Estatística Aplicada à Administração

Marcelo Menezes Reis

R375e Reis, Marcelo Menezes

Estatística aplicada à administração / Marcelo Menezes Reis. – Florianópolis : Departamento de Ciências da Administração /UFSC, 2008.

300p.: il.

Inclui bibliografia

Curso de Graduação em Administração a Distância

1. Estatística. 2. Administração – Métodos estatísticos. 3. Amostragem (Estatística). 4. Probabilidades. 5. Variáveis aleatórias. 6. Testes de hipóteses estatísticas. 7. Educação a distância. I. Título.

CDU: 519.2:65

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

PRESIDENTE DA REPÚBLICA

Luiz Inácio Lula da Silva

MINISTRO DA EDUCAÇÃO

Fernando Haddad

SECRETÁRIO DE EDUCAÇÃO A DISTÂNCIA

Carlos Eduardo Bielschowsky

DIRETOR DO DEPARTAMENTO DE POLÍTICAS EM EDUCAÇÃO A DISTÂNCIA – DPEAD

Hélio Chaves Filho

SISTEMA UNIVERSIDADE ABERTA DO BRASIL

Celso Costa

UNIVERSIDADE FEDERAL DE SANTA CATARINA

REITOR

Lúcio José Botelho

VICE-REITOR

Ariovaldo Bolzan

PRÓ-REITOR DE ENSINO DE GRADUAÇÃO

Marcos Laffin

DIRETORA DE EDUCAÇÃO A DISTÂNCIA

Araci Hack Catapan

CENTRO SÓCIO-ECONÔMICO

DIRETOR

Maurício Fernandes Pereira

VICE-DIRETOR

Altair Borgert

DEPARTAMENTO DE CIÊNCIAS DA ADMINISTRAÇÃO

CHEFE DO DEPARTAMENTO

João Nilo Linhares

SUBCHEFE DO DEPARTAMENTO

Raimundo Nonato de Oliveira Lima

COORDENADOR DE CURSO

Alexandre Marino Costa

COMISSÃO DE PLANEJAMENTO, ORGANIZAÇÃO E FUNCIONAMENTO

Alexandre Marino Costa – Presidente

Gilberto de Oliveira Moritz

João Nilo Linhares

Luiz Salgado Klaes

Marcos Baptista Lopez Dalmau

Maurício Fernandes Pereira

Raimundo Nonato de Oliveira Lima

Sumário

UNIDADE 6 – Variáveis aleatórias

UNIDADE 7 – Modelos probabilísticos mais comuns

UNIDADE 8 – Inferência estatística e distribuição amostral

UNIDADE 9 – Estimação de parâmetros

UNIDADE 10 – Testes de hipóteses

Apresentação

Caro estudante!

Toda vez que alguém ouve a palavra “Estatística”, as reações costumam combinar aversão, medo, negação da importância, restrições ideológicas até, e sempre a noção 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.

Você está iniciando a disciplina de Estatística Aplicada à Administração. Os métodos estatísticos são ferramentas primordiais para o administrador de qualquer organização, pois possibilitam obter informações confiáveis, sem as quais a tomada de decisões seria mais difícil ou mesmo impossível. E, não se esqueça, a essência de administrar é tomar decisões. Por este motivo, esta disciplina faz parte do currículo do curso de Administração.

Nesta disciplina, você aprenderá como obter dados confiáveis (conceitos de planejamento de pesquisa estatística e amostragem), como resumi-los e organizá-los (análise exploratória de dados) e, aplicando técnicas apropriadas (probabilidade aplicada e inferência estatística), generalizar os resultados encontrados para tomar decisões. Procurei apresentar exemplos concretos de aplicação, usando ferramentas computacionais simples (como as planilhas eletrônicas, com as quais você teve um primeiro contato na disciplina de Informática Básica). O domínio dos métodos estatísticos dará a você um grande diferencial, pois permitirá tomar melhores decisões, o que, em essência, é o objetivo primordial de qualquer organização.

Sucesso em sua caminhada.

Prof. Marcelo Menezes Reis

UNIDADE

1

Introdução à Estatística e ao planejamento estatístico

Objetivo

Nesta Unidade, você vai identificar o conceito de Estatística, sua importância para o administrador e os principais aspectos do planejamento estatístico.

Definição e subdivisões da Estatística

Caro estudante, seja bem-vindo!

Convido-o a adentrar comigo nesse universo amplo, porém desafiador e instigante que é a discussão/reflexão sobre a **Estatística**. A partir da leitura do material, podemos juntos construir e socializar olhares articulando teoria e prática. Que rico esse movimento!

Bem, como você percebeu, o campo de debate é fértil e terá muito a discutir. Este será um espaço de socialização e construção do conhecimento. Não esqueça que dúvidas e indagações são sempre pertinentes, pois são delineadoras para o processo ao qual estamos nos dispondo coletivamente nesta disciplina.

Não é possível tomar decisões corretas sem dados confiáveis. Os governantes do Egito antigo e da Suméria (seus administradores) já sabiam disso, portanto, 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 a expansão do conhecimento, os métodos estatísticos tornaram-se mais sofisticados, com a adoção de modelos probabilísticos, inferência estatística e, nos últimos trinta anos, a aplicação 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.

Hoje em dia, todo administrador precisa usar métodos estatísticos. Para tanto, ele precisa conhecê-los, a começar por suas definições e subdivisões. Veremos isso nesta Unidade, além de apresentarmos os conceitos de planejamento estatístico: como obter dados confiáveis.

Conceito de Estatística

> "Estatística é a ciência que permite obter conclusões a partir de dados." (Paul Velleman)

Estatística é uma 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 quer verificar a qualidade de um lote inteiro de peças fornecidas através de uma pequena amostra;
- um político quer saber qual é o percentual de eleitores que votarão nele nas próximas eleições;
- um empresário deseja saber se há mercado potencial para abrir uma casa noturna em um determinado bairro da cidade; e
- em quais ações devo investir para obter maior rendimento?

GLOSSÁRIO

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

Variabilidade

O principal problema que surge ao se 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" e os gêmeos idênticos ("clones" naturais) somente apresentam um código genético comum; se forem submetidos a experiências de vida diferentes, terão um desenvolvimento distinto. Sendo assim, a variabilidade é **inevitável** e **inerente** à vida.

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

Você tem as mesmas preferências musicais que tinha há dez anos (muitos, sim, mas muitos, não)? Você tem a mesma aparência que tinha há dez anos? Você votaria no mesmo candidato a deputado federal em que votou na última eleição (caso você se lembre...)? Você tem o mesmo peso que tinha há dez anos? Imagine, então, as diferenças de pessoa para pessoa, de cidade para cidade, de povo para povo...

A Estatística permite descrever, identificar as fontes e mesmo 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 uma pergunta do mundo real. Para respondê-la, é preciso estudar uma ou mais características de uma **População** de interesse. População é o conjunto de medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m). Se, por exemplo, estamos avaliando as opiniões de eleitores sobre os candidatos a presidente, a população da pesquisa seria constituída pelas opiniões declaradas pelos eleitores em questão.

Maiores detalhes, você vai estudar ainda nesta Unidade.

GLOSSÁRIO

*Amostra – um subconjunto finito e representativo da população. Fonte: Barbetta (2006).

Como o interesse maior está na população, o ideal seria pesquisar toda a população, 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; torna-se, então, necessário pesquisar apenas uma **amostra***, um subconjunto finito e representativo da população.

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

Mais detalhes ainda nesta Unidade.

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

Tema da Unidade 2.

Uma das principais subdivisões da Estatística justamente 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.

Tema das Unidades 3 e 4.

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.

Estatística Indutiva, tema das Unidades 8, 9, e 10.

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

Tema das Unidades 5, 6 e 7.

A Figura 1 ilustra a subdivisão da Estatística. Veja:

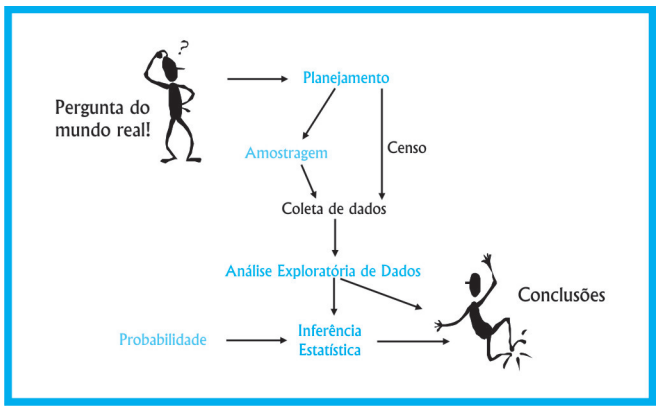


Figura 1: Subdivisões da Estatística
Fonte: elaborada pelo autor

GLOSSÁRIO

*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: elaborado pelo autor a partir de Lopes (1999).

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, precisa interpretá-los, levando em conta a variabilidade inerente e inevitável em todos os fenômenos. Como a Estatística fornece os meios para todas estas etapas, trata-se de um conhecimento indispensável para o administrador.

Não se esqueça: em qualquer profissão, é preciso analisar dados (verificando se sua fonte é confiável) e relacioná-los ao contexto no qual estão estão inseridos, e várias vezes compará-los com dados passados e fazer previsões sobre seu comportamento futuro. Veja o exemplo a seguir (Figura 2), extraído de um jornal de grande circulação.

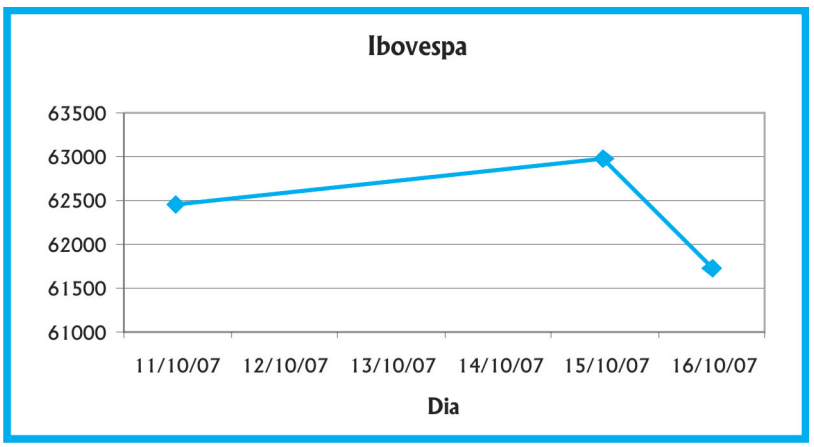


Figura 2: Variação do Ibovespa de 11/10/2007 a 16/10/2007
Fonte: Diário Catarinense de 17/10/2007

Através de um simples gráfico de linhas, podemos observar uma "queda" no índice da Bolsa de Valores de São Paulo, entre 15 e 16 de outubro de 2007, motivada pelo temor de queda nas bolsas internacionais, por sua vez causada pelo possível agravamento da crise imobili-

Na Unidade 3, você vai estudá-lo com mais detalhes.

ária nos EUA: constatação da situação atual, relacionada ao contexto internacional. Ao mesmo tempo em que se verifica queda, sabe-se que o Ibovespa nunca antes havia ultrapassado 60 mil pontos até recentemente: relacionamento com dados do passado. Os investidores em ações negociadas na Bovespa provavelmente tomarão suas novas decisões de compra e venda baseadas nestas informações.

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

GLOSSÁRIO

*Planejamento estatístico de pesquisa – conjunto de métodos cuja implementação visa a garantir a confiabilidade dos dados coletados. Fonte: Barbetta, Reis e Bornia (2004).

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

Antes de 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: isso nada tem a ver com Estatística, mas 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". Somente bacharéis em Administração (graduados em cursos de Administração) podem registrar-se no CRA. O CRA preocupa-se muito com a qualidade dos cursos de Administração, e freqüentemente apresenta sugestões para aperfeiçoar currículos e disciplinas, visando à melhoria da formação dos profissionais.

Estas e outras informações, você encontra em http://www.crasc.org.br/index.php?pg=inicial/oque.htm, acessado em 17/10/2007.

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 de Santa Catarina, 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 estes itens serão temas das próximas seções.

Objetivos da pesquisa

Como você já sabe, há dois tipos de objetivo: o geral e os específicos. A pesquisa pode ter apenas **um objetivo geral**. Este objetivo inclui o propósito que motivou a pesquisa, 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, no qual explicamos o que queremos medir (preferências, opiniões sobre fatos ou pessoas, resultados de experimentos, entre outros).

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 de Santa Catarina, 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 como um todo, o efeito seria benéfico, por contribuir para a formação de quadros 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 infra-estrutura dos seus cursos (salas, bibliotecas, laboratórios, ventilação, limpeza, iluminação); e
  - 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. E, através dos objetivos específicos, vamos chegar às variáveis, que você vai estudar mais à frente. O próximo passo é definir quem será pesquisado, ou seja, a população da pesquisa.

População

A definição de população foi vista no início desta Unidade, você se lembra?

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, estamos avaliando as opiniões de eleitores sobre os candidatos a presidente, a população da pesquisa seria constituída pelas opiniões declaradas pelos eleitores em questão. A população pode se referir a seres humanos, animais e mesmo objetos: alturas de pessoas adultas

do sexo masculino, peso de bois adultos, diâmetros dos parafusos produzidos em uma fábrica.

É muito importante também ter alguma noção do 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.

E, para o nosso exemplo, da pesquisa do CRA, qual seria a população?

- Conjunto das opiniões dos registrados no CRA de Santa Catarina, graduados no Estado, sobre os seus cursos.
- Tamanho da população: em 24/10/2007, havia 11.676 registrados no CRA de Santa Catarina. Vamos supor que 9.000 foram graduados em faculdades catarinenses.

Com estes aspectos definidos, podemos partir para a definição das variáveis, o que efetivamente será medido.

Variáveis

Quando um certo fenômeno é estudado, determinadas características são analisadas: as **variáveis**. É através das variáveis que se torna 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 um 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 altura, em metros, quando você tinha 12 anos? Naquele momento, a variável altura tinha apenas um valor possível. No ano seguinte, **em outro momento**, provavelmente a altura já era diferente, por sua vez, não deve ser a mesma que você tem hoje. Mas em cada momento, para você, ela 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). Veja na Figura 3, a seguir, a classificação das variáveis por nível de mensuração.

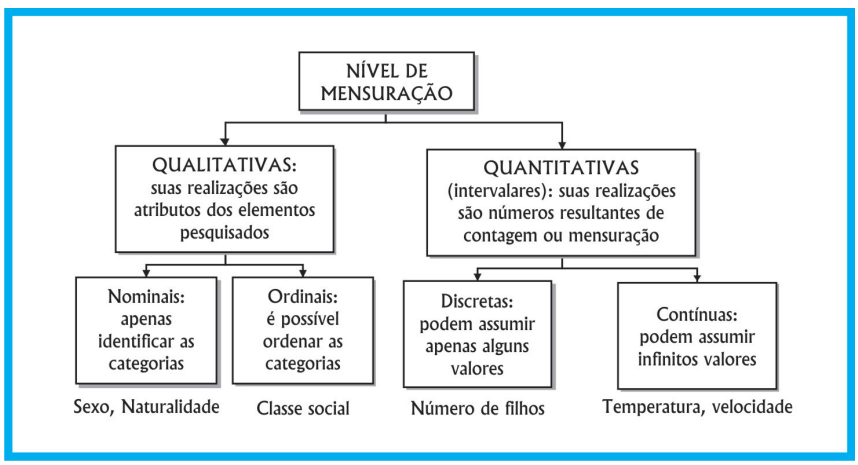


Figura 3: Classificação das variáveis por nível de mensuração
Fonte: elaborada pelo autor

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 em termos de quais itens pertencem a diferentes categorias, mas não se pode quantificar nem mesmo ordenar tais categorias. Por exemplo, pode-se dizer que dois indivíduos são diferentes em termos da variável A (sexo, por exemplo), mas não se pode dizer qual deles tem mais da qualidade representada pela variável. Exemplos típicos de variáveis nominais: sexo, naturalidade, entre outros; e
- as qualitativas **ordinais** permitem ordenar os itens medidos em termos de qual tem menos e qual tem mais da qualidade representada pela variável, mas ainda não permitem que se diga o quanto mais. Um exemplo típico de uma variável ordinal é o *status* socioeconômico das famílias residentes em uma localidade; sabe-se que média-alta é mais alta do que média, mas não se pode dizer, por exemplo, que é 18% mais alta.

Já 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); e
- as quantitativas **contínuas** são aquelas que podem assumir teoricamente qualquer valor em um intervalo (velocidade, peso).

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 idéia de quão mais pesado. 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 vai influenciar a qualidade dos resultados da pesquisa, **os custos.**

Veremos nas Unidades 3, 4, 8, 9 e 10 quais serão as técnicas estatísticas mais apropriadas para analisar os dados.

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

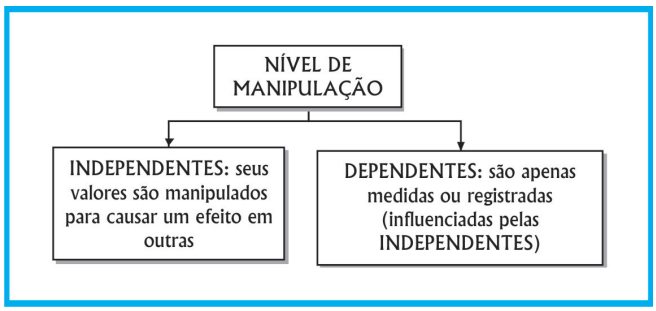


Figura 4: Classificação das variáveis por nível de manipulação
Fonte: elaborada pelo autor

Variáveis **independentes** são aquelas que são manipuladas, enquanto as **dependentes** são apenas medidas ou registradas, como resultado da manipulação das variáveis independentes. Esta distinção confunde muitas pessoas, que dizem que "todas as variáveis dependem de alguma coisa". Entretanto, uma vez que se esteja acostumado a esta distinção, ela se torna indispensável.

Os termos variável dependente e independente aplicam-se principalmente à pesquisa experimental, na qual algumas variáveis são manipuladas, e, neste 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. Ou seja, elas dependem do que as unidades experimentais farão em resposta.

Veremos mais detalhes nas próximas Unidades.

Contrariando um pouco 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 preexistentes dos próprios sujeitos.

Muitas vezes, fazemos a pesquisa para tentar identificar o relacionamento existente entre variáveis. Em uma pesquisa eleitoral para presidente do Brasil, por exemplo, 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. Neste caso, para o nosso exemplo da pesquisa com os registrados no CRA de Santa Catarina, as variáveis a serem medidas devem definir pelo menos uma variável para cada objetivo específico, conforme a seguir.

Para 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.

Veja que cada um destes quatro aspectos podem ser segmentados em outros para obter maiores detalhes. E então, como mensurá-los? Neste caso, devemos utilizar uma escala ordinal. Veja a pergunta:

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

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

Repare que, para cada acadêmico, em um dado momento, há apenas um resultado possível para a pergunta (ou assim limitamos no enunciado da questão). Poderíamos construir perguntas semelhantes para os outros três itens e para cada objetivo específico.

Passaremos agora à definição do delineamento da pesquisa, momento no qual 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 conduzida como **levantamento**, em que o pesquisador usualmente apenas registra os dados, através de um questionário ou qualquer outro instrumento de pesquisa. Procura-se responder às perguntas da pesquisa, através da identificação de associações entre as variáveis ou entre grupos de elementos da população, mas o pesquisador não tem controle sobre as variáveis. Por este motivo, para que os resultados sejam confiáveis, costuma ser necessário obter um grande conjunto de dados.

Nossa pesquisa com os registrados no Conselho Regional de Administração (CRA) de Santa Catarina poderia ser conduzida como um levantamento, através da aplicação de um questionário aos acadêmicos de Administração. Veja a Figura 5:

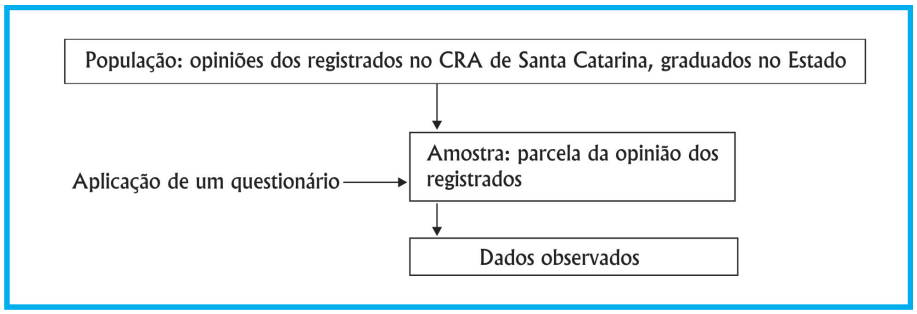


Figura 5: Pesquisa por levantamento
Fonte: adaptada pelo autor a partir de Barbetta (2006)

Quando há absoluta necessidade (e viabilidade) de provar relações de causa e efeito, o delineamento apropriado é o **experimento**. Neste tipo de delineamento, podemos manipular algumas variáveis para observar o efeito em outras, removendo (ou tentando remover) todas as outras variáveis que poderiam influenciar o resultado final: assim, se o experimento for adequadamente conduzido, será possível provar que a variação nos valores de uma ou mais variáveis causou as mudanças, entre outras. 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:

- pesquisas de opinião (eleitoral ou não), de mercado, de desemprego, de produção industrial, entre outras, são implementadas como levantamentos; e
- pesquisas na indústria farmacêutica (sobre eficácia e segurança de medicamentos), na indústria química (quais fatores vão propiciar um maior rendimento nas reações químicas), na indústria siderúrgica (qual é a composição necessária de uma liga de aço para obter a dureza especificada), entre outras, são conduzidas como experimentos.

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, e, portanto, precisamos ter um quadro muito completo da situação atual. É exatamente isso que o IBGE faz cada dez anos no Brasil com o censo demográfico. Mas há também os censos industrial, agropecuário, entre outros.

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

Nas pesquisas por amostragem, 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. A economia de tempo e dinheiro é evidente ao utilizar amostragem, bem como se torna obrigatório o seu uso em casos em que há a destruição ou exaustão dos elementos pesquisados, como em testes destrutivos: imagine o indivíduo que quer 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.

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

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. Sua 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 a Pesquisa Nacional por Amostragem de Domicílios (PNAD), não haveria 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 de Santa Catarina: 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.

GLOSSÁRIO

*Instrumento de pesquisa – dispositivo usado para coletar os valores das variáveis nos elementos da população. Fonte: Barbetta, Reis e Bornia (2004).

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

É através do **instrumento de pesquisa*** que coletamos os valores das variáveis, os dados da pesquisa. É importante ressaltar que ele está intrinsecamente relacionado às variáveis da pesquisa. Portanto, no seu projeto precisamos deixar claro qual é o relacionamento 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 dos tipos de instrumento de pesquisa, e em muitas situações ele não é o mais apropriado.

Imagine que queremos registrar o movimento em lojas de um shopping center, com a finalidade de saber quais apresentam clientela suficiente para continuarem a merecer a permanência. Não precisamos aplicar um questionário aos clientes, que podem se recusar a responder, ou aos lojistas, que podem ser “criativos demais” nas respos-

tas. Basta registrar em uma **planilha** quantas pessoas entraram 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, umidade relativa do ar, velocidade do vento: obviamente, não precisamos de um questionário para isso.

O questionário torna-se quase 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 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 é procurar obter as informações sem induzir ou confundir o respondente. As perguntas precisam ser claras, afirmativas ou interrogativas, evitando negações e coerentes com o nível intelectual dos elementos da população.

Em uma cidade de Santa Catarina, foi implementado um sistema integrado de transporte coletivo; foi feita uma pesquisa de opinião com os usuários, através de questionário; uma das questões perguntava se o usuário estava satisfeito com o itinerário dos ônibus; houve grande número de respostas em branco ou incoerentes com as outras perguntas; muitos respondentes não sabiam o que era itinerário.

Na nossa pesquisa, precisaríamos aplicar alguma espécie de questionário. O CRA dispõe de várias informações sobre os registrados, incluindo endereço postal, e talvez até telefone e endereço eletrônico. Poderíamos enviar os questionários por um destes três meios.

Saiba mais...

- Para saber mais sobre experimentos, consulte MOORE, D.S. et al. *A prática da Estatística empresarial*: como usar dados para tomar decisões. Rio de Janeiro: LTC, 2006, na seção 3.2.

- Para saber mais sobre elaboração de questionários, consulte BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006, capítulo 2.

RESUMO

O resumo desta Unidade está esquematizado na Figura 6.

Veja:

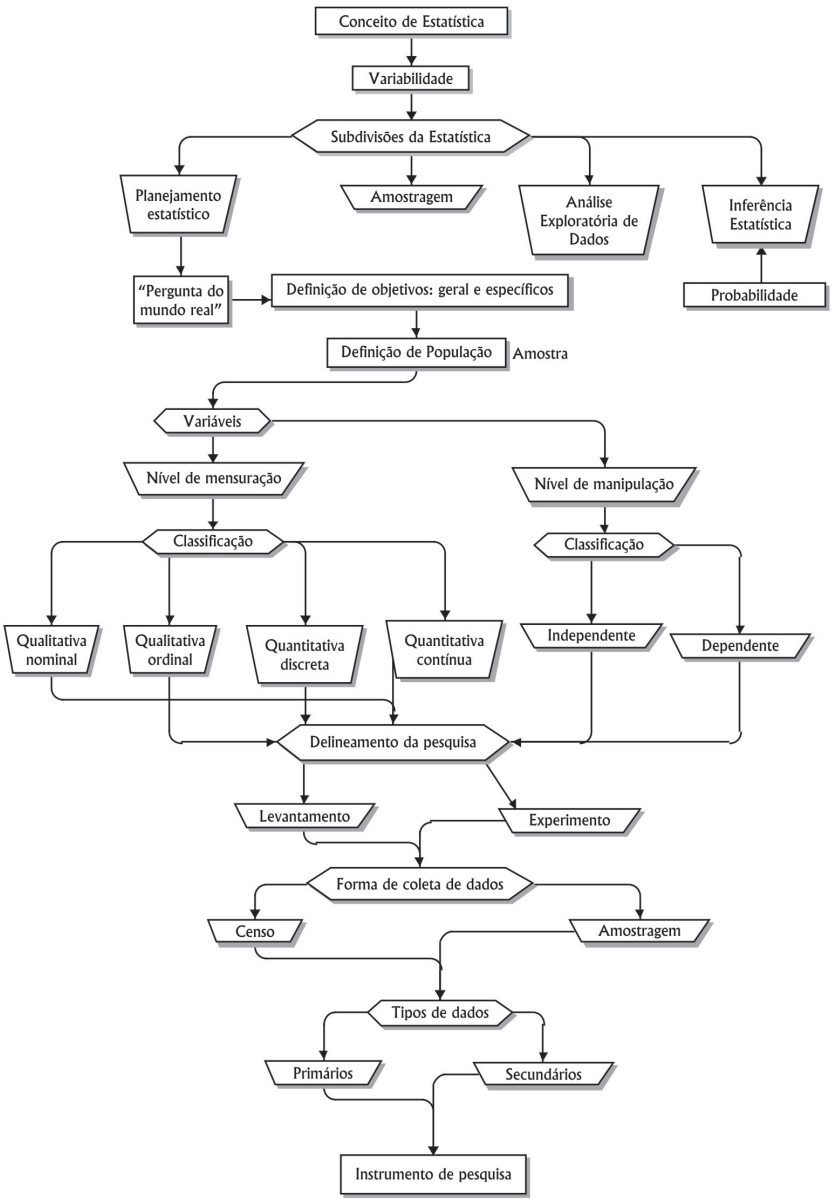


Figura 6: Resumo da Unidade 1
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Caro estudante!

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

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

Saiba que você não está sozinho neste processo, e que existe uma equipe que lhe dará base e suporte em todas as necessidades para a construção do seu conhecimento.

UNIDADE

2

Técnicas de amostragem

Objetivo

Nesta Unidade, você vai compreender em detalhes o que é amostragem, quando deve usá-la, as suas principais técnicas, a definição do plano de amostragem, e aprenderá a 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 observamos também que se trata de uma das subdivisões da Estatística, cujo conhecimento é indispensável para o administrador. Tenha em mente que estamos interessados em obter dados confiáveis para a tomada de decisões, e muitas vezes precisaremos realizar pesquisas para coletar tais dados. Convidamos você a conhecer um pouco mais sobre esta técnica de pesquisa e seus diferentes métodos de aplicação.

Há vários argumentos para justificar a utilização da amostragem, mas há casos em que seu uso pode não ser a melhor opção. O administrador precisa conhecer tais argumentos, para que, confrontando com os recursos disponíveis e os objetivos da pesquisa, possa tomar a melhor decisão sobre a forma de coleta dos dados.

Se o administrador decidir por amostragem, é preciso delinear o plano de amostragem, indicando como ela será implementada e qual será o seu tamanho, item crucial e que vai influenciar muito nos custos da pesquisa. Vamos ver isso em detalhes nesta Unidade.

O que é amostragem?

Amostragem é a subdivisão 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 (veja na Unidade 1) já

GLOSSÁRIO

*Censo – forma de coleta de dados em que a pesquisa é realizada com todos os elementos da população. Fonte: Barbetta, Reis e Bornia (2004).

*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 (2006).

*População – é o conjunto de medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m). Fonte: Andrade e Ogliari (2007).

*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.

foram cumpridas e que o administrador agora precisa decidir se coletará os dados por **censo*** ou por **amostragem***.

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 deste processo é mostrada na Figura 7:

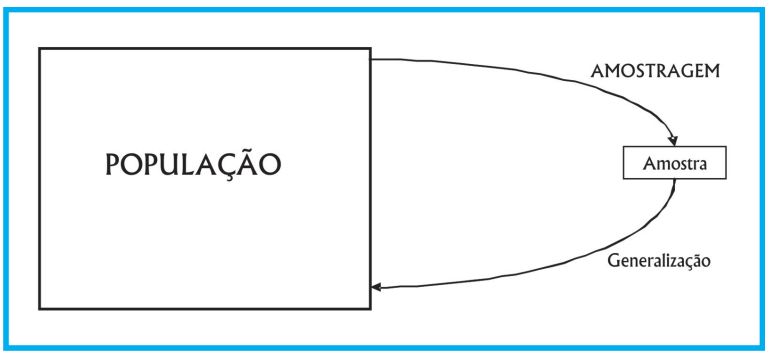


Figura 7: Processo de Amostragem e Generalização
Fonte: elaborada pelo autor

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

Quando devemos usar amostragem

Podemos enumerar, basicamente, três motivos para usar amostragem em uma pesquisa: economia, rapidez de processamento e quando há a necessidade de **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 cada dez anos.
- **Rapidez de processamento**: como a quantidade de dados coletada é muito menor do que a produzida em um censo, especialmente para grandes populações, o seu processamento

é mais rápido. Os resultados ficam disponíveis em pouco tempo, permitindo tomar decisões em seguida. 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, exigindo 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:

Economia Rapidez de processamento Testes destrutivos

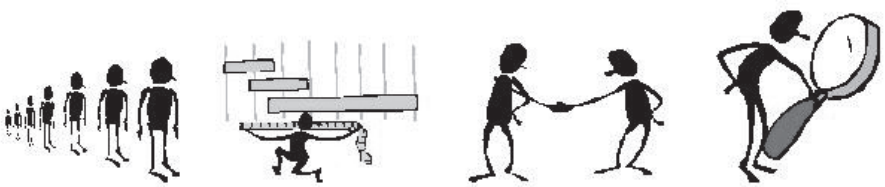
Figura 8: Os três motivos para usar amostragem em uma pesquisa
Fonte: Microsoft Office (2007)

Após reconhecer os motivos de utilizar a amostragem, pense em algumas situações em que seria recomendável utilizar esta técnica.

Quando NÃO devemos usar amostragem

Existem situações em que a utilização de amostragem pode não ser a melhor opção. Neste caso, podemos enumerar basicamente quatro motivos: população pequena, característica de fácil mensuração, necessidades políticas e necessidade de alta precisão.

- **População pequena**: quando é utilizada uma amostra probabilística (aleatória) e a população é pequena (digamos, menos de cem elementos), o tamanho mínimo de amostra para obter bons resultados será quase igual ao próprio tamanho da população (veremos isso mais adiante, ainda nesta Unidade). Vale a pena, então, 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. Neste caso, a coleta dos dados seria bastante simples, possibilitando avaliar todos os elementos da população. Outro caso freqüente na indústria são os 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 vai 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 cada dez anos? Porque as informações demográficas têm que ser precisas, para orientar políticas governamentais, e somente dessa maneira esse objetivo pode ser atingido. A Figura 9 sintetiza os motivos:



População pequena/Fácil mensuração/Necessidades políticas/Alta precisão

Figura 9: Situações em que a utilização de amostragem pode não ser a melhor opção
Fonte: Microsoft Office (2007)

Exercite a mente! Pense em algumas situações nas quais seja aconselhável usar um censo. Você deve se lembrar da pesquisa que esboçamos na Unidade 1: "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". Vimos que o número de registrados no CRA, com graduação em Santa Catarina, foi suposto igual a 9.000. Além disso, há uma listagem com os registrados, para fins de cobrança de anuidade inclusive, 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). Com base no que foi dito até agora, você sabe responder se a pesquisa deve ser conduzida por censo ou por amostragem? Vamos ver juntos, então!

GLOSSÁRIO

*Amostra representativa – é aquela que representa na sua composição todas as subdivisões da população, procurando retratar da melhor maneira possível a sua variabilidade. Fonte: elaborado pelo autor.

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: ela precisa ser uma "cópia reduzida" da população. 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: devemos 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.

Vamos aprender ainda nesta Unidade uma fórmula simplificada para o cálculo do tamanho de amostra, e na Unidade 9 veremos uma expressão mais completa. Em ambos os casos, porém, veremos que o tamanho de amostra também dependerá da precisão que queremos para o nosso resultado.

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 homogênea 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, permitindo a atribuição de números a cada um deles, e faz-se o sorteio. Idealmente, nós escreveríamos os números dos elementos da população em pequenos papéis, depositaríamos em uma urna, misturaríamos os papéis, e, de olhos vendados, escolheríamos os números, selecionando a amostra. Para grandes populações, esse procedimento é inviável, e com a disponibilidade de recursos computacionais, contraproducente.

O sorteio pode ser realizado através de **tabelas de números aleatórios** ou **algoritmos de geração de números pseudo-aleatórios***. (Algoritmos de geração de números pseudo-aleatórios são programas computacionais que geram números aleatórios (pseudo-aleatórios, pois têm uma regra de formação), procurando simular os sorteios manuais de números de 0 a 9 e garantir que todo número com a mesma quantidade de algarismos tenha a mesma probabilidade de ocorrência.)

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 realmente 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 pseudo-aleató-

GLOSSÁRIO

*Amostra suficiente – é aquela que tem um tamanho tal que permite representar adequadamente a variabilidade da população. Fonte: elaborado pelo autor.

*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 (2004).

rios, 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, caso o algoritmo seja bom, somente ocorre após milhões ou bilhões de sorteios, quantidade muitíssimo superior àquela usada nas nossas pesquisas. Alguns estatísticos construíram tabelas de números pseudo-aleatórios e as deixaram disponíveis para o público em geral.

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

Nos dias de hoje, com todas as facilidades da informática, são cada vez mais comuns bases de dados armazenadas em meio digital, desde uma simples planilha do Microsoft Excel até grandes bancos de dados.

Então, pergunta-se: 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 pseudo-aleatórios?

Trata-se de programas computacionais que procuram simular os sorteios reais de números. As grandes vantagens do seu uso são a possibilidade de adaptar facilmente o sorteio ao tamanho da população envolvida e, obviamente, a velocidade de processamento. Veja um exemplo de números aleatórios de quatro dígitos (de 0001 a 9000) gerados pelo Microsoft Excel®:

Na seção "Para saber mais", vamos disponibilizar um link que explica como gerar números pseudo-aleatórios com este aplicativo.

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

No caso da nossa pesquisa para o CRA de Santa Catarina, em que temos 9.000 registrados graduados em Santa Catarina e há uma listagem da população, pense: como seria o sorteio?

No caso mais simples de amostragem aleatória, o registrado de número 3.439 seria sorteado, seguido pelo 907 e pelo 5.369, e assim por diante, até completar o tamanho de amostra. Usualmente, cria-se automaticamente uma nova base de dados com os elementos sorteados.

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

Toda a teoria de inferência estatística pressupõe que a amostra, a partir da qual será feita a generalização (veja a Figura 1 desta Unidade), foi retirada de forma aleatória.

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 unidades selecionadas para 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 de Santa Catarina, 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 se chega 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, vemos o resumo dos diversos tipos de amostragem:

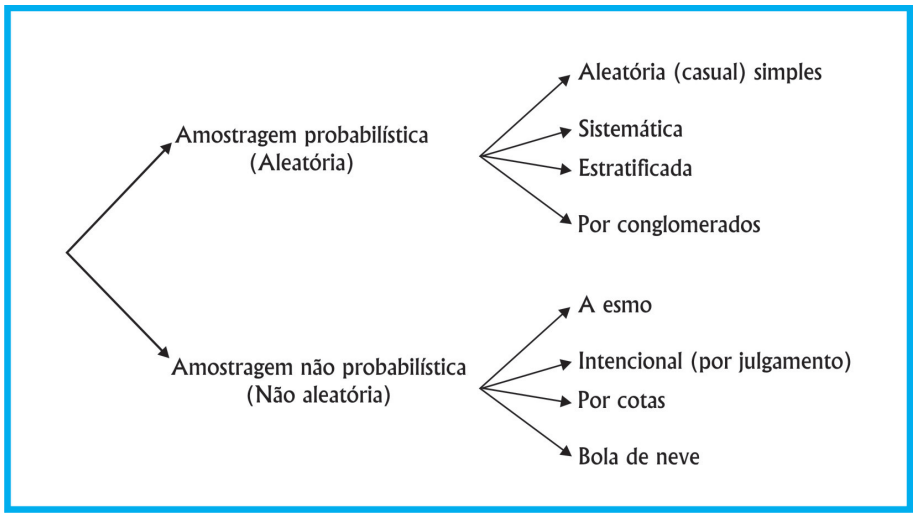


Figura 10: Tipos de Amostragem
Fonte: elaborada pelo autor

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. 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: "mas, afinal, o que significa 'estatisticamente'?". Significa que podemos associar aos resultados uma 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 uso da amostragem probabilística é que todos os elementos da população tenham uma probabilidade

maior do que zero de pertencer à amostra. Tal condição é materializada se:

- 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, que dispõe de uma relação com os 9.000 registrados que se graduaram em Santa Catarina; e
- os elementos da amostra são selecionados através de alguma forma de sorteio não viciado: tabelas de números aleatórios, números pseudo-aleatórios 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 têm probabilidade de pertencer à amostra.

GLOSSÁRIO

* 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 (2004).

* 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 (2006).

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 se admite 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 pseudo-aleatórios gerados por computador, os integrantes da amostra são selecionados. Neste tipo de amostragem probabilística, todos os elementos da população têm a mesma probabilidade de pertencer à amostra. Foi exatamente o que fizemos no final do tema "Aspectos necessários para o sucesso da amostragem" para a nossa pesquisa do CRA.

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 pseudo-aleatórios, 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 queremos obter uma amostra de idades de uma listagem justamente ordenada desta forma, neste 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;
- encontra-se o intervalo de retirada k = N/n:
  - se k for fracionário, deve-se aumentar n até tornar o resultado inteiro; e
  - 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 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 deste processo é retratado na Figura 11, veja:

GLOSSÁRIO

*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 (2006).

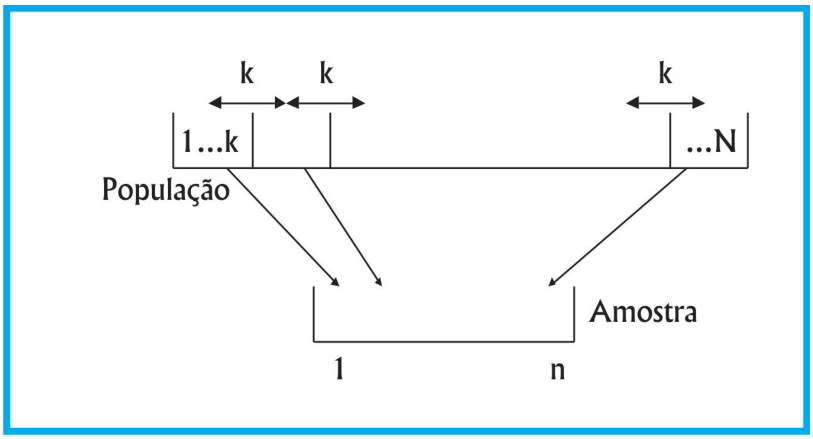


Figura 11: Processo de amostragem sistemática
Fonte: elaborada pelo autor

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.

A operadora dispõe de uma lista ordenada alfabeticamente com todos os seus assinantes. O intervalo de retirada será:

k = N/n = 25.037/800 = 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 = 24.800/800 = 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 que a população de uma pesquisa seja heterogênea em relação aos objetivos da pesquisa. No caso de uma pesquisa eleitoral para governador, por exemplo, podemos esperar que a opinião deva ser diferente dependendo da região onde o eleitor mora, classe social e mesmo profissão dos entrevistados. Contudo, podemos supor que haja certa homogeneidade de opinião dentro de cada grupo. Então, supõe-se que haja heterogeneidade entre os estratos, mas homogeneidade dentro dos estratos, 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-se usar a **amostragem estratificada***, como representa a Figura 12:

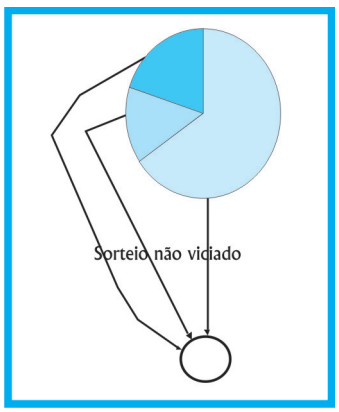


Figura 12: Amostragem estratificada
Fonte: elaborada pelo autor

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

GLOSSÁRIO

*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.

* 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 (2006).

- uniforme, quando os mesmos números de elementos são selecionados de cada estrato.

A amostragem estratificada proporcional possibilita resultados melhores, mas exige um grande 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, você acredita que a população é heterogênea em relação aos objetivos da pesquisa? Será que a região do Estado, o fato de ter estudado em faculdade pública ou particular pode influenciar as opiniões dos registrados sobre os cursos nos quais 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 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 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);
- pesquisam-se todos os elementos dos conglomerados sorteados ou sorteiam-se elementos deles.

GLOSSÁRIO

*Amostragem por conglomerados – é a amostragem probabilística em que a população é subdividida em grupos definidos por conveniência (usualmente geográfica), e alguns destes grupos são selecionados por sorteio, e elementos dos grupos sorteados podem também ser sorteados para compor a amostra. Fonte: Barbetta (2006).

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

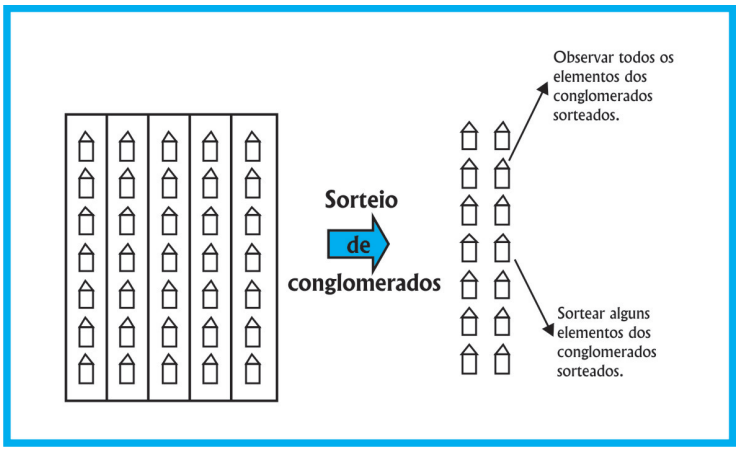


Figura 13: Amostragem por conglomerados
Fonte: elaborada pelo autor

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.

Mais informações em http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad98/saude/metodologia.shtm

Você deve estar se perguntando: "e quando não for possível garantir a probabilidade de todo elemento da população pertencer à amostra?" Então, este 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. Então, pode-se recorrer à **amostragem não probabilística***.

Ao usar a amostragem não probabilística, o pesquisador não sabe qual é a probabilidade de que um elemento da população tem 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 nos quais a população de trabalho não pode ser enumerada.

Veja que existem ainda vários tipos de amostragem não probabilística e que serão descritos na seqüência.

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 cem. Se fôssemos realizar uma amostragem aleatória simples, o processo talvez fosse trabalhoso demais. Então, simplesmente retiramos os elementos a esmo. Este 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.

GLOSSÁRIO

*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 (2006).

*Erro amostral – é o valor máximo que o pesquisador admite errar na estimativa de uma característica da população a partir de uma amostra aleatória desta mesma população. Fonte: Barbetta (2006).

Amostragem por julgamento (intencional)

Na amostragem por julgamento, o pesquisador deliberadamente escolhe alguns elementos para fazer parte da amostra, com base no seu julgamento de que aqueles seriam representativos da população. Este tipo de amostragem é bastante usado em estudos qualitativos. Obviamente, o risco de 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 parece 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 dividir em um grande número para compensar a falta de aleatoriedade, e seleciona-se uma cota de cada subgrupo, proporcional ao seu tamanho.

Em uma 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 que qualquer um que não esteja passando pelo local no exato momento da pesquisa possa ser selecionado).

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

Leia um texto muito interessante sobre o tema, que se encontra disponível em: http://www.ime.unicamp.br/~dias/falaciaPesquisaEleitoral.pdf

No exemplo apresentado no Quadro 1, imagine que queremos 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:

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 pelo autor de Lakatos e Marconi (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, em pontos de movimento do referido bairro, que se aproximassem dos critérios e entrevistá-las, recolhendo suas opiniões sobre o governo municipal. Usualmente, os resultados são generalizados estatisticamente para a população, empregando as técnicas que serão vistas na Unidade 9 deste livro-texto, mas rigorosa-

mente 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 que é identificado e entrevistado, pede-se que identifique outros que possam ser qualificados como respondentes. Pode levar a amostras compostas apenas por "amigos" dos primeiros entrevistados, o que pode causar distorções 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.

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 obter o 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);

- estamos fazendo uma estimativa exagerada do tamanho de amostra;
- não vamos nos preocupar com 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 $\mathbf{E_0}$. Este erro é o valor máximo que o pesquisador admite ter 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% ± 2%. Além disso, há uma probabilidade de que este intervalo não contenha o valor real do parâmetro, ou seja, o percentual de eleitores que declaram o voto no candidato, pelo fato de que estamos usando uma amostra, embora isso raramente seja dito na mídia, especialmente na televisão.

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:

$$n = \frac{N \times n_0}{N + n_0}$$

Pense, com esse exemplo, 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; e

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 a primeira estimativa, n0 não depende do tamanho da população, e o erro amostral é 2% para ambas, podemos calculá-lo apenas uma vez. Devemos dividir o 2% por 100 antes de substituir 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. Então, a amostra deverá ter pelo menos 186 elementos para garantir um erro amostral de 2%. Observe que a amostra representa 93% da população. Será que um censo não seria mais aconselhável neste caso?

- 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 grande mudança. Este exemplo prova que não precisamos de grandes amostras para obter uma boa precisão nos resultados.

A Figura 14 mostra um gráfico relacionando tamanhos de amostra para diferentes tamanhos de população, considerando um erro amostral tolerável igual a 2%.

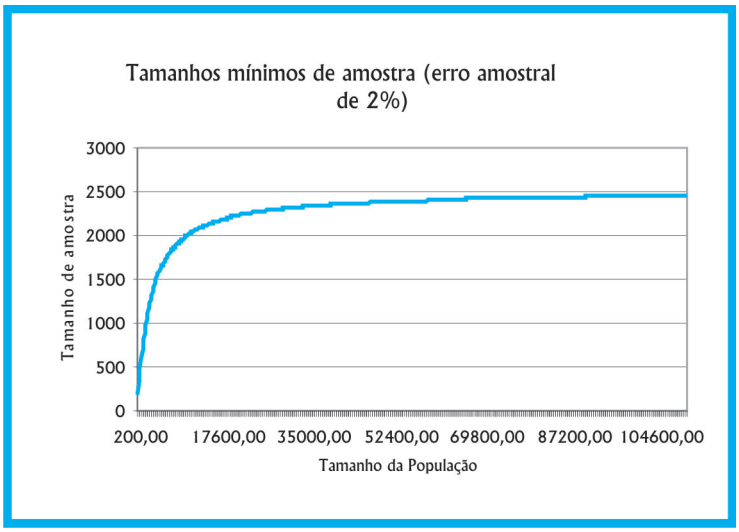


Figura 14: Tamanho de amostra x tamanho da população (e0 = 2%)
Fonte: elaborada pelo autor a partir de Microsoft

Observe na Figura 14 que, a partir de um determinado tamanho de população, para o mesmo erro amostral, o ritmo de crescimento do tamanho da amostra vai diminuindo; 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 ter uma boa amostra.

É importante alertar que, ao calcular o tamanho de amostra para amostragem estratificada, deve-se 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, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006, capítulo 3.

- Sobre características de fácil mensuração, consulte em LAGO NETO, J.C. *O efeito da autocorrelação em gráficos de controle para variável contínua*: um estudo de caso. 1999. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC. Florianópolis.

- Sobre pesquisas eleitorais, consulte SOUZA, J. *Pesquisas eleitorais*: críticas e técnicas. Brasília: Centro Gráfico do Senado Federal, 1990.

- Sobre como gerar números pseudo-aleatórios ou obter amostras aleatórias simples no Microsoft Excel , leia o texto "*Como gerar uma amostra aleatória simples com o Microsoft Excel®*", no Ambiente Virtual de Ensino-Aprendizagem.

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

RESUMO

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

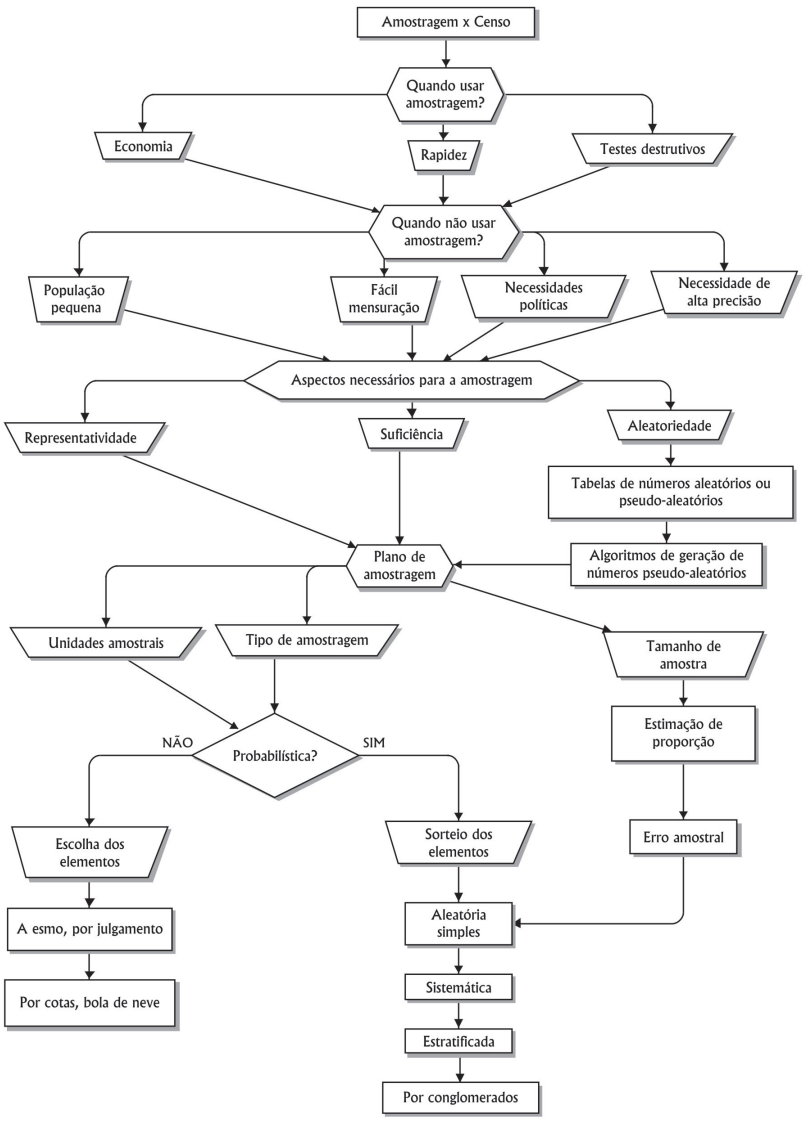


Figura 15: Resumo da Unidade 2
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Caro estudante!

Chegamos ao final da Unidade 2. Nela estudamos amostragem e censo, e suas formas de utilização, habilidades necessárias para um bom administrador. Esta Unidade foi repleta de figuras, quadros, representações, e exemplos de utilização das técnicas e das diferentes formas de utilização, na íntegra de suas especificidades, e deu sustentação para as discussões das próximas unidades. Releia, caso necessário, todos os exemplos, leia as indicações do Saiba mais e discuta com seus colegas. Na realização das atividades de aprendizagem, você colocará em prática os ensinamentos repassados. Conte sempre com o acompanhamento da tutoria e das explicações do professor. Lembre-se que não está sozinho. Conte com a gente!

UNIDADE

3

Análise Exploratória de Dados I

Objetivo

Nesta Unidade, você vai compreender a definição de Análise Exploratória de Dados e aprenderá como realizar a descrição tabular e gráfica de conjuntos de dados referentes a variáveis qualitativas e quantitativas.

O que é 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 da pesquisa. O passo inicial para isso é usar os conceitos e técnicas da 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.

O primeiro passo da Análise Exploratória de Dados é organizar os dados, para que seja possível resumi-los e, posteriormente, interpretá-los. 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.

Ainda nesta Unidade, vamos estudar como realizar a análise exploratória de dados através de tabelas e gráficos para cinco casos, 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 Análise Exploratória de Dados: sem isso, a sua capacidade de tomada de decisões ficará seriamente comprometida.

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

No passado, a Análise Exploratória de Dados era chamada de Estatística Descritiva, por preocupar-se com a descrição dos dados tão-somente.

GLOSSÁRIO

*Variáveis estatísticas – são características que podem ser observadas ou medidas em cada elemento pesquisado, sob as mesmas condições. Para cada variável, para cada elemento pesquisado, em um dado momento, há um e apenas um resultado possível. Fonte: Barbetta (2006).

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 mais **variáveis estatísticas***, ou do seu relacionamento.

O processo da Análise Exploratória de Dados consiste em organizar, resumir e interpretar as medidas das variáveis da melhor maneira possível. Para tanto, 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 para possibilitar sua análise. Atualmente, os dados costumam ser armazenados em meio computacional, seja em grandes bases de dados, programas estatísticos ou mesmo planilhas eletrônicas, sejam oriundos de pesquisa de campo ou apenas registros de operações financeiras, 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 nome do aluno, data de nascimento, número de matrícula, Índice de Aproveitamento Acumulado (IAA), Índice de Aproveitamento Corrigido (IAP) e outras informações, ou uma operadora de cartão de crédito, que armazena as transações efetuadas, contendo o número do cartão, nome do titular, hora da transação, valor do crédito, bem ou serviço adquirido.

Os casos constituem cada indivíduo ou registro. Para a base do DAE, João Ninguém nasceu em 20 de fevereiro de 1985, matrícula

02xxxxxxx-01, IAA = 3,5, IAP = 6,0. Para a operadora de cartão de crédito, cartão número xxxxxxxxx-84, José Nenhum, R$ 200, 14h28min – 11 de dezembro de 2003, supermercado.

Exemplo 1: a Megamontadora Toyord regularmente conduz pesquisas de mercado com os clientes que compraram carros zero km diretamente de suas concessionárias. O objetivo é avaliar a satisfação dos clientes em relação aos diferentes modelos, seu design, adequação ao perfil do cliente. A última pesquisa foi terminada em julho de 2007: 250 clientes foram entrevistados entre o 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 que já estão no mercado há dez anos. As seguintes variáveis foram obtidas:

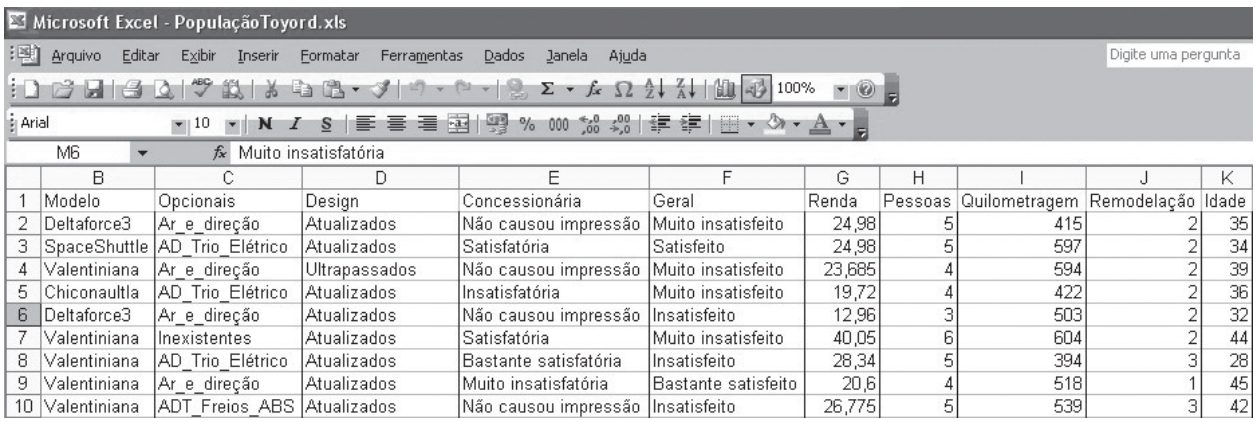
Trata-se de uma empresa fictícia e de uma pesquisa fictícia.

- **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 imprevistos)**: 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** de há quantos anos o veículo comprado teve a sua última remodelação de design: em anos completos (se há menos de um ano o entrevistador anotou zero); e
- **idade do cliente** em anos completos.

Imagine que você é *trainee* da Toyord. Sua missão é analisar os resultados da pesquisa apresentando um relatório. Dependendo do seu desempenho, você poderá ser contratado em definitivo ou dispensado (sem carta de recomendação). Como deve ser estruturada a base de dados para permitir a análise?

Digamos que você dispõe dos 250 questionários que foram aplicados e você vai tabulá-los em uma planilha eletrônica, como o Microsoft Excel®. Há dez variáveis, a base de dados deve ter, então, dez colunas e 250 linhas (no Excel , 251, já que a primeira será 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 pelo autor de Microsoft

Veja que cada uma das variáveis é registrada em uma coluna específica e que nas linhas se encontram os registros de cada funcionário. Por exemplo, o respondente 1 adquiriu um modelo DeltaForce3, com os opcionais ar-condicionado e direção hidráulica, considera o

design do veículo atualizado, diz que o atendimento da concessionária onde comprou o veículo não causou impressão, 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 com este veículo, crê 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. Juntamente com ele, está disponibilizado o texto "Como realizar análise exploratória de dados no Microsoft Excel".

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

A grande maioria dos programas estatísticos, gerenciadores de bases de dados e 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 se quiser, respeitando, porém, as capacidades dos programas. O Microsoft Excel®, por exemplo, admite 65.000 linhas, o que é suficiente para muitas aplicações.

Uma vez 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 freqüências, como veremos a seguir.

GLOSSÁRIO

*Distribuições de freqüências – organizações dos dados de acordo com as ocorrências dos diferentes resultados observados. Fonte: Barbetta, Reis e Bornia (2004).

Distribuição de freqüências

O processo de resumo e organização dos dados busca basicamente registrar as ocorrências dos possíveis valores das variáveis que caracterizam o fenômeno, em suma, consistem em elaborar **distribuições de freqüências*** das variáveis, para que o conjunto de dados possa ser reduzido, possibilitando a sua análise.

A construção da distribuição de freqüências exige que os possíveis valores da variável sejam discriminados e seja contado o número

GLOSSÁRIO

*Freqüência absoluta – registro dos valores da variável por meio de contagem das ocorrências no conjunto de dados. Fonte: Barbetta, Reis e Bornia (2004).

*Freqüência 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 (2004).

de vezes em que cada valor ocorreu no conjunto de dados. Para grandes arquivos de dados, tal processo somente é viável utilizando meios computacionais.

Uma distribuição de freqüências 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. Vamos ver cinco casos: quando há apenas uma variável qualitativa, quando há apenas uma variável quantitativa, quando há duas variáveis (sendo ambas qualitativas, ambas quantitativas, ou uma qualitativa e a outra 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. Esta contagem pode ser registrada em números absolutos, **freqüência absoluta***, ou em números relativos, **freqüência relativa ou percentual***. Ambos os registros devem ser feitos e apresentados: a freqüência absoluta permite avaliar se os resultados são sólidos (é temerário tomar decisões com base em pequenas quantidades de dados); já a freqüência relativa possibilita comparar os resultados da distribuição de freqüências com outros conjuntos de tamanhos diferentes. A distribuição de freqüências pode ser apresentada em forma de tabela ou gráfico.

Se alguém diz que 33,33% (percentual) das mulheres de um curso se casaram com professores, você poderia ter uma má impressão destas moças. Mas se alguém diz que das três mulheres (dados brutos) deste curso, uma delas casou-se com um professor, o efeito já não será tão grande. Fonte: Anedota extraída do livro *Como mentir com Estatística*, de Darrel Huff. Rio de Janeiro: Ediouro, 1992.

Exemplo 2: imagine que você está interessado em descrever a variável opinião sobre a concessionária (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 concessionária.

Valores	Freqüência	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

Podemos concluir, neste 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, neste 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 freqüências. Quando a variável é qualitativa, podemos usar dois tipos de gráficos: **em barras** ou **em setores**.

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

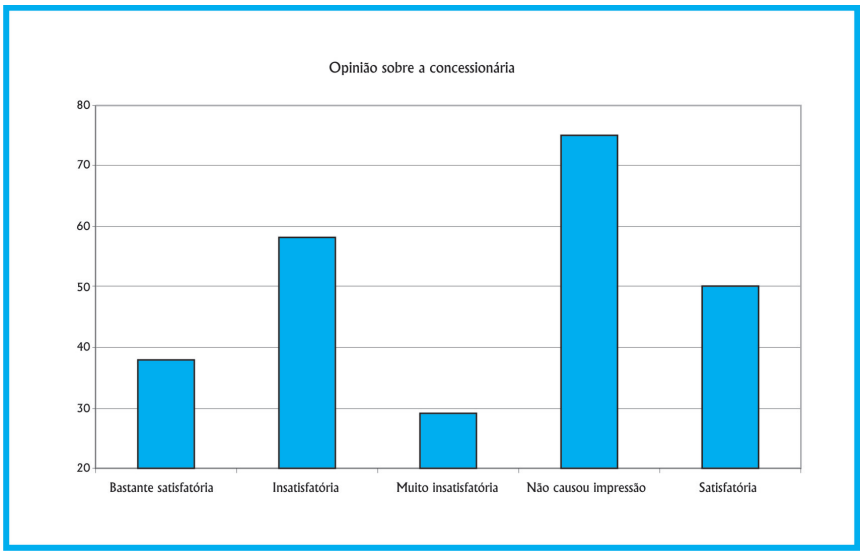


Figura 17: Gráfico em barras para Opinião sobre as concessionárias
Fonte: adaptada pelo autor a partir de Microsoft Office (2007)

Trata-se da mesma distribuição de freqüências 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 freqüência.

Contudo, você consegue identificar alguma particularidade neste gráfico? Olhe bem!

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

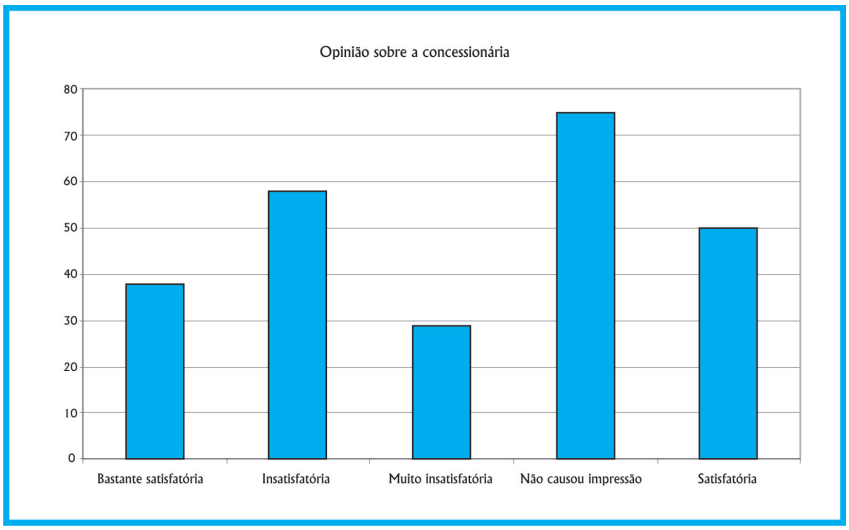


Figura 18: Gráfico em barras para Opinião sobre as concessionárias
Fonte: adaptada pelo autor de Microsoft

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 através de uma regra de três simples, na qual a freqüência total (ou o percentual total 100%) corresponderia aos 360°, e a freqüência ou a proporção de cada categoria corresponderia a um valor desconhecido em graus.

$$\textbf{Graus de uma categoria} = \frac{\textbf{360}^{\textbf{o}} \times \textbf{freqüência (proporção) da categoria}}{\textbf{freqüência (proporção) total}}$$

Observe os valores em graus correspondentes aos resultados do Quadro 1 (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

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

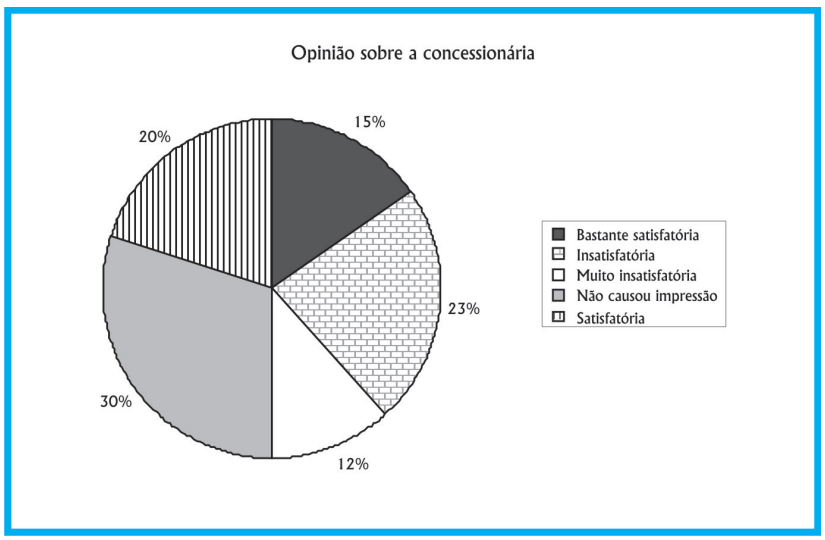


Figura 19: Gráfico em setores para a Opinião sobre as concessionárias
Fonte: adaptada pelo autor de Microsoft®

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

Caso de uma variável quantitativa

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

Se a variável for quantitativa discreta e puder assumir apenas alguns valores, a abordagem será semelhante à das variáveis qualitativas. A diferença reside na substituição de atributos por números, gerando uma **distribuição de freqüência para dados não agrupados**. Vamos ver um exemplo.

Neste terceiro exemplo – 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 4, uma apresentação tabular (em forma de tabela) da variável número de pessoas transportadas.

Valores	Freqüência	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 4: Número de pessoas usualmente transportadas no veículo
Fonte: elaborado pelo autor

Pela observação do Quadro 4, 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 mesmo idade dos consumidores.

E como representar a distribuição de freqüências para variáveis quantitativas discretas graficamente? O Quadro 3 poderia ser representado através de um histograma, um gráfico de barras justapostas, em que as áreas das barras são proporcionais às freqüências de cada valor. Vamos ver (Figura 20):

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

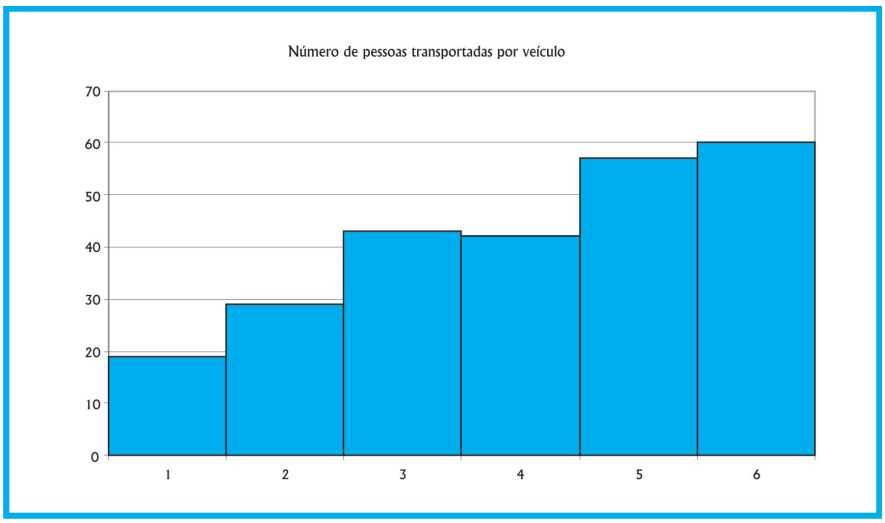


Figura 20: Histograma do Número de pessoas transportadas por veículo
Fonte: adaptada pelo autor de Microsoft

Neste caso, eu poderia 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 esta começa em zero.

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

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

Mais informações, veja a seção Saiba mais.

Se o conjunto for grande, é preciso representar os dados através 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, quantos valores do conjunto se encontram em cada faixa. Há duas maneiras de fazer isso:

- através da **categorização*** (recodificação) da variável, por exemplo, todos que ganham até 4 salários mínimos (R$ 1.520) pertencem à classe baixa, todos que ganham entre 4,01 e 20 salários mínimos (até R$ 7.600) pertencem à classe média, e acima disso pertencem à classe alta – esta abordagem é largamente utilizada na mídia; e
- através de uma **distribuição de freqüências 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 freqüências 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: $\text{Nº de classes} = \sqrt{\text{Nº de elementos}}$. Neste 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 neste 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 freqüências de cada classe; e

GLOSSÁRIO

*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

*Distribuição de freqüências para dados agrupados – distribuição de freqüências na qual os valores da variável são agrupados em faixas de ocorrência, e as freqüências, contadas para cada faixa, para facilitar o resumo do conjunto de dados, usualmente empregado para variáveis quantitativas contínuas. Fonte: Barbetta, Reis e Bornia (2004).

- determinar os pontos médios de cada classe através da média dos dois limites (serão os representantes das classes).

Vamos ver exemplos de ambas as abordagens.

A resolução passo a passo deste problema está na seção Saiba mais desta Unidade, que explica como realizar análise exploratória de dados no Excel®. Aqui apresentaremos apenas os resultados finais.

Exemplo 4: 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:

a) se optássemos por categorizar a variável da seguinte forma: todos que ganham até 4 salários mínimos (R$ 1.520) pertencem à classe baixa, todos que ganham entre 4,01 e 20 salários mínimos (até R$ 7.600) pertencem à classe média, e acima disso pertencem à classe alta?; e

b) se optássemos por uma distribuição de freqüências para dados agrupados?

No caso do item a, a categorização levará à criação de uma nova variável, agora qualitativa, permitindo uma abordagem semelhante à que vimos anteriormente. No Quadro 5 e na Figura 21, estão os resultados: tabela de freqüências e gráfico em setores.

Valores	Freqüência	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 5: Renda categorizada em classe social
Fonte: elaborado pelo autor

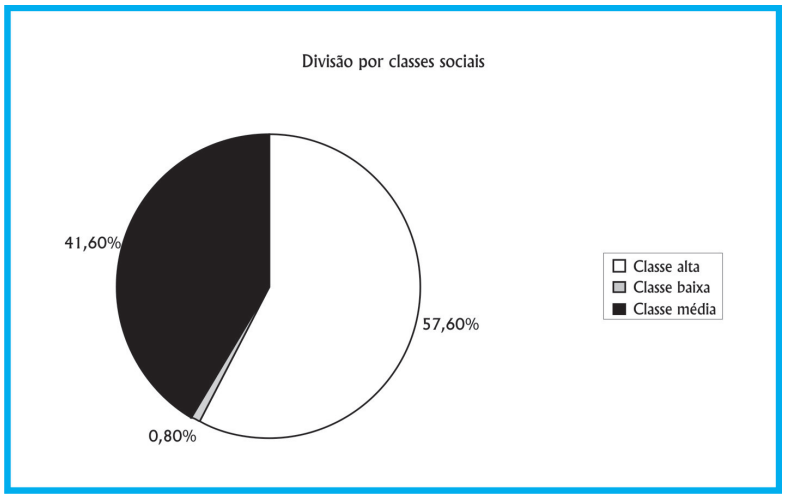


Figura 21: Gráfico em setores para a Renda categorizada em classes
Fonte: adaptada pelo autor de Microsoft Office (2007)

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 neste 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:

- Intervalo = Maior – 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);

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

- Nº de classes = $\sqrt{\text{Nº de elementos}} = \sqrt{250} = 15{,}81 \cong 16$. Por este 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 5 classes.

  Amplitude das classes = 86,015/5 = 16,844 (valor exato)

  A amplitude das classes pode ser ligeiramente maior do que a obtida acima, poderíamos, novamente procurando a simplificação do problema, usar amplitude igual a 16,85. Se a

amplitude não for um valor exato, deve sempre ser arredondada para cima, garantindo que as classes conterão 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 (neste 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: (limite inferior + limite superior)/2 (os pontos médios calculados estão no quadro abaixo); e
- freqüências de cada classe (Quadro 6):

Classes	Freqüência	Percentuais	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 6: Renda agrupada em classes
Fonte: elaborado pelo autor

Observe que perdemos informação sobre o conjunto original: sabe-se 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 freqüências das classes passam a ser as freqüências dos pontos médios. Podemos afirmar que quase 80% dos clientes têm renda até 35,495 salários mínimos.

O Quadro 6 também pode ser representado através de um histograma (Figura 22), uma vez que a variável permanece sendo formalmente quantitativa. Mas o histograma para uma tabela de dados agrupados é um pouco diferente do 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.

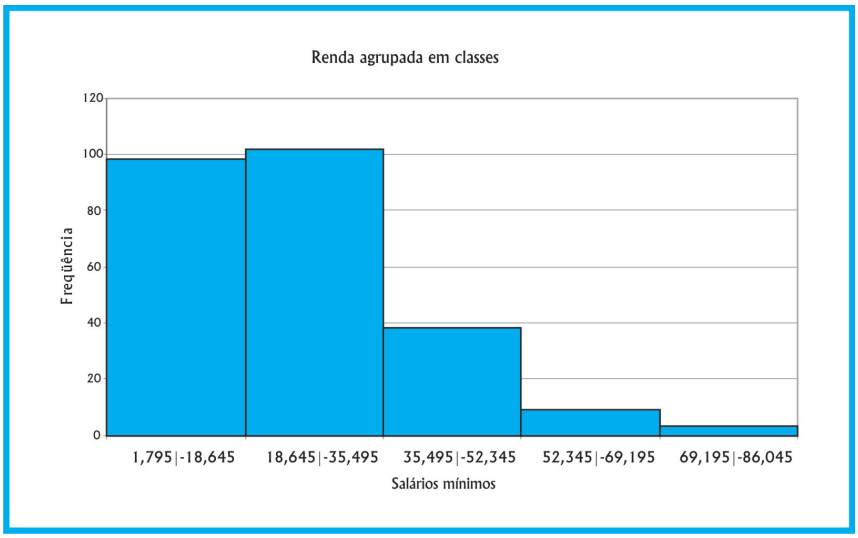


Figura 22: Histograma para Renda agrupada em classes
Fonte: adaptada pelo autor de Microsoft Office (2007)

Note que a interpretação é mais direta quando olhamos o gráfico apresentado na Figura 20.

Mas aqui surge um fato interessante. Parece haver contradição com a interpretação do item a, na qual concluímos que a maioria absoluta dos clientes é de classe alta. Isso ocorre devido à definição arbitrária das classes, e à ainda mais arbitrária definição de classes baixa, média e alta. Pense em como você resolveria esta 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 uma grande quantidade de dados, existe uma maior probabilidade de que estas estimativas correspondam exatamente aos verdadeiros valores. Outro problema são as medidas estatísticas calculadas com base na distribuição de freqüências para dados agrupados: serão apenas estimativas dos valores reais devido à perda de informação referida acima.

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. Começaremos com duas variáveis qualitativas.

Caso de duas variáveis qualitativas

O administrador freqüentemente precisa estudar o relacionamento entre duas ou mais variáveis, para tomar decisões. Por exemplo, há relação entre o sexo do consumidor e a preferência por um modelo de carro, ou entre a escolaridade do eleitor e o candidato a presidente escolhido, entre outras.

Quando as duas variáveis são qualitativas (originalmente ou quantitativas categorizadas), usualmente é construída uma distribuição conjunta de freqüências, também chamada de **tabela de contingências*** ou dupla classificação. Nela são contadas as freqüências 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 exemplo.

Para a mesma situação do Exemplo 1. Agora, você está interessado em observar o relacionamento entre a variável Modelo adquirido e a Opinião geral do cliente sobre o veículo, e expressá-lo 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 freqüências. Para grandes bases de dados, mesmo para o nosso exemplo em que há apenas 250 casos, seria um processo tedioso e sujeito a erros. Portanto, o mais inteligente é utilizar alguma ferramenta computacional, mesmo uma planilha eletrônica como o Microsoft Excel®.

Usando uma ferramenta computacional, chegaremos ao Quadro 7.

GLOSSÁRIO

*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 freqüências de todos os cruzamentos possíveis. Fonte: Barbetta (2006).

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 7: Tabela de contingências de modelo por opinião geral (apenas freqüências)
Fonte: elaborado pelo autor

Observe a última coluna e a última linha do quadro acima: são os chamados **totais marginais***, isto é, as freqüências 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 freqüentes 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 grandes problemas. Apenas quando a quantidade ultrapassa 5% da base de dados, há motivo para preocupação, pois 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.

Conforme vimos na Unidade 1.

GLOSSÁRIO

*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 (2002).

O Quadro 7 pode ser apresentado de forma gráfica, através de um gráfico de barras múltiplas (Figura 23).

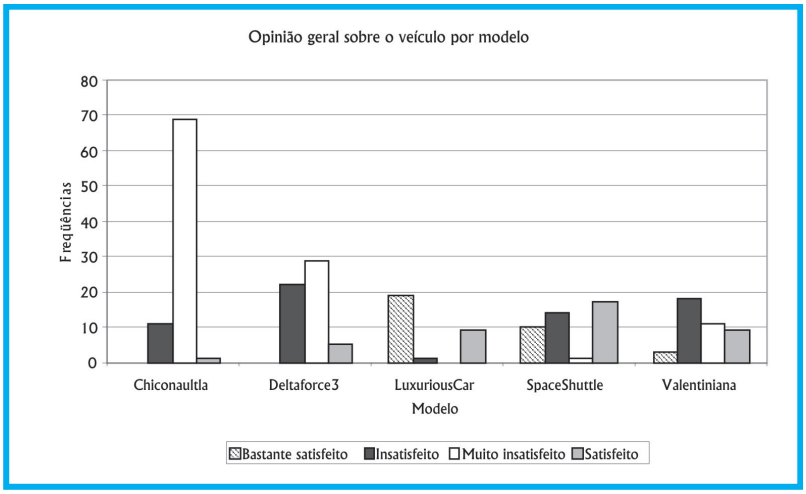


Figura 23: Gráfico de barras múltiplas: Opinião geral por modelo
Fonte: adaptado pelo autor de Microsoft Office (2007)

As freqüências absolutas podem ser insuficientes para a interpretação dos resultados, especialmente quando comparando os resultados com outros conjuntos de dados de tamanhos diferentes. Assim, podemos calcular percentuais em relação aos totais de cada coluna, aos totais de cada linha ou ao total geral da tabela. Vamos apresentar apenas um dos percentuais possíveis, em relação aos totais das linhas. No texto "Como realizar análise exploratória de dados no Microsoft Excel®", são apresentados todos os resultados (Quadro 8):

Opinião geral sobre o veículo									
Modelo	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 8: Tabela de contingência de Opinião geral por modelo (com % por linha)
Fonte: elaborado pelo autor

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, e cada vez mais esse profissional precisa avaliar mais de duas variáveis, o que exige métodos matemáticos sofisticados, implementados computacionalmente. Veremos mais sobre esse tema a seguir.

Caso de duas variáveis quantitativas

Muitas vezes, também estamos interessados em avaliar o relacionamento entre variáveis quantitativas, sejam elas discretas, sejam 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) desta correlação, caso ela exista.

Uma das variáveis é chamada de independente. Esta 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 uma outra variável, chamada de dependente. Seus valores são resultado da variação dos valores das variáveis independentes.

Reveja as definições de variáveis na Unidade 1.

Esta denominação costuma levar à 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 uma certa "força", ou seja, correlação não significa causalidade.

GLOSSÁRIO

*Correlação – medida de associação entre duas variáveis quantitativas. Fonte: Barbetta, Reis e Bornia (2004).

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).

GLOSSÁRIO

*Observações emparelhadas – medidas de duas ou mais variáveis que foram realizadas na mesma unidade experimental/amostral, no mesmo momento. Fonte: elaborado pelo autor

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

Se estivermos analisando duas variáveis quantitativas, cujas observações constituem pares ordenados, chamando estas variáveis de **X** (independente) e **Y** (dependente), podemos 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, até mesmo uma planilha eletrônica como o Microsoft Excel®.

Saiba mais no texto "Como realizar análise exploratória de dados no Microsoft Excel®".

Através do diagrama de dispersão, podemos ter uma idéia 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 natureza da correlação (se é possível ajustar uma reta, parábola, exponencial, aos pontos).

Vamos a um exemplo para ilustrar.

Para a mesma situação do Exemplo 1, gostaríamos de saber como é o relacionamento entre a variável Renda mensal do cliente e a Quilometragem média mensal por ele percorrida.

As duas variáveis de interesse (Renda e Quilometragem) são quantitativas. Quem pode influenciar quem? É mais lógico imaginar que, quanto maior a renda familiar, haverá mais dinheiro para comprar combustível, e, portanto, maior a quilometragem percorrida com o veículo. Sendo assim, a variável renda será posta no eixo horizontal (X) do diagrama de dispersão, e a quilometragem, no eixo vertical (Y). Veja a Figura 24:

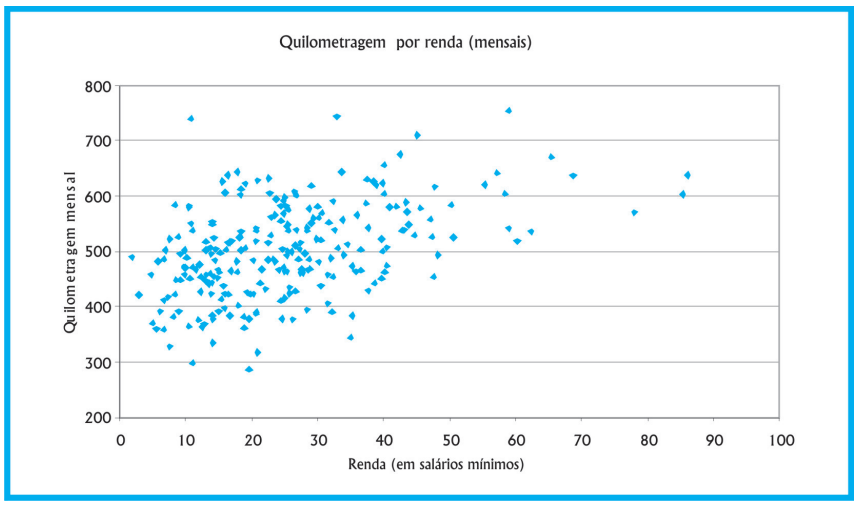


Figura 24: Diagrama de dispersão de Quilometragem por Renda
Fonte: adaptada pelo autor de Microsoft®

Aparentemente, a correlação entre as variáveis é positiva, como era esperado: maiores valores de renda correspondem a maiores valores de quilometragem. A correlação não parece ser muito forte, pois os pontos não estão muito próximos. Quanto à natureza, é difícil afirmar, talvez seja linear, mas é apenas um palpite neste caso.

Para esclarecer, vamos ver outro exemplo.

Neste caso, uma empresa agroindustrial processa soja para obter óleo. A direção quer estudar o relacionamento 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 York. Para tanto,

coletou um conjunto de 400 pares de observações e plotou o diagrama de dispersão exposto na Figura 25.

Observando o diagrama (Figura 25), é possível afirmar que o relacionamento entre as variáveis é fortemente linear?

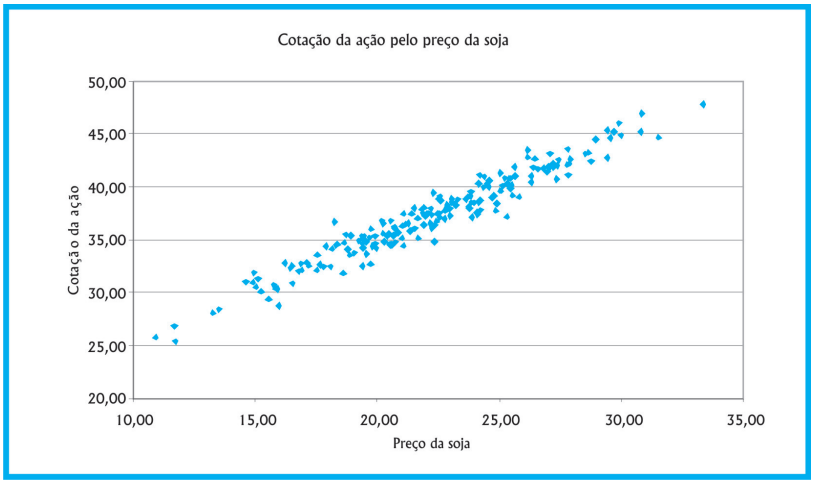


Figura 25: Diagrama de dispersão de Cotação da ação por Preço da soja
Fonte: adaptada pelo autor de Microsoft

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, pode-se observar que seria possível ajustar uma reta entre os pontos. Portanto, conclui-se que o relacionamento 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, trimestres), teremos uma **série temporal***.

Você já deve ter visto em algum lugar uma tabela ou um gráfico mostrando a 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 dos candidatos a presidente em cada pesquisa eleitoral. O objetivo da análise de uma série temporal é identificar a existência de padrões que nos auxiliem a tomar decisões.

GLOSSÁRIO

*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 tenha que lidar com séries temporais.

Vamos agora ao último caso desta Unidade, muito importante para o administrador, pois é bastante comum ele ter que estudar o relacionamento entre uma variável qualitativa e outra quantitativa.

Caso de uma variável qualitativa e uma quantitativa

Usualmente, pressupõe-se que analisaremos a variável quantitativa em função dos valores da variável qualitativa, visto que esta última 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 estudamos 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. Neste 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. Neste caso, gostaríamos de avaliar se existe algum relacionamento entre a renda do consumi-

dor e o modelo adquirido. Espera-se que exista tal relacionamento, pois os modelos Chiconaultla e DeltaForce3 são os mais baratos, e o sofisticado LuxuriousCar é o mais caro de todos.

Neste caso, podemos obter distribuições de freqüências da variável Renda para cada valor da variável Modelo. Seria uma situação semelhante à 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 freqüências em cada modelo, resultando os dados ilustrados no Quadro 9:

RENDA	MODELO					Total
	Chiconautla	DeltaForce3	Valentiniana	SpaceShuttle	LuxuriousCar	
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 freqüências de Renda agrupadas em classe por Modelo
Fonte: elaborado pelo autor

Há 249 dados no quadro, porque o dado perdido (descoberto no Quadro 7) foi removido do conjunto.

Observe a semelhança do quadro mostrado acima com o Quadro 7. Da mesma forma que lá fizemos, é possível calcular percentuais em relação aos totais das linhas, colunas ou total geral.

Podemos perceber que o relacionamento esperado entre as variáveis foi confirmado: 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.

O Quadro 9 poderia ser expresso através de um gráfico, um **histograma categorizado**. Infelizmente, tal gráfico não pode ser feito em uma planilha eletrônica (como o Excel®) sem consideráveis manipulações. Mas, através de um software estatístico, no nosso caso, o Statsoft Statistica 6.0®, isso é possível (Figura 26):

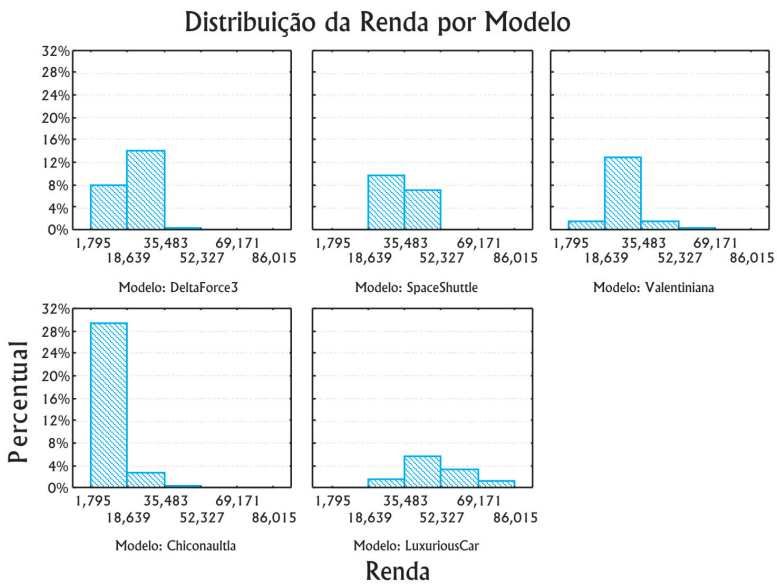


Figura 26: Histograma categorizado da Renda por Modelo adquirido
Fonte: adaptada pelo autor de Statsoft®

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

Saiba mais...

- Sobre correlação entre variáveis (quantitativas): BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006, capítulo 13.

- Sobre correlação entre variáveis (quantitativas): MOORE, D. S.; et al. *A prática da Estatística empresarial*: como usar dados para tomar decisões. Rio de Janeiro: LTC, 2006, capítulo 2.

- Sobre análise de séries temporais: LEVINE, D. M.; et al. *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 através 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.

RESUMO

O resumo desta Unidade está mostrado na Figura 27:

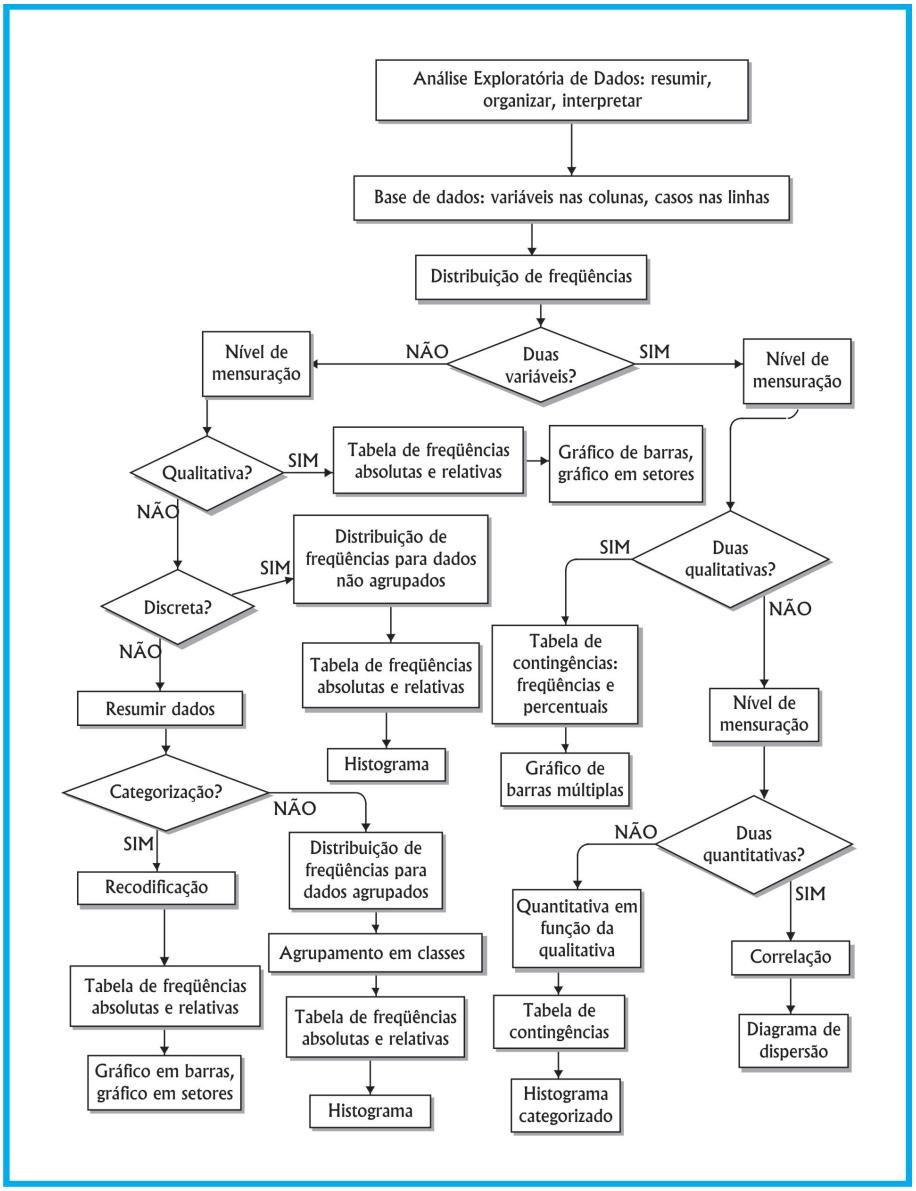


Figura 27: Resumo da Unidade 3
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Caro estudante!

Esta Unidade foi importantíssima para você entender a Análise Exploratória de Dados. Vimos como organizar, interpretar e resumir as informações coletadas, os níveis de mensuração e o 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. Leia e releia quantas vezes sejam necessárias os variados exemplos propostos para cada categoria estudada. As figuras, os quadros, as representações e os exemplos são grandes aliados nesse processo de aprendizagem.

Interaja com sua turma e responda as atividades. A tutoria está pronta a lhe auxiliar, e o professor, ansioso em reconhecer suas habilidades desenvolvidas a partir do conhecimento deste conteúdo. Vamos em frente!

UNIDADE

4

Análise Exploratória de Dados II

Objetivo

Nesta Unidade, você vai conhecer mais uma maneira de descrever e analisar um conjunto de dados referente a uma variável quantitativa (discreta ou contínua): através das medidas de síntese.

Medidas de posição ou de tendência central

Caro estudante!

Na Unidade 3, estudamos como fazer a descrição tabular e gráfica das variáveis, seja isoladamente, seja relacionadas a outras, e interpretar os resultados obtidos. Além daquelas 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. Sua utilização pode ser feita de forma complementar às 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 sempre em mente que é indispensável que o administrador conheça as medidas de síntese para que possa realizar Análise Exploratória de Dados através delas. Vamos ver 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.

GLOSSÁRIO

*Média aritmética simples – medida de posição que é o 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 (2006).

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 xi 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.

No Microsoft Excel®, a média aritmética simples é implementada através da função MÉDIA( ).

Vamos ver um exemplo que vai nos acompanhar por algum tempo. 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.

Com o objetivo é calcular a média de cada turma, ao somar os valores teremos o mesmo resultado: 48. Como cada turma tem oito 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 que as três turmas têm desempenhos idênticos. Será? Observe atentamente o Quadro 10.

Veja que na primeira turma temos realmente os dados distribuídos regularmente em torno da média, com a mesma variação tanto abaixo quanto acima. Já na segunda, vemos uma distorção maior; embora a maioria das notas seja alta, algumas notas baixas "puxam" a média para um valor menor. E, no terceiro grupo, há apenas uma nota baixa, mas seu valor é tal que realmente 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 é a posição em que um suporte equilibra esta tábua.

Vamos ver como os valores do exemplo distribuem-se em um diagrama apropriado (Figura 28):

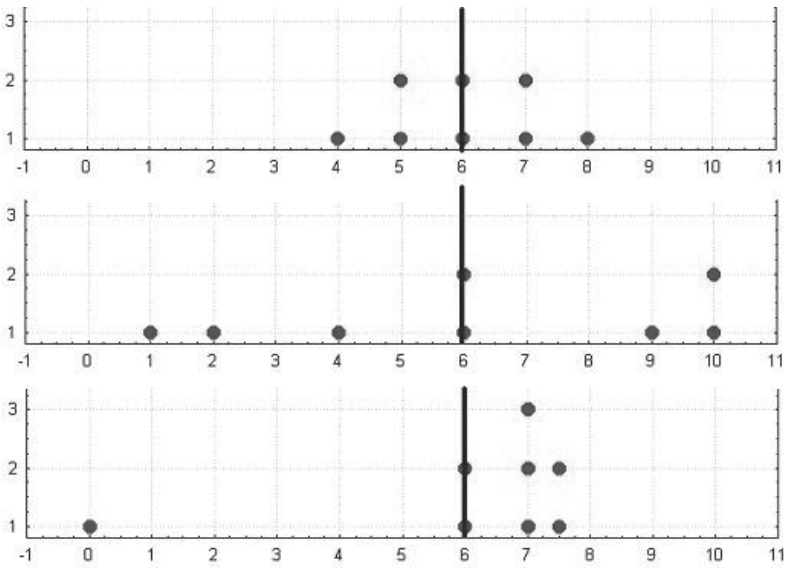


Figura 28: Interpretação do valor da média
Fonte: adaptada pelo autor de Microsoft Office e Statsoft®

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

GLOSSÁRIO

*Valores discrepantes – 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 pelo autor de Bussab e Morettin (2002).

Essa era a grande crítica que era feita nas décadas de 1960 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.

distribuição já é mais irregular, com valores mais "distantes" na parte de baixo, e no o terceiro grupo, a distribuição 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.

Outro aspecto importante a ressaltar é que a média pode ser um valor que a variável não pode assumir. Isto 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 diariamente pela mídia em vários países.

É extremamente comum calcular médias de variáveis quantitativas a partir de distribuições de freqüências representadas em tabelas: simplesmente, multiplica-se cada valor (ou o ponto médio da classe) pela freqüência 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 freqüências 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 freqüência de um valor qualquer da variável discreta ou de uma classe qualquer, e n é o número total de elementos do conjunto.

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

GLOSSÁRIO

*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 (2006).

Valores	Freqüência	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

Precisamos multiplicar a coluna de valores x_i pela da freqüência 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	Freqüência 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	1.019

Quadro 12: Número de pessoas usualmente transportadas no veículo
Fonte: elaborado pelo autor

Agora, podemos calcular a média:

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

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

Veja novamente a Figura 18 da Unidade 3 e observe como o valor da média permite equilibrar os pesos e as freqüências dos vários valores da variável.

Se quisermos calcular a média aritmética simples a partir de uma distribuição de freqüências 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 freqüências 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 freqüências 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 freqüência 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.

Ela 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.

Neste 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.

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

Por esse motivo, os dados precisam estar ordenados crescentemente.

Turma A: Md = (6 + 6)/ 2 = 6

Turma B: Md = (6 + 6)/ 2 = 6

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 = 5ª. Como o conjunto tem um número ímpar de valores, o valor da mediana será igual ao valor que estiver na 5ª posição.

Mediana = 15 Média = 20,89

Observe que, neste caso, média e mediana são diferentes, pois a média foi distorcida pelos valores mais altos 35 e 60, que constituem uma minoria. Neste 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**.

Veremos no Excel que 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 freqüências, 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 freqüência no conjunto. Pode, então, ser considerada a 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 freqüente, podendo este 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. Este 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 (Quadro 14).

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.

A turma A tem três 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 à 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% 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% 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, primeiramente, calcular as posições dos quartis.

Posição do quartil inferior = **(n + 1)/4**

Posição do quartil superior = **[3×(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 anteriores e imediatamente posteriores à posição calculada. Se os dados estiverem dispostos em uma distribuição de freqüências, utilizar o mesmo procedimento observando as freqüências associadas a cada valor (variável discreta) ou ponto médio de classe.

No Excel®, os quartis são implementados através da função QUARTIL(;1) para quartil inferior e QUARTIL(;3) para quartil superior.

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

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

Há 29 elementos no conjunto, que já está ordenado crescentemente. Podemos calcular as posições dos quartis.

Posição do quartil inferior = (n + 1)/4 = (29 + 1)/4 = 7,5ª.
Posição do quartil superior = [3×(n+1)]/4 = [3 × (29 + 1)]/4 = 22,5ª.

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

Qi = (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 idéia do nível de renda daquela população.

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

Qs = (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.

GLOSSÁRIO

*Medidas de dispersão – medidas numéricas que visam a avaliar a variabilidade do conjunto de dados, sintetizando-a em um número. Fonte: elaborado pelo autor

Medidas de dispersão ou de variabilidade

O objetivo das **medidas de dispersão*** é mensurar quão próximos uns dos outros estão os valores de um grupo (e algumas medem 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 idéia 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.

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, percebe-se facilmente que a dispersão dos dados tem comportamento diferente nas duas tur-

mas, e essa é a principal desvantagem do uso do intervalo como medida de dispersão.

Colocaremos os dados do Exemplo 7 em um diagrama apropriado (Figura 29):

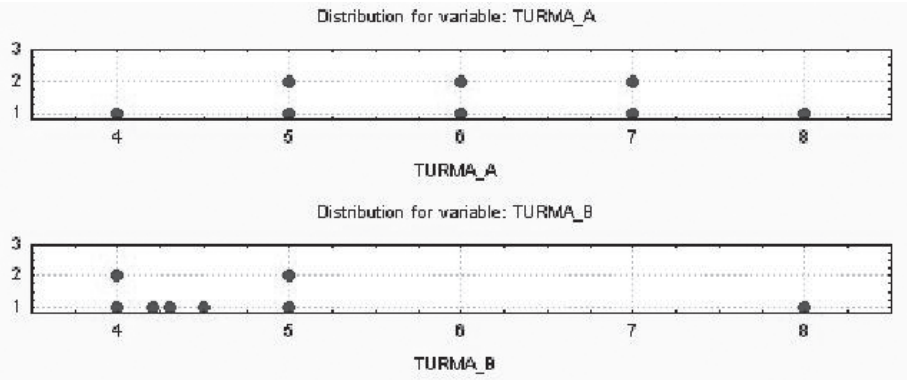


Figura 29: Desvantagem do uso do intervalo como medida de dispersão
Fonte: adaptada pelo autor de Statsoft® e Microsoft®

Observa-se 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 idéia de como os dados estão distribuídos entre os extremos (não permite identificar que o valor 8 na turma B é um valor discrepante).

No Excel®, podemos obter o intervalo através das funções MÁXIMO ( ) e MÍNIMO ( ) .

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.

A razão dessa distinção será explicada mais adiante, na Unidade 7. Pode-se 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.

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? Deixo 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 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 kg^2, 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 se pode afirmar, porém, é que, quanto maior a variância, mais dispersos os dados estão em torno da média (maior a dispersão do conjunto).

No Excel®, a variância populacional é obtida através da função VARP( ), e a variância amostral, através da função VAR( ).

Para fins de Análise Exploratória de Dados, caracterizar a dispersão através da variância não é muito adequado. Costuma-se usar a raiz quadrada positiva da variância, o desvio-padrão. Vamos ver mais sobre isso? Continuemos, então, a estudar!

Desvio-padrão (s)

É a raiz quadrada positiva da variância, apresentando a mesma unidade dos dados e da média, permitindo 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 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 ter uma idéia do valor típico e da distribuição dos dados em torno dele.

Deixo como exercício para você elevar os desvios obtidos com os dados das turmas, expressos no Quadro 10, Exemplo 1, ao quadrado, somá-los e dividi-los por 7 (suponha que se trata de uma amostra). Assim, você obterá os desvios-padrão das notas das turmas.

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, e este erro 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:

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

Primeiramente, cada valor (x_i) do conjunto é elevado ao quadrado, e somam-se todos os resultados obtendo $\sum_{i=1}^{n} x_i^2$. Somam-se também todos os valores do conjunto para obter $\sum_{i=1}^{n} x_i$, somatório este que será elevado ao quadrado. Os somatórios e o valor de **n** (número de elementos no conjunto) são substituídos na fórmula para obter os resultados.

É desta forma que os programas computacionais calculam o desvio-padrão.

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

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

Onde $\mathbf{x_i}$ é o valor da variável ou ponto médio da classe, $\mathbf{f_i}$ é a freqüência 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.

No Excel®, podemos obter o desvio-padrão populacional através da função DESVPADP( ), e amostral, através da função DESVPAD ( ).

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

Há 29 elementos no conjunto, n = 29.

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

Elevando cada valor ao quadrado e somando-os, vamos obter:

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

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

$$s = \sqrt{\frac{\sum_{i=1}^{29}(x_i^2) - \left[\frac{\left(\sum_{i=1}^{29} x_i\right)^2}{29}\right]}{29-1}} = \sqrt{\frac{17497{,}919125 - \left[\frac{(654{,}935)^2}{29}\right]}{29-1}} = \sqrt{\frac{17497{,}919125 - 14791{,}02946}{28}}$$

$s \cong 9{,}83$ 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 idéia de criar uma medida de dispersão que relacionasse média e desvio-padrão, o coeficiente de variação percentual, que veremos a seguir.

GLOSSÁRIO

*Coeficiente de variação percentual – resultado da divisão do desvio-padrão pela média do conjunto, multiplicado por 100, permite avaliar o quanto o desvio-padrão representa da média. Fonte: Barbetta, Reis e Bornia (2004); Anderson, Sweeney e Williams (2007).

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-padrão).

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

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.

Neste exemplo, 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.

Para a turma A: $\bar{x} = 6$ s = 1,31 c.v.% = (1,31/6) x 100 = 21,82%
Para a turma B: $\bar{x} = 6$ s = 3,51 c.v.% = (3,51/6) x 100 = 58,42%
Para a turma C: $\bar{x} = 6$ s = 2,49 c.v.% = (2,49/6) x 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 anteriormente, a comparação ficou ainda mais simples, pois as médias dos grupos eram iguais, bastaria avaliar apenas os desvios-padrão 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: "mas por que é tão importante calcular a média e o desvio-padrão dos valores de uma variável registrados em um conjunto de dados?". Argumentam que talvez a mediana seja uma melhor medida de posição e que os quartis permitem ter uma boa idéia da dispersão. Contudo, há um teorema que permite, a partir da média e do desvio-padrão, obter estimativas dos extremos do conjunto, especialmente quando se trata de uma amostra: é o teorema de Chebyshev, também chamado de Desigualdade de Chebyshev.

Teorema de Chebyshev

A proporção (ou fração) de **qualquer** conjunto de dados a menos de K desvios-padrão a contar da média é sempre ao menos $1 - 1/K^2$, onde K é um número positivo maior do que 1. Provavelmente, você não entendeu nada... Vamos tentar esclarecer.

Vamos supor que K fosse igual a 2 ou igual a 3:

- para K = 2, pelo teorema de Chebyshev, $1 – 1/K^2 = 0,75$; então, ao menos 3/4 (75%) de todos os elementos do conjunto estão no intervalo que vai de dois desvios-padrão abaixo da média a dois desvios-padrão acima da média;
- para K = 3, pelo teorema de Chebyshev, $1 – 1/K^2 = 0,89$; então, ao menos 8/9 (89%) de todos os elementos do conjunto estão no intervalo que vai de três desvios-padrão abaixo da média a três desvios-padrão acima da média.

Uma pesquisa por amostragem obteve que a renda mensal de um Estado apresenta média de 800 reais e desvio-padrão de 200 reais. Neste décimo exemplo, usando o teorema de Chebyshev, vamos identificar os limites estimados onde estão 75% das rendas.

Conforme visto anteriormente, se a proporção de interesse é 0,75 (75%), então K será igual a 2. Assim, podemos encontrar os valores que estão a dois desvios-padrão da média:

- 2 desvios-padrão abaixo = 800 – 2.200 = 400 reais
- 2 desvios-padrão acima = 800 + 2.200 = 1.200 reais.

Então, pelo menos 75% das rendas mensais devem estar entre 400 e 1.200 reais. Isso possibilita avaliar a distribuição de renda sem a necessidade de um censo (ver Unidades 1 e 2).

Na prática, as proporções reais costumam ser maiores do que os valores calculados pelo Teorema de Chebyshev. Mas o Teorema apresenta a vantagem de ser válido para todos os casos e não exigir o conhecimento da distribuição seguida pelos dados para estimar as proporções, basta apenas o cálculo da média e do desvio-padrão.

Mas precisamos combinar várias medidas para uma análise mais elaborada, especialmente no que se refere à assimetria e à simetria da distribuição dos valores da variável quantitativa no conjunto de dados, que veremos a seguir.

Assimetria das distribuições

Identificar se a distribuição de uma variável quantitativa em um determinado conjunto de dados é simétrica ou assimétrica pode ser de grande valia por vários motivos:

1) se os dados são provenientes de uma amostra, identificar a simetria ou não da distribuição pode ser necessário para selecionar o modelo probabilístico mais adequado para descrever a variável na população;

Na Unidade 6, você vai estudar alguns modelos.

2) no caso de um experimento em que todas as causas de variação indesejadas são suprimidas, a ocorrência de assimetria quando era esperada simetria ou o contrário pode indicar que houve algum erro de planejamento ou de medição; e

3) nos casos em que são comparadas distribuições da mesma variável quantitativa em situações diferentes, a identificação de um comportamento assimétrico ou simétrico, inesperado ou diferenciado pode alertar para aspectos anteriormente despercebidos ou existência de erros.

Alguns programas computacionais calculam uma medida de assimetria ("*skewness*"): quando este valor é exatamente igual a zero, a distribuição em questão é perfeitamente simétrica. Mas a forma ideal de analisar a simetria de uma distribuição é combinar a avaliação das medidas e de um gráfico, seja um histograma, seja um diagrama em caixas. As Figuras 30, 31 e 32 apresentam gráficos de distribuições que poderiam ser ajustados a histogramas.

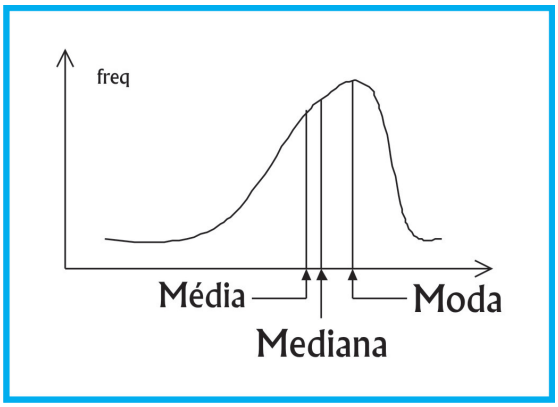


Figura 30: Distribuição assimétrica negativa (assimétrica para a esquerda)
Fonte: elaborada pelo autor

Observe que o "pico" da distribuição, identificado pela moda, está à direita do gráfico, indicando que "falta algo" à esquerda, justificando a denominação "assimétrica à esquerda". Observe também que a mediana é *maior* do que a média. Há uma medida estatística de assimetria que calcula a diferença entre média e mediana: quando a diferença é negativa (mediana maior do que a média), a distribuição é "assimétrica negativa". Este tipo de distribuição poderia retratar as idades em alguns países europeus, onde a taxa de natalidade dos naturais do país é muito baixa, e, devido à qualidade de vida, a longevidade é grande.

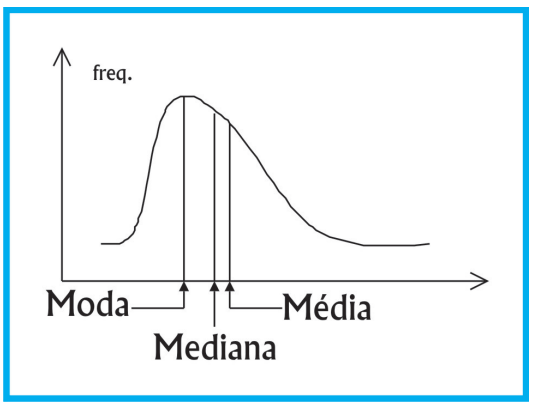


Figura 31: Distribuição assimétrica positiva (assimétrica para a direita)
Fonte: elaborada pelo autor

Observe que o "pico" da distribuição, identificado pela moda, está à esquerda do gráfico, indicando que "falta algo" à direita, justifi-

cando a denominação “assimétrica à direita”. Observe também que a média é *menor* do que a mediana. Agora, a diferença entre média e mediana será positiva: quando a diferença é positiva, a distribuição é “assimétrica negativa”. Este tipo de distribuição é razoavelmente comum na prática, pois é fácil obter valores excepcionalmente altos, sendo o caso mais típico a variável renda.

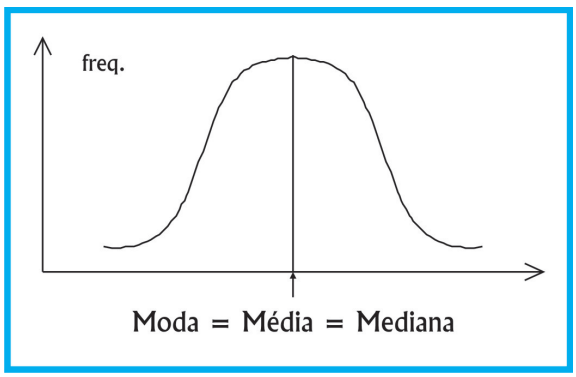


Figura 32: Distribuição simétrica
Fonte: elaborada pelo autor

Observe que as três medidas de posição coincidem. E que aproximadamente metade dos dados está abaixo do centro, e a outra metade, acima, ou seja, a distribuição é “simétrica” em relação às suas medidas de posição. A diferença entre média e mediana é igual a zero. Muitas variáveis apresentam distribuição simétrica, especialmente aquelas resultantes de medidas corpóreas, mas não somente. As Figuras a seguir apresentam histogramas de distribuições assimétricas e simétrica.

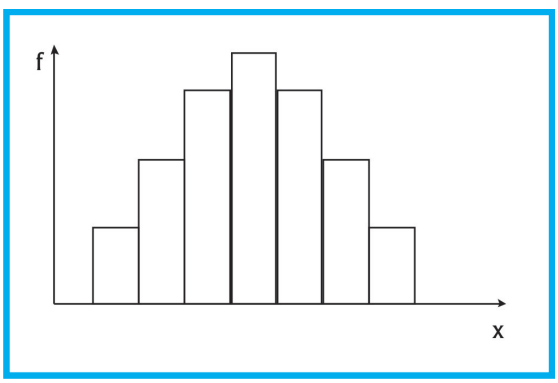


Figura 33: Histograma de distribuição simétrica
Fonte: elaborada pelo autor

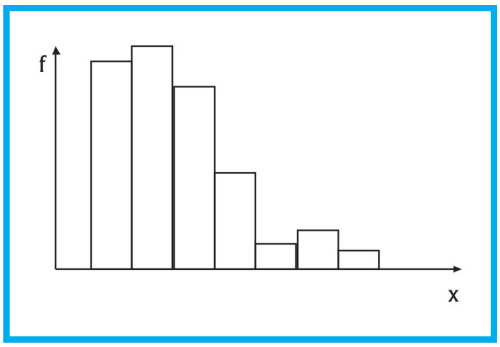


Figura 34: Histograma de distribuição assimétrica para a direita (negativa)
Fonte: elaborada pelo autor

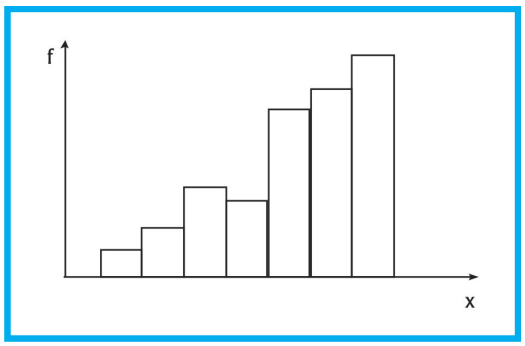


Figura 35: Histograma de distribuição assimétrica para a esquerda (positiva)
Fonte: elaborada pelo autor

Podemos utilizar a mediana e os quartis para avaliar não só a simetria, mas também a dispersão de um conjunto de dados. O procedimento para verificar a existência de assimetria consiste em avaliar a diferença existente entre os quartis e a mediana: se os quartis inferior e superior estiverem à mesma distância da mediana, a distribuição do conjunto pode ser considerada simétrica. A avaliação da dispersão depende da existência de um padrão para comparação, seja um outro conjunto de dados, seja alguma especificação. Um conjunto de dados apresentará maior dispersão do que outro se os seus quartis estiverem mais distantes da mediana. Observe as Figuras a seguir.

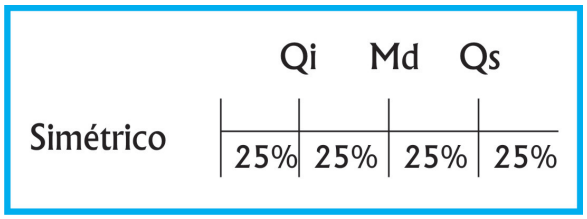


Figura 36: Quartis de uma distribuição simétrica – 1º caso
Fonte: elaborada pelo autor

Observe que a diferença Qs – Md é igual à diferença Md – Qi, o que indica a simetria do conjunto. É importante lembrar que os quartis dividem o conjunto em quatro partes iguais (25% dos dados).

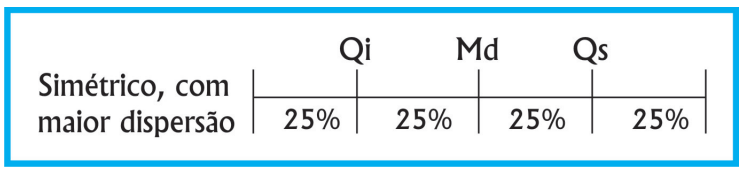


Figura 37: Quartis de uma distribuição simétrica – 2º caso
Fonte: elaborada pelo autor

Observe que a diferença Qs – Md continua igual à diferença Md – Qi, o que indica a simetria do conjunto. Mas agora a dispersão do conjunto é maior, quando comparada ao 1º caso: os quartis estão mais distantes da mediana (as diferenças Qs – Md e Md – Qi serão maiores do que as obtidas no 1º caso).

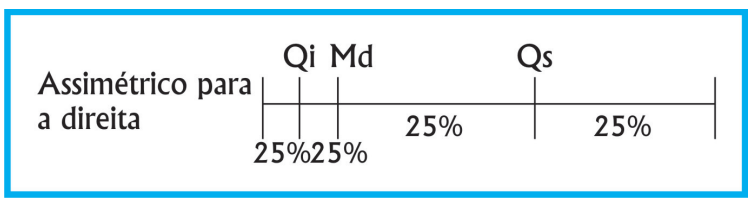


Figura 38: Quartis de uma distribuição assimétrica para a direita
Fonte: elaborada pelo autor

Na Figura 38, é fácil perceber que as diferenças são claramente desiguais: há assimetria. E como Qs – Md é maior do que Md – Qi, é para a direita. O conjunto apresenta uma dispersão mais elevada nos valores maiores. Isso fez com que o quartil superior aumentasse de

valor (deslocando-o para a direita) e ficasse mais distante da mediana do que o inferior, significando assimetria para a direita (ou positiva).

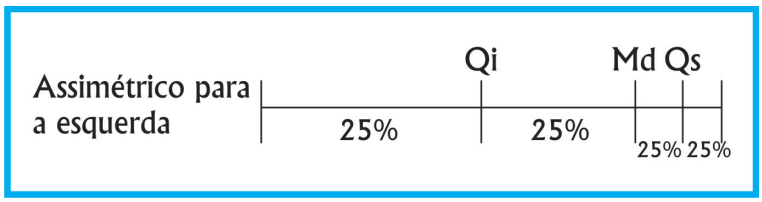


Figura 39: Quartis de uma distribuição assimétrica para a esquerda
Fonte: elaborada pelo autor

Na Figura 39, novamente as diferenças são claramente desiguais: há assimetria. E como Md – Qi é maior do que Qs – Md, é para a esquerda. Neste caso, ocorre o oposto da Figuras 36. Há maior dispersão nos valores mais baixos, fazendo com que o quartil inferior aumentasse de valor e ficasse mais distante da mediana do que o superior, significando assimetria para a esquerda (ou negativa).

A avaliação de assimetria e dispersão também pode ser feita por meio de uma ferramenta gráfica, o diagrama em caixas, que não será apresentado aqui.

Outro aspecto muito interessante das medidas de síntese é a possibilidade de calculá-las para subgrupos do conjunto de dados, em função dos valores de uma outra variável do conjunto. Veremos isso a seguir.

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 que você veja novamente o oitavo exemplo da Unidade 3. Verá que construímos distribuições de freqüências 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! Vamos ver o exemplo a seguir.

Para a mesma situação dos Exemplos 1 e 8 da Unidade 3, gostaríamos de avaliar, neste décimo primeiro exemplo, se existe algum relacionamento entre a renda do consumidor e o modelo adquirido. Espera-se que exista tal relacionamento, pois os modelos Chiconaultla e DeltaForce3 são os mais baratos, e o sofisticado LuxuriousCar é o mais caro de todos.

Através do Microsoft Excel®, 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). Ao realizar este procedimento, usando os dados do arquivo AmostraToyord.xls, vamos obter (Quadro 19):

Modelo	Medida	Valor
Chiconaultla	Freqüência	81
	Mínimo	1,795
	Máximo	40,160
	Média	12,704
	Desvio-padrão (amostral)	6,038
DeltaForce3	Freqüência	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

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

Quadro 19: Medidas de síntese de Renda por Modelo
Fonte: elaborado pelo autor

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, o relacionamento 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		57,777%

Quadro 20: Coeficientes de Variação Percentual de Renda por Modelo
Fonte: elaborado pelo autor

Aparentemente, a relação existente entre a 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, LuxuriousCar, cujos compradores têm a média mais alta, não apresenta o menor coeficiente de variação percentual. O modelo cujos 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 a um 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 um software estatístico, podemos calcular outras medidas além das mostradas nos Quadros anteriores. No nosso caso, usando o Statsoft Statistica 6.0®, podemos obter:

Modelo	Medidas							
	$\bar{X}$	Freq.	s	Mín	Máx	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: adaptado pelo autor de Statsoft®

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

Proponho que você faça um exercício para calcular as diferenças entre quartil superior e mediana, e entre mediana e quartil inferior para avaliar se há ou não assimetria (veja as Figuras 36 a 39 para se orientar na análise).

Saiba mais...

- Sobre medidas de síntese, assimetria, diagramas em caixa e outros aspectos, procure em BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006, capítulo 6.

- Sobre outros tipos de média (harmônica, geométrica), SPIEGEL, M. 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, D. R.; SWEENEY, D.J.; WILLIAMS, T.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, D. M.; et al. *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 e na Unidade 4 através 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.

RESUMO

O resumo desta Unidade está demonstrado na Figura 40:

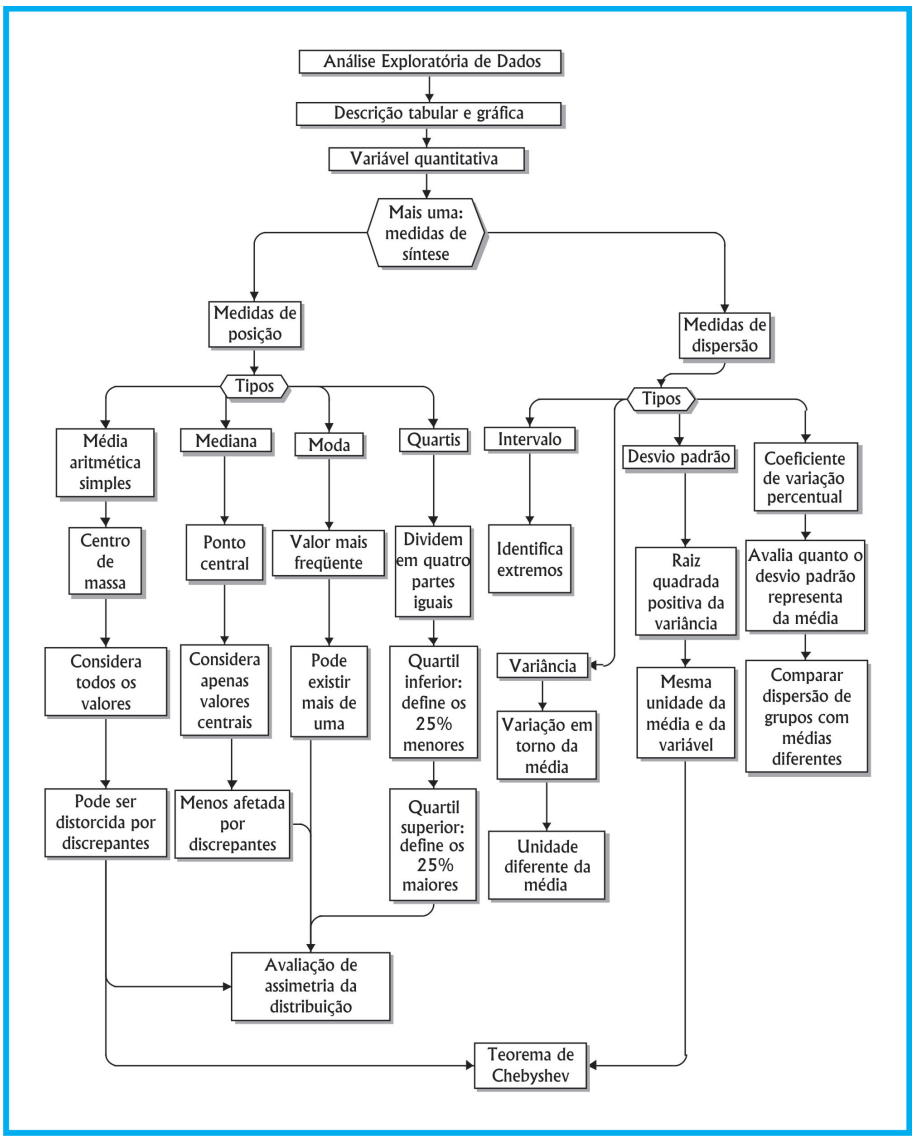


Figura 40: Resumo da Unidade 4
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Com este tópico, finalizamos a Análise Exploratória de Dados. É extremamente importante que você faça todos os exercícios, entre em contato com a tutoria para tirar dúvidas, pois não há outra forma de aprender a não ser praticando. Na Unidade 5, veremos os conceitos de Probabilidade, que são indispensáveis para compreender o processo de inferência (generalização) estatística. Vamos em frente, e ótimos estudos!

UNIDADE

5

Conceitos básicos de Probabilidade

Objetivo

Nesta Unidade, você vai compreender os conceitos de Probabilidade e a importância do uso do raciocínio probabilístico para auxiliar 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, e os administradores diariamente precisam tomar decisões, muitas delas extremamente sérias, em um 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, e surge, então, a Probabilidade como uma das abordagens de tratamento da incerteza.

A utilização de métodos probabilísticos proporciona um 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.

Nas Unidades 3 e 4, foi utilizado um raciocínio predominantemente indutivo. Os dados foram coletados, e, através da sua organização em distribuições de freqüências 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 meninas nascidos em um Estado brasileiro. Consultando dados do IBGE, provenientes de censos e levantamentos anteriores (portanto, distribuições de freqüências da variável qualitativa sexo dos recém-nascidos), há interesse em prever qual será o percentual de nascimentos no ano de 2009: em suma, será usado um raciocínio dedutivo; a partir de algumas suposições sobre o problema (a definição dos resultados possíveis, os percentuais registrados em anos anteriores), tenta-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 um pouco acima ou um 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 uma certa regularidade nos percentuais de nascimentos (verificados anteriormente), mas não é possível responder qual será o resultado exato antes de o fenômeno ocorrer.

A regularidade citada (que foi observada para um grande número de nascimentos) permite que seja calculado o grau de certeza ou confiabilidade da previsão feita, que recebe o nome de **Probabilidade***. Haverá uma 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.

Sem saber, montamos um **modelo probabilístico*** para o problema em questão:

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

GLOSSÁRIO

*Variabilidade – diferenças encontradas por sucessivas medições realizadas em pessoas, animais ou objetos, em tempos ou situações diferentes. Fonte: Montgomery (2002).

*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.

* Modelo probabilístico – modelo matemático para descrever a certeza de ocorrência de eventos, no qual são definidos os eventos possíveis e uma regra de ocorrência para calcular quão provável é cada evento ou conjunto de eventos. Fonte: Barbetta (2006).

O modelo probabilístico permite expressar o grau de incerteza 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, e se este dado não é viciado, espera-se que cada face ocorra em 1/6 do total de lançamentos: se o dado for lançado um grande número de vezes, isso provavelmente ocorrerá, mas um resultado diferente poderia ser obtido sem significar que o dado está viciado, principalmente se forem feitos poucos lançamentos.

Neste ponto, é importante ressaltar que os modelos probabilísticos não têm razão de ser para fenômenos (experimentos) não aleatórios (**determinísticos**): aqueles em que, usando teorias e fórmulas apropriadas, se pode prever exatamente qual será o seu resultado antes de o fenômeno ocorrer, por exemplo, o lançamento de uma pedra de 5 kg de uma altura de dez metros, havendo interesse em cronometrar o tempo para que ela atinja o chão. Conhecendo 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, não há necessidade sequer de realizar o experimento.

Para construir ou utilizar modelos probabilísticos, é necessário que haja um 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%).

Para prosseguirmos, precisamos de algumas definições importantes para estudar os modelos probabilísticos. Precisamos definir exatamente as condições em que devemos usar modelos probabilísticos, e isso exige saber o que são experimento aleatório, espaço amostral e eventos. Vamos ver?

Definições prévias

Experimento aleatório

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

- não se pode afirmar, antes de fazer o experimento, qual será o resultado de uma realização, mas é possível determinar o conjunto de resultados possíveis; e
- quando é realizado um grande número de vezes (replicado), apresentará uma regularidade que permitirá construir um modelo probabilístico para analisar o experimento.

São experimentos aleatórios:

a) o lançamento de um dado e a observação da face voltada para cima; não se sabe exatamente qual face vai ocorrer, apenas que será uma das seis, e que, se o dado for não viciado, e o lançamento, imparcial, todas as faces têm a mesma chance de ocorrer;

b) a observação dos diâmetros, em mm, de eixos produzidos em uma metalúrgica; sabe-se que as medidas devem estar próximas de um valor nominal, mas não se sabe exatamente qual é o diâmetro de cada eixo antes de efetuar as mensurações; e

c) o número de mensagens que são transmitidas corretamente por dia em uma rede de computadores; sabe-se que o mínimo possível é zero, mas não se sabe nem sequer 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:

a) o lançamento de um dado e a observação da face voltada para cima: Ω = {1, 2, 3, 4, 5, 6};

b) a retirada de uma carta de um baralho comum (52 cartas) e a observação do naipe: Ω = {copas, espadas, ouros, paus};

c) o número de mensagens que são transmitidas corretamente por dia em uma rede de computadores: Ω = {0, 1, 2, 3, ...};

d) a observação do diâmetro, em mm, de um eixo produzido em uma metalúrgica: Ω = {D, tal que D > 0}; e

e) as vendas mensais, em unidades, de determinado modelo de veículo: Ω = {0, 1, ...}

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

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

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; ou
- **infinito**, formado por intervalos de números reais, como no caso d.

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

A construção do modelo probabilístico dependerá do tipo de espaço amostral, como será visto mais 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 ocorre! Geralmente, há interesse em calcular a probabilidade de que um determinado evento venha a ocorrer, e este evento pode ser definido de forma verbal, precisando ser "traduzido" para as definições da Teoria de Conjuntos, que veremos a seguir.

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

Sejam o experimento aleatório lançamento de um dado não viciado e 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

$E_3 = \{1, 3\}$

serão apresentadas as definições de **evento união**, **evento intersecção**, **eventos mutuamente exclusivos** e **evento complementar**.

Evento **união** de E_1 com E_2 ($E_1 \cup E_2$): evento que ocorre se E_1 ou E_2 ou ambos ocorrem.

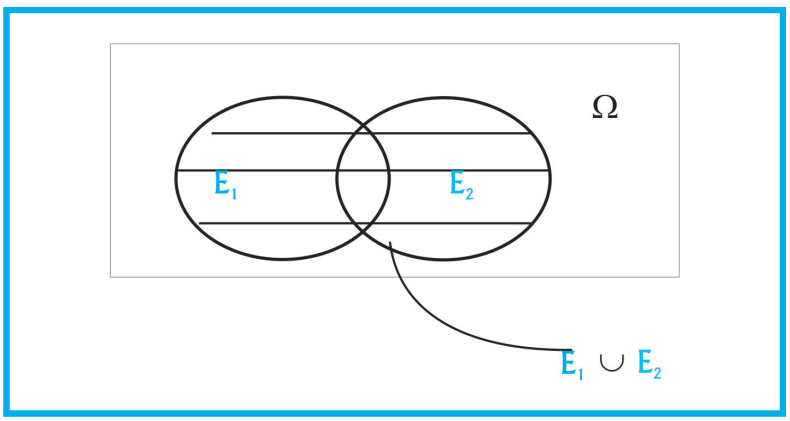


Figura 41: Evento união
Fonte: elaborada pelo autor

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

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.

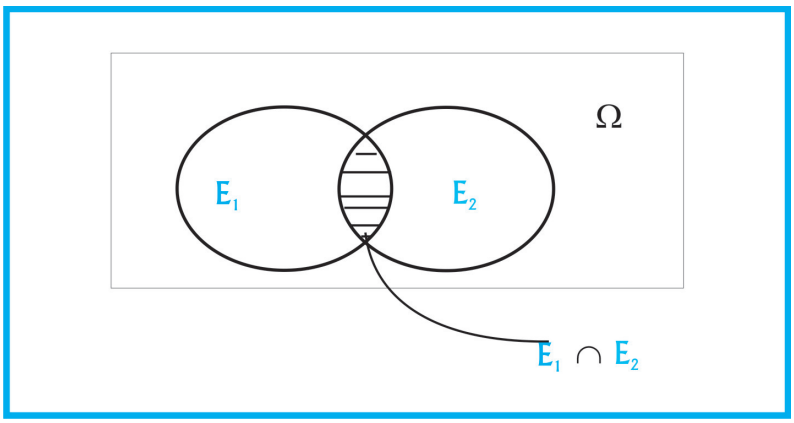


Figura 42: Evento intersecção
Fonte: elaborada pelo autor

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

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

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

> $E_3 = \{1, 3\}$ $E_1 = \{2, 4, 6\}$ $E_1 \cap E_3 = \varnothing$ => 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 são o conjunto vazio.

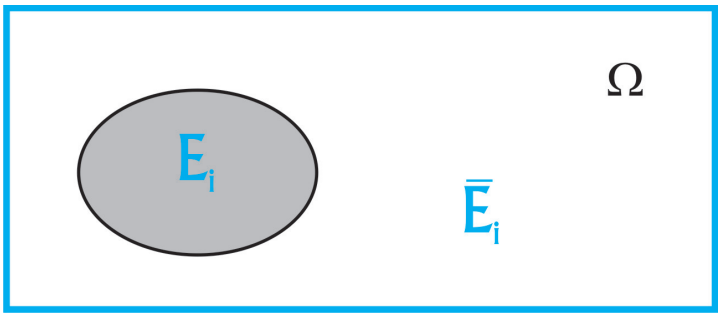


Figura 43: Evento complementar
Fonte: elaborada pelo autor

$E_i \cup \overline{E}_i = \Omega$ $\quad$ $E_i \cap \overline{E}_i = \varnothing$

$E_1 = \{2, 4, 6\}$ $\quad$ $\overline{E}_1 = \{1, 3, 5\}$

$E_2 = \{3, 4, 5, 6\}$ $\quad$ $\overline{E}_2 = \{1, 2\}$

A compreensão das definições anteriores será extremamente útil quando calcularmos probabilidades, pois as expressões poderão ser deduzidas ou simplificadas se identificarmos que se trata de evento união, intersecção, ou se os eventos de interesse são mutuamente exclusivos ou complementares. Conhecido isso, podemos agora passar à definição de probabilidade ou, mais especificamente, às definições de probabilidade, que são complementares.

Definições de probabilidade

Por que usamos plural, “definições”, ao invés de “definição”? Porque ao longo dos séculos várias definições de probabilidade foram apresentadas, e elas se complementam.

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 nestes 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 é a chance de você ganhar?

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.

Intuitivamente, você responderia que há 50% (1/2) de chances de ganhar, uma vez que há apenas duas faces (resultados) possíveis. Mesmo sem saber o que é 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 neste segundo exemplo: qual é 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. Em cada realização do experimento, 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 ocorrer: tanto cara quanto coroa tem 50% de chance de ocorrer, todos os quatro naipes (copas, espadas, ouros e paus) têm 25% de chance de ocorrer. Sem que você soubesse, você 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 $\mathbf{n}_{Ei}$ destes resultados referem-se ao evento $\mathbf{E}_i$, então a probabilidade de o evento $\mathbf{E}_i$ ocorrer será:

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

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 técnicas de análise combinatória (que serão vistas posteriormente) ou por considerações teóricas (“bom senso”).

Seja o seguinte experimento aleatório: lançamento de um dado não viciado e observação da face voltada para cima. Neste Exemplo 3, vamos calcular as probabilidades de ocorrência dos seguintes eventos:

a) face 1;

b) face par; e

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, resultando em $\mathbf{n} = 6$. 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, $\mathbf{n_{Ei}} = 1$, e a probabilidade de ocorrer a face 1 será:

$$P(Ei) = \frac{n_{Ei}}{n} = \frac{1}{6}$$

O evento “face par” tem três resultados associados: {2, 4, 6}. Então, $\mathbf{n_{Ei}} = 3$, e a probabilidade de ocorrer face par será:

$$P(Ei) = \frac{n_{Ei}}{n} = \frac{3}{6} = \frac{1}{2}$$

O evento “face menor ou igual a 2” tem dois resultados associados: {1, 2}. Então, $\mathbf{n_{Ei}} = 2$, e a probabilidade de ocorrência de face menor ou igual a 2 será:

$$P(Ei) = \frac{n_{Ei}}{n} = \frac{2}{6} = \frac{1}{3}$$

Como viu 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 desta 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 e 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 (na França, que tem, em média, 225 dias nublados por ano) é a mesma na cidade de Sevilha (na Espanha, 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 que é repetido **n** vezes, e $\mathbf{E_i}$, um evento associado.

A freqüência relativa do evento $\mathbf{E_i}$: $f_{REi} = \frac{n_{Ei}}{n} = \frac{\text{nº vezes que } E_i \text{ ocorreu}}{\text{total de tentativas}}$

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

Quando não há outra maneira de obter as probabilidades dos eventos, é necessário realizar o experimento (veja novamente a Unidade 1) várias vezes para que seja possível obter um número tal de tentativas que permita que as freqüências relativas estimem as probabilidades, para que se passa 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 através de testes de sobrevivência realizados pelos fabricantes.

Mas, em alguns casos, não é possível realizar experimentos, a maioria dos fenômenos socioeconômicos e climáticos, por exemplo. Neste caso, precisamos estimar as probabilidades através das freqüências relativas históricas.

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

Axiomas e propriedades da probabilidade

Alguns autores chamam estes axiomas e propriedades de definição axiomática da probabilidade.

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

a) $0 \leq \mathbf{P(E_i)} \leq 1{,}0$

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

b) $\mathbf{P\ (\Omega)} = 1{,}0$

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**.

c) Se E_1, E_2, ..., En são eventos mutuamente exclusivos, então $P(E_1 \cup E_2 \cup ... \cup E_n) = P(E_1) + P(E_2) + ... + P(E_n)$

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

GLOSSÁRIO

*Evento impossível – evento com probabilidade de ocorrer igual a 0%, é o conjunto vazio. Fonte: Barbetta, Reis e Bornia (2004).

Além dos axiomas, há algumas propriedades básicas da probabilidade:

a) $P\ (\varnothing) = 0$

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) $\Sigma\ \mathbf{P(E_i)} = 1{,}0$

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) $\mathbf{P(E_i)} = 1 - \mathbf{P(\overline{E}_i)}$

A probabilidade de ocorrência de um evento qualquer será igual à 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).

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

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

Veja, neste quarto exemplo, 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: $\mathbf{E_1}$ = face 1 = {1}, $\mathbf{E_2}$ = face par = {2, 4, 6} e $\mathbf{E_3}$ = face ≤ 2 {1, 2}, cujas probabilidades já foram calculadas.

Calcular a probabilidade de ocorrência dos seguintes eventos:

a) complementar de $\mathbf{E_1}$;

b) complementar de $\mathbf{E_2}$;

c) união de $\mathbf{E_2}$ e $\mathbf{E_3}$; e

d) união de $\mathbf{E_1}$ e $\mathbf{E_2}$.

No Exemplo 2, obtiveram-se $P(\mathbf{E_1}) = 1/6$, $P(\mathbf{E_2}) = 3/6$ e $P(\mathbf{E_3}) = 2/6$.

Usando as propriedades:

$\mathbf{P(E_1)} = 1 - \mathbf{P(\overline{E}_1)}$ então $\mathbf{P(\overline{E}_1)} = 1 - \mathbf{P(E_1)} = 1 - 1/6 = 5/6$
$\overline{\mathbf{E}}_1 = \{2, 3, 4, 5, 6\}$

$\mathbf{P(E_2)} = 1 - \mathbf{P(\overline{E}_2)}$ então $\mathbf{P(\overline{E}_2)} = 1 - \mathbf{P(E_2)} = 1 - 3/6 = 3/6$
$\overline{\mathbf{E}}_2 = \{1, 3, 5\}$

$\mathbf{P(E_2 \cup E_3)} = \mathbf{P(E_2)} + \mathbf{P(E_3)} - \mathbf{P(E_2 \cap E_3)}$ Observe que há apenas um elemento em comum entre os eventos $\mathbf{E_2}$ e $\mathbf{E_3}$: apenas um resultado associado => $\mathbf{P(E_2 \cap E_3)} = 1/6$

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

$\mathbf{P(E_1 \cup E_2)} = \mathbf{P(E_1)} + \mathbf{P(E_2)} - \mathbf{P(E_1 \cap E_2)}$ Não há elementos em comum entre os eventos $\mathbf{E_1}$ e $\mathbf{E_2}$: eles são mutuamente exclusivos, sua

intersecção é o conjunto vazio, e a probabilidade de ocorrência do conjunto vazio é nula. $\mathbf{P(E_1 \cup E_2)} = 1/6 + 3/6 - 0 = 4/6$

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 Petrobrás. Supõe-se que o preço do barril do petróleo subirá cerca de 10% nos próximos dias, há uma probabilidade estimada de isso acontecer. E, sabendo disso, você gostaria de saber qual é a probabilidade de que as ações da empresa subam também 10% na Bovespa. Este caso, em que queremos 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 um outro evento B. Por exemplo, qual é a probabilidade de chover amanhã em Florianópolis, sabendo-se que hoje choveu? Ou qual é 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 é a probabilidade de que as ações da Petrobrás aumentem 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 a sua expressão será:

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

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).

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

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

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

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

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

GLOSSÁRIO

* Operação comutativa – operação em que a seqüência de realização não modifica o resultado, "a ordem dos fatores não altera o produto". Fonte: elaborado pelo autor.

Sejam o lançamento de dois dados não viciados, um após o outro, e a observação das faces voltadas para cima. Neste quinto exemplo, vamos calcular as probabilidades:

- de que as faces sejam iguais, supondo-se que sua soma é menor ou igual a 5; e
- 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 todo problema de probabilidade, é preciso montar o espaço amostral. Neste caso, serão os pares de faces dos dados, e como os dados são lançados um após o outro, a ordem das faces é importante:

$$\Omega = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

Figura 44: Espaço amostral do Exemplo 5
Fonte: elaborada pelo autor

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

- "Faces iguais, sabendo-se que sua soma é menor ou igual a 5" significa dizer probabilidade de ocorrência de faces iguais supondo-se que **já ocorreram** faces cuja soma é menor ou igual a 5; chamando o evento faces iguais de $\mathbf{E_1}$ e o evento soma das faces menor ou igual a 5 de $\mathbf{E_2}$, estamos procurando $P(\mathbf{E_1} \mid \mathbf{E_2})$, probabilidade de ocorrência de $\mathbf{E_1}$ condicionada à ocorrência PRÉVIA de $\mathbf{E_2}$.

Usando a fórmula:

$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$, é preciso encontrar os valores das probabilidades.

Primeiramente, definir 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:

$\mathbf{E_1}$ = {(1,1) (2,2) (3,3) (4,4) (5,5) (6,6)} – faces iguais, 6 resultados, $\mathbf{nE_1}$ = 6.

$\mathbf{E_2}$ = {(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, $\mathbf{nE_2}$ = 10.

Os elementos em comum formarão o evento intersecção: $\mathbf{E_1 \cap E_2}$ = {(1,1) (2,2)} – faces iguais e soma das faces ≤5, 2 resultados, $\mathbf{nE_1 \cap E_2}$ = 2.

$$\mathbf{P(E_2) = nE_2 / n = 10/36} \qquad \mathbf{P(E_1 \cap E_2) = nE_1 \cap E_2 / n = 2/36}$$

Tendo as probabilidades acima, é possível calcular a probabilidade condicional:

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

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

Este resultado poderia ser obtido de outra forma. Se a soma das faces é menor ou igual a 5, o evento $\mathbf{E_2}$ já ocorreu previamente, então o espaço amostral **modificou-se**, passando a ser o conjunto de resultados do evento $\mathbf{E_2}$:

novo Ω = {(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 dez resultados, novo **n** = 10.

O número de resultados do evento "faces iguais ($\mathbf{E_1}$) no novo espaço amostral" é igual a 2, novo n$\mathbf{E_1}$ = 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 $\mathbf{E_1}$ no novo espaço amostral, ou seja, a probabilidade de ocorrência do evento $\mathbf{E_1}$ **condicionada** à ocorrência prévia do evento $\mathbf{E_2}$, P($\mathbf{E_1}$ | $\mathbf{E_2}$), será:

> P($\mathbf{E_1}$ | $\mathbf{E_2}$) = novo **n**$\mathbf{E_1}$/ novo **n** = 2/10 = 0,2 (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 $\mathbf{E_1}$ e o evento soma das faces menor ou igual a 5 de $\mathbf{E_2}$, estamos procurando P($\mathbf{E_2}$ | $\mathbf{E_1}$), probabilidade de ocorrência de $\mathbf{E_2}$ condicionada à ocorrência PRÉVIA de $\mathbf{E_1}$.

Houve uma mudança no evento que ocorreu previamente.

Usando 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 tenham soma menor ou igual a 5, sabendo-se 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 $\mathbf{E_1}$ já ocorreu previamente, então o espaço amostral **modificou-se**, passando a ser o conjunto de resultados do evento $\mathbf{E_1}$:

novo Ω = { (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 ($\mathbf{E_2}$)" no novo espaço amostral é igual a 2, novo n$\mathbf{E_2}$ = 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 $\mathbf{E_2}$ no novo espaço amostral, ou seja, a probabilidade de ocorrência do evento $\mathbf{E_2}$ **condicionada** à ocorrência prévia do evento $\mathbf{E_1}$, $P(\mathbf{E_2} \mid \mathbf{E_1})$, será:

$P(\mathbf{E_2} \mid \mathbf{E_1})$ = novo $\mathbf{nE_2}$/ novo $\mathbf{n}$ = 2/6 = 0,33 (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 freqüências de dados históricos ou experimentais para obtê-las.

Regra do produto

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

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

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

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

Neste 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 seqüência dos eventos para montar as expressões acima: analisar corretamente que evento já ocorreu.

No Exemplo 6, digamos que uma urna contém duas bolas brancas e três vermelhas. Retiram-se duas bolas ao acaso, uma após a outra. Veremos nos itens abaixo se a retirada foi feita **sem reposição**.

a) Qual é a probabilidade de que as duas bolas retiradas sejam da mesma cor?

b) Qual é a probabilidade de que as duas bolas retiradas sejam 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 podemos 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 $E_3 = V_1 \cap B_2$;
- 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: em 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 (duas bolas brancas no total de cinco), $P(B_1) = 2/5$; e
- a probabilidade de retirar bola vermelha na 1ª retirada é de 3/5 (três bolas vermelhas em cinco), $P(V_1) = 3/5$.

Se a primeira bola retirada foi branca (o evento B_1 ocorreu previamente), restaram quatro bolas, uma branca e três vermelhas:

Repare que o número de bolas, número de resultados, diminuiu de cinco para quatro, porque as retiradas são feitas sem reposição.

- a probabilidade de retirar uma bola branca na 2ª retirada se na 1ª foi extraída uma branca, é de 1/4 (uma bola branca em quatro), $P(B_2 | B_1) = ¼$; e
- a probabilidade de retirar uma bola vermelha na 2ª retirada, se na 1ª foi extraída uma branca é de 3/4 (três bolas vermelhas em quatro), $P(V_2 | B_1) = 3/4$.

Se a primeira bola retirada foi vermelha (o evento V_1 ocorreu previamente), restaram quatro bolas, duas brancas e duas vermelhas:

- a probabilidade de retirar uma bola branca na 2ª retirada, se na 1ª foi extraída uma vermelha, é de 2/4 (duas bolas brancas em quatro), $P(B_2 | V_1) = 2/4$; e
- a probabilidade de retirar uma bola vermelha na 2ª retirada, se na 1ª foi extraída uma vermelha, é de 2/4 (duas bolas vermelhas em quatro), $P(V_2 | V_1) = 2/4$.

a) O evento que nos interessa: “bolas da mesma cor”: brancas ou vermelhas, evento união brancas-vermelhas.

Chamando bolas da mesma cor de evento F:

$$F = [(B_1 \cap B_2) \cup (V_1 \cap V_2)]$$

Usando as propriedades da 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 de o conjunto vazio ocorrer é igual a zero (ver seção 5.3.3); então, simplesmente:

$$P(F) = P\,[(B_1 \cap B_2) \cup (V_1 \cap V_2)]= P(B_1 \cap B_2) + P(V_1 \cap V_2)$$

Usando a regra do produto:

$P(B_1 \cap B_2) = P(B_1) \times P(B_2 \mid B_1) = (2/5) \times (1/4) = 2/20 = 1/10$
$P(V_1 \cap V_2) = P(V_1) \times P(V_2 \mid V_1) = (3/5) \times (2/4) = 6/20 = 3/10$

Substituindo na expressão:

$P(F) = P\,[(B_1 \cap B_2) \cup (V_1 \cap V_2)] = P(B_1 \cap B_2) + P(V_1 \cap V_2) = 1/10 + 3/10 = 4/10 = 0{,}4\ (40\%)$

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, sabe-se que as duas bolas são da mesma cor (o evento F acima JÁ OCORREU), e há interesse em saber a probabilidade de que as duas bolas sejam vermelhas:

$$\mathbf{P[(V_1 \cap V_2) \mid F] = P\{(V_1 \cap V_2) \mid [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\}}$$

Usando a expressão de probabilidade condicional:

$$\mathbf{P\{(V_1 \cap V_2) \mid [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\}} = \frac{P\{(V_1 \cap V_2) \cap [(B_1 \cap B_2) \cup (V_1 \cap V_2)]\}}{P[(B_1 \cap B_2) \cup (V_1 \cap V_2)]}$$

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 $\mathbf{[(B_1 \cap B_2) \cup (V_1 \cap V_2)]}$, 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:

$(V_1 \cap V_2) \cap \mathbf{[(B_1 \cap B_2) \cup (V_1 \cap V_2)]} = (V_1 \cap V_2)$;
$P\{(V_1 \cap V_2) \cap \mathbf{[(B_1 \cap B_2) \cup (V_1 \cap V_2)]}\} = P\,(V_1 \cap V_2) = 3/10$.

Sabendo que $P\{(V_1 \cap V_2) \mid \mathbf{[(B_1 \cap B_2) \cup (V_1 \cap V_2)]}\} = 4/10$ (do item a.1) e substituindo os valores na fórmula:

$$P\{(V_1 \cap V_2) \mid [(\mathbf{B_1 \cap B_2}) \cup (\mathbf{V_1 \cap V_2})]\} = \frac{P(V_1 \cap V_2)}{P[(B_1 \cap B_2) \cup (V_1 \cap V_2)]} = \frac{3/10}{4/10} = \frac{3}{4}$$

$P\{(V_1 \cap V_2) \mid [(\mathbf{B_1 \cap B_2}) \cup (\mathbf{V_1 \cap V_2})]\} = 0{,}75\ (75\%)$

Então, se as retiradas forem feitas sem reposição, e as duas bolas forem da mesma cor, a probabilidade de que sejam vermelhas será igual a 0,75 (75%).

As retiradas e as probabilidades podem ser representadas através de um diagrama chamado de "Árvore de probabilidades":

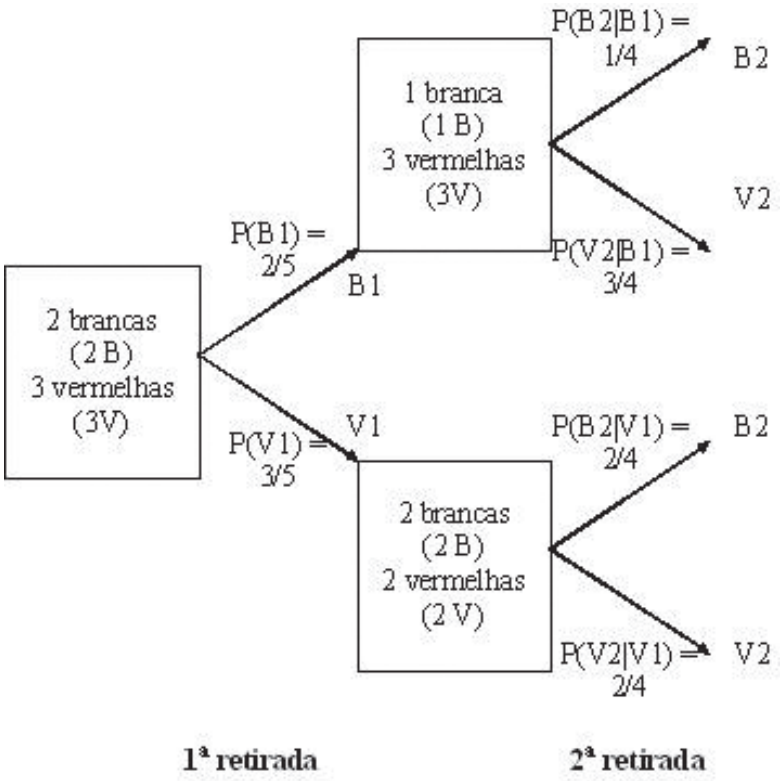


Figura 44: Árvore de probabilidades – Retiradas sem reposição
Fonte: elaborada pelo autor

Observe que, através da árvore de probabilidades, podemos chegar aos mesmos resultados obtidos anteriormente. Partindo 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: uma bola branca e três vermelhas, se na 1ª retirada obteve-se uma bola branca, ou duas bolas brancas e duas 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: neste 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 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 todos os resultados. Nestes casos, usa-se 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 dos eventos 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:

$P(A \mid B) = P(A)$ e $P(B \mid A) = P(B)$

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

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

*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: para a mesma situação do Exemplo 6. Uma urna contém duas bolas brancas e três vermelhas. Retiram-se duas bolas ao acaso, uma após a outra. Resolva os itens abaixo, se a retirada for feita **com reposição**.

a) Qual é a probabilidade de que as duas bolas retiradas sejam da mesma cor? R.: 0,52(52%).

b) Qual é a probabilidade de que as duas bolas retiradas sejam vermelhas, supondo-se que são da mesma cor? R.: 0,69 (69%).

A seguir, veremos como resolver um problema quando for inviável usar a árvore de probabilidades.

Probabilidade combinatória

Como já vimos, em muitos casos a resolução dos problemas de probabilidade enumerando todos os resultados possíveis torna-se extremamente difícil. Há uma forma mais rápida de enumerar os resultados: as técnicas de análise combinatória.

Relembremos a definição clássica de probabilidade, que consistia em calcular o quociente entre o número de resultados associados ao evento e o número total de resultados possíveis. O cálculo desses números de resultados pode ser feito utilizando análise combinatória, tanto para os casos em que os eventos são dependentes quanto quando há independência.

As técnicas de análise combinatória buscam basicamente calcular o número de maneiras de dispor um certo número de "objetos" em um número limitado de "espaços" distintos (menor do que o número de objetos), sendo um objeto em cada espaço. Se o número de "objetos" é, teoricamente, infinito (ou ilimitado), temos a análise combinatória com repetição ilimitada (situação de independência): é o que ocorre nos casos em que há reposição. Se, porém, o número de "objetos" é limitado, temos a análise combinatória sem repetição (situação de dependência): casos em que não há reposição.

Análise combinatória com reposição ilimitada

Há **n** objetos disponíveis em número ilimitado; em outras palavras, há reposição, de quantas maneiras diferentes é possível preencher **k** espaços distintos com os objetos, cada espaço com um objeto?

Tem-se um espaço e **n** objetos, há **n** maneiras de dispô-los no espaço. Tem-se dois espaços, e os mesmos **n** objetos disponíveis para cada um, haverá $\mathbf{n^2}$ maneiras: as **n** maneiras do primeiro espaço multiplicadas pelas **n** maneiras do segundo. Se houver três espaços, haverá $\mathbf{n^3}$ maneiras, e assim por diante.

Generalizando, se há **n** objetos disponíveis em número **ilimitado** para preencher **k** espaços distintos, cada espaço com um objeto, há $\mathbf{n^k}$ maneiras de fazê-lo, e cada preenchimento é independente dos outros.

Veremos neste oitavo exemplo quantas palavras de cinco letras podem ser escritas com as 26 letras do alfabeto, sem se preocupar com o significado.

Veja que, primeiramente, é preciso identificar os objetos e os espaços.

Os objetos, neste caso, são as letras do alfabeto, 26, então **n** = 26. Como não há preocupação com o significado das palavras, os objetos estão disponíveis em número ilimitado.

Os espaços são as letras da palavra: cada palavra deve ter cinco letras, então **k = 5**.

Usando a expressão de análise combinatória com repetição ilimitada, o número de palavras será: $\mathbf{n^k} = 26^5 = 11.881.376$ palavras.

Uma urna contém duas bolas brancas e três vermelhas. Retiram-se ao acaso, uma após a outra, com reposição. Qual a probabilidade de que as duas bolas sejam da mesma cor? Neste nono exemplo, utilizaremos a análise combinatória.

Este exemplo é uma repetição do item a do Exemplo 7. Aqui chegaremos ao mesmo resultado usando análise combinatória.

O evento de nosso interesse: bolas da mesma cor = F = $[(B_1 \cap B_2) \cup (V_1 \cap V_2)]$. Vimos que os eventos $(B_1 \cap B_2)$ e $(V_1 \cap V_2)$ são mutuamente exclusivos, então:

$$P(F) = P[(B_1 \cap B_2) \cup (V_1 \cap V_2)] = P(B_1 \cap B_2) + P(V_1 \cap V_2)$$

Vamos calcular, então, as probabilidades necessárias.

$P(B_1 \cap B_2)$ = (Nº de resultados para duas bolas brancas) / (Nº total de resultados)

$P(V_1 \cap V_2)$ = (Nº de resultados para duas bolas vermelhas) / (Nº total de resultados)

Os denominadores serão os mesmos para os dois quocientes: há um total de cinco bolas ("objetos") disponíveis em número ilimitado (porque há reposição) para extrair em duas retiradas ("espaços"), resultando **n** = 5 e **k** = 2, então:

$$\text{Nº total de resultados} = \mathbf{n^k} = 5^2 = 25$$

Nº de resultados para duas bolas brancas: há um total de duas bolas brancas ("objetos") disponíveis em número ilimitado (porque há reposição) para extrair em duas retiradas ("espaços"), resultando **n** = 2 e **k** = 2, então:

$$\text{Nº total de resultados} = \mathbf{n^k} = 2^2 = 4$$

Nº de resultados para duas bolas vermelhas: há um total de três bolas vermelhas ("objetos") disponíveis em número ilimitado (porque há reposição) para extrair em duas retiradas ("espaços"), resultando **n** = 3 e **k** = 2, então:

Nº total de resultados = $\mathbf{n^k} = 3^2 = 9$

Então:

$P(B_1 \cap B_2)$ = (Nº de resultados para duas bolas brancas) / (Nº total de resultados) = 4 /25

$P(V_1 \cap V_2)$ = (Nº de resultados para duas bolas vermelhas) / (Nº total de resultados) = 9/25

Substituindo na fórmula:

$P(F) = P[(B_1 \cap B_2) \cup (V_1 \cap V_2)] = P(B_1 \cap B_2) + P(V_1 \cap V_2) = 4/25 + 9/25 = 13/25 = 0{,}52$ (52%)

Então, se as retiradas forem feitas com reposição, a probabilidade de que as duas bolas sejam da mesma cor será igual a 0,52 (52%). Observe que é exatamente o mesmo resultado obtido no Exemplo 7. Claro que, para este caso extremamente simples (apenas duas retiradas com dois resultados possíveis em cada uma), o uso de análise combinatória não é necessário, mas permite chegar aos mesmos resultados que seriam obtidos com as técnicas anteriores. Se, porém, houver muitas retiradas e/ou muitas opções, se tornará indispensável.

Análise combinatória sem reposição

Continua havendo **n** objetos para colocar em **k** espaços, mas os objetos não estão mais disponíveis em número ilimitado: não há repetição ou não há reposição. A seleção de um dos objetos modifica a probabilidade de seleção dos outros: há dependência. Para calcular o número de maneiras possíveis de preencher os espaços, é preciso relembrar os conceitos de arranjos e combinações.

Os arranjos são utilizados para calcular o número de maneiras de dispor os **n** objetos nos **k** espaços, quando a **ordem** e a **natureza** dos objetos são importantes para o problema. O número de arranjos de **n** objetos distintos tomados **k** a **k** será:

$A_{n,k} = \frac{n!}{(n-k)!}$ n! significa fatorial de n: n x (n-1) x (n-2) x....x 1; lembrando que 0! = 1.

Cinco carros, disputando os três primeiros lugares em uma corrida. Há quantas maneiras diferentes de classificá-los? Vejamos no Exemplo 10.

Observe que há cinco objetos a dispor em três espaços, então **n** = 5 e **k** = 3. Os objetos não estão disponíveis em número ilimitado: uma vez definido o primeiro colocado, ele não pode simultaneamente ocupar a terceira posição. Outro aspecto importante é que importam tanto a **ordem** quanto a **natureza** dos objetos: há diferença se o corredor A não chegar entre os três primeiros, mas também há diferença se o corredor chegar em primeiro ou segundo. Sendo assim, serão usados arranjos.

$$A_{n,k} = \frac{n!}{(n-k)!} = \frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times 2!}{2!} = 60 \text{ maneiras.}$$

Então, há 60 maneiras de classificar os cinco carros nos três primeiros lugares.

As combinações são utilizadas para calcular o número de maneiras de dispor os **n** objetos nos **k** espaços, quando apenas a natureza dos objetos é importante para o problema. O número de combinações de **n** objetos distintos tomados **k** a **k** será:

$$C_{n,k} = \frac{n!}{k! \times (n-k)!}$$

Vamos ver, no exemplo a seguir (Exemplo 11), de quantas maneiras diferentes podemos selecionar três dentre cinco pessoas para uma tarefa?

Observe que, novamente, há cinco objetos a dispor em três espaços, então **n** = 5 e **k** = 3. Os objetos não estão disponíveis em número ilimitado: uma vez que uma pessoa seja selecionada, não poderá novamente ser escolhida. Neste caso, importa apenas a natureza dos objetos, apenas definir as pessoas que serão selecionadas. Sendo assim, serão usadas combinações.

$$C_{n,k} = \frac{n!}{k! \times (n-k)!} = \frac{5!}{3! \times (5-3)!} = \frac{5 \times 4 \times 3!}{3! \times 2!} = 10 \text{ maneiras.}$$

Então, há dez maneiras de selecionar três dentre cinco pessoas.

Exemplo 12: uma urna contém 18 bolas brancas, 15 vermelhas e dez azuis. Serão retiradas X bolas, sem reposição, e observadas suas cores.

a) Seja X = 8 (oito bolas). Qual a probabilidade de que as bolas sejam da mesma cor?

b) Seja X = 6 (seis bolas). Qual é a probabilidade de que duas sejam brancas, duas sejam vermelhas e duas sejam azuis?

Este problema seria extremamente trabalhoso para resolver usando uma árvore de probabilidades, por possuir várias retiradas com três resultados cada. Observe que não há reposição, portanto deve-se usar análise combinatória sem repetição: repare que não há interesse na ordem das bolas retiradas (tanto no item a quanto no item b), mas apenas na cor das bolas (na sua "natureza"), sendo assim devem-se usar combinações para calcular o número de resultados necessários para calcular as probabilidades.

a) Há uma grande quantidade de resultados possíveis para este problema, deve-se identificar o evento de interesse: oito bolas da mesma cor. Neste caso, oito bolas brancas, ou oito bolas vermelhas ou oito bolas azuis, evento união oito brancas com oito vermelhas com oito azuis. Chamando o evento oito bolas da mesma cor de F:

F = (8 brancas $\cup$ 8 vermelhas $\cup$ 8 azuis).

Observe que os três eventos acima são mutuamente exclusivos: as oito bolas retiradas não podem ser brancas e azuis simultaneamente. Então:

P(F) = P (8 brancas $\cup$ 8 vermelhas $\cup$ 8 azuis) = P(8 brancas) + P(8 vermelhas) + P(8 azuis)

Para calcular as probabilidades dos eventos, pode-se usar a definição clássica de probabilidade:

P(8 brancas) = (Nº resultados para 8 brancas) / (Nº total de resultados)

P(8 vermelhas) = (Nº resultados para 8 vermelhas) / (Nº total de resultados)

P(8 azuis) = (Nº resultados para 8 azuis) / (Nº total de resultados)

O denominador será o mesmo para todas as expressões. Há um total de 43 bolas (43 objetos, **n** = 43) para colocar em oito espaços (8 retiradas, **k** = 8), usando combinações:

$$\text{Nº total de resultados} = C_{n,k} = \frac{n!}{k!\times(n-k)!} = \frac{43!}{8!\times(43-8)!} = 145008513$$

Para as bolas brancas. Há 18 bolas brancas (18 objetos, **n** = 18) para colocar em oito espaços (8 retiradas, **k** = 8), usando combinações:

$$\text{Nº de resultados para oito brancas} = C_{n,k} = \frac{n!}{k\times(n-k)!} = \frac{18!}{8!\times(18-8)!} = 43758$$

Para as bolas vermelhas. Há 15 bolas vermelhas (15 objetos, **n** = 15) para colocar em oito espaços (8 retiradas, **k** = 8), usando combinações:

$$\text{Nº de resultados para oito vermelhas} = C_{n,k} = \frac{n!}{k!\times(n-k)!} = \frac{15!}{8!\times(15-8)!} = 6435$$

Para as bolas azuis. Há dez bolas azuis (10 objetos, **n** = 10) para colocar em oito espaços (8 retiradas, **k** = 8), usando combinações:

$$\text{Nº de resultados para oito azuis} = C_{n,k} = \frac{n!}{k!(n-k)!} = \frac{10!}{8!\times(10-8)!} = 45$$

Substituindo os valores diretamente na fórmula geral:

P(F) = P(8 brancas) + P(8 vermelhas) + P(8 azuis)

$$P(F) = \frac{43758}{145008513} + \frac{6435}{145008513} + \frac{45}{145008513} = 0{,}000346$$

Este valor tão baixo era esperado, devido à quantidade de bolas e ao número total de combinações possíveis.

Arredondando, a probabilidade de que as oito bolas retiradas sejam da mesma cor é igual a 0,0003 (0,03%).

b) Neste caso, há interesse em calcular a probabilidade de que duas bolas sejam brancas, e duas sejam vermelhas e duas sejam azuis, evento intersecção duas brancas com duas vermelhas com duas azuis. Chamando este evento de G: G = (2 brancas ∩ 2 vermelhas ∩ 2 azuis).

Para os casos de intersecção, o cálculo do número de resultados associados precisa ser feito da seguinte forma: os números de resultados possíveis associados a cada "subevento" componente devem ser multiplicados para obter o número de resultados da intersecção.

Saiba que isso, porém, não significa que os eventos sejam independentes!

P(G) = (Nº res. 2 brancas x Nº res. 2 vermelhas x Nº res. 2 azuis)/ (Nº total de resultados)

Há um total de 43 bolas (43 objetos, **n** = 43) para colocar em seis espaços (6 retiradas, **k** = 6), usando combinações:

$$\text{Nº de resultados} = C_{n,k} = \frac{n!}{k!\times(n-k)!} = \frac{43!}{6!\times(43-6)!} = 6096454$$

Nº de res. duas brancas: há 18 bolas brancas (18 objetos, **n** = 18) para colocar em dois espaços (2 retiradas, **k** =2), usando combinações:

$$\text{Nº de resultados para duas brancas} = C_{n,k} = \frac{n!}{k!\times(n-k)!} = \frac{18!}{2!\times(18-2)!} = 153$$

Nº de res. duas vermelhas: há 15 bolas vermelhas (15 objetos, **n** = 15) para colocar em dois espaços (2 retiradas, **k** =2), usando combinações:

$$\text{Nº de resultados para duas vermelhas} = C_{n,k} = \frac{n!}{k!\times(n-k)!} = \frac{15!}{2!\times(15-2)!} = 105$$

Nº de res. dois azuis: há dez bolas azuis (10 objetos, **n** = 10) para colocar em dois espaços (2 retiradas, **k** =2), usando combinações:

$$\text{Nº de resultados para duas azuis} = C_{n,k} = \frac{n!}{k!\times(n-k)!} = \frac{10!}{2!\times(10-2)!} = 45$$

Substituindo na fórmula de P(G):

P(G) = (153 x 105 x 45)/ (6.096.454) = 0,11858

Arredondando, a probabilidade de que duas bolas sejam brancas, e duas vermelhas e duas azuis é igual a 0,12 (12%).

Saiba mais...

- Sobre conceitos básicos de Probabilidade, BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006, capítulo 7.
- Também sobre conceitos básicos de Probabilidade STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 3.
- LOPES, P. A. *Probabilidades e Estatística*. Rio de Janeiro: Reichmann e Affonso Editores, 1999, capítulo 3.

RESUMO

O resumo desta Unidade está mostrado na Figuras 45:

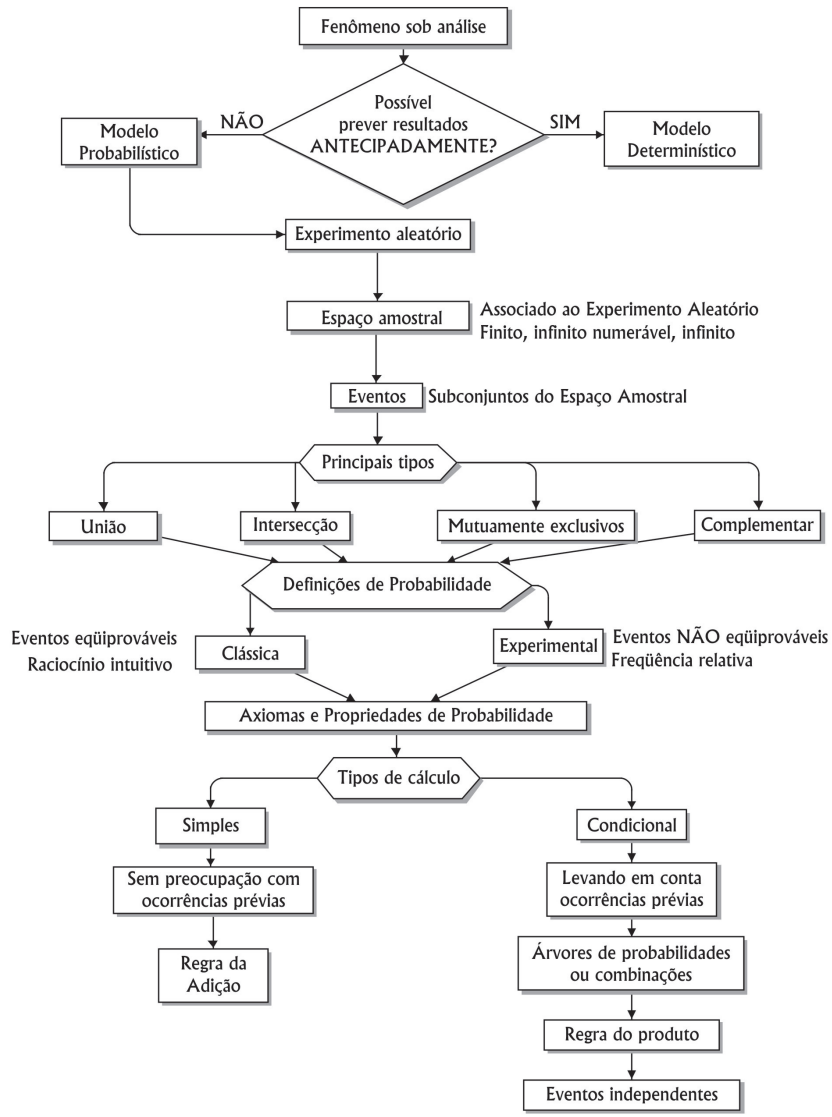


Figura 46: Resumo da Unidade 5
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Chegamos ao final de Unidade 5. Esperamos que você tenha aprendido todos os conceitos trabalhados e, com os exemplo propostos, tenha colocado em prática as informações adquiridas. Neles propomos que você reconhecesse os modelos probabilísticos, modelos determinísticos, principais tipos de evento e os diferentes tipos de cálculo. Na Unidade 6, vamos prosseguir aprendendo o conceito de variável aleatória, que será indispensável para as Unidades 7, 8 e 9. Veremos, ainda, nas Unidades seguintes, a expansão do estudo para o conceito de variável aleatória e alguns dos modelos probabilísticos mais empregados. Tudo isso para chegarmos às Unidades 8 e 9, nas quais aplicaremos os conceitos de probabilidade no processo de inferência estatística, conforme já foi dito na Unidade 1.

Não desanime, caso tenha ficado alguma dúvida. Estamos com você sempre! Interaja, solicite auxílio e, caso necessário, releia o material. Realize a atividade de aprendizagem e entenda todo o processo amplamente.

Ótimos estudos!

UNIDADE

6

Variáveis aleatórias

Objetivo

Nesta Unidade, você vai conhecer e compreender o conceito de variável aleatória e seu relacionamento com os modelos probabilísticos. Vai aprender a interpretar também que estes modelos podem ser construídos para as variáveis aleatórias.

Conceito de variável aleatória

Caro estudante!

Uma pergunta que é normalmente feita a todos que trabalham com Ciências Exatas: "por que a obsessão em reduzir tudo a números"?. Vimos em Análise Exploratória de Dados que uma variável quantitativa, geralmente – porque nem tudo pode ser reduzido a números, como a inteligência e a criatividade – apresenta mais informação que uma variável qualitativa, pode ser resumida não somente através de tabelas e gráficos, mas também através de medidas de síntese.

Nos exemplos sobre probabilidade apresentados na Unidade 5, os eventos foram geralmente definidos de forma verbal: bolas da mesma cor, duas 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, através de uma função.

Esta função é chamada de variável aleatória. Os modelos probabilísticos podem, então, ser construídos para as variáveis aleatórias. O administrador precisa conhecer estes conceitos, porque eles proporcionam maior objetividade na obtenção das probabilidades, o que torna o processo de tomada de decisões mais seguro. Vamos conhecer esses conceitos nesta Unidade?

Uma definição inicial de variável aleatória poderia ser: 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 um experimento aleatório. Se o espaço amostral for finito ou infinito

GLOSSÁRIO

*Espaço amostral – é o conjunto de todos os resultados possíveis de um experimento aleatório. Fonte: Barbetta, Reis e Bornia (2004).

numerável, a variável aleatória é dita discreta. Se o espaço amostral for infinito, a variável aleatória é dita contínua.

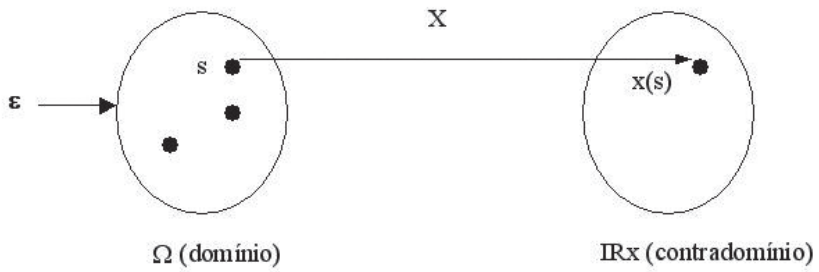


Figura 46: Variável aleatória
Fonte: elaborada pelo autor

Por exemplo, imaginemos o **experimento aleatório*** jogar uma moeda honesta duas vezes e observar a face voltada para cima. O espaço amostral seria finito:

Ω = {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***.

GLOSSÁRIO

*Experimento aleatório – é um 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 pelo autor 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 (2004).

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 duas 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 um determinado site num certo 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 2: 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

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, se você quiser conferir os resultados, some as probabilidades; se elas não somarem 1, há algo 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 tiver errado a anterior, 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 abaixo (onde A é acertar a cobrança, e E significa errar).

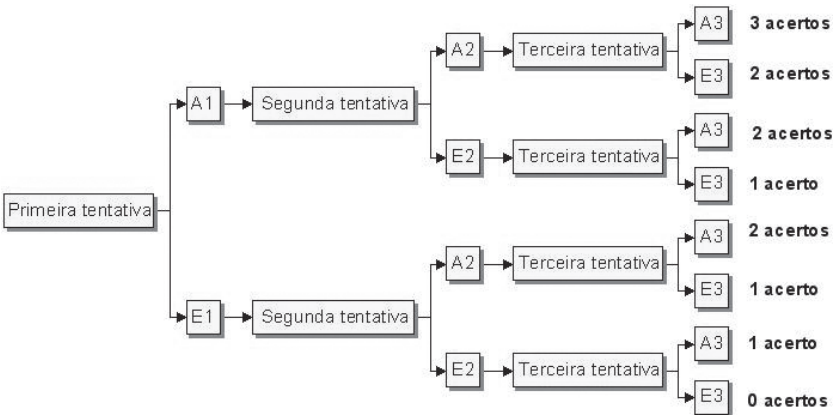


Figura 47: Árvore de eventos
Fonte: elaborada pelo autor

Observe que todos os eventos são mutuamente exclusivos, o jogador não pode, na mesma seqüência de três 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 abaixo:

$$X = 0 \Leftrightarrow [E_1 \cap E_2 \cap E_3]$$
$$X = 1 \Leftrightarrow [(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$$
$$X = 2 \Leftrightarrow [(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap A_3) \cup (A_1 \cap E_2 \cap A_3)]$$
$$X = 3 \Leftrightarrow [A_1 \cap A_2 \cap A_3]$$

Então:

$$P(X=0) = P[E_1 \cap E_2 \cap E_3]$$
$$P(X=1) = P[(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$$
$$P(X=2) = P[(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap A_3) \cup (A_1 \cap E_2 \cap A_3)]$$
$$P(X=3) = [A_1 \cap A_2 \cap A_3]$$

Assume-se que na primeira tentativa o jogador tem 50% de chance de acertar, então:

E_1 (errar a primeira cobrança) é o evento complementar de A_1 (acertar a primeira cobrança).

$$P(A_1) = 0{,}5 \text{ e } P(E_1) = 0{,}5$$

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 estas probabilidades estabelecidas, lembrando da regra do produto e considerando o fato de que os eventos são mutuamente ex-

clusivos, 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:

$$P(X=0) = P(E_1) \text{ x } P(E_2|\ E_1) \text{ x } P(E_3|\ E_2) = 0,5 \times 0,7 \times 0,7 = 0,245\ (24,5\%)$$

Estendendo o procedimento acima para os outros valores:

$$P(X=1) = P[(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$$

$$P(X=2) = P[(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (A_1 \cap E_2 \cap A_3)]$$

$$P(X=3) = P[A_1 \cap A_2 \cap A_3]$$

Como os eventos são mutuamente exclusivos:

$$P(X=1) = P(A_1 \cap E_2 \cap E_3) + P(E_1 \cap A_2 \cap E_3) + P(_1 \cap E_2 \cap A_3)$$

$$P(X=1) = P(A_1) \times P(E_2|A_1) \times P(E_3|E_2) + P(E_1) \times P(A_2|E_1) \times P(E_3|A_2) + P(E_1) \times P(E_2|E_1) \times P(A_3|E_2)$$

$$P(X=1) = 0,5 \times 0,4 \times 0,7 + 0,5 \times 0,3 \times 0,4 + 0,5 \times 0,7 \times 0,3 = 0,305$$

$$P(X=2) = P(A_1 \cap A_2 \cap E_3) + P(E_1 \cap A_2 \cap E_3) + P(A_1 \cap E_2 \cap A_3)$$

$$P(X=2) = P(A_1) \times P(A_2|A_1) \times P(E_3|A_2) + P(E_1) \times P(A_2|E_1) \times P(A_3|A_2) + P(A_1) \times P(E_2|A_1) \times P(A_3|E_2)$$

$$P(X=2) = 0,5 \times 0,6 \times 0,4 + 0,5 \times 0,3 \times 0,6 + 0,5 \times 0,4 \times 0,3 = 0,27\ (27\%)$$

$$P(X=3) = P[A_1 \cap A_2 \cap A_3] = P(A_1) \times P(A_2|A_1) \times P(A_3|A_2) = 0,5 \times 0,6 \times 0,6 = 0,18\ (18\%)$$

Com os valores calculados acima, é possível construir a Tabela 3 com os pares valores-probabilidades.

Tabela 3: Distribuição de probabilidades: número de acertos em três cobranças

$X = x_i$	$p(X = x_i)$
0	0,245
1	0,305
2	0,270
3	0,180
Total	1,0

Fonte: elaborada pelo autor

Ao longo dos séculos, matemáticos e estatísticos deduziram modelos matemáticos para tornar mais simples a obtenção de distribuição de probabilidades para uma variável aleatória discreta. Alguns destes modelos serão vistos na Unidade 7.

Vamos agora passar para a análise das variáveis aleatórias contínuas.

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 uma pesquisa;
- demanda por um produto em um mês; e
- tempo de vida de uma lâmpada incandescente.

Uma variável aleatória contínua está associada a um espaço amostral infinito. Assim, a probabilidade de que a variável assuma

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 ≥ 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:

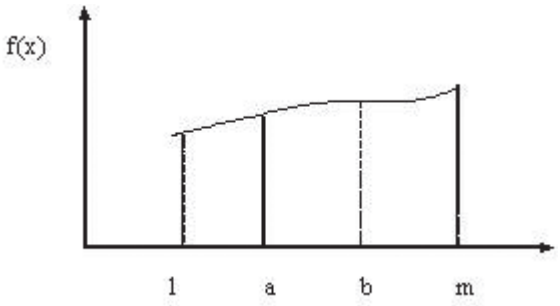


Figura 48: Função densidade de probabilidades
Fonte: elaborada pelo autor

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 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 estão dispostas em tabelas, que são encontradas em praticamente todos os livros de Estatística, e que serão vistas na Unidade 7.

Agora, vamos ver 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), utilizando probabilidades ao invés de freqüências no cálculo. Analogamente, a variância (simbolizada por V(X)) é a variância vista anteriormente, utilizando probabilidades. Da mesma forma que em Análise Exploratória de Dados, é também comum trabalhar 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 recordando que se trata de uma variável aleatória e estão sendo usadas probabilidades, e não freqüências.

Para uma variável aleatória discreta, o valor esperado e a variância podem ser calculados da seguinte forma:

$$E(X) = \sum_{i=1}^{n} x_i \times p(x_i) \qquad V(X) = E\left(X^2\right) - [E(X)]^2, \text{ onde } E\left(X^2\right) = \sum_{i=1}^{n} x_i^2 \times p(x_i)$$

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 na comparação de propostas. Suponha que os valores de uma variável aleatória sejam lucros ou prejuízos, advindos de decisões tomadas, por exemplo, decidir por uma proposta de compra do cliente A ou do cliente B. Associados aos valores, há probabilidades; como decidir qual é 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 especificamen-

te 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. 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

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 por 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.

Agora, vamos ver 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 ao Quadro 2 com os valores e probabilidades, para poder calcular o valor de $E(X^2)$:

Tabela 4: Distribuição de probabilidades do Exemplo 1 (com coluna $x_i^2 \times p(X = x_i)$)

X	$p(X = x_i)$	$x_i \times p(X = x_i)$	$x_i^2 \; p(X = x_i)$
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

Substituindo nas expressões de valor esperado e variância:

$$E(X) = \sum_{i=1}^{n} x_i \times p(x_i) = 1{,}385 \text{ acertos}$$

$$V(X) = \sum_{i=1}^{n} x_i^2 \times p(x_i) - \left[\sum_{i=1}^{n} x_i \times p(x_i)\right]^2 = 3{,}005 - (1{,}385)^2 = 1{,}087 \text{ acertos}$$

$$\sigma(X) = \sqrt{V(X)} = \sqrt{1{,}087} = 1{,}042 \text{ acertos}$$

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, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulos 5 e 6.

- Sobre as propriedades de valor esperado e variância, BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulos 5 e 6.

- Também sobre variáveis aleatórias, STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulos 5 e 6.

- Sobre teoria estatística da decisão: BEKMAN, O. R.; COSTA NETO, P. O. *Análise Estatística da Decisão*. São Paulo: Edgard Blücher, 1980, 4ª reimpressão, 2006.

RESUMO

O resumo desta Unidade está demonstrado na Figura 49:

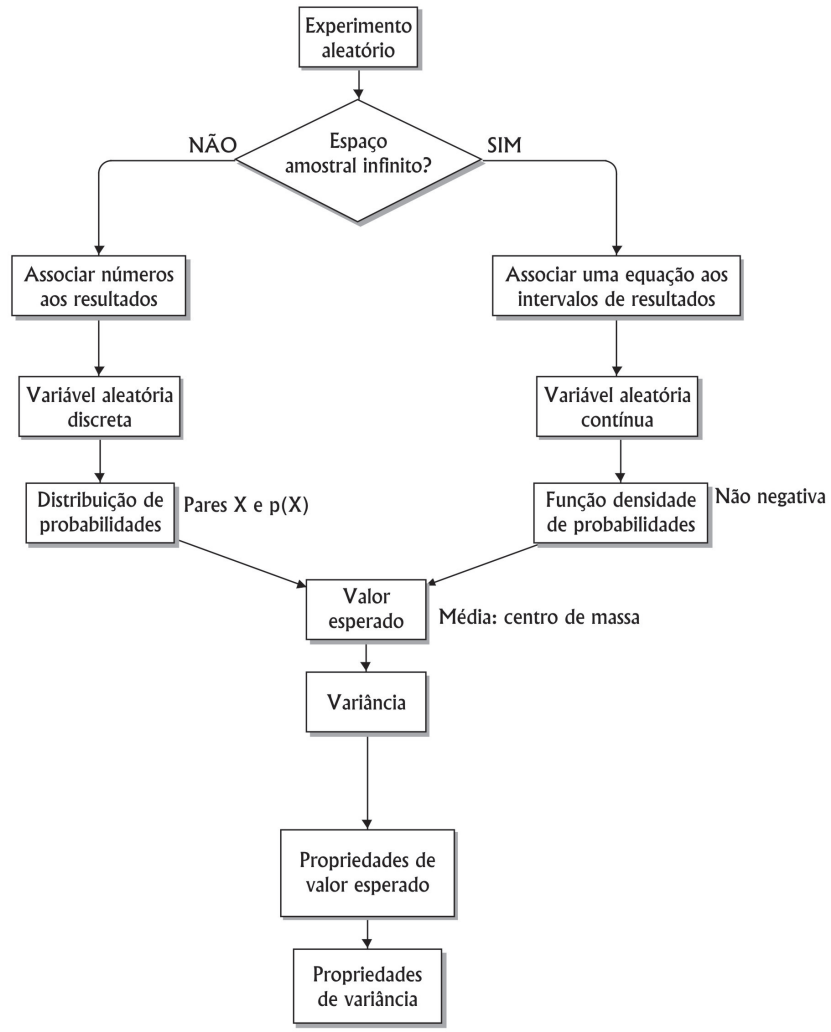


Figura 49: Resumo da Unidade 6
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Chegamos ao final de mais uma Unidade. Veremos mais sobre os temas abordados na Unidade 7, quando estudaremos várias distribuições de probabilidade (modelos probabilísticos) que são extremamente úteis para modelar muitas situações práticas, auxiliando na tomada de decisões. Estes conhecimentos serão depois aplicados nas Unidades 8 e 9.

UNIDADE

7

Modelos probabilísticos mais comuns

Objetivo

Nesta Unidade, você vai conhecer os modelos probabilísticos mais importantes para variáveis aleatórias discretas e contínuas. Você aprenderá a identificar as situações reais em que podem ser usados para o cálculo de probabilidades e a importância disso 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 vamos reinventar a roda e correndo o risco de ela sair quadrada...

Por que não usar os conhecimentos prévios desenvolvidos ao longo de centenas de anos de pesquisa e 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.

Através 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).

GLOSSÁRIO

*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 (2004).

*Espaço amostral finito – é aquele formado por um número limitado de resultados possíveis. Fonte: Barbetta, Reis e Bornia (2004).

*Espaço amostral infinito numerável – é aquele formado por um número infinito de resultados, mas que podem ser listados. Fonte: Barbetta, Reis e Bornia (2004).

Aqui é importante avaliar com cuidado a **variável aleatória*** discreta.

É preciso identificar se o **espaço amostral** é **finito*** ou **infinito numerável***: alguns modelos são apropriados para um caso, e não para o outro.

Vamos ver os dois modelos mais importantes: binomial e Poisson.

Modelo binomial

Seja um experimento aleatório qualquer que apresenta 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" (estão entre aspas, porque a definição de sucesso não quer necessariamente algo "positivo", e também porque poderá significar um grupo de resultados); e
- os ensaios são independentes entre si, apresentando probabilidades de "sucesso" (**p**) e de "fracasso" (**1–p**) constantes.

GLOSSÁRIO

*Variável aleatória discreta – o espaço amostral ao qual ela está associada é finito ou infinito numerável. Fonte: Barbetta, Reis e Bornia (2004).

Neste 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 **p** = probabilidade de "sucesso"

Com esses dois parâmetros, é possível calcular as probabilidades de um determinado número de sucessos, bem como obter o valor esperado e a variância da variável X:

$$E(X) = n \times p \qquad V(X) = n \times p \times (1-p)$$

Exemplo 1: experimentos binomiais:

a) observar o número de caras em três lançamentos imparciais de uma moeda honesta: n=3; p=0,5;

b) observar o número de meninos nascidos em três partos de uma família: n=3; p = x; e

c) observar o número de componentes defeituosos em uma amostra de dez componentes de um grande número de peças que apresentaram anteriormente 10% de defeituosos: n = 10; p= 0,1.

Vamos ver com maiores 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, e sabendo pela história da família que P(H) = 0,52 e P(M) = 0,48 (então **p** = 0,52 e **1– p** = 0,48), quais serão as probabilidades obtidas para a variável aleatória número de meninos em três nascimentos? Vamos obter a distribuição de probabilidades.

Usando os conceitos gerais de probabilidade, é preciso primeiramente determinar o espaço amostral, como poderão ser os sexos das três crianças:

$$\mathbf{\Omega = \{H \cap H \cap H,\ H \cap H \cap M,\ H \cap M \cap H,\ M \cap H \cap H,\ H \cap M \cap M,\ M \cap H \cap M,\ M \cap M \cap H,\ M \cap M \cap M\}}$$

Supondo que os nascimentos sejam independentes, podemos calcular as probabilidades de cada intersecção simplesmente multiplicando as probabilidades individuais de seus componentes:

$P\{H \cap H \cap H\} = P(H) \times P(H) \times P(H) = p \times p \times p = p^3$

$P\{H \cap H \cap M\} = P(H) \times P(H) \times P(M) = p \times p \times (1-p) = p^2 (1-p)$

$P\{H \cap M \cap H\} = P(H) \times P(M) \times P(H) = p \times (1-p) \times p = p^2 \times (1-p)$

$P\{M \cap H \cap H\} = P(M) \times P(H) \times P(H) = (1-p) \times p \times p = p^2 \times (1-p)$

$P\{H \cap M \cap M\} = P(H) \times P(M) \times P(M) = p \times (1-p) \times (1-p) = p \times (1-p)^2$

$P\{M \cap H \cap M\} = P(M) \times P(H) \times P(M) = (1-p) \times p \times (1-p) = p \times (1-p)^2$

$$P\{M \cap M \cap H\} = P(M) \times P(M) \times P(H) = (1-p) \times (1-p) \times p = p \times (1-p)^2$$

$$P\{M \cap M \cap M\} = P(M) \times P(M) \times P(M) = (1-p) \times (1-p) \times (1-p) = (1-p)^3$$

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 dois "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 um "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. Supondo 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 três ensaios (**n** =3):

- para dois sucessos (**k** =2) $C_{3,2} = \frac{3!}{2! \times (3-2)!} = 3$ (o mesmo resultado obtido por enumeração);
- para um 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 $\mathbf{n \geq k}$), permitindo 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 um certo valor **k** ($0 \leq \mathbf{k} \leq \mathbf{n}$) 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.

Neste segundo exemplo, admitamos que a probabilidade de que uma companhia não entregue seus produtos no prazo é igual a 18%. Quais são as probabilidades de que, em três entregas, uma, duas ou todas as três entregas sejam feitas no prazo? Calcular também 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 três terá distribuição binomial com parâmetros

n = 3 e **p** = 0,82 (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 todas as probabilidades, o resultado é igual a 1, como teria que ser. O valor esperado, variância e o desvio-padrão serão:

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 um número diferente de 0 elevado a zero é igual a 1.

$$E(X) = n \times p = 3 \times 0{,}82 = 2{,}46 \text{ entregas}$$

$$V(X) = n \times p \times (1 - p) = 3 \times 0{,}82 \times 0{,}18 = 0{,}4428 \text{ entrgas}^2.$$

$$\sigma(X) = \sqrt{V(X)} = \sqrt{0{,}4428} = 0{,}665 \text{ entregas}$$

A média é quase igual ao número de operações, devido à alta probabilidade de sucesso.

Estudos anteriores mostraram que há 73% de chance de consumidores do sexo feminino apresentarem uma reação positiva a anúncios

publicitários com crianças. Uma agência está conduzindo um estudo, apresentando um novo anúncio para cinco consumidoras. Vamos ver nesse Exemplo 3 qual é a probabilidade de que pelo menos três das cinco consumidoras apresentem reação positiva? Calcular também o valor esperado, a variância e o desvio-padrão do número de consumidoras que apresentam reação positiva.

Para cada consumidora ("ensaio"), há apenas dois resultados: reação positiva ou não. Há um número limitado de realizações, **n** = 5. Definindo "sucesso" como reação positiva, e supondo as consumidoras "independentes", a variável aleatória X, número de consumidoras com reação positiva em cinco que assistiram ao novo anúncio terá distribuição binomial com parâmetros

n = 5 e **p** = 0,73 (e **1 – p** = 0,27).

O evento de interesse é a recuperação de pelo menos três ratos (3 ou mais): **P(X ≥ 3)**.

$$\mathbf{P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)}$$

É preciso calcular as três probabilidades acima e somá-las, então:

$$P(X=3) = C_{5,3} \times 0{,}73^3 \times (0{,}27)^2 = \frac{5!}{3! \times (5-3)!} \times 0{,}73^3 \times (0{,}27)^2 = 0{,}284$$

$$P(X=4) = C_{5,4} \times 0{,}73^4 \times (0{,}27)^1 = \frac{5!}{4! \times (5-4)!} \times 0{,}73^4 \times (0{,}27)^1 = 0{,}383$$

$$P(X=5) = C_{5,5} \times 0{,}73^5 \times (0{,}27)^0 = \frac{5!}{5! \times (5-5)!} \times 0{,}73^5 \times (0{,}27)^0 = 0{,}207$$

$$\mathbf{P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) = 0{,}284 + 0{,}383 + 0{,}207 = 0{,}874}$$

A probabilidade de que pelo menos três das cinco consumidoras apresentem reação positiva é igual a 0,874 (87,4%).

Há duas outras formas de chegar ao mesmo resultado:

- através do complementar:

$$\mathbf{P(X \geq 3) = 1 - P(X<3) = 1 - [\ P(X=0) + P(X=1) + P(X=2)];}$$

- mudando a definição de sucesso, de reação positiva para **reação negativa** (**p** = 0,27), se pelo menos três consumidoras

apresentam reação positiva, então, no máximo duas apresentam reação negativa.

Mas se o espaço amostral fosse infinito numerável? Teríamos que usar o modelo de Poisson. Você conhece este modelo? Sabe como tirar proveito de suas facilidades? Vamos estudar juntos para aprender ou para relembrar!

Modelo de Poisson

Vamos supor um experimento binomial, com apenas dois resultados possíveis, mas com a seguinte característica: apesar de a probabilidade **p** ser constante, o valor de **n** teoricamente é infinito.

Na situação acima, o modelo binomial não poderá ser utilizado. Nestes casos, deve ser utilizado o modelo de Poisson.

Como seria a solução para o caso acima?

Como **n** é "infinito", deve-se fazer a análise das ocorrências em um período contínuo (de tempo, de espaço, entre outros) subdividido em um 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 obter a 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** ≅ 2,71. E m = n x p ou m = λ x 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 neste Exemplo 4 os experimentos e fenômenos que seguem a distribuição de Poisson.

a) Número mensal de acidentes de tráfego em um cruzamento. Observe que é uma variável aleatória discreta, 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, um 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 muitos 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 contí-

nuo (no caso, período de tempo, uma 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 à dos acidentes em um cruzamento, só que o "grau de aleatoriedade" deste 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.

Neste Exemplo 5, uma telefonista recebe cerca de 0,20 chamadas por minuto (valor obtido de medições anteriores).

a) Qual é a probabilidade de receber exatamente cinco chamadas nos primeiros dez minutos?

b) Qual é a probabilidade de receber até duas chamadas nos primeiros 12 minutos?

c) Qual é 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 um limite para o número de chamadas no período (sabe-se apenas que o número mínimo pode ser 0): por esse motivo, a utilização da binomial é inviável. Contudo, há uma taxa de ocorrência (λ = 0,20 chamadas/minuto), e isso permite utilizar a distribuição de Poisson.

a) Neste caso, o período **t** será igual a 10 minutos (**t** = 10 min.), P(X = 5)?

m = λ x t = 0,20 x 10 = 2 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 que a telefonista receba exatamente cinco chamadas em dez minutos é igual a 0,0361 (3,61%).

b) Neste caso, o período **t** será igual a 12 minutos (**t** = 12 minutos). O evento de interesse é até duas chamadas em 12 minutos ($X \leq 2$).

$m = \lambda \times t = 0{,}20 \times 12 = 2{,}4$ 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 que a telefonista receba até duas chamadas em 12 minutos é igual a 0,5697 (56,97%).

c) Neste caso, o período **t** será igual a 30 minutos (**t** = 30 minutos). Primeiro, calcula-se a variância:

$V(X) = m = \lambda \times t = 0{,}2 \times 30 = 6$ chamadas²

O desvio-padrão é a raiz quadrada positiva da variância:

$$\sigma(X) = \sqrt{V(X)} = \sqrt{6} \cong 2,45 \text{ chamadas}$$

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 que uma variável aleatória contínua assuma exatamente um 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, calcula-se a probabilidade de ocorrência de eventos formados por intervalos de valores, através de uma função densidade de probabilidades (ver Unidade 6). Uma outra conseqüência disso é que os símbolos $>$ e $\geq$ ($<$ e $\leq$ também) são equivalentes para variáveis aleatórias contínuas.

O modelo mais simples para variáveis aleatórias contínuas é o 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 50:

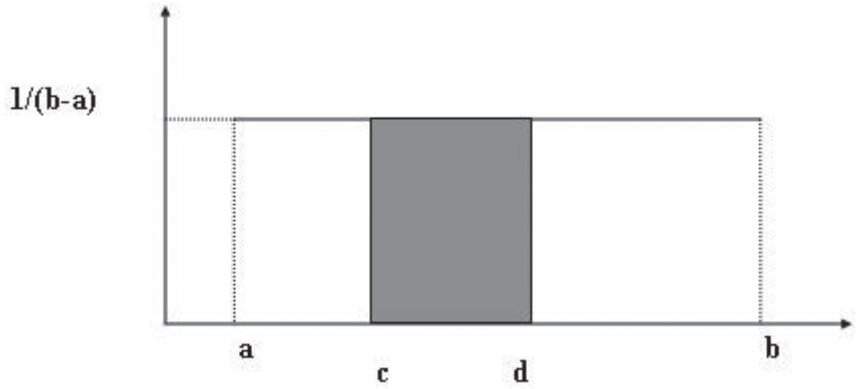


Figura 50: Modelo uniforme
Fonte: elaborada pelo autor

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 de a 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 do que **b**), se sua função densidade de probabilidades for tal como a da Figuras 50.

A probabilidade de que a variável assuma 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 a variância são:

$$E(X) = \frac{a + b}{2} \qquad V(X) = \frac{(b - a)^2}{12}$$

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 pró-

ximos 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.

Neste Exemplo 6, a temperatura T de destilação do petróleo é crucial para determinar a qualidade final do produto. Suponha que T seja considerada uma variável aleatória contínua com distribuição uniforme de 150 a 300° C, e que o custo para produzir um galão de petróleo seja de 50 u.m. Se o óleo é destilado a menos de 200° C, o galão é vendido a 75 u.m. Se a temperatura for superior a 200° C, o produto é vendido a 100 u.m.

Adaptado de BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 4. ed. São Paulo: Atual, 1987.

a) Fazer o gráfico da função densidade de probabilidade de T.

b) Qual é o lucro médio esperado por galão?

a) Os parâmetros a e b definem completamente uma distribuição uniforme; para fazer o gráfico, basta encontrá-los no enunciado acima. Identifica-se que o limite inferior, a, vale 150° C, e o superior, b, vale 300° C, resultando no gráfico a seguir:

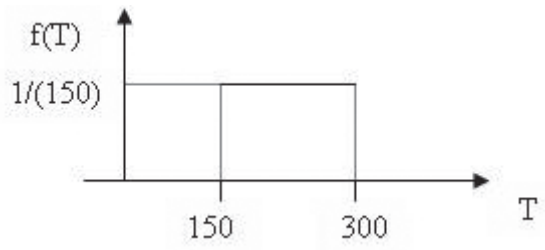


Figura 51: Temperatura de destilação do petróleo
Fonte: elaborada pelo autor

b) A variável aleatória de interesse, lucro, é discreta, somente pode assumir dois valores: 25 u.m. (caso o óleo seja destilado a menos de 200° C, posto que o galão custa 50 u.m. para ser produzido e será vendido a 75 u.m. nestas condições), ou 50 u.m. (caso o óleo seja destilado a mais de 200° C, posto que o galão custa 50 u.m. para ser produzido e será vendido a 100 u.m.). Sendo assim, seus valores possíveis serão: {25, 50}, sendo os resultados mutuamente exclusivos.

Lembrando das definições de distribuições de probabilidades, e de valor esperado e variância para variáveis aleatórias discretas (Unidade 6), para obter o lucro médio (valor esperado da variável lucro), é preciso obter as probabilidades de ocorrência dos seus dois valores (25 e 50). Relacionando com os valores de T:

$P(\text{Lucro} = 25) = P(T \leq 200)$ $P(\text{Lucro} = 25) = P(T > 200)$

Os valores das probabilidades acima correspondem às áreas abaixo da curva da função densidade de probabilidades para cada intervalo, calculando as áreas:

$$P(T \leq 200) = (200 - 150) \times \frac{1}{(300 - 150)} = \frac{50}{150}$$

$$P(T > 200) = (300 - 200) \times \frac{1}{(300 - 150)} = \frac{100}{150}$$

Então, a distribuição de probabilidades da variável lucro será (Quadro 22):

Lucro	Probabilidade
25	50/150
50	100/150
Total	1,0

Quadro 22: Distribuição de probabilidades da variável lucro
Fonte: elaborado pelo autor

Calculando o valor esperado:

$$E(\text{Lucro}) = \sum \text{Lucro}_i \times P(\text{Lucro}_i) \qquad E(\text{Lucro}) = 25 \times \frac{50}{150} + 50 \times \frac{100}{150} = 41{,}67 \text{ u.m.}$$

O lucro médio é de 41,67 u.m. Repare que a variável lucro não pode assumir este valor, o que significa que o valor esperado (a média) não é o valor mais provável. Neste problema, o valor mais provável, a moda (ver Unidade 4), vale 50 u.m., pois tem a maior probabilidade de ocorrência (66,67%).

Agora, vamos passar ao modelo mais importante para variáveis aleatórias contínuas.

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 esta probabilidade diminui à medida que os valores se afastam deste centro (para valores menores ou maiores). O modelo probabilístico contínuo mais adequado talvez seja o modelo normal ou gaussiano.

O matemático alemão Gauss utilizou amplamente este modelo no tratamento de erros experimentais, embora não tenha sido o seu "descobridor".

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 aleatória contínua que siga o modelo normal (distribuição normal) será como a Figura 52:

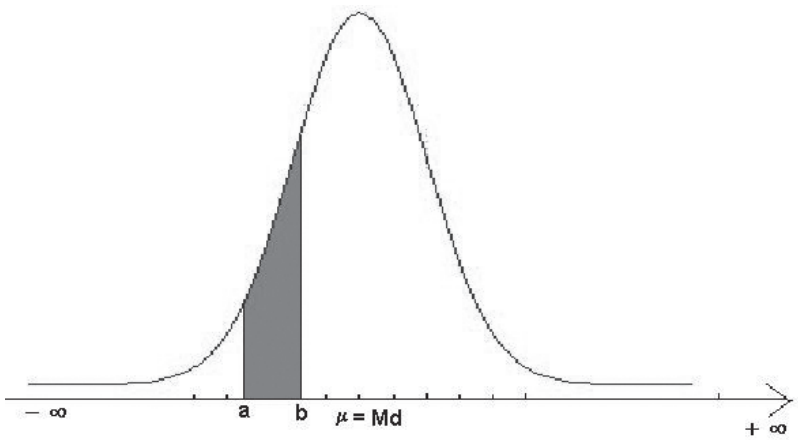


Figura 52: Distribuição normal
Fonte: elaborada pelo autor

Características do modelo normal0:

- 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) => **X: N (μ , σ^2)** significa que a variável X tem distribuição normal com média **μ** e variância σ^2;
- a área escura na Figura 3 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; e
- a função normal citada acima tem a seguinte (e aterradora...) fórmula para sua função densidade de probabilidade:

É comum a utilização de letras do alfabeto grego para representar algumas medidas. Não se esqueça que o desvio-padrão ( ) é a raiz quadrada positiva da variância.

$$f(x) = \frac{1}{\sqrt{2 \times \pi \times \sigma^2}} \times e^{\left(\frac{-1}{2} \times \left[\frac{x-\mu}{\sigma}\right]^2\right)} \quad -\infty < x < +\infty$$

Saiba que não existe solução analítica para uma integral da expressão acima: qualquer integral precisa ser resolvida usando métodos numéricos de integração, que são extremamente trabalhosos quando implementados manualmente (somente viáveis se são usados 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.

Gauss e todas as outras pessoas que usavam a distribuição normal para calcular probabilidades até recentemente resolviam as integrais usando métodos numéricos manualmente.

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-padrão ($\mu - 2\sigma$) e a média mais dois desvios-padrão ($\mu + 2\sigma$); e
- 99,7% dos dados estão situados entre a média menos três desvios-padrão ($\mu - 3\sigma$) e a média mais três desvios-padrão ($\mu + 3\sigma$).

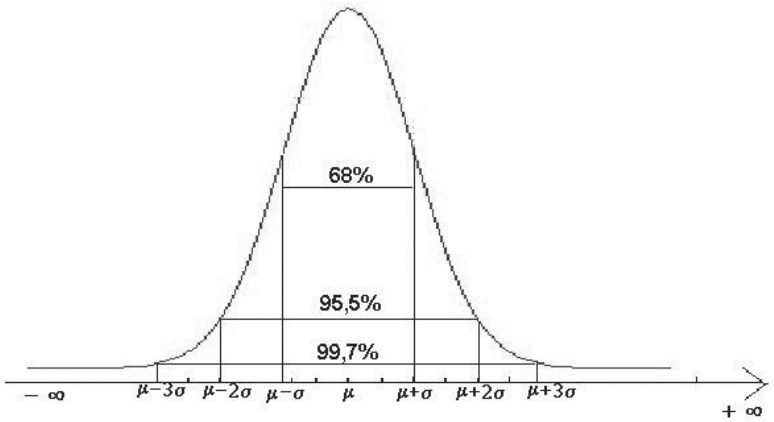


Figura 53: Percentuais de dados e número de desvios-padrão
Fonte: elaborada pelo autor

Por causa dessas características, alguém teve a idéia 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** ~(, **1**)]. As probabilidades foram calculadas para esta 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-padrão, a contar da média, a que está um valor da variável, através da seguinte expressão:

$$Z = \frac{x - \mu}{\sigma}$$

Z – número de desvios-padrão a partir da média;
x – valor de interesse;
μ – média da distribuição normal de interesse;
σ – desvio-padrão da distribuição normal.

Z é um valor relativo: será negativo para valores de **x** menores do que a média e será positivo para valores de **x** maiores do que a média. Pela transformação, uma distribuição normal qualquer **X: N (μ , σ²)** passa a ser equivalente à distribuição normal-padrão Z: **N˜** (, **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 AVEA, apresentamos uma 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 54 mostra como podemos usar essa Tabela para encontrar, por exemplo, a área sob a cauda superior da curva, além de z = 0,21.

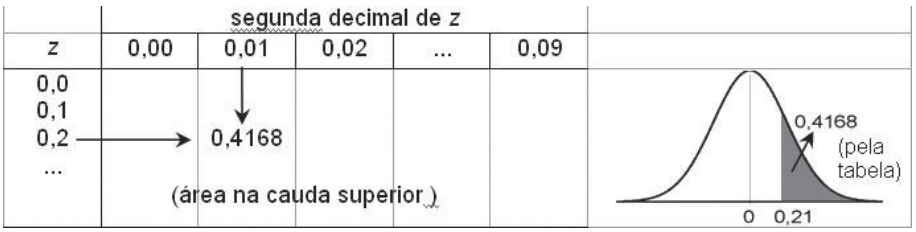


Figura 54: Ilustração do uso da tabela da distribuição normal-padrão (Tabela III do apêndice) para encontrar a área na cauda superior relativa ao valor de z = 0,21
Fonte: Barbetta, Reis, Bornia (2004)

No Exemplo 7, 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 55 e 56, se pode ver a correspondência entre as duas distribuições:

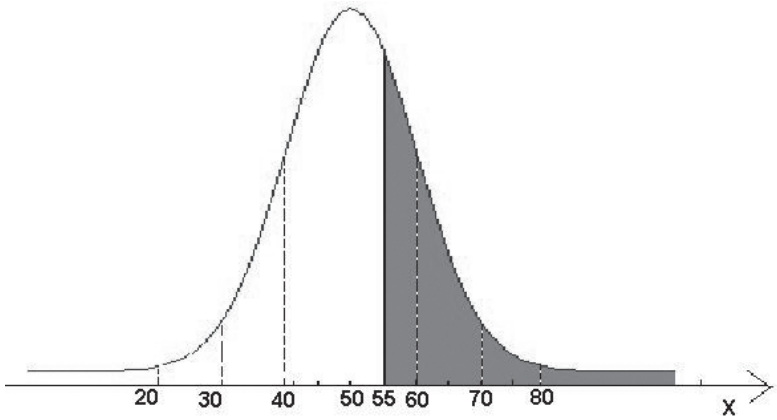


Figura 55: Distribuição normal N(50,102)
Fonte: elaborada pelo autor

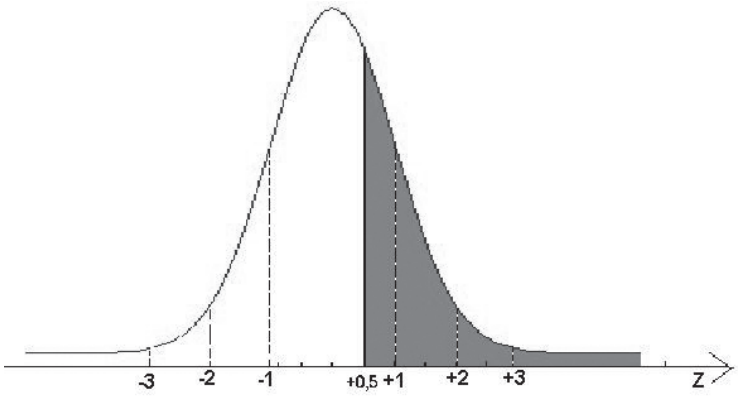


Figura 56: Distribuição normal-padrão
Fonte: elaborada pelo autor

O evento P (X>55) é equivalente ao evento P (Z> 0,5). Este valor pode ser obtido na tabela da distribuição normal-padrão (ver Ambiente Virtual). Os valores de Z são apresentados com duas 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 do tipo das figuras acima [P(Z> z1)]. 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%.

Neste oitavo exemplo, supondo 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 57 e 58, podemos ver a correspondência entre as duas distribuições:

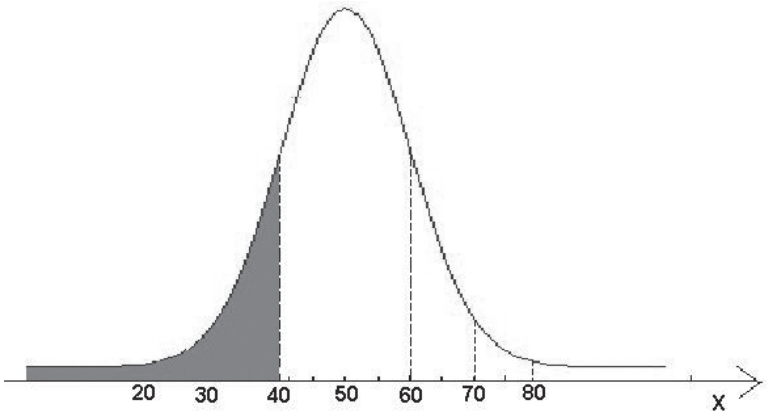


Figura 57: Distribuição normal N(50,102)
Fonte: elaborada pelo autor

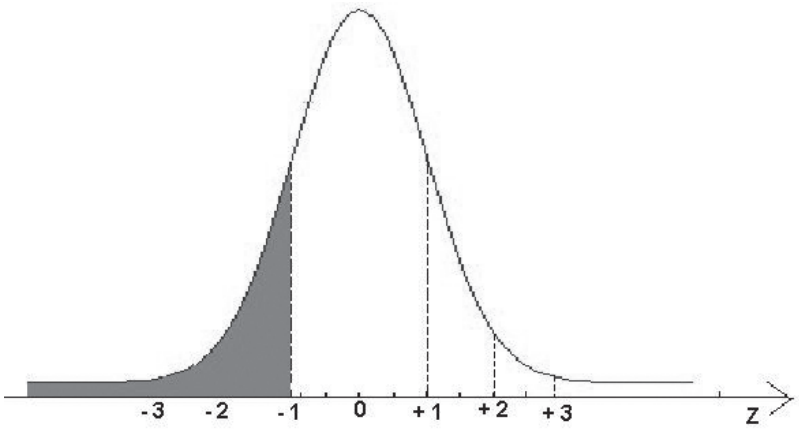


Figura 58: Distribuição normal-padrão
Fonte: elaborada pelo autor

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 as figuras da distribuição normal para a direita, teremos o seguinte resultado (Figura 59):

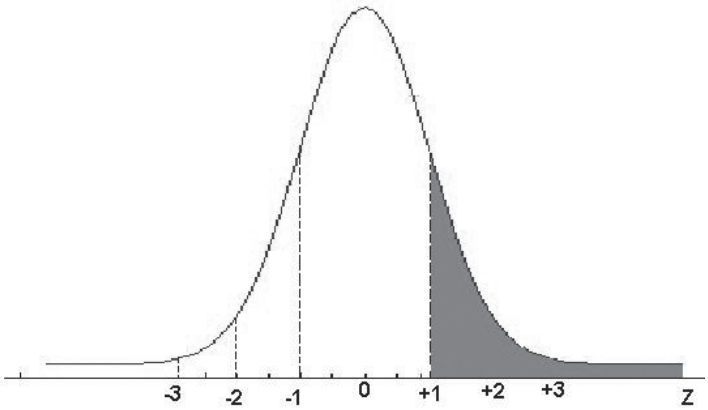


Figura 59: Distribuição normal-padrão
Fonte: elaborada pelo autor

Ou seja, a área P(Z < –1) = P(Z > 1). Esta probabilidade, nós podemos encontrar diretamente pela tabela, fazendo o cruzamento do valor 1,0 (na coluna) com o valor 0,00 (na linha do topo) encontramos o valor 0,1587 (15,87%). Portanto, P(X<40) = P(Z<-1) = P(Z>1), que é igual a 0,1587.

No nono exemplo, supondo a mesma variável aleatória X com média 50 e desvio-padrão 10. Agora, há interesse em calcular a probabilidade de que X seja maior do que 35.

Primeiro, precisamos calcular o valor de Z correspondente a 35. Z = (35 – 50)/ 10 = –1,50.

Pelas Figuras 60 e 61, se pode ver a correspondência entre as duas distribuições:

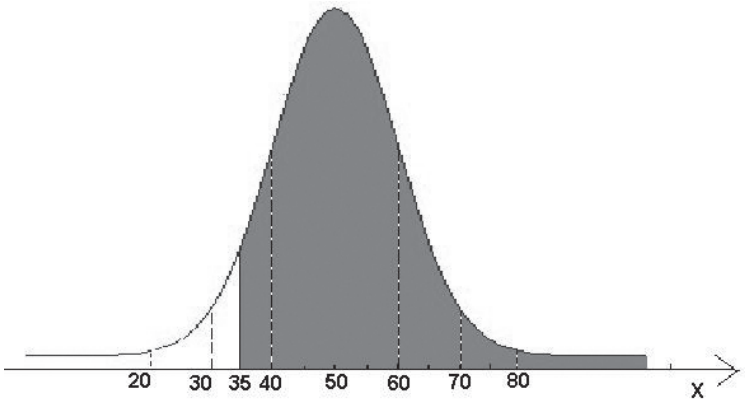


Figura 60: Distribuição Normal N(50,102)
Fonte: elaborada pelo autor

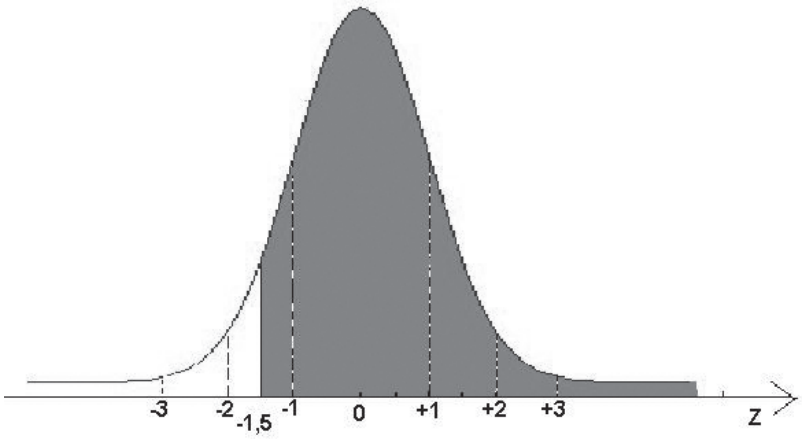


Figura 61: Distribuição normal-padrão
Fonte: elaborada pelo autor

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 que P(Z > –1,50) = 1 - P(Z < –1,50). Usando o raciocínio descrito no Exemplo 8 (rebatendo as figuras para a direita), vamos obter: 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 (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%.

No Exemplo 10, supondo a mesma variável aleatória X com média 50 e desvio-padrão 10. Agora, há interesse em calcular a probabilidade de que X assuma valores entre 48 e 56.

Calcular P (48 < X < 56). Veja a Figura 62 abaixo:

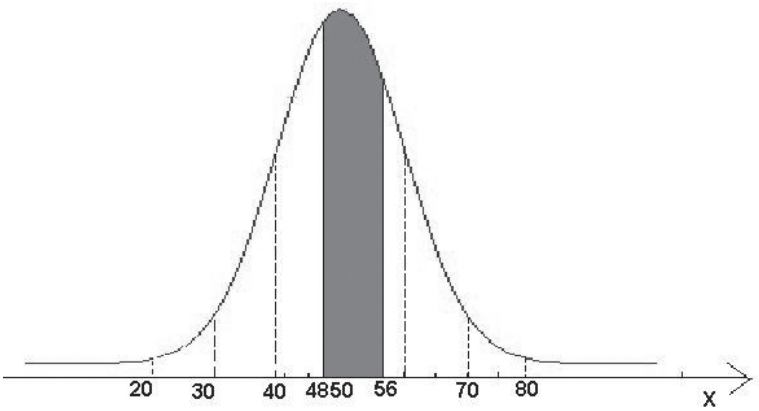


Figura 62: Distribuição normal N(50, 10^2)
Fonte: elaborada pelo autor

Novamente, precisamos calcular os valores de Z correspondentes a 48 e a 56.

$Z_1 = (48 - 50)/ 10 = - 0,20$ $Z_2 = (56 - 50)/ 10 = 0,60$

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 $+\infty$ 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 de a 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.

No Exemplo 11, supondo 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 neste 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 63:

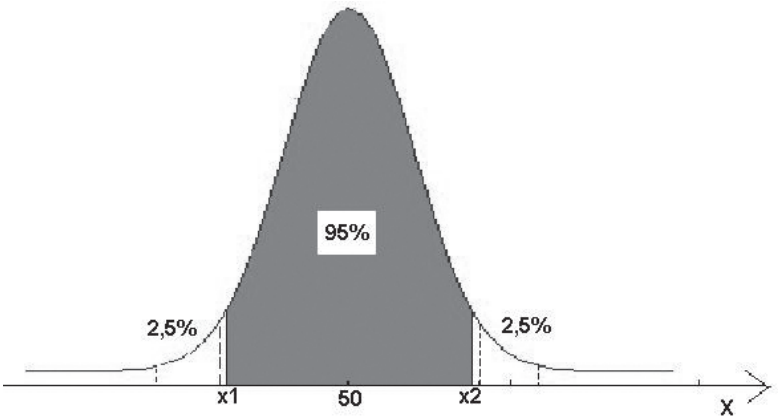


Figura 63: Distribuição normal N (50, 10^2)
Fonte: elaborada pelo autor

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 $\mathbf{x}_1$ e $\mathbf{x}_2$. Passando para a distribuição normal-padrão $\mathbf{x}_1$, corresponderá a um valor $\mathbf{z}_1$, e $\mathbf{x}_2$ a um valor $\mathbf{z}_2$, como na Figura 64:

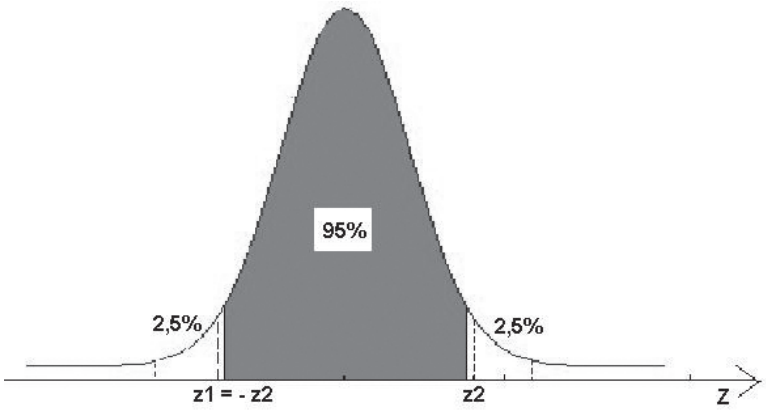


Figura 64: Distribuição normal-padrão
Fonte: elaborada pelo autor

Repare que a média da distribuição normal-padrão é igual a zero, fazendo com que $\mathbf{z_1}$ e $\mathbf{z_2}$ sejam iguais em módulo. Podemos encontrar $\mathbf{z_2}$, já que $P(Z > \mathbf{z_2}) = 0,025$

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 $\mathbf{z_2}$, ao procurar pela probabilidade 0,025, encontramos o valor exato 0,025, e, por conseguinte, o valor de $\mathbf{z_2}$, que é igual a 1,96: $P (Z > 1,96) = 0,025$.

Como $\mathbf{z_1} = -\mathbf{z_2}$, encontramos facilmente o valor de z1: $\mathbf{z_2} = -1,96$. $P (Z < -1,96) = 0,025$.

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:

$$Z = \frac{x - \mu}{\sigma} \quad => \quad x = \mu + Z \times \sigma$$

E assim obteremos os valores de $\mathbf{x_1}$ e $\mathbf{x_2}$, que correspondem a $\mathbf{z_1}$ e $\mathbf{z_2}$, respectivamente:

É muito importante que se preste atenção no sinal do valor de z ao obter o valor de x.

$$\mathbf{x_1 = \mu + (z_1 \times \sigma = 50 + [(-1,96) \times 10] = 30,4}$$

$$\mathbf{x_2 = \mu + (z_2 \times \sigma = 50 + (1,96 \times 10) = 69,6}$$

GLOSSÁRIO

*Modelo binomial – modelo probabilístico para variáveis aleatórias discretas que descreve o número de sucessos em n experimentos independentes (sendo n finito e conhecido). Os experimentos podem ter apenas dois resultados possíveis, e a probabilidade de sucesso permanece constante durante os n experimentos. Fonte: Barbetta, Reis e Bornia (2004) e Lopes (1999).

Observe que os resultados obtidos são coerentes: 30,4 está abaixo da média (1,96 desvios-padrão), e 69,6, acima (também 1,96 desvios-padrão). O intervalo definido por estes dois valores compreende 95% dos resultados da variável X.

Todo este trabalho poderia ter sido poupado, 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 a sua capacidade de aproximar 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***.

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 sejam 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.

Para os que pensam que o advento dos computadores eliminou este problema, um alerta: em alguns casos, os números envolvidos são tão grandes que sobrepujam suas capacidades.

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)

variância = $\sigma^2 = n \times p \times (1 - p)$ (variância do modelo binomial)

Usando 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. Este intervalo deve ser

centrado no valor discreto e 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. **Esta 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 fazer a aproximação pela normal, será: P(**k – 0,5** < X < **k + 0,5**).

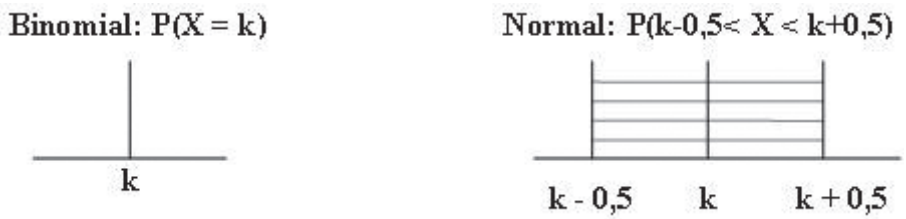


Figura 65: Correção de continuidade da aproximação do modelo binomial pelo normal – 1º caso
Fonte: elaborada pelo autor

2) Há interesse em calcular a probabilidade de X assumir valores menores ou iguais a um valor **k** genérico, P(X ≤ **k**); ao fazer a aproximação pela normal, será: P(X < **k** + **0,5**), todo o intervalo referente a k será incluído.

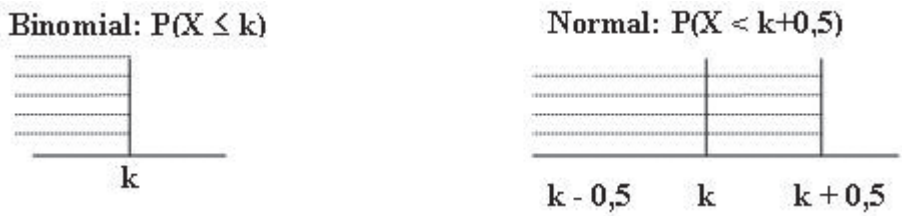


Figura 66: Correção de continuidade da aproximação do modelo binomial pelo normal – 2º caso
Fonte: elaborada pelo autor

3) Há interesse em calcular a probabilidade de X assumir valores maiores ou iguais a um valor **k** genérico, P(X ≥ **k**); ao fazer a aproximação pela normal, será: P(X > **k – 0,5**), todo o intervalo referente a k será incluído.

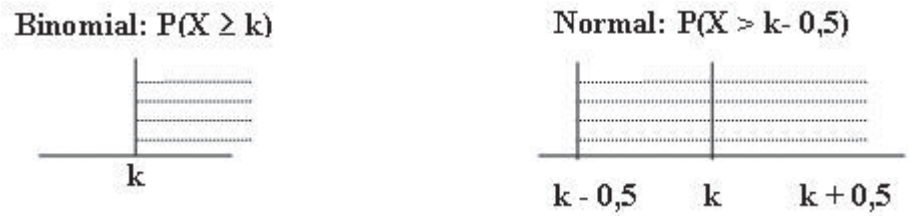


Figura 67: Correção de continuidade da aproximação do modelo binomial pelo normal – 3º caso
Fonte: elaborada pelo autor

4) Há interesse em calcular a probabilidade de X assumir valores menores do que um valor **k** genérico, P(X < **k**); ao fazer a aproximação pela normal, será: P(X < **k – 0,5**), todo o intervalo referente a k será excluído.

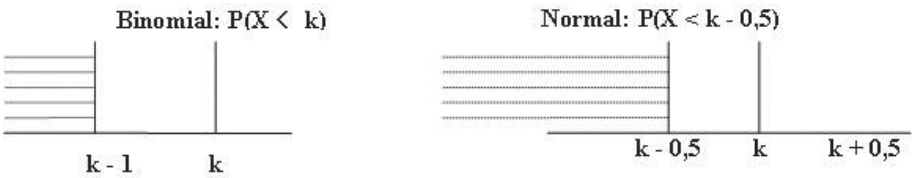


Figura 68: Correção de continuidade da aproximação do modelo binomial pelo normal – 4º caso
Fonte: elaborada pelo autor

5) Há interesse em calcular a probabilidade de X assumir valores maiores do que um valor **k** genérico, P(X > **k**); ao fazer a aproximação pela normal, será: P(X > **k + 0,5**), todo o intervalo referente a k será excluído.

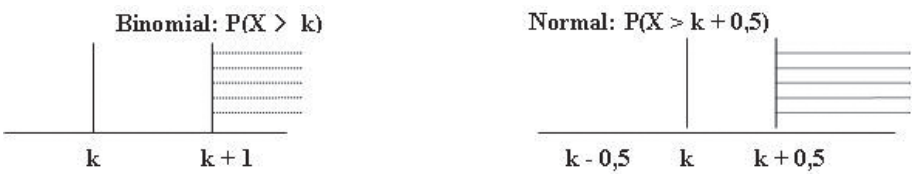


Figura 69: Correção de continuidade da aproximação do modelo binomial pelo normal – 5º caso
Fonte: elaborada pelo autor

Um município tem 40.000 eleitores. Para uma pesquisa de opinião eleitoral, uma amostra aleatória de 1.500 pessoas foi selecionada.

Vamos ver, nesse décimo segundo exemplo, qual é a probabilidade de que pelo menos 500 dos eleitores sejam menores de 25 anos se 35% dos 40.000 são menores do que 25 anos?

Este problema poderia ser resolvido usando 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 1.500 eleitores da amostra, e há independência entre as retiradas, pois a amostra foi retirada de forma aleatória (e a amostra representa menos de 5% dos 40.000 eleitores).

Então: "sucesso" = menor de 25 anos

p = 0,35 **1 - p** = 0,65 **n** = 1.500

A variável aleatória discreta X, número de eleitores menores de 25 anos em 1.500, 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:

$$P(X \geq 500) = P(X = 500) + P(X = 501) + ...+ P(X = 1.500)$$

Há cerca de 1.000 expressões binomiais.

Vamos ver se é possível aproximar pelo modelo normal.

O valor de **n** é grande:

$$\mathbf{n} \times \mathbf{p} = 1.500 \times 0,35 = 525 > 5 \text{ e } \mathbf{n} \times (\mathbf{1} - \mathbf{p}) = 1.500 \times 0,65 = 975 > 5.$$

Como as condições foram satisfeitas, é possível aproximar por um modelo normal:

média = $\square$ = **n × p = 1.500 x 0,35 = 525**

desvio-padrão = $\square = \sqrt{n \times p \times (1-p)} = \sqrt{1500 \times 0,35 \times 0,65} = 18,47$

Pelo modelo binomial: P (X ≥ 500). Pelo modelo normal, será: P (X ≥ 499,5).

$$P(X \geq 499,5) = P(Z > z_1)$$

$$z_1 = (499,5 - 525)/18,47 = -1,38$$

$$P(Z > -1,38) = 1 - P(Z > 1,38)$$

Procurando na tabela da distribuição normal-padrão:

$P(Z > 1{,}38) = 0{,}0838$

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 que pelo menos 500 dos eleitores da amostra sejam 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 Guiness, 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 (veremos isso na Unidade 8). 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 seu trabalho e quis publicá-las, mas a Guiness apenas permitiu após ele adotar o pseudônimo “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 um determinado número de elementos da amostra poderá ter seus valores variando livremente; este 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, apresentando maior variância para menores valores de amostra. Quanto maior o tamanho da amostra, mais a variância de **t** aproxima-se de 1,00 (variância da normal-padrão). A distribuição t de Student está na Figura 70 para vários graus de liberdade:

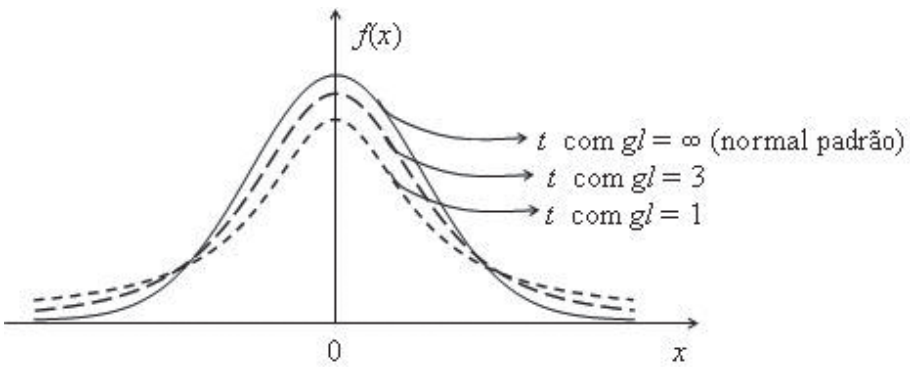


Figura 70: Distribuição t de Student para vários graus de liberdade
Fonte: Barbetta, Reis, Bornia (2004)

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.

Neste décimo terceiro exemplo, imagine a situação do Exemplo 12, obter os valores de t simétricos em relação à média que contêm 95% dos dados, supondo uma amostra de dez elementos.

Esta é a correção propriamente dita, pois, ao usar pequenas amostras, o risco de que a variância amostral da variável seja diferente da variância populacional é maior, podendo levar a intervalos de confiança que não correspondem à realidade. A não-utilização desta correção foi a fonte de muitos erros no passado e, infelizmente, ainda de alguns erros no presente.

Para tamanhos de amostra maiores do que 30, supõe-se que a variância de **t** é igual a 1: por isso, a aproximação do item b.1.

Temos que encontrar os valores $\mathbf{t_1}$ e $\mathbf{t_2}$, simétricos em relação à média que definem o intervalo que contém 95% dos dados. Como supomos uma amostra de dez 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 $\mathbf{t_1}$ e $\mathbf{t_2}$ sejam iguais em módulo. Podemos encontrar $\mathbf{t_2}$, já que $P(t > \mathbf{t_2}) = 0{,}025$. Veja a Figura 71:

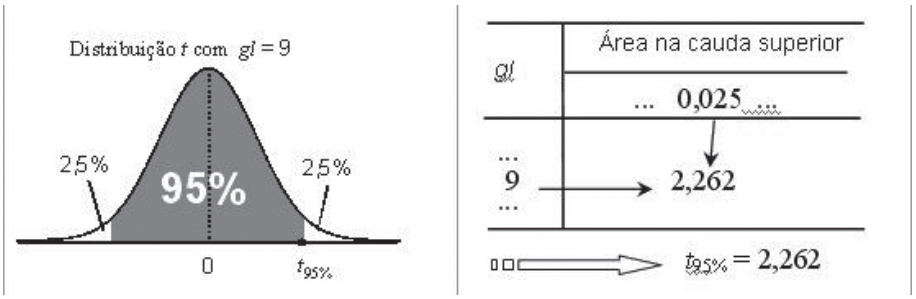


Figura 71: Uso da tabela da distribuição t de Student. Ilustração com gl = 9 e área na cauda superior de 2,5%
Fonte: Barbetta, Reis, Bornia (2004)

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, embora não vamos discutir como se dá esta derivação aqui.

Na Unidade 3, estudamos como descrever os relacionamentos entre duas variáveis qualitativas, geralmente expressos através de uma tabela de contingências. No Exemplo 5 da Unidade 3, analisamos o relacionamento 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 freqüências observadas de cada cruzamento entre os valores de duas variáveis qualitativas,

expressas em uma tabela de contingências, com as freqüências esperadas desses mesmos cruzamentos, se as duas variáveis não tivessem qualquer relacionamento entre si. Esta estatística é chamada de qui-quadrado, χ^2, e caso a hipótese seja de que as variáveis não se relacionam, ela seguirá o modelo qui-quadrado com um 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.

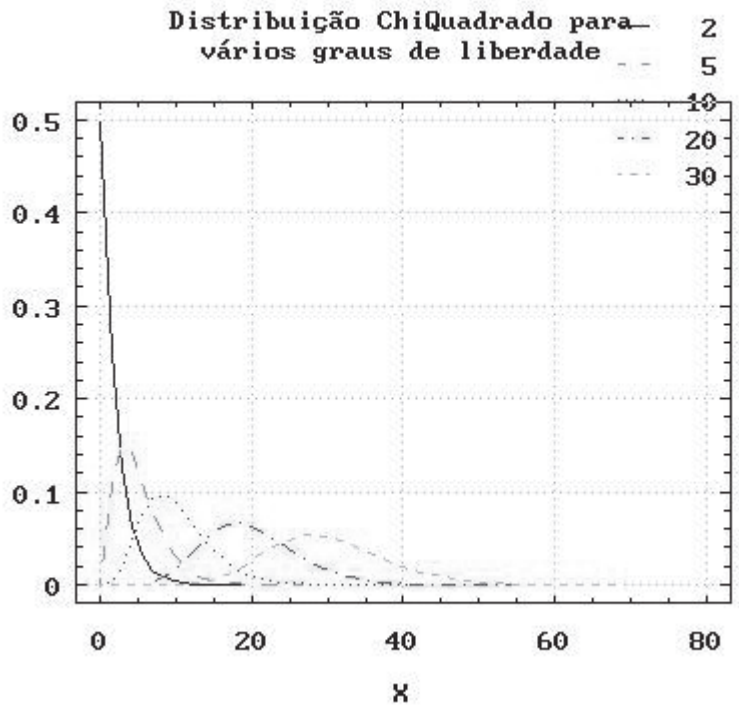


Figura 72: Modelo qui-quadrado com 2, 5, 10, 20 e 30 graus de liberdade
Fonte: adaptada pelo autor de Stagraphics®

A Figura 72 mostra as curvas do modelo (distribuição) qui-quadrado para 2, 5, 10, 20 e 30 graus de liberdade. Observe que a figura é assimétrica e como varia de forma, dependendo do número de graus de liberdade da estatística.

A tabela da distribuição qui-quadrado encontra-se no Ambiente Virtual de Ensino-Aprendizagem, para vários graus de liberdade e valores de probabilidade. Vamos ver um exemplo.

Neste décimo quarto exemplo, imagine que queremos encontrar o valor da estatística qui-quadrado, para três graus de liberdade, deixando uma área na cauda superior de 5%.

O valor da estatística qui-quadrado que define uma área na cauda superior de 5% pode ser encontrado através da Tabela, cruzando a linha de três graus de liberdade com a coluna de área na cauda superior igual a 0,05. Veja a Figura a seguir:

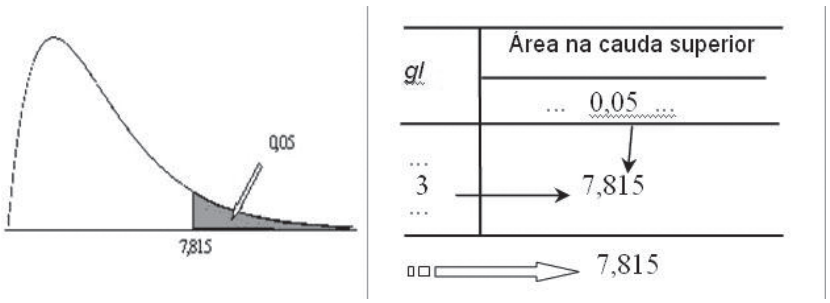


Figura 73: Uso da tabela da distribuição qui-quadrado. Ilustração com gl = 3 e área na cauda superior de 5%
Fonte: adaptado pelo autor de Barbetta, Reis, Bornia (2004)

Com este tópico, terminamos a Unidade 7. Na Unidade 8, você verá 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, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006, capítulo 7.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 5.

STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 4.

- Sobre modelos probabilísticos para variáveis aleatórias contínuas:

BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006. capítulo 8.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, 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, D. M. et al. *Estatística*: teoria e aplicações – usando Microsoft Excel em português. 5. ed. Rio de Janeiro: LTC, 2006, capítulos 4 e 5.

RESUMO

O resumo desta Unidade está mostrado na Figura 74:

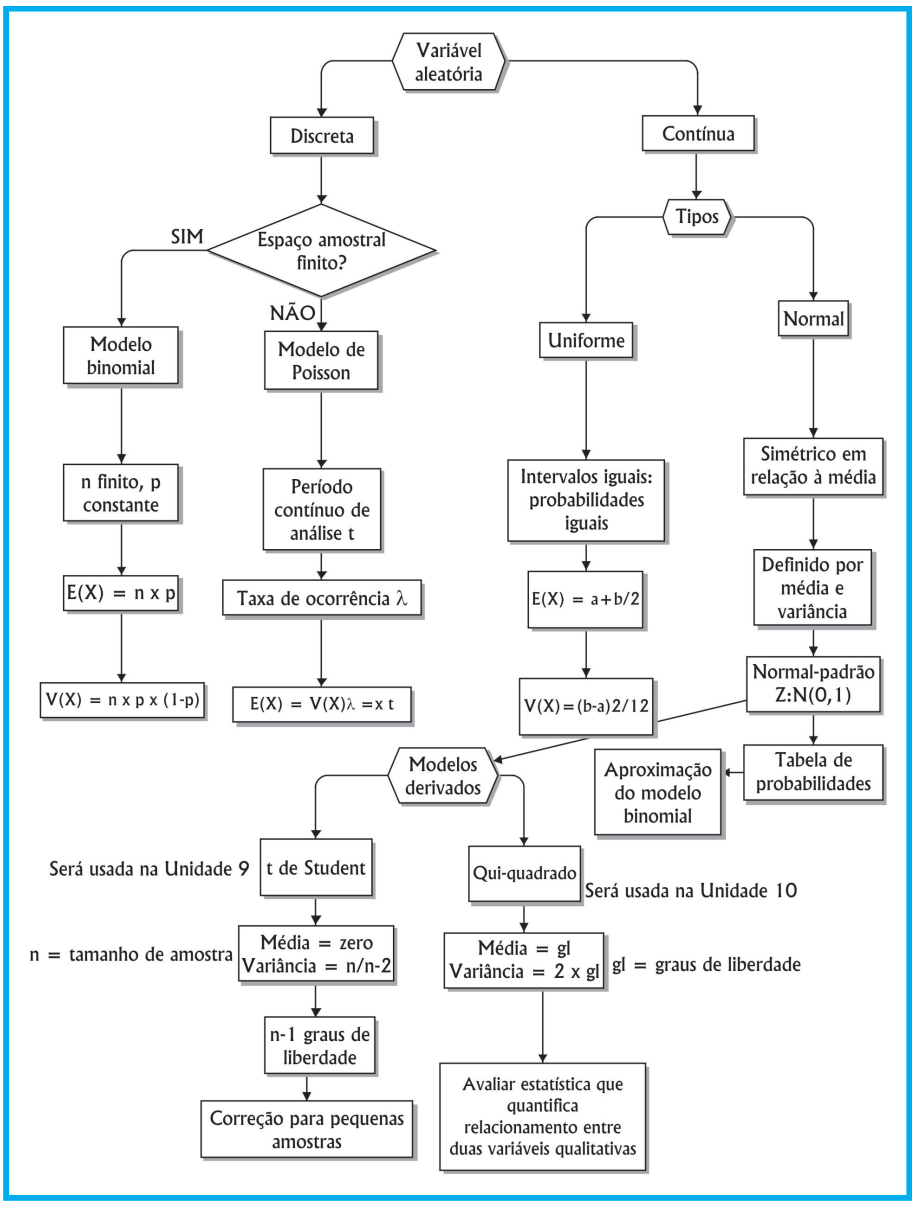


Figura 74: Resumo da Unidade 7
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Caro estudante!

Chegamos ao final da Unidade 7 do nosso livro. Nela estudamos os modelos probabilísticos mais comuns. Esta 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 as atividades de aprendizagem e visite o Ambiente Virtual de Ensino-Aprendizagem. Conte sempre com o acompanhamento da tutoria e das explicações do professor. Ótimos estudos!

UNIDADE

8

Inferência estatística e distribuição amostral

Objetivo

Nesta Unidade, você vai conhecer os conceitos de inferência estatística e de distribuição amostral, que são a base para o processo de generalização usado pelos administradores em várias tomadas 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 variáveis aleatórias, Unidades 5, 6 e 7), podemos generalizar os resultados de uma pesquisa por amostragem (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 (Unidade 5).

Uma vez retirada a amostra, fazemos Análise Exploratória dos Dados (Unidades 3 e 4): por exemplo, calculamos a média de uma variável quantitativa. Esta média e todas as demais estatísticas serão variáveis aleatórias (pois estão associadas ao **espaço amostral** de um experimento aleatório), e poderemos tentar identificar o modelo probabilístico mais apropriado para elas (Unidades 6 e 7). Mas, neste caso, o modelo probabilístico de uma estatística da amostra é chamado de **distribuição amostral**.

Conhecer a distribuição amostral das principais estatísticas vai nos ser muito útil quando estudarmos os tipos particulares de inferência estatística: estimação de parâmetros (Unidade 9) e testes de hipóteses (Unidade 10).

Vamos continuar aprendendo? É muito bom ter você conosco!

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

Na Unidade 2 enumeramos as principais razões para usar amostragem.

das medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m).

Uma população pode ser representada através de um modelo probabilístico: este apresenta condições para uso, forma para a distribuição de probabilidades e parâmetros.

Os dados necessários para a obtenção do modelo podem ser obtidos através de 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 uma 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 uma 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 que os métodos estatísticos permitem calcular essa probabilidade de erro. E 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 é o valor da média de uma variável que segue uma distribuição normal?;

GLOSSÁRIO

*Amostra aleatória, casual ou probabilística – 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 (2006).

*Parâmetros – característica numérica do modelo probabilístico da variável de interesse na população, tais como média, variância, proporção. Fonte: Barbetta, Reis e Bornia (2004).

*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: adaptado pelo autor de Barbetta, Reis e Bornia (2004).

- qual é 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 certo valor?;
- o modelo probabilístico da população é uma distribuição normal?; e
- 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.

GLOSSÁRIO

*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: adaptado pelo autor de Barbetta, Reis e Bornia (2004).

Parâmetros e estatísticas

Vamos imaginar uma pesquisa como a da Unidade 1, opinião dos registrados no CRA-SC sobre os cursos em que se graduaram, desde que tenham se graduado 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, uma vez que tenha sido realizada por amostragem probabilística, os dados podem ser estatisticamente generalizados para a população.

Uma vez tendo 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 freqüências, que relacionam os valores que a variável pode assumir com a freqüência (contagem) com que foram encontrados naquele conjunto. Esta distribuição pode ser apresentada na forma de uma tabela ou através de um gráfico (estes dois métodos podem ser usados tanto para variáveis qualitativas quanto para variáveis quantitativas).

Esta última está relacionada aos percentuais de ocorrência dos valores em uma distribuição de freqüências de uma variável qualitativa.

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.

GLOSSÁRIO

*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 (2004).

Atenção, vamos relembrar o que cada uma dessas significa.

***Média**: média aritmética simples (ver Unidade 4, seção 4.1.1). Trata-se de uma estatística que caracteriza o "centro de massa" do conjunto de dados (valor esperado – ver Unidade 6, seção 6.4). 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, seção 4.2.2) que mede a dispersão em torno da média do conjunto (em torno do valor esperado – Ver Unidade 6.4), 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, tendo, portanto, uma unidade que é igual à unidade da média, sendo muitas vezes preferida para efeito de mensuração da dispersão. Quando é o valor populacional recebe o símbolo **σ**; e 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 5 resume os parâmetros e as estatísticas.

Tabela 5: 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

Onde N é o número de elementos da população, n é o número de elementos da amostra, e f_a é a freqüência 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. Este modelo recebe o nome de distribuição amostral.

Distribuição amostral

Seja uma população qualquer com um parâmetro **θ** de interesse, correspondendo a uma estatística **T** em uma amostra. Amostras aleatórias são retiradas da população; e, para cada amostra, calcula-se o valor **t** da estatística **T**.

NÃO confundir com o t da distribuição t de Student, seção 7.2.4, 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 abaixo, constituída pelos pesos em kg de oito pessoas adultas:

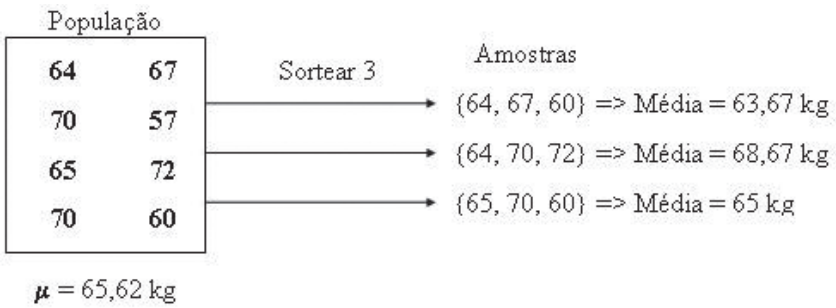


Figura 75: Distribuição amostral – Exemplo 1
Fonte: elaborada pelo autor

Observe que foram retiradas três amostras. Para cada amostra, foi calculada a média, visando a estimar a média populacional, que vale 65,62 kg. Observe que há uma variação na estatística média, pois o processo de amostragem é aleatório: é um experimento aleatório. Esta variação precisa ser considerada quando são realizadas as inferências sobre os parâmetros.

Assim sendo, o conhecimento das distribuições amostrais das principais estatísticas é necessário para fazer inferências sobre os parâmetros do modelo probabilístico da população. Por ora, basta co-

nhecer as distribuições amostrais das estatísticas médias de uma variável quantitativa qualquer e a proporção de um dos dois únicos resultados de uma variável qualitativa.

Distribuição amostral da média

Vamos observar as particularidades da distribuição amostral da média.

Neste segundo exemplo, suponha uma variável quantitativa cujos valores constituem uma população com os seguintes valores: (**2, 3, 4**, **5**).

Para esta população, que tem uma distribuição uniforme, podemos observar que os parâmetros são: $\mu = 3{,}5$ $\sigma^2 = 1{,}25$ (usou-se **n** no denominador por ser uma população).

Se retirarmos todas as amostras aleatórias de dois elementos (com reposição) possíveis desta 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 abaixo:

$$\overline{X}\begin{Bmatrix} (2,0) & (2,5) & (3,0) & (3,5) \\ (2,5) & (3,0) & (3,5) & (4,0) \\ (3,0) & (3,5) & (4,0) & (4,5) \\ (3,5) & (4,0) & (4,5) & (5,0) \end{Bmatrix}$$

Se estas médias forem plotadas em um histograma (Figura 76):

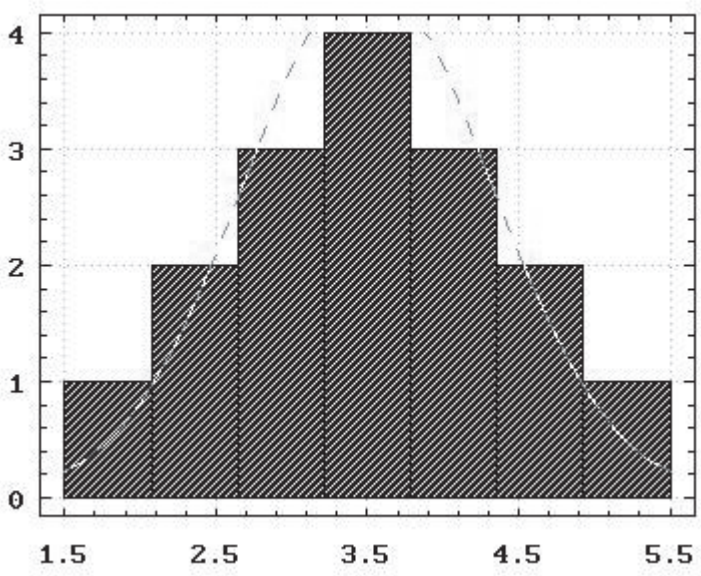


Figura 76: Histograma de médias amostrais
Fonte: adaptada pelo autor de Statsoft®

Se forem calculadas a média e a variância das médias de todas as amostras, o resultado será:

$$\overline{X} = 56/16 = 3{,}5 = \mu \qquad V(\overline{x}) = 0{,}625 = \frac{1{,}25}{2} = \frac{\sigma^2}{n}$$

Observe 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 acima vai se aproximar de um modelo normal, independentemente do formato da distribuição da variável na população.

Vamos ver outro exemplo.

Na Figura 77, temos a distribuição populacional de uma variável quantitativa qualquer de interesse. Ela apresenta média populacional (μ) igual a **416,99**, e variância populacional (σ^2) igual a **89554,51264**.

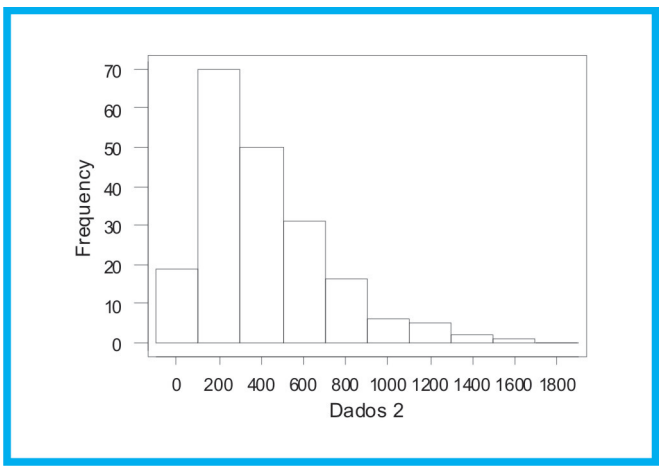


Figura 77: Distribuição populacional de uma variável quantitativa
Fonte: adaptada pelo autor de Minitab®

Observe que a distribuição é assimétrica, ou seja, não é normal! Vamos imaginar que seja possível retirar várias amostras aleatórias (com reposição) desta população, medir os valores da variável e calcular a média da variável em cada amostra. Posteriormente, construiremos um histograma das médias das amostras, e calcularemos a média das médias e a variância das médias.

Vamos começar com 40 amostras aleatórias de dois elementos cada. Veja a Figura 78.

A retirada das amostras foi efetuada através do pacote estatístico Minitab®.

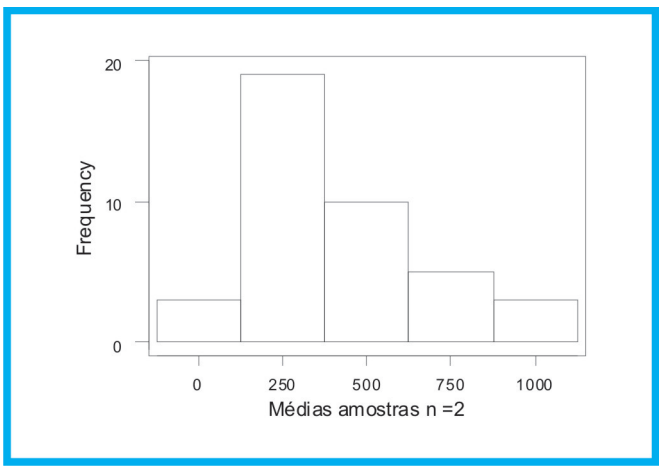


Figura 78: Distribuição amostral da média (n = 2)
Fonte: adaptada pelo autor de Minitab®

A média das médias amostrais vale 423,8875, e a variância das médias amostrais vale 67528,98666. E, lembrando do exemplo anterior, podemos calcular o quociente variância populacional pelo tamanho da amostra: $\sigma^2/n = 89554,51264/2 = 44777,25632$.

Observando o histograma, vemos que a distribuição das médias, para amostras de dois elementos, continua assimétrica, e o valor da média das médias amostrais (423,8875) não está muito próximo da média populacional (416,99), bem como a variância das médias amostrais (67528,98666), distante de 2/n = 44777,25632.

Obviamente, o tamanho da amostra utilizada (dois elementos) ainda não foi grande o bastante para levar aos resultados obtidos no Exemplo 2 (provavelmente porque a distribuição da população é assimétrica). Vamos agora ver os resultados obtidos para 40 amostras aleatórias de quatro elementos cada. O histograma das médias está na Figura 79.

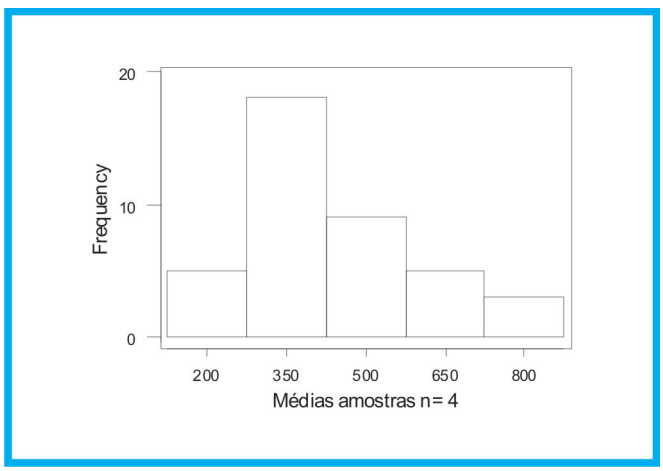


Figura 79: Distribuição amostral da média (n = 4)
Fonte: adaptada pelo autor de Minitab®

A média das médias amostrais vale 444,5375, e a variância das médias amostrais vale 26464,3269. E, lembrando do exemplo anterior, podemos calcular o quociente variância populacional pelo tamanho da amostra: 2/n = 89554,51264/4 = 22388,62816.

Observando o histograma, vemos que a distribuição das médias, para amostras de quatro elementos, continua assimétrica, e o valor da média das médias amostrais (444,5375) não está muito próximo da

média populacional (416,99), mas a variância das médias amostrais (26464,3269) aproxima-se mais de $\sigma^2/n = 22388,62816$.

Novamente, o tamanho da amostra utilizada (quatro elementos) ainda não foi o bastante para levar aos resultados obtidos no Exemplo 2. Vamos agora ver os resultados obtidos para 40 amostras aleatórias de 16 elementos cada. O histograma das médias está na Figura 80.

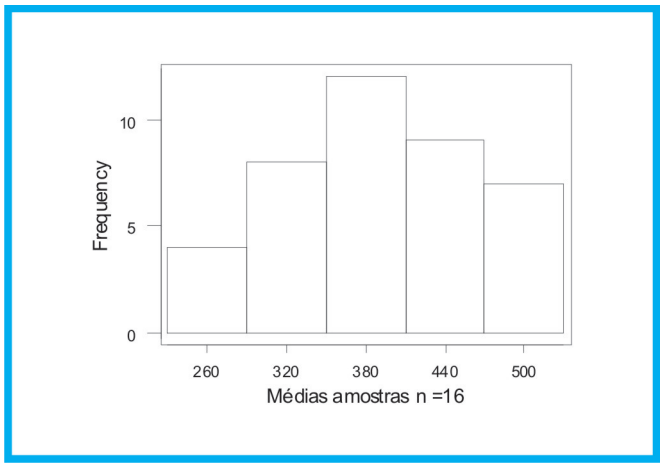


Figura 80: Distribuição amostral da média (n = 16)
Fonte: adaptada pelo autor de Minitab®

A média das médias amostrais vale 394,4922, e a variância das médias amostrais vale 5568,3945. E, lembrando do exemplo anterior, podemos calcular o quociente variância populacional pelo tamanho da amostra: $\sigma^2/n = 89554,51264/16 = 5597,1577$.

Observando o histograma, vemos que a distribuição das médias, para amostras de 16 elementos, está mais próxima da simetria, o valor da média das médias amostrais (394,4922) está mais próximo da média populacional (416,99), e a variância das médias amostrais (5568,3945) aproxima-se bastante de $\sigma^2/n = 5597,1577$.

Estamos muito próximos de obter um comportamento simétrico e aproximadamente normal para o histograma das médias amostrais. Se retirarmos mais 40 amostras, mas agora com 30 elementos em cada, o resultado poderá ser visto na Figura 81.

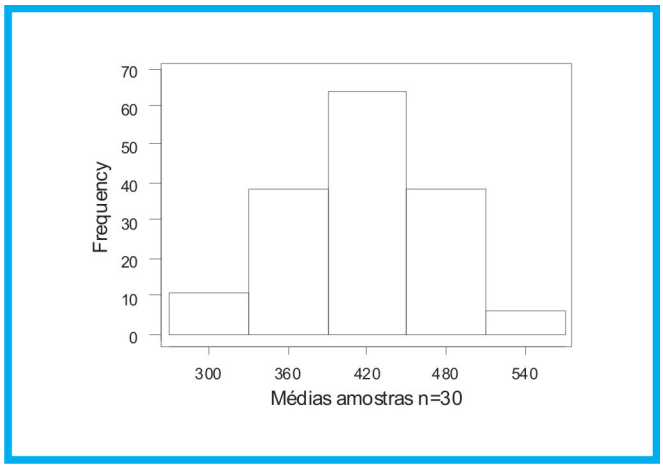


Figura 81: Distribuição amostral da média (n = 16)
Fonte: adaptada pelo autor de Minitab®

A média das médias amostrais vale 421,9217, e a variância das médias amostrais vale 2945,1326. E, lembrando do exemplo anterior, podemos calcular o quociente variância populacional pelo tamanho da amostra: $\sigma^2/n = 89554,51264/30 = 2985,1508$.

Observando o histograma, vemos que a distribuição das médias, para amostras de 30 elementos, é virtualmente normal, o valor da média das médias amostrais (421,9217) está bem próximo da média populacional (416,99), e a variância das médias amostrais (2945,1326) também é muito próxima de $\sigma^2/n = 2985,1508$.

Podemos, então, enunciar os teoremas:

Teorema das Combinações Lineares

Se a variável de interesse segue uma distribuição normal na população, a distribuição amostral das médias de amostras aleatórias retiradas desta população também será normal, independentemente do tamanho destas 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 desta população será normal se o tamanho destas 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.

Este "suficientemente grande" varia de distribuição para distribuição; 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" aquelas que possuem mais de 30 elementos; a partir deste tamanho, a aproximação poderia ser feita sem maiores preocupações.

Distribuição amostral da proporção

Vamos estudar as particularidades da distribuição amostral da proporção através de um exemplo.

Neste Exemplo 4, pense agora em uma variável qualitativa que pode assumir apenas dois valores e que constitui a seguinte população: (□, □, □, □, ■)

Vamos supor que há interesse no valor ¦ (este valor seria o nosso "sucesso"). A proporção deste valor na população (o valor do parâmetro) será $\pi = 1/5$.

Se retirarmos todas as amostras aleatórias de dois elementos (com reposição) possíveis desta população ($n = 2$), teremos os seguintes resultados:

Há 25 amostras possíveis.

(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, □)	(□, □)	(□, □)	(□, □)	(□, ■)
(□, ■)	(□, ■)	(□, ■)	(□, ■)	(■, ■)

Figura 82: Amostras de tamanho 2 para proporção
Fonte: elaborada pelo autor

Observe 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 esta proporção amostral de **p**, teremos os seguintes resultados:

$$p = \begin{matrix} (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (1/2) & (1/2) & (1/2) & (1/2) & (1) \end{matrix}$$

Calculando a média (valor esperado) e a variância das proporções acima, teremos:

$$\bar{X} = E(p) = \frac{1}{5} = \pi \qquad s^2 = 0{,}08 = \frac{\left(\frac{1}{5}\right) \times \left(1 - \frac{1}{5}\right)}{2} = \frac{\pi \times (1 - \pi)}{n}$$

Voltaremos a analisar o significado deste resultado quando estudarmos estimação por ponto.

Observe 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.

Isto também é decorrência do Teorema Central do Limite.

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** × **p**) e o produto do número de realizações pela probabilidade de "fracasso" (**n** × [**1 – p**]) forem ambos maiores ou iguais a 5. E esta distribuição normal teria média igual a **n** × **p** e variância igual a **n** × **p** × (**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** × (**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 o seu estudo é imprescindível. Ressalte-se, porém, que a sua aplicação costuma se resumir 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 assumir a validade do Teorema Central do Limite, é preciso usar técnicas de inferência não paramétrica (que nós não veremos nesta disciplina).

Saiba mais...

- Sobre distribuição amostral:

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 7.

STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: 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, D. M.; et al. *Estatística*: teoria e aplicações – usando Microsoft Excel em português. 5. ed. Rio de Janeiro: LTC, 2006, capítulo 5.

RESUMO

O resumo desta Unidade está mostrado na Figura 83:

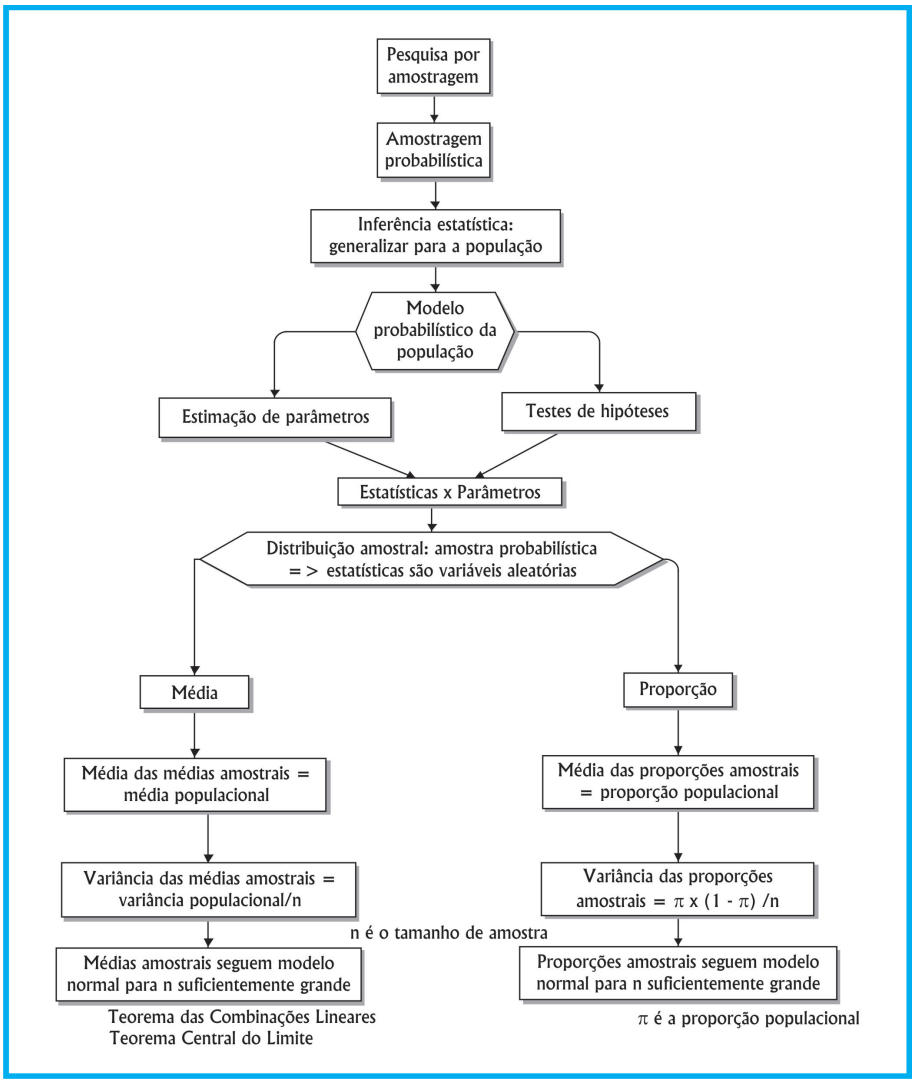


Figura 83: Resumo da Unidade 8
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Caro estudante!

Esta Unidade foi muito importante para o seu aprendizado, pois lhe dará base para chegar à inferência estatística propriamente dita, assunto que será tema de discussão nas Unidades 9 e 10. Vimos até agora sobre a inferência estatística e distribuição amostral, seu modelo probabilístico e testes de hipóteses. Chegamos ao final desta Unidade e à continuidade da aprendizagem proposta desde o início deste material. Interaja com seus colegas, responda as atividades de aprendizagem e visite o Ambiente Virtual de Ensino-Aprendizagem, espaço este que contemplará suas possíveis dúvidas. Procure seu tutor e solicite todas as informações necessárias para o seu aprendizado. Bons estudos!

UNIDADE

9

Estimação de parâmetros

Objetivo

Nesta Unidade, você vai conhecer e aplicar os conceitos de estimação de parâmetros por ponto e por intervalo de média e proporção, e aprenderá como calcular o tamanho mínimo de amostra necessário para a estimação por intervalo.

Estimação por ponto de parâmetros

Prezado estudante!

Na Unidade 8, você viu o conceito de distribuição amostral e observou a importância do modelo normal. Nesta Unidade, você vai aprender como aplicar estes 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 ou características dos modelos probabilísticos que descrevem as variáveis na população. Por ser inviável ou inconveniente pesquisar toda a população, coletamos uma amostra para estudá-la. Os resultados da amostra podem ser, então, usados para fazer afirmações probabilísticas sobre o parâmetro de interesse: definir um intervalo possível para os valores do parâmetro e calcular a probabilidade de que o valor real do parâmetro esteja 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 (para esta última, apresentamos uma primeira expressão de cálculo 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 da 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.

Como os parâmetros serão estimados através das estatísticas, estimadores, 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 **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 **T** é um estimador **justo** (não tendencioso) de um parâmetro **θ** quando o valor esperado de **T** é igual ao valor do parâmetro **θ** a ser estimado: E(T) = **θ**.

2) Um estimador 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: $l_{imn \to \infty} V(T) = 0$.

3) Se há dois estimadores justos de um parâmetro, o mais **eficiente** é aquele que apresentar a menor variância.

GLOSSÁRIO

*Estimação por ponto – tipo de estimação de parâmetros que procura identificar qual é o melhor estimador para um parâmetro populacional a partir das várias estatísticas amostrais disponíveis, seguindo alguns critérios. Fonte: Barbetta, Reis e Bornia (2004) e Silva (1999).

Conforme foi dito na introdução desta Unidade, estamos interessados em estimar dois parâmetros: média e proporção populacional. Vamos, então, buscar 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 estes parâmetros, usando os critérios definidos acima.

Lembrando dos Exemplos 2, 3 e 4 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?

Lembre-se de que, após retirar todas as amostras aleatórias possíveis daquela população, calculamos a média de cada amostra e, posteriormente, a média dessas médias. Constatou-se 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:

$$E(\bar{x}) = \mu \qquad V(\bar{x}) = \frac{\sigma^2}{n}$$

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; e
- 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 4. Qual das estatísticas disponíveis seria o melhor estimador?

Lembre-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. Constatou-se 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:

$$E(p) = \pi \qquad V(p) = \frac{\pi \times (1 - \pi)}{n}$$

O melhor estimador da média populacional **μ** 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. Este 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 extraindo a raiz quadrada da variância amostral.

O problema da estimação por ponto é que geralmente só dispomos de uma amostra aleatória. Intuitivamente, qual será a probabilidade de que a média ou proporção amostral, de uma amostra aleatória, coincida exatamente com o valor do parâmetro? É como pescar usando uma lança de bambu... É preciso muita habilidade para pegar o peixe... Mas, se você puder usar uma rede, ficará bem mais fácil. Esta "rede" é a estimação por intervalo.

Estimação por intervalo de parâmetros

Fazer uma estimação por intervalo de um parâmetro é efetuar uma afirmação probabilística sobre este parâmetro, indicando uma faixa de possíveis valores.

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 uma determinada população.

Intuitivamente, percebemos que as estatísticas calculadas nessa única amostra, mesmo sendo os melhores estimadores para os parâmetros de interesse, terão uma 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 uma certa probabilidade chamada de nível de confiança (que costuma ser simbolizado como $\mathbf{1 - \alpha}$) de conter o valor real do parâmetro e a probabilidade de que esta faixa realmente contenha o valor real do parâmetro. A probabilidade de que o intervalo de confiança não contenha o valor real do parâmetro é chamada de nível de significância ($\boldsymbol{\alpha}$), e o valor desta probabilidade será o complementar do **nível de confiança***. É comum definir o nível de significância como uma probabilidade máxima de erro, um risco máximo admissível.

A determinação do intervalo de confiança para um determinado parâmetro resume-se basicamente a definir o limite inferior e o limite superior do intervalo, supondo um determinado **nível de 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 $\boldsymbol{\mu}$ e proporção populacional π), a distribuição amostral dos estimadores (média amostral $\bar{x}$ e proporção amostral $\mathbf{p}$, respectivamente) pode ser aproximada por uma distribuição normal: o intervalo de confiança será, então, simétrico em relação ao valor calculado

GLOSSÁRIO

*Intervalo de confiança – faixa de valores da estatística usada como estimador, dentro da qual há uma probabilidade conhecida de que o verdadeiro valor do parâmetro esteja. Sinônimo de estimação por intervalo. Fonte: Barbetta, Reis e Bornia (2004).

*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 Sclovel (2006).

*Nível de significância – complementar do nível de confiança, a probabilidade de que o intervalo de confiança não contenha o valor real do parâmetro. Fonte: Barbetta, Reis e Bornia (2004).

da estimativa (média ou proporção amostral), com base na amostra aleatória coletada (Figura 84):

Figura 84: Intervalo de confiança para um modelo normal
Fonte: elaborada pelo autor

Onde: $\mathbf{L_i}$ é o limite inferior, e $\mathbf{L_s}$ é o limite superior do intervalo de confiança; $\mathbf{1 - \alpha}$ é o nível de confiança estabelecido, observando que o valor do nível de significância $\boldsymbol{\alpha}$ é dividido igualmente entre os valores abaixo de $\mathbf{L_i}$ e acima de $\mathbf{L_s}$.

Para obter os limites em função do nível de confiança, devemos utilizar a distribuição normal-padrão (variável **Z** com média zero e variância 1): fixar um certo valor de probabilidade, obter o valor de **Z** correspondente, e substituir o valor em **Z** = (**x** – "média") "desvio-padrão", para obter o valor x (valor correspondente ao valor de Z para a probabilidade fixada). Observe a Figura 85:

Foram colocados entre aspas, porque dependerão dos parâmetros sob análise e de outros fatores.

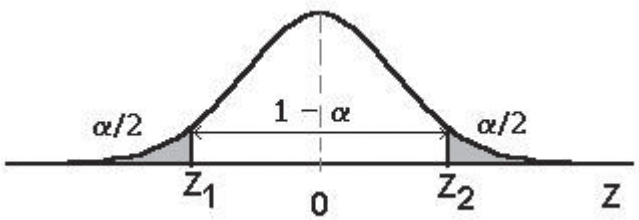


Figura 85: Intervalo de confiança para a distribuição normal-padrão
Fonte: elaborada pelo autor

O limite $\mathbf{L_i}$ (inferior) corresponde a $\mathbf{Z_1}$, e o limite $\mathbf{L_s}$ (superior) corresponde a $\mathbf{Z_2}$. 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 $\mathbf{Z_1}$ e $\mathbf{Z_2}$ serão iguais em módulo ($\mathbf{Z_1}$ será negativo, e $\mathbf{Z_2}$, positivo):

$\mathbf{Z_1}$ será um valor de $\mathbf{Z}$ tal que $P(Z \leq Z_1) = \frac{\alpha}{2}$, e $\mathbf{Z_2}$ será um valor tal que $P(Z \leq Z_2) = 1 - \frac{\alpha}{2}$

Então, obteremos os valores dos limites através das expressões:

$Z_1 = (L_i - \text{“média”})/\text{“desvio-padrão”} \Rightarrow L_i = \text{“média”} \times Z_1 \times \text{“desvio-padrão”}$
$Z_2 = (L_s - \text{“média”})/\text{“desvio-padrão”} \Rightarrow L_S = \text{“média”} + Z_2 \times \text{“desvio-padrão”}$

Como $\mathbf{Z_1 = -Z_2}$, podemos substituir:

$L_i = \text{“média”} - Z_2 \times \text{“desvio-padrão”}$
$L_s = \text{“média”} + Z_2 \times \text{“desvio-padrão”}$

E este valor $\mathbf{Z_2}$ costuma ser chamado de $\mathbf{Z_{crítico}}$, porque corresponde aos limites do intervalo:

$L_i = \text{“média”} - Z_{crítico} \times \text{“desvio-padrão”}$
$L_s = \text{“média”} + Z_{crítico} \times \text{“desvio-padrão”}$

Reparem que o mesmo valor é somado e subtraído da “média”. Este valor é chamado de semi-intervalo ou precisão do intervalo, ou margem de erro, $\mathbf{e_0}$:

$e_0 = Z_{crítico} \times \text{“desvio-padrão”}$

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 das expressões anteriores:

$L_i = \text{“média”} - Z_{crítico} \times \text{“desvio-padrão”} = \text{“média”} - e_0$
$L_s = \text{“média”} + Z_{crítico} \times \text{“desvio-padrão”} = \text{“média”} + e_0$

Neste 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 $\mathbf{e_0}$ dependerá de outros aspectos.

a) Se a variância populacional $\boldsymbol{\sigma^2}$ da variável (cuja média populacional queremos estimar) for conhecida.

Neste caso, a variância amostral da média poderá ser calculada através da expressão:

$V(\bar{x}) = \frac{\sigma^2}{n}$, e, por conseguinte, o "desvio-padrão" será

desvio-padrão $= \frac{\sigma}{\sqrt{n}}$

E $\mathbf{e_0}$ será: $e_0 = Z_{critico} \times \frac{\sigma}{\sqrt{n}}$

Bastará, então, fixar o nível de confiança (ou de significância) para obter $\mathbf{Z_{crítico}}$ através da Tabela disponível no Ambiente Virtual e calcular $\mathbf{e_0}$.

b) Se a variância populacional $\boldsymbol{\sigma^2}$ da variável for desconhecida.

Naturalmente, este é o caso mais encontrado na prática. Como se deve proceder? Dependerá do tamanho da amostra.

b.1) Grandes amostras (mais de 30 elementos)

Nestes casos, procede-se como no item anterior, apenas fazendo com que $\boldsymbol{\sigma = s}$, 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)

Nestes 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 7.

Quando a variância populacional da variável é desconhecida e a amostra tem até 30 elementos, substitui-se $\boldsymbol{\sigma}$ por **s** e **Z** por $\mathbf{t_{n-1}}$ em todas as expressões para determinação dos limites do intervalo de confiança, obtendo:

L_i = “média” – $t_{n-1\ crítico}$ x “desvio-padrão” = “média” – e_0
L_s = “média” + $t_{n-1\ crítico}$ x “desvio-padrão” = “média” + e_0

E $\mathbf{e_0}$ será:

$$e_0 = t_{n-1,critico} \times \frac{s}{\sqrt{n}}$$

Os valores de $\mathbf{t_{n-1,crítico}}$ podem ser obtidos de forma semelhante aos de $\mathbf{Z_{crítico}}$, definindo o nível de confiança (ou de significância), mas precisam também da definição do número de graus de liberdade (**n – 1**): tendo estes 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 e0 precisam ser corrigidos. Caso contrário, os limites dos intervalos não serão acusados. A correção é mostrada na equação a seguir:

$$e_{0_{corrigido}} = e_0 \times \sqrt{\frac{N-n}{N-1}}$$

Vamos ver um exemplo.

Neste primeiro exemplo, retirou-se uma amostra aleatória de quatro elementos de uma produção de cortes bovinos no intuito de estimar a média do peso do corte. Obtiveram-se média de 8,2 kg e desvio-padrão de 0,4 kg. Supondo população normal; determinar 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 **α** = 0,01, e 1 – **α** = 0,99.

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%.

As estatísticas disponíveis são:

média amostral = 8,2 kg

s = 0,4 kg

n = quatro elementos.

Definição da variável de teste: 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 $\mathbf{t_{n-1}}$.

Encontrar o valor de $\mathbf{t_{n-1,crítico}}$: como o intervalo de confiança para a média é bilateral, teremos uma situação semelhante à da Figura 86:

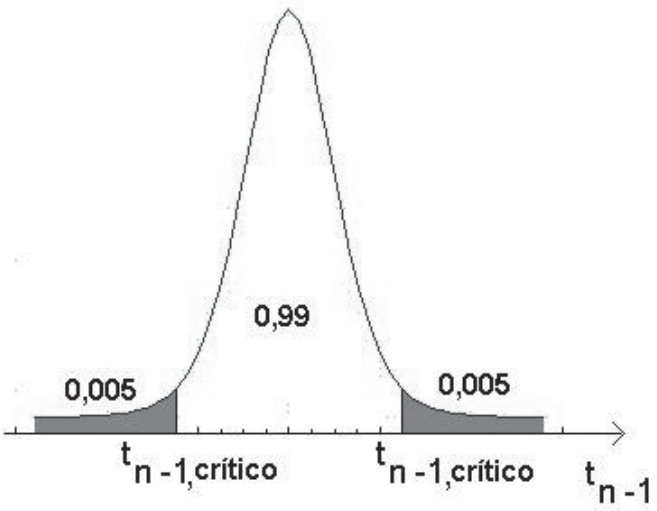


Figura 86: Distribuição t de Student para 99% de confiança
Fonte: elaborada pelo autor

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 acima: os valores críticos serão $\mathbf{t_{3;0,005}}$ e $\mathbf{t_{3;0,995}}$, os quais serão iguais em módulo. E o valor de $\mathbf{t_{n-1,crítico}}$ será igual a **5,84** (em módulo).

Determinam-se os limites do intervalo através da expressão abaixo (cujo resultado será somado e subtraído da média amostral) para determinar os limites do intervalo:

$$e_0 = \frac{t_{n-1,crítico} * s}{\sqrt{n}} = \frac{5,84 * 0,4}{\sqrt{4}} = 1,168kg$$

$L_I = \bar{x} - e_0 = 8,2 - 1,168 = 7,032kg$

$L_S = \bar{x} + e_0 = 8,2 + 1,168 = 9,368kg$

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 que a verdadeira média populacional do peso de corte esteja entre 7,032 e 9,368 kg.

Estimação por intervalo da proporção populacional

Anteriormente, declaramos que o melhor estimador para a proporção populacional $\boldsymbol{\pi}$ é a proporção amostral **p**. E que esta proporção amostral teria média igual a $\boldsymbol{\pi}$ e variância igual a $[\boldsymbol{\pi} \times (\mathbf{1} - \boldsymbol{\pi})]/\mathbf{n}$, onde **n** é o tamanho da amostra aleatória. A distribuição da proporção amostral **p** é binomial, e sabe-se que a distribuição binomial pode ser aproximada por uma normal se algumas condições forem satisfeitas:

Se $\mathbf{n} \times \boldsymbol{\pi} \geq 5$ E $\mathbf{n} \times (1 - \boldsymbol{\pi}) \geq 5$.

Ora, se $\boldsymbol{\pi}$ fosse conhecido, não estaríamos aqui nos preocupando com a sua estimação por intervalo; assim, vamos verificar se é possível aproximar a distribuição binomial de **p** por uma normal se:

$\mathbf{n} \times \mathbf{p} \geq 5$ E $\mathbf{n} \times (\mathbf{1} - \mathbf{p}) \geq 5$, ou seja, usando o próprio valor da proporção amostral observada (trata-se de uma aproximação razoável).

Se e somente se estas duas condições forem satisfeitas, poderemos usar as expressões abaixo (lembrando das expressões anteriores):

L_i = "média" – $Z_{crítico}$ x "desvio-padrão" = "média" – e_0
L_s = "média" + $Z_{crítico}$ x "desvio-padrão" = "média" + e_0

Neste 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 $\boldsymbol{\pi}$ é desconhecido, usaremos a proporção amostral **p** como aproximação.

Então, $\mathbf{e_0}$ será:

$$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}}$$

Bastará, então, fixar o nível de confiança (ou de significância), $\mathbf{Z_{crítico}}$, e calcular $\mathbf{e_0}$.

Novamente, precisamos corrigir o valor de e0 para o caso de população finita:

$$e_{0_{corrigido}} = e_0 \times \sqrt{\frac{N-n}{N-1}}$$

Em suma, a estimação por intervalo da média e da proporção populacional consiste basicamente em calcular a amplitude do semi-intervalo (o $\mathbf{e_0}$), de acordo com as condições do problema sob análise.

- Para a média, observar se é viável considerar que a distribuição da variável na população é normal, ou que a amostra seja 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 tem até 30 elementos, usar a variável $\mathbf{t_{n-1}}$ da distribuição de Student.
- 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 $\boldsymbol{\pi}$ de peças defeituosas.

Adotou-se um nível de significância de 5%; então, $\boldsymbol{\alpha}$ = 0,05, e 1 – $\boldsymbol{\alpha}$ = 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** x **p** = 1.000 x 0,035 = 35 > 5, e **n** x (**1– p**) = 1.000 x 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 $\mathbf{Z}_{\text{crítico}}$: como o intervalo de confiança para a média é bilateral, teremos uma situação semelhante à da figura abaixo:

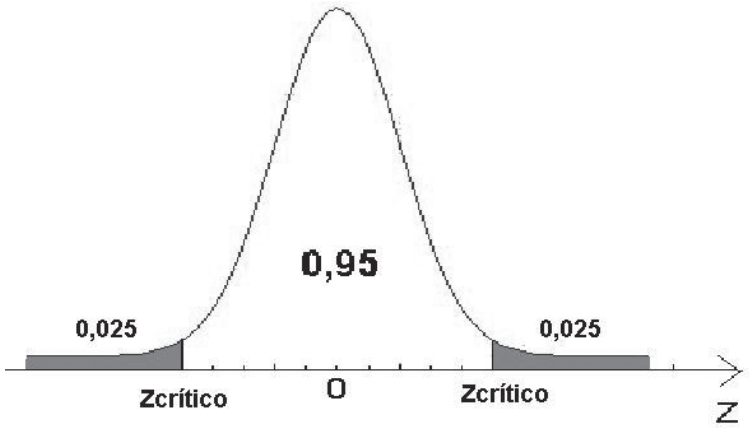


Figura 87: Distribuição normal-padrão para 95% de confiança
Fonte: elaborada pelo autor

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 87 acima: os valores críticos serão $\mathbf{Z}_{0,025}$ e $\mathbf{Z}_{0,975}$, os quais serão iguais em módulo. E o valor de $\mathbf{Z}_{\text{crítico}}$ será igual a 1,96 (em módulo).

Passa-se agora à determinação dos limites do intervalo, através da expressão abaixo, cujo resultado será somado e subtraído da proporção amostral de peças defeituosas, para determinar os limites do intervalo:

$$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}} = 1{,}96 \times \sqrt{\frac{0{,}035 \times 0{,}965}{1000}} = 0{,}0114$$

$L_I = p - e_0 = 0{,}035 - 0{,}0114 = 0{,}0126$

$L_S = p + e_0 = 0{,}035 + 0{,}0114 = 0{,}0464$

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 que a verdadeira proporção populacional de plantas atacadas pelo fungo esteja entre 2,36% e 4,64%.

Tamanho mínimo de amostra para estimação por intervalo

Como foi observado nos itens anteriores, a determinação dos limites de um intervalo de confiança (determinação do $\mathbf{e_0}$) 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 $\mathbf{e_0}$) para o intervalo de confiança: é muito comum querermos estabelecer previamente qual será a faixa de variação de um determinado parâmetro, com uma 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 $\mathbf{e_0}$, resultando em um intervalo de confiança mais "largo", com menor precisão; e
- se resolvermos aumentar a precisão (menor valor de $\mathbf{e_0}$) e obter um intervalo de confiança mais "estreito", teremos uma queda no nível de confiança.

A solução para o dilema acima é obter um **tamanho mínimo de amostra** capaz de atender simultaneamente ao nível de confiança (ou de significância) e à precisão ($\mathbf{e_0}$) especificados. Como as expressões de $\mathbf{e_0}$ 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 ($\mathbf{e_0}$).

Tamanho mínimo de amostra para estimação por intervalo da média populacional

a) Variância populacional conhecida

$$e_0 = Z_{critico} \times \frac{\sigma}{\sqrt{n}} \text{ isolando } \mathbf{n}\text{: } n = \left(\frac{Z_{critico} \times \sigma}{e_0} \right)^2$$

Neste caso, basta especificar o valor de $\mathbf{e_0}$ (na **mesma unidade** do desvio-padrão populacional □), e o nível de confiança (que será usado para encontrar o $\mathbf{Z_{crítico}}$) e calcular o tamanho mínimo de amostra.

b) Variância populacional desconhecida

$$e_0 = t_{n-1,critico} \times \frac{s}{\sqrt{n}} \text{ isolando } \mathbf{n}\text{: } n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2$$

O procedimento, neste 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 desta amostra, são calculadas as estatísticas (entre elas, o desvio-padrão amostral **s**) que são substituídas na expressão acima.

Se **n** ≤ **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 exigidas, 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, seja para o intervalo de confiança de média ou proporção, seja através da seguinte fórmula:

$$n_{corrigido} = \frac{N \times n}{N + n}, \text{ onde N é o tamanho da população}$$

Assim procedendo, evitamos o inconveniente de obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

GLOSSÁRIO

*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).

Considere, neste Exemplo 3, os dados do Exemplo 1. Para estimar a média, com 1% de significância e precisão de 0,2 kg, esta 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 $\mathbf{t_{n-1}}$. Assim, será usada a seguinte expressão para calcular o tamanho mínimo de amostra para a estimação por intervalo da média populacional:

$$n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2$$

O nível de significância é o mesmo do item a. Sendo assim, o valor crítico continuará sendo o mesmo: $\mathbf{t_{n-1,crítico}} = 5{,}84$. O desvio-padrão amostral vale 0,4 kg, e o valor de $\mathbf{e_0}$, a precisão, foi fixado em 0,2 kg. Basta, então, substituir os valores na expressão:

$$n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2 = \left(\frac{5{,}84 \times 0{,}4}{0{,}2} \right)^2 = 136{,}42 \cong 137 \text{ elementos}$$

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:

$$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}} \text{ isolando n: } n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1-p)$$

É necessário especificar o nível de confiança (ou de significância) que será usado para encontrar o $\mathbf{Z_{crítico}}$, e o valor de $\mathbf{e_0}$ (tomando o cuidado de que tanto $\mathbf{e_0}$ 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** x (**1 – p**), que ocorrerá quando ambas as proporções forem iguais a 0,5 (50%).

Esta 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 poderiam prejudicar muito o resultado da pesquisa. Paga-se, então, o preço de ter uma amostra substancialmente maior do que talvez fosse necessário.

Conforme visto na Unidade 2, se o tamanho da população for conhecido, é recomendável corrigir o tamanho da amostra obtida, seja para o intervalo de confiança de média ou proporção, seja através da seguinte fórmula:

$n_{corrigido} = \frac{N \times n}{N + n}$, onde N é o tamanho da população

Assim procedendo, evitamos o inconveniente de obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

Neste quarto exemplo, considere o caso do Exemplo 2. Supondo 99% de confiança e precisão de 1%, esta amostra é suficiente para estimar a proporção populacional?

De acordo com o Exemplo 2, é possível utilizar a aproximação pela distribuição normal. A expressão para o cálculo do tamanho mínimo de amostra para a proporção populacional será:

$$n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1 - p)$$

Os valores de **p** e **1 – p** já são conhecidos:

p = 0,035 **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 $\mathbf{Z_{0,005}}$ e $\mathbf{Z_{0,995}}$, os quais serão iguais em módulo. E o valor de $\mathbf{Z_{crítico}}$ será igual a 2,58 (em módulo).

A precisão foi fixada em 1% (0,01). Substituindo os valores na expressão acima:

$$n = \left(\frac{Z_{crítico}}{e_0} \right)^2 \times p \times (1 - p) = \left(\frac{2,58}{0,01} \right)^2 \times 0,035 \times 0,965 = 2.248,14 \cong 2.249$$

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 a precisão exigidas. Recomenda-se o retorno à população para a retirada aleatória de mais 1.249 peças.

Visto tudo o que estudamos, agora 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 declaram 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 23)?

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 23: Resultados de uma pesquisa eleitoral municipal
Fonte: fictícia, elaborado pelo autor.

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, P. A.; REIS, M. M.; BORNIA, A.C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 7.

- Sobre estimadores e intervalos de confiança para variância:
TRIOLA, M. *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 de Ensino-Aprendizagem.

- Sobre a utilização do Microsoft Excel para realizar estimação por intervalo:
LEVINE, D. M.; et al. *Estatística:* teoria e aplicações – usando Microsoft Excel em português. 5. ed. Rio de Janeiro: LTC, 2006, capítulo 6.

RESUMO

O resumo desta Unidade está mostrado na Figura 88:

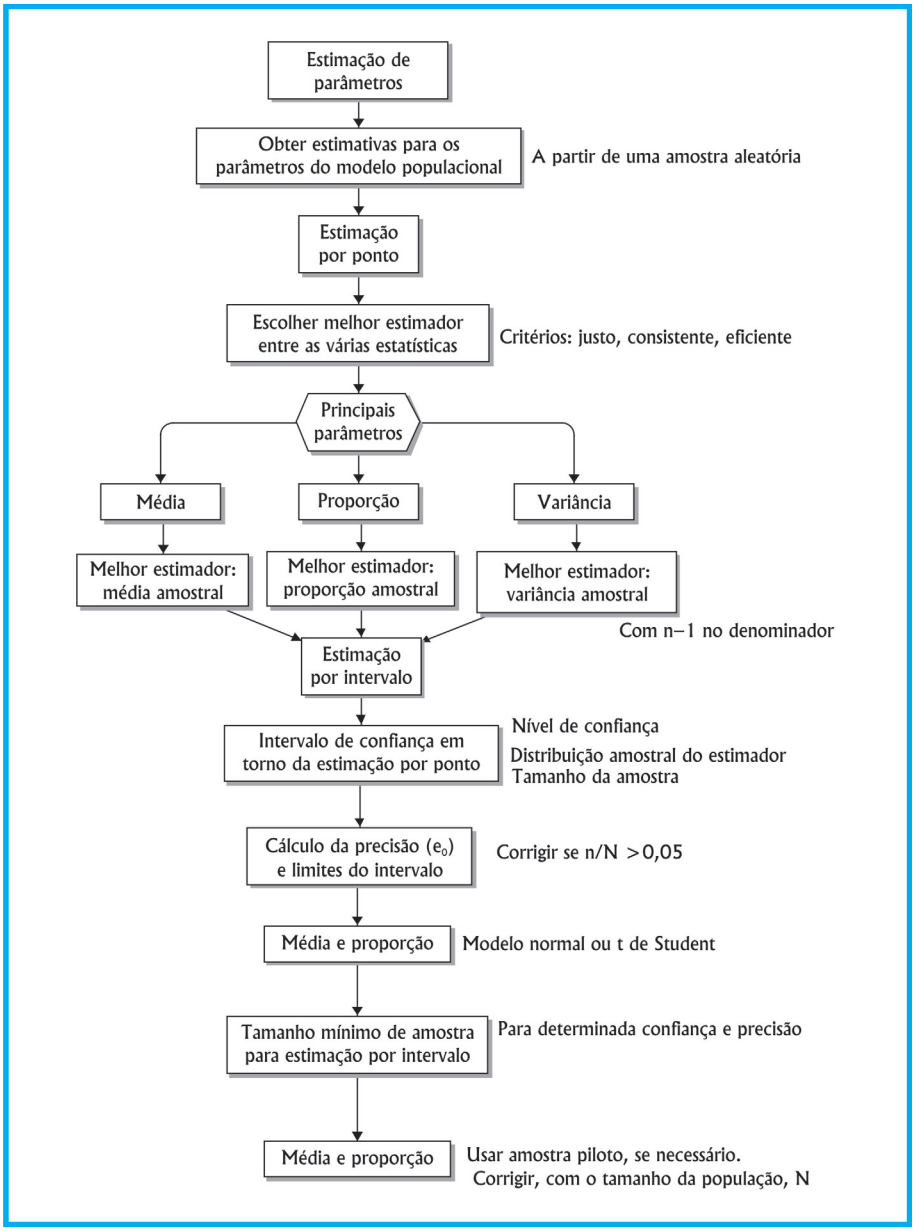


Figura 88: Resumo da Unidade 9
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

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 este 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 as Atividades de aprendizagem.

UNIDADE

10

Testes de hipóteses

Objetivo

Nesta Unidade, você vai compreender e aplicar 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. Você aprenderá também qual é a importância de tais conceitos para o dia-a-dia do administrador.

Lógica dos testes de hipóteses

Caro estudante, você viu anteriormente que uma determinada população pode ser descrita através de um modelo probabilístico, que apresenta características e parâmetros. Muitas vezes, estes parâmetros são desconhecidos, e há interesse em estimá-los para obter um melhor conhecimento sobre a população: retira-se, então, uma amostra aleatória da população, e através das técnicas de **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 probabi-lística desta mesma população – Unidade 9) procura-se obter uma estimativa de algum parâmetro de interesse, e associamos uma probabilidade de que a estimativa esteja correta. Nesta 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 probabilístico da população a partir de uma amostra aleatória desta população), a outra grande subdivisão constitui os **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). Vamos saber mais!

Para saber mais

*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 e "testes livres de distribuição", porque os testes não paramétricos não exigem que os dados tenham uma aderência a um certo modelo.

Contrariamente à estimação de parâmetros, os testes de hipóteses permitem fazer inferências sobre outras características do modelo

probabilístico da população além dos parâmetros (como, por exemplo, a forma do modelo probabilístico da população). Quando os testes são feitos sobre os parâmetros da população, são chamados de testes paramétricos, e quando são feitos sobre outras características, são chamados de **testes não paramétricos***. Não obstante, vamos nos restringir aos **testes paramétricos***: de uma média de uma variável quantitativa e de uma proporção de um dos valores de uma variável qualitativa.

A você, estudante interessado em testes não paramétricos, recomendo a referência: SIEGEL, S. *Estatística Não Paramétrica (para as Ciências do Comportamento)*. São Paulo: McGraw-Hill, 1975. É uma boa referência no assunto, em português.

Imagine-se que um 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 um determinado comportamento. Formula, então, uma hipótese estatística sobre a característica da população, e esta 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? Esta 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**.

GLOSSÁRIO

*Testes não paramétricos – testes de hipóteses sobre outros aspectos do modelo probabilístico da variável sob análise, ou alternativa aos testes paramétricos quando as condições para uso destes não forem satisfeitas.

*Testes paramétricos – testes de hipóteses sobre parâmetros do modelo probabilístico da variável sob análise.

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 estar cometendo um erro, mas com a utilização de métodos estatísticos, é possível calcular o valor desta probabilidade.

Usando outros métodos (empíricos), não há como ter idéia da chance de erro (pode ser um erro de 0% ou de 5.000%...).

O nível de significância é uma probabilidade; portanto, é um número real que varia de 0 a 1 (0 a 100%), e como é a probabilidade de se cometer um erro no teste, é interessante que seja o mais próximo possível de zero: valores típicos são 10%, 5%, 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 ele indica a confiabilidade do resultado obtido, a probabilidade de que a decisão tomada esteja correta.

GLOSSÁRIO

*Nível de significância – probabilidade arbitrada pelo pesquisador, valor máximo de erro admissível para rejeitar a hipótese nula sendo ela verdadeira. Fonte: Barbetta, Reis e Bornia (2004) e Moore, Mc Cabe, Dukworth e Sclove (2006)

Você deve estar lembrado destes dois conceitos de estimação de parâmetros: nível de confiança é a probabilidade de que o intervalo de confiança contivesse o valor real do parâmetro, e nível de significância, complementar daquele, é a probabilidade de que o intervalo não contivesse o parâmetro, em suma, a probabilidade de a estimação estar correta ou não, respectivamente.

Tipos de hipótese

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 $\mathbf{H_0}$) e a **hipótese alternativa** (denotada por $\mathbf{H_1}$ ou $\mathbf{H_a}$). Enunciar as hipóteses é o primeiro e, possivelmente, mais importante passo de um teste de hipóteses, pois todo o procedimento dependerá dele.

A hipótese nula ($\mathbf{H_0}$) 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 ($\mathbf{H_1}$), que será uma hipótese complementar de $\mathbf{H_0}$, fornecerá uma alternativa à hipótese nula: muitas vezes, é justamente o que o pesquisador quer provar (ou o 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 ($\mathbf{H_0}$): vai-se aceitar ou não a hipótese até então considerada verdadeira.

É importante ter a noção exata do que significa aceitar ou rejeitar a hipótese nula ($\mathbf{H_0}$). A decisão é tomada sobre esta 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. Já quando a decisão é por 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 $\mathbf{H_0}$ seria comparável a um veredito de não culpado, "*not guilty*", ou seja, não há provas suficientes para condenar o réu. Por outro lado, rejeitar $\mathbf{H_0}$ 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 $\mathbf{H_0}$, sendo $\mathbf{H_0}$ verdadeira**.

Tipos de teste paramétrico

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), este será **unilateral**: somente há interesse se as diferenças entre os dados da amostra e o valor de teste forem em uma determinada direção. Se a formulação da hipótese alternativa indicar que o parâmetro é diferente do valor de teste, este 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 ≠ valor de teste

A escolha do tipo de teste dependerá das condições do problema sob estudo. Sejam as três situações abaixo:

a) um novo protocolo de atendimento foi implementado no Banco RMG, visando a 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;

b) cerca de 2.000 formulários de pedidos de compra estão sendo analisados. Os clientes podem ficar insatisfeitos se

GLOSSÁRIO

*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 (2004)

*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 (2004)

houver erros nos formulários. Neste caso, admite-se que a proporção máxima de formulários com erros seja de 5%. Ou seja, um valor maior do que 5% causaria problemas. Um teste **unilateral** seria o adequado; e

c) uma peça automotiva precisa ter 100 mm de diâmetro, exatamente. Neste 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, jamais de uma eventual evidência amostral disponível.

A decisão de aceitar ou rejeitar $\mathbf{H_0}$ dependerá das **regiões de aceitação e rejeição de $\mathbf{H_0}$**, 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 (**?**) ou nível de confiança (**?1 –**) adotado; e
- de um valor crítico da estatística ou variável de teste a partir do qual a hipótese será rejeitada, e este valor dependerá, por sua vez, do nível de significância, do tipo de teste e da **distribuição amostral** do parâmetro.

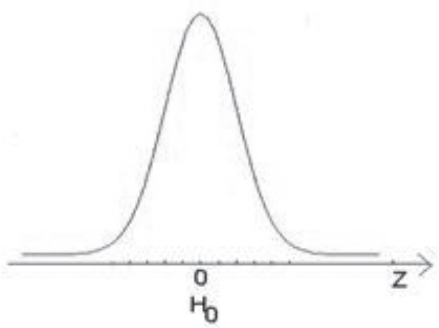
A **região de aceitação de H_0*** será a faixa de valores da estatística (ou da variável de teste) associada ao parâmetro em que as diferenças entre o que foi observado na amostra e o que era esperado não são significativas.

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**.

Esta 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 entender 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), $\mathbf{H_0}$: μ = **b**. Supondo que a distribuição amostral do parâmetro (distribuição de $\bar{x}$) seja normal, será centrada em **b**: é possível fazer a conversão para a distribuição normal-padrão (média zero e desvio-padrão 1, variável **Z**) (Figuras 89 e 90).



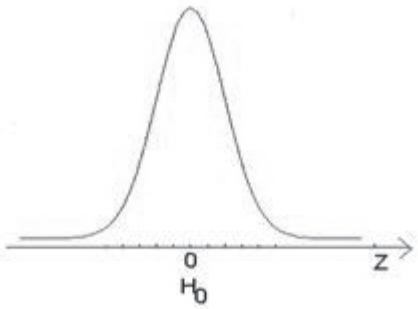
$\mathbf{H_0}$: μ = **b**

Figura 89: Hipótese nula: média populacional = b
Fonte: elaborada pelo autor

GLOSSÁRIO

*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. Fonte: Costa Neto (2002)

*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. Fonte: Costa Neto (2002)



$\mathbf{H_0}$: $\mu = \mathbf{0}$

Figura 90: Hipótese nula normal-padrão
Fonte: elaborada pelo autor

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 $\mathbf{H_0}$.

Se a hipótese alternativa fosse $\mathbf{H_1}$: $\boldsymbol{\mu} < \mathbf{b}$ ($\mathbf{H_1}$: $\boldsymbol{\mu} < \mathbf{0}$), ou seja, se o teste fosse unilateral à esquerda, a região de rejeição de $\mathbf{H_0}$ seria (Figuras 91 e 92):

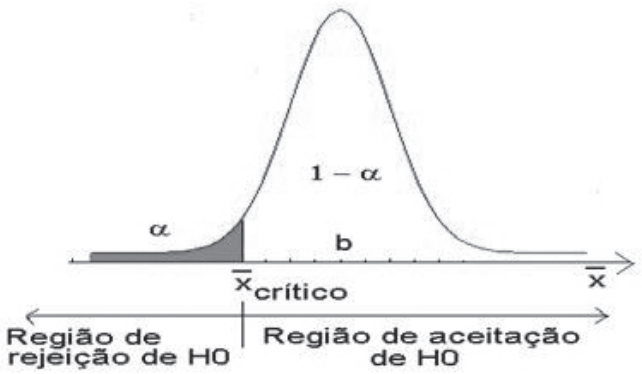


Figura 91: H_1: $\mu < \mathbf{b}$
Fonte: elaborada pelo autor

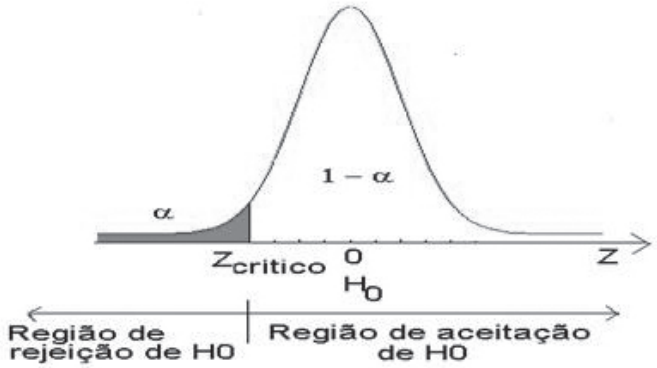


Figura 92: H_1: $\mu < \mathbf{0}$
Fonte: elaborada pelo autor

Observe que há um valor crítico de : 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 $\bar{x}$. $\boldsymbol{\mu_0}$ é o valor sob teste (**b** no exemplo), e $\boldsymbol{\sigma}$ é um desvio-padrão (cujo valor será explicitado posteriormente).

A decisão será tomada comparando o valor da média amostral $\bar{x}$ com o valor crítico desta mesma média: se for menor do que o valor crítico $\bar{x}_{crítico}$ (ou seja, está na região de **rejeição de H_0**), então 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 $\mathbf{Z_{crítico}}$ 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 $\boldsymbol{\alpha}$ 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 $\boldsymbol{\alpha}$ é a probabilidade de rejeitar $\mathbf{H_0}$ sendo $\mathbf{H_0}$ verdadeira.

Probabilidade de tomar uma decisão errada fixada pelo pesquisador.

Neste ponto, é importante ressaltar um item 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 $\mathbf{Z}_{\text{crítico}}$ será NEGATIVO, uma vez que a região de rejeição de $\mathbf{H}_0$ está à ESQUERDA de 0 (menor do que zero). No teste unilateral, à direita, que veremos a seguir, o valor de $\mathbf{Z}_{\text{crítico}}$ será positivo, pois a região de rejeição de $\mathbf{H}_0$ estará à DIREITA de 0 (maior do que zero). Se, por exemplo, o nível de significância fosse de 5% (0,05), o valor de $\mathbf{Z}_{\text{crítico}}$ 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 $Z = (\bar{x} - \boldsymbol{\mu}_0)/\sigma$ substituindo **Z** por – 1,645.

O sinal correto é importante para que o valor seja coerente com a posição da região de rejeição de $\mathbf{H}_0$.

Se a hipótese alternativa fosse $\mathbf{H}_1$: $\boldsymbol{\mu} > \mathbf{b}$ ($\mathbf{H}_1$: $\boldsymbol{\mu} > \mathbf{0}$), ou seja, se o teste fosse unilateral à direita, a região de rejeição de $\mathbf{H}_0$ seria (Figuras 93 e 94)

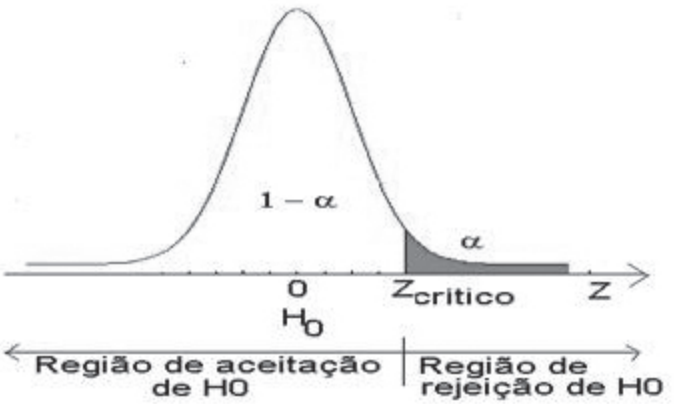


Figura 93: H_1: média populacional > b
Fonte: elaborada pelo autor

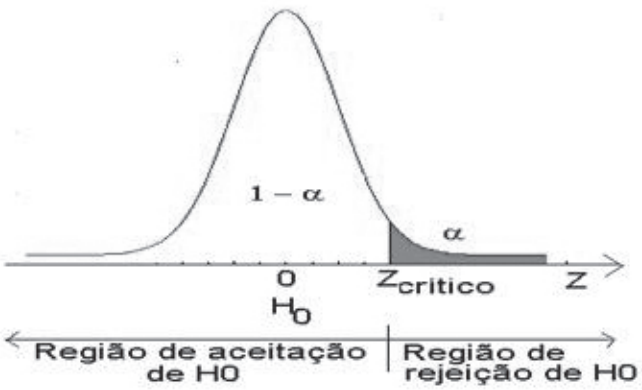


Figura 94: H_1: média > 0
Fonte: elaborada pelo autor

Neste caso, o valor crítico está à direita: se a média amostral $\bar{x}$ ou a variável de teste **Z** tiver valor superior aos respectivos valores críticos, a hipótese nula será rejeitada, pois os valores "caíram" na região de rejeição de $\mathbf{H_0}$. Como foi notado anteriormente, o valor de $\mathbf{Z_{crítico}}$ será positivo, pois é maior do que zero: usando o mesmo nível de significância de 5%, o valor de $\mathbf{Z_{crítico}}$ 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 $\mathbf{H_1}$: $\boldsymbol{\mu \neq b}$ ($\mathbf{H_1}$: $\boldsymbol{\mu \neq 0}$), ou seja, o teste fosse unilateral à direita, a região de rejeição de $\mathbf{H_0}$ seria: (Figuras 95 e 96)

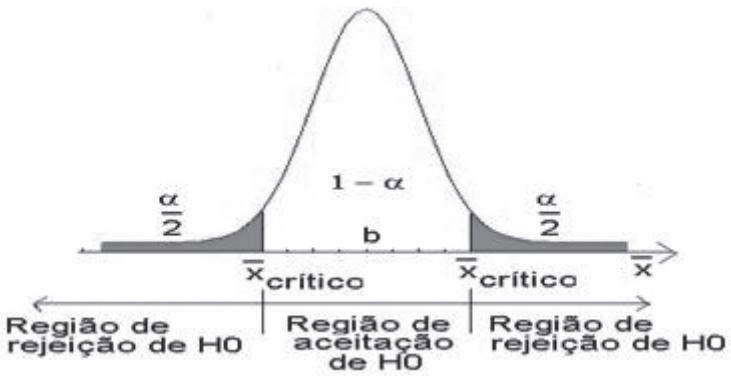


Figura 95: H_1: média populacional ≠ b
Fonte: elaborada pelo autor

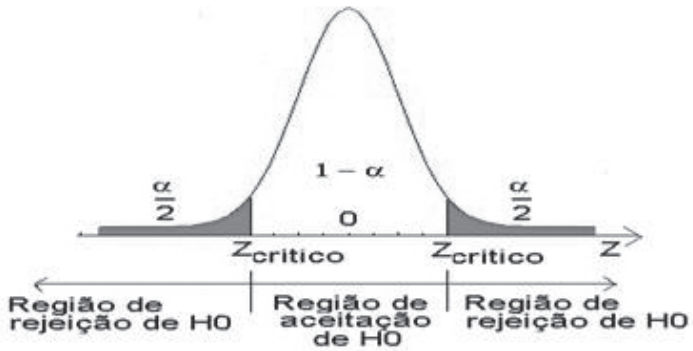


Figura 96: H_1: média ≠ 0
Fonte: elaborada pelo autor

Neste 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** tiver valor 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, visando a 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 fosse menor do que 30, resultando:

Teste unilateral à esquerda

$\mathbf{H_0}$: $\boldsymbol{\mu} = 30$ onde $\boldsymbol{\mu_0} = 30$ (valor de teste).

$\mathbf{H_1}$: $\boldsymbol{\mu} < 30$ 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. Neste caso, admite-se que a proporção máxima de formulários com erros seja de 5%. Ou seja, um valor maior do que 5% causaria problemas. Um teste **unilateral** seria o adequado. Neste caso, um teste de proporção, o problema será um valor maior do que 5%, resultando:

Teste unilateral à direita

$\mathbf{H_0}$: $\boldsymbol{\pi} = 5\%$ onde $\boldsymbol{\pi_0} = 5\%$ (valor de teste)

$\mathbf{H_1}$: $\boldsymbol{\pi} > 5\%$

c) Uma peça automotiva precisa ter 100 mm de diâmetro, exatamente. Neste 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, resultando:

Teste bilateral

$\mathbf{H_0}$: $\boldsymbol{\mu}$ = 100 mm onde $\boldsymbol{\mu_0}$ = 100 mm (valor de teste)

$\mathbf{H_1}$: $\boldsymbol{\mu} \neq$ 100 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 do teste: 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

Neste caso, há interesse em testar a hipótese de que o parâmetro média populacional (**μ**) de uma certa variável quantitativa seja maior, menor ou diferente de um certo valor. Para a realização deste teste, é necessário que uma das duas condições 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 desta população é considerada “suficien-

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.

temente grande", o que, de acordo com o Teorema Central do Limite, permite concluir que a distribuição amostral da média é normal; e

- supõe-se também 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 queremos testar) for conhecida.

Neste caso, a variância amostral da média poderá ser calculada através da expressão:

$V(\bar{x}) = \frac{\sigma^2}{n}$, e, por conseguinte, o "desvio-padrão" será:

O desvio-padrão é a raiz quadrada positiva da variância.

$$\text{desvio-padrão} = \frac{\sigma}{\sqrt{n}}$$

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) 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 ($Z_{crítico}$ que depende do nível de significância adotado), de acordo com o tipo de teste:

Neste caso, $Z_{crítico}$ será negativo, já que a região de rejeição de H_0 está à esquerda de zero.

Se H_1 μ: $> \mu_0$	**Rejeitar H_0 se $Z > Z_{crítico}$**	**($\bar{x} > \bar{x}_{crítico}$)**
Se H_1 μ: $< \mu_0$	**Rejeitar H_0 se $Z < Z_{crítico}$**	**($\bar{x} < \bar{x}_{crítico}$)**
Se H_1 μ: $\neq \mu_0$	**Rejeitar H_0 se $\lvert Z \rvert > \lvert Z_{crítico} \rvert$**	

b) Se a variância populacional σ^2 da variável for desconhecida.

Naturalmente, este é o caso mais encontrado na prática. Como se deve proceder? Dependerá do tamanho da amostra.

b.1) Grandes amostras (mais de 30 elementos)

Nestes 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); e

b.2) Pequenas amostras (até 30 elementos).

Nestes 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**. Esta distribuição já é conhecida (ver Unidades 7 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.

E talvez este seja o motivo de se considerar mais de 30 elementos como sendo uma amostra suficientemente grande.

A variável de teste será, então, $\mathbf{t_{n-1}}$ (**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 ($\mathbf{t_{n-1,crítico}}$ que depende do nível de significância adotado), de acordo com o tipo de teste:

Neste caso, $t_{n-1,crítico}$ será negativo, já que a região de rejeição de H_0 está à esquerda de zero.

Se H_1 μ: > ?$_0$	**Rejeitar H_0 se $t_{n-1} > t_{n-1,crítico}$** $(\bar{x} > \bar{x}_{crítico})$
Se H_1 μ: < ?$_0$	**Rejeitar H_0 se $t_{n-1} < t_{n-1,crítico}$** $(\bar{x} < \bar{x}_{crítico})$
Se H_1 μ: ≠ ?$_0$	**Rejeitar H_0 se $\lvert t_{n-1}\rvert > \lvert t_{n-1,crítico}\rvert$**

A correta identificação dos valores críticos, decorrente da correta identificação da região de rejeição de $\mathbf{H_0}$, 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 neste primeiro exemplo.

Enunciar as hipóteses.

Conforme visto na seção anterior, o teste mais adequado para este caso é um teste bilateral:

$\mathbf{H_0}$: $\boldsymbol{\mu}$ = 100 mm onde $\boldsymbol{\mu_0}$ = 100 mm (valor de teste)

$\mathbf{H_1}$: $\boldsymbol{\mu} \neq$ 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 $\mathbf{t_{n-1}}$ da distribuição **t** de Student.

Definir a região de aceitação de $\mathbf{H_0}$ (Figura 97).

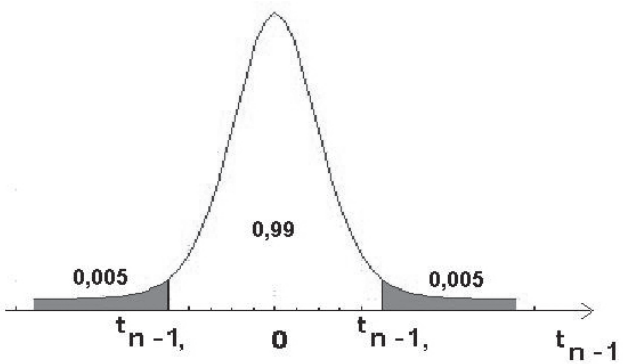


Figura 97: Regiões de rejeição e aceitação da hipótese nula – teste bilateral de média
Fonte: elaborada pelo autor

Observe que, por ser um teste bilateral, o nível de significância ? foi dividido em dois, metade para cada região de rejeição de $\mathbf{H_0}$. 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 acima: os valores críticos serão $\mathbf{t_{14;0,005}}$ e $\mathbf{t_{14;0,995}}$, os quais serão iguais em módulo. E o valor de $\mathbf{t_{n-1,crítico}}$ será igual a 2,977 (em módulo).

Através dos valores da amostra, avaliar o valor da variável.

Neste 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 $\boldsymbol{\mu_0}$ 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 mm^2. Substituindo na equação acima:

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = t_{15-1} = t_{14} = \frac{100{,}7 - 100}{\sqrt{0{,}01}/\sqrt{15}} = 27{,}11, \text{ então } |\mathbf{t_{14}}| = 27{,}11$$

Decidir pela aceitação ou rejeição de $\mathbf{H_0}$.

Como se trata de um teste bilateral:

Rejeitar $\mathbf{H_0}$ se $|\mathbf{t_{n-1}}| > |\mathbf{t_{n-1,crítico}}|$

Como $|\mathbf{t_{14}}| = 27{,}11 > |\mathbf{t_{n-1,crítico}}| = |\mathbf{t_{14,0,995}}| = 2{,}977$

Rejeitar $\mathbf{H_0}$ 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.

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 $\mathbf{H_0}$ pode ocorrer por acaso. Sempre que apresentar o resultado, recomende que uma nova amostra seja retirada para avaliar novamente o problema. Mas, neste caso, rejeita-se H_0 com folga.

Um novo protocolo de atendimento foi implementado no Banco RMG, visando a 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. 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 minu-

tos. O protocolo pode ser considerado satisfatório a 5% de significância? Vejamos neste segundo exemplo.

Este problema já foi estudado anteriormente. Seguindo o roteiro do Apêndice.

Enunciar as hipóteses. Conforme visto na seção 10.4, o teste mais adequado para este caso é um teste unilateral à esquerda:

$\mathbf{H_0 : \mu} = 30$ onde $\boldsymbol{\mu_0} = 30$ (valor de teste)

$\mathbf{H_1} : \boldsymbol{\mu} < 30$

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 $\mathbf{H_0}$ (Figura 98).

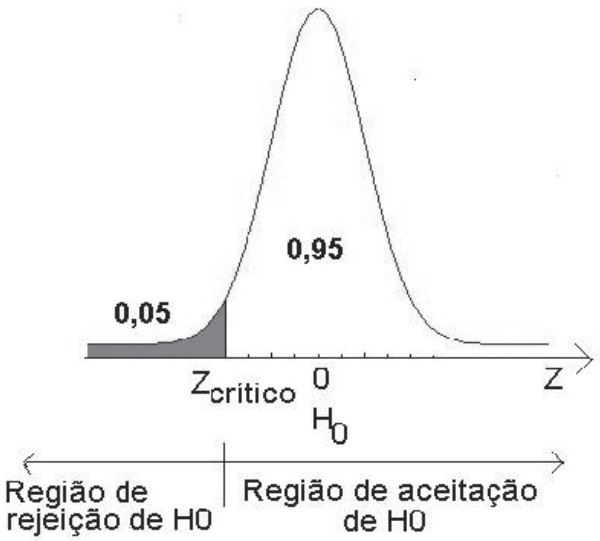


Figura 98: Regiões de aceitação e de rejeição – teste unilateral à esquerda
Fonte: elaborada pelo autor

Observe que, por ser um teste unilateral à esquerda, o nível de significância $\boldsymbol{\alpha}$ está todo concentrado em um dos lados da distribuição, definindo a região de rejeição de $\mathbf{H_0}$. Para encontrar o valor crítico, devemos procurar, na tabela da distribuição normal, pela probabi-

lidade acumulada 0,95. Ou procurar a probabilidade complementar 0,05 e mudar o sinal do valor encontrado, pois o $\mathbf{Z_{crítico}}$ aqui é menor do que zero. O valor crítico será igual a –1,645.

Através dos valores da amostra, avaliar o valor da variável.

Neste ponto, é preciso encontrar o valor da variável de teste:

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

O valor de teste $\boldsymbol{\mu_0}$ 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 acima:

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{29 - 30}{5 / \sqrt{35}} = -1{,}183$$

Decidir pela aceitação ou rejeição de $\mathbf{H_0}$. Como se trata de um teste unilateral à direita:

Rejeitar H_0 se $Z < Z_{crítico}$ Como $Z = -1{,}185 > Z_{crítico} = -1{,}645$

Aceitar $\mathbf{H_0}$ 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

Neste caso, há interesse em testar a hipótese de que o parâmetro proporção populacional ($\boldsymbol{\pi}$) de um dos valores de uma certa variável seja maior, menor ou diferente de um certo valor. Para a realização deste teste tal como será descrito, é necessário que duas condições sejam satisfeitas:

- que o produto $\boldsymbol{\pi}$ **n x** $_0$ seja maior ou igual a 5; e
- que o produto $\boldsymbol{\pi}$ n x (1 – $_0$) seja maior ou igual a 5.

Onde **n** é o tamanho da amostra, e $\boldsymbol{\pi_0}$ 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:

$$\text{Média} = \boldsymbol{\pi_0} \qquad \text{Desvio-padrão} = \sqrt{\frac{\pi_0 \times (1-\pi_0)}{n}}$$

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 acima; 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 ($\mathbf{Z_{crítico}}$ que depende do nível de significância adotado), de acordo com o tipo de teste:

Neste caso, $Z_{crítico}$ será negativo, já que a região de rejeição de H_0 está à esquerda de zero.

Se H_1 π: $> \pi_0$ Rejeitar H_0 se $Z > Z_{crítico}$ ($p > p_{crítico}$)
Se H_1 π: $< \pi_0$ Rejeitar H_0 se $Z < Z_{crítico}$ ($p < p_{crítico}$)
Se H_1 π: $\neq \pi_0$ Rejeitar H_0 se $|Z| > |Z_{crítico}|$

Cerca de 2.000 formulários de pedidos de compra estão sendo analisados. Os clientes podem ficar insatisfeitos se houver erros nos formulários. Neste caso, admite-se que a proporção máxima de formulários com erros seja de 5%. Suponha que, entre os 2.000 formulá-

rios, 7% apresentavam erros. A proporção máxima foi ultrapassada a 1% de significância? Vejamos neste terceiro exemplo.

Enunciar as hipóteses. Conforme visto na seção 10.4, o teste mais adequado para este caso é um teste unilateral à direita:

$\mathbf{H_0}$: $\boldsymbol{\pi}$ = 5% onde $\boldsymbol{\pi_0}$ = 5% (valor de teste)

$\mathbf{H_1}$: $\boldsymbol{\pi} > \mathbf{5\%}$

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:

$\boldsymbol{\pi}$ **n x** $_0$ **= 2.000 x 0,05 = 100** e **n x (1** $\boldsymbol{\pi}$ **–** $_0$**) = 2.000 x 0,95 = 1.900**. Como ambos são maiores do que 5, é possível fazer uma aproximação pela normal, e a variável de teste será **Z**.

Definir a região de aceitação de $\mathbf{H_0}$ (Figura 99).

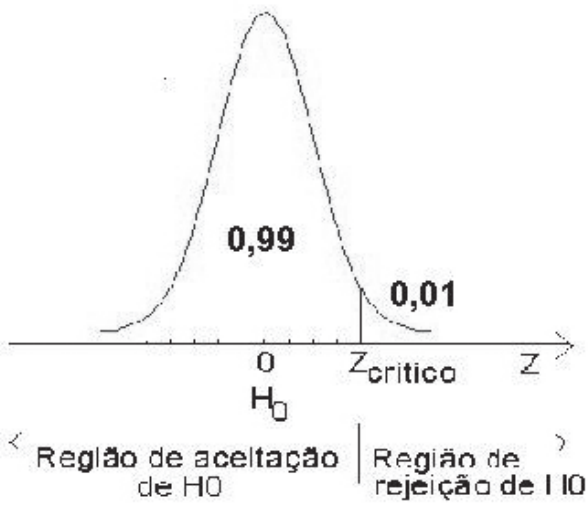


Figura 99: Regiões de aceitação e de rejeição – teste unilateral à direita
Fonte: elaborada pelo autor

Observe que, por ser um teste unilateral à direita, o nível de significância $\boldsymbol{\alpha}$ está todo concentrado em um dos lados da distribuição, definindo a região de rejeição de $\mathbf{H_0}$. Para encontrar o valor críti-

co, devemos procurar, na tabela da distribuição normal, pela probabilidade acumulada 0,01 (o $\mathbf{Z}_{\text{crítico}}$ aqui é maior do que zero). O valor crítico será igual a 2,326.

Através dos valores da amostra, avaliar o valor da variável. Neste 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 $\boldsymbol{\pi_0}$ 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 acima:

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}} = \frac{0{,}07 - 0{,}05}{\sqrt{\frac{0{,}05 \times (0{,}95)}{2000}}} = 4{,}104$$

Decidir pela aceitação ou rejeição de $\mathbf{H_0}$. Como se trata de um teste unilateral à direita:

Rejeitar $\mathbf{H_0}$ se $\mathbf{Z} > \mathbf{Z}_{\text{crítico}}$ Como $\mathbf{Z} = 4{,}104 > \mathbf{Z}_{\text{crítico}} = 2{,}326$

Rejeitar $\mathbf{H_0}$ 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 qui-quadrado).

Teste de associação de qui-quadrado

O **teste do qui-quadrado***, também chamado de teste de independência de qui-quadrado, está vinculado à análise de duas variáveis qualitativas. Vamos ver alguns conceitos antes de apresentar o teste de associação de qui-quadrado.

GLOSSÁRIO

*Teste de associação (independência) de quiquadrado – teste que permite avaliar se duas variáveis qualitativas, cujas freqüências estão dispostas em uma tabela de contingências, apresentam associação significativa ou não. Fonte: Barbetta, Reis e Bornia (2004)

Variáveis qualitativas e tabelas de contingências

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 as perguntas do parágrafo anterior.

Variáveis qualitativas são aquelas cujas realizações são atributos, categorias (Unidades 1 e 3). Como exemplo de variáveis qualitativas, tem-se: sexo de uma pessoa (duas categorias, masculino e feminino), grau de instrução (analfabeto, Ensino Fundamental incompleto, etc.), opinião sobre um assunto (favorável, desfavorável, indiferente).

Em estudos sobre variáveis qualitativas, é extremamente comum registrar as freqüências de ocorrência de cada valor que as variáveis podem assumir, e quando há duas variáveis envolvidas, é comum registrar-se a freqüência de ocorrência dos cruzamentos entre valores: por exemplo, quantas pessoas do sexo masculino são favoráveis a uma certa proposta de lei, quantas são desfavoráveis, quantas pessoas do sexo feminino são favoráveis. E, para facilitar a análise dos resulta-

dos, estes costumam ser dispostos em uma tabela de contingências. Esta 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 24 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 24: Tabela de contingências de função por sexo
Fonte: elaborado pelo autor

Podemos apresentar os percentuais calculados em relação aos totais das colunas no Quadro 25:

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 25: Tabela de contingências de função por sexo: percentuais por colunas
Fonte: elaborado pelo autor

Seria interessante saber se as duas variáveis são estatisticamente dependentes e o quão forte é esta 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. Neste 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 freqüências observadas da variável são representadas em uma tabela de contingências, e a hipótese nula ($\mathbf{H_0}$) do teste será que as duas variáveis não diferem em relação às freqüências com que ocorre uma característica particular, ou seja, as variáveis são independentes, que será testada contra a hipótese alternativa ($\mathbf{H_1}$) 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 qui-quadrado (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 $E_{ij} = \dfrac{\text{total da linha i} \times \text{total da coluna j}}{\text{total geral}}$

Onde:

- E_{ij} é a freqüência esperada, sob a condição de independência entre as variáveis, em uma célula qualquer da tabela de contingências. Todas as freqüências esperadas precisam ser maiores ou iguais a 5 para que o resultado do teste seja válido;
- O_{ij} é a freqüência observada em uma célula qualquer da tabela de contingências;

Se isso não ocorrer, recomenda-se agrupar as categorias (de uma ou outra variável, ou de ambas) até obter todas as freqüências pelo menos iguais a 5.

- L é o número total de linhas da tabela de contingências (número de valores que uma das variáveis pode assumir); e
- 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 freqüência 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 freqüência esperada. Somam-se os valores de todas as células e obtêm-se o valor da estatística: quanto maior $\mathbf{q^2}$, mais o observado se afasta do esperado; portanto, maior a dependência.

Sob a hipótese de independência, a estatística $\mathbf{q^2}$ seguirá o modelo qui-quadrado, que estudamos na Unidade 7, prometendo usá-la aqui na Unidade 10.

O teste do qui-quadrado, para avaliar se duas variáveis são independentes, será unilateral: ou seja, a hipótese nula será rejeitada se $\mathbf{q^2} > \mathbf{q^2}_{crítico}$, para um certo número de graus de liberdade. Por exemplo, para o caso em que há 3 graus de liberdade e o nível de significância fosse de 5% (a região de rejeição de H_0 ficará À DIREITA), o valor crítico seria (lembre-se da Unidade 7) (Figura 100):

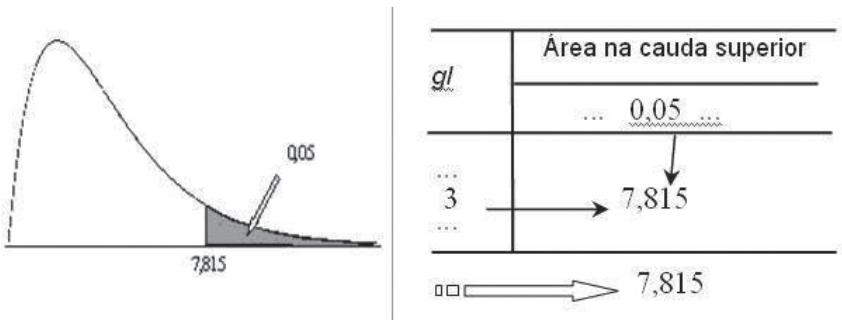


Figura 100: Uso da tabela da distribuição qui-quadrado. Ilustração com gl = 3 e área na cauda superior de 5%
Fonte: adaptado pelo autor de Barbetta, Reis e Bornia (2004)

O número de graus de liberdade da estatística é calculado da seguinte forma:

graus de liberdade = **(L – 1) x (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 este valor, porque para calcular as freqüências 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) x (2–1)]: bastaria calcular a freqüência 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 neste quinto exemplo se as variáveis são independentes a 1% de significância.

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) (Quadro 26):

Sexo	Função			
	Escritório	Serviços gerais	Gerência	Total
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Quadro 26: Tabela de contingências de função por sexo
Fonte: elaborado pelo autor

No quadro acima, se encontram os totais marginais e o total geral:

L1 = total Masculino = 258 L2 = total Feminino = 216 C1 = total Escritório = 157

C2 = total S.Gerais = 27 C3 = total Gerência = 84 N = total geral =474

Repare que, somando os totais das linhas, o resultado é o total geral, e que somando os totais das colunas, o resultado é o total geral também.

Calcular as freqüências esperadas.

Calculando as freqüências esperadas de acordo com a fórmula vista anteriormente:

Masculino – Escritório	E = (258 363)/ 474 = 197,58
Masculino – Serviços Gerais	E = (258 27)/ 474 = 14,70
Masculino – Gerência	E = (258 84)/ 474 = 45,72
Feminino – Escritório	E = (216 363)/ 474 = 165,42
Feminino – Serviços Gerais	E = (216 27)/ 474 = 12,30
Feminino – Gerência	E = (216 84)/ 474 = 38,28

Observe que os resultados são os mesmos obtidos no Exemplo 3.2.

Calculando a estatística $\mathbf{q^2}$ para cada célula:

Agora, podemos calcular as diferenças entre as freqüências e as demais operações, que serão mostradas nos Quadros 27 , 28 e 29.

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 27: Diferença entre freqüências observadas e esperadas de função por sexo

Fonte: elaborado pelo autor

$(O-E)^2$	Função		
Sexo	**Escritório**	**Serviços gerais**	**Gerência**
Masculino	1646,921	151,383	799,672
Feminino	1646,921	151,383	799,672

Quadro 28: Diferença entre freqüências observadas e esperadas de função por sexo elevadas ao quadrado
Fonte: elaborado pelo autor

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 29: Estatísticas q^2 de função por sexo
Fonte: elaborado pelo autor

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:

$$(\textbf{número de linhas} - 1) \times (\textbf{número de colunas} - 1) = (2-1) \times (3-1) = 2$$

Então, $\mathbf{q}^2_2 = 79{,}227$

O $\mathbf{q}^2$ crítico será: procurando na Tabela 3 do Ambiente Virtual de Ensino-Aprendizagem, ou em um programa, para 2 graus de liberdade e 99% de confiança (1% de significância): $\mathbf{q}^2_{2,\text{crítico}} = 9{,}21$

Ver Figura 101:

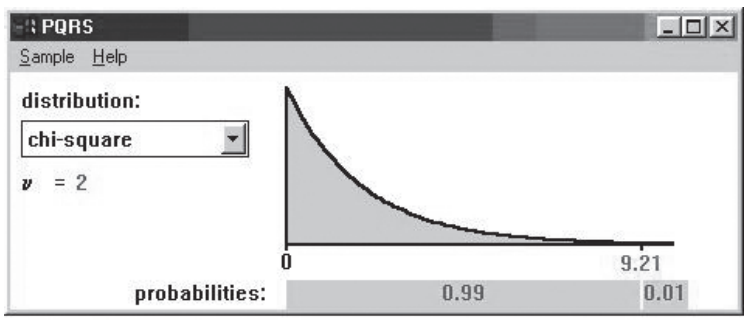


Figura 101: Valor crítico de q^2 para 2 graus de liberdade e 1% de significância
Fonte: adaptada pelo autor de PQRS®

Como $\mathbf{q}^2_2$ é maior do que $\mathbf{q}^2_{2,\text{crítico}}$, rejeitamos $\mathbf{H}_0$ 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 freqüências da tabela.

No tópico Saiba Mais, você terá indicações de vários outros tipos de hipótese que não foram mencionados nesta Unidade. As referências lá citadas serão extremamente valiosas se você tiver que:

- 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 um determinado modelo probabilístico.

Com este tópico, terminamos nossa jornada. Agora, é com vocês. Boa sorte!

Saiba mais...

Sobre tipos de erro, poder, em testes de hipóteses:

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 8.

STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Ed. Harbra, 2001, capítulo10.

Sobre testes de uma variância:

BARBETTA, P. A.; REIS, M. M.; BORNIA, A.C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 8.

TRIOLA, M. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999, capítulo 7.

- Sobre testes de comparação de duas médias:

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 9.

- Sobre testes de comparação de duas proporções:

MOORE, D. S.; et al. *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, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 9.
STEVENSON, W. J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 11.
MOORE, D. S.; McCABE, G. P.; DUCKWORTH, W. M.; SCLOVE, S. L. *A prática da Estatística Empresarial*: como usar dados para tomar decisões. Rio de Janeiro: LTC, 2006, capítulos 14 e 15.
TRIOLA, M. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999, capítulo 11.

- Sobre testes não paramétricos:

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004, capítulo 10.
STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 13.
SIEGEL, S. *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, D. M.; et al. *Estatística*: teoria e aplicações – usando Microsoft Excel em português. 5. ed. Rio de Janeiro: LTC, 2006, capítulo 6.

RESUMO

O resumo desta Unidade está demonstrado na Figura 102:

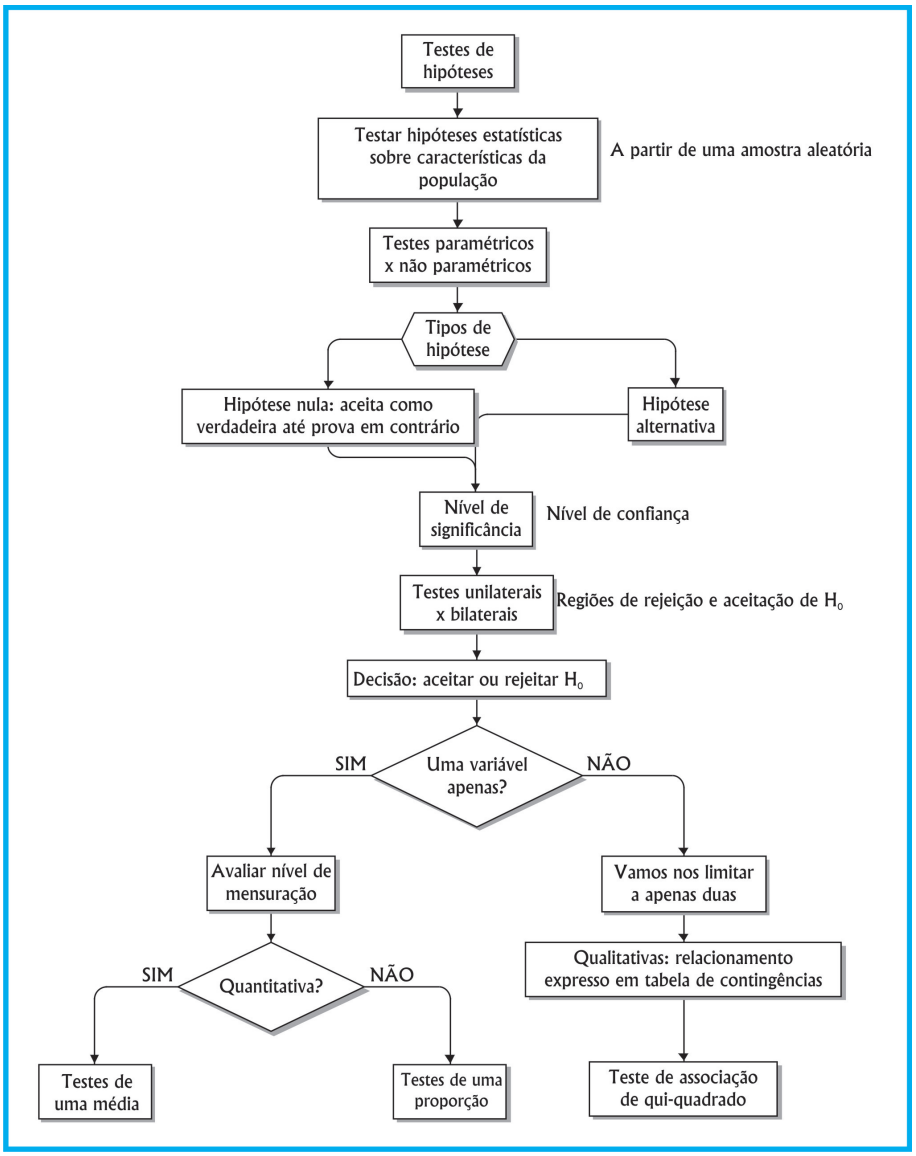


Figura 102: Resumo da Unidade 10
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Chegamos ao final da disciplina de Estatística Aplicada à Administração. Estudamos, nesta última Unidade, os testes de hipóteses, tipos de hipótese e suas variáveis. A Unidade foi explorada com figuras, exemplos e quadros para melhor representar o conteúdo oferecido. Além do material produzido pelo professor, você tem em mãos uma riquíssima fonte de referências para saber mais sobre o assunto. Explore os conhecimentos propostos. Não tenha esta Unidade como fim, mas sim o começo de uma nova trajetória em sua vida acadêmica. Bons estudos e boa sorte!

REFERÊNCIAS

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. *Estatística Aplicada à Administração e Economia*. 2. ed. São Paulo: Thomson Learning, 2007.

ANDRADE, D. F.; OGLIARI, P. J. *Estatística para as ciências agrárias e biológicas*: com noções de experimentação. Florianópolis: Ed. da UFSC, 2007.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para Cursos de Engenharia e Informática*. São Paulo: Atlas, 2004.

BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 5. ed. São Paulo: Saraiva, 2002.

COSTA NETO, P. L. da O. *Estatística*. 2. ed. São Paulo: Edgard Blücher, 2002.

LOPES, P. A. *Probabilidades e Estatística*. Rio de Janeiro: Reichmann e Affonso Editores, 1999.

LAKATOS, E. M.; MARCONI, M. de A. *Técnicas de Pesquisa*. 5. ed. São Paulo: Atlas, 2003.

MONTGOMERY, D. C. *Introdução ao Controle Estatístico da Qualidade*. 4. ed. Rio de Janeiro: LTC, 2004.

MOORE, D. S.; et al. *A prática da Estatística Empresarial*: como usar dados para tomar decisões. Rio de Janeiro: LTC, 2006.

STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001.

TRIOLA, M. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999.

VIRGILITTO, S. 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. Tem ministrado disciplinas de Estatística em vários cursos de graduação e pós-graduação da Universidade, incluindo as disciplinas de Estatística 1 e 2 do curso de Administração.