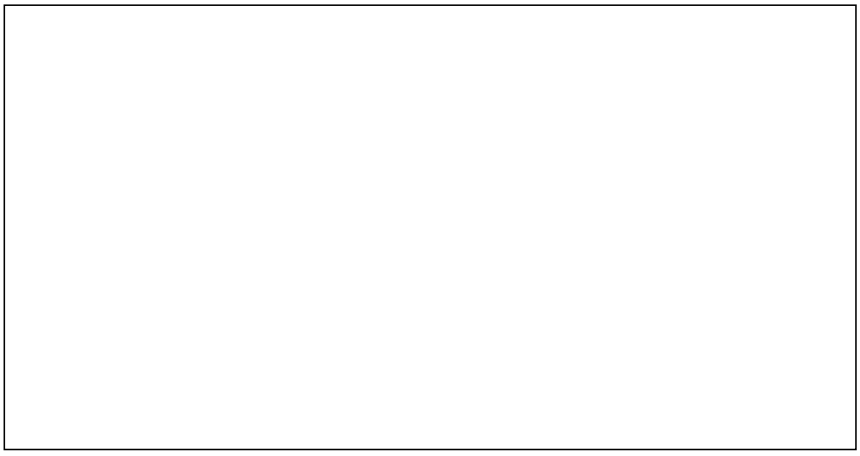
Estatística Aplicada à Administração I

Marcelo Menezes Reis



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

PRESIDENTE DA REPÚBLICA
Dilma Vana Roussef

MINISTRO DA EDUCAÇÃO
Aloizio Mercadante

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

UNIVERSIDADE FEDERAL DE SANTA CATARINA
REITOR
Lúcio José Botelho
VICE-REITOR
Ariovaldo Bolzan
PRÓ-REITOR DE ENSINO DE GRADUAÇÃO
Marcos Lafim
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

COORDENADOR DE CURSO
Alexandre Marino Costa

COMISSÃO DE PLANEJAMENTO, ORGANIZAÇÃO E FUNCIONAMENTO
Alexandre Marino Costa
Gilberto de Oliveira Moritz
João Nilo Linhares
Luiz Salgado Klaes

Comentado [MMR1]: Modificar os nomes onde necessário.

Sumário

- 5.3 - Sazonalidade: obtenção pelo método da razão para a média móvel.
- 5.4 - Obtenção das componentes cíclica e irregular.
- 5.5 - Recomposição e avaliação da acuracidade do modelo recomposto.
- 5. 6 – Outros modelos de séries temporais

UNIDADE 6 – Probabilidade

- Modelos determinísticos e probabilísticos: conceitos, abordagem da incerteza, teoria da decisão.
- Experimento aleatório, espaço amostral e eventos.
- Conceitos de probabilidade: clássico, experimental.
- Axiomas e propriedades de probabilidade.
- Probabilidade condicional: aplicações à decisão.

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 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 I. Os métodos estatísticos são ferramentas cruciais 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, juntamente com a disciplina Estatística Aplicada à Administração II.

Nesta disciplina e na próxima você aprenderá como obter dados confiáveis (conceitos de planejamento de pesquisa estatística), 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 entender o conceito de Estatística, sua importância para o administrador e os principais aspectos do planejamento estatístico para garantir a obtenção de dados confiáveis.

1.1 - **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 que 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 quarenta 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, materializando-se naquilo que os profissionais chamam de "Big data", o que vem constituindo o campo de conhecimento de mineração de dados e inteligência empresarial.

Comentado [MMR2]: Glossário. Big Data é um termo popular usado para descrever o crescimento, a disponibilidade e o uso exponencial de informações estruturadas e não estruturadas. Fonte: SAS Institute, disponível em http://www.sas.com/pt_br/insights/big-data/what-is-big-data.html, acessado em 14/10/2015. Fim Glossário.

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.

1.1.1 - 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;
- em quais ações devo investir para obter maior rendimento?

1.1.2 – Variabilidade

O principal problema que surge ao tentar responder essas perguntas é que todas as medidas feitas para tal, por mais acurados que sejam os instrumentos de medição, apresentarão sempre uma **variabilidade**, ou seja, não há respostas perfeitas. Glossário: Variabilidade: diferenças encontradas por sucessivas medições realizadas em pessoas, animais ou objetos, em tempos ou situações diferentes Fonte: Montgomery, 2004 Fim Glossário. 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 à experiências de vida diferentes terão um desenvolvimento distinto. Sendo assim, variabilidade é **inevitável** e **inerente** à vida.

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

Você tem as mesmas preferências musicais que tinha há dez anos atrás (muitos sim, mas muitos não)? Você tem a mesma aparência que tinha há dez anos atrás? 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 atrás? 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.

1.1.3 - 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 LINK Maiores detalhes, você vai estudar, ainda nesta Unidade FIM DO LINK. 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.

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** Glossário: Amostra: um subconjunto finito e representativo da população. Fonte: Barbetta, 2014 Fim Glossário da população, um subconjunto finito e representativo da população. LINK Mais detalhes ainda nesta unidade FIM DO LINK.

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

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

Uma das principais subdivisões da Estatística justamente é a **Amostragem** LINK Tema da Unidade 2 da disciplina Estatística Aplicada à Administração II, FIM DO LINK, que reúne os métodos necessários para coletar adequadamente amostras representativas e suficientes para que os resultados obtidos possam ser generalizados para a população de interesse.

Após a coleta dos dados, por censo ou amostragem, a **Análise Exploratória de Dados** LINK Tema das Unidades 2 e 3 FIM DO LINK 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. Os métodos de correlação e regressão (Unidade 4) e de análise de séries temporais (Unidade 5) também contribuem descrever e modelar a variabilidade. Se a pesquisa foi feita por censo basta realizar a análise exploratória de dados para obter as conclusões.

Posteriormente, através da **Inferência Estatística** LINK Estatística Indutiva, tema das Unidades 4, 5, e 6 de Estatística Aplicada à Administração II FIM DO LINK é possível generalizar as conclusões dos dados para a população quando os dados forem provenientes de uma **amostra**, utilizando a **probabilidade** 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: Lopes, 1999 Fim Glossário LINK Tema da Unidade 6 desta disciplina e Unidades 1 e 2 de Estatística Aplicada à Administração II FIM DO LINK para calcular a confiabilidade das conclusões obtidas.

A Figura 1 ilustra a subdivisão da estatística. Veja:

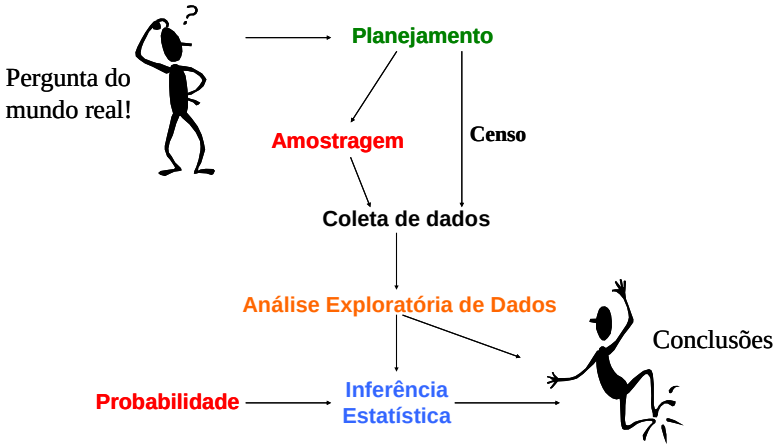


Figura 1 - Subdivisões da Estatística

Fonte: elaborada pelo autor

1.2 - 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 onde 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 site de economia.

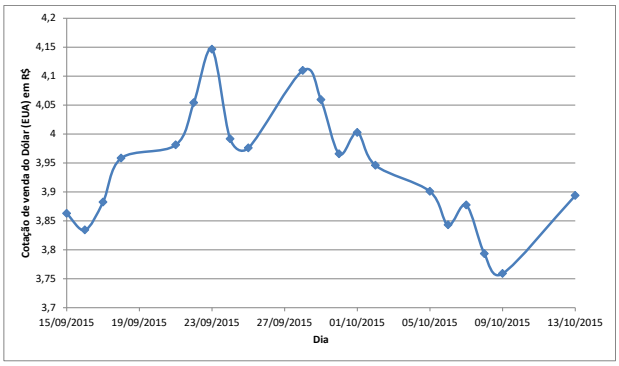


Figura 2 – Variação da cotação do dólar comercial (venda) de 15/09/2015 a 13/10/2015

Fonte: UOL Economia Cotações, disponível em http://economia.uol.com.br/cotacoes/cambio/dolar-comercial-estados-unidos/?historico, acessado em 14/10/2015, adaptada pelo autor de Microsoft®.

Através de um simples gráfico de linhas LINK Nas Unidade 2 e 5, você vai estudá-lo com mais detalhes FIM DO LINK podemos observar a cotação ultrapassando R$ 4 em 23/09/2015, em decorrência da crise política, da apresentação de um orçamento deficitário ao Congresso Nacional por parte do Governo da União, seguida por uma queda provavelmente devida à ação do Banco Central, mas ultrapassando novamente os R$ 4 após o dia 27/09/2015. Mais uma atuação do Banco Central deve ter sido a causa da redução da cotação para pouco mais de R$ 3,75 em 09/10/2015, para subir mais vez a R$ 3,9 em 13/10/2015. Os problemas na economia brasileira, o rebaixamento do grau de investimento dos títulos brasileiros feito pela agência Standard & Poor's, a incerteza sobre o sucesso do ajuste fiscal proposto pelo Governo da União, tudo isso pode ter contribuído para o aumento da cotação do dólar, que pode trazer grandes prejuízos a todos que importam produtos, mas podem ser uma grande oportunidade para todos que exportam.

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

1.3 - Planejamento Estatístico de Pesquisa

O planejamento estatístico da pesquisa Glossário Planejamento estatístico da pesquisa: conjunto de métodos cuja implementação visa garantir a confiabilidade dos dados coletados. Fonte: Barbetta, Reis e Bornia, 2010. Fim Glossário é parte do planejamento geral da pesquisa.

Antes de se pensar em qualquer abordagem estatística é preciso definir o que se quer pesquisar, em qualquer campo do conhecimento. "Como poderemos escolher o melhor caminho se não sabemos para onde ir"? Em outras palavras é preciso definir corretamente a "**pergunta do mundo real**" que queremos responder: 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" LINK Tô afim de saber: Estas e outras informações você encontra em http://www.crasc.org.br/index.php?pg=inicial/oque.htm, acessado em 17/10/2007. FIM DO LINK. 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.

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.

1.3.1 - Objetivos da pesquisa

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

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

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

- **Objetivo geral**: avaliar a opinião dos registrados no CRA 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).
  - 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.

1.3.2 – População

Uma parte importante do delineamento de qualquer pesquisa é a definição da população. Tal definição dependerá obviamente dos objetivos da pesquisa, das características a mensurar, dos recursos disponíveis. LINK A definição de população foi vista no inicio desta Unidade, você se lembra? FIM DO LINK

"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 11676 registrados no CRA de Santa Catarina. Vamos supor que 9000 foram graduados em faculdades catarinenses.

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

1.3.3 – Variáveis

Quando um determinado 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.

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

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, que 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 a Figura (3) a seguir e entenda a classificação das variáveis por nível de mensuração.

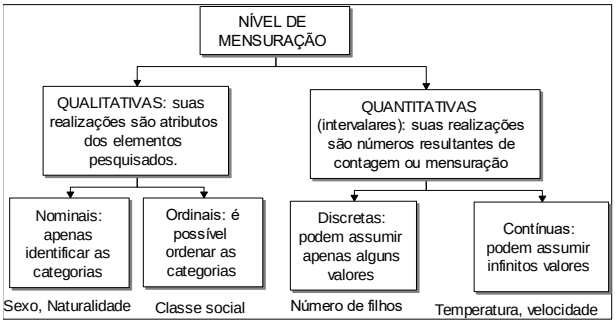


Figura 3 - Classificação das variáveis por nível de mensuração
Fonte: elaborada pelo autor.

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

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

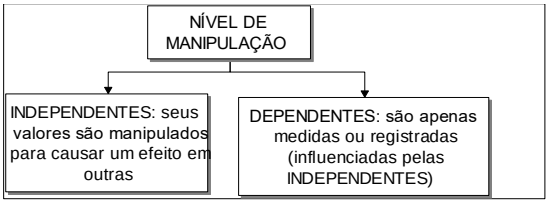


Figura 4 - Classificação das variáveis por nível de manipulação
Fonte: elaborada pelo autor.

Variáveis **independentes** são aquelas que são manipuladas enquanto que 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 LINK Veremos mais detalhes nas próximas Unidades LINK, onde 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.

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 pré-existentes 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
- 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 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 onde as preocupações lógicas e teóricas das fazes anteriores cedem lugar as questões mais práticas de verificação.

1.3.4 - 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.

A 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:

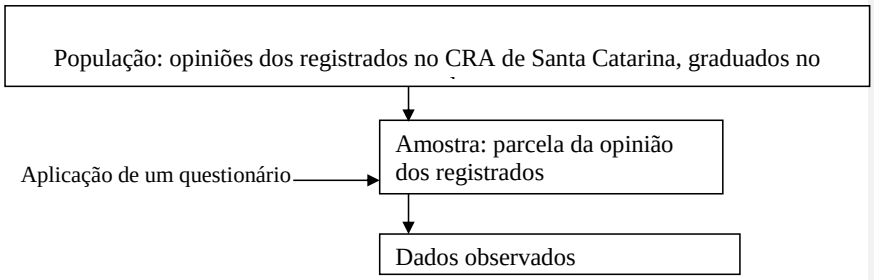


Figura 5 - Pesquisa por levantamento
Fonte: adaptado pelo autor a partir de Barbetta, 2014

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.
- pesquisas na indústria farmacêutica (sobre eficácia e segurança de medicamentos), na indústria química (quais fatores irão 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.

1.3.5 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 a 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 3000 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. LINK

Na Unidade 2 de Estatística Aplicada à Administração II você vai estudar as formas de minimizar tal risco. FIM DO LINK

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

Os dados secundários são dados existentes, coletados por outros pesquisadores e disponíveis em relatórios ou publicações. A sua utilização pode reduzir muito os custos de uma pesquisa. Se fosse necessário obter informações demográficas poderíamos utilizar os relatórios do IBGE referentes ao último censo, ou à pesquisa nacional por amostragem de domicílios (PNAD), 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 de Estatística Aplicada à Administração II vamos apresentar os vários tipos de amostragem.

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

1.3.6 Instrumento de pesquisa

É através do instrumento de pesquisa Glossário: Instrumento de pesquisa: dispositivo usado para coletar os valores das variáveis nos elementos da população. Fonte: Barbetta, Reis e Bornia, 2010. Fim Glossário 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 recusar-se a responder, ou aos lojistas, que podem ser "criativos demais" nas respostas. 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 que indispensável quando precisamos mensurar ou avaliar atitudes, preferências, crenças e comportamentos que exigem a manifestação dos pesquisados. Pesquisas de mercado, acerca da aceitação de um produto ou 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.

DESTAQUE 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; grande número de respostas em branco, ou incoerentes com as outras

perguntas; muitos respondentes não sabiam o que era itinerário. FIM DO DESTAQUE

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.

Tô afim de saber:

- Para saber mais sobre experimentos, consulte 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, na seção 3.2.
- Para saber mais sobre elaboração de questionários, consulte BARBETTA,P. A. *Estatística Aplicada às Ciências Sociais*. 9ª. ed. – Florianópolis: Ed. da UFSC, 2014, capítulo 2.

Atividades de Aprendizagem

Confira se você teve bom entendimento do que tratamos nesta Unidade, respondendo as questões conforme os conceitos estudados, e encaminhe-as para seu tutor através do Ambiente Virtual de Ensino-Aprendizagem.
Boa sorte! Se precisar de auxílio, não deixe de fazer contato com seu tutor.

1) A direção do CED (Centro de Educação da UFSC), e os departamentos do centro, têm interesse em avaliar se a biblioteca setorial está atendendo adequadamente os alunos de graduação, pós-graduação, pessoal docente e técnico-administrativo do CED. Há preocupação com o acervo em si (atualização, composição, número de cópias disponíveis, adequação às necessidades de cada curso), e com o atendimento aos usuários (número de atendentes, horário, “cortesia"). Faça o planejamento da pesquisa estabelecendo
a) Objetivo geral da pesquisa.
b) Objetivos específicos da pesquisa.
c) Tipo de pesquisa (Levantamento ou Experimento).
d) População da pesquisa.

e) Quais são as variáveis da pesquisa? Por quê? Como serão medidas? Por quê?

f) Como serão coletados os dados: secundários ou primários, censo ou amostragem? Por quê?

g) Escolha do instrumento de pesquisa. Justificativa.

i) Elaboração do instrumento de pesquisa. Justificativa dos itens e opções escolhidas.

Resumo

O resumo desta Unidade está esquematizado na Figura 6. Veja:

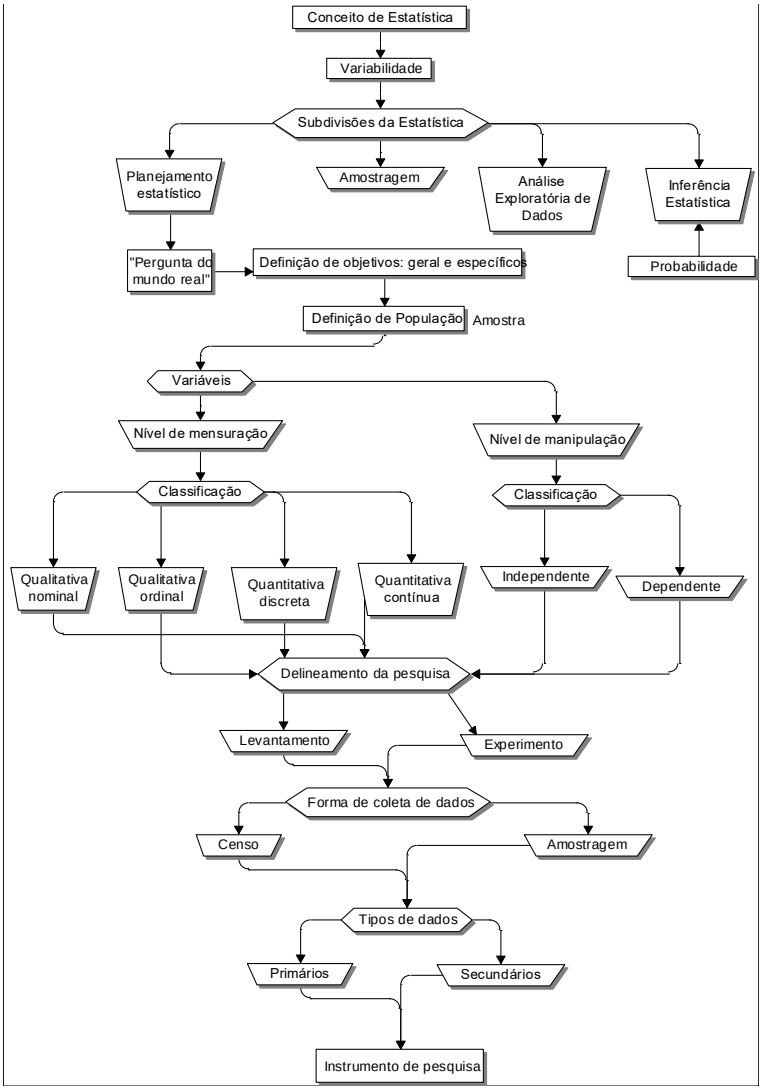


Figura 6- Resumo da Unidade 1

Fonte: elaborada pelo autor.

Caro estudante,

Fazer com que você entenda 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 consequências 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

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.

Caro estudante, na Unidade anterior estudamos o planejamento de uma pesquisa. 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.

2.1 – Conceitos básicos

A **Análise Exploratória de Dados**, antigamente chamada apenas de Estatística Descritiva, LINK No passado a Análise Exploratória de Dados era chamada de Estatística Descritiva, por preocupar-se com a descrição dos dados tão somente. LINK 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).

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** 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, 2014. Fim Glossário, 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.

2.1.1 – 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), 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 1995, matrícula 13xxxxxxx-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 setembro de 2015, supermercado.

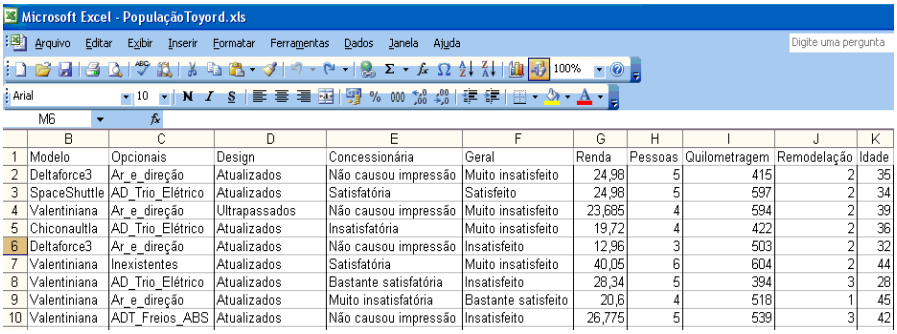
Exemplo 1 - A Megamontadora TOYORD LINK Trata-se de uma empresa fictícia, e de uma pesquisa fictícia.LINK regularmente conduz pesquisas de mercado com os clientes que compraram carros zero km diretamente de suas concessionárias. O objetivo é

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 2015: 250 clientes foram entrevistados entre o total de 30.000 que compraram veículos novos entre maio de 2014 e maio de 2015. A pesquisa foi restringida aos modelos mais vendidos, e que já estão no mercado há 10 anos. As seguintes variáveis foram obtidas:

- **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).
- **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 contratados em definitivo ou dispensados (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 irá tabulá-los em uma planilha eletrônica, como o Br.Office Calc ® ou o Microsoft Excel ®. Há dez variáveis, a base de dados deve ter então 10 colunas, e 250 linhas (no Calc ®, 251, já que a primeira será usada para pôr o nome das variáveis). Veja o resultado, com as primeiras linhas (casos) na Figura 7:



	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 7 - 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 encontram-se 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 5 pessoas no veículo, trafega em média 415 km por mês com este veículo, crê que a última remodelação foi feita há 2 anos atrás 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 7 está disponível no Ambiente Virtual de Ensino-Aprendizagem. Juntamente com ele estão disponibilizados os textos “Como realizar análise exploratória de dados no Br.Office Calc ®” e “Como realizar análise exploratória de dados no Microsoft Excel ®). LINK Veja a seção “Saiba mais” desta Unidade. O arquivo de dados e o texto servirão para as Unidades 2 e 3. LINK

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 7. Podemos ter tantas colunas e linhas quantas se quiser, respeitando, porém, as capacidades dos programas, o Microsoft Excel ®, por exemplo, admite 1.048.576 linhas por 16.384 colunas, o que é suficiente para muitas aplicações.

Comentado [MMR3]: Fonte: Suporte do Office, disponível em https://support.office.com/pt-BR/article/Especifica%C3%A7%C3%B5es-e-limites-do-Microsoft-Excel-1672b34d-7043-467e-8e27-269d656771c3, acessado em 14/10/2015.

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 frequências, como veremos a seguir.

2.1.2 – Distribuição de frequências

O processo de resumo e organização dos dados busca basicamente registrar as ocorrências dos possíveis valores das variáveis que caracterizam o fenômeno, em suma consiste em elaborar **Distribuições de Frequências** das variáveis Glossário Distribuições de Frequências: organizações dos dados de acordo com as ocorrências dos diferentes resultados observados. Fonte: Barbetta, Reis e Bornia, 2010. Fim Glossário para que o conjunto de dados possa ser reduzido, possibilitando a sua análise.

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

Uma distribuição de frequências pode ser expressa através de tabelas ou de gráficos, que terão algumas particularidades dependendo do nível de mensuração da variável e de quantas variáveis serão analisadas. 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).

2.2 – Distribuição de frequências para 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, **frequência absoluta,** Glossário Frequência absoluta: registro dos valores da variável por meio de contagem das ocorrências no conjunto de dados. Fonte: Barbetta, Reis e Bornia, 2010. Fim Glossário ou em números relativos, **frequência relativa ou percentual**. Glossário Frequência relativa ou percentual: registro dos valores da variável por meio de proporção (relativa) ou percentagem (percentual) do total das ocorrências do conjunto de dados. Fonte: Barbetta, Reis e Bornia, 2010. Fim Glossário Ambos os registros devem ser feitos e apresentados: a frequência absoluta permite avaliar se os resultados são sólidos (é temerário tomar decisões com base em pequenas quantidades de dados); já a frequência relativa possibilita comparar os resultados da distribuição de frequências com outros conjuntos de tamanhos diferentes. A distribuição de frequências pode ser apresentada em forma de tabela ou gráfico.

LINK Se alguém diz que 33,33% (percentual) das mulheres de um curso casaram-se 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. LINK

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

Quadro 2 - Opinião dos clientes sobre as concessionárias Toyord

Fonte: elaborado pelo autor

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 frequências. Quando a variável é qualitativa podemos usar dois tipos de gráficos: **em barras** ou **em setores**.

No gráfico de barras (Figura 8) em um dos eixos são colocadas as categorias da variável e no outro as frequências ou percentuais de cada categoria. As barras podem ser horizontais ou verticais (preferencialmente estas). Para os dados do segundo exemplo, usando as frequências:

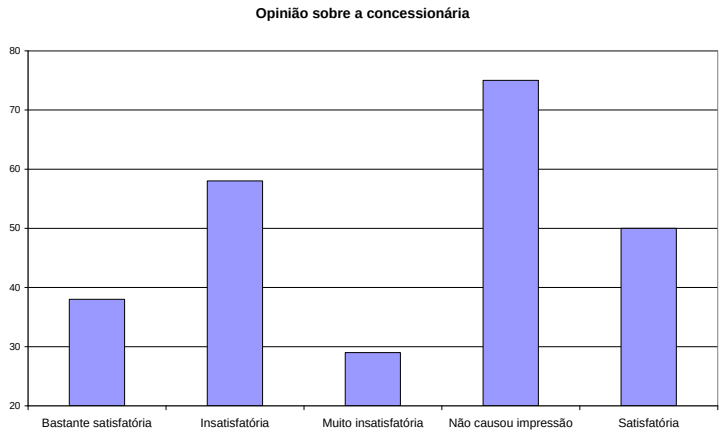


Figura 8 - Gráfico em barras para Opinião sobre as concessionárias

Fonte: adaptada pelo autor a partir de Microsoft ®

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

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

A escala começa em 20, e não em zero. Sendo assim, as diferenças relativas entre as frequências 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 9, com a escala correta.

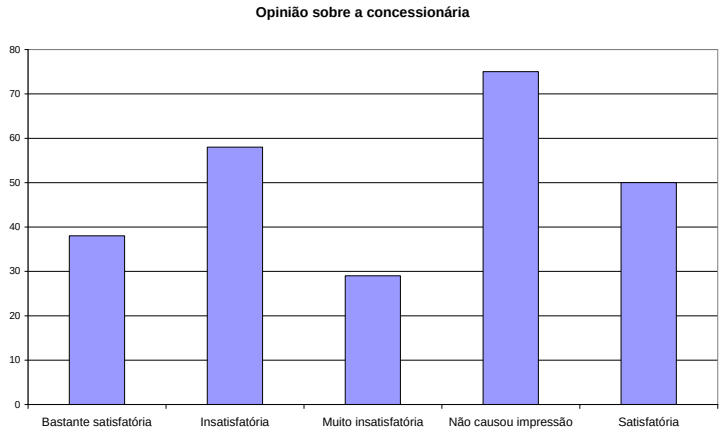


Figura 9 - Gráfico em barras para Opinião sobre as concessionárias

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

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

Observe os valores em graus correspondentes aos resultados do Quadro 1 (Quadro 3).

Valores	Frequência	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 10:

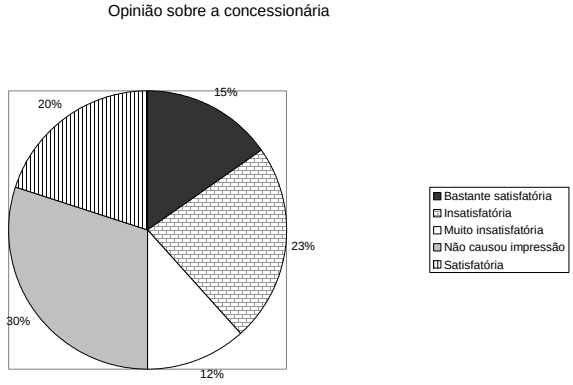


Figura 10 - Gráfico em setores para a Opinião sobre as concessionárias

Fonte: adaptada pelo autor de Microsoft ®

Pela observação dos percentuais é possível perceber o predomino 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.

2.3 – Distribuição de frequências para duas variáveis qualitativas

O administrador frequentemente 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 frequências, também chamada de **tabela de contingências**, ou dupla classificação. 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 frequências de todos os cruzamentos possíveis. Fonte: Barbetta, 2014. Fim Glossário Nela são contadas as frequências de cada cruzamento possível entre os valores das variáveis. A expressão pode incluir o cálculo de percentuais em relação ao total das linhas, colunas ou total geral da tabela. A representação gráfica também é possível. Vamos ver um exemplo.
Para a mesma situação do Exemplo 1. Agora você está interessado em observar relacionamento entre a variável modelo adquirido e a opinião geral do cliente sobre o veículo, e expressa-lo de forma tabular e gráfica.

A variável modelo apresenta 5 resultados possíveis (5 modelos foram considerados nesta pesquisa), e a variável opinião geral pode assumir 4 resultados (Bastante satisfeito, satisfeito, insatisfeito e muito insatisfeito). Isso significa que podemos ter até 20 cruzamentos possíveis para os quais precisamos contar as frequências. Para grandes bases de dados, 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 ® ou o Br.Office Calc®.

Usando uma ferramenta computacional chegaremos ao Quadro 4.

	Opinião geral sobre o veículo				
Modelo	Muito insatisfeito	Insatisfeito	Satisfeito	Bastante satisfeito	Total
		1			1
Chiconaultla	69	11	1	0	81
DeltaForce3	29	22	5	0	56
Valentiniana	11	18	9	3	41
SpaceShuttle	1	14	17	10	42
LuxuriousCar	0	1	9	19	29
Total	110	67	41	32	250

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

Fonte: elaborado pelo autor

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

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

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

digitação na tabulação dos dados, ou o instrumento de pesquisa LINK Conforme vimos na Unidade 1 LINK foi mal projetado, pois muitos elementos da população não forneceram as informações desejadas.

O Quadro 4 pode ser apresentado de forma gráfica, através de um gráfico de barras múltiplas (Figura 11).

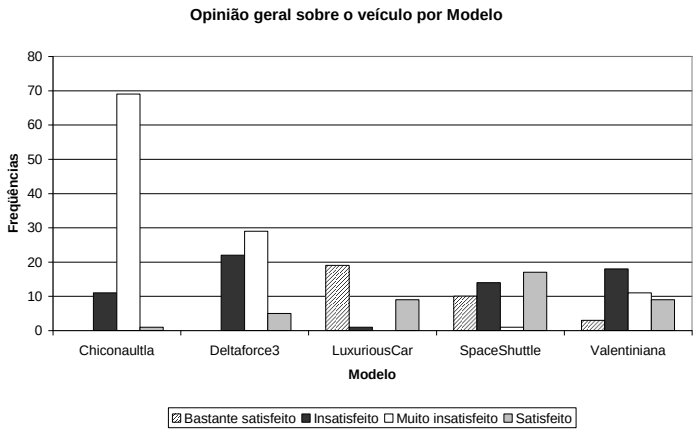


Figura 11 - Gráfico de barras múltiplas: Opinião geral por Modelo

Fonte: adaptado pelo autor de Microsoft ®

As frequências 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, ou aos totais de cada linha ou ao total geral da tabela. Vamos apresentar apenas um dos percentuais possíveis, em relação aos totais das linhas (maiores detalhes nos textos "Como realizar análise exploratória de dados no Microsoft Excel ®" e "Como realizar análise exploratória de dados com o Br.Office Calc ®"):

	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 5 - 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.

2.4 – Distribuição de frequências para uma variável quantitativa

A construção das distribuições de frequências para variáveis quantitativas é semelhante ao caso das variáveis qualitativas: relacionar os valores da variável com as suas

ocorrências no conjunto de dados, 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 frequência para dados não agrupados**. Vamos ver um exemplo.

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

Valores	**Frequência**	**Percentual**
1	19	7,60%
2	29	11,60%
3	43	17,20%
4	42	16,80%
5	57	22,80%
6	60	24,00%
Total	250	100%

Quadro 6 – Número de pessoas usualmente transportadas no veículo

Fonte: elaborado pelo autor

Pela observação do Quadro 6 podemos concluir que os veículos têm uso predominantemente "familiar" (várias pessoas transportadas usualmente). Sabendo disso o administrador pode decidir por direcionar o *marketing* ou mesmo a produção de modelos visando o 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 frequências para variáveis quantitativas discretas graficamente? O Quadro 6 poderia ser representado através de um **Histograma**, um gráfico de barras justapostas, em que as áreas das barras são proporcionais LINK A maioria dos programas (estatísticos ou não) que constroem histogramas para variáveis quantitativas discretas costuma ignorar isso. LINK às frequências de cada valor. Vamos ver (Figura 12):

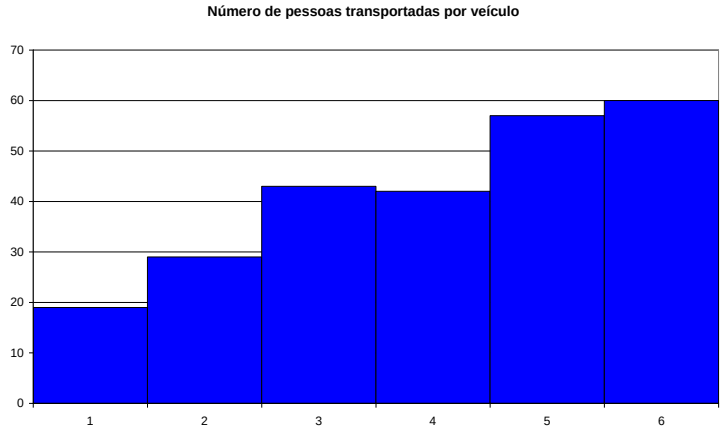


Figura 12 - Histograma do Número de pessoas transportadas por veículo

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 frequência baixa, o que resultaria em uma tabela enorme.

Se o conjunto de dados for pequeno, até 100 observações, é possível usar ferramentas gráficas como o diagrama de pontos e o ramo em folhas. LINK Mais informações veja a seção “Tô afim de saber”. LINK

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 encontram-se em cada faixa. Há duas maneiras de fazer isso:

- através da **categorização** (recodificação) da variável, 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. Fim Glossário por exemplo, todos que ganham até 4 salários mínimos (R$ 1520) 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.
- através de uma **distribuição de frequências para dados agrupados** (ou agrupada em classes), Glossário Distribuição de frequências para dados agrupados: distribuição de frequências na qual os valores da variável são agrupados em faixas de ocorrência, e as frequências contadas para cada faixa, para facilitar o resumo do conjunto de dados, usualmente empregado para variáveis quantitativas contínuas. Fonte: Barbetta, Reis e Bornia, 2010. Fim Glossário processo mais elaborado, e mais “estatístico”, veremos o procedimento a seguir.

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

  - Determinar o intervalo do conjunto (diferença entre o maior e o menor valor do conjunto).
  - Dividir o intervalo em um número conveniente de classes, onde: $\text{No de classes} = \sqrt{\text{No de elementos}}$ 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

Li |--| Ls ambos incluídos

Determinar as frequências de cada classe.

Determinar os pontos médios de cada classe através da média dos 2 limites (serão os representantes das classes).

Vamos ver exemplos de ambas as abordagens.

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

a) Se optássemos por categorizar a variável da seguinte forma: todos que ganham até 4 salários mínimos (R$ 1520) pertencem à classe baixa, todos que ganham entre 4,01 e 20 salários mínimos (até R$ 7600) pertencem à classe média, e acima disso pertencem à classe alta?

b) Se optássemos por uma distribuição de frequências para dados agrupados?

LINK 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. LINK

No caso do item a, a categorização levará à criação de uma nova variável, agora qualitativa, permitindo uma abordagem semelhante a que vimos anteriormente. No Quadro 7 e na Figura 13 estão os resultados: tabela de frequências e gráfico em setores.

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

Quadro 7 - Renda categorizada em classe social

Fonte: elaborado pelo autor

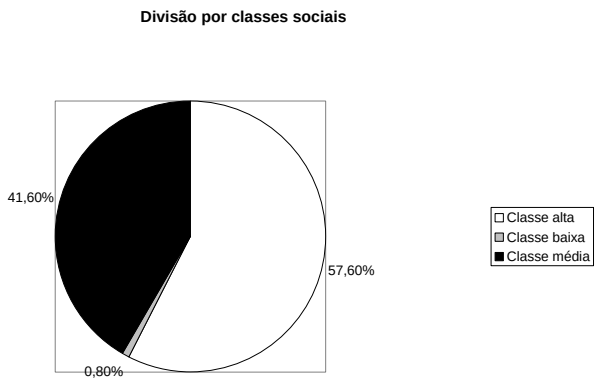


Figura 13 - Gráfico em setores para a Renda categorizada em classes

Fonte: adaptada pelo autor de Microsoft ®

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, LINK Estes valores foram obtidos no arquivo de dados citado no início desta Unidade LINK as classes devem englobar do menor ao maior valor).
- $\text{No de classes} = \sqrt{\text{No de elementos}} = \sqrt{250} = 15{,}81 \cong 16$ 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 arredondado 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)
- Frequências de cada classe (Quadro 8):

Classes	Frequência	Percentual	Pontos médios
1,795|-18,645	98	39,2%	10,22
18,645|-35,495	102	40,8%	27,07
35,495|-52,345	38	15,2%	43,92
52,345|-69,195	9	3,6%	60,77
69,195|-86,045	3	1,2%	77,62
Total	250	100%	-

Quadro 8 - Renda agrupada em classes

Fonte: elaborado pelo autor

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 se mais quais são os seus valores exatos, ou seja, as frequências das classes passam a ser as frequências dos pontos médios. Podemos afirmar que quase 80% dos clientes têm renda até 35,495 salários mínimos.

O Quadro 8 também pode ser representado através de um histograma, (Figura 14) 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, e o ponto inicial de cada barra é o limite inferior da classe, e o ponto final é o limite superior.

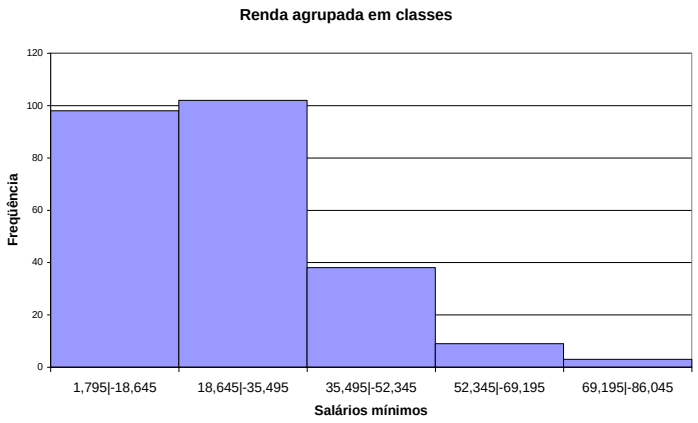


Figura 14 - Histograma para renda agrupada em classes

Fonte: adaptada pelo autor de Microsoft ®

Note que a interpretação é mais direta quando olhamos o gráfico apresentado na Figura 14.

Mas aqui surge um fato interessante. Parece haver contradição com a interpretação do item a, onde concluímos que a maioria absoluta dos clientes é de classe alta. Isso ocorre devido à definição arbitrária das classes, e a 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 frequências para dados agrupados: serão apenas estimativas dos valores reais devido à perda de informação referida acima. LINK A tendência atual é NÃO CALCULAR medidas estatísticas com base em tabelas de dados grupados.LINK

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

2.5 – Distribuição de frequências para 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 estivermos estudando duas variáveis, uma qualitativa e outra quantitativa, a qualitativa será considerada independente (ou de agrupamento) e a quantitativa a dependente. Vejamos dois exemplos rápidos.

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

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

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

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

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

RENDA	**MODELO**					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 frequências de Renda agrupadas em classe por Modelo

Fonte: elaborado pelo autor

Observe a semelhança da tabela mostrada no quadro acima com aquela do Quadro 7. Da mesma forma que lá fizemos, é possível calcular percentuais em relação aos totais

das linhas, colunas ou total geral. LINK Há 249 dados na tabela porque o dado perdido (descoberto no Quadro 4) foi removido do conjunto. LINK

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.

A tabela do Quadro 9 poderia ser expressa através de um gráfico, um **histograma categorizado**. Infelizmente tal gráfico não pode ser feito em uma planilha eletrônica (como o Br.Office Calc ®) sem consideráveis manipulações. Mas, através de um software estatístico, no nosso caso o Statsoft Statistica 6.0 ®, isso é possível (Figura 15):

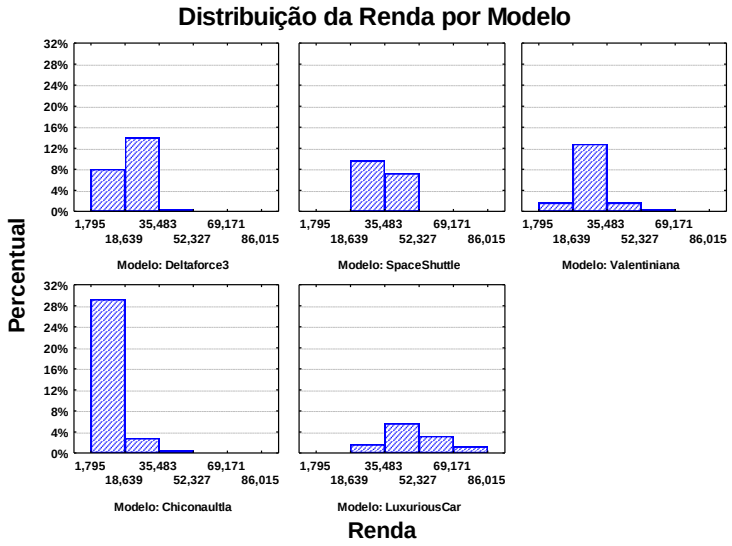


Figura 15 - Histograma categorizado da Renda por Modelo adquirido

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 3 as medidas de síntese e na Unidade 4 estudaremos como analisar duas variáveis quantitativas.

To afim de saber...

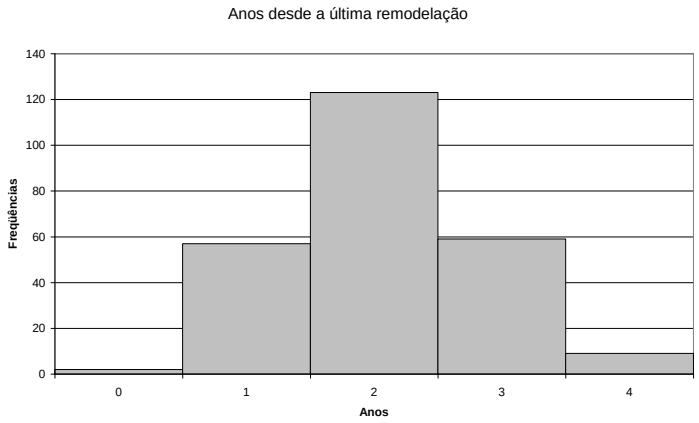
- Sobre como realizar as análises descritas nesta Unidade e na Unidade 3 através do Microsoft Excel ® consulte "Como realizar análise exploratória de dados no Microsoft Excel ®", disponível no ambiente virtual assim como o arquivo de dados usado nos exemplos apresentados.
- Sobre como realizar as análises descritas nesta Unidade e na Unidade 3 através do Br.Office Calc ® consulte "Como realizar análise exploratória de dados com o Br.Office Calc ®" disponível no ambiente virtual assim como o arquivo de dados usado nos exemplos apresentados.

Atividades de aprendizagem

Chegamos ao fim da Unidade 2 da disciplina de Estatística Aplicada a Administração. Agora, chegou o momento de verificar se você teve bom entendimento. Para saber, responda as atividades propostas e encaminhe-as para seu tutor através do Ambiente Virtual de Aprendizagem. As atividades devem ser feitas usando o Microsoft Excel ®, através do arquivo AmostraToyord.xls que está no ambiente virtual, salvo onde indicado.

1) Construa a distribuição de frequências para a variável opinião sobre design (Design) dos veículos da Toyord. Analisando os resultados, os clientes, de uma forma geral, tem boa impressão sobre o design dos veículos da TOYORD? JUSTIFIQUE.

2) A variável anos de remodelação dos veículos (na percepção do cliente) está representada no histograma a seguir:



Fonte: adaptada pelo autor de Microsoft ®

O departamento de marketing alega que precisa de mais orçamento para "convencer" os clientes que os veículos da TOYORD têm design avançado, pois eles creem que a maioria dos clientes acha que eles foram remodelados "há vários anos atrás". Os dados confirmam a crença do departamento de marketing? JUSTIFIQUE!

3) Construa a distribuição de frequências agrupada em classes para a variável quilometragem média mensal percorrida com o veículo. Você considera que os clientes da Toyord usam bastante o veículo ou não? JUSTIFIQUE!

4) Construa a tabela de contingências para modelo por opinião sobre concessionárias (concessionária). Com base nela avalie se os clientes de todos os modelos estão satisfeitos com os serviços prestados. JUSTIFIQUE a sua resposta.

5) Os executivos da Toyord creem que seus clientes mais abastados são mais críticos, tendem a ser mais insatisfeitos com seus veículos. Para verificar se isso é verdade construíram a tabela a seguir. Com base nela, a crença dos executivos é verificada (calcule os percentuais que achar necessários)? JUSTIFIQUE!

Renda	Opinião geral sobre o veículo				
	Bastante satisfeito	Insatisfeito	Muito insatisfeito	Satisfeito	Total
1,795|-18,645	0	16	78	4	98
18,645|-35,495	7	47	26	22	102
35,495|-52,345	15	4	5	14	38
52,345|-69,195	7	0	1	1	9
69,195|-86,045	3	0	0	0	3
Total geral	32	67	110	41	250

Fonte: adaptado pelo autor de Microsoft ®

Resumo

O resumo desta Unidade está mostrado na Figura 16:

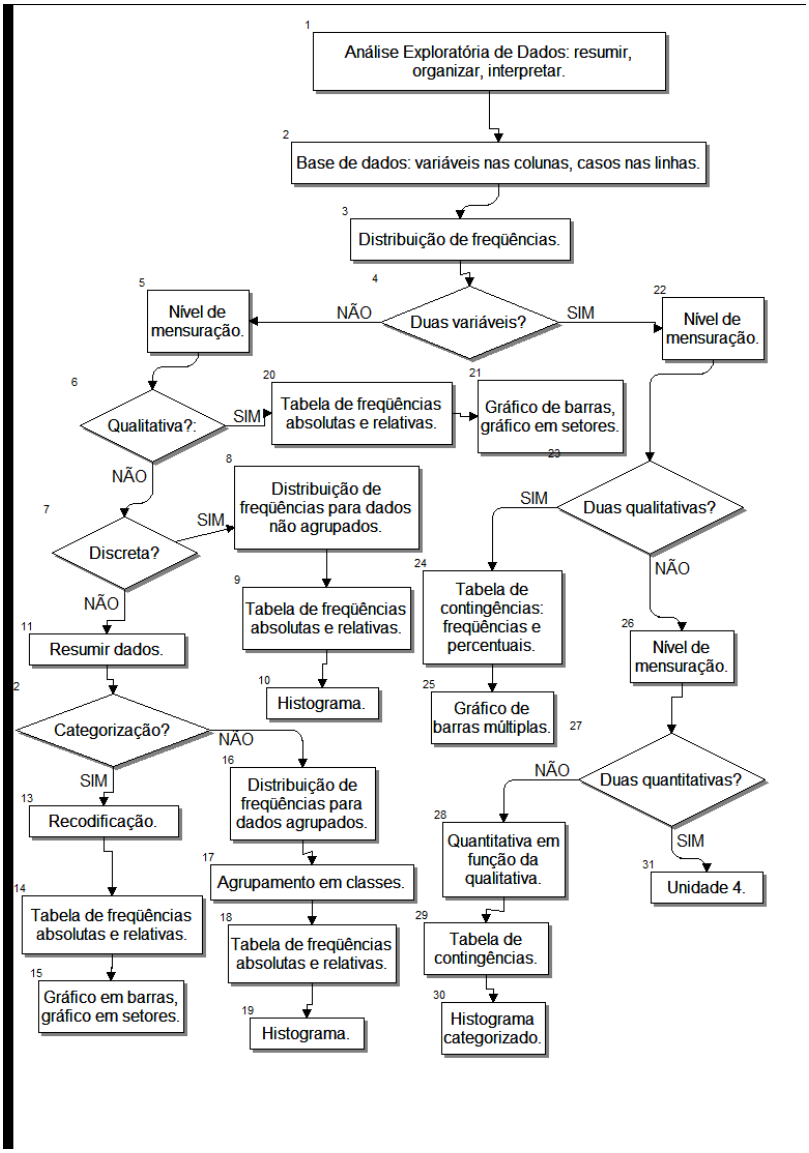


Figura 16 - Resumo da Unidade 2

Fonte: elaborado pelo autor

Caro estudante,

Essa Unidade foi importantíssima para você entender a Análise Exploratória de Dados. Vimos como organizar, interpretar e resumir as informações coletadas, níveis de mensuração e número de variáveis. 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 necessários os variados exemplos propostos para cada categoria estudada. As Figuras, quadros, representações e 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 3

Análise exploratória de dados II

Objetivo

Nesta Unidade você vai aprender mais uma maneira de descrever e analisar um conjunto de dados referente a uma variável quantitativa (discreta ou contínua): através das medidas de síntese. Serão apresentadas as medidas de posição e de dispersão que permitem sintetizar o comportamento da variável individualmente ou em função dos valores de outra variável.

Caro estudante!

Na Unidade 2 estudamos como fazer a descrição tabular e gráfica das variáveis, seja isoladamente ou 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. A sua utilização pode ser feita de forma complementar às técnicas vistas na Unidade 2, 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.

3.1 – Medidas de Posição ou de Tendência Central

As Medidas de Posição procuram caracterizar a tendência central do conjunto, um valor numérico que "represente" o conjunto. 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.

3.1.1 - Média ($\bar{X}$)

A Média aqui citada é a **média aritmética simples**, 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, 2014. Fim GLOSSÁRIO a soma dos valores observados dividida pelo número desses valores. Seja um conjunto de **n** valores de uma variável quantitativa X, a média do conjunto será:

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ Onde x_i é um valor qualquer do conjunto, $\sum_{i=1}^{n} x_i$ é a soma dos valores do conjunto, e n é o tamanho do conjunto. LINK No Microsoft Excel ® e no Br.Office Calc ® a média aritmética simples é implementada através da função MÉDIA( ). LINK

Vamos ver um exemplo que irá 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 8 alunos, as três turmas terão a mesma média: 6.

No exemplo que acabamos de ver as três turmas têm a mesma média (6), então se apenas essa medida fosse utilizada para caracterizá-las poderíamos ter a impressão 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") GLOSSÁRIO Valores discrepantes (*outliers*): valores de uma variável quantitativa que se distanciam muito (para cima ou para baixo) da maioria das observações. Por exemplo, a renda de Bill Gates é um valor discrepante da variável renda de pessoas morando nos EUA. Fonte: adaptado pelo autor de Bussab e Morettin, 2002. Fim GLOSSÁRIO 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 17):

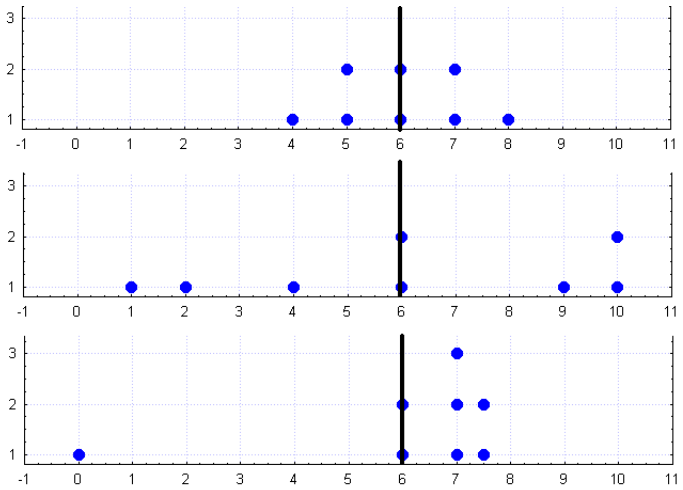


Figura 17 - Interpretação do valor da média

Fonte: adaptada pelo autor de Microsoft® 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 distribuição já é mais irregular, com valores mais "distantes" na parte de baixo, e o terceiro grupo é claramente assimétrica 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, 2014 Fim GLOSSÁRIO em relação à média (que foi distorcida pelo valor discrepante 0). Portanto muito cuidado ao caracterizar um conjunto apenas por sua média. LINK Essa era a grande crítica que era feita nas décadas de 1960 e 1970 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. LINK

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.

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

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

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

Onde k é o número de valores da variável discreta, ou o número de classes da variável agrupada, x_i é um valor qualquer da variável discreta, ou o ponto médio de uma classe

qualquer, f_i é a frequência de um valor qualquer da variável discreta ou de uma classe qualquer, e n é o número total de elementos do conjunto.

Neste segundo exemplo vamos calcular a média do número de pessoas usualmente transportadas no veículo, através da distribuição de frequências obtida no terceiro exemplo exposto na Unidade 2 (Quadro 11).

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

Quadro 11 – Número de pessoas usualmente transportadas no veículo

Fonte: elaborado pelo autor

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

Valores x_i	**Frequência f_i**	**$x_i \times f_i$**
1	19	19
2	29	58
3	43	129
4	42	168
5	57	285
6	60	360

Total	250	1019

Quadro 12 – Número de pessoas usualmente transportadas no veículo

Fonte: elaborado pelo autor

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 LINK Veja novamente a Figuras 12 da Unidade 2, e observe como o valor da média permite equilibrar os pesos, as frequências, dos vários valores da variável. LINK

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

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

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

3.1.2 - 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.

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 4 ª e na 5 ª posição do conjunto. LINK Por esse motivo os dados precisam estar ordenados crescentemente. LINK

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

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

3.1.3 – Moda (Mo)

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

É a medida de posição de obtenção mais simples, e também pode ser usada para variáveis qualitativas, pois apenas registra qual é o valor mais frequente, podendo 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 3 modas: os valores 5, 6 e 7 ocorrem duas vezes cada. A turma B tem duas modas: os valores 6 e 10 ocorrem duas vezes cada. A turma C tem uma moda apenas: o valor 7 ocorre 3 vezes.

3.1.4 – Quartis

Para alguns autores os **quartis** não são medidas de posição, são separatrizes. Porém, como sua forma de cálculo é semelhante a da mediana, resolvemos incluí-los no tópico de Medidas de Posição. Os quartis são medidas que dividem o conjunto em 4 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 anterior, e imediatamente posterior à posição calculada. Se os dados estiverem dispostos em uma distribuição de frequências, utilizar o mesmo procedimento observando as frequências associadas a cada valor (variável discreta) ou ponto médio de classe. LINK No Microsoft

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

No exemplo 6 iremos 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 posição 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$ 6478,05 pelo salário mínimo de maio de 2007). Com base nisso poderíamos ter uma ideia do nível de renda daquela população.

Para encontrar o quartil superior precisamos calcular a média dos valores que estão na 22ª e 23ª posição do conjunto: no caso, 15,820 e 18,275, 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$ 10233,40 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.

3.2 – Medidas de dispersão ou de variabilidade

O objetivo das medidas de dispersão GLOSSÁRIO Medidas de dispersão: medidas numéricas que visam avaliar a variabilidade do conjunto de dados, sintetizando-a em um número. Fonte: elaborado pelo autor. Fim GLOSSÁRIO é medir quão próximos uns dos outros estão os valores de um grupo (e algumas mensuram a dispersão dos dados em torno de uma medida de posição). Com isso é obtido um valor numérico que sintetiza a variabilidade.

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

3.2.1 – Intervalo

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

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

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

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

Turma	Valores

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

Quadro 16 – Notas das turmas A e B

Fonte: elaborado pelo autor.

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 turmas, e essa á principal desvantagem do uso do intervalo como medida de dispersão.

Se colocarmos os dados do Exemplo 7 em um diagrama apropriado (Figura 18):

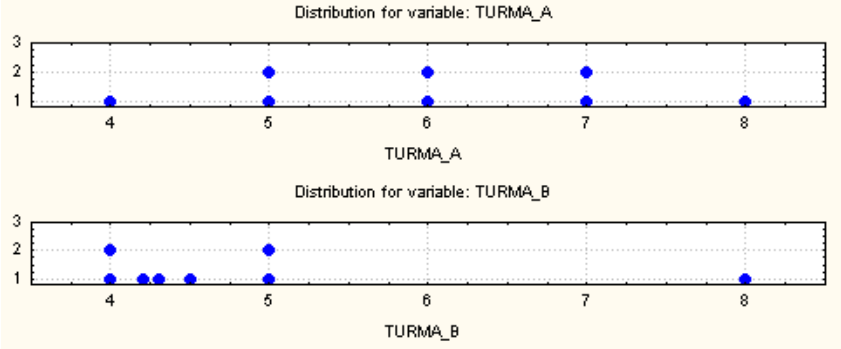


Figura 18 - Desvantagem do uso do intervalo como medida de dispersão

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 ideia de como os dados estão distribuídos entre os extremos (não permite identificar que o valor 8 na turma B é um valor discrepante). LINK No Microsoft Excel ® e no Br.Office Calc ® podemos obter o Intervalo através das funções MÁXIMO ( ) e MÍNIMO ( ) . LINK

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.

3.2.2 - 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.

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

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. LINK A razão dessa distinção será explicada mais adiante na Unidade 5 de Estatística Aplicada à Administração II. 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. LINK

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 2 de Estatística Aplicada à Administração II).

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

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!

3.2.3 - 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 ideia 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 padrões 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:

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

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

Tal como no caso da média pode haver interesse em calcular o desvio padrão de variáveis quantitativas a partir de distribuições de frequências representadas em tabelas. 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 frequências:

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

Onde $\mathbf{x_i}$ é o valor da variável ou ponto médio da classe, $\mathbf{f_i}$ a frequência associada, k é o número de valores da variável discreta (ou o número de classes da variável agrupada), e n é o número de elementos do conjunto. LINK No Microsoft Excel ® e no Br.Office Calc ®

podemos obter o desvio padrão populacional através da função DESVPADP( ) e amostral através da função DESVPAD( ). LINK

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:

$$s = \sqrt{\frac{\sum_{i=1}^{29}\left(x_i^{\,2}\right) - \left[\frac{\left(\sum_{i=1}^{29} x_i\right)^2}{29}\right]}{29-1}} = \sqrt{\frac{17497{,}919125 - \left[\frac{(654{,}935)^2}{29}\right]}{29-1}} = \sqrt{\frac{17497{,}919125 - 14791{,}029456}{28}}$$

$s \cong 9{,}83$ salários mínimos.

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

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

relacionasse média e desvio padrão, o coeficiente de variação percentual, que veremos a seguir.

3.2.4 - Coeficiente de Variação Percentual (c.v.%)

O coeficiente de variação percentual 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, 2010; Anderson, Sweeney e Williams, 2007. Fim GLOSSÁRIO é uma medida de dispersão relativa, pois permite comparar a dispersão de diferentes distribuições (com diferentes médias e desvios padrões).

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

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 padrões dos grupos, mas para comparar a dispersão de distribuições com médias diferentes é imprescindível a utilização do coeficiente de variação percentual.

Você deve se perguntar, mas porque 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 ideia 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. LINK Tô afim de saber: há um pequeno texto sobre o Teorema de Chebyshev no ambiente virtual. Fim LINK.

3.3 - 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 olhe novamente o oitavo exemplo da Unidade 3 verá que construímos distribuições de frequências agrupadas em classes, para a variável renda (quantitativa), em função dos valores da variável modelo (qualitativa). Poderíamos fazer o mesmo com as medidas de síntese! Vamos ver o exemplo a seguir.

Para a mesma situação dos Exemplos 1 e 7 da Unidade 2. Gostaríamos de avaliar, neste 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 ® e do Br.Office Calc ® podemos calcular várias medidas de síntese da variável renda, em função dos modelos de veículos. O Excel ® permite obter as seguintes medidas em função dos valores de outra variável: média, desvio padrão (amostral e populacional), variância (amostral e populacional), mínimo e máximo (infelizmente não permite cálculo de mediana ou quartis). O Calc permite obter as mesmas medidas, mas para cada uma telas é necessário acionar o assistente de dados, enquanto no Excel é possível agrupá-las em uma única tabela. Ao realizar este procedimento, usando os dados do arquivo AmostraToyord.xls vamos obter (Quadro 19):

Modelo	**Medida**	**Valor**
Chiconaultla	Frequência	81
	Mínimo	1,795
	Máximo	40,160
	Média	12,704
	Desvio padrão (amostral)	6,038
DeltaForce3	Frequência	56
	Mínimo	10,820
	Máximo	48,220
	Média	22,063
	Desvio padrão (amostral)	6,956
LuxuriousCar	Frequência	29
	Mínimo	29,800
	Máximo	86,015
	Média	50,932
	Desvio padrão (amostral)	14,922
SpaceShuttle	Frequência	42
	Mínimo	18,865
	Máximo	47,300

	Média	33,050
	Desvio padrão (amostral)	7,620
Valentiniana	Frequência	41
	Mínimo	13,055
	Máximo	65,390
	Média	27,353
	Desvio padrão (amostral)	8,383
Frequência Total		249
Mínimo Total		1,795
Máximo Total		86,015
Média Total		25,105
Desvio padrão (amostral) Total		14,505

Quadro 19 - Medidas de síntese de Renda por Modelo

Fonte: elaborado pelo autor.

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 Total		57,777%

Quadro 20 - Coeficientes de Variação Percentual de Renda por Modelo

Fonte: elaborado pelo autor.

Aparentemente, a relação existente entre renda média e os modelos não se reproduz completamente no que tange à dispersão. Embora o Chiconaultla (modelo mais barato e cujos compradores tem 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 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, quartis inferiores e superiores comportam-se de forma semelhante às médias. A propósito, médias e medianas são próximas, o que indicaria simetria das distribuições das rendas para todos os modelos.

Tô afim de saber:

- Sobre medidas de síntese, assimetria, diagramas em caixa e outros aspectos, procure em BARBETTA, P. A. Estatística Aplicada às Ciências Sociais. 9ª. ed. – Florianópolis: Ed. da UFSC, 2014, capítulo 7.
- Sobre outros tipos de médias (harmônica, geométrica), SPIEGEL, M. R. Estatística. 3ª 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. 2ª ed. – São Paulo: Thomson Learning, 2007, Capítulo 3.
- Sobre Análise Exploratória de Dados utilizando o Excel, LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª 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.
- Sobre como realizar as análises descritas nesta Unidade e na Unidade 4 através do Br.Office Calc ® consulte "Como realizar análise exploratória de dados com o Br.Office Calc ®" disponível no ambiente virtual assim como o arquivo de dados usado nos exemplos apresentados.

Atividades de aprendizagem

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

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

Anos de remodelação	Frequências
0	2
1	57
2	123
3	59
4	9
Total	250

Fonte: elaborado pelo autor.

a) Calcule a média, mediana, moda e quartis da variável anos de remodelação.

b) A direção da Toyord acredita que se uma parcela considerável dos clientes perceber que seus modelos são atualizados (foram remodelados há no máximo 2 anos) o design e o marketing dos veículos estão coerentes. Com base nos resultados da letra a, os dados mostram isso? Justifique.

c) Calcule o intervalo, desvio padrão e coeficiente de variação percentual da variável anos de remodelação.

d) Com base nos resultados dos itens a e c, você considera que os dados estão fortemente concentrados em torno da média? Justifique.

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

a) Com base nos resultados os executivos estão certos? Justifique.

b) Compare com as conclusões que você obteve na questão 5 da Unidade 2.

Resumo

O resumo desta Unidade está demonstrado na Figura 19:

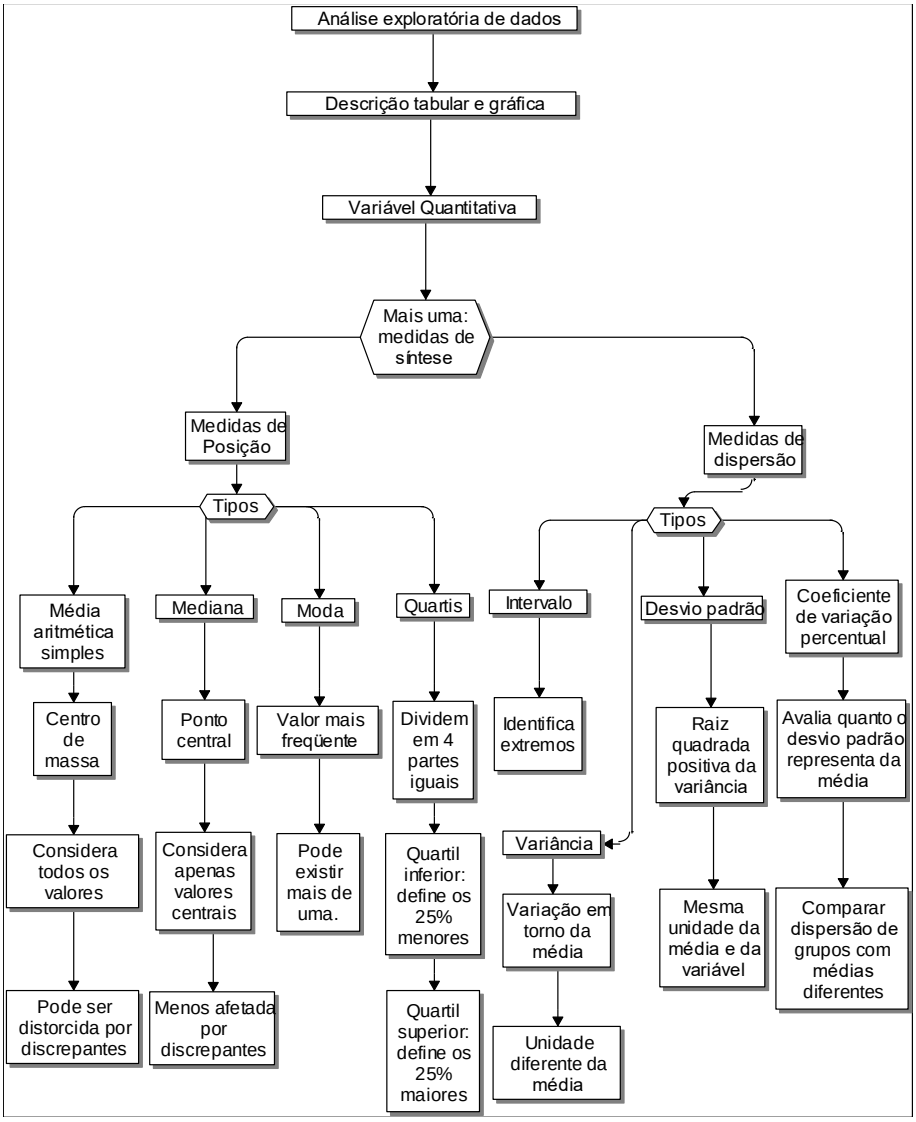


Figura 19 - Resumo da Unidade 3

Fonte: elaborado pelo autor

Com este tópico finalizamos 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 4, veremos os conceitos de correlação e regressão que permitem estudar o relacionamento entre duas variáveis quantitativas e realizar previsões dos valores de uma delas através dos da outra, por meio de um modelo matemático. Vamos em frente e ótimos estudos!!!

Unidade 4

Correlação e Regressão

Objetivo

Nesta Unidade vamos estudar como analisar o relacionamento entre dias variáveis quantitativas (discretas ou contínuas) através de gráficos e medidas de correlação, de maneira a construir modelos de regressão que permitam prever os valores de uma variável em função dos de outra.

Caro estudante,

Na Unidade 3 estudamos as medidas de síntese, mais uma maneira de resumir de resumir um conjunto de dados de uma variável quantitativa: medidas de posição e de dispersão.

Nesta Unidade vamos aprender como analisar o relacionamento entre duas variáveis quantitativas: a força e a direção do relacionamento podem ser mensuradas por uma medida de correlação ou de através de um gráfico; se as variáveis apresentam correlação é possível sugerir um modelo de regressão, que permita prever os valores de uma delas, com base nos da outra, através de uma equação. Um modelo de regressão é uma importante ferramenta de previsão, o que será muito útil na atividade do administrador.

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 Glossário: Correlação: medida de associação entre duas variáveis quantitativas. Fonte: Barbetta, Reis e Bornia, 2010. Fim Glossário 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.

Comentado [MMR4]: A variável independente também pode ser chamada de preditora ou regressora em alguns textos ou aplicativos computacionais.

Há uma outra variável, chamada de dependente, seus valores são resultado da variação dos valores das variáveis Independentes. LINK Reveja as definições de variáveis na Unidade 1. LINK

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

Por exemplo, pode haver correlação entre a pluviosidade mensal (em mm) em Florianópolis e o número de ratos exterminados por mês na cidade de Sidney, na Austrália,

mas seria um pouco forçado imaginar que uma coisa “causou" a outra. É necessário usar bom senso.

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

Para que seja possível avaliar o relacionamento entre duas variáveis (neste caso quaisquer, não apenas quantitativas) os dados devem provir de **observações emparelhadas** Glossário Observações emparelhadas: medidas de duas ou mais variáveis que foram realizadas na mesma unidade experimental/amostral, no mesmo momento. Fonte: elaborado pelo autor. Fim Glossário 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. Além disso, espera-se que haja uma quantidade suficiente de dados para garantir a qualidade da análise.

Comentado [MMR5]: Veja a seguinte matéria: “Quando Nicolas Cage faz filmes, mortes são evitadas”. Disponível em http://www1.folha.uol.com.br/ciencia/2015/05/1626815-quando-nicolas-cage-faz-filmes-mortes-sao-evitadas-veja-outras-correlacoes-estatisticas-que-mentem.shtml, acessado em 14/10/2015.

4.1 – Diagrama de Dispersão

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 ® ou o Br.Office Calc ® LINK Saiba mais no texto “Como realizar análise exploratória de dados no Br.Office Calc ®” LINK.

Através do diagrama de dispersão podemos ter uma ideia inicial de como as variáveis estão relacionadas: a direção da correlação (isto é, quando os valores de **X** aumentam, os valores de **Y** aumentam também ou diminuem), a força da correlação (em que “taxa” os valores de **Y** aumentam ou diminuem em função de **X**) e a natureza da correlação (se é possível ajustar uma reta, parábola, exponencial, aos pontos).

Vejamos alguns exemplos.

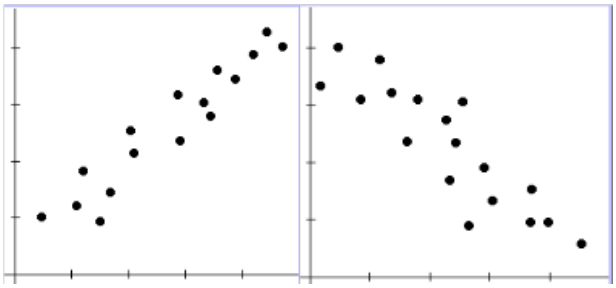


Figura 21 – Diagramas de Dispersão (casos de correlação linear)

Fonte: elaborado pelo autor

No diagrama à esquerda da Figura 21 percebemos duas características: à medida que a variável X aumenta, os valores de Y tendem a aumentar também; seria perfeitamente possível ajustar uma reta crescente que passasse por entre os pontos (obviamente a reta não poderia passar por todos eles). Concluímos então que há correlação linear (porque é possível ajustar uma reta aos dados) positiva (porque as duas variáveis aumentam seus valores conjuntamente). No diagrama à direita da Figura 21 também percebemos duas características: à medida que a variável X aumenta, os valores de Y tendem a diminuir; seria perfeitamente possível ajustar uma reta decrescente que passasse por entre os pontos. Concluímos então que há correlação linear (porque é possível ajustar uma reta aos dados) negativa (porque quando uma das variáveis aumenta seus valores e a outra diminui).

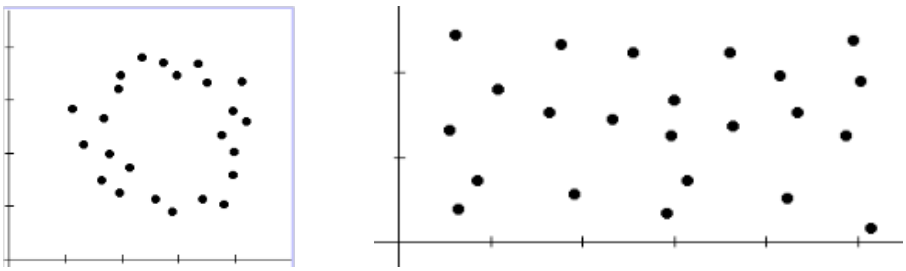


Figura 22 – Diagramas de Dispersão (correlação não linear e ausência de correlação)

Fonte: elaborado pelo autor

No caso do diagrama à esquerda da Figura 22 é óbvio que há alguma espécie de correlação entre as variáveis: os pontos apresentam claramente um padrão, semelhante a um círculo. Contudo, não se trata de uma relação linear, pois seria totalmente inadequado ajustar uma reta aos dados. Assim, há correlação, mas não é linear. No caso do diagrama à direita da Figura 22 há uma situação totalmente diversa dos casos anteriores. NÃO HÁ padrão nos pontos, linear ou não linear, os pontos parecem distribuir-se de forma aleatória. Então, conclui-se que NÃO HÁ CORRELAÇÃO entre as duas variáveis.

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 Iorque. Para tanto coletou um conjunto de 400 pares de observações, e plotou o diagrama de dispersão exposto na Figura 23.

Observando o diagrama (Figura 23) é possível afirmar que o relacionamento entre as variáveis é fortemente linear?

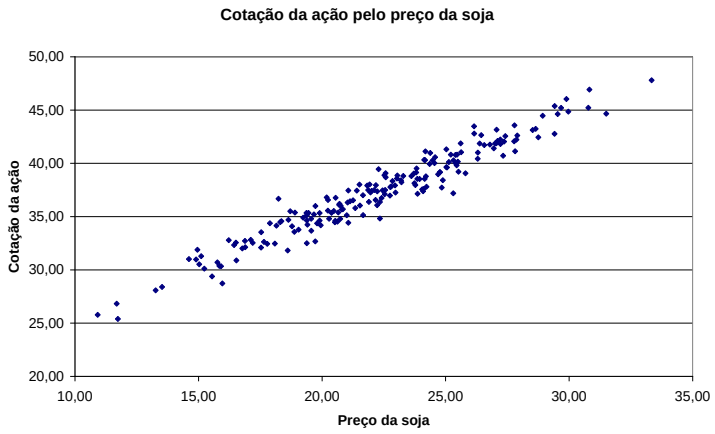


Figura 23 - Diagrama de dispersão de Cotação da ação por Preço da soja

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, é possível 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**. Glossário Série temporal: conjunto de observações de uma variável quantitativa, ordenado no tempo (diário, semanal, mensal, anual). Fonte: Moore, McCabe, Duckworth e Sclove, 2006. Fim Glossário As séries temporais serão estudadas na Unidade 5.

Quando a correlação entre as variáveis é linear é possível quantificar a força desta correlação através de uma medida de síntese, o coeficiente de correlação linear de Pearson.

4.2 – Coeficiente de Correlação Linear de Pearson

Através do diagrama de dispersão é possível identificar se há correlação linear, e se a correlação linear é positiva ou negativa. Quanto mais o diagrama de dispersão aproximar-se de uma reta mais forte será a correlação linear.

É interessante notar que alguns erroneamente confundem "inexistência de correlação linear" com inexistência de correlação entre as duas variáveis. Duas variáveis podem apresentar uma forte correlação não linear, conforme visto na seção anterior.

Se após observar o diagrama de dispersão decidir-se que é razoável considerar que as variáveis possuem um relacionamento linear é possível mensurar a direção e a força desse relacionamento através de um coeficiente de correlação: o coeficiente de correlação

linear de Pearson. Trata-se de um coeficiente adimensional, amostral, que pode ser expresso por:

$$r = \frac{Cov(X,Y)}{s_X \times s_Y} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{n-1}}{s_X \times s_Y}$$

O numerador da expressão é chamado de Covariância de X e Y, que permite mensurar o relacionamento entre as variáveis. A Covariância é dividida pelos desvios padrões de X e Y para que seja eliminado o efeito que uma variável com maiores valores numéricos causaria no resultado, e n é o número de observações.

A covariância permite mensurar o relacionamento entre X e Y:
- quando os valores de X e Y são ambos grandes ou ambos pequenos (as distâncias em relação às médias têm o mesmo sinal) a covariância será grande e positiva.
- quando o valor de X é alto e o de Y é baixo (ou vice-versa) a covariância será grande e negativa.
- dividindo-a por n-1 o seu valor não será mais afetado pelo tamanho da amostra.

Apesar de válida, a expressão mostrada anteriormente costuma levar a resultados que apresentam substanciais erros de arredondamento. A forma do coeficiente de correlação linear de Pearson mais utilizada (inclusive em calculadoras, programas estatísticos e planilhas eletrônicas) é:

$$r = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{\sqrt{\left[n \times \sum_{i=1}^{n}(x_i^{\,2}) - \left(\sum_{i=1}^{n} x_i\right)^2\right]} \times \sqrt{\left[n \times \sum_{i=1}^{n}(y_i^{\,2}) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Para fazer os cálculos é preciso calcular a soma dos valores de X, a soma dos valores de Y, a soma dos valores do produto XY, a soma dos quadrados dos valores de X, a soma dos quadrados dos valores de Y e o número de valores da amostra (n).

O coeficiente de correlação linear de Pearson pode variar de -1 a +1 (passando por zero), e é adimensional: se r = -1 significa que há uma correlação linear negativa perfeita entre as variáveis (todos os pontos no diagrama de dispersão estariam dispostos em uma reta decrescente); se r = +1 significa que há uma correlação linear positiva perfeita entre as variáveis (todos os pontos no diagrama de dispersão estariam dispostos em uma reta crescente); e se r = 0 significa que não há correlação linear entre as variáveis. Admite-se que se |r| > 0,7 a correlação linear pode ser considerada forte.

Comentado [MMR6]: GLOSSÁRIO: Adimensional: sem unidade. Elaborado pelo autor.

Novamente, um alto coeficiente de correlação linear de Pearson (próximo a +1 ou a -1) não significa uma relação de causa e efeito entre as variáveis, apenas que as duas variáveis apresentam aquela tendência de variação conjunta.

Exemplo 1 - Seja um conjunto de dados referente ao número de clientes e as vendas semanais (em $1000) de filiais de uma empresa de entrega de encomendas. Considerando que o número de clientes seja a variável independente (X) e as vendas semanais seja a dependente (Y), pois se imagina que o número de clientes pode influenciar as vendas semanais. O diagrama de dispersão mostrando as duas variáveis é mostrado na Figura 23.

Comentado [MMR7]: Fonte: LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

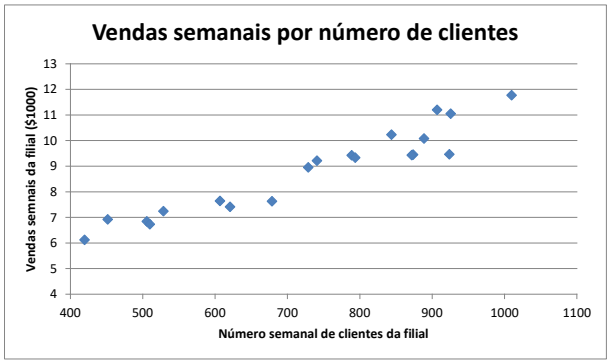


Figura 24 - Diagrama de dispersão de Vendas semanais por número de clientes

Fonte: adaptada pelo autor de Microsoft ®

A correlação entre as variáveis é claramente positiva: maiores valores do número de clientes correspondem a maiores valores de vendas semanais, o que parece plausível. A correlação parece ser muito forte, pois os pontos estão muito próximos. Quanto à natureza,

é possível observar que seria possível ajustar uma reta entre os pontos. Portanto, conclui-se que o relacionamento entre as variáveis é fortemente linear. Para quantificar a força desta relação podemos calcular o coeficiente de correlação linear de Pearson, o que exigirá a obtenção dos somatórios apresentados na fórmula vista anteriormente, cujos cálculos necessários estão no Quadro 22.

Filial	Clientes (X)	Vendas (Y)	X^2	Y^2	XY
1	907	11,2	822649	125,44	10158,4
2	926	11,05	857476	122,1025	10232,3
3	506	6,84	256036	46,7856	3461,04
4	741	9,21	549081	84,8241	6824,61
5	789	9,42	622521	88,7364	7432,38
6	889	10,08	790321	101,6064	8961,12
7	874	9,45	763876	89,3025	8259,3
8	510	6,73	260100	45,2929	3432,3
9	529	7,24	279841	52,4176	3829,96
10	420	6,12	176400	37,4544	2570,4
11	679	7,63	461041	58,2169	5180,77
12	872	9,43	760384	88,9249	8222,96
13	924	9,46	853776	89,4916	8741,04
14	607	7,64	368449	58,3696	4637,48
15	452	6,92	204304	47,8864	3127,84
16	729	8,95	531441	80,1025	6524,55
17	794	9,33	630436	87,0489	7408,02
18	844	10,23	712336	104,6529	8634,12
19	1010	11,77	1020100	138,5329	11887,7
20	621	7,41	385641	54,9081	4601,61
Somatório	14623	176,11	11306209	1602,097	134127,9

Quadro 22 – Dados de Número de clientes e Vendas semanais de filiais de uma empresa de entrega de encomendas.

Fonte: adaptado pelo autor de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Recordando a segunda fórmula do coeficiente de correlação linear de Pearson:

$$r = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{\sqrt{\left[n \times \sum_{i=1}^{n}\left(x_i^{\,2}\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right]} \times \sqrt{\left[n \times \sum_{i=1}^{n}\left(y_i^{\,2}\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Os valores necessários:

- n = 20 (pois há 20 filiais, portanto 20 observações emparelhadas).

- $\sum_{i=1}^{20} x_i = 14623$ $\sum_{i=1}^{20} y_i = 176{,}11$ $\sum_{i=1}^{20} x_i^2 = 11306209$ $\sum_{i=1}^{20} y_i^2 = 1602{,}097$

- $\sum_{i=1}^{20} x_i \times y_i = 134127{,}9$

Substituindo-os na fórmula:

$$r = \frac{20 \times 134127{,}9 - 14623 \times 176{,}11}{\sqrt{\left[20 \times 11306209 - (14623)^2\right]} \times \sqrt{\left[20 \times 1602{,}097 - (176{,}11)^2\right]}} = 0{,}9549$$

Corroborando nossas conclusões anteriores, o coeficiente de correlação linear de Pearson teve resultado positivo, e próximo de 1, indicando forte correlação linear positiva entre as variáveis ao menos para estas filiais. E o valor do coeficiente corresponde à disposição dos pontos no diagrama mostrado na Figura 24, onde era possível ajustar uma reta crescente aos pontos.

O passo lógico seria obter uma equação que permitisse expressar o relacionamento das variáveis, de maneira que seja possível fazer previsões sobre a variável dependente a partir dos valores da variável independente.

4.3 – Regressão Linear Simples

A Análise de Regressão tem por finalidade obter uma função de regressão: uma função matemática que exprima o relacionamento entre duas ou mais variáveis. Se apenas duas variáveis estão envolvidas chama-se de regressão **simples**, se há mais de uma variável independente (e apenas uma dependente) chama-se de regressão **múltipla**.

Comentado [MMR8]: Este nome vem do trabalho de Francis Galton (1822-1911), que identificou que pais altos geravam filhos cujas alturas *regrediam* a um valor, que depois se identificou como sendo a média. Fonte: BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010.

"A função de regressão 'explica' grande parte da variação de **Y** com **X**. Uma parcela da variação permanece sem ser explicada, e é atribuída ao acaso". As mesmas suposições gerais utilizadas na análise de correlação são necessárias: a existência de uma teoria que "explique" o relacionamento entre as variáveis, observações emparelhadas, a

quantidade suficiente de dados. Além desses, para realizar a Análise de Regressão, seja linear (reta), exponencial, logarítmica, polinomial, etc., alguns pressupostos básicos são necessários:

- supõe-se que há uma função que justifica **em média**, a variação de uma variável em função da variação da outra;
- os pares x,y terão uma variação em torno da linha representativa desta função, devido a uma variação aleatória adicional, chamada de **variância residual** ou **resíduo**;
- a variável **X** (variável INDEPENDENTE) é suposta **sem erro**.
- a variável **Y** (variável DEPENDENTE) terá uma variação nos seus valores "dependente[1]" de **X** se houver regressão.
- a função de regressão será: $\mathbf{Y} = \boldsymbol{\varphi}(\mathbf{X}) + \boldsymbol{\varepsilon}$ onde $\boldsymbol{\varphi}(\mathbf{X})$ é a função de regressão propriamente dita e $\boldsymbol{\varepsilon}$ é a componente aleatória de **Y**, devida ao acaso (e que SEMPRE existirá).
- a variação residual de **Y** em torno da linha teórica de regressão segue uma distribuição com média zero e desvio padrão constante (independente dos valores de **X**).
- para se decidir pela utilização de um modelo de regressão devem existir evidências NÃO ESTATÍSTICAS que indiquem relação causal entre as variáveis (alguma lei da física por exemplo, como a Lei de Hook).

Uma vez conhecida a forma da linha de regressão o problema resume-se a **estimar seus parâmetros**.

Na regressão linear simples há apenas duas variáveis envolvidas, e o modelo mais usado, por ser o mais simples, é o modelo linear (reta). Este modelo é bastante difundido porque muitos relacionamentos entre variáveis podem ser descritos através de uma reta, seja utilizando os dados originais, seja após aplicar alguma transformação (logarítmica, exponencial, etc.) a eles que cause a *linearização* da curva.

[1] Foi colocado entre aspas porque a existência de regressão NÃO IMPLICA necessariamente em que Y depende de X, apenas que elas têm uma variação relacionada, que pode ser causada por uma outra variável.

A reta teórica será $\mathbf{Y} = \boldsymbol{\beta} \times \mathbf{X} + \boldsymbol{\alpha}$ e os coeficientes $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ serão estimados através dos valores amostrais **b** e **a** respectivamente,: $\hat{Y} = b \times X + a$, onde $\hat{\mathbf{Y}}$ é a estimativa de **Y**, b é o coeficiente angular da reta (a sua inclinação), e a é o coeficiente linear (o ponto onde a reta toca o eixo Y).

A "melhor reta", a que passará o mais próximo possível dos pontos observados, será encontrada pelo método dos mínimos quadrados: são encontrados os coeficientes que minimizam os quadrados dos desvios (resíduos) de cada ponto do diagrama de dispersão em relação a uma reta teórica. Temos os seguintes valores de **b** e **a**:

$$b = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n \times \sum_{i=1}^{n}(x_i^{\,2}) - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad a = \frac{\sum_{i=1}^{n} y_i - b \times \sum_{i=1}^{n} x_i}{n}$$

Muitas calculadoras já têm estas fórmulas programadas em um módulo estatístico (juntamente com a fórmula do coeficiente de correlação linear de Pearson). Além disso, planilhas eletrônicas e programas estatísticos também fazem tais cálculos.

Exemplo 2 – Para os dados do Exemplo 1, obter os coeficientes da reta de mínimos quadrados.

Na Figura 24 vimos que o relacionamento entre as variáveis número de clientes e vendas podia ser considerado linear, e que se tratava de uma relação positiva, ou seja, seria possível ajustar uma reta crescente aos dados. Isso foi corroborado pelo valor do coeficiente de correlação linear de Pearson, próximo de 1. Todos os somatórios necessários já foram calculados:

- $\sum_{i=1}^{20} x_i = 14623$ $\sum_{i=1}^{20} y_i = 176{,}11$ $\sum_{i=1}^{20} x_i^{\,2} = 11306209$ $\sum_{i=1}^{20} y_i^{\,2} = 1602{,}097$
- $\sum_{i=1}^{20} x_i \times y_i = 134127{,}9$ n = 20

Substituindo os valores na fórmula do coeficiente b:

Comentado [MMR9]: Glossário: Coeficiente angular indica a inclinação da reta e coeficiente linear o ponto onde a reta corta o eixo de Y. Fonte: BARBETTA, P. A. Estatística Aplicada às Ciências Sociais. 9ª. ed. – Florianópolis: Ed. UFSC, 2014

Comentado [MMR10]: Glossário. Método dos mínimos quadrados: método de estimação dos coeficientes de uma equação que consiste em obter os coeficientes que minimizem a soma dos quadrados dos erros (diferenças entre o valor real e o valor predito pela equação). Fonte: BARBETTA, P.A., REIS, M.M., BORNIA, A.C. Estatística para Cursos de Engenharia e Informática. 3ª ed. São Paulo: Atlas, 2010.

Comentado [MMR11]: Glossário. Resíduo é a diferença entre o valor real da variável dependente (Y) e o valor dela predito pelo modelo de regressão a partir do valor da variável independente ou X. Fonte: elaborado pelo autor.

$$b = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n \times \sum_{i=1}^{n}(x_i^{\,2}) - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{20 \times 134127{,}9 - (14623 \times 176{,}11)}{20 \times 11306209 - (14623)^2} = 0{,}008729$$

Agora é possível obter o valor de a:

$$a = \frac{\sum_{i=1}^{n} y_i - b \times \sum_{i=1}^{n} x_i}{n} = \frac{176{,}11 - 0{,}008729 \times 14623}{20} = 2{,}423$$

A equação da reta será então: $\hat{Y} = 0{,}008729 \times X + 2{,}423$. Isso significa que para cada acréscimo de 1 cliente na filial espera-se um aumento de 0,008729 ($1000) nas vendas, e que 2,423 é o ponto onde a reta tocará o eixo Y.

Vejamos como ficaria o diagrama de dispersão com a reta acima traçada sobre ele, na Figura 25.

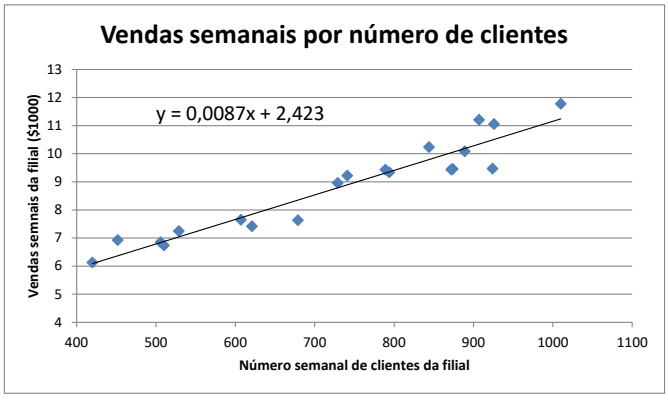


Figura 25 - Diagrama de dispersão de Vendas semanais por número de clientes com equação de reta

Fonte: adaptada pelo autor de Microsoft ®

Diversos programas estatísticos e mesmo planilhas eletrônicas (como o Microsoft Excel ®) permitem obter os coeficientes de mínimos quadrados para vários modelos de regressão, como o linear, polinômios de vários graus, logarítmico, exponencial, potência, entre outros, abaixo alguns deles, permitindo estimar os valores de Y através dos valores de

X (a estimativa de Y é denotada como $\hat{Y}$):

- linear (reta) - $\hat{Y} = b \times X + a$;

- polinômio de segundo grau - $\hat{Y} = c \times X^2 + b \times X + a$

- logarítmico - $\hat{Y} = b \times Ln(X) + a$;

- potência - $\hat{Y} = b \times X^a$;

- exponencial - $\hat{Y} = b \times e^{a \times X}$

Os aplicativos computacionais têm algoritmos que obtêm os coeficientes de mínimos quadrados dos vários modelos a partir dos dados fornecidos, facilitando em muito o processo de análise.

Se houver mais de uma variável independente (preditora), a regressão é chamada de múltipla, e devido à complexidade matemática para obtenção dos coeficientes do modelo, mesmo para o caso linear, é recomendável a utilização de aplicativos computacionais (mais detalhes em "Saiba mais").

4.4 – Coeficiente de determinação

Alguns símbolos precisam ser relembrados:

$\bar{Y}$ é a média aritmética dos valores **observados** de **Y**.

$\hat{Y}_i$ é um valor genérico **predito** de **Y** através do modelo de regressão (qualquer modelo).

$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$: medida da variabilidade **total** dos dados em torno da média de **Y**.

$\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$: medida da parcela da variabilidade dos dados em torno da média de **Y** "**explicada**" pela regressão.

$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$: medida da parcela da variabilidade dos dados em torno da média de **Y** "**não explicada**" pela regressão, chamada também de variação **residual**.

Comentado [MMR12]: Glossário: Ln(X) é o logaritmo neperiano de X, ou seja, o logaritmo com base e, sendo e a constante de Euler, que vale aproximadamente 2,71. Fonte: elaborado pelo autor.

E: $\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$ (a soma da variação explicada com a variação residual resulta na variação total).

Neste ponto é interessante introduzir coeficiente de determinação, o $\mathbf{r^2}$. Este coeficiente descreve a proporção da variabilidade média de **Y** que é explicada pela variação de **X** através do modelo de regressão (QUALQUER modelo). Sua fórmula geral é:

$$r^2=\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}=\frac{variânciaexplicada}{\text{var}\,iância\,total}$$

Para o caso linear (reta) o coeficiente de determinação será simplesmente o quadrado do coeficiente de correlação linear de Pearson (**r**), e como ele será um valor adimensional, mas pode variar apenas de 0 a +1.O coeficiente de determinação é uma boa medida da aderência do modelo de regressão aos dados, quanto mais próximo de +1 maior a parcela da variabilidade média total de **Y** que é explicada pela variação de **X** através do modelo.

A partir de que valor o modelo de regressão é adequado? Para coeficientes de determinação superiores a 0,5 (mais de 50% da variabilidade média total de Y é explicada pela variação de X através do modelo de regressão). Para o caso *linear* isso significa que o **módulo** do coeficiente de correlação linear deve ser maior do que 0,7 para que a regressão linear seja uma boa opção.

Exemplo 3 – Para os dados do Exemplo 1 obter o coeficiente de determinação para o modelo linear e interprete o resultado.

Como se trata de um modelo linear, podemos obter o coeficiente de determinação elevando o coeficiente de correlação linear de Pearson (calculado no Exemplo 1) ao quadrado.

$$r^2 = 0{,}9549^2 \cong 0{,}9118$$

Em média 91,18% da variabilidade de Y pode ser "explicada" pela variabilidade de X através do modelo linear $\hat{Y}=0{,}0087\times X+2{,}423$.

O valor do r^2 é substancialmente maior do que 0,5, indicando que o modelo linear apropriado para os dados (corroborando as conclusões dos Exemplos 1 e 2).

Embora útil para regressão simples o r^2 não é apropriado para regressão múltipla, pois à medida que são acrescentadas mais variáveis independentes (preditoras, regressoras) ao modelo o seu valor vai "inflando", mesmo que as variáveis não acrescentem muito à qualidade do resultado final. Mesmo fazendo algumas modificações que resultam no r^2 ajustado, o coeficiente é insuficiente para atestar a validade do modelo e de que as suposições necessárias são atendidos. Para isso é preciso fazer a análise dos resíduos do modelo.

Comentado [MMR13]: Ver WALPOLE, R.E., MYERS, R.H., MYERS, S. L., YE, K. Probabilidade e Estatística para Engenharia e Ciências. São Paulo: Pearson Prentice Hall, 2009. Seção 12.6

4.5 – Análise de Resíduos

Idealmente a adequação de um modelo de regressão é realizada através da análise dos seus resíduos. Os resíduos são as diferenças entre os valores ***observados*** da variável independente e os valores ***preditos*** da variável independente através do modelo de regressão.

$$Resíduo_i = Y_i - \hat{Y}_i$$

Conforme visto anteriormente, se o modelo de regressão é adequado os resíduos devem ter média zero, e sua variância deve ser constante: variando os valores de X, ou das previsões, os resíduos apresentam variabilidade semelhante (o modelo não será apropriado, por exemplo, se para pequenos valores de X os resíduos forem pequenos, e para valores maiores eles forem grandes, ou vice-versa).

Para tornar a análise mais confiável, sem que as grandezas dos resíduos venham prejudicá-la recomenda-se ***padronizar*** os resíduos: calcula-se o desvio padrão dos resíduos e divide-se cada um deles pelo desvio padrão (não há necessidade de subtrair a média, pois supõe-se que ela vale zero).

$$Resíduo\ padronizado_i = \frac{Y_i - \hat{Y}_i}{s_{Resíduos}}$$

Para fazer a análise de resíduos precisamos construir pelo menos dois diagramas de dispersão:

- um que relacione os resíduos padronizados com os próprios valores preditos da variável independente;
- outro que relacione os resíduos padronizados com os valores da variável independente.

Se o modelo de regressão é adequado os resíduos padronizados não podem apresentar quaisquer padrões, eles devem distribuir-se de forma aleatória nos dois diagramas, atendendo os seguintes critérios:

- a quantidade de resíduos padronizados positivos deve ser aproximadamente igual à quantidade de negativos.
- a grandeza dos resíduos padronizados positivos deve ser aproximadamente igual a dos negativos, para todos os valores preditos da variável dependente, e para todos os valores da variável independente.
- não pode haver padrões não aleatórios (tendências crescentes ou decrescentes, curvas, etc.) em nenhum dos diagramas; em outras palavras é preciso que os pontos sejam dispostos em "**nuvem**".

Somente se ***todas*** estas condições forem satisfeitas é que podemos considerar o modelo de regressão apropriado. Se houver dois ou mais modelos apropriados, escolhemos o mais simples, ou aquele que apresentar o mais alto coeficiente de determinação. Os diagramas deveriam ser como a Figura 26.

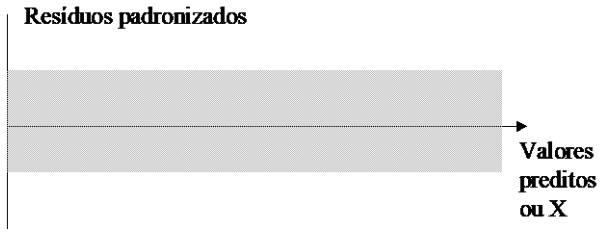


Figura 26 – Formato esperado dos resíduos de um modelo de regressão apropriado

Fonte: elaborado pelo autor

A análise de resíduos pode ser usada para regressão simples ou múltipla. Enquanto na regressão simples o diagrama de dispersão permite avaliar facilmente a correlação entre

as duas variáveis, quando há mais de três variáveis envolvidas a visualização do relacionamento não é mais possível, mas os resíduos podem ser avaliados.

Exemplo 3 – Fazer a análise dos resíduos do modelo linear obtido no Exemplo 2, para os dados do Exemplo 1.

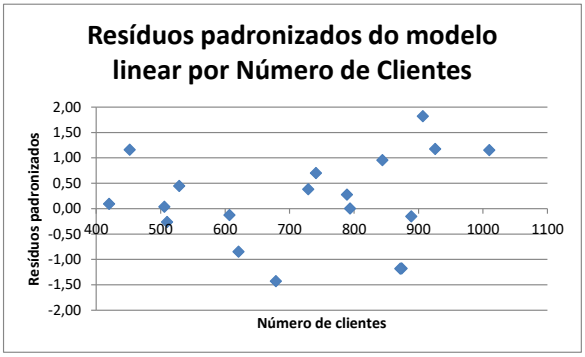
No Exemplo 2 o modelo obtido para prever as vendas (em $1000) das filiais em função do número de clientes foi: $\hat{Y} = 0{,}008729 \times X + 2{,}423$. Para obter os resíduos é preciso fazer as previsões para os dados apresentados no Exemplo 1, calcular os resíduos para cada um deles, depois calcular o desvio padrão dos resíduos, e obter os resíduos padronizados, tal como no Quadro 23.

Filial	Clientes (X)	Vendas (Y)	$\hat{Y} = 0{,}008729 \times X + 2{,}423$	$Resíduo_i = Y_i - \hat{Y}_i$	Resíduo Padron. $_i$
1	907	11,2	$\hat{Y} = 0{,}008729 \times 907 + 2{,}423 = 10{,}3139$	0,8861	1,8152
2	926	11,05	$\hat{Y} = 0{,}008729 \times 926 + 2{,}423 = 10{,}4792$	0,5708	1,1693
3	506	6,84	$\hat{Y} = 0{,}008729 \times 506 + 2{,}423 = 6{,}8252$	0,0148	0,0303
4	741	9,21	$\hat{Y} = 0{,}008729 \times 741 + 2{,}423 = 8{,}8697$	0,3403	0,6971
5	789	9,42	$\hat{Y} = 0{,}008729 \times 789 + 2{,}423 = 9{,}2873$	0,1327	0,2718
6	889	10,08	$\hat{Y} = 0{,}008729 \times 889 + 2{,}423 = 10{,}1573$	-0,0773	-0,1584
7	874	9,45	$\hat{Y} = 0{,}008729 \times 874 + 2{,}423 = 10{,}0268$	-0,5768	-1,1816
8	510	6,73	$\hat{Y} = 0{,}008729 \times 510 + 2{,}423 = 6{,}86$	-0,1300	-0,2663
9	529	7,24	$\hat{Y} = 0{,}008729 \times 529 + 2{,}423 = 7{,}0253$	0,2147	0,4398
10	420	6,12	$\hat{Y} = 0{,}008729 \times 420 + 2{,}423 = 6{,}077$	0,0430	0,0881
11	679	7,63	$\hat{Y} = 0{,}008729 \times 679 + 2{,}423 = 8{,}3303$	-0,7003	-1,4346
12	872	9,43	$\hat{Y} = 0{,}008729 \times 872 + 2{,}423 = 10{,}0094$	-0,5794	-1,1869
13	924	9,46	$\hat{Y} = 0{,}008729 \times 924 + 2{,}423 = 10{,}4618$	-1,0018	-2,0522
14	607	7,64	$\hat{Y} = 0{,}008729 \times 607 + 2{,}423 = 7{,}7039$	-0,0639	-0,1309
15	452	6,92	$\hat{Y} = 0{,}008729 \times 452 + 2{,}423 = 6{,}3554$	0,5646	1,1566
16	729	8,95	$\hat{Y} = 0{,}008729 \times 729 + 2{,}423 = 8{,}7653$	0,1847	0,3784
17	794	9,33	$\hat{Y} = 0{,}008729 \times 794 + 2{,}423 = 9{,}3308$	-0,0008	-0,0016
18	844	10,23	$\hat{Y} = 0{,}008729 \times 844 + 2{,}423 = 9{,}7658$	0,4642	0,9509
19	1010	11,77	$\hat{Y} = 0{,}008729 \times 1010 + 2{,}423 = 11{,}21$	0,5600	1,1472
20	621	7,41	$\hat{Y} = 0{,}008729 \times 621 + 2{,}423 = 7{,}8257$	-0,4157	-0,8516

Quadro 23 – Dados de Número de clientes e Vendas semanais de filiais de uma empresa de entrega de encomendas, previsões por um modelo linear, resíduos e resíduos padronizados.

Fonte: adaptado pelo autor de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

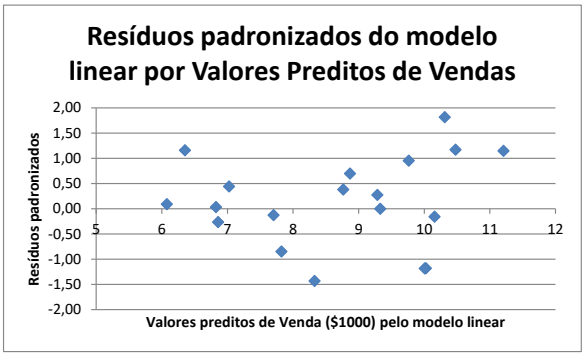
O desvio padrão dos resíduos (necessário para obter os resíduos padronizados) vale 0,4881 ($1000). Os diagramas de dispersão dos resíduos padronizados do modelo linear pelo número de clientes (variável independente) e para os valores preditos de Vendas são mostrados nas Figuras 27 e 28, e as análises dos resíduos são apresentadas ao lado.



A quantidade de resíduos positivos (12) é semelhante a de negativos (8). Todos os resíduos (exceto um) encontram-se a 1,5 desvios padrões acima ou abaixo de zero. Os pontos parecem distribuir-se aleatoriamente para todo o intervalo de valores do Número de Clientes. Por estas razões o modelo linear (que gerou os resíduos) parece apropriado para descrever o relacionamento entre o Número de clientes e as Vendas.

Figura 27 – Resíduos padronizados do modelo linear por Número de clientes

Fonte: adaptada pelo autor de Microsoft ®



A quantidade de resíduos positivos (12) é semelhante a de negativos (8). Todos os resíduos (exceto um) encontram-se a 1,5 desvios padrões acima ou abaixo de zero. Os pontos parecem distribuir-se aleatoriamente para todos os valores preditos de Venda pelo modelo linear. Por estas razões o modelo linear (que gerou os resíduos) parece apropriado para descrever o relacionamento entre o Número de clientes e as Vendas.

Figura 28 – Resíduos padronizados do modelo linear por valores preditos de Venda

Fonte: adaptada pelo autor de Microsoft ®

Quanto mais dados (pares de observações) estiverem disponíveis, melhor a qualidade do modelo obtido e mais conclusiva a análise dos resíduos, como pode ser visto no próximo exemplo.

Exemplo 4 - Estamos avaliando o relacionamento entre as variáveis venda de refrigerantes (em R$ 1000) e temperatura ambiente (em graus Celsius) nos meses de verão. Na Figura 29 vemos o diagrama de dispersão das duas variáveis (temperatura é a independente e vendas a dependente), com dois modelos ajustados através do Microsoft Excel:
reta (Y=255,17×X-6451,7) e parábola (polinômio de 2º grau: $Y=23,039\times X^2-1220,1\times X+17074$).

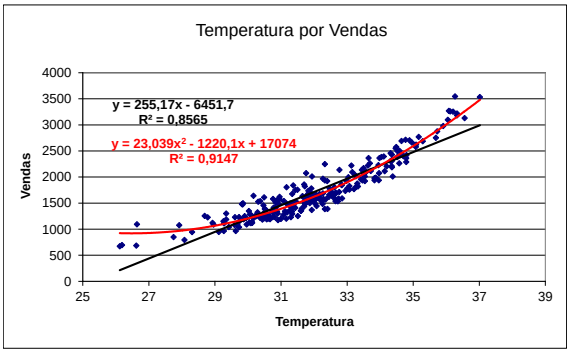


Figura 29 - Diagrama de dispersão de Vendas de refrigerantes por Temperatura com equações de reta e de parábola e coeficientes de determinação

Fonte: adaptada pelo autor de Microsoft ®

Observando o diagrama podemos ver que a parábola (polinômio de 2º grau) aparenta melhor ajuste aos dados, pois "segue" melhor o seu comportamento do que a reta. Os resíduos do modelo de parábola provavelmente serão menores do que os da reta, o que pode ser constatado também pelo seu coeficiente de determinação (0,8631), que é maior do que o da reta (0,8049).

Devido à grande quantidade de dados o procedimento de obtenção dos resíduos padronizados dos modelos (tal como visto no Exemplo 3) não será mostrado aqui, passando diretamente para a apresentação dos diagramas de dispersão dos resíduos dos modelos de reta e de parábola. Na Figura 30 os resíduos padronizados para o modelo de reta.

Comentado [MMR14]: Nas atividades de aprendizagem serão disponibilizados dados completos para a análise de regressão, mas deverá ser realizada através de uma planilha eletrônica.

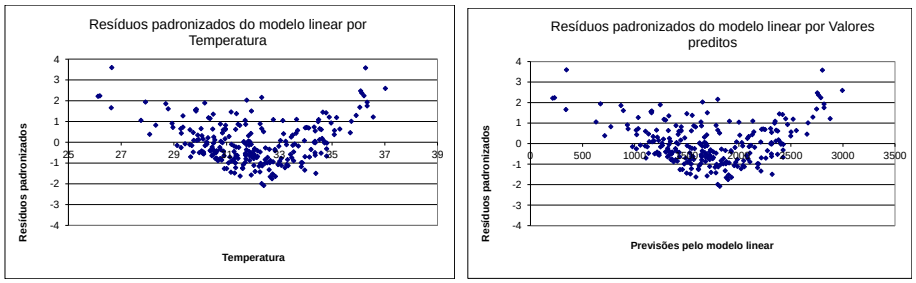


Figura 30 – Diagramas de dispersão dos resíduos padronizados do modelo de reta por Temperatura e valores preditos de Venda de refrigerante.

Fonte: adaptada pelo autor de Microsoft ®

A quantidade de resíduos positivos parece semelhante a de negativos (deveríamos contá-los por meio de algum procedimento computacional), a linha do zero parece "dividir" o número de pontos em duas partes iguais em ambos os diagramas. A maioria esmagadora dos pontos positivos concentra-se abaixo de 2 desvios padrões (linha do 2), e maioria dos negativos também (acima da linha -2), em ambos os diagramas. Há claramente padrão em ambos os diagramas. Para valores menores de temperatura e valores preditos os resíduos são positivos e maiores. À medida que a temperatura e os valores preditos vão aumentando os valores dos resíduos vão diminuindo, tornando-se negativos, até que passam a subir novamente. Em outras palavras, o comportamento dos resíduos do modelo da reta NÃO É ALEATÓRIO. Por estas razões o modelo de reta (que gerou os resíduos) NÃO parece apropriado para descrever o relacionamento entre as variáveis.

Na Figura 31 os resíduos padronizados para o modelo de parábola.

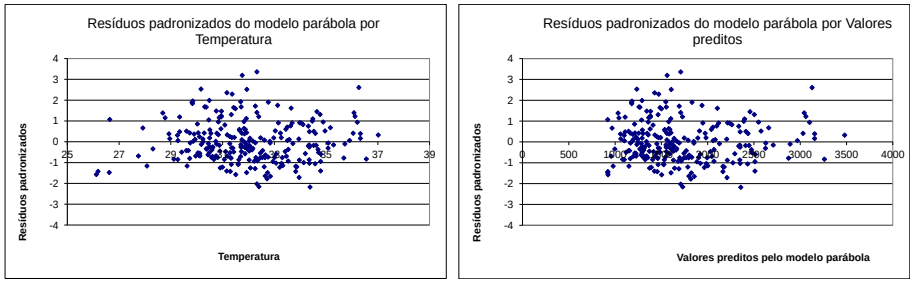


Figura 31 – Diagramas de dispersão dos resíduos padronizados do modelo de parábola por Temperatura e valores preditos de Venda de refrigerante.

Fonte: adaptada pelo autor de Microsoft ®

A quantidade de resíduos positivos e negativos é aparentemente bastante semelhante em ambos os diagramas (a linha do zero divide os pontos em duas "metades" similares). Em ambos os diagramas os resíduos positivos e negativos têm grandezas semelhantes, distantes no máximo a 2 desvios padrões do zero, para a maioria dos pontos. Em ambos os diagramas NÃO são identificados padrões, os pontos parecem distribuir-se de forma aleatória, formando uma "nuvem". Por estas razões o modelo de parábola (que gerou os resíduos) parece apropriado para descrever o relacionamento entre a Temperatura e as Vendas de refrigerantes.

"E se todos os modelos de regressão produzirem resíduos padronizados com comportamento semelhante ao da Figura 31?" Como escolher o melhor modelo? Neste caso devemos selecionar aquele que apresentar o maior coeficiente de determinação. "E se os coeficientes de determinação dos modelos forem semelhantes (menos de 5% de diferença)?" Neste caso devemos usar a REGRA DA PARCIMÔNIA: escolher o modelo de regressão mais simples: modelo mais simples têm menos coeficientes para estimar, e/ou sua equação é mais simples de operar[2].

Comentado [MMR15]: Regra da Parcimônia: escolher o modelo mais simples sempre que dois ou mais forem apropriados. Fonte: LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Saiba mais...

Sobre diagramas de dispersão, correlação e regressão linear simples procure em BARBETTA, P. A. Estatística Aplicada às Ciências Sociais. 9ª. ed. – Florianópolis: Ed. da UFSC, 2014, capítulo 13.

Sobre análise de resíduos e análise de correlação e regressão simples e múltipla no Microsoft Excel ® procure em LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Para saber como realizar as análises descritas nesta Unidade através do Microsoft ® consulte *Como realizar análise de regressão no Microsoft Excel ®,* disponível no Ambiente Virtual assim como os arquivos de dados usados nos exemplos apresentados.

[2] O modelo linear (reta) é considerado como o mais simples por ter apenas dois coeficientes que precisam ser estimados, e por sua facilidade de cálculo. Além disso, outros modelos podem ser linearizados através de algumas transformações matemáticas.

O resumo desta Unidade está demonstrado na Figura 32.

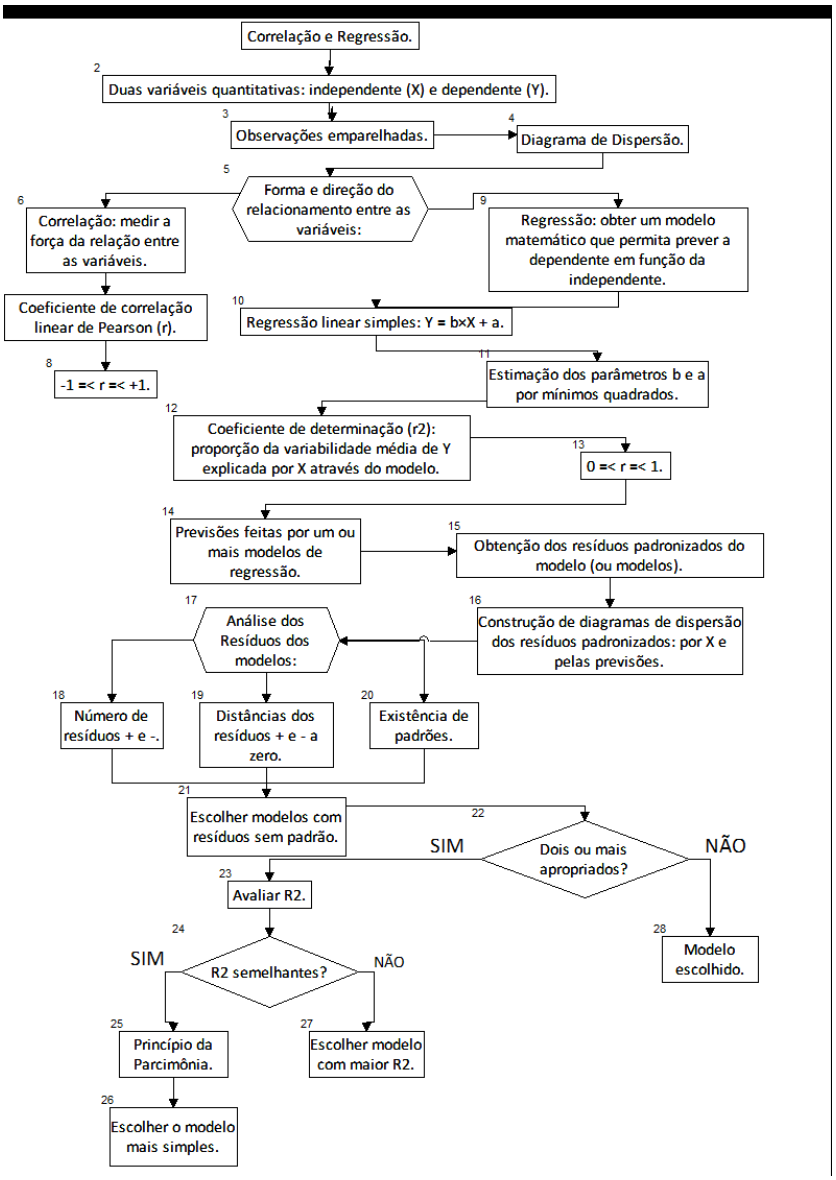


Figura 32 - Resumo da Unidade 4

Fonte: elaborado pelo autor

Atividades de aprendizagem

Os dados para todas as atividades, além dos usados nos Exemplos 1, 2, 3 e 4, encontram-se no arquivo Bidimensional.xls que está no Ambiente Virtual de Ensino-Aprendizagem. As atividades devem ser feitas usando o Microsoft Excel ® ou o Br.Office Calc ®.

1) Uma rede de lojas de vendas por atacado quer avaliar o desempenho de suas filiais, e de quebra verificar a viabilidade de implantar uma nova loja em Joinville, SC. Produziu a tabela a seguir, relacionando o número de clientes com as vendas em milhares de reais em um determinado mês. Com base nela responda as questões apresentadas.

Filial	Número de clientes	Vendas (R$ mil)
1	423	88
2	898	192
3	1095	196
4	1001	191
5	597	100
6	1200	240
7	862	169
8	1300	240
9	845	157
10	440	120
11	922	160
12	620	135
13	876	155
14	745	141
15	1345	250
16	865	172
17	1170	203
18	692	138
19	955	182
20	913	177
21	845	164
22	1004	189
23	1003	208
24	1200	201
25	712	118

a) Qual é a variável independente? Qual é a variável dependente? JUSTIFIQUE sua resposta.

b) Construa um diagrama de dispersão. Com base no diagrama você sugere a adoção de um modelo linear (reta) para o relacionamento entre as variáveis? JUSTIFIQUE sua resposta.
c) Calcule os coeficientes de correlação linear de Pearson e de determinação com os dados da tabela ao lado. Com base nos resultados você sugere a adoção de um modelo linear para o relacionamento entre as variáveis? JUSTIFIQUE sua resposta. Compare com a resposta do item b.
d) Calcule os coeficientes angular e linear da melhor reta que pode ser ajustada aos dados. Interprete o significado do coeficiente angular.
e) Faça a análise de resíduos: calcule os valores preditos de Y pelo modelo, calcule os resíduos (diferença entre valores reais e preditos de Y), calcule os resíduos padronizados (dividindo cada resíduo pelo desvio padrão de todos os resíduos), e construa o diagrama de dispersão dos resíduos padronizados pelos valores preditos de Y. Com base neste diagrama de dispersão você sugere a adoção de um modelo linear para o relacionamento entre as variáveis? JUSTIFIQUE sua resposta/ Compare com as respostas dos itens b e c.
f) Uma pesquisa de mercado detectou que uma nova loja em Joinville teria cerca de 900 clientes em potencial. O custo operacional mensal de uma loja capaz de atender tal número está por volta de 190 mil reais. Usando o modelo desenvolvido no item d estime o valor de vendas para 900 clientes e decida se a loja deve ser aberta ou não.
Adaptado de LEVINE, D.M., BERENSON, M.L., STEPHAN, D., Estatística: Teoria e Aplicações usando Microsoft ® Excel em Português. Rio de Janeiro: LTC, 2000.

2) Como corretor incansável você decidiu estudar um pouco mais os valores de avaliação e de venda dos últimos imóveis negociados pela sua imobiliária. Há interesse em construir um modelo relacionando preço de venda e valor de avaliação do imóvel, com o objetivo de fazer algumas previsões. Você somente negocia com casas, e obteve os valores de avaliação e de venda de 30 unidades, descritas na tabela a seguir. Com base nela responda as questões apresentadas.
a) Qual é a variável independente? Qual é a variável dependente? JUSTIFIQUE sua resposta.

Casa	Preço de avaliação (R$ 1000)	Preço de venda (R$ 1000)
1	94,1	78,17
2	101,9	80,24
3	88,65	74,03
4	115,5	86,31
5	87,5	75,22
6	72	65,54
7	91,5	72,43
8	113,9	85,61
9	69,34	60,8
10	96,9	81,88
11	96	79,11
12	61,9	59,93
13	93	75,27
14	109,5	85,88
15	93,75	76,64
16	106,7	84,36
17	81,5	72,94
18	94,5	76,5
19	69	66,28
20	96,9	79,74
21	86,5	72,78
22	97,9	77,9
23	83	74,31
24	97,3	79,85
25	100,8	84,78
26	97,9	81,61
27	90,5	74,92
28	97	79,98
29	92	77,96
30	95,9	79,07

b) Construa um diagrama de dispersão. Com base no diagrama você sugere a adoção de um modelo linear (reta) para o relacionamento entre as variáveis? JUSTIFIQUE sua resposta.

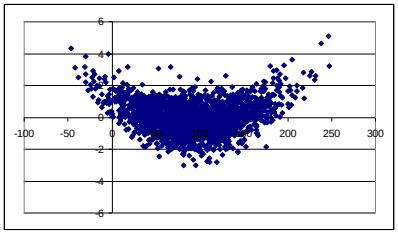
c) Calcule os coeficientes de correlação linear de Pearson e de determinação com os dados da tabela ao lado. Com base nos resultados você sugere a adoção de um modelo linear para o relacionamento entre as variáveis? JUSTIFIQUE sua resposta. Compare com a resposta do item b.

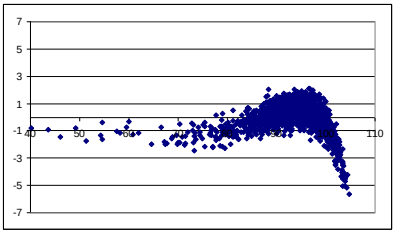
d) Calcule os coeficientes angular e linear da melhor reta que pode ser ajustada aos dados. Interprete o valor do coeficiente angular.

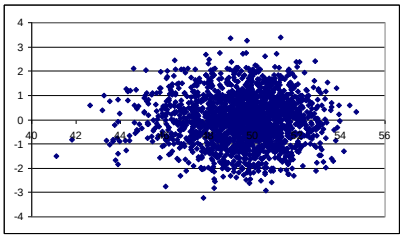
e) Faça a análise de resíduos: calcule os valores preditos de Y pelo modelo, calcule os resíduos (diferença entre valores reais e preditos de Y), calcule os resíduos padronizados

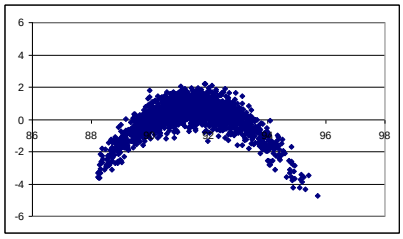
(dividindo cada resíduo pelo desvio padrão de todos os resíduos), e construa o diagrama de dispersão dos resíduos padronizados pelos valores preditos de Y. Com base neste diagrama de dispersão você sugere a adoção de um modelo linear para o relacionamento entre as variáveis? JUSTIFIQUE sua resposta/ Compare com as respostas dos itens b e c.
Adaptado de LEVINE, D.M., BERENSON, M.L., STEPHAN, D., Estatística: Teoria e Aplicações usando Microsoft ® Excel em Português. Rio de Janeiro: LTC, 2000.

3) Os diagramas de dispersão mostram no eixo horizontal os valores preditos pelo modelo de regressão e no eixo vertical os resíduos padronizados. Faça a análise dos resíduos e emita uma opinião sobre a adequação do modelo obtido para o relacionamento entre as variáveis. JUSTIFIQUE a sua resposta.

a)


b)


c)


d)


Unidade 5

Análise de Séries Temporais

Objetivo

Nesta Unidade você aprenderá como identificar padrões em uma série temporal (ordenada no tempo: em anos, meses, dias) de uma variável quantitativa, de maneira a prever o comportamento futuro desta variável e auxiliar no processo de tomada de decisão.

Conceito de Série Temporal

Caro estudante!

Na Unidade 4 estudamos Correlação e Regressão, sendo que ao final nos aprofundamos na construção de modelos de regressão, que podem ser usados para previsão.

Quando a variável independente é uma medida de tempo temos uma série temporal.

Nesta Unidade vamos aprender como decompor as componentes de uma série temporal de acordo com o modelo clássico em tendência, variações sazonais, ciclos e componentes irregulares. Tal conhecimento permitirá fazer previsões sobre o comportamento futuro da variável, auxiliando na tomada de decisão.

Comentado [MMR16]: Glossário - Modelo clássico de série temporal: modelo em que a série é suposta como resultado da agregação de quatro componentes (tendência, ciclos, sazonalidade e componentes irregulares). Fonte: Fonte: LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

"Série Temporal é um conjunto de observações sobre uma variável, ordenado no tempo", e registrado em períodos regulares Glossário Série temporal: conjunto de observações de uma variável quantitativa, ordenado no tempo (diário, semanal, mensal, anual). Fonte: Moore, McCabe, Duckworth e Sclove, 2006. Fim Glossário. Podemos enumerar os seguintes exemplos de séries temporais: temperaturas máximas e mínimas diárias em uma cidade, vendas mensais de uma empresa, valores mensais do IPC-A, valores de fechamento diários do IBOVESPA, resultado de um eletroencefalograma, gráfico de controle de um processo produtivo.

A suposição básica que norteia a análise de séries temporais é que há um sistema causal mais ou menos constante, relacionado com o tempo, que exerceu influência sobre os dados no passado e pode continuar a fazê-lo no futuro. Este sistema causal costuma atuar criando padrões não aleatórios que podem ser detectados em um gráfico da série temporal, ou mediante algum outro processo estatístico.

O objetivo da análise de séries temporais é identificar padrões não aleatórios na série temporal de uma variável de interesse, e a observação deste comportamento passado pode permitir fazer previsões sobre o futuro, orientando a tomada de decisões.

Vamos ver um gráfico de uma série temporal.

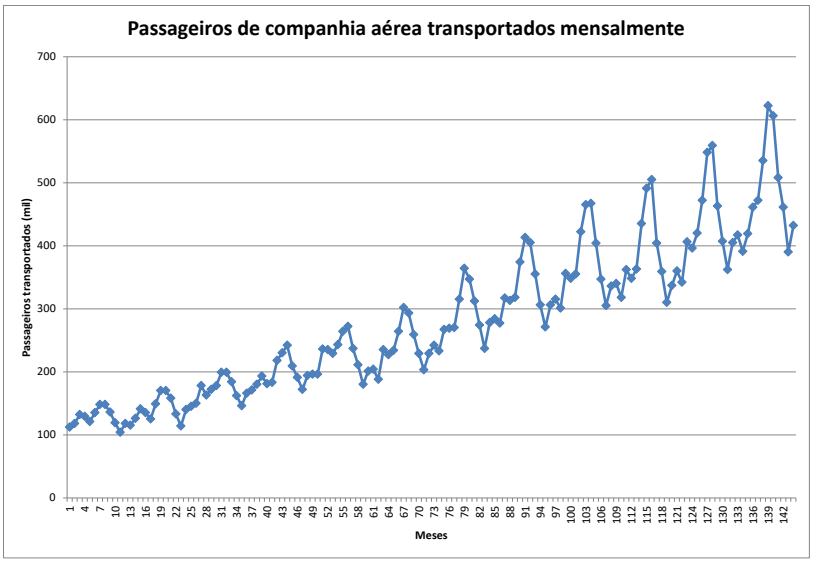


Figura 33 – Série mensal do número de passageiros (em mil) transportados por uma companhia aérea

Fonte: adaptado pelo autor de Microsoft ® a partir de dados de Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). Time Series Analysis, Forecasting and Control, 3rd ed. Prentice Hall, Englewood Clifs, NJ.

Quais padrões podemos identificar na Figura 33?

- observe que há uma tendência crescente no número de passageiros transportados (ou pelo menos havia antes de 11 de setembro de 2001...);
- há uma sucessão regular de "picos e vales" no número de passageiros transportados, isso deve ser causado pelas oscilações devido a feriados, períodos de férias escolares, etc., que estão geralmente relacionados às estações do ano, e que se repetem todo ano (com maior ou menor intensidade).

Em outras palavras, identificamos dois padrões que podem tornar a ocorrer no futuro: crescimento no número de passageiros transportados e flutuações sazonais. Tais padrões poderiam ser incorporados a um modelo estatístico, possibilitando fazer previsões que auxiliarão na tomada de decisões.

Vamos observar mais um conjunto de dados, a produção mensal de veículos no Brasil entre janeiro de 1997 e dezembro de 2014.

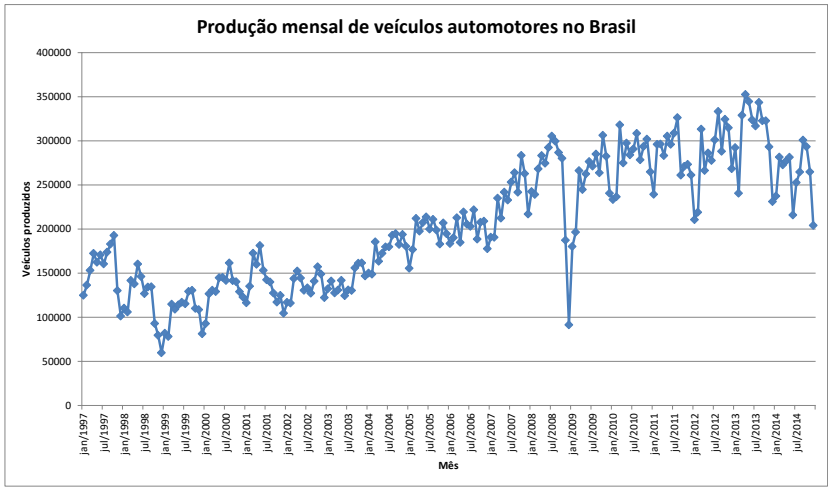


Figura 34 – Série mensal da produção de veículos automotores no Brasil de janeiro de 1997 a dezembro de 2014

Fonte: adaptado pelo autor de Microsoft ® a partir de dados da ANFAVEA – Associação Nacional dos Fabricantes de Veículos Automotores, disponíveis em http://www.anfavea.com.br/tabelas.html, acessados em 13/11/2015.

Comentado [MMR17]: Estes dados estão no arquivo Séries Temporais, disponível no ambiente virtual.

Quais padrões podemos identificar na Figura 34?

- observe que há uma tendência crescente no número de veículos produzidos (começando em cerca de 125000 em janeiro de 1997 e terminando em 200000 em dezembro de 2014);
- as flutuações (picos e vales) não são tão regulares quanto as identificadas na Figura 33;
- observa-se uma queda na produção no mês de janeiro de 2009, em fins de 2008 a produção mensal estava em torno de 300000 veículos, e caiu para menos de 100000 naquele mês (provavelmente por causa da crise mundial no último trimestre de 2008).

Nas Figuras 33 e 34 era claramente visível um comportamento crescente na série, mas isso pode não acontecer: o comportamento pode ser decrescente, ou pode ser estável no longo prazo. Para este último caso há o exemplo da Figura 35.

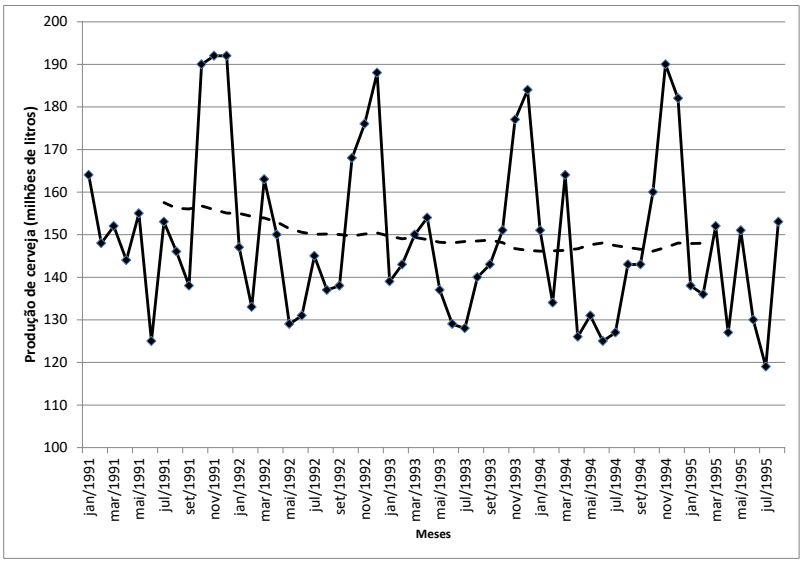


Figura 35 – Produção mensal de cerveja em milhões de litros da Austrália de Janeiro de 1991 a Agosto de 1995.

Fonte: adaptado pelo autor de Microsoft ® a partir de MAKRIDAKIS, S., WHEELWRIGHT, S.C., HYNDMAN, R.J. Forecasting: methods and applications. 3rd ed.- New York: Wiley, 1998.

No caso da Figura 35 os valores de produção parecem flutuar em torno de uma média mais ou menos constante (a linha pontilhada, veremos posteriormente que se trata de uma média móvel), de cerca de 150 Ml. Parece haver sazonalidade também, pois os picos de produção de cerveja repetem-se nos meses de novembro e dezembro de cada ano.

O problema fundamental é utilizar um modelo que permita incluir os vários tipos de padrões, possibilitando realizar previsões. O ponto de partida é realizar a ***decomposição*** da série em padrões. Uma das maneiras de fazer isso é através da decomposição no modelo clássico das séries temporais.

Comentado [MMR18]: GLOSSÁRIO: decomposição de uma série temporal consiste em realizar operações matemáticas e gráficas para identificar seus padrões de longo prazo, curto prazo e eventos fortuitos. Fonte: elaborado pelo autor.

5.1 – Modelo clássico das séries temporais

Segundo o modelo clássico todas as séries temporais são compostas de quatro padrões:

- tendência (**T**), que é o comportamento de longo prazo da série, que pode ser causada pelo crescimento demográfico, ou mudança gradual de hábitos de consumo, ou qualquer outro aspecto que afete a variável de interesse no longo prazo;
- variações cíclicas ou ciclos (C), flutuações nos valores da variável com duração ***superior*** a um ano, e que se repetem com certa periodicidade, que podem ser resultado de variações da economia como períodos de crescimento ou recessão, ou fenômenos climáticos como o El Niño (que se repete com periodicidade superior a um ano);
- variações sazonais ou sazonalidade (S), flutuações nos valores da variável com duração ***inferior*** a um ano, e que se repetem todos os anos, geralmente em função das estações do ano (ou em função de feriados ou festas populares, ou por exigências legais, como o período para entrega da declaração de Imposto de Renda); se os dados forem registrados ***anualmente*** NÃO haverá influência da sazonalidade na série;
- variações irregulares (I), que são as flutuações inexplicáveis, resultado de fatos fortuitos e inesperados como catástrofes naturais, atentados terroristas como o de 11 de setembro de 2001, decisões intempestivas de governos, etc.

Comentado [MMR19]: Alguns autores não incluem as variações cíclicas no modelo clássico da série temporal porque às vezes é preciso uma série muito longa para visualizá-los, o que pode ser difícil de obter, ou as condições que influenciaram a série no passado distante podem não mais existir, ou ainda porque sua periodicidade pode variar muito, dificultando a sua inclusão no modelo de previsão.

Aqui é importante salientar que nem sempre uma série temporal, mesmo que o modelo clássico seja considerado apropriado para analisá-la, irá apresentar todos os componentes citados acima:

- a série pode apresentar apenas variações irregulares: não se percebe comportamento crescente ou decrescente de longo prazo (tendência), ou flutuações sazonais ou cíclicas.
- a série pode apresentar apenas tendência e variações irregulares: não são identificadas flutuações sazonais ou cíclicas, apenas o comportamento crescente/decrescente de longo prazo e as variações aleatórias.
- a série pode apresenta apenas variações sazonais e irregulares: o comportamento de longo prazo da série é aproximadamente constante, mas observam-se flutuações dentro dos períodos de um ano, que se repetem todos os anos.

- quaisquer outras combinações possíveis.

A decomposição da série permitirá identificar quais componentes estão atuando naquele conjunto em particular, além de possibilitar obter índices e/ou equações para realizar previsões para períodos futuros da série.

A questão crucial do modelo clássico é decidir como será a equação que relaciona as componentes com a variável. Há duas opções: o modelo aditivo ou o modelo multiplicativo:

- No modelo **aditivo** o valor da série (Y) será o resultado da soma dos valores das componentes (que apresentam a mesma unidade da variável):

Y = T + C + S + I ou **Y = T + C + I** (se os dados forem registrados anualmente)

Nas *previsões* não temos como incluir a componente irregular no modelo, pois ela é resultado de fatos fortuitos, teoricamente imprevisíveis. Todas as componentes têm a mesma unidade da série: se esta estiver em milhões de reais, todas também terão tal unidade.

- Pode ser usado também o modelo **multiplicativo**, no qual o *produto* das componentes resultará na variável da série:

Y = T × C × S × I ou **Y = T × C × I** (se os dados forem registrados anualmente)

Novamente, não incluímos a componente irregular nas previsões. Há, porém, uma diferença crucial: apenas a tendência tem a mesma unidade da variável. As demais componentes têm valores que modificam a tendência: assumem valores em torno de 1 (se maiores do que 1 aumentam a tendência, se menores diminuem a tendência, se exatamente iguais a 1 não causam efeito).

Chamando a variável de interesse de Y, a equação de sua série temporal seria:

$$Y = f(T,C,S,I)$$

Qual é o melhor modelo? Dependerá dos dados da própria série, das características intrínsecas do problema. Apresentaremos posteriormente medidas que possibilitam avaliar a adequação das previsões feitas por um modelo.

5.2 – Obtenção da tendência de uma série temporal

A tendência descreve o comportamento da variável retratada na série temporal no longo prazo. Há três objetivos básicos na sua identificação: avaliar o seu comportamento para utilizá-lo em previsões, removê-la da série para facilitar a visualização das outras componentes, ou ainda identificar o nível da série (o valor ou faixa típica de valores que a variável pode assumir, se não for observado comportamento crescente ou decrescente no longo prazo). A tendência será a mesma tanto para o modelo aditivo quanto para o multiplicativo.

Neste texto veremos a obtenção da tendência por duas formas: através de um modelo de regressão (como o modelo linear - reta) ou através de médias móveis.

Comentado [MMR20]: Uma outra forma popular de obtenção de tendência é o ajuste, ou ajustamento exponencial, que não deixa de ser uma média móvel, mas exponencialmente ponderada: as medidas mais distantes têm um peso exponencialmente menor do que as medidas mais próximas do ponto onde está sendo calculado o valor ajustado. Fonte: LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

5.2.1 – Obtenção de tendência por mínimos quadrados

O procedimento é semelhante ao usado na regressão linear simples (ver Unidade 4, seção 4.3), mas agora a variável independente será *sempre* o tempo. Para uma série registrada anualmente, por exemplo, de 2005 a 2014, a variável independente assumiria os valores dos anos. Para uma série registrada mensalmente, por exemplo, com 60 meses, a variável independente poderia assumir os valores de 1 a 60. As equações podem ser as mesmas usadas anteriormente (a estimativa do valor da série, Y, é denotada como $\hat{Y}$), e que também podem ter seus coeficientes obtidos por aplicativos computacionais:

- linear (reta) - $\hat{Y} = b \times x + a$;
- polinômio de segundo grau - $\hat{Y} = c \times x^2 + b \times x + a$
- logarítmico - $\hat{Y} = b \times Ln(x) + a$;
- potência - $\hat{Y} = b \times x^a$;
- exponencial - $\hat{Y} = b \times e^{a \times x}$

Para o caso da reta pode ser a mesma equação utilizada na seção 4.3:

$$b = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n \times \sum_{i=1}^{n}(x_i^{\,2}) - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad a = \frac{\sum_{i=1}^{n} y_i - b \times \sum_{i=1}^{n} x_i}{n}$$

Para os dados mostrados na Figura 33, através do Microsoft Excel ® é possível ajustar as cinco tendências mostradas acima, o resultado está na Figura 36.

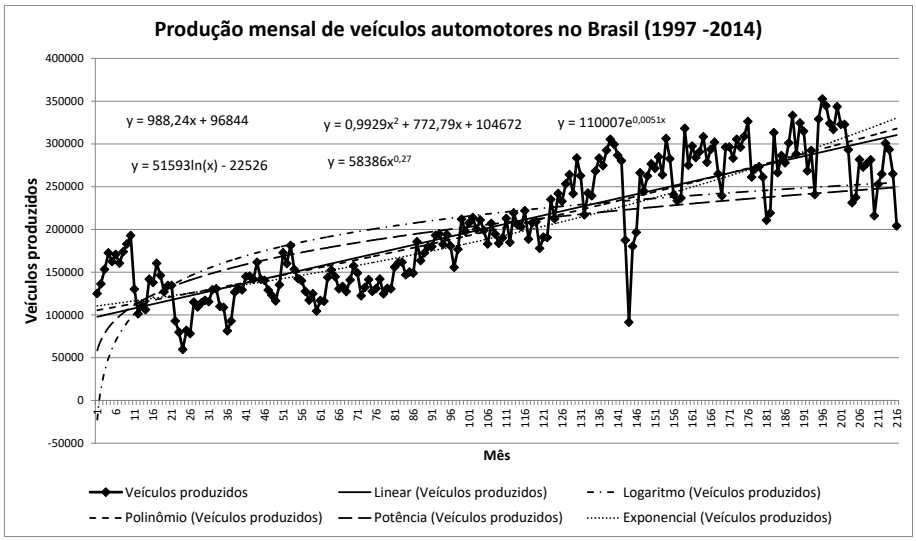


Figura 36 – Série mensal da produção de veículos automotores no Brasil de janeiro de 1997 a dezembro de 2014 com cinco modelos de tendência obtidos por mínimos quadrados

Fonte: adaptado pelo autor de Microsoft ® a partir de dados da ANFAVEA – Associação Nacional dos Fabricantes de Veículos Automotores, disponíveis em http://www.anfavea.com.br/tabelas.html, acessados em 13/11/2015.

O Microsoft Excel ® (e outros aplicativos computacionais) consegue obter as equações de vários modelos, e desenhar as respectivas curvas sobre os dados da série original. Mas como decidir qual é a melhor tendência? Para o caso das séries temporais é comum utilizar medidas de acuracidade.

Comentado [MMR21]: Os procedimentos para a construção de gráficos de séries temporais no Microsoft Excel ® e obtenção de tendências por mínimos quadrados são descritos no texto Análise de séries temporais (modelo clássico) com o Microsoft Excel ®, disponível no ambiente virtual.

Comentado [MMR22]: GLOSSÀRIO: medidas de acuracidade são estatísticas que permitem avaliar o ajuste de uma previsão aos dados originais, por meio do cálculo de médias das diferenças (erros) entre os dados originais e as previsões em cada período da série temporal. Fonte: MAKRIDAKIS, S., WHEELWRIGHT, S.C., HYNDMAN, R.J. Forecasting: methods and applications. 3rd ed.- New York: Wiley, 1998.

Dentre as várias disponíveis destacam-se três, usadas inclusive por softwares estatísticos como o Minitab ®: Erro Absoluto Médio (EAM), Erro Quadrático Médio (EQM) e Erro Percentual Absoluto Médio (EPAM). Todas se baseiam nos cálculos dos erros: as diferenças entre os valores da série e os valores preditos pelas equações de tendência para cada período t da série.

Comentado [MMR23]: Ver http://support.minitab.com/pt-br/minitab/17/topic-library/modeling-statistics/time-series/time-series-models/what-are-mape-mad-and-msd/, (em inglês) acessado em 17/11/2015.

Comentado [MMR24]: Estas mesmas medidas serão usadas posteriormente para avaliar qual modelo (aditivo ou multiplicativo) é mais apropriado para descrever o comportamento da série.

Erro absoluto médio (EAM): $$EAM = \frac{1}{n} \times \sum_{x=1}^{n} |e_x|$$

Erro quadrático médio (EQM): $$EQM = \frac{1}{n} \times \sum_{x=1}^{n} e_x^{\,2}$$

Erro percentual absoluto médio (EPAM): $$EPAM = \frac{1}{n} \times \sum_{x=1}^{n} \left| \left(\frac{e_x}{Y_x} \right) \times 100 \right|$$

Onde e_x é o erro (diferença entre o valor da série, Y_x, e o valor previsto por um modelo de tendência $\hat{Y}_x$ em um período genérico x). As duas primeiras medidas dependem da escala dos valores da série, o que dificulta a comparação com outras séries ou mesmo diferentes intervalos de tempo na mesma série. A última, EPAM, por ser relativa, não apresenta aqueles problemas[3]. Não obstante, por apresentar divisão pelos valores da série, pode ser inapropriada quando a série tiver valores iguais ou próximos a zero. A segunda medida, EQM, semelhante ao desvio padrão, dá maior ênfase a grandes erros do que EAM[4]. Pode-se usar todas, o que é fácil de implementar em uma planilha eletrônica, ou já faz parte dos programas estatísticos. O melhor modelo será o que apresentar os valores mais próximos de zero.

Comentado [MMR25]: Como o Minitab ®.

Exemplo 1 – O Quadro 24 apresenta a produção mensal de veículos no Brasil para os meses de Janeiro a Dezembro de 1997 (correspondem aos valores de x, período, de 1 e 12, respectivamente), extraídos dos dados usados nas Figuras 34 e 36, e as previsões feitas para os mesmos meses pelas equações de tendência mostradas na Figura 36.

[3] MAKRIDAKIS, S., WHEELWRIGHT, S.C., HYNDMAN, R.J. Forecasting: methods and applications . John Wiley & Sons, 3rd edition, 1998, páginas 42-44.
[4] CAMM, J. D., EVANS, J. R. Management Science and decision technology. South-Western College Publishing, 2000, página 103.

x	Prod. (Y_x) veículos	$\widehat{Y}_x =$				
		988,24x + 96844	$0,9929x^2 + 772,79x + 104672$	51593ln(x) - 22526	$58386x^{0,27}$	$110007e^{0,0051x}$
1	124889	97832,24	105445,7829	-22526	105445,8	58386
2	136323	98820,48	106221,5516	13235,54	106219,6	70402,3
3	153164	99808,72	106999,3061	34154,7	106993,3	78547,34
4	172391	100796,96	107779,0464	48997,08	107767,1	84891,64
5	162310	101785,2	108560,7725	60509,73	108540,9	90163,47
6	170685	102773,44	109344,4844	69916,25	109314,7	94712,99
7	160400	103761,68	110130,1821	77869,34	110088,5	98738,2
8	173863	104749,92	110917,8656	84758,63	110862,3	102363
9	182952	105738,16	111707,5349	90835,41	111636	105670,6
10	192829	106726,4	112499,19	96271,27	112409,8	108719,8
11	130140	107714,64	113292,8309	101188,6	113183,6	111553,9
12	101255	108702,88	114088,4576	105677,8	113957,4	114205,7

Quadro 24 – Série mensal da produção de veículos automotores no Brasil de Janeiro a Dezembro de 1997, com cinco modelos de tendência obtidos por mínimos quadrados.

Fonte: adaptado pelo autor de Microsoft ® a partir de dados da ANFAVEA – Associação Nacional dos Fabricantes de Veículos Automotores, disponíveis em http://www.anfavea.com.br/tabelas.html, acessados em 13/11/2015.

Comentado [MMR26]: Os procedimentos para realizar o cálculo das tendências pelos cinco modelos citados no Quadro 24 são mostrados no texto Análise de séries temporais (modelo clássico) com o Microsoft Excel ®, disponível no ambiente virtual.

Substituindo o valor de t nas equações mostradas no Quadro 24 é possível calcular as tendências por mínimos quadrados para todos os períodos da série. Para o período 2, por exemplo, as tendências são:

- linear: $\widehat{Y}_x = 988,24 \times 2 + 96844 = 98820,48$;

- polinômio de segundo grau: $\widehat{Y}_x = 0,9929 \times 2^2 + 772,79 \times 2 + 104672 = 106219,6$;

- logarítmico: $\widehat{Y}_x = 51593 \times \ln(2) - 22526 = 13235,54$;

- potência: $\widehat{Y}_x = 58386 \times 2^{0,27} = 106219,6$;

- exponencial: $\widehat{Y}_x = 110007 \times e^{0,0051 \times 2} = 70402,3$.

No Quadro 25 mostra-se como realizar o cálculo dos erros para a tendência linear para os primeiros doze meses da série da Figura 36 (Janeiro a Dezembro de 1997).

x	Prod. (Y_x) veículos	Equação da tendência $\hat{Y}_x = 988{,}24x + 96844$	Erro $e_x = Y_x - \hat{Y}_x$	Módulo do erro $\lvert e_x \rvert$	Erro quadrático e_x^2	Erro percentual $\lvert (e_x / Y_x) \times 100 \rvert$
1	124889	97832,24	27056,76	27056,76	732068261,7	21,66
2	136323	98820,48	37502,52	37502,52	1406439006	27,51
3	153164	99808,72	53355,28	53355,28	2846785904	34,84
4	172391	100796,96	71594,04	71594,04	5125706564	41,53
5	162310	101785,2	60524,8	60524,8	3663251415	37,29
6	170685	102773,44	67911,56	67911,56	4611979982	39,79
7	160400	103761,68	56638,32	56638,32	3207899292	35,31
8	173863	104749,92	69113,08	69113,08	4776617827	39,75
9	182952	105738,16	77213,84	77213,84	5961977088	42,20
10	192829	106726,4	86102,6	86102,6	7413657727	44,65
11	130140	107714,64	22425,36	22425,36	502896771,1	17,23
12	101255	108702,88	-7447,88	7447,88	55470916,49	7,36

Quadro 25 – Série mensal da produção de veículos automotores no Brasil de Janeiro a Dezembro de 1997, com tendência linear obtida por mínimos quadrados, erros, módulos de erro, erros quadráticos e erros percentuais.

Fonte: adaptado pelo autor de Microsoft ® a partir de dados da ANFAVEA – Associação Nacional dos Fabricantes de Veículos Automotores, disponíveis em http://www.anfavea.com.br/tabelas.html, acessados em 13/11/2015.

Comentado [MMR27]: Os procedimentos para realizar o cálculo dos erros citados no Quadro 25, e das medidas de acuracidade do Quadro 26 são mostrados no texto Análise de séries temporais (modelo clássico) com o Microsoft Excel ®, disponível no ambiente virtual.

Realizando o mesmo procedimento para as outras equações de tendência, para todos os períodos da série mostrada na Figura 36, podem-se obter as medidas de acuracidade de cada modelo, conforme o Quadro 26.

Medida	Modelo				
	Linear	Polinômio de 2º grau	Logarítmico	Potência	Exponencial
EAM	27928,64	**27752,31**	42944,76	39481,05	28195,36
EQM	1247075874	**1235156012**	2618630743	2222579666	1306864744
EPAM	15,83	15,63	25,37	22,01	**15,35**

Quadro 26 – Medidas de acuracidade dos modelos de tendências por mínimos quadrados da produção de veículos automotores no Brasil de Janeiro a Dezembro de 1997.

Fonte: adaptado pelo autor de Microsoft ® a partir de dados da ANFAVEA – Associação Nacional dos Fabricantes de Veículos Automotores, disponíveis em http://www.anfavea.com.br/tabelas.html, acessados em 13/11/2015.

No Quadro 26 os menores valores das medidas de acuracidade são mostrados em negrito. A tendência por polinômio de segundo grau tem os menores valores de EAM e

EQM, mas a tendência por exponencial tem o menor EPAM. Por maioria, escolhe-se o polinômio de segundo grau como o melhor modelo para representar a tendência da série por mínimos quadrados. Podemos usar este modelo para fazer a previsão da tendência da série nos doze meses de 2015, que seriam os períodos 217 a 228 da série, o que é mostrado no Quadro 27.

Mês	Período	Previsão tendência (polinômio de 2º grau) (veículos)	
Janeiro 2015	217	$\hat{Y}_x = 0{,}9929\times217^2 + 772{,}79\times217 + 104672 =$	319122,0981
Fevereiro 2015	218	$\hat{Y}_x = 0{,}9929\times218^2 + 772{,}79\times218 + 104672 =$	320326,7996
Março 2015	219	$\hat{Y}_x = 0{,}9929\times219^2 + 772{,}79\times219 + 104672 =$	321533,4869
Abril 2015	220	$\hat{Y}_x = 0{,}9929\times220^2 + 772{,}79\times220 + 104672 =$	322742,16
Maio 2015	221	$\hat{Y}_x = 0{,}9929\times221^2 + 772{,}79\times221 + 104672 =$	323952,8189
Junho 2015	222	$\hat{Y}_x = 0{,}9929\times222^2 + 772{,}79\times222 + 104672 =$	325165,4636
Julho 2015	223	$\hat{Y}_x = 0{,}9929\times223^2 + 772{,}79\times223 + 104672 =$	326380,0941
Agosto 2015	224	$\hat{Y}_x = 0{,}9929\times224^2 + 772{,}79\times224 + 104672 =$	327596,7104
Setembro 2015	225	$\hat{Y}_x = 0{,}9929\times225^2 + 772{,}79\times225 + 104672 =$	328815,3125
Outubro 2015	226	$\hat{Y}_x = 0{,}9929\times226^2 + 772{,}79\times226 + 104672 =$	330035,9004
Novembro 2015	227	$\hat{Y}_x = 0{,}9929\times227^2 + 772{,}79\times227 + 104672 =$	331258,4741
Dezembro 2015	228	$\hat{Y}_x = 0{,}9929\times228^2 + 772{,}79\times228 + 104672 =$	332483,0336

Quadro 27 – Previsão da tendência da série de produção de veículos automotores no Brasil de Janeiro a Dezembro de 2015, através de modelo de polinômio de segundo grau obtido por mínimos quadrados.

Fonte: adaptado pelo autor de Microsoft ® a partir de dados da ANFAVEA – Associação Nacional dos Fabricantes de Veículos Automotores, disponíveis em http://www.anfavea.com.br/tabelas.html, acessados em 13/11/2015.

5.2.2 – Obtenção de tendência por médias móveis

As médias móveis são uma forma alternativa de obtenção da tendência ou nível de uma série temporal. Calcula-se a média dos primeiros k períodos da série, colocando o resultado no período exatamente no centro deles. Progressivamente, vamos acrescentando

Comentado [MMR28]: GLOSSÀRIO. Média móvel é um procedimento em que se calcula a média de um certo número de observações, e à medida que uma nova observação torna-se disponível uma nova média é calculada incorporando a nova observação e descartando a mais antiga da série. Fonte: MAKRIDAKIS, S., WHEELWRIGHT, S.C., HYNDMAN, R.J. Forecasting: methods and applications. 3rd ed.- New York: Wiley, 1998.

um período seguinte e desprezando o primeiro da média imediatamente anterior, e calculando novas médias, que vão se movendo até o fim da série. O número de períodos (**k**) é chamado de ordem da série. O processo permite "alisar" a série, seu resultado é menos influenciado pelas variações irregulares, e, se a série for registrada com periodicidade inferior a 1 ano, o efeito das variações sazonais também é removido.

Exemplo 2 - Os dados no Quadro 28 representam as vendas anuais das fábricas (em milhões de unidades), em todo o mundo, de carros, caminhões e ônibus fabricados pela General Motors Corporation (GM) de 1970 a 1992. Obtenha a tendência da série por médias móveis de 3, 5 e 7 períodos, e plote-as em um gráfico junto com os dados originais.

Ano	Vendas	Ano	Vendas	Ano	Vendas
1970	5,3	1978	9,5	1986	8,6
1971	7,8	1979	9,0	1987	7,8
1972	7,8	1980	7,1	1988	8,1
1973	8,7	1981	6,8	1989	7,9
1974	6,7	1982	6,2	1990	7,5
1975	6,6	1983	7,8	1991	7,0
1976	8,6	1984	8,3	1992	7,2
1977	9,1	1985	9,3		

Quadro 28 – Vendas mundiais da General Motors Corporation, em milhões de veículos, de 1970 a 1992.

Fonte: adaptado pelo autor de : LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Comentado [MMR29]: Estes dados estão no arquivo Séries Temporais, disponível no ambiente virtual

Primeiramente vamos apresentar um gráfico da série original (Figura 37), para observar se não seria possível ajustar algum dos modelos anteriores (seção 5.2.1) como tendência da série.

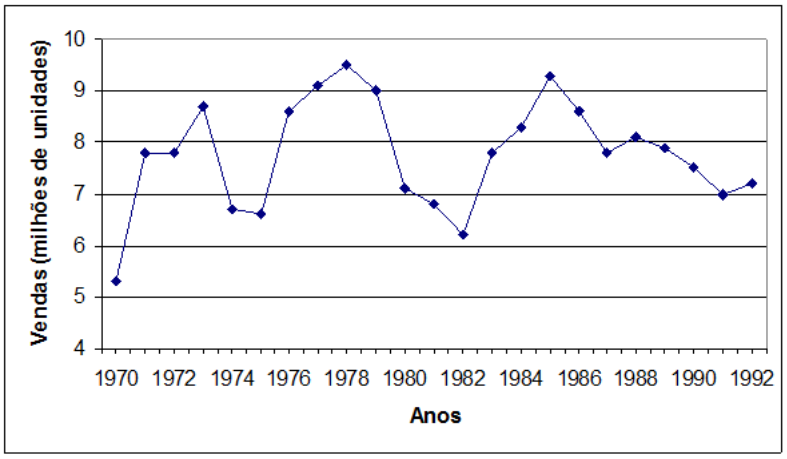


Figura 37 – Vendas mundiais da General Motors Corporation, em milhões de veículos, de 1970 a 1992

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Observando a Figura 37 não parece haver um comportamento crescente, ou decrescente, no longo prazo. Poderia se afirmar que a série não tem tendência, mas sim que está estacionária, e que não seria apropriado ajustar alguma das equações de tendência vistas na seção 5.2.1 aos dados. Não obstante, há interesse em obter o nível da série, em que patamar as vendas estão.

Comentado [MMR30]: GLOSSÁRIO Série temporal estacionária é uma série cujos valores flutuam em torno de uma média constante. Fonte: MAKRIDAKIS, S., WHEELWRIGHT, S.C., HYNDMAN, R.J. Forecasting: methods and applications. 3rd ed.- New York: Wiley, 1998.

Vamos aplicar médias móveis de 3, 5 e 7 períodos e observar os resultados.

Médias Móveis de 3 períodos (ordem 3)

Devemos juntar os períodos de 3 em 3, sempre acrescentando o próximo e desprezando o primeiro do grupo anterior, colocando o resultado no período central (2º período):

1970 - 1971 - 1972 com resultado em 1971; 1971 - 1972 - 1973 com resultado em 1972;

1972 - 1973 - 1974 com resultado em 1973; e assim por diante, até chegar a

1990 - 1991 - 1992 com resultado em 1991.

O Quadro 29 apresenta os resultados:

Ano	Vendas (Y) - em milhões	Total Móvel 3 períodos	Média Móvel 3 períodos
1970	5,3	-	-
1971	7,8	20,9	6,97
1972	7,8	24,3	8,10
1973	8,7	23,2	7,73
1974	6,7	22	7,33
1975	6,6	21,9	7,30
1976	8,6	24,3	8,10
1977	9,1	27,2	9,07
1978	9,5	27,6	9,20
1979	9	25,6	8,53
1980	7,1	22,9	7,63
1981	6,8	20,1	6,70
1982	6,2	20,8	6,93
1983	7,8	22,3	7,43
1984	8,3	25,4	8,47
1985	9,3	26,2	8,73
1986	8,6	25,7	8,57
1987	7,8	24,5	8,17
1988	8,1	23,8	7,93
1989	7,9	23,5	7,83
1990	7,5	22,4	7,47
1991	7	21,7	7,23
1992	7,2	-	-

Quadro 29 – Médias móveis de três períodos (ordem 3) das vendas mundiais da General Motors Corporation, em milhões de veículos, de 1970 a 1992.

Fonte: adaptado pelo autor de: LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Comentado [MMR31]: Os procedimentos para realizar o cálculo das médias móveis dos Quadro 29, 30 e 31são mostrados no texto Análise de séries temporais (modelo clássico) com o Microsoft Excel ®, disponível no ambiente virtual.

Observe que ao calcularmos médias móveis alguns períodos ficam sem tendência, porque os resultados das médias são postos no centro dos períodos.

Média móvel de 5 períodos (ordem 5)

Devemos juntar os períodos de 5 em 5, sempre acrescentando o próximo e desprezando o primeiro do grupo anterior, colocando o resultado no período central (3º período):

1970 - 1971 - 1972 - 1973 - 1974 com resultado em 1972;

1971 - 1972 - 1973 - 1974 - 1975 com resultado em 1973;

1972 – 1973 – 1974 – 1975 – 1976 com resultado em 1974; e assim por diante, até chegar a

1988 - 1989 - 1990 - 1991 - 1992 com resultado em 1990.

O Quadro 30 apresenta os resultados.

Ano	Vendas (Y) - em milhões	Total Móvel 5 períodos	Média Móvel 5 períodos
1970	5,3	-	-
1971	7,8	-	-
1972	7,8	36,3	7,26
1973	8,7	37,6	7,52
1974	6,7	38,4	7,68
1975	6,6	39,7	7,94
1976	8,6	40,5	8,1
1977	9,1	42,8	8,56
1978	9,5	43,3	8,66
1979	9	41,5	8,3
1980	7,1	38,6	7,72
1981	6,8	36,9	7,38
1982	6,2	36,2	7,24
1983	7,8	38,4	7,68
1984	8,3	40,2	8,04
1985	9,3	41,8	8,36
1986	8,6	42,1	8,42
1987	7,8	41,7	8,34
1988	8,1	39,9	7,98
1989	7,9	38,3	7,66
1990	7,5	37,7	7,54
1991	7	-	-
1992	7,2	-	-

Quadro 30 – Médias móveis de cinco períodos (ordem 5) das vendas mundiais da General Motors Corporation, em milhões de veículos, de 1970 a 1992.

Fonte: adaptado pelo autor de : LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Novamente, alguns períodos ficam sem tendência, porque os resultados das médias são postos no centro dos períodos. Aqui, como as médias agrupam 5 períodos, dois ficam sem tendência no início e dois ao final da série.

Média móvel de 7 períodos

Devemos juntar os períodos de 7 em 7, sempre acrescentando o próximo e desprezando o primeiro do grupo anterior, colocando o resultado no período central (5o período):

1970 - 1971 - 1972 - 1973 - 1974 - 1975 - 1976 com resultado em 1973;

1971 - 1972 - 1973 - 1974 - 1975 - 1976 - 1977 com resultado em 1974;

1972 - 1973 - 1974 - 1975 - 1976 - 1977 - 1978 com resultado em 1975;

e assim por diante, até chegar a

1986 - 1987 - 1988 - 1989 - 1990 - 1991 - 1992 com resultado em 1989.

O Quadro 31 apresenta os resultados.

Ano	Vendas (Y) - em milhões	Total Móvel 5 períodos	Média Móvel 5 períodos
1970	5,3	-	-
1971	7,8	-	-
1972	7,8	-	-
1973	8,7	51,5	7,36
1974	6,7	55,3	7,90
1975	6,6	57	8,14
1976	8,6	58,2	8,31
1977	9,1	56,6	8,09
1978	9,5	56,7	8,10
1979	9	56,3	8,04
1980	7,1	55,5	7,93
1981	6,8	54,7	7,81
1982	6,2	54,5	7,79
1983	7,8	54,1	7,73
1984	8,3	54,8	7,83
1985	9,3	56,1	8,01
1986	8,6	57,8	8,26
1987	7,8	57,5	8,21
1988	8,1	56,2	8,03
1989	7,9	54,1	7,73
1990	7,5	-	-
1991	7	-	-
1992	7,2	-	-

Quadro 31 – Médias móveis de sete períodos (ordem 7) das vendas mundiais da General Motors Corporation, em milhões de veículos, de 1970 a 1992.

Fonte: adaptado pelo autor de : LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

No Quadro 31, como as médias agrupam 7 períodos, três ficam sem tendência no início e três ao final da série.

Comentado [MMR32]: Este é um dos inconvenientes das médias móveis, nunca serão obtidas estimativas para todos os períodos da série.

A Figura 38 apresenta o gráfico da série original com as médias móveis de 3, 5 e 7 períodos.

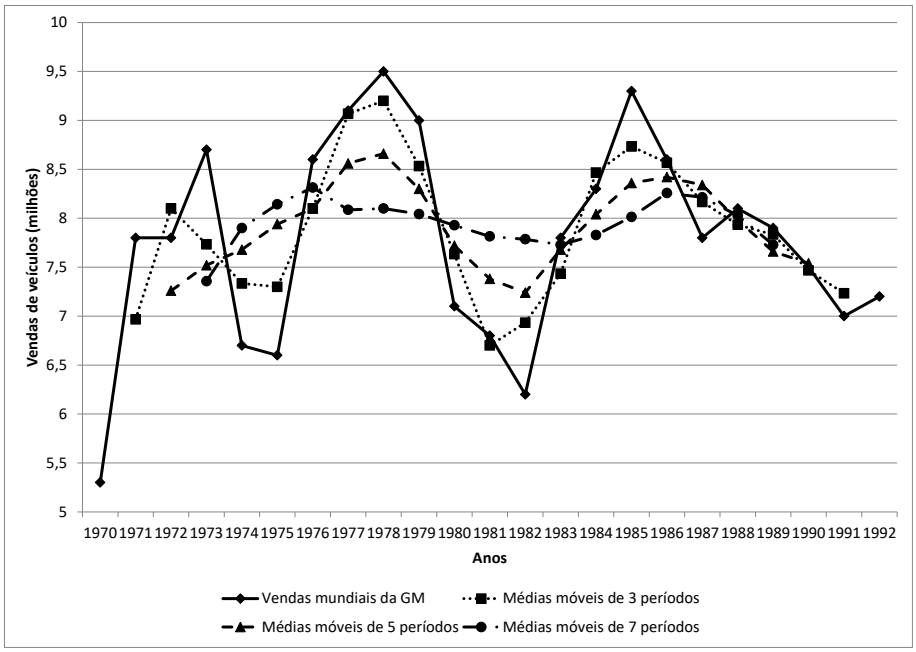


Figura 38 – Vendas mundiais da General Motors Corporation, em milhões de veículos, de 1970 a 1992: série original e médias móveis de 3, 5 e 7 períodos

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Quanto maior o número de períodos da série agrupados pela média móvel mais "alisada" fica a linha de tendência (média móvel de 7 períodos): esta representa melhor o comportamento de longo prazo, indicando uma ligeira oscilação em torno de 8 milhões de unidades vendidas (este é o nível da série). E quanto menor o número de períodos, mais a tendência acompanhará o comportamento dos dados originais (média móvel de 3 períodos). Por este motivo, quando uma série apresenta muitas irregularidades é comum "alisá-la" através de médias móveis.

Mas o que aconteceria se o número de períodos fosse par? Se possível, devemos escolher um número ímpar de períodos, para que o resultado seja colocado em um período

central que tem correspondente na série temporal. Contudo, se a série temporal for registrada trimestralmente, e queremos obter a sua tendência por médias móveis, devemos utilizar médias móveis de 4 períodos (porque há 4 trimestres no ano), para que possamos obter a tendência sem influência da sazonalidade. Se a série for registrada mensalmente, devemos utilizar médias móveis de 12 períodos. Nestes dois casos os períodos "centrais" (que começariam em 2,5º e 6,5º respectivamente) não têm correspondente na série original, o que tornará impossível remover a tendência da série para observar outras componentes. As médias móveis precisam ser centralizadas: calculam-se novas médias móveis, a partir das calculadas com 4 ou 12 períodos, mas agora de 2 períodos, colocando seus resultados em períodos que têm correspondentes na série.

Comentado [MMR33]: GLOSSÁRIO. Médias móveis centralizadas são médias móveis calculadas a partir de outras duas médias móveis que tenham um número PAR de períodos, para que o resultado das centralizadas seja localizado em um período que tenha correspondência aos da série temporal original. Fonte: MAKRIDAKIS, S., WHEELWRIGHT, S.C., HYNDMAN, R.J. Forecasting: methods and applications. 3rd ed.- New York: Wiley, 1998.

Exemplo 3 – Os dados do Quadro 32 mostram as vendas trimestrais em milhões de dólares da loja de departamentos JC Penney de 1996 a 2001.

Trimestre	Vendas (US$ milhões)	Trimestre	Vendas (US$ milhões)
1996-I	4452	1999-I	7339
1996-II	4507	1999-II	7104
1996-III	5537	1999-III	7639
1996-IV	8157	1999-IV	9661
1997-I	6481	2000-I	7528
1997-II	6420	2000-II	7207
1997-III	7208	2000-III	7538
1997-IV	9509	2000-IV	9573
1998-I	6755	2001-I	7522
1998-II	6483	2001-II	7211
1998-III	7129	2001-III	7729
1998-IV	9072	2001-IV	9542

Quadro 32 – Vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001 (em milhões de dólares).

Fonte: 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.

A Figura 39 mostra o gráfico de linhas para os dados do Quadro 32.

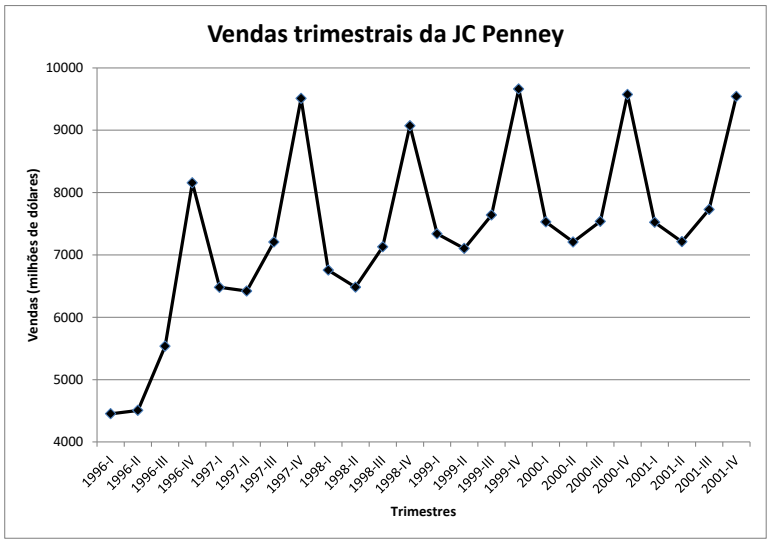


Figura 39 – Vendas trimestrais da loja de departamentos JC Penney, em milhões de dólares, de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Na Figura 39 podemos perceber a existência de picos e vales, o que indica que deve existir influência de sazonalidade na série. Talvez o leitor também perceba que embora haja um crescimento nas vendas de 1996 a 1998, daí em diante ocorre uma estabilização. Poderia ser ajustado um dos modelos usados na seção 5.2.1 para obter a tendência da série, mas também podemos usar uma média móvel.

Como a série é registrada trimestralmente, e a tendência deve ser obtida por médias móveis, é preciso calcular médias móveis de 4 períodos, pois há 4 trimestres no ano. Contudo, como este número de períodos é par, médias móveis de 2 períodos, calculadas a partir daquelas de 4 períodos, precisam ser obtidas para obter resultados centrados.

No Quadro 33 são apresentados os cálculos necessários.

Trimestre	Vendas (US$ milhões)	Médias móveis (4 períodos)	Médias móveis (2 períodos centradas)
1996-I	4452		
1996-II	4507		
		5663,25	
1996-III	5537		5916,875
		6170,5	
1996-IV	8157		6409,625
		6648,75	
1997-I	6481		6857,625
		7066,5	
1997-II	6420		7235,5
		7404,5	
1997-III	7208		7438,75
		7473	
1997-IV	9509		7480,875
		7488,75	
1998-I	6755		7478,875
		7469	
1998-II	6483		7414,375
		7359,75	
1998-III	7129		7432,75
		7505,75	
1998-IV	9072		7583,375
		7661	
1999-I	7339		7724,75
		7788,5	
1999-II	7104		7862,125
		7935,75	
1999-III	7639		7959,375
		7983	
1999-IV	9661		7995,875
		8008,75	
2000-I	7528		7996,125
		7983,5	
2000-II	7207		7972,5
		7961,5	
2000-III	7538		7960,75
		7960	
2000-IV	9573		7960,5
		7961	
2001-I	7522		7984,875
		8008,75	
2001-II	7211		8004,875
		8001	
2001-III	7729		
2001-IV	9542		

Quadro 33 – Vendas mensais da loja de departamentos JC Penney (em milhões de dólares) e as médias móveis de 4 períodos e médias móveis centradas de 2 períodos.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

As linhas mais escuras no Quadro 33 indicam os períodos "centrais" das médias móveis de ordem 4, que não têm correspondente na série original. Calculam-se as médias móveis de 4 períodos:

- os primeiros 4 períodos são os 4 trimestres de 1996: 1996 I, 1996 II, 1996 III, 1996 IV; a média móvel deles (22653/4 ≅ 5663,25) deve ficar no centro destes períodos, ou seja entre 1996 II e 1996 III, que é um período inexistente na série original;
- em seguida desprezamos 1996 I e incluímos 1997 I: 1996 II, 1996 III, 1996 IV, 1997 I; a média móvel (24682/4 ≅ 6170,5) deve ficar entre 1996 III e 1996 IV, novamente inexistente na série original;
- prosseguimos até os quatro últimos períodos: 2001 I, 2001 II, 2001 III, 2001 IV; a média móvel (32004/4 = 8001) deve ficar entre 2001 II e 2001 III.

Agora precisamos obter as médias móveis centradas, juntando 2 médias móveis de 4 períodos calculadas anteriormente:

- a média móvel de 4 períodos que está entre 1996 II e 1996 III, com a que está entre 1996 III e 1996 IV, cujo resultado ((5663,25+6170,5)/2 ≅ 5916,875) deverá ficar em 1996 III (passando a ter correspondente na série original);
- a média móvel de 4 períodos que está entre 1996 III e 1996 IV, com a que está entre 1996 IV e 1997 I, cujo resultado ((6170,5+6648,75)/2 ≅ 6409,625) deverá ficar em 1996 IV (passando a ter correspondente na série original);
- prosseguimos até as últimos duas médias móveis de 4 períodos: entre 2001 I e 2001 II, e entre 2001 II e 2001 III, cujo resultado ((8008,75+8001)/2 ≅ 8004,875) deverá ficar em 2001 II.

Comentado [MMR34]: Todo o procedimento para obtenção de médias móveis e médias móveis centradas utilizando o Microsoft Excel ® é mostrado no arquivo Análise de Séries Temporais (modelo clássico) com o Microsoft Excel ®, disponível no ambiente virtual.

Repare que faltam médias móveis para exatamente 2 períodos no início da série e para exatamente 2 no final, porque as médias móveis iniciais envolvem 4 períodos (porque há 4 trimestres no ano). Se a série fosse mensal faltariam 6 períodos no início e 6 no final.

Vamos ver como ficam a série original e a tendência em um gráfico, mostrado na Figura 40.

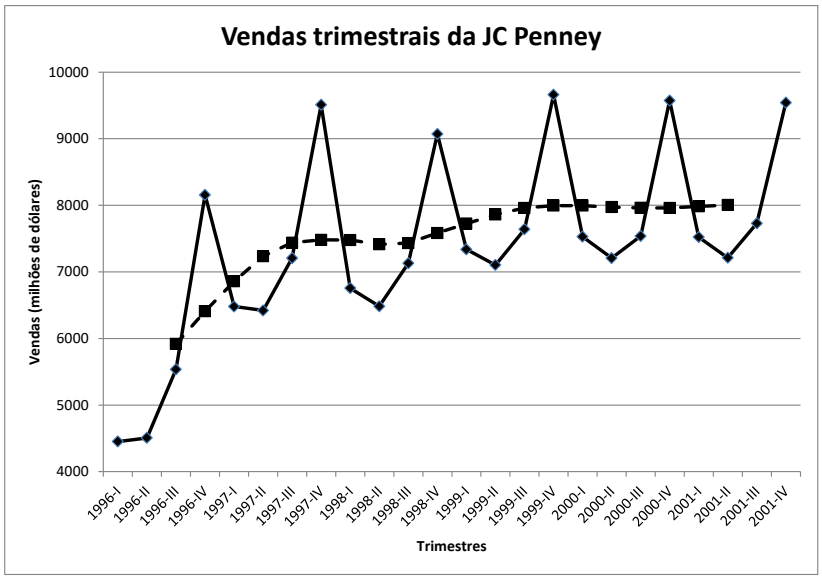


Figura 40 – Vendas trimestrais da loja de departamentos JC Penney, em milhões de dólares e médias móveis centradas, de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Observe que o efeito da sazonalidade não aparece nas médias móveis, e que realmente o valor das vendas realmente está se estabilizando em torno de 8000 milhões (8 bilhões) de dólares nos últimos trimestres da série.

Se a série fosse registrada mensalmente, como a do Exemplo 1, as médias deveriam ter 12 períodos (a mesma ordem da sazonalidade), e depois também precisariam ser centradas, o que iria acarretar que os seis meses iniciais e os seis finais da série ficariam sem médias móveis calculadas.

5.2.3 – Remoção da Tendência

Uma vez identificada a tendência, seja por equações ou por médias móveis, ela pode ser removida da série, para facilitar a visualização das outras componentes:

$Y - T = C + S + I$ para um modelo aditivo $\frac{Y}{T} = C \times S \times I$ para um modelo multiplicativo

Vejamos como ficaria a série mostrada na Figura 36 com a remoção da tendência obtida por médias, pelos modelos aditivo e multiplicativo (ambas supondo uma tendência por polinômio de 2º grau): nas Figuras 41 e 42.

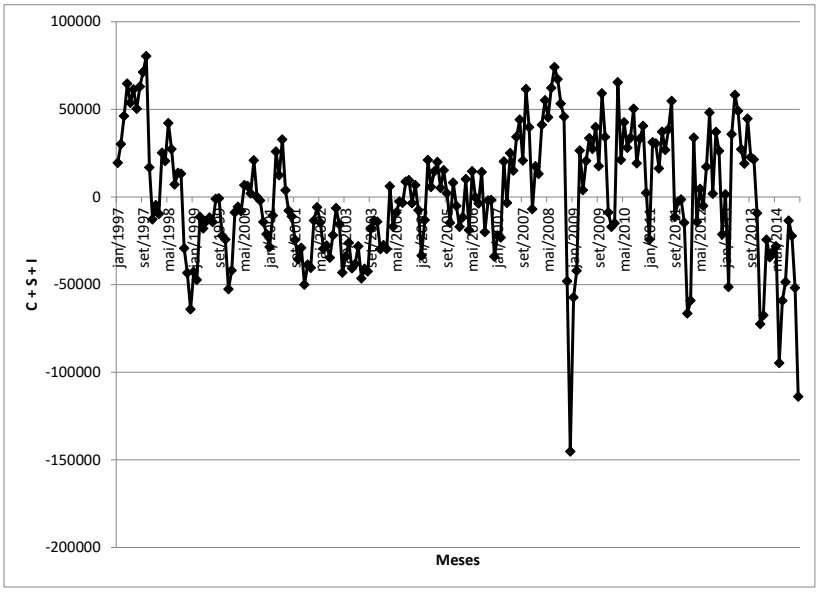


Figura 41 – Série mensal da produção de veículos automotores no Brasil com tendência por polinômio de 2º grau removida – modelo aditivo

Fonte: adaptado pelo autor de Microsoft ® a partir de dados da ANFAVEA – Associação Nacional dos Fabricantes de Veículos Automotores, disponíveis em http://www.anfavea.com.br/tabelas.html, acessados em 13/11/2015

Na Figura 41 os valores oscilam em torno de zero: se maiores do que zero indicam componentes (C + S + I) que aumentam a tendência, se menores que diminuem.

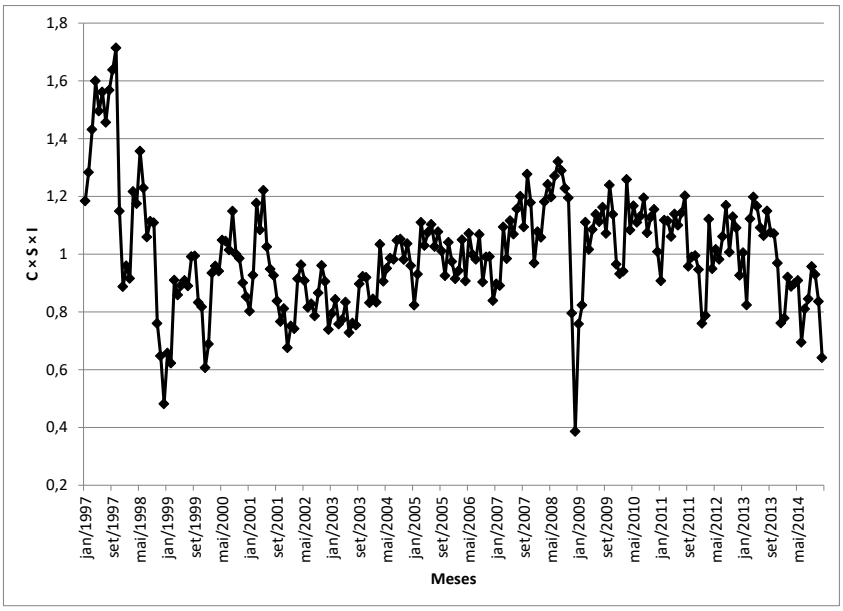


Figura 42 – Série mensal da produção de veículos automotores no Brasil com tendência por polinômio de 2º grau removida – modelo multiplicativo

Fonte: adaptado pelo autor de Microsoft ® a partir de dados da ANFAVEA – Associação Nacional dos Fabricantes de Veículos Automotores, disponíveis em http://www.anfavea.com.br/tabelas.html, acessados em 13/11/2015

Na Figura 42 os valores oscilam em torno de 1: a tendência foi removida, restaram apenas as componentes cíclicas, sazonais e irregulares (C × S × I) que modificam a tendência em um modelo multiplicativo.

5.3 – Obtenção da componente sazonal de uma série temporal

Conforme visto na seção 5.1 a componente sazonal é uma oscilação de curto prazo, que ocorre sempre dentro do ano, e que se repete sistematicamente ano após ano. Obviamente uma série temporal registrada anualmente (ou seja, os valores dos dias, meses, trimestres, são resumidos em um valor anual) não tem componente sazonal.

Nos modelos aditivo e multiplicativo as componentes sazonais são representadas pelos índices sazonais, ou fatores sazonais, um para cada período em que o ano é dividido (se a série é registrada mensalmente há 12 índices, se trimestralmente há 4 índices, etc.). Os índices sazonais modificam a tendência, ao serem somados (modelo aditivo) ou multiplicados por ela:

- no modelo aditivo, se todos os índices forem próximos ou exatamente iguais a zero então as componentes sazonais parecem não exercer grande efeito sobre a série; se os índices forem substancialmente diferentes de zero, tanto positivos como negativos, o valor da tendência será modificado por eles, indicando influência das componentes sazonais na série.
- no modelo multiplicativo, se todos os índices sazonais forem aproximadamente iguais a 1 então as componentes sazonais parecem não exercer grande efeito sobre a série; se os índices forem substancialmente diferentes de 1, pelo menos 5% acima ou abaixo em alguns dos meses ou trimestres, o valor da tendência será modificado por eles, indicando que as componentes sazonais afetam a série.

Quando se usa o modelo aditivo, a soma de todos os índices sazonais precisa ser igual, ou muito próxima, de zero. Quando se usa o modelo multiplicativo a soma precisa ser igual ao período da sazonalidade: se a série é trimestral deve ser igual a 4 (4 trimestres no ano), se é mensal deve ser igual a 12, e assim por diante. Em alguns casos é preciso fazer pequenas correções para garantir tal comportamento.

Para obter os índices sazonais recomenda-se que a série temporal tenha, no mínimo, quatro anos completos (16 trimestres), se trimestral, ou cinco anos completos (60 meses), se mensal. Caso contrário, será mais difícil confirmar a existência da regularidade inerente à componente sazonal (alguns programas estatísticos simplesmente não apresentam os resultados para séries menores).

Comentado [MMR35]: Como o Statistica ® da Statsoft.

Há vários métodos para a obtenção dos índices sazonais, entre eles o método da razão para a média móvel (ou método da média móvel percentual). Ele consiste em:

1) obter médias móveis de ordem igual ao número de períodos sazonais (4 se a série é trimestral, 12 se é mensal);
2) obter médias móveis de 2 períodos, centradas, a partir calculadas no passo 1;
3) obter os índices sazonais para cada período:
- no modelo ADITIVO, subtraindo dos valores originais da série as médias móveis centradas calculadas no passo 2;
- no modelo MULTIPLICATIVO, dividindo os valores originais da série pelas médias móveis centradas calculadas no passo 2;
4) obter medidas de síntese dos índices calculados no passo 3, que representarão cada período sazonal.
- no modelo ADITIVO, calcular a média aritmética simples dos valores correspondentes ao período sazonal (média dos índices obtidos em todos os janeiros da série, por exemplo);
- no modelo MULTIPLICATIVO, calcular a média aritmética simples dos valores correspondentes ao período sazonal, sem incluir os valores máximo e mínimo – também chamada de média interna.
5) Somar todos os índices calculados no passo 4.
- no modelo ADITIVO verificar se a soma é igual a zero.
- no modelo MULTIPLICATIVO verificar se a soma é igual à ordem da sazonalidade (4 se a série for trimestral, 12 se for mensal, etc.).
6) Fazer as correções necessárias para que a soma dos índices seja coerente (igual a zero para o aditivo e igual à ordem da sazonalidade no multiplicativo):
- no modelo ADITIVO, somar todos os índices calculados no passo 4 e dividir a soma pela ordem da sazonalidade (4 se trimestral, 12 se mensal, etc.); o resultado (Fator) deverá ser subtraído de cada uma das médias dos índices, garantindo que a soma deles seja igual a zero.
- no modelo MULTIPLICATIVO, somar todos os índices calculados no passo 4, subtrair da soma a ordem da sazonalidade (4 se trimestral, 12 se mensal, etc.), e dividir a subtração pela ordem da sazonalidade (novamente, 4 se trimestral, 12 se mensal, etc.), obtendo o Excesso; subtrair o Excesso de 1; o resultado (Fator) deverá ser multiplicado por cada uma das médias internas dos índices, garantindo que a soma deles seja igual à ordem da sazonalidade.

Os passos 1 e 2 são virtualmente idênticos ao procedimento para obtenção de tendência por médias móveis visto na seção 5.2.2 (quando o número de períodos é par). Para ajudar a compreender o conteúdo os procedimentos acima são expostos nos fluxogramas mostrados nas Figuras 43 e 44.

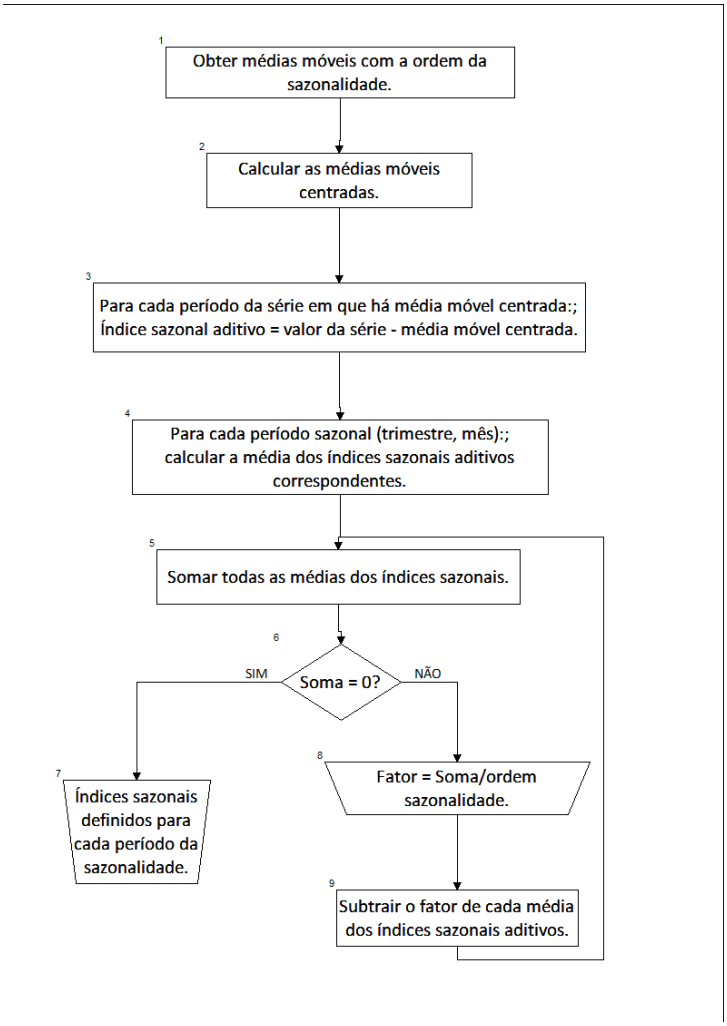


Figura 43 – Fluxograma de obtenção da componente sazonal – modelo aditivo

Fonte: elaborado pelo autor.

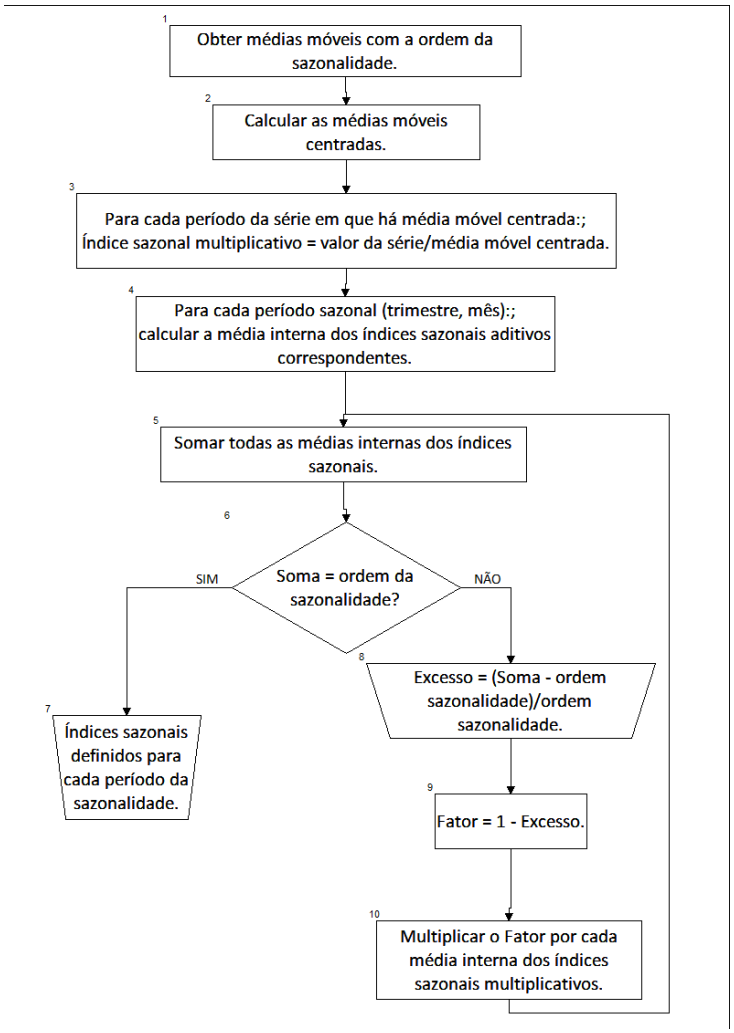


Figura 44 – Fluxograma de obtenção da componente sazonal – modelo multiplicativo

Fonte: elaborado pelo autor.

Exemplo 4 - Obtenha a componente sazonal, tanto pelo modelo aditivo quanto pelo multiplicativo, para a série vendas da loja JC Penney apresentada no Exemplo 3. Interprete os resultados encontrados. Podemos usar os dados do Quadro 33 para obter os índices sazonais aditivos e multiplicativos, os resultados estão no Quadro 34.

Trimestre	Vendas (US$ milhões)	Médias móveis (2 períodos centradas)	Índices sazonais aditivos	Índices sazonais multiplicativos
1996-I	4452			
1996-II	4507			
1996-III	5537	5916,875	-379,875	0,936
1996-IV	8157	6409,625	1747,375	1,273
1997-I	6481	6857,625	-376,625	0,945
1997-II	6420	7235,5	-815,5	0,887
1997-III	7208	7438,75	-230,75	0,969
1997-IV	9509	7480,875	2028,125	1,271
1998-I	6755	7478,875	-723,875	0,903
1998-II	6483	7414,375	-931,375	0,874
1998-III	7129	7432,75	-303,75	0,959
1998-IV	9072	7583,375	1488,625	1,196
1999-I	7339	7724,75	-385,75	0,950
1999-II	7104	7862,125	-758,125	0,904
1999-III	7639	7959,375	-320,375	0,960
1999-IV	9661	7995,875	1665,125	1,208
2000-I	7528	7996,125	-468,125	0,941
2000-II	7207	7972,5	-765,5	0,904
2000-III	7538	7960,75	-422,75	0,947
2000-IV	9573	7960,5	1612,5	1,203
2001-I	7522	7984,875	-462,875	0,942
2001-II	7211	8004,875	-793,875	0,901
2001-III	7729			
2001-IV	9542			

Quadro 34 – Vendas mensais da loja de departamentos JC Penney (em milhões de dólares) e as médias móveis de 4 períodos e médias móveis centradas de 2 períodos.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

As médias móveis centradas já haviam sido calculadas no Exemplo 3.

Para obter os índices sazonais **aditivos** devemos subtrair dos valores originais da série as médias móveis centradas, a partir de 1996 III até 2001 II, cujos resultados estão na penúltima coluna do Quadro 33. Os índices para cada trimestre serão:

Trimestre I =>	-376,625	-723,875	-385,75	-468,125	-462,875
Trimestre II =>	-815,5	-931,375	-758,125	-765,5	-793,875
Trimestre III=>	-379,875	-230,75	-303,75	-320,375	-422,75
Trimestre IV=>	1747,375	2028,125	1488,625	1665,125	1612,5

Como é um modelo aditivo precisamos calcular a média de cada trimestre. Então as médias dos índices sazonais aditivos (em milhões de dólares) serão:

Trimestre I = -483,45 Trimestre II = -812,875

Trimestre III = -331,5 Trimestre IV = 1708,35

Observe que há uma diferença considerável entre as médias. Nos três primeiros trimestres as vendas caem bastante em relação à média trimestral (até 812 milhões de dólares...), para subir muito no último (1708 milhões de dólares em relação à média trimestral). Estas oscilações são grandes demais para ter ocorrido por acaso, há influência da sazonalidade na série de vendas. Somando os índices vamos obter 80,525, indicando que é preciso realizar uma correção. Como a sazonalidade tem ordem 4, divide-se a soma por 4 obtendo 20,131. Subtraindo de cada índice este valor:

Trimestre I = -483,45 – (20,131) = -503,581

Trimestre II = -812,875 – (20,131) = -833,006

Trimestre III = -331,5 – (20,131) = -351,631

Trimestre IV = 1708,35 – (20,131) = 1688,219

E a soma dos quatro índices é virtualmente igual a zero. E a interpretação da componente sazonal pelo modelo aditivo:

- há influência de sazonalidade na série de vendas da JC Penney pelo modelo aditivo, pois os valores das médias dos índices sazonais aditivos para cada trimestre afastam-se significativamente de zero;
- nos três primeiros trimestres as médias caem substancialmente em relação à média trimestral de vendas, são, então, períodos de *baixa* nas vendas, sendo o Trimestre II o de

baixa mais acentuada (pois a média corrigida dos índices aditivos para este trimestre vale -833,006, indicando uma queda nas vendas de 833 milhões de dólares em relação à média trimestral);

- no quarto trimestre (Trimestre IV) as vendas aumentam muito em relação à média trimestral de vendas, é, então, um período de *alta* nas vendas (pois a média corrigida dos índices aditivos para este trimestre vale 1688,219, indicando um aumento nas vendas de 1688,219 milhões de dólares em relação à média trimestral).

Comentado [MMR36]: Lembre-se que no modelo aditivo a componente sazonal tem a mesma unidade da série, neste caso é uma série em milhões de dólares, então a componente sazonal será também em milhões de dólares.

Para obter os índices sazonais **multiplicativos** devemos dividir os valores originais da série pelas médias móveis centradas, a partir de 1996 III até 2001 II, cujos resultados estão na última coluna do Quadro 33. Os índices para cada trimestre serão:

Trimestre I => 0,945 **0,903** **0,950** 0,941 0,942

Trimestre II => 0,887 **0,874** **0,904** 0,904 0,901

Trimestre III=> **0,936** **0,969** 0,959 0,960 0,947

Trimestre IV=> **1,273** 1,271 **1,196** 1,208 1,203

Como é o modelo multiplicativo é preciso calcular as médias internas para cada semestre, excluindo os valores extremos, mostrados em negrito acima. Os resultados estão a seguir:

Trimestre I => (0,945+0,941+0,942)/3 ≅ 0,943

Trimestre II => (0,887+0,904+0,901) ≅ 0,897

Trimestre III=> (0,959+0,960+0,947) ≅ 0,955

Trimestre IV=> (1,271+1,208+1,203) ≅ 1,227

Somando os índices obtém-se 4,022, quando deveria ser 4, indicando que é necessária uma correção. Como a sazonalidade é trimestral:

Excesso = (4,022 – 4)/ 4 = 0,0055 Fator = 1 – 0,0055 = 0,9945

Multiplicando o fator pelos índices obtidos anteriormente:

Trimestre I => 0,943× 0,9945 ≅ 0,938

Trimestre II => 0,897× 0,9945 ≅ 0,892

Trimestre III=> 0,955× 0,9945 ≅ 0,950

Trimestre IV=> 1,227× 0,9945 ≅ 1,220

E a soma resulta 4.

E a interpretação da componente sazonal pelo modelo multiplicativo:

- há influência de sazonalidade na série de vendas da JC Penney pelo modelo multiplicativo, pois os valores dos índices sazonais multiplicativos para cada trimestre afastam-se mais de 5% de 1;
- nos três primeiros trimestres as vendas caem substancialmente em relação à média trimestral de vendas, são, então, períodos de *baixa* nas vendas, sendo o Trimestre II o de baixa mais acentuada (pois o índice sazonal multiplicativo para este trimestre vale 0,892, indicando uma queda nas vendas de 10,8% em relação à média trimestral);
- no quarto trimestre (Trimestre IV) as vendas aumentam muito em relação à média trimestral de vendas, é, então, um período de *alta* nas vendas (pois o índice sazonal multiplicativo para este trimestre vale 1,220, indicando um aumento nas vendas de 22% em relação à média trimestral).

Comentado [MMR37]: Lembre-se que no modelo multiplicativo a componente sazonal NÃO tem a mesma unidade da série, sendo um índice que modifica o valor da tendência (esta sim, tanto no modelo aditivo quanto no multiplicativo tem a mesma unidade da série).

Se a série temporal fosse registrada mensalmente haveria 12 índices sazonais. Há um exercício nas atividades de aprendizagem sobre a obtenção de componentes sazonais da série mensal mostrada na Figura 35. E no ambiente virtual o texto Análise de Séries Temporais (modelo clássico) com o Microsoft Excel ® mostra como realizar todo o procedimento com os dados mostrados na Figura 36.

5.3.1 – Remoção da Sazonalidade

Em diversas situações pode haver interesse em observar a série temporal sem a presença de componentes sazonais. Por exemplo, em uma série de índices de preços ao consumidor sabe-se que os preços sofrem variações consideráveis dentro do ano, devido à influência de safra/entressafra, feriados e temporadas de férias. Tais flutuações podem tornar difícil a identificação do efeito das outras componentes. Podemos, então remover a sazonalidade, utilizando os índices sazonais aditivos e multiplicativos: no aditivo *subtraem-se* os índices dos valores da série e no multiplicativo os valores da série são *divididos* pelos índice sazonais.

Comentado [MMR38]: Isso ocorre especialmente com as séries de índices de preços ao consumidor. Maiores detalhes em http://www.ibge.gov.br/home/estatistica/indicadores/precos/inpc_ipca/metsazon.shtm, acessado em 26/11/2015.

Para os dados da Figura 40, utilizando os índices sazonais obtidos no Exemplo 4, a série de vendas sem sazonalidade é mostrada no Quadro 35, Figura 45 e Figura 46.

Trimestre	Vendas	Índ. sazonais aditivos	Índ. sazonais multiplicativos	Série sem sazonalidade (aditivo)	Série sem sazonalidade (multiplicativo)
1996-I	4452	-503,581	0,938	4955,581	4746,268657
1996-II	4507	-833,006	0,892	5340,006	5052,690583
1996-III	5537	-351,631	0,950	5888,631	5828,421053
1996-IV	8157	1688,219	1,220	6468,781	6686,065574
1997-I	6481	-503,581	0,938	6984,581	6909,381663
1997-II	6420	-833,006	0,892	7253,006	7197,309417
1997-III	7208	-351,631	0,950	7559,631	7587,368421
1997-IV	9509	1688,219	1,220	7820,781	7794,262295
1998-I	6755	-503,581	0,938	7258,581	7201,492537
1998-II	6483	-833,006	0,892	7316,006	7267,93722
1998-III	7129	-351,631	0,950	7480,631	7504,210526
1998-IV	9072	1688,219	1,220	7383,781	7436,065574
1999-I	7339	-503,581	0,938	7842,581	7824,093817
1999-II	7104	-833,006	0,892	7937,006	7964,125561
1999-III	7639	-351,631	0,950	7990,631	8041,052632
1999-IV	9661	1688,219	1,220	7972,781	7918,852459
2000-I	7528	-503,581	0,938	8031,581	8025,586354
2000-II	7207	-833,006	0,892	8040,006	8079,596413
2000-III	7538	-351,631	0,950	7889,631	7934,736842
2000-IV	9573	1688,219	1,220	7884,781	7846,721311
2001-I	7522	-503,581	0,938	8025,581	8019,189765
2001-II	7211	-833,006	0,892	8044,006	8084,080717
2001-III	7729	-351,631	0,950	8080,631	8135,789474
2001-IV	9542	1688,219	1,220	7853,781	7821,311475

Quadro 35 – Vendas mensais da loja de departamentos JC Penney (em milhões de dólares), índice sazonais aditivos e multiplicativos e séries com a sazonalidade removida.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Pode-se perceber no Quadro 35 que as séries com a sazonalidade removida (as duas últimas colunas à direita) não são exatamente iguais, e nem poderiam ser, pois os dois modelos (aditivo e multiplicativo) têm estruturas diferentes de cálculo, mas são semelhantes, na mesma ordem de grandeza. Gráficos da série de vendas com a sazonalidade removida seriam muito semelhantes à Figura 40, com a série sem sazonalidade ocupando o lugar das médias móveis (a diferença é que na série sem sazonalidade há medidas para todos os períodos): vejamos as Figuras 45 e 46.

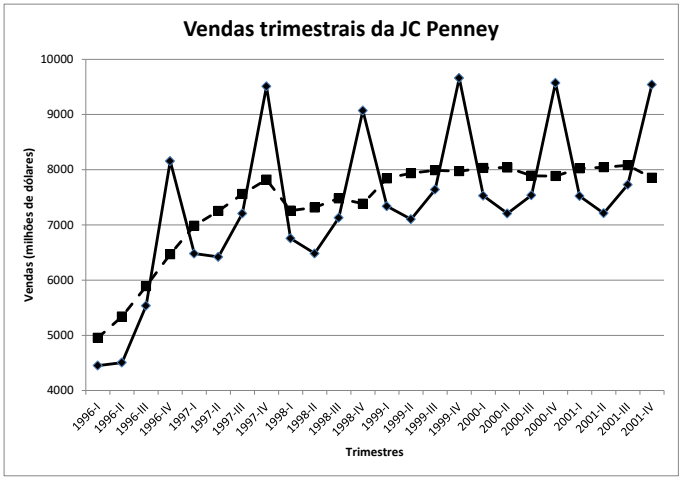


Figura 45 – Vendas trimestrais da loja de departamentos JC Penney, em milhões de dólares e série sem sazonalidade (modelo aditivo), de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

A linha pontilhada na Figura 45 mostra a série com a sazonalidade removida pelo modelo aditivo. Observa-se claramente um aumento das vendas no período de 1996 (começando em 5000 milhões) a 1999 (chegando a 8000 milhões), havendo uma estagnação até o fim da série (em 8000 milhões). Na Figura 46 a linha pontilhada mostra a série com a sazonalidade removida pelo modelo multiplicativo, sendo o comportamento muito semelhante.

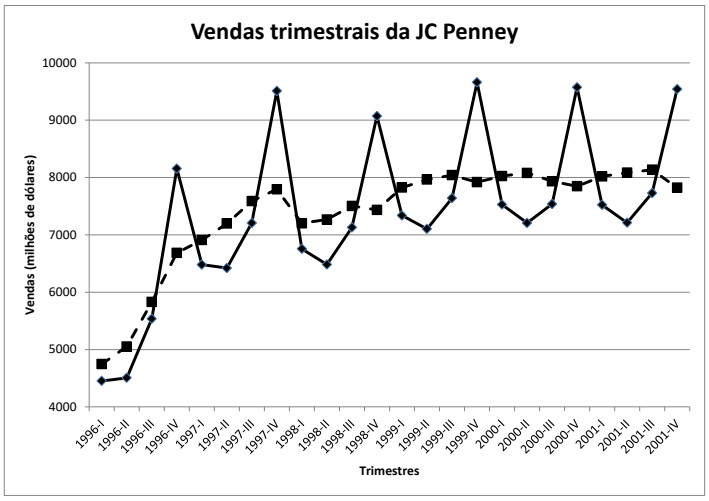


Figura 46 – Vendas trimestrais da loja de departamentos JC Penney, em milhões de dólares e série sem sazonalidade (modelo multiplicativo), de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

5.4 – Obtenção das componentes cíclica e irregular de uma série temporal

Geralmente as componentes cíclica e irregular são avaliadas em conjunto. Conforme visto anteriormente os ciclos são padrões de longo prazo (superiores a um ano), como por exemplo períodos de crescimento e recessão da economia. Já a componente irregular é resultado de fatos fortuitos, inesperados. Alguns autores não mencionam a componente cíclica porque em certos casos a série temporal precisa abranger décadas para que seja possível identificar os ciclos, e, especialmente em séries socioeconômicas os dados mais antigos podem estar realmente ultrapassados e contribuir para a construção de um modelo irreal. Não obstante, optou-se por levá-las em conta neste texto para obter um modelo completo.

As componentes cíclica e irregular são obtidas através da remoção das componentes tendência e sazonalidade (esta última apenas se os dados *não* forem anuais).

No modelo aditivo: $CI = Y - T - S$

No modelo multiplicativo: $CI = \frac{Y}{(T \times S)}$

Onde Y é o valor original da série, T é a tendência, e S é a componente sazonal (representada através dos índices sazonais).

Podemos construir um gráfico de linhas com as componentes cíclica e irregular, através do qual verificamos se os ciclos realmente influenciam a série, qual é sua periodicidade, e ainda se o efeito da componente irregular é muito grande (e se é possível relacioná-lo com fatos específicos). Às vezes estar tornam difícil a visualização dos ciclos, o que pode exigir a aplicação de médias móveis às componentes cíclica e irregular para "alisá-las", de modo a facilitar a identificação de ciclos.

Para identificar se há ciclos na série os seguintes padrões devem ser observados no gráfico das componentes cíclica e irregular:

- no modelo **aditivo**, se há alternâncias sistemáticas entre valores maiores e menores do que **zero** ao longo dos períodos, e se os valores permanecem predominantemente maiores/menores do que zero durante pelo menos 1 ano (por exemplo: 2 anos acima de zero, seguido por 3 abaixo de zero, e assim sucessivamente);
- no modelo **multiplicativo**, se há alternâncias sistemáticas entre valores maiores e menores do que **1** ao longo dos períodos, e se os valores permanecem predominantemente maiores/menores do que 1 durante pelo menos 1 ano (por exemplo: 2 anos acima de 1, seguido por 3 abaixo de 1, e assim sucessivamente);

Os valores zero e 1 são os pontos neutros nos modelos aditivo e multiplicativo respectivamente, se as variações não se afastarem muito de zero (no modelo aditivo) ou de 1 (no modelo multiplicativo) elas não causarão modificações tangíveis na tendência, e portanto não influenciarão na série. A alternância sistemática precisa ser identificada, caso contrário o efeito dos ciclos ou é inexistente ou é inferior ao da componente irregular.

Mesmo que seja identificada a presença de ciclos na série é difícil incorporá-los no modelo de previsão porque sua periodicidade pode não ser constante, especialmente se for uma série socioeconômica: três anos de alta pode ser seguidos por três de alta, ou por quatro de alta, ou mesmo por dois ou sete de alta. Embora haja a alternância, ela pode não apresentar a regularidade necessária para incorporá-la ao modelo de previsão.

Exemplo 5 - Os dados a seguir representam as vendas líquidas (em bilhões de dólares), e a tendência (obtida por uma equação polinômio de 2º grau) da Kodak. Remova a tendência da série usando os modelos aditivo e multiplicativo. Você identifica influência de ciclos?

Comentado [MMR39]: O polinômio de 2º grau foi considerado a melhor tendência por mínimos quadrados para estes dados após o cálculo das medidas de acuracidade vistas na seção 5.2.1.

Ano	Período	Vendas	Tendência = 0,0128×Período^2+0,5168×Período+1,1961
1978	1	1,6	1,7257
1979	2	2	2,2809
1980	3	2,7	2,8617
1981	4	3,7	3,4681
1982	5	4,6	4,1001
1983	6	4,62	4,7577
1984	7	5	5,4409
1985	8	5,78	6,1497
1986	9	6,3	6,8841
1987	10	8	7,6441
1988	11	10,25	8,4297
1989	12	10,5	9,2409
1990	13	11,9	10,0777
1991	14	10,2	10,9401
1992	15	10,6	11,8281
1993	16	10,6	12,7417
1994	17	11,5	13,6809
1995	18	13,3	14,6457
1996	19	17	15,6361
1997	20	18,4	16,6521
1998	21	18,9	17,6937
1999	22	18,9	18,7609
2000	23	18,94	19,8537

Quadro 36 – Vendas líquidas da Kodak (em US$ Bilhões) de 1978 a 2000 com tendência calculada pelo modelo de polinômio de 2º grau.
Fonte: adaptado pelo autor de Microsoft Excel ® a partir de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Como a série é anual NÃO HÁ influência da sazonalidade. Podemos simplesmente subtrair a Tendência das Vendas (modelo aditivo) ou dividir as Vendas pela Tendência (modelo multiplicativo), obtendo as componentes cíclica e irregular. Os resultados estão no Quadro 37.

Ano	Período	Vendas	Tendência	CI Aditivo = Vendas - Tendência	CI Multiplicativo = Vendas/Tendência
1978	1	1,6	1,7257	-0,1257	0,9272
1979	2	2	2,2809	-0,2809	0,8768
1980	3	2,7	2,8617	-0,1617	0,9435
1981	4	3,7	3,4681	0,2319	1,0669
1982	5	4,6	4,1001	0,4999	1,1219
1983	6	4,62	4,7577	-0,1377	0,9711
1984	7	5	5,4409	-0,4409	0,9190
1985	8	5,78	6,1497	-0,3697	0,9399
1986	9	6,3	6,8841	-0,5841	0,9152
1987	10	8	7,6441	0,3559	1,0466
1988	11	10,25	8,4297	1,8203	1,2159
1989	12	10,5	9,2409	1,2591	1,1363
1990	13	11,9	10,0777	1,8223	1,1808
1991	14	10,2	10,9401	-0,7401	0,9323
1992	15	10,6	11,8281	-1,2281	0,8962
1993	16	10,6	12,7417	-2,1417	0,8319
1994	17	11,5	13,6809	-2,1809	0,8406
1995	18	13,3	14,6457	-1,3457	0,9081
1996	19	17	15,6361	1,3639	1,0872
1997	20	18,4	16,6521	1,7479	1,1050
1998	21	18,9	17,6937	1,2063	1,0682
1999	22	18,9	18,7609	0,1391	1,0074
2000	23	18,94	19,8537	-0,9137	0,9540

Quadro 37 – Vendas líquidas da Kodak (em US$ Bilhões) de 1978 a 2000 com tendência calculada pelo modelo de polinômio de 2º grau, componentes cíclica e irregular pelos modelos aditivo e multiplicativo.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Antes de mostrar os gráficos das componentes cíclica e irregular podemos observar o comportamento da série de vendas líquidas da Kodak de 1978 a 2000 na Figura 47.

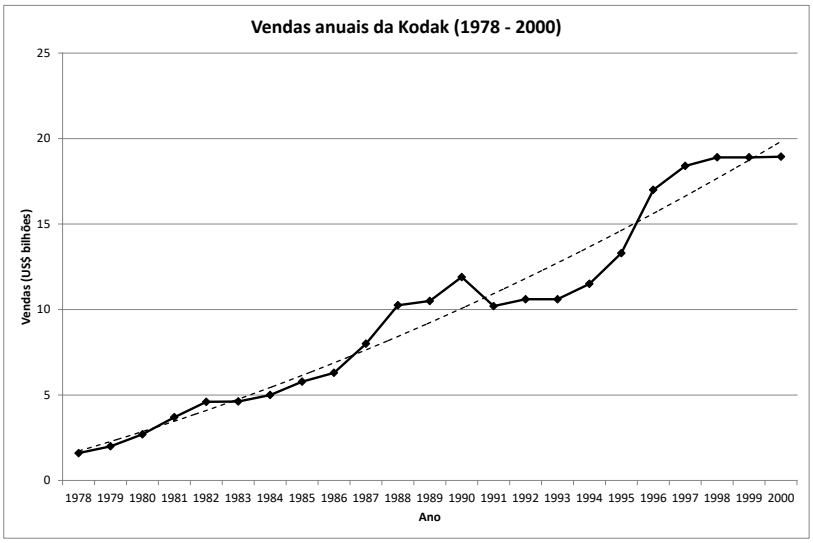


Figura 47 – Vendas líquidas da Kodak (em US$ bilhões) de 1978 a 2000 com tendência por polinômio de 2º grau.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Assumindo que a série de vendas teve o efeito da inflação removido, observa-se uma tendência de crescimento nas vendas líquidas no período analisado: em 1978 as vendas estavam abaixo de US$ 2 bilhões, chegando a mais de US$ 18 bilhões em 2000. Observa-se também o bom ajuste da curva (pontilhada) calculada pelo modelo de polinômio de 2º grau.

Comentado [MMR40]: Caso contrário o aumento real das vendas pode ser menor, dependendo da inflação acumulada no período.

Nas Figuras 48 e 49 são apresentados os gráficos das componentes cíclica e irregular para o modelo aditivo e para o multiplicativo, respectivamente, com os valores calculados nas duas últimas colunas do Quadro 37.

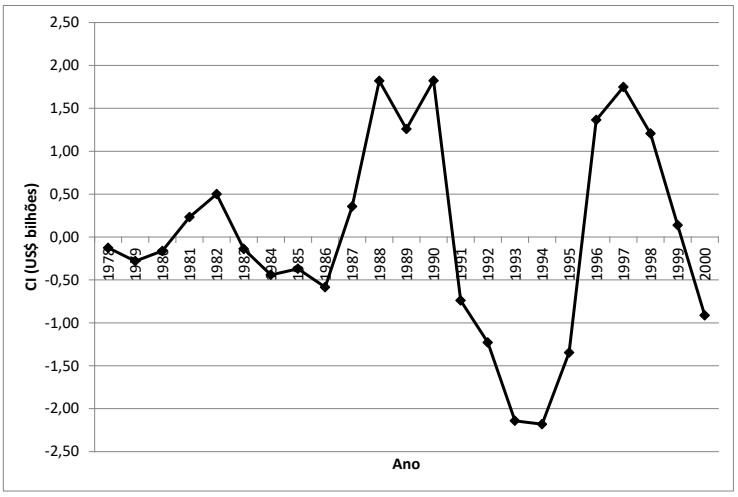


Figura 48 – Componentes cíclica e irregular da série de Vendas líquidas da Kodak (em US$ bilhões) de 1978 a 2000 pelo modelo clássico aditivo.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Observando a Figura 48, a escala do eixo vertical chama a atenção: como o modelo é aditivo a escala é a mesma da Figura 47, US$ milhões.

Comentado [MMR41]: No modelo aditivo todas as componentes têm a mesma escala da série temporal.

Pela Figura 48 é possível identificar uma variação sistemática: nos anos de 1978 a 1980 (3 anos) as componentes cíclica e irregular têm valores MENORES DO QUE ZERO (em torno de -0,25, queda de US$ bilhões, configurando um ciclo de baixa nas vendas). Em 1981 e 1982 os valores passam a ser MAIORES DO QUE ZERO (em torno de 0,25 a 0,50, aumento entre US$ 0,25 e 0,50 bilhões, configurando um ciclo de alta nas vendas). De 1983 a 1986 (4 anos), os valores são MENORES DO QUE ZERO (em torno de -0,50, queda de US$ 0,50 bilhões, outro ciclo de baixa). Em 1987 ocorre outra inversão, valores MAIORES DO QUE ZERO (em torno de 1,50, aumento de US$ 1,5 bilhões, outro ciclo de alta) até 1990. Em 1991, as componentes cíclica e irregular voltam a ficar MENORES DO QUE ZERO (flutuando de -1,0 a -2,0, queda entre US$ 1 e 2 bilhões, outro ciclo de baixa),

permanecendo assim até 1995 (5 anos). No ano de 1996 ocorre outra inversão da série, com os valores tornando a ser MAIORES DO QUE ZERO (em torno de 1,50, aumento de US$ 1,50 bilhões, outro ciclo de alta) até o ano 1999. Em 2000 as componentes cíclica e irregular voltam a ser MENORES DO QUE ZERO: presume-se que nos anos seguintes devem continuar assim, configurando outro ciclo de baixa. Conclui-se então que HÁ VARIAÇÃO CÍCLICA nesta série, pois se pode perceber uma alternância entre valores maiores e menores do que zero com duração superior a um ano. Porém, não há como incluir a componente cíclica no modelo de previsão, porque não há regularidade nas durações dos ciclos de alta e baixa: 3 anos de baixa, 2 de alta, 4 de baixa, 4 de alta, 5 de baixa, 4 alta, 1 de baixa (este último talvez incompleto). Além disso, a influência da componente irregular não parece ter sido significativa, pois os valores não apresentaram grandes flutuações dentro de cada ciclo de alta e de baixa.

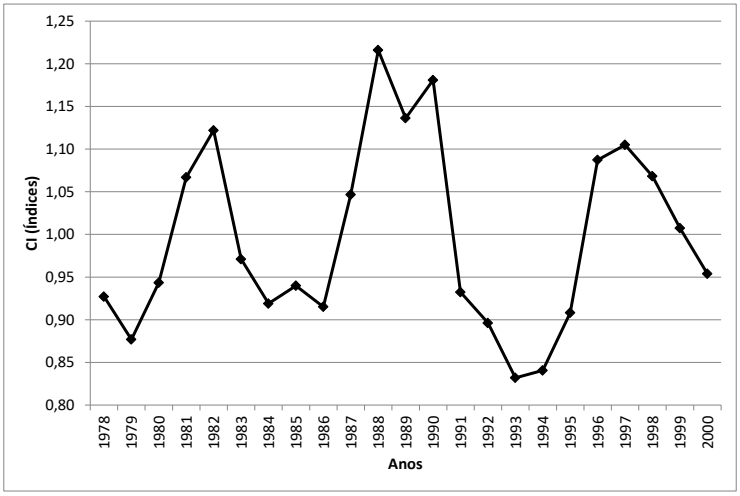


Figura 49 – Componentes cíclica e irregular da série de Vendas líquidas da Kodak (em US$ bilhões) de 1978 a 2000 pelo modelo clássico multiplicativo.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 2005.

Observando a Figura 49, a escala do eixo vertical *não* é a mesma da Figura 47. No modelo multiplicativo as componentes sazonal, cíclica e irregular são índices que, quando maiores do que 1 aumentam a tendência da série, e quando menores do que 1 reduzem a tendência.

Pela Figura 49 é possível identificar uma variação sistemática: nos anos de 1978 a 1980 (3 anos) as componentes cíclica e irregular têm valores MENORES DO QUE UM (valores em torno de 0,90, queda de 10% nas vendas em relação à média anual, configurando um ciclo de baixa nas vendas). Em 1981 e 1982 os valores passam a ser MAIORES DO QUE UM (valores em torno de 1,10, aumento de 10% nas vendas, configurando um ciclo de alta nas vendas). De 1983 a 1986 (4 anos), os valores são MENORES DO QUE UM (valores em torno de 0,95, queda de 5% nas vendas, outro ciclo de baixa). Em 1987 ocorre outra inversão, valores MAIORES DO QUE UM (valores em torno de 1,15, aumento de 15% nas vendas, outro ciclo de alta) até 1990. Em 1991, as componentes cíclica e irregular voltam a ficar MENORES DO QUE UM (valores entre 0,90 e 0,85, queda entre 10% e 15% nas vendas, outro ciclo de baixa), permanecendo assim até 1995 (5 anos). No ano de 1996 ocorre outra inversão da série, com os valores tornando a ser MAIORES DO QUE UM (valores entre 1,05, aumento de 5% nas vendas, outro ciclo de alta) até o ano 1999. Em 2000 as componentes cíclica e irregular voltam a ser MENORES DO QUE UM: presume-se que nos anos seguintes devem continuar assim, configurando outro ciclo de baixa. Conclui-se então que HÁ VARIAÇÃO CÍCLICA nesta série, pois se pode perceber uma alternância entre valores maiores e menores do que um com duração superior a um ano. Porém, não há como incluir a componente cíclica no modelo de previsão, porque não há regularidade nas durações dos ciclos de alta e baixa: 3 anos de baixa, 2 de alta, 4 de baixa, 4 de alta, 5 de baixa, 4 alta, 1 de baixa (este último talvez incompleto). Da mesma forma que no modelo aditivo a componente irregular parece ter pouca influência sobre a série temporal de vendas da Kodak no período estudado.

Se houvesse regularidade nos ciclos eles deveriam ser levados em conta na recomposição da série (para identificar qual é o modelo mais apropriado para a série: aditivo ou multiplicativo) e posterior previsão. Na recomposição devemos calcular a

Comentado [MMR42]: Quando o índice é menor do que 1, subtrai-se 1 dele para obter o valor da queda ou redução: 0,90 – 1 = -0,10, multiplicando por 100, queda de 10%.

Comentado [MMR43]: Quando o índice é maior do que 1, subtrai-se 1 dele para obter o valor do aumento: 1,10 – 1 = 0,10, multiplicando por 100, aumento de 10%.

Comentado [MMR44]: Apenas para constar, a Kodak pediu concordata nos EUA em 2012 (ver http://g1.globo.com/economia/noticia/2012/01/kodak-pede-concordata-nos-eua.html, acessado em 02/12/2015), conseguiu livrar-se temporariamente da falência vendendo várias de suas patentes e investindo no mercado de impressoras a jato de tinta, mas parece que não foi o suficiente (ver https://www.youtube.com/watch?v=aC2NdpvyGgs, acessado em 02/12/2015).

mediana das componentes cíclica e irregular de *todos* os períodos de alta e baixa (para os modelos aditivo e multiplicativo). Para a previsão é preciso identificar se o período de interesse será de alta ou de baixa, e o recomendável é utilizar a mediana das componentes cíclica e irregular do último período *completo* de alta ou de baixa.

5.5 – Recomposição da série temporal e avaliação da acuracidade do modelo recomposto

A recomposição consiste em agregar todas as componentes identificadas na análise de séries temporais. Ao fazer a recomposição, levando em conta todas as componentes que causam influência na série é possível avaliar qual modelo, aditivo ou multiplicativo, apresenta melhores resultados.

No modelo aditivo: $\hat{Y} = T + S + CI$

No modelo multiplicativo: $\hat{Y} = T \times S \times CI$

Onde T é a tendência (definida por uma equação ou por médias móveis - seção 5.2), S é a componente sazonal (definida pelos índices sazonais - seção 5.3), e CI representa as componentes cíclica e irregular (definida por índices - seção 5.4).

Comentado [MMR45]: Formalmente, seria apenas a componente cíclica, pois apenas esta poderia ser incluída no modelo, pois a componente irregular é por definição imprevisível.

Para avaliar qual modelo (aditivo ou multiplicativo) é mais apropriado para descrever uma série temporal podemos usar as medidas de acuracidade vistas na seção 5.2.1 (EAM – Erro absoluto médio, EQM – Erro quadrático médio e EPAM – Erro percentual absoluto médio). A série é recomposta pelos dois modelos e depois as medidas são calculadas, aquele que obtiver os menores valores é escolhido para realizar previsões sobre a série.

Exemplo 5 – Para os dados do Exemplo 3 - vendas trimestrais da loja de departamentos JC Penney – fazer a decomposição da série encontrando a componente tendência por mínimos quadrados, a componente sazonal e as componentes cíclica e irregular (as três últimas pelos modelos aditivo e multiplicativo). Posteriormente recompor a série com as componentes que tiverem sua influência identificada pelos modelos aditivo e multiplicativo, calcular as

medidas de acuracidade EAM, EQM e EPAM para cada modelo e decidir qual deles é o mais apropriado para as previsões das vendas futuras.

Os dados originais encontram-se no Quadro 32, e a Figura 39 mostra o gráfico de linhas das vendas trimestrais. Na Figura 50 é apresentado o gráfico das vendas com as cinco tendências por mínimos quadrados obtidas através do Microsoft Excel ®.

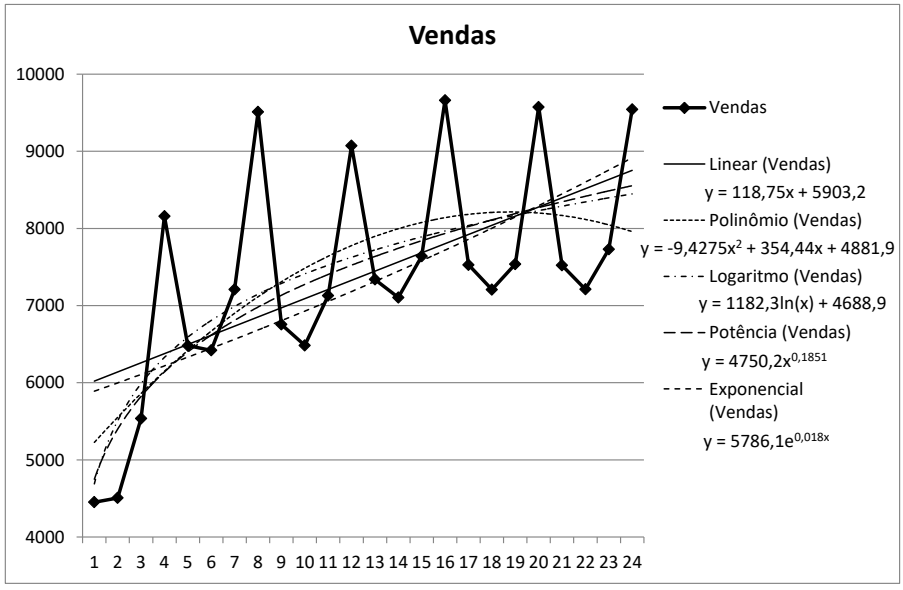


Figura 50 – Vendas trimestrais da loja de departamentos JC Penney, em milhões de dólares e cinco modelos de tendência por mínimos quadrados, de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

No modelo clássico quando há interesse de fazer previsões para vários períodos à frente deve-se usar tendência por mínimos quadrados. Vários aplicativos computacionais como o Microsoft Excel ® e o Br.Office Calc ® obtêm as equações dos modelos, onde x é um período genérico da série (no presente exemplo, variando de 1 a 24). Para avaliar qual

dos cinco modelos mostrados na Figura 50 é o mais apropriado é preciso calcular as tendências de cada um, e obter as medidas de acuracidade (EAM, EQM e EPAM) vistas na seção 5.2.1, tal como feito no Exemplo 1. As previsões feitas pelos cinco modelos mostrados na Figura 50 estão no Quadro 38.

x	$Previsão = \hat{T} =$				
	118,75x + 5903,2	$-9{,}4275x^2 + 354{,}44x + 4881{,}9$	$1182{,}3\ln(x) + 4688{,}9$	$4750{,}2x^{0{,}1851}$	$5786{,}1e^{0{,}018x}$
1	6021,95	4688,9	5226,913	4750,2	5891,193
2	6140,7	5508,408	5553,07	5400,482	5998,194
3	6259,45	5987,789	5860,373	5821,395	6107,139
4	6378,2	6327,916	6148,82	6139,786	6218,063
5	6496,95	6591,738	6418,413	6398,693	6331,002
6	6615,7	6807,297	6669,15	6618,319	6445,992
7	6734,45	6989,55	6901,033	6809,881	6563,07
8	6853,2	7147,424	7114,06	6980,296	6682,275
9	6971,95	7286,679	7308,233	7134,149	6803,645
10	7090,7	7411,246	7483,55	7274,647	6927,219
11	7209,45	7523,932	7640,013	7404,124	7053,038
12	7328,2	7626,805	7777,62	7524,339	7181,142
13	7446,95	7721,44	7896,373	7636,649	7311,573
14	7565,7	7809,057	7996,27	7742,125	7444,373
15	7684,45	7890,628	8077,313	7841,631	7579,585
16	7803,2	7966,932	8139,5	7935,869	7717,253
17	7921,95	8038,608	8182,833	8025,424	7857,421
18	8040,7	8106,187	8207,31	8110,784	8000,136
19	8159,45	8170,11	8212,933	8192,363	8145,442
20	8278,2	8230,754	8199,7	8270,515	8293,387
21	8396,95	8288,439	8167,613	8345,545	8444,02
22	8515,7	8343,439	8116,67	8417,717	8597,388
23	8634,45	8395,995	8046,873	8487,264	8753,543
24	8753,2	8446,313	7958,22	8554,389	8912,533

Quadro 38 Vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001 (em milhões de dólares) com cinco tendências obtidas por mínimos quadrados.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Realizando os cálculos dos erros absolutos, quadráticos e percentuais absolutos, para cada modelo obtemos os Quadros 39, 40 e 41. Lembrando que:

Erro absoluto = |Vendas – Previsão|; Erro quadrático = (Vendas – Previsão)^2

$$\text{Erro \% absoluto} = \left|\frac{Vendas - \text{Previsão}}{\text{Vendas}}\right| \times 100$$

x	Erros absolutos = |Vendas – Previsão|				
	Reta	Logaritmo	Polinômio de 2º grau	Potência	Exponencial
1	1569,95	236,90	774,91	298,20	1439,19
2	1633,70	1001,41	1046,07	893,48	1491,19
3	722,45	450,79	323,37	284,39	570,14
4	1778,80	1829,08	2008,18	2017,21	1938,94
5	15,95	110,74	62,59	82,31	150,00
6	195,70	387,30	249,15	198,32	25,99
7	473,55	218,45	306,97	398,12	644,93
8	2655,80	2361,58	2394,94	2528,70	2826,73
9	216,95	531,68	553,23	379,15	48,64
10	607,70	928,25	1000,55	791,65	444,22
11	80,45	394,93	511,01	275,12	75,96
12	1743,80	1445,19	1294,38	1547,66	1890,86
13	107,95	382,44	557,37	297,65	27,43
14	461,70	705,06	892,27	638,13	340,37
15	45,45	251,63	438,31	202,63	59,41
16	1857,80	1694,07	1521,50	1725,13	1943,75
17	393,95	510,61	654,83	497,42	329,42
18	833,70	899,19	1000,31	903,78	793,14
19	621,45	632,11	674,93	654,36	607,44
20	1294,80	1342,25	1373,30	1302,49	1279,61
21	874,95	766,44	645,61	823,54	922,02
22	1304,70	1132,44	905,67	1206,72	1386,39
23	905,45	666,99	317,87	758,26	1024,54
24	788,80	1095,69	1583,78	987,61	629,47

Quadro 39 Vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001 (em milhões de dólares) erros absolutos para os modelos de tendência de reta, logaritmo, polinômio de 2º grau, potência e exponencial.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Podemos calcular o EAM – Erro Absoluto Médio para cada modelo de previsão da tendência por mínimos quadrados.

x	Erros quadráticos = (Vendas – Previsão)^2				
	Reta	Logaritmo	Polinômio de 2º grau	Potência	Exponencial
1	2464743,00	56121,61	600489,38	88923,24	2071275,91
2	2668975,69	1002817,81	1094262,44	798310,63	2223660,72
3	521934,00	203211,00	104569,77	80880,30	325059,00
4	3164129,44	3345548,93	4032786,91	4069154,29	3759475,58
5	254,40	12263,00	3917,20	6774,52	22499,45
6	38298,49	149999,14	62075,72	39330,40	675,57
7	224249,60	47720,59	94229,05	158498,49	415934,64
8	7053273,64	5577042,46	5735737,60	6394342,89	7990374,56
9	47067,30	282682,15	306066,20	143754,07	2366,33
10	369299,29	861641,30	1001100,30	626704,21	197330,86
11	6472,20	155970,95	261133,78	75693,20	5770,18
12	3040838,44	2088588,21	1675419,58	2395255,11	3575342,18
13	11653,20	146260,07	310664,10	88594,72	752,22
14	213166,89	497106,05	796145,75	407203,88	115854,00
15	2065,70	63316,53	192117,85	41059,21	3530,11
16	3451420,84	2869867,59	2314962,25	2976076,09	3778151,87
17	155196,60	260720,67	428805,60	247430,71	108518,47
18	695055,69	808536,41	1000620,10	816825,53	629064,05
19	386200,10	399563,31	455533,88	428190,92	368985,61
20	1676507,04	1801623,61	1885952,89	1696467,48	1637408,60
21	765537,50	587428,55	416815,50	678225,85	850120,74
22	1702242,09	1282419,20	820238,15	1456166,51	1922072,96
23	819839,70	444882,08	101042,93	574964,49	1049687,58
24	622205,44	1200529,91	2508359,09	975375,46	396228,69

Quadro 40 - Vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001 (em milhões de dólares) erros quadráticos para os modelos de tendência de reta, logaritmo, polinômio de 2º grau, potência e exponencial.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Podemos calcular o EQM – Erro Quadrático Médio para cada modelo de previsão da tendência por mínimos quadrados.

x	Erros percentuais absolutos = $\left\|\frac{\text{Vendas - Previsão}}{Vendas}\right\| \times 100$				
	Reta	Logaritmo	Polinômio de 2º grau	Potência	Exponencial
1	35,26	5,32	17,41	6,70	32,33
2	36,25	22,22	23,21	19,82	33,09
3	13,05	8,14	5,84	5,14	10,30
4	21,81	22,42	24,62	24,73	23,77
5	0,25	1,71	0,97	1,27	2,31
6	3,05	6,03	3,88	3,09	0,40
7	6,57	3,03	4,26	5,52	8,95
8	27,93	24,84	25,19	26,59	29,73
9	3,21	7,87	8,19	5,61	0,72
10	9,37	14,32	15,43	12,21	6,85
11	1,13	5,54	7,17	3,86	1,07
12	19,22	15,93	14,27	17,06	20,84
13	1,47	5,21	7,59	4,06	0,37
14	6,50	9,92	12,56	8,98	4,79
15	0,59	3,29	5,74	2,65	0,78
16	19,23	17,54	15,75	17,86	20,12
17	5,23	6,78	8,70	6,61	4,38
18	11,57	12,48	13,88	12,54	11,01
19	8,24	8,39	8,95	8,68	8,06
20	13,53	14,02	14,35	13,61	13,37
21	11,63	10,19	8,58	10,95	12,26
22	18,09	15,70	12,56	16,73	19,23
23	11,71	8,63	4,11	9,81	13,26
24	8,27	11,48	16,60	10,35	6,60

Quadro 41 - Vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001 (em milhões de dólares) erros percentuais absolutos para os modelos de tendência de reta, logaritmo, polinômio de 2º grau, potência e exponencial.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Podemos calcular o EQM – Erro Quadrático Médio para cada modelo de previsão da tendência por mínimos quadrados.

Para decidir qual é o melhor modelo de tendência por mínimos quadrados é preciso comparar suas medidas de acuracidade, conforme mostradas no Quadro 42.

Medidas de acuracidade	Modelo de previsão de tendência				
	Reta	Logaritmo	Polinômio de 2º grau	Potência	Exponencial
EAM	882,73	832,30	878,80	**820,50**	870,41
EQM	1254192,763	**1006077,547**	1091793,585	1052675,091	1310422,495
EPAM	12,22	10,88	11,66	**10,60**	11,86

Quadro 42 – Medidas de acuracidade dos modelos de previsão de tendência por mínimos quadrados das vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

No Quadro 42 as menores medidas estão e negrito: o menor EAM foi obtido para a tendência do modelo Potência, o menor EQM foi obtido para o modelo Logaritmo e o menor EPAM para o modelo Potência. Por ter apresentado menores EAM e EPAM o modelo Potência deve ser escolhido como a tendência por mínimos quadrados para a série de vendas trimestrais da JC Penney.

No Exemplo 4 obtivemos os índices sazonais das vendas, tanto para o modelo aditivo quanto para o multiplicativo. Podemos ver os resultados no Quadro 43.

Trimestre	Índice sazonal aditivo	Índice sazonal multiplicativo
I	-503,581	0,943
II	-833,006	0,897
III	-351,631	0,955
IV	1688,219	1,227

Quadro 43 – Índices sazonais aditivos e multiplicativos para as vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Agora precisamos avaliar se há influência de ciclos na série. Para tanto vamos empregar as expressões vistas na seção 5.4:

No modelo aditivo: $CI = Y - T - S$

No modelo multiplicativo: $$CI = \frac{Y}{(T \times S)}$$

Onde Y é o valor das vendas trimestrais, T é a tendência calculada pelo modelo potência, e S é a componente sazonal (representada através dos índices sazonais, aditivos ou multiplicativos). Os resultados estão no Quadro 44.

Período	Trimestre	Vendas	T(potência)	S (aditivo)	S (multiplicativo)	CI Aditivo	CI (multiplicativo)
1	1996-I	4452	4750,200	-503,581	0,938	205,381	1,000
2	1996-II	4507	5400,482	-833,006	0,892	-60,476	0,935
3	1996-III	5537	5821,395	-351,631	0,950	67,237	1,001
4	1996-IV	8157	6139,786	1688,219	1,220	328,996	1,089
5	1997-I	6481	6398,693	-503,581	0,938	585,889	1,080
6	1997-II	6420	6618,319	-833,006	0,892	634,687	1,087
7	1997-III	7208	6809,881	-351,631	0,950	749,750	1,114
8	1997-IV	9509	6980,296	1688,219	1,220	840,485	1,116
9	1998-I	6755	7134,149	-503,581	0,938	124,432	1,010
10	1998-II	6483	7274,647	-833,006	0,892	41,360	0,999
11	1998-III	7129	7404,124	-351,631	0,950	76,507	1,014
12	1998-IV	9072	7524,339	1688,219	1,220	-140,558	0,988
13	1999-I	7339	7636,649	-503,581	0,938	205,933	1,025
14	1999-II	7104	7742,125	-833,006	0,892	194,881	1,029
15	1999-III	7639	7841,631	-351,631	0,950	149,001	1,026
16	1999-IV	9661	7935,869	1688,219	1,220	36,912	0,998
17	2000-I	7528	8025,424	-503,581	0,938	6,157	1,001
18	2000-II	7207	8110,784	-833,006	0,892	-70,778	0,996
19	2000-III	7538	8192,363	-351,631	0,950	-302,732	0,969
20	2000-IV	9573	8270,515	1688,219	1,220	-385,734	0,948
21	2001-I	7522	8345,545	-503,581	0,938	-319,963	0,961
22	2001-II	7211	8417,717	-833,006	0,892	-373,711	0,960
23	2001-III	7729	8487,264	-351,631	0,950	-406,633	0,959
24	2001-IV	9542	8554,389	1688,219	1,220	-700,608	0,914

Quadro 44 – Obtenção das componentes cíclica e irregular pelos modelos aditivo e multiplicativo das vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001, a partir da tendência pelo modelo potência e índices sazonais aditivos e multiplicativos.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Ao construir gráficos de linhas para os resultados das duas últimas colunas do Quadro 44 podemos ver as Figuras 51 e 52.

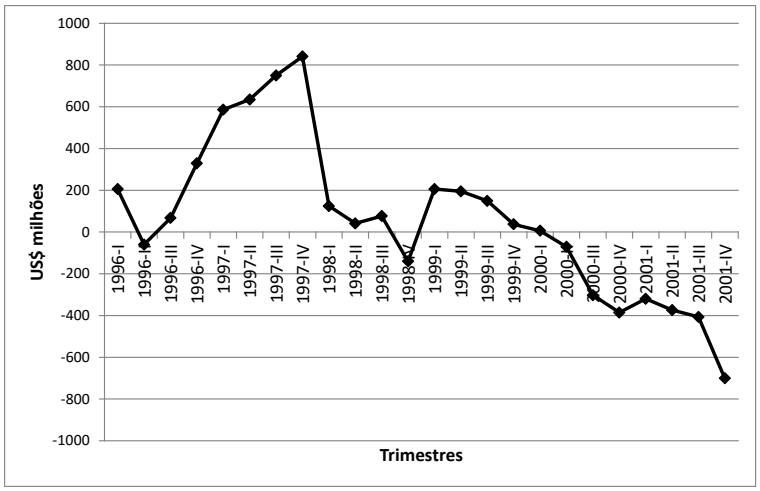


Figura 51 – Componentes cíclica e irregular (modelo aditivo) das vendas trimestrais da loja de departamentos JC Penney (em US$ milhões) de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Comentado [MMR46]: No modelo aditivo as componentes têm a mesma unidade da série.

Embora os valores variem em torno de zero, não há uma variação sistemática (uma sequência definida de anos de alta e baixa), portanto não devemos incluir ciclos na recomposição das vendas da JC Penney. Muito provavelmente ou não há realmente ciclos, ou a influência das variações irregulares suplanta os seus efeitos.

Comentado [MMR47]: O zero é o "neutro" do modelo aditivo, já que é o valor que somado a outro não causa modificação, indicando a ausência de influência da componente sob análise.

Raciocínio semelhante pode ser feito com o gráfico da Figura 52, para o modelo multiplicativo, mas agora observando que embora haja variação em torno de 1, não há variação sistemática que permita identificar ciclos estáveis que possam ser incluídos na recomposição das vendas da JC Penney.

Comentado [MMR48]: O 1 é o "neutro" do modelo multiplicativo, já que é o valor que somado multiplicado por outro não causa modificação, indicando a ausência de influência da componente sob análise.

Conclui-se então que não evidência de influência de ciclos na série de vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001, e que a recomposição (tanto pelo modelo aditivo quanto pelo multiplicativo) deve incluir apenas a tendência (modelo potência) e os índices sazonais.

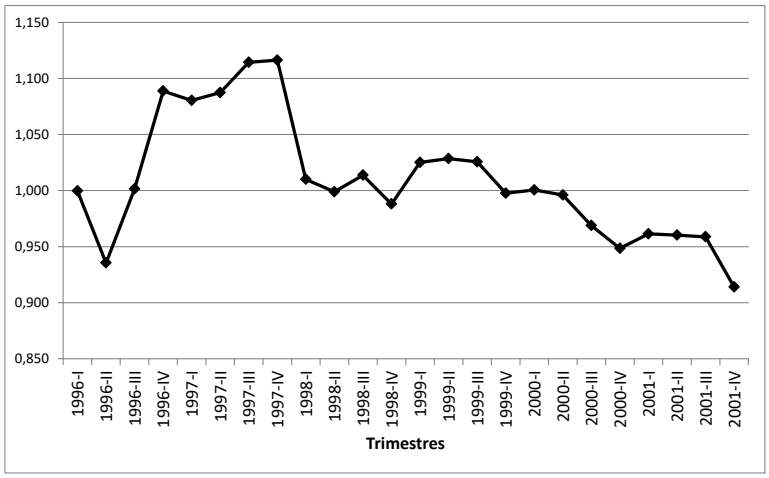


Figura 52 – Componentes cíclica e irregular (modelo multiplicativo) das vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Para fazer a recomposição é preciso então agregar tendência e componente sazonal, e depois calcular as medidas de acuracidade para os modelos aditivo e multiplicativo para identificar qual deles é o mais apropriado para série sob análise.

Pelo modelo aditivo: $\hat{Y} = \hat{T} + S$, onde $\hat{Y}$ é o valor recomposto da série (neste caso pelo modelo aditivo), $\hat{T}$ é a tendência obtida pelo modelo potência e S os índices sazonais pelo modelo aditivo (um índice para cada trimestre, repetindo-se nos respectivos semestres ao longo da série). Os resultados podem ser vistos no Quadro 45.

Comentado [MMR49]: A mesma tendência é usada nos dois modelos.

Pelo modelo multiplicativo: $\hat{Y} = \hat{T} \times S$, onde $\hat{Y}$ é o valor recomposto da série (neste caso pelo modelo multiplicativo), $\hat{T}$ é a tendência obtida pelo modelo potência e S os índices sazonais pelo modelo multiplicativo (um índice para cada trimestre, repetindo-se nos respectivos semestres ao longo da série). Os resultados podem ser vistos no Quadro 45.

Comentado [MMR50]: A mesma tendência é usada nos dois modelos.

Período	Trimestre	Vendas	T(potência)	S (aditivo)	S (multiplicativo)	$\hat{Y}$ (aditivo)	$\hat{Y}$ (multiplicativo)
1	1996-I	4452	4750,200	-503,581	0,938	4246,619	4453,390
2	1996-II	4507	5400,482	-833,006	0,892	4567,476	4818,037
3	1996-III	5537	5821,395	-351,631	0,950	5469,763	5529,451
4	1996-IV	8157	6139,786	1688,219	1,220	7828,004	7492,726
5	1997-I	6481	6398,693	-503,581	0,938	5895,111	5998,878
6	1997-II	6420	6618,319	-833,006	0,892	5785,313	5904,529
7	1997-III	7208	6809,881	-351,631	0,950	6458,250	6468,365
8	1997-IV	9509	6980,296	1688,219	1,220	8668,515	8518,448
9	1998-I	6755	7134,149	-503,581	0,938	6630,568	6688,381
10	1998-II	6483	7274,647	-833,006	0,892	6441,640	6490,071
11	1998-III	7129	7404,124	-351,631	0,950	7052,493	7032,806
12	1998-IV	9072	7524,339	1688,219	1,220	9212,558	9182,374
13	1999-I	7339	7636,649	-503,581	0,938	7133,067	7159,482
14	1999-II	7104	7742,125	-833,006	0,892	6909,119	6907,132
15	1999-III	7639	7841,631	-351,631	0,950	7489,999	7448,372
16	1999-IV	9661	7935,869	1688,219	1,220	9624,088	9684,588
17	2000-I	7528	8025,424	-503,581	0,938	7521,843	7523,966
18	2000-II	7207	8110,784	-833,006	0,892	7277,778	7236,030
19	2000-III	7538	8192,363	-351,631	0,950	7840,732	7781,515
20	2000-IV	9573	8270,515	1688,219	1,220	9958,734	10092,974
21	2001-I	7522	8345,545	-503,581	0,938	7841,963	7824,084
22	2001-II	7211	8417,717	-833,006	0,892	7584,711	7509,861
23	2001-III	7729	8487,264	-351,631	0,950	8135,633	8061,627
24	2001-IV	9542	8554,389	1688,219	1,220	10242,608	10439,402

Quadro 45 – Recomposição pelos modelos aditivo e multiplicativo das vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001, a partir da tendência pelo modelo potência e índices sazonais aditivos e multiplicativos.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Para calcular as medidas EAM, EQM e EPAM é necessário calcular os erros para cada período, para os dois modelos. Independente do modelo os erros serão calculados da seguinte forma, e os resultados para o aditivo estão no Quadro 46 e para o multiplicativo no Quadro 47:

$$\text{Erro absoluto} = \left| Vendas - \hat{Y} \right|$$

$$\text{Erro quadrático} = \left(Vendas - \hat{Y} \right)^2$$

$$\text{Erro percentual absoluto} = \left| \left(\frac{Vendas - \hat{Y}}{Vendas} \right) \times 100 \right|$$

Período	Trimestre	Vendas	$\hat{Y}$ (aditivo)	$\left\|Vendas - \hat{Y}\right\|$	$\left(Vendas - \hat{Y}\right)^2$	$\left\| \left(\frac{Vendas - \hat{Y}}{Vendas} \right) \times 100 \right\|$
1	1996-I	4452	4246,619	205,381	42181,46	4,613
2	1996-II	4507	4567,476	60,476	3657,35	1,342
3	1996-III	5537	5469,763	67,237	4520,77	1,214
4	1996-IV	8157	7828,004	328,996	108238,20	4,033
5	1997-I	6481	5895,111	585,889	343265,57	9,040
6	1997-II	6420	5785,313	634,687	402827,99	9,886
7	1997-III	7208	6458,250	749,750	562124,97	10,402
8	1997-IV	9509	8668,515	840,485	706415,11	8,839
9	1998-I	6755	6630,568	124,432	15483,35	1,842
10	1998-II	6483	6441,640	41,360	1710,63	0,638
11	1998-III	7129	7052,493	76,507	5853,36	1,073
12	1998-IV	9072	9212,558	140,558	19756,43	1,549
13	1999-I	7339	7133,067	205,933	42408,24	2,806
14	1999-II	7104	6909,119	194,881	37978,59	2,743
15	1999-III	7639	7489,999	149,001	22201,15	1,951
16	1999-IV	9661	9624,088	36,912	1362,50	0,382
17	2000-I	7528	7521,843	6,157	37,91	0,082
18	2000-II	7207	7277,778	70,778	5009,49	0,982
19	2000-III	7538	7840,732	302,732	91646,51	4,016
20	2000-IV	9573	9958,734	385,734	148790,44	4,029
21	2001-I	7522	7841,963	319,963	102376,60	4,254
22	2001-II	7211	7584,711	373,711	139659,91	5,183
23	2001-III	7729	8135,633	406,633	165350,30	5,261
24	2001-IV	9542	10242,608	700,608	490851,24	7,342

Quadro 46 – Erros absolutos, quadráticos e percentuais absolutos da recomposição pelo modelo aditivo das vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Período	Trimestre	Vendas	$\hat{Y}$ (multiplicativo)	$\lvert Vendas - \hat{Y} \rvert$	$(Vendas - \hat{Y})^2$	$\left\lvert \left(\frac{Vendas - \hat{Y}}{Vendas} \right) \times 100 \right\rvert$
1	1996-I	4452	4453,390	1,390	1,93	0,031
2	1996-II	4507	4818,037	311,037	96743,73	6,901
3	1996-III	5537	5529,451	7,549	56,99	0,136
4	1996-IV	8157	7492,726	664,274	441260,53	8,144
5	1997-I	6481	5998,878	482,122	232441,18	7,439
6	1997-II	6420	5904,529	515,471	265710,72	8,029
7	1997-III	7208	6468,365	739,635	547059,81	10,261
8	1997-IV	9509	8518,448	990,552	981193,20	10,417
9	1998-I	6755	6688,381	66,619	4438,09	0,986
10	1998-II	6483	6490,071	7,071	50,00	0,109
11	1998-III	7129	7032,806	96,194	9253,21	1,349
12	1998-IV	9072	9182,374	110,374	12182,39	1,217
13	1999-I	7339	7159,482	179,518	32226,54	2,446
14	1999-II	7104	6907,132	196,868	38757,13	2,771
15	1999-III	7639	7448,372	190,628	36338,97	2,495
16	1999-IV	9661	9684,588	23,588	556,37	0,244
17	2000-I	7528	7523,966	4,034	16,28	0,054
18	2000-II	7207	7236,030	29,030	842,76	0,403
19	2000-III	7538	7781,515	243,515	59299,64	3,231
20	2000-IV	9573	10092,974	519,974	270373,43	5,432
21	2001-I	7522	7824,084	302,084	91254,77	4,016
22	2001-II	7211	7509,861	298,861	89317,66	4,145
23	2001-III	7729	8061,627	332,627	110640,71	4,304
24	2001-IV	9542	10439,402	897,402	805330,38	9,405

Quadro 47 – Erros absolutos, quadráticos e percentuais absolutos da recomposição pelo modelo multiplicativo das vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Os resultados com as medidas de acuracidade para os dois modelos estão no Quadro 48.

Modelo	Medidas de acuracidade		
	EAM	EQM	EPAM
Aditivo	292,033	144321,17	3,896
Multiplicativo	300,434	171889,43	3,915

Quadro 48 – Medidas de acuracidade da recomposição pelos modelos aditivo e multiplicativo das vendas trimestrais da loja de departamentos JC Penney de 1996 a 2001.

Fonte: adaptado pelo autor de Microsoft Excel ® a partir de dados de 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.

Como o modelo multiplicativo apresenta as menores medidas de acuracidade, indicando menores erros, deve ser o escolhido para realizar previsões das vendas trimestrais da loja de departamentos JC Penney.

Uma vez escolhido o modelo mais apropriado como deve ser feita a previsão para os períodos seguintes ao término da série. No caso do Exemplo 5, em que a série tem 24 períodos, a previsão para os quatro trimestres seguintes (25 a 28), pelo modelo multiplicativo deveriam ser feitas incluindo apenas a tendência (por meio da equação de potência) e os índices sazonais, resultando:

$Y_{25} = (4750{,}2 \times 25^{0{,}1851}) \times 0{,}938 = 8080{,}877$ (milhões de dólares)

$Y_{26} = (4750{,}2 \times 26^{0{,}1851}) \times 0{,}892 = 7745{,}706$ (milhões de dólares)

$Y_{27} = (4750{,}2 \times 27^{0{,}1851}) \times 0{,}950 = 8304{,}477$ (milhões de dólares)

$Y_{28} = (4750{,}2 \times 28^{0{,}1851}) \times 1{,}220 = 10741{,}560$ (milhões de dólares)

5.6 – Outros modelos de séries temporais

Além do modelo clássico apresentado neste capítulo podem ser usados os métodos de Holt-Winters para modelos multiplicativos e aditivos.

Há também outras abordagens diferentes do modelo clássico. Entre estes modelos devem ser citados os modelos Autoregressivos (AR), os modelos de Médias Móveis Autoregressivos de (ARMA), os modelos de Médias Móveis Integrados Autoregressivos (ARIMA) e os modelos de Médias Móveis Integrados Autoregressivos Sazonais (SARIMA). Tais tópicos geralmente são vistos em cursos de pós-graduação.

Tô afim de saber:

- Sobre uso de análise de séries temporais para previsões, SAMOHYL, R. W. ; SOUZA, G. P. ; MIRANDA, R. G. . **Métodos simplificados de previsão empresarial**. 1. ed. Rio de Janeiro: Ciência Moderna, 2008. v. 1. 182p.
- Sobre o uso do modelo clássico de séries temporais, STEVENSON, Willian J. **Estatística Aplicada à Administração**. – São Paulo: Harbra, 2001.
- Para análise de séries temporais para previsão com dados anuais usando o Microsoft Excel ®, LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. **Estatística: Teoria e Aplicações - Usando Microsoft Excel em Português**. 5ª ed. – Rio de Janeiro: LTC, 2005.
- Sobre modelos de Holt-Winters de séries temporais, SOARES, J. F., FARIAS, A. A., CESAR, C. C. – **Introdução à Estatística**, LTC, Rio de Janeiro, 1991.
- Sobre modelos de médias móveis (MA) e autoregressivos (AR) usando aplicativos computacionais, 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.
- Para tópicos mais avançados de modelos de médias móveis, autoregressivos, autoregressivos de médias móveis (ARMA), autoregressivos integrados de médias móveis (ARIMA), MORETTIN, P. A. ; Toloi, C.M.C.. **Analise de Séries Temporais**. 2. ed. São Paulo: Editora Edgard Blucher, 2006. 535p
- Para saber como realizar as análises descritas nesta Unidade através do Microsoft ® consulte *Como realizar análise de séries temporais (modelo clássico) no Microsoft Excel ®*, disponível no Ambiente Virtual assim como os arquivos de dados usados nos exemplos apresentados.

Resumo

O resumo desta Unidade está demonstrado na Figura 53:

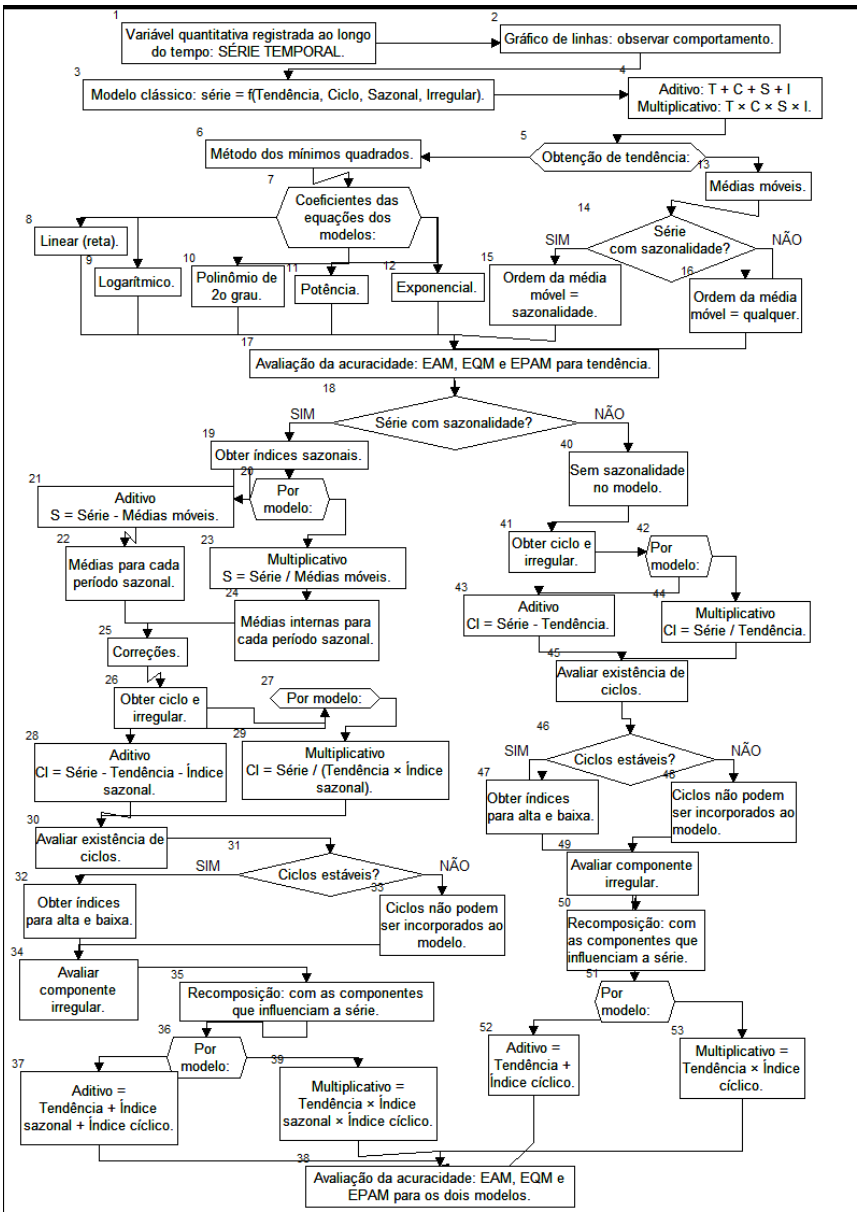


Figura 53 – Resumo da Unidade 5

Fonte: elaborado pelo autor

Nesta Unidade estudamos como analisar dados de uma variável quantitativa registrada ao longo do tempo. 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 6, veremos os conceitos básicos de probabilidade que serão imprescindíveis para a disciplina de Estatística Aplicada à Administração II. Vamos em frente e ótimos estudos!!!

Atividades de aprendizagem

1) A que componentes de uma série temporal (pelo modelo clássico) estariam principalmente associados cada um dos seguintes eventos. JUSTIFIQUE suas respostas.
a) Uma recessão.
b) Um acréscimo na oferta de empregos durante os meses de verão.
c) O declínio da taxa de mortalidade decorrente do progresso da medicina.
d) Uma greve na indústria do aço.
e) Uma procura continuamente crescente por automóveis pequenos.
f) O efeito nas vendas de cigarros das crescentes restrições ao fumo em lugares fechados e a divulgação de mais pesquisas mostrando os malefícios do tabagismo.
g) Maior procura por roupas de lã.
h) O fenômeno climático "El Niño".
i) Um terremoto em Taiwan que danificou várias fábricas de memórias RAM para computadores.
j) Maior procura por artigos de papelaria e livros escolares.
k) Aumento no volume total de benefícios pagos pelo INSS.
(Adaptado de SPIEGEL, M.R., Estatística, 3ª edição – São Paulo: Makron Books, 1993, pg. 468).

2) Considere os dados mostrados na Figura 34, série mensal da produção de veículos automotores no Brasil de 1997 a 2014 (disponível no ambiente virtual). Naquele exemplo a

melhor tendência por mínimos quadrados encontrada foi o modelo polinômio de 2º grau (ver as medidas de acuracidade no Quadro 26), com a seguinte equação:

$\widehat{Y}_x = 0{,}9929 \times \text{período}^2 + 772{,}79 \times \text{período} + 104672$.

Sabendo disso responda os itens seguir (é extremamente recomendado que seja usada uma planilha eletrônica para isso).

a) Faça a previsão da tendência pelo modelo polinômio de 2º grau descrito acima para os períodos 1 (janeiro de 1997) a 216 (dezembro de 2014).

b) Obtenha os 12 índices sazonais da série para o modelo *aditivo* usando o procedimento visto na seção 5.3. Com base nos valores obtidos há evidência de influência de sazonalidade na série temporal pelo modelo aditivo? JUSTIFIQUE.

c) Obtenha os 12 índices sazonais da série para o modelo *multiplicativo* usando o procedimento visto na seção 5.3. Com base nos valores obtidos há evidência de influência de sazonalidade na série temporal pelo modelo multiplicativo? JUSTIFIQUE.

d) Obtenha as componentes cíclica e irregular da série para o modelo *aditivo* usando o procedimento visto na seção 5.4. Construa um gráfico de linhas com os resultados obtidos. Com base neste gráfico há evidência de influência de ciclos na série temporal pelo modelo aditivo? JUSTIFIQUE. Se houver regularidade nos ciclos identifique a extensão dos períodos de alta e de baixa e calcule índices cíclicos que os representem (medianas de todos os períodos de alta e de todos os períodos de baixa, respectivamente).

e) Obtenha as componentes cíclica e irregular da série para o modelo *multiplicativo* usando o procedimento visto na seção 5.4. Construa um gráfico de linhas com os resultados obtidos. Com base neste gráfico há evidência de influência de ciclos na série temporal pelo modelo aditivo? JUSTIFIQUE. Se houver regularidade nos ciclos identifique a extensão dos períodos de alta e de baixa e calcule índices cíclicos que os representem (medianas de todos os períodos de alta e de todos os períodos de baixa, respectivamente).

f) Utilizando todas as componentes que influenciam a série temporal faça a sua recomposição pelo modelo aditivo: use o modelo de mínimos quadrados com os menores valores das medidas de acuracidade, os índices sazonais pelo modelo aditivo obtidos no item b (se houver influência da sazonalidade) e os índices cíclicos obtidos no item d (se houver influência e regularidade nos ciclos).

g) Utilizando todas as componentes que influenciam a série temporal faça a sua recomposição pelo modelo multiplicativo: use o modelo de mínimos quadrados com os menores valores das medidas de acuracidade, os índices sazonais pelo modelo aditivo obtidos no item c (se houver influência da sazonalidade) e os índices cíclicos obtidos no item e (se houver influência e regularidade nos ciclos).

h) Calcule as medidas de acuracidade EAM, EQM e EPAM para as recomposições para os modelos aditivo e multiplicativo. Qual dos dois modelos é o mais apropriado para representar a série mensal da produção de veículos automotores no Brasil de 1997 a 2014? JUSTIFIQUE.

i) Usando o modelo escolhido no item h, com as componentes tendência (o modelo com os menores valores das medidas de acuracidade), sazonalidade (usando os índices sazonais apropriados, se houver sua influência na série) e ciclos (identificando se 2015 será de baixa ou de alta, e usando os índices cíclicos apropriados – do último período completo de baixa ou de alta - se houver sua influência na série) faça as previsões para os doze meses de 2015 (períodos 217 a 228 da série).

2) A corretora DEFICITCERTO opera no mercado de ações do país asiático Chung Kuo, um dos principais "Tigres Asiáticos". Há interesse em avaliar o comportamento do índice Chun da bolsa da capital de Chung Kuo, que tem uma estrutura semelhante ao IBOVESPA, sendo medido em pontos. Há uma série de valores mensais em milhares de pontos (de fechamento dos meses), de janeiro de 1998 até dezembro de 2015 (216 períodos), disponível no ambiente virtual. Vocês precisam fazer previsões confiáveis sobre os valores do índice em 2016. RECOMENDAÇÃO IMPORTANTE: use uma planilha eletrônica para resolver este exercício.

a) Construa um gráfico de linhas da série do índice Chun.

b) Adicione as tendências dos modelos linear, logarítmico, polinômio de 2º grau, potência e exponencial ao gráfico do item a (use uma planilha eletrônica como o Microsoft Excel ®).

c) Faça a previsão da tendência de janeiro de 1998 (período 1) a dezembro de 2015 (período 216) pelos cinco modelos obtidos no item b.

d) Calcule as medidas de acuracidade dos cinco modelos (EAM, EQM, e EPAM, o erro será a diferença entre o valor da série e as previsões de tendência através de cada modelo) e

escolha qual (linear, logarítmico, polinômio de 2º grau, potência ou exponencial) é o mais apropriado para representar a tendência da série do índice Chun. JUSTIFIQUE.

e) Obtenha os 12 índices sazonais da série para o modelo *aditivo* usando o procedimento visto na seção 5.3. Com base nos valores obtidos há evidência de influência de sazonalidade na série temporal pelo modelo aditivo? JUSTIFIQUE.

f) Obtenha os 12 índices sazonais da série para o modelo *multiplicativo* usando o procedimento visto na seção 5.3. Com base nos valores obtidos há evidência de influência de sazonalidade na série temporal pelo modelo multiplicativo? JUSTIFIQUE.

g) Obtenha as componentes cíclica e irregular da série para o modelo *aditivo* usando o procedimento visto na seção 5.4. Construa um gráfico de linhas com os resultados obtidos. Com base neste gráfico há evidência de influência de ciclos na série temporal pelo modelo aditivo? JUSTIFIQUE. Se houver regularidade nos ciclos identifique a extensão dos períodos de alta e de baixa e calcule índices cíclicos que os representem (medianas de todos os períodos de alta e de todos os períodos de baixa, respectivamente).

h) Obtenha as componentes cíclica e irregular da série para o modelo *multiplicativo* usando o procedimento visto na seção 5.4. Construa um gráfico de linhas com os resultados obtidos. Com base neste gráfico há evidência de influência de ciclos na série temporal pelo modelo aditivo? JUSTIFIQUE. Se houver regularidade nos ciclos identifique a extensão dos períodos de alta e de baixa e calcule índices cíclicos que os representem (medianas de todos os períodos de alta e de todos os períodos de baixa, respectivamente).

i) Utilizando todas as componentes que influenciam a série temporal faça a sua recomposição pelo modelo aditivo: use o modelo de mínimos quadrados com os menores valores das medidas de acuracidade, os índices sazonais pelo modelo aditivo obtidos no item b (se houver influência da sazonalidade) e os índices cíclicos obtidos no item d (se houver influência e regularidade nos ciclos).

j) Utilizando todas as componentes que influenciam a série temporal faça a sua recomposição pelo modelo multiplicativo: use o modelo de mínimos quadrados com os menores valores das medidas de acuracidade, os índices sazonais pelo modelo aditivo obtidos no item c (se houver influência da sazonalidade) e os índices cíclicos obtidos no item e (se houver influência e regularidade nos ciclos).

k) Calcule as medidas de acuracidade EAM, EQM e EPAM para as recomposições para os modelos aditivo e multiplicativo. Qual dos dois modelos é o mais apropriado para representar a série mensal do índice Chun de janeiro de 1998 a dezembro de 2015? JUSTIFIQUE.

l) Usando o modelo escolhido no item h, com as componentes tendência (o modelo com os menores valores das medidas de acuracidade), sazonalidade (usando os índices sazonais apropriados, se houver sua influência na série) e ciclos (identificando se 2016 será de baixa ou de alta, e usando os índices cíclicos apropriados – do último período completo de baixa ou de alta - se houver sua influência na série) faça as previsões para os doze meses de 2016 (períodos 217 a 228 da série).

Unidade 6

Conceitos básicos da 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.

6 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 frequências e medidas de síntese foi possível caracterizar a variabilidade GLOSSÁRIO Variabilidade: diferenças encontradas por sucessivas medições realizadas em pessoas, animais ou objetos, em tempos ou situações diferentes. Fonte: Montgomery, 2004. Fim GLOSSÁRIO do fenômeno observado, e elaborar hipóteses ou conjecturas a respeito.

Suponha que estamos 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 frequências da variável qualitativa sexo dos recém-nascidos) há interesse em prever qual será o percentual de nascimentos no ano de 2030: 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 2020? 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 certa regularidade nos percentuais de nascimentos (verificados anteriormente), mas não é possível responder qual será o resultado exato antes do fenômeno ocorrer.

A regularidade citada (que foi observada para um grande número de nascimentos) permite que seja calculado o grau de certeza, ou confiabilidade, da previsão feita, que recebe o nome de **Probabilidade**. GLOSSÁRIO 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. GLOSSÁRIO 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 GLOSSÁRIO Modelo Probabilístico: modelo matemático para descrever a certeza de ocorrência de eventos, onde são definidos os eventos possíveis e uma regra de ocorrência para calcular quão provável é cada evento ou conjunto de eventos. Fonte: Barbetta, 2007. Fim GLOSSÁRIO 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.

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

A regra citada foi definida a partir de observações anteriores do fenômeno, mas também poderia ser formulada com base em considerações teóricas. Por exemplo, se há interesse em estudar as proporções de ocorrências das faces de um dado, 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. LINK 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 24000 vezes (!) para verificar a validade dessa lei: obteve 12012 caras, praticamente o valor esperado (12000, 50%). LINK

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 pode-se prever exatamente qual será o seu resultado antes do fenômeno ocorrer, por exemplo, o lançamento de uma pedra de 5 kg de uma altura de 10 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 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 é experimento aleatório, espaço amostral e eventos. Vamos ver?

6.2 – Definições Prévias

6.2.1 – Experimento Aleatório

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.

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.

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.

6.2.2 – 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) Lançamento de um dado e observação da face voltada para cima: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

b) Retirada de uma carta de um baralho comum (52 cartas) e observação do naipe: $\Omega =$ {copas, espadas, ouros, paus}.

c) O número de mensagens que são transmitidas corretamente por dia em uma rede de computadores: $\Omega = \{0, 1, 2, 3, ...\}$. LINK Note que não há um limite superior conhecido, mas somente é possível a ocorrência de valores inteiros. LINK

d) A observação do diâmetro, em mm, de um eixo produzido em uma metalúrgica: $\Omega = \{D$, tal que $D > 0\}$. LINK Não há um limite superior e, teoricamente, pode haver uma infinidade de valores. LINK

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

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.

6.2.3 – 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, LINK Embora nem todos os autores concordem com esta abordagem, ela auxilia bastante na compreensão dos conceitos. LINK que veremos a seguir.

Seja 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.

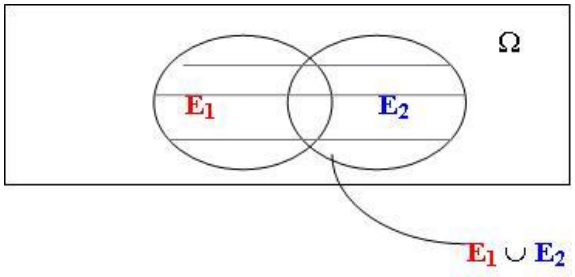


Figura 54 - Evento União

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.

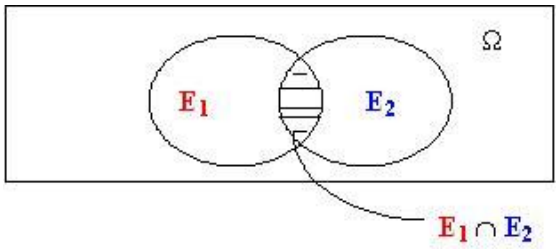


Figura 55 - Evento intersecção

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).

Dentre 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 é o conjunto vazio.

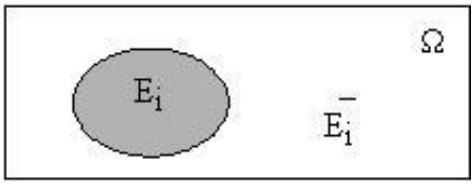


Figura 56 - Evento Complementar

Fonte: elaborada pelo autor

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

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

$$\mathbf{E}_2 = \{3, 4, 5, 6\} \quad \bar{\mathbf{E}}_2 = \{1, 2\}$$

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.

6.3 – Definições de Probabilidade

Por que usamos plural, definições ao invés de definição? Porque ao longo dos séculos várias definições de probabilidade foram apresentadas, sendo que 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.

6.3.1 – 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*. LINK 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. LINK Qual é a chance de você ganhar?

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. A 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 têm 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 do 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.

c) Face menor ou igual a 2.

O Espaço Amostral deste experimento será: Ω = {1, 2, 3, 4, 5, 6}. Sendo assim há um total de 6 resultados possíveis, resultando em **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 ou a retirada da carta do baralho não eram mais experimentos aleatórios.

O segundo problema decorre da pergunta: será que em todos os experimentos aleatórios todos os eventos terão a mesma probabilidade de ocorrer? Será que a probabilidade de chover no mês de novembro na cidade de Brest (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**.

6.3.2 – Definição experimental de probabilidade

Seja um experimento aleatório que é repetido **n** vezes, e $\mathbf{E_i}$ um evento associado.

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

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 frequência relativa. Lembre-se da Unidade 3, a descrição de um fenômeno pode ser feita por distribuição de frequências.

Quando não há outra maneira de 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 frequências relativas estimem as probabilidades, para que seja possível construir um modelo probabilístico para o experimento. Isso pode ser feito em laboratório, em condições controladas, por exemplo, a vida útil das lâmpadas vendidas no comércio é definida 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 frequências relativas históricas.

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

6.3.3 – Axiomas e Propriedades de Probabilidade

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

Seja um experimento aleatório e um espaço amostral Ω 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) $0 \leq$ **P(Ei)** $\leq 1{,}0$

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

b) **P (Ω)** = 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, ..., E_n são eventos mutuamente exclusivos, então

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

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

Além dos axiomas há algumas propriedades básicas da Probabilidade:

a) **P (∅)** = 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**. GLOSSÁRIO Evento Impossível: evento com probabilidade de ocorrer igual a 0%, é o conjunto vazio. Fonte: Barbetta, Reis, e Bornia, 2008. Fim GLOSSÁRIO.

b) Σ **P(E_i)** = 1,0

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

c) **P(Ei)** = 1 - **P($\bar{E}$i)**

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

d) Sejam **E_i** e **E_j** dois eventos quaisquer: P (**E_i** ∪ **E_j**) = **P(E_i)** + **P(E_j)** - P(**E_i** ∩ **E_j**)

A probabilidade de ocorrência do evento União de dois outros eventos será igual a soma das probabilidades de cada evento menos a probabilidade de ocorrência do evento Intersecção dos mesmos dois eventos. 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 $\mathbf{E_3}$.

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

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

Usando as propriedades:

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

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

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

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

$\mathbf{P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)}$ 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 tal evento 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 á ocorrência de outro, somente poderá ser resolvido por **Probabilidade Condicional**, que veremos a seguir.

6.4 – 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:

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

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

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

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

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**, 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. Fim GLOSSÁRIO implicando em:

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

Seja o lançamento de 2 dados não viciados, um após o outro, e a observação das faces voltadas para cima. Neste quinto exemplo iremos calcular as probabilidades:

a) de que as faces sejam iguais supondo-se que sua soma é menor ou igual a 5.

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

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

Como 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 57 - 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:

a) "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:

$P(E_1 \mid E_2) = \dfrac{P(E_1 \cap E_2)}{P(E_2)}$ 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, $\mathbf{n_{E1}} = 6$.

$\mathbf{E_2}$ = **{**(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, $\mathbf{n_{E2}} = 10$.

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, $\mathbf{n_{E1 \cap E2}}$ = 2.

$P(E_2) = n_{E2} / n = 10/36$ **$P(E_1 \cap E_2) = n_{E1 \cap E2} / n = 2/36$**

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 são 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 10 resultados, novo **n** = 10.

O número de resultados do evento faces iguais ($\mathbf{E_1}$) no novo Espaço Amostral é igual a 2, novo $\mathbf{n_{E1}}$ = 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 $\mathbf{E_2}$, $P(\mathbf{E_1} \mid \mathbf{E_2})$, será:

$P(\mathbf{E_1} \mid \mathbf{E_2})$ = novo $\mathbf{n_{E1}}$/ 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; LINK Houve uma mudança no evento que ocorreu previamente. LINK 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 $P(\mathbf{E_2} \mid \mathbf{E_1})$, probabilidade de ocorrência de $\mathbf{E_2}$ condicionada à ocorrência PRÉVIA de $\mathbf{E_1}$.

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

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

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 se outra forma. Se as faces são iguais, o evento **E_1** já ocorreu previamente, então o Espaço Amostral modificou-se, passando a ser o conjunto de resultados do evento **E_1**:

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

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

P(**E_2**| **E_1**) = novo **n_{E2}**/ novo **n** = 2/6 = 0,33 (33%) o mesmo resultado obtido anteriormente.

DESTAQUE **É extremamente importante lembrar que, conceitualmente P(A|B) ≠ P(B|A), pois os eventos que ocorreram previamente são diferentes.** DESTAQUE

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

6.4.1 – Regra do Produto

Uma das consequências 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.

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

Neste caso o evento A ocorreu previamente, e o segundo valor é a probabilidade de ocorrência de B dado que A ocorreu. LINK Não se esqueça de que a intersecção é comutativa. LINK

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

No exemplo seis, digamos que uma urna contém 2 bolas brancas e 3 vermelhas. Retiram-se 2 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 2 bolas retiradas sejam da mesma cor?

b) Qual é a probabilidade de que as 2 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 2 cores e 2 retiradas, então podemos ter:

- a 1ª e a 2ª bolas brancas (2 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 (2 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 retirada forem feitas sem reposição elas serão dependentes, pois o Espaço Amostral será modificado: a cada retirada, as probabilidades de ocorrência são modificadas porque as bolas não são repostas.

- a probabilidade de retirar bola branca na 1ª retirada é de 2/5(2 bolas brancas no total de 5), $P(B_1) = 2/5$;

- a probabilidade de retirar bola vermelha na 1ª retirada é de 3/5 (3 bolas vermelhas em 5), $P(V_1) = 3/5$.

Se a primeira bola retirada foi branca (o evento B_1 ocorreu previamente), restaram 4 bolas, 1 branca e 3 vermelhas:

- a probabilidade de retirar uma bola branca na 2ª retirada se na 1ª foi extraída uma branca é de 1/4 (1 bola branca em 4) LINK Repare que o número de bolas, número de resultados, diminuiu de 5 para 4 porque as retiradas são feitas sem reposição. LINK, $P(B_2| B_1) = 1/4$.

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

Se a primeira bola retirada foi vermelha (o evento V_1 ocorreu previamente), restaram 4 bolas, 2 brancas e 2 vermelhas:

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

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

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

Chamando bolas da mesma cor de evento F: $F = [(B_1 \cap B_2) \cup (V_1 \cap V_2)]$

Usando as propriedades de probabilidade:

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

Os eventos $(B_1 \cap B_2)$ e $(V_1 \cap V_2)$ são mutuamente exclusivos, se as bolas são da mesma cor ou são brancas ou são vermelhas, então a intersecção entre eles é o conjunto vazio, e a probabilidade do conjunto vazio ocorrer é igual a zero (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:

$P(B_1 \cap B_2) = P(B_1) \times P(B_2| B_1) = (2/5) \times (1/4) = 2/20 = 1/10$

$P(V_1 \cap V_2) = P(V_1) \times P(V_2| V_1) = (3/5) \times (2/4) = 6/20 = 3/10$

Substituindo na expressão:

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

Então, se as retiradas forem feitas sem reposição a probabilidade de que as 2 bolas sejam da mesma cor será igual a 0,4 (40%).

b) - Neste caso sabe-se que as 2 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:

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

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 2 bolas vermelhas e 2 bolas da mesma cor? <u>O próprio evento 2 bolas vermelhas</u> $(V_1 \cap V_2)$, 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:

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

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

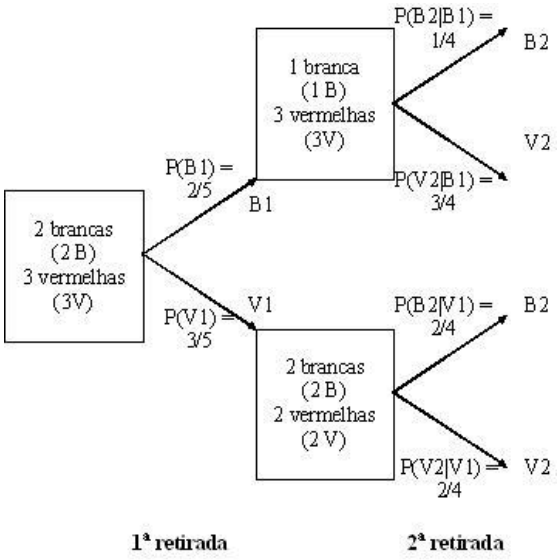


Figura 58 - Árvore de Probabilidades - Retiradas sem reposição

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: 1 bola branca e 3 vermelhas se na 1ª retirada obteve-se uma bola branca, ou 2 bolas brancas e 2 vermelhas se na 1ª retirada obteve-se uma bola vermelha.

A partir dos novos Espaços Amostrais é possível calcular as probabilidades condicionais para cada caso, e depois substituí-las nas fórmulas adequadas. Contudo, a árvore será inútil se o evento para o qual se deseja calcular a probabilidade não for definido adequadamente: 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 $\{(V_1 \cap V_2) \mid \mathbf{[(B_1 \cap B_2) \cup (V_1 \cap V_2)]}\}$.

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?

6.5 – 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:

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

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

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

DESTAQUE As expressões acima são válidas se os eventos **A** e **B** forem independentes. DESTAQUE

Em situações práticas dois eventos são independentes quando a ocorrência de um deles não modifica, ou modifica muito pouco, o Espaço Amostral do Experimento Aleatório. É o que ocorria na Unidade 2 quando fazíamos amostragem aleatória simples: naquele momento não foi dito que a amostragem era com reposição, que dificilmente é feita

na prática, mas admite-se que sendo o tamanho da população muito grande, a retirada de uma pequena amostra não modificará muito as proporções dos eventos.

Exemplo 7 resolva o Exemplo 6, mas agora supondo que as retiradas foram feitas **com reposição**.

a) Qual é a probabilidade de que as 2 bolas retiradas sejam da mesma cor? R.: 0,52(52%)

b) Qual é a probabilidade de que as 2 bolas retiradas sejam vermelhas, supondo-se que são da mesma cor? R.: 0,69 (69%).

Tô afim de saber:

- Sobre conceitos básicos de Probabilidade, BARBETTA,P. A. Estatística Aplicada às Ciências Sociais. 7ª. ed. – Florianópolis: Ed. da UFSC, 2007, capítulo 7.
- Também sobre conceitos básico de Probabilidade STEVENSON, Willian J. Estatística Aplicada à Administração. São Paulo: Ed. Harbra, 2001, capítulo 3.
- LOPES, P. A. Probabilidades e Estatística. Rio de Janeiro: Reichmann e Affonso Editores, 1999, capítulo 3.

Atividades de aprendizagem

1) Numa eleição para a prefeitura de uma cidade, 30% dos eleitores pretendem votar no candidato A, 50% no candidato B e 20% em branco ou nulo. Sorteia-se um eleitor na cidade e verifica-se o candidato de sua preferência.

a) Construa um modelo probabilístico para o problema.

b) Qual é a probabilidade de o eleitor sorteado votar em um dos dois candidatos? (R.: 0,8)

Adaptado de BARBETTA, P. A. Estatística Aplicada às Ciências Sociais. 7ª ed. Florianópolis: Ed. da UFSC, 2007.

2) Extraem-se ao acaso duas cartas de um baralho de 52 cartas, uma após a outra SEM reposição. Calcule as seguintes probabilidades:

a) Ambas as cartas são vermelhas. (R.: 0,245) b) Ambas as cartas são de paus. (R.: 0,058)

c) Ambas as cartas são de “Figuras” (ás, rei, dama ou valete). (R.: 0,0905)
d) Uma carta de paus e outra de copas. (R.: 0,1274)
Adaptado de STEVENSON, W.J. Estatística Aplicada à Administração, São Paulo: Harper do Brasil, 1981, página 76.

3) Repita o exercício 2 supondo que as retiradas fossem feitas COM reposição.
Adaptado de STEVENSON, W.J. Estatística Aplicada à Administração, São Paulo: Harper do Brasil, 1981, página 75.

a) R.: 0,25 b) R.: 0,0625 c) R.: 0,0947 d) 0,125

4) Para um determinado telefone a probabilidade de se conseguir linha é de 0,75 em dias normais e 0,25 em dias de chuva. A probabilidade de chover em um dia é 0,1. Além disso tendo-se conseguido linha, a probabilidade de que um número esteja ocupado é 11/21.
a) Qual é a probabilidade de que um telefone tenha sua ligação completada? (R.: 0,333)
b) Dado que um telefonema foi completado, qual é a probabilidade de estar chovendo? (R.: 0,0357)

Resumo

O resumo desta Unidade está mostrado na Figura 59:

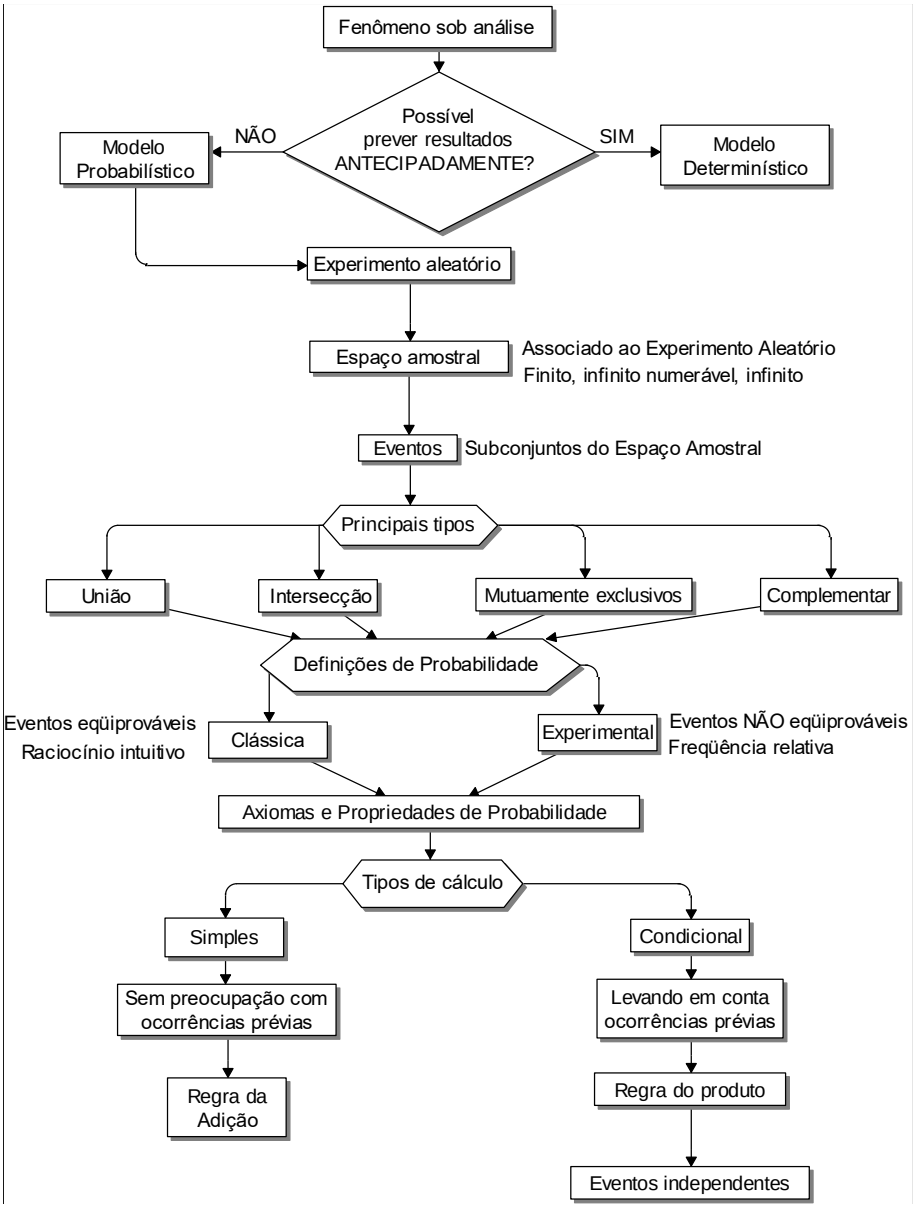


Figura 59 - Resumo da Unidade 6

Fonte: elaborado pelo autor

Chegamos ao final de Unidade 6. Esperamos que você tenha aprendido todos os conceitos trabalhados e com os exemplos propostos, tenha colocado em pratica as informações adquiridas. Neles propomos que você reconhecesse os modelos probabilísticos, modelos determinísticos, principais tipos de eventos, e os diferentes tipos de cálculos. Na disciplina de Estatística Aplicada à Administração vamos prosseguir, aprendendo o conceito de variável aleatória, e alguns dos modelos probabilísticos mais empregados. Tudo isso para aplicar 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 auxilio e caso necessário releia o material. Realize a atividade de aprendizagem e entenda todo o processo amplamente.

Ótimos estudos!

Chegamos ao final da disciplina de Estatística Aplicada à Administração I. Estudamos nessa última os conceitos básicos de probabilidade, que serão imprescindíveis na disciplina de Estatística Aplicada à Administração II. 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**. 2ª ed. – São Paulo: Thomson Learning, 2007

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010.

BARBETTA,P. A. **Estatística Aplicada às Ciências Sociais**. 9ª. ed. – Florianópolis: Ed. da UFSC, 2014.

COSTA NETO, P.L. da O. Estatística. 2ª ed, São Paulo: Edgard Blücher, 2002.

LOPES, P. A. **Probabilidades e Estatística**. Rio de Janeiro: Reichmann e Affonso Editores, 1999.

MARCONI, Marina de Andrade, LAKATOS, Eva Maria. Técnicas de Pesquisa - 5a ed. São Paulo: Atlas, 2003.

MONTGOMERY, D. C. Introdução ao Controle Estatístico da Qualidade. 4.ed. Rio de Janeiro: LTC, 2004.

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.

STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. 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.

Minicurrículo e foto do autor

Minicurrículo:

MARCELO MENEZES REIS é 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 Associado, 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 os de Administração.

Foto: