$$

Onde **e** é uma constante: **e** $\cong$ 2,71. E **m** = **n** $\times$ **p** ou **λ** $\times$ **t**.

Uma particularidade interessante da distribuição de Poisson é que o Valor Esperado e a Variância de uma variável aleatória que siga tal distribuição serão iguais:

$$E(X) = m = \lambda \times t$$

$$V(X) = m = \lambda \times t$$

O Modelo de Poisson é muito utilizado para modelar fenômenos envolvendo filas: filas de banco, filas de mensagens em um servidor, filas de automóveis em um cruzamento. Vejamos, no exemplo 3, os experimentos e fenômenos que seguem a distribuição de Poisson:

- a) Número mensal de acidentes de trânsito em um cruzamento.

Observe que é uma variável aleatória discreta, que pode assumir apenas valores inteiros (0, 1, 2, 3,...). Cada realização do “experimento” (acidente) pode ter apenas dois resultados: ocorre o acidente ou não ocorre o acidente. Mas, o número máximo de realizações é desconhecido! Assim, a distribuição binomial não pode ser usada, e a análise do número de acidentes precisa ser feita em um período contínuo (no caso, período de tempo: 1 mês), exigindo o uso da distribuição de Poisson.

- b) Número de itens defeituosos produzidos por hora em uma indústria.

Novamente, uma variável aleatória discreta (valores inteiros: 0, 1, 2, 3, ...), cada realização só pode ter dois resultados possíveis (peça sem defeito ou peça defeituosa). Se o número máximo de realizações for conhecido, provavelmente a probabilidade de uma peça ser defeituosa será reduzida e apesar de ser possível a utilização da distribuição binomial o uso da distribuição de Poisson obterá resultados muito próximos. Se o número máximo de realizações for desconhecido, a distribuição binomial não pode ser usada e a análise do número de acidentes precisa ser feita em um período contínuo (no caso, período de tempo: 1 hora), exigindo o uso da distribuição de Poisson.

- c) Desintegração dos núcleos de substâncias radioativas: contagem do número de pulsações radioativas a intervalos de tempo fixos.

Situação semelhante a dos acidentes em um cruzamento, só que o “grau de aleatoriedade” desse experimento é muito maior. O número máximo de pulsações também é desconhecido, obrigando a realizar a análise em um período contínuo, utilizando a distribuição de Poisson.

Exemplo 4: uma telefonista recebe cerca de 0,20 chamadas por minuto (valor obtido de medições anteriores).

- Qual a probabilidade de receber exatamente 5 chamadas nos primeiros 10 minutos?
- Qual a probabilidade de receber até 2 chamadas nos primeiros 12 minutos?
- Qual o desvio padrão do número de chamadas em meia hora?

Há interesse no número de chamadas ocorridas em um período contínuo (de tempo no caso). Para cada “ensaio” há apenas dois resultados possíveis: a chamada ocorre ou não. Observe que não há limite para o número de chamadas no período (sabemos apenas que o número mínimo pode ser 0): por esse motivo a utilização da binomial é inviável... Contudo, há uma taxa de ocorrência (taxa de ocorrência $\lambda = 0,20$ chamadas/minuto) e isso permite utilizar a distribuição de Poisson.

- Nesse caso, o período t será igual a 10 minutos ($t = 10$ min.), $P(X = 5)$?

$$m = \lambda \times t = 0,20 \times 10 = 2 \text{ chamadas}$$

$$P(X = k) = \frac{e^{-m} \times m^k}{k!} = P(X = 5) = \frac{e^{-2} \times 2^5}{5!} = 0,0361$$

Então a probabilidade de a telefonista receber exatamente 5 chamadas em 10 minutos é igual a 0,0361 (3,61%).

- Nesse caso, o período t será igual a 12 minutos ($t = 12$ min.). O evento de interesse é até 2 chamadas em 12 minutos ($X \leq 2$).

$$m = \lambda \times t = 0,20 \times 12 = 2,4 \text{ chamadas}$$

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$P(X = 0) = \frac{e^{-2,4} \times 2,4^0}{0!} = 0,0907$$

$$P(X = 1) = \frac{e^{-2,4} \times 2,4^1}{1!} = 0,2177$$

$$P(X = 2) = \frac{e^{-2,4} \times 2,4^2}{2!} = 0,2613$$

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0,0907 + 0,2177 + 0,2613 = 0,5697$$

Então, a probabilidade de a telefonista receber até 2 chamadas em 12 minutos é igual a 0,5697 (56,97%).

- c) Nesse caso, o período t será igual a 30 minutos ($t = 30$ minutos). Primeiro calcula-se a variância:

$$V(X) = m = \lambda \times t = 0,2 \times 30 = 6 \text{ chamadas}^2$$

O Desvio Padrão é a raiz quadrada positiva da variância:

$$\sigma(X) = \sqrt{V(X)} = \sqrt{6} \cong 2,45 \text{ entregas.}$$

Há vários outros modelos para variáveis aleatórias discretas: hipergeométrico, geométrico, binomial negativo.

Na próxima seção vamos ver os principais modelos de variáveis aleatórias contínuas.

Modelos para Variáveis Aleatórias Contínuas

Nesta seção, estudaremos os modelos uniforme, normal, t e qui-quadrado.

Modelo Uniforme

Quando o Espaço Amostral associado a um Experimento Aleatório é infinito torna-se necessário o uso de uma Variável Aleatória Contínua para associar números reais aos resultados. Os modelos probabilísticos vistos anteriormente não podem ser empregados: a probabilidade de uma variável aleatória contínua assumir exatamente determinado valor é **zero**.

Para entender melhor a declaração acima, vamos relembrar a definição clássica de probabilidade: a probabilidade de ocorrência de um evento será igual ao quociente entre o número de resultados associados ao evento pelo número total de resultados possíveis. Ora, se o número total de resultados é infinito, ou tende ao infinito para ser mais exato,

a probabilidade de ocorrência de um valor específico é igual a zero. Por esse motivo, quando se lida com Variáveis Aleatórias Contínuas, calculamos a probabilidade de ocorrência de eventos formados por intervalos de valores, através de uma função densidade de probabilidades (ver Unidade 6). Outra consequência disso é que os símbolos: $>$, $\geq$, $<$ e $\leq$ são equivalentes para variáveis aleatórias contínuas.

O modelo mais simples para variáveis aleatórias contínuas é o modelo uniforme.

Seja uma variável aleatória contínua qualquer X que possa assumir valores entre A e B . Todos os valores entre A e B têm a mesma probabilidade de ocorrer, resultando no gráfico apresentado na Figura 40:

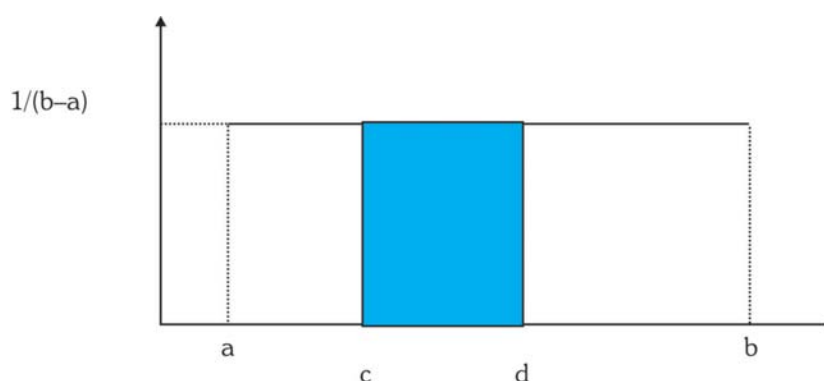


Figura 40: Modelo uniforme

Fonte: Elaborada pelo autor deste livro

Para que a área entre a e b seja igual a 1, o valor da ordenada precisa ser igual a $1/(b - a)$, constante; portanto, para todo o intervalo. A área escura representa a probabilidade da variável X assumir valores no intervalo $c - d$. Trata-se do modelo uniforme.

Dois intervalos de valores da variável aleatória contínua, que tenham o mesmo tamanho, têm a mesma probabilidade de ocorrer (desde que dentro da faixa de valores para os quais a função de densidade de probabilidades não é nula). Formalmente, uma variável aleatória contínua X tem distribuição uniforme, com parâmetros a e b

reais (sendo **a** menor que **b**), se sua função densidade de probabilidades for como a da Figura 40.

A probabilidade de a variável assumir valores entre **c** e **d** (sendo **a** < **c** < **d** < **b**) é a área compreendida entre **c** e **d**:

$$P(c < X < d) = (d - c) \times \frac{1}{(b - a)}$$

Seu valor esperado e variância são:

$$E(X) = \frac{a + b}{2} \quad V(X) = \frac{(b - a)^2}{12}$$

Intuitivamente, podemos supor que muitas variáveis aleatórias contínuas terão um comportamento diferente do caso acima: em algumas delas haverá maior probabilidade de ocorrências de valores próximos ao limite inferior ou superior: para cada caso deverá ser ajustado um modelo probabilístico contínuo adequado.

O modelo uniforme é bastante usado para gerar números pseudo-aleatórios em processos de amostragem probabilística.

Agora vamos passar ao modelo mais importante para variáveis aleatórias contínuas.

No ambiente virtual, temos um exemplo (5) resolvido de modelo uniforme, adaptado de BUSSAB, Wilton O.; MORETTIN, Pedro A. *Estatística Básica*. 5. ed. São Paulo: Saraiva, 2003.

Modelo Normal

Há casos em que há maior probabilidade de ocorrência de valores situados em intervalos centrais da função densidade de probabilidades da variável aleatória contínua, e essa probabilidade diminui, à medida que os valores se afastam desse centro (para valores menores ou maiores) e, nesses casos, o modelo probabilístico contínuo mais adequado corresponde ao modelo Normal ou gaussiano.

Isso é especialmente encontrado em variáveis biométricas, resultantes de medidas corpóreas em seres vivos.

O Modelo Normal é extremamente adequado para medidas numéricas em geral, descrevendo vários fenômenos, e permitindo fazer aproximações de modelos discretos. É extremamente importante também para a Estatística Indutiva (mais detalhes na próxima Unidade). O gráfico da distribuição de probabilidades de uma variável ale-

O matemático alemão Gauss utilizou amplamente este modelo no tratamento de erros experimentais, embora não tenha sido o seu “descobridor”.

atória contínua que siga o modelo Normal (distribuição Normal) será como a Figura 41:

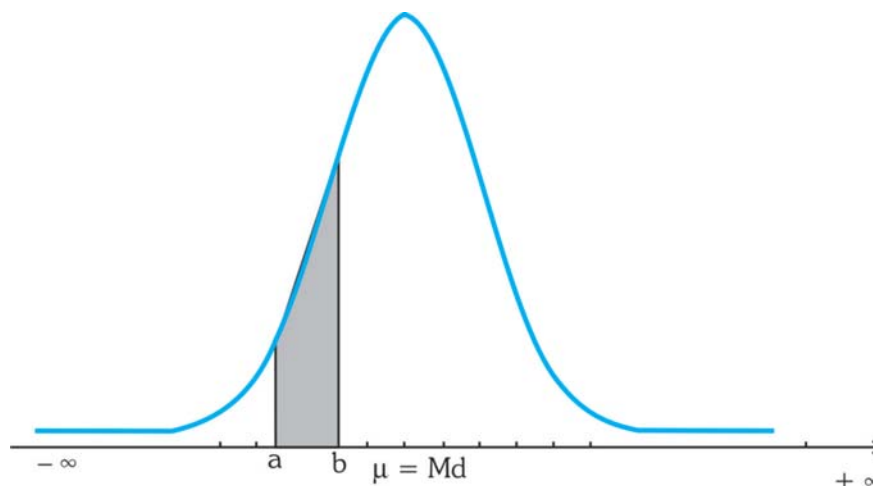


Figura 41: Distribuição normal
Fonte: Elaborada pelo autor deste livro

Características:

A curva apresenta forma de sino, há maior probabilidade de a variável assumir valores próximos do centro.

Os valores de média (μ) e de mediana (**Md**) são iguais, significando que a curva é simétrica em relação à média.

Teoricamente a curva prolonga-se de $-\infty$ a $+\infty$ (menos infinito a mais infinito), então a área total sob a curva é igual a 1 (100%).

Qualquer distribuição normal é perfeitamente especificada por seus parâmetros média (μ) e **variância** (σ^2) $\Rightarrow X: N(\mu, \sigma^2)$. Significa que a variável X tem distribuição normal com média μ e variância σ^2 .

A área escura na Figura 41 é a probabilidade de uma variável que siga a distribuição normal assumir valores entre **a** e **b**: esta área é calculada através da integral da função normal de **a** a **b**.

Cada combinação (μ, σ^2) resulta em uma distribuição Normal diferente, portanto há uma família infinita de distribuições.

A função normal citada acima tem a seguinte (e aterradora!) fórmula para sua função densidade de probabilidade:

$$f(x) = \frac{1}{\sqrt{2 \times \pi \times \sigma^2}} \times e^{\left(\frac{-1}{2} \times \left[\frac{x - \mu}{\sigma} \right]^2 \right)} \quad -\infty < x < +\infty$$

É comum a utilização de letras do alfabeto grego para representar algumas medidas. Não se esqueça de que o desvio padrão (σ) é a raiz quadrada positiva da variância.

Saiba que não existe solução analítica para uma integral da expressão acima: qualquer integral precisa ser resolvida empregando-se métodos numéricos de integração, que são extremamente trabalhosos quando implementados manualmente (somente viáveis se usarem meios computacionais). De Moivre, Laplace e Gauss desenvolveram seus trabalhos entre a metade do século XVIII e início do século XIX, e os computadores começaram a se popularizar a partir da década de 60, do século XX.

Porém, todas as distribuições normais apresentam algumas características em comum, independentemente de seus valores de média e de variância:

- 68% dos dados estão situados entre a média menos um desvio padrão ($\mu - \sigma$) e a média mais um desvio padrão ($\mu + \sigma$);
- 95,5% dos dados estão situados entre a média menos dois desvios padrões ($\mu - 2\sigma$) e a média mais dois desvios padrões ($\mu + 2\sigma$);
- 99,7% dos dados estão situados entre a média menos três desvios padrões ($\mu - 3\sigma$) e a média mais três desvios padrões ($\mu + 3\sigma$).

Gauss, e todas as outras pessoas que usavam a distribuição Normal para calcular probabilidades até recentemente resolviam as integrais com o emprego de métodos numéricos manualmente.

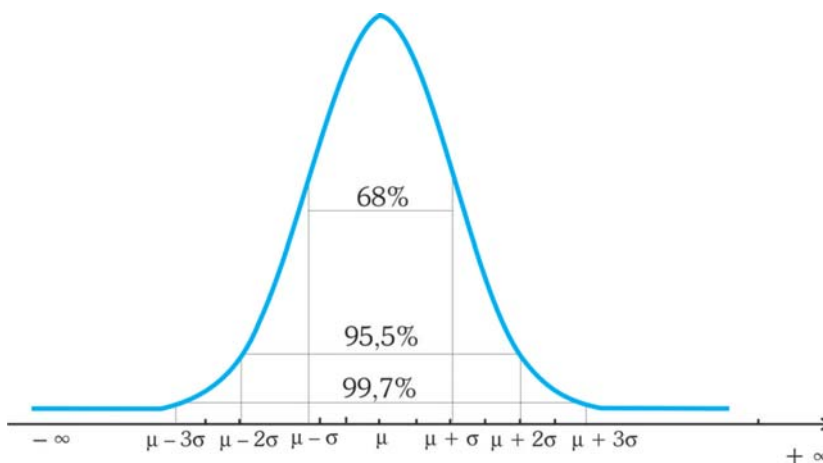


Figura 42: Percentuais de dados e número de desvios padrões

Fonte: Elaborada pelo autor deste livro

Por causa dessas características alguém teve a ideia de criar um modelo normal padrão: uma variável Z com distribuição normal de média igual a zero e desvio padrão igual a 1 [$Z: N(0, 1)$]. As probabilidades foram calculadas para essa distribuição padrão e registradas

em uma tabela. Através de uma transformação de variáveis chamada padronização é possível converter os valores de qualquer distribuição Normal em valores da distribuição Normal padrão e assim obter suas probabilidades. Calcular o número de desvios padrões, a contar da média que está um valor da variável, através da seguinte expressão:

$$Z = \frac{x - \mu}{\sigma}$$

Z – número de desvios padrões a partir da média;

x – valor de interesse;

μ – média da distribuição normal de interesse; e

σ – desvio padrão da distribuição normal.

Z é um valor relativo: será negativo para valores de **x** menores do que a média e positivo para valores de **x** maiores do que a média. Pela transformação, uma distribuição Normal qualquer **X: N (μ , σ^2)** passa a ser equivalente à distribuição Normal padrão **Z: N(0,1)**, um valor de interesse **x** pode ser convertido em um valor **z**.

As probabilidades de uma variável com distribuição normal podem ser representadas por áreas sob a curva da distribuição normal padrão. No ambiente virtual, apresentamos a Tabela, que relaciona valores positivos de **z**, com áreas sob a cauda superior da curva. Os valores de **z** são apresentados com duas decimais. A primeira decimal fica na coluna da esquerda e a segunda decimal na linha do topo da tabela. A Figura 43 mostra como podemos usar essa tabela para encontrar, por exemplo, a área sob a cauda superior da curva, além de $z = 0,21$.



Figura 43: Ilustração do uso da tabela da distribuição normal padrão (Tabela III do apêndice) para encontrar a área na cauda superior relativa ao valor de $z = 0,21$

Fonte: Barbetta, Reis e Bornia (2008)

Exemplo 6: suponha uma variável aleatória X com média 50 e desvio padrão 10. Há interesse em calcular a probabilidade do evento $X > 55$.

Primeiro, calculamos o valor de Z correspondente a 55. $Z = (55 - 50) / 10 = + 0,5$.

Pelas Figuras 44 e 45 podemos ver a correspondência entre as duas distribuições:

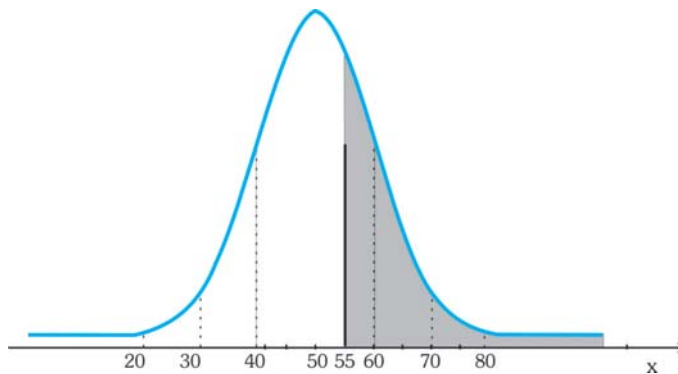


Figura 44: Distribuição Normal $N(50, 102)$

Fonte: Elaborada pelo autor deste livro

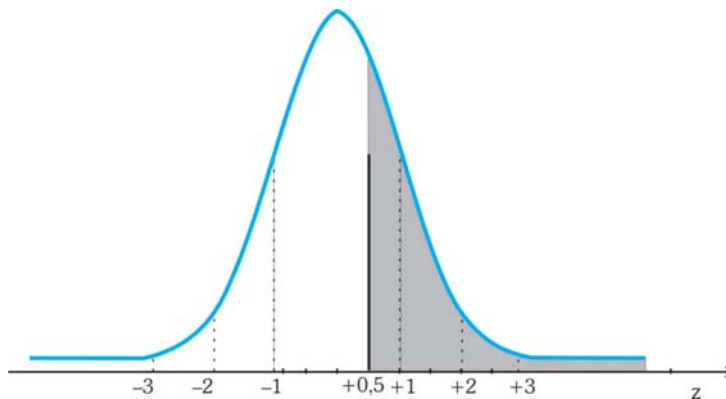


Figura 45: Distribuição normal padrão

Fonte: Elaborada pelo autor deste livro

O evento $P(X > 55)$ é equivalente ao evento $P(Z > 0,5)$. Esse valor pode ser obtido na tabela da distribuição normal padrão (ver ambiente virtual). Os valores de Z são apresentados com dois decimais: o primeiro na coluna da extrema esquerda e o segundo na linha do topo da tabela. Observe, pelas Figuras que estão no alto da tabela, que as probabilidades são para eventos equivalentes aos da Figura acima [$P(Z > z_1)$]. Assim, poderíamos procurar a probabilidade do evento ($Z > 0,5$): fazendo o cruzamento do valor 0,5 (na coluna) com o

valor 0,00 (na linha do topo) encontramos o valor 0,3085 (30,85%). Portanto, $P(X > 55)$ é igual a 0,3085. Observe a coerência entre o valor encontrado e as áreas nas Figuras: a área é menor do que a metade das Figuras (metade das Figuras significaria 50%), e a probabilidade encontrada vale 30,85%.

Exemplo 7: suponha a mesma variável aleatória X com média 50 e desvio padrão 10. Agora, há interesse em calcular a probabilidade de que X seja menor do que 40.

Primeiro, precisamos calcular o valor de Z correspondente a 40.

$$Z = (40 - 50) / 10 = -1,00$$

Pelas Figuras 46 e 47, podemos ver a correspondência entre as duas distribuições:

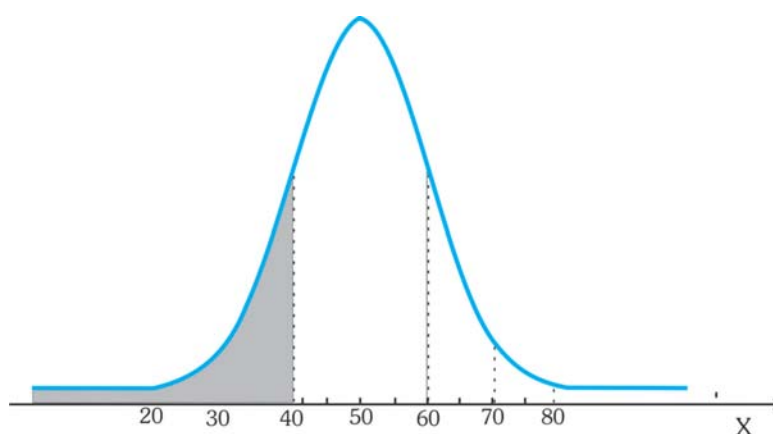


Figura 46: Distribuição Normal $N(50,10^2)$

Fonte: Elaborada pelo autor deste livro

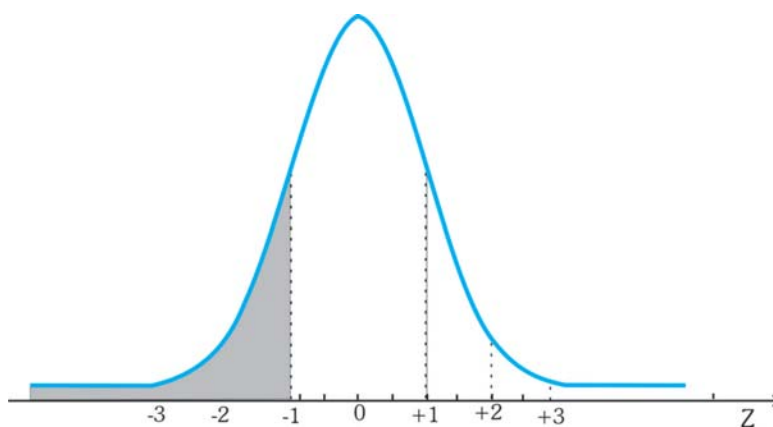


Figura 47: Distribuição normal padrão

Fonte: Elaborada pelo autor deste livro

O evento $P(X < 40)$ é equivalente ao evento $P(Z < -1,00)$. Repare, porém, que queremos encontrar $P(Z < -1,00)$ e a tabela nos

apresenta valores apenas para $P(Z > 1,00)$. Contudo, se rebatermos a Figura da distribuição normal para a direita, teremos o seguinte resultado (Figura 48):

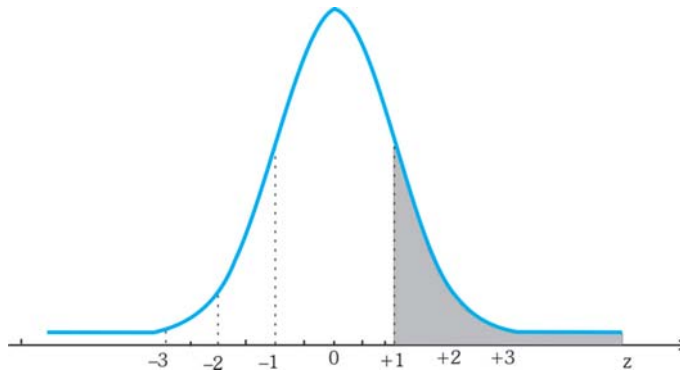


Figura 48: Distribuição normal padrão
Fonte: Elaborada pelo autor deste livro

Ou seja, a área $P(Z < -1) = P(Z > 1)$. Essa probabilidade podemos encontrar diretamente pela tabela. Fazendo o cruzamento do valor 1,0 (na coluna) com o valor 0,00 (na linha do topo), encontraremos o valor 0,1587 (15,87%). Portanto:

$$P(X < 40) = P(Z < -1) = P(Z > 1), \text{ que é igual a } 0,1587.$$

Exemplo 8: suponha a mesma variável aleatória X com média 50 e desvio padrão 10. Agora há interesse em calcular a probabilidade de X ser maior do que 35.

Primeiro, precisamos calcular o valor de Z correspondente a 35.

$$Z = (35 - 50) / 10 = -1,50$$

Pelas Figuras 49 e 50, podemos ver a correspondência entre as duas distribuições:

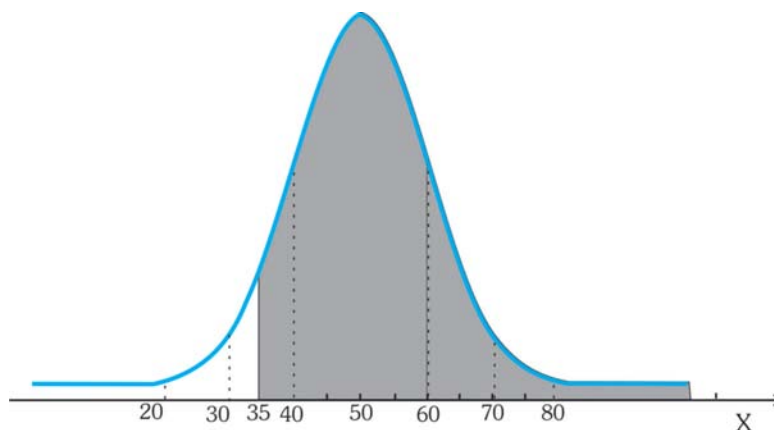


Figura 49: Distribuição Normal $N(50, 102)$
Fonte: Elaborada pelo autor deste livro

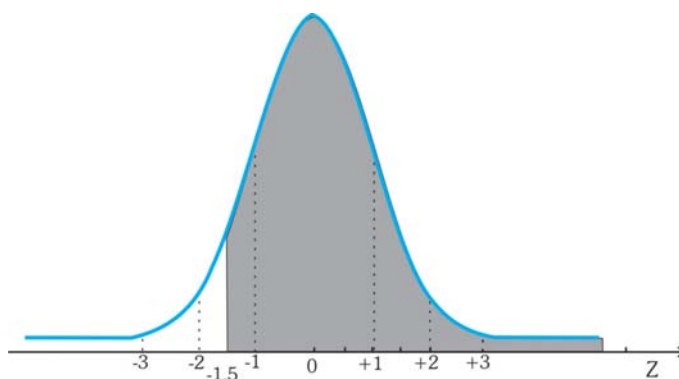


Figura 50: Distribuição normal padrão

Fonte: Elaborada pelo autor deste livro

Não podemos obter a probabilidade $P(Z > -1,50)$ diretamente, pois a tabela do Ambiente Virtual apresenta apenas resultados para valores positivos de Z . Sabemos que a probabilidade total vale 1,0; podemos, então, considerar:

$$P(Z > -1,50) = 1 - P(Z < -1,50)$$

Adotando-se o raciocínio descrito no Exemplo 7 (rebatendo as Figuras para à direita), obteremos: $P(Z < -1,50) = P(Z > 1,50)$

Esta última probabilidade pode ser facilmente encontrada na tabela da distribuição normal padrão: $P(Z > 1,50) = P(Z < -1,50) = 0,0668$

Basta substituir na expressão:

$$P(Z > -1,50) = 1 - P(Z < -1,50) = 1 - 0,0668 = 0,9332 \text{ (93,32\%)}$$

Observe novamente a coerência entre as áreas das Figuras acima e o valor da probabilidade: a área nas Figuras compreende mais do que 50% da probabilidade total, aproximando-se do extremo inferior da distribuição, perto de 100%; e a probabilidade encontrada realmente é próxima de 100%.

Exemplo 9: suponha a mesma variável aleatória X com média 50 e desvio padrão 10. Agora, há interesse em calcular a probabilidade de X assumir valores entre 48 e 56. Calcular $P(48 < X < 56)$. Veja a Figura 51:

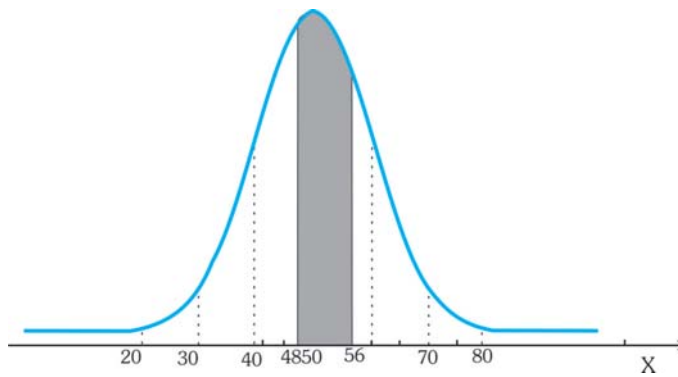


Figura 51: Distribuição normal $N(50, 10^2)$

Fonte: Elaborada pelo autor deste livro

Novamente, precisamos calcular os valores de Z correspondentes a 48 e a 56.

$$Z_1 = (48 - 50)/10 = -0,20 \quad Z_2 = (56 - 50)/10 = 0,60$$

$$\text{Então: } P(48 < X < 56) = P(-0,20 < Z < 0,60)$$

Repare que a área entre 48 e 56 é igual à área de 48 até + MENOS a área de 56 até $+\infty$:

$$P(48 < X < 56) = P(X > 48) - P(X > 56) = P(-0,20 < Z < 0,60) = P(Z > -0,20) - P(Z > 0,60)$$

E os valores acima podem ser obtidos na tabela da distribuição normal padrão:

$$P(Z > 0,60) = 0,2743$$

$$P(Z > -0,20) = 1 - P(Z > 0,20) = 1 - 0,4207 = 0,5793$$

$$P(48 < X < 56) = P(-0,20 < Z < 0,60) = P(Z > -0,20) - P(Z > 0,60) = 0,5793 - 0,2743 = 0,3050$$

Então, a probabilidade da variável X assumir valores entre 48 e 56 é igual a 0,305 (30,5%).

A distribuição Normal também pode ser utilizada para encontrar valores da variável de interesse correspondentes a uma probabilidade fixada.

Exemplo 10: suponha a mesma variável aleatória X com média 50 e desvio padrão 10. Encontre os valores de X , situados à mesma distância abaixo e acima da média, que contém 95% dos valores da variável.

Como a distribuição Normal é simétrica em relação à média e, como nesse problema, os valores de interesse estão situados à mesma distância da média, “sobram” 5% dos valores; 2,5% na cauda inferior e 2,5% na superior, como na Figura 52:

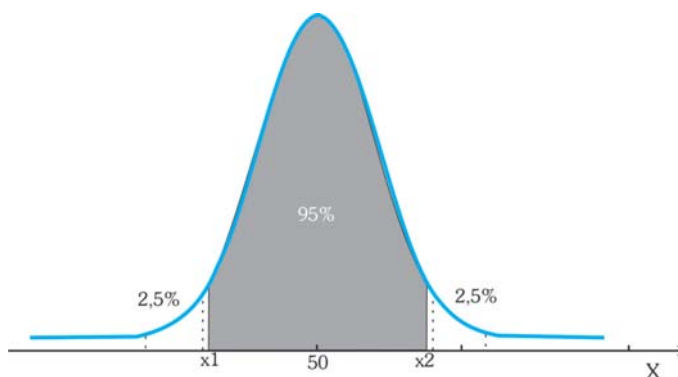


Figura 52: Distribuição normal $N(50, 10^2)$

Fonte: Elaborada pelo autor deste livro

É preciso encontrar os valores de Z (na tabela da distribuição Normal padrão) correspondentes às probabilidades da Figura acima e, a partir daí, obter os valores de x_1 e x_2 . Passando para a distribuição Normal padrão x_1 , corresponderá a um valor z_1 , e x_2 a um valor z_2 , como na Figura 53:

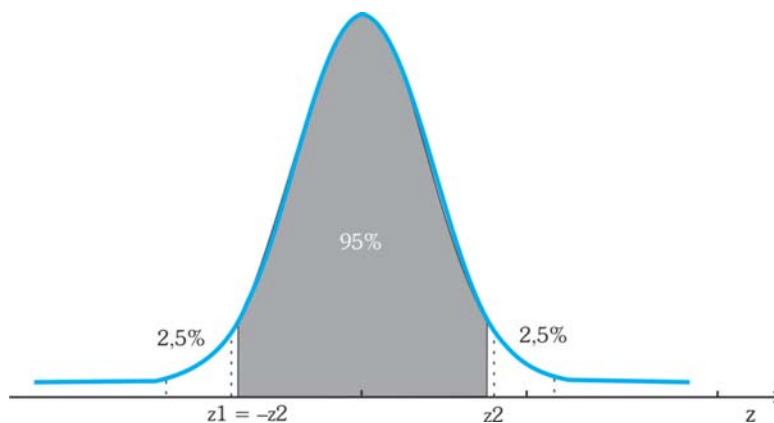


Figura 53: Distribuição normal padrão

Fonte: Elaborada pelo autor deste livro

Repare que a média da distribuição Normal padrão é igual a zero, fazendo com que z_1 e z_2 sejam iguais em módulo. Podemos encontrar z_2 , já que $P(Z > z_2) = 0,025$.

É necessário encontrar o valor da probabilidade na tabela da distribuição Normal padrão (ou o valor mais próximo) e obter o valor de Z associado.

Para o caso de z_2 , ao procurarmos pela probabilidade 0,025, encontramos o valor exato 0,025 e, por conseguinte, o valor de z_2 que é igual a 1,96: $P(Z > 1,96) = 0,025$.

Como $z_1 = -z_2$, encontramos facilmente o valor de z_1 : $z_1 = -1,96$. $P(Z < -1,96) = 0,025$.

Observe que os valores são iguais em módulo, mas corresponderão a valores diferentes da variável X . A expressão usada para obter o valor de Z , em função do valor da variável X , pode ser usada para o inverso:

$$Z = \frac{x - \mu}{\sigma} \Rightarrow x = \mu + Z \times \sigma$$

E, desse modo, obteremos os valores de x_1 e x_2 . Observe se o resultado obtido corresponde a z_1 e z_2 , respectivamente:

$$x_1 = \mu + (z_1 \times \sigma) = 50 + [(-1,96) \times 10] = 30,4$$

$$x_2 = \mu + (z_2 \times \sigma) = 50 + (1,96 \times 10) = 69,6$$

Repare que os resultados obtidos são coerentes: 30,4 está abaixo da média (1,96 desvios padrões) e 69,6 acima (também 1,96 desvios padrões). O intervalo definido por esses dois valores compreende 95% dos resultados da variável X .

Todo esse trabalho poderia ter sido evitado se houvesse um programa computacional que fizesse esses cálculos. Há vários *softwares* disponíveis no mercado, alguns deles de domínio público, que calculam as probabilidades associadas a determinados eventos, como também os valores associados a determinadas probabilidades.

Uma das características mais importantes do modelo normal é sua capacidade de se aproximar de outros modelos, permitindo muitas vezes simplificar os cálculos de probabilidade. Na próxima seção vamos ver como o modelo normal pode ser usado para aproximar o binomial.

É muito importante prestar atenção no sinal do valor de z ao obter o valor de x .

Modelo binomial – modelo probabilístico para variáveis aleatórias discretas que descreve o número de sucessos em n experimentos independentes (sendo n finito e conhecido), com os experimentos podendo ter apenas dois resultados possíveis, e a probabilidade de sucesso permanece constante durante os n experimentos. Fonte: Barbetta, Reis e Bornia (2008); Lopes (1999).

Modelo Normal como Aproximação do Binomial

O modelo Binomial (discreto) pode ser aproximado pelo modelo Normal (contínuo) se certas condições forem satisfeitas:

- quando o valor de **n** (número de ensaios) for tal que os cálculos binomiais tornam-se trabalhosos demais.
- quando o produto **n × p** (o valor esperado do modelo Binomial) e o produto **n × (1 – p)** forem ambos maiores ou iguais a 5.

Se isso ocorrer, uma binomial, de parâmetros **n** e **p**, pode ser aproximada por uma normal com:

- **média** = $\mu = n \times p$ (valor esperado do modelo Binomial); e
- **variância** = $\sigma^2 = n \times p \times (1 - p)$ (variância do modelo Binomial).

Adotando-se o modelo Normal (contínuo) para aproximar o Binomial (discreto), é necessário fazer uma correção de continuidade: associar um intervalo ao valor discreto, para que o valor da probabilidade calculada pelo modelo contínuo seja mensurável. Esse intervalo deve ser centrado no valor discreto, e deve ter uma amplitude igual à diferença entre dois valores consecutivos da variável discreta: se, por exemplo, a diferença for igual a 1 (a variável somente pode assumir valores inteiros), o intervalo deve ter amplitude igual a 1 – 0,5 abaixo do valor e 0,5 acima. **Essa correção de continuidade precisa ser feita para garantir a coerência da aproximação.**

Seja uma variável aleatória **X** com distribuição Binomial.

- 1) Há interesse em calcular a probabilidade de **X** assumir um valor **k** genérico, $P(X = k)$. Ao se fazer a aproximação pela Normal, teremos: $P(k - 0,5 < X < k + 0,5)$

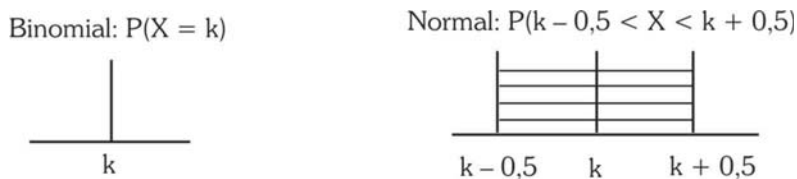


Figura 54: Correção de continuidade da aproximação do modelo Binomial pelo Normal – 1º caso

Fonte: Elaborada pelo autor deste livro

Para os que pensam que o advento dos computadores eliminou esse problema, um alerta: em alguns casos os números envolvidos são tão complexos que sobrepõem a capacidade da máquina (*softwares*).

- 2) Há interesse em calcular a probabilidade de X assumir valores menores ou iguais a um valor k genérico, $P(X \leq k)$. Ao se fazer a aproximação pela Normal, teremos: $P(X < k + 0,5)$, todo o intervalo referente a k será incluído.

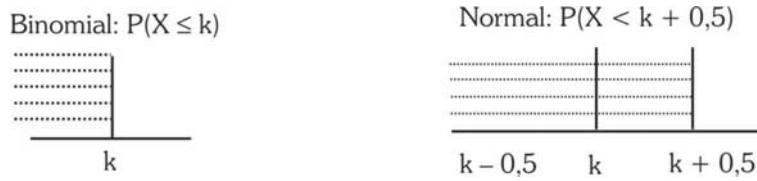


Figura 55: Correção de continuidade da aproximação do modelo binomial pelo Normal – 2º caso

Fonte: Elaborada pelo autor deste livro

- 3) Há interesse em calcular a probabilidade de X assumir valores maiores ou iguais a um valor k genérico, $P(X \geq k)$. Ao se fazer a aproximação pela Normal, teremos: $P(X > k - 0,5)$, todo o intervalo referente a k será incluído.

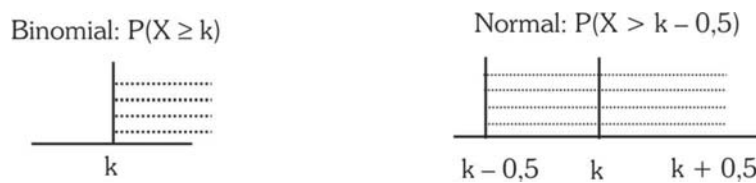


Figura 56: Correção de continuidade da aproximação do modelo binomial pelo Normal – 3º caso

Fonte: Elaborada pelo autor deste livro

- 4) Há interesse em calcular a probabilidade de X assumir valores menores do que um valor k genérico, $P(X < k)$. Ao se fazer a aproximação pela Normal, teremos: $P(X < k - 0,5)$, todo o intervalo referente a k será excluído.

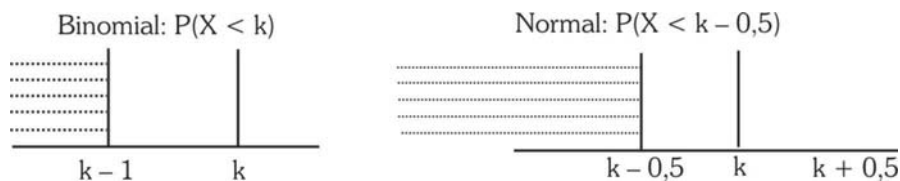


Figura 57: Correção de continuidade da aproximação do modelo binomial pelo Normal – 4º caso

Fonte: Elaborada pelo autor deste livro

- 5) Há interesse em calcular a probabilidade de X assumir valores maiores do que um valor k genérico, $P(X > k)$. Ao se fazer a aproximação pela Normal, teremos: $P(X > k + 0,5)$, todo o intervalo referente a k será excluído.

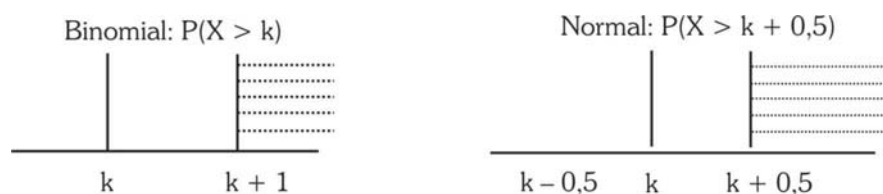


Figura 58: Correção de continuidade da aproximação do modelo binomial pelo Normal – 5º caso

Fonte: Elaborada pelo autor deste livro

Um município tem 40.000 eleitores. Para uma pesquisa de opinião eleitoral uma amostra aleatória de 1.500 pessoas foi selecionada. Vamos ver, no exemplo 11 a seguir, qual a probabilidade de, pelo menos, 500 dos eleitores serem menores de 25 anos se 35% dos 40.000 são menores de 25 anos?

Esse problema poderia ser resolvido empregando-se o modelo Binomial. Há apenas dois resultados possíveis para cada eleitor: menor de 25 anos (“sucesso”) e maior ou igual a 25 anos (“fracasso”). Existe um limite superior de realizações, no caso os 1500 eleitores da amostra, e há independência entre as retiradas, pois a amostra foi extraída de forma aleatória (e a amostra representa menos de 5% dos 40.000 eleitores).

Então: “sucesso” = menor de 25 anos.

$$p = 0,35 \quad 1 - p = 0,65 \quad n = 1.500$$

A variável aleatória discreta X , número de eleitores menores de 25 anos em 1500, terá distribuição binomial com parâmetros $n = 1.500$ e $p = 0,35$.

O evento “pelo menos 500 menores de 25 anos” seria definido como 500 ou mais eleitores:

$$P(X \geq 500) = P(X = 500) + P(X = 501) + \dots + P(X = 1500)$$

Há cerca de 1000 expressões binomiais.

Vamos ver se é possível a aproximação pelo modelo Normal.

O valor de n é grande: $n \times p = 1500 \times 0,35 = 525 > 5$ e $n \times (1 - p) = 1500 \times 0,65 = 975 > 5$.

Como as condições foram satisfeitas, é possível aproximar por um modelo Normal:

$$\text{média} = \mu = n \times p = 1500 \times 0,35 = 525$$

$$\text{desvio padrão} = \sigma = \sqrt{n \times p \times (1-p)} = \sqrt{1.500 \times 0,35 \times 0,65} = 18,47$$

Pelo modelo Binomial: $P(X \geq 500)$

Pelo modelo Normal será: $P(X \geq 499,5)$

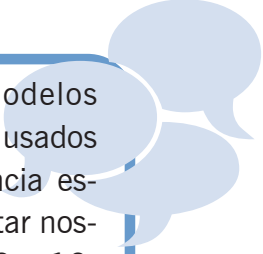
$$P(X \geq 499,5) = P(Z > z_1) \quad z_1 = (499,5 - 525)/18,47 = -1,38$$

$$P(Z > -1,38) = 1 - P(Z > 1,38)$$

Procurando na tabela da distribuição Normal padrão: $P(Z > 1,38) = 0,0838$

$$\text{Então: } P(X \geq 500) \cong P(X \geq 499,5) = P(Z > -1,38) = 1 - P(Z > 1,38) = 1 - 0,0838 = 0,9162$$

A probabilidade de, pelo menos, 500 dos eleitores da amostra serem menores de 25 anos é igual a 0,9162 (91,62%).



Nas próximas duas seções vamos ver modelos probabilísticos derivados do modelo normal, usados predominantemente em processos de inferência estatística. Vamos introduzi-los agora para facilitar nosso trabalho quando chegarmos às Unidades 9 e 10.

Modelo (distribuição) t de *Student*

Havia um matemático inglês, William Gosset, que trabalhava para a cervejaria Guinness, em Dublin, Irlanda, no início do século XX. Ele atuava no controle da qualidade do cultivo de ingredientes para a fabricação de cerveja.

Nessa época alguns estatísticos usavam a distribuição normal no estabelecimento de intervalos de confiança para a média a partir de pequenas amostras, tema que estudaremos na Unidade 8. Eles calculavam média aritmética simples e variância da amostra e generalizavam os resultados através do modelo normal, como fizemos no Exemplo 11.

Gosset descobriu que o modelo normal não funcionava direito para pequenas amostras e desenvolveu um novo modelo probabilístico derivado do normal, introduzindo uma correção para levar em conta justamente o tamanho de amostra. Ele aplicou suas descobertas em

Essa é a correção propriamente dita, pois ao usar pequenas amostras o risco de a variância amostral da variável ser diferente da variância populacional é maior, podendo levar a intervalos de confiança que não correspondem à realidade. A não utilização dessa correção representou a fonte de muitos erros no passado e, infelizmente, de alguns erros no presente.

seu trabalho, e quis publicá-las; a Guinness, no entanto, para tal publicação, exigiu que ele adotasse o pseudônimo de “Student”. Por isso, o seu modelo é conhecido como **t** de *Student* para $n - 1$ graus de liberdade.

O valor $n - 1$ (tamanho da amostra menos 1) é chamado de número de **graus de liberdade** da estatística. Quando a variância amostral é calculada, supõe-se que a média já seja conhecida; assim apenas determinado número de elementos da amostra poderá ter seus valores variando livremente, esse número será igual a $n - 1$, porque um dos valores não poderá variar livremente, pois terá que ter um valor tal que a média permaneça a mesma calculada anteriormente. Assim, a estatística terá $n - 1$ graus de liberdade.

Trata-se de uma distribuição de probabilidades que apresenta média igual a zero (como a normal padrão), é simétrica em relação à média, mas apresenta uma variância igual a $n/(n - 2)$, ou seja, seus valores dependem do tamanho da amostra e apresenta maior variância para menores valores desta. Quanto maior o tamanho da amostra, mais a variância de **t** se aproxima de 1,00 (variância da normal padrão). A distribuição **t** de *Student* está na Figura 59, a seguir, para vários graus de liberdade (Para tamanhos de amostra maiores que 30, supõe-se que a variância de **t** seja igual a 1: por isso a aproximação do item b.1.):

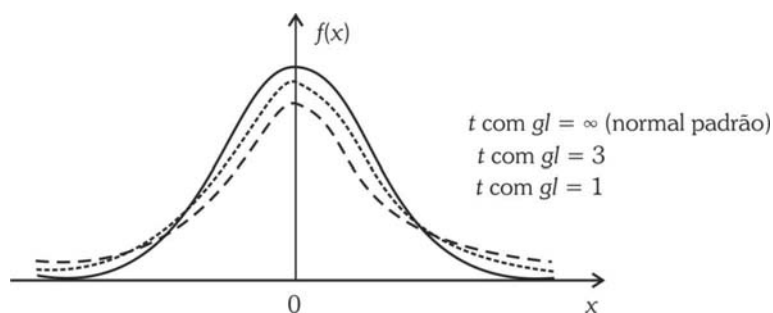


Figura 59: Distribuição **t** de *Student* para vários graus de liberdade
 Fonte: Barbeto, Reis e Bornia (2008)

Observe que, tal como a distribuição normal padrão, a distribuição **t** de *Student* é **simétrica** em relação à média (que é igual a zero).

A tabela da distribuição **t** de *Student* encontra-se no ambiente virtual, para vários graus de liberdade e valores de probabilidade.

Vamos ver um exemplo.

Exemplo 11: imagine a situação do Exemplo 10 e obtenha os valores de t simétricos em relação à média que contêm 95% dos dados, supondo uma amostra de 10 elementos.

Temos que encontrar os valores t_1 e t_2 , simétricos em relação à média que definem o intervalo que contém 95% dos dados. Como supomos uma amostra de 10 elementos, a distribuição **t de Student** terá $10 - 1 = 9$ graus de liberdade. Repare que a média da distribuição **t de Student** é igual a zero, fazendo com que t_1 e t_2 sejam iguais em módulo. Podemos encontrar t_2 , já que $P(t > t_2) = 0,025$. Veja a Figura 60:

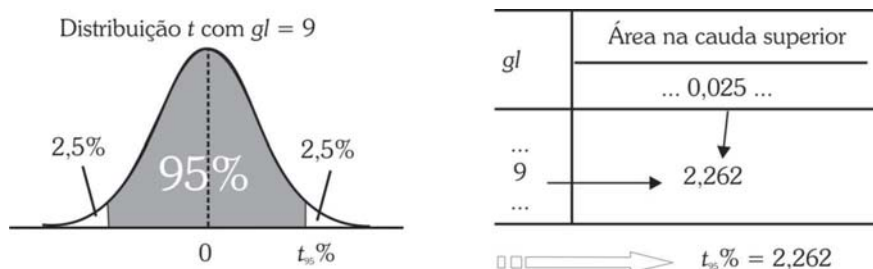


Figura 60: Uso da tabela da distribuição t de Student. Ilustração com $gl = 9$ e área na cauda superior de 2,5%
Fonte: Barbeto, Reis e Bornia (2008)

Vamos utilizar bastante a distribuição **t de Student** nas Unidades 9 e 10.

Modelo Qui-quadrado

Trata-se de mais um modelo derivado da distribuição normal e, por questões didáticas, não iremos discutir como se dá essa derivação aqui.

Na Unidade 3, estudamos como descrever os relacionamentos entre duas variáveis qualitativas, geralmente expresso através de uma tabela de contingências. No Exemplo 5, da Unidade 3, analisamos a relação entre modelo e opinião geral sobre os veículos da Toyord. Havíamos concluído que havia relacionamento, pois os modelos mais baratos apresentavam maiores percentuais de insatisfeitos do que os mais caros.

Na Unidade 10, vamos aprender a calcular uma estatística que relacionará as frequências observadas de cada cruzamento entre os valores de duas variáveis qualitativas, expressas em uma tabela de contingências, com as frequências esperadas desses mesmos cruzamentos, se as duas variáveis não tiverem qualquer relação entre si. Essa estatística é chamada de quiquadrado, χ^2 ; e, na hipótese de as variáveis não se relacionarem, ela seguirá o modelo qui-quadrado com certo número de graus de liberdade.

O número de graus de liberdade dependerá das condições da tabela: para o caso que será visto na Unidade 10, será o produto do número de linhas da tabela – 1 pelo número de colunas da tabela – 1. É uma distribuição assimétrica, sempre positiva, que tem valores diferentes dependendo do seu número de graus de liberdade. Sua média é igual ao número de graus de liberdade, e a variância é igual a duas vezes o número de graus de liberdade.

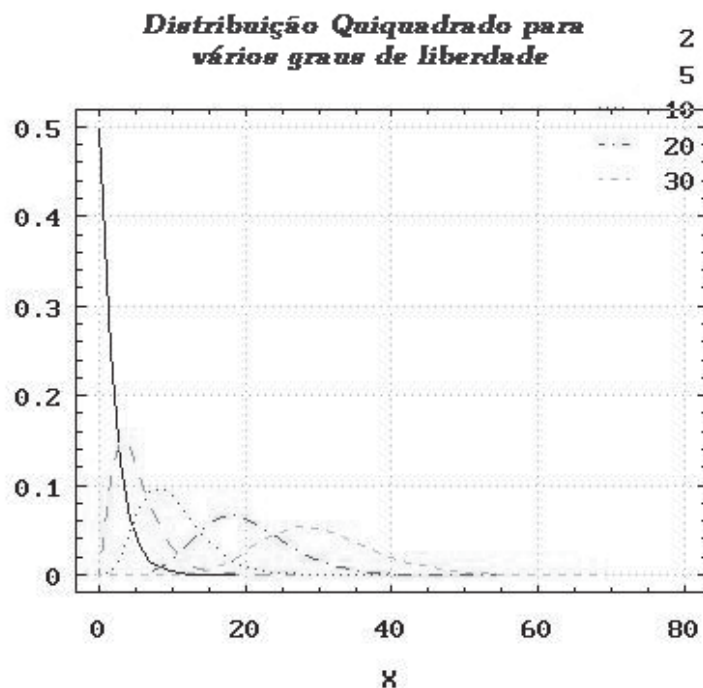


Figura 61: Modelo quiquadrado com 2, 5, 10, 20 e 30 graus de liberdade
Fonte: Elaborada pelo autor deste livro

A Figura 61 mostra as curvas do modelo (distribuição) qui-quadrado para 2, 5, 10, 20 e 30 graus de liberdade. Observe como variam de forma dependendo do número de graus de liberdade da estatística.

A tabela da distribuição quiquadrado encontra-se no Ambiente Virtual de Ensino-Aprendizagem, para vários graus de liberdade e valores de probabilidade. Procuremos entendê-la por meio de mais um exemplo, o 12°.

Imagine que queiramos encontrar o valor da estatística qui-quadrado, para 3 graus de liberdade, e, deixar uma área, na cauda superior, de 5%.

O valor da estatística quiquadrado que define uma área na cauda superior de 5% pode ser encontrado através da Tabela, cruzando a linha de 3 graus de liberdade com a coluna de área na cauda superior igual a 0,05. Veja a Figura a seguir:

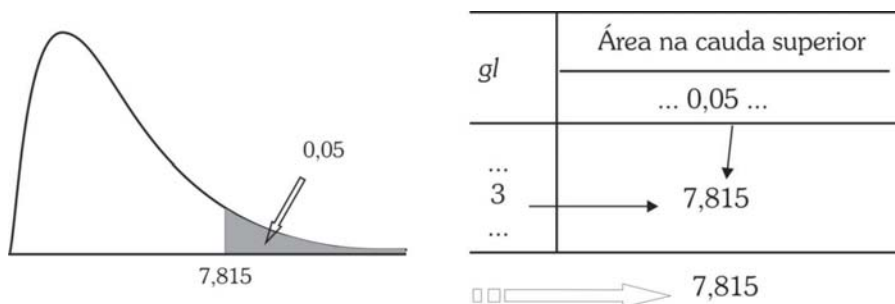


Figura 62: Uso da tabela da distribuição quiquadrado. Ilustração com $gl = 3$ e área na cauda superior de 5%

Fonte: Adaptada de Barbeto, Reis e Bornia (2008)

Com esse tópico, terminamos a Unidade 7. Na Unidade 8, você irá se deparar com o importante conceito de distribuição amostral, que é indispensável para o processo de generalização (inferência) estatística que será estudado nas Unidades 9 e 10.

Saiba mais...

Sobre modelos probabilísticos para variáveis aleatórias discretas:

BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Ed. da UFSC, 2007, capítulo 7; BARBETTA, Pedro A. REIS, Marcelo M., BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, capítulo 5; STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 4.

Sobre modelos probabilísticos para variáveis aleatórias contínuas:

BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Ed. da UFSC, 2007, capítulo 8; BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, capítulo 6; STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 5.

Sobre a utilização do Microsoft Excel® para cálculo de probabilidades para os principais modelos probabilísticos veja: LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística: teoria e aplicações – usando microsoft excel em português*. 5. ed. Rio de Janeiro: LTC, 200, capítulos 4 e 5.

Resumindo



A Figura 63, a seguir, representa o resumo desta Unidade.

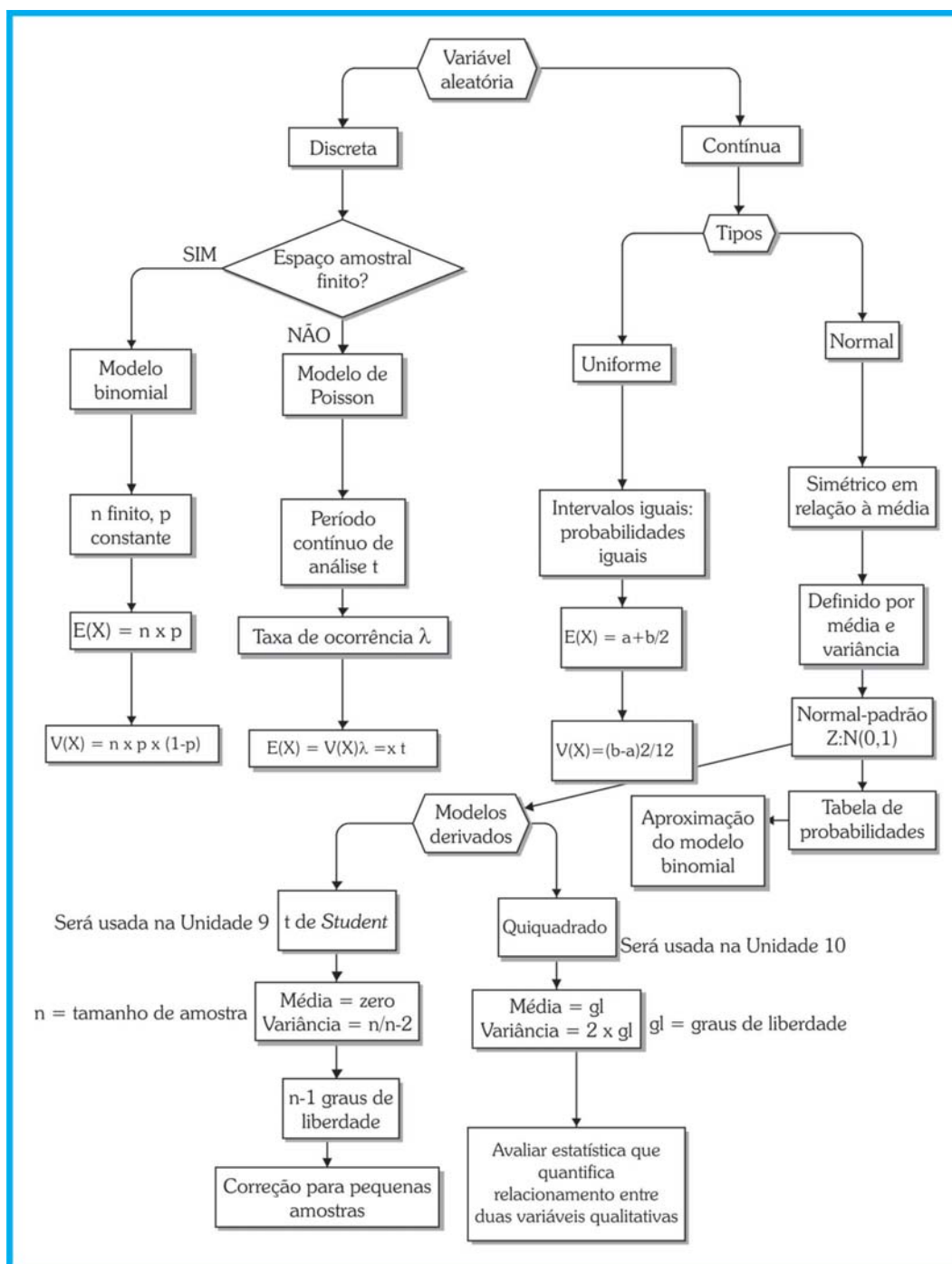


Figura 63: Resumo da Unidade 7
Fonte: Elaborada pelo autor deste livro



Caro estudante,
Chegamos ao final da Unidade 7 do nosso livro. Nela estudamos os modelos probabilísticos mais comuns. Essa Unidade foi repleta de Figuras, Quadros, representações e exemplos de utilização das técnicas e das diferentes formas de utilização destes modelos. Releia, caso necessário, todos os exemplos, leia as indicações do *Saiba mais* e discuta com seus colegas. Responda à atividade de aprendizagem e visite o Ambiente Virtual de Ensino-Aprendizagem. Conte sempre com o acompanhamento da tutoria e das explicações do professor. Ótimos estudos!



Atividades de aprendizagem

1. Em determinado processo de fabricação, 10% das peças são defeituosas. As peças são acondicionadas em caixas com 5 unidades cada uma. As caixas serão aceitas apenas se apresentarem no máximo uma peça defeituosa. Perguntamos:
 - a) Qual o modelo teórico mais adequado para esse caso? Por quê?
 - b) Qual a probabilidade de haver exatamente 3 peças defeituosas em uma caixa?
 - c) Qual a probabilidade de uma caixa ser aceita?
 - d) Qual a probabilidade de, num lote de 10 caixas, pelo menos 8 serem aceitas?
2. Uma comissão, responsável pelo recebimento de equipamentos em uma empresa, faz testes em equipamentos selecionados aleatoriamente, dentre os que chegam. Para avaliar determinada marca de transformadores de pequeno porte, a comissão selecionou aleatoriamente 18 dentre os que chegaram e classificará a marca como satisfatória se não existir nenhum defeituoso nessa amostra. Sabe-se

que a produção desses equipamentos apresenta um percentual de 6% de defeituosos.

- a) Qual a probabilidade de a marca vir a ser considerada satisfatória?
 - b) Qual a probabilidade de uma amostra, no máximo - de um grupo de 8 amostras desses transformadores (cada amostra com 18 transformadores) - ser considerada satisfatória?
3. Uma operadora de pedágios está preocupada com o dimensionamento de uma de suas praças. Muitos motoristas reclamam das filas, pois há apenas duas gôndolas em operação durante o período. Estudos mostraram que em média 4 carros chegam na praça de pedágio a cada 15 minutos.
- a) Qual a probabilidade de mais de 2 carros chegarem à praça em 30 minutos?
 - b) Qual a probabilidade de chegada de até 2 carros em um período de uma hora?
 - c) Você recomenda que a empresa aumente o número de gôndolas? Por quê?
4. Sabe-se que a precipitação anual de chuva, cuja altura é medida em cm, em certa localidade, é uma variável aleatória normalmente distribuída com altura média igual a 29,5 cm e desvio padrão de 2,5 cm de chuva.
- a) Qual a altura de chuva ultrapassada em cerca de 5% das medições?
 - b) Se em mais de 45% das vezes a altura de chuva ultrapassar 32 cm, torna-se viável a instalação de um sistema para coleta e para armazenamento de água da chuva (como complemento à atual malha de abastecimento). É viável instalar o sistema na localidade?
5. O tempo de vida de determinado componente eletrônico distribui-se normalmente com média de 250 horas e variância de 49 horas. Você adquire um desses componentes.
- a) Qual a probabilidade de seu tempo de vida ultrapassar as 260 horas?

- b) Qual deveria ser o prazo de garantia para esses componentes para que o serviço de reposição atendesse a somente 5% dos componentes adquiridos?
6. Imagine que a UFSC tivesse antecipado os resultados a seguir, referentes aos candidatos não eliminados, antes de divulgar a relação com as notas de todos os candidatos.

PONTUAÇÃO FINAL VESTIBULAR UFSC – 2002		
	Economia	Administração
Média	50,92	55,11
Desvio padrão	9,09	8,22
vagas/Candidatos	0,370	0,412

Admitindo-se que as notas são normalmente distribuídas:

- a) O que você responderia para um candidato à Economia que estimasse ter conseguido 50 pontos? Em sua opinião, ele conseguiria se classificar? E se ele estimasse ter conseguido 60 pontos?
- b) O que você responderia para candidatos aos cursos de Economia e Administração que estimassem ter conseguido, respectivamente, 55 e 58 pontos?
- c) Imagine que você tenha de responder a dezenas de vestibulandos. Para poupar trabalho, estime a nota mínima para classificação em cada curso.

8

UNIDADE

Inferência Estatística e Distribuição Amostral



Ao finalizar esta Unidade, você deverá ser capaz de compreender os conceitos de inferência estatística e de distribuição amostral, fundamental para o processo de generalização empregado pelos administradores na tomada de decisão.

Conceito de Inferência Estatística

Caro estudante, vamos relembrar um pouco nossa trajetória ao longo da disciplina de estatística:

Na Unidade 1, vimos que através da **Inferência Estatística**, usando os conceitos de Probabilidade e de variáveis aleatórias, estas últimas estudadas nas Unidades 5, 6 e 7, podemos generalizar os resultados de uma pesquisa por amostragem, assunto tratado na Unidade 2, para a população da qual a amostra foi retirada.

Lembre-se: estamos supondo que a amostra foi retirada por meio de **amostragem probabilística ou aleatória**! Temos, então, um **experimento aleatório**: não sabemos quem fará parte da amostra antes do sorteio, conforme Unidade 5.

Uma vez retirada a amostra, fazemos a análise exploratória dos dados – temas explanados nas Unidades 3 e 4 – e calculamos a média de uma variável quantitativa. Essa média e todas as demais estatísticas serão variáveis aleatórias (pois estão associadas ao **Espaço Amostral** de um experimento aleatório); desse modo, poderemos tentar a identificação do modelo probabilístico mais apropriado para elas – de acordo com nossos estudos das Unidades 6 e 7 – e, nesse caso, o modelo probabilístico de uma estatística da amostra é chamado de **Distribuição Amostral**.

Conhecer a Distribuição Amostral das principais estatísticas será muito útil quando estudarmos os tipos particulares de Inferência Estatística: Estimação de Parâmetros, vistos na Unidade 9, e Testes de Hipóteses, conforme Unidade 10. Continuemos nossos estudos!

Na Unidade 2 enumeramos as principais razões para usar amostragem.

Amostra aleatória, casual ou probabilística

– amostra retirada por meio de um sorteio não viciado, que garante que cada elemento da população terá probabilidade maior do que zero de pertencer à amostra. Fonte: Elaborado pelo autor deste livro.

Parâmetros – alguma medida descritiva (média, variância, proporção) dos valores $x_1, x_2, x_3, \dots$, associados à população. Fonte: Barbetta, Reis e Bornia (2008).

Estimação de Parâmetros

– forma de inferência estatística que busca estimar os parâmetros do modelo probabilístico da variável de interesse na população, a partir de dados de uma amostra probabilística desta mesma população. Fonte: Barbetta, Reis e Bornia (2008).

Estatística é a ciência que se ocupa de organizar, descrever, analisar e interpretar dados para que seja possível a tomada de decisões e/ou a validação científica de uma conclusão. Os dados são coletados para estudar uma ou mais características de uma População: conjunto das medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m).

Uma população pode ser representada por meio de um modelo: este apresenta condições para uso, forma para a distribuição, e parâmetros.

Os dados necessários para a obtenção do modelo podem ser obtidos mediante um censo (pesquisa de toda a população), ou **através de uma amostra** (subconjunto finito) da população.

A amostra deve ser: representativa da população, suficiente (para que o resultado tenha confiabilidade) e aleatória (retirada por sorteio não viciado).

A Inferência Estatística consiste em fazer *afirmações probabilísticas* sobre as características do modelo probabilístico, que se supõe representar uma população, a partir dos dados de uma amostra aleatória (probabilística) desta mesma população.

Fazer uma afirmação probabilística sobre uma característica qualquer é associar à declaração feita probabilidade de que tal declaração esteja correta (e, portanto, a probabilidade complementar de que esteja errada). Quando se usa uma amostra da população, sempre haverá a probabilidade de se estar cometendo um erro (justamente por ser usada uma amostra): a diferença entre os métodos estatísticos e os outros reside no fato de os métodos estatísticos permitirem calcular essa probabilidade de erro. Para que isso seja possível a amostra da população precisa ser aleatória.

As afirmações probabilísticas sobre o modelo da população podem ser basicamente:

- estimar quais são os possíveis valores dos parâmetros – Estimação de Parâmetros;
- qual o valor da média de uma variável que segue uma distribuição normal?

- qual o valor da proporção de um dos dois resultados possíveis de uma variável que segue uma distribuição binomial.
- testar hipóteses sobre as características do modelo: parâmetros, forma da distribuição de probabilidades, entre outros – Testes de Hipóteses.
- o valor da média de uma variável que segue uma distribuição é maior do que um determinado valor?
- o modelo probabilístico da população é uma distribuição normal?
- o valor da média de uma variável que segue uma distribuição normal em uma população é diferente da mesma média em outra população?

Estudaremos Estimação de Parâmetros na Unidade 9 e Testes de Hipóteses na Unidade 10.

Parâmetros e Estatísticas

Vamos recorrer a mesma pesquisa sugerida na Unidade 1: opinião dos registrados no CRA-SC sobre os cursos em que se graduaram, desde que graduados em Santa Catarina. Naquela Unidade, e depois na Unidade 2, declaramos que era possível realizar uma amostragem probabilística, e vimos um exemplo de como fazer isso.

Independente da pesquisa ter sido realizada por amostragem probabilística, os dados podem ser estatisticamente generalizados para a população.

Uma vez coletado os dados, é preciso resumi-los e organizá-los de maneira a permitir uma primeira análise e posterior uso das informações. As técnicas estatísticas que se ocupam desses aspectos constituem a Análise Exploratória de Dados, que estudamos detalhadamente nas Unidades 3 e 4.

O conjunto de dados pode ser resumido (e apresentado) através das distribuições de frequências, que relacionam os valores que a

Testes de hipóteses – forma de inferência estatística que busca testar hipóteses sobre características (parâmetros, forma do modelo) do modelo probabilístico da variável de interesse na população, a partir de dados de uma amostra probabilística desta mesma população. Fonte: Barbetta, Reis e Bornia (2008).

A proporção está relacionada aos percentuais de ocorrência dos valores em uma distribuição de frequências de uma variável qualitativa.

Estatísticas – medidas de síntese da variável calculadas com base nos resultados de uma amostra da população. Se a amostra for probabilística (aleatória) as estatísticas podem ser consideradas variáveis aleatórias. Fonte: Barbetta, Reis e Bornia (2008).

variável pode assumir com a frequência (contagem) com que foram encontrados naquele conjunto. Essa distribuição pode ser apresentada na forma de uma tabela, ou por meio de um gráfico (esses dois métodos podem ser usados tanto para variáveis qualitativas quanto para variáveis quantitativas).

Há uma terceira forma de resumir o conjunto de dados, quando a variável sob análise é quantitativa: as medidas de síntese ou estatísticas. As principais estatísticas são a média, o desvio padrão, a variância e a proporção.

Atenção! Vamos relembrar o que cada uma significa:

Média: média aritmética simples (ver Unidade 4) trata-se de uma estatística que caracteriza o “centro de massa” do conjunto de dados (Valor Esperado – ver Unidade 6). Quando é a média populacional recebe o símbolo μ , quando é a média amostral recebe o símbolo $\bar{x}$.

Variância: trata-se de uma estatística (ver Unidade 4): mede a dispersão em torno da média do conjunto, (em torno do valor esperado – possuindo uma unidade que é o quadrado da unidade da média (e dos valores do conjunto)). Quando é a variância populacional recebe o símbolo σ^2 quando é a variância amostral recebe o símbolo s^2 .

Desvio padrão: é a raiz quadrada positiva da variância, que possui uma unidade que é igual à unidade da média; preferida, muitas vezes, para efeito de mensuração da dispersão. Quando é o valor populacional, recebe o símbolo σ ; quando é o amostral recebe o símbolo s .

Proporção: consiste em calcular a razão entre o número de ocorrências do valor de interesse de uma variável qualitativa e o número total de ocorrências registradas no conjunto (de todos os valores que a variável pode assumir). Quando é uma proporção populacional, recebe o símbolo π ; quando é uma proporção amostral, recebe o símbolo p .

Os valores das medidas de síntese, além de resumirem o conjunto de dados, constituem uma indicação dos prováveis valores dos parâmetros. Assim, em estudos baseados em amostras, é comum utilizar tais medidas de síntese como estatísticas que serão utilizadas para estimar os parâmetros do modelo probabilístico que descreve a população.

A Tabela 4 resume os parâmetros e as estatísticas:

Tabela 4: Parâmetros e Estatísticas mais comuns

MEDIDAS DE SÍNTESE	PARÂMETROS (POPULAÇÃO)	ESTATÍSTICAS (AMOSTRA)
Média	$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^n X_i}{n}$
Variância	$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$
Proporção	$\pi = \frac{f_a}{N}$	$p = \frac{f_a}{n}$

Fonte: Elaborada pelo autor deste livro

Onde: N = o número de elementos da população; n = o número de elementos da amostra; f_a = a frequência de ocorrência de um dos valores de uma variável qualitativa na população ou na amostra.

As Estatísticas são variáveis aleatórias, pois seus valores podem variar dependendo do resultado da amostra. Se forem variáveis aleatórias, podem ser caracterizadas através de algum modelo probabilístico. Esse modelo recebe o nome de distribuição amostral.

Distribuição Amostral

Seja uma população qualquer com um parâmetro θ de interesse, que corresponda a uma estatística $\mathbf{T}$ em uma amostra. Amostras aleatórias são retiradas da população e para cada amostra calcula-se o valor $\mathbf{t}$ da estatística $\mathbf{T}$.

NÃO confundir com o t da distribuição o t de Student (Unidade 7).

Os valores de **t** formam uma nova população que segue uma distribuição de probabilidades que é chamada de **distribuição amostral de T**.

Vamos ver um exemplo.

Exemplo 1 – Seja a população a seguir constituída pelos pesos em kg de oito pessoas adultas:

População			Amostras
64	67	Sortear 3	
70	57		{64, 67, 60} ⇒ Média = 63,67 kg
65	72		{65, 70, 60} ⇒ Média = 65 kg
70	60		{65, 70, 60} ⇒ Média = 65 kg
$\mu = 65,62 \text{ kg}$			

Figura 64: Distribuição Amostral – Exemplo 1

Fonte: Elaborada pelo autor deste livro

Observe que foram retiradas três amostras. Para cada amostra foi calculada a média, com objetivo de estimar a média populacional, que vale 65,62 kg. Repare que há uma variação na estatística média, pois o processo de amostragem é aleatório: é um experimento aleatório. Essa variação precisa ser considerada quando são realizadas as inferências sobre os parâmetros.

Dessa maneira, o conhecimento das distribuições amostrais das principais estatísticas é necessário para fazer inferências sobre os parâmetros do modelo probabilístico da população. Por hora, basta conhecer as distribuições amostrais das estatísticas média de uma variável quantitativa qualquer e a proporção de um dos dois únicos resultados de uma variável qualitativa.

Distribuição Amostral da Média

Atentemos para as particularidades da distribuição amostral da média.

Exemplo 2 – suponha uma variável quantitativa cujos valores constituem uma população com os seguintes valores: **(2, 3, 4, 5)**.

Para essa população, que tem uma distribuição uniforme, podemos perceber que os parâmetros são: $\mu = 3,5$ e $\sigma^2 = 1,25$ (usou-se n no denominador por ser uma população)

Se retirarmos todas as amostras aleatórias de dois elementos (com reposição) possíveis dessa população ($n = 2$), teremos os seguintes resultados:

Há 16 amostras possíveis.

(2, 2)	(2, 3)	(2, 4)	(2, 5)
(3, 2)	(3, 3)	(3, 4)	(3, 5)
(4, 2)	(4, 3)	(4, 4)	(4, 5)
(5, 2)	(5, 3)	(5, 4)	(5, 5)

O cálculo das médias de todas as amostras acima resultará na matriz a seguir:

$$\bar{X} \begin{Bmatrix} (2,0) & (2,5) & (3,0) & (3,5) \\ (2,5) & (3,0) & (3,5) & (4,0) \\ (3,0) & (3,5) & (4,0) & (4,5) \\ (3,5) & (4,0) & (4,5) & (5,0) \end{Bmatrix}$$

Se essas médias forem plotadas em um histograma:

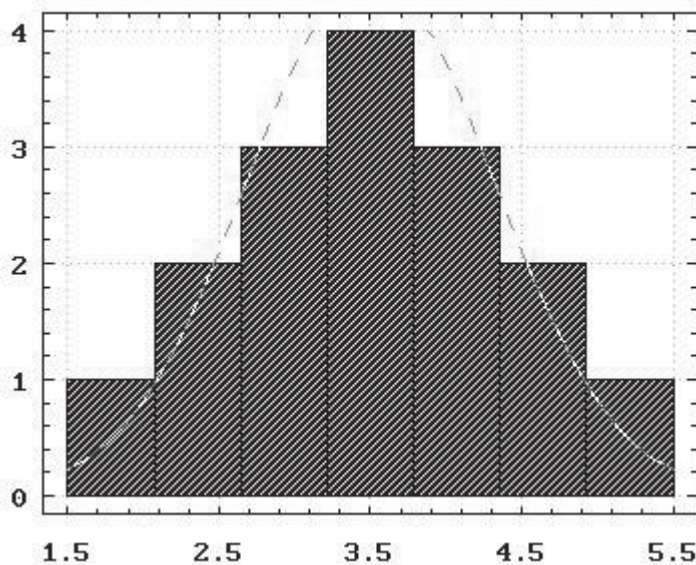


Figura 65: Histograma de médias amostrais

Fonte: Elaborada pelo autor deste livro

Se forem calculadas a média e a variância das médias de todas as amostras, o resultado será:

$$\bar{X} = 56/16 = 3,5 = \mu \qquad V(\bar{x}) = 0,625 = \frac{1,25}{2} = \frac{\sigma^2}{n}$$

Note como a distribuição das médias amostrais da variável pode ser aproximada por um modelo normal (não obstante a distribuição da variável na população não ser normal), e que o valor esperado das médias amostrais (média das médias) é igual ao valor da média populacional da variável e a variância das médias amostrais é igual ao valor da variância populacional da variável dividida pelo tamanho da amostra. Quanto maior o tamanho da amostra (quanto maior **n**) mais o histograma anterior vai se aproximar de um modelo normal, independentemente do formato da distribuição da variável na população.



Podemos então enunciar os teoremas.

Com relação à “suficientemente grande”, importante ressaltar que isso varia de distribuição para distribuição e, como foi visto, uma distribuição uniforme precisa de uma amostra pequena ($n = 2$ no caso) para que a aproximação seja possível; outras distribuições precisam de amostras maiores. Alguns autores costumam chamar de “grandes amostras” àquelas que possuem mais de 30 elementos, a partir desse tamanho a aproximação poderia ser feita sem muitas preocupações.

Teorema das Combinações Lineares

Se a variável de interesse segue uma distribuição normal na população a distribuição amostral das médias de amostras aleatórias retiradas da população também será normal, independentemente do tamanho das amostras.

Teorema Central do Limite

Se a variável de interesse não segue uma distribuição normal na população (ou não se sabe qual é a sua distribuição), a distribuição amostral das médias de amostras aleatórias retiradas da população será normal se o tamanho das amostras for **suficientemente grande**, com uma média igual à média populacional e uma variância igual à variância populacional dividida pelo tamanho da amostra.

Para o caso da Proporção podemos chegar a uma conclusão semelhante.

Distribuição Amostral da Proporção

Vamos estudar as particularidades da distribuição amostral da proporção através de um exemplo.

Exemplo 3: pense agora em uma variável qualitativa que pode assumir apenas dois valores, e que constitui a seguinte população: (□, □, □, □, ■)

Suponhamos que há interesse no valor ■ (valor que seria o nosso “sucesso”). A proporção desse valor na população (o valor do parâmetro) será $\pi = 1/5$

Se retirarmos todas as amostras aleatórias de dois elementos (com reposição) possíveis dessa população ($n = 2$), teremos os seguintes resultados:

Há 25 amostras possíveis.

(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, ■)	(□, ■)	(□, ■)	(□, ■)	(■, ■)

Figura 66: Amostras de tamanho 2 para proporção

Fonte: Elaborada pelo autor deste livro

Repare que se definirmos a variável como o número de “sucessos” (número de ■), esta seguirá um modelo binomial: há apenas dois resultados possíveis para cada realização, há um número limitado de realizações ($n = 2$ no caso), e cada realização independe da outra (porque a amostra é aleatória com reposição).

Calculando a proporção de | em cada uma das amostras, e chamando essa proporção amostral de p , teremos os seguintes resultados:

$$p = \begin{matrix} & (0) & (0) & (0) & (0) & (1/2) \\ & (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (0) & (1/2) \\ & (0) & (0) & (0) & (0) & (1/2) \\ (1/2) & (1/2) & (1/2) & (1/2) & (1/2) & (1) \end{matrix}$$

Calculando a média (valor esperado) e a variância das proporções acima teremos:

$$\bar{X} = E(p) = \frac{1}{5} = \pi \quad s^2 = 0,08 = \frac{\left(\frac{1}{5}\right) \times \left(1 - \frac{1}{5}\right)}{2} = \frac{\pi \times (1 - \pi)}{n}$$

Perceba que o valor esperado (média) das proporções amostrais é igual ao valor da proporção populacional de ■, e que a **variância das proporções amostrais é igual ao produto da proporção populacional de ■ por seu complementar, dividido pelo tamanho da amostra**.

Lembre-se de que um modelo binomial pode ser aproximado por um modelo normal se algumas condições forem satisfeitas: se o produto do número de realizações pela probabilidade de “sucesso” ($n \times p$) e o produto do número de realizações pela probabilidade de “fracasso” ($n \times [1 - p]$) forem ambos maiores ou iguais a 5. Essa distribuição normal teria média igual a $n \times p$ e variância igual a $n \times p \times (1 - p)$. Se estivermos interessados apenas na proporção (probabilidade de “sucesso”) e não no número de “sucessos”, as expressões anteriores podem ser divididas por n (o tamanho da amostra): média = p e variância = $[p \times (1 - p) / n]$.

Por causa do Teorema Central do Limite é que o modelo normal é tão importante. É claro que ele representa muito bem uma grande variedade de fenômenos, mas é devido à sua utilização em Inferência Estatística que seu estudo é imprescindível. Ressalte-se, porém, que sua aplicação costuma resumir-se ao que se chama de Inferência Paramétrica, inferências sobre os parâmetros dos modelos probabilísticos que descrevem as variáveis na população. Para fazer inferências sobre outros aspectos que não os parâmetros, ou quando as amostras utilizadas não forem suficientemente grandes para se assumir a validade do Teorema Central do Limite, é preciso usar técnicas de Inferência Não Paramétrica, conteúdo este não contemplado nesta disciplina.

Voltaremos a analisar o significado desse resultado quando estudarmos Estimação por Ponto.

Isso também é decorrência do Teorema Central do Limite.

Saiba mais...

Sobre distribuição amostral veja: BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulo 7.

STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Ed. Harbra, 2001, Capítulo 7.

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. *Estatística Aplicada à Administração e Economia*. 2. ed. São Paulo: Thomson Learning, 2007, Capítulo 7.

Sobre a utilização do Microsoft Excel para estudar distribuições amostrais veja LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística: Teoria e Aplicações – Usando Microsoft Excel em Português*. 5. ed. Rio de Janeiro: LTC, 200, Capítulo 5.

Resumindo

A Figura 67, a seguir, representa o resumo desta Unidade.

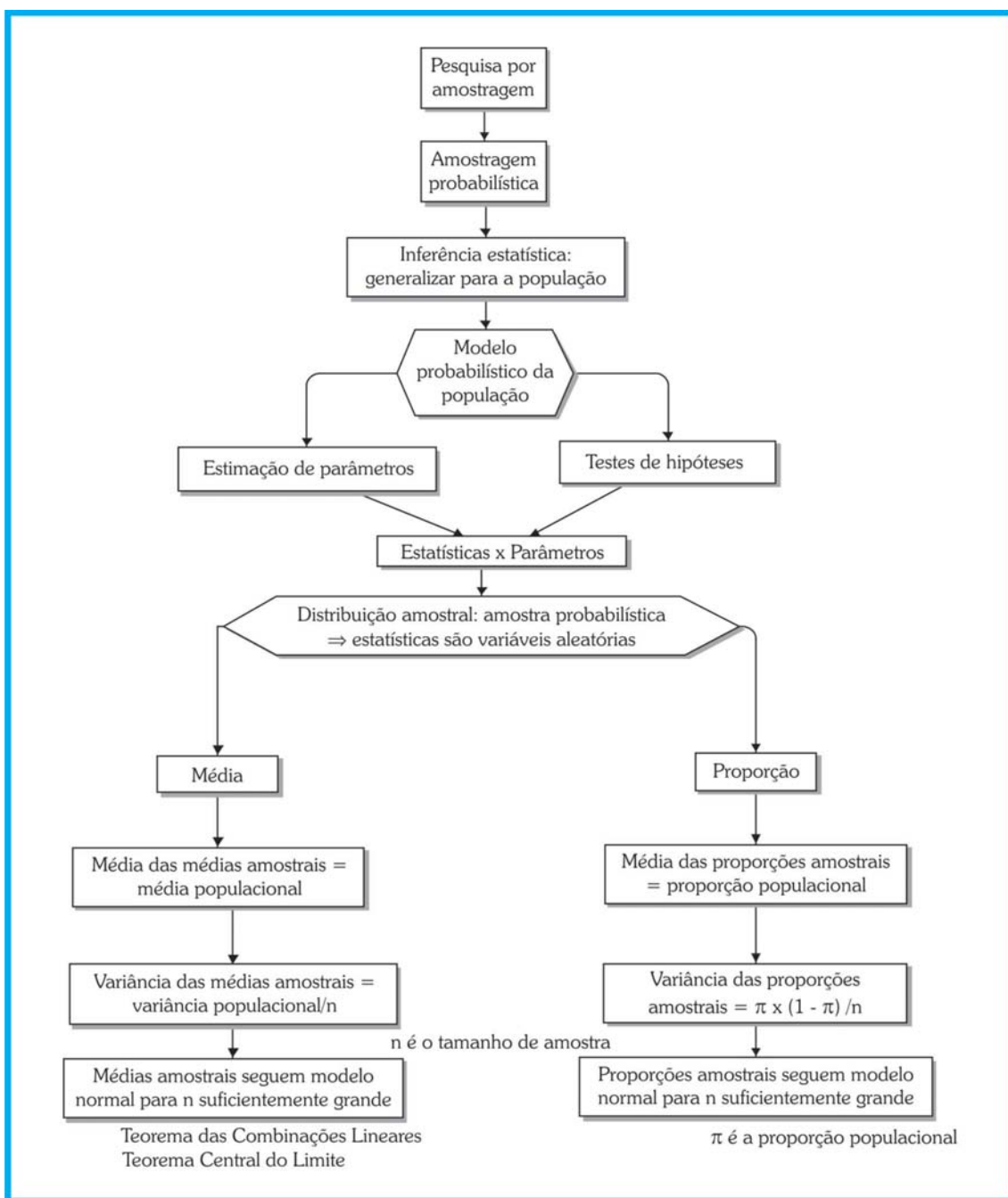


Figura 67: Resumo da Unidade 8
Fonte: Elaborada pelo autor deste livro

Caro estudante,

Esta Unidade foi muito importante para o seu aprendizado, pois lhe deu bases importantes que permitirão avançar na disciplina. Vimos, até agora, sobre a inferência estatística e distribuição amostral, seu modelo probabilístico e testes de hipóteses. Ao final da Unidade, sugerimos que interaja com seus colegas, responda às atividades de aprendizagem e visite o Ambiente Virtual de Ensino-Aprendizagem, espaço no qual poderá resolver suas dúvidas com o auxílio de seu tutor. Bons estudos!



Atividades de aprendizagem

1. Uma fundição produz blocos para motor de caminhões. Os furos para as camisas devem ter diâmetro de 100 mm, com tolerância de 5 mm. Para verificar qual o diâmetro médio no processo, a empresa vai retirar uma amostra com 36 blocos e medir os diâmetros de 36 furos (1 a cada bloco). Suponha que o desvio padrão (populacional) dos diâmetros seja conhecido e igual a 3 mm.

- a) Qual o desvio padrão da distribuição da média amostral?
- b) Qual a probabilidade da média amostral diferir da média populacional (desconhecida) em mais de 0,5 mm (para mais ou para menos)?
- c) Qual a probabilidade da média amostral diferir da média populacional (desconhecida) em mais de 1 mm (para mais ou para menos)?
- d) Se alguém afirmar que a média amostral não se distanciará da média populacional em mais de 0,98 mm, qual a probabilidade dessa pessoa acertar?

Camisa – mecânica: invólucro de metal que protege peças ou partes de certas máquinas e motores. Fonte: Aulete (2008).

- e) Se alguém afirmar que a média amostral não se distanciará da média populacional em mais de 1,085 mm, qual a probabilidade dessa pessoa errar?

9

UNIDADE

Estimação de Parâmetros



Ao finalizar esta Unidade, você deverá ser capaz de compreender e de aplicar os conceitos de Estimação de Parâmetros por Ponto e por Intervalo de Média e Proporção; e de calcular o tamanho mínimo de amostra necessário para a Estimação por Intervalo.

Estimação por Ponto de Parâmetros

Caro estudante,

Na Unidade 8, você viu o conceito de Distribuição Amostral e constatou a importância do modelo normal. Nesta Unidade, você vai aprender como aplicar esses conceitos no primeiro tipo particular de inferência estatística: a **Estimação de Parâmetros**, por ponto e por intervalo.

Parâmetros são medidas de síntese de variáveis quantitativas na população que estamos pesquisando. Por ser inviável ou inconveniente pesquisar toda a população, coletamos uma amostra para estudá-la. Os resultados da amostra podem ser usados, então, para fazer afirmações probabilísticas sobre o parâmetro de interesse: definir um intervalo possível para os valores do parâmetro e calcular a probabilidade de o valor real do parâmetro estar dentro dele (esta é a Estimação por Intervalo).

Vamos aprender como estimar os parâmetros média de uma variável quantitativa e proporção de um dos valores de uma variável qualitativa. Além disso, você vai ver como é possível definir, de forma mais acurada, o tamanho mínimo de uma amostra aleatória para estimar média e proporção (tema iniciado na Unidade 2).

Uma vez tendo decidido que modelo probabilístico é mais adequado para representar a variável de interesse na população, resta obter os seus parâmetros. Nos estudos feitos com base em amostras, é preciso escolher qual das estatísticas de amostra será o melhor estimador para cada parâmetro do modelo.

A Estimação por Ponto consiste em determinar qual será o melhor estimador para o parâmetro de interesse.

Estimação por ponto – tipo de estimação de parâmetros que procura identificar qual o melhor estimador para um parâmetro populacional a partir das várias estatísticas amostrais disponíveis, seguindo alguns critérios. Fonte: Barbetta, Reis e Bornia (2008).

Como os parâmetros serão estimados através das estatísticas de uma amostra aleatória; e como para cada amostra aleatória as estatísticas apresentarão diferentes valores, os estimadores também terão valores aleatórios. Em outras palavras, um estimador é uma variável aleatória que pode ter um modelo probabilístico para descrevê-la.

Naturalmente haverá várias estatísticas $\mathbf{T}$ que poderão ser usadas como estimadores de um parâmetro θ qualquer. Como escolher qual das estatísticas será o melhor estimador para o parâmetro?

Há basicamente três critérios para a escolha de um estimador: o estimador precisa ser justo, consistente e eficiente.

- 1) Um estimador $\mathbf{T}$ é um estimador **justo** (não tendencioso) de um parâmetro θ quando o valor esperado de $\mathbf{T}$ é igual ao valor do parâmetro θ a ser estimado: $E(\mathbf{T}) = \theta$
- 2) Um estimador $\mathbf{T}$ é um estimador **consistente** de um parâmetro θ quando além ser um estimador justo a sua variância tende a zero à medida que o tamanho da amostra aleatória aumenta: $\lim_{n \rightarrow \infty} V(\mathbf{T}) = 0$.
- 3) Se há dois estimadores justos de um parâmetro o mais **eficiente** é aquele que apresentar a menor variância.



Conforme foi dito na introdução desta Unidade, estamos interessados em estimar dois parâmetros: média e proporção populacional. Vamos buscar, então, os estimadores mais apropriados para ambos.

Estimação por Ponto dos Principais Parâmetros

Os principais parâmetros que vamos avaliar aqui são: média de uma variável que segue um modelo normal (ou qualquer modelo se a amostra for suficientemente grande) em uma população (média populacional – μ) e proporção de ocorrência de um dos valores de uma variável que segue um modelo binomial em uma população (proporção populacional – π). Em suma escolher quais estatísticas

amostrais são mais adequadas para estimar esses parâmetros, usando os critérios definidos acima.

Considerando os exemplos 2 e 3 da Unidade 8, algumas constatações que lá foram feitas passarão a fazer sentido agora.

Vamos supor que houvesse a intenção de estimar a média populacional da variável do Exemplo 2. Qual das estatísticas disponíveis seria o melhor estimador?

Lembrem-se de que, após retirar todas as amostras aleatórias possíveis daquela população, calculamos a média de cada amostra e, posteriormente, a média dessas médias. Constatamos que o valor esperado das médias amostrais (média das médias) é igual ao valor da média populacional da variável e a variância das médias amostrais é igual ao valor da variância populacional da variável dividida pelo tamanho da amostra:

$$E(\bar{x}) = \mu \quad V(\bar{x}) = \frac{\sigma^2}{n}$$

O melhor estimador da média populacional μ é a média amostral $\bar{x}$, pois se trata de um estimador justo e consistente:

- justo porque o valor esperado da média amostral será a média populacional;
- consistente porque, se o tamanho da amostra n tender ao infinito a variância da média amostral (do Estimador) tenderá a zero.

Agora, vamos supor que houvesse a intenção de estimar a proporção populacional do valor ■ da variável do Exemplo 3. Qual das estatísticas disponíveis seria o melhor estimador?

Lembrem-se de que, após retirar todas as amostras aleatórias possíveis daquela população, calculamos a proporção de ■ em cada amostra e, posteriormente, a média dessas proporções. Constatamos que o valor esperado das proporções amostrais (média das proporções) é igual ao valor da proporção populacional do valor ■ da variável e a variância das proporções amostrais é igual ao valor do produto da proporção populacional do valor ■ da variável pela sua complementar dividida pelo tamanho da amostra:

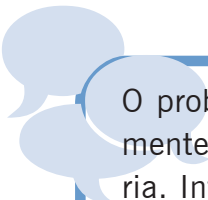
$$E(p) = \pi \quad V(p) = \frac{\pi \times (1 - \pi)}{n}$$

O melhor estimador da proporção populacional π é a proporção amostral p , pois se trata de um estimador **justo** e **consistente**:

- Justo porque o valor esperado da proporção amostral será a proporção populacional; e
- Consistente porque se o tamanho da amostra n tender ao infinito a variância da proporção amostral (do Estimador) tenderá a zero.

Poderíamos fazer um procedimento semelhante para estimar outros parâmetros como, por exemplo, a variância populacional de uma variável. Esse procedimento não será demonstrado, mas o melhor estimador da variância populacional será a variância amostral se for usado $n - 1$ no denominador da expressão de cálculo. Somente assim a variância amostral será um estimador justo (não viciado) da variância populacional.

Como o desvio padrão é a raiz quadrada da variância, é comum estimar o desvio padrão populacional extraíndo a raiz quadrada da variância amostral.



O problema da Estimação por Ponto é que, geralmente, dispomos apenas de uma amostra aleatória. Intuitivamente, qual será a probabilidade de a média ou proporção amostral, de uma amostra aleatória, coincidir com o valor do parâmetro? É como pescar usando-se uma lança de bambu, que requer muita habilidade... Porém, se você puder usar uma rede, ficará bem mais fácil. Essa “rede” é a Estimação por Intervalo.

Estimação por Intervalo de Parâmetros

Geralmente, uma inferência estatística é feita com base em uma única amostra: na maior parte dos casos é totalmente inviável retirar todas as amostras possíveis de determinada população.

Intuitivamente, percebemos que as estatísticas calculadas nessa única amostra, mesmo sendo os melhores estimadores para os parâmetros de interesse, terão probabilidade infinitesimal de coincidir exatamente com os valores reais dos parâmetros. Então a Estimação por Ponto dos parâmetros é insuficiente, e as estimativas assim obtidas servirão apenas como referência para a Estimação por Intervalo.

A Estimação por Intervalo consiste em colocar um Intervalo de Confiança (I.C.) em torno da estimativa obtida através da Estimação por Ponto.

O Intervalo de Confiança terá certa probabilidade, chamada de Nível de confiança (que costuma ser simbolizado como $1 - \alpha$), de conter o valor real do parâmetro e a probabilidade de esta faixa realmente abranger o valor real do parâmetro. A probabilidade de o Intervalo de Confiança não conter o valor real do parâmetro é chamada de Nível de Significância (α), e o valor dessa probabilidade será o complementar do Nível de Confiança. É comum definir o Nível de Significância como probabilidade máxima de erro, um risco máximo admissível.

A determinação do Intervalo de Confiança para determinado parâmetro resume-se basicamente a definir o Limite Inferior e o Limite Superior do intervalo, supondo determinado Nível de Confiança ou Significância.

A definição dos limites dependerá também da distribuição amostral da estatística usada como referência para o intervalo e do tamanho da amostra utilizada.

Para os dois parâmetros em que temos maior interesse (média populacional μ e proporção populacional π) a distribuição amostral dos estimadores (média amostral $\bar{x}$ e proporção amostral p , respectivamente) pode ser aproximada por uma distribuição normal: o Intervalo de Confiança será então simétrico em relação ao valor calculado da estimativa (média ou proporção amostral), com base na amostra aleatória coletada (Figura 68):

Fazer uma Estimação por Intervalo de um parâmetro é efetuar uma afirmação probabilística sobre este parâmetro, indicando uma faixa de possíveis valores.

Intervalo de Confiança

– faixa de valores da estatística usada como estimador, dentro da qual existe a probabilidade conhecida de estar o verdadeiro valor do parâmetro. Sinônimo de estimação por intervalo. Fonte: Barbetta, Reis e Bornia (2008).

Nível de Significância

– complementar do nível de confiança, a probabilidade de que o intervalo de confiança não contenha o valor real do parâmetro. Probabilidade de erro espera-se que seja um valor baixo, de no máximo 10%. Fonte: Barbetta, Reis e Bornia (2008).

Nível de Confiança

– probabilidade de que o intervalo de confiança contenha o valor real do parâmetro a estimar, espera-se que seja um valor alto, de no mínimo 90%. Fonte: Moore, McCabe, Duckworth e Sclove (2006).



Figura 68: Intervalo de confiança para um modelo normal

Fonte: Elaborada pelo autor deste livro

Onde: L_i é o limite inferior e L_s é o limite superior do Intervalo de Confiança; $1 - \alpha$ é o Nível de Confiança estabelecido, observando que o valor do Nível de Significância α é dividido igualmente entre os valores abaixo de L_i e acima de L_s .

Para obter os limites em função do Nível de Confiança, devemos utilizar a distribuição normal padrão (variável Z com média zero e variância um): fixar certo valor de probabilidade, obter o valor de Z correspondente, e substituir o valor em $Z = (x - \text{“média”}) / \text{“desvio padrão”}$ para obter o valor x (valor correspondente ao valor de Z para a probabilidade fixada). Observe a Figura 69:

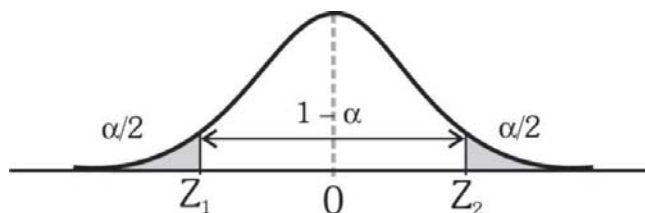


Figura 69: Intervalo de confiança para a distribuição normal padrão

Fonte: Elaborada pelo autor deste livro

O limite L_i (inferior) corresponde a Z_1 e o limite L_s (superior) corresponde a Z_2 . O ponto central 0 (zero) corresponde ao valor calculado da Estimativa. Como a variável Z tem distribuição normal com média igual a zero (lembrando que a distribuição normal é simétrica em relação à média) os valores de Z_1 e Z_2 serão iguais em módulo (Z_1 será negativo e Z_2 positivo):

$$Z_1 \text{ será um valor de } Z \text{ tal que } P(Z \leq Z_1) = \frac{\alpha}{2}, \text{ e } Z_2 \text{ será um valor tal que } P(Z \leq Z_2) = 1 - \frac{\alpha}{2}$$

Então, obteremos os valores dos limites através das expressões:

Foram colocados entre aspas porque os valores dependerão dos parâmetros sob análise e de outros fatores.

$$Z_1 = (L_i - \text{"média"}) / \text{"desvio padrão"} \Rightarrow L_i = \text{"média"} + Z_1 \times \text{"desvio padrão"}.$$

$$Z_2 = (L_s - \text{"média"}) / \text{"desvio padrão"} \Rightarrow L_s = \text{"média"} + Z_2 \times \text{"desvio padrão"}.$$

Como $Z_1 = -Z_2$, podemos substituir:

$$L_i = \text{"média"} - Z_2 \times \text{"desvio padrão"}.$$

$$L_s = \text{"média"} + Z_2 \times \text{"desvio padrão"}.$$

Este valor Z_2 costuma ser chamado de $Z_{\text{crítico}}$, porque corresponde aos limites do intervalo:

$$L_i = \text{"média"} - Z_{\text{crítico}} \times \text{"desvio padrão"}.$$

$$L_s = \text{"média"} + Z_{\text{crítico}} \times \text{"desvio padrão"}.$$

Reparem que o mesmo valor é somado e subtraído da “média”. Esse valor é chamado de semi-intervalo ou precisão do intervalo (ou margem de erro) e_0 :

$$e_0 = Z_{\text{crítico}} \times \text{"desvio padrão"}.$$

Resta agora definir corretamente o valor da “média” e do “desvio padrão” para cada um dos parâmetros em que estamos interessados (média e proporção populacional). Com base nas conclusões obtidas na Estimação por Ponto isso será simples. Contudo, há alguns outros aspectos que precisarão ser esmiuçados.

Estimação por Intervalo da Média Populacional

Lembrando as expressões anteriores:

$$L_i = \text{"média"} - Z_{\text{crítico}} \times \text{"desvio padrão"} = \text{"média"} - e_0.$$

$$L_s = \text{"média"} + Z_{\text{crítico}} \times \text{"desvio padrão"} = \text{"média"} + e_0.$$

Nesse caso, a “média” será a média amostral $\bar{x}$ (ou mais precisamente o seu valor):

$$P(\bar{x} - e_0 \leq \mu \leq \bar{x} + e_0) = 1 - \alpha$$

O valor de e_0 dependerá de outros aspectos.

a) Se a variância populacional σ^2 da variável (cuja média populacional queremos estimar) for conhecida.

Nesse caso a variância amostral da média poderá ser calculada através da expressão:

$$V(\bar{x}) = \frac{\sigma^2}{n}, \text{ e, por conseguinte, o "desvio padrão" será } = \frac{\sigma}{\sqrt{n}}$$

$$\text{E } e_0 \text{ será: } e_0 = Z_{\text{crítico}} \times \frac{\sigma}{\sqrt{n}}$$

Bastará então fixar o Nível de Confiança (ou de Significância) para se obter $Z_{\text{crítico}}$ por meio da Tabela disponível no Ambiente Virtual e calcular e_0 .

b) Se a variância populacional σ^2 da variável for desconhecida.

Naturalmente esse é o caso mais encontrado na prática. Como deveremos proceder? Dependerá do tamanho da amostra.

b.1 – Grandes amostras (mais de 30 elementos).

Nesses casos, procede-se como no item anterior, apenas fazendo com que $\sigma = s$, ou seja, considerando que o desvio padrão da variável na população é igual ao desvio padrão da variável na amostra (suposição razoável para grandes amostras).

b.2 – Pequenas amostras (até 30 elementos).

Nesses casos a aproximação do item b.1 não será viável. Terá que ser feita uma correção na distribuição normal padrão (Z) através da distribuição **t de Student**, que estudamos na Unidade 6.

Quando a variância populacional da variável é desconhecida e a amostra tem até 30 elementos substitui-se σ por s e Z por t_{n-1} em todas as expressões para determinação dos limites do intervalo de confiança, obtendo-se:

$$L_i = \text{"média"} - t_{n-1, \text{crítico}} \times \text{"desvio padrão"} = \text{"média"} - e_0$$

$$L_s = \text{"média"} + t_{n-1, \text{crítico}} \times \text{"desvio padrão"} = \text{"média"} + e_0$$

$$\text{E } e_0 \text{ será: } e_0 = t_{n-1, \text{crítico}} \times \frac{s}{\sqrt{n}}$$

Os valores de $t_{n-1, \text{crítico}}$ podem ser obtidos de forma semelhante aos de $Z_{\text{crítico}}$, definindo o Nível de Confiança (ou de Significância), mas precisam também da definição do número de graus de liberdade ($n - 1$): de posse desses valores basta procurar o valor da Tabela 2 do Ambiente Virtual ou em um programa computacional.

Se o tamanho da amostra (n) for superior a 5% do tamanho da população (N), os valores de e_0 precisam ser corrigidos. Caso contrário os limites dos intervalos não serão conhecidos. A correção é mostrada na equação a seguir:

$$e_{0_{\text{corrigido}}} = e_0 \times \sqrt{\frac{N-n}{N-1}}$$

Vejamos por meio de um exemplo.

Exemplo 1: retirou-se uma amostra aleatória de quatro elementos de uma produção de cortes bovinos no intuito de estimar a média do peso do corte. Obteve-se média de 8,2 kg e desvio padrão de 0,4 kg. Suponha população normal e determine um intervalo de confiança para a média populacional com 1% de significância.

O parâmetro de interesse é a média populacional μ do peso do corte.

Adotou-se um nível de significância de 1%, então $\alpha = 0,01$ e $1 - \alpha = 0,99$.

As estatísticas disponíveis são: **média amostral** = 8,2 kg **s** = 0,4 kg **n** = 4 elementos.

Definição da variável de teste: como a variância populacional é **desconhecida** e o tamanho da amostra é menor que 30 elementos, não obstante a população ter distribuição normal, a distribuição amostral da média será **t** de *Student*, e a variável de teste será: **t_{n-1}**

Encontrar o valor de **t_{n-1, crítico}**: como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da Figura 70 a seguir.

Este valor pode ser arbitrado pelo usuário ou pode ser uma exigência do problema sob análise, ou até mesmo uma exigência legal. Os níveis de significância mais comuns são de 1%, 5% ou mesmo 10%.

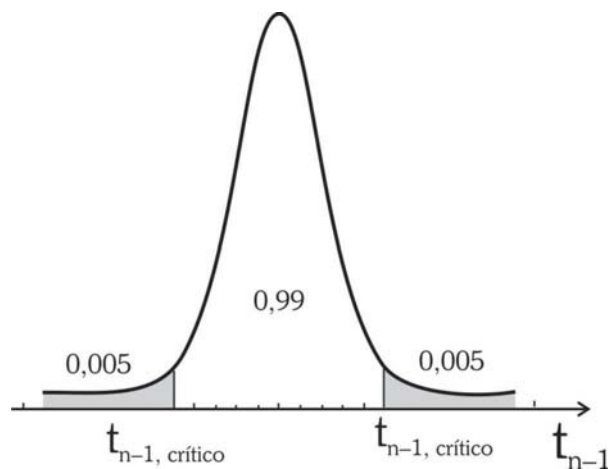


Figura 70: Distribuição t de *Student* para 99% de confiança

Fonte: Elaborada pelo autor deste livro

Para encontrar o valor crítico, devemos procurar na tabela da distribuição de *Student*, na linha correspondente a $n-1$ graus de liberdade, ou seja, em $4 - 1 = 3$ graus de liberdade. O valor da probabilidade pode ser visto na Figura anterior: os valores críticos serão $t_{3;0,005}$ e $t_{3;0,995}$ os quais serão iguais em módulo. E o valor de $t_{n-1,\text{crítico}}$ será igual a **5,84** (em módulo).

Determinamos os limites do intervalo, através da expressão seguinte (cujo resultado será somado e subtraído da média amostral), para estabelecer os limites do intervalo:

$$e_0 = \frac{t_{n-1,\text{crítico}} * s}{\sqrt{n}} = \frac{5,84 * 0,4}{\sqrt{4}} = 1,168 \text{ kg}$$

$$L_1 = \bar{x} - e_0 = 8,2 - 1,168 = 7,032 \text{ kg}$$

$$L_5 = \bar{x} + e_0 = 8,2 + 1,168 = 9,368 \text{ kg}$$

Então o intervalo de 99% de confiança para a média populacional da dimensão é [7,032;9,368] kg. Interpretação: há 99% de probabilidade de a verdadeira média populacional do peso de corte estar entre 7,032 kg e 9,368 kg.

Estimação por Intervalo da Proporção Populacional

Anteriormente, declaramos que o melhor estimador para a proporção populacional π é a proporção amostral p ; e que essa proporção amostral teria média igual a π e variância igual a $[\pi \times (1 - \pi)]/n$. Onde n é o tamanho da amostra aleatória. A distribuição da proporção amostral p é binomial, e sabemos que a distribuição binomial pode ser aproximada por uma normal se algumas condições forem satisfeitas:

$$\text{Se } n \times \pi \geq 5 \text{ E } n \times (1 - \pi) \geq 5$$

Ora, se π fosse conhecido não estaríamos aqui nos preocupando com a Estimação dele por Intervalo. Assim, vamos verificar se é possível aproximar a distribuição binomial de p por uma normal se: $n \times p \geq 5$ e $n \times (1 - p) \geq 5$, ou seja, usando o próprio valor da proporção amostral observada (trata-se de uma aproximação razoável).

Se e somente se essas duas condições forem satisfeitas poderemos usar as expressões a seguir:

$$L_i = \text{“média”} - Z_{\text{crítico}} \times \text{“desvio padrão”} = \text{“média”} - e_0$$

$$L_s = \text{“média”} + Z_{\text{crítico}} \times \text{“desvio padrão”} = \text{“média”} + e_0$$

Nesse caso, a “média” será a proporção amostral (ou mais precisamente o seu valor):

$$P(p - e_0 \leq \mu \leq p + e_0) = 1 - \alpha$$

E o valor do “desvio padrão” será igual a: $\sqrt{\frac{\pi \times (1 - \pi)}{n}}$

Novamente, como π é desconhecido, usaremos a proporção amostral p como aproximação.

$$\text{Então } e_0 \text{ será: } e_0 = Z_{\text{crítico}} \times \sqrt{\frac{p \times (1 - p)}{n}}$$

Bastará, então, fixar o Nível de Confiança (ou de Significância), $Z_{\text{crítico}}$ e calcular e_0 .

Novamente, precisamos corrigir o valor de e_0 para o caso de população finita:

$$e_{0_{\text{corrigido}}} = e_0 \times \sqrt{\frac{N - n}{N - 1}}$$

Em suma, a Estimação por Intervalo da média e da proporção populacional consiste basicamente em calcular a amplitude do semi-intervalo (o e_0), de acordo com as condições do problema sob análise.

- Para a média, observar se é viável considerar a distribuição da variável na população como normal, ou considerar a amostra como suficientemente grande para que a distribuição das médias amostrais possa ser considerada normal;
- se isso for verificado, identificar se a variância populacional da variável é conhecida. Caso seja, deverá ser usada a variável Z da distribuição normal padrão para qualquer tamanho de amostra;
- se variância populacional da variável é desconhecida, há duas possibilidades: para amostras com mais de 30 elementos usar a variável Z , e fazer a variância populacional igual à variância amostral da variável; se a amostra possuir até 30 elementos, usar a variável t_{n-1} da distribuição de Student; e

- para a proporção, observar se é possível fazer a aproximação pela distribuição normal.

Vamos ver um exemplo.

No exemplo 2, retirou-se uma amostra aleatória de 1.000 peças de um lote. Verificou-se que 35 eram defeituosas.

Determinar um intervalo de confiança de 95% para a proporção peças defeituosas no lote.

O parâmetro de interesse é a proporção populacional π de peças defeituosas.

Adotou-se um nível de significância de 5%, então $\alpha = 0,05$ e $1 - \alpha = 0,95$

As estatísticas são: proporção amostral de peças defeituosas $p = 35/1.000$ $n = 1.000$ elementos.

Definição da variável de teste: precisamos verificar se é possível fazer a aproximação pela normal, então $n \times p = 1000 \times 0,035 = 35 > 5$ e $n \times (1 - p) = 1000 \times 0,965 = 965 > 5$. Como ambos os produtos satisfazem as condições para a aproximação podemos usar a variável Z da distribuição normal padrão.

Encontrar o valor de $Z_{\text{crítico}}$: como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da figura a seguir:

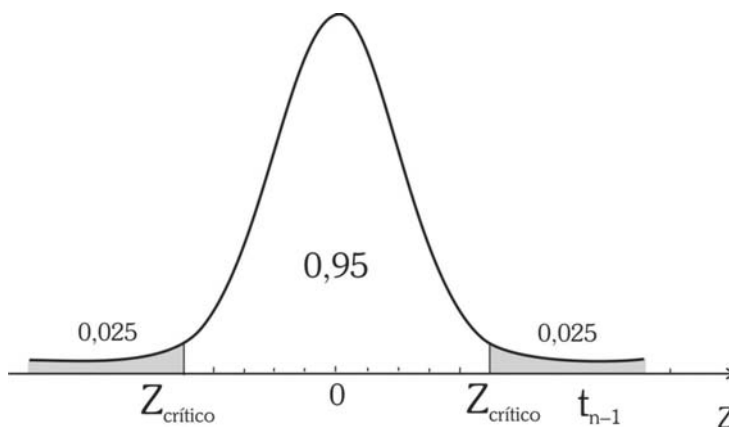


Figura 71: Distribuição normal padrão para 95% de confiança

Fonte: Elaborada pelo autor deste livro

Para encontrar o valor crítico, devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,975 ($0,95 + 0,025$). O valor da probabilidade pode ser visto na Figura anterior: os valores críticos serão $Z_{0,025}$ e $Z_{0,975}$ os quais serão iguais em módulo. E o valor de $Z_{\text{crítico}}$ será igual a 1,96 (em módulo).

Passamos agora à determinação dos limites do intervalo, através da expressão seguinte, cujo resultado será somado e subtraído da proporção amostral de peças defeituosas, para determinar os limites do intervalo:

$$e_0 = Z_{\text{crítico}} \times \sqrt{\frac{p \times (1-p)}{n}} = 1,96 \times \sqrt{\frac{0,035 \times 0,965}{1000}} = 0,0114$$

$$L_1 = p - e_0 = 0,035 - 0,0114 = 0,0236$$

$$L_s = p + e_0 = 0,035 + 0,0114 = 0,0464$$

Então, o intervalo de 95% de confiança para a proporção populacional de peças defeituosas é [2,36%;4,64%]. Interpretação: há 95% de probabilidade de a verdadeira proporção populacional de plantas atacadas pelo fungo estar entre 2,36% e 4,64%.

Tamanho Mínimo de Amostra para Estimação por Intervalo

Como observado nos itens anteriores, a determinação dos limites de um Intervalo de Confiança (determinação do e_0) depende do tamanho da amostra aleatória coletada, além do Nível de Confiança e da distribuição amostral do estimador utilizado. Nada podemos fazer quanto à distribuição amostral do estimador; o Nível de Confiança nós podemos controlar, seria interessante definir então uma **precisão** (um valor para e_0) para o Intervalo de Confiança: é muito comum querermos estabelecer previamente qual será a faixa de variação de determinado parâmetro, com certa confiabilidade.

Contudo, para um mesmo tamanho de amostra:

- Se aumentarmos o Nível de Confiança (reduzirmos o Nível de Significância) teremos um valor crítico maior, o que aumentará o valor de e_0 , resultando em um Intervalo de Confiança mais “largo”, com menor precisão.
- Se resolvermos aumentar a precisão (menor valor de e_0), isto é, obter-se um Intervalo de Confiança mais “estrito”, teremos uma queda no Nível de Confiança.

A solução para o dilema acima é a obtenção de um **tamanho mínimo de amostra** capaz de atender simultaneamente ao Nível de Confiança (ou de Significância) e à precisão (e_0) especificados. Como as expressões de e_0 são em função do tamanho de amostra (n), seria razoável pensar em reordená-las de forma a fazer com que o tamanho de amostra seja função do Nível de Confiança e da precisão (e_0).

Tamanho Mínimo de Amostra para Estimação por Intervalo da Média Populacional

a) Variância populacional conhecida:

$$e_0 = Z_{\text{crítico}} \times \frac{\sigma}{\sqrt{n}} \quad \text{isolando } n: n = \left(\frac{Z_{\text{crítico}} \times \sigma}{e_0} \right)^2$$

Nesse caso, basta especificar o valor de e_0 (na **mesma unidade** do desvio padrão populacional σ) e o Nível de Confiança (que será usado para encontrar o $Z_{\text{crítico}}$) e calcular o tamanho mínimo de amostra.

b) Variância populacional desconhecida

$$e_0 = t_{n-1, \text{crítico}} \times \frac{s}{\sqrt{n}} \quad \text{isolando } n: n = \left(\frac{t_{n-1, \text{crítico}} \times s}{e_0} \right)^2$$

O procedimento, nesse caso, seria semelhante, exceto por um pequeno problema: se estamos calculando o tamanho da amostra como podemos conhecer $n - 1$ e o desvio padrão amostral s ?

Quando a variância populacional da variável é desconhecida, o usual é retirar uma amostra piloto com um tamanho n^* arbitrário. A partir dos resultados dessa amostra, são calculadas as estatísticas (entre elas o desvio padrão amostral s), que são substituídas na expressão anterior.

Se $n \leq n^*$, então a amostra piloto é suficiente para o Nível de Confiança e a precisão exigidos.

Se $n > n^*$, então a amostra piloto é insuficiente para o Nível de Confiança e a precisão exigidos, sendo então necessário retornar à população e retirar os elementos necessários para completar o tamanho mínimo de amostra. O processo continua até que a amostra seja considerada suficiente.

Conforme visto na Unidade 2, se o tamanho da população for conhecido é recomendável corrigir o tamanho da amostra obtida, para

Amostra piloto – amostra teste, de tamanho arbitrado pelo pesquisador, a partir da qual são calculadas estatísticas necessárias para a determinação do tamanho mínimo de amostra. Fonte: Costa Neto (2002).

o intervalo de confiança de média ou proporção, através da seguinte fórmula:

$$n_{\text{corrigido}} = \frac{N \times n}{N + n} \text{ onde } N \text{ é o tamanho da população}$$

Assim procedendo, evitaremos o inconveniente de se obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

Considerem, neste exemplo 3, os dados do Exemplo 1. Para estimar a média, com 1% de significância e precisão de 0,2 kg, essa amostra é suficiente

Como a variância populacional é desconhecida, e o tamanho da amostra é menor do que 30 elementos, não obstante a população ter distribuição normal, a distribuição amostral da média será **t de Student**, e a variável de teste será: t_{n-1}

Assim, será usada a seguinte expressão para calcular o tamanho mínimo de amostra para a estimação por intervalo da média populacional.

$$n = \left(\frac{t_{n-1, \text{crítico}} \times s}{e_0} \right)^2$$

O nível de significância é o mesmo do item a. Sendo assim, o valor crítico continuará sendo o mesmo: $t_{n-1, \text{crítico}} = 5,84$. O desvio padrão amostral vale 0,4 kg; e o valor de e_0 , a precisão, foi fixada em 0,2 kg. Basta, então, substituir os valores na expressão:

$$n = \left(\frac{t_{n-1, \text{crítico}} \times s}{e_0} \right)^2 = \left(\frac{5,84 \times 0,4}{0,2} \right)^2 = 136,42 \cong 137 \text{ elementos}$$

Conclui-se que a amostra retirada é insuficiente, pois é menor do que o valor calculado acima.

Tamanho Mínimo de Amostra para Estimação por Intervalo da Proporção Populacional

Para a proporção populacional teremos:

$$e_0 = Z_{\text{crítico}} \times \sqrt{\frac{p \times (1-p)}{n}} \text{ isolando } n: n = \left(\frac{Z_{\text{crítico}}}{e_0} \right)^2 \times p \times (1-p)$$

Essa solução somente é usada quando a natureza da pesquisa é tal que não é possível retirar uma amostra piloto: a retirada de uma amostra piloto, e a eventual retirada de novos elementos da população, poderia prejudicar muito o resultado da pesquisa. Paga-se, então, o preço por se ter uma amostra substancialmente maior do que o necessário.

É necessário especificar o Nível de Confiança (ou de Significância) que será usado para encontrar o $Z_{\text{crítico}}$ e o valor de e_0 (tomando o cuidado de que tanto e_0 quanto p e $1-p$ estejam **todos** como proporções adimensionais ou como percentuais) para que seja possível calcular o valor do tamanho mínimo de amostra.

Da mesma forma que no caso da Estimação da média, quando a variância populacional é desconhecida teremos que recorrer a uma amostra piloto. No cálculo do tamanho mínimo de amostra para a Estimação por Intervalo da proporção populacional há, porém, uma solução alternativa: utiliza-se uma estimativa exagerada da amostra, supondo o máximo valor possível para o produto $p \times (1-p)$, que ocorrerá quando ambas as proporções forem iguais a 0,5 (50%).

Conforme visto na Unidade 2, se o tamanho da população for conhecido, é recomendável corrigir o tamanho da amostra obtida, para o intervalo de confiança de média ou proporção, mediante a seguinte fórmula:

$$n_{\text{corrigido}} = \frac{N \times n}{N + n}, \text{ onde } N \text{ é o tamanho da população}$$

Assim procedendo, evitamos o inconveniente de se obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

Exemplo 4: considere o caso do Exemplo 2.

Suponha 99% de confiança e precisão de 1%, essa amostra é suficiente para estimar a proporção populacional.

De acordo com o Exemplo 2, é possível se utilizar a aproximação pela distribuição normal. A expressão para o cálculo do tamanho mínimo de amostra para a proporção populacional será:

$$n = \left(\frac{Z_{\text{crítico}}}{e_0} \right)^2 \times p \times (1-p)$$

Os valores de p e $1-p$ já são conhecidos:

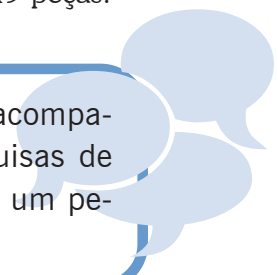
$$p = 0,035 \quad 1-p = 0,965$$

O nível de confiança exigido é de 99%: para encontrar o valor crítico devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,995 (0,99+0,005); os valores críticos serão $Z_{0,005}$ e $Z_{0,995}$, os quais serão iguais em módulo. E o valor de $Z_{\text{crítico}}$ será igual a 2,58 (em módulo).

A precisão foi fixada em 1% (0,01). Substituindo os valores na expressão anterior:

$$n = \left(\frac{Z_{\text{crítico}}}{e_0} \right)^2 \times p \times (1-p) = \left(\frac{2,58}{0,01} \right)^2 \times 0,035 \times 0,965 = 2.248,14 \cong 2.249$$

Observe que o tamanho mínimo de amostra necessário para atender a 99% de confiança e precisão de 1% deveria ser de 2.249 elementos. Como a amostra coletada possui apenas 1.000 elementos, ela é insuficiente para a confiança e precisão exigidas. Recomendamos o retorno à população para a retirada aleatória de mais 1.249 peças.



Depois do que estudamos, você já pode acompanhar atentamente os resultados das pesquisas de opinião veiculadas na mídia. Apenas mais um pequeno adendo.

“Empate Técnico”

Estamos acostumados a ouvir declarações do tipo “os candidatos A e B estão tecnicamente empatados na preferência eleitoral”. O que significa isso? Geralmente as pesquisas de opinião eleitoral consistem em obter as proporções de entrevistados que declara votar neste ou naquele candidato, naquele momento. Posteriormente as proporções são generalizadas estatisticamente para a população, através do cálculo de intervalos de confiança para as proporções de cada candidato. Se os intervalos de confiança das proporções de dois ou mais candidatos apresentam grandes superposições declara-se que há um “empate técnico”: as diferenças entre eles devem-se provavelmente ao acaso, e para todos os fins estão em condições virtualmente iguais, naquele momento.

Neste exemplo 5, imagine que uma pesquisa de opinião eleitoral apresentasse os seguintes resultados (intervalos de confiança para a proporção que declara votar no candidato) sobre a prefeitura do município de Tapioca. Quais candidatos estão tecnicamente empatados (Quadro 22)?

OPINIÃO	LIMITE INFERIOR %	LIMITE SUPERIOR %
Godofredo Astrogildo	31%	37%
Filismino Arquibaldo	14%	20%
Urraca Hermengarda	13%	19%
Salustiano Quintanilha	22%	28%
Indecisos	11%	17%

Quadro 22: Resultados de uma pesquisa eleitoral municipal (fictícia)

Fonte: Elaborado pelo autor deste livro

Filismino e Urraca estão tecnicamente empatados, pois seus intervalos de confiança apresentam grande sobreposição. Godofredo está muito na frente, pois o limite inferior de seu intervalo é maior do que o limite superior de Salustiano, que está em segundo lugar. É importante ressaltar que o número de indecisos é razoável, variando de 11 a 17%, quando eles se decidirem poderão mudar completamente o quadro da eleição, ou garantir a vitória folgada de Godofredo.

Saiba mais...

Sobre propriedades e características desejáveis de um estimador, BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulo 7.

Sobre estimadores e intervalos de confiança para variância, TRIOLA, Mario. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999, Capítulo 6.

Para entender melhor o conceito de distribuição amostral e sua relação com estimação de parâmetros, veja o arquivo Estima.xls, e suas instruções, no ambiente virtual;

Sobre a utilização do Microsoft Excel para realizar estimação por intervalo, LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística: Teoria e Aplicações – Usando Microsoft Excel em Português*. 5. ed. Rio de Janeiro: LTC, 200, Capítulo 6.



O resumo desta Unidade está mostrado na Figura 72:

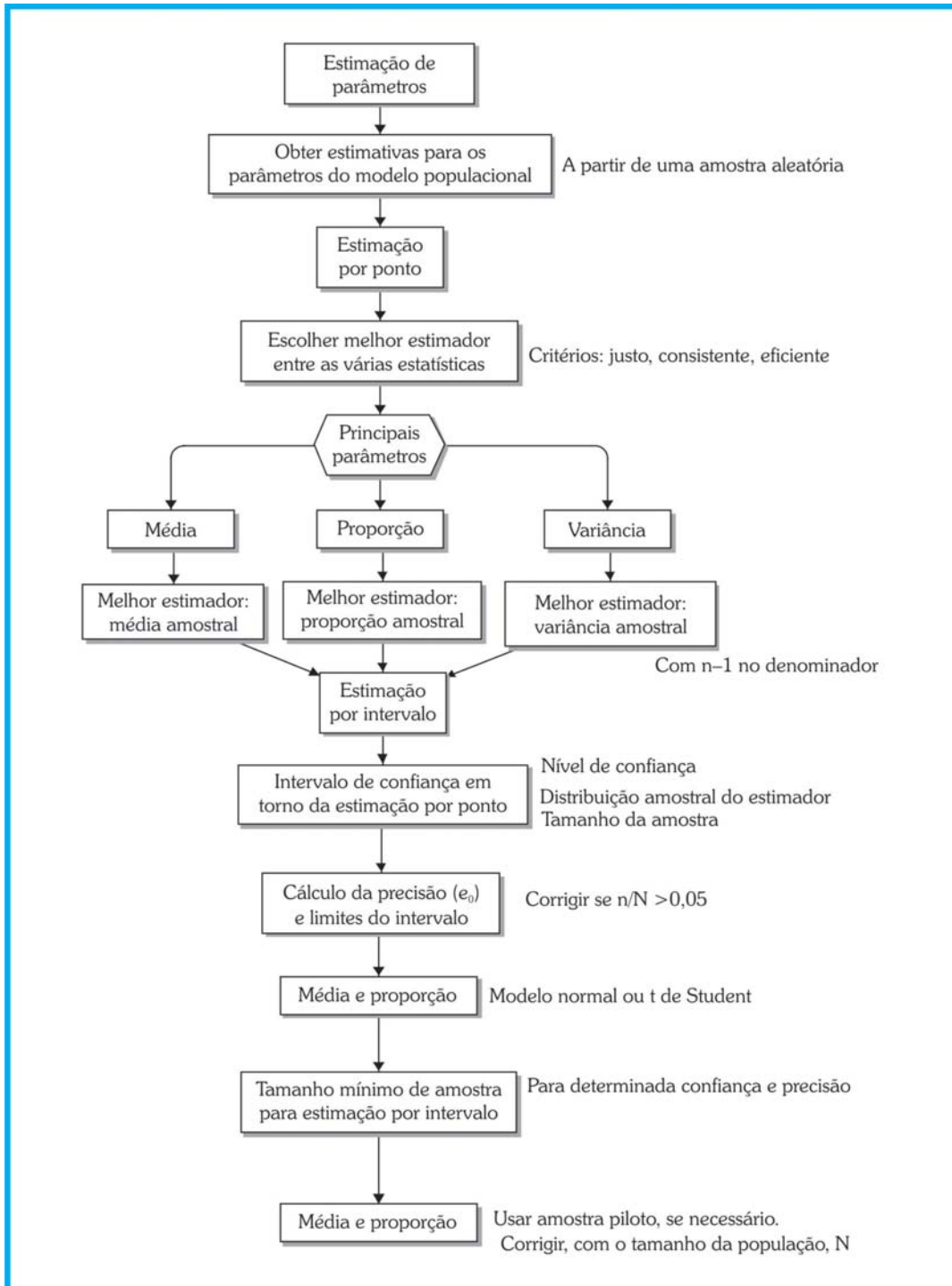


Figura 72: Resumo da Unidade 9
Fonte: Elaborada pelo autor deste livro

Vimos, nesta Unidade, os conceitos de Estimação de Parâmetros. Aprendemos a estimar os parâmetros média de uma variável quantitativa e proporção de um dos valores de uma variável qualitativa além de definir o tamanho mínimo de uma amostra aleatória para estimar média e proporção. Veremos mais sobre esse assunto na última Unidade deste livro. Estamos próximos do final do nosso material e é de suma importância a continuidade da interação com seus colegas e professor. Não deixe de ver as tabelas indicadas no livro e disponíveis no Ambiente Virtual de Ensino-Aprendizagem e de realizar a atividade de aprendizagem.



Atividades de aprendizagem

1. Buscando melhorar a qualidade do serviço, uma empresa estuda o tempo de atraso na entrega dos pedidos recebidos. Supondo que o tempo de atraso se encontra normalmente distribuído, e conhecendo o tempo de atraso dos últimos 20 pedidos, descritos a seguir (em dias), determine:

5 1 0 3 6 10 2 3 4 1 5 3 1 6 6 9 0 0 1 0

- a) Estime o atraso médio na entrega dos pedidos com confiança de 90%.
- b) Se fosse conhecido que a população possui desvio padrão igual a 2 dias, como ficaria a resposta do item a)?
- c) Para a situação do item a (variância populacional desconhecida), o tamanho da amostra é suficiente, se é necessária uma precisão de 0,5 dias, para o mesmo nível de confiança?

2. A satisfação da população em relação a determinado governo foi pesquisada através de uma amostra com a opinião de 1.000 habitantes do estado. Destes, 585 se declararam insatisfeitos com a administração estadual. Admitindo-se um nível de significância de 5%, solucione os itens a seguir.

a) Estime o percentual da população que está insatisfeita com a administração estadual.

b) Qual o tamanho da amostra necessária para a estimação se a empresa responsável pela pesquisa estipulou uma folga máxima de 2,5%?

3. Os índices apresentados pelos alunos do curso de Economia e de Administração estão sendo questionados por eles, no sentido de definirem se há diferença entre os cursos. Para tanto foram analisados os índices de 10 alunos de cada curso, escolhidos aleatoriamente dentre os regularmente matriculados e anotados seus valores, onde se obteve:

Economia	– média 7,3	desvio padrão 2,6
Administração	– média 7,1	desvio padrão 3,1

a) Estime os valores médios dos índices de cada curso com 95% de confiança.

b) Para o mesmo nível de confiança de a. Será que 10 alunos é uma amostra suficiente, em ambos os cursos, para estimar seus índices médios, com uma precisão igual a 1?

4. O CRA-SC está conduzindo uma pesquisa acerca da opinião dos acadêmicos de administração sobre seus respectivos cursos. Suspeita-se que haja diferença entre as proporções de satisfeitos de instituições públicas e privadas: os acadêmicos das públicas seriam mais satisfeitos. Para avaliar essa suposição, foi conduzida uma pesquisa por amostragem, entrevistando alunos de duas instituições públicas, SHUFSC e GASE, e de três privadas, PATÁPIO de SÁ, UNIMALI e UNILUS. Os resultados estão na tabela a seguir:

MEDIDAS	UNIVERSIDADES				
	SHUFSC	GASE	PATÁPIO	UNIMALI	UNILUS
n	120	165	185	194	189
p	0,55	0,48	0,32	0,49	0,25
N (população)	890	900	1500	1200	1800

Usando 1 % de significância responda os itens a seguir:

- Estime a proporção populacional de satisfeitos com o curso, em cada universidade.
- Para uma margem de erro de 2%, qual deveria ser o tamanho de amostra para estimar a proporção de satisfeitos em cada universidade?

10

UNIDADE

Testes de Hipóteses



Ao finalizar esta Unidade, você deverá ser capaz de compreender os conceitos de Testes de Hipóteses, especialmente para média de uma variável quantitativa; proporção de um dos valores de uma variável quantitativa; e associação entre duas variáveis qualitativas; e, ainda, capaz de aplicá-los e de perceber a importância deles no dia a dia do administrador.

Caro estudante, você viu anteriormente que determinada população pode ser descrita através de um modelo, que apresenta características e parâmetros. Muitas vezes, esses parâmetros são desconhecidos e há interesse em estimá-los para se obter melhor conhecimento sobre a população. Então, retira-se uma amostra aleatória da população e, através das técnicas de Estimação de Parâmetros (Unidade 9), procuramos obter a estimativa de algum parâmetro de interesse e, a esta, associamos probabilidade de a estimativa estar correta. Nessa última e importantíssima Unidade, veremos que a Estimação de Parâmetros é uma subdivisão da Inferência Estatística (que consiste em fazer afirmações probabilísticas sobre o modelo da população a partir de uma amostra aleatória dessa população), a outra grande subdivisão constitui os **Testes de Hipóteses**. Vamos saber mais!

Testes paramétricos – testes de hipóteses sobre parâmetros do modelo da variável sob análise Fonte: Elaborado pelo autor deste livro.

Testes não paramétricos – testes de hipóteses sobre outros aspectos do modelo da variável sob análise ou alternativas aos testes paramétricos quando as condições para uso destes não forem satisfeitas. Fonte: Elaborado pelo autor deste livro.

Contrariamente à Estimação de Parâmetros, os Testes de Hipóteses permitem fazer inferências sobre outras características do modelo da população além dos parâmetros, como a forma do modelo da população, por exemplo. Quando os Testes são feitos sobre os parâmetros da população, são chamados de Testes Paramétricos; e, quando feitos sobre outras características, chamados de Testes Não Paramétricos. Não obstante, vamos nos restringir aos Testes Paramétricos: da média de uma variável quantitativa e da proporção de um dos valores de uma variável qualitativa.

Vimos que determinada população pode ser descrita através de um modelo probabilístico, que apresenta características e parâmetros.

Tô a fim de saber

Na realidade a denominação correta deveria ser “Testes dependentes de distribuição de referência” (porque para fazer inferências sobre os parâmetros devemos supor que o modelo probabilístico populacional é normal, por exemplo, ou que a distribuição amostral do parâmetro pode ser aproximada por uma normal). Já os não paramétricos deveriam ser “Testes livres de distribuição” (porque os Testes Não Paramétricos não exigem que os dados tenham uma aderência a certo modelo).

Muitas vezes, esses parâmetros são desconhecidos e há interesse em estimá-los para obtermos melhor conhecimento sobre a população. Nesse caso, retiramos uma amostra aleatória da população e, por intermédio das técnicas de **Estimação de Parâmetros**, procuramos obter a estimativa de algum parâmetro de interesse e, a ela, associamos a probabilidade de a estimativa estar correta. A Estimação de Parâmetros compreende uma das subdivisões da Inferência Estatística, que realiza afirmações probabilísticas sobre o modelo probabilístico da população a partir de uma amostra aleatória dessa população. Os **Testes de Hipóteses** constituem a outra grande subdivisão.

Imagine que determinado pesquisador está interessado em alguma característica de uma população. Devido a estudos prévios, ou simplesmente por bom senso (melhor ponto de partida para o estudo), ele estabelece que a característica terá determinado comportamento. Formula então uma hipótese estatística sobre a característica da população e essa hipótese é aceita como válida até prova estatística em contrário.

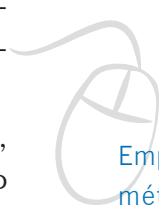
Para testar a hipótese, é coletada uma amostra aleatória representativa da população, sendo calculadas as estatísticas necessárias para o teste. Naturalmente, devido ao fato de ser utilizada uma amostra aleatória, haverá diferenças entre o que se esperava, sob a condição da hipótese verdadeira, e o que realmente foi obtido na amostra. A questão a ser respondida é: as diferenças são significativas o bastante para que a hipótese estatística estabelecida seja rejeitada? Essa não é uma pergunta simples de responder: dependerá do que está sob teste (que parâmetro, por exemplo), da confiabilidade desejada para o resultado, entre outros. Basicamente, porém, será necessário comparar as diferenças com uma referência, a distribuição amostral de um parâmetro, por exemplo, que supõe que a hipótese sob teste é verdadeira: a comparação costuma ser feita através de uma estatística de teste que envolve os valores da amostra e os valores sob teste.

A tomada de decisão é feita da seguinte forma:

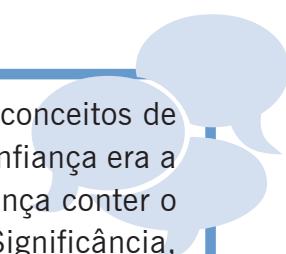
- se a diferença entre o que foi observado na amostra e o que era esperado (sob a condição da hipótese verdadeira) não for **significativa**, a hipótese será **aceita**; e
- se a diferença entre o que foi observado na amostra e o que era esperado (sob a condição da hipótese verdadeira) for significativa, a hipótese será **rejeitada**.

O valor a partir do qual a diferença será considerada significativa será determinado pelo Nível de Significância do teste. O Nível de Significância geralmente é fixado pelo pesquisador, muitas vezes de forma arbitrária, e também será a probabilidade de erro do Teste de Hipóteses: a probabilidade de cometer um erro no teste, rejeitando uma hipótese válida. Como a decisão do teste é tomada a partir dos dados de uma amostra aleatória da população, há sempre a probabilidade de se cometer um erro, mas com a utilização de métodos estatísticos é possível calcular o valor dessa probabilidade.

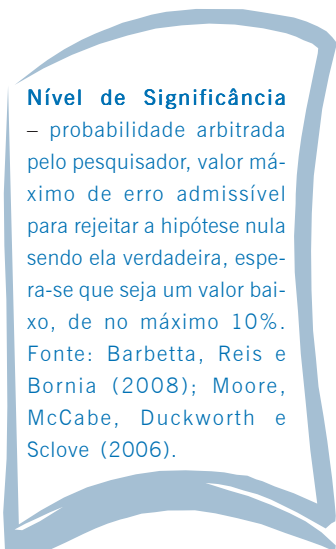
O **Nível de Significância** é uma probabilidade. Portanto, trata-se de um número real que varia de 0 a 1 (0 a 100%). Como existe a probabilidade de se cometer um erro no teste, é interessante que seja o mais próximo possível de zero: valores típicos são 5%, 10%, 1% e até menores dependendo do problema sob análise. Contudo, não é possível usar um Nível de Significância igual a zero porque, devido ao uso de uma amostra aleatória, sempre haverá chance de erro; a não ser que a amostra fosse do tamanho da população. O complementar do Nível de Significância é chamado de **Nível de Confiança**, pois este indica a confiabilidade do resultado obtido, a probabilidade de a decisão tomada estar correta.



Empregando-se outros métodos (empíricos) não há como ter ideia da chance de erro (pode ser um erro de 0% ou de 5.000%...).



Você deve estar lembrado destes dois conceitos de Estimação de Parâmetros: Nível de Confiança era a probabilidade de o Intervalo de Confiança conter o valor real do parâmetro, e Nível de Significância, complementar daquele, diz respeito a probabilidade de o Intervalo não conter o parâmetro. Em suma, a probabilidade da Estimação estar correta ou não.



Nível de Significância
– probabilidade arbitrada pelo pesquisador, valor máximo de erro admissível para rejeitar a hipótese nula sendo ela verdadeira, espera-se que seja um valor baixo, de no máximo 10%.
Fonte: Barbetta, Reis e Bornia (2008); Moore, McCabe, Duckworth e Sclove (2006).

Tipos de Hipóteses

Para realizar um Teste de Hipóteses, é necessário definir (enunciar) duas Hipóteses Estatísticas complementares (que abrangem todos os resultados possíveis): a chamada **Hipótese Nula** (denotada por H_0) e a **Hipótese Alternativa** (denotada por H_1 ou H_a). Enunciar as hipóteses é o primeiro e possivelmente mais im-

portante passo de um Teste de Hipóteses, pois todo o procedimento dependerá dele.

A Hipótese Nula (H_0) é a hipótese estatística aceita como verdadeira até prova estatística em contrário: pode ser o ponto de partida mais adequado para o estudo, ou exatamente o contrário do que o pesquisador quer provar (ou o contrário daquilo que o preocupa).

A Hipótese Alternativa (H_1), que será uma hipótese complementar de H_0 , fornecerá uma alternativa à hipótese nula: muitas vezes é justamente aquilo que o pesquisador quer provar (ou aquilo que o preocupa).

Quando as hipóteses são formuladas sobre os parâmetros do modelo probabilístico da população, o Teste de Hipóteses é chamado de Paramétrico. Quando as hipóteses são formuladas sobre outras características do modelo, o Teste é chamado de Não Paramétrico.

A decisão do teste consiste em aceitar ou rejeitar a Hipótese Nula (H_0): aceitar-se ou não a hipótese até então considerada verdadeira.

É importante ter a noção do que significa aceitar ou rejeitar a Hipótese Nula (H_0). A decisão é tomada sobre essa hipótese e não sobre a Hipótese Alternativa, porque é a Hipótese Nula que é considerada verdadeira (até prova em contrário). Quando se aceita a Hipótese Nula, significa que não há provas suficientes para rejeitá-la. Quando, no entanto, a decisão é por se rejeitar a Hipótese Nula, há evidências suficientes de que as diferenças obtidas (entre o que era esperado e o que foi observado na amostra) não ocorreram por acaso. Usando uma analogia com o direito dos EUA, aceitar H_0 seria comparável a um veredito de não culpado “*not guilty*”, ou seja, não há provas suficientes para condenar o réu. Por outro lado, rejeitar H_0 seria comparável a um veredito de culpado “*guilty*”, ou seja, as provas reunidas são suficientes para condenar o réu. O Nível de Significância será a probabilidade assumida de **Rejeitar H_0 sendo H_0 verdadeira**.

Tipos de Testes Paramétricos

A formulação das hipóteses é o ponto inicial do problema, e deve depender única e exclusivamente das conclusões que se pretende obter com o teste. A formulação da hipótese alternativa determinará o tipo de teste: se Unilateral ou Bilateral.

Se a formulação da hipótese alternativa indicar que o parâmetro é maior ou menor do que o valor de teste (valor considerado verdadeiro até prova em contrário), o teste será **Unilateral**: somente há interesse se as diferenças entre os dados da amostra e o valor de teste forem em determinada direção. Se a formulação da hipótese alternativa indicar que o parâmetro é diferente do valor de teste, o teste será **Bilateral**: há interesse nas diferenças em qualquer direção. As hipóteses então seriam:

Testes Unilaterais

H_0 : parâmetro = valor de teste.

H_1 : parâmetro < valor de teste.

H_0 : parâmetro = valor de teste.

H_1 : parâmetro > valor de teste.

Testes Bilaterais

H_0 : parâmetro = valor de teste.

H_1 : parâmetro $\neq$ valor de teste.

A escolha do tipo de teste dependerá das condições do problema sob estudo. Sejam as três situações a seguir:

- Um novo protocolo de atendimento foi implementado no Banco RMG com objetivo de reduzir o tempo que as pessoas passam na fila do caixa. O protocolo será considerado satisfatório se a média do tempo de fila for menor do que 30 minutos. Um teste **Unilateral** seria o adequado.
- Cerca de 2.000 formulários de pedidos de compra estão sendo analisados. Os clientes podem ficar insatisfeitos se houver erros nos formulários. Nesse caso, admite-se que a proporção máxima de formulários com erros seja de 5%.

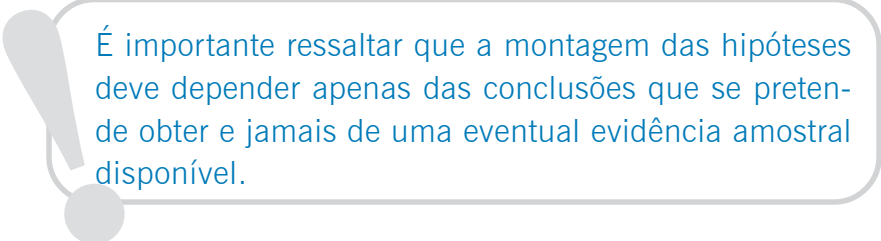
Teste unilateral – teste no qual a região de rejeição da hipótese nula está concentrada em apenas um dos lados da distribuição amostral da variável de teste. Fonte: Barbetta, Reis e Bornia (2008).

Teste bilateral – teste no qual a região de rejeição da hipótese nula está dividida em duas partes, em cada um dos lados da distribuição amostral da variável de teste. Fonte: Barbetta, Reis e Bornia (2008).

Ou seja, um valor maior do que 5% causaria problemas. Um teste **Unilateral** seria o adequado.

- c) Uma peça automotiva precisa ter 100 mm de diâmetro, exatamente. Nesse caso, a dimensão não pode ser maior ou menor do que 100 mm (em outras palavras não pode ser diferente de 100 mm), pois isso indicará que a peça não está de acordo com as especificações. Um teste **Bilateral** seria o adequado.

Após definir as hipóteses, é coletada uma amostra aleatória da população para testá-las.



É importante ressaltar que a montagem das hipóteses deve depender apenas das conclusões que se pretende obter e jamais de uma eventual evidência amostral disponível.

A decisão de aceitar ou de rejeitar H_0 dependerá das **regiões de aceitação e rejeição de H_0** , que por sua vez dependem dos seguintes fatores:

- do parâmetro sob teste (e da estatística ou variável de teste usada para testá-lo).
- do tipo de teste, se Unilateral ou Bilateral.
- do valor de teste (valor do parâmetro considerado verdadeiro até prova em contrário).
- do Nível de Significância (σ) ou Nível de Confiança ($1 - \sigma$) adotado.
- de um valor crítico da estatística ou variável de teste a partir do qual a hipótese será rejeitada, e esse valor dependerá por sua vez do Nível de Significância, do tipo de teste e da **Distribuição Amostral** do parâmetro.

A **Região de Aceitação de H_0** será a faixa de valores da estatística (ou da variável de teste) associada ao parâmetro em que as diferenças entre o que foi observado na amostra e o que era esperado não são significativas.

A **Região de Rejeição de H_0** será a faixa de valores da estatística (ou da variável de teste) associada ao parâmetro em que as diferenças entre o que foi observado na amostra e o que era esperado **são significativas**.

Essa abordagem é chamada de abordagem clássica dos testes de hipóteses. Há também a do valor-p, muito usada por programas computacionais. Mas, neste texto, vamos usar apenas a clássica por considerá-la mais clara.

Para entendermos melhor os conceitos acima, observe a situação a seguir:

Há interesse em realizar um teste de hipóteses sobre o comprimento médio de uma das dimensões de uma peça mecânica. O valor nominal da média (aceito como verdadeiro até prova em contrário) é igual a “**b**” (valor genérico), $H_0: \mu = \mathbf{b}$. Supondo que a distribuição amostral do parâmetro (distribuição de $\bar{x}$) seja normal, e será centrada em **b**, é possível fazer a conversão para a distribuição normal padrão (média zero e desvio padrão 1, variável Z). Repare nas Figuras 73 e 74 a seguir.

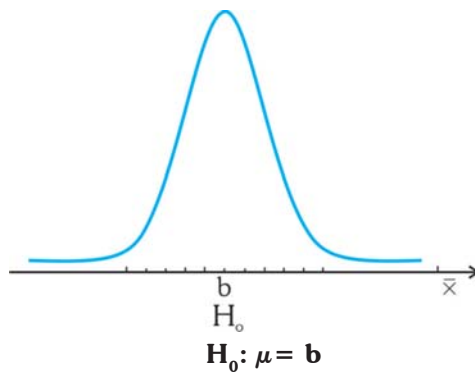


Figura 73: Hipótese nula: média populacional = b
Fonte: Elaborada pelo autor deste livro

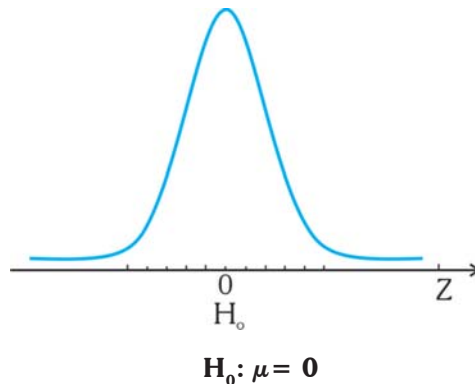


Figura 74: Hipótese nula normal padrão
Fonte: Elaborada pelo autor deste livro

O valor de **b** (média da dimensão e média de $\bar{x}$) corresponde a zero, média da variável **Z**. Dependendo da formulação da Hipótese Alternativa, haveria diferentes Regiões de Rejeição de **H₀**.

Se a Hipótese Alternativa fosse: **H₁: $\mu < b$** (**H₁: $\mu < 0$**), ou seja, se o teste fosse Unilateral à esquerda, a Região de Rejeição de **H₀** seria:

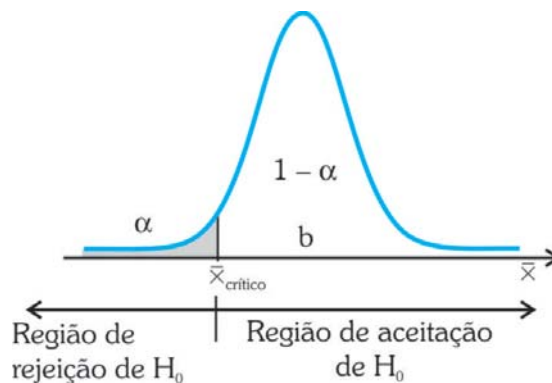


Figura 75: **H₁: $\mu < b$**

Fonte: Elaborada pelo autor deste livro

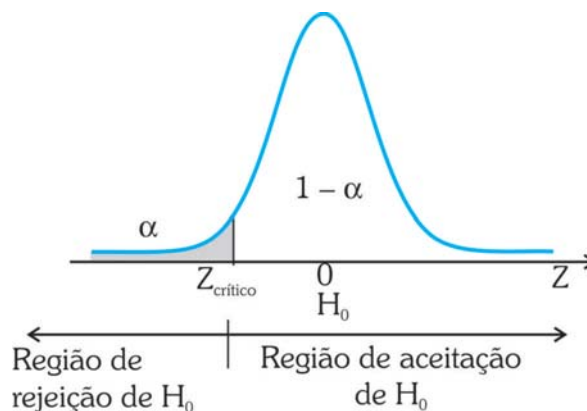


Figura 76: **H₁: $\mu < 0$**

Fonte: Elaborada pelo autor deste livro

Observe que há um valor crítico de $\bar{x}$: abaixo dele a Hipótese Nula será rejeitada, acima será aceita. A determinação do valor é feita com base no Nível de Significância, a área abaixo da curva normal até o valor crítico de $\bar{x}$. Geralmente obtém-se o valor crítico da variável de teste (**Z** neste caso, na segunda figura) através de uma tabela, que corresponde ao valor crítico de $\bar{x}$, faz-se a transformação de variáveis $Z = \frac{(\bar{x} - \mu_0)}{\sigma}$ e obtém-se o valor crítico de $\bar{x}$. μ_0 é o valor

sob teste (**b** no exemplo) e σ é um desvio padrão (cujo valor será explicitado posteriormente).

A decisão será tomada comparando-se valor da média amostral $\bar{x}$ com o valor crítico dessa mesma média: se for menor do que o valor crítico $\bar{x}_{\text{crítico}}$, ou seja, está na região de **Rejeição de H_0** ; desse modo, rejeita-se a Hipótese Nula. É muito comum também tomar a decisão comparando o valor da variável de teste (**Z** neste caso), obtido com base nos dados da amostra, com o valor crítico $Z_{\text{crítico}}$ desta mesma variável (obtido de uma tabela ou programa computacional): se for menor do que o valor crítico, rejeita-se a Hipótese Nula. Observe que o valor do Nível de Significância α é colocado na curva referente à Hipótese Nula, porque é esta que é aceita como válida até prova em contrário. Observe também que a faixa de valores da região de Rejeição pertence à curva da Hipótese Nula, assim o valor α é a probabilidade de Rejeitar **H_0** sendo **H_0** verdadeira.

Nesta altura, é importante ressaltar um ponto que costuma passar despercebido. Se a decisão for tomada com base na variável de teste (**Z**, por exemplo) é interessante notar que, como o teste é Unilateral à esquerda, o valor $Z_{\text{crítico}}$ será **negativo**, uma vez que a região de Rejeição de **H_0** está à **esquerda** de 0 (menor do que zero). No teste Unilateral à direita, que veremos a seguir, o valor de $Z_{\text{crítico}}$ será positivo, pois a região de Rejeição de **H_0** estará à **direita** de 0 (maior do que zero). Se, por exemplo, o Nível de Significância fosse de 5% (0,05), o valor de $Z_{\text{crítico}}$ para o teste Unilateral à esquerda seria -1,645. Se houvesse interesse em obter o valor de $\bar{x}_{\text{crítico}}$ correspondente, bastaria usar a expressão $Z = (\bar{x} - \mu_0)/\sigma$ substituindo **Z** por -1,645.

Se a Hipótese Alternativa fosse **$H_1: \mu > b$** (**$H_1: \mu > 0$**), ou seja, se o teste fosse Unilateral à direita, a Região de Rejeição de **H_0** seria:

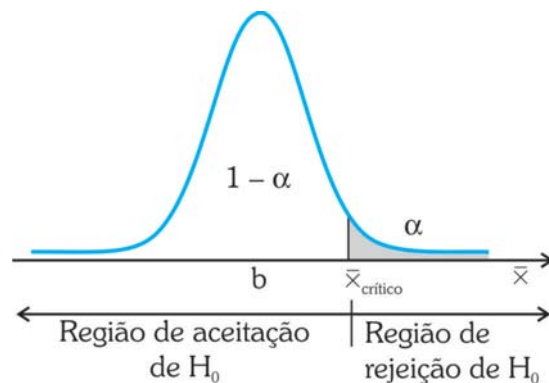


Figura 77: $H_1: \mu > b$

Fonte: Elaborada pelo autor deste livro

Probabilidade de tomar uma decisão errada fixada pelo pesquisador.

O sinal correto é importante para que o valor seja coerente com a posição da região de Rejeição de **H_0** .

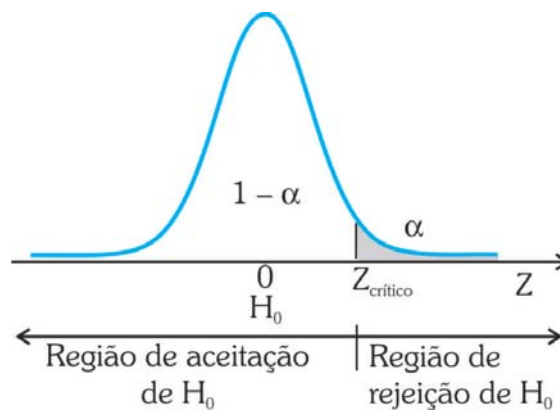


Figura 78: $H_1: \mu > 0$

Fonte: Elaborada pelo autor deste livro

Nesse caso, o valor crítico está à direita: se a média amostral $\bar{x}$ ou a variável de teste Z tiverem valores superiores aos respectivos valores críticos, a Hipótese Nula será rejeitada, pois os valores “caíram” na região de Rejeição de H_0 . Como foi notado anteriormente, o valor de $Z_{\text{crítico}}$ será positivo, pois é maior do que zero: adotando-se o mesmo Nível de Significância de 5%, o valor de $Z_{\text{crítico}}$ seria 1,645, igual em módulo ao anterior, uma vez que a distribuição normal padrão é simétrica em relação à sua média, que é igual a zero.

Se a Hipótese Alternativa fosse $H_1: \mu \neq b$ ($H_1: \mu \neq 0$), ou seja, o teste fosse Unilateral à direita, a Região de Rejeição de H_0 seria:

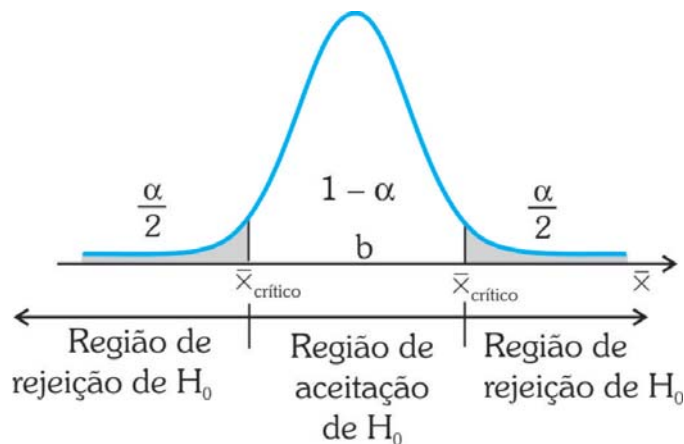


Figura 79: $H_1: \mu \neq b$

Fonte: Elaborada pelo autor deste livro

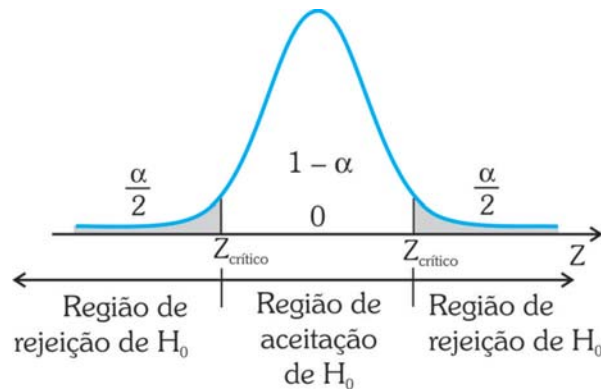


Figura 80: $H_1: \mu \neq 0$

Fonte: Elaborada pelo autor deste livro

Nesse caso, a região de Rejeição se divide em duas iguais (probabilidades iguais à metade do Nível de Significância α), semelhante ao que acontece na Estimação por Intervalo. Existirão dois valores críticos, um abaixo do valor de teste e outro acima: se a média amostral $\bar{x}$ ou a variável de teste Z tiverem valores acima do valor crítico “superior” ou abaixo do valor crítico “inferior”, a Hipótese Nula será rejeitada, pois os valores “caíram” em uma das duas regiões de Rejeição. Se for usada a variável de teste Z , os valores críticos serão iguais em módulo, pois estão à mesma distância do valor sob teste (zero).

Recordando as três situações que foram abordadas anteriormente, seria interessante definir completamente as Hipóteses Estatísticas. Nos dois primeiros casos, optou-se por um Teste Unilateral e, no terceiro, por um Teste Bilateral.

- a) Um novo protocolo de atendimento foi implementado no Banco RMG com objetivo de reduzir o tempo que as pessoas passam na fila do caixa. O protocolo será considerado satisfatório se a média do tempo de fila for menor do que 30 minutos. Um teste **Unilateral** seria o adequado. Mas Unilateral à Esquerda ou à Direita? Como está grifado na frase anterior, haverá problema se a média do tempo for menor que 30, resultando:

Teste **unilateral à esquerda**:

$H_0: \mu = 30$ onde $\mu_0 = 30$ (valor de teste); e

$H_1: \mu < 30$ Teste Unilateral à Esquerda.

- b) Cerca de 2.000 formulários de pedidos de compra estão sendo analisados. Os clientes podem ficar insatisfeitos se houver erros nos formulários. Assim, admite-se que a proporção máxima de formulários com erros seja de 5%. Isto é, um valor maior que 5% causaria problemas. Um teste **Unilateral** seria o adequado. Nesse caso, um teste de proporção; o problema será um valor maior que 5%, resultando:

Teste unilateral à direita:

$H_0: \pi = 5\%$, onde $\pi_0 = 5\%$ (valor de teste); e

$H_1: \pi > 5\%$.

- c) Uma peça automotiva precisa ter 100 mm de diâmetro, exatamente. Nesse caso, a dimensão não pode ser maior ou menor que 100 mm, pois isso indicará que a peça não está de acordo com as especificações. Um teste **Bilateral** seria o adequado, resultando:

Teste Bilateral:

$H_0: \mu = 100 \text{ mm}$, onde $\mu_0 = 100 \text{ mm}$ (valor de teste); e

$H_1: \mu \neq 100 \text{ mm}$.

Para a definição correta das hipóteses, é imprescindível a correta identificação do valor de teste, pois se trata de um dos aspectos mais importantes deste: o resultado da amostra será comparado ao valor de teste.

Lembrando novamente que a tomada de decisão depende da correta determinação da região de Rejeição (e, por conseguinte, de Aceitação) da Hipótese Nula, e isso, por sua vez, depende diretamente da formulação das Hipóteses Estatísticas.

Testes de Hipóteses sobre a Média de uma Variável em uma População

Nesse caso, há interesse em testar a hipótese de que o parâmetro média populacional (μ) de certa variável quantitativa seja maior, menor ou diferente de certo valor. Para a realização desse teste, é necessário que uma das condições a seguir seja satisfeita:

- sabe-se, ou é razoável supor, que a variável de interesse segue um modelo normal na população: isso significa que a distribuição amostral da média também será normal, permitindo realizar a inferência estatística paramétrica;
- a distribuição da variável na população é desconhecida, mas a amostra retirada dessa população é considerada “suficientemente grande”; o que, de acordo com o Teorema Central do Limite, permite concluir que a distribuição amostral da média é normal; e
- supõe-se que a amostra é representativa da população e foi retirada de forma aleatória.

Tal como na Estimação de Parâmetros por Intervalo, existirão diferenças nos testes, dependendo do conhecimento ou não da variância populacional da variável.

a) Se a variância populacional (σ^2) da variável (cuja média populacional se quer testar) for conhecida.

Nesse caso, a variância amostral da média poderá ser calculada através da expressão:

$$V(\bar{x}) = \frac{\sigma^2}{n}, \text{ e, por conseguinte, o “desvio padrão” será: } \frac{\sigma}{\sqrt{n}}$$

A variável de teste será a variável Z da distribuição normal padrão, lembrando que:

$$Z = \frac{\text{valor} - \text{“média”}}{\text{“desvio-padrão”}}$$

A “**média**” será o valor de teste (suposto verdadeiro até prova em contrário) denotado por μ_0 . O **valor** (obtido pela amostra) será a média amostral (que é o melhor estimador da média populacional)

Há muita controvérsia a respeito do que seria uma amostra “suficientemente grande”, mas geralmente uma amostra com pelo menos 30 elementos costuma ser considerada grande o bastante para que a distribuição amostral da média possa ser aproximada por uma normal.

O desvio padrão é a raiz quadrada positiva da variância.

Neste caso Z crítico será negativo, já que a região de Rejeição de H_0 está à esquerda de zero.

E talvez este seja o motivo de se considerar mais de 30 elementos como sendo uma amostra suficientemente grande.

denotada por $\bar{x}$, e o “desvio padrão” será o valor obtido anteriormente. Sendo assim, a expressão da variável de teste Z :

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Compara-se o valor da variável de teste com o valor crítico ($Z_{\text{crítico}}$ que depende do Nível de Significância adotado) de acordo com o tipo de teste:

Se $H_1 \mu: > \mu_0$ Rejeitar H_0 se $Z > Z_{\text{crítico}}$ ($\bar{x} > \bar{x}_{\text{crítico}}$);

Se $H_1 \mu: < \mu_0$ Rejeitar H_0 se $Z < Z_{\text{crítico}}$ ($\bar{x} < \bar{x}_{\text{crítico}}$); e

Se $H_1 \mu: \neq \mu_0$ Rejeitar H_0 se $|Z| \neq |Z_{\text{crítico}}|$.

b) Se a variância populacional σ^2 da variável for desconhecida.

Naturalmente esse é o caso mais encontrado na prática. Como devemos proceder? Dependerá do tamanho da amostra.

b.1 – Grandes amostras (mais de 30 elementos):

Nesses casos, procede-se como no item anterior, apenas fazendo com que $\sigma = s$, ou seja, considerando que o desvio padrão da variável na população é igual ao desvio padrão da variável na amostra (suposição razoável para grandes amostras);

b.2 – Pequenas amostras (até 30 elementos):

Nesses casos, a aproximação do item b.1 não será viável. Terá que ser feita uma correção na distribuição normal padrão (Z) através da distribuição **t de Student**. Essa distribuição já é conhecida (ver Unidades 6 e 9). Trata-se de uma distribuição de probabilidades que possui média zero (como a distribuição normal padrão, variável Z), mas sua variância é igual a $n/(n-2)$, ou seja, a variância depende do tamanho da amostra. Quanto maior for o tamanho da amostra mais o quociente acima se aproxima de 1 (a variância da distribuição normal padrão), e mais a distribuição **t de Student** aproxima-se da distribuição normal padrão. A partir de $n = 30$, já é possível considerar a variância da distribuição **t de Student** aproximadamente igual a 1.

A variável de teste será então t_{n-1} (t com $n - 1$ graus de liberdade):

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

onde s é o desvio padrão amostral e os outros valores têm o mesmo significado da expressão anterior.

Compara-se o valor da variável de teste com o valor crítico ($t_{n-1, \text{crítico}}$ que depende do Nível de Significância adotado) de acordo com o tipo de teste:

Se $H_1: \mu > \mu_0$ Rejeitar H_0 se $t_{n-1} > t_{n-1, \text{crítico}}$ ($\bar{x} > \bar{x}_{\text{crítico}}$);

Se $H_1: \mu < \mu_0$ Rejeitar H_0 se $t_{n-1} < t_{n-1, \text{crítico}}$ ($\bar{x} < \bar{x}_{\text{crítico}}$); e

Se $H_1: \mu \neq \mu_0$ Rejeitar H_0 se $|t_{n-1}| \geq |t_{n-1, \text{crítico}}|$.

A correta identificação dos valores críticos, decorrente da correta identificação da região de rejeição de H_0 , por sua vez decorrente da adequada formulação das hipóteses estatísticas, é indispensável para que o resultado obtido seja coerente.

Uma peça automotiva precisa ter 100 mm de diâmetro, exatamente. Foram medidas 15 peças, aleatoriamente escolhidas. Obteve-se média de 100,7 mm e variância de 0,01 mm². Supõe-se que a dimensão segue distribuição normal na população. A peça está dentro das especificações? Usar 1% de significância? Vejamos o primeiro exemplo:

Enunciar as hipóteses.

Conforme visto anteriormente, o teste mais adequado para esse caso é um Teste Bilateral:

$H_0: \mu = 100$ mm, onde $\mu_0 = 100$ mm (valor de teste); e

$H_1: \mu \neq 100$ mm.

Nível de significância.

O problema declara que é necessário usar 1% de significância (se não fosse especificado, outro valor poderia ser usado).

Variável de teste.

Uma vez que a variância populacional da variável é **desconhecida** (o valor fornecido é a variância amostral) e a amostra retirada apresenta apenas 15 elementos (portanto menos de 30), a variável de teste será t_{n-1} da distribuição **t de Student**.

Nesse caso, $t_{n-1, \text{crítico}}$ será negativo já que a região de Rejeição de H_0 está à esquerda de zero.

Definir a região de aceitação de H_0 :

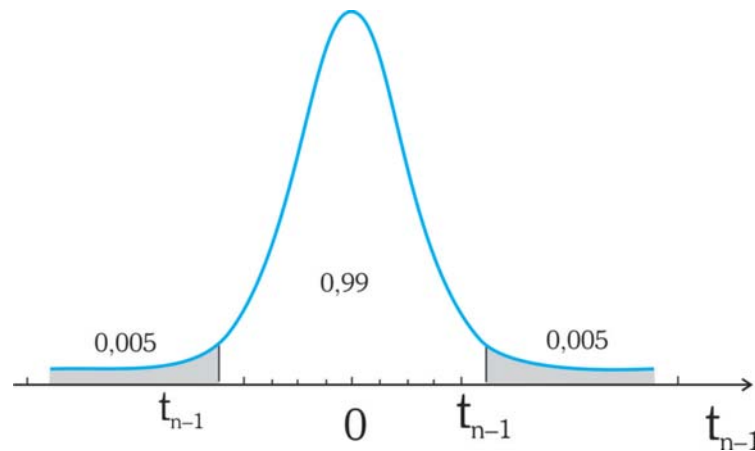


Figura 81: Regiões de rejeição e aceitação da hipótese nula – Teste bilateral de média

Fonte: Elaborada pelo autor deste livro

Observe que por ser um teste Bilateral, o Nível de Significância α foi dividido em dois, metade para cada região de rejeição de H_0 . Para encontrar o valor crítico, devemos procurar na tabela da distribuição de **Student**, na linha correspondente a $n-1$ graus de liberdade, ou seja, em $15 - 1 = 14$ graus de liberdade. O valor da probabilidade pode ser visto na figura anterior: os valores críticos serão $t_{14;0,005}$ e $t_{14;0,995}$ os quais serão iguais em módulo. E o valor de $t_{n-1,\text{crítico}}$ será igual a 2,977 (em módulo).

Através dos valores da amostra, avaliar o valor da variável.

Nesse ponto, é preciso encontrar o valor da variável de teste:

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

O valor de teste μ_0 é igual a 100 mm, a média amostral $\bar{x}$ vale 100,7 mm, o tamanho de amostra n é igual a 15 e o desvio padrão amostral s é a raiz quadrada de 0,01 mm². Substituindo na equação, teremos:

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = t_{15-1} = t_{14} = \frac{100,7 - 100}{\sqrt{0,01} / \sqrt{15}} = 27,11 \text{ então } |t_{14}| = 27,11$$

Decidir pela aceitação ou rejeição de H_0 .

Como se trata de um teste bilateral:

Rejeitar H_0 se $|t_{n-1}| \neq |t_{n-1,\text{crítico}}|$; e

Como $|t_{14}| = 27,11 > |t_{n-1,\text{crítico}}| = |t_{14,0,995}| = 2,977$.

Rejeitar H_0 a 1% de Significância (há 1% de chance de erro).

Interpretar a decisão no contexto do problema. Há provas estatísticas suficientes de que a dimensão da peça não está dentro das especificações.

Um novo protocolo de atendimento foi implementado no Banco RMG com objetivo de reduzir o tempo que as pessoas passam na fila do caixa. O protocolo será considerado satisfatório se a média do tempo de fila for **menor** que 30 minutos. Suponha que o tempo que 35 clientes (selecionados aleatoriamente) passaram na fila foi monitorado, resultando em uma média de 29 minutos e desvio padrão de 5 minutos. O protocolo pode ser considerado satisfatório a 5% de significância? Vejamos o segundo exemplo:

Esse problema já foi estudado anteriormente.

Enunciar as hipóteses. Conforme visto anteriormente o teste mais adequado para esse caso é um Teste Unilateral à Esquerda:

$H_0: \mu = 30$, onde $\mu_0 = 30$ (valor de teste); e

$H_1: \mu < 30$.

Nível de significância. O problema declara que é necessário usar 5%.

Variável de teste. Uma vez que a variância populacional da variável é **desconhecida** (o valor fornecido é o desvio padrão **amostral**), mas a amostra retirada apresenta 35 elementos (portanto mais de 30), a variável de teste será **Z** da distribuição normal.

Definir a região de aceitação de H_0 :

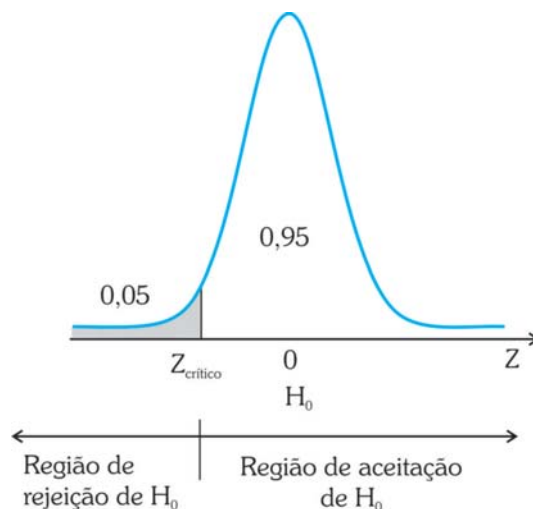


Figura 82: Regiões de aceitação e de rejeição – teste unilateral à esquerda

Fonte: Elaborada pelo autor deste livro

Cuidado com os casos de **fronteira**, em que o valor da variável de teste é muito próximo do valor crítico. Nesses casos, a rejeição ou aceitação de H_0 pode ocorrer por acaso. Sempre que apresentar o resultado recomende que uma nova amostra seja retirada para avaliar novamente o problema. Mas, nesse caso, rejeita-se H_0 com folga.

Observe que por ser um teste Unilateral à Esquerda, o Nível de Significância α está todo concentrado em um dos lados da distribuição, definindo a região de rejeição de H_0 . Para encontrar o valor crítico, devemos procurar na tabela da distribuição normal, pela probabilidade acumulada 0,95. Ou procurar a probabilidade complementar 0,05 e mudar o sinal do valor encontrado, pois o $Z_{\text{crítico}}$ aqui é menor do que zero. O valor crítico será igual a $-1,645$.

Através dos valores da amostra, avaliar o valor da variável.

Nesse ponto, é preciso encontrar o valor da variável de teste:

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

O valor de teste μ_0 é igual a 30, a média amostral $\bar{x}$ vale 29, o tamanho de amostra n é igual a 35 e o desvio padrão amostral s é 5. Substituindo na equação:

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{29 - 30}{5 / \sqrt{35}} = -1,183$$

Decidir pela aceitação ou rejeição de H_0 . Como se trata de um teste Unilateral à Direita:

Rejeitar H_0 se $Z < Z_{\text{crítico}}$. Como $Z = -1,185 > Z_{\text{crítico}} = -1,645$

Aceitar H_0 a 5% de Significância (há 5% de chance de erro).

Interpretar a decisão no contexto do problema. Não há provas estatísticas suficientes para concluir que o protocolo tem um desempenho satisfatório.

Testes de Hipóteses sobre a Proporção de uma Variável em uma População

Nesse caso, há interesse em testar a hipótese de que o parâmetro proporção populacional (π) de um dos valores de certa variável seja maior, menor ou diferente de determinado valor. Para a realização desse teste, tal como será descrito, é necessário que duas condições sejam satisfeitas:

- que o produto $n \times \pi_0$ seja maior ou igual a 5; e

- que o produto $n \times (1 - \pi_0)$ seja maior ou igual a 5.

Onde n é o tamanho da amostra e π_0 é a proporção sob teste (de um dos valores da variável). Se ambas as condições forem satisfeitas, a distribuição amostral da proporção que é binomial (uma Bernoulli repetida n vezes) pode ser aproximada por uma normal. Obviamente supõe-se que a amostra é representativa da população e foi retirada de forma aleatória, e que a variável pode assumir apenas dois valores, aquele no qual há interesse e o seu complementar.

Se as condições acima forem satisfeitas, a distribuição amostral da proporção poderá ser aproximada por uma normal com:

$$\text{Média} = \pi_0 \quad \text{Desvio Padrão} = \sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}$$

Lembrando da expressão da variável Z :

$$Z = \frac{\text{valor} - \text{“média”}}{\text{“desvio-padrão”}}$$

O **valor** será a proporção amostral (que é o melhor estimador da proporção populacional) do valor da variável denotada por p . A “média” e o “desvio padrão” são os valores definidos anteriormente, então a expressão de Z será:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}}$$

Compara-se o valor da variável de teste com o valor crítico ($Z_{\text{crítico}}$ que depende do Nível de Significância adotado) de acordo com o tipo de teste:

Se $H_1 : \mu > \mu_0$ Rejeitar H_0 se $Z > Z_{\text{crítico}}$ ($p > p_{\text{crítico}}$);

Se $H_1 : \mu < \mu_0$ Rejeitar H_0 se $Z < Z_{\text{crítico}}$ ($p < p_{\text{crítico}}$); e

Se $H_1 : \mu \neq \mu_0$ Rejeitar H_0 se $|Z| \neq |Z_{\text{crítico}}|$.

Cerca de 2.000 formulários de pedidos de compra estão sendo analisados. Os clientes podem ficar insatisfeitos se houver erros nos formulários. Nesse caso, admite-se que a proporção máxima de formulários com erros seja de 5%. Suponha que dentre os 2.000 formulários 7% apresentassem erros. A proporção máxima foi ultrapassada a 1% de significância? Vejamos o terceiro exemplo:

Nesse caso, $Z_{\text{crítico}}$ será negativo, já que a região de Rejeição de H_0 está à esquerda de zero.

Enunciar as hipóteses. Conforme visto anteriormente, o teste mais adequado para esse caso é o Teste Unilateral à Direita:

$H_0: \mu = 5\%$, onde $\mu_0 = 5\%$ (valor de teste); e

$H_1: \mu > 5\%$.

Nível de significância. O problema declara que é necessário usar 1% de significância (se não fosse especificado, outro valor poderia ser usado).

Variável de teste. Como se trata de um teste de proporção, é necessário verificar o valor dos produtos:

$n \times \pi_0 = 2000 \times 0,05 = 100$ e $n \times (1 - \pi_0) = 2000 \times 0,95 = 1900$. Como ambos são maiores que 5, é possível fazer uma aproximação pela normal, e a variável de teste será Z .

Definir a região de aceitação de H_0 :

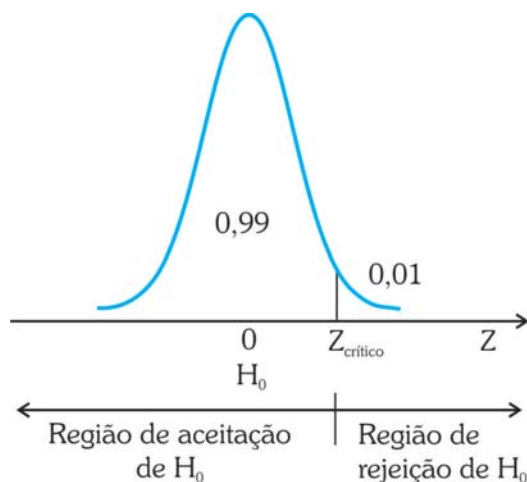


Figura 83: Regiões de aceitação e de rejeição – teste unilateral à direita
Fonte: Elaborada pelo autor deste livro

Observe que por ser um teste Unilateral à Direita, o Nível de Significância α está todo concentrado em um dos lados da distribuição, definindo a região de rejeição de H_0 . Para encontrar o valor crítico, devemos procurar na tabela da distribuição normal, pela probabilidade acumulada 0,01 (o $Z_{\text{crítico}}$ aqui é maior do que zero). O valor crítico será igual a 2,326.

Através dos valores da amostra, avaliar o valor da variável. Nesse ponto, é preciso encontrar o valor da variável de teste:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}}$$

O valor de teste π_0 é igual a 0,05 (5%), a proporção amostral p vale 0,07 (7%), e o tamanho de amostra n é igual a 2.000. Substituindo na equação:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}} = \frac{0,07 - 0,05}{\sqrt{\frac{0,05 \times (0,95)}{2000}}} = 4,104$$

Decidir pela aceitação ou rejeição de H_0 . Como se trata de um teste Unilateral à Direita:

Rejeitar H_0 se $Z > Z_{\text{crítico}}$ Como $Z = 4,104 > Z_{\text{crítico}} = 2,326$

Rejeitar H_0 a 1% de Significância (há 1% de chance de erro).

Interpretar a decisão no contexto do problema. Há provas estatísticas suficientes de que a proporção está acima do máximo admitido. Provavelmente, os vendedores/compradores precisarão passar por novo treinamento.

Este não é um caso de fronteira

Agora, vamos ver um tipo de teste estatístico muito utilizado pelos administradores para avaliar o relacionamento entre duas variáveis qualitativas: o teste de associação (independência de quiquadrado).

Teste de Associação de Quiquadrado

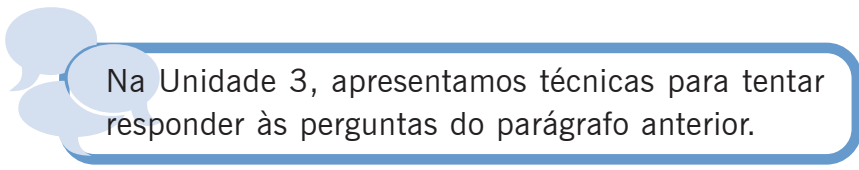
O teste do quiquadrado, também chamado de teste de independência de quiquadrado, está vinculado à análise de duas variáveis qualitativas. Vamos ver alguns conceitos antes de apresentar o teste de associação de quiquadrado.

Teste de associação (independência) de quiquadrado – teste que permite avaliar se duas variáveis qualitativas, cujas frequências estão dispostas em uma tabela de contingências, apresentam associação significativa ou não. Fonte: Barbetta, Reis e Bornia (2008).

Variáveis Qualitativas e Tabelas de Contingência

É comum haver interesse em saber se duas variáveis quaisquer estão relacionadas, e o quanto estão relacionadas, seja na vida prática, seja em trabalhos de pesquisa, por exemplo:

- se a satisfação com um produto está relacionada à faixa etária do consumidor; e
- se a função exercida por uma pessoa em uma organização está associada a seu gênero.



Na Unidade 3, apresentamos técnicas para tentar responder às perguntas do parágrafo anterior.

Variáveis Qualitativas são as variáveis cujas realizações são atributos, categorias (Unidades 1 e 3). Como exemplo de variáveis qualitativas temos: sexo de uma pessoa (duas categorias, masculino e feminino), grau de instrução (analfabeto, primeiro grau incompleto etc.), opinião sobre um assunto (favorável, desfavorável, indiferente).

Em estudos sobre variáveis qualitativas é extremamente comum registrar as frequências de ocorrência de cada valor que as variáveis podem assumir, e quando há duas variáveis envolvidas é comum registrar-se a frequência de ocorrência dos cruzamentos entre valores. Por exemplo: quantas pessoas do sexo masculino são favoráveis a determinada proposta de lei; quantas são desfavoráveis; quantas pessoas do sexo feminino são favoráveis. E, para facilitar a análise dos resultados, eles costumam ser dispostos em uma Tabela de Contingências. A Tabela de Contingências relaciona os possíveis valores de uma variável qualitativa com os possíveis valores da outra, registrando quantas ocorrências foram verificadas de cada cruzamento.

Exemplo 4 – O Quadro 23 mostra uma tabela de contingências relacionando as funções exercidas e o sexo de 474 funcionários de uma organização.

SEXO	FUNÇÃO			
	ESCRITÓRIO	SERVIÇOS GERAIS	GERÊNCIA	TOTAL
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Quadro 23: Tabela de contingências de função por sexo

Fonte: Elaborado pelo autor deste livro

Podemos apresentar os percentuais calculados em relação aos totais das colunas no Quadro 24:

SEXO	FUNÇÃO			
	ESCRITÓRIO	SERVIÇOS GERAIS	GERÊNCIA	TOTAL
Masculino	43,25%	100%	88,10%	54%
Feminino	56,75%	0%	11,90%	46%
Total	100%	100%	100%	100%

Quadro 24: Tabela de contingências de função por sexo:
percentuais por colunas

Fonte: Elaborado pelo autor deste livro

Seria interessante saber se as duas variáveis são estatisticamente dependentes, e o quão forte é essa associação. Repare que os percentuais de homens e mulheres em cada função são diferentes dos percentuais marginais (de homens e mulheres no total de funcionários) e, em duas funções, as diferenças são bem grandes.

O teste de associação de qui-quadrado é uma das ferramentas estatísticas mais utilizadas quando se deseja estudar o relacionamento entre duas variáveis qualitativas. Permite verificar se duas variáveis qualitativas são independentes, se as proporções de ocorrência dos valores das variáveis observadas estão de acordo com o que era esperado etc. Nesse texto, haverá interesse em usar o teste para avaliar se duas variáveis qualitativas são independentes.

Como todo teste de hipóteses, o teste de associação de qui-quadrado consiste em comparar os valores observados em uma amostra com os valores de uma referência (referência esta que supõe que a hipótese nula seja válida).

As frequências observadas da variável são representadas em uma tabela de contingências, e a Hipótese Nula (H_0) do teste será

Se isso não ocorrer recomenda-se agrupar as categorias (de uma ou outra variável, ou de ambas) até a obtenção de todas as frequências pelo menos iguais a 5.

aquela na qual as duas variáveis não se diferem em relação às frequências com que ocorre uma característica particular, ou seja, as variáveis são independentes; e tal hipótese será testada contra a Hipótese Alternativa (H_1) de que as variáveis não são independentes.

O teste pode ser realizado porque o grau de dependência pode ser quantificado descritivamente através de uma estatística, que se chama justamente quiquadrado (χ^2) na população, mas na amostra é chamada de q^2 , cuja expressão é:

$$q^2 = \sum_{i=1}^L \sum_{j=1}^C \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

Sendo $E_{ij} = \frac{\text{total da linha } i \times \text{total da coluna } j}{\text{total geral}}$

Onde:

- E_{ij} é a frequência esperada, sob a condição de independência entre as variáveis, em uma célula qualquer da tabela de contingências. Todas as frequências esperadas precisam ser maiores ou iguais a 5 para que o resultado do teste seja válido.
- O_{ij} é a frequência observada em uma célula qualquer da tabela de contingências.
- L é o número total de linhas da tabela de contingências (número de valores que uma das variáveis pode assumir).
- C é o número total de colunas da tabela (número de valores que a outra variável pode assumir).

Então, para cada célula da tabela de contingências, calcula-se a diferença entre a frequência observada e a esperada. Para evitar que as diferenças positivas anulem as negativas, as diferenças são elevadas ao quadrado. E para evitar que uma diferença grande em termos absolutos, mas pequena em termos relativos, “inflacione” a estatística, ou que uma diferença pequena em termos absolutos, mas grande em termos relativos, tenha sua influência reduzida, divide-se o quadrado da diferença pela frequência esperada. Somam-se os valores de todas as células e obtêm-se o valor da estatística: quanto maior q^2 , mais o Observado se afasta do Esperado, portanto maior a dependência.

Sob a hipótese de independência a estatística q^2 seguirá o modelo quiquadrado, que estudamos na Unidade 7, prometendo usá-la aqui na Unidade 10.

O Teste do quiquadrado para avaliar se duas variáveis são independentes será unilateral: a Hipótese Nula será rejeitada se $q^2 > q^2_{\text{crítico}}$, para certo número de graus de liberdade. Por exemplo, para o caso em que há três graus de liberdade e o Nível de Significância fosse de 5% (a região de Rejeição de H_0 ficará **à direita**), o valor crítico será:

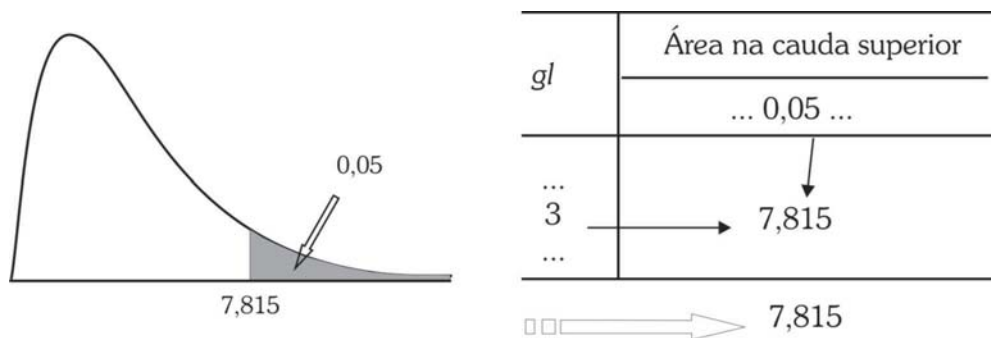


Figura 84: Uso da tabela da distribuição quiquadrado. Ilustração com $gl = 3$ e área na cauda superior de 5%

Fonte: Adaptada de Barbeto, Reis e Bornia (2008)

O número de graus de liberdade da estatística é calculado da seguinte forma:

$$\text{graus de liberdade} = (L - 1) \times (C - 1)$$

Sendo o número de linhas e o número de colunas referentes à tabela de contingências (são os números de valores que cada variável pode assumir).

O número de graus de liberdade assume esse valor porque para calcular as frequências esperadas não é necessário calcular os valores de todas as células, as últimas podem ser calculadas por diferença já que os totais são fixos. Por exemplo, para duas variáveis que somente podem assumir dois valores cada, o número de graus de liberdade seria igual a 1 $[(2-1) \times (2-1)]$: bastaria calcular a frequência esperada de uma célula e obter as outras por diferença em relação ao total.

Para o conjunto do Exemplo 4, supondo que os resultados são provenientes de uma amostra aleatória, vamos verificar se as variáveis são independentes a 1% de significância, neste quinto exemplo.

Enunciar as Hipóteses:

H_0 : as variáveis sexo e função são independentes; e

H_1 : as variáveis sexo e função não são independentes.

Nível de significância: determinado pelo problema, $\alpha = 0,01$; $1 - \alpha = 0,99$.

Retirar as amostras aleatórias e montar a tabela de contingências (isso já foi feito!). Veja o Quadro 25:

SEXO	FUNÇÃO			
	ESCRITÓRIO	SERVIÇOS GERAIS	GERÊNCIA	TOTAL
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Quadro 25: Tabela de contingências de função por sexo
Fonte: Elaborado pelo autor deste livro

No Quadro 25 encontramos os totais marginais e o total geral:

$$L_1 = \text{total Masculino} = 258 \quad L_2 = \text{total Feminino} = 216$$

$$C_1 = \text{total Escritório} = 157$$

$$C_2 = \text{total S.Gerais} = 27 \quad C_3 = \text{total gerência} = 84$$

$$N = \text{total geral} = 474$$

Repare que, somando os totais das linhas, o resultado é o total geral; e somando os totais das colunas, o resultado é o total geral também.

Calcular as frequências esperadas:

Calculando as frequências esperadas de acordo com a fórmula vista anteriormente:

$$\text{Masculino – Escritório} \quad E = (258 \times 363) / 474 = 197,58$$

$$\text{Masculino – Serviços Gerais} \quad E = (258 \times 27) / 474 = 14,70$$

$$\text{Masculino – Gerência} \quad E = (258 \times 84) / 474 = 45,72$$

$$\text{Feminino – Escritório} \quad E = (216 \times 363) / 474 = 165,42$$

$$\text{Feminino – Serviços Gerais} \quad E = (216 \times 27) / 474 = 12,30$$

$$\text{Feminino – Gerência} \quad E = (216 \times 84) / 474 = 38,28$$

Vamos calcular a estatística q^2 para cada célula, pois agora podemos avaliar as diferenças entre as frequências e as demais operações, que serão mostradas nos Quadros 26, 27 e 28.

O – E	FUNÇÃO		
SEXO	ESCRITÓRIO	SERVIÇOS GERAIS	GERÊNCIA
Masculino	157 – 197,58	27 – 14,70	74 – 45,72
Feminino	206 – 165,42	0 – 12,30	10 – 38,28

Quadro 26: Diferença entre frequências observadas e esperadas de função por sexo

Fonte: Elaborado pelo autor deste livro

(O – E) ²	FUNÇÃO		
SEXO	ESCRITÓRIO	SERVIÇOS GERAIS	GERÊNCIA
Masculino	1646,921	151,383	799,672
Feminino	1646,921	151,383	799,672

Quadro 27: Diferença entre frequências observadas e esperadas de função por sexo elevadas ao quadrado

Fonte: Elaborado pelo autor deste livro

Finalmente:

(O – E) ² /E	FUNÇÃO		
SEXO	ESCRITÓRIO	SERVIÇOS GERAIS	GERÊNCIA
Masculino	8,336	10,301	17,490
Feminino	9,956	12,304	20,891

Quadro 28: Estatísticas q^2 de função por sexo

Fonte: Elaborado pelo autor deste livro

Agora, podemos somar os valores:

$$q^2 = 8,336 + 10,301 + 17,490 + 9,956 + 12,304 + 20,891 = 79,227$$

Os graus de liberdade: **(número de linhas – 1) × (número de colunas – 1) = (2 – 1) × (3 – 1) = 2**

$$\text{Então } q^2_2 = 79,227$$

O q^2 crítico será: procurando na Tabela 3 do ambiente, ou em um programa, para 2 graus de liberdade e 99% de confiança (1% de significância): $q^2_{2,\text{crítico}} = 6,63$

Repare na Figura 85 a seguir.

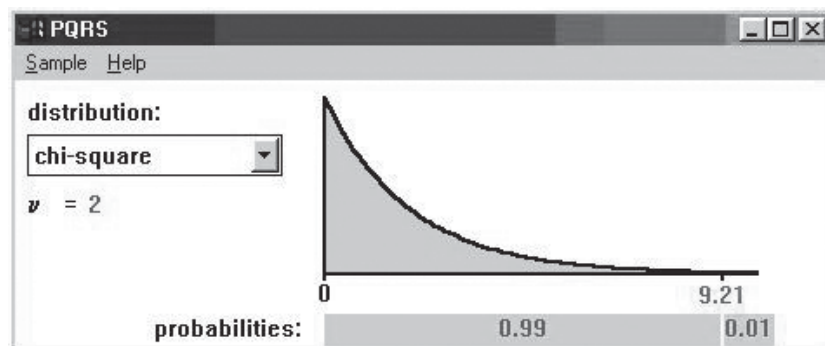


Figura 85: Valor crítico de q^2 para 2 graus de liberdade e 1% de significância
Fonte: Elaborada pelo autor deste livro

Como q^2_2 é maior do que $q^2_{2,\text{crítico}}$, rejeitamos H_0 a 1% de significância. Há evidência estatística suficiente que indica que as variáveis: função e sexo não são independentes. Isso confirma nossas suspeitas iniciais, devido às grandes diferenças nas frequências da tabela.

No tópico *Tô a fim de saber*, você terá indicações de vários outros tipos de hipóteses que não foram mencionados nesta Unidade. As referências lá citadas serão extremamente valiosas se você precisar:

- aplicar testes para avaliar se há diferenças entre médias de duas ou mais populações;
- aplicar testes para avaliar se há diferenças entre proporções de duas populações; e
- aplicar testes não paramétricos, por exemplo, testes de aderência dos dados a determinado modelo probabilístico.

Com este tópico, terminamos nossa jornada... Agora, é com você. Boa sorte!

Sobre tipos de erro em testes de hipóteses: BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulo 8; STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, Capítulo 10.

Sobre testes de uma variância: BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulo 8; TRIOLA, Mário. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999, Capítulo 7.

Sobre testes de comparação de duas médias: BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulo 9.

Sobre testes de comparação de duas proporções: MOORE, David S.; McCABE, George P.; DUCKWORTH, William M.; SCLOVE, Stanley L. *A prática da estatística empresarial: como usar dados para tomar decisões*. Rio de Janeiro: LTC, 2006, Capítulo 8.

Sobre Análise de Variância, comparação de várias médias: BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulo 9; STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, Capítulo 11; MOORE, David S.; McCABE, George P.; DUCKWORTH, William M.; SCLOVE, Stanley L. *A prática da estatística empresarial: como usar dados para tomar decisões*. Rio de Janeiro: LTC, 2006, Capítulos 14 e 15; TRIOLA, Mário. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999, Capítulo 11.

Sobre testes não paramétricos: BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulo 10; STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, Capítulo 13; SIEGEL, Stanley. *Estatística Não Paramétrica (para as Ciências do Comportamento)*. São Paulo: McGraw-Hill, 1975.

Sobre a utilização do Microsoft Excel para realizar testes de hipóteses: LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística: Teoria e Aplicações – Usando Microsoft Excel em Português*. 5. ed. Rio de Janeiro: LTC, 2000, Capítulo 6.



O resumo desta Unidade está demonstrado na Figura 86:

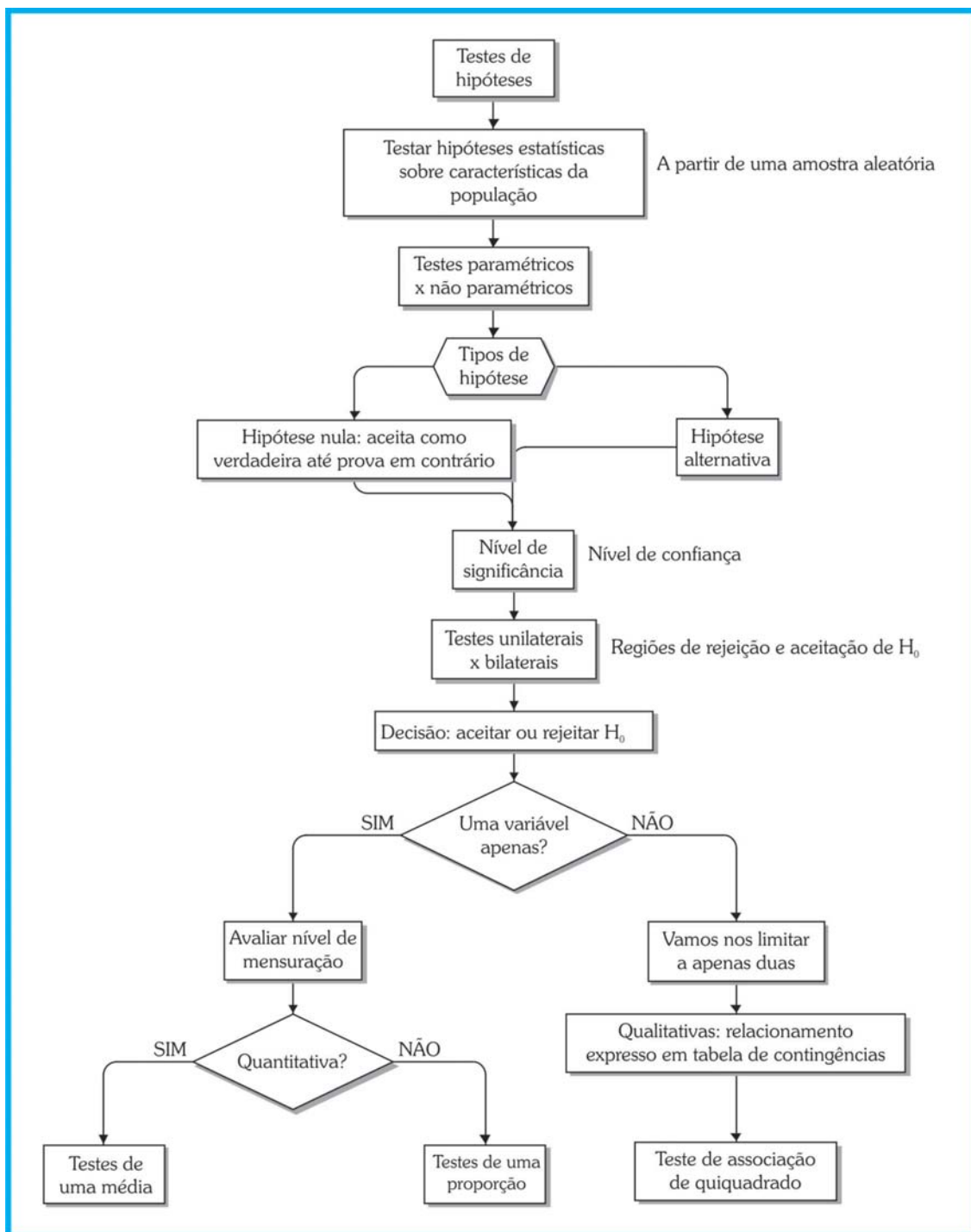


Figura 86: Resumo da Unidade 10
Fonte: Elaborada pelo autor deste livro

Chegamos ao final da disciplina de *Estatística Aplicada à Administração* e, por meio dela, estudamos os testes de hipóteses, tipos de hipóteses e suas variáveis. A Unidade foi explorada com Figuras, Quadros, Tabelas e muitos exemplos para melhor representar o conteúdo oferecido. Além do material produzido pelo professor, você tem em mãos uma riquíssima fonte de referências para aprender mais sobre o assunto. Explore os conhecimentos propostos. Não tenha esta Unidade como fim, mas como o começo de uma nova trajetória em sua vida acadêmica. Bons estudos!



Atividades de aprendizagem

1. O tempo médio de atendimento em uma agência lotérica está sendo analisado por técnicos. Uma amostra de 40 clientes foi sistematicamente monitorada em relação ao tempo que levavam para ser atendidos, obtendo-se as seguintes estatísticas: tempo médio de atendimento de 195 segundos e desvio padrão de 15 segundos.

Considerando-se que o tempo de utilização segue uma distribuição normal:

O dono da agência garante que o tempo médio de atendimento é de 3 minutos (se for maior ele se compromete a contratar mais um atendente). Com base nos dados da amostra, a afirmação do dono é verdadeira, ou ele deve contratar um novo atendente? Use um nível de significância de 1%?

2. Com o intento de melhorar a qualidade do serviço, uma empresa estuda o tempo de atraso na entrega dos pedidos recebidos. Supondo que esse tempo se encontra normalmente distribuído, e conhecendo o tempo de atraso dos últimos 20 pedidos, descritos a seguir (em dias), determine:

5 1 0 3 6 10 2 3 4 1 5 3 1 6 6 9 0 0 1 0

Um cliente enfurecido quer comprovar estatisticamente a hipótese (declarada pela empresa) de que o atraso médio será de no máximo 1 dia. Ele argumenta que deve ser maior, e quer uma confiança de 99%. O cliente tem razão na sua reclamação?

3. Uma máquina produz peças classificadas como boas ou defeituosas. Retirou-se uma amostra de 1.000 peças da produção, verificando-se que 35 eram defeituosas. O controle de qualidade para a linha de produção a fim de rearranjar os equipamentos envolvidos quando o percentual de defeituosos é superior a 3%.

Baseado nos dados amostrais, a linha de produção deve ser parada?

4. Queremos saber se há diferenças significativas entre três meios de comunicação, em termos de lembrança do consumidor da propaganda veiculada. O resultado de um estudo sobre propaganda mostrou:

CAPACIDADE DE LEMBRANÇA	MEIO DE COMUNICAÇÃO			
	REVISTA	TV	RÁDIO	TOTAL
Lembram da propaganda	25	93	7	125
Não lembram da propaganda	73	10	108	191
Total	98	103	115	316

- a) Usando 1% de significância é possível concluir que há associação entre a capacidade de lembrança e o meio de comunicação usado?
- b) Observando os resultados acima, qual meio de comunicação você recomendaria para veicular uma propaganda, para maximizar a capacidade de lembrança do público alvo? Por quê?

Adaptado de LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística: Teoria e Aplicações usando Microsoft Excel® em Português*. Rio de Janeiro: LTC, 2000.

Referências



ANDERSON, David R.; SWEENEY, Dennis J.; WILLIAMS, Thomas A. *Estatística Aplicada à Administração e Economia*. 2. ed. São Paulo: Thomson Learning, 2007.

ANDRADE, Dalton F.; OGLIARI, Paulo J. *Estatística para as Ciências Agrárias e Biológicas com noções de experimentação*. Florianópolis: Ed. da UFSC, 2007.

AULETE, Caldas. *Dicionário digital contemporâneo da língua portuguesa*. Rio de Janeiro: Lexikon editora digital, 2008. Disponível em: < www.aulete.com.br >. Acesso em: 9 abr. 2012.

BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008.

BARBETTA, Pedro A. *Estatística Aplicada às Ciências Sociais*. 7. ed. Florianópolis: Ed. da UFSC, 2007.

BUSSAB, Wilton O.; MORETTIN, Pedro A. *Estatística Básica*. 5. ed. São Paulo: Saraiva, 2003.

COSTA NETO, Paulo L. da O. *Estatística*. 2. ed. São Paulo: Edgard Blücher, 2002.

DIÁRIO Catarinense. *Indicadores Econômicos*, 17 de outubro de 2007.

HUFF, Daniel. *Como mentir com Estatística*. Rio de Janeiro: Ediouro, 1992.

LOPES, Paulo A. *Probabilidades e Estatística*. Rio de Janeiro: Reichmann e Affonso Editores, 1999.

MARCONI, Marina de Andrade, LAKATOS, Eva Maria. *Técnicas de Pesquisa*. 5. ed. São Paulo: Atlas, 2003.

MONTGOMERY, Douglas C. *Introdução ao Controle Estatístico da Qualidade*. 4. ed. Rio de Janeiro: LTC, 2004.

MOORE, David S.; McCABE, George P.; DUCKWORTH, William M.; SCLOVE, Stanley L. *A prática da estatística empresarial: como usar dados para tomar decisões*. Rio de Janeiro: LTC, 2006.

StatSoft. Inc. STATISTICA (data analysis software system), version 7, 2004. Disponível em: <www.statsoft.com>. Acesso em: 19 fev. 2009.

VIRGILITTO, Salvatore B. *Estatística Aplicada: técnicas básicas e avançadas para todas as áreas do conhecimento*. São Paulo: Alfa-Omega, 2003.

Marcelo Menezes Reis

Formado em Engenharia Elétrica pela Universidade Federal de Santa Catarina – UFSC, Bacharel em Administração de Empresas pela Universidade para o Desenvolvimento de Santa Catarina – UDESC, registro no CRA-SC 4049, Especialização em Seis Sigma (Beyond Six Sigma Certification Program) na University of South Florida – USF (EUA), Mestre em Engenharia Elétrica pela Universidade Federal de Santa Catarina, e Doutor em Engenharia de Produção pela Universidade Federal de Santa Catarina. Professor Adjunto, lotado no Departamento de Informática e Estatística da Universidade Federal de Santa Catarina, desde 1995. Ministra disciplinas de estatística em vários cursos de graduação e pós-graduação da Universidade, incluindo os de Administração.

