Métodos Quantitativos Estatísticos

Paulo Ricardo Bittencourt Guimarães

1.ª edição

XXX Guimarães, Paulo Ricardo Bittencourt.

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

245 p.

ISBN: XXX-XX-XXXX-XXX-X

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

CDD XXX.XXXX

Inteligência Educacional e Sistemas de Ensino

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

Paulo Ricardo Bittencourt Guimarães

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

Sumário

sumário

Apresentação

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

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

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

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

Conceitos e Aplicações

Introdução

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

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

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

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

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

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

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

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

Conceitos básicos

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

Estatística Descritiva

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

Inferência Estatística

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

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

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

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

População

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

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

Exemplos:

- Estudo da inadimplência dos clientes do banco X no Brasil

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

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

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

Amostra aleatória

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

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

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

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

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

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

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

Veja os exemplos:

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

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

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

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

Técnicas de Amostragem

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

A amostragem será probabilística se todos os elementos da população tiverem probabilidade conhecida, e diferente de zero, de pertencer à amostra. Caso contrário, a amostragem será não-probabilística. Uma amostragem não-probabilística é obtida quando o acesso a informações não é tão simples ou os recursos forem limitados, assim o pesquisador faz uso de dados que estão mais a seu alcance, é a chamada amostragem por conveniência.

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

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

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

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

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

Amostragem aleatória simples

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

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

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

Amostragem sistemática

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

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

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

Amostragem estratificada

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

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

Amostragem por conglomerados

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

Amostragem de conveniência (não-probabilística)

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

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

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

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

Exemplos de planos amostrais:

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

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

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

Tipos de variáveis

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

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

Exemplo 1: (variável qualitativa nominal)

População: moradores de uma cidade.

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

Exemplo 2: (variável qualitativa ordinal)

População: moradores de um condomínio.

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

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

Exemplo 3: (variável quantitativa discreta)

População: hospitais de uma determinada cidade.

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

Exemplo 4: (variável quantitativa contínua)

População: moradores de uma determinada cidade.

Variável: estatura dos indivíduos.

Ampliando seus conhecimentos

(MATTAR, 2001)

Pesquisa de mercado

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

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

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

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

Etapas de uma pesquisa

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

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

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

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

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

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

Atividades de aplicação

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

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

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

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

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

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

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

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

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

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

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

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

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

Análise Exploratória de Dados

Introdução

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

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

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

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

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

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

Tabela 1: Principais técnicas de estatística descritiva

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

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

Tabelas

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

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

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

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

Exemplo:

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

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

Fonte: IBGE, 1988.

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

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

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

Observação:

Nas casas ou células devemos colocar:

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

Tabelas de contingência

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

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

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

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

Fonte: IBGE, 1988.

Tabelas de distribuição de freqüências

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

Como exemplo, considere os dados da tabela abaixo:

Tabela 4: Rendimento mensal de fundos de investimento

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gráficos

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

Os principais tipos de gráficos usados na representação estatística são:

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

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

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

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

Descrição gráfica das variáveis qualitativas

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

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

Tabela 5: Distribuição de fundos relativos às regiões do Brasil

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

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

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

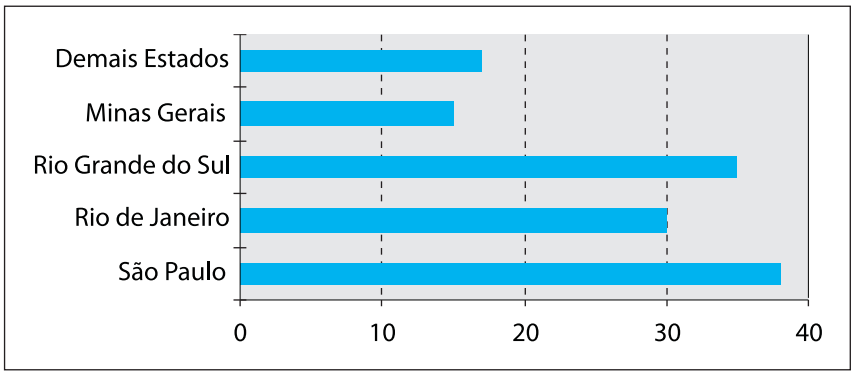


Figura 1: Gráfico de barras

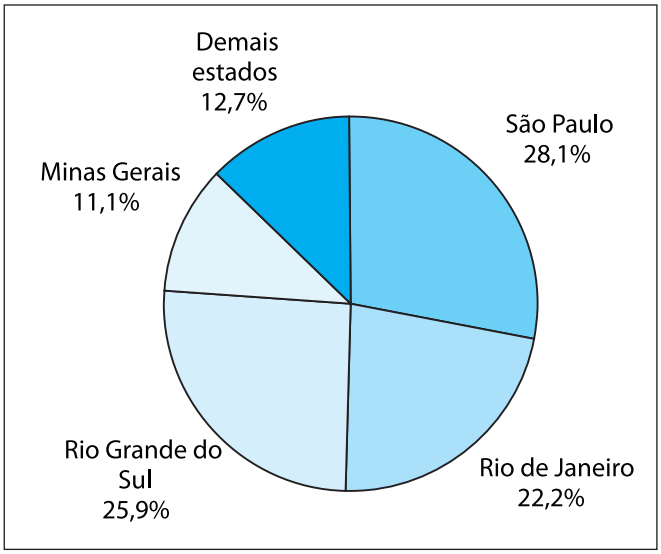


Figura 2: Gráfico de setores

Descrição gráfica das variáveis quantitativas discretas

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

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

Sejam os seguintes valores obtidos:

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

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

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

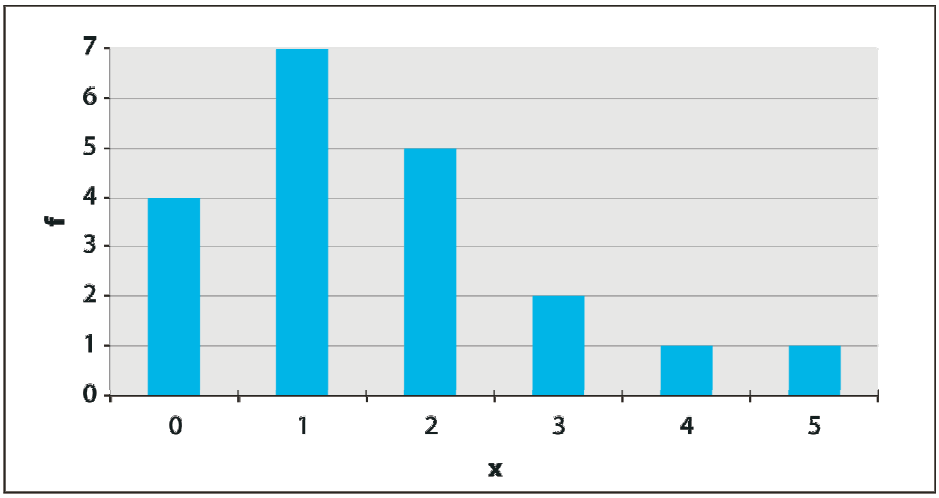


Figura 3: Gráfico de barras

Descrição gráfica das variáveis quantitativas contínuas – classes de freqüências

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

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

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

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

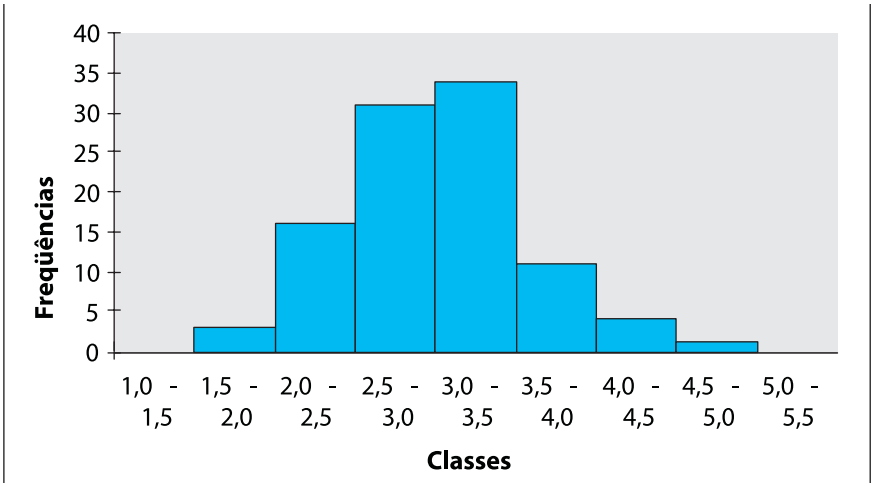


Figura 4: Histograma

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

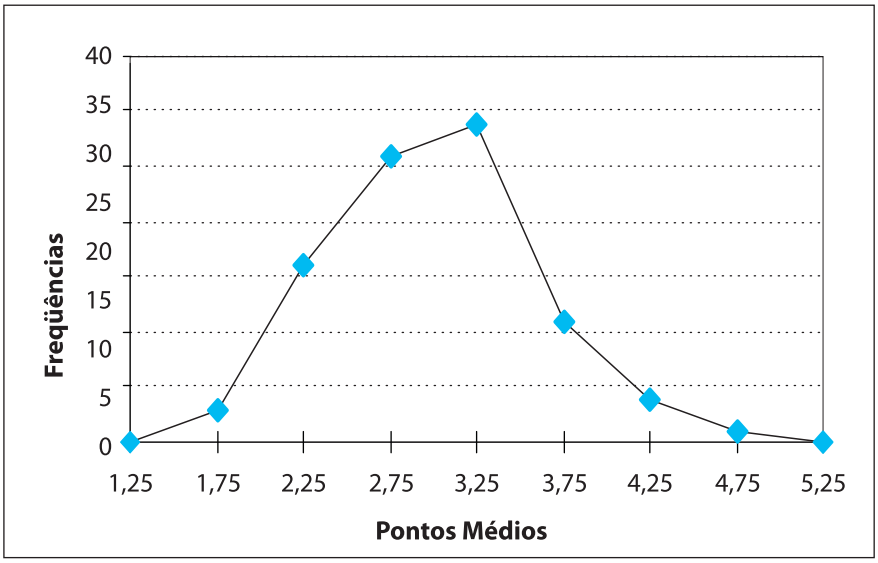


Figura 5: Polígono de freqütências

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

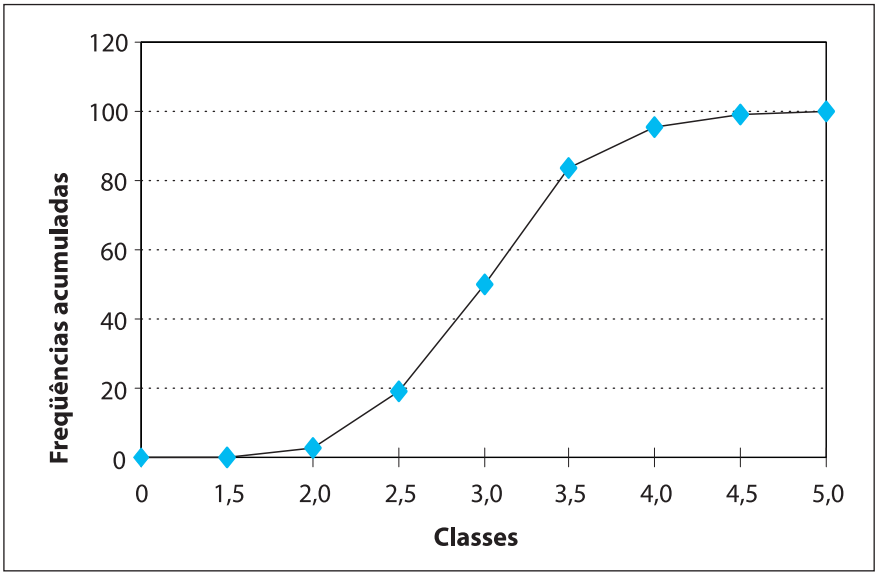


Figura 6: Ogiva de Galton Crescente

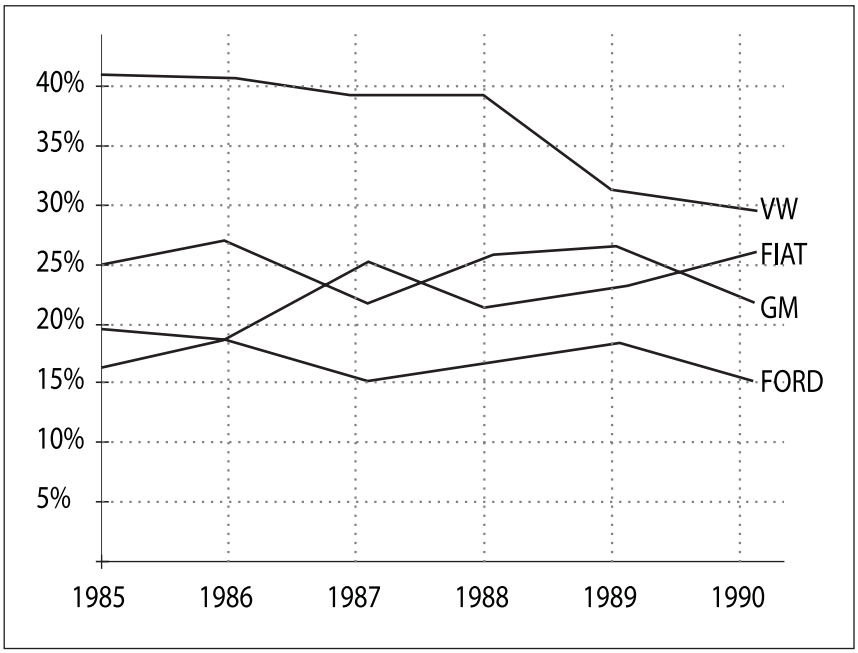


Figura 7: Gráfico de linhas

Ramo-e-folhas

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

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

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

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

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

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

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

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

Figura 9. Diagrama de ramos e folhas

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

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

Ampliando seus conhecimentos

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

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

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

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

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

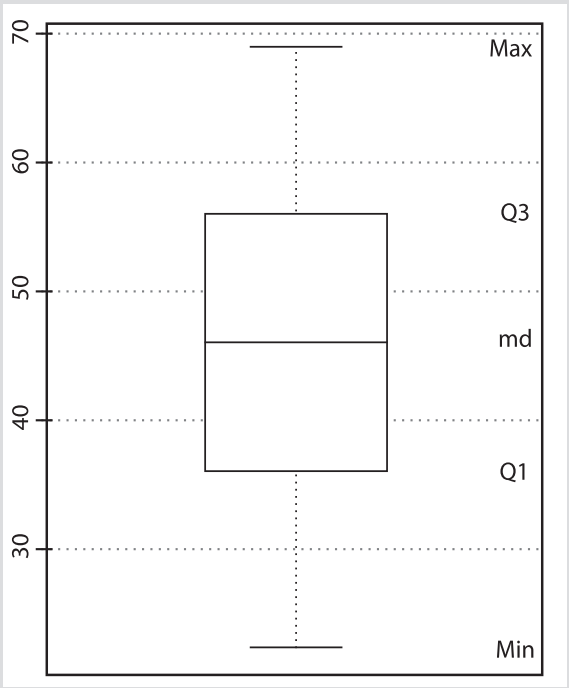


Figura 1: *Box-plot*

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

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

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

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

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

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

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

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

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

  em que $d_F = Q_3 - Q_1$

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

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

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

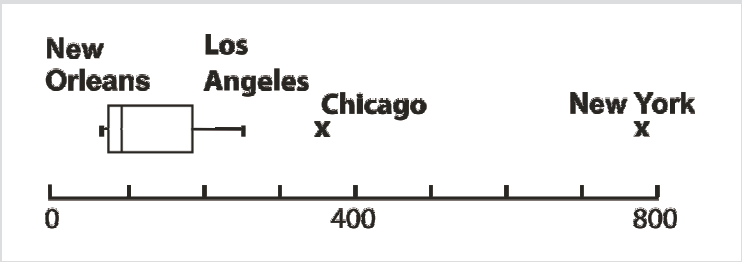


Figura 2: *Box-plot* – População das principais capitais (1960)

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

Atividades de aplicação

Resolva as questões abaixo utilizando as definições vistas neste capítulo.

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

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

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

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

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

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

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

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

a) O limite superior da 2ª classe.

b) O limite inferior da 5ª classe.

c) A amplitude do intervalo da 3ª classe.

d) A amplitude total.

e) O ponto médio da 4ª classe.

f) A freqüência da 1ª classe.

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

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

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

j) A freqüência percentual da última classe.

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

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

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

a) Distribuição das alturas dos brasileiros adultos.

b) Distribuição das alturas dos suecos adultos.

c) Distribuição das alturas dos japoneses adultos.

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

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

Que tipo de informação revelam esses dados ?

Medidas de Posição e Variabilidade

Introdução

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

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

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

Medidas de Posição ou de Tendência Central

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

As principais medidas de tendência central são:

- média aritmética;
- mediana;
- moda.

Média Aritmética ($\overline{X}$)

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

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

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

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

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

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

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

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

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

a média aritmética será:

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

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

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

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

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

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

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

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

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

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

Mediana (Md)

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

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

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

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

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

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

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

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

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

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

24, 37, 41, 65, 68 e 82

A mediana será, por convenção:

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

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

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

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

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

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

Exemplo: Se na graduatória

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

em lugar de 82 houvesse 1000 casos, isto é,

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

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

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

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

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

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

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

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

Moda (Mo)

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

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

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

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

essas 8 medidas não definem uma moda.

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

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

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

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

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

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

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

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

$L = 30$

$t = 10$

$f_1 = 2$

$f_2 = 10$

e, portanto,

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

= 31 anos e 8 meses = 31 anos completos.

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

Medidas de Dispersão

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

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

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

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

Amplitude de Variação (R)

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

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

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

Por exemplo, nos dois conjuntos de valores:

4, 6, 6, 6, 8

4, 5, 6, 7, 8

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

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

Amplitude Semiquartil ou Desvio Quartil

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

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

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

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

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

Desvio Padrão e Variância

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

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

Se, por exemplo, as observações tivessem os valores:

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

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

Tabela 3: Diferenças entre as observações e a respectiva média

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

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

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

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

No primeiro caso ter-se-ia:

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

e a medida de variabilidade seria

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

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

No segundo caso, ter-se-ia:

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

e a medida de variabilidade seria

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Considerações finais sobre o desvio-padrão:

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

Coeficiente de Variação

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

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

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

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

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

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

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

Ampliando seus conhecimentos

(MATTAR, 1996)

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

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

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

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

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

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

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

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

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

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

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

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

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

Atividades de aplicação

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

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

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

2. A média aritmética é a razão entre:

a) o número de valores e o somatório deles.

b) o somatório dos valores e o número deles.

c) os valores extremos.

d) os dois valores centrais.

e) nenhuma das alternativas anteriores.

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

a) 50

b) 60

c) 66

d) 90

e) nenhuma das anteriores.

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

a) a moda.

b) a média.

c) a mediana.

d) o elemento mediano.

e) nenhuma das anteriores.

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

a) positiva.

b) negativa.

c) zero.

d) diferente de zero.

e) nenhuma das alternativas anteriores.

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

a) a média e a moda.

b) a média e a mediana.

c) a mediana e a moda.

d) a média, a mediana e a moda.

e) nenhuma das alternativas anteriores.

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

a) moda.

b) média.

c) mediana.

d) qualquer das anteriores.

e) nenhuma das anteriores.

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

a) desvio padrão e média.

b) média e desvio padrão.

c) mediana e amplitude interquartílica.

d) desvio padrão e moda.

e) nenhuma das alternativas anteriores.

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

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

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

Com esses resultados podemos afirmar:

a) a turma 2 apresentou maior dispersão absoluta.

b) a dispersão relativa é igual à dispersão absoluta.

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

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

e) nenhuma das alternativas anteriores.

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

a) R$ 1.050,00

b) R$ 1.012,50

c) R$ 405,00

d) R$ 245,00

e) nenhuma das alternativas anteriores.

11. O cálculo da variância supõe o conhecimento da:

a) média.

b) mediana.

c) moda.

d) ponto médio.

e) desvio padrão.

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

a) negativo.

b) positivo.

c) a unidade.

d) zero.

e) nenhuma das alternativas anteriores.

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

a) ao desvio padrão de X.

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

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

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

e) nenhuma das alternativas anteriores.

Introdução à Probabilidade

Introdução

O termo probabilidade é usado de modo muito amplo, em nosso cotidiano para sugerir um certo grau de incerteza sobre o que ocorreu no passado, o que ocorrerá no futuro ou o que está ocorrendo no presente.

A idéia de probabilidade desempenha papel importante em muitas situações que envolvem uma tomada de decisão. Suponhamos que um empresário deseja lançar um novo produto no mercado. Ele precisará de informações sobre a "probabilidade" de sucesso para seu novo produto. Os modelos probabilísticos podem ser úteis em diversas áreas do conhecimento humano, tais como: Administração de empresas, Economia, Psicologia, Biologia e outros ramos da ciência.

Probabilidade é uma coleção ampla de conceitos que trata dos estudos de *experimentos aleatórios* ou *não-determinísticos*. Probabilidade pode significar também, um número num intervalo de 0 a 1, o qual fornece um significado ao avaliar a ocorrência de um resultado num experimento.

Em resumo, probabilidade é responsável pelos estudos do comportamento dos fenômenos aleatórios.

Conceitos iniciais de probabilidade

Experimento Aleatório (E)

Define-se por experimento qualquer processo de observação. Um experimento é dito aleatório quando seus resultados estão sujeitos unicamente ao acaso. Quando o experimento é executado repetidas vezes, os resultados surgirão seguindo uma configuração definida ou regularidade. É essa regularidade que torna possível construir um modelo matemático preciso com o qual se analisará o processo.

Exemplos:

E_1 : Em uma linha de produção, fabrique peças em série e conte o número de peças defeituosas produzidas em um período de 24 horas.

E_2 : Uma asa de avião é fixada por um grande número de rebites. Conte o número de rebites defeituosos.

E_3 : Uma lâmpada é fabricada. Em seguida é ensaiada quanto à duração da vida, pela colocação em um soquete e anotação do tempo decorrido (em horas) até queimar.

E_4 : A resistência à tração de uma barra metálica é medida.

O que os experimentos acima têm em comum? Os seguintes traços são pertinentes à caracterização de um experimento aleatório:

- cada experimento poderá ser repetido indefinidamente sob condições essencialmente inalteradas;
- muito embora não sejamos capazes de afirmar que um resultado particular ocorrerá, seremos capazes de descrever o conjunto de todos os possíveis resultados do experimento;
- quando o experimento for repetido um grande número de vezes, uma configuração definida ou regularidade surgirá.

Espaço Amostral (S)

Para cada experimento aleatório E, define-se o *espaço amostral* como o conjunto formado por todos os resultados possíveis do experimento aleatório E.

Exemplos:

Vamos considerar cada um dos experimentos acima e descrever um espaço amostral para cada um deles. O espaço amostral S_i se referirá ao experimento E_i.

$S_1 = \{ 0, 1, 2, ..., N \}$, onde N é o número máximo que pode ser produzido em 24 horas.

$S_2 = \{ 0, 1, 2, ..., M \}$, onde M é o número de rebites empregados.

$S_3 = \{ t / t \geq 0 \}$

$S_4 = \{ T / T \geq 0 \}$

Observação: Os elementos de S são chamados de pontos amostrais e, são denotados por $w_1, w_2, ... \in S$.

Evento Aleatório

Evento aleatório (relativo a um particular espaço amostral S, associado a um experimento E) é simplesmente um conjunto (combinações) de resultados possíveis.

Na terminologia dos conjuntos, um evento é um subconjunto do espaço amostral S.

Dizemos que um determinado evento A ocorre se ocorrer um de seus resultados.

Exemplo: Novamente, referimo-nos aos experimentos relacionados anteriormente: A_i se referirá ao evento associado ao experimento E_i.

A_1: "todas a peças são perfeitas", isto é, { 0 }

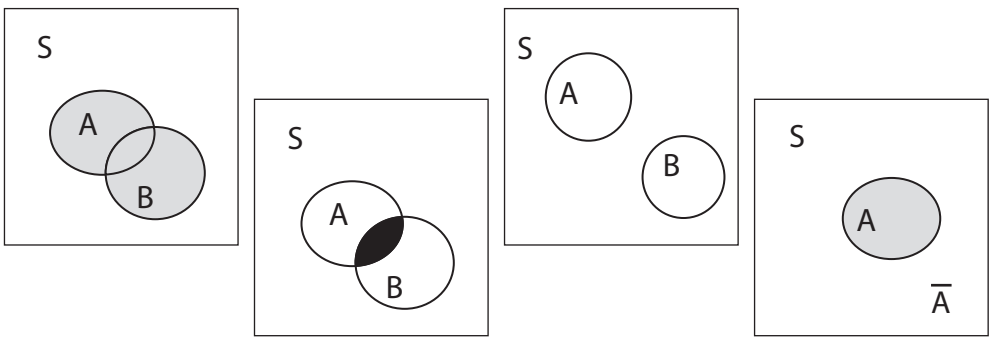
A_2: "mais do que dois rebites eram defeituosos", isto é, { 3, 4, 5, ..., M}

A_3: "a lâmpada queima em menos de 3 horas", isto é, {t / t < 3}

Operações com eventos

Estas operações podem ser graficamente representadas pelo *diagrama de Venn* por meio da definição da região sombreada.

Como evento é um conjunto, poderemos realizar com elas as operações costumeiras de união e interseção de conjuntos. Assim:



1º diagrama: União: $A \cup B$

$A \cup B$ é o evento que ocorre se A ocorrer ou B ocorrer ou ambos ocorrerem. É a união de todos os elementos que pertencem a A, pertencem a B *ou* a ambos os conjuntos.

2º diagrama: Interseção: $A \cap B$

$A \cap B$ é o evento que ocorre se A e B ocorrerem. $A \cap B$ corresponde à área escura do 2º diagrama de Venn, ou seja, é um novo conjunto formado por todos os elementos que pertencem a A *e* pertencem a B.

3º diagrama: Exclusão: $A \cap B = \varnothing$

Eventos mutuamente exclusivos: Dois eventos A e B são denominados mutuamente exclusivos se eles não puderem ocorrer simultaneamente, isto é, A interseção B = conjunto vazio. A e B são mutuamente exclusivos, pois a ocorrência de A impede a ocorrência de B e vice-versa: $A \cap B = \varnothing$ (evento impossível).

4º diagrama: Negação ou evento complementar

A negação do evento A, denotada por A^c ou $\overline{A}$ (lê-se A complementar ou A traço) é o evento que ocorre se A não ocorrer. Corresponde à área em branco do 4º diagrama.

Exemplo:

1) Seja E o experimento "sortear um cartão dentre dez cartões numerados de 1 a 10". Sejam os eventos $\overline{A}$ = {sair o número 7} e B = {sair um número par}, então, se S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, teremos: A = {7} e B = { 2, 4, 6, 8, 10}.

$A \cup B$ = {7, 2, 4, 6, 8, 10}; $A \cap B = \varnothing$ (evento impossível)

O complementar de A será: $\overline{A}$ = {1, 2, 3, 4, 5, 6, 8, 9, 10};

O complementar de B será: $\overline{B}$ = { 1, 3, 5, 7, 9}

$A \cup \overline{A} = S$; $A \cap \overline{A} = \varnothing$; $B \cup \overline{B} = S$; $B \cap \overline{B} = \varnothing$.

Eventos independentes

Dois eventos são considerados independentes quando a ocorrência de um deles não depende ou não está vinculada com a ocorrência do outro, isto é, P(A/B) = P(A) e P(B/A) = P(B).

Logo, a regra do produto para dois eventos independentes é dada por:

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

Exemplo: Aplicação da regra do produto.

1) Retira-se, com reposição, duas cartas de um baralho com 52 cartas. Qual a probabilidade de que ambas sejam de "paus"?

Solução: Sejam os eventos:

A = {a primeira carta é de "paus"}

B = {a segunda carta é de "paus"}

Como A e B são independentes, a ocorrência de um deles não está vinculada à ocorrência do outro.

Observem que, como o processo é com reposição, o espaço amostral não é alterado para o cálculo da probabilidade do outro evento. Assim:

$P(A \cap B) = P(A).\ P(B) = 13/52\ .\ 13/52\ =\ 1/16\ = 0{,}0625$ ➔ 6,25%

Definições de Probabilidades e Propriedades

Definição frequentista

Repetindo-se *n* vezes o experimento aleatório E, o evento A ocorrerá um certo número *m* de vezes; *m* é a freqüência com que o evento A ocorre e $\frac{m}{n}$ é a freqüência relativa de ocorrência de A.

Chama-se de probabilidade de ocorrência do evento A, e denota-se por P(A), o valor limite da freqüência relativa para uma seqüência muito grande de realizações do experimento ($n \to \infty$), ou seja,

$$P(A) = \lim_{n \to \infty} \left(\frac{m}{n} \right)$$

Suponha, como exemplo, que uma locadora de automóveis queira estimar a probabilidade de ocorrerem acidentes com a sua frota de veículos. Para isso, verifica quantos acidentes ocorreram em determinadas vezes que os automóveis da frota foram locados. Pode ser que se *n* (número de locações) for igual a 10, a probabilidade de ocorrerem acidentes não represente

fielmente a realidade. No entanto, se for observado um número maior de locações (1 000, por exemplo), aos poucos surge uma estimativa da probabilidade de ocorrerem acidentes cada vez mais próxima da realidade.

Definição clássica

Seja E um experimento aleatório e S o espaço amostral associado a E. Suponha que S seja finito e que todos os resultados de S sejam igualmente prováveis.

Considere, ainda, o evento $A \subset S$. Se n_S e n_A são respectivamente o número de elementos de S e de A, a probabilidade de ocorrência do evento A é um número real definido por:

$$P(A) = \frac{n_A}{n_S}$$

Definição Axiomática

Seja E um experimento e S um espaço amostral associado a E. A cada evento A associaremos um número real representado por P(A) e denominado Probabilidade de A, que satisfaça as seguintes propriedades:

(1) $0 \leq P(A) \leq 1$

(2) $P(S) = 1$

(3) Se A e B forem eventos mutuamente exclusivos, $P(A \cup B) = P(A) + P(B)$

(4) Se $A_1, A_2, ..., A_n, ...$ forem, dois a dois, eventos mutuamente exclusivos, então,

$$P(U_{i=1}^{\infty} A_i) = P(A_1) + P(A_2) + L + \cdots + (A_n) + \cdots$$

Observação: Caso A e B sejam dois eventos quaisquer, então

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

Na verdade, a utilização da definição de Probabilidade e das operações com eventos servem para organizar o raciocínio do Cálculo de Probabilidades, mais ou menos como é feito com um fluxograma.

Agora aproveitaremos as operações de conjuntos descritas anteriormente para o cálculo de probabilidades que envolvem eventos de nosso interesse. Tentemos responder intuitivamente a questão abaixo para depois formalizar o procedimento de cálculo:

a) Para ter a certeza do nascimento de pelo menos um menino, um casal planeja ter 5 bebês. Qual a chance de sucesso?

Respondendo de forma intuitiva, a probabilidade do casal ter pelo menos 1 menino será igual a probabilidade de ter 1, 2, 3, 4 ou 5 meninos que é equivalente ao complementar da probabilidade de não ter nenhum menino, ou seja, 1 – P("5 meninas") = 1 – $(1/2)^5$ = 0,96875 ou 96,875% se presumirmos que a probabilidade de nascimento de meninos e meninas é igual.

b) Peças que saem de uma linha de produção são marcadas defeituosas (D) ou não defeituosas (N). As peças são inspecionadas e sua condição registrada. Isto é feito até que duas peças defeituosas consecutivas sejam fabricadas ou que todas as quatro peças do lote tenham sido inspecionadas, aquilo que ocorrer em primeiro lugar. Calcule a probabilidade do experimento ser interrompido antes do lote inteiro ter sido inspecionado.

Para que o experimento seja interrompido antes do lote inteiro ser inspecionado, devemos observar duas peças defeituosas entre as 3 primeiras peças inspecionadas. Isto pode ocorrer quando as duas primeiras peças inspecionadas forem defeituosas e aí então o experimento é finalizado. Pode ocorrer também que se a 2ª peça defeituosa ocorrer na 3ª peça inspecionada, então entre as duas primeiras inspeções, haverá certamente 1 peça defeituosa. Sendo assim, a probabilidade solicitada seria a soma da probabilidade de 3 situações: P(1ª peça defeituosa e 2ª peça defeituosa) + P(1ª peça defeituosa, 2ª peça perfeita e 3ª peça defeituosa) + P(1ª peça perfeita, 2ª peça defeituosa e 3ª peça defeituosa).

Como se pode observar, das resoluções acima, existe a necessidade de se estruturar, de forma organizada, o raciocínio de cálculo. Para isso, devemos seguir alguns passos:

1. Descrever o espaço amostral e o seu tamanho (n);

2. Definir o evento de interesse no problema (A);

3. Verificar o número de eventos que são favoráveis ao evento de interesse (n_A);

4. Calcular $P(A) = \frac{n_A}{n}$

Mas atenção: Isto só vale se todos os resultados do espaço amostral forem equiprováveis!

Caso os eventos A e B não sejam equiprováveis use:

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

Outros procedimentos de organização são utilizados como: regras de Multiplicação, regras de Adição, Permutações e Arranjos, e Combinações. São os chamados Métodos de Enumeração.

Probabilidade Condicionada

Se A e B são eventos de um espaço amostral S, com P(B) diferente de zero, então a probabilidade condicional do evento A, tendo ocorrido o evento B, é indicada por P(A/B) e definida pela relação:

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

Para o cálculo da probabilidade condicional de A em relação a B, P(A/B), basta contarmos o número de casos favoráveis ao evento $A \cap B$ e dividirmos pelo número de casos favoráveis do evento B:

$$P(A/B) = \frac{N.C.F.a\ A \cap B}{N.C.F.a\ B}$$

Observação: N.C.F. – número de casos favoráveis

Exemplo: Aplicação da regra do produto.

1. Retira-se, sem reposição, duas peças de um lote de 10 peças, onde 4 são boas. Qual a probabilidade de que ambas sejam defeituosas ?

Solução: Sejam os eventos:

A = {a primeira peça ser defeituosa};

B = {a segunda peça ser defeituosa}.

Precisamos, então, avaliar $P(A \cap B)$.

$P(A \cap B) = P(A).P(B/A) \rightarrow P(A \cap B) = 6/10 \,.\, 5/9 = 1/3 = 0{,}3333... \rightarrow 33{,}33\,\%$

Observe que P(B/A) é a probabilidade de a segunda peça ser defeituosa, dado que a primeira foi defeituosa.

2. Uma urna contém 5 bolas brancas e 3 pretas. Duas bolas são retiradas sem reposição. Qual a probabilidade de que:

a) 1ª seja branca e a 2ª seja preta?

$P(B_1 \cap P_2) = P(B_1).P(P_2/B_1) = 5/8 \,.\, 3/7 = 15/56 = 26{,}79\%$

b) as duas sejam brancas?

$P(B_1 \cap B_2) = P(B_1).P(B_2/B_1) = 5/8 \,.\, 4/7 = 20/56 = 35{,}71\%$

c) as duas sejam pretas?

$P(P_1 \cap P_2) = P(P_1).P(P_2/P_1) = 3/8 \,.\, 2/7 = 6/56 = 10{,}71\%$

d) sejam uma de cada cor?

$P(P_1 \cap B_2) + P(B_1 \cap P_2) = (3/8 \,.\, 5/7) + (5/8 \,.\, 3/7) = 30/56 = 53{,}57\%$

e) sejam ambas da mesma cor?

$P(P_1 \cap P_2) + P(B_1 \cap B_2) = (3/8 \,.\, 2/7) + (5/8 \,.\, 4/7) = 26/56 = 46{,}43\%$

Regra de Bayes

Sejam A_1, A_2, A_3, ..., A_n, *n* eventos mutuamente exclusivos tais que $A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n = \mathbf{S}$. Sejam $\mathbf{P(A_i)}$ as probabilidades conhecidas de todos os eventos A_i e B um evento qualquer de **S** tal que conhecemos todas as probabilidades condicionais $\mathbf{P(B/A_i)}$. Então para cada "i" teremos:

$$P(A_i/B) = \frac{P(A_i).P(B/A_i)}{P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + \ldots + P(A_n).P(B/A_n)}$$

O resultado acima é bastante importante, pois, como vimos, relaciona probabilidades *a priori*: $\mathbf{P(A_i)}$ com probabilidades *a posteriori*: $\mathbf{P(A_i/B)}$, probabilidade de ocorrer $\mathbf{A_i}$ depois que ocorrer **B.**

Suponhamos a seguinte configuração:

Cor	Urna 1	Urna 2	Urna 3	Total
Preta	3	4	2	9
Branca	1	3	3	7
Vermelha	5	2	3	10
Total	**9**	**9**	**8**	**26**

Escolheu-se uma urna ao acaso e dela extraiu-se uma bola ao acaso, verificando-se que a bola é branca. Qual a probabilidade de a bola ter vindo da urna 2?

Solução:

Probabilidades *a priori*: $P(U_1) = 1/3$; $P(U_2) = 1/3$; $P(U_3) = 1/3$;

Probabilidades *a posteriori*: $P(br/U_1) = 1/9$; $P(br/U_2) = 1/3$; $P(br/U_3) = 3/8$;

$$P(U_2/br) = \frac{P(U_2).P(br/U_2)}{P(U_1).P(br/U_1) + P(U_2).P(br/U_2)+P(U_3).P(br/U_3)} =$$

$$= \frac{1/3.1/3}{1/3.1/9 + 1/3.1/3+1/3.3/8} = 0{,}4067$$

Variável Aleatória Unidimensional (v. a.)

Na maioria dos experimentos dados até agora, ao descrevermos o espaço aleatório, não especificamos que um resultado individual, necessariamente, seja um número. Por exemplo: ao descrever uma peça manufaturada, podemos usar apenas as categorias "defeituosas" e "não defeituosas". Contudo, em muitas situações experimentais, estaremos interessados na mensuração de alguma coisa e no seu registro como um número. Mesmo no exemplo mencionado, poderemos atribuir um número a cada resultado não numérico do experimento. Por exemplo: podemos atribuir o valor 1 às peças não defeituosas e 0 às peças defeituosas.

Exemplo: Em uma linha de montagem de engrenagens, inspecionam-se 4 peças da produção diária para se controlar a produção de engrenagens defeituosas.

Representando por:

d: engrenagem com defeito e

b: engrenagem perfeita.

Temos o seguinte espaço amostral S para esse experimento:

S = {dddd, dddb, ddbd, dbdd, bddd, ddbb, dbbd, dbdb, bddb, bdbd, bbdd, dbbb, bdbb, bbdb, bbbd, bbbb}

Seja X uma variável aleatória que conta o número de engrenagens com defeito dentre as 4 inspecionadas. Temos então:

$$X = 0, 1, 2, 3, 4$$

Variável Aleatória Discreta e sua função de probabilidade

Uma variável aleatória será discreta se o número de resultados possíveis que ela pode assumir for finito ou infinito enumerável.

Exemplo: Contagem da ocorrência de um fenômeno em um certo número de repetições ou em um certo espaço de tempo.

Seja X uma variável aleatória discreta. A cada possível resultado x_i associaremos um número real $p(x_i) = P(X = x_i)$, denominado de probabilidade de x_i. A função *p* é denominada de *função de probabilidade da variável aleatória discreta* X. Sendo *p* uma função de probabilidade, devemos ter satisfeitas as condições:

(i) $p(x_i) \geq 0$, para todo i

(ii) $\Sigma_i\, p(x_i) = 1$

O conjunto de pares $[x_i, p(x_i)]$ é denominado *distribuição de probabilidade da variável aleatória* X.

Variável Aleatória Contínua e sua função densidade de probabilidade

Uma variável aleatória será contínua se o número de resultados possíveis que ela poderá assumir for infinito não-enumerável, ou seja, se o conjunto de valores que ela pode assumir for um intervalo ou uma reunião de intervalos

Exemplo: Seja X a duração da vida (em horas) de um certo dispositivo eletrônico. Então, o conjunto dos valores que X pode assumir poderá ser representado da seguinte forma: $\{x \in R / x \geq 0\}$, onde R é o conjunto dos números reais.

Seja X uma variável aleatória contínua. Define-se *função densidade de probabilidade* (f.d.p.) como sendo a função *f* que satisfaz às seguintes condições:

(i) $f(x) \geq 0$ para todo $x \in R$

(ii) $\int_{R_x} f(x)\, dx = 1$

A propriedade (ii) indica que a área total limitada pela curva que representa a função f(x) e o eixo das abscissas é igual a 1.

Seja o intervalo $[a, b)$ $x \in R_x$. Então, a probabilidade de um certo valor X pertencer a esse intervalo é dada por:

$$Pr(a \leq X \leq b) = \int_a^b f(x)\,dx,$$

que representa a área sob a curva no gráfico da função densidade de probabilidade, entre $x = a$ e $x = b$. Para isso se usa o recurso da integração.

Algumas variáveis que podem ser consideradas contínuas: salários (em R$), espessura de vigas metálicas (em mm), taxa de colesterol no sangue (em mg/dl). Dessa forma, podemos estar interessados em saber, por exemplo, a probabilidade de alguém receber um salário superior a R$ 10.000,00 ou a probabilidade da espessura da viga estar dentro das especificações ou ainda, a probabilidade da taxa do colesterol estar dentro da normalidade.

Esperança Matemática, Média ou Valor Esperado

É bastante útil descrever uma distribuição de probabilidade em termos de sua média e de sua variância. A média, denotada por E(X), é chamada valor esperado da distribuição de probabilidade. Considere X uma variável aleatória. A esperança matemática, média ou valor esperado de X é a média ponderada de todos os possíveis valores da variável com os respectivos valores de probabilidade tomados como pesos.

Exemplo no caso discreto:

Considere a seguinte variável discreta e sua respectiva função de probabilidade.

x	0	1	2
p(x)	½	¼	¼

Assim, teremos a esperança E(X) = (0.1/2) + (1.1/4) + (2.1/4) = 3/4

Variância

A variância de uma variável aleatória X, denotada por V(X), é calculada como uma medida de dispersão dos dados em relação à média E(X). Pode ser calculada fazendo-se

$$\sigma^2 = Var(X) = E[X - E(X)]^2$$

ou ainda,

$$Var(X) = E(X^2) - [E(X)]^2$$

Considerando os mesmos exemplos vistos acima, teremos:

Variável discreta: $E(X) = 3/4$ e $E(X^2) = (0^2.1/2) + (1^2.1/4) + (2^2.1/4) = 5/4$

$Var(X) = 5/4 - (3/4)^2 = 11/16$

Ampliando seus conhecimentos

Risco e Probabilidade

(Wikipédia)

O que é Risco?

É o resultado objetivo da combinação entre a probabilidade de ocorrência de um determinado evento e o impacto resultante.

O simples fato de existir uma atividade, abre a possibilidade da ocorrência de eventos ou situações cujas conseqüências constituem oportunidades para obter vantagens (lado positivo) ou então ameaças ao sucesso (lado negativo).

O risco pode ser definido como a combinação da probabilidade de um acontecimento e das suas conseqüências.

O que é Análise de Riscos?

Processo pelo qual são relacionados os eventos, os impactos e avaliadas as probabilidades destes se tornarem reais.

Geralmente, se executa uma análise de riscos dentro de organizações que estão planejando ou desenvolvendo projetos específicos ou para negócios (finanças, compra e venda etc). Sendo a abordagem de negócios a mais utilizada.

Como orientação da confecção de uma análise de riscos, temos os seguintes passos e cuidados:

a) **Construir a Matriz de Impacto**

Esta matriz envolve um conjunto de itens que influenciam no dimensionamento do impacto no caso de ocorrência de uma determinada ameaça, sendo, então, relacionados abaixo:

- Determinar os elementos críticos do negócio que poderão ser afetados por falhas e erros no processo;
- Levantar as ameaças/eventos decorrentes da execução dos passos do processo de negócio, que podem afetar ou causar um determinado impacto sobre algum elemento crítico do negócio relacionado;
- Definir o impacto para o negócio no caso de ocorrência das ameaças/eventos relacionadas.

b) Construir a Matriz de Probabilidade

Esta matriz envolve alguns aspectos que influenciam na probabilidade de ocorrência de uma determinada ameaça/evento, sendo, então, relacionados abaixo:

- Levantar os controles ou proteções existentes que poderiam prevenir ou minimizar a ocorrência das ameaças/eventos relacionadas;
- Definir as fraquezas ou fragilidades que possam existir nos controles relacionados, de forma a obter uma avaliação da sua efetividade;
- Definir qual a probabilidade da ameaça/evento vir a se realizar devido a falha do controle (ou este ser sobrepujado) e o impacto previsto acontecer.

c) Definir os Riscos

Esta etapa envolve a sumarização dos impactos relacionados e as suas respectivas probabilidades, de forma a que seja calculado o risco real de um determinado evento (e o seu impacto) vir a ocorrer.

Atividades de aplicação

1. Defina o espaço amostral de cada um dos seguintes experimentos:

a) lançamento simultâneo de três moedas;

b) distribuição de sexo de uma família com três filhos;

c) lançamento simultâneo de dois dados (não viciados);

d) retirada de duas cartas de um baralho com 8 cartas, sendo 4 damas e 4 valetes;

e) retirada de duas bolas sucessivamente, de uma urna com cinco bolas, sendo três brancas e duas amarelas.

2. Dois dados são lançados. Pede-se:

a) enumere o evento A = {a soma dos pontos é 9};

b) enumere o evento B = {a soma dos pontos é 7};

c) calcule a probabilidade do evento A;

d) calcule a probabilidade do evento B;

e) calcule a probabilidade de ocorrer A ou B;

f) calcule a probabilidade de ocorrer A e B;

3. São dadas duas urnas:

Cor	Urna A	Urna B	Total
Preta	2	3	5
Branca	5	12	17
Vermelha	3	5	8
Total	**10**	**20**	**30**

a) Calcular a probabilidade de retirar uma bola branca da urna "A";

b) Determine a probabilidade de retirarmos uma bola branca ou vermelha da urna "A";

c) Determine a probabilidade de retirarmos uma bola branca da urna "A" e uma bola vermelha da urna "B";

d) Qual a probabilidade de serem retiradas duas bolas vermelhas da urna "A", com reposição?;

e) Qual a probabilidade de serem retiradas duas bolas pretas da urna "B"? (sem reposição);

4. A probabilidade de o aluno "X" resolver este problema é de 3/5, e de o aluno "Y" é de 4/7.

Qual a probabilidade de que o problema seja resolvido por eles?

5. Um grupo de 100 pessoas apresenta, de acordo com o sexo e qualificação a seguinte composição:

Sexo	Especializados	Não especializados	Total
Homens	21	39	60
Mulheres	14	26	40
Total	**35**	**65**	**100**

Calcular:

a) A probabilidade de um escolhido ser homem.

b) A probabilidade de um escolhido ser mulher e não especializada.

c) Qual a porcentagem dos não especializados?

d) Qual a porcentagem dos homens não especializados?

e) Se o sorteado é especializado, qual a probabilidade de ser mulher?

f) Se o sorteado for homem, qual a probabilidade de ser não especializado?

6. Uma urna contém quatro bolas brancas, cinco azuis e seis pretas em uma outra temos cinco bolas brancas, seis azuis e duas pretas. Extrai-se uma bola de cada urna, na seqüência estabelecida anteriormente, qual a probabilidade:

a) de que ambas sejam da mesma cor?

b) da primeira ser azul e a segunda ser preta?

c) de uma ser azul e a outra ser preta?

d) da primeira ser branca e a segunda não ser branca?

7. A probabilidade da classe "A" comprar um carro é 3/4, da "B" é 1/6 e da "C", 1/20.

A probabilidade de o indivíduo da classe "A" comprar um carro da marca "W" é 1/10; de B comprar da marca "W" é 3/5 e de C é 3/10. Em certa loja um indivíduo comprou um carro da marca "W".

Qual a probabilidade de que o indivíduo:

a) Da classe "A" o tenha comprado?

b) Da classe "B" o tenha comprado?

c) Da classe "C" o tenha comprado?

8. Três máquinas M_1, M_2 e M_3 produzem respectivamente 40%, 50% e 10% do total de peças de uma fábrica. A porcentagem de peças defeituosa nas respectivas máquinas é 3%, 5% e 2%. Uma peça é sorteada ao acaso e verifica-se que é defeituosa. Qual a probabilidade de que a peça tenha vindo da máquina:

a) M_1

b) M_2

c) M_3

9. A empresa de construção "Tijolo S.A." vai apresentar uma proposta de construção de um armazém do tipo A. Considere a variável aleatória X, que representa o número de dias para construir um armazém do tipo A, e a respectiva função de probabilidade:

X	20	21	22	23	24
P(x)	k/2	0,15	3k	0,1	0,05

a) Determine o valor da constante k, justificando.

b) Qual a probabilidade do tempo de construção demorar mais de 22 dias?

c) Qual a probabilidade do tempo de construção demorar entre 21 e 23 dias (inclusive)?

d) Quantos dias espera a empresa demorar para construir o referido armazém?

e) Calcule o valor de Var (X).

f) Os custos de construção são os seguintes:

- Materiais: 16.000 euros
- Mão de obra: 750 euros por cada dia de construção

Os responsáveis pela empresa pretendem obter um valor esperado do lucro de 2.500 euros. Atendendo aos custos que constam na tabela anterior, calcule o valor que deve ser apresentado na proposta de construção.

Distribuição Binomial, Distribuição Poisson e Distribuição Normal

Introdução

A distribuição de probabilidade é uma função que determina probabilidades para eventos ou proposições. Para qualquer conjunto de eventos ou proposições, existem muitas maneiras de determinar probabilidades, de forma que a escolha de uma ou outra distribuição é equivalente a criar diferentes hipóteses sobre os eventos ou proposições em questão. A distribuição de probabilidade de uma variável descreve como as probabilidades estão distribuídas sobre os valores da variável aleatória.

Há várias formas equivalentes de se especificar uma distribuição de probabilidade. Uma distribuição é chamada de *distribuição discreta* se for definida em um conjunto contável e discreto, tal como o subconjunto dos números inteiros; ou é chamada de *distribuição contínua* se tiver uma função distribuição contínua, tal como uma função polinomial ou exponencial.

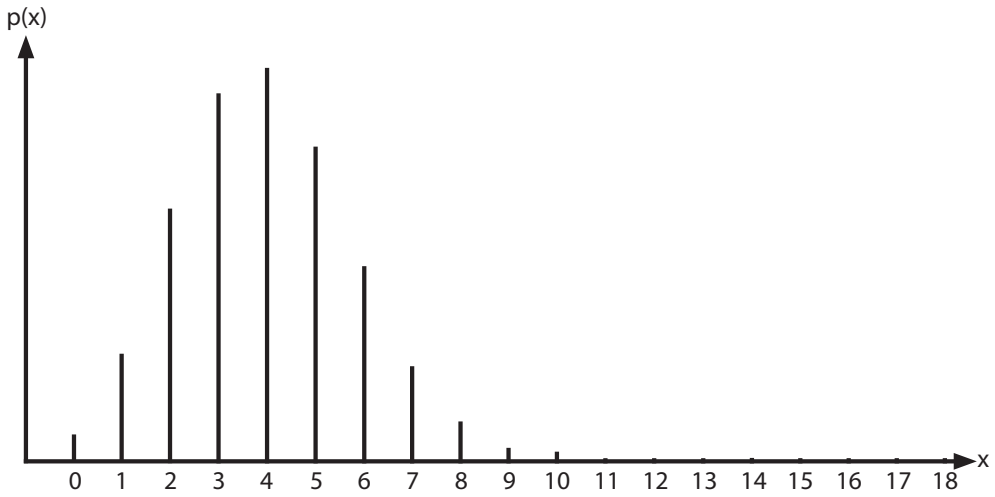
A seguir, veremos as principais distribuições de probabilidade: Binomial e Poisson para variáveis aleatórias discretas e a distribuição Normal para uma variável aleatória contínua.

Analisemos a definição de variável aleatória discreta: seja X uma variável aleatória discreta e x_i um certo valor de X. A probabilidade de ocorrência de x_i é dada por $P(X = x_i) = p(x_i)$, onde:

- $p(x_i) \geq 0$
- a soma de todos os $p(x_i)$ é igual a 1.

Como as variáveis aleatórias discretas X assumem valores inteiros (geralmente), as probabilidades associadas a esses valores (x_i) são pontuais de forma que a distribuição de probabilidade é representada por quantidades de massa localizadas nos pontos x_i.

Figura 1: Esboço de uma função de probabilidade discreta.



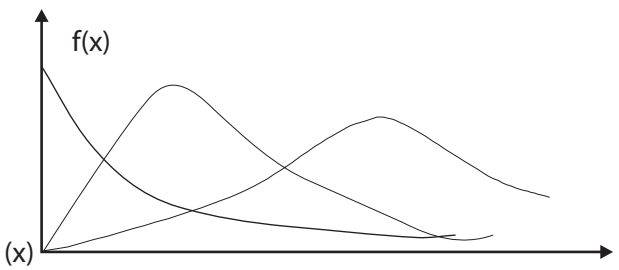
Por outro lado, a probabilidade de ocorrência de uma variável aleatória contínua dentro de um determinado intervalo (a,b), é dada por:

$$\Pr(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Onde $\int_a^b$ é a notação que se usa para representar a integração de uma determinada função em um intervalo de *a* até *b*. Utilizada para cálculo de áreas e aqui será utilizada para cálculo de probabilidades.

As variáveis aleatórias contínuas X assumem valores dentro de um intervalo contínuo, e as probabilidades associadas a esses valores podem ser consideradas "áreas abaixo de uma curva".

Figura 2: Esboço de algumas funções densidades de probabilidade contínuas.



Distribuição de probabilidade Binomial

Antes de introduzirmos a distribuição de probabilidade Binomial, vamos definir outra distribuição, a distribuição Bernoulli, que dá origem a ela. Na distribuição Bernoulli:

a) Cada experimento é dito ser uma tentativa. Em cada tentativa, existem dois resultados possíveis: sucesso ou falha.

b) A probabilidade de sucesso é igual a algum valor constante para todas as tentativas.

c) Os resultados sucessivos são estatisticamente independentes. A probabilidade de sucesso na próxima tentativa não pode variar, não importando quantos sucessos ou falhas tenham sido obtidos.

O processo de Bernoulli é comumente utilizado em aplicações envolvendo controle de qualidade. Cada novo item criado no processo de produção pode ser considerado como uma tentativa resultando em uma unidade com ou sem defeito. Esse processo não se limita a objetos; podendo ser usado em pesquisas eleitorais e de preferências dos consumidores por determinados produtos.

Consideremos agora **n** tentativas independentes de ensaios de Bernoulli. Cada tentativa admite apenas dois resultados complementares: sucesso com probabilidade **p** ou fracasso com probabilidade **q**, de modo a se ter p + q = 1. As probabilidades de sucesso e fracasso são as mesmas para cada tentativa. A variável aleatória X, que conta o número total de sucessos, é denominada Binomial.

Exemplo: suponha que peças saiam de uma linha de produção e sejam classificadas como defeituosas (D) ou como não-defeituosas (N). Admita que 3 dessas peças sejam escolhidas ao acaso. Se a probabilidade de que uma peça seja defeituosa é de 0,2, calcule a probabilidade de obtermos 0, 1, 2 ou 3 peças defeituosas.

Então teremos: n = 3 (número de repetições do experimento); p = 0,2 (probabilidade de "sucesso", ou de obter uma peça defeituosa).

Considere, agora, a seguinte definição:

Seja E um experimento e A um evento associado a E. Considere ainda P(A) = p, denominada Probabilidade de ocorrência de A, que satisfaça as seguintes propriedades:

- ocorrem **n** repetições independentes do experimento E;
- a probabilidade **p** é sempre constante para cada repetição;
- a variável aleatória X será definida como sendo o número de vezes que o evento A ocorre;
- $P(A^C) = 1 - P(A) = q$

Então,

$$P(X = k) = \binom{n}{k}.p^{k}.q^{n-k}, k = 0, 1, 2, ..., n.$$

em que $\binom{n}{k}$ é a combinação de **n** elementos divididos em **k** grupos. Pode ser desenvolvida fazendo-se: $\binom{n}{k} = \frac{n!}{k!.(n-k)!} = \frac{n.(n-1).(n-2)\ldots(n-k+1)}{k.(k-1).(k-2)\ldots 1}$

Agora a resolução da questão acima fica muito mais simples. Basta definirmos:

- n = 3
- p = 0,2

$$P(X = 0) = \binom{3}{0}.p^{0}.q^{3} = \frac{3!}{0!3!}.1\,.0,8^{3} = 0,512$$

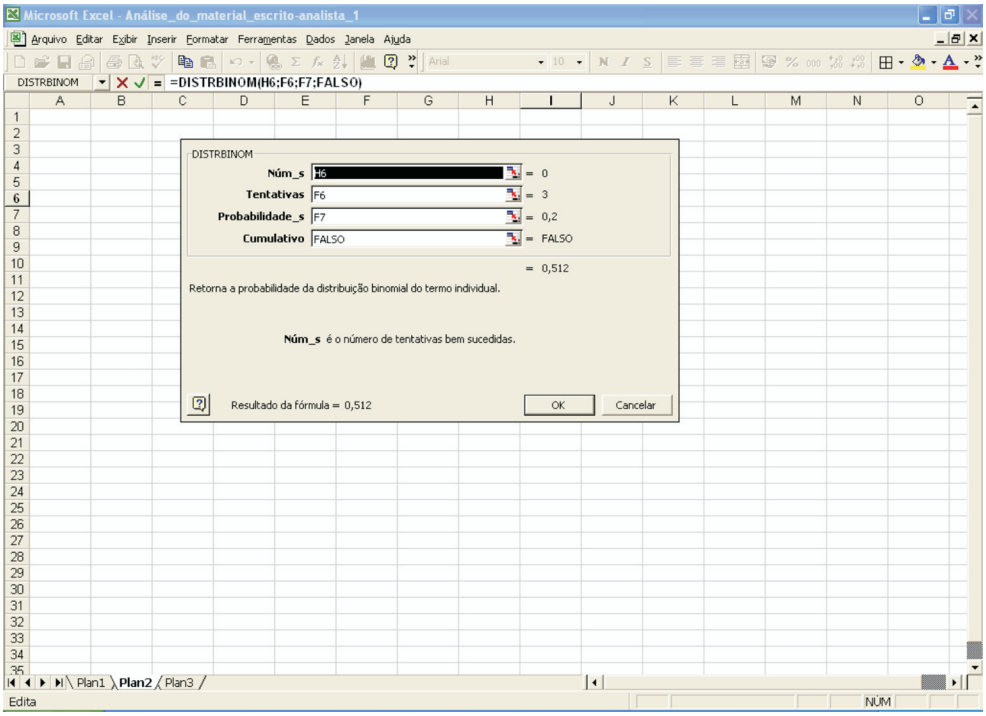
$$P(X = 1) = \binom{3}{1}.p^{1}.q^{2} = \frac{3!}{1!2!}.0,2^{1}\,.\,0,8^{2} = 0,384$$

$$P(X = 2) = \binom{3}{2}.p^{2}.q^{1} = \frac{3!}{2!1!}.0,2^{2}\,.\,0,8^{1} = 0,096$$

$$P(X = 3) = \binom{3}{3}.p^{3}.q^{0} = \frac{3!}{3!0!}.0,2^{3}\,.\,0,8^{0} = 0,008$$

Utilizando a planilha eletrônica *Excel*, podemos resolver o problema acima de uma forma muito fácil, simplesmente utilizando as funções. Então, utilizaríamos a função DISTRBINOM considerando:

- **Num_s** (número de tentativas bem-sucedidas) – é o valor que X assume, pode ser 0, 1, 2 ou 3, dependendo da probabilidade que se deseja calcular;
- **Tentativas** – é o tamanho da amostra, no caso n = 3;
- **Probabilidade_s** – é a probabilidade de sucesso, no caso, p = 0,2;
- **Cumulativo** – é a opção que fornece a probabilidade acumulada ou a probabilidade individual. No caso, preencher o campo com FALSO para considerar a probabilidade individual.



Notação: X ~ b(n; p)

Isto significa que a variável aleatória X tem distribuição Binomial com parâmetros n e p.

A *esperança* e a *variância* para uma variável aleatória com distribuição Binomial são dadas por:

$$\mu = E(X) = n.p$$

$$\sigma^2 = Var(X) = n.p.(1 - p)$$

Distribuição de Probabilidade Poisson

Na distribuição Binomial, a variável aleatória X é o número de "sucessos" que ocorrem em *n* tentativas independentes do experimento. Podemos considerar agora uma variável aleatória X igual ao número de "sucessos" que ocorrem num intervalo contínuo.

Por exemplo:

- número de chamadas X que uma telefonista recebe no intervalo de uma hora;

- o número de falhas em 1 m^2 de tecidos;
- o número de vezes que um computador "trava" em um intervalo de 8 horas.

Uma variável aleatória assim, assume valores inteiros, ou seja, X = 0, 1, 2, 3, 4, ...

Um fenômeno ou experimento de Poisson tem as seguintes características:

- o número de sucessos que ocorrem num intervalo contínuo é independente daqueles que ocorrem em qualquer outro intervalo disjunto;
- em intervalos de mesmo comprimento a probabilidade de ocorrência de um mesmo número de "sucessos" é igual;
- em intervalos muito pequenos, a probabilidade de mais de um "sucesso" é desprezível.

Nessas condições, a variável aleatória X = número de sucessos que ocorrem num determinado intervalo contínuo de tem distribuição de Poisson com parâmetro λ e função de probabilidade dada por:

$$p(x) = Pr(X = x) = \frac{e^{-\lambda}.\lambda^{x}}{x!}, \text{ para } x = 0, 1, 2, ...$$

em que λ é a média de sucessos no intervalo considerado e **e** é a constante exponencial que é igual a 2,718281828.

Notação: $X \sim P(\lambda)$

Isso significa que a variável aleatória X tem distribuição Poisson com parâmetro λ.

A *esperança* e a *variância* para uma variável aleatória com distribuição de Poisson são dadas por:

$\mu = E(X) = \lambda$

$\sigma^2 = Var(X) = \lambda$

Exemplo: Clientes em potencial chegam a um posto de gasolina de acordo com um processo de Poisson com taxa de 20 carros por hora. Então, a função de probabilidade associada é dada por:

$$p(x) = \frac{e^{-20}.20^{x}}{x!}, \text{para } x = 0,1,2,...$$

A probabilidade de chegarem em 1 hora:

a) Exatamente 10 carros:

$$P(X = 10)\frac{e^{-20}.20^{10}}{10!} = 0,0058 \text{ ou } 0,58\%$$

b) 10 carros ou menos:

$$P(X \leq 10)\sum_{x=0}^{10}\frac{e^{-20}.20^{x}}{x!} = 0,0108 \text{ ou } 1,08\% = 0,0108 \text{ ou } 1,08\%$$

c) Mais de 20 carros:

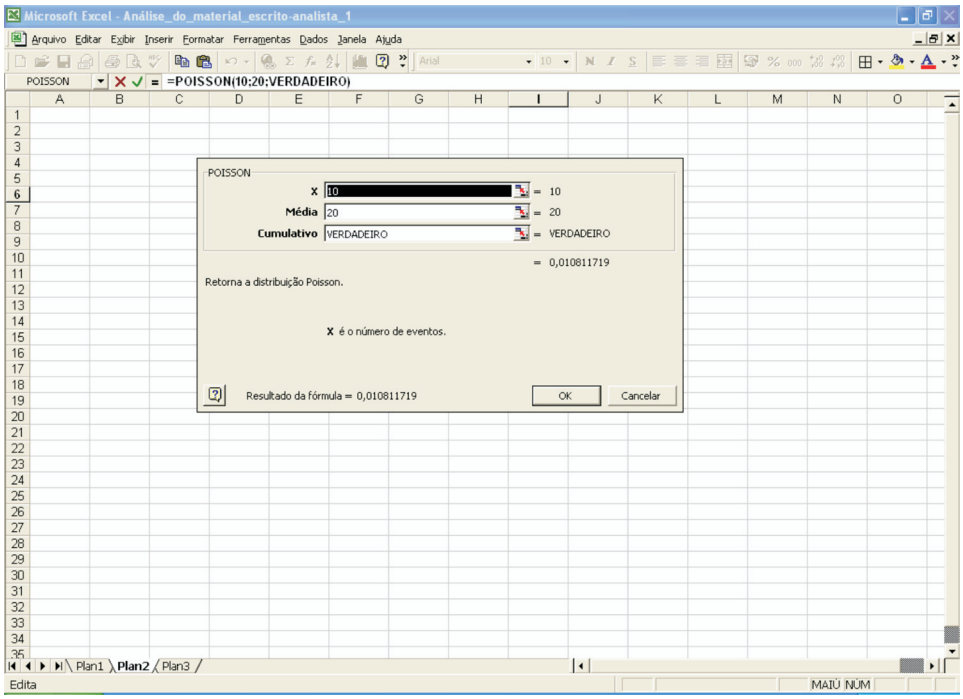
$$P(X > 20) = \sum_{x=21}^{\infty}\frac{e^{-20}.20^{x}}{x!} = 1 - \sum_{x=0}^{20}\frac{e^{-20}.20^{x}}{x!} = 0,441 \text{ ou } 44,1\%$$

d) Entre 11 e 20 carros:

$P(11 \leq X \leq 20) = P(X \leq 20) - P(X \leq 10) = \sum_{x=11}^{20}\frac{e^{-20}.20^{x}}{x!} = 0,559 - 0,0108 =$ 0,548 ou 54,8%

Utilizando o *Excel*, utilizaríamos a função POISSON considerando:

- **X** (número de eventos) – é o valor que X assume, pode ser 0, 1, 2 etc, até infinito dependendo da probabilidade que se deseja calcular.
- **Média** – é o valor do parâmetro λ.
- **Cumulativo** – é a opção que fornece a probabilidade acumulada ou a probabilidade individual. No caso, preencher o campo com VERDADEIRO para considerar a probabilidade acumulada.



Distribuição de Probabilidade Normal

A distribuição normal foi estudada inicialmente no século 18, quando uma análise de erros experimentais levou a uma curva em forma de sino. Embora ela tenha aparecido pela primeira vez em 1733 por DeMoivre, a distribuição normal recebe o nome de distribuição gaussiana, em homenagem ao cientista alemão Karl Friedrick Gauss, que foi o primeiro a utilizá-la em 1809.

Nos séculos 18 e 19, matemáticos e físicos desenvolveram uma função densidade de probabilidade que descrevia bem os erros experimentais obtidos em medidas físicas. Essa função densidade de probabilidade resultou na bem conhecida curva em forma de sino, chamada de distribuição normal ou gaussiana. Essa distribuição fornece uma boa aproximação de curvas de freqüência para medidas de dimensões e características humanas, como a altura de uma população.

A distribuição normal é a mais importante das *distribuições contínuas* de probabilidade, e tem sua origem associada aos erros de mensurações. A distribuição normal desempenha papel preponderante na estatística, e os processos de inferência nela baseados têm larga aplicação.

A distribuição normal tem sua função densidade de probabilidade (f.d.p.) dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

em que:

- μ – é a média da variável X;
- σ – é o desvio padrão da variável X;
- π – é uma constante numérica igual a 3,141593.

Notação: $X \sim N(\mu; \sigma^2)$

Isso significa que a variável aleatória X tem distribuição Normal com parâmetros μ e σ^2.

São propriedades da distribuição normal:

1) A distribuição é simétrica em relação a $x = \mu$, ou seja, nesse ponto a curva se divide em duas partes iguais.

2) A função f(x) tem um ponto de máximo para $x = \mu$.

3) As "caudas" da função f(x) são chamadas "assintóticas", ou seja, só atingem o ponto f(x) = 0 quando x tende a + infinito ou – infinito. Isso quer dizer que a curva jamais cruza o eixo x.

4) A função f(x) tem dois pontos de inflexão para $x = \mu + \sigma$ e $x = \mu - \sigma$. Nestes pontos a função acentua sua curvatura.

5) A função de distribuição acumulada é dada por

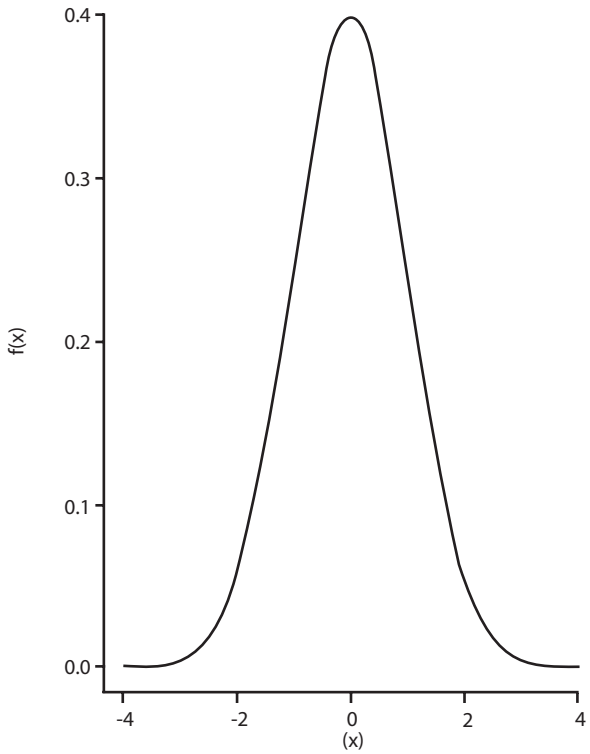
$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

A função F(x), dada acima, pode ser colocada numa forma mais simples, considerando-se a transformação:

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

que é a *variável normal padronizada* ou *reduzida Z*.

Figura 3: Curva da distribuição Normal padrão.



Notamos que a transformação utilizada consiste em adotarmos uma nova distribuição normal de média $\mu = 0$ e variância $\sigma^2 = 1$ ou desvio padrão $\sigma = 1$. Portanto,

$$Z \sim N(0; 1).$$

Isso significa que a variável aleatória Z assume uma distribuição Normal com média zero e variância 1.

Assim, a *f.d.p.* da variável normal padronizada será dada por

$$g(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}, -\infty \leq z \leq \infty$$

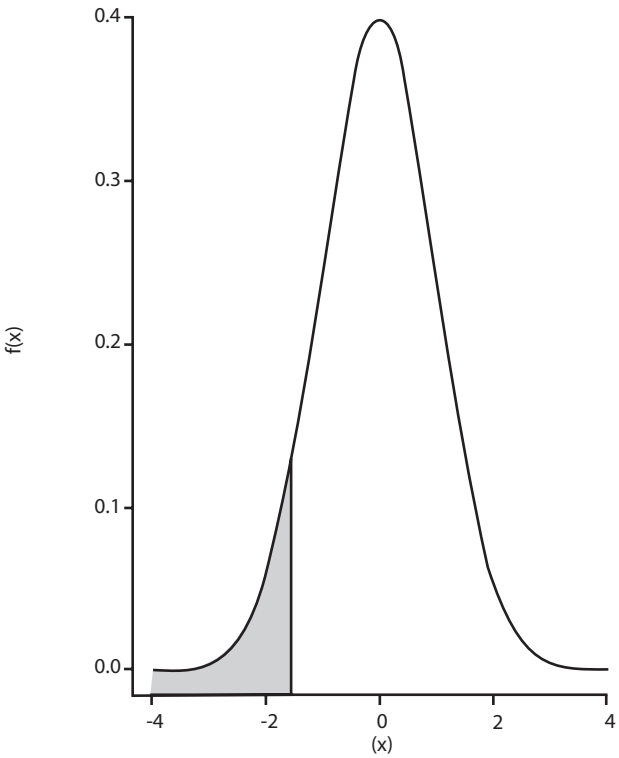
A distribuição normal padronizada pode ser tabulada utilizando-se métodos de integração numérica.

Exemplo: Uma indústria fabrica peças mecânicas cujas medidas dos diâmetros externos são normalmente distribuídas com média 40,0 mm e desvio padrão de 2,0 mm. Vamos calcular a percentagem de peças defeituosas

fabricadas, sabendo-se que o setor de controle de qualidade dessa indústria classifica como defeituosas aquelas peças cujos diâmetros externos:

a) são inferiores a 37,0 mm.

$P(X<37) = P(Z<(37–40)/2) = P(Z<–1,5) = 0,067$ ou 6,7%.

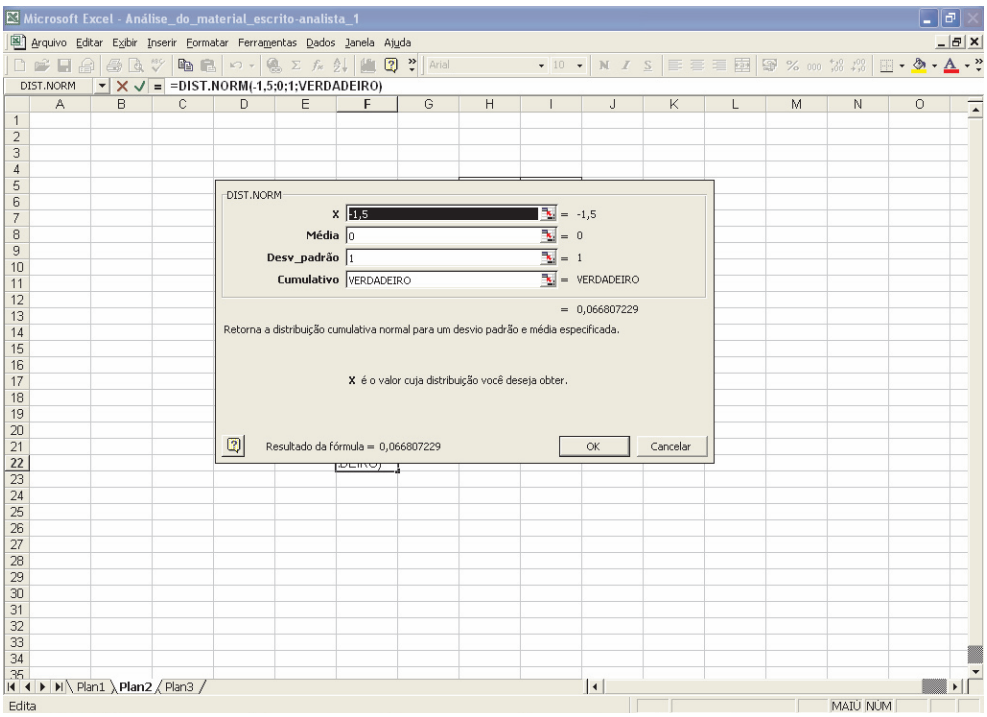


Consultando a tabela da distribuição normal padrão (anexo 1), iremos procurar a linha referente ao valor 1,5 e a coluna referente ao valor zero (1,5 + 0,00 = 1,50). Cruzando esses dois valores, obteremos, no corpo da tabela, 0,4332. Esse valor, como a figura ilustra na tabela de valores críticos, nos dá o tamanho da área entre o ponto zero e o ponto 1,5. Utilizando as propriedades de simetria da curva normal, teremos que $P(Z<–1,5) = 0,5 – 0,4332 = 0,067$ que é o tamanho da área assinalada em cinza, na figura acima, pois o valor de X nesse caso é negativo.

Usando a planilha do *Excel*, utilizaríamos a função DIST.NORM:

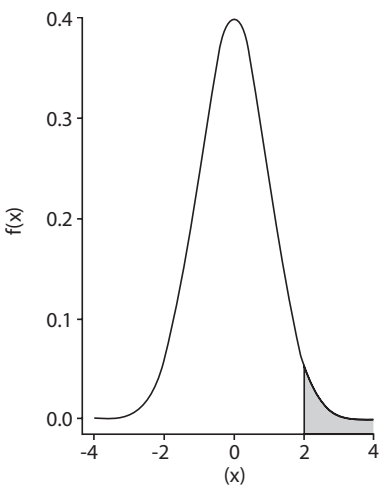
- **X** – é o valor cuja probabilidade se deseja calcular;
- **Média** – é o valor do parâmetro μ da distribuição;

- **Desv_padrão** – é o valor de σ;
- **Cumulativo** – é a opção que fornece a probabilidade acumulada ou a probabilidade individual. No caso, sempre preencher o campo com VERDADEIRO.



b) São superiores a 44,0 mm.

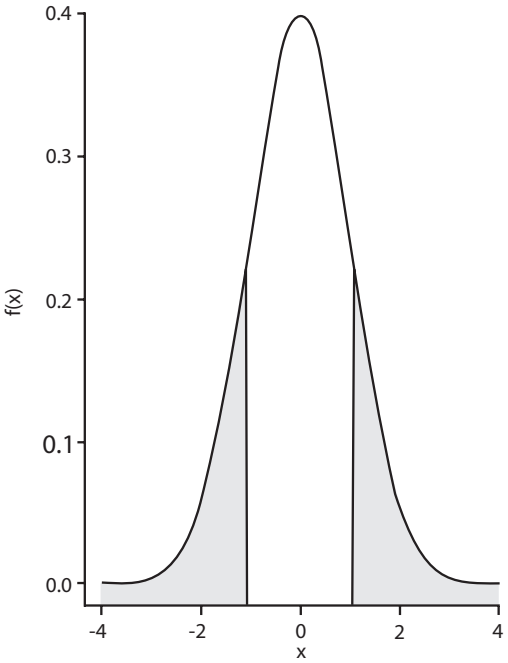
$$P(X>44) = P(Z>(44-40)/2) = P(Z>2) = 0,023 \text{ ou } 2,3\%.$$



c) Desviam-se mais de 2,0 mm da média.

$P(X<38) + P(X>42) = P(Z<(38-40)/2) + P(Z>(42-40)/2)$

$= P(Z<-1) + P(Z>1) = 0{,}1586 + 0{,}1586 = 0{,}3164$ ou 31,64%.



Testes para a Distribuição Normal

Muitos testes usados em estatística partem do princípio que os dados são provenientes de uma população normal. Ou seja, só podem ser utilizados se for comprovada a suposição de normalidade dos dados. Dessa forma, testes estatísticos devem ser feitos para verificar esse fato.

Existem os testes qualitativos e quantitativos. Dentre os *testes qualitativos*, existem três representações gráficas que são comumente utilizadas: o gráfico de probabilidade normal (*normal probability plot*), o da probabilidade normal positiva (*half-normal probability plot*) e o da probabilidade normal sem tendências (*detrended normal probability plot*).

As Figuras 4 a 6 apresentam esses gráficos gerados pelo *software Statistica*, e selecionando-se a variável Pressão. Caso os pontos caiam próximos à linha reta, pode-se dizer que os dados seguem uma distribuição normal. No caso da Figura 6, fica claro que não há qualquer tendência característica de normalidade para o comportamento dos dados de pressão.

Figura 4: Gráfico da Probabilidade Normal.

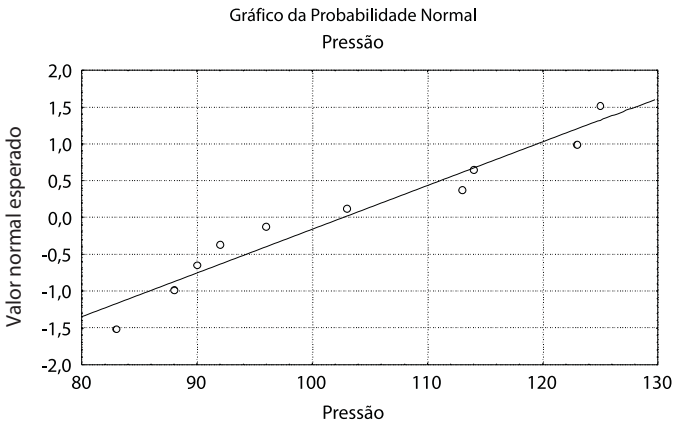


Figura 5: Gráfico da Probabilidade Normal Positiva.

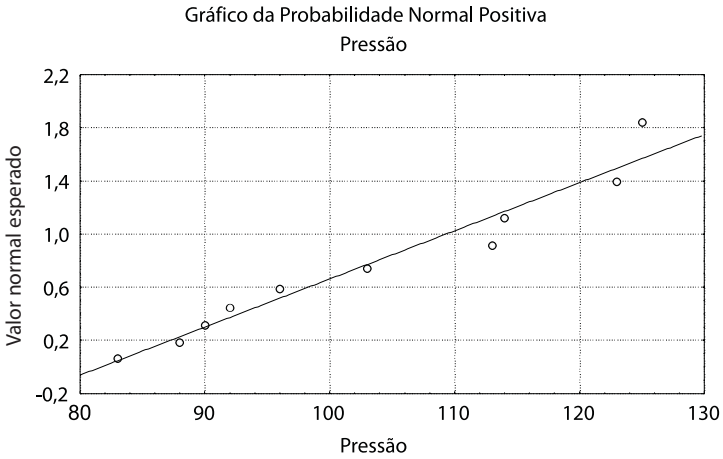
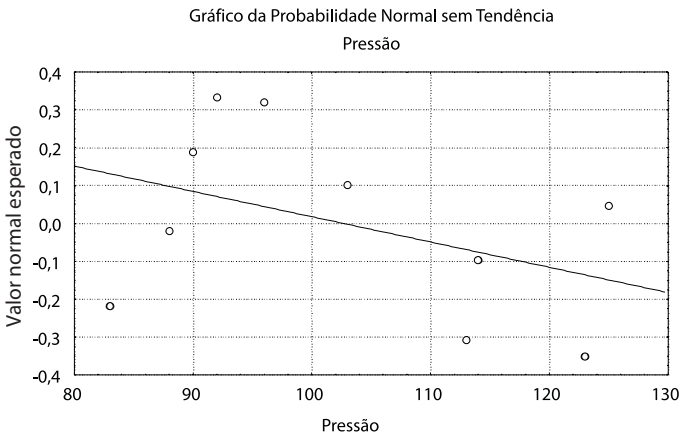


Figura 6: Gráfico da Probabilidade Normal sem Tendência.



Os *testes quantitativos* são mais eficientes, pois independem de qualquer interpretação subjetiva. Eles consistem em calcular uma estatística, característica de cada teste, e verificar se o seu valor é significativo, dependendo do nível de significância escolhido. Caso seja, então a hipótese de que os dados seguem uma distribuição normal deve ser rejeitada. Os testes mais usados para verificar normalidade são:

- *Kolmogorov-Smirnov* – usado quando a média e o desvio-padrão da distribuição normal são conhecidos e não estimados a partir dos dados. Entretanto, geralmente esses parâmetros são calculados a partir dos dados reais.
- *Lilliefors* – usado quando a média e o desvio-padrão da população são desconhecidos e acabam sendo estimados a partir dos dados da amostra.
- *Shapiro-Wilks (W)* – outra opção para verificação de normalidade, em que se trabalha com os dados ordenados, geralmente quando se tem menos de 50 observações.

Caso seja verificado que a população não seja normal, transformações da variável podem ser feitas, a fim de torná-la normal. A transformação de Box-Cox é uma das transformações mais utilizadas. Ela consiste em extrair a raiz quadrada ou aplicar o logaritmo nos valores da variável em estudo.

Outra alternativa, caso a suposição de normalidade não seja atingida, é realizar um teste estatístico que não necessita de comprovação de normalidade dos dados, os chamados "testes não-paramétricos". Apresentaremos a seguir o teste não-paramétrico de Lilliefors para testar a suposição de normalidade.

Teste de Lilliefors

No caso em que se deseja testar normalidade e a média e a variância não são previamente especificadas, mas sim estimados por meio dos dados da amostra. Deve-se utilizar o teste de Lilliefors. Esse teste tem procedimento análogo ao teste Kolmogorov-Smirnov, porém utiliza uma tabela de valores críticos própria e mais adequada a esse tipo de situação.

Esse teste de aderência avalia a concordância entre a distribuição observada da amostra e uma determinada distribuição teórica. Para isso, utilizamos a

função distribuição acumulada observada, compara-se com a teórica, determina-se o ponto em que essas distribuições mais divergem, e testamos se essa divergência é aleatória ou não.

Seja $F_0(X)$ uma distribuição teórica acumulada e $S_n(X)$ uma distribuição observada em uma amostra de **n** observações (distribuição empírica).

Encontra-se a seguir o maior valor das diferenças entre $F_0(X)$ e $S_n(X)$, ou seja,

$$D = \text{máx}\left| F_0(X) - S_n(X) \right|$$

Compara-se o valor observado com o valor crítico que se encontra na tabela em anexo. Se o valor calculado for inferior ao valor tabelado, então podemos considerar que os dados se ajustam bem a uma distribuição Normal.

Exemplo: As produções médias (sacas) obtidas em um experimento envolvendo um novo adubo em plantações de milho encontram-se tabuladas abaixo:

Classes	f_i	x_i	$F(x_i)$	$S(x_i)$	$\lvert F(x_i) - S(x_i) \rvert$
2 700 |— 3 000	13	2 850	0,045	0,113	0,068
3 000 |— 3 300	18	3 150	0,155	0,269	0,114
3 300 |— 3 600	24	3 450	0,371	0,478	0,107
3 600 |— 3 900	32	3 750	0,639	0,756	**0,117**
3 900 |— 4 200	17	4 050	0,851	0,904	0,053
4 200 |— 4 500	11	4 350	0,958	1,000	0,042
	115				

Podemos admitir que a produção média segue uma distribuição normal?

A coluna S(x) apresenta as probabilidades acumuladas, por exemplo, o primeiro valor, 0,113, foi obtido pela razão: 13/115 e os demais valores foram obtidos sempre acumulando o valor das classes anteriores, até a última classe em que S =1. Os valores de F(X) são as probabilidades acumuladas de uma distribuição normal. Mas para esse cálculo, precisamos dos valores dos

parâmetros da distribuição. Como esses valores não são conhecidos, devem ser estimados. A estimativa do parâmetro μ é a média amostral e a estimativa do parâmetro σ^2 é a variância amostral. Assim, teremos a estimativa de μ = 3 593,5 sacas (para calcular a média, nesse caso, primeiro multiplica-se o ponto médio de cada classe, pela sua respectiva freqüência. A partir disso, soma-se todos os resultados obtidos e divide-se pelo número de elementos – 115) e a estimativa da variância = 191 601,8 (obtida através da fórmula da variância: $\frac{\sum \left(x_i - \overline{X}\right)^2 .f_i}{n-1}$). Assim, já é possível obtermos as probabilidades acumuladas.

Dessa forma, as probabilidades acumuladas para as classes da tabela acima são calculadas sempre em função de seu ponto médio (x_i):

$P(X \leq 2\,850) = P(Z \leq -1,7) = 0,045$

$P(X \leq 3\,150) = P(Z \leq -1,01) = 0,156$

$P(X \leq 3\,450) = P(Z \leq -0,33) = 0,371$

$P(X \leq 3\,750) = P(Z \leq 0,36) = 0,639$

$P(X \leq 4\,050) = P(Z \leq 1,04) = 0,851$

$P(X \leq 4\,350) = P(Z \leq 1,73) = 0,958$

Agora, basta calcularmos as diferenças entre a distribuição acumulada observada pelos dados e a distribuição acumulada teórica, calculada por meio da distribuição Normal. Essas diferenças são apresentadas na última coluna. A maior das diferenças encontrada foi 0,117. Assim, precisamos verificar se essa diferença pode ou não ser considerada significativa. Consultando a tabela de valores críticos, a um nível de significância de 5% precisaremos informar o tamanho da amostra (n). Nesse caso, n = 115 e usamos a última linha da tabela que aponta $\frac{0,886}{\sqrt{n}} = 0,082$. Como o valor calculado (0,117) é superior ao valor crítico tabelado (0,082) rejeitamos a hipótese nula e temos indícios suficientes para afirmar que a distribuição normal, nesse caso, não

se ajusta aos dados.

(Wikipédia)

Jakob Bernoulli, (1654 em Basiléia - 1705 idem)

Foi professor de matemática em Basiléia, tendo sido importantíssima sua contribuição à geometria analítica, à teoria das probabilidades e ao cálculo de variações.

Em 1713, depois de sua morte, foi publicado seu grande tratado sobre a teoria das probabilidades *Ars Conjectandi* que ainda oferece interesse prático na aplicação da teoria da probabilidade no seguro e na estatística.

Siméon Denis Poisson (Pithiviers em 1781 - Sceaux em 1840)

Engenheiro e matemático francês, considerado o sucessor de *Laplace* no estudo da mecânica celeste e da atração de esferóides. Entrou para a École Polytechnique (1798), em Palaiseau, onde se formou, estudando com professores como *Joseph Louis Lagrange*, *Pierre Simon Laplace e Jean Baptiste Fourier*.

Em *Recherches sur la probabilité des jugements* (1837) apareceu a famosa *distribuição de Poisson* de intensa aplicação em estatística. Na teoria de probabilidades, descobriu a forma limitada da distribuição Binomial que posteriormente recebeu o seu nome e hoje é considerada uma das mais importantes distribuições na probabilidade.

Abraham de Moivre (Vitry em 1667 – Londres em 1754)

Matemático francês que fez carreira profissional na Inglaterra, onde foi professor particular e tornou-se um destacado pesquisador com grandes contribuições no campo da teoria das probabilidades, porém sem se tornar professor universitário por causa de sua nacionalidade. Pioneiro do desenvolvimento de Geometria Analítica e a Teoria de Probabilidade, publicou o célebre *Doctrine of Chances* (1718), sobre a Teoria do Acaso, onde expôs a definição de independência estatística junto com muitos problemas com dados e outros jogos. Também pesquisou estatísticas de mortalidade e fundou a teoria de anuidades.

Johann Carl Friedrich Gauss (Braunschweig em 1777 – Göttingen em 1855)

Trabalhou em diversos campos da Matemática e da Física dentre eles a Teoria dos Números, Geometria Diferencial, Magnetismo, Astronomia e Óptica. Seu trabalho influenciou imensamente outras áreas.

Em probabilidade e estatística, ficou famoso pelo desenvolvimento do método dos mínimos quadrados e pela descoberta da distribuição normal, agora também conhecida como a *Distribuição Gaussiniana*, a conhecida lei de probabilidade, definida graficamente por meio da chamada *Curva de Gauss*.

Ampliando seus conhecimentos

Atividades de aplicação

1. Seja X uma variável aleatória com distribuição Binomial, baseada em 10 repetições de um experimento. Se p = 0,3, calcule as seguintes probabilidades:

a) $P(X \leq 8)$

b) $P(X=7)$

c) $P(X \geq 6)$

2. Um jogador de basquetebol acerta um arremesso com probabilidade 0,9. Em cinco arremessos, a probabilidade de o jogador acertar todos é:

a) 0,59

b) 0,9

c) 0,81

d) 0,9 x 5

e) 0,45

3. Suponha que 5% de todas as peças que saiam de uma linha de produção sejam defeituosas. Se 10 dessas peças forem escolhidas e inspecionadas, qual será a probabilidade de que no máximo 2 defeituosas sejam encontradas?

4. O número de navios petroleiros que chegam a determinada refinaria, a cada dia, tem distribuição de Poisson, com parâmetro $\lambda = 2$. As atuais instalações do porto podem atender a três petroleiros por dia. Se mais de 3 navios aportarem por dia, os excedentes devem seguir para outro porto.

a) Em um dia, qual é a probabilidade de se ter de mandar petroleiros para outro porto?

b) De quanto as atuais instalações devem ser aumentadas para permitir manobrar todos os petroleiros, em aproximadamente 90% dos dias?

c) Qual é o número esperado de petroleiros a chegar por dia?

d) Qual é o número mais provável de petroleiros a serem atendidos diariamente?

e) Qual é o número esperado de petroleiros a serem atendidos diariamente?

f) Qual é o número esperado de petroleiros que voltarão a outros portos diariamente?

5. O número de clientes que chegam a fila de um banco durante o intervalo de uma hora é uma variável aleatória com distribuição de Poisson com média igual a 5. A probabilidade de não haver chegada de clientes durante esse intervalo é :

a) e^{-0}

b) 0

c) 0,0067

d) 0,034

e) 1

6. Em uma curva Normal Padrão, a área entre -1,96 e 1,96 corresponde a 0,95. Para uma variável aleatória X normalmente distribuída com média 10 e variância 100, a área correspondente a 95% centrais dessa curva está situada entre:

a) –9,6 e 29,6

b) –8,6 e 10,6

c) –9,6 e 11,6

d) 18,6 e 20,6

e) –186 e 206

7. Suponha que a distribuição de salários de uma empresa americana segue uma distribuição normal, com média mensal de US$ 15.000,00 e desvio padrão de US$ 2.000,00. Calcule a probabilidade de alguém ganhar menos de US$ 5.000,00.

8. A força (em Newton) com que um tecido sintético se parte é representa-

Estimação de parâmetros

Introdução

É muito comum, quando estudamos uma população, conhecermos a distribuição da característica em estudo e não conhecermos os parâmetros dessa distribuição. Então, com base numa amostra aleatória dessa população, nós deveremos estimar um valor aproximado para os parâmetros da população. *Estimação* é o processo que consiste em utilizar dados amostrais para estimar os valores de parâmetros populacionais.

Lembremos que, *parâmetros* são funções de valores populacionais, enquanto que *estatísticas* são funções de valores amostrais.

Inicialmente, vejamos a questão de estimação de um modo mais geral. Consideremos uma amostra $(X_1, X_2, ...,X_n)$ de uma variável aleatória que descreve uma característica de interesse de uma população. Seja θ um parâmetro que desejamos estimar, como por exemplo a média μ ou a variância σ^2 .

Definição 1: ***Estimador e Estimativa***

Um *estimador* do parâmetro θ é qualquer função das observações $X_1, X_2, ...,X_n$, isto é, $g(X_1, X_2, ...,X_n)$. O valor que *g* assume, isto é, $g(x_1, x_2, ..., x_n)$, é referido como uma *estimativa* de θ e é usualmente escrito assim: $\hat{\theta} = g(x_1, x_2, ..., x_n)$.

Note que, segundo esta definição, um estimador é qualquer estatística cujos valores são usados para estimar θ (ou uma função de θ).

O problema da estimação é, então, determinar uma função $T = g(X_1, X_2, ...,X_n)$ que seja "próxima" de θ, segundo algum critério. Esses critérios são vistos mais adiante.

Notação: θ : parâmetro a ser estimado

T : um estimador de θ

$\hat{\theta}$: uma estimativa de θ

Estimadores pontuais (ou por ponto)

A estimação pontual (por ponto) consiste simplesmente em, à falta de melhor informação, adotar a estimativa disponível como sendo o valor do parâmetro. A idéia é, em sua essência, extremamente simples, porém a qualidade dos resultados depende fundamentalmente da conveniente escolha do estimador. Assim, dentre os vários estimadores razoáveis que poderemos imaginar para um determinado parâmetro, devemos ter a preocupação de escolher aquele que melhor satisfaça às propriedades de um bom estimador. Essas propriedades são dadas logo a seguir.

Definição 2: ***Estimador pontual***

Seja $X_1, X_2, ...,X_n$ uma amostra aleatória de uma variável aleatória X que descreve uma característica de interesse de uma população com uma distribuição $f_x(x; \theta)$. Então, qualquer estatística $T = g(X_1, X_2, ...,X_n)$ é um *estimador pontual* de θ.

Notação: $\hat{\theta} = T(x) = g(x_1, x_2, ..., x_n)$ é a estimativa pontual de θ.

Propriedades dos estimadores pontuais

Estimador não-viesado (não-viciado)

O estimador T é dito um estimador não-viesado de θ se, sua média (ou esperança) for o próprio parâmetro que se pretende estimar, isto é,

$$E(T) = \theta.$$

Isso significa que os valores aleatórios de T ocorrem em torno do valor do parâmetro θ, o que é, obviamente, desejável.

Eficiência

Se T e T' são dois estimadores não-viesados de um mesmo parâmetro θ, e ainda

$$Var(T) < Var(T'),$$

então, o estimador T é dito *mais eficiente* do que o estimador T'.

Erro médio quadrático (erro quadrático médio - EQM)

Chamaremos de

$$e = T - \theta$$

o *erro amostral* que cometemos ao estimar o parâmetro θ da distribuição da variável aleatória X do estimador $T = g(X_1, X_2, ...,X_n)$, baseado na amostra $X_1, X_2, ...,X_n$.

Chama-se de *erro quadrático* médio (EQM) o valor

$$EQM(T) = E(e^2) = E[(T - \theta)^2].$$

Ou seja, EQM é a esperança do quadrado dos resíduos (a diferença entre a estimativa e o verdadeiro valor do parâmetro). Esta quantidade nos ajuda a avaliar a qualidade do estimador utilizado para estimar θ.

Assim, chamando de *precisão* à proximidade de cada observação de sua própria média enquanto que, a *acurácia* mede a proximidade de cada observação ao valor-alvo que se procura atingir; temos que, um estimador preciso tem variância pequena, mas pode ter EQM grande. Por outro lado, um estimador acurado é não-viesado e tem variância pequena, o que implica EQM pequeno.

Métodos para encontrar estimadores pontuais

Veremos agora alguns critérios propostos com a finalidade de resolver o problema de como escolher os estimadores mais adequados. Dentre eles, citaremos os métodos (ou princípios) da máxima verossimilhança e dos momentos.

Método da máxima verossimilhança

Este método desenvolvido por Ronald Fisher em 1920 é bastante empregado e funciona de forma a encontrar aquele valor do parâmetro θ que maximiza a probabilidade de obter a amostra observada, na ordem particular em que os elementos da mesma aparecem.

Exemplo: Suponha que temos *n* provas de Bernoulli com Pr(sucesso) = *p*, $0 < p < 1$ e X = número de sucessos. Devemos tomar como estimador aquele valor de *p* que torna a amostra observada a mais provável de ocorrer.

Suponha, por exemplo, que $n = 3$ e obtemos 2 sucessos e 1 fracasso. A função de verossimilhança é

$$L(p) = \Pr(2 \text{ sucessos e } 1 \text{ fracasso}) = p^2(1 - p).$$

Agora precisamos obter o máximo desta função. Isto é obtido através de derivação:

$$L'(p) = \frac{\partial p^2 (1-p)}{\partial p} = 2p(1 - p) - p^2 \rightarrow p(2 - 3p) = 0$$

do que seguem $p = 0$ ou $p = 2/3$. É fácil ver que o ponto de máximo é $\hat{p} = \frac{2}{3}$, que é o *estimador de máxima verossimilhança* (E.M.V.) de p.

Definição 3: ***Função de verossimilhança e estimador de máxima verossimilhança***

Uma variável aleatória X tem densidade f(x), e $x_1, x_2, ..., x_n$ os valores amostrais. Definimos a função de verossimilhança, L, como

$$L = f(X_1; \theta). f(X_2; \theta). ... . f(X_n; \theta)$$

Ou seja, o produto de cada uma das funções de probabilidade (ou funções de densidade) das variáveis $X_1, X_2, ..., X_n$.

O estimador de máxima verossimilhança de θ, baseado na amostra $X_1, X_2, ...,X_n$, é o valor de $\hat{\theta}$ de θ que maximiza L, considerada como uma função de θ para uma dada amostra $X_1, X_2, ...,X_n$.

Observação: Para se encontrar $\hat{\theta}$, podemos recorrer às técnicas de cálculo diferencial integral ou fazermos por inspeção da função L. Ao recorrermos às técnicas de cálculo, na maioria das vezes, torna-se mais fácil trabalhar com a transformação ln[L], e o valor que maximiza L é o mesmo que maximiza o ln[L].

Exemplo: Considerando o exemplo anterior, de modo geral, o EMV do parâmetro p de uma binomial, com X sucessos em n provas é $\hat{p} = \frac{X}{n}$.

Para se chegar nesse estimador, observe que a função de verossimilhança neste caso é

$$L(p) = p^x (1 - p)^{n-x},$$

e que o máximo dessa função ocorre no mesmo ponto que ln[L(p)]. Portanto,

$$\ln[L(p)] = x . \ln(p) + (n - x) . \ln(1 - p),$$

e derivando

$$\ln'[L(p)] = \frac{x}{p} - \frac{n - x}{n - p} = 0\,,$$

de onde obtemos $\hat{p} = \frac{X}{n}$.

Método dos momentos

Este método foi o primeiro a ser proposto e usado. Consiste em supor que os momentos da distribuição da população coincidem com os da amostra. Expressando os parâmetros populacionais a estimar em função dos momentos de ordem menor, obtém-se um sistema de equações cuja solução fornece as estimativas desejadas. Esse método produz, em geral, estimadores consistentes, mas que, muitas vezes, não são os mais eficientes.

Então basicamente o que se faz é montar um sistema de equações com tantas equações quantos forem os parâmetros a estimar. Assim, temos μ'_r o r-ésimo momento em torno de zero, isto é,

$$\mu'_r = E(X^r),$$

e M'_j o j-ésimo momento amostral em torno de zero, isto é,

$$M'_j = \frac{1}{n} \cdot \sum_i X_i^j$$

Podemos formar o conjunto de equações:

$$M'_1 = \mu'_1$$

$$M'_2 = \mu'_2$$

$$\vdots$$

$$M'_k = \mu'_k\,,$$

ou seja, $M'_r = \mu'_r$, com r = 1, 2, ..., k.

À solução desse sistema de equações chamamos de estimador de θ obtido pelo método dos momentos.

Exemplo: Considere uma amostra aleatória $X_1, X_2, ...,X_n$ de $X \sim N(\mu; \sigma^2)$ (Leia: X tem distribuição Normal com parâmetros $(\mu; \sigma^2)$). Faça $\theta = (\theta_1, \theta_2) = (\mu; \sigma^2)$. Estime μ e σ^2.

Solução:

$$\mu'_1 = E(X^1) = E(X) = \mu = \theta_1$$

$$\mu'_2 = E(X^2) = \sigma^2 + \mu^2 = \theta_2 + \theta_1^2$$

$$(I)\ M'_1 = \mu'_1 \rightarrow \frac{1}{n} \cdot \sum_i X_i = \mu \Rightarrow \hat{\mu} = \overline{X}$$

$$(II)\ M'_2 = \mu'_2 \rightarrow \frac{1}{n} \cdot \sum_i X_i^2 = \sigma^2 + \mu^2 \rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_i \left(X_i - \overline{X}\right)^2$$

Intervalos de Confiança (I.C.)

Até agora, todos os estimadores apresentados foram *estimadores pontuais*, isto é, especificam um único valor para o estimador. Este procedimento não permite julgar qual a possível magnitude do erro que estamos cometendo. Daí surge a idéia de construir os *intervalos de confiança* em torno da estimativa pontual, de modo que esse intervalo tenha uma probabilidade conhecida de conter o verdadeiro valor do parâmetro.

Ao intervalo que, com probabilidade conhecida, deve conter o valor real do parâmetro chamaremos *intervalo de confiança* para esse parâmetro. À probabilidade, que designaremos por $\gamma = 1 - \alpha$, de que um intervalo de confiança contenha o valor do parâmetro chamaremos *nível de confiança* ou *grau de confiança* do respectivo intervalo. Veremos que α é a *probabilidade de erro na estimação por intervalo*, isto é, a probabilidade de errarmos ao afirmar que o valor do parâmetro está contido no intervalo de confiança.

Exemplo: A estimativa pontual da média populacional μ é feita por um valor $\overline{X}$. Qualquer que seja a amostra, teremos um erro que será $\overline{X} - \mu$. De acordo com o Teorema do Limite Central, teremos

$$e = \overline{X} - \mu \sim N\left(0; \sigma^2_{\overline{x}}\right)$$

com $\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$. Daqui podemos determinar qual a probabilidade de conter erros de determinada magnitude. Por exemplo,

$$Pr(|e| < 1{,}96\ \sigma_x) = 0{,}95$$

ou

$$Pr(|\overline{X} - \mu| < 1{,}96\ \sigma_x) = 0{,}95$$

que é equivalente a

$$Pr(\mu - 1{,}96\ \sigma_x < X < \mu + 1{,}96\ \sigma_x) = 0{,}95 \qquad \textbf{(I)}$$

Esta afirmação probabilística pode ser escrita do seguinte modo:

$$Pr(\overline{X} - 1{,}96\ \sigma_x < \mu < \overline{X} + 1{,}96\ \sigma_x) = 0{,}95 \qquad \textbf{(II)}$$

Convém lembrar que μ não é uma variável aleatória mas um parâmetro, e a expressão (II) deve ser interpretada do seguinte modo: construídos todos os intervalos da forma $\overline{X} \pm 1{,}96\,\sigma_{\overline{x}}$, 95% deles conterão o verdadeiro valor do parâmetro μ.

Sorteada uma amostra e encontrada sua média $\overline{X}$, e admitindo conhecido $\sigma_{\overline{x}}$, podemos construir o intervalo

$$\overline{X} \pm 1{,}96\ \sigma_{\overline{x}}\,.$$

Este intervalo pode ou não conter o parâmetro μ, mas pelo exposto acima temos 95% de confiança, de que contenha.

Definição 4: Seja $(X_1, X_2, ..., X_n)$ uma amostra aleatória de uma população e θ o parâmetro de interesse. Se T é um estimador de θ, e conhecida distribuição amostral de T, sempre é possível achar dois valores t_1 e t_2, tal que

$$Pr(t_1 < \theta < t_2) = 1 - \alpha = \gamma$$

sendo γ um valor fixado e $0 < \gamma < 1$.

Para uma dada amostra, teremos dois valores fixos t_1 e t_2, e o intervalo de confiança para θ com nível de confiança γ é indicado do seguinte modo:

$$IC(\theta: \gamma) = [t_1, t_2]\,.$$

Intervalo de confiança para μ com $\sigma_2 = \sigma_0^2$ conhecido

O intervalo de confiança para μ com $100\gamma\%$ de confiança é dado por:

$$I.C.(\mu : \gamma) = \left[\overline{X} - z_{\alpha/2} . \frac{\sigma_0}{\sqrt{n}} ; \overline{X} + z_{\alpha/2} . \frac{\sigma_0}{\sqrt{n}} \right],$$

com $Pr(Z < -z) = Pr(Z > z) = \dfrac{\alpha}{2}$.

Lembrando que $z_{\alpha/2}$ é o valor da distribuição Normal padrão cuja área à direita é igual a $\frac{\alpha}{2}$.

Exemplo: Um metalúrgico fez quatro determinações do ponto de fusão do manganês resultando em (graus centígrados): 1 269, 1 271, 1 263 e 1 265. Vamos construir o intervalo de confiança para a média μ desta população assumindo que a amostra é aleatória e que o ponto de fusão do manganês é uma variável aleatória com distribuição normal (μ,25), ou seja, a variância é conhecida e igual a 25. Use $\alpha = 0{,}01$.

Assim, basta substituirmos as informações do problema em $\left[\overline{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \overline{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right]$. Temos $\overline{X}$ = 1 267 e o valor de Z obtido é igual a 2,576 fazendo a consulta à tabela da distribuição normal padrão.

Este valor é obtido através do valor $\alpha/2$ = 0,005 que é o tamanho da área à direita (ou à esquerda) da curva normal. Como a tabela solicita o tamanho da área sob a curva normal que vai do ponto central (zero) até o limite, devemos fazer 0,5 – 0,005 = 0,495 .

Assim,

$$\left[1\,267 - 2{,}576 \cdot \frac{5}{\sqrt{4}}; 1\,267 + 2{,}576 \cdot \frac{5}{\sqrt{4}}\right] = \left[1\,267 - 6{,}44; 1\,267 + 6{,}44\right] =$$

= (1 260,56 ; 1 273,44) são os limites do intervalo de confiança.

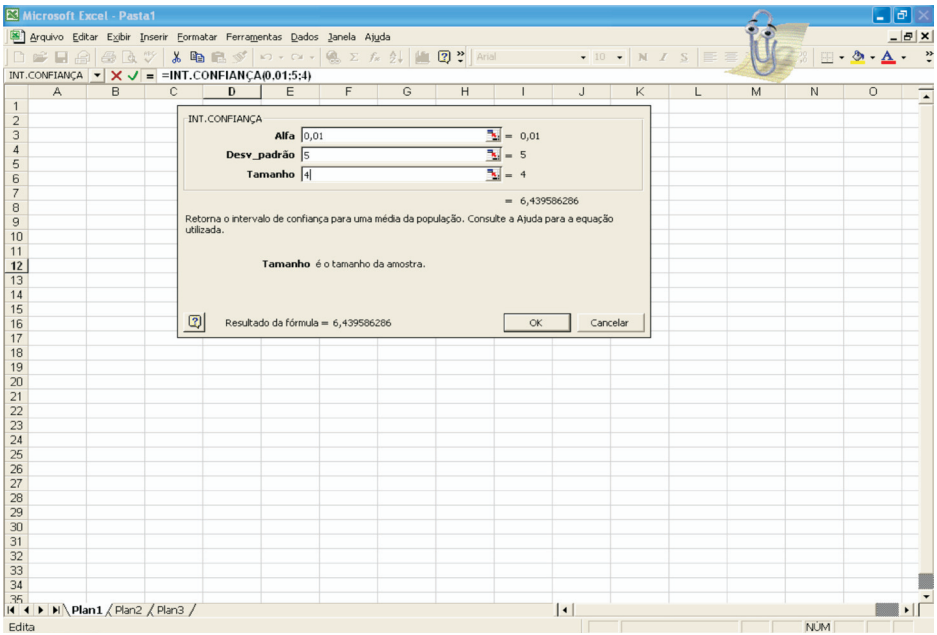
Resolvendo o problema com a planilha *Excel* poderíamos usar a função INT.CONFIANÇA fazendo:

Alfa, nível de significância empregado, neste caso igual a 0,01;

Desv_Padrão, desvio padrão conhecido, neste caso igual a 5;

Tamanho, tamanho da amostra, aqui temos n = 4 observações.

O resultado fornecido pela planilha é o *erro de estimativa* ou seja, $z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}$.



Intervalo de confiança para μ com σ² desconhecido

O intervalo de confiança para μ com 100γ % de confiança é dado por:

$$I.C.(\mu : \gamma) = \left[\bar{X} - t \cdot \frac{S}{\sqrt{n}} ; \bar{X} + t \cdot \frac{S}{\sqrt{n}} \right]$$

com $Pr(t_{(n-1)} < -t) = Pr(t_{(n-1)} > t) = \frac{\alpha}{2}$.

Lembrando que t_{n-1} é o valor da distribuição t de Student com n–1 graus de liberdade cuja área à direita é igual a $\frac{\alpha}{2}$.

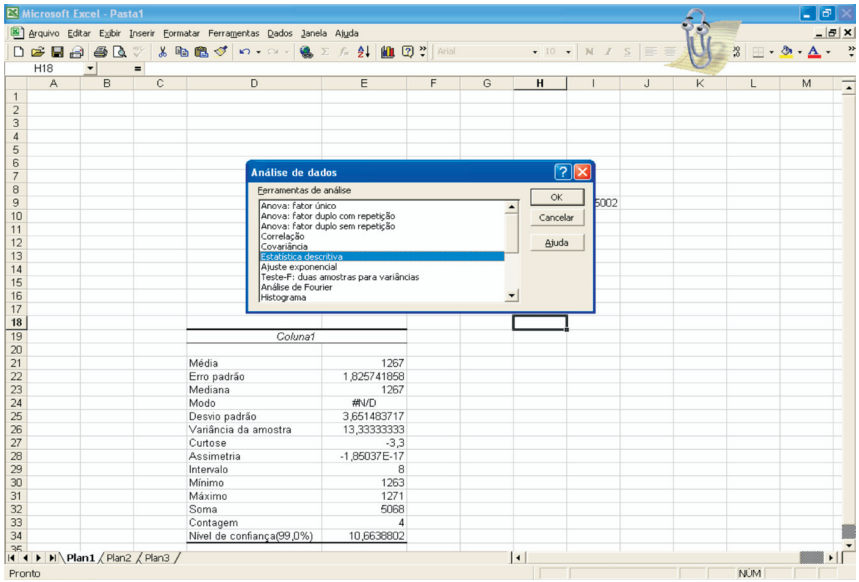
Portanto, agora, com a variância desconhecida usamos a tabela t de Student em vez da tabela Z.

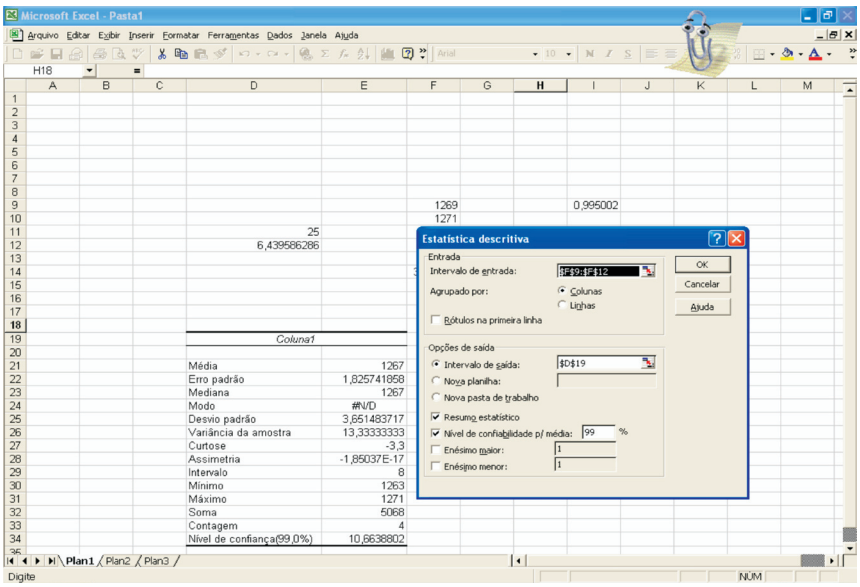
Exemplo: suponhamos agora, usando o problema resolvido acima, que a variância fosse na verdade desconhecida. Assim, teríamos que obter uma estimativa com base na amostra. Assim, teríamos S = 3,6514 (por meio da fórmula de variância: $\frac{\sum (x_i - \bar{X})^2 . f_i}{n-1}$) e o intervalo seria um pouco modificado pois t_{n-1} = 5,8408 consultando a tabela t de Student com nível de significância de 1%. Então o intervalo será:

$$\left[1\,267-5,8408\cdot\frac{3,6514}{\sqrt{4}};1\,267+5,8408\cdot\frac{3,6514}{\sqrt{4}}\right]=(1\,267-10,66;1\,267+10,66)=$$

$= (1\,256,34;\ 1\,277,66)$. Observe que neste caso o erro de estimativa é maior que quando consideramos a variância conhecida.

Na planilha *Excel*, uma forma de obter o intervalo acima é utilizando a ferramenta de Análise de dados (Estatística Descritiva), que fornece uma série de resultados a respeito da amostra:





Intervalo de confiança para a razão de variâncias σ_1^2/σ_2^2

O intervalo de confiança para σ_1^2/σ_2^2 com $100\gamma\%$ de confiança é dado por:

$$I.C.(\sigma^2 : \gamma) = \left[\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_2} ; \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_1} \right],$$

onde F_1 e F_2 são tais que, $Pr(F_{n_1-1;n_2-1} < F_1) = Pr(F_{n_1-1;n_2-1} > F_2) = \frac{\alpha}{2}$. Este intervalo é muito útil para verificarmos se duas populações são homogêneas. Para encontrar F_1, fazemos $Pr(F_{n_2-1;n_1-1} > \frac{1}{F_1}) = \frac{\alpha}{2}$.

Lembrando que $F_{n_1-1;n_2-1}$ é o valor da distribuição F com n_1-1 e n_2-1 graus de liberdade.

Exemplo: Queremos verificar se duas máquinas produzem peças com a mesma homogeneidade quanto à resistência à tensão. Para isso, sorteamos duas amostras de 6 peças de cada máquina, e obtivemos as seguintes resistências:

Máquina A	145	127	136	142	141	137
Máquina B	143	128	132	138	142	132

Vamos obter o intervalo de confiança para a razão das variâncias considerando um nível de significância de 10%. Primeiramente obtemos as variâncias dos dados acima. $S_1^2 = 40$ e $S_2^2 = 36{,}97$. Consultando a tabela F em anexo temos $F_1 = 0{,}198$ e $F_2 = 5{,}05$.

F_2 foi obtido primeiro alimentando a tabela com $n_1-1=5$ e $n_2-1=5$ graus de liberdade e o valor fornecido foi 5,05. Para obter F_1 fazemos $\frac{1}{5{,}05} = 0{,}198$.

Assim, temos o intervalo $\left[\frac{40}{36{,}97} \cdot \frac{1}{5{,}05} ; \frac{40}{36{,}97} \cdot \frac{1}{0{,}198} \right] = (0{,}214 ; 5{,}46)$.

Como o valor 1 está incluído no intervalo, isto significa que os dois grupos são homogêneos ou seja, as variâncias podem ser consideradas iguais.

Intervalo de confiança para proporção

Vamos agora obter um intervalo de confiança para *p*. Sabemos que X = número de sucessos nas *n* provas de Bernoulli, então X tem uma distribuição aproximadamente normal, com média $\mu = np$ e variância $\sigma^2 = n.p.(1-p)$. Conseqüentemente,

$$Z = \frac{X - n.p}{\sqrt{n.p.(1-p)}} \sim N(0; 1),$$

ou ainda,

$$Z = \frac{\frac{X}{n} - p}{\sqrt{p.(1-p)/n}} = \frac{\hat{p} - p}{\sqrt{p.(1-p)/n}} \sim N(0; 1).$$

Assim, o intervalo para P será

$$\left[\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}\right], \textbf{(III)}$$

onde z é tal que $Pr(Z < -z) = Pr(Z > z) = \frac{\alpha}{2}$.

Exemplo: Suponha que em n = 400 provas, obtemos k = 80 sucessos. Vamos obter um intervalo de confiança para *p*, com $\gamma = 0{,}90$.

Neste caso, $\hat{p} = 80/400 = 0{,}2$ e $(1-\hat{p}) = 1 - 0{,}2 = 0{,}8$, então, o intervalo de confiança, utilizando a expressão (III), é dado por:

$$0{,}2 \pm (1{,}64).\sqrt{\frac{(0{,}2).(0{,}8)}{400}} = 0{,}2 \pm 0{,}033,$$

ou seja,

$$IC(p: 90\%) = [\,0{,}167;\, 0{,}233].$$

Note que o valor $Z_{\alpha/2} = 1{,}64$ foi obtido consultando a tabela Z (Normal padrão) para um nível de significância de 10% ($1 - \gamma$). Distribui-se, neste caso, 5% de significância para cada lado do intervalo de confiança. Assim, na tabela, devemos procurar o valor 0,45 (0,5 – 0,05) que irá ser encontrado na linha 1,6 e na coluna 0,04, então $Z_{\alpha/2} = 1{,}64$.

Erro de Estimação e Tamanho das amostras

Acabamos de ver como construir intervalos de confiança para os principais parâmetros populacionais. Em todos os casos, supusemos dado o nível de confiança desses intervalos. Evidentemente, o nível de confiança deve ser fixado de acordo com a probabilidade de acerto que se deseja ter na estimação por intervalo. Sendo conveniente, o nível de confiança pode ser aumentado até tão próximo de 100% quanto se queira, mas isso resulta em intervalos de amplitude cada vez maiores, o que significa perda de precisão na estimação.

É claro que seria desejável termos intervalos com alto nível de confiança e pequena amplitude, o que corresponderia a estimarmos o parâmetro em questão com pequena probabilidade de erro e grande precisão. Isso, porém, requer uma amostra suficientemente grande, pois, para *n* fixo, confiança e precisão variam em sentido opostos.

Veremos a seguir como determinar o erro de estimação e o tamanho das amostras necessárias nos casos de estimação da média ou de uma proporção populacional.

O erro num intervalo de estimação diz respeito à diferença entre a média amostral e a verdadeira média da população. Como o intervalo tem centro na média amostral, o *erro máximo provável* é igual à metade da amplitude do intervalo (semi-amplitude).

Vimos que o intervalo de confiança para a média μ da população normal quando σ é conhecido tem semi-amplitude dada por:

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \textbf{(IV)}$$

Fixando *e* e *n* na expressão acima, podemos determinar α, o que equivale a determinar a confiança de um intervalo de amplitude conhecida. Podemos também, fixados α e *e*, determinar *n*, que é o problema da determinação do tamanho da amostra necessária para se realizar a estimação por intervalo com confiança e a precisão desejadas. Deste modo temos que,

$$n = \left(\frac{z_{\alpha/2}\cdot\sigma}{e}\right)^2 . \quad (A)$$

Está será a expressão usada para a determinação do tamanho da amostra necessária, se σ for conhecido.

Não conhecendo o desvio-padrão da população, deveríamos substituí-lo por sua estimativa S e usar a distribuição t de Student, ou seja, substituir σ por S e usar t de Student na expressão (IV). Ocorre, porém, que, não tendo ainda sido retirada a amostra, não dispomos, em geral, do valor de S. Se não conhecemos nem ao menos uma limitação superior para σ, a única solução é, então, colher uma *amostra-piloto* de tamanho *n'* e, com base nela, obtermos uma estimativa S, empregando, a seguir, a expressão

$$n = \left(\frac{t_{n'-1} . S}{e} \right)^2 . \text{(B)}$$

Se $n \leq n'$, a amostra-piloto é suficiente para a estimação. Caso contrário, deveremos retirar, ainda, da população, os elementos necessários à complementação do tamanho mínimo da amostra.

Procedemos de forma análoga se desejamos estimar uma proporção populacional com determinada confiança e dada precisão. Da expressão (III) podemos obter

$$n = \left(\frac{z_{\alpha/2}}{e} \right)^2 . p.(1-p) . \textbf{(V)}$$

O obstáculo à determinação do tamanho da amostra por meio da expressão (V) está em desconhecermos *p* e tampouco dispormos de sua estimativa $\hat{p}$, pois a amostra ainda não foi retirada. Essa dificuldade pode ser resolvida por meio de uma amostra-piloto, analogamente ao caso descrito na estimação de μ, ou analisando-se o comportamento do fator $p.(1-p)$ para $0 \leq p \leq 1$. Pode-se observar facilmente que $p.(1-p)$ é a expressão de uma parábola cujo ponto máximo é $p = ½$.

Desse modo, se substituirmos, na expressão (V), p.(1-p) por seu valor máximo, ¼, seguramente o tamanho de amostra obtido será suficiente para a estimação, qualquer que seja *p*. Isso equivale a considerar

$$n = \left(\frac{z_{\alpha/2}}{e} \right)^2 . \frac{1}{4} = \left(\frac{z_{\alpha/2}}{2e} \right)^2 . \textbf{(VI)}$$

Pelo mesmo raciocínio, se sabemos que seguramente $p \leq p_0 \leq ½$ ou $p \geq p_0 \geq ½$, podemos usar o limite p_0 em vez de *p*, na expressão (VI), obtendo um tamanho de amostra suficiente, pois teremos então $p.(1-p) \leq p_0.(1-p_0)$.

Evidentemente, usando-se a expressão (VI), corre-se o risco de dimensionar uma amostra bem maior do que a realmente necessária. Isso ocorrerá se *p* for, na realidade, próximo de 0 ou 1. Se o custo envolvido for elevado e proporcional ao tamanho da amostra, será desejável evitar que tal fato ocorra, sendo mais prudente a tomada de uma amostra-piloto. Inversamente, em muitos casos, é preferível, por simplificação, proceder conforme indicado, com base em uma limitação superior para o fator p.(1-p).

Exemplo: Qual o tamanho de amostra necessária para se estimar a média de uma população infinita cujo desvio-padrão é igual a 4, com 98% de confiança e precisão de 0,5?

Ao definirmos a precisão da estimativa desejada, estamos estabelecendo o erro máximo que desejamos cometer, com a confiança dada. Logo, essa precisão equivale numericamente à própria semi-amplitude do intervalo de confiança. Portanto, utilizando a experssão A dado que o desvio padrão é conhecido, temos:

$$n = \left(\frac{z_{\alpha/2} . \sigma}{e} \right)^2 = \left(\frac{2,33 . 4}{0,5} \right)^2 = 347,50.$$

O valor de $Z_{\alpha/2}$ =2,33 foi obtido consultando a tabela Z da distribuição normal padrão considerando $\alpha/2 = 0,01$. Devemos encontrar, na tabela, portanto, o valor referente à área 0,50 – 0,01 = 0,49.

Logo, necessitamos de uma amostra de 348 elementos.

Exemplo: Qual o tamanho de amostra suficiente para estimarmos a proporção de defeituosos fornecidos por uma máquina, com precisão de 0,02 e 95% de confiança, sabendo que essa proporção seguramente não é superior a 0,20?

Agora estamos estimando uma proporção e precisamos dimensionar uma amostra com 95% de confiança e margem de erro de 2%.

Então usando a expressão V, temos

$$n = \left(\frac{z_{\alpha/2}}{e} \right)^2 . p_0 . (1 - p_0) = \left(\frac{1,960}{0,02} \right)^2 . 0,20 . 0,80 = 1536,64$$

O valor de $Z_{\alpha/2}$ = 1,96 foi encontrado na tabela da distribuição normal padrão a partir do valor 0,5 – 0,025 = 0,475. Somando a linha 1,90 mais a coluna 0,06, obtemos 1,96 como sendo o valor crítico.

Logo, será suficiente uma amostra de 1 537 elementos.

Ampliando seus conhecimentos

Técnica Bootstrap

(BARROS, 2005)

O método Bootstrap foi originalmente proposto por Bradley Efron em um influente artigo publicado no Annals of Statistics, em 1979. Este método de simulação se baseia na construção de distribuições amostrais por reamostragem, e é muito utilizado para estimar intervalos de confiança de parâmetros, em circunstâncias em que outras técnicas não são aplicáveis, em particular no caso em que o número de amostras é reduzido. Esta técnica foi extrapolada para a resolução de muitos outros problemas de difícil resolução por meio de técnicas de análise estatística tradicionais (baseadas na hipótese de um elevado número de amostras). Pode ser utilizado, por exemplo, para estimar o viés e a variância de estimadores ou de testes de hipóteses calibrados. O método tem por base a idéia de que o pesquisador pode tratar sua amostra como se ela fosse a população que deu origem aos dados e usar amostragem com reposição da amostra original para gerar pseudoamostras. A partir destas pseudo-amostras, é possível estimar características da população, tais como média, variância, percentis, etc. Vários esquemas diferentes de simulação Bootstrap têm sido propostos na literatura e muitos deles apresentam bom desempenho em uma ampla variedade de situações.

Suponha disponível um conjunto de observações e o interesse em fazer inferências a respeito do parâmetro μ. Sabe-se que o estimador não viciado de μ é a média amostral $\overline{x}$ cujo erro padrão pode ser calculado por:

$$\text{Erro padrão da média} = \left[\frac{1}{n.(n-1)} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 \right]^{1/2}$$

Por outro lado, suponha o interesse em fazer inferência para algum outro parâmetro, como, por exemplo, o coeficiente de correlação, não há nenhuma fórmula analítica simples que permite calcular o seu erro padrão. O método de Bootstrap foi projetado para fazer simulações para este tipo de problema. A idéia básica da simulação Bootstrap é amostrar os dados originais com reposição, obtendo-se dados analíticos, em que, destes dados, calcula-se a estatística de interesse.

Repete-se este processo inúmeras vezes até a obtenção de B valores. Calcula-se o erro padrão destes valores e então, tem-se o erro padrão da estatística. Dado o custo alto e a escassez conseqüente de dados em muitas aplicações, combinadas com o custo reduzido e abundância do poder da computação, o método de Bootstrap se torna uma técnica muito atraente por extrair informações de dados empíricos.

Atividades de aplicação

1. Para encontrar o estimador de máxima verossimilhança de um parâmetro, devemos maximizar a função de verossimilhança através de que procedimento?

a) Derivando a função de verossimilhança.

b) Maximizando a probabilidade de sucesso.

c) Obtendo o valor da função que minimiza o erro.

d) Encontrando um estimador que não é tendencioso.

e) Aumentando o tamanho da amostra.

2. Foram sorteadas 15 famílias com filhos num certo bairro e observado o número de crianças de cada família, matriculadas na escola. Os dados foram: 1, 1, 2, 0, 2, 0, 2, 3, 4, 1, 1, 2, 0, 0, e 2. Obtenha as estimativas correspondentes aos seguintes estimadores da média de crianças na escola nesse bairro:

μ_1 = (mínimo + máximo)/2

$\mu_2 = (X_1 + X_2)/2$

$\mu_3 = \overline{X}$

Qual deles é o melhor estimador da média e por quê?

3. Suponha que X tenha distribuição N(μ,4). Uma amostra de tamanho 25 fornece média amostral $\overline{X} = 78{,}3$. Determine um intervalo de confiança de 99% para μ.

4. Registraram-se os valores 0,28; 0,30; 0,27; 0,33; 0,31 segundos, obtidos em 5 medições de tempo de reação de um indivíduo a um certo estímulo. Determine os limites de confiança de:

a) 95%;

b) 99% para o tempo médio de reação.

5. O fabricante de uma droga medicinal declarou que ela era 90% eficaz em curar uma alergia, em um período de 8 horas. Em uma amostra de 200 pessoas que tinham a alergia, a droga curou 160 pessoas. Determinar se a declaração do fabricante é legítima. Considere $\alpha = 0{,}01$.

6. O comprimento de certo tipo de eixo, produzido por uma indústria automobilística tem uma pequena variação de peça por peça. Sabe-se que o desvio padrão é de 4 mm. Uma amostra aleatória de 100 desses eixos forneceu um comprimento médio de 4,52 mm.

a) Construa o intervalo de confiança 90%, para a média do comprimento desses eixos.

b) Dê a sua interpretação para o intervalo encontrado. Será que podemos dizer que o intervalo encontrado tem probabilidade de 0,90 de conter a verdadeira média?

7. Interprete e comente as afirmações abaixo:

a) "A média de salário inicial para recém formados em Administração está entre 8 e 10 salários mínimos com 95% de confiança".

b) "Quanto maior for o tamanho da amostra, maior é a probabilidade da média amostral estar próxima da verdadeira média".

8. Desejamos coletar uma amostra de uma variável aleatória X com distribuição Normal de média desconhecida e variância 30. Qual deve ser o tamanho da amostra para que, com 0,92 de probabilidade, a média amostral não difira da média da população por mais de 3 unidades?

9. Numa pesquisa de mercado, desejamos estimar a proporção de pessoas que compram determinada marca de xampu.

a) Que tamanho de amostra deveremos ter para que, com probabilidade 0,90, a proporção amostral não se desvie do verdadeiro valor por mais de 0,05?

b) Se tivermos a informação adicional de que a aceitação do xampu é no mínimo 0,8, qual deve ser então o tamanho da amostra?

Testes de Hipóteses: conceitos

Introdução

Os testes estatísticos são regras de decisões, vinculadas a um fenômeno da população, que nos possibilitam avaliar, com o auxílio de uma amostra, se determinadas hipóteses (suposições, conjecturas, algo qualquer que um pesquisador esteja estabelecendo) podem ser rejeitadas, ou não.

No campo da Inferência Estatística, a busca por respostas acerca de certas características de uma população estudada é de fundamental impotância. Apenas com base nessas características é que se devem estabelecer regras e tomar decisões sobre qualquer hipótese formulada no que se refere à população. Dessa forma, escolhida uma variável X e colhida uma amostra aleatória da população, podemos estar interessados em inferir a respeito de alguns de seus parâmetros (média, variância e proporção, por exemplo) e, também, sobre o comportamento da variável (a sua distribuição de probabilidade). A realização de testes de hipóteses nos fornece meios para que possamos, com determinado grau de certeza, concluir se os valores dos parâmetros ou mesmo a distribuição associados à população considerada, podem representá-la de forma satisfatória. Nesse contexto, temos os Testes Paramétricos, vinculados à estimação dos valores dos parâmetros e os Testes de Aderência, associados à busca da distribuição de X. Na verdade, quando realizamos Testes Paramétricos, esses estão intimamente ligados aos Testes de Aderência, pois, para se obter a "determinada certeza" citada, é necessário que saibamos qual a distribuição de probabilidade que melhor se ajusta às estimativas observadas por intermédio das amostras.

A maior parte das ciências se utiliza da técnica estatística denominada Teste de Hipóteses. Podemos citar algumas suposições: a roleta de certo cassino é honesta; a propaganda de um produto veiculada na televisão surtiu o efeito desejado; uma ração desenvolvida para certo animal proporcionou um ganho maior de peso do que aquela já utilizada há anos; vale a pena trocar as máquinas desta indústria por outras, mais modernas; qual medicamento é mais eficaz no tratamento de certa doença; a metodologia empregada na educação infantil está associada ao aprendizado.

A teoria geral da construção e análise de testes de hipóteses é um capítulo muito importante da Estatística. Seus fundamentos teóricos foram desenvolvidos por Neyman e Pearson, e o método usual de obtenção de testes é o *método da razão de verossimilhança*.

Vamos supor que exista uma hipótese, a qual é considerada válida até prova em contrário, referente a um dado parâmetro da população. Essa hipótese é testada com base em resultados amostrais, sendo aceita ou rejeitada, conforme veremos a seguir.

Sob diversos aspectos, o problema dos testes de hipóteses é o oposto ao da estimação, mas há também vários pontos que são comuns aos dois casos. A estimação é feita com base em uma variável convenientemente escolhida, função dos elementos da amostra, denominada *estimador*. Nos problemas de teste de hipóteses, nossas conclusões baseiam-se em variáveis calculadas a partir da amostra ou amostras disponíveis. Os mesmos critérios para a escolha de bons estimadores, em problemas de estimação, vão agora nos orientar na escolha da variável de teste adequada. Por exemplo, vimos que a média amostral $\overline{X}$ é o estimador da média populacional μ. Então, pelas mesmas razões, se desejarmos testar uma hipótese referente ao verdadeiro valor da média μ da população, a variável de teste mais adequada será $\overline{X}$.

A seguir, introduzimos a idéia de teste de hipóteses por meio de um exemplo hipotético que, partindo de uma situação simples, será gradualmente ampliado para atender à situação geral de teste de hipóteses.

Exemplo: Suponha que uma indústria compre de certo fabricante parafusos cuja carga média de ruptura por tração é especificada em 50 kg e o desvio padrão das cargas de ruptura é suposto igual a 4 kg e independente do valor médio.

O comprador deseja verificar se um grande lote de parafusos recebidos deve ser considerado satisfatório. Existe alguma razão para se temer que esse lote possa ser formado por parafusos, cuja carga média de ruptura seja inferior a 50 kg, o que seria indesejável. Por outro lado, o fato de a carga média de ruptura ser eventualmente superior a 50 kg não preocupa o comprador, pois, nesse caso, os parafusos seriam de qualidade superior à especificada.

Então, o comprador adota o seguinte critério para decidir se concorda em comprar o lote ou se prefere devolvê-lo ao fabricante: tomar uma amostra

aleatória de 25 parafusos do lote e submetê-los a ensaio de ruptura; se a carga média de ruptura observada nessa amostra for maior ou igual a 48kg, ele comprará o lote; caso contrário, ele se recusará a comprar.

Esse comprador está testando a hipótese de que a carga média de ruptura dos parafusos do lote seja 50kg, contra a alternativa de que ela seja inferior a 50kg.

Suponha que, depois de realizado o teste, nós afirmássemos que a *população* dos valores da carga de ruptura tem realmente $\mu = 50$kg. Poderíamos estar errados nessa afirmação? A resposta é sim, o que levaria o comprador a aceitar um lote abaixo das especificações exigidas. Então, para melhor entendermos a regra de decisão adotada, é interessante estudarmos os tipos de erros que podemos cometer.

Podemos cometer dois tipos de erro:

Erro tipo I : rejeitar o lote de parafusos quando, na verdade, o lote era satisfatório, isto é, rejeitar quando realmente $\mu = 50$ kg.

Erro tipo II : aceitar o lote de parafusos quando, na verdade, o lote não era satisfatório, isto é, aceitar quando $\mu < 50$ kg.

O erro tipo I, levaria o comprador a deixar de adquirir um lote perfeitamente satisfatório e o erro tipo II, levaria o comprador a adquirir um lote insatisfatório, com prejuízo à produção.

Conceitos Fundamentais

Consideremos uma amostra $(X_1, X_2, ..., Xn)$ de uma variável aleatória que descreve uma característica de interesse de uma população. Seja $\hat{\theta}$ um estimador (uma estatística) de um parâmetro θ dessa população.

Hipótese nula e Hipótese alternativa

Uma *hipótese estatística*, que denotaremos por *H*, é qualquer afirmação sobre a população em estudo. Em geral, o que nos interessa são as afirmações sobre os parâmetros da população.

Usualmente, vamos decidir entre duas hipóteses, uma bastante específica a respeito do valor do parâmetro, chamada de *hipótese nula* e denotada

por H_0; e a segunda fornecendo uma alternativa mais geral, chamada de *hipótese alternativa* e denotada por H_1.

Suponha, por exemplo, que desejamos testar a afirmação de que o parâmetro θ da população é igual a um valor qualquer θ_0. Neste caso, as hipóteses são definidas de acordo com o interesse da pesquisa e podemos estabelecer testes específicos conforme o objetivo do pesquisador. Por exemplo:

a) Teste Bilateral (Bicaudal) : $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

Note que o objetivo desse teste é decidir se o parâmetro populacional não difere de θ_0, não importando se θ é maior ou menor do que θ_0.

ou

b) Teste Unilateral à Direita: $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$

Esse teste tem por finalidade verificar se, o parâmetro não só difere de θ_0, mas também, se é maior do que θ_0. Objetivamente, poderíamos citar uma pesquisa que visa verificar se um determinado candidato a prefeito, conseguiu aumentar sua intenção de votos após a realização de um debate com seu adversário realizado pela televisão.

ou ainda

c) Teste Unilateral à Esquerda $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$

Esse teste tem por finalidade verificar se o parâmetro não só difere de θ_0, mas, também, se é menor do que θ_0. Nesse contexto, poderíamos estabelecer uma Regra de Decisão para verificar, por exemplo, se o retorno de investimento de determinado fundo é menor do que θ_0 . Pois, se for menor, não é recomendado continuarmos investindo nesse fundo.

Erros Tipo I e Tipo II

A hipótese nula, H_0, pode ser falsa ou verdadeira. Entretanto, o processo de sua rejeição ou aceitação é diferente daquele usado para provar uma proposição matemática que também é falsa ou é verdadeira. Em contraste, há sempre um grau de incerteza na decisão tomada a respeito de uma hipótese estatística. Esse é o preço a ser pago por estarmos trabalhando em uma situação em que a variabilidade é inerente.

- *Erro tipo I*: rejeitar H_0 quando esta é verdadeira.

- *Erro tipo II*: não rejeitar H_0 quando esta é falsa.

A probabilidade de se cometer um erro tipo I depende dos valores dos parâmetros da população e é designada por α. O valor de α, para H_0 verdadeira, é chamado *nível de significância do teste*; isto é, o nível de significância de um teste é a probabilidade com que desejamos correr o risco de um erro tipo I. O resultado da amostra é cada vez mais significante para rejeitar H_0 quanto menor for o nível α. Usualmente, esses valores são fixados em 5%, 1% ou 0,1%.

A probabilidade de se cometer um erro tipo II é designada por β. A determinação do valor β já é mais difícil, pois, usualmente não se especificam valores fixos para o parâmetro na situação alternativa. Podemos atribuir alguns valores, escolhidos dentro do caso alternativo, e encontrar o valor correspondente de β.

O esquema a seguir mostra os erros que podemos cometer e suas probabilidades.

Situação específica na população (realidade)			
Decisão		H_0 verdadeira	H_0 falsa
	aceita H_0	correto	erro tipo II
		$(1-\alpha)$	(β)
	rejeita H_0	erro tipo I	correto
		(α)	$(1-\beta)$

Deve-se notar que as probabilidades α e β são condicionadas à realidade. Fica claro, também, no esquema, que o erro tipo I só pode ser cometido se H_0 for verdadeira, e o erro tipo II, se H_0 for falsa. Da mesma forma, o erro tipo I só pode ser cometido se H_0 for rejeitada e o erro tipo II, se H_0 for aceita.

O erro tipo I é controlado pelo pesquisador, e é ele que define a margem de erro que está disposto a correr. Existem vários fatores que influenciam na escolha do nível de significância. Em pesquisas, como nas ciências exatas, biológicas, agronômicas, em que as variáveis são mais fáceis de mensurar, os instrumentos de medida são confiáveis, o controle de fatores intervenientes é razoável, o conhecimento da área é maior, a gravidade das conseqüências do erro menor, entre outros, permitem um maior rigor e, portanto, pode-se ser mais exigente, diminuindo o nível de significância. Contudo, em pesquisas, nas ciências humanas, que lida com pessoas, com construtos polêmicos, instrumentos ainda não testados, as conseqüências do erro não são tão graves, podendo ser mais flexível. Via de regra, usa-se o nível de 5%.

Região Crítica

A faixa de valores da variável de teste que leva à rejeição de H_0 é denominada *região crítica* (RC) do teste. A faixa restante constitui a *região de aceitação*.

Esta região é construída de modo que P($\theta \in$ RC dado que H_o verdadeira) seja igual a α, um número fixado.

Se o valor observado da estatística pertençe a RC, rejeitamos H_0; caso contrário, não rejeitamos H_0.

Poder de um teste

Definida uma hipótese H_0 sobre um parâmetro $\theta = \theta_0$, e determinada a região crítica RC para sua estatística $\hat{\theta}$, a *função poder do teste* $\beta(\theta)$ indica a probabilidade de uma decisão correta, segundo as diversas alternativas do parâmetro, e pode ser usada para se decidir entre dois testes, indicando qual deles é melhor para testar uma mesma hipótese.

Regra de Decisão

Vamos tomar o seguinte exemplo referente ao rendimento bruto de um certo fundo de investimentos. Poderíamos criar uma Regra de Decisão com base em $\alpha = 0,01$ e H_1: $\mu < 1,71\%$. Assim, poderíamos estabelecer a seguinte regra: caso coletarmos uma amostra cujo resultado observado for menor do que 1,67%, decidiremos por rejeitar H_0, pois a probabilidade disso ocorrer é menor do que $\alpha = 0,01$. Ou seja, sob a referência (α=0,01), a amostra coletada deverá ser vista como rara se a hipótese nula for verdadeira (H_0 : $\mu = 1,71\%$). Conseqüentemente, seria mais conveniente optarmos por afirmar que $\mu < 1,71\%$.

É interessante refletir sobre a seguinte pergunta: o valor 1,67% é menor do que 1,71%? Obviamente que perguntando desta forma todos diriam que sim. Porém, antes que saibamos como esses resultados foram obtidos, a melhor resposta seria: depende. Considere, então, as seguintes reflexões:

1. Se medíssemos os rendimentos de dois fundos do tipo A e B, da mesma maneira e obtivéssemos, respectivamente, 1,67% e 1,71%. Concluiríamos que A é, de fato, pior do que B;

2. Se o interesse for descobrir e comparar o rendimento médio de dois fundos (A e B), poderíamos obter essas médias de várias maneiras. Vejamos dois casos:

a) com a coleta das duas populações, as médias obtidas seriam as médias verdadeiras, ou seja, os valores paramétricos (μ_A e μ_B). Assim, diríamos novamente que 1,67% é menor do que 1,71%.

b) coletando-se a população de A e uma amostra de B, e obtidas as médias $\mu_A = 1{,}67\%$ e $\bar{x}_B = 1{,}71\%$, não poderíamos afirmar com absoluta certeza que 1,67% é menor do que 1,71%. Pois, sabemos que $\bar{X}$ é uma variável aleatória e apenas com base no comportamento de $\bar{X}_A$ é que poderíamos decidir se, provavelmente, $\mu_A < \mu_B$. Assim, se tanto no fundo A quanto no fundo B, ou nos dois, forem coletadas amostras, a resposta para a questão proposta sempre dependerá do comportamento das estimativas das possíveis amostras. Comportamento esse, representado por meio de uma distribuição de probabilidades e, portanto, toda decisão a respeito da questão virá acompanhada de um grau de incerteza. A Inferência Estatística, por intermédio do Teste de Hipóteses, visa responder a essa questão.

Passos para a construção de um teste de hipóteses

Daremos abaixo, uma seqüência que pode ser usada sistematicamente para qualquer teste de hipóteses sobre um parâmetro populacional θ.

Passo 1: Definir qual a hipótese nula, H_0, a ser testada e qual a hipótese alternativa H_1.

Passo 2: Escolher a estatística de teste (estimador) adequada que será usada para julgar a hipótese nula H_0.

Passo 3: Escolher o nível de significância α e estabelecer a região crítica.

Passo 4: Calcular o valor da estatística de teste com base em uma amostra de tamanho *n* extraída da população.

Passo 5: Rejeitar H_0 se o valor calculado da estatística pertencer à região crítica. Não rejeitar H_0 se o valor calculado da estatística não pertencer à região crítica.

Valor p (p-valor)

É a probabilidade de cometer o erro de tipo I (rejeitar H_0 quando ela é verdadeira), com os dados de uma amostra específica. Este valor é calculado pelo *software* estatístico, assim o comparamos com o nível de significância escolhido e tomamos a decisão. Se o p-valor for menor que o nível de significância escolhido rejeitamos H_o, caso contrário, não rejeitamos H_0.

Testes de hipóteses não-paramétricos

A Estatística não-paramétrica pode ser definida como uma coleção de métodos estatísticos aplicada a conjuntos de dados em que as suposições distribucionais necessárias para aplicação de uma técnica clássica (Intervalo de Confiança, Teste de Hipótese) não são satisfatoriamente atendidas. É também bastante útil no tratamento de dados nos quais o nível de mensuração das observações não é dos melhores.

Tais procedimentos são usados há muitos anos, embora não com o nome atual. O rei Nabucodonossor aplicou informalmente o teste da permutação, 600 anos a.C. Cálculos da probabilidade binomial foram feitos em 1710 pelo médico inglês Arbuthnott.

O primeiro livro-texto denotado aos métodos não-paramétricos foi escrito por Siegel (1956). No entanto, Savage designa o ano de 1936 como o verdadeiro início da Estatística não-paramétrica, marcado pela publicação do artigo de Hotelling e Pabst sobre correlação por postos.

O tema central em Estatística é a chamada *Inferência Estatística* que aborda dois tipos de problemas fundamentais: a estimação de parâmetros de uma população, e o teste de hipóteses. Na Inferência Estatística procuramos tirar conclusões sobre um grande número de eventos com base na observação de apenas parte deles. Os testes relacionados a Inferência Estatística nos dizem qual a margem de diferença que deve ser encontrada na amostra para que possamos afirmar que elas representam realmente diferenças nos tratamentos (grupos). Como nesses procedimentos, na verdade são testadas hipóteses a respeito dos parâmetros populacionais, esses são chamados de "Paramétricos".

Algumas técnicas não são tão rigorosas na especificação de condições acerca dos parâmetros da população da qual a amostra foi obtida.

Conseqüentemente, as conclusões não são tão poderosas quanto às obtidas por técnicas paramétricas. Essas técnicas são chamadas de "distribuição livre" ou "não-paramétricas".

Vantagens e desvantagens

Vantagens

- Dispensam normalidade dos dados.
- O p-valor é exato (no caso paramétrico o cálculo do p-valor se baseia numa distribuição de probabilidade teórica).
- São testes mais simples.
- São úteis quando é difícil estabelecer uma escala de valores quantitativos para os dados.
- São mais eficientes que os paramétricos quando não existe normalidade.

Desvantagens

- Proporcionam um desperdício de informações, já que em geral não consideram a magnitude dos dados.
- Quando as suposições do modelo estatístico são atendidas são menos eficientes que os paramétricos.
- A utilização das tabelas dos testes é mais complicada.

Escolha do teste estatístico adequado

É importante a definição de critérios que nos ajudem a decidir qual o teste ideal para determinado problema.

Um desses critérios, sem dúvida, é o Poder do Teste $(1-\beta)$. O teste que apresenta uma maior probabilidade de rejeitar H_0 quando H_0 é falsa, entre todos os testes de nível α, deve ser escolhido. Mas só isto não basta e nem sempre é simples de ser obtido, portanto precisamos de outras informações para escolher o teste mais adequado:

- Como foi obtida a amostra, ou seja, o plano experimental.

- Natureza da população (pessoas, objetos, áreas, animais, etc.).
- Tipo de mensuração dos dados (escala de mensuração).

Quando se usa um teste paramétrico existe uma série de pressupostos a serem verificados, além do nível mínimo de mensuração exigido ser a escala intervalar.

Quando essas suposições não são verificadas é possível que o teste nos leve a resultados errôneos.

No caso não-paramétrico, o primeiro critério a ser verificado deve ser o nível de mensuração dos dados.

Nível de Mensuração

a) Escala Nominal

É o mais baixo nível de mensuração. Utiliza símbolos ou números simplesmente para distinguir elementos em diferentes categorias (como um nome), não havendo entre eles, geralmente, possibilidade de comparação do tipo maior-menor, melhor-pior.

Exemplos:

- Masculino (M), Feminino (F)
- Perfeito (1), Defeituosa (0)
- Europeu (1), Americano(2), Africano (3), Asiático(4)

b) Escala Ordinal

Utiliza números apenas para classificar elementos numa ordem crescente ou decrescente. Existe assim algum tipo de relação entre as categorias embora a diferença entre elas seja de difícil quantificação.

Exemplos*:*

- Classes sócio-econômicas – (A, B, C, D, E)
- Patentes do Exército – (soldado, cabo, sargento, etc)
- Opinião de um determinado produto – (Ruim, Regular, Bom, Muito bom, Excelente)

c) Escala Intervalar (Intervalo de medida)

Ocorre quando a escala tem as características da escala ordinal e ainda é possível quantificar a diferença entre dois números desta escala.

Exemplo: Temperatura, Peso, Altura, Rendimentos

Observação: Alguns autores apontam ainda a existência de outra escala: a Escala de Razão, equivalente a escala intervalar, porém o valor zero é o verdadeiro ponto de origem.

Principais planos experimentais

Existem algumas situações que podem ser consideradas as mais freqüentes no cotidiano de quem aplica técnicas estatísticas para analisar dados amostrais. São os planos experimentais que orientam o pesquisador à condução do seu estudo, seguindo os princípios da metodologia científica. Podemos considerar abaixo, os planos mais comuns:

Caso de uma amostra

Neste plano nosso interesse é verificar se determinada amostra pode provir de uma população especificada. São usualmente conhecidos como testes de aderência ou bondade do ajuste. Neste caso, retira-se uma amostra aleatória e compara-se a distribuição amostral com uma distribuição de interesse. Os principais testes utilizados nesse caso são:

- Teste Z;
- Teste t de Student;
- Teste Qui-quadrado;
- Teste de Kolmogorov-Smirnov;
- Teste de Lilliefors.

Caso de duas amostras relacionadas

Muitas vezes estamos interessados na comparação de dois tratamentos. No entanto é muito comum ocorrer uma grande disparidade entre os elementos dos grupos. Para evitar que um grupo de indivíduos seja natu-

ralmente superior ao outro, é comum proceder algum tipo de pareamento entre os indivíduos. O tipo mais comum de pareamento é utilizando cada indivíduo como seu próprio controle, submetendo-o aos dois tratamentos em ocasiões diferentes. Outro tipo de pareamento é tentar selecionar, para cada par, indivíduos que sejam tão semelhantes quanto possível. Por exemplo: gêmeos, órgãos (ouvidos, braços, pés etc.). São também conhecidos como testes do tipo "antes-depois". Os principais testes são:

- Teste t para amostras dependentes;
- Teste de McNemar;
- Teste de Wilcoxon.

Caso de duas amostras independentes

Estes testes se aplicam a planos amostrais em que se deseja comparar dois grupos independentes. Esses grupos podem ter sido formados de duas maneiras diferentes:

a) Extraiu-se uma amostra da população A e outra amostra da população B.

b) Indivíduos da mesma população foram alocados aleatoriamente a um dos dois tratamentos em estudo.

Diferente do caso de dados pareados, não se exige que as amostras tenham o mesmo tamanho. Os principais testes são:

- Teste Z;
- Teste t de Student para amostras independentes;
- Teste Qui-quadrado;
- Teste de Mann-Whitney.

Caso de k amostras relacionadas

Neste tipo de plano são comparados 3 ou mais grupos (tratamentos) relacionados entre si. Imagine que *n* indivíduos sejam observados, cada um, em 3 ou mais momentos tendo sido registrada a sua respectiva evolução. Então teremos a seguinte estrutura de dados:

	Tratamentos				
Indivíduo	**1**	**2**	**3**	**...**	**k**
1	X_{11}	X_{21}	X_{31}	...	X_{k1}
2	X_{12}	X_{22}	X_{32}	...	X_{k2}
3	X_{13}	X_{23}	X_{33}	...	X_{k3}
...	...	...	...	...	...
n	X_{1n}	X_{2n}	X_{3n}	...	X_{kn}

Onde as unidades amostrais utilizadas no experimento foram avaliadas sob as *k* condições de avaliação ou tratamentos (tempo, dietas, distância etc.). Os principais testes são:

- Análise de Dados Longitudinais;
- Teste de Friedman.

Caso de k amostras independentes

Neste tipo de plano são comparados 3 ou mais grupos (tratamentos) independentes entre si, cada grupo pode ter um número diferente de observações. Os principais testes são:

- Análise de Variância (ANOVA);
- Teste de Kruskal-Wallis.

Ampliando seus conhecimentos

Apresentação dos resultados dos testes

(CAMPOS. 2007)

Uma vez realizados os testes adequados, estes dão o seu parecer, sob a forma de um valor numérico, apresentado (conforme o teste) como valor de ***F*** (análise de variância), de ***t*** (teste t, de Student), ***U*** (Mann-Whitney), ***Q*** (teste de Cochran), χ^2 (letra grega qui, testes diversos, que usam o chamado qui-quadrado), z (McNemar e Wilcoxon), ***H*** (Kruskal-Wallis), ou ρ (letra grega rho, utilizada nos testes de correlação).

Não-significância estatística (H_0)

Em todos os casos, o valor numérico calculado pelo teste deve ser confrontado com valores críticos, que constam em tabelas apropriadas a cada teste. Essas tabelas geralmente solicitam duas informações, que permitem localizar o valor crítico tabelado: nível de significância (usualmente 5 % ou 1 %), e o número de graus de liberdade das amostras comparadas.

Valores menores que o tabelado indicam que ele não pode ser considerado diferente do que se obteria se as amostras comparadas fossem iguais. Enfim, estaria configurado o que se chama de não-significância estatística, ou de aceitação da hipótese nula (H_0).

Significância estatística (H_1)

Porém, se o valor calculado for igual ou maior que o tabelado, aceita-se a chamada hipótese alternativa (H_1), ou seja, a hipótese de que as amostras comparadas não podem ser consideradas iguais, pois o valor calculado supera aquele que se deveria esperar, caso fossem iguais, lembrando sempre que a igualdade, em Estatística, não indica uma identidade. Isso quer dizer que pode eventualmente haver alguma diferença, mas esta não deve ultrapassar determinados limites, dentro dos quais essa diferença decorre apenas da variação natural do acaso, típica da variação entre as repetições do ensaio. No caso de o valor calculado ser maior do que o valor tabelado, diz-se que há significância estatística, que pode ser ao nível de 5 %, se o valor calculado for maior que o valor tabelado para 5 %. Ou ao nível de 1 %, caso o valor calculado seja igual ou maior que o valor tabelado para 1 %.

Abaixo segue uma tabela que resume as conclusões que devem ser tomadas em relação a cada p-valor observado:

P ≥ 0,10	Não existe evidência contra H_0
P < 0,10	Fraca evidência contra H_0
P < 0,05	Evidência signficativa
P < 0,01	Evidência altamente significativa
P < 0,001	Evidência extremamente significativa

Atividades de aplicação

1. Nas situações descritas abaixo, descreva qual é a população, a amostra, o parâmetro de interesse e o tipo de teste que poderiam ser usados para estimar o parâmetro de interesse:

a) Para avaliar a proporção de alunos do Curso X favoráveis a eliminação da disciplina de Estatística do currículo, selecionou-se aleatoriamente 80 alunos do curso.

b) Para avaliar a eficácia de um curso que orienta como fazer boa alimentação e exercícios físicos, selecionou-se uma amostra aleatória de 20 pessoas obesas de uma certa cidade.

c) Para avaliar uma campanha contra o fumo, conduzida pela prefeitura de uma cidade, acompanhou-se uma amostra aleatória de 100 fumantes.

2. Com o objetivo de avaliar se o desempenho de um certo candidato, numa apresentação em público, foi positivo, selecionou-se uma amostra de uma grande platéia, indagando a cada um, sua opinião sobre o candidato, antes e depois da apresentação: se melhorou ou piorou.

a) Apresente as hipóteses nula e alternativa.

b) Se, numa amostra de 11 pessoas, 8 passaram a ter uma opinião mais favorável, enquanto 3 passaram a ter opinião menos favorável sobre o candidato, o que se pode afirmar com base somente nessas informações?

c) Se, numa amostra de 200 pessoas, 130 passaram a ter melhor impressão, enquanto 70 pioraram sua impressão sobre o candidato, o que se pode afirmar?

d) Qual o tipo de teste mais adequado para analisar estes dados?

3. Para avaliar o efeito de um brinde nas vendas de determinado produto, planeja-se comparar as vendas em lojas que vendem o produto com o brinde, com as vendas em lojas que não oferecem o brinde. Para reduzir o efeito de variações devidas a outros fatores, as lojas foram agrupadas em pares, de tal forma que as lojas de um mesmo par são as mais similares possíveis, em termos, por exemplo, do volume

de vendas, localidade, identidade de preços etc. Em cada par de lojas, uma passou a oferecer o brinde e a outra não.

a) Apresente as hipóteses nula e alternativa;

b) Os resultados das vendas, em quantidade de unidades vendidas, foram os constantes na tabela a seguir. Com base nesses dados, responda se os mesmos mostram alguma evidência para se afirmar que a oferta do brinde aumentou as vendas.

Par de loja	Vendas sem brinde	Vendas com brinde
1	33	43
2	43	39
3	26	33
4	19	32
5	37	43
6	27	46

c) Qual o tipo de teste mais adequado para analisar estes dados?

4. Fez-se uma pesquisa junto a 83 diretores das maiores agências de propaganda canadenses, a fim de se determinar a eficácia relativa de comerciais de 15 segundos em relação à dos comerciais de 30 segundos. Em uma escala de 5 pontos (1 = excelente e 5 = fraco), os entrevistados avaliaram os comerciais de TV de 15 e 30 segundos quanto a conscientização da marca, memorização da idéia principal, persuasão da capacidade de relatar uma história emocional. Observe a tabela abaixo com os resultados do estudo e responda as seguintes perguntas:

a) Qual a hipótese nula e a hipótese alternativa?

b) Que testes estatísticos poderiam ser aplicados nesse caso e qual o nível de significância mais indicado?

c) O que se pode observar a respeito dos resultados obtidos?

Classificação média de comerciais de 15 e 30 segundos quanto às 4 variáveis de comunicação

	Conscientização da marca		Memorização da idéia básica		Persuasão		Capacidade de relatar uma história emocional	
Comerciais	15 s	30 s	15 s	30 s	15 s	30 s	15 s	30 s
Escore médio	2,5	1,9	2,7	2,0	3,7	2,1	4,3	1,9

Testes de Hipóteses

Introdução

Apresentaremos, neste capítulo, os testes de hipóteses mais utilizados do ponto de vista paramétrico e não-paramétrico. Os testes paramétricos exigem que seja verificada a pressuposição de que os dados coletados sejam normalmente distribuídos enquanto que os testes não-paramétricos não fazem essa exigência e por isso são considerados menos consistentes, sendo, porém, uma alternativa a ser usada caso os pressupostos de normalidade não sejam observadas ou, ainda, quando o tamanho da amostra não é suficientemente grande. No caso paramétrico, como o nome já diz, o objetivo é testar hipóteses acerca de parâmetros, com base em dados amostrais. No caso não-paramétrico, as hipóteses não são formuladas em termos de parâmetros, já que não há preocupação com a distribuição que os dados seguem. Para cada tipo de plano experimental existem testes específicos a serem utilizados. Nos preocuparemos aqui com os seguintes planos: a) comparação de duas amostras independentes; b) comparação de duas amostras relacionadas; c) comparação de três ou mais amostras independentes; d) teste de aderência.

Comparação de duas amostras independentes

Neste caso estamos interessados em comparar duas populações, representadas cada uma por suas respectivas amostras. Não necessariamente as duas amostras têm o mesmo tamanho. Os principais testes são:

- Teste t de Student para médias;
- Teste Z para proporções;
- Teste Mann-Whitney (não-paramétrico)

Teste t de Student para comparação de médias

A média de uma população é uma de suas características mais importantes. É muito comum desejarmos tomar decisões a seu respeito, por exemplo,

quando são comparadas duas amostras ou dois tratamentos. Considere as seguintes hipóteses:

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 < \mu_2$$

ou

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 > \mu_2$$

ou ainda

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

As duas primeiras situações definem os chamados testes *unilaterais*, por que a região de rejeição está somente em uma das caudas da distribuição. A última situação define os testes *bilaterais*, no qual a região de rejeição se distribui igualmente em ambas as caudas da distribuição.

Assim, se estivermos interessados em mostrar que um parâmetro é significativamente superior ou inferior a um determinado valor, teremos que realizar um teste unilateral e teremos uma única região de rejeição, do tamanho do nível de significância fixado. Mas se, no entanto, estivermos interessados em mostrar que um determinado parâmetro é diferente de um determinado valor (sem especificar se inferior ou superior) teremos que realizar um teste bilateral e a região de rejeição será dividida em duas partes iguais, nas extremidades da curva do teste, em que cada região de rejeição terá metade do nível de significância.

Dessa forma, para realização do teste, deveremos primeiramente estimar a média e o desvio padrão de cada uma das amostras envolvidas e calcular a estatística do teste:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad \textbf{(1)}$$

a qual tem distribuição t de Student com $n_1 + n_2 - 2$ graus de liberdade. Nesse caso, supõe-se que *as variâncias amostrais são diferentes*. Caso as variâncias *não sejam diferentes*, devemos usar:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \textbf{(2)}$$

onde:

– $\overline{X}_1$ e $\overline{X}_2$ são as médias amostrais do grupo 1 e 2 respectivamente;

– S_1 e S_2 são os desvios padrões do grupo 1 e 2 respectivamente;

– n_1 e n_2 são os tamanhos de amostra do grupo 1 e 2 respectivamente;

$$S_p^2 = \frac{(n_1-1).S_1^2+(n_2-1).S_2^2}{n_1+n_2-2}$$

A tabela abaixo resume o procedimento a ser seguido:

Tabela 1. Decisão nos testes de comparação de médias

Hipóteses	Decisão
$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 < \mu_2$	rejeita H_0 se, $t < -t(\alpha)_{n_1+n_2-2}$
$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 > \mu_2$	rejeita H_0 se, $t > t(\alpha)_{n_1+n_2-2}$
$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$	rejeita H_0 se, $\lvert t \rvert > t(\alpha/2)_{n_1+n_2-2}$

Exemplo: Um teste de resistência a ruptura feito em seis cabos usualmente utilizados acusou resistência média de 3 530kg com variância de 660kg. Um novo cabo foi testado e verificou-se uma resistência média de 3 560kg e variância de 600kg em uma amostra de tamanho 8. Compare as médias dos dois cabos, ao nível de significância α = 5%. E se a variância do cabo novo fosse 850kg?

Assim, queremos testar se $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$. O teste é bilateral pois se deseja verificar se os dois cabos diferem em relação à resistência média, sem especificar para que lado. Usaremos a expressão (2), pois vamos considerar as variâncias "iguais" (ou seja, muito próximas). Rigorosamente, essa verificação deveria ser feita através da aplicação do teste F para razão de variâncias. Considerando válida essa suposição de igualdade das variâncias, teremos:

$$S_p^2 = \frac{(6-1).660+(8-1).600}{6+8-2} = 625 \quad \text{e} \quad t = \frac{(3530-3560)}{25\sqrt{\frac{1}{6}+\frac{1}{8}}} = -2,22.$$

O valor crítico $t(\alpha/2)_{n_1+n_2-2}$ para α = 5% é dado por 2,179. Este valor é encontrado na tabela t de Student consultando a coluna 0,025 (pois o teste é bilateral) e a linha 12 ($n_1 + n_2 - 2$). Assim, teremos 2 valores críticos, –2,179 e

+2,179. Como t < –2,179, rejeitamos a hipótese nula e afirmamos que existe diferença significativa entre os dois tipos de cabo. Os dois cabos diferem significativamente em relação à resistência média.

Agora, considerando que $S_2^2 = 850$kg teremos,usando a expressão (1):

$$t = \frac{(3530 - 3560)}{\sqrt{\frac{660}{6} + \frac{850}{8}}} = -2{,}04$$

e, neste caso, a nossa decisão será exatamente o contrário do que obtivemos, ou seja, como t > –2,179 não rejeitamos a hipótese nula e não observamos diferença entre os cabos.

Teste Z para comparação de proporções

Em alguns estudos, o interesse está em comparar duas proporções provenientes de amostras distintas. Nesse caso, obtém-se n_1 observações da população 1 e n_2 observações da população 2. Verifica-se em cada uma das amostras o total x_1 e x_2, respectivamente, de "sucessos" e calculam-se as proporções amostrais $p_1 = \frac{x_1}{n_1}$ e $p_2 = \frac{x_2}{n_2}$. As hipóteses testadas são as seguintes:

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 < P_2$$

ou

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 > P_2$$

ou ainda

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 \neq P_2$$

A estatística do teste é dada por:

$$Z = \frac{p_1 - p_2}{S_p} \quad \textbf{(3)}$$

Onde $S_p = \sqrt{\frac{p.(1-p)}{n_1} + \frac{p.(1-p)}{n_2}}$ **(4)** e $p = \frac{n_1.p_1 + n_2.p_2}{n_1 + n_2}$ **(5)**

Exemplo: Em uma cidade do interior realizou-se uma pesquisa eleitoral com 200 eleitores, na qual o candidato a presidente X aparece com 35%

das intenções de voto. A mesma pesquisa também foi realizada na cidade vizinha, com 500 eleitores, e o mesmo candidato surge com 28% das intenções de voto. Podemos afirmar estatisticamente que na primeira cidade o candidato X apresenta uma maior intenção de voto? (nível de significância $\alpha = 0,05$)

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 > P_2$$

É um teste unilateral pois está claramente verificado se na primeira pesquisa foi encontrada uma proporção maior do que na segunda cidade.

Pela expressão (5) temos $p = \frac{(200.0,35)+(500.0,28)}{200+500} = 0,3$ e pela expressão (4)

$$S_p = \sqrt{\frac{0,3.(1-0,3)}{200}+\frac{0,3.(1-0,3)}{500}} = 0,038 \text{ e finalmente:}$$

$$Z = \frac{0,35-0,28}{0,038} = 1,84$$

Ao nível de significância de 5% temos Z (α) = 1,64. Este valor crítico é obtido na tabela da distribuição normal padrão, considerando uma área marcada em cinza de tamanho 0,45, ou seja, 0,5 – 0,05. Localizando o valor 0,45 no corpo da tabela (ou o valor mais próximo), veremos que ele se localiza na linha 1,6 e na coluna 0,04. Então, somamos os dois valores e obtemos 1,64.

Como a estatística Z calculada é superior ao valor crítico, rejeitamos a hipótese nula. Existem evidências para admitir que na primeira cidade o candidato X apresenta uma proporção significativamente superior de intenção de voto.

Teste não-paramétrico de Mann-Whitney

Esse teste se aplica na comparação de dois grupos independentes, para se verificar se pertencem ou não à mesma população. É a alternativa a ser usada quando as suposições de normalidade não são verificadas. Considere, portanto, duas amostras de tamanho n_1 e $n_{2,}$ respectivamente. O teste consiste basicamente na substituição dos dados originais pelos seus respectivos postos ordenados (*ranks*) e cálculo da estatística do teste. Além disso, o

procedimento de teste depende do tamanho das amostras. Considere o grupo 2 aquele com o maior número de observações:

- Quando $9 \leq n_2 \leq 20$, calcula-se:

  $U = n_1.n_2 + \frac{n_1.(n_1+1)}{2} - R_1$, onde R_1 é a soma dos postos atribuídos aos valores do grupo 1.

- $n_2 > 20$

  Utiliza-se nesse caso a aproximação normal dada por:

$$\mu_U = \frac{n_1.n_2}{2} \qquad \sigma_U = \sqrt{\frac{n_1.n_2.(n_1+n_2+1)}{12}} \qquad z = \frac{U-\mu_U}{\sigma_U}$$

Os valores da estatística calculada são comparados com os valores críticos obtidos a partir de uma tabela (Mann Whitney). Caso a estatística U calculada seja inferior ao valor crítico deveremos rejeitar a hipótese nula.

Exemplo: Dois tipos de solução química, A e B, foram ensaiadas para determinação de Ph. As análises de amostras de cada solução estão apresentadas na tabela que segue. Verifique se há diferença entre elas.

$H_0: Ph_A = Ph_B$

$H_a: Ph_A > Ph_B$

A	Posto (A)	B	Posto (B)
7,49	13	7,28	2
7,35	4,5	7,35	4,5
7,54	19	7,52	17,5
7,48	11	7,50	14,5
7,48	11	7,38	7
7,37	6	7,48	11
7,51	16	7,31	3
7,50	14,5	7,22	1
7,52	17,5	7,41	8
		7,45	9
	R_A = 112,5		**R_B =77,5**

$$U = (9.10) + \frac{(9.10)}{2} - 112,5 = 22,5$$

O valor crítico para $n_1 = 9$ e $n_2 = 10$ em que $\alpha = 0{,}05$ (teste unilateral) será $U_c = 24$. Como o valor calculado da estatística é inferior ao valor crítico então iremos rejeitar H_0. Assim, temos evidências suficientes para afirmar que a solução química A apresenta Ph superior à solução química B.

Comparação de duas amostras relacionadas

Neste caso estamos interessados em comparar uma amostra extraída em dois momentos distintos. Deseja-se verificar se a diferença observada entre os dois momentos (efeito do tratamento) é significativa. Os principais testes são:

- Teste t de Student para dados pareados;
- Teste de Wilcoxon (não-paramétrico)

Teste t para dados pareados

Para observações pareadas, o teste apropriado para a diferença entre as médias das duas amostras consiste em primeiro determinar a diferença **d** entre cada par de valores e então testar a hipótese nula de que a média das diferenças na população é zero. Então, do ponto de vista de cálculo, o teste é aplicado a uma única amostra de valores **d**.

A diferença média para um conjunto de observações pareadas é $\bar{d} = \frac{\sum d}{n}$

e o desvio padrão das diferenças das observações pareadas é dado por:

$$S_d = \sqrt{\frac{\sum d^2 - n\bar{d}^2}{n-1}}$$

e a estatística do teste será: $t = \frac{\bar{d}}{S_d / \sqrt{n}}$ **(6)**

Essa estatística deve ser comparada com o valor crítico do teste t de Student para determinado nível de significância α e n–1 graus de liberdade.

Exemplo: Considere o experimento realizado com 10 automóveis de certa fábrica. Os veículos foram avaliados com dois tipos de combustíveis. Primeiramente, um combustível sem aditivo e em seguida o mesmo combustível com aditivo. Deseja-se verificar se os automóveis conseguem uma quilo-

metragem maior com a utilização do combustível com aditivo. Seguem os dados abaixo:

Automóvel	Quilometragem sem aditivo (B)	Quilometragem com aditivo (A)	d (A–B)
1	26,2	26,7	0,5
2	25,2	25,8	0,6
3	22,3	21,9	-0,4
4	19,6	19,3	-0,3
5	18,1	18,4	0,3
6	15,8	15,7	-0,1
7	13,9	14,2	0,3
8	12,0	12,6	0,6
9	11,5	11,9	0,4
10	10,0	10,3	0,3
Total	**174,6**	**176,8**	**2,2**

$$H_0: \mu_A = \mu_B \quad vs \quad H_a: \mu_A < \mu_B$$

Pelos dados da tabela temos $\bar{d}$ =0,22 e Sd = 0,361

Assim, $t = \dfrac{0,22}{\dfrac{0,361}{\sqrt{10}}} = 1,927$ e comparando com o valor crítico t (0,05) com 9 graus de liberdade que é 1,833, podemos concluir que o valor calculado se encontra dentro da região de rejeição, ou seja, existe diferença significativa entre as quilometragens obtidas com e sem aditivo. A quilometragem obtida com aditivo é significativamente superior.

Note que o valor crítico 1,833 foi encontrado na tabela t de Student na coluna 0,05 (pois o teste é unilateral) e linha 9.

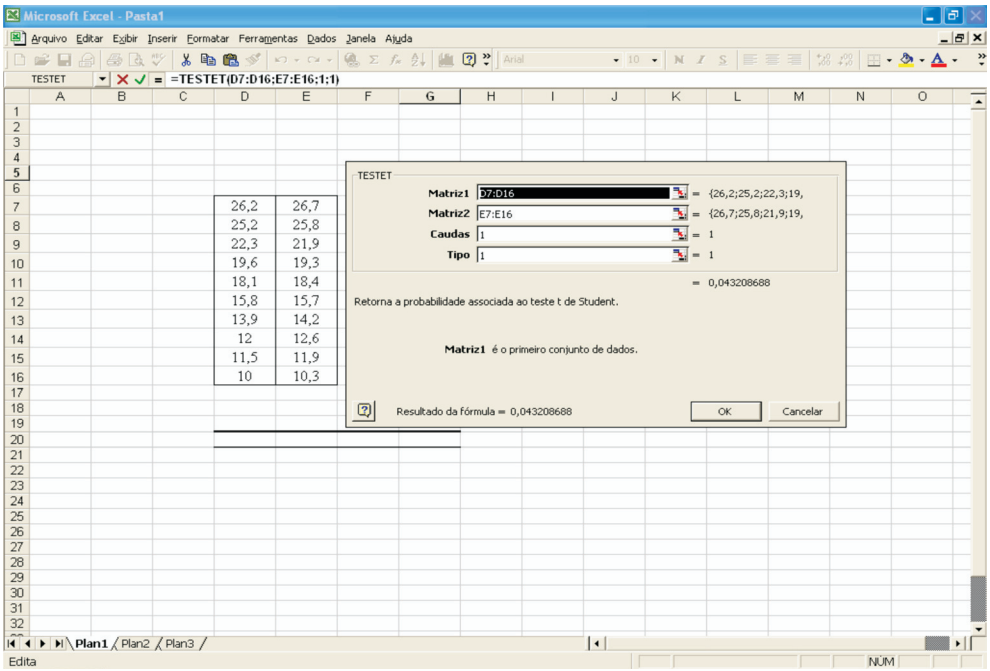
Com a planilha *Excel*, é possível realizar diversos testes de significância estatística, desde que se possuam os dados brutos. Para resolver esse exemplo, usaríamos a função TESTET, considerando:

Matriz 1: conjunto de dados referente ao primeiro grupo;

Matriz 2: conjunto de dados referente ao segundo grupo;

Caudas: indica se o teste é unilateral (1) ou bilateral (2). No caso, aqui o teste é unilateral;

Tipo: indica o tipo do teste, se é pareado (1) ou de amostras independentes (2 ou 3). No caso, aqui o teste é pareado.



Observe que a planilha irá fornecer p–valor = 0,0432, que, comparado com o nível de significância de 0,05, indica a existência de diferença significativa.

Teste de Wilcoxon

Neste teste não-paramétrico, devemos considerar as diferenças di's, onde $di = Y_i - X_i$. Devemos ordenar os di's, atribuindo postos do menor para o maior, sem considerar o sinal da diferença (em módulo). A continuação do teste, a partir daqui, depende do tamanho da amostra:

- $n < 25$

Considere T sendo a menor soma dos postos de mesmo sinal. Compara-se então o valor de T calculado com aqueles tabelados. O objetivo é testar se a mediana é nula, ou seja,

H_0: Mediana = 0

H_a: Mediana > 0

Mediana < 0

Mediana ≠ 0

Iremos rejeitar a hipótese nula quando o valor calculado de T for inferior ao valor crítico definido pelo nível de significância.

- $n \geq 25$

Nesse caso, T tem distribuição aproximadamente normal e podemos usar a aproximação considerando:

$$\mu_T = \frac{N.(N+1)}{4} \qquad e \qquad \sigma_T = \sqrt{\frac{N.(N+1).(2N+1)}{24}}$$

Calcula-se assim a estatística $z = \frac{T-\mu_T}{\sigma_T}$ e compara-se com os valores tabelados da distribuição de Z (Normal Padrão).

Podem ocorrer alguns empates. Nesse caso, deveremos considerar duas situações:

- Quando $X_i = Y_i$, ou seja, a informação pré equivale à informação pós para um mesmo indivíduo, descarta-se esse par da análise e redefinimos **n** como sendo o número de pares, tais que $X_i \neq Y_i$ para i = 1, 2, 3, ... , n.
- Quando duas ou mais di's tem o mesmo valor, atribui-se como posto a média dos postos que seriam atribuídos a eles caso não ocorresse empate.

Exemplo:

Di	|di|	Postos	Cálculo para Empates
-5	5	2*	$\rightarrow \frac{1+2+3}{3}$
5	5	2*	
5	5	2*	
7	7	4	
10	10	5	
-13	13	6,5**	$\rightarrow \frac{6+7}{2} = 6,$
13	13	6,5**	
15	15	8	

Exemplo: Numa pesquisa realizada em dois momentos distintos em 11 empresas operadoras de telefonia celular, investigou-se o % de clientes que avaliaram positivamente cada uma delas:

	% de avaliação positiva		di	|di|	p
Operadora	1° momento	2° momento			
1	8,7	7,7	1,0	1,0	4
2	18,6	9,6	9,0	9,0	9
3	8,0	16,0	–8,0	8,0	6
4	12,9	13,4	–0,5	0,5	2
5	10,9	9,6	1,3	1,3	5
6	13,4	13,0	0,4	0,4	1
7	11,9	23,7	–11,8	11,8	11
8	14,3	6,2	8,1	8,1	7
9	20,0	9,6	10,4	10,4	10
10	14,4	13,8	0,6	0,6	3
11	6,6	15,1	–8,5	8,5	8

Aplicando o teste de Wilcoxon, testaremos as seguintes hipóteses:

$$H_0 : \mu_T = 0 \text{ vs } H_a : \mu_T \neq 0$$

Somando-se os postos associados a diferenças negativas, teremos T = 6 + 2 + 11 + 8 = 27. O valor crítico, consultando a linha n = 11 e $\alpha = 0,05$ é igual a 13 (na verdade, o nível de significância aqui acaba sendo um valor próximo de 0,05, mais precisamente, 0,0471). Assim, não podemos rejeitar H_0, ou seja, a porcentagem de avaliação positiva não se modificou nos dois momentos.

Comparação de 3 ou mais amostras independentes

Esse tipo de plano é uma extensão do caso em que duas amostras independentes estão sendo comparadas, mas agora para o caso de 3 ou mais amostras. Se houver pelo menos um par de amostras diferentes, o teste irá apontar diferença significativa. No caso paramétrico, a opção é o teste F de Snedecor, também chamado de Análise de variância ou Anova. Mais uma vez aqui não há necessidade de os grupos que estarão sendo comparados terem tamanhos de amostras iguais. Consideremos, então, a seguinte estrutura de dados:

Tratamentos				
1	**2**	**3**	**...**	**k**
X_{11}	X_{21}	X_{31}	...	X_{K1}
X_{12}	X_{22}	X_{32}	...	X_{K2}
X^{13}	X_{23}	X_{33}	...	X_{K3}
..	...	...	...	...
X_{1n_1}	X_{2n_2}	X_{3n_3}	...	X_{Kn_K}

Análise de Variância

Uma análise de variância permite que vários grupos sejam comparados a um só tempo, utilizando variáveis contínuas. O teste é paramétrico (a variável de interesse deve ter distribuição normal) e os grupos têm que ser independentes. As hipóteses testadas são as seguintes:

$H_0 : \mu_1 = \mu_2 = ... = \mu_k$ vs H_1 : pelo menos um par $\mu_i \neq \mu_j$, para $i \neq j$

Os elementos que compõem o cálculo da Anova são sumarizados na tabela abaixo:

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Entre grupos	SQA	k – 1	QMA = $SQA/k-1$	QMA/QME
Erro amostral	SQE	N – k	QME = $SQE/N-k$	
Total	**SQT**	**N – 1**		

$$SQA = \sum \frac{T_k^2}{n_K} - \frac{T^2}{N} \quad \textbf{(7)} \quad \text{e} \quad SQT = \sum_{i=1}^{n}\sum_{k=1}^{k} X^2 - \frac{T^2}{N} \quad \textbf{(8)} \quad \text{e} \quad SQE = SQT - SQA$$

- T_k é a soma dos valores de um certo tratamento k;
- n_k é o número de observações no tratamento k;
- T^2 é a soma de todos os valores amostrados elevada ao quadrado;
- N é o número total de observações;
- X é cada observação amostrada.

O valor calculado de F é comparado com o valor crítico, definido pelo nível de significância e pelos graus de liberdade k – 1 e N – k. Caso $F_{cal} > F_{crit}$, devemos rejeitar a hipótese nula.

Exemplo: Quinze pessoas que participaram de um programa de treinamento são colocadas, de forma aleatória, sob três diferentes tipos de ensino. Os graus obtidos no exame de conclusão do treinamento são apresentados abaixo. Teste a hipótese de que não existe diferença significativa entre os 3 métodos de instrução, a um nível de significância de 5%.

Métodos de instrução		
A_1	A_2	A_3
86	90	82
79	76	68
81	88	73
70	82	71
84	89	81

$H_0 : \mu_1 = \mu_2 = \mu_3$ vs H_1 : pelo menos um par $\mu_i \neq \mu_j$, para $i \neq j$ $i, j = 1, 2, 3$.

Analisando a tabela acima, obtemos as seguintes informações:

$n_1 = n_2 = n_3 = 5$

$T_1 = 400$ $T_2 = 425$ $T_3 = 375$ $T = 1\,200$

$T_1^2 = 160\,000$ $T_2^2 = 180\,625$ $T_3^2 = 140\,625$ $T = 1\,440\,000$

Calculando as expressões (7) e (8):

$$SQA = \sum \frac{T_k^2}{n_K} - \frac{T^2}{N} = \left(\frac{160\,000}{5} + \frac{180\,625}{5} + \frac{140\,625}{5} \right) - \frac{1\,440\,000}{15} = 250$$

$$SQT = \sum_{i=1}^{n} \sum_{k=1}^{k} X^2 - \frac{T^2}{N} = 96\,698 - 96\,000 = 698$$

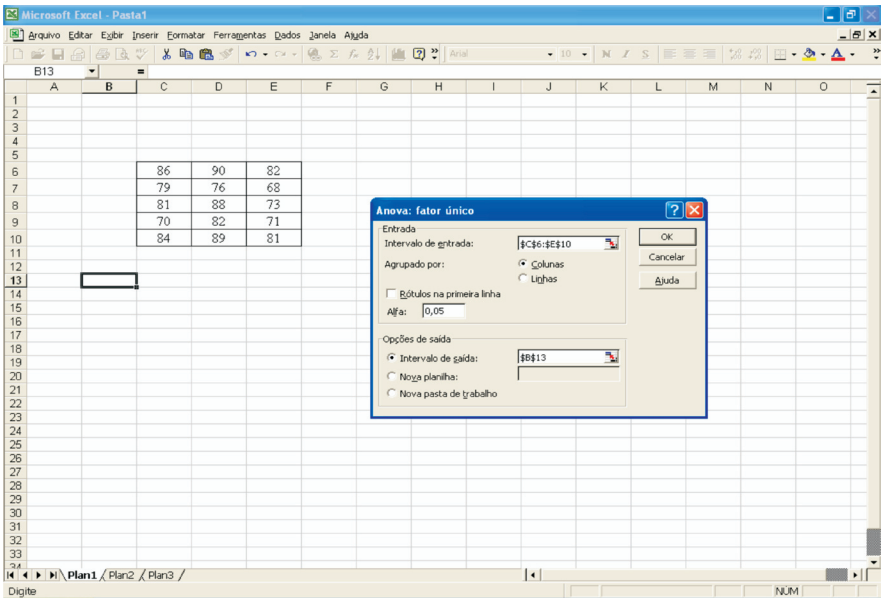
$SQE = 698 - 250 = 448$

A tabela da Anova fica então:

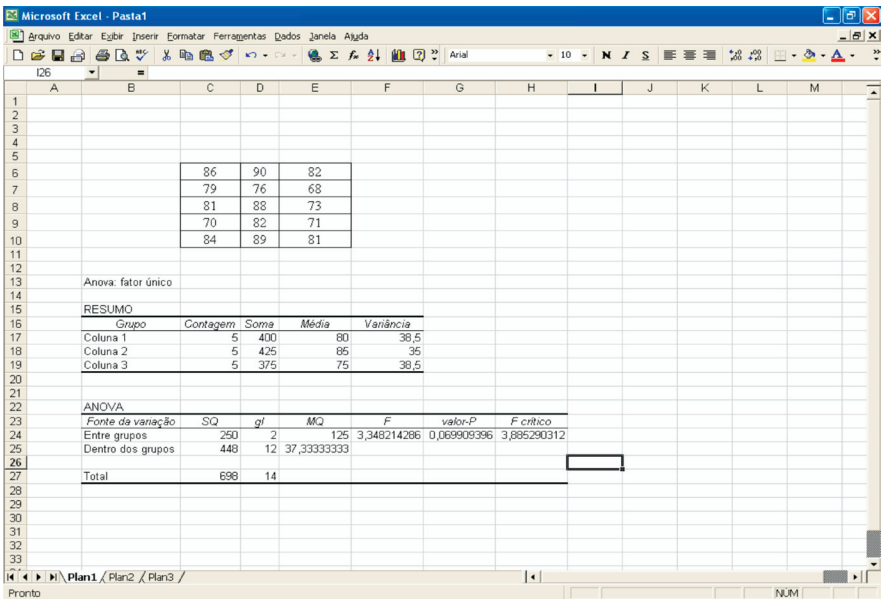
Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Entre grupos	250	2	125	3,35
Erro amostral	448	12	37,33	
Total	**698**	**14**		

Comparando o valor de F calculado com o valor crítico de 3,89, que é obtido considerando-se $\alpha = 0,05$ e cruzando a coluna $n_1 = 2$ e linha $n_2 = 12$ (graus de liberdade), podemos concluir que não há diferença significativa entre os métodos de instrução.

Com a planilha *Excel*, selecionamos FERRAMENTAS E ANÁLISE DE DADOS e selecionamos a opção: Anova: fator único.



A planilha nos fornecerá o seguinte resultado:



Teste de Kruskal-Wallis

Outro teste útil na comparação de **k** tratamentos independentes é o teste de Kruskal-Wallis. Ele nos indica se há diferença entre pelo menos dois deles. É na verdade uma extensão do teste de Wilcoxon para duas amostras independentes e se utiliza dos postos atribuídos aos valores observados.

Primeiramente, deve-se atribuir um posto a cada valor observado, sempre atribuindo o menor posto ao menor valor e o maior posto ao maior valor. Após se efetuar a soma dos postos para cada tratamento (R_j) calcula-se a estatística H:

$$H = \frac{12}{N.(N+1)} \cdot \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3 \cdot (N+1)$$

onde $\mathbf{n_j}$ é o número de observações do j-ésimo tratamento, **N** é o total de observações e $\mathbf{R_j}$ é a soma de postos do tratamento j.

Compara-se o valor calculado H com o valor crítico, que é definido pelo nível de significância e pelos tamanhos de amostra n_1, n_2, ..., n_k. Caso o valor de H calculado seja superior ao valor crítico, rejeita-se H_0.

Exemplo: Numa pesquisa sobre qualidade de vinho, foram provados três tipos por cinco degustadores. Cada degustador provou 12 amostras (4 de cada tipo) e atribuiu a cada uma delas uma nota de zero a dez. As médias das notas atribuídas pelos 5 degustadores a cada uma das amostras foram:

Tipo 1	Posto	Tipo 2	Posto	Tipo 3	Posto
5,0	1	8,3	7	9,2	11
6,7	2	9,3	12	8,7	9
7,0	4	8,6	8	7,3	5
6,8	3	9,0	10	8,2	6

Vamos verificar se há preferência dos degustadores por algum dos tipos de vinho.

H_0: não existe preferência por algum tipo de vinho

H_1: existe pelo menos uma diferença nas comparações realizadas entre os vinhos.

Calculando-se a estatística do teste, considerando $R_1 = 10$, $R_2 = 37$ e $R_3 = 31$

$$H = \frac{12}{12.13} \cdot 607,5 - 3.(12+1) = 7,73$$

O valor crítico ao nível de significância de 5% é 5,6923. Este valor é obtido na tabela fazendo $n_1 = 4$, $n_2 = 4$ e $n_3 = 4$. O nível de significância é precisamente 0,049. Desta forma, rejeitamos a hipótese nula. Certamente o vinho tipo 1 é considerado inferior pelos degustadores.

Testes de aderência

Estes testes são úteis para verificar se determinada amostra pode provir de uma população ou distribuição de probabilidade especificada. São usualmente conhecidos como testes de aderência ou bondade do ajuste. Nesse caso, retira-se uma amostra aleatória e compara-se à distribuição amostral com a distribuição de interesse.

Teste Qui-quadrado

É um teste amplamente utilizado em análise de dados provenientes de experimentos, em que o interesse está em observar freqüências em diversas categorias (pelo menos duas).

É uma prova de aderência útil para comprovar se a freqüência observada difere significativamente da freqüência esperada. Está geralmente especificada por uma distribuição de probabilidade.

Para utilizar o teste, não devemos ter mais de 20% das freqüências esperadas abaixo de 5 e nenhuma freqüência esperada igual a zero. Para evitar freqüências esperadas pequenas, devem-se combinar as categorias até que as exigências sejam atendidas.

Após definirmos a hipótese nula, testamos se as freqüências observadas diferem muito das freqüências esperadas da seguinte forma:

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \text{ em que}$$

- k = número de categorias (classes)
- o_i = freqüência observada na categoria i
- e_i = freqüência esperada na categoria i

Quanto maior o valor de X^2, maior será a probabilidade de as freqüências observadas estarem divergindo das freqüências esperadas.

A estatística do teste X^2 tem distribuição Qui-Quadrado com k – 1 graus de liberdade. Depois de calculada a estatística do teste, deve-se compará-la com o seu respectivo valor crítico, definido pelo nível de significância e graus de liberdade.

Exemplo: Deseja-se testar se a posição de largada de um cavalo (por dentro ou por fora) influencia o resultado de uma corrida de cavalos.

Posição	1	2	3	4	5	6	7	8
Número de Vitórias	29	19	18	25	17	10	15	11
	18*	18*	18*	18*	18*	18*	18*	18*

* Resultado esperado pela hipótese nula

$$H_0 : f_1 = f_2 = \cdots = f_8 \quad \text{versus} \quad H_a : f_1 \neq f_2 \neq \cdots \neq f_8$$

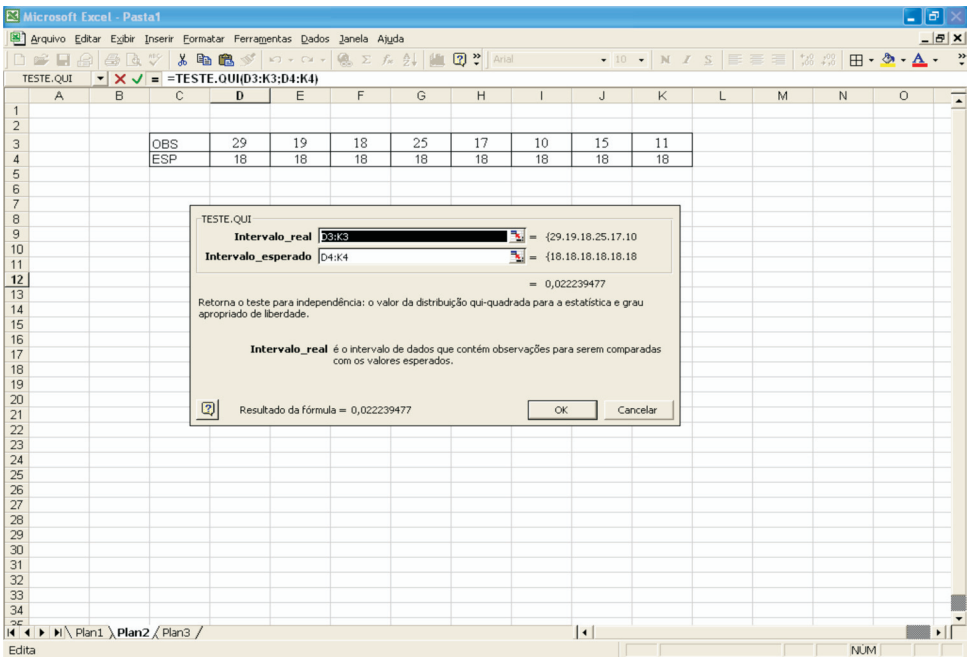
$$X^2 = \sum_{k=1}^{8} \frac{(o_i - e_i)^2}{e_i} = \frac{(29-18)^2}{18} + \frac{(19-18)^2}{18} + \cdots + \frac{(11-18)^2}{18} = 16,3$$

A tabela Qui-quadrado com 7 graus de liberdade indica que o valor 14,06 está associado a um nível de significância de 5%. Este valor é obtido na tabela, cruzando as informações da coluna 0,05 e linha 7. Nota-se que o valor calculado do qui-quadrado é superior ao valor crítico, o que nos leva a rejeitar a hipótese nula. Portanto, temos evidência de que a posição de largada dos cavalos influencia no resultado da corrida.

Com a planilha *Excel*, usaríamos a função TESTE.QUI, considerando:

Intervalo_real: posição das freqüências observadas na planilha;

Intervalo_esperado: posição das freqüências esperadas na planilha;



Observe que a planilha irá fornecer o p–valor = 0,022 que sendo menor que o nível de significância (0,05) nos leva à rejeição da hipótese nula.

Ampliando seus conhecimentos

Mineração de dados

(GONÇALVES, 2001)

Mineração de dados, ou *data mining*, é definida como uma etapa na descoberta do conhecimento em bancos de dados que consiste no processo de analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis. Para isso são utilizadas técnicas que envolvem métodos estatísticos que visam descobrir padrões e regularidades entre os dados pesquisados.

Em um mundo globalizado, sem fronteiras geográficas, onde as empresas competem mundialmente, a informação torna-se um fator crucial na busca pela competitividade. O fato de uma empresa dispor de certas informações possibilita-lhe aumentar o valor agregado de seu produto ou reduzir seus custos em relação àquelas que não possuem o mesmo tipo de informação. As informações e o conhecimento compõem um recurso estratégico essencial para o sucesso da adaptação da empresa em um ambiente de concorrência. Toda empresa tem informações que proporcionam sustentação para suas decisões, entretanto apenas algumas conseguem otimizar o seu processo decisório e aquelas que estão nesse estágio evolutivo seguramente possuem vantagem empresarial.

As ferramentas de mineração de dados, por definição, devem trabalhar com grandes bases de dados e retornar, como resultado, conhecimento novo e relevante; porém devemos ser céticos quanto a essa afirmação, pois esse tipo de ferramenta irá criar inúmeras relações e equações, o que pode tornar impossível o processamento desses dados.

A grande promessa da mineração de dados resume-se na afirmação de que ela 'vasculha' grandes bases de dados em busca de padrões escondidos, que extrai informações desconhecidas e relevantes e as utiliza para tomar decisões críticas de negócios. Outra promessa em relação a essa tecnologia de informação diz respeito à forma como elas exploram as inter-relações entre os dados. As ferramentas de análise disponíveis dispõem de um método basea-

do na verificação, isto é, o usuário constrói hipóteses sobre inter-relações específicas e então verifica ou refuta essas hipóteses por meio do sistema. Esse modelo torna-se dependente da intuição e habilidade do analista em propor hipóteses interessantes, em manipular a complexidade do espaço de atributos e em refinar a análise, baseado nos resultados de consultas potencialmente complexas ao banco de dados. Já o processo de mineração de dados, para o autor, seria responsável pela geração de hipóteses, garantindo mais rapidez, acurácia e completude dos resultados.

A cada ano, companhias acumulam mais e mais dados em seus bancos de dados. Esses dados muitas vezes são mantidos mesmo depois de esgotados seus prazos legais de existência, como no caso de notas fiscais. Com o passar do tempo, esse volume de dados passa a armazenar internamente o histórico das atividades da organização. Como conseqüência, esses bancos de dados passam a conter verdadeiros 'tesouros' de informação sobre vários procedimentos dessas companhias. Toda essa informação pode ser usada para melhorar os procedimentos da empresa, permitindo que ela detecte tendências e características disfarçadas e reaja rapidamente a um evento que ainda pode estar por vir. No entanto, apesar do enorme valor desses dados, a maioria das organizações é incapaz de aproveitar totalmente o que está armazenado em seus arquivos.

Essa informação está implícita, escondida sob uma montanha de dados, e não pode ser descoberta utilizando-se sistemas de gerenciamento de banco de dados convencionais. A quantidade de informação armazenada está explodindo e ultrapassa a habilidade técnica e a capacidade humana na sua interpretação.

Por isso, diversas ferramentas têm sido usadas para examinar os dados que as empresas possuem, no entanto, a maioria dos analistas tem reconhecido que existem padrões, relacionamentos e regras escondidos nesses dados, os quais não podem ser encontrados por meio da utilização de métodos tradicionais. A resposta é usar softwares de mineração de dados que utilizam algoritmos matemáticos avançados para examinar grandes volumes de dados detalhados.

A necessidade de transformar a 'montanha' de dados armazenados em informações significativas é óbvia, entretanto, sua análise ainda é demorada, dispendiosa, pouco automatizada e sujeita a erros, mal entendidos e falta de precisão. A automatização dos processos de análise de dados, com a utilização de softwares ligados diretamente à massa de informações, tornou-se uma necessidade. Esse motivo deve ser o responsável pelo crescimento do mercado de tecnologias de informação.

Atividades de aplicação

1. Um experimento foi realizado em 115 propriedades para verificar a eficácia de um novo adubo para plantações de milho. As produções médias das propriedades com o novo adubo encontram-se tabuladas abaixo. Compare com as produções médias garantidas pelo fabricante nas especificações técnicas do produto. Considere α = 0,05.

Classes (sacas/hectare)	f_i	e_i
2 700 |— 3 000	13	12
3 000 |— 3 300	18	20
3 300 |— 3 600	24	25
3 600 |— 3 900	32	25
3 900 |— 4 200	17	20
4 200 |— 4 500	11	13
Total	**115**	**115**

2. Em um exame a que se submeteram 117 estudantes de escolas públicas, a nota média foi 74,5 e o desvio padrão 8. Em uma escola particular, em que 200 estudantes foram submetidos a esse mesmo exame, a nota média foi de 75,9 com desvio padrão 10. A escola particular apresenta um melhor rendimento no exame? Considere α = 0,05.

3. Um médico-cientista imagina ter inventado uma droga revolucionária que baixa a febre em 1 minuto. Quinze voluntários foram selecionados (pacientes de uma clínica, com febre acima de 37°c) e os resultados foram os seguintes (em graus Celsius):

Paciente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Diferença*	1	0	3	4	3	2	1	1	4	1	0	0	2	3	3

* diferença de temperatura: o quanto a temperatura baixou em 1 minuto.

A droga inventada pelo médico é verdadeiramente eficiente?

4. Um criador verificou em uma amostra do seu rebanho (500 cabeças) 50 animais com verminose. Em seguida, avaliou outras 100 cabeças de

gado, mas antes solicitou ao veterinário uma solução para o problema. O veterinário alterou a dieta dos animais e acredita que a doença diminuiu de intensidade. Um exame nesse grupo de 100 cabeças do rebanho, escolhidas ao acaso, indicou 4 delas com verminose. Ao nível de significância de 1%, há indícios de que a proporção é menor?

5. Queremos comparar três hospitais, com relação à satisfação demonstrada por pacientes quanto ao atendimento durante o período de internação. Para tanto, foram selecionados, aleatoriamente, pacientes com grau de enfermidade semelhante. Cada paciente preencheu um questionário e as respostas geraram índices variando de 0 a 100, indicando o grau de satisfação. Os resultados foram:

	Hospital		
Pacientes	A	B	C
1	93	60	70
2	86	58	75
3	85	47	77
4	90	62	72
5	91	58	78
6	82	61	78
7	88	63	70
8	86	64	71
9	87	68	68
10	85	58	73
11		57	74
12		67	80
13		61	68
14		56	
15		58	

Baseando-se nos dados apresentados, teste se as médias populacionais são iguais. Qual sua conclusão? Use $\alpha = 0,05$.

Análise de Correlação e Medidas de Associação

Introdução

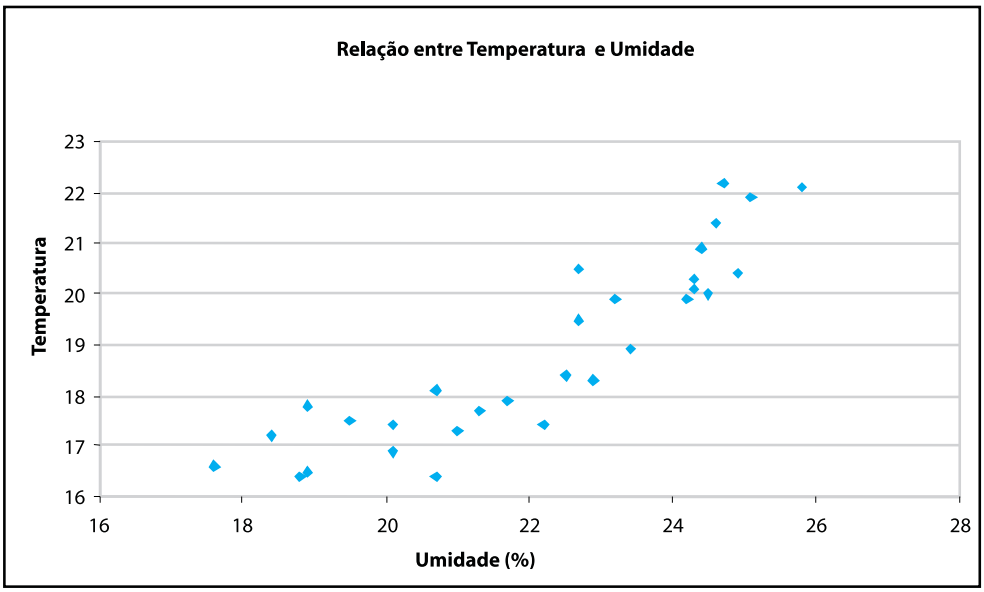
Muitas vezes, precisamos avaliar o grau de relacionamento entre duas ou mais variáveis. É possível descobrir, com precisão, o quanto uma variável interfere no resultado de outra. As técnicas associadas à Análise de Correlação representam uma ferramenta fundamental de aplicação nas Ciências Sociais e do comportamento, da Engenharia e das Ciências Naturais. A importância de se conhecer os diferentes métodos e suas suposições de aplicação é exatamente pelo cuidado que se deve ter para não se utilizar uma técnica inadequada. Existem diversos critérios de avaliação dessa relação, alguns próprios para variáveis que seguem uma distribuição normal e outros para variáveis que não seguem uma distribuição teórica conhecida. É comum a utilização do Coeficiente de Correlação de Pearson. No entanto, existem situações em que o relacionamento entre duas variáveis não é linear, ou uma delas não é contínua ou as observações não são selecionadas aleatoriamente. Nesses casos, outras alternativas de coeficientes devem ser aplicadas. Entre as diversas alternativas, veremos aqui algumas das mais importantes: Coeficiente de Spearman e Coeficiente de Contingência.

Segundo o dicionário Aurélio, *correlação* significa *relação mútua entre dois termos*, qualidade de correlativo, correspondência. Correlacionar, significa estabelecer relação ou correlação entre; ter correlação. Enquanto que a palavra *regressão* significa *ato ou efeito de regressar*, de voltar, retorno, regresso; dependência funcional entre duas ou mais variáveis aleatórias. A palavra regredir significa ir em marcha regressiva, retroceder.

Mas, onde e como surgiram os termos correlação e regressão? Foi Francis Galton (1822-1911), primo de Charles Darwin, quem usou pela primeira vez esses termos, cujo trabalho influenciou a Estatística e a Psicologia. Galton publicou o livro *Gênio Hereditário*, em 1869, no qual aplicou conceitos estatísticos a problemas da hereditariedade. O primeiro relato em que Galton usou o termo "co-relações" foi em 1888.

Diagramas de Dispersão

Um dos métodos mais usados para a investigação de pares de dados é a utilização de diagramas de dispersão cartesianos (ou seja, os conhecidos diagramas x-y). Geometricamente, um diagrama de dispersão é simplesmente uma coleção de pontos num plano cujas duas coordenadas cartesianas são os valores de cada membro do par de dados. E para quê fazemos um diagrama de dispersão? Este é o melhor método de examinar os dados no que se refere à ocorrência de tendências (lineares ou não), agrupamentos de uma ou mais variáveis, mudanças de espalhamento de uma variável em relação à outra e verificar a ocorrência dos valores discrepantes. Observe o exemplo a seguir:



Podemos notar pela análise da figura acima, a relação linear entre as duas variáveis. Os coeficientes apresentados a seguir nos auxiliam na quantificação do grau de relacionamento entre as variáveis de interesse.

A Covariância e o Coeficiente de Correlação de Pearson

Quando estudamos a relação entre duas variáveis X e Y, devemos primeiramente compreender o conceito de covariância. Se a variância é uma estatística por meio da qual chegamos ao desvio padrão que é uma medida de dispersão, da mesma maneira a covariância é uma estatística pela qual che-

gamos ao coeficiente de correlação que mede o grau de associação "linear" entre duas variáveis aleatórias X e Y.

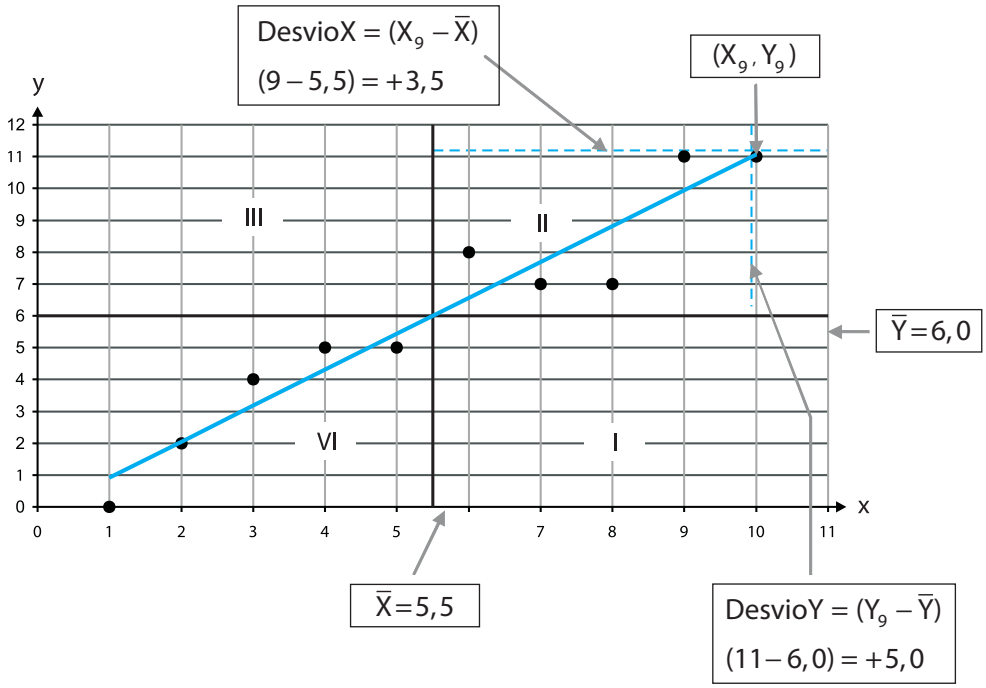
Observe o exemplo abaixo. Sejam X e Y duas variáveis aleatórias quaisquer, que tomam os seguintes valores:

Tabela 1. Cálculo do Coeficiente de Correlação de Pearson

X	Y	DesvioX $(X_i - \bar{X})^2$	DesvioY $(Y_i - \bar{Y})$	D X D Y $(X_i - \bar{X}).(Y_i - Y)$	Desvio X^2 $(X_i - \bar{X})^2$	Desvio Y^2 $(Y_i - \bar{Y})^2$	PRE_1 Y=a+bX
1	0	–4,50	–6,00	27,00	20,25	36,00	0,92727
2	2	–3,50	–4,00	14,00	12,25	16,00	2,05455
3	4	–2,50	–2,00	5,00	6,25	4,00	3,18182
4	5	–1,50	–1,00	1,50	2,25	1,00	4,30909
5	5	–0,50	–1,00	0,50	0,25	1,00	5,43636
6	8	0,50	2,00	1,00	0,25	4,00	6,56364
7	7	1,50	1,00	1,50	2,25	1,00	7,69091
8	7	2,50	1,00	2,50	6,25	1,00	8,81818
9	11	3,50	5,00	17,50	12,25	25,00	9,94545
10	11	4,50	5,00	22,50	20,25	25,00	11,07273
55	**60**	**0**	**0**	**93,00**	**82,50**	**114,00**	**60,0000**

Na tabela anterior está uma ilustração dos cálculos dos componentes da covariância e correlação.

A figura a seguir mostra a relação entre as duas variáveis X e Y, bem como a linha ajustada a esses valores pelo método de mínimos quadrados. Observe que a média de X é 5,5 e a média de Y é 6,0, e que elas estão formadas pelas linhas paralelas ao eixo Y e ao eixo X respectivamente. Vejamos agora o que significa os desvios de cada ponto em relação à média. Observe que cada ponto está formado pelo par ordenado (X_i,Y_i), onde X_i indica o valor da variável X e Y_i o valor da variável Y naquele ponto.



Tome, agora, por exemplo,

$$DesvioX = (X_9 - \bar{X}) = (9 - 5{,}5) = +3{,}5 \text{ e } DesvioY = (Y_9 - Y) = (11 - 6{,}0) = +5{,}0$$

O produto dos desvios:

$$DesviosX \,.\, DesvioY = (X_9 - \bar{X}).(Y_9 - \bar{Y}) = (9 - 5{,}5).(11 - 6{,}0) = (+3{,}5).(+5{,}0) = 17{,}5$$

Se calcularmos esses produtos para todos os valores de X e Y e somarmos temos o numerador da covariância de X e Y:

$$C(X,Y) = \frac{\sum (X_i - \bar{X}).(Y_i - \bar{Y})}{n} = \frac{93}{10} = 9{,}3 \qquad \textbf{(1)}$$

Logo, covariância significa co-variação, como as duas variáveis variam de forma conjunta. Agora, vejamos o que acontece se os pontos estivessem no quadrante I. Neste caso, os desvios de X seriam todos positivos, enquanto que os desvios de Y seriam todos negativos, logo, os produtos tomam valores negativos. O mesmo vai acontecer com os pontos do quadrante III, nele os desvios de X tomam valores negativos e os desvios de Y, valores positivos, logo, os produtos tomam valores negativos. Assim, se a maioria dos pontos caem nos quadrantes I e III, a covariância toma valores negativos, indicando que essas duas

variáveis se relacionam de forma negativa ou inversa, ou seja, quando uma cresce a outra diminui e vice-versa.

Quando os pontos se distribuem nos quatro quadrantes, haverá valores positivos e negativos, logo a soma tende para zero, e nesse caso, afirmamos que não existe relação linear entre essas variáveis. Observamos que esta estatística tende para zero, mesmo havendo uma relação que não seja linear, por exemplo se os dados tivessem o formato de uma parábola, ou relação quadrática.

Apesar de a covariância ser uma estatística adequada para medir relação linear entre duas variáveis, ela não é adequada para comparar graus de relação entre variáveis, dado que ela está influenciada pelas unidades de medida de cada variável, que pode ser metros, quilômetro, quilogramas, centímetros etc. Para evitar a influência da ordem de grandeza e unidades de cada variável, dividimos a covariância pelo desvio padrão de X e de Y, dando origem ao coeficiente de correlação de Pearson:

Notação:

Coeficiente de correlação amostral: **r**

Coeficiente de correlação populacional: **ρ**

$$r = \frac{C(X,Y)}{S_Y . S_X} \quad \textbf{(2)}$$

$$r = \frac{9,3}{2,8723 \,.\, 3,3764} = 0,95896$$

Onde: $S^2_x = 82,5 / 10 = 8,25 \rightarrow S_x = 2,8723$

$S_y^2 = 114,0 / 10 = 11,4 \rightarrow S_y = 3,3764$

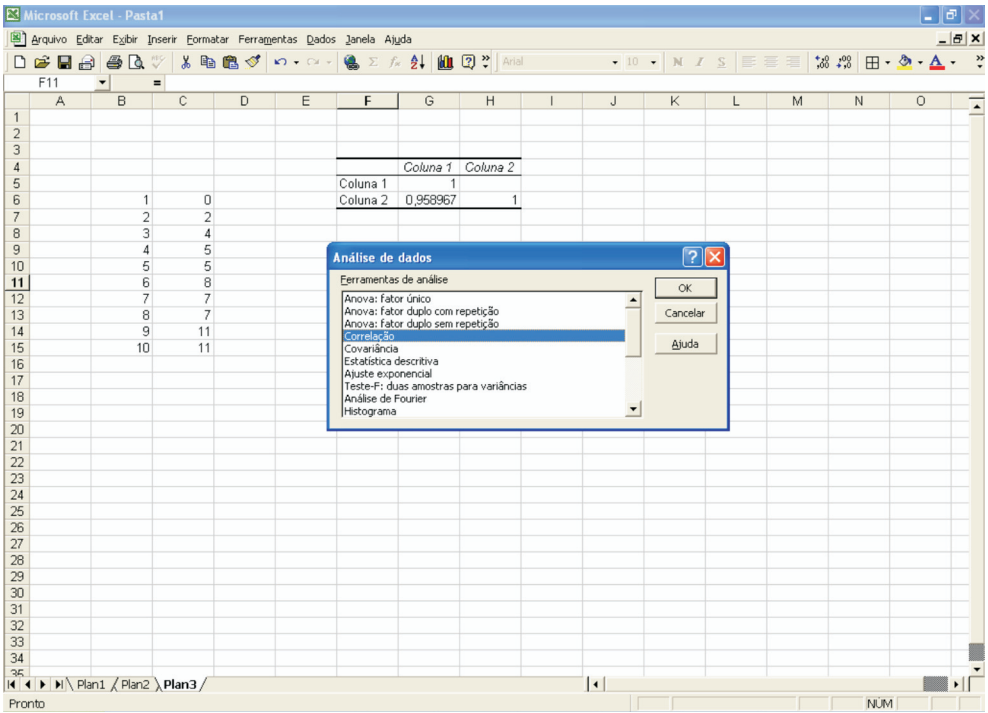
Como o coeficiente de correlação está isento de unidades e da ordem de grandeza das variáveis, este toma valores entre –1 e 1.

Relação positiva ➔ **r** tomará o valor 1 quando a relação é perfeita.

Relação negativa ➔ **r** tomará o valor –1 quando a relação é perfeita.

Relação difusa ou não linear ➔ **r** será igual a 0.

No *Excel*, usando a opção Correlação em "Análise de dados", obtemos:



O coeficiente de Determinação

Outro coeficiente amplamente utilizado para mensurar o grau de correlação entre duas variáveis é o *coeficiente de determinação*. É definido elevando o valor do coeficiente de Pearson ao quadrado e denotado por $\mathbf{r^2}$. Pode ser interpretado como a proporção da variação de Y que é explicada pela variável X (e vice versa).

Muito embora o coeficiente de determinação seja relativamente fácil de interpretar, ele não pode ser testado estatisticamente. Contudo, a raiz quadrada do coeficiente de determinação, que é o coeficiente de correlação (r), pode ser testada estatisticamente, pois está associada a uma estatística de teste que é distribuída segundo uma distribuição t de Student, quando a correlação populacional $\rho = 0$.

O coeficiente de correlação para dados populacionais é:

População: $\rho = \sqrt{\rho^2}$

O coeficiente de correlação para dados amostrais é:

Amostra: $r = \sqrt{r^2}$

Significância do coeficiente de correlação

Para comprovarmos se o coeficiente de correlação é significativo, devemos realizar o seguinte teste de hipóteses:

Hipóteses:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

A estatística de teste é $t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

com n-2 graus de liberdade na tabela t de Student. Caso o valor de t_c seja superior ao valor crítico de t, devemos rejeitar a hipótese nula. Se a hipótese nula, ao nível de significância α, for rejeitada podemos concluir que efetivamente existe uma relação significativa entre as variáveis.

Exemplo 1: Para estudar a poluição de um rio, um cientista mediu a concentração de um determinado composto orgânico (Y) e a precipitação pluviométrica na semana anterior (X):

X	Y
0,91	0,10
1,33	1,10
4,19	3,40
2,68	2,10
1,86	2,60
1,17	1,00

Existe alguma relação entre o nível de poluição e a precipitação pluviométrica? Teste sua significância, ao nível de 5%.

Calculando a média de X e de Y temos $\overline{X} = 2{,}023$ e $\overline{Y} = 1{,}717$.

Calculando a covariância entre X e Y pela expressão (1),

$$C(X,Y) = \frac{(0{,}91-2{,}023).(0{,}10-1{,}717)+(1{,}33-2{,}023).(1{,}10-1{,}717)+...+(1{,}17-2{,}023).(1{,}00-1{,}717)}{6}$$

$$C(X,Y) = 1{,}0989$$

Calculando os desvios padrões de X e Y temos: $S_x = 1{,}125$ e $S_y = 1{,}10$

E assim, pela expressão (2),

$$r = \frac{C(X,Y)}{S_y.S_x} = \frac{1{,}0989}{1{,}125.1{,}1} = 0{,}888$$

Testando a significância do coeficiente,

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}888\sqrt{6-2}}{\sqrt{1-(0{,}888)^2}} = 3{,}86$$

O valor crítico de t para n– 2 = 4 graus de liberdade e 5% de nível de significância é 2,78. Note que o teste de significância do coeficiente será sempre bilateral.

Como o valor calculado de t é superior ao valor crítico, podemos concluir que existem evidências suficientes para afirmar que o composto orgânico (Y) e a precipitação pluviométrica (X) estejam correlacionados.

Exemplo 2: Procurando quantificar os efeitos da escassez de sono sobre a capacidade de resolução de problemas simples, um agente tomou ao acaso 10 sujeitos e os submeteu a experimentação. Deixou-os sem dormir por diferentes números de horas, após o que solicitou que os mesmos resolvessem os itens "contas de adicionar" de um teste. Obteve, assim, os seguintes dados:

Nº de erros - Y	Horas sem dormir - X
8	8
6	8
6	12
10	12
8	16
14	16
14	20
12	20
16	24
12	24

Calcule o coeficiente de correlação linear de Pearson e teste a sua significância ao nível de 1%.

Calculando a média de X e de Y temos $\overline{X} = 16$ e $\overline{Y} = 10,6$.

Calculando a covariância entre X e Y pela expressão (1),

$$C(X,Y) = \frac{(8-16).(8-10,6)+(8-16).(6-10,6)+...+(24-16).(12-10,6)}{10} = 15,2$$

Calculando os desvios padrões de X e Y temos:

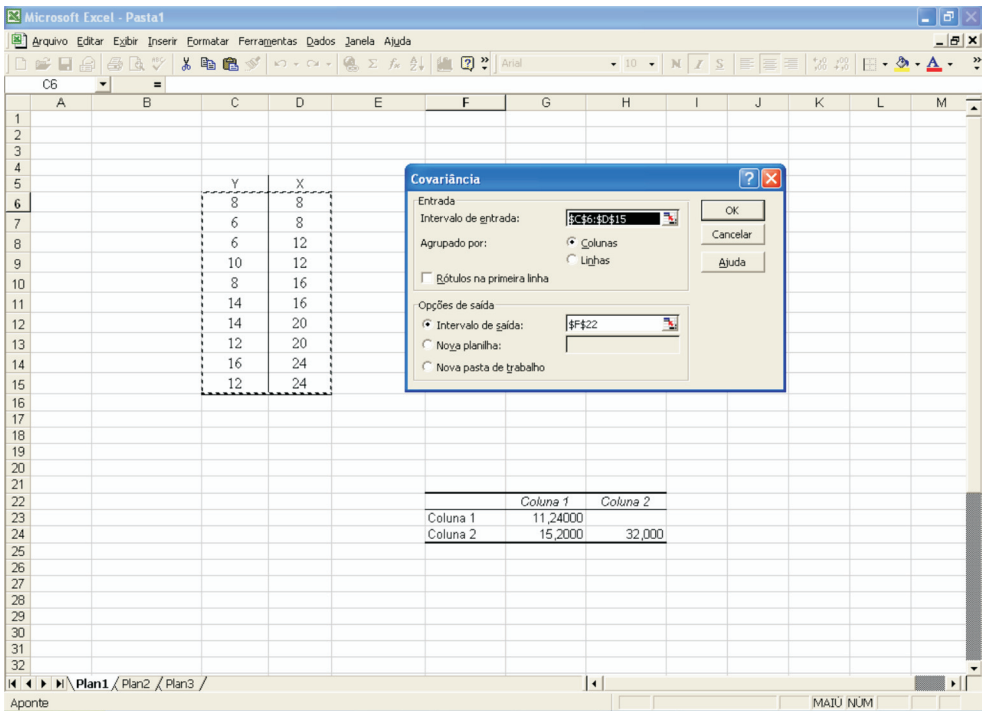
$S_x = 5,656854$ e $S_y = 3,352611$

E assim, pela expressão (2),

$$r = \frac{C(X,Y)}{S_y.S_x} = \frac{15,2}{5,656854.3,352611} = 0,801467$$

Observação: procure sempre usar o maior número de casas decimais possível.

Usando a planilha *Excel* poderemos também obter uma matriz de covariância, que nos fornece a covariância entre X e Y além da variância de X e de Y.



Y	X
8	8
6	8
6	12
10	12
8	16
14	16
14	20
12	20
16	24
12	24

Agora testando a significância do coeficiente,

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}801467\sqrt{10-2}}{\sqrt{1-(0{,}801467)^2}} = 3{,}79$$

O valor crítico de t para n–2 = 8 graus de liberdade e 1% de nível de significância é 3,355 (bilateral).

Como o valor calculado de t é superior ao valor crítico, podemos concluir que existem evidências suficientes para afirmar que o número de horas sem dormir (X) influencia significativamente o número de erros (Y).

Medidas de Associação

Freqüentemente, estamos interessados em verificar a existência de associação entre dois conjuntos de escores e também o grau desta associação. No caso paramétrico, a medida usual é o coeficiente de correlação **r** de Pearson que exige mensuração dos escores no mínimo ao nível intervalar. Ainda, se estivermos interessados em comprovar a significância de um valor observado de **r** de Pearson deveremos supor que os escores provenham de uma distribuição normal. Quando estas suposições não são atendidas, podemos utilizar um dos coeficientes de correlação não-paramétricos e suas respectivas provas de significância.

Coeficiente de Contingência C

Este coeficiente mede a associação entre dois conjuntos de atributos quando um ou ambos os conjuntos são medidos em escala nominal.

Considere uma tabela de contingência k x r, que representa as freqüências cruzadas dos escores A (divididos em k categorias) e escores B (divididos em r categorias). O grau de associação entre dois conjuntos de atributos é calculado por:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$$ onde χ^2 é a estatística Qui-quadrado.

O p-valor associado ao valor da estatística Qui-quadrado com (r-1) x (k-1) graus de liberdade é a prova de significância do coeficiente de contingência **C**.

O coeficiente **C** se caracteriza por assumir valor zero quando há inexistência de associação porém nunca será igual à 1. O limite superior do coeficiente é dado por $\sqrt{\frac{k-1}{k}}$ (quando k = r). Note que para calcular o coeficiente **C**, a tabela de contingência deve satisfazer as restrições do teste Qui-quadrado.

Exemplo: Estudantes de escolas particulares e de escolas públicas selecionados aleatoriamente foram submetidos a testes padronizados de conhecimento, e produziram os resultados abaixo. Verifique o grau de associação entre as variáveis mensuradas e teste a significância ao nível de 5%.

	Escores			
Escola	**0 – 275**	**276 – 350**	**351 – 425**	**426 – 500**
Particular	6	14	17	9
Pública	30	32	17	3

Queremos aqui verificar o grau de associação entre as variáveis "Escola" e "Escore de conhecimento". A variável Escola é mensurada em nível nominal, o que inviabiliza a utilização do coeficiente **r** de Pearson.

Obtendo então o coeficiente de Contingência, necessitamos inicialmente calcular o valor da estatística χ^2:

Freq. Obs.	6	14	17	9
	30	32	17	3
Freq. Esp.	12,94	16,53	12,22	4,31
	23,06	29,47	21,78	7,69

$$\chi^2 = \frac{(6-12,94)^2}{12,94} + \frac{(14-16,53)^2}{16,53} + \ldots + \frac{(3-7,69)^2}{7,69} = 17,28$$

O coeficiente de contingência é:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{17,28}{128+17,28}} = 0,345$$

Para testar a significância do coeficiente, precisamos verificar o valor crítico de χ^2 considerando α=0,05 e (r–1) x (k–1) = 3 graus de liberdade. Esse valor é igual a 7,81. Comparando com o valor calculado de 17,28, podemos admitir a existência de associação significativa entre a escola e o escore de

conhecimento. Analisando atentamente, poderíamos acrescentar que o fato de um estudante pertencer a uma escola particular faz com que ele obtenha um escore de conhecimento mais alto.

Coeficiente de correlação de Spearman

É uma medida de associação que exige que ambas as variáveis se apresentem em escala de mensuração pelo menos ordinal. Basicamente, equivale ao coeficiente de correlação de Pearson aplicado a dados ordenados. Assim,

$$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}} = r_s$$

ou seja, o coeficiente de correlação de Spearman se utiliza da expressão do coeficiente de Pearson, porém calculado com postos. Esta expressão equivale à

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$ onde $d_i = x_i - y_i$ a diferença de postos dos escores X e Y.

Para verificar a significância do valor observado de r_s, podemos usar a expressão de t de Student

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$ onde **t** tem n–2 graus de liberdade.

Exemplo: As notas obtidas por 10 estudantes de Administração e o seu QI (quociente de inteligência) são apresentadas no quadro abaixo:

Notas	8	9,5	10	9,1	6,5	9	9,5	5,2	9,1	9,3
QI	127	149	150	135	122	129	142	100	136	139

Utilize o coeficiente de Spearman para verificar se as variáveis estão associadas e qual o seu grau de associação.

Inicialmente, ordenamos os valores originais, transformando-os em postos. Aqui então substituímos os valores originais pelos seus respectivos postos, ou seja, o menor valor da variável em questão será substituído pelo valor 1 e assim por diante. Em seguida, calculamos as diferenças de postos:

Notas	3	8,5	10	5,5	2	4	8,5	1	5,5	7
QI	3	9	10	5	2	4	8	1	6	7
di	0	–0,5	0	0,5	0	0	0,5	0	–0,5	0
(di)2	0	0,25	0	0,25	0	0	0,25	0	0,25	0

Calculando o coeficiente:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6.\left(0^2 + 0{,}25^2 + \ldots + 0^2\right)}{10^3 - 10} = 1 - \frac{6.0{,}25}{990} = 0{,}998$$

Verificando a significância estatística do coeficiente:

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} = 0{,}998\sqrt{\frac{8}{1-(0{,}998)^2}} = 0{,}998\sqrt{\frac{8}{0{,}004}} = 44{,}63$$

O valor crítico da estatística t de Student é obtido definindo-se n–2 = 8 graus de liberdade e o nível de significância, que admitiremos igual a 1%. Este valor é igual a 3,36. Mais uma vez temos aqui um teste bilateral pois estamos verificando se o coeficiente é diferente de zero.

Assim, podemos comprovar que o coeficiente de associação é altamente significativo, ou seja, existem fortes indícios que apontam para notas altas obtidas por aqueles que possuem maiores quocientes de inteligência.

Ampliando seus conhecimentos

Teste de Kappa

(LANDIS; KOCH, 1977)

O Teste de Kappa é uma medida de concordância interobservador e mede o grau de concordância, além do que seria esperado tão-somente pelo acaso.

Para descrevermos se há ou não concordância entre dois ou mais avaliadores, ou entre dois métodos de classificação, utilizamos a medida Kappa que é baseada no número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre os avaliadores. Esta medida de concordância assume valor máximo igual a 1, que representa total concordância ou, ainda,

pode assumir valores próximos e até abaixo de 0, os quais indicam nenhuma concordância.

O coeficiente Kappa é calculado a partir da seguinte fórmula:

$$\text{Kappa} = \frac{P_0 - P_E}{1 - P_E}$$

$$\text{onde} P_0 = \frac{\textit{número de concordâncias}}{\textit{número de concordâncias} + \textit{número de discordâncias}}$$

$$\text{e } P_E = \sum_{i=1}^{n} (p_{i1}.p_{i2}) \text{ sendo que:}$$

- n é o número de categorias;
- i é o índice da categoria (que vale de 1 a n);
- p_{i1} é a proporção de ocorrência da categoria i para o avaliador 1;
- p_{i2} é a proporção de ocorrência da categoria i para o avaliador 2.

Para avaliar se a concordância é razoável, Landis, JR e Koch, GG (1977) sugerem a seguinte interpretação:

Valores obtidos de Kappa	Interpretação
<0	Nenhuma concordância
0 – 0,19	Concordância pobre
0,20 – 0,39	Concordância leve
0,40 – 0,59	Concordância moderada
0,60 – 0,79	Concordância substancial
0,80 – 1,00	Concordância quase perfeita

Fonte: Landis JR, Koch GG. The measurement of observer agreement for categorical data. **Biometrics** 1977; **33**: 159-174

Exemplo: Em certo órgão de financiamento, em cada edital aberto, se apresentam diversos pesquisadores que enviam projetos, solicitando recursos para desenvolvê-los. Estes projetos recebem uma avaliação, muitas vezes subjetiva, baseada na opinião de um consultor.

Considere a tabela a seguir, que resume as avaliações feitas por dois avaliadores a 30 projetos que concorrem ao financiamento. O interesse deste estudo é saber qual é a concordância entre estes dois profissionais e se há alguma classificação com concordância maior do que as demais.

		AVALIADOR 2			
		A	B	C	**Total**
AVALIADOR 1	A	14 (0,47)	1 (0,03)	1 (0,03)	**16 (0,53)**
	B	3 (0,10)	3 (0,10)	2 (0,07)	**8 (0,27)**
	C	0 (0,00)	1 (0,03)	5 (0,17)	**6 (0,20)**
	Total	**17 (0,57)**	**5 (0,16)**	**8 (0,27)**	**30 (1,00)**

* entre parênteses as proporções

Calculando o coeficiente Kappa:

$$P_0 = \frac{14+3+5}{30} = \frac{22}{30} = 0,7333$$

$$P_E = \sum_{i=1}^{n} (p_{i1}.p_{i2}) = (0,57 \,.\, 0,53) + (0,16 \,.\, 0,27) + (0,27 \,.\, 0,20) = 0,3021 + 0,0432$$

$$+ 0,054 = 0,3993$$

$$\text{Kappa} = \frac{0,733 - 0,3993}{1 - 0,3993} = 0,556$$

Note que a concordância geral pode ser considerada apenas moderada. Avaliando cada uma das três classificações, notamos que a concordância é alta quando os avaliadores atribuem o conceito A e o conceito C. No entanto, para atribuir o conceito B, um conceito intermediário, a concordância já não é tão satisfatória.

Atividades de aplicação

1. Foi tomada uma amostra aleatória de 10 carregamentos recentes feitos por caminhão de uma companhia, anotada a distância em quilômetros e o tempo de entrega. Os dados seguem abaixo:

Carregamento	1	2	3	4	5	6	7	8	9	10
Distância em Km (X)	825	215	1 070	550	480	920	1 350	325	670	1 215
Tempo de entrega em dias (Y)	3,5	1,0	4,0	2,0	1,0	3,0	4,5	1,5	3,0	5,0

a) Construa o diagrama de dispersão.

b) Calcule o coeficiente de correlação de Pearson para os dados desta amostra.

c) Calcule o coeficiente de determinação.

d) Verifique se o coeficiente de correlação é significativo (α=0,05).

2. Para uma amostra de n = 10 tomadores de empréstimos em uma companhia financeira, o coeficiente de correlação entre a renda familiar média e débitos a descoberto de curto prazo foi calculado r = 0,50.

Teste a hipótese de que não existe correlação entre as duas variáveis, usando um nível de significância de 5%.

3. Para avaliar a relação entre habilidade verbal e habilidade matemática, escores de 8 estudantes foram obtidos, gerando a tabela abaixo:

	Estudantes							
Escore	1	2	3	4	5	6	7	8
Matemática	80	50	36	58	72	60	56	68
Verbal	65	60	35	39	48	44	48	61

Calcule o coeficiente de correlação e teste sua significância.

4. Em um estudo conduzido com 10 pacientes, estes foram colocados sob uma dieta de baixas gorduras e altos carboidratos. Antes de iniciar a dieta, as medidas de colesterol e de triglicerídeos foram registradas para cada indivíduo .

a) Construa um gráfico de dispersão para esses dados.

b) Há alguma evidência de relação linear entre os níveis de colesterol e de triglicerídeos?

c) Calcule o coeficiente de correlação de Spearman e teste sua significância.

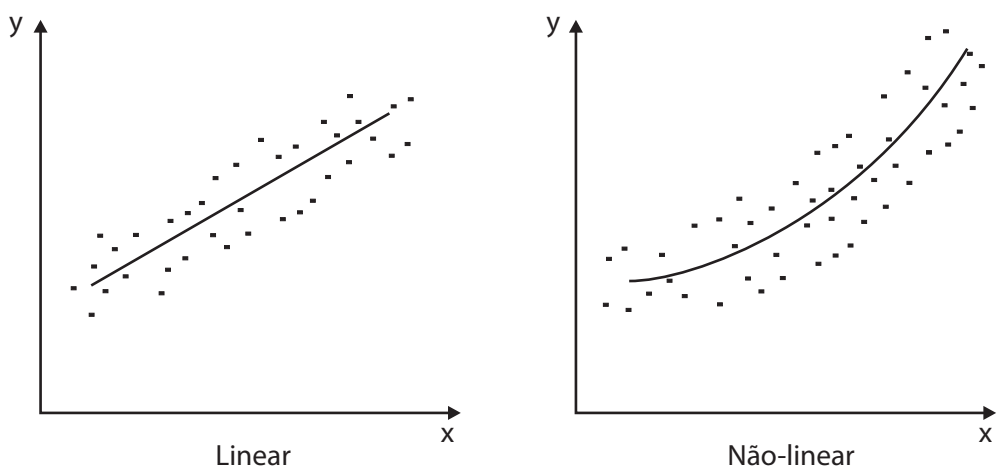
Paciente	Colesterol (mmol/l)	Triglicerídeos (mmol/l)
1	5,12	2,30
2	6,18	2,54
3	6,77	2,95
4	6,65	3,77
5	6,36	4,18
6	5,90	5,31
7	5,48	5,53
8	6,02	8,83
9	10,34	9,48
10	8,51	14,20

Análise de Regressão

Introdução

Os modelos de regressão são largamente utilizados em diversas áreas do conhecimento, tais como: computação, administração, engenharias, biologia, agronomia, saúde, sociologia etc. O principal objetivo desta técnica é obter uma equação que explique satisfatoriamente a relação entre uma variável resposta e uma ou mais variáveis explicativas, possibilitando fazer predição de valores da variável de interesse. Este relacionamento pode ser por uma equação linear ou uma função não-linear, conforme figura abaixo:

Figura 1: Formas lineares e não lineares de relação entre pares de variáveis



Regressão linear simples

Se uma relação linear é válida para sumarizar a dependência observada entre duas variáveis quantitativas, então a equação que descreve esta relação é dada por:

$$Y = a + b.X$$

Esta relação linear entre X e Y é determinística, ou seja, ela "afirma" que todos os pontos caem exatamente em cima da reta de regressão. No entanto este fato raramente ocorre, ou seja, os valores observados não caem todos

exatamente sobre esta linha reta. Existe uma diferença entre o valor observado e o valor fornecido pela equação. Esta diferença, denominada erro e representada por ε, é uma variável aleatória que quantifica a falha do modelo em ajustar-se aos dados exatamente. Tal erro pode ocorrer devido ao efeito, dentre outros, de variáveis não consideradas e de erros de medição. Incorporando esse erro à equação acima temos:

$$Y = a + b.X + \varepsilon$$

que é denominado modelo de regressão linear simples. *a* e *b* são os parâmetros do modelo.

A variável X, denominada variável regressora, explicativa ou independente, é considerada uma variável controlada pelo pesquisador e medida com erro desprezível. Já Y, denominada variável resposta ou dependente, é considerada uma variável aleatória, isto é, existe uma distribuição de probabilidade para Y em cada valor possível de X. É muito freqüente, na prática, encontrarmos situações em que Y tenha distribuição normal. Este é um dos principais pressupostos para aplicação desta técnica.

Exemplo 1: O preço de aluguel de automóveis de uma agência é definido pela seguinte equação: $Y = 8 + 0{,}15.X$, onde Y = Taxa de aluguel (R$); X = distância percorrida (km).

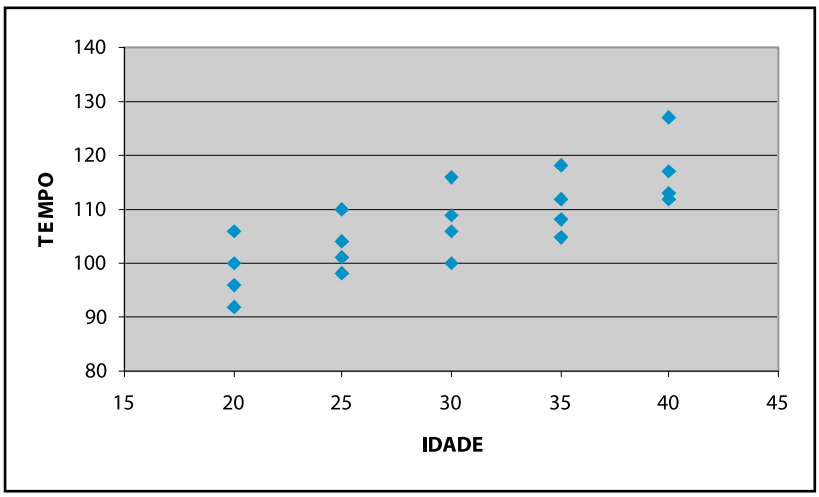
Assim, a taxa de aluguel inicia com o preço de R$ 8,00 e vai aumentando à medida que a distância percorrida aumenta. Assim, se fosse percorrida uma distância de 100 km, a taxa de aluguel seria de $8 + 0{,}15 \times 100 =$ R$ 23,00. No entanto, como essa equação foi obtida baseada em dados de automóveis de diversas marcas, certamente haverá uma variação no preço, por causa de diversos outros fatores. Assim, essa equação terá uma margem de erro, que é devida a esses inúmeros fatores que não foram controlados.

Exemplo 2: Um psicólogo investigando a relação entre o tempo que um indivíduo leva para reagir a um certo estímulo e sua idade obteve os seguintes resultados:

Tabela 1: Idade (em anos) e tempo de reação à um certo estímulo (em segundos)

Y - Tempo de reação (segundos)	X - Idade (em anos)
96	20
92	20
106	20
100	20
98	25
104	25
110	25
101	25
116	30
106	30
109	30
100	30
112	35
105	35
118	35
108	35
113	40
112	40
127	40
117	40

Figura 2: Diagrama de dispersão entre a idade (X) e o tempo de reação (Y)



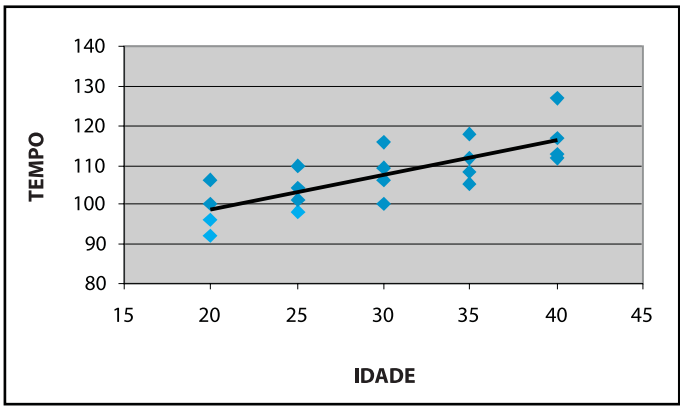
A partir da representação gráfica desses dados, mostrada na figura 2, é possível visualizar uma relação linear positiva entre a idade e o tempo de reação. O coeficiente de correlação de Pearson para esses dados resultou em r = 0,768, bem como seu respectivo teste de significância em t_{cal} = 5,09, que comparado ao valor tabelado $t_{tab,5\%}$ = 2,1 , fornece evidências de relação linear entre essas duas variáveis, ou seja, há evidências de considerável relação linear positiva entre idade e tempo de reação.

Podemos, então, usar um modelo de regressão linear simples para descrever essa relação. Para isso, é necessário estimar, com base na amostra observada, os parâmetros desconhecidos *a e b* deste modelo. O método de estimação denominado Mínimos Quadrados Ordinários (MQO) é freqüentemente utilizado em regressão linear, para esta finalidade, e será apresentado mais adiante.

Continuando a análise dos dados do exemplo, é possível obter o seguinte modelo de regressão linear simples ajustado:

Y = 80,5 + 0,9.X

Figura 3: Reta de regressão ajustada aos dados



Como a variação dos dados em X não inclui x = 0, não há interpretação prática do coeficiente *a* = 80,5. Por outro lado, *b* = 0,9 significa que a cada aumento de 1 ano na idade das pessoas, o tempo de reação médio (esperado) aumenta em 0,9 segundos.

Assim, se: X = 20 anos, teremos Y = 98,5 seg.

Para X = 21 anos, Y = 99,4 seg.

X = 22 anos, Y = 100,3 seg.

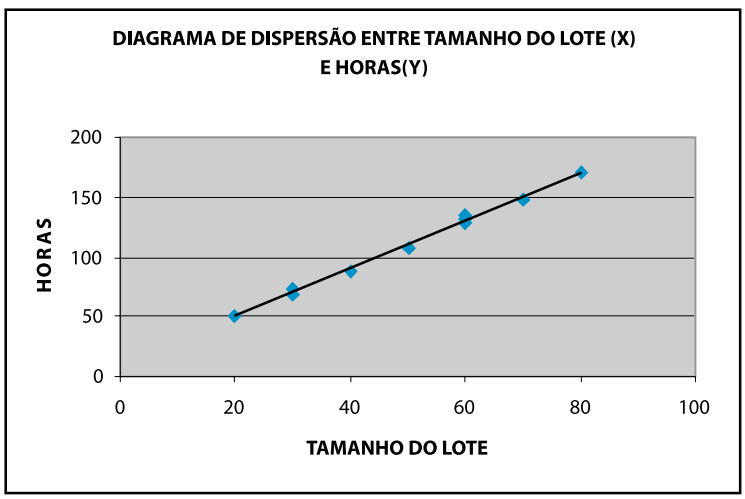
Dessa maneira, de ano para ano, o aumento no tempo de reação esperado é de 0,9 segundos.

Exemplo 3: Uma certa peça é manufaturada por uma companhia, uma vez por mês, em lotes, que variam de tamanho de acordo com as flutuações na demanda. A tabela abaixo contém dados sobre tamanho do lote e número de horas gastas na produção de 10 recentes lotes produzidos sob condições similares. Estes dados são apresentados graficamente na Figura 4, tomando-se horas-homem como variável *dependente* ou variável *resposta* (Y) e o tamanho do lote como variável *independente* ou *preditora* (X).

Tabela 2: Tamanho de lote e número de horas gastas na produção de cada lote.

Lote (i)	Horas (Y_i)	Tamanho do lote (X_i)
1	73	30
2	50	20
3	128	60
4	170	80
5	87	40
6	108	50
7	135	60
8	69	30
9	148	70
10	132	60

Figura 4: Relação estatística entre Y e X, referente aos dados da Tabela 2.



A Figura 4 sugere claramente que há uma relação linear positiva entre o tamanho do lote e o número de horas, de modo que, maiores lotes tendem a corresponder a maiores números de horas-homem consumidas. Porém, a relação não é perfeita, ou seja, há uma dispersão de pontos sugerindo que alguma variação no número de horas não é dependente do tamanho do lote. Por exemplo, dois lotes de 30 unidades (1 e 8) demandaram quantidades um pouco diferentes de horas. Na Figura 4, foi traçada uma linha (reta) de relacionamento descrevendo a relação estatística entre horas e tamanho do lote. Ela indica a tendência geral da variação em horas-homem quando há trocas no tamanho do lote.

Observa-se que grande parte dos pontos da figura não cai diretamente sobre a linha de relacionamento estatístico. A dispersão dos pontos em torno da linha de relacionamento representa a variação em horas que não é associada ao tamanho do lote, e que é usualmente considerada aleatória. Relações estatísticas são geralmente úteis, mesmo não tendo uma relação funcional exata.

Método dos mínimos quadrados ordinários (MQO)

Para estimar os parâmetros do modelo, é necessário um método de estimação. O método estatístico utilizado e recomendado pela sua precisão é o método dos mínimos quadrados que ajusta a melhor "equação" possível aos dados observados.

Com base nos *n* pares de observações (y_1, x_1), (y_2, x_2), ..., (y_n, x_n), o método de estimação por MQO consiste em escolher *a* e *b* de modo que a soma dos quadrados dos erros, ε_i (i=10 ,..., n), seja mínima.

Para minimizar esta soma, que é expressa por:

$$SQ = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{I-1}^{n} .(y_i - a - b.x_i)$$

devemos, inicialmente, diferenciar a expressão com respeito a "a" e "b" e, em seguida, igualar a zero as expressões resultantes. Feito isso, e após algumas operações algébricas, os estimadores resultantes são:

$$b = \frac{\sum x_i.y_i - n.\overline{y}.\overline{x}}{\sum x_i^2 - n.\overline{x}^2}$$

$$a = \overline{y} - b.\overline{x}$$

onde $\overline{Y}$ é a média amostral dos y_i's e $\overline{x}$ a média amostral dos x_i's.

Logo, $E(Y|x) = a + b.x$ é o modelo de regressão linear simples ajustado, em que $E(Y|x)$, denotado também $\hat{Y}$ por simplicidade, é o valor médio predito de Y para qualquer valor $X = x$ que esteja na variação observada de X.

No exemplo 2, as estimativas dos parâmetros resultaram em $a = 80,5$ e $b = 0,9$. Veja como esses valores foram obtidos:

$$\sum X_i = 2\,150 \qquad \sum Y_i = 600 \qquad n = 20 \qquad \sum X_i Y_i = 65\,400$$

$$\overline{X} = 30 \qquad \overline{Y} = 107,5 \qquad \sum X_i^2 = 19\,000$$

$$b = \frac{\sum x_i y_i - n.\overline{y}.\overline{x}}{\sum x_i^2 - n.\overline{x}^2} = \frac{65\,400 - 20.107,5.30}{19\,000 - 20.(30)^2} = \frac{900}{1\,000} = 0,9$$

$$a = \overline{y} - b.\overline{x} = 107,5 - 0,9.30 = 80,5$$

No exemplo 3, as estimativas dos parâmetros *a* e *b* são:

$$\sum X_i = 500 \qquad \sum Y_i = 1\,100 \qquad n = 10 \qquad \sum X_i Y_i = 61\,800$$

$$\overline{X} = 50 \qquad \overline{Y} = 110 \qquad \sum X_i^2 = 28\,400$$

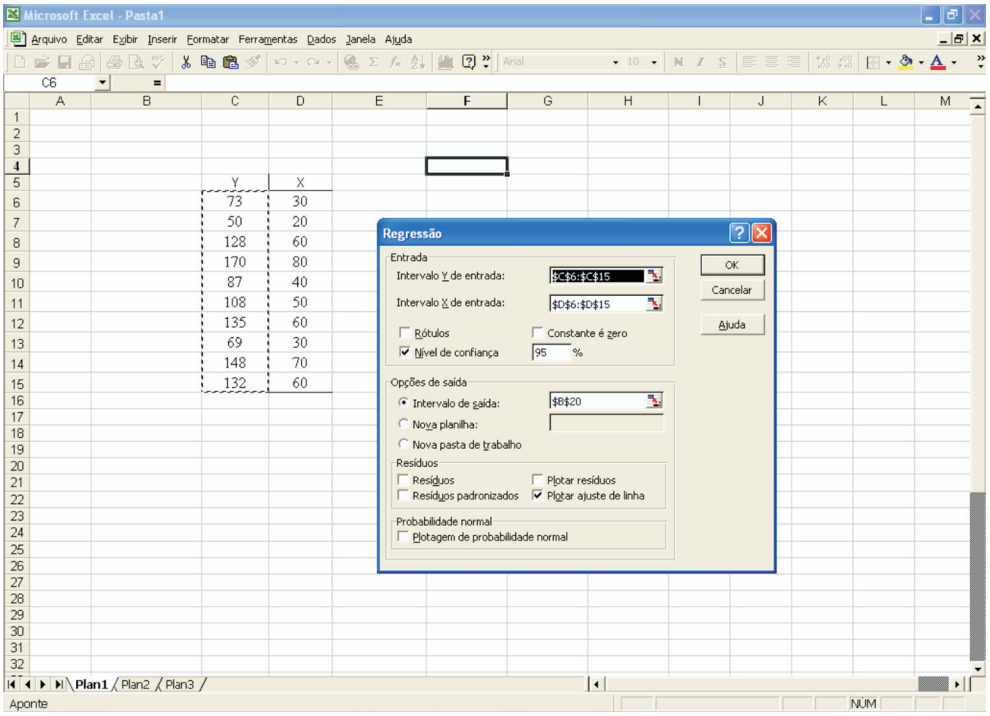
$$b = \frac{\sum x_i.y_i - n.\overline{y}.\overline{x}}{\sum x_i^2 - n.\overline{x}^2} = \frac{61\,800 - 10.110.50}{28\,400 - 10.(50)^2} = \frac{6\,800}{3\,400} = 2$$

Assim, a equação de regressão linear entre X e Y será dada por:

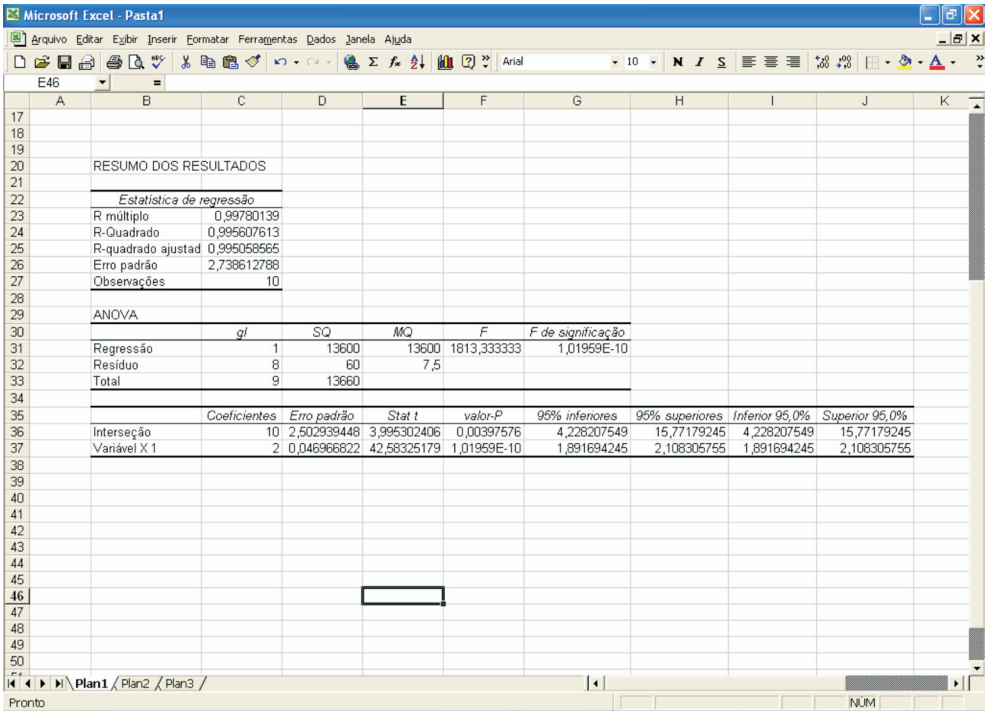
$$Y = 10 + 2.X + \varepsilon$$

Interpretando o modelo acima, poderemos observar que, aumentando o tamanho do lote em uma unidade, o número de horas gastas na produção será aumentado em 2 horas.

Obtendo a reta de regressão com ajuda da planilha *Excel*, teremos que selecionar a opção REGRESSÃO no módulo de Análise de dados (em ferramentas):



A saída fornecida pela planilha é a seguinte:



Observe que o *Excel* fornece, além dos coeficientes de correlação, a Anova da regressão para testar a sua significância e os coeficientes estimados com seus respectivos testes de significância.

Análise de Variância da Regressão

Para verificar a adequação do modelo aos dados, algumas técnicas podem ser utilizadas. A "análise de variância da Regressão" é uma das técnicas mais usadas. Assim, podemos analisar a adequação do modelo pela ANOVA da regressão a qual é geralmente apresentada como na tabela abaixo:

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	p-1	SQreg	SQreg/p-1	QMreg/QMres	
Resíduos	n-p	SQres	SQres/n-p		
Total	**n-1**	**SQtotal**	**Sqtotal/n-1**		

Onde:

- SQreg = soma dos quadrados devido à regressão:

$$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2$$

- SQres = soma dos quadrados devido aos erros:

$$SQres = SQtotal - Sqreg = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2$$

- SQtotal = soma dos quadrados totais:

$$SQtotal = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- p = número de variáveis do modelo
- n = numero de observações.

Caso o *p-valor* seja inferior ao nível de significância estabelecido, então consideramos a regressão como significativa.

Uma maneira auxiliar de medir o "ganho" relativo introduzido pelo modelo é usar o coeficiente de determinação o qual é definido por R^2 que é calculado por SQreg/SQtotal.

Para os exemplos 2 e 3, a tabela da Anova seria construída de seguinte forma:

Exemplo 2:

$$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2 = \sum_{i=1}^{n} (80,5 + 0,9x_i - 107,5)^2 = 810$$

Para obter a soma de quadrados acima, deveremos substituir em X_i todos os valores de idade da Tabela 1.

$$SQtotal = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - 107,5)^2 = 1\ 373$$

Para obter a soma de quadrados acima, deveremos substituir em Y_i todos os valores de tempo de reação da Tabela 1.

$$SQres = 1\ 373 - 810 = 563$$

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	1	810	810	25,90	< 0,01
Resíduos	18	563	31,27		
Total	**9**	**1 373**	**72,26**		

O que indica que a regressão entre X e Y é significativa. O modelo Y = 80,5 +0,9.X pode ser considerado adequado para realizar predições de Y. O coeficiente r^2 de determinação para esse modelo é de 0,59 o que representa um poder apenas razoável de explicação dos valores de tempo de reação pela idade. Muito provavelmente outras variáveis estejam influenciando o tempo de reação.

Exemplo 3:

$$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2 = \sum_{i=1}^{n} (10 + 2x_i - 110)^2 = 13\ 600$$

Para obter a soma de quadrados acima, deveremos substituir em X_i todos os valores do tamanho do lote da Tabela 2.

$$SQtotal = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - 107,5)^2 = 13\ 660$$

Para obter a soma de quadrados acima, deveremos substituir em Y_i todos os valores de números de horas gastas da Tabela 2.

$$SQres = 13\,660 - 13\,600 = 60$$

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	1	13 600	13 600	1 813,33	< 0,01
Resíduos	8	60	7,5		
Total	**9**	**13 660**	**1 517,78**		

O que indica que a regressão entre X e Y é significativa. O modelo Y = 10 + 2.X pode ser considerado de boa qualidade para realizar predições de Y. O coeficiente r^2 de determinação para esse modelo é de 0,996.

Erro padrão de estimação e intervalos de predição

O erro padrão da estimação é um desvio padrão condicional, na medida em que indica o desvio padrão da variável dependente Y, dado um valor específico da variável dependente X. O erro padrão baseado em dados amostrais é dado por:

$$\hat{\sigma}_u = \sqrt{\frac{\sum (y - \hat{Y})^2}{n-2}}$$

Para fins de cálculo, é mais conveniente uma versão alternativa da fórmula:

$$\hat{\sigma}_u = \sqrt{S_y^2 . \left(1 - r^2\right)}$$

onde $S_y^2 = \frac{\sum_{i=1}^{n} \left(y - \overline{y}\right)^2}{n}$

O erro padrão pode ser usado para estabelecer um intervalo de predição para a variável dependente, dado um valor específico da variável independente.

Uma vez que o erro padrão de estimação está baseado em dados de amostra, é apropriado o uso da distribuição *t* de Student com n-2 graus de liberdade. Assim, um intervalo de predição para a variável dependente Y, em análise de regressão simples é:

$$\left[Y \pm t_{n-2;\alpha/2} . \hat{\sigma}_u\right]$$

Para os dados do exemplo 2, teríamos o erro padrão da estimação dado por:

Dado que $S_y^2 = 68{,}65$ e $r^2 = 0{,}59$ então

$$\hat{\sigma}_u = \sqrt{S_y^2 . \left(1 - r^2\right)} = \sqrt{68{,}65 . \left(1 - 0{,}59\right)} = 5{,}30$$

E o intervalo de predição, com 95% de confiança, para um valor de Y=112 seria:

$$[\overline{Y} \pm t_{n-2;\alpha/2} . \hat{\sigma}_u] = [112 \pm 2{,}10 . 5{,}30] = [100{,}87 , 123{,}13]$$

Ou seja, para uma pessoa com 35 anos, o tempo de reação predito estaria entre 100,87 e 123,13 segundos, com 95% de confiança.

Para os dados do exemplo 3 teríamos o erro padrão da estimação dado por:

Dado que $S_y^2 = 1\,366$ e $r^2 = 0{,}996$ então

$$\hat{\sigma}_u = \sqrt{S_y^2 . \left(1 - r^2\right)} = \sqrt{1\,366 . \left(1 - 0{,}996\right)^2} = 2{,}34$$

E o intervalo de predição, com 95% de confiança, para um valor predito de Y = 110 seria:

$$[\overline{Y} - t_{n-2;\alpha/2} . \hat{\sigma}_u] = [110 - 2{,}31 . 2{,}34] = [104{,}59; 115{,}41]$$

Ou seja, para um lote de tamanho 50, seriam necessárias de 104,59 a 115,41 horas, com 95% de confiança.

Análise de Resíduos

Os desvios $e_i = y_i - \hat{y}_i$ ($i = 1, ..., n$) são denominados resíduos e são considerados uma amostra aleatória dos erros. Por este fato, uma análise gráfica dos resíduos é, em geral, realizada para verificar as suposições assumidas para os erros ε_i.

Para verificação dos pressupostos necessários para ajuste de um modelo de regressão é necessário realizar uma Análise de Resíduos. Os 3 tipos de resíduos mais comumente utilizados são:

- Resíduos brutos;
- Resíduos padronizados;
- Resíduos estudentizados.

Ampliando seus conhecimentos

Análise de Regressão Múltipla

A regressão múltipla envolve três ou mais variáveis, ou seja, uma única variável dependente, porém duas ou mais variáveis independentes (explicativas).

A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples. Mesmo quando estamos interessados no efeito de apenas uma das variáveis, é aconselhável incluir as outras capazes de afetar Y, efetuando uma análise de regressão múltipla, por 2 razões:

a) Para reduzir os resíduos. Reduzindo-se a variância residual (erro padrão da estimativa), aumenta a força dos testes de significância;

b) Para eliminar a tendenciosidade que poderia resultar se simplesmente ignorássemos uma variável que afeta Y substancialmente.

Uma estimativa é tendenciosa quando, por exemplo, numa pesquisa em que se deseja investigar a relação entre a aplicação de fertilizante e o volume de safra, atribuímos erroneamente ao fertilizante os efeitos do fertilizante, mais a precipitação pluviométrica.

O ideal é obter o mais alto relacionamento explanatório com o mínimo de variáveis independentes, sobretudo em virtude do custo na obtenção de dados para muitas variáveis e também pela necessidade de observações adicionais para compensar a perda de graus de liberdade decorrente da introdução de mais variáveis independentes.

A equação da regressão múltipla tem a forma seguinte:

$Y = a + b_1x_1 + b_2x_2 + \ldots + b_k x_k + e_i$, onde:

- a = intercepto do eixo y;
- b_i = coeficiente angular da i-ésima variável;
- k = número de variáveis independentes.

Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis resulta um plano, e um problema de k variáveis resulta um hiperplano.

Também na regressão múltipla, as estimativas dos mínimos quadrados são obtidas pela escolha dos estimadores que minimizam a soma dos quadrados dos desvios entre os valores observados Y_i e os valores ajustados $\hat{Y}$.

Na regressão simples:

b = aumento em Y, decorrente de um aumento unitário em X.

Na regressão múltipla:

b_i = aumento em Y se X_i for aumentado de 1 unidade, mantendo-se constantes todas as demais variáveis X_j.

Atividades de aplicação

1. Os encargos diários com o consumo de gás propano (Y) de uma empresa dependem da temperatura ambiente (X). A tabela seguinte apresenta o valor desses encargos em função da temperatura exterior:

Temperatura (°C)	5	10	15	20	25
Encargos (dólares)	20	17	13	11	9

Seja $Y = \beta_0 + \beta_1 X + \varepsilon$ o correspondente modelo de regressão linear.

a) Determine, usando o método dos mínimos quadrados, a respectiva reta de regressão e represente-a no diagrama de dispersão.

b) Quantifique a qualidade do ajuste obtido e interprete.

c) Determine um intervalo de confiança a 95% para os encargos médios com gás propano num dia em que a temperatura ambiente é de 17ºC.

2. Suponha que um analista toma uma amostra aleatória de 9 carregamentos feitos recentemente por caminhões de uma companhia. Para cada carregamento, registra-se a distância percorrida em km (X) e o respectivo tempo de entrega (Y). Obteve-se:

$$\sum x_i = 6\,405, \sum y_i = 23,5,\ \sum x_i^2 = 56\,280\,75, \sum y_i^2 = 74,75, \sum x_i y_i = 20\,295.$$

a) Estime, usando o modelo de regressão linear, o tempo esperado de entrega para uma distância de 1 050km.

b) Comente a afirmação "o tempo de entrega é explicado em aproximadamente 94% pela distância percorrida".

3. Seja Y o número de chamadas telefônicas atendidas num determinado serviço de atendimento a clientes decorridos X minutos após as 8h30. Em determinado dia da semana observaram-se os seguintes pares de valores:

Tempo após 8h30(min)	1	3	4	5	6
Número de chamadas atendidas	2	5	10	11	12

Seja $Y = \beta_0 + \beta_1 X + \varepsilon$ o correspondente modelo de regressão linear.

a) Estime β_0 e β_1 usando o método dos mínimos quadrados e represente a correspondente reta de regressão no diagrama de dispersão.

b) Determine o correspondente coeficiente de determinação, bem como o coeficiente de correlação; como você interpreta os valores obtidos?

c) Estime a variância do erro.

d) Seja E [Y (2)] = E [Y | x = 2]. Estime E [Y (2)]; determine um intervalo de confiança para E [Y (2)] com 95% de confiança.

Gabaritos

Capítulo 1 – Conceitos e Aplicações

1.

a) É uma estratégia adequada. Se a amostra coletada for representativa da população, os resultados serão bastante confiáveis.

b) Também pode ser considerada uma estratégia adequada. A pesquisa atingirá, nos locais de venda, o público-alvo do novo produto e apresentará resultados confiáveis.

c) Esta é uma estratégia mais qualitativa, denominada discussão em grupo (grupo focal). Os resultados obtidos apresentam muitas informações em profundidade, porém sem muita representatividade, pelo número reduzido da amostra.

2.

a) Esta é uma estratégia adequada, pois compara dois grupos de pacientes homogêneos e possibilita avaliar o efeito do novo medicamento. É preciso, no entanto, garantir que o número de pacientes escolhidos seja em número satisfatório.

b) Não é uma estratégia adequada. Não se devem disponibilizar medicamentos novos no mercado sem que antes tenham sido avaliados em laboratório e outros experimentos controlados. E nada garante que será atingida a população alvo de interesse do estudo.

c) É uma estratégia parcialmente adequada. Deve-se avaliar se os pacientes deste hospital representam de forma satisfatória a população alvo ou se é apenas uma escolha por conveniência. Pode ser que os pacientes hospitalizados sejam pacientes em estado mais grave, o que poderá viesar os resultados do estudo.

3.

a) É uma estratégia adequada. Escolhendo uma amostra representativa do lote conseguiremos, com uma boa margem de confiança, avaliar a qualidade do lote.

b) Não é adequado. Não devemos liberar mercadorias para o comércio sem que antes a sua qualidade tenha sido avaliada.

c) Não é adequado. Avaliar 10% do lote pode ser exaustivo ou insuficiente, dependendo do tamanho do lote. Existem maneiras definidas de calcular o número de amostras que vão representar satisfatoriamente a população.

Capítulo 2 – Análise Exploratória de Dados

1. Construindo-se a tabela de freqüência dos dados considerando 5 classes:

$k = 1 + 3{,}3.\log(n)$ $\quad h_i = \frac{AT}{k}$ $\quad AT = 119 - 50$

$k = 1 + 3{,}3.\log(20)$ $\quad h_i = \frac{69}{5}$ $\quad AT = 69$

$k = 1 + 3{,}3 \,.\, 1{,}30103$ $\quad \mathbf{h_i = 13{,}80}$

k = 5,29

Para facilitar a construção da tabela de freqüências, utilizaremos classe igual a 5 e intervalo de classe igual a 15.

Classe	Freqüência	%
50 |— 65	8	40
65 |— 80	7	35
80 |— 95	4	20
95 |— 110	0	0
110 |— 125	1	5

Podemos observar que a grande maioria das instituições (75%) apresentou lucro de até 80 milhões de dólares enquanto que uma delas apresentou um lucro muito superior às demais (119 milhões de dólares).

2. Construindo a tabela com os dados do problema obteremos:

i	Pesos (kg)	f_i	Pm_i	fr_i	%
1	48 |— 53	10	50,5	0,20	20
2	53 |— 58	7	55,5	0,14	14
3	58 |— 63	5	60,5	0,10	10
4	63 |— 68	7	65,5	0,14	14
5	68 |— 73	5	70,5	0,10	10
6	73 |— 78	6	75,5	0,12	12
7	78 |— 83	6	80,5	0,12	12
8	83 |— 88	1	85,5	0,02	2
9	88 |— 93	1	90,5	0,02	2
10	93 |— 98	2	95,5	0,04	4
–	**TOTAL**	**50**		**1**	**100**

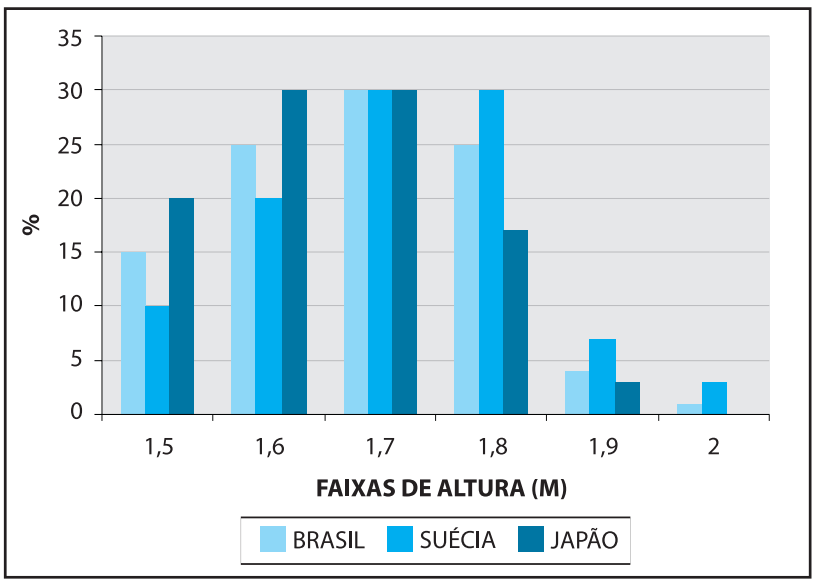
Fazendo a leitura da tabela:

a) 58 **b)** 68 **c)** 5 **d)** 50

e) 65,5 **f)** 10 **g)** 29 **h)** 16

i) 23 **j)** 4% **k)** 34% **l)** 20%

3. Um possível gráfico para representar a distribuição de altura da população dos 3 países poderia ser um histograma:



4. Podemos observar, pela interpretação dos ramos-e-folhas, que as duas corretoras apresentam porcentagens médias de lucros semelhantes, por volta de 5,0%. Por outro lado, a corretora B apresenta uma variabilidade muito menor que a corretora A. A corretora B, portanto apresenta um desempenho muito mais homogêneo que a corretora A.

Capítulo 3 – Medidas de Posição e Variabilidade

1. A. O mais provável seria ganhar menos, pois se o terceiro quartil é de R$ 5.000,00, significa que 75% dos salários são inferiores a este valor, a despeito da média ser de R$ 10.000,00 muito provavelmente influenciada por salários muito elevados dos altos cargos desta empresa.

B. Apresentaria-me na empresa Y, pois lá é praticamente certo que meu salário seria muito próximo da média de R$ 7.000,00 dado que os salários praticamente não apresentam variabilidade; quase todos recebem o mesmo salário.

2. B. O somatório dos valores e o número deles.

3. B. 60.

4. C. a mediana.

5. C. zero.

6. B. a média e a mediana.

7. A. moda.

8. A. desvio padrão e média.

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

10. A. R$ 1.050,00

11. A. média

12. D. zero

13. B. ao desvio padrão de X, multiplicado pela constante 5

$$\overline{X}_x = \frac{-2-1+0+1+2}{5} = 0$$

$$\overline{X}_Y = \frac{220+225+230+235+240}{5} = \frac{1\,150}{5} = 230$$

$\overline{X}_x = 0$

x_i	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2 \cdot f_i$
−2	−2	4	4
−1	−1	1	1
0	0	0	0
1	1	1	1
2	2	4	4
TOTAL			**10**

$$S^2 = \frac{10}{4} \rightarrow S^2 = 2{,}5$$

$$S = \sqrt{2{,}5} \rightarrow S = 1{,}58$$

$\overline{X}_Y = 230$

x_i	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2 \cdot f_i$
220	−10	100	100
225	−5	25	25
230	0	0	0
235	5	25	25
240	10	100	100
TOTAL			**25**

$$S^2 = \frac{250}{4} \rightarrow S^2 = 62{,}5$$

$$S = \sqrt{62{,}5} \rightarrow S = 7{,}905$$

$$\frac{7{,}905}{1{,}58} = 5 \text{ (constante)}$$

Capítulo 4 – Introdução à Probabilidade

1.

a) S={KKK, KKC, KCK, CKK, KCC, CKC, CCK, CCC}

b) S={MMM, MMF, MFM, FMM, MFF, FMF, FFM, FFF}

c) S={(1,1), (1,2), …, (1,6), (2,1), …, (2,6), ...,(6,1), ..., (6,6)}

d) S={DD, DV, VD, VV}

e) S={BB, BA, AB, AA}

2.

a) A={(3,6), (4,5), (5,4), (6,3)}

b) B={(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}

c) P(A) = 4/36

d) P(B) = 6/36

e) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/36 + 6/36 - 0 = 10/36$

f) $P(A \cap B) = 0$

3.

a) P(retirar uma bola branca da urna "A") = 5/10

b) P(retirar uma bola branca ou uma vermelha da urna "A") = 8/10

c) P(retirar uma bola branca e uma vermelha da urna "A") = 0

d) P(retirar duas bolas vermelhas da urna "A", com reposição) = (3/10).(3/10) = 9/100

e) P(retirar duas bolas pretas da urna "A", sem reposição) = (2/10).(1/10) = 2/100

4.

$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y) = 3/5 + 4/7 - (3/5 \,.\, 4/7) = 29/35 = 82{,}86\%$

5.

a) P(H) = 60/100 = 0,6 ou 60%.

b) $P(M \cap NE) = 26/100 = 0{,}26$ ou 26%.

c) P(NE) = 65/100 = 0,65 ou 65%

d) $P(H \cap NE) = 39/100 = 0{,}39$ ou 39%.

e) P(M/E) = 14/35 = 0,4 ou 40%

f) P(NE/H) = 39/60 = 0,65 ou 65%

6.

a) $P((B1 \cap B2) \cup (A1 \cap A2) \cup (P1 \cap P2)) = (4/15 \cdot 5/13) + (5/15 \cdot 6/13) + (6/15 \cdot 2/13) = 62/195$

b) $P(A1 \cap P2) = 5/15 \cdot 2/13 = 10/195$

c) $P((A1 \cap P2) \cup (P1 \cap A2)) = (5/15 \cdot 2/13) + (6/15 \cdot 6/13) = 46/195$

d) $P(B1\ C\ B2^{C}) = 4/15 \cdot 8/13 = 32/195$

7.

$P(W) = (1/10 \cdot 3/4) + (3/5 \cdot 1/6) + (3/10 \cdot 1/20) = 3/40 + 3/30 + 3/200 = 0,19$

a) $P(A/W) = P(W \cap A)/ P(W) = P(A) \cdot P(W/A) / P(W) = (1/10 \cdot 3/4)/0,19 = 0,3947$

b) $P(B/W) = P(W \cap B)/ P(W) = P(B) \cdot P(W/B) / P(W) = (3/5 \cdot 1/6)/0,19 = 0,5263$

c) $P(C/W) = P(W \cap C)/ P(W) = P(C) \cdot P(W/C) / P(W) = (3/10 \cdot 1/20)/0,19 = 0,0789$

8.

$P(D) = (0,4 \cdot 0,03) + (0,5 \cdot 0,05) + (0,1 \cdot 0,02) = 0,012 + 0,025 + 0,002 = 0,039$

a) $P(M_1/D) = P(M_1 \cap D)/ P(D) = P(M_1) \cdot P(D/M_1) / P(D) = (0,4 \cdot 0,03)/0,039 = 0,3077$

b) $P(M_2/D) = P(M_2 \cap D)/ P(D) = P(M_2) \cdot P(D/M_2) / P(D) = (0,5 \cdot 0,05)/0,039 = 0,6410$

c) $P(M_3/D) = P(M_3 \cap D)/ P(D) = P(M_3) \cdot P(D/M_3) / P(D) = (0,1 \cdot 0,02)/0,039 = 0,0513$

9.

a) Sabemos que $\Sigma_i\ p(x_i) = 1$, assim: $k/2 + 0,15 + 3k + 0,1 + 0,05 = 1$, ou seja, $3,5k + 0,30 = 1$ e isto implica que $k = 0,2$

b) $P(X>22) = P(X=23) + P(X=24) = 0,15$ ou 15%

c) $P(20<X<24) = P(X=21) + P(X=22) + P(X=23) = 0,85$ ou 85%

d) Pela definição de esperança de uma variável aleatória discreta:

$E(X) = \sum_{i=1}^{\infty} x_i.p.(x_i)$.

Assim,

$E(X) = (20 \,.\, 0,1) + (21 \,.\, 0,15) + (22 \,.\, 0,6) + (23 \,.\, 0,1) + (24 \,.\, 0,05) = 21,85$ dias

e) Pela definição de variância, temos que: $Var(X) = E(X^2) - [E(X)]^2$

Temos que $E(X^2) = (20^2 \,.\, 0,1) + (21^2 \,.\, 0,15) + (22^2 \,.\, 0,6) + (23^2 \,.\, 0,1) + (24^2 \,.\, 0,05) = 478,25$ e assim $Var(X) = 478,25 - (21,85^2) = 0,8275$

f) Custo da obra: 16.000 + (750 . 21,85) = 32.387,50 euros.

Custo da obra + lucro = 34.887,50 euros.

Capítulo 5 – Distribuição Binomial, Distribuição Poisson e Distribuição Normal

1.

a) $P(X \leq 8) = \sum_{x=0}^{8} \binom{10}{x}.0,3^x.0,7^{10-x} = 0,999$

b) $P(X=7) = \binom{10}{7}.0,3^7.0,7^3 = 0,009$

c) $P(X \geq 6) = \sum_{x=7}^{10} \binom{10}{x}.0,3^x.0,7^{10-x} = 0,047$

2.

a) $0,9^5 = 0,59$

3. P(no máximo duas peças defeituosas) =

$P(X=0) + P(X=1) + P(X=2) = \sum_{x=0}^{2} \binom{10}{x}.0,05^x.0,95^{10-x} = 0,9885$ ou 98,85%

4. O número de navios petroleiros que chegam a determinada refinaria, a cada dia, tem distribuição de Poisson, com parâmetro $\lambda = 2$. As atuais instalações do porto podem atender a três petroleiros por dia. Se mais de 3 navios aportarem por dia, os excedentes devem seguir para outro porto.

a) $P(X > 3) = 1 - \sum_{x=0}^{3} \frac{e^{-\lambda}.\lambda^{x}}{x!} = 1 - 0,857 = 0,143$

b) Se as instalações forem ampliadas para permitir mais um petroleiro, teremos:

$P(X \leq 4) = \sum_{x=0}^{4} \frac{e^{-\lambda}.\lambda^{x}}{x!} = 0,947$

c) $E(X) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}.\lambda^{x}}{x!} = \sum_{x=0}^{\infty} x \frac{e^{-2}.2^{x}}{x!} = 2$

d) 1 ou 2 petroleiros. $P(X=1) = P(X=2) = 0,2707$

e) Qual é o número esperado de petroleiros a serem atendidos diariamente?

Se chegarem 0, 1, 2 ou 3 petroleiros todos serão atendidos. Se vierem mais de 3 petroleiros, somente 3 serão atendidos. Dessa forma:

Número esperado:

$0.P(X=0) + 1.P(X=1) + 2.P(X=2) + 3.P(X \geq 3) = 1,78$

f) Se vierem 0,1, 2 ou 3 petroleiros nenhum precisará ir a outros portos. Caso mais de 3 petroleiros cheguem, apenas 3 podem ser recebidos. Assim:

Número esperado:

$1.P(X=4) + 2.P(X=5) + 3.P(X=6) + 4.P(X=7) + ... = 0,22$

5.

c) $P(X=0) = \frac{e^{-5}.5^{0}}{0!} = 0,0067$

6.

a) –9,6 e 29,6

Para obtermos o valor padronizado 1,96, faremos: $\frac{X-10}{10} = 1,96$

Assim, $X = 29,6$

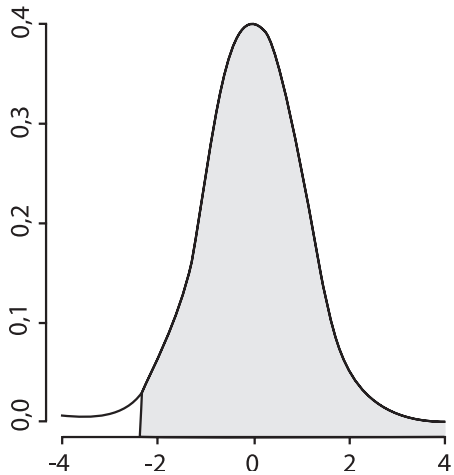
Para obtermos o valor padronizado –1,96, faremos: $\frac{X-10}{10} = -1,96$

Assim, $X = -9,6$

7. $P(X < 5\,000) = P\left(Z < \dfrac{5\,000 - 15\,000}{2\,000}\right) = P(Z < -5) = 0{,}0000002871$

8. P(X≥772N)

$$= P\left(Z \geq \frac{772 - 800}{\sqrt{144}}\right) = P(Z \geq -2{,}33) = 1 - P(Z \leq -2{,}33) = 1 - 0{,}0098 = 0{,}99$$



Capítulo 6 – Estimação de Parâmetros

1.

a) derivando a função de verossimilhança.

2. $\mu_1 = 2$

$\mu_1 = 1$

$\mu_3 = \bar{x} = \sum \dfrac{x}{n} = \dfrac{21}{15} = 1{,}4$

μ_3 é o melhor estimador porque leva em consideração todos os valores da amostra, proporcionando um resumo de dados e por isso pode ser considerado mais confiável.

3. Os limites do intervalo são obtidos a partir da seguinte expressão:

$$\left[\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = \left[78{,}3 - 2{,}58 \cdot \frac{2}{\sqrt{25}}; 78{,}3 + 2{,}58 \cdot \frac{2}{\sqrt{25}}\right] = [77{,}27; 79{,}33]$$

4.

a) 95%

$$\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right]$$

$$= \left[0,298 - 2,78 \cdot \frac{0,024}{\sqrt{5}}; 0,298 + 2,78 \cdot \frac{0,024}{\sqrt{5}}\right] = [0,268; 0,328]$$

b) 99%

$$\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right] =$$

$$= \left[0,298 - 4,60 \cdot \frac{0,024}{\sqrt{5}}; 0,298 + 4,60 \cdot \frac{0,024}{\sqrt{5}}\right] = [0,248; 0,348]$$

5. $$\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}\right) =$$

$$= \left(0,80 - 2,58.\sqrt{\frac{0,80.(0,20)}{200}}; \quad 0,80 + 2,58.\sqrt{\frac{0,80.(0,20)}{200}}\right)$$

$$= (0,723 ; 0,873)$$

O valor 0,90 declarado pelo fabricante, não está incluído no intervalo. Portanto, não temos evidências de que a declaração do fabricante seja legítima, ao nível de significância de 1%.

6.

a) $$\left[\bar{X} - z_{0,05} . \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{0,05} . \frac{\sigma_0}{\sqrt{n}}\right] =$$

$$\left[4,52 - 1,64 . \frac{4}{\sqrt{100}}; 4,52 + 1,64 . \frac{4}{\sqrt{100}}\right] = (3,864; 5,176)$$

b) Sim, a probabilidade do verdadeiro valor da média (valor populacional) estar incluído nos limites do intervalo encontrado é de 90%.

7.

a) O verdadeiro valor do salário inicial médio estará entre 8 e 10 salários mínimos com probabilidade de 95%.

b) Quanto maior o tamanho da amostra, menor é o erro de estimativa e portanto a média amostral estará mais próxima da média populacional. Veja, por exemplo em

$\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right]$ o erro de estimativa $z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}$ é menor a medida que se aumenta o valor de n.

8. Queremos obter uma amostra para estimar a média de uma distribuição normal que respeite a seguinte probabilidade:

$$P\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}\right] = 0{,}92$$

O valor de Z na tabela será obtido encontrando a área 0,5 – $\alpha/2$ = 0,5 – 0,04 = 0,46. Este valor é 1,75.

Assim, $P\left[\bar{X} - 1{,}75 \cdot \frac{\sqrt{30}}{\sqrt{n}}; \bar{X} + 1{,}75 \cdot \frac{\sqrt{30}}{\sqrt{n}}\right] = 0{,}92$

Como o erro de estimativa, segundo o enunciado, não deve ser superior a 3 unidades, então:

$1{,}75 \cdot \frac{\sqrt{30}}{\sqrt{n}} = 3$. Isolando n, teremos que ele será maior que 10,28.

9. Neste problema, o nível de confiança fixado é de 90% e conseqüentemente, o nível de significância é de 10%.

a) Como não temos uma estimativa prévia da proporção amostral, consideramos p=0,05. Desta forma, teremos:

$$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 \cdot \frac{1}{4} = \left(\frac{z_{\alpha/2}}{2e}\right)^2 \rightarrow n = \left(\frac{1{,}64}{2.0{,}05}\right)^2 = 268{,}96$$

b) Agora temos uma informação prévia sobre a proporção amostral (0,8) e assim o cálculo da amostra será:

$$n = \left(\frac{z_{\alpha/2}}{e}\right)^2 . p_0 . (1 - p_0) = \left(\frac{1{,}64}{0{,}05}\right)^2 . 0{,}20 . 0{,}80 = 172{,}13$$

Capítulo 7 – Testes de Hipóteses: conceitos

1.

a) A população é a totalidade de alunos do Curso X. A amostra é composta pelos 80 alunos do Curso, selecionados aleatoriamente. O parâmetro de interesse é a proporção de alunos favoráveis a eliminação da disciplina de Estatística do currículo. O teste adequado seria para testar a proporção de uma amostra.

b) A população é a totalidade de pessoas obesas com certa idade. A amostra é composta pelas 20 pessoas obesas daquela faixa etária, selecionadas aleatoriamente. O parâmetro de interesse é a média de perda de peso, ou seja peso antes – peso depois (do curso). O teste adequado seria para comparar amostras relacionadas.

c) A população é a totalidade de moradores fumantes da cidade. A amostra é composta pelas 100 pessoas fumantes, selecionadas aleatoriamente. Um dos parâmetros de interesse pode ser a média de cigarros consumidos. O teste adequado seria para testar a média de uma amostra.

2.

a) H_0 = opinião antes = opinião depois

H_a = opinião antes ≠ opinião depois

b) Embora a maioria das pessoas tenha se manifestado mais favorável ao candidato, não seria prudente afirmarmos que este resultado possa ser considerado estatisticamente significativo.

c) Com este tamanho de amostra já é possível realizar um teste de significância. Muito provavelmente, iremos rejeitar a hipótese nula, de igualdade das opiniões. Poderemos, se o teste comprovar, inferir os resultados para toda a população e afirmar com um certo nível de confiança, que se passou a ter melhor impressão sobre o candidato após a apresentação.

d) Um teste para comparação da proporção de duas amostras relacionadas (antes e depois da apresentação).

3.

a) H_0 = vendas sem brinde = vendas com brinde

H_a = vendas sem brinde ≠ vendas com brinde

b) Com exceção de uma loja, todas as 5 demais apresentaram maiores índices de venda ao oferecer o brinde. É um forte indicativo de maiores vendas com oferta do brinde, embora o número de lojas participantes deste experimento possa ser considerado baixo.

c) O tipo de teste mais adequado seria um teste para comparação de médias de duas amostras independentes, embora pudesse ser utilizado também um teste para comparação de médias de duas amostras relacionadas, desde que bem justificado o critério de pareamento das unidades observadas.

4.

a) H_0 = eficácia relativa comerciais de 15 segundos = eficácia relativa comerciais de 30 segundos

H_a = eficácia relativa comerciais de 15 segundos < eficácia relativa comerciais de 30 segundos

b) Caso o tamanho de amostra seja satisfatório e a suposição de normalidade seja comprovada, pode ser aplicado um teste paramétrico para comparação de duas amostras independentes. Caso os pressupostos para aplicação de um teste paramétrico não sejam atendidos, podemos recorrer a um teste não paramétrico para comparação de duas amostras independentes. O nível de significância mais indicado seria de 1% ou 5%.

c) Nas 4 variáveis avaliadas podemos observar que os comerciais de 30 segundos apresentaram uma melhor avaliação em relação aos comerciais de 15 segundos.

Capítulo 8 – Testes de Hipóteses

1. As hipóteses a serem testadas são:

H_0: As produções médias de milho estão de acordo com a especificação do fabricante;

H_a: A produção média de milho não se ajusta à distribuição especificada pelo fabricante.

Aplicando o teste Qui-quadrado para testar a aderência dos dados à distribuição especificada pelo fabricante, temos:

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \frac{(13-12)^2}{12} + \frac{(18-20)^2}{20} + \ldots + \frac{(11-13)^2}{13} = 3{,}04$$

Consultando a tabela de valores críticos, considerando k–1 = 5 graus de liberdade e $\alpha = 0{,}05$, temos $x^2 = 11{,}1$. Como o valor calculado é inferior ao valor crítico, não rejeitamos a hipótese nula e podemos concluir que os dados se ajustam satisfatoriamente à distribuição especificada pelo fabricante.

2. As hipóteses a serem testadas são:

H_0: a nota média dos estudantes de escola pública não difere da nota média dos estudantes da escola particular;

Ha: a nota média dos estudantes de escola pública difere da nota média dos estudantes da escola particular.

Aplicando o teste t de Student para comparação de duas amostras independentes, temos que verificar primeiramente se as variâncias podem ser consideradas iguais. Construindo o intervalo de confiança para a razão de variâncias temos:

$$\left[\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_2}; \frac{S_1^2}{S_2^2}\cdot\frac{1}{F_1}\right] = \left[\frac{64}{100}\cdot\frac{1}{1{,}4833}; \frac{64}{100}\cdot 1{,}4833\right] = (0{,}43\,;\,0{,}94)$$

Desta forma as variâncias não são iguais.

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(75{,}9 - 74{,}5)}{\sqrt{\frac{64}{117} + \frac{100}{200}}} = 1{,}3682$$

Consultando a tabela de valores críticos, considerando $n_1 + n_2 - 2 = 315$ graus de liberdade e $\alpha = 0{,}05$, temos $t_c = 1{,}96$. Como o valor calculado é inferior ao valor crítico, não rejeitamos a hipótese nula e podemos concluir que as notas médias das duas escolas não diferem.

3. As hipóteses a serem testadas são:

H_0: a nova droga não baixa a febre, ou seja, Diferença = 0;

H_a: a nova droga baixa a febre, ou seja, Diferença ≠ 0.

Aplicando o teste t de Student para comparação de duas amostras relacionadas, temos:

$$S_d = \sqrt{\frac{\sum d^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{80 - \left(15.(1{,}866)^2\right)}{14}} = 1{,}408$$ e a estatística do teste será:

$$t = \frac{1{,}866}{1{,}408 / \sqrt{15}} = 5{,}131$$

Consultando a tabela de valores críticos, considerando n–1 = 14 graus de liberdade e $\alpha = 0{,}05$ (bilateral), temos $t_c = 2{,}14$. Como o valor calculado é superior ao valor crítico, rejeitamos a hipótese nula e podemos concluir que a nova droga baixa a febre significativamente.

4. As hipóteses a serem testadas são:

H_0: a proporção de animais com verminose é igual nos dois grupos;

H_a: a proporção de animais com verminose é inferior no grupo que teve alteração da dieta.

O teste, portanto, é unilateral e aplicando o teste Z para proporção, temos:

$$p = \frac{n_1.p_1 + n_2.p_2}{n_1 + n_2} = \frac{(500.0{,}10) + (100.0{,}04)}{600} = 0{,}09$$

$$S_p = \sqrt{\frac{p.(1-p)}{n_1} + \frac{p.(1-p)}{n_2}} = \sqrt{\frac{0{,}09.0{,}91}{500} + \frac{0{,}09.0{,}91}{100}} = 0{,}031$$

$$Z = \frac{p_1 - p_2}{S_p} = \frac{0{,}10 - 0{,}04}{0{,}031} = 1{,}93$$

Consultando a tabela de valores críticos da distribuição normal padrão, considerando $\alpha = 0{,}01$, temos $Z_c = 2{,}33$. Como o valor calculado é inferior ao valor crítico, não rejeitamos a hipótese nula e podemos concluir que a doença não diminuiu significativamente de intensidade.

5. As hipóteses a serem testadas são:

H_0: não existe diferença de satisfação entre os 3 hospitais;

H_a: existe pelo menos uma diferença entre os hospitais, com relação à média de satisfação.

Realizando o Teste F, de Análise de Variâncias, temos:

$$SQA = \sum \frac{T_k^2}{n_K} - \frac{T^2}{N} = \frac{(873)^2}{10} + \frac{(898)^2}{15} + \frac{(954)^2}{13} - \frac{(2725)^2}{38} =$$
$$= 76\,212,9 + 53\,760,267 + 70\,008,92 - 195\,411,1842 = 4\,570,9$$

$$SQT = \sum_{i=1}^{n}\sum_{k=1}^{k} X^2 - \frac{T^2}{N} = 200\,623 - 195\,411,1842 = 5\,211,82$$

e SQE = SQT – SQA = 5 211,82 – 4 570,9 = 640,92

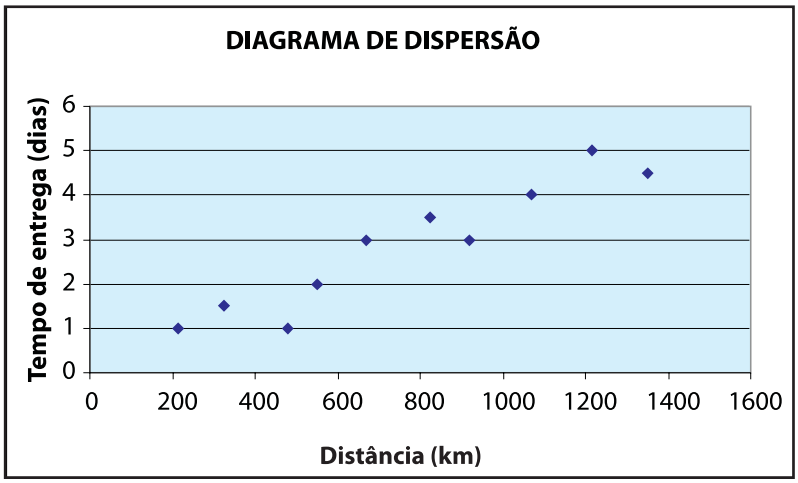
Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Entre grupos	4 570,90	2	2 285,450	124,8
Erro amostral	640,92	35	18,312	
Total	**5 211,82**	**37**		

O valor crítico de F, definido pelo nível de significância ($\alpha = 0,05$) e pelos graus de liberdade 2 e 35 é igual a 3,30. Como $F_{cal} > F_{crit}$ devemos rejeitar a hipótese nula. Os hospitais diferem em relação à satisfação média.

Capítulo 9 – Análise de Correlação e Medidas de Associação

1.

a)



b) $C(X,Y)=\frac{\sum(X_i-\bar{X}).(Y_i-\bar{Y})}{n}=\frac{4\,653}{10}=465,3$

$r=\frac{C(X,Y)}{S_Y.S_X}=\frac{465,3}{360,26.1,36}=0,9497$

c) $r^2=(r)^2=(0,9497)^2=0,9019$

d) $t_c=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{0,9497\sqrt{8}}{\sqrt{1-0,9019}}=8,576$

Comparando o valor calculado com o valor crítico, considerando 8 graus de liberdade e 5% de significância temos $t_{crítico} = 2,31$. Assim, podemos considerar o coeficiente de correlação altamente significativo.

2. $t_c=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{0,50\sqrt{8}}{\sqrt{1-0,25}}=1,63$

Comparando o valor calculado com o valor crítico, considerando 8 graus de liberdade e 5% de significância temos $t_{crítico} = 2,31$. Assim, não podemos considerar o coeficiente de correlação significativo. Não existe correlação entre a renda familiar e os débitos a descoberto de curto prazo.

3. $C(X,Y)=\frac{\sum(X_i-\overline{X}).(Y_i-\overline{Y})}{n}=\frac{654}{8}=81{,}75$

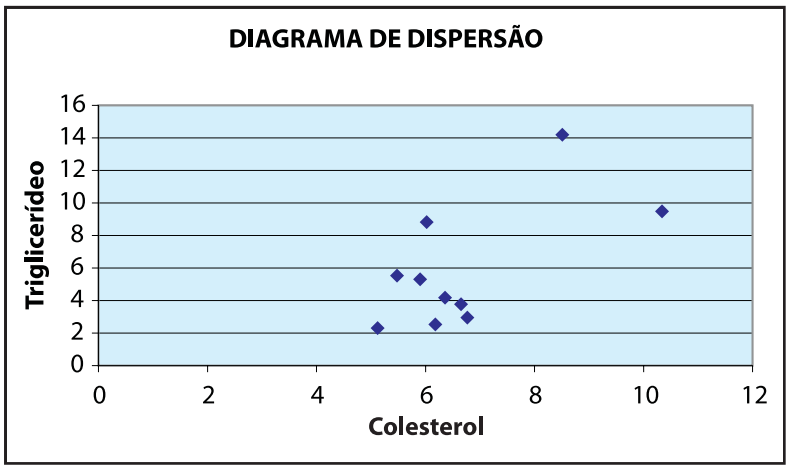
$$r=\frac{C(X,Y)}{S_Y.S_X}=\frac{81{,}75}{12{,}77.10{,}22}=0{,}626$$

$$t_c=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{0{,}626\sqrt{6}}{\sqrt{1-0{,}392}}=1{,}967$$

Comparando o valor calculado com o valor crítico, considerando 6 graus de liberdade e 5% de significância temos $t_{crítico}$ = 2,45. Assim, podemos considerar o coeficiente de correlação não significativo, ou seja, não existem evidências de correlação significativa entre habilidade verbal e habilidade matemática.

4.

a)



b) baseado no diagrama acima, não está muito clara a existência de relação linear entre colesterol e triglicerídeos.

Paciente	Colesterol (mmol/l)	Triglicerídeos (mmol/l)	Postos Colesterol	Postos Triglicerídeos	d_i	d_i^2
1	5,12	2,30	1	1	0	0
2	6,18	2,54	5	2	3	9
3	6,77	2,95	8	3	5	25
4	6,65	3,77	7	4	3	9

Paciente	Colesterol (mmol/l)	Triglicerídeos (mmol/l)	Postos Colesterol	Postos Triglicerídeos	d_i	d_i^2
5	6,36	4,18	6	5	1	1
6	5,90	5,31	3	6	–3	9
7	5,48	5,53	2	7	–5	25
8	6,02	8,83	4	8	–4	16
9	10,34	9,48	10	9	1	1
10	8,51	14,20	9	10	–1	1
Soma						**96**

c) $$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6.96}{1000 - 10} = 0{,}418$$

Para verificar a significância do valor observado de r_s podemos usar a expressão de t de Student

$$t = r_s . \sqrt{\frac{n-2}{1-r_s^2}} = 0{,}418 . \sqrt{\frac{8}{1-0{,}1748}} = 1{,}30$$

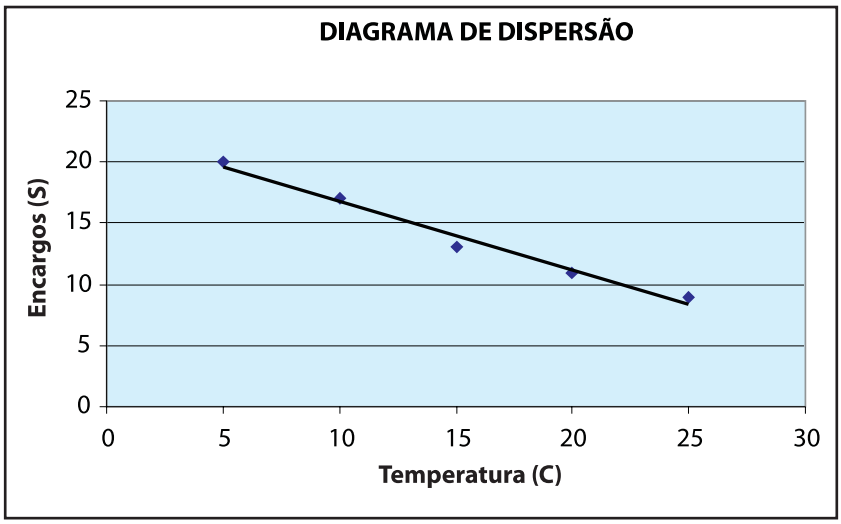
Comparando o valor calculado com o valor crítico, considerando 8 graus de liberdade e 5% de significância temos $t_{crítico} = 2{,}31$. Assim, podemos considerar o coeficiente de associação significativo, ou seja, existem evidências de correlação significativa entre colesterol e triglicerídeos.

Capítulo 10 – Análise de Regressão

1. $\hat{\beta}_1 = \dfrac{\sum x_i.y_i - n.\bar{y}.\bar{x}}{\sum x_i^2 - n.\bar{x}^2} = \dfrac{910 - 5.14.15}{1375 - 5.225} = -0{,}56$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1.\bar{x} = 14 - (-0{,}56).15 = 22{,}4$

Então $\hat{Y} = 22{,}4 - 0{,}56X$.



b) Dado que $\bar{y} = \dfrac{70}{5} = 14$

$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2 = \sum_{i=1}^{n} (22{,}4 - 0{,}65x_i - 14)^2 = 78{,}4$

$SQres = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (y_i - 22{,}4 - 0{,}65x_i)^2 = 1{,}6$

SQtotal = 78,4 + 1,6 = 80

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	1	78,4	78,4	147	< 0,001
Resíduos	3	1,6	0,53		
Total	**4**	**80**	**20**		

A regressão pode ser considerada altamente significativa ($p < 0,001$). O coeficiente de determinação calculado a partir dos dados da Anova, $r^2 = 78,4/80 = 0,98$. Pode se considerar bastante satisfatória a qualidade do ajuste.

c) $S_y^2 = \frac{\sum_{i=1}^{n}(y-\bar{y})^2}{n} = \frac{80}{5} = 16$

$\hat{\sigma} = \sqrt{S_y^2.(1-r^2)} = \sqrt{16.(1-0,98)} = 0,565$

$\hat{} = 22,4 - 0,56 \,.\, 17 = 12,88$

$[\hat{Y} \pm t_{n-2;\,\alpha/2}.\hat{\sigma}_u] = [12,88 \pm 3,18\,.\,0,565] = [11,08; 14,68]$

2.

a) $\hat{\beta}_1 = \frac{\sum x_i.y_i - n.\bar{y}.\bar{x}}{\sum x_i^2 - n.\bar{x}^2} = \frac{20\,295 - 9.2,61.711,67}{5\,628\,075 - 9.(711,66)^2} = \frac{3\,577,87}{106\,993,4} = 0,00334$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1.\bar{x} = 2,611 - 0,00334.711,66 = 0,234$

Então $\hat{Y} = 0,234 + 0,00334.X = 0,234 + 0,00334 \,.\, 1\,050 = 3,741$ dias

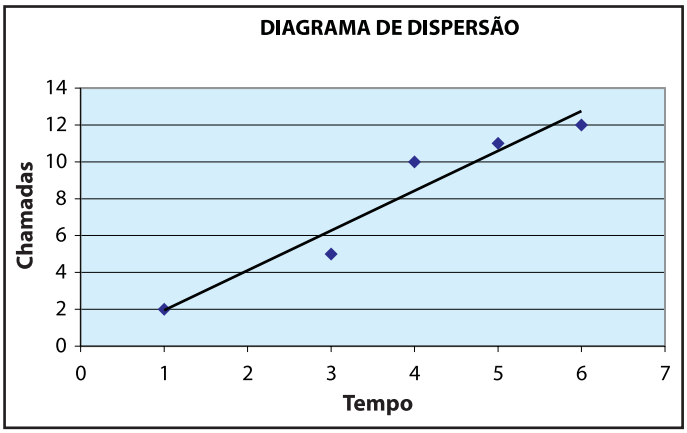
b) Isto significa que 94% da variação do tempo de entrega está associada à distância a ser percorrida e outras variáveis como: região urbana ou rural, clima durante o percurso, treinamento do motorista etc, são responsáveis pelos demais 6%. No entanto, essas variáveis não foram observadas nesse estudo.

3.

a) $\hat{\beta}_1 = \frac{\sum x_i.y_i - n.\bar{y}.\bar{x}}{\sum x_i^2 - n.\bar{x}^2} = \frac{184 - 5.8.3,8}{87 - 5.(3,8)^2} = \frac{32}{14,8} = 2,16$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1.\bar{x} = 8 - 2,16.3,8 = -0,21$

Então $\hat{Y} = -0,21 + 2,16.X$



b) $SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2 = \sum_{i=1}^{n} (-0{,}21 + 2{,}16x_i - 8)^2 = 69{,}05$

$SQres = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (y_i + 0{,}21 - 2{,}16x_i)^2 = 4{,}8109$

$SQtotal = 69{,}05 + 4{,}8109 = 73{,}8609$

Assim $r^2 = \dfrac{SQres}{SQtotal} = \dfrac{69{,}05}{73{,}86} = 0{,}9349$ e $r = \sqrt{r^2} = 0{,}9668$

O coeficiente de determinação calculado nos indica que é bastante satisfatória a qualidade do ajuste. A relação entre as duas variáveis pode ser considerada bastante forte, pela análise do coeficiente de correlação.

c) $\hat{\sigma}_u = \sqrt{\dfrac{\sum (y - \hat{Y})^2}{n-2}} = \sqrt{\dfrac{4{,}8109}{3}} = 1{,}266$

d) $E[Y(2)] = -0{,}21 + 2{,}16 \cdot 2 = 4{,}11$

$[\hat{Y} \pm t_{n-2;\,\alpha/2} . \hat{\sigma}_u] = [4{,}11 \pm 3{,}18 . 1{,}266] = [0{,}08;\ 8{,}13]$

Anexo I

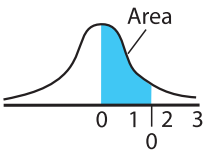


Tabela de valores críticos – Normal

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3112	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.1965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0..4982	0.4982	0.4983	0.4983	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4988	0.4989	0.4989	0.4990	0.4990
3.1	0.49903									

Anexo II

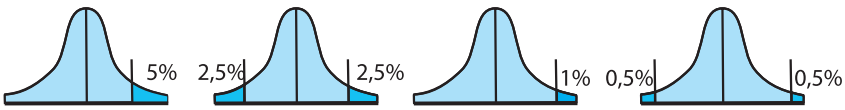


Tabela de valores críticos – t de Student

df	0.05	0.025	0.01	0.005
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	.3.55
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
50	1.676	2.009	2.403	2.678
100	1.660	1.984	2.364	2.626
∞	1.645	1.960	2.326	2.576

Anexo III

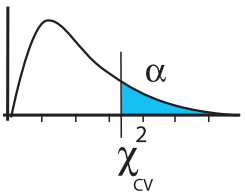


Tabela de valores críticos – Qui-quadrado				
df	**0.05**	**0.025**	**0.01**	**0.005**
1	3.84	5.02	6.63	7.88
2	5.99	7.38	9.21	10.60
3	7.82	9.35	11.35	12.84
4	9.49	11.14	13.28	14.86
5	11.07	12.83	15.09	16.75
6	12.59	14.45	16.81	18.55
7	14.07	16.01	18.48	20.28
8	15.51	17.54	20.09	21.96
9	16.92	19.02	21.66	23.59
10	18.31	20.48	23.21	25.19
11	19.68	21.92	24.72	26.75
12	21.03	23.34	26.21	28.30
13	22.36	24.74	27.69	29.82
14	23.69	26.12	29.14	31.31
15	25.00	27.49	30.58	32.80
16	26.30	28.85	32.00	34.27
17	27.59	30.19	33.41	35.72
18	28.87	31.53	34.81	37.15
19	30.14	32.85	36.19	38.58
20	31.41	34.17	37.56	40.00
21	32.67	35.48	38.93	41.40
22	33.93	36.78	40.29	42.80
23	35.17	38.08	41.64	44.18
24	36.42	39.37	42.98	45.56
25	37.65	40.65	44.32	46.93
26	38.89	41.92	45.64	48.29
27	40.11	43.20	46.96	49.64
28	41.34	44.46	48.28	50.99
29	42.56	45.72	49.59	52.34
30	43.77	46.98	50.89	53.67
40	55.75	59.34	63.71	66.80
50	67.50	71.42	76.17	79.52
100	124.34	129.56	135.82	140.19

Anexo IV

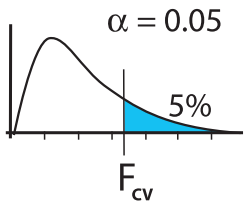


Tabela de valores críticos – F de Snedecor

Degrees of Freedom for the F-Ratio numerator

Degrees of Freedom for the F-Ratio denominator	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1
2	18.51	19.00	19.16	19.25	19.30	19.36	19.35	19.37	19.38	19.40
3	10.13	9.55	9.328	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85
1000	3.85	3.01	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84

Anexo V

Tabela de valores críticos – Mann Whitney																				
1- tail test at $\alpha = 0.025$ or 2- tail test at $\alpha = 0.05$																				
	N_1																			
N_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	21	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	95	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Anexo V – Continuação

1- tail test at $\alpha = 0.05$ or 2- tail test at $\alpha = 0.10$																				
	N_1																			
N_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2					0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3			0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4			0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5		0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6		0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7		0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
8		1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
9		1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
10		1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
11		1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
12		2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
13		2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
14		2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
15		3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
16		3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
17		3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
18		4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

$N_1 < N_2$

Anexo VI

Tabela de valores críticos – Lilliefors		
n	**α= 0,05**	**α=0,01**
5	0,337	0,405
10	0,258	0,294
15	0,220	0,257
20	0,190	0,231
25	0,173	0,200
30	0,161	0,187
>30	$0,886/\sqrt{n}$	$1,031/\sqrt{n}$

Anexo VII

Tabela de valores críticos – Wilcoxon								
Number of pairs N	T	.05 α	T	.025 α	T	.01 α	T	.005 α
5	0	.0313						
	1	.0625						
6	2	.0469	0	.0156				
	3	.0781	1	.0313				
7	3	.0391	2	.0234	0	.0078		
	4	.0547	3	.0391	1	.0156		
8	5	.0391	3	.0195	1	.0078	0	.0039
	6	.0547	4	.0273	2	.0117	1	.0078
9	8	.0488	5	.0195	3	.0098	1	.0039
	9	.0645	6	.0273	4	.0137	2	.0059
10	10	.0420	8	.0244	5	.0098	3	.0049
	11	.0527	9	.0322	6	.0137	4	.0068
11	13	.0415	10	.0210	7	.0093	5	.0049
	14	.0508	11	.0269	8	.0122	6	.0068
12	17	.0461	13	.0212	9	.0081	7	.0046
	18	.0549	14	.0261	10	.0105	8	.0061
13	21	.0471	17	.0239	12	.0085	9	.0040
	22	.0549	18	.0287	13	.0107	10	.0052
14	25	.0453	21	.0247	15	.0083	12	.0043
	26	.0520	22	.0290	16	.0101	13	.0054
15	30	.0473	25	.0240	19	.0090	15	.0042
	31	.0535	26	.0277	20	.0108	16	.0051
16	35	.0467	29	.0222	23	.0091	19	.0046
	36	.0523	30	.0253	24	.0107	20	.0055
17	41	.0492	34	.0224	27	.0087	23	.0047
	42	.0544	35	.0253	28	.0101	24	.0055
18	47	.0494	40	.0241	32	.0091	27	.0045
	48	.0542	41	.0269	33	.0104	28	.0052
19	53	.0478	46	.0247	37	.0090	32	.0047
	54	.0521	47	.0273	38	.0102	33	.0054
20	60	.0487	52	.0242	43	.0096	37	.0047
	61	.0527	53	.0266	44	.0107	38	.0053

Anexo VIII

Tabela de valores críticos – Kruskal Wallis														
n1	n2	n3	H	P	n1	n2	n3	H	P	n1	n2	n3	H	P
2	1	1	2,7000	0,500	4	4	1	6,6667	0,010	5	4	1	6,9545	0,008
2	2	1	3,6000	0,200				6,1667	0,022				6,8400	0,011
2	2	2	4,5714	0,067				4,9667	0,048				4,9855	0,044
			3,7143	0,200				4,8667	0,054				4,8600	0,056
3	1	1	3,2000	0,300				4,1667	0,082				3,9873	0,098
3	2	1	4,2857	0,100				4,0667	0,102				3,9600	0,102
			3,8571	0,133	4	4	2	7,0364	0,006	5	4	2	7,2045	0,009
3	2	2	5,3572	0,029				6,8727	0,011				7,1182	0,010
			4,7143	0,148				5,4545	0,046				5,2727	0,049
			4,5000	0,067				5,2364	0,052				5,2682	0,050
			4,4643	0,105				4,5545	0,098				4,5409	0,098
3	3	1	5,1429	0,043				4,4455	0,103				4,5182	0,101
			4,5714	0,100	4	4	3	7,1439	0,010	5	4	3	7,4449	0,010
			4,0000	0,129				7,1364	0,011				7,3949	0,011
3	3	2	6,2500	0,011				5,5985	0,049				5,6564	0,049
			5,3611	0,032				5,5758	0,051				5,6308	0,050
			5,1389	0,061				4,5455	0,099				4,5487	0,099
			4,5556	0,100				4,4773	0,102				4,5231	0,103
			4,2500	0,012	4	4	4	7,6538	0,008	5	4	4	7,7604	0,009
3	3	3	7,2000	0,004				7,5385	0,011				7,7440	0,011
			6,4889	0,011				5,6923	0,049				5,6571	0,049
			5,6889	0,029				5,6538	0,054				5,6176	0,050
			5,6000	0,050				4,6539	0,097				4,6187	0,100
			5,0667	0,086				4,5001	0,104				4,5527	0,102
			4,6222	0,100	5	1	1	3,8571	0,143	5	5	1	7,3091	0,009
4	1	1	3,5714	0,200	5	2	1	5,2500	0,036				6,8364	0,011
4	2	1	4,8214	0,057				5,0000	0,048				5,1273	0,046
			4,5000	0,076				4,4500	0,071				4,9091	0,053
			4,0179	0,114				4,2000	0,095				4,1091	0,086
4	2	2	6,0000	0,014				4,0500	0,119				4,0364	0,105
			5,3333	0,033	5	2	2	6,5333	0,008	5	5	2	7,3385	0,010
			5,1250	0,052				6,1333	0,013				7,2692	0,010
			4,4583	0,100				5,1600	0,034				5,3385	0,047
			4,1667	0,105				5,0400	0,056				5,2462	0,051
4	3	1	5,8333	0,021				4,3733	0,090				4,6231	0,970
			5,2083	0,050				4,2933	0,122				4,5077	0,100
			5,0000	0,057	5	3	1	6,4000	0,012	5	5	3	7,5780	0,010
			4,0556	0,093				4,9600	0,048				7,5429	0,010
			3,8889	0,129				4,8711	0,052				5,7055	0,046
4	3	2	6,4444	0,008				4,0178	0,095				5,6264	0,510
			6,3000	0,011				3,8400	0,123				4,5451	0,100
			5,4444	0,046	5	3	2	6,9091	0,009				4,5363	0,102
			5,4000	0,051				6,8218	0,010	5	5	4	7,8229	0,100
			4,5111	0,098				5,2509	0,049				7,7914	0,010
			4,4444	0,102				5,1055	0,052				5,6657	0,049
								4,6509	0,091				5,6429	0,050
								4,4945	0,101				4,5229	0,099
					5	3	3	7,0788	0,009				4,5200	0,101
								6,9818	0,011	5	5	5	8,0000	0,009
								5,6485	0,049				7,9800	0,010
													5,7800	0,049

Referências

BUSSAB, W. O.; MORETIN, P. A. **Estatística Básica**. 4. ed. São Paulo: Saraiva, 2003.

BARROS, Emilio. **Aplicações e Simulações Monte Carlo e *Bootstrap***. Monografia (Bacharelado em Estatística) – Universidade Estadual de Maringá, Maringá, 2005. Disponível em: <http://www.des.uem.br/graduacao/Monografias/Monografia_Emilio.pdf.>. Acesso em: 23 nov. 2007.

CAMPOS, G. M. **Estatística Prática para Docentes e Pós-Graduados**. Disponível em: <http://www.forp.usp.br/restauradora/gmc/gmc_livro/gmc_livro_cap14.html>. Acesso em: 23 nov. 2007.

COSTA NETO, P. L. de O. **Estatística Básica**. 2. ed. São Paulo: Edgard Blücher, 2002.

GONÇALVES, Lóren Pinto Ferreira. **Avaliação de Ferramentas de Mineração de Dados como Fonte de Dados Relevantes para a Tomada de Decisão**: aplicação na Rede Unidão de Supermercados. Dissertação (Mestrado Interinstitucional em Administração) – Universidade da Região da Campanha (Urcamp), São Leopoldo, 2001. Disponível em: <http://volpi.ea.ufrgs.br/teses_e_dissertacoes/td/000410.pdf>

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. **Análise Exploratória de Dados** – Técnicas Robustas. Lisboa: Edições Salamandra, 1983.

HOEL, PORT & STONE. **Introdução à Teoria da Probabilidade**. Rio de Janeiro: Editora Interciência ,1981.

KAZMIER, L. J. **Estatística Aplicada à Economia e Administração**. 4. ed. São Paulo: Bookman 2007.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. et al. **Estatística**: Teoria e Aplicações – Usando Microsoft Excel. 3. ed. Rio de Janeiro: LTC, 2005.

MATTAR, F. N. **Pesquisa de Marketing**. São Paulo: Atlas, 2001.

_____. São Paulo: Atlas, 1996. (Edição compacta).

MEYER, P. L. **Probabilidade**: Aplicações à Estatística. 2. ed. Rio de Janeiro: LTC, 2000.

SIEGEL, S.; CASTELLAN JR., N. J. **Estatística Não-Paramétrica para Ciências do Comportamento**. Porto Alegre: Artmed, 2006.

TRIOLA, M. F. **Introdução à Estatística**. 9. ed. Rio de Janeiro: LTC, 2005.

VIEIRA, S., WADA, R. **O que é Estatística?** 3. ed. São Paulo: Brasiliense, 1991.

WONNACOT, T. H. WONNACOTT, R. J. **Estatística Aplicada à Economia e à Administração**. Rio de Janeiro: LTC, 1981.