Notas de Aula de

Estatística Aplicada à Engenharia

Monica Aparecida Tomé Pereira
Paulo José Pereira

O material da apostila foi retirado da seguinte bibliografia:

[1] A. C. Pedroso de Lima; M. Magalhães. **Noções de Probabilidade e Estatística**. 3 ed. IME - USP.

[2] ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. **Estatística Aplicada à Administração e Economia**. Tradução de Luiz Sérgio de Castro Paiva. São Paulo: Pioneira Thomson Learning, 2002.

[3] BARBIN, D. **Planejamento e Análise Estatística de Experimentos Agronômicos**. Arapongas: MIDAS, 2003. 208p. ISBN 85.89687-01-5.

[4] BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. São Paulo: Saraiva, 2003. 526p. ISBN 85-02-03497-9.

[5] DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 3 ed. Jonn Wiley & Sons (Wiley series in probability and statistics), 1998. 706p. ISBN 9812-53-034-7.

[6] FARIAS, A . A ., SOARES, J. F., CÉSAR, C. C. **Introdução à Estatística**. Segunda edição. Rio de Janeiro: LTC. 2003

[7] FURTADO, F., DANIEL. **Estatística Básica**. Editora UFLA. 2005. Hines, W.W. et. al. Probabilidade e Estatística na Engenharia 4ª edição. LTC, 2006.588P.

[8] HOFFMANN, RODOLFO. **Estatística para Economistas**. Terceira edição. São Paulo: Pioneira, Thomson Learning, 1998

[9] LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. **Estatística: teoria e aplicações usando o Excel.** Rio de Janeiro: LTC, 2005.

[10] MARTINS, G. A. **Estatística Geral e Aplicada**. São Paulo: Atlas, 2001

[11] MEYER, P. L. **Probabilidade: aplicações à estatística.** 2ª. ed. Rio de Janeiro: Livros Técnicos e Científicos, 1983. 426p.

[12] MONTGOMERY, D. C., RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros.** Rio de Janeiro: LTC. 2ª edição, 2003. ISBN 85-216-1360-1.

[13] MOORE, D. S.; McCABE, G. P. **Introdução à Prática da Estatística**. Rio de Janeiro:LTC. 3ª Edição, 2002. ISBN 85-216-1324-5.

[14] MORETTIN, L. G. , Estatística Básica: Probabilidade. São Paulo: McGraw- Hill, 1999.

[15] MORETTIN, L. G. , Estatística Básica – Volume 2: Inferência. São Paulo: McGraw-Hill, 2000.

[16] NAVIDI, W. **Probabilidade e Estatística para Ciências Exatas**. AMGH Editora, 2012. 604p.

[17] PIMENTEL-GOMES, F.; GARCIA, C. H. **Estatística Aplicada a Experimentos Agronômicos e Florestais**. Piracicaba: FEALQ, 2002. 309p. ISBN 85-7133-014-X.

[18] RODRIGUES, P. C. **Bioestatística**. 3ª Edição. Niterói. EDUFF. 2002. 339p.

[19] SILVA, H. M. da et al. **Estatística**. Vol. 2. São Paulo: Atlas, 1999.

[20] SPIEGEL, M. R. SCHILLER, J.; SRINIVASAN, R. A. **Teoria e Problemas de Probabilidade e Estatística**. 2 Ed. Tradução: Sandra Ianda Correa Carmona. Porto Alegre Bookman (Coleção Schaum), 2004. 398p. ISBN 85-363-0297-6.

[21] STEVENSON, W. J. **Estatística Aplicada à Administração**. São Paulo: Harbra, 1986.

[22] STEVENSON, W. J. **Estatística Aplicada à Administração**. São Paulo: Harbra, 2001.

[23] TOLEDO e OVALE. **Estatística Básica**. Editora ATLAS.

[24] TRIOLA, M. F. **Introdução à Estatística**. Nona edição. Tradução Alfredo Alves Farias. Rio de Janeiro: LTC. 2005.

[25] VIEIRA, S. **Princípios de estatística**. São Paulo: Pioneira, 1999.

[26] WALPOLE, R. E. & et. al. **Probabilidade e estatística para engenharia e ciências**. 8ª ed. Pearson Prentice Hall, 2009. 491 p.

Sumário

1 Introdução

1.1 O que é Estatística?

Estatística é uma ciência definida como o conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimento, realizados em qualquer área do conhecimento. Entende-se por dados um (ou mais) conjunto de valores, numéricos ou não. A grosso modo pode-se dividir a Estatística em três grandes áreas: Estatística Descritiva, Probabilidade e Inferência Estatística.

Estatística Descritiva

A Estatística Descritiva é, em geral, utilizada na etapa inicial da análise, quando tomamos contato com os dados pela primeira vez. Pode ser definida como um conjunto de técnicas destinadas a descrever e resumir os dados, a fim de que possamos tirar informações e conclusões a respeito de características de interesse.

Exemplos:

- **Perfil do Consumidor.** A informação obtida pelos prestadores de serviços é guardada em grandes bancos de dado, usados pra a construção de perfis de consumidores. Esses perfis são usados, por exemplo, para identificar roubos de cartões de crédito para a criação de listas de clientes de potenciais. Métodos computacionais intensivos são usados para a construção desses perfis.

- **Anuário Estatístico Brasileiro.** O IBGE publica esse anuário a cada ano, apresentando em várias tabelas os mais diversos dados sobre o Brasil: educação, saúde, transporte, economia, cultura, etc. Embora simples e fáceis de serem entendidas, as tabelas são fruto de um processo demorado e extremamente dispendioso de coleta e apuração de dados.

Probabilidade

A Probabilidade é a base matemática sob a qual a Estatística é construída. Fornece métodos para quantificar a incerteza existente em determinada situação, usando ora um número ora uma função matemática. Foi desenvolvida a partir de problemas apresentados por jogadores, nobres franceses, a grandes matemáticos da época, como Blaise Pascal.

Exemplo:

- **O cálculo do prêmio do seguro.** O cálculo das probabilidades é fundamental para se conviver de forma inteligente com o risco, inerente a tantos processos sociais. No caso de seguros, isso é evidente. Uma companhia de seguros deve saber calcular o valor a se cobrar para segurar, por exemplo, a saúde de um indivíduo. Se seu valor é alto demais, ela não terá clientes; se é baixo demais, pode não ter recursos para honrar seus compromissos. Esses cálculos são feitos por profissionais chamados atuários.

Inferência Estatística

É o estudo de técnicas que possibilitam a extrapolação, a um grande conjunto de dados, das informações e conclusões obtidas a partir de subconjuntos de valores, usualmente de dimensão muito menor. Deve ser notado que se tivermos acesso a todos os elementos que desejamos estudar, não é necessário o uso das técnicas de inferência estatística; entretanto, elas são indispensáveis quando existe a impossibilidade de acesso a todo o conjunto de dados, por razões de natureza econômica, ética ou física.

Exemplos:

- **Comparação: Testes sobre medicamentos.** Um experimento médico testa um novo analgésico para ver se ele é melhor que o produto padrão correspondente. Dez pessoas selecionadas aleatoriamente tomam o novo medicamento, e as dez outras tomam o remédio padrão. O experimento é do tipo "duplo cego", isto é nem o paciente e nem o médico sabem qual dos dois remédios está sendo administrados. Essa informação é do conhecimento apenas do estatístico e de outros que vão analisar os dados. Suponhamos que os resultados tenham sido os seguintes:

Remédio	**Quantos relataram diminuição da dor**
Novo	8
Padrão	5

  É correto concluir que o novo remédio é melhor só porque mais pessoas , no grupo das que o tomaram, relataram diminuição da dor? Trata-se de um problema estatístico da maior relevância. É preciso saber se a margem de 8 para 5 é real ou se os dois remédios são igualmente eficientes, tendo a diferença ocorrido apenas por uma variação aleatória.

- **Previsão: Demanda por produtos e serviços.** Os planejadores tanto do Estado quanto do setor privado precisam estimar a demanda por serviços . Quantos leitos hospitalares serão necessários? Quantas vagas nas diferentes séries escolares? Quantos médicos cardiologistas devem prestar serviços em uma comunidade? Quanto um supermercado venderá nas festas de fim de ano? Para isso, a Estatística desenvolveu métodos de previsão, peças fundamentais na solução desses problemas. Um exemplo particularmente interessante é usado para o monitoramento de custos de energia em grandes consumidores. O preço pago pela energia à companhia de eletricidade pelos grandes consumidores aumenta bastante se o consumo ultrapassa certa cota acertada entre as duas partes. É de interesse do consumidor evitar essa ocorrência. Utilizando um método estatístico de previsão, cria-se um sistema que permite às organizações conhecerem com antecedência o momento em que deve desligar algum equipamento não-essencial, ou adiar algum processo, reduzindo o consumo e, conseqüentemente, os custos.

- **Explicação de resultados.** A Estatística, por intermédio de seus modelos, permite o conhecimento de fatores determinantes de vários eventos. Assim, na pesquisa clínica, a Estatística ajuda na determinação de fatores de prognóstico para pacientes e de fatores de risco para doenças. Na avaliação de sistemas educacionais, a Estatística permite, através de análise dos dados, encontrar os determinantes de desempenho escolar. São usados os chamados modelos de regressão. É importante frisar que os métodos estatísticos não determinam as causas de um fenômeno.

1.2 Por que estudar Estatística?

Devido a muitos aspectos da prática de engenharia envolverem o trabalho com dados, obviamente algum conhecimento de estatística é importante para qualquer engenheiro. Especificamente, técnicas estatísticas podem ser uma ajuda poderosa no planejamento de novos produtos e sistemas, melhorando projetos existentes e planejando, desenvolvendo e melhorando os processos de produção. A importância de se estudar Estatística pode ser encontrada, de maneira geral, em quatro situações apresentadas a seguir:

i) Possibilidade de crescimento profissional;

ii) Tomar decisões corretas e evitar ser iludido por certas apresentações viciosas;

iii) Cursos posteriores utilizam a análise estatística;

iv) Revistas profissionais possuem freqüentemente referências a estudos estatísticos.

1.3 Algumas aplicações da Estatística

- Uma firma que está se preparando para lançar um novo produto precisa conhecer as preferências dos consumidores no mercado de interesse. Para isso, pode fazer uma pesquisa de mercado entrevistando um número de residências escolhidas aleatoriamente. Poderá, então, usar os resultados para estimar as preferências de toda a população;

- Um auditor deve verificar os livros de uma firma para se certificar de que os lançamentos refletem efetivamente a situação financeira da companhia. O auditor deve examinar pilhas de documentos originais, como notas de venda, ordens de compra e requisições. Seria um trabalho incalculável consultar todos os documentos originais; em vez disso o auditor pode verificar uma amostra de documentos escolhidos aleatoriamente e, com base nessa amostra, fazer inferências sobre toda a população;

- Se estivermos recebendo um grande embarque de mercadorias de um fornecedor, teremos de certificar-nos de que o produto realmente satisfaz os requisitos de qualidade acordados. Seria por demais dispendioso fazer uma verificação de cada item; mas aqui, mais uma vez, as técnicas estatísticas vêm em nosso auxílio, permitindo-nos fazer inferências sobre a qualidade de todo o lote mediante inspeção de uma amostra de itens escolhidos aleatoriamente.

1.4 Definições Importantes em Estatística

Definição 1.4.1 **População:** É o grande conjunto de dados que contém a característica que temos interesse em determinado estudo. O tamanho da população é representado por N.

Definição 1.4.2 **Amostra:** É uma parte da população, ou seja, um subconjunto da população. Em geral tem dimensão sensivelmente menor que a população. O tamanho da amostra é representado por n.

Definição 1.4.3 **Variável:** É uma característica de interesse do estudo. Existem dois tipos de variáveis:

i) **Qualitativas:** variável com dados que fornecem rótulos , qualidades, nomes, para uma característica em estudo. Elas podem ser:

   1. Ordinais: variáveis que têm uma ordenação natural, indicando intensidades crescentes de realização;
   2. Nominais: variáveis em que não é possível estabelecer uma ordem natural entre seus valores.

   Então, variáveis tais como Turma (A ou B), Sexo (feminino ou masculino) são variáveis qualitativas nominais. Por outro lado, variáveis como Tamanho (pequeno, médio ou grande), Classe Social (baixa, média ou alta) são variáveis qualitativas ordinais.

ii) **Quantitativas:** Uma variável com dados que indicam a quantidade de alguma coisa. É sempre numérica. Elas podem ser:

   1. Discretas: variáveis resultantes de contagens, assumindo assim, em geral, valores inteiros;
   2. Contínuas: variáveis que assumem valores em intervalos dos números reais e, geralmente, são provenientes de uma mensuração.

   Por exemplo, Número de Irmãos (0, 1, 2, ...) e Número de Defeitos (0, 1, 2,...) são quantitativas discretas, enquanto Peso e Altura são quantitativas contínuas.

Definição 1.4.4 **Dados brutos:** é uma seqüência de valores numéricos não organizados, obtidos diretamente da observação de um fenômeno coletivo.

Definição 1.4.5 **Rol:** é uma seqüência ordenada dos dados brutos.

2 Medidas numéricas

A partir de agora suponha que os dados observados na amostra são x_1, x_2, ..., x_n. Note que n é o tamanho da amostra. A partir dos x's vamos encontrar números que resumem as características da amostra. Vamos estar interessados em 2 tipos principais de medidas numéricas: as que caracterizam a localização do centro da amostra e as que caracterizam a dispersão dos dados.

2.1 Medidas de Tendência Central

Média Aritmética

A média aritmética é uma medida que indica onde está o "centro"da sua amostra ou da população. A média amostral é representada por $\overline{x}$ e a média populacional é representada por μ. As fórmulas de cálculo são apresentadas a seguir:

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

Considere agora a sua amostra x_1, x_2, ..., x_n e suponha que você ordene a amostra, de tal forma que x_1 é o menor elemento da amostra, x_2 é o segundo menor elemento, ..., x_n é o maior elemento da amostra. Os valores x_1, x_2,..., x_n são chamados de estatísticas de ordem da amostra. Outras medidas de tendência central e de dispersão serão definidas a partir das estatísticas de ordem.

Propriedades da Média:

i) A soma algébrica dos desvios em relação à média aritmética é nula:

$$\sum_{i}^{n} (x_i - \overline{x}) = 0$$

ii) A soma dos quadrados dos desvios de um conjunto de dados em relação a sua média e um valor mínimo.

$$D = \sum_{i}^{n} (x_i - \overline{x})^2$$

representa um valor mínimo.

Demonstração: Fazendo

$$D = \sum_{i}^{n} (x_i - A)^2$$

Expandindo o somatório e derivando D em relação a A tem-se:

$$\begin{aligned} D = \sum_{i}^{n} (x_i - A)^2 &= \sum_{i}^{n} (x_i^2 - 2Ax_i + A^2) \\ &= \sum_{i}^{n} x_i^2 - \sum_{i}^{n} 2Ax_i + \sum_{i}^{n} A^2 \\ \frac{\partial D}{\partial A} &= -2 \sum_{i=1}^{n} x_i + 2nA \end{aligned}$$

Igualando a derivada a zero, tem-se:

$$\begin{aligned} -2 \sum_{i=1}^{n} x_i + 2nA &= 0 \\ 2nA &= 2 \sum_{i=1}^{n} x_i \\ A &= \frac{\sum_{i}^{n} x_i}{n} = \overline{x} \end{aligned}$$

Portanto, o ponto ótimo obtido igualando a primeira derivada a zero, pode ser um ponto de máximo ou de mínimo. Para certificar que o valor de D, quando A é igual à média amostral, é um valor mínimo basta mostrar que a segunda derivada é positiva. A segunda derivada de D em relação a A é dada por:

$$\frac{\partial^2 D}{\partial A^2} = 2n > 0$$

Verifica-se que para qualquer tamanho de amostra o valor $2n$ será positivo.

iii) A média de um conjunto de dados acrescido (ou subtraído) em cada elemento por uma constante é igual a média original mais (ou menos) essa constante.

$$\overline{x}' = \overline{x} \pm k$$

Em que é a média do novo conjunto de dados.

iv) Multiplicando todos os dados por uma constante a nova média será igual ao produto da média anterior pela constante.

$$\overline{x}' = k \cdot \overline{x}$$

v) A média é influenciada por valores extremos

Mediana

A mediana amostral é definida a partir das estatísticas de ordem como:

$m = X_{\frac{n+1}{2}}$ sendo n o tamanho da amostra ímpar ou,

$m = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$, sendo n par.

Por exemplo, se existem 10 observações na amostra, a mediana equivale à média entre x_5, x_6. Se a amostra contém 11 elementos, a mediana é x_6.

Analogamente ao caso da média, também podemos definir uma mediana para a população.

A mediana amostral tem uma vantagem sobre a média amostral : ela é menos influenciada por observações extremas do que a média amostral.

Por exemplo, suponha os seguintes dados para uma amostra: 1, 3, 4, 2, 7, 6, 8.

A média amostral é $4,43$, e a mediana é 4.

Agora se os dados forem: 1, 3, 4, 2, 7, 2519, 8.

A média amostral é $363,43$, mas a mediana continua sendo 4.

É claro que este exemplo é radical, mas ilustra bem o fato da mediana ser mais "robusta"ao encontrar observações discrepantes do resto da amostra.

Moda

A moda amostral é simplesmente a observação mais freqüente na amostra. Se os dados são: 1, 4, 8, 12, 5, 4, 4, 7, a moda é 4, o valor que ocorreu mais vezes. Também é possível definir a moda de uma população.

Observação 2.1.1 A moda, ao contrário das outras medidas de tendência central, pode ser obtida mesmo que a variável seja qualitativa.

Mediana *versus* Média

De modo geral, o uso da mediana é indicado quando :

i) Os valores para a variável em estudo têm distribuição de freqüências assimétrica (verificada através das ferramentas gráficas);

ii) O conjunto de dados possui algumas poucas observações extremas (valores muito mais altos ou muito mais baixos que os outros);

iii) Não conhecemos exatamente o valor de algum elemento, mas temos alguma informação sobre a ordem que ele ocupa no conjunto de dados. Por exemplo, no caso de salários, se alguém não quisesse informar o quanto recebe, mas apenas dissesse que ganha mais (ou menos) do que um certo valor, de modo que conseguíssemos determinar uma ordem para essa pessoa, poderíamos calcular a mediana, mas não a média.

Medidas Separatrizes

São números reais que dividem a seqüência ordenada de dados em partes que contêm a mesma quantidade de elementos da série.

Desta forma, a mediana que divide a seqüência ordenada em dois grupos, cada um deles contendo 50% dos valores da seqüência, é também uma medida separatriz.

Além da mediana, as outras medidas separatrizes que destacaremos são os percentis e os quartis.

Percentil e Quartil

O percentil de ordem k (onde k é qualquer valor entre 0 e 100), denotado por P_k, é o valor tal que $K\%$ dos valores do conjunto de dados são menores ou iguais a ele. Assim, o percentil de ordem 10, o P_{10}, é o valor da variável tal que 10% dos valores são menores ou iguais a ele; o percentil de ordem 65 deixa 65% dos dados menores ou iguais a ele, etc.

Os percentis de ordem 10, 20, 30, ... 90 dividem o conjunto de dados em dez partes com mesmo número de observações e são chamados de *decis*.

Os percentis de ordem 25, 50 e 75 dividem o conjunto de dados em quatro partes com o mesmo número de observações. Assim, estes três percentis recebem o nome de **quartis – primeiro quartil (Q1), segundo quartil (Q2) e terceiro quartil (Q3)**, respectivamente. O segundo quartil é a já conhecida mediana.

Existem vários processos para calcular os percentis. Vamos ficar com um método mais simples encontrado em ANDERSON, SWEENEY & WILLIAMS (2002). As diferenças serão muito pequenas e desaparecerão à medida que aumenta o número de dados.

O seguinte procedimento pode ser usado para calcular o p-ésimo percentil:

1º Arranje os dados na ordem ascendente (ordem de classificação do menor valor para o maior).

2º Calcule um índice i:

$$i = \left(\frac{k}{100}\right) n$$

onde p é o percentil de interesse e n é o número de observações.

3º Se i não for um no inteiro, arredonde para cima. O próximo inteiro maior que i denota a posição do p-ésimo percentil.

Se i é um inteiro, o p-ésimo percentil é a média dos valores de dados nas posições i e $i+1$.

2.2 Medidas de Variabilidade

As medidas de tendência central não são as únicas medidas necessárias para caracterizar uma amostra (ou população). Métodos estatísticos também são usados para nos ajudar a entender a **variabilidade**. Por variabilidade, entende-se que sucessivas observações de um sistema ou fenômeno não produzem o mesmo resultado, então precisamos também saber o quanto as observações na amostra estão "espalhadas". As medidas de dispersão ou de variabilidade são medidas que informam sobre a dispersão dos dados e são necessárias para, junto com a média, representar bem um conjunto de observações. A seguir estudaremos algumas medidas de variabilidade que são importantes.

Desvio Médio

Considerando que num conjunto de dados cada valor apresenta em relação à média aritmética um afastamento, o desvio médio será a média aritmética destes afastamentos, levando-se em conta

os valores absolutos desses desvios.

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

Apesar de seu aspecto atrativo, essa medida, devido ao fato de usar valores absolutos, conduz a dificuldades teóricas em problemas de inferência estatística por isso dificilmente é usada.

Variância

É a medida mais comum de dispersão. A variância amostral, denotada por s^2 é definida como:

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

onde $\overline{x}$ é a média amostral, já definida.
A variância populacional será denotada por σ^2. Então, temos que:

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

Observação 2.2.1 A variância (da amostra ou da população) é sempre maior ou igual a zero.

A unidade de medida da variância é o quadrado da unidade de medida das observações. Assim, se os dados estão em metros, à variância é expressa em metros quadrados. Isso dificulta a interpretação da variância amostral. Para evitar isso trabalhamos com o desvio padrão, definido a seguir.

Desvio Padrão

O desvio padrão amostral, denotado por s, é definido como a raiz quadrada positiva da variância amostral. Pelos comentários acima concluímos que s é sempre expresso nas mesmas unidades de medida que as observações na amostra. O desvio padrão da população é definido como a raiz quadrada da variância da população, e denotado por σ. Logo,

No caso de uma amostra:	No caso de uma população:
$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$

Na interpretação do desvio padrão, lembre-se de que ele é a medida de quanto uma entrada típica se desvia da média. Quanto mais espalhadas estiverem as entradas, maior será o desvio padrão.

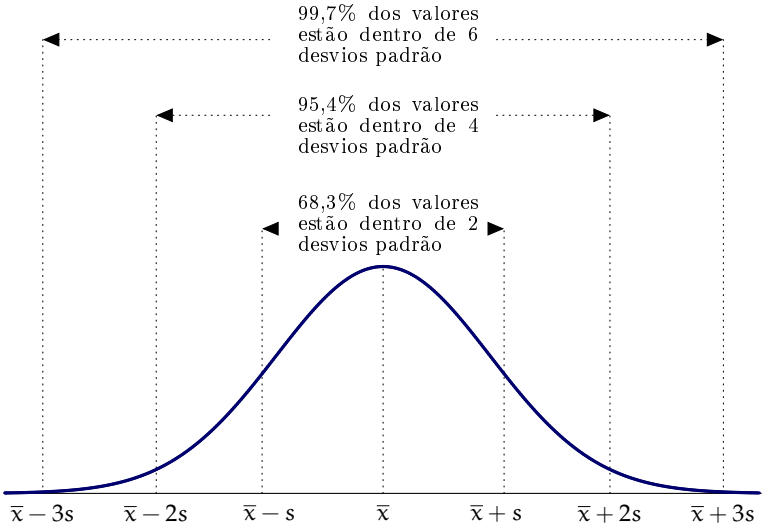
Interpretação do Desvio Padrão

Regra Empírica: Para distribuição amostral em forma de sino (simétrica) com média $\bar{x}$ e desvio padrão s, tem-se:

i) O intervalo $\bar{x} \pm s$ contém aproximadamente 68% de todas as observações amostrais;

ii) O intervalo $\bar{x} \pm 2s$ contém aproximadamente 95% das observações amostrais pra distribuições simétricas;

iii) O intervalo $\bar{x} \pm 3s$ contém aproximadamente 99,7% das observações amostrais pra distribuições simétricas



Coeficiente de Variação

O coeficiente de variação amostral é definido como:

$$CV = \frac{s}{\bar{x}} \cdot 100$$

onde s é o desvio padrão amostral e $\bar{x}$ é a média amostral. A definição do coeficiente de variação para a população é análoga, substituindo s por σ e $\bar{x}$ por μ.

O coeficiente de variação é uma medida de variabilidade relativa. Refere-se à variabilidade dos dados em relação à média.

Observação 2.2.2 O desvio padrão ou a variância permitem a comparação da variabilidade entre conjuntos numéricos que possuem a mesma média e a mesma unidade de medida ou grandeza. Diz-seque o desvio padrão é uma medida de dispersão absoluta. Nos casos em que os conjuntos possuem diferentes unidades e/ou possuem médias diferentes, uma medida de dispersão relativa, como o coeficiente de variação (CV), é indispensável para se comparar à variabilidade.

Erro Padrão da Média

Quando procedemos a uma investigação científica em que utilizamos dados de uma fração representativa de uma população (amostra), a média aritmética determinada apresentará, em relação à média populacional, um afastamento.

Se outras amostras fossem retiradas da população, apresentariam médias aritméticas que teriam outros afastamentos em relação à média populacional. Para se determinar a média destes afastamnetos utilizamos o erro padrão da média, cuja estimativa pode ser encontrada por:

$$s_x = \frac{s}{\sqrt{n}}$$

Amplitude Total

É a diferença entre a maior observação e a menor observação do conjunto dados:

$$A = maior\ valor - menor\ valor$$

Amplitude Interquartil

Uma medida da variabilidade que supera a dependência dos valores extremos é a *amplitude interquartil (AIQ)*. Essa medida de variabilidade é simplesmente a diferença entre o terceiro quartil, Q_3 e o primeiro quartil Q_1. Em outras palavras, a amplitude interquartil é o intervalo para 50% dos dados do meio.

$$AIQ = Q_3 - Q_1$$

2.3 Outras Medidas Conhecidas

Média Aritmética Ponderada

No cálculo da média aritmética não ponderada todos os valores observados foram somados atribuindo-se o mesmo peso a todas observações. Agora veremos uma nova forma de calcular a média. Consideremos um exemplo familiar de cálculo da média de notas de estudantes, quando o exame final vale duas vezes mais do que as duas provas comuns realizadas no decorrer do semestre. Se um determinado obter as notas 7, 5 e 8 a sua média ponderada final será:

$$\frac{1 \times (7) + 1 \times (5) + 2 \times (8)}{1 + 1 + 2}$$

Em termos gerais, a fórmula para a média aritmética ponderada é:

$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i}$$

onde w_i é o peso da observação i e n é o número de observações.

A soma dos pesos não pode ser igual a zero. Fora disto, não existe restrição para os valores dos pesos. Se todos os pesos forem iguais a 1, a média ponderada recai em seu caso particular, a média aritmética não ponderada. O mesmo ocorre se todos os pesos forem iguais a uma constante c. Portanto, a média aritmética não ponderada na realidade é uma média aritmética ponderada com pesos iguais.

Média Geométrica

A média geométrica de uma amostra é definida como a raiz enésima do produto nos n valores amostrais.

$$G = \sqrt[n]{(x_1)(x_2)\dots(x_n)}$$

Por exemplo, a média geométrica de 5, 9 e 13 é:

$$G = \sqrt[n]{(5)(9)(13)} = 8,36$$

Para a mesma série de dados a média é 9. É sempre verdade que a média aritmética é maior do que a média geométrica para qualquer série de valores positivos, com exceção do caso em que os valores da série são todos iguais, quando as duas médias coincidem.

Média Harmônica

A média harmônica é o inverso da média aritmética dos inversos dos valores observados. Simbolicamente, para uma amostra, temos:

$$H = \frac{\frac{1}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}}{n} = \frac{\frac{1}{\sum\left(\frac{1}{x}\right)}}{n} = \frac{n}{\sum\left(\frac{1}{x}\right)}$$

A média harmônica dos três valores 4, 10 e 16 é:

$$H = \frac{3}{\frac{1}{4} + \frac{1}{10} + \frac{1}{16}} = 7,273$$

$$H = 7,273$$

Para os mesmos dados a média aritmética é 10 e a média geométrica é 8,62. Para qualquer série de dados cujos valores não são todos os mesmos e que não incluem o zero, a média harmônica é sempre menor que tanto a média aritmética como a média geométrica.

3 Organização e Apresentação de Dados

3.1 Distribuição de Frequências

Nesta seção serão apresentados alguns procedimentos que podem ser utilizados para organizar e descrever um conjunto de dados, seja em uma população ou em uma amostra.

A questão inicial é: dado um conjunto de dados como "tratar" os valores numéricos ou não, a fim de se extrair informações a respeito de uma ou mais características de interesse? Basicamente faremos uso da distribuição de freqüência, que pode ser apresentada sob forma gráfica ou tabular.

Suponha, por exemplo, que um questionário foi aplicado aos alunos do primeiro ano de uma faculdade fornecendo informações tais como identificação do aluno, sexo, idade, horas de atividades físicas praticadas por semana e número de vezeas que vai ao cinema por mês. O conjunto de informações disponíveis, após a tabulação do questionário ou pesquisa de campo, é denominado de **tabela de dados brutos** e contém os dados da maneira que foram coletados inicialmente. Um arranjo de dados brutos em ordem crescente ou decrescente de grandeza é chamado de **rol**.

Apesar de conter muita informação, a tabela de dados brutos pode não ser prática para respondermos as questões de interesse. Portanto, à partir da tabela de dados brutos, pode-se construir uma nova tabela com as informações resumidas, para cada variável. Essa tabela é denominada distribuição de freqüência ou tabela de freqüências e, como o nome indica, conterá os valores da variável e suas respectivas contagens, as quais são denominadas **freqüências**, que são elas **freqüência de classe**, **freqüência relativa**, **freqüência acumulada**, **freqüência percentual**.

Portanto uma distribuição de freqüência é um agrupamento de dados em classes, exibindo o número ou percentagem de observações em cada classe.

No caso especifico de variáveis qualitativas, a tabela de freqüências consiste em listar os valores possíveis da variável, numéricos ou não e fazer a contagem na tabela de dados brutos do número de suas ocorrências.

Exemplo 3.1.1 A seqüência abaixo representa a observação dos problemas que levaram 25 pacientes de uma clínica a procurar ajuda psicológica.

FOBIAS	FOBIAS	FOBIAS	AGRESSIVIDADE	DEPRESSÃO
AGRESSIVIDADE	ALCOOLISMO	DEPRESSÃO	PÂNICO	FOBIAS
DEPRESSÃO	ALCOOLISMO	ESQUIZOFRENIA	AGRESSIVIDADE	PÂNICO
PÂNICO	ESQUIZOFRENIA	DEPRESSÃO	ALCOOLISMO	DEPRESSÃO
FOBIAS	ESQUIZOFRENIA	DEPRESSÃO	AGRESSIVIDADE	ALCOOLISMO

As observações distintas são: agressividade, alcoolismo, depressão, esquizofrenia, fobias e pânico.

As freqüências simples respectivas são: 4, 4, 6, 3, 5 e 3.

Portanto a tabela de freqüências para esta variável qualitativa nominal será dada da seguinte maneira:

Tabela 3.1.1: Problemas que levaram 25 pacientes a procurar auxílio psicológico

Problema	f_i	f_r	$f_p(\%)$	f_{ac}
Agressividade	4	0,16	16	5
Alcoolismo	4	0,16	16	8
Depressão	6	0,24	24	14
Esquizofrenia	3	0,12	12	17
Fobias	5	0,20	20	22
Pânico	3	0,12	12	25
Total	25	1,0	100	

Observação 3.1.1 Este tipo de tabela também pode ser utilizada quando temos uma variável quantitativa discreta em que aparecem uma pequena quantidade valores que se repetem várias vezes.

Exemplo 3.1.2 A seqüência abaixo representa a observação do números de acidentes por dia, em uma rodovia, durante 20 dias.

Nº	0	2	0	1	1	0	0	0	3	2	1	0	1	2	0	1	3	2	2	0

Os valores distintos da seqüência são: 0, 1, 2, 3.

As freqüências simples respectivas são: 8, 5, 5, 2.

Portanto a tabela de freqüências para esta variável discreta será dada da seguinte maneira:

Tabela 3.1.2: Número de acidentes em uma rodovia, durante 20 dias

Número de Acidentes	f_i	f_r	$f_p(\%)$	f_{ac}
0	8	0,40	40	8
1	5	0,25	25	13
2	5	0,25	25	18
3	5	0,10	10	20
Total	20	1,00	100	

No caso de variáveis quantitativas contínuas vamos apresentar a seguir uma maneira que nos parece mais indicada para construção de uma tabela de freqüências.

1º Passo: Determinar a amplitude total dos dados (A), ou seja, o maior valor menos o menor o valor do conjunto de dados.

2º Passo: Determinar o nº de classes (k) a usar. Duas regras práticas consistem em:

Para $n < 100$, então $k = \sqrt{n}$;
Para $n > 100$, então $k \cong 1 + 3,32 \log n$.

Caso k dê um valor não inteiro, arredonde sempre para o próximo inteiro.

3º Passo: Dividir a amplitude por k–1, para obter a amplitude de classes (c):

$$c = \frac{A}{k-1}$$

4º Passo: Determinar o limite inferior inicial que será:

$$LI_1 = menor\ valor - \frac{c}{2}$$

5º Passo: Calcular o ponto médio de cada intervalo de classe:

$$P_i = \frac{LS_i + LI_i}{2}$$

sendo que LS_i é o limite superior de classe e LI_i é o limite inferior de classe.

Exemplo 3.1.3 O tempo de utilização de caixas eletrônicos depende de cada usuário e das operações efetuadas. Foram coletadas 16 medidas desse tempo (em minutos).

1,5	1,1	1,2	1,7	0,9	1,3	1,4	1,8
1,4	1,3	1,7	1,6	1,2	1,2	1,0	0,9

1º Passo: $A = 1,8–0,9 = 0,9$

2º Passo: $k = \sqrt{16} = 4$

3º Passo: $c = \frac{0,9}{4-1} = 0,30$

4º Passo: $LI_1 = 0,9 - \frac{0,30}{2} = 0,75$

O **5º passo** está feito direto na tabela de freqüências.

Tabela 3.1.3: Tempo de utilização de caixas eletrônicos

Tempos	f_i	f_r	$f_p(\%)$	f_{ac}	P_i
$0,75 \vdash 1,05$	3	0,1875	18,75	3	0,9
$1,05 \vdash 1,35$	6	0,3750	37,50	9	1,20
$1,35 \vdash 1,65$	4	0,2500	25,00	13	1,50
$1,65 \vdash 1,95$	3	0,1875	18,75	16	1,80
Total	16	1,00	100		

Exemplo 3.1.4 Um rigoroso Centro de Atendimento Psicológico de uma Universidade resolveu fazer um estudo sobre o tempo de atraso para o atendimento dos estagiários em psicologia, para procurar melhorar esse atendimento. Durante uma semana foram coletadas 25 medidas desse tempo (em minutos) em que os pacientes ficaram esperando fora do horário.

11,5	10,2	10,2	11,7	10,9	12,3	15,4	16,0	17,0	16,5	14,0	12,8	14,5
13,4	13,0	11,7	13,6	12,9	15,2	15,0	14,9	16,8	15,7	15,0	13,0	

1º Passo: $A = 17,0–10,2 = 6,9$

2º Passo: $k = \sqrt{25} = 5$

3º Passo: $c = \frac{6,8}{5-1} = 1,7$

4º Passo: $LI_1 = 10,2 - \frac{1,7}{2} = 9,35$

O **5º passo** está feito direto na tabela de freqüências.

Tabela 3.1.4: Tempo de atraso para o atendimento feito por estagiários de um Centro de Atendimento Psicológico

Tempos	f_i	f_r	$f_p(\%)$	f_{ac}	P_i
$9,35 \vdash 11,05$	3	0,12	12,0	3	10,2
$11,05 \vdash 12,75$	4	0,16	16,0	7	11,9
$12,75 \vdash 14,45$	7	0,28	28,0	14	13,6
$14,45 \vdash 16,15$	8	0,32	32,0	22	15,3
$16,15 \vdash 17,85$	3	0,12	12,0	25	17,0
Total	25	1,00	100		

Observação 3.1.2

1. Há situações em que a variável é por natureza discreta mas o conjunto de possíveis valores é muito grande. Por exemplo, supondo que o número de horas que pessoas assistem TV, durante a semana tem valores inteiros entre 0 e 30, então uma tabela representando seus valores e respectivas freqüências seria muito extensa e pouco prática.
2. Na realidade, as classes não precisam necessariamente ter a mesma amplitude como no exemplo acima. Porém, sempre que possível, devemos trabalhar com classes de mesma amplitude. Isto facilita os cálculos posteriores.
3. Note que usamos para representar as classes, intervalos reais semiabertos à direita. Isto significa que o intervalo contém o limite inferior, mas não contém o limite superior, ou seja o intervalo de classe $9,35 \vdash 11,05$ contém os valores reais maiores ou iguais a $9,35$ e menores que $11,05$.

Tabela de Contingência – (Tabela de Dupla Entrada):

Na análise de dados há situações que se precisam representar duas variáveis consideradas qualitativas simultaneamente. As tabelas que comportam as informações destas variáveis são chamadas de tabelas de contingência:

Tabela 3.1.5: Número de pacientes internados no Hospital São Sebastião, por clínica e por convênio, em 2002 – Viçosa (MG)

	Clínicas				
Classificação	Pediátrica	Médica	Obstétrica	Cirúrgica	**Total**
Particulares	14	106	30	82	232
SUS	635	1330	1014	1326	4305
Outros	137	1293	168	822	2420
Total	786	2729	1212	2230	6957

Fonte: Hospital São Sebastião (2003)

3.2 Medidas De Posição e de Variabilidade Para Dados Agrupados

Sempre que possível, as medidas estatísticas devem ser calculadas antes de os dados serem agrupados. Não raro, entretanto, só conhecemos a tabela de distribuição de freqüências, ou seja, os dados estando agrupados. Para calcular as principais medidas descritivas da Estatística têm-se que utilizar as seguintes fórmulas:

i) Medidas de Tendência Central

Média Aritmética:

$$\overline{x} = \frac{\sum_{i=1}^{k} P_i f_i}{n} \text{ ou } \mu = \frac{\sum_{i=1}^{k} P_i f_i}{N}$$

Mediana:

$$M_d = LI_{M_d} + \left(\frac{\frac{n}{2} - F_A}{f_{M_d}} \right) \cdot c_{M_d}$$

sendo que:
LI_{M_d} é o limite inferior da classe mediana;
f_{M_d} é a freqüência simples da classe mediana;
F_A é a freqüência acumulada das classes anteriores à classe mediana;
c_{M_d} é a amplitude da classe mediana.

Moda:

$$M_0 = LI_{M_0} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \cdot c_{M_0}$$

sendo que:
LI_{M_0} é o limite inferior da classe modal;
Classe modal é aquela que possui maior freqüência simples.
Δ_1 é a diferença entre as freqüências simples da classe modal e a classe anterior;

Δ_2 é a diferença entre as freqüências simples da classe modal e a classe posterior;
c_{M_0} é a amplitude da classe modal.

ii) Medidas Separatrizes

Percentis:

$$P_m = LI_{P_m} + \left(\frac{m \cdot \frac{n}{100} - F_A}{f_{P_m}} \right) \cdot c_{P_m}$$

sendo que:
m é o número de ordem do percentil
LI_{P_m} é o limite inferior da classe do percentil de ordem m;
f_{P_m} é a freqüência simples da classe do percentil de ordem m;
F_A é a freqüência acumulada das classes anteriores à classe do percentil de ordem m;
c_{P_m} é a amplitude da classe do percentil de ordem m.
A classe do percentil de ordem m é aquela que contém o valor situado de tal modo que apenas (100–m) dos dados são maiores que ele.

iii) Medidas de Variabilidade

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^{k} P_i^2 f_i - \frac{\left(\sum_{i=1}^{k} P_i f_i \right)^2}{n} \right]$$ é a variância amostral.

$$\sigma^2 = \frac{1}{N} \left[\sum_{i=1}^{k} P_i^2 f_i - \frac{\left(\sum_{i=1}^{k} P_i f_i \right)^2}{N} \right]$$ é a variância populacional.

O cálculo do desvio padrão continua sendo mesmo, ou seja, a raiz quadrada da variância.

Exemplo 3.2.1 Uma empresa estabelece o salário de seus vendedores com base na produtividade. Uma amostra de salários mensais dos vendedores desta empresa revelou a seguinte tabela de freqüências:

Salários (R$)	f_i	f_r	$f_p(\%)$	f_{ac}	P_i
$550 \vdash 610$	3	0,12	12,0	3	580
$610 \vdash 670$	3	0,12	12,0	6	640
$670 \vdash 730$	4	0,16	16,0	10	700
$730 \vdash 790$	9	0,36	36,0	19	760
$790 \vdash 850$	6	0,24	24,0	25	820
Total	25	1,00	100		

a) Calcule a média, mediana e a moda para essa variável;

b) Calcule o primeiro e terceiro quartil;

c) Calcule o percentil 80;

d) Calcule a variância e o coeficiente de variação.

3.3 Apresentações Gráficas

Quando as distribuições de freqüências têm como principal objetivo condensar grandes conjuntos de dados em uma forma fácil de assimilar, é melhor apresentar as distribuições graficamente.

Gráficos de Colunas, Barras e de Setores

Um gráfico de barras é um dispositivo gráfico bastante utilizado para retratar dados qualitativos que foram sintetizados em uma distribuição de freqüência simples, em uma distribuição de freqüência relativa ou em uma distribuição de freqüência percentual. No eixo horizontal do gráfico, especificamos os rótulos que são usados para as classes. Uma escala de freqüência, simples, relativa ou percentual pode ser usada para o eixo vertical do gráfico.

Então, usando-se uma barra de largura fixa desenhada acima de cada rótulo de classe, estendemos a altura da barra até atingir a freqüência, simples, relativa ou percentual como indicado pelo eixo vertical. As barras são separadas para enfatizar o fato de que cada classe é uma categoria em separado.

O gráfico de setores (ou de pizza) é um círculo dividido em setores, cujos tamanhos são proporcionais as freqüências ou percentagens correspondentes.

Exemplo 3.3.1 A seguir é apresentada uma tabela de freqüências que sintetiza a variável SEXO dos funcionários da empresa XYZ.

Tabela 3.3.1: Frequências de acordo com o sexo dos funcionários da empresa XYZ

Sexo	Frequência	Percentual
Masculino	258	54,43%
Feminino	216	45,57%
Total	474	100%

Gráfico 3.3.1: Frequência percentual dos funcionários da empresa XYZ de acordo com o sexo dos funcionários da empresa XYZ(%)

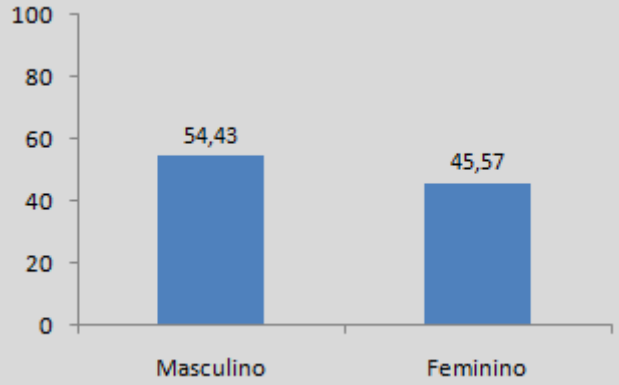
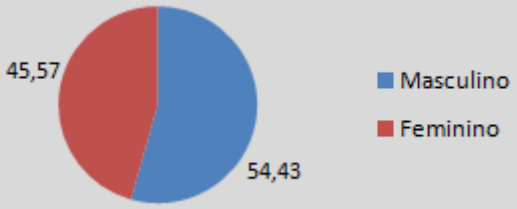
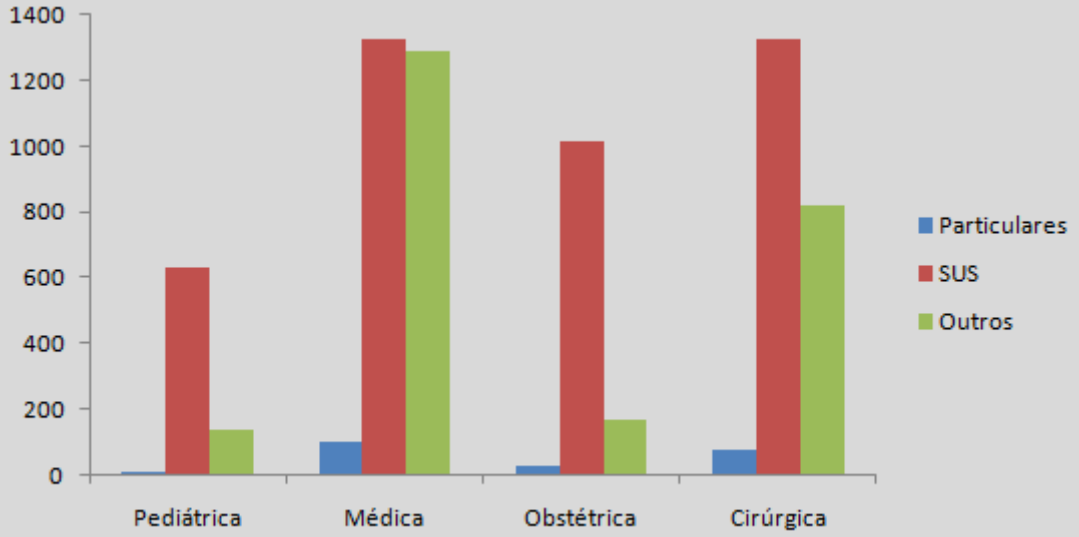


Gráfico 3.3.2: Frequência percentual dos funcionários da empresa XYZ de acordo com o sexo dos funcionários da empresa XYZ(%)



Para os casos de tabelas de dupla entrada como a Tabela 3.1.5, os dados podem ser representados graficamente da seguinte forma:

Gráfico 3.3.3: Número de pacientes internados no Hospital São Sebastião, por tipo de clínica e por tipo de convênio, no ano de 2002



O Boxplot

O *Boxplot* é um gráfico proposto para a detecção de valores discrepantes (*outliers*), que são aqueles valores muito diferentes do restante do conjunto de dados. Esses valores discrepantes podem representar erros no processo de coleta ou de processamento dos dados, e, nesse caso, devem ser corrigidos ou excluídos do banco de dados.

No entanto, os *outliers* podem ser valores corretos, que, por alguma razão, são muito diferentes dos demais valores. Nesse caso, a análise desses dados deve ser cuidadosa, pois, como sabemos, algumas estatísticas descritivas, como a média e o desvio padrão, são influenciadas por valores extremos.

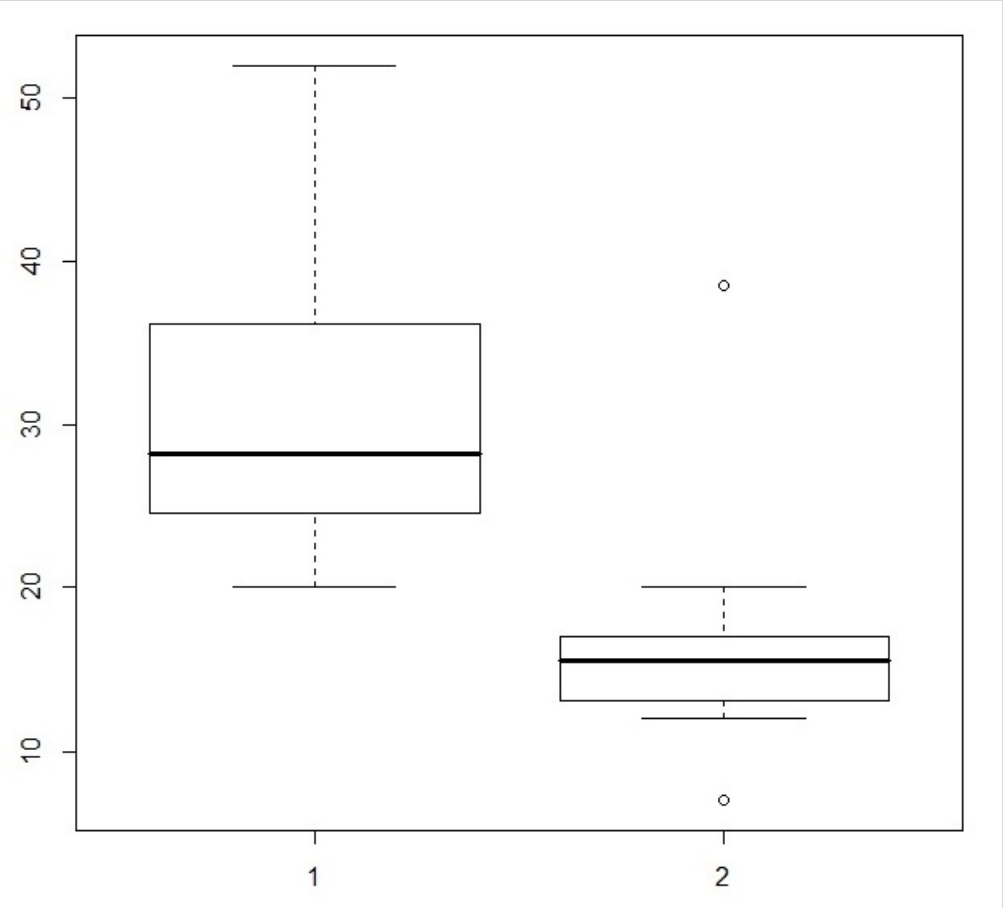
Na construção do *Boxplot*, utilizamos alguns percentis (mediana, primeiro e terceiro quartis), que são pouco influenciados por valores extremos. Além disso, precisamos saber quais são os valores mínimo e máximo do conjunto de dados.

O *Boxplot* é constituído por uma caixa atravessada por uma linha, construído usando um eixo com uma escala de valores. O fundo da caixa é marcado na escala de valores na altura do primeiro quartil (Q1). O topo da caixa é marcado na altura do terceiro quartil(Q3). Uma linha é traçada dentro da caixa na altura da mediana, que não precisa estar necessariamente no meio da caixa. Como sabemos, entre o primeiro e o terceiro quartis, temos 50% dos dados. Podemos pensar, então, que essa caixa contém metade dos dados do conjunto.

A altura da caixa é dada pela **amplitude interquartil** ($AIQ = Q3 - Q1$).

A partir da caixa, para cima, segue uma linha até o ponto mais remoto que não exceda $LS = Q3 + (1,5) \cdot AIQ$, chamado *limite superior*. De modo similar, da parte inferior da caixa, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que $LI = Q1 - (1,5) \cdot AIQ$, chamado *limite inferior*. Os valores compreendidos entre esses dois limites são chamados *valores adjacentes*. As observações que estiverem acima do limite superior ou abaixo do limite inferior estabelecidos serão chamadas pontos exteriores (atípicos, discrepantes) e representadas por asteriscos.

Exemplo 3.3.2



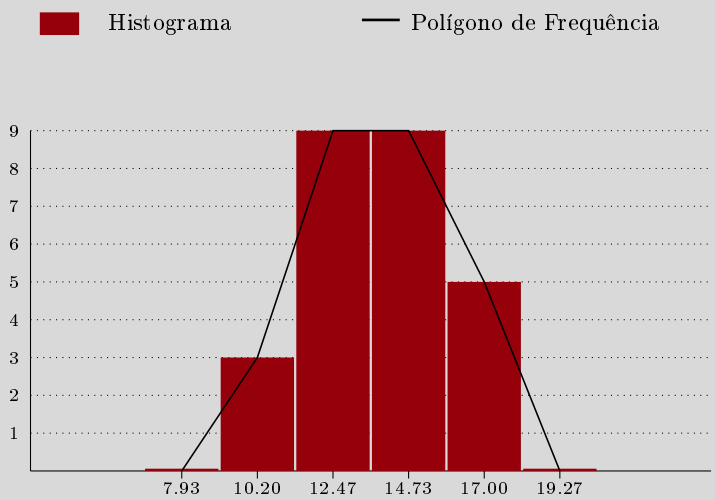
Histograma e Polígono de Freqüência

Para as distribuições de freqüência para dados agrupados em intervalos de classe, a forma mais comum de apresentação gráfica é o histograma. Um histograma é construído, representando-se as medidas ou observações que são agrupados em uma escala horizontal e as freqüências de classe ou percentuais em uma escala vertical; traçam-se então retângulos, cujas bases são iguais aos intervalos de classe e cujas alturas são as freqüências de classe ou percentuais correspondentes. Note que os retângulos de um histograma vão de uma fronteira de classe a outra.

Outra forma de apresentação gráfica para as distribuições de freqüência para dados agrupados em intervalos de classe é o polígono de freqüência. Aqui, as freqüências de classe são marcadas nos pontos médios, e os valores sucessivos são unidos por segmentos retilíneos. O polígono se inicia no ponto médio da classe anterior a 1ª classe e se encerra no ponto médio da classe posterior ao último intervalo de classe.

Exemplo 3.3.3 A seguir é apresentado um exemplo de histograma e um polígono de freqüências para uma variável tempo de estudo.

Figura 3.3.1: Histograma e polígono de frequencias para horas de estudo semanal fora de sala de aula



4 Assimetria e Curtose

Duas distribuições também podem diferir uma da outra em termos de assimetria ou achatamento, ou ambas. Como veremos, assimetria e achatamento (o nome técnico utilizado para esta última característica de forma da distribuição é *curtose*) têm importância devido a considerações teóricas relativas à inferência estatística que são freqüentemente baseadas na hipótese de populações distribuídas normalmente. Medidas de assimetria e de curtose são, portanto, úteis para se precaver contra erros aos estabelecer esta hipótese.

4.1 Assimetria

Posições Relativas da Média, Mediana e Moda em Função da Assimetria das Distribuições

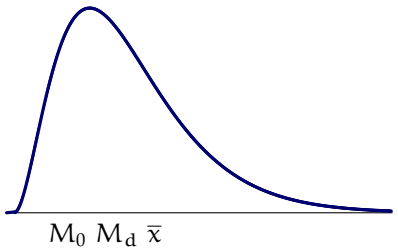


Figura 4.1.1: Distribuição Assimétrica à Direita

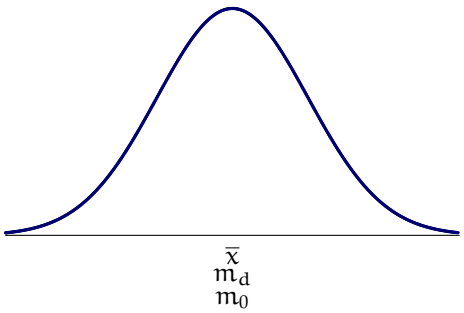


Figura 4.1.2: Distribuição Simétrica

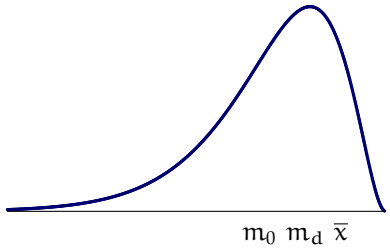


Figura 4.1.3: Distribuição Assimétrica à Esquerda

A seguir apresentaremos histogramas de distribuições assimétricas e simétrica:

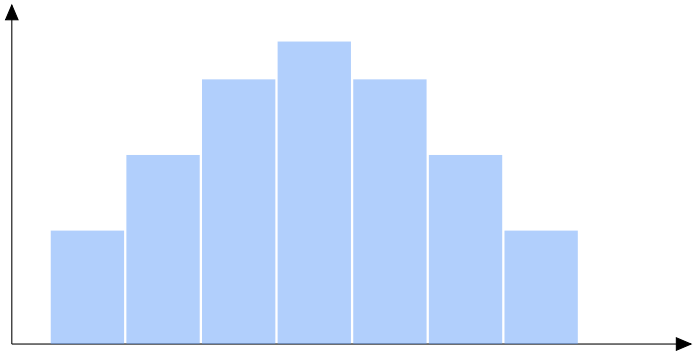


Figura 4.1.4: Histograma de distribuição simétrica

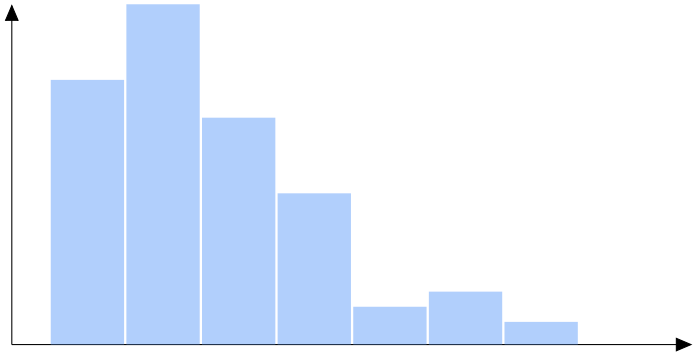


Figura 4.1.5: Histograma de distribuição asssimétrica para a direita

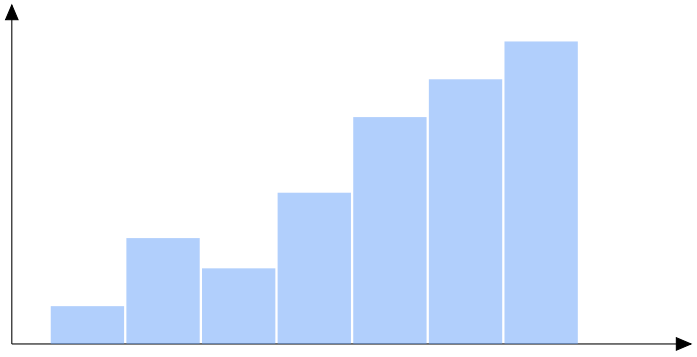


Figura 4.1.6: Histograma de distribuição asssimétrica para a esquerda

Podemos medir também a assimetria por meio de duas medidas: *coeficiente de assimetria de Pearson*, denotado por A_s e o *coeficiente momento de assimetria*, denotado por a_3.

O coeficiente de assimetria de Pearson:

$$A_s = \frac{3(\overline{x} - Mediana)}{s}$$

Então, temos que:
$A_s = 0$, temos uma distribuição simétrica;
$A_s < 0$, temos uma distribuição assimétrica à esquerda;
$A_s > 0$, temos uma distribuição assimétrica à direita.

O coeficiente assimetria de momento:

$$a_3 = \frac{m_3}{m_2\sqrt{m_2}}$$

Sendo:

$$m_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \qquad m_3 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n}$$

Então, temos que:
$a_3 = 0$, temos uma distribuição simétrica;
$a_3 < 0$, temos uma distribuição assimétrica à esquerda;
$a_3 > 0$, temos uma distribuição assimétrica à direita.

Observação 4.1.1 Identificar se a distribuição de uma variável quantitativa em um determinado conjunto de dados é simétrica ou assimétrica pode ser de grande valia por vários motivos:

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

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

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

4.2 Curtose: uma medida de achatamento

A Curtose indica até que ponto a curva de freqüências de uma distribuição se apresenta mais afilada ou mais achatada do que uma curva padrão, denominada normal. Apresentaremos agora

duas medidas de achatamento das distribuições, *o coeficiente percentílico de curtose*, denotado por C e o *coeficiente momento de curtose*, denotado por a_4.

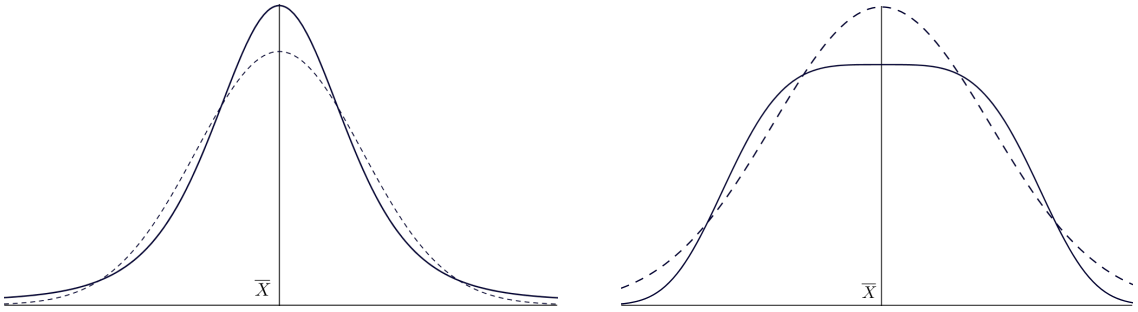
O coeficiente percentílico de curtose:

Esta medida é algebricamente tratável e geometricamente interpretável.

$$C = \frac{(Q_3 - Q_1)}{2(P_{90} - P_{10})}$$

Por meio do coeficiente de curtose, classificamos diferentes graus de achatamento em três categorias: leptocúrtica, platicúrtica e mesocúrtica (ver Figura 4.2.1).

Figura 4.2.1: Curtose leptocúrtica, platicúrtica e mesocúrtica



Na Figura 4.2.1 compara-se a curtose de duas distribuições com a curtose de uma distribuição mesocúrtica (em linha tracejada). Na figura da esquerda temos uma distribuição leptocúrtica (linha cheia) e na figura da direita temos uma distribuição platicúrtica (linha cheia).

Temos então que:

$C = 0,263$, a curva é mesocúrtica;

$C > 0,263$, a curva é platicúrtica;

$C < 0,263$, a curva é leptocúrtica.

O coeficiente momento de curtose:

$$a_4 = \frac{m_4}{m_2^2}$$

Sendo:

$$m_4 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{n}$$

Temos então que:

$a_4 = 3$, a curva é mesocúrtica;

$a_4 < 3$, a curva é platicúrtica;

$a_4 > 3$, a curva é leptocúrtica.

4.3 Exercícios

4.3.1 Um artigo reportou dados sobre um experimento, investigando o efeito de muitas variáveis de processos de oxidação, em fase vapor, de naftaleno. Uma amostra de conversão percentual molar de naftaleno em anidrido maléico resulta em:
4,2 ; 4,7 ; 4,7 ; 5,0 ; 3,8 ; 3,6 ; 3,0 ; 5,1 ; 3,1 ; 3,8 ; 4,8 ; 4,0 ; 5,2 ; 4,3 ; 2,8 ; 2,0 ; 2,8 ; 3,3 ; 4,8 ; 5,0 ; 4,8 ; 3,9 ; 5,3 ; 5,0 ; 4,7 ; 3,6 ; 3,8 ; 3,0 ; 3,2 ; 4,2 ; 4,5 ; 4,7 ; 4,9 ; 4,0 ; 4,1 ; 4,4 ; 5,0

a) Encontre a média, a mediana e a moda;

b) Encontre o 45º percentil e interprete-o;

c) Organize os dados em uma tabela de frequências adequada.

4.3.2 Acredita-se que a resistência à tensão da borracha siliconizada seja uma função da temperatura de cura. Um estudo foi realizado, no qual amostras de 12 espécimes de borracha foram preparadas usando temperaturas de cura de 20°C e 45°C. Os dados mostram os valores de resistência à tensão, em megapascals:

20°C	2,07	2,14	2,22	2,03	2,21	2,03	2,05	2,18	2,09	2,14	2,11	2,02
45°C	2,52	2,15	2,49	2,03	2,37	2,05	1,99	2,42	2,08	2,42	2,29	2,01

a) Identifique a variável em estudo e classifique-a;

b) Faça uma análise descritiva comparando os dois grupos e interpretando os resultados a partir de:

b.1) Média, mediana e o 70º percentil e, usando a medida adequada, em termos de variabilidade;

4.3.3 Uma equipe de Higiene e Segurança do Trabalho de uma empresa de aviação, preocupada com o número de horas trabalhadas pelos funcionários, observou em seus registros recentes, uma amostra com o tempo de mão-de-obra gasto na revisão completa de um motor de jato. A tabela 4.3.1 foi obtida:

Tabela 4.3.1: Tempo de mão de obra gasto na revisão de um motor de jatos

Tempo (horas)	f_i
$0,00 \vdash 4,00$	1
$4,00 \vdash 8,00$	5
$8,00 \vdash 2,00$	10
$12,00 \vdash 16,00$	15
$16,00 \vdash 20,00$	4

a) Determine o número médio de horas de mão de obra necessário para revisão de cada motor;

b) Com base nesta informação (item a)), qual deve ser o tempo total de mão de obra para a revisão de dez motores que aguardam revisão?

c) Se a empresa dispõe no momento de dois homens trabalhando 12 horas por dia nestas revisões, conseguirá provavelmente revisar estes dez motores em quatro dias? Por que?

4.3.4 Foram anotados os níveis de colesterol (em $\mathrm{mg/100ml}$) para uma amostra de trinta pacientes de uma clínica cardíaca. As medidas se referem a homens entre 40 e 60 anos de idade que foram à clínica fazer um *check-up*.

160	161	160	170	167	163	172	172	173	177
178	182	181	181	186	185	194	197	199	203
205	203	206	206	211	209	208	214	218	225

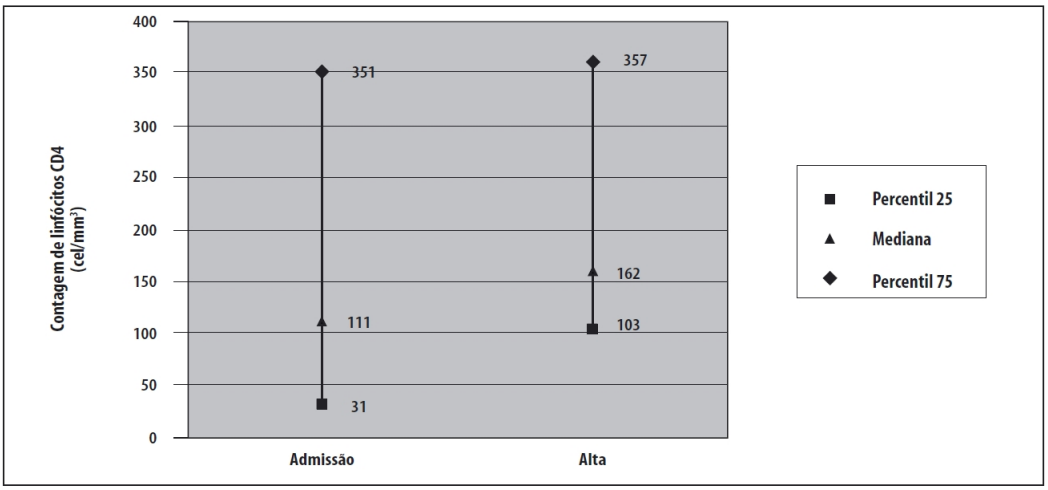
a) Indique a variável em estudo e classifique-a;

b) Calcule o nível de colesterol mediano e calcule e interprete o percentil 67;

c) Organize os dados em uma tabela de frequência completa com intervalos de classes de tamanho 10, iniciando do nível 160;

d) Refaça o item b) usando as informações da tabela de frequencias obtida em b) e comente as diferenças encontradas entre os valores das medidas calculadas em b) e d).

4.3.5 No trabalho Perfil dos Pacientes com AIDS acompanhados pelo Serviço de Assistência Domiciliar Terapêutica do Município de Contagem, Estado de Minas Gerais, Brasil, 2000-2003, tem-se que Serviço de Assistência Domiciliar Terapêutica (ADT) do Município de Contagem, Estado de Minas Gerais, Brasil, atende pacientes com AIDS impossibilitados de comparecer ao ambulatório ou que apresentam dificuldade de adesão ao tratamento. O objetivo do estudo é avaliar as indicações mais freqüentes de ADT, as características dos pacientes atendidos e sua evolução.



Contagem de linfócitos $\mathrm{CD4}^+$ na admissão e na alta de 34 pacientes assistidos pelo Serviço de Assistência Domiciliar Terapêutica do Município de Contagem, Estado de Minas Gerais. Brasil, agosto de 2000 a dezembro de 2003

a) Quais os valores dos percentis 25, 50 e 75 na admissão (na entrada) dos sujeitos no ADT?

b) Quais os valores dos percentis 25, 50 e 75 na alta (na saída) dos sujeitos no ADT?

c) Interprete e compare os resultados de a) e b).

4.3.6 Como parte de um estudo de controle de qualidade que visa melhorar uma linha de produção, os pesos (em onças) de uma amostra de 40 barras de sabão foram medidos. Os resultados estão abaixo:

11,6	14,3	15,8	16,5	12,7	12,8	16,5	13,7	13,3	14,3
14,6	15,9	15,2	15,6	15,6	15,8	16,2	16,5	16,5	17,3
15,9	17,1	18,3	18,8	20,6	17,4	17,6	12,6	18,3	18,5
17,7	17,0	16,1	15,8	16,4	19,2	20,3	14,6	14,8	15,0

a) Qual a variável em estudo. Classifique-a.

b) Represente estes dados por meio de um boxplot.

c) Calcule e interprete o terceiro quartil

4.3.7 Um a empresa estabelece o salário de seus vendedores com base na produtividade. Desta forma, 10% é fixo e 90% são comissões sobre a venda. Uma amostra de salários mensais nesta empresa revelou Tabela 4.3.2 abaixo. Se a empresa decidir, a nível de incentivo, fornecer uma cesta básica para 5% dos vendedores que pior desempenho tiveram durante o próximo mês com base nesta amostra, qual será o maior salário que receberá esta cesta básica?

Tabela 4.3.2: Salários dos Vendedores

US$	f_i
$70,00 \vdash 120,00$	8
$120,00 \vdash 170,00$	28
$170,00 \vdash 220,00$	54
$220,00 \vdash 270,00$	32
$270,00 \vdash 320,00$	12
$320,00 \vdash 370,00$	6

4.3.8 A média e o desvio padrão da produtividade de duas cultivares de milho são respectivamente $\overline{x}_A = 4,0$ t/ha e $s_A = 0,80$ t/ha para a variedade de polinização aberta A e $\overline{x}_B = 8,0$ t/ha e $s_B = 1,20$ t/ha para o híbrido simples B. Qual das cultivares possui maior uniformidade de produção?

4.3.9 Os agentes de fiscalização de certo município realizam, periodicamente, uma vistoria nos bares e restaurantes para apurar possíveis irregularidades na venda de seus produtos. A seguir, são apresentados dados de uma vistoria sobre os pesos (em gramas) de uma amostra de 10 bifes, constantes de um cardápio de um restaurante como "bife de 200 gramas":

170 175 180 185 190 195 200 200 200 205

a) Se o peso mediano da amostra for inferior 185 gramas e o terceiro quartil não ultrapassar 195 o estabelecimento recebe uma advertência por meio de uma notificação. Avaliando resultado desta amostra, o restaurante avaliado receberá uma advertência?

b) Em um outro município do mesmo porte a equipe de fiscalização utiliza outro procedimento: aplicar a advertência naqueles estabelecimentos em que o primeiro quartil for inferior a 180 gramas e o peso médio for menor que 190 gramas. Desta maneira os dados da amostra apresentados anteriormente levariam o estabelecimento a receber a advertência nesse outro município?

4.3.10 Uma distribuidora de refrigerantes fez um levantamento sobre o consumo semanal (em litros) por pessoa, em jan/2005, em uma cidade do litoral, obtendo a tabela abaixo:

CONSUMO	Nº DE PESSOAS
$0,0 \vdash 0,5$	10
$0,5 \vdash 1,0$	25
$1,0 \vdash 1,5$	9
$1,5 \vdash 2,0$	7
$2,0 \vdash 2,5$	6

a) Determine e interprete o consumo médio.

b) Qual o percentual de pessoas que consomem menos de 1 litro por semana?

c) Determine e interprete o consumo modal e o consumo mediano.

d) Se a empresa tem um lucro de R$0,50 por litro, qual o lucro médio por pessoa?

e) Calcule o coeficiente de assimetria de Pearson e classifique a distribuição dos dados.

4.3.11 (Problema 33 do Capítulo 3 do livro Estatística Básica de Bussab e Morettin) Um órgão do governo do estado está interessado em determinar padrões sobre o investimento em educação, por habitante, realizado pelas prefeituras. De um levantamento amostral com 10 cidades, foram obtidos os valores da tabela abaixo:

Cidade	A	B	C	D	E	F	G	H	I	J
Investimento	20	16	14	8	19	15	14	16	19	18

Nesse caso, será considerado como investimento básico a média final das observações calculada da seguinte maneira:

1. Obtém-se uma média inicial.
2. Eliminam-se do conjunto aquelas observações que forem superiores à média inicial mais duas vezes o desvio padrão, ou inferiores à média inicial menos duas vezes o desvio padrão. (Este procedimento tem a finalidade de eliminar do conjunto a cidade cujo investimento é muito diferente dos demais)
3. Calcula-se a média final com o novo conjunto de observações.

Qual o investimento básico que você daria como resposta?

4.3.12 Mostre que:

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

4.3.13 Foi realizada na região Oeste do Paraná, no município de Marechal Cândido Rondon, em 1992, um levantamento da produtividade leiteira diária de 20 produtores rurais, atendidos pelo plano "Panela Cheia"(Roesler, 1997). Os resultados dos intervalos de parto (em meses) dos 20 produtores estão apresentados a seguir.

11,80	11,90	12,00	12,30	12,80	12,99	13,10	13,50	13,80	14,10
14,55	14,65	14,70	15,00	15,10	15,20	15,50	15,80	15,90	15,96

Obtenha as seguintes estimativas das medidas de dispersão:

a) Amplitude total;

b) Variância e desvio padrão;

c) Coeficiente de variação;

d) Em cada caso anterior comentar, sobre o significado da estimativa obtida e sobre a forma que devem ser aplicadas;

e) Com a relação à Curtose, qual é a classificação destes dados?

f) Se cada dado for dividido por 12, para se obter o intervalo de partos em anos, qual será os novos valores da amplitude, variância, desvio padrão, CV e erro padrão da média?

g) Se você fosse solicitado a representar os dados por duas medidas, quais você usaria e por que?

h) Após o programa Panela Cheia o intervalo de partos apresentou média de 13,85 e desvio padrão de 2,00 meses. Qual é na sua opinião a situação que apresentou maior variabilidade, antes ou após o Programa?

4.3.14 Abaixo estão representados os dados referentes a um grupo de animais avaliados pela idade, sexo, espécie, e nível de infestação por protozoários.

Quadro 4.3.1: Animais de diversas espécies avaliados quanto a idade, sexo, espécie e nível de infestação na Fazenda Passa Quatro – Cordeiro – RJ, no período de 2006

Animal	Idade	Sexo	Espécie	Nível de Infestação
1	2	Macho	Bovino	Alta
2	2	Macho	Suíno	Baixa
3	1	Fêmea	Caprino	Média
4	4	Fêmea	Bovino	Média
5	5	Fêmea	Caprino	Alta
6	2	Fêmea	Caprino	Alta
7	4	Macho	Suíno	Média
8	2	Fêmea	Suíno	Média
9	5	Macho	Bovino	Baixa
10	5	Fêmea	Bovino	Média
11	3	Macho	Caprino	Alta
12	3	Macho	Caprino	Média
13	5	Macho	Bovino	Média
14	1	Fêmea	Suíno	Baixa

Fonte: Fazenda Passa Quatro – Cordeiro - RJ

a) Identifique e classifique todas as variáveis descritas no banco de dados acima;

b) Construa tabelas de distribuição de freqüência para as variáveis idade e sexo;

c) Construa uma tabela de dupla entrada (ou tabela de contingência) para as variáveis espécie e nível de infestação e faça uma análise da tabela;

d) Represente com gráfico adequado as variáveis *espécie* e *nível de infestação*.

5 Probabilidade

Definição 5.0.1 Probabilidade é a possibilidade, ou chance, de que um determinado evento venha a ocorrer, podendo ser, por exemplo, a chance de retirar uma carta preta de um baralho de cartas, a chance de um indivíduo preferir um produto em relação a outro ou ainda a chance de um novo produto de mercado obter sucesso.

Além de sua aplicação na metodologia estatística, a teoria da probabilidade vem adquirindo importância crescente como instrumento analítico em uma sociedade que é forçada a medir incertezas. Por exemplo, antes de ativar uma usina nuclear, devemos analisar s probabilidade de um acidente. Antes de armar um artefato nuclear, devemos analisar a probabilidade de uma detonação acidental. Antes de aumentar o limite de velocidade m rodovias, devemos procurar estimar a probabilidade do aumento de acidentes fatais.

5.1 Conceitos Básicos

Ao lidarmos com problemas de probabilidade, vamos encontrar experimentos, eventos e a coleção de todos os resultados possíveis. A seguir temos alguns conceitos importantes dentro da probabilidade:

- **Experimento:** É qualquer processo que permite ao pesquisador fazer observações. Ele pode ser determinístico ou aleatório.

- **Espaço amostral:** Consiste em um conjunto de todos os resultados possíveis de um experimento, será representado pela letra grega Ω (ômega).

- **Evento:** É um subconjunto do espaço amostral de um experimento aleatório, será representado letras maiúsculas ($A, B, C, ...$).

- A **União** de dois eventos A e B, denotada por $A \cup B$, representa a ocorrência de pelo menos um dos eventos A ou B.

- A **Intersecção** do evento A com B, denotada por $A \cap B$, é a ocorrência simultânea de A e B.

- Dois eventos A e B são **disjuntos** ou **mutuamente exclusivos** ou **mutuamente excludentes** quando não têm elementos em comum. Isto é,

$$A \cap B = \varnothing.$$

- Dizemos que A e B são **complementares** se sua união é o espaço amostral e sua intersecção é vazia. O complementar de A será representado por A^C e temos que:

$$A \cup A^C = \Omega \qquad \text{e} \qquad A \cap A^C = \varnothing.$$

Vamos considerar probabilidade como sendo uma função $P(.)$ que atribui valores numéricos aos eventos do espaço amostral, conforme definição a seguir:

Definição 5.1.1 Uma função $P(.)$ é denominada probabilidade se

i) $0 \leqslant P(A) \leqslant 1, \forall\, A \subset \Omega$;

ii) $P(\Omega) = 1$;

iii) $P\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} P(A_j)$, onde $A_i \cap A_j = \varnothing$ para $i \neq j$.

Abordaremos apenas o conceito de probabilidade relacionada com a ocorrência de um evento em relação a todas as possibilidades possíveis. Se o evento A pode ocorrer de $n(A)$ maneiras diferentes num total de $n(\Omega)$ modos possíveis, então a probabilidade de ocorrência de A é definida por:

$$P(A) = \frac{n(A)}{n(\Omega)}.$$

Exemplo 5.1.1 Uma fábrica produz determinado artigo. Da linha de produção são retirados três artigos, e cada um é classificado como bom (B) ou defeituoso (D) de acordo com a ordem de seleção. Um espaço amostral deste experimento é:

$$\Omega = \{BBB, BBD, BDB, BDD, DBB, DBD, DDB, DDD\}$$

Seja A o evento que consiste em obter exatamente dois artigos defeituosos, então:

$$A = \{BDD, DBD, DDB\}$$

Exemplo 5.1.2 Considere o experimento que consiste em retirar uma lâmpada de um lote e medir seu "tempo de vida" antes de queimar. Um espaço amostral conveniente é:

$$\Omega = \{t \in \mathbb{R} : t \geqslant 0\}$$

Se A indicar o evento "o tempo de vida da lâmpada é inferior a 20 horas", então:

$$A = \{t : 0 \leqslant t \leqslant 20\}.$$

As operações da união, intersecção e complementação entre eventos possuem propriedades análogas àquelas válidas para operações entre conjuntos:

i) $(A \cap B)^c = A^c \cup B^c$

ii) $(A \cup B)^c = A^c \cap B^c$

iii) $A \cap \varnothing = \varnothing$, $A \cap \Omega = A$

iv) $\varnothing^c = \Omega$, $\Omega^c = \varnothing$

v) $A \cap A^c = \varnothing$

vi) $A \cup A^c = \Omega$

vii) $A \cup \varnothing = A$, $A \cup \Omega = \Omega$

viii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Observação 5.1.1 Temos as seguintes fórmulas para o cálculo de probabilidades:

i) $P(A^c) = 1 - P(A)$

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

iii) $P(A \cup B) = P(A) + P(B)$, quando A e B forem mutuamente exclusivos.

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

5.2 Probabilidade Condicional, Indepedência e Teorema de Bayes

A avaliação das chances de que um evento ocorra pode ser muito diferente, dependendo da informação que temos. Uma estimativa da probabilidade de sua casa ruir amanhã seria claramente muito maior se um terremoto violento estiver sendo esperado do que se não houvesse qualquer razão para esperar uma atividade sísmica.

Probabilidade Condicional

Para dois eventos quaisquer A e B, sendo $P(B) > 0$, definimos a probabilidade condicional de A dado B, $P(A\backslash B)$, como sendo:

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

Se quisermos definir a probabilidade condicional de B dado A, $P(B\backslash A)$, sendo $P(A) > 0$, temos:

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

Exemplo 5.2.1 A seguir temos dados referentes a alunos matriculados em quatro cursos de mestrado do Departamento de Ciências Exatas de uma grande universidade no ano de 2005. Selecionando um aluno aleatoriamente, sabe-se que ele está matriculado em Estatística. Calcule a probabilidade deste aluno ser do sexo feminino.

	SEXO	
CURSO	HOMENS	MULHERES
Matemática	70	40
Matemática Aplicada	15	15
Estatística	10	20
Ciência da Computação	20	10

Se A e B indicam, respectivamente, os eventos "aluno é mulher"e "aluno matriculado em Estatística", então $P(A\backslash B)$ será:

$$P(A\backslash B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{20}{200}}{\frac{30}{200}} = \frac{2}{3}$$

Regra da Multiplicação:

A definição de probabilidade condicional pode ser rescrita para fornecer uma expressão geral pra a probabilidade da intersecção de dois eventos:

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

Exemplo 5.2.2 A probabilidade de que uma bateria de automóvel, sujeita a alta temperatura no compartimento do motor, sofra baixa corrente de carga é $0,7$. A probabilidade da bateria estar sujeita a alta temperatura no compartimento do motor é $0,05$.

Faça A denotar o evento em que a bateria sofra baixa corrente de carga e faça B denotar o evento em que a bateria esteja sujeita a alta temperatura no compartimento do motor. A probabilidade da bateria estar sujeita a baixa corrente de carga e a alta temperatura no compartimento do motor é:

$$P(A \cap B) = P(A\backslash B) \cdot P(B) = 0,7 \cdot 0,05 = 0,035.$$

Independência:

Dois eventos A e B são independentes se e somente se:

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

então têm-se que:

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

Vejamos agora o conceito de independência pra três eventos: dizemos que os eventos A, B, C são independentes se, e somente se:

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

ii) $P(A \cap C) = P(A) \cdot P(C)$

iii) $P(B \cap C) = P(B) \cdot P(C)$

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

Se apenas as três primeiras relações estiverem satisfeitas, dizemos que os eventos A, B, C são mutuamente independentes. É possível que três eventos sejam mutuamente independentes, mas não sejam completamente independentes.

Teorema de Bayes:

Freqüentemente, começamos a análise com um cálculo de probabilidade *inicial* ou *prévia* para eventos de interesse específico. Então, a partir de fontes tais como uma amostra, um relatório especial ou um interesse específico. Então, a partir de fontes tais como uma amostra, um relatório especial ou um teste de produto, obtemos informação adicional sobre os eventos. Dada essa nova informação, atualizamos os valores prévio da probabilidade calculando as probabilidades adicionais, denotadas *probabilidades posteriores*. O **Teorema de Bayes** fornece um meio de fazer esses cálculos de probabilidade.

Teorema 5.2.1 (Teorema de Bayes) *Suponha que A_1, A_2, ..., A_n são eventos mutuamente exclusivos cuja união é o espaço amostral Ω. Então, se A é um evento qualquer, temos que:*

$$P(A_k \backslash A) = \frac{P(A_k) \cdot P(A \backslash A_k)}{\sum_{j=1}^{n} P(A_j) \cdot P(A \backslash A_j)}.$$

Isto nos permite encontrar as probabilidades dos vários eventos A_1, A_2, ..., A_n que podem ser a causa de ocorrência de A.

Demonstração: Da definição de probabilidade condicional temos:

$$P(A_k \backslash A) = \frac{P(A_k \cap A)}{P(A)}$$

O numerador desta expressão pode ser reescrito pela regra do produto, condicionado à A_k, isto é,

$$P(A_k \cap A) = P(A \cap A_k) = P(A_k) \cdot P(A \backslash A_k)$$

Para completar a demonstração note que:

$$P(A) = \sum_{j=1}^{n} P(A \cap A_j) = \sum_{j=1}^{n} P(A_j) \cdot P(A \backslash A_j)$$

Exemplo 5.2.3 Para selecionar seus funcionários, uma empresa oferece aos candidatos um curso de treinamento durante uma semana. No final do curso, eles são submetidos a uma prova e 25% são classificados como bons (B), 50% como médios (M) e os restantes 25% como fracos (F). Para facilitar a seleção, a empresa pretende substituir o treinamento por um teste contendo questões referentes a conhecimentos gerais e específicos. Para isso, gostaria de conhecer qual é a probabilidade de um indivíduo aprovado no teste ser considerado fraco, caso

fizesse o curso. Assim, neste ano, antes do início do curso, os candidatos foram submetidos ao teste e receberam o conceito aprovado (A) ou reprovado (R). No final do curso, obtiveram-se as seguintes probabilidades condicionais:

$$P(A\backslash B) = 0,80 \quad P(A\backslash M) = 0,50 \quad P(A\backslash F) = 0,20$$

Pelo Teorema de Bayes a probabilidade pedida será calculada por:

$$P(F\backslash A) = \frac{P(F) \cdot P(A\backslash F)}{P(B) \cdot P(A\backslash B) + P(M) \cdot P(A\backslash M) + P(F) \cdot P(A\backslash F)} = \\ = \frac{0,25 \cdot 0,20}{0,25 \cdot 0,80 + 0,50 \cdot 0,50 + 0,25 \cdot 0,20} = 0,10$$

Então, apenas 10% dos aprovados é que seriam classificados como fracos durante o curso.

5.3 Variável Aleatória e Distribuição de Probabilidade

Variável Aleatória

Suponha que cada ponto de um espaço amostral seja atribuído um número. Temos então uma função definida em um espaço amostral denominada *variável aleatória* (v.a.). Essa variável geralmente é denotada por uma letra maiúscula como X ou Y.

Uma quantidade X, associada a cada possível resultado do espaço amostral, é denominada de **variável aleatória discreta** se assume valores num conjunto enumerável, com certa probabilidade. Por outro lado, será denominada **variável aleatória contínua** se seu conjunto de valores é qualquer intervalo dos números reais, o que seria um conjunto não enumerável.

Exemplo 5.3.1 Uma moeda é lançada duas vezes e é observada sua face. O espaço amostral é:

$$\Omega = \{KK, KC, CK, CC\}$$

Uma variável aleatória de interesse poderia ser: X = nº de caras. A cada evento simples, ou ponto de Ω, associamos um nº, que é o valor assumido para a variável aleatória X.

Evento	KK	KC	CK	CC
X	0	1	1	2

Os valores de X são então 0, 1 e 2.

Distribuição de Probabilidade

Uma distribuição de probabilidade é um modelo matemático que relaciona o valor da variável com a probabilidade de ocorrência daquele valor na população.

Quando o parâmetro sendo medido só pode assumir certos valores, tais como os inteiros 0, 1, 2, ..., a distribuição de probabilidade é chamada **distribuição discreta**. Por exemplo, a distribuição do nº de defeitos em placas de circuito seria uma variável discreta.

Quando a variável aleatória contém um nº infinito não-enumerável de pontos, temos assim as **distribuições contínuas** de probabilidade.

Distribuição Discreta de Probabilidade

A função que atribui a cada valor da variável aleatória discreta sua respectiva probabilidade é denominada de função discreta de probabilidade ou simplesmente função de probabilidade. Essa função é representada na seguinte tabela:

X	x_1	x_2	x_3	...
$P(X = x_i)$	p1	p2	p3	...

Sendo que

$$0 \leqslant p_i \leqslant 1 \text{ e } \sum_{i}^{k} p_i = 1$$

Exemplo 5.3.2 No exemplo 5.3.1 do lançamento da moeda duas vezes teremos a seguinte tabela que pode representar a distribuição de probabilidade discreta:

X (nº de caras)	0	1	2
$P(X = x_i)$	0,25	0,50	0,25

Média e Variância de uma Variável Aleatória Discreta

A média de uma variável aleatória X usa o modelo de probabilidade para ponderar os valores possíveis de X. A *Média* ou *Valor Esperado* ou ainda *Esperança Matemática* de X, denotado por $E(X)$ ou μ, é:

$$E(x) = \mu = \sum_{i=1}^{k} x_i P(X = x_i)$$

A Esperança Matemática possui as seguintes propriedades:

i) $E(K) = K$, K = constante;

ii) $E(KX) = KE(X)$;

iii) $E(X + Y) = E(X) + E(Y)$;

iv) $E(X - Y) = E(X) - E(Y)$;

v) $E(X \pm K) = E(X) \pm K$;

vi) Se X e Y são independentes, então $E(X \cdot Y) = E(X) \cdot E(Y)$

A variância de uma variável aleatória X é uma medida de dispersão ou espalhamento nos valores possíveis para X. A variância de X, denotada por $V(X)$ ou σ^2 é:

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

sendo que:

$$E(X^2) = \sum_{i=1}^{k} x_i^2 P(X = x_i)$$

Observação 5.3.1 O **desvio padrão** como visto anteriormente é a raiz quadrada da variância.

Exemplo 5.3.3 O número de mensagens enviadas por hora, através de uma rede de computadores, tem a seguinte distribuição:

$X = n^o$ de mensagens	10	11	12	13	14	15
$P(X = x_i)$	0,08	0,15	0,30	0,20	0,20	0,07

Determine a média e o desvio padrão do número de mensagens enviadas por hora.

$$E(X) = 10 \cdot (0,08) + 11 \cdot (0,15) + 12 \cdot (0,30) + 13 \cdot (0,20) + 14 \cdot (0,20) + 15 \cdot (0,07) = 12,5$$

$$E(X^2) = 10^2 \cdot (0,08) + 11^2 \cdot (0,15) + 12^2 \cdot (0,30) + 13^2 \cdot (0,20) + 14^2 \cdot (0,20) + 15^2 \cdot (0,07) = 158,1$$

$$V(X) = 158,1 - [12,5]^2 = 1,85$$

$$\sigma = \sqrt{1,85} = 1,36$$

Função de Distribuição Acumulada

Dada a variável aleatória X, chamaremos a função de distribuição acumulada ou simplesmente função de distribuição $F(x)$ à função

$$F(x) = P(X \leqslant x)$$

5.4 Exercícios

5.4.1 Identifique o que está errado nas afirmações seguintes:

a) A probabilidade de um experimento químico ser bem sucedido é $0,44$ e a probabilidade de falhar é $0,53$;

b) De acordo com um médico, a probabilidade de um paciente contrair gripe é $1,2$;

c) A probabilidade de dois eventos mutuamente exclusivos ocorrerem simultaneamente é sempre igual a 1.

5.4.2 Os problemas de assédio sexual têm recebido muita atenção nos últimos anos. Em uma pesquisa, 420 trabalhadores (240 dos quais homens) consideram uma simples batida no ombro como uma forma de assédio sexual, enquanto que 580 trabalhadores (380 dos quais homens) não consideram isso como assédio (com base nos dados de *Bruskin/Goldrin Research*). Escolhido aleatoriamente um dos trabalhadores pesquisados, determine a probabilidade de obter um homem que não considere um simples tapa no ombro como um forma de assédio sexual.

5.4.3 Quatro executivos têm a responsabilidade de decidir se uma nova filial da empresa deve ser instalada no interior de São Paulo. De acordo com a ordem de questionamento e indicando C, para quem concorda com a nova filial e D, para quem discorda da implantação da nova filial, faça:

a) Monte o espaço amostral;

b) Monte o seguinte evento: "pelo menos dois concordam com a nova filial" e calcule sua probabilidade;

c) Monte o evento complementar do item b).

5.4.4 (Magalhães & Lima. Noções de Probabilidade e Estatística) Dois processadores tipo A e B são colocados em teste por 50 mil horas. A probabilidade de que um erro de cálculo aconteça em um processador do tipo A é de 1/30, no tipo B, 1/80 e ambos, 1/1000. Qual a probabilidade de que:

a) Pelo menos um dos processadores tenha apresentado erro?

b) Nenhum processador tenha apresentado erro?

c) Apenas o processador A tenha apresentado erro?

5.4.5 Uma amostra de 500 famílias foi selecionada em uma grande área metropolitana para determinar várias informações acerca do comportamento do consumidor. Entre as questões indagadas, estava "Você gosta de comprar roupas?". De 240 homens, 136 responderam que sim. De 260 mulheres, 224 responderam que sim.

a) Represente as informações dadas anteriormente por meio de uma tabela de dupla entrada;

b) Qual é a probabilidade de que um entrevistado, aleatoriamente selecionado,

b.1 seja um homem e não goste de comprar roupas?

b.2 seja uma mulher ou goste de comprar roupas?

b.3 não goste de comprar roupas?

b.4 goste de comprar roupa, dado que o entrevistado seja um homem?

b.5 seja um amulher, dado que o enterevistado não goste de comprar roupas?

5.4.6 Em uma indústria de enlatados, as linhas de Produção I, II, III respondem por 50%, 30%, 20% da produção, respectivamente. As proporções de latas com defeito de produção nas linhas I, II, e III são 0,4%, 0,6% e 1,2%. Qual a probabilidade de uma lata defeituosa (descoberta ao final da inspeção do produto acabado) provir da linha I?

5.4.7 Se $P(A) = 0,4$ e $P(B) = 0,5$, que se pode dizer quanto a $P(A \cup B)$, se A e B não são mutuamente exclusivos?

5.4.8 Um estudo de hábitos de fumantes compreende 200 casados (54 dos quais fumam), 100 divorciados (38 dos quais fumam) e 50 adultos que nunca se casaram (11 dos quais fumam) [com base em dados do *Department of Health and Human Services*]. Escolhdio aleatoriamente um indivíduo dessa mostra, determine a probabilidade de obter alguém divorciado ou fumante.

5.4.9 Os trabalhadores de uma fábrica são encorajados constantemente para que se pratique a tolerância zero a acidentes de trabalho. Os acidentes podem ocorrer devido ao ambiente de trabalho ou a condições que não são seguras. Por outro lado, eles podem ocorrer por descuido ou erro humano. Além disso, os turnoa de trabalho dos funcionários, que são das 7h às 15h de trabalho (turno matutino), das 15 às 23h (turno vespertino) e das 23 às 7h (turno noturno), podem ser outro fator de acidentes. Durante o ano passado, ocorreram 300 acidentes. As porcentagens de acidentes para as combinações de condições são:

Turno	Condições Inseguras	Erro humano
Matutino	5%	32%
Vespertino	6%	25%
Noturno	2%	30%

Se um acidente reportado é selecionado aleatoriamente dentre os 300,

a) Qual é a probabilidade de que o acidente tenha ocorrido durante o turno noturno?

b) Qual é a probabilidade de que o acidente tenha ocorrido devido a erro humano?

c) Qual é a probbilidade de que o acidente tenha ocorrido no turno vespertino ou no turno noturno?

5.4.10 Um a empresa de sementes fiscalizadas vende pacotes com 20 Kg cada. As máquinas A, B, e C enchem 25%, 35% e 40% do total produzido, respectivamente. Da produção de cada máquina 5%, 4% e 2% respectivamente, são pacotes fora do peso aceitável. Escolhe-se ao acaso um pacote e verifica-se que está fora do peso aceitável. Qual a probabilidade de que o pacote tenha vindo da máquina A?

5.4.11 As probabilidades prévias para os eventos A_1 e A_2 são $P(A_1) = 0,40$ e $P(A_2) = 0,60$. Sabe-se também que $P(A_1 \cap A_2) = 0$. Suponha que $P(B \backslash A_1) = 0$ e $P(B \backslash A_2) = 0,05$.

a) A_1 e A_2 são mutuamente exclusivos? Por quê?

b) Calcule $P(A_1 \cap B)$ e $P(A_2 \cap B)$

5.4.12 Em rebanhos bovino investigados em pequenas propriedades no interior de Goias, sabe-se que a probabilidade de um animal ter Rinotraqueíte Infecciosa Bovina (IBR) é 0,6, de ter e a Tristeza Parasitária Bovina (TPB) é 0,75 e de ter (IBR e TPB) é 0,5. Selecionado ao acaso um animal, determine:

a) A probabilidade de não ter IBR;

b) A probabilidade de ter IBR ou TPB.

5.4.13 Um novo teste para o diagnóstico precoce de uma doença infecciosa está em estudo. O teste foi aplicado a um conjunto de animais para os quais se conhecia à priori se estavam infectados ou não. Um teste positivo indica a detecção da infecção. Dos resultados da aplicação do teste obteve-se o seguinte quadro:

Característica	Resultado do Teste	
	Teste deu Positivo (P)	Teste deu Negativo (N)
Indivíduo Infectado (I)	68	83
Indivíduos Saudáveis (S)	52	97

Determine:

a) A sensibilidade do teste (probabilidade do teste dar positivo, sabendo que os indivíduos estavam infectados);

b) A especificidade do teste (probabilidade do teste dar negativo sabendo que os indivíduos eram saudáveis);

c) Os falsos positivos (probabilidade do teste dar positivo sabendo que os indivíduos eram saudáveis);

d) Os falsos negativos (probabilidade do teste dar negativo sabendo que os indivíduos estavam infectados).

5.4.14 Num estudo de patologias esqueléticas traumáticas em 280 cavalos de corrida verificou-se o seguinte:

Lesões	Sexo	
	Fêmea	Macho
Articulares	93	86
Tendinosas	26	29
Musculares	19	13
Fraturas	2	12

a) Escolhido um animal ao acaso qual a probabilidade de ser fêmea?

b) Sabendo-se que um animal teve uma lesão articular, qual a probabilidade de ser macho?

c) Considere os acontecimentos "ser macho"e “teve fratura”. Verifique a independência destes dois eventos.

5.4.15 Três máquinas fabricam moldes não-ferrosos (anti-ferrugem). A máquina A produz 1% de defeituosos, a máquina B 2%, e a máquina C 5%. Cada maquina é responsável por 1/3 da produção total. Um inspetor examina um molde e constata que está perfeito. Calcule a probabilidade de ele ter sido produzido por cada uma das máquinas.

5.4.16 Um fazendeiro estima que, quando uma pessoa experiente planta árvores, 90% sobrevivem, mas quando um novato as planta, apenas 50% sobrevivem. Se uma árvore plantada não sobrevive, determine a probabilidade de ela ter sido plantada por um novato, sabendo-se que 2/3 das árvores são plantadas por novatos?

5.4.17 Um pesquisador desenvolve sementes de quatro tipos de plantas, P1, P2, P3 e P4. Plantados canteiros-pilotos destas sementes, a probabilidade de todas germinarem é de 40% para P1, 30% para P2, 25% para P3 e 50% para P4.

a) Escolhido um canteiro ao acaso, verificou-se que nem todas as sementes haviam germinado. Calcule a probabilidade de que o canteiro escolhido seja o de sementes de P3.

b) Escolhido um canteiro ao acaso, verificou-se que todas as sementes haviam germinado. Calcule a probabilidade de que o canteiro escolhido seja o de sementes de P1.

5.4.18 Num mercado, três corretoras A, B, C são responsáveis por 20%, 50% e 30% do volume total de contratos negociados, respectivamente. Do volume de cada corretora, 20%, 5% e 2%, respectivamente, são contratos futuros em dólares. Um contrato é escolhido ao acaso e este é futuro em dólares. Qual é a probabilidade de ter sido negociado pela corretora A? E pela corretora C?

5.4.19 (Navidi, W. Probabilidade e Estatísitca para Ciências Exatas) Para um determinado tipo de placa de circuito impresso, 50% não possuem defeitos, 25% possuem um defeito, 12% possuem dois defeitos, 8% contêm três defeitos e os 5% restantes possui quatro defeitos. Seja Y, a $\nu.a.$ que indica o número de defeitos em uma placa escolhida aleatoriamente. A $\nu.a.$ Y é discreta ou contínua? Por que? Apresente a distribuição de probabilidade de Y e encontre seu desvio padrão.

5.4.20 A s probabilidades de um investidor vender uma propriedade com um lucro de R$ 10.500,00, de R$ 5.500,00, de R$ 3.000,00 ou com prejuízo de R$ 4.000,00 são 0,22, 0,36, 0,28 e 0,14, respectivamente. Qual é o lucro esperado (esperança) do investidor?

5.4.21 O *Forbes 1993 Subscriber Study* e o *Fortune 1994 National Subscriber Portrait* reportaram as seguintes distribuições de probabilidades para o número de veículos por família de assinante.

X	0	1	2	3	4
$P(X = x_i)$ - *Forbes*	0,045	0,23	0,449	0,169	0,107
$P(X = x_i)$ - *Fortune*	0,028	0,165	0,489	0,185	0,133

a) Qual a média de veículos por família para cada grupo de assinantes?

b) Calcule F(2) para as duas revistas;

c) Qual é a variância do número de veículos por família para cada grupo de assinantes?

d) Usando suas respostas aos itens a) e b), que comparações você pode fazer sobre o número de veículos por família para os assinantes da *Forbes* e da *Fortune*?

5.4.22 Um negociante espera vender um automóvel até sexta-feira. A expectativa de que venda na segunda-feira é de 50%. Na terça-feira é de 30%, na quarta feira é de 10%, na quinta-feira é de 5% e na sexta-feira é de 5%. Seu lucro é de R$ 3.000,00 se vender na segunda-feira e diminui 40% a cada dia.

a) Calcule o valor esperado de lucro deste negociante nesta venda;

b) Calcule sua variância e seu desvio padrão.

5.4.23 Uma máquina fabrica placas de papelão que podem apresentar nenhum, um, dois, três ou quatro defeitos, com probabilidade 90%, 5%, 3%, 1% e 1%, respectivamente. O preço de venda de uma placa perfeita é de R$ 10,00 e á medida que apresente defeitos, o preço cai 50% para cada defeito apresentado. Qual é o preço médio de vendas destas placas?

5.4.24 Mostre que:

$$\sum_{i=1}^{k}(x_i-\mu)^2P(X=x_i)=E(X^2)-[E(X)]^2$$

5.4.25 Dada a variável aleatória

X	−1	2	5	8
$P(X=x_i)$	0,2	0,3	0,4	0,1

Calcule a média e o desvio padrão da variável $Y=\frac{4}{3}X-3$

5.4.26 Um florista faz estoque de uma flor de curta duração que lhe custa \$0,50 e que ele vende a \$1,50 no primeiro dia em que a flor está na loja. Toda flor que não é vendida nesse primeiro dia não serve mais e é jogada fora. Seja X a variável aleatória que denota o número de flores que os fregueses compram em um dia casualmente escolhido. O florista descobriu que a função de probabilidade de X é dada pela tabela abaixo:

x	0	1	2	3
p(x)	0,1	0,4	0,3	0,2

Quantas flores deveria o florista ter em estoque a fim de maximizar a média (valor esperado) do seu lucro?

5.5 Modelos Discretos de Probabilidade

Distribuição Uniforme Discreta

É considerado o caso mais simples de variável aleatória discreta, em que cada valor possível ocorre com a mesma probabilidade.

A variável aleatória discreta X, assumindo os valores $x_1, x_2, \ldots, x_k$, tem distribuição uniforme discreta se, e somente se,

$$P(X=x_i)=p=\frac{1}{k}$$

para todo $i=1, 2, \ldots, k$.

A média e a variância estão abaixo:

$$E(X)=\frac{1}{k}\sum_{i=1}^{k}x_i$$

$$V(X) = \frac{1}{k}\left\{\sum_{i=1}^{k} x_i^2 - \frac{\left(\sum_{i=1}^{k} x_i^2\right)}{k}\right\}$$

Exemplo 5.5.1 Seja X uma *v.a.* que indica o "número de pontos marcados na face superior de um dado", quando ele é lançado. Calcule a $P(X = 2)$, a $E(X)$ e $V(X)$.

$$P(X = 2) = \frac{1}{6}$$

$$E(X) = \frac{1}{6} \cdot [1 + 2 + 3 + 4 + 5 + 6] = \frac{21}{6} = 3,5$$

$$V(X)\frac{1}{6} \cdot \left[(1 + 4 + \cdots + 36) - \frac{(21)^2}{6}\right] = \frac{35}{12} = 2,9$$

Distribuição de Bernoulli

Em um experimento que é possível definir uma *v.a.* X, que assume os valores 1, se ocorrer sucesso, e 0 se ocorrer fracasso, podemos chamar esta variável de variável aleatória de Bernoulli. Temos então que:

$$P(X = 0) = 1 - p$$

$$P(X = 1) = p$$

Sua média e sua variância são dadas por:

$$E(X) = p;$$

$$V(X) = p - p^2 = p(1 - p)$$

Observação 5.5.1 Experimentos que resultam numa *v.a.* de Bernoulli são chamados de ensaios ou provas de Bernoulli.

Exemplo 5.5.2 Suponha que um dado é lançado: ou ocorre face 5 ou não. Calcule $P(X = 0)$ e $E(X)$.

$$P(X = 1) = p = \frac{1}{6}$$

$$P(X = 0) = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}$$

$$E(X) = p = \frac{1}{6}$$

Distribuição Binomial

Considere um processo consistindo de uma seqüência de n provas independentes. Por provas independentes queremos dizer que o resultado de cada prova não depende, de qualquer maneira, dos resultados das provas anteriores. Quando o resultado de cada prova é ou "sucesso"ou "fracasso", as provas são chamadas **provas de Bernoulli**. Se a probabilidade de "sucesso"em qualquer prova – digamos, p – é constante, então o n^o de "sucessos"x em n provas de Bernoulli independentes tem distribuição Binomial com parâmetros n e p, definida como segue:

A distribuição Binomial com parâmetros $n \geqslant 0$ e $0 < p < 1$ é:

$$P(X = x) = \binom{n}{x} p^x \cdot (1-p)^{n-x},\ x = 0,\ 1,\ \ldots,\ n$$

Sua média e sua variância são:

$$E(X) = n \cdot p$$

$$V(X) = n \cdot p \cdot (1-p)$$

Exemplo 5.5.3 Se 20% dos parafusos produzidos por uma máquina são defeituosos, determine a probabilidade de que, em 4 parafusos escolhidos aleatoriamente, apenas 1 seja defeituoso.

X = números de parafusos defeituosos produzidos por uma máquina

$$P(X = 1) = \binom{4}{1} 0,2^1 \cdot (1-0,2)^{4-1} = 0,4096$$

Distribuição de Poisson

A distribuição de Poisson é uma distribuição discreta de probabilidade aplicável a ocorrências de um evento em um intervalo especificado. A variável aleatória X é o n^o de ocorrências do evento em um intervalo. O intervalo pode ser o tempo, a distância, a área, o volume ou outra unidade análoga. A probabilidade de o evento ocorrer x vezes em um intervalo é dada por:

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!},\ x = 0,\ 1,\ \ldots \text{ onde } e \approx 2,71828$$

A distribuição de Poisson exige:

i) Que a $v.a.$ X seja o n^o de ocorrências de um evento em um intervalo;

ii) Que as ocorrências sejam aleatórias;

iii) Que as ocorrências sejam independentes umas das outras;

iv) Que as ocorrências sejam distribuídas uniformemente sobre o intervalo considerado.

Sua média e sua variância são:

$$E(X) = \lambda \qquad V(X) = \lambda$$

Exemplo 5.5.4 Suponha que dados históricos mostram que o número médio de chegadas a uma caixa automática (tipo drive-thru) de um banco durante um período de 15 minutos nas manhãs de fins de semana é de 10 carros. Calcule a probabilidade de exatamente cinco chegas em 15 minutos neste fim de semana.

$X =$ O número de carros que chegam em qualquer período de 15 minutos

$$P(X = 5) = \frac{e^{-10} \cdot 10^5}{5!} = 0,0378$$

A distribuição de Poisson difere da Binomial em dois aspectos importantes:

i) A distribuição Binomial é afetada pelo tamanho amostral n e pela probabilidade p, enquanto a distribuição de Poisson é afetada apenas pelo parâmetro λ;

ii) Em uma distribuição Binomial, os valores possíveis da variável aleatória X são $0, 1, 2, \ldots, n$, enquanto em uma distribuição de Poisson os valores possíveis de X são $0, 1, \ldots$ sem limite superior.

Distribuição Hipergeométrica

Considere uma população finita composta de N itens. Algum n^o, digamos $k (k \leqslant N)$, destes itens pertence a uma determinada classe de interesse. Uma amostra aleatória de n itens é retirada da população sem reposição e o n^o de itens na amostra que se situa na classe de interesse – digamos, x – é observado. Então X é uma variável aleatória com distribuição Hipergeométrica definida como segue:

$$P(X = x) = \frac{\binom{k}{x} \cdot \binom{N-k}{n-x}}{\binom{N}{n}}$$

A média e a variância da distribuição são:

$$E(X) = \frac{nk}{N} \qquad V(X) = \frac{nk}{N}\left(1 - \frac{k}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Exemplo 5.5.5 Uma batelada de peças contém 100 peças de um fornecedor local de tubos e 200 peças de um fornecedor de tubos de um estado vizinho. Se quatro peças forem selecionadas, ao acaso e sem reposição, qual será a probabilidade de que elas sejam todas provenientes do fornecedor local?

$X =$ número de peças na amostra do fornecedor local.

$$P(X=4)=\frac{\binom{100}{4}\cdot\binom{200}{0}}{\binom{300}{4}}=0,0119$$

Distribuição Geométrica

Em certas situações em que estamos interessados na probabilidade de o primeiro sucesso ocorrer na *x-ésima* prova, essa probabilidade é dada pela distribuição Geométrica.

Sendo p e $(1-p)$ as probabilidades de "sucesso"e "fracasso", respectivamente, e considerando que o primeiro sucesso há de ser precedido por $x-1$ fracassos, temos que:

$$P(X=x)=p\cdot(1-p)^{x-1}.$$

A média e a variância são:

$$E(X)=\frac{1}{p}\qquad\qquad V(X)=\frac{1-p}{p^2}.$$

Exemplo 5.5.6 A probabilidade de uma pastilha conter uma partícula grande de contaminação é de $0,01$. Se for considerado que as pastilhas sejam independentes, qual será a probabilidade de que exatamente 125 pastilhas necessitem ser analisadas antes que uma partícula grande seja detectada?

$X=$ O número de amostras analisadas até que uma partícula grande seja detectada.

$$P(X=126)=0,01\cdot(1-0,01)^{125}=0,0028$$

Distribuição Multinomial

Suponha que os eventos A_1, A_2, ..., A_k sejam mutuamente exclusivos e ocorram com probabilidades p_1, p_2, ..., p_k onde $p_1+p_2+\ldots+p_k=1$. Se X_1, X_2, ..., X_k são respectivas variáveis aleatórias resultando no número de vezes que A_1, A_2, ..., A_k ocorrem em um total de n ensaios, de modo que $X_1+X_2+\cdots+X_k=n$, então:

$$P(X_1=n_1,\,X_2=n_2,\,\ldots,\,X_k=n_k)=\frac{n!}{n_1!n_2!\ldots n_k!}\cdot p_1^{n_1}p_2^{n_2}\ldots p_k^{n_k}$$

onde $n_1+n_2+\cdots+n_k=n$, é a função de probabilidade conjunta das variáveis aleatórias X_1, X_2, ..., X_k.

Os números esperados de vezes que A_1, A_2, ..., A_k ocorrem em n ensaios são, respectivamente, np_1, np_2, ..., np_k, isto é,

$$E(X_1)=np_1,\,E(X_2)=np_2,\,\ldots,\,E(X_k)=np_k.$$

Exemplo 5.5.7 Se um dado honesto for lançado 12 vezes, a probabilidade de obter-se 1, 2, 3, 4, 5 e 6 pontos exatamente duas vezes, cada um, é:

$$P(X_1 = 2, X_2 = 2, \ldots, X_6 = 2) = \frac{2!}{2!2!2!2!2!2!} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 = 0,00344$$

Aproximação da Distribuição Binomial pela Poisson

Na distribuição Binomial, se $n \to \infty$ enquanto a probabilidade de p de ocorrência de um evento é próxima de zero, de modo que $(1-p)$ é próximo de 1, o evento é dito um *evento raro*. Na prática, vamos considerar um evento como raro se o número de ensaios é pelo menos 50 ($n \geqslant 50$) $np < 5$. Em tais casos, a distribuição binomial é muito bem aproximada pela distribuição de Poisson com $\lambda = np$.

Exemplo 5.5.8 Das ferramentas produzidas por um certo processo de fabricação, 10% apresentam algum defeito. Encontre a probabilidade de, em uma amostra de 10 ferramentas escolhidas aleatoriamente, exatamente 2 sejam defeituosas, usando a aproximação de Poisson para a distribuição Binomial.

Temos $\lambda = np = 10 \cdot 0,1 = 1$.

Então, de acordo com a distribuição de Poisson,

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1^2 e^{-1}}{2!} = 0,1839$$

5.6 Exercícios

5.6.1 Em momentos de pico, a chegada de aviões a um aeroporto se dá com uma média de 1 avião a cada 3 minutos.

a) Determine a probabilidade de duas chegadas em três minutos durante o horário de pico;

b) Determine a probabilidade de no máximo uma chegada em três minutos.

5.6.2 Uma grande empresa patrocina um programa de investimentos em ações para seus empregados. Sabendo que a probabilidade de que um empregado participe desse programa é de 0,4 e sendo que 10 empregados foram escolhidos aleatoriamente. Utilizando a distribuição Binomial, calcule:

a) A probabilidade de que exatamente 5 destes empregados participem deste programa;

b) A esperança e a variância.

5.6.3 Um funcionário dos correios deve remeter, por via aérea, para a Europa, seis pacotes de um lote de 15. Acontece que ele mistura todos e carimba de "via aérea"aleatoriamente seis pacotes.

a) Qual a probabilidade de que apenas três dos pacotes que devem ir via aérea sigam realmente por essa via?

b) Encontre a esperança e a variância.

5.6.4 Dos 16 caminhões de entrega de uma loja de departamentos, cinco emitem excesso de poluentes. Selecionados aleatoriamente para inspeção oito dos 16 caminhões.

a) Qual é a probabilidade de essa amostra incluir no máximo 2 dos caminhões que emitem excesso de poluentes?

b) Qual é a probabilidade de essa amostra incluir 4 ou 5 caminhões que emitem excesso de poluentes?

c) Encontre a esperança e a variância.

5.6.5 Uma confecção de roupa infantil suspeita que 30% de sua produção apresentam algum defeito. Se tal suspeita é correta, determine a probabilidade de que, numa amostra de quatro peças, sejam encontradas:

a) No mínimo três peças com defeitos;

b) Menos de três peças boas;

c) Encontre a esperança e o seu desvio padrão.

5.6.6 Um estudo cuidadoso de uma fita magnética de dados de computadores mostra uma incidência de 2 defeitos para cada 500 pés de fita. Determine:

a) A probabilidade de mais de um defeito em 500 pés de fita selecionados aleatoriamente;

b) A esperança matemática e a variância dessa variável aleatória e seu desvio padrão.

5.6.7 Um embarque de 10 itens tem duas unidades com defeitos e oito unidades sem defeito. Na inspeção de embarque, uma amostra de unidades será selecionada e testada. Se uma unidade com defeito for encontrada, o embarque de 10 unidades será rejeitado. Se uma amostra de três itens é selecionada, qual é a probabilidade de que o embarque seja rejeitado?

5.6.8 Uma caixa contém 5 bolas vermelhas, 4 bolas brancas e 3 bolas azuis. Uma bola é selecionada ao acaso da caixa, sua cor é anotada, e então a bola é recolocada na caixa. Encontre a probabilidade de que, em 6 bolas selecionadas desta maneira, 3 sejam vermelhas, 2 sejam brancas e uma seja azul.

5.6.9 Um vendedor programa seis visitas e acredita que a probabilidade de ele ser recebido pelo encarregado de compras das empresas visitadas é de 80%.

a) Qual a probabilidade de ele completar pelo menos quatro visitas?

b) Qual é a probabilidade de ele ser recebido por todos os encarregados de compra?

c) Se ele acredita que completando uma visita suas despesas do dia estão cobertas, qual é a probabilidade de ele ter prejuízo nesse dia?

5.6.10 A taxa de chegada de clientes em uma agência bancária é de quatro clientes por minuto. Determine a probabilidade de chegarem mais que 14 clientes nos próximos dois minutos.

5.6.11 Na manufatura de certo artigo, é sabido que um entre dez dos artigos é defeituoso. Qual a probabilidade de que uma amostra casual de tamanho quatro contenha:

a) Nenhum defeituoso?

b) Não mais do que dois defeituosos?

5.6.12 Um contador eletrônico de bactérias registra, em média, cinco bactérias por cm^3 de um líquido. Determine:

a) O desvio padrão do número de bactérias por cm^3;

b) A probabilidade de que pelo menos duas bactérias ocorrem num volume líquido de 1 cm^3?

5.6.13 A probabilidade de que haja alguma falha no lançamento de uma nave espacial é 10%. Qual é a probabilidade de que para lançar a nave seja necessário:

a) Duas tentativas?

b) No máximo 3 tentativas.

5.6.14 Suponha que a probabilidade de um componente de computador ser defeituoso é de $0,2$. Numa mesa de testes, uma batelada é posta à prova, um a um. Determine a probabilidade do primeiro defeito encontrado ocorrer no sétimo componente testado.

5.6.15 A probabilidade de que um bit transmitido através de um canal de transmissão digital seja recebido com erro é $0,1$. Suponha que as transmissões são eventos independentes. Calcule a probabilidade de o primeiro erro ocorra no quinto bit transmitido.

5.6.16 Geólogos estimam o tempo decorrido desde o resfriamento mais recente de um mineral contando o número de vestígios de fissões de urânio na superfície do mineral. Um determinado tipo de mineral tem uma idade tal que deve ter uma média de 6 vestígios por cm^2 da área superficial. Considere que o número de vestígios em uma área segue um a distribuição de Poisson. Seja X o número de vestígios contado em 1 cm^2 de área superficial. Determine:

a) $P(X = 7)$

b) $P(X \geqslant 3)$

5.6.17 Suponha que determinado medicamento, usado para o diagnóstico precoce da gravidez, é capaz de confirmar casos positivos em 90% de mulheres muito jovens . Nestas condições, qual é a probabilidade de, em uma amostra de 9 gestantes muito jovens que fizeram uso deste medicamento:

a) Duas delas não terem confirmado precocemente a gravidez?

b) No máximo três delas terem confirmado precocemente a gravidez?

5.6.18 Um experimento de genética envolve 6 genótipos mutuamente excludentes identificados por A, B, C, D, E e F, todos igualmente prováveis. Testados 20 indivíduos, determine a probabilidade de obter exatamente: 5 A; 4 B; 3 C; 2 D; 3 E; 3 F.

5.6.19 Os seguintes eventos podem ocorrer com um pacote enviado pelo correio: chegar em perfeito estado, chegar danificado ou perder-se pelo caminho. As probabilidades desses eventos são, respectivamente $0,7$, $0,2$ e $0,1$. Foram enviados recentemente 10 pacotes pelo correio. Qual a probabilidade de 6 chegarem corretamente ao destino, 2 serem perdidos e os outros 2 avariados?

5.6.20 Sabe-se pela experiência, que 2% das chamadas recebidas por uma mesa telefônica são para números errados. Com a aproximação de Poisson da distribuição Binomial, determine a probabilidade de três dentre 200 chamadas recebidas pela mesa serem para número errado.

5.6.21 $0,6\%$ dos detonadores fornecidos a um arsenal são defeituosos, utilize a aproximação de Poisson para distribuição Binomial para determinar a probabilidade de que, em uma amostra aleatória de 500 detonadores, quatro sejam defeituosos.

5.6.22 Os registros mostram que há uma probabilidade de $0,0012$ de uma pessoa se intoxicar na lanchonete de um parque de diversões. Com a aproximação de Poisson para a Binomial, determine a probabilidade de que, de 1000 pessoas que visitam o parque, no máximo duas se intoxicarem.

5.6.23 Em certa cidade, $3,2\%$ de todos os motoristas habitados se envolvem em, ao menos, um acidente de carro em uma ano. Com o auxílio da aproximação de Poisson para a distribuição Binomial , determine a probabilidade de que, dentre 200 motoristas escolhidos aleatoriamente nessa cidade:

a) Exatamente seis se envolvam em ao menos um acidente em um ano;

b) No máximo oito se envolvam em ao menos um acidente em um ano;

c) Cinco ou mais se envolvam em ao menos um acidente em um ano.

5.6.24 Os defeitos em determinada máquina ocorrem aproximadamente na mesma frequência. Dependendo do tipo de defeito, o técnico leva 1, 2, 3, 4 ou 5 horas para consertar a máquina

a) Descreva o modelo probabilístico apropriado para representar a duração do tempo de reparo da máquina;

b) Qual é o tempo médio de reparo desta máquina? E o desvio-padrão deste tempo de reparo?

c) São 15 horas e acaba de ser entregue uma máquina para reparo. A jornada normal de trabalho do técnico termina às 17 horas. Qual é a probabilidade de que o técnico não precise fazer hora extra para terminar o conserto desta máquina?

5.6.25 Numa placa de microscópio, com área dividida em quadrantes de $1\ mm^2$, encontram-se em média cinco colônias por mm^2. Considerando que as colônias distribuem-se aleatoriamente na placa, encontre:

a) A probabilidade de um quadrante ter exatamente uma colônia;

b) A probabilidade de se encontrar pelo menos duas colônias num quadrante;

c) A probabilidade de se encontrar oito colônias em $2\ mm^2$.

5.6.26 Um produtor de sementes vende pacotes com 20 sementes cada. Os pacotes que apresentarem mais de uma semenete sem germinar serão indenizados. A probabilidade de uma semente germinar é de $0,98$.

a) Calcule a média e avariância da variável aleatória "número de sementes que não germinam por pacote";

b) Qual é a probabilidade de um pacote não ser indenizado?

c) Se o produtor vende 1.000 pacotes, qual é o número esperado de pacotes indenizados?

5.6.27 Um agricultor planta seis sementes escolhidas aleatoriamente de uma caixa com cinco sementes de tulipa e quatro de crisântemo. Qual é a probabilidade de ele plantar:

a) Duas sementes de crisântemo e quatro de tulipa?

b) No mínimo três sementes de tulipa?

5.7 Variáveis Aleatórias Contínuas

Uma função X, definida sobre o espaço amostral e assumindo valores num intervalo de números reais, é dita uma *variável aleatória contínua.*

A característica principal de uma **v.a.** contínua é que, sendo resultado de uma mensuração, o seu valor pode ser pensado como pertencendo a um intervalo ao redor do valor efetivamente observado, onde a função densidade de probabilidade (**f.d.p.**) $f(x)$ tem as propriedades

i) $f(x) \geqslant 0$;

ii) $\int_{-\infty}^{\infty} f(x)dx = 1.$

Da definição e das propriedades acima segue que, se X é uma variável aleatória contínua, então a probabilidade de X assumir um valor particular é zero, e a probabilidade de um intervalo, isto é, de X assumir um valor entre dois diferentes valores, digamos, a e b, é dada por

$$P(a \leqslant X \leqslant b) = \int_{b}^{a} f(x)dx,$$

que é igual área sob $f(x)$ de a e b para qualquer a e b.

Exemplo 5.7.1 Faça a variável aleatória contínua X denotar o diâmetro de um orifício perfurado em uma placa com um componente metálico. O diâmetro que se quer atingir, o chamado diâmetro alvo, é $12,5$ milímetros. A maioria dos distúrbios aleatórios no processo resulta em diâmetros maiores. Dados históricos mostram que a distribuição de X pode ser modelada por uma função densidade de probabilidade $f(x) = 20e^{-20(x-12,5)}$, $x \geqslant 12,5$. Se uma peça com um diâmetro maior que $12,60$ milímetros for descartada, qual será a proporção de peças descartadas?

$$P(X > 12,60) = \int_{12,60}^{\infty} f(x)dx = \int_{12,60}^{\infty} f(x) = 20e^{-20(x-12,5)}$$
$$= -e^{-20(x-12,5)} \Big|_{12,6}^{\infty} = 0,135$$

Sua função de distribuição acumulada $F(x)$ pode ser representada como

$$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(u)du, \ (-\infty < x < \infty).$$

Vemos então que $0 \leqslant F(x) \leqslant 1$, para todo x real; além disso, $F(x)$ não é decrescente e possui as seguintes propriedades:

i) $\lim_{x\to\infty} F(x) = 0$;

ii) $\lim_{x\to\infty} F(x) = 1$.

Para as variáveis aleatórias contínuas, o seguinte resultado é importante:

$$F'(x) = \frac{dF(x)}{dx} = f(x)$$

Exemplo 5.7.2 Suponha que

$$F(x) = \begin{cases} 0 & , \quad x < 0 \\ 1 - e^{-x} & , \quad x \geqslant 0 \end{cases}$$

seja a função distribuição acumulada de uma v.a. X. Encontre a f.d.p. de X.
$\frac{dF(x)}{dx} = e^{-x}$, então a f.d.p. de X é dada por:

$$f(x) = \begin{cases} 0 & , \quad x < 0 \\ e^{-x} & , \quad x \geqslant 0 \end{cases}$$

Outro resultado importante: Se a e b forem dois números reais quaisquer,

$$P(a < X \leqslant b) = F(b) - F(a)$$

Esse resultado não será afetado se incluirmos ou não os extremos a e b na desigualdade entre parênteses.

A esperança e a variância de uma variável aletória X contínua é encontrada da seguinte forma:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

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

sendo que:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx$$

5.8 Exercícios

5.8.1 O diâmetro X de um cabo elétrico supõe-se ser uma variável aleatória contínua X, com f.d.p.:

$$f(x) = \begin{cases} 6x(1-x) & , \quad 0 \leqslant x \leqslant 1 \\ 0 & , \quad \text{caso contrário} \end{cases}$$

a) Verifique que essa expressão é uma f.d.p para X;

b) Calcule $P\left(X > \frac{1}{5}\right)$;

c) Encontre F(x) e o desvio padrão de X;

d) Calcule a seguinte probabilidade condicional $P\left(X \leqslant \frac{1}{2} / \frac{1}{3} < X < \frac{2}{3}\right)$.

5.8.2 A *v.a.* contínua X tem a seguinte f.d.p.:

$$f(t) = \begin{cases} 3x^2 & , \quad -1 \leqslant x \leqslant 0 \\ 0 & , \quad \text{caso contrário} \end{cases}$$

Calcule $E(X)$ e $V(X)$.

5.8.3 Suponha que a *v.a.* X tem a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} \dfrac{4x(9-x^2)}{81} & , \quad 0 \leqslant x \leqslant 3 \\ 0 & , \quad \text{caso contrário} \end{cases}$$

a) Encontre $F(x)$;

b) Calcule $P(X > 2,1)$;

c) Encontre o desvio padrão da variável.

5.8.4 Seja X com a seguinte f.d.p

$$f(x) = \begin{cases} c(1-x^2) & , \quad \text{se } -1 \leqslant x \leqslant 1 \\ 0 & , \quad \text{caso contrário} \end{cases}$$

a) Encontre o valor de c;

b) Calcule a média e a variância de X.

5.8.5 A demanda diária de arroz num supermercado, em centenas de quilos, é uma *v.a.* com f.d.p

$$f(x) = \begin{cases} \dfrac{2}{3x} & , \quad \text{se } 0 \leqslant x < 1 \\ \dfrac{-x}{3} + 1 & , \quad \text{se } 1 \leqslant x < 3 \\ 0 & , \quad \text{se } x < 0 \text{ ou } x > 3 \end{cases}$$

Qual a probabilidade de se vender mais do que $150Kg$, num dia escolhido ao acaso?

5.8.6 A seguir temos a seguinte função:

$$f(x) = \begin{cases} cx^2 & , \quad \text{se } 0 < x < 3 \\ 0 & , \quad \text{caso contrário} \end{cases}$$

a) Calcule o valor da constante c, sabendo que $f(x)$ é uma função densidade de probabilidade;

b) Calcule $P(1 < X < 2)$.

5.8.7 Encontre a função de distribuição acumulada da variável aleatória do Exercício 5.8.6.

5.8.8 A função de distribuição acumulada de uma variável aleatória X é:

$$F(x) = \begin{cases} 1 - e^{-2x} & , \quad x \geqslant 0 \\ 0 & , \quad x < 0 \end{cases}$$

Encontre:

a) A f.d.p. de X;

b) A probabilidade de $X > 2$;

c) A probabilidade de $-3 < X \leqslant 4$

5.8.9 Uma variável aleatória tem a seguinte f.d.p.:

$$f(x) = \begin{cases} cx^2 & , \quad \text{se } 1 \leqslant x \leqslant 2 \\ cx & , \quad \text{se } 1 < x < 3 \\ 0 & , \quad \text{caso contrário} \end{cases}$$

Determine:

a) A constante c;

b) $P(X > 2)$;

c) $P\left(\frac{1}{2} < X < \frac{3}{2}\right)$.

Principais Modelos Contínuos

Nesta seção, apresentamos os principais modelos para variáveis aleatórias contínuas. Os diversos modelos serão caracterizados pela sua função densidade probabilidade e, em vários casos, apresentamos também a função de distribuição.

Distribuição Uniforme

Dizemos que X segue o modelo Uniforme, no intervalo $[a, b] \subset \mathbb{R}$, se todos os sub-intervalos de $[a, b]$ com mesmo comprimento tiverem a mesma probabilidade. Sua f.d.p é dada por

$$f(x) = \begin{cases} \dfrac{1}{b-a} & , \quad \text{se } a \leqslant x \leqslant b \\ 0 & , \quad \text{caso contrário} \end{cases}$$

A função distribuição acumulada da Uniforme é facilmente encontrada e é dada por:

$$F(x) = \begin{cases} 0 & , \quad \text{se } x < a \\ \dfrac{x-a}{b-a} & , \quad \text{se } a \leqslant x < b \\ 1 & , \quad \text{se } x \geqslant b \end{cases}$$

Sua esperança e variância são dadas da seguinte maneira:

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

Distribuição Exponencial

Um modelo com aplicação em diversas áreas de engenharia e matemática é o modelo Exponencial. Tempo de vida de equipamentos, intervalos entre chegadas de mensagens eletrônicas ou de chamadas telefônicas a uma central, são algumas das quantidades que têm sido bem modeladas com essa distribuição.

A variável aleatória X segue o modelo Exponencial de parâmetro λ, $\lambda > 0$, se tiver densidade dada por:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \quad \text{se } x \geqslant 0 \\ 0 & , \quad \text{caso contrário} \end{cases}$$

O parâmetro λ indica a taxa de ocorrência por unidade de medida, que pode ser tempo, distância ou volume, entre outras.

A expressão da função de distribuição acumulada é dada por:

$$F(x) = \begin{cases} 0 & , \quad \text{se } x < 0 \\ (1 - e^{-\lambda x}) & , \quad \text{se } x \geqslant 0 \end{cases}$$

Exemplo 5.8.1 Um serviço de atendimento ao consumidor recebe chamadas telefônicas num intervalo de tempo, em horas, que segue uma distribuição Exponencial com $\lambda = 5$. O parâmetro λ pode ser interpretado como sendo uma taxa de 5 chamadas por hora. Calcule a probabilidade de um intervalo entre chegadas ter duração inferior a 30 minutos.

$$P\left(X < \frac{1}{2}\right) = P\left(X \leqslant \frac{1}{2}\right) = F\left(\frac{1}{2}\right) = 1 - e^{-\frac{5}{2}} = 0{,}918$$

A esperança e a variância da distribuição exponencial são dadas da seguinte forma:

$$E(X) = \frac{1}{\lambda} \qquad \text{e} \qquad V(X) = \frac{1}{\lambda^2}$$

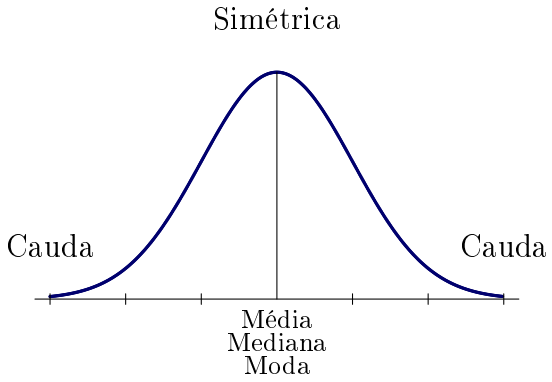
Distribuição Normal

A distribuição Normal é, provavelmente, a mais importante distribuição de probabilidade, tanto na teoria quanto na prática da estatística. Se X é uma variável aleatória normal, então a distribuição de probabilidade de X é definida como se segue:

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

A equação da curva Normal é especificada usando 2 parâmetros: a **média** populacional μ, e o **desvio padrão** populacional σ, ou equivalentemente a **variância** populacional σ^2. Denotamos $N(\mu, \sigma^2)$ à curva Normal com média μ e variância σ^2. A média refere-se ao centro da distribuição e o desvio padrão ao espalhamento de curva.

A distribuição normal é simétrica em torno da média, o que implica que e média, a mediana e a moda são todas coincidentes. Esta distribuição tem uma aparência visual de uma curva simétrica, unimodal, em forma de sino, como apresentado abaixo. Ela é assintótica, ou seja, a curva aproxima-se cada vez mais do eixo x, mas nunca o toca efetivamente.

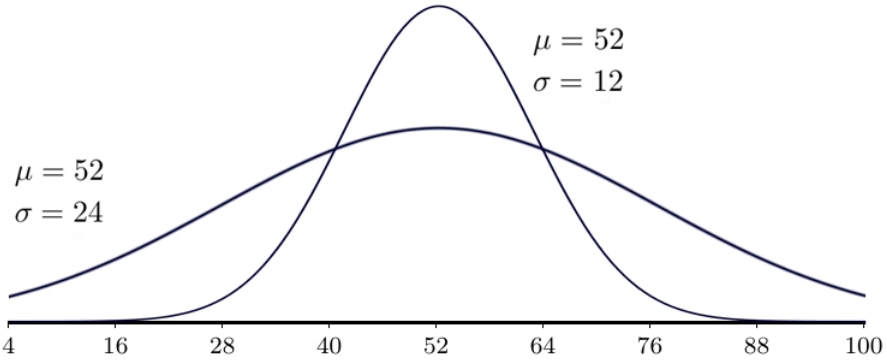


A área sob a curva normal é 1. Então, para quaisquer dois valores específicos podemos determinar a probabilidade da área sob a curva entre esses dois valores. Para a distribuição Normal, a probabilidade de valores caindo dentro de um, dois, ou três desvios padrão da média é:

Intervalo	Probabilidade
$\mu \pm 1\sigma$	68,27%
$\mu \pm 2\sigma$	95,45%
$\mu \pm 3\sigma$	99,73%

Podemos também ter distribuições normais com o mesmo desvio padrão, mas com distintas médias ou com médias e desvios padrões distintos. Na realidade a distribuição normal é um nome genérico para definir uma família de infinitas distribuições normais particulares, cada uma com os seus valores específicos de média e desvio padrão. O que caracteriza, portanto, e diferencia uma distribuição normal de outra são os valores destes dois parâmetros, como mostra a Figura 5.8.1:

Figura 5.8.1: Duas Distribuições Normais com mesma média, mas desvios padrões diferentes



No cálculo de probabilidades para variáveis contínuas, devemos resolver a integral da função densidade no intervalo de interesse, isto é:

$$P(a \leqslant X \leqslant b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} dx$$

Entretanto, a integral acima só pode ser resolvida de modo aproximado e por métodos numéricos. Por essa razão as probabilidades pra o modelo Normal são calculadas com o auxílio de

tabelas. Para se evitar a multiplicação desnecessária de tabelas para cada par de valores (μ, σ^2), utiliza-se uma transformação que conduz sempre ao cálculo de probabilidades com uma variável de parâmetros $(0,1)$, isto é, média 0 e variância 1.

A Distribuição Normal Padrão

Considere $X \backsim N(\mu, \sigma^2)$ e defina uma nova variável Z, sendo:

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

A criação de uma nova variável aleatória por essa transformação é referida como padronização. A variável aleatória Z representa a distância de X a partir de sua média em termos dos desvios padrões. Pode-se ainda verificar que essa transformação não afeta a normalidade e, assim, a variável aleatória Z terá distribuição $N(0,1)$ e será denominada de Normal Padrão ou Normal Reduzida. Para determinar a probabilidade de $X \in [a, b]$, procedemos da seguinte forma:

$$P(a \leqslant X \leqslant b) = P\left(\frac{a - \mu}{\sigma} \leqslant Z \leqslant \frac{b - \mu}{\sigma}\right),$$

e, portanto, quaisquer que sejam os valores de μ e σ, utilizamos a Normal Padrão para obter probabilidades com a distribuição Normal.

Os valores para $P(0 \leqslant Z \leqslant z)$, são apresentados em tabela anexa. Com a simetria da densidade Normal podemos calcular valores de probabilidades em outros intervalos. Note que a simetria também implica que a probabilidade de estar acima (ou abaixo) de zero é $0,5$.

Exemplo 5.8.2 O faturamento mensal de uma loja segue uma distribuição normal com média R$ 20.000,00 e desvio padrão R$ 4.000,00. Calcule:

a) A probabilidade de que, num determinado mês, o faturamento esteja entre R$ 20.000,00 e R$ 25.000,00;

b) A probabilidade de que, num determinado mês, o faturamento esteja ente R$ 15.000,00 e R$ 20.000,00;

c) A probabilidade de que, num determinado mês, o faturamento esteja entre R$ 19.000,00 e R$ 25.000,00.

a)

$$P(20.000 < X < 25.000) = P\left(\frac{20.000 - 20.000}{4.000} < Z < \frac{25.000 - 20.000}{4.000}\right) = P(0 < Z < 1,25)$$
$$= 0,3944$$

b)

$$P(15.000 < X < 20.000) = P\left(\frac{15.000 - 20.000}{4.000} < Z < \frac{20.000 - 20.000}{4.000}\right) = P(-1,25 < Z < 0)$$
$$= 0,3944$$

c)

$$P(19.000 < X < 25.000) = P\left(\frac{19.000 - 20.000}{4.000} < Z < \frac{25.000 - 20.000}{4.000}\right) =$$
$$= P(-0,25 < Z < 0) + P(0 < Z < 1,25) =$$
$$= 0,0987 + 0,3944 = 0,4931$$

Aproximação da Distribuição Binomial pela Normal

Se n é grande e nem p e nem $(1-p)$ estão muito próximos de zero, a distribuição Binomial pode ser bastante aproximada por uma distribuição Normal através da variável aleatória padronizada dada por:

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

Aqui X é a variável aleatória dando o número de sucessos em n ensaios de Bernoulli e p é a probabilidade de sucesso. A aproximação se torna melhor com n crescendo e é exata no caso do limite. Na prática, a aproximação é muito boa se ambos, np e $n(1-p)$, são maiores do que 5. O fato de que a distribuição Binomial se aproxima da distribuição Normal pode ser descrita como:

$$\lim_{n\to\infty} P\left(a \leqslant \frac{X - np}{\sqrt{np(1-p)}} \leqslant b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{\frac{-u^2}{2}}\, du$$

Conceitualmente, dizemos que a variável aleatória padronizada $\dfrac{X - np}{\sqrt{np(1-p)}}$ é assintoticamente normal.

Observação 5.8.1 Ao empregar da aproximação normal à distribuição binomial, estaremos aproximando a distribuição de uma variável aleatória discreta com a distribuição de uma variável aleatória contínua. Por isso,, algum cuidado deve ser tomado com os pontos extremos dos intervalos considerados. Por exemplo, para uma variável aleatória contínua, $P(X = 3) = 0$, enquanto para uma variável aleatória discreta esta probabilidade pode ser não nula.

As seguintes **correções de continuidade** melhoram a aproximação:

a) $P(X = k) \simeq P\left(k - \frac{1}{2} \leqslant X \leqslant k + \frac{1}{2}\right)$;

b) $P(a \leqslant X \leqslant b) \simeq P\left(a - \frac{1}{2} \leqslant X \leqslant \frac{1}{2} + b\right)$.

Exemplo 5.8.3 Suponha-se que um sistema seja formado por 100 componentes, cada um dos quais tenha confiabilidade igual a $0,95$ (isto é, a probabilidade de que o componente funcione adequadamente durante um período especificado é igual a $0,95$). Se esses componentes funcionarem independentemente um do outro, e se o sistema completo funcionar adequadamente quando ao menos 80 componentes funcionarem, qual será a confiabilidade do sistema?

X = número de componentes que funcionem;

$$E(X) = np = 100 \cdot 0,95 = 95$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \cdot 0,95 \cdot 0,05} = 2,1794$$

Empregando a aproximação da binomial pela normal e usando a correção de continuidade temos que:

$$P(80 \leqslant X \leqslant 100) \simeq P\left(80 - \frac{1}{2} \leqslant X \leqslant \frac{1}{2} + 100\right) = P(79,5 \leqslant X \leqslant 100,5)$$

$$P(79,5 \leqslant X \leqslant 100,5) = P\left(\frac{79,5-95}{2,1734} \leqslant Z \leqslant \frac{100,5-95}{2,1794}\right) = P(-7,11 \leqslant Z \leqslant 2,52) =$$

$$= P(0 \leqslant Z \leqslant 2,52) + P(-7,11 \leqslant Z \leqslant 0) =$$

$$= 0,9941$$

Aproximação da Distribuição Poisson pela Normal

Desde que existe uma relação entre as distribuições Binomial e Normal e entre as distribuições Binomial e de Poisson, espera-se que também deva haver uma relação entre as distribuições de Poisson e Normal. Isto de fato acontece. Podemos mostrar que se X é a variável aleatória de Poisson e $Z = \frac{X-\lambda}{\sqrt{\lambda}}$ é a variável aleatória padronizada correspondente, então:

$$\lim_{\lambda \to \infty} P\left(a \leqslant \frac{X-\lambda}{\sqrt{\lambda}} \leqslant b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{\frac{-u^2}{2}} du$$

Isto é, a distribuição de Poisson se aproxima da distribuição Normal quando $\lambda \to \infty$ ou $\frac{X-\lambda}{\sqrt{\lambda}}$ é assintoticamente normal.

Exemplo 5.8.4 Suponha-se que, em uma determinada central telefônica, as chamadas cheguem com taxa de 2 por minuto. Qual é a probabilidade de 22 ou menos chamadas sejam recebidas durante um período de 15 minutos?

X = número de chamadas recebidas

$$E(X) = 2 \cdot 15 = 30$$

$$P(X \leqslant 22) \simeq P\left(Z \leqslant \frac{(22+0,5)-30}{\sqrt{30}}\right) = P(Z \leqslant -1,37) = 0,50 - P(-1,37 \leqslant Z \leqslant 0) = 0,0853$$

5.9 Exercícios

5.9.1 Seja X uma $\nu.a.$ exponencial tal que $E(X) = 3$. Calcule:

a) $P(3 < X < 6)$

b) $P(X > 6 \backslash X > 3)$

5.9.2 Mostre que para uma $\nu.a.$ X com distribuição uniforme definida no intervalo $[a, b]$:

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

5.9.3 O tempo de vida (em horas) de um transistor pode ser considerado uma variável aleatória com distribuição exponencial com média de 500 horas. Calcule a probabilidade de que o transistor dure mais que a média. Encontre sua função distribuição acumulada.

5.9.4 A temperatura T de destilação do petróleo é crucial na determinação da qualidade final do produto. Suponha que T seja considerada uma $v.a.$ com distribuição uniforme no intervalo $(150, 300)$. Calcule:

a) A temperatura média

b) $P(T < 200)$

c) $P(T > 250)$

5.9.5 Um fabrica de carros sabe que os motores de sua fabricação têm duração normal, com média de 150.000 km e desvio padrão de 5000 km. Qual a probabilidade de que um carro, escolhido ao acaso, dos fabricados por essa firma, tenha um motor que dure:

a) Menos de 170.000 km?

b) Entre 140.000 km e 165.000 km?

5.9.6 Um carro de corrida é um dos muitos brinquedos fabricados pela Mack Corporation. Os tempos de montagem para esse brinquedo seguem uma distribuição normal, com uma média aritmética de 55 minutos e um desvio padrão de 4 minutos. A empresa fecha às 17 horas todos os dias. Caso um trabalhador comece a montar um carro de corrida às 16 horas, qual é a probabilidade de que ele venha a terminar esta tarefa antes do horário de encerramento do dia?

5.9.7 O período de falta ao trabalho em um mês por causa de doenças dos empregados é normalmente distribuído com uma média de 100 horas e desvio padrão de 20 horas.

a) Qual é a probabilidade desse período no próximo mês estar entre 50 e 80 horas?

b) Qual é a probabilidade desse período ser maior que 70 horas?

5.9.8 Um salário semanal dos operários da indústria de construção civil é distribuído normalmente em torno de uma média de R$ 480,00, com desvio padrão de R$ 50,00, encontre a probabilidade de um operário ter um salário semanal situado:

a) Entre R$ 480 e R$ 483;

b) De uma amostra de 500 operários da indústria de construção, quantos esperaríamos que ganhassem salários acima de R$ 485,00?

c) Qual é o valor do salário para escolhermos 10% dos operários com maiores remunerações? $(x = ?)$

5.9.9 As temperaturas registradas por um termômetro quando colocado em água fervente (temperatura real de 100°C) tem distribuição normal com média 99,8°C e desvio padrão 0,1°C.

a) Qual é a probabilidade do termômetro indicar uma temperatura maior que 100°C?

b) Qual é a probabilidade do termômetro indicar uma temperatura dentro de $\pm 0,05$°C da temperatura verdadeira?

5.9.10 Estudos meteorológicos indicam que a precipitação pluviométrica mensal, em períodos de seca numa certa região, pode ser considerada como seguindo a distribuição Normal de média 30 mm e variância 16 mm^2.

a) Qual seria o valor da precipitação pluviométrica de modo que exista apenas 10% de probabilidade de haver uma precipitação inferior a esse valor?

b) Admitindo esse modelo correto para os próximos 50 meses, em quantos deles esperaríamos uma precipitação pluviométrica superior a 34 mm?

5.9.11 Para X uma variável aleatória Normal com média μ e variância σ^2, encontre:

a) $P(X \geqslant \mu + 2\sigma)$

b) O número a tal que $P(\mu - a\sigma \leqslant X \leqslant \mu + a\sigma) = 0,99$

c) O número a tal que $P(X > a) = 0,90$

5.9.12 Uma empresa produz um equipamento cuja vida útil admite distribuição normal com média 300h e desvio padrão 20h. Se a empresa garantiu uma vida útil de pelo menos 280h para uma das unidades vendidas, qual a probabilidade de ela ter que repor essa unidade?

5.9.13 Os balancetes semanais realizados em uma empresa mostraram que o lucro realizado distribui- se normalmente com média 48.000 u.m. e desvio padrão 8.000 u.m.. Qual a probabilidade de que:

a) Na próxima semana o lucro seja maior que 50.000 u.m.?

b) Na próxima semana o lucro esteja entre 40.000 u.m. e 45.000 u.m.?

c) Na próxima semana haja prejuízo?

5.9.14 O Departamento de Marketing da empresa resolve premiar 5% dos seus vendedores mais eficientes. Um levantamento das vendas individuais por semana mostrou que elas se distribuíam normalmente com média 240.000 u.m. e desvio padrão 30.000 u.m.. Qual o volume de vendas mínimo que um vendedor deve realizar para ser premiado?

5.9.15 Uma máquina produz um tubo de plástico rígido cujo diâmetro admite distribuição normal de probabilidade, com média 100 mm e desvio padrão 0,5 mm. Os tubos com diâmetro menor que 98,2 mm ou maior que 100,6 mm são considerados defeituosos e devem ser reciclados. Qual a proporção que deverá ser reciclada?

5.9.16 Uma máquina produz 10% de seus parafusos de porca com defeito. Encontre, utilizando a aproximação normal para a distribuição binomial, a probabilidade de uma amostra aleatória de 400 parafusos de porca produzidos por esta máquina apresentar:

a) No máximo 30 com defeito;

b) Entre 30 e 50 com defeito;

c) Entre 35 e 45 com defeito;

d) 65 ou mais de parafusos com defeito.

5.9.17 Considere que o número de partículas de asbestos em um centímetro quadrado de poeira siga a distribuição de Poisson com uma média de 1.000. Se um centímetro quadrado de poeira for analisado, qual será a probabilidade de que menos de 950 partículas sejam encontradas? (Use a aproximação da Poisson pela Normal).

5.9.18 Um produto eletrônico para escritório contém 200 componentes eletrônicos. Suponha que a probabilidade de cada componente operar sem falhar durante a vida útil do produto seja $0,999$ e suponha que os componentes falhem independentemente. Utilizando a aproximação normal para a distribuição binomial calcule a probabilidade de cinco ou mais dos 200 componentes originais falharem durante a vida útil do produto.

5.9.19 Suponha que dados históricos mostram que o número médio de chegadas a uma caixa automática (tipo drive-thru) de um banco durante um período de 15 minutos nas manhãs de fins de semana é de 10 carros. Use a aproximação Normal para a Poisson e calcule a probabilidade de no máximo 110 cheguem em 3 horas.

5.10 Distribuição de Probabilidade Conjunta

Chamamos de **Conjunta** a probabilidade que se refere a duas (ou mais) variáveis aleatórias, discretas ou contínuas simultaneamente. Podemos ainda dizer que é distribuição de probabilidade de um **vetor aleatório** (X, Y) – para o caso bidimensional, isto é, com duas variáveis.

Distribuição Conjunta de Variáveis Discretas

Suponha um time de vôlei que vai disputar um campeonato muito equilibrado, em que a probabilidade de ganhar ou perder uma partida seja $0,5$. O técnico pede ao analista de estatísticas da equipe que faça uma análise das probabilidades das trÊs primeiras partidas, consideradas vitais para o restante da competição. Em particular, a vitória na primeira partida é considerada decisiva pela comissão técnicas.

O analista, então, define duas variáveis aleatórias:

$X =$ Número de vitórias obtidas nos três primeiros jogos; $Y = 1$, caso ocorra vitória no primeiro jogo, e $Y = 0$ caso ocorra o contrário.

Os possíveis resultados, V (vitória) e D (derrota) – e os correspondentes valores de X e Y estão na tabela a seguir:

Resultados Possíveis	X	Y
VVV	3	1
VVD	2	1
VDV	2	1
VDD	1	1
DVV	2	0
DDV	1	0
DVD	1	0
DDD	0	0

A seguir, o analista constrói uma tabela que apresenta as probabilidades conjuntas de X e Y. Assim, na posição da tabela que corresponde a $X = 2$ e $Y = 1$, devemos colocar a probabilidade de isso ocorrer, isto é, $P(X = 2 \text{ e } Y = 1)$. Pela tabela acima, verificamos que, em oito resultados

possíveis, temos dois em que há duas vitóriaS ($X = 2$) e há vitória no primeiro jogo ($Y = 1$). Portanto, $P(X = 2, Y = 1) = \frac{2}{8}$. Procedendo desta maneira teremos a Tabela 5.10.1:

Tabela 5.10.1: Probabilidades conjuntas de X e Y

Variável Y	**Variável X**			
	0	**1**	**2**	**3**
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Como se nota, a probabilidade de X (só de X, sem considerar o que ocorre com Y) é dada pela soma das probabilidades ao longo da coluna, ou seja, somando-se as probabilidades de todos os valores de Y. E a distribuição de probabilidade só de Y é obtida da mesma forma, ou seja, somando-se as probabilidades ao longo da linha, isto é, somam-se todos os valores possíveis de X.

Na tabela a 5.10.2, além da distribuição conjunta de X e Y, mostramos também a **distribuições marginais** de X e de Y, representadas por P(X) e P(Y):

Tabela 5.10.2: Distribuições Marginais de X e Y

Variável Y	**Variável X**				
	0	**1**	**2**	**3**	P(Y)
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{1}{2}$
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
P(X)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

O número 1 no canto inferior direito da tabela representa a soma das probabilidades marginais (e da conjunta também), que tem de ser, obviamente, igual a 1.

É possível utilizar a Tabela 5.10.2 para calcular as probabilidades condicionais, embora elas não possam ser obtidas diretamente dessa fonte. Suponhamos que queiramos saber qual a probabilidade de X ser igual a 1, dado que Y é 1 (isto é, se acontecer uma vitória no primeiro jogo, qual a probabilidade de que só aconteça uma vitória nos três jogos).

Pela definição de probabilidade condicional, temos:

$$P(X = 1 \backslash Y = 1) = \frac{P(X = 1 \text{ e } Y = 1)}{P(Y = 1)} = \frac{\frac{1}{8}}{\frac{1}{2}} = \frac{1}{4}$$

Calcularemos agora, a esperança e a variância das variáveis aleatórias X e Y. Para calcular E(X) e V(X), usaremos as probabilidades dadas pela distribuição marginal de X, que pode asumir os valores 0, 1, 2 e 3:

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1{,}5$$

$$E(X^2) = 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{8} = 3$$

$$V(X) = E(X^2) - [E(X)]^2 = 0,75$$

Para Y, vale o mesmo raciocínio:

$$E(Y) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0,5$$

$$E(Y^2) = 0^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{2} = 0,5$$

$$V(Y) = E(Y^2) - [E(Y)]^2 = 0,5 - 0,25 = 0,25$$

Para variáveis aleatórias X e Y, vale sempre que $E(X+Y) = E(X) + E(Y)$.

Se X e Y são independentes $\Rightarrow E(XY) = E(X) \cdot E(Y)$.

No entanto $E(XY) = E(X) \cdot E(Y)$ não implica em X e Y independentes.

Se as variáveis são dependentes, a relação entre elas pode ser de vários tipos e, no caso de ser linear, vamos definir uma medida dessa dependência, a **Covariância**.

Covariância de duas variáveis aleatórias

É uma medida de dependência linear entre X e Y que é dada por:

$$covar(X,Y) = E(XY) - E(X) \cdot E(Y)$$

Em palavras, a covariância é o valor esperado do produto dos desvios de cada variável em relação à sua média.

Para o nosso exemplo temos que

$$E(XY) = 0 \cdot \frac{4}{8} + 1 \cdot \frac{1}{8} + 2 \cdot \frac{2}{8} + 3 \cdot \frac{1}{8} = 1$$

$$covar(X,Y) = 1 - 1,5 \cdot 0,5 = 0,25$$

Observe que, no caso em que X e Y serem independentes, temos $covar(X,Y) = 0$, uma vez que o valor esperado do produto se torna igual ao produto dos valores esperados. A partir da Covariância, definimos uma nova medida de dependência linear.

Correlação Entre Variáveis Aleatórias

O coeficiente de correlação linear entre as variáveis discretas X e Y é calculado pela seguinte expressão:

$$\rho_{X,Y} = \frac{covar(X,Y)}{\sqrt{V(X) \cdot V(Y)}}$$

Pela definição acima, o coeficiente de correlação é o quociente entre a covariância e o produto os desvios-padrão de X e Y.

Observação 5.10.1 $-1 \leqslant \rho_{X,Y} \leqslant 1$

A interpretação de sua expressão segue os mesmos passos da covariância, sendo que valores de $\rho_{X,Y}$, próximos de ± 1 indicam correlação forte.

Para o nosso exemplo, o coeficiente de correlação será:

$$\rho_{X,Y} = \frac{0,25}{\sqrt{0,75 \cdot 0,25}} = 0,5774$$

5.11 Distribuições Amostrais

Definição 5.11.1 A distribuição de probabilidades de uma estatística é chamada de uma distribuição amostral. Por exemplo, a distribuição de probabilidades de $\overline{x}$ é chamada de distribuição amostral da média.

A distribuição amostral de uma estatística depende da distribuição da população, do tamanho da amostra e do método de seleção da amostra.

Distribuição Amostral da Média

Um dos procedimentos estatísticos mais comuns é o uso de uma média da amostra $\overline{x}$ para fazer inferências sobre a média da população μ.

Definição 5.11.2 A distribuição amostral de $\overline{x}$ é a distribuição de probabilidade de todos os valores possíveis da média da amostra, $\overline{x}$.

Considere a determinação da distribuição amostral da média $\overline{x}$ da amostra. Suponha que uma amostra aleatória de tamanho n seja retirada de uma população normal com média μ e variância σ^2. Então, temos que

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

tem uma distribuição normal com média

$$\mu_x = \frac{\mu + \mu + \mu + \cdots + \mu}{n} = \mu$$

e se uma população for infinita e amostragem for aleatória, ou se a população for finita e a amostragem for com reposição, então a variância da distribuição amostral das médias será dada por

$$\sigma_x^2 = \frac{\sigma^2 + \sigma^2 + \sigma^2 + \cdots + \sigma^2}{n} = \frac{\sigma^2}{n}$$

e se a população for de tamanho N, se a amostragem for sem reposição, então a variância será dada com a inclusão do chamado **fator de correção da população finita**:

$$\sigma_x^2 = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1}\right)$$

Observação 5.11.1 O valor de $\sigma_{\bar{x}}$ (o desvio padrão de $\bar{x}$) é útil para determinar a distância a que a média da amostra pode estar da média populacional. Devido ao papel que $\sigma_{\bar{x}}$ desempenha em calcular possíveis erros de amostragem, $\sigma_{\bar{x}}$ é denominado **erro-padrão da média**.

Se estivermos amostrando de uma população que tenha uma distribuição desconhecida de probabilidades, a distribuição amostral da média da amostra será aproximadamente normal, com média μ e variância $\frac{\sigma^2}{n}$, se o tamanho n da amostra for grande. Esse é um dos mais úteis teoremas em Estatística, o chamado Teorema Central do Limite, que tem o seguinte enunciado:

Teorema Central do Limite:

Se $x_1, x_2, \ldots, x_3$ for uma amostra aleatória de tamanho n, retirada de uma população, com média μ e variância σ^2, e se $\bar{x}$ for a média da amostra, então:

$$\lim_{n\to\infty} Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \simeq N(0,1)$$

Este teorema mostra que quando o tamanho da amostra aumenta, independentemente da forma da distribuição da população, a distribuição amostral de $\bar{x}$ aproxima-se cada vez mais de uma distribuição normal.

Exemplo 5.11.1 Uma companhia eletrônica fabrica resistores que têm uma resistência média de 100 Ω e um desvio padrão de 10 Ω. A distribuição de resistência é normal. Encontre a probabilidade de uma amostra aleatória de $n = 25$ resistores ter uma resistência média menor que 95 Ω.

Note que a distribuição amostral $\bar{x}$ é normal, com média $\mu_{\bar{x}} = 100\ \Omega$ e um desvio padrão de

$$\sigma_{\bar{x}} = \frac{\sigma}{n} = \frac{10}{\sqrt{25}} = 2$$

$$\Rightarrow Z = \frac{95-100}{2} = -2,5$$

$$\therefore P(\bar{x} < 95) = P(Z < -2,5) = 0,50 - 0,4938 = 0,0062$$

Distribuição Amostral de uma Proporção

Em muitas situações, nos negócios e na economia, usamos a proporção $\bar{p}$ para fazer inferências estatísticas sobre a proporção da população p.

Definição 5.11.3 A distribuição amostral de $\bar{p}$ é a distribuição de probabilidade de todos os valores possíveis da proporção da amostra $\bar{p}$.

Proporção amostral é a fração dos indivíduos com uma dada característica em uma amostra de tamanho n, isto é,

$$\overline{p} = \frac{n^o \text{ de indivíduos na amostra com dada característica}}{n}$$

Se construirmos para *i-ésimo* indivíduo uma variável aleatória Y_i tal que

$$Y_i = \begin{cases} 1, & \text{se o indivíduo apresenta a característica;} \\ 0, & \text{caso contrário.} \end{cases}$$

Podemos reescrever a proporção amostra como

$$\overline{p} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} = \sum_{i=1}^{n} \frac{Y_i}{n} = \overline{Y}$$

Logo, a proporção amostral nada mais é do que a média de variáveis aleatórias convenientemente definidas. Assumindo que a proporção de indivíduos com a dada característica na população é p, e que os indivíduos são selecionados aleatoriamente, temos que $Y_1, \ldots, Y_n$ formam uma seqüência de variáveis aleatórias independentes com distribuição de Bernoulli. Assim, $E(Y_i) = p$ e $V(Y_i) = p(1-p)$. Logo,

$$E(\overline{p}) = E\left(\sum_{i=1}^{n} \frac{Y_i}{n}\right) = p$$

$$V(\overline{p}) = V\left(\sum_{i=1}^{n} \frac{Y_i}{n}\right) = \frac{p \cdot (1-p)}{n}$$

Tendo em vista o Teorema Central do Limite temos que para n suficientemente grande,

$$\frac{\overline{Y} - E(\overline{Y})}{\sqrt{V(Y)}} = \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{n \to \infty} \sim N(0,1)$$

Exemplo 5.11.2 Suponha que a proporção de peças fora da especificação em um lote é de 40%. Tomada uma amostra de tamanho 30, qual será a probabilidade desta amostra fornecer uma proporção de peças defeituosas menor que 0,50?

$$P(\overline{p} < 0,50) = P\left(\frac{X}{30} < 0,50\right) = P(X < 15)$$

Considerando a aproximação Normal, temos, como conseqüência do Teorema Central do Limite

$$\overline{p} \sim N\left(0,40, \frac{0,40(1-0,40)}{30}\right)$$

Assim,

$$P(\overline{p} < 0,50) \simeq P\left(\frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0,50 - 0,40}{\sqrt{\frac{0,40(1-0,40)}{30}}}\right) = P(Z < 1,12) = 0,8686$$

Observação 5.11.2 A distribuição amostral $\overline{p}$ pode ser aproximada por uma distribuição normal de probabilidade sempre que o tamanho da amostra é grande. Com $\overline{p}$, o tamanho da amostra pode ser considerado grande sempre que as seguintes duas condições são satisfeitas:

i) $n.p \geqslant 5$;

ii) $n.(1-p) \geqslant 5$.

5.12 Exercícios

5.12.1 A tabela abaixo mostra a distribuição conjunta das variáveis aleatórias discretas U e V. Encontre suas distribuições marginais e calcule a covariância das duas variáveis e o respectivo coeficiente de correlação.

V \ U	0	1	2
-1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
0	$\frac{1}{8}$	0	$\frac{1}{8}$
1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

5.12.2 Num estudo sobre a rotatividade de mão de obra, foram definidas para certa população as v.a. X = número de empregos que um funcionário teve no último ano e Y = salário. Obteve-se a seguinte distribuição conjunta:

X \ Y	1	2	3	4
800	0	0	0,10	0,10
1.200	0,05	0,05	0,10	0,10
2.000	0,05	0,20	0,05	0
5.000	0,10	0,05	0,05	0

São dados: $E(X) = 2,5$, $DP(X) = 1,0$, $E(Y) = 2.120$, $DP(Y) = 1.505,2$

a) Calcule $P(X = 2)$ e $P(X = 2/Y = 1.200)$;

b) Obtenha o coeficiente de correlação entre X e Y.

5.12.3 Uma variável aleatória X tem distribuição normal, com média 100 e desvio padrão 10.

a) Qual a $P(90 < X < 110)$?

b) Se $\overline{x}$ for a média de uma amostra de 16 elementos retirados dessa população, calcule $P(90 < \overline{x} < 110)$.

5.12.4 A capacidade máxima de um elevador é de 500 Kg. Se a distribuição X dos pesos dos usuários for suposta $N(70, 100)$:

a) Qual é a probabilidade de sete passageiros ultrapassarem esse limite?

b) Qual é a probabilidade de seis passageiros ultrapassarem esse limite?

5.12.5 O presidente da *Doerman Distritors, Inc.*, acredita que 30% dos pedidos de compra da empresa venham de clientes novos ou de primeira vez. Uma amostra aleatória simples de 100 pedidos será usada para estimar a proporção de clientes novos ou de primeira vez. Os resultados da amostra serão usados para verificar a reivindicação do presidente de $p = 0,30$.

a) Qual é a probabilidade de que a proporção da amostra estará entre $0,20$ e $0,40$?

b) Qual é a probabilidade de que a proporção da amostra estará dentro de $\pm 0,05$ da proporção da população $p = 0,30$.

5.12.6 Considere que 15% dos itens produzidos em uma operação de linha de montagem são defeituosos, mas que o gerente de produção da empresa não está ciente dessa situação. Considere, além disso, que 50 peças são testadas pelo departamento de garantia da qualidade para determinar a qualidade da operação de montagem. Seja $\overline{p}$ a proporção da amostra defeituosa encontrada pelo teste de garantia de qualidade.

a) Qual é a probabilidade de que a proporção da amostra estará dentro de $\pm 0,03$ da proporção da população que está defeituosa?

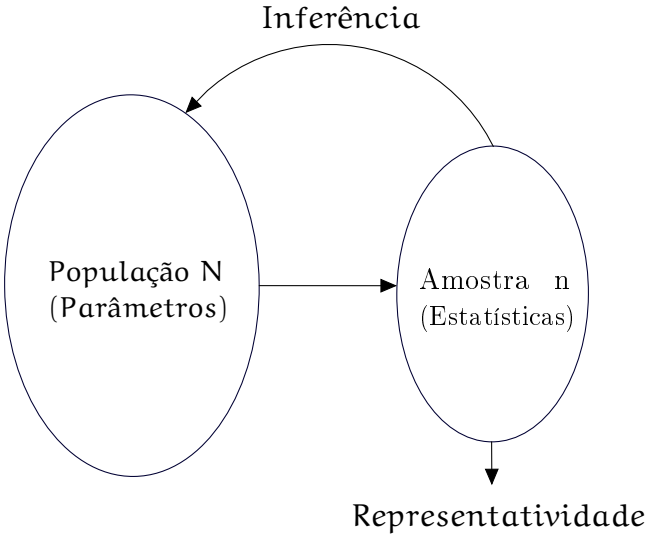
b) Se o teste mostra que $\overline{p} = 0,10$ ou mais, a operação da linha de montagem será paralisada para se verificar a causa dos defeitos. Qual é a probabilidade de que a amostra de 50 peças levará à conclusão de que a linha de montagem deva ser paralisada?

6 Processos de Amostragem

De forma geral, as populações ou universos nos quais o pesquisador está interessado são grandes demais para serem estudados na sua totalidade. O tempo necessário para estudar toda a população, as despesas e o número de pessoas envolvidas é de tal monta que tornam o estudo proibitivo. Por isso, o mais comum é se estudarem amostras retiradas da população de interesse.

Amostragem é então a parte da Estatística que estuda os diversos processos de obtenção de amostras, com objetivo que elas sejam **representativas** da população em estudo.

Observação 6.0.1 **Representativa** significa que guarda ou reproduza as mesmas propriedades da população.



Na execução de uma pesquisa, na maioria das vezes, é impossível avaliar todos os elementos de uma população, devido a problemas de custo e tempo. Quando esse é o caso, é preferível conhecer a população a partir de uma parte dela, chamada **amostra**. A importância da amostragem está em

que mediante a informação contida numa amostra é possível fazer inferências (análise e conclusões) sobre as características da população.

População é o conjunto de todos os elementos ou resultados sob investigação. **Amostra** é qualquer subconjunto da população.

Observação 6.0.2 O tamanho da população pode ser **FINITO** ou **INFINITO**:

População Finita: Aquela em que todos os seus elementos podem ser identificados e enumerados. Exemplo: Os associados de um clube esportivo.

As populações finitas são freqüentemente definidas por listas tais como relação nominal de membros de organização, registros de matrícula de estudantes, listagens de contas de cartão de crédito, números de produtos de inventário e assim por diante.

População Infinita: Não é possível identificar (ou numerar) todos os seus elementos. Exemplo: As árvores pertencentes a um determinado tipo de eucalipto que existem no mundo.

As populações infinitas também são freqüentemente definidas por um processo contínuo em que os elementos da população consistem de itens gerados como se o processo operasse indefinidamente sob as mesmas condições. Em tais casos, é impossível obter uma lista de todos os elementos na população. Por exemplo, populações que consistem de todas as peças possíveis de serem manufaturadas, todas as visitas possíveis de cliente, todas as possíveis transações bancárias e assim por diante, podem ser classificadas como populações infinitas.

Levantamentos Amostrais: A amostra é obtida de uma população bem definida, por meio de processos bem protocolados e controlados pelo pesquisador. Podemos, ainda, subdivi-los em dois subgrupos: **amostras probabilísticas** e **amostras não- probabilísticas**. O primeiro reúne todas as aquelas técnicas que usam mecanismos aleatórios de seleção dos elementos de uma amostra, atribuindo a cada um deles uma probabilidade, conhecida *a priori*, de pertencer à amostra. No segundo grupo estão os demais procedimentos, tais como: amostras intencionais, nas quais os elementos são selecionados com o auxílio de especialistas, e amostras de voluntários, como ocorre em alguns testes sobre novos medicamentos e vacinas. Aqui vamos discutir os principais processos de amostragem probabilística.

Amostragem Probabilística:

Definição 6.0.1 Se todos os elementos da população tiverem probabilidade conhecida e não nula de pertencer à amostra o **processo de amostragem** é denominado probabilístico.

É aplicável quando é possível enumerar todos os elementos da população.

Principais Processos de Amostragem Probabilísticos:

a) **Amostragem simples ao acaso (ASA)** Tem objetivo de obter uma amostra representativa, quando os elementos da população são todos homogêneos.

Neste processo de amostragem todos os elementos da população têm a mesma probabilidade de serem coletados. Todos os elementos da população são numerados $\{1, \ldots, N\}$ e sorteia-se n elementos através de um dispositivo aleatório, por exemplo a "tabela de números aleatórios"(anexa no fim da apostila) ou programa de computadores.

O número de diferentes amostras aleatórias simples de tamanho n que pode ser selecionado de uma população finita de tamanho N é:

$$\frac{N!}{n!(N-n)!}$$

b) Amostragem Sistemática

Tem o objetivo de aumentar a representatividade da amostra, dando maior cobertura a população. É usada quando os elementos da população estão ordenados de alguma maneira (em listas, filas, prateleiras, linhas de produção).

Procedimento: Os elementos da população são numerados $(1, 2, \ldots, N)$, e o primeiro elemento da amostra é sorteado, por exemplo i. Os demais são retirados em uma progressão aritmética, saltando r elementos, até completar o total da amostra (n), isto é , $i + r$, $i + 2r$, $i + 3r$, etc. O valor de r, que é conhecido como "passos de amostragem"é determinado pela seguinte razão:

$$r = \frac{N}{n}$$

O primeiro elemento é sorteado entre 1 e o valor de r.

A amostragem sistemática também é freqüentemente utilizada em pesquisas de opinião, realizadas em locais públicos, quando não se dispõe de uma relação da população.

Se o tamanho da população é desconhecido, não podemos determinar exatamente o valor de r. Escolheremos intuitivamente um valor razoável para r.

Às vezes a amostragem sistemática é preferida à amostragem aleatória simples porque é mais fácil de executar e proporciona mais informações com menor custo.

c) Amostragem Estratificada

Tem o objetivo de melhorar a representatividade da amostra quando os elementos da população são heterogêneos, porém, podem ser agrupados em subpopulações (estratos) contendo elementos homogêneos.

Procedimento: A população é dividida em grupos de elementos chamados estratos, contendo elementos homogêneos, tais que cada elemento na população pertence a um e somente um estrato. Depois que os estratos são formados, toma-se uma amostra aleatória simples de cada estrato.

Quanto ao tamanho das sub-amostras retiradas (n_i), é classificada em:

i) Uniforme:
   Quando de K estratos, retiram-se amostras de mesmo tamanho n, independentemente do tamanho do estrato;

ii) Proporcional:
   Quando o tamanho da amostra retirado em cada estrato (n_i) é proporcional ao tamanho do estrato.

Exemplo 6.0.1 Temos que $N = 4000$. Essa população foi dividida em 3 estratos (estrato $1 = 2000$, estrato $2 = 1200$ e estrato $3 = 800$) para ser realizado o levantamento amostral. O tamanho da amostra (n) calculado é igual a 60. As respectivas amostras de cada estrato serão:

$$n_1 = \frac{2000}{4000} \cdot 60 = 30 \qquad n_2 = \frac{1200}{4000} \cdot 60 = 18 \qquad n_3 = \frac{800}{4000} \cdot 60 = 12$$

A base para a formação do estrato, tal como um departamento, local, idade, tipo de indústria e assim por diante, está a critério do planejador da amostra.

d) Amostragem por Conglomerados:

Na amostragem por conglomerado, a população é dividida primeiro em grupos de elementos separados chamados de conglomerados. Cada elemento da população pertence a um e somente um conglomerado. Uma amostra aleatória simples dos conglomerados é então tomada. Todos os elementos dentro de cada conglomerado amostrado formam a amostra. A amostragem por conglomerado tende a fornecer os melhores resultados quando os elementos são heterogêneos. No caso ideal, cada conglomerado é uma versão em pequena escala representativa da população inteira. O valor da amostragem por conglomerado depende da representatividade que ele tem da população inteira. Se todos os conglomerados são parecidos a esse respeito, amostrar um pequeno número de conglomerados fornecerá boas estimativas dos parâmetros da população.

Uma das aplicações primárias da amostragem por conglomerado é amostragem de área, por onde os conglomerados são blocos de cidade ou outras áreas bem definidas. A amostragem por conglomerado geralmente exige um tamanho maior de amostra total do que a amostragem aleatória simples e a amostragem estratificada. No entanto, pode resultar em economia de custo por causa do fato de que quando um entrevistador é enviado para um conglomerado amostrado (por exemplo, um quarteirão de uma cidade), muitas observações da amostra podem ser obtidas em um tempo relativamente curto. Por isso, um tamanho maior de amostra pode ser obtido com custo total significativamente mais baixo.

Procedimento: Consiste em subdividir a população em componentes (grupos ou conglomerados de elementos) que reproduzem bem as características da população, sorteia- se um número determinado desses conglomerados ($\mathfrak{m}$) e todos os elementos destes vão compor a amostra.

Fontes de Erro em Pesquisas por Amostragem

O erro amostral tolerável (o valor máximo que o pesquisador admite errar na estimativa do parâmetro) considera que a amostra foi retirada seguindo rigorosamente o plano de amostragem, e que não houve viés por parte do pesquisador. Caso contrário, ou seja, se ocorrem **erros não amostrais**, o erro amostral não pode mais ser garantido. Estes erros poderiam ser:

- Problemas no instrumento de pesquisa (questionário ambíguo, opções não conseguem medir as respostas do respondente);
- Problemas com as pessoas que aplicam a pesquisa (entrevistadores mal treinados, cansados ou simplesmente inadequados para função);
- Falta de resposta (uma parcela da amostra pode recusar-se a participar da pesquisa, ao menos em um primeiro momento);
- Erro de cobertura (bastante comum em pesquisas que usam questionários por correio, drop-off ou on-line, em suma é a diferença entre a população alvo e a acessível, somente os interessados respondem à pesquisa, o que pode causar tendência nos resultados).

Lidando com os Erros

Há métodos estatísticos disponíveis, que possibilitam a estimativa dos tamanhos prováveis dos erros amostrais. Tudo o que podemos fazer com erros alheios à amostragem é tentar minimiza-los,

no estágio de desenho da pesquisa. Uma **pesquisa-piloto** pode ser uma opção, já que sua função é testar uma pesquisa, em um grupo relativamente pequeno, para tentar identificar os possíveis problemas no desenho da pesquisa, antes de conduzir a própria pesquisa.

6.1 Tamanho da Amostra para Estimar a Média de uma Variável de uma População Infinita no Processo de Amostragem Simples ao Acaso

$$n = \frac{Z^2_{\frac{\alpha}{2}}\,\sigma^2}{e^2}$$

onde e = erro amostral expresso na unidade da variável. O erro amostral é a máxima diferença que o investigador admite suportar entre μ e $\overline{x}$.

Se σ é desconhecido, você poderá estimá-lo usando s, desde que aplique uma amostra preliminar.

Exemplo 6.1.1 Suponha que um gerente de marketing deseje estimar a média aritmética da população, do consumo anual de óleo para a calefação residencial, entre ± 50 galões de distância e relação ao verdadeiro valor, e deseje estar 95% confiante de estar corretamente estimando a verdadeira média aritmética. Com base em um estudo realizado no ano anterior, ele acredita que o desvio padrão pode ser estimado em 325 galões. Encontre o tamanho de amostra necessário.

$$n = \frac{Z^2_{\frac{\alpha}{2}}\,\sigma^2}{e^2} = \frac{1,96^2 \cdot 325^2}{50^2} = 162,31 = 163$$

Observação 6.1.1 A regra geral é sempre arredondar para o próximo valor inteiro acima, no sentido de "supersatisfazer"os critérios desejados.

6.2 Tamanho da Amostra para Estimar a Proporção de uma Variável de uma População Infinita no Processo de Amostragem Simples ao Acaso

$$n = \frac{Z^2_{\frac{\alpha}{2}}\,\widehat{p}(1-\widehat{p})}{e^2}$$

$\widehat{p}$ = estimativa da verdadeira proporção de um dos níveis da variável escolhida. Por exemplo, se a variável escolhida for o porte da empresa, $\widehat{p}$ poderá ser a estimativa da verdadeira proporção de grandes empresas do setor que está sendo estudado. Será expresso em decimais. Assim, se $\widehat{p} = 30\%$, teremos $\widehat{p} = 0,30$.

Caso não se tenha estimativas prévias para $\widehat{p}$, admita $\widehat{p} = 0,50$ obtendo assim o maior tamanho de amostra possível considerando constantes os valores de $Z_{\frac{\alpha}{2}}$ e e.

e = erro amostral expresso em decimais. O erro amostral nesse caso será a máxima diferença que o investigador admite suportar entre p e $\widehat{p}$.

Exemplo 6.2.1 Um gerente de operações deseja ter 90% de confiança de estimar a proporção de jornais que estão fora dos padrões de conformidade, dentro dos limites de $\pm 0,05$, em relação a seu verdadeiro valor. Além disso, uma vez que o editor do jornal não realizou anteriormente uma pesquisa deste tipo, nenhuma informação se encontra disponível a partir de dados anteriores. Determine o tamanho de amostra necessário.

$$n = \frac{Z^2_{\frac{\alpha}{2}} \widehat{p}(1-\widehat{p})}{e^2} = \frac{1,645^2 \cdot 0,50 \cdot 0,50}{0,05^2} = 270,60 = 271$$

6.3 Tamanho da Amostra para Estimar a Média de uma Variável de uma População Finita no Processo de Amostragem Simples ao Acaso

$$n = \frac{Z^2_{\frac{\alpha}{2}} \sigma^2 N}{e^2(N-1) + Z^2_{\frac{\alpha}{2}} \sigma^2}$$

onde e = erro amostral expresso na unidade da variável. O erro amostral é a máxima diferença que o investigador admite suportar entre μ e $\overline{x}$.

6.4 Tamanho da Amostra para Estimar a Proporção (p) de uma Variável de uma População Finita no Processo de Amostragem Simples ao Acaso

$$n = \frac{Z^2_{\frac{\alpha}{2}} \widehat{p}(1-\widehat{p}) N}{e^2(N-1) + Z^2_{\frac{\alpha}{2}} \widehat{p}(1-\widehat{p})}$$

6.5 Tamanho da Amostra para Estimar a Média de uma Variável de uma População Finita no Processo de Amostragem Estratificada

$$n = \frac{\sum_{i=1}^{k} \left(\frac{N_i^2 \sigma_i^2}{w_i} \right)}{N^2 \frac{e^2}{Z^2_{\frac{\alpha}{2}}} + \sum_{i=1}^{k} N_i \sigma_i^2}$$

onde:
N_i = número de elementos do estrato i;
K = número de estratos;
N = número de elementos da população;
$w_i = \frac{N_1}{N}$;

σ_i^2 = variância populacional do estrato i. Poderá ser avaliado de, pelo menos, uma das três maneiras: especificações técnicas, resgatar o valor de estudos semelhantes ou fazer conjecturas com base em amostras-piloto.

6.6 Tamanho da Amostra para Estimar a Proporção de uma Variável de uma População Finita no Processo de Amostragem Estratificada

$$n = \frac{\sum_{i=1}^{k} \left(\frac{N_i^2 \cdot \widehat{p}_i(1-\widehat{p}_i)}{w_i} \right)}{N^2 \frac{e^2}{Z^2_{\frac{\alpha}{2}}} + \sum_{i=1}^{k} N_i \widehat{p}_i(1-\widehat{p}_i)}$$

Onde:

$\widehat{p}_i$ = estimativa da verdadeira proporção do estrato i. Caso não se tenha estimativas prévias para $\widehat{p}_i$, admita $\widehat{p}_i = 0,50$, obtendo assim o maior tamanho de amostra possível do estrato i, considerando constantes os valores de $Z^2_{\frac{\alpha}{2}}$ e e.

6.7 Exercícios

6.7.1 Uma socióloga deseja saber as opiniões de mulheres adultas empregadas, sobre verbas governamentais para creches. Para isto, ela obtém uma relação dos 520 membros de uma firma local e do clube das mulheres profissionais, e envia um questionário a 100 dessas mulheres selecionadas aleatoriamente. Apenas 48 questionários são devolvidos. Qual é a população nesse estudo? Qual é a amostra da qual se obtêm efetivamente as informações? Qual é a taxa (percentagem) de não-resposta?

6.7.2 A *Statewide Insurance Company* usou uma amostra aleatória simples de 36 proprietários de apólice para estimar a idade média da população de proprietários de apólice. À confiança de 95%, a margem de erro foi de 2,35 anos. Esse resultado foi baseado em um desvio padrão da amostra de 7,2 anos, Que tamanho de amostra aleatória simples seria necessário para reduzir a margem de erro para 2 anos? E para 1,5 anos? E para 1 ano?

6.7.3 Um levantamento de mulheres executivas realizado por *Louis Harris & Associates* mostrou que 33% das pessoas pesquisadas avaliaram suas próprias empresas como um excelente lugar para as executivas trabalharem (*Working Woman*, novembro de 1994). Suponha que a *Working Woman* queira realizar um levantamento anual para monitorar essa proporção. Com $p = 0,33$ como um valor planejado para a proporção da população, quantas executivas deverão ser amostradas para cada uma das seguintes margens de erro? Assuma que todas as estimativas por intervalo são realizadas em um nível de confiança de 90%.

a) 10%

b) 5%

c) 2%

d) Em geral, o que acontece ao tamanho da amostra quando a margem de erro diminui?

6.7.4 Um fabricante de tintas usa uma máquina para encher as latas.

a) O fabricante quer estimar o volume médio de tinta que a máquina despeja nas latas com erro de $0,25$ onça. Determine o mínimo tamanho necessário da amostra para construir um intervalo de confiança de 98% para a média populacional. Suponha que o desvio padrão populacional seja $0,85$ onça.

b) Repita a parte a) usando um tolerância ao erro de $0,15$ onça. Qual o erro requer um tamanho de amostra maior? Justifique.

6.7.5 Sabe-se que 20% de um determinado produto de uma loja continham defeitos. Realizou-se uma amostra de 200 itens em uma outra loja, que contém 1.200 itens desse produto, para determinar a quantidade de produtos defeitos. Com 95% de confiança, qual a margem de erro da pesquisa?

6.7.6 Em uma população de 245.465 elementos, qual o tamanho mínimo de uma amostra aleatória simples para estimar proporção, com um nível de confiança de 95% e erro de $3,5$%?

6.7.7 Uma amostra estratificada deve ser retirada da empresa *XYZ*, para que uma pesquisa de opinião sobre a nova política de benefícios mensais. Esta empresa está dividida da seguinte maneira:

Departamento	N_i
Administrativo	100
Produção	400
Vendas	100
Serviços Gerais	80

a) Calcule o tamanho da amostra para estimar proporção usando um erro de 5% e 95% de confiança.

b) Verifique os tamanhos das amostras de cada estrato se você utilizasse a alocação uniforme e depois faça o mesmo se você utilizasse a alocação proporcional.

6.7.8 Sabendo que $e = Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\cdot\widehat{q}}{n}}\sqrt{\frac{N-n}{N-1}}$, mostre que:

$$n = \frac{Z^2_{\frac{\alpha}{2}}\cdot\widehat{p}\cdot\widehat{q}\cdot N}{e^2(N-1)+Z^2_{\frac{\alpha}{2}}\cdot\widehat{p}\cdot\widehat{q}}$$

7 Inferência Estatística

O campo da inferência estatística consiste naqueles métodos usados para tomar decisões ou tirar conclusões acerca de uma população. Esses métodos utilizam a informação contida em uma amostra da população para tirar conclusões. A inferência pode ser dividida em duas grandes áreas: **estimação de parâmetros e teste de hipóteses**.

Definições Importantes:

Definição 7.0.1 **População**: Consiste na totalidade das observações em que estamos interessados. Ela pode ser:

i) População Finita: Aquela em que todos os seus elementos podem ser identificados e enumerados. Exemplo: As propriedades rurais do Brasil

ii) População Infinita: Não é possível identificar (ou enumerar) todos os seus elementos. Exemplo: As árvores pertencentes a um determinado tipo de eucalipto que existem no mundo.

Definição 7.0.2 **Amostra**: É um subconjunto de observações selecionadas a partir de uma população.

Definição 7.0.3 **Parâmetros**: Constantes inerentes a populações relacionadas a uma determinada variável de interesse (X). Em muitas vezes são desconhecidos. Toda distribuição de probabilidade estudada nos capítulos anteriores depende de parâmetros, que determinam sua função específica, por exemplo, o p da binomial, o λ de Poisson, a μ e o σ da Normal, etc. Diferentes valores dos parâmetros conduzem a valores distintos das probabilidades.

Definição 7.0.4 **Estimador ou Estatística**: É uma variável aleatória que é função dos elementos amostrais X_i, $i = 1, 2, \ldots, n$.

Exemplo: A média amostral ($\overline{x}$) é um estimador de μ (parâmetro populacional). Ela é uma variável aleatória que depende dos resultados obtidos em cada amostra particular. Uma vez que uma estatística é uma variável aleatória, ela tem distribuição de probabilidades.

Definição 7.0.5 **Estimativa**: É o valor numérico do estimador em uma determinada amostra.

7.1 Estimação

Quando estamos interessados em determinado parâmetro de uma população, lançamos mão de uma amostra extraída dessa população, estudamos seus elementos e procuramos, através dessa amostra, estimar o parâmetro populacional.

Exemplo 7.1.1 Um candidato a prefeito pode quere avaliar a proporção de eleitores de seu município que o favorecem, consultando uma amostra de 100 eleitores. A proporção de eleitores da amostra favoráveis a ele servirá como estimativa da correspondente proporção populacional, que só será conhecida após as eleições.

A estimação de um parâmetro populacional comporta dois tipos: **estimação pontual** e **estimação intervalar**. A estimação pontual procura fixar um valor do parâmetro único que esteja satisfatoriamente próximo do verdadeiro valor do parâmetro. A estimação intervalar procura determinar intervalos com limites aleatórios, que abranjam o valor do parâmetro populacional, com uma margem de segurança prefixada.

Estimação Pontual

Representamos o parâmetro populacional de interesse por θ (theta). Para estimá-lo, extraímos uma amostra de tamanho n da população e procuramos construir uma função desses valores, ou seja, uma **estatística**, tal que seu valor, calculado com base nos dados amostrais, reflita, tão aproximadamente quanto possível, o valor do parâmetro populacional θ. Uma estatística desse tipo, destinada a estimar um parâmetro populacional θ, é chamada de estimador de θ. Designa-se por $\widehat{\theta}$. E, como diferentes amostras originam valores distintos para o estimador, $\widehat{\theta}$ é, ele próprio, uma variável aleatória.

Exemplo 7.1.2 Suponha que a variável aleatória X seja normalmente distribuída com uma média desconhecida μ. A média da amostra é um estimador da média desconhecida μ da população. Isto é, $\widehat{\mu} = \overline{x}$. Depois da amostra ter sido selecionada, o valor numérico de $\overline{x}$ é a estimativa de μ. Assim, se $x_1 = 25$, $x_2 = 30$, $x_3 = 29$, $x_4 = 31$, então a estimativa de μ é:

$$\overline{x} = \frac{25 + 30 + 29 + 31}{4} = 28,75$$

Propriedades dos Estimadores

i) **Não-tendenciosidade**

É razoável exigir que um bom estimador tenha sua distribuição de valores de algum modo centrada no verdadeiro valor θ do parâmetro a ser estimado. E, como a média, ou esperança,

de uma variável aleatória é uma medida de centro da mesma, uma exigência razoável para um estimador $\widehat{\theta}$ é que $E(\widehat{\theta}) = \theta$. Isto é, sua média deve ser igual ao valor do parâmetro. Um estimador que possui esta propriedade é chamado de **estimador não-tendencioso**. A não-tendenciosidade implica que os diversos valores de $\widehat{\theta}$ se distribuam em torno do verdadeiro valor θ sem ocasionar subestimação ou sobrestimação sistemática de θ.

ii) Variância mínima

Dois estimadores $\widehat{\theta}_1$ e $\widehat{\theta}_2$ não-tendencioso de θ podem causar dispersões diferentes em torno do verdadeiro valor de θ. Naturalmente, quanto menor for essa dispersão, melhor o estimador refletirá aquele valor. Então, a segunda exigência é: entre os estimadores não-tendeciosos de θ, escolhe-se aquele que tenha menor variância. Tal estimador, se existir, chama-se **estimador não-tendencioso de variância mínima** de θ.

Exemplo 7.1.3 Um pesquisador deseja estimar a produção média de um processo químico com base na observação da produção de três realizações: X_1, X_2, X_3 de um experimento. Considere os dois estimadores da média:

$$\widehat{\theta}_1 = \overline{x} = \frac{X_1 + X_2 + X_3}{3} \text{ Média Amostral}$$

$$\widehat{\theta}_2 = \overline{x}_p = \frac{X_1 + 2X_2 + X_3}{4} \text{ uma Média Ponderada}$$

Qual deve ser o preferido?

$$E(\widehat{\theta}_1) = E\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3}E[X_1 + X_2 + X_3] = \frac{1}{3}[\mu + \mu + \mu] = \mu$$

$$E(\widehat{\theta}_2) = E\left[\frac{X_1 + 2X_2 + X_3}{4}\right] = \frac{1}{4}E[X_1 + 2X_2 + X_3] = \frac{1}{4}[\mu + 2\mu + \mu] = \mu$$

Os dois são não-tendenciosos.

$$V(\widehat{\theta}_1) = V\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3^2}V[X_1 + X_2 + X_3] = \frac{1}{9}[\sigma^2 + \sigma^2 + \sigma^2] = \frac{3\sigma^2}{9} = \frac{\sigma^2}{3}$$

$$V(\widehat{\theta}_2) = V\left[\frac{X_1 + 2X_2 + X_3}{4}\right] = \frac{1}{4^2}V[X_1 + 2X_2 + X_3] = \frac{1}{16}[\sigma^2 + 4\sigma^2 + \sigma^2] = \frac{6\sigma^2}{16} = \frac{3\sigma^2}{8}$$

Logo, $\widehat{\theta}_1$ é melhor estimador que $\widehat{\theta}_2$.

iii) Consistência

Um estimador $\widehat{\theta}$ é consistente se, à medida que o tamanho da amostra aumenta, seu valor esperado converge para o parâmetro de interesse e sua variância converge para zero. Ou seja, $\widehat{\theta}$ é consistente se as duas propriedades seguintes são satisfeitas:

i) $\lim_{n\to\infty} E(\widehat{\theta}) = \theta$

ii) $\lim_{n\to\infty} V(\widehat{\theta}) = 0$

Note que, na definição de consistência, estamos implicitamente usando o fato que o estimador depende de n, o tamanho da amostra. Na definição do vício, o resultado deve valer para qualquer que seja n, isto é, $E(\widehat{\theta}) = \theta$, para todo n. Na definição de consistência, o estimador necessita ser não viciado apenas para valores grandes de n.

Erro Padrão

Quando o valor numérico ou estimativa de um parâmetro é reportado, é geralmente desejável dar alguma idéia da precisão da estimação. A medida da precisão geralmente empregada é o erro-padrão do estimador que está sendo usado. O erro-padrão é uma medida de variabilidade da distribuição de um estimador de θ e é dado pelo desvio padrão do estimador e representado por $EP(\widehat{\theta})$.

Exemplo 7.1.4 Para o exemplo 7.1.3, encontre o erro-padrão para os dois estimadores.

$$EP(\widehat{\theta}_1) = \sigma\sqrt{\frac{1}{3}} \qquad EP(\widehat{\theta}_2) = \sigma\sqrt{\frac{3}{3}}$$

Intervalos de Confiança

Em muitas situações, uma estimativa de um parâmetro não fornece informação completa para o profissional. Um estimador pontual com base em uma amostra produz um único número como estimativa do parâmetro. Muitas vezes, entretanto, queremos considerar, conjuntamente, o estimador e a precisão com que se estima o parâmetro. A forma usual de se fazer isso é através dos chamados **Intervalos de Confiança (I.C.)**.

Sejam então $x_1, x_2, \ldots, x_n$ uma amostra aleatória de uma população e θ o parâmetro de interesse. Sejam $\widehat{\theta}_0$, $\widehat{\theta}_1$ estatísticas tais que $P(\widehat{\theta}_0 < \theta < \widehat{\theta}_1) = 1 - \alpha$.

Então o intervalo $[\widehat{\theta}_1, \widehat{\theta}_2]$ é chamado intervalo de confiança de nível $100(1-\alpha)\%$ para o parâmetro θ. Usualmente toma-se $1-\alpha$ como $0,95$ ou $0,99$. Muitos estatísticos consideram a construção de intervalos de confiança o principal método de estudo de um parâmetro populacional através de uma amostra.

Exemplo 7.1.5 Consideremos uma população Normal com média μ e variância conhecida σ^2 e uma amostra dessa população. Sabemos, que a média dessa amostra tem distribuição Normal com média μ e variância $\frac{\sigma^2}{n}$, ou seja:

$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Fixando α em $0,05$, vemos, pela Tabela 10.1.1, disponível em Anexos 10, da distribuição Normal padronizada Z, que:

$$P(-1,96 < Z < 1,96) = 0,95$$

isto é,

$$P\left[-1,96 < \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1,96\right] = 0,95$$

Reescrevendo as desigualdades entre parênteses, obtemos:

$$P\left[\overline{x} - 1,96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1,96 \cdot \frac{\sigma}{\sqrt{n}}\right] = 0,95$$

Nesse caso, portanto, $\theta = \mu$, $\widehat{\theta}_0 = \overline{x} - 1,96 \cdot \frac{\sigma}{n}$, $\widehat{\theta}_1 = \overline{x} + 1,96 \cdot \frac{\sigma}{n}$

É importante observar que o nível de confiança se aplica ao processo de construção de intervalos, e não a um intervalo, e não a um intervalo específico. Para explicitar o conceito de intervalo de confiança, suponha que retiremos um grande número de amostras de tamanho n, fixo, da população e m estudo e que para cada amostra construamos um intervalo. Os limites dos intervalos resultantes serão diferentes. O verdadeiro valor do parâmetro estará contido, em média, em $100(1-\alpha)\%$ desses intervalos, ou seja, $100(1-\alpha)\%$ dos intervalos construídos abrangerão o verdadeiro valor do parâmetro (no caso, μ) mas cada intervalo contém, ou não contém, o parâmetro.

7.2 Intervalos de Confiança para a Média Populacional

A média é uma importante característica da população e por isso, é de interesse sua estimação via intervalos de confiança.

Uma notação útil será denominarmos $Z_{\frac{\alpha}{2}}$ como o valor de Z tal que:

$$P\left[-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

Exemplo 7.2.1 Se $\alpha = 0,05$, então, usando a Tabela 10.1.1, encontramos $Z_{0,025} = 1,96$. Se $\alpha = 0,05$, $Z_{0,005} = 2,58$.

i) Para a população Normal, com σ conhecido

Sabemos que: $\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Temos que:

$$P\left[-Z_{\frac{\alpha}{2}} < \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

expressão que nos dará o seguinte intervalo de confiança:

$$\left[\overline{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] \quad \text{ou} \quad \left[\overline{x} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}; \overline{x} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

A situação aqui considerada é, entretanto, muito artificial, pois, embora a hipótese de normalidade seja razoável em muitos casos práticos, dificilmente se conhece a variância de uma população quando sua média é desconhecida.

Exemplo 7.2.2 Para uma amostra de 50 observações de uma população normal com média desconhecida e desvio padrão $\sigma = 6$, seja $20,5$ a média amostral $\overline{x}$. Construir um intervalo de 95% de confiança para a média populacional.

$$\overline{x} = 20,5 \qquad n = 50 \qquad \sigma = 6$$

$$\left[20,5 - 1,96 \cdot \frac{6}{\sqrt{50}}; 20,5 + 1,96 \cdot \frac{6}{\sqrt{50}}\right] \Rightarrow [18,84;\ 22,16] \text{ ou } [18,84 \leqslant \mu \leqslant 22,16]$$

O resultado obtido $[18,84;\ 22,16]$ é um intervalo de 95% de confiança para a média populacional μ, calculado com base na amostra observada.

ii) Para a população Normal, com σ desconhecido

Nesse caso, o intervalo de confiança é calculado utilizando-se uma estatística:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

em que s é o estimador do desvio padrão. Essa estatística tem a distribuição conhecida como **t de *Student*** com $\nu = n - 1$ **graus de liberdade**, sendo n o tamanho da amostra e ν o parâmetro da distribuição. A forma da distribuição t é parecida com a Normal; sua forma mais aberta reflete a maior variabilidade esperada em pequenas amostras. Essa distribuição tem média 0 e seu desvio padrão varia com o tamanho da amostra, e é maior do que 1. Aumentando-se n, a distribuição t tende para a Normal.

Observação 7.2.1 O número de graus de liberdade para uma coleção de dados amostrais é o número de valores amostrais que podem variar depois que certas restrições tiverem sido impostas aos dados amostrais.

Para cada valor de ν temos uma distribuição diferente, e assim, em princípio, precisaríamos de um grande número de tabelas dessa distribuição. Uma situação intermediária consiste em apresentar, apenas para algumas combinações de valores de ν e α, os valores $t_{\frac{\alpha}{2},(\nu)}$ tais que:

$$P\left[-t_{\frac{\alpha}{2},(\nu)} < t_{(\nu)} < t_{\frac{\alpha}{2},(\nu)}\right] = 1 - \alpha$$

Com essas aplicações, o intervalo de confiança para μ é dado por:

$$\left[\overline{x} - t_{\frac{\alpha}{2},(n-1)} \cdot \frac{s}{\sqrt{n}};\ \overline{x} + t_{\frac{\alpha}{2},(n-1)} \cdot \frac{s}{\sqrt{n}}\right]$$

Exemplo 7.2.3 Em teste de sensitividade levado a efeito em 18 válvulas de certa marca, obtiveram-se média de $3,2$ microvolts e variância $0,20$ microvolt. Determinar um intervalo de 95% de confiança para a sensitividade média da população de válvulas.

$$n = 18 \qquad 1-\alpha = 0,95 \qquad \frac{\alpha}{2} = 0,025 \qquad \nu = 18-1 = 17 \qquad t_{0,025(17)} = 2,11$$

$$\left[3,2 - 2,11 \cdot \frac{0,477}{\sqrt{18}};\ 3,2 + 2,11 \cdot \frac{0,477}{\sqrt{18}}\right] \Rightarrow [2,98;\ 3,42]$$

7.3 Intervalo de Confiança para a Diferença entre duas Médias Populacionais de duas Distribuições Normais

i) Com variâncias populacionais conhecidas

Se $\overline{x}_1 - \overline{x}_2$ forem as médias de duas amostras aleatórias independentes de tamanhos n_1 e n_2, provenientes de populações com variâncias conhecidas σ_1^2 e σ_2^2, respectivamente, então um intervalo de confiança de $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ é:

$$\left[(\overline{x}_1 - \overline{x}_2) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};\ (\overline{x}_1 - \overline{x}_2) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Exemplo 7.3.1 Testes de resistência à tensão foram feitos em duas estruturas contendo dois teores de alumínio. Essas estruturas foram usadas na fabricação das asas e um avião comercial. De experiências passadas com o processo de fabricação dessas estruturas e com o procedimento de testes, os desvios padrões das resistências à tensão são considerados conhecidos e iguais $\sigma_1 = 1,0$ e $\sigma_2 = 1,5$. Os dados obtidos com amostras $n_1 = 10$ e $n_2 = 12$ mostraram que $\overline{x}_1 = 87,6$ e $\overline{x}_2 = 74,5$. Se μ_1 e μ_2 denotarem as resistências médias verdadeiras à tensão para os dois tipos (dois teores diferentes) da estrutura, então podemos achar um intervalo de confiança de 90% para a diferença na resistência média $\mu_1 - \mu_2$, conforme segue:

$$\left[(87,6-74,5) - 1,645 \cdot \sqrt{\frac{(1,0)^2}{10} + \frac{(1,5)^2}{12}};\ (87,6-74,5) + 1,645 \cdot \sqrt{\frac{(1,0)^2}{10} + \frac{(1,5)^2}{12}}\right]$$

$$\Rightarrow [12,22;\ 13,98]$$

ii) Com variâncias populacionais desconhecidas, porém iguais:

Se $\overline{x}_1$, $\overline{x}_2$, s_1^2, s_2^2 forem as médias e as variâncias de duas amostras aleatórias de tamanhos n_1 e n_2, respectivamente, provenientes de duas populações normais independentes, com variâncias desconhecidas, porém iguais, então um intervalo de confiança de $100(1-\alpha)\%$ para a diferença nas médias $\mu_1 - \mu_2$ é:

$$\left[(\overline{x}_1 - \overline{x}_2) - t_{\frac{\alpha}{2},(n_1+n_2-2)} s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; (\overline{x}_1 - \overline{x}_2) + t_{\frac{\alpha}{2},(n_1+n_2-2)} s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

em que

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

é a estimativa combinada do desvio padrão comum da população comum.

Exemplo 7.3.2 Um artigo no jornal *Hazardous Waste and Hazardous Materials* (Vol. 6, 1989) reportou os resultados de uma análise do peso de cálcio em cimento padrão e em cimento contendo chumbo. Níveis reduzidos de cálcio indicaram que o mecanismo de hidratação do cimento foi bloqueado, permitindo à água atacar várias localizações na estrutura de cimento. Dez amostras de cimento padrão tiveram um teor médio percentual em peso de cálcio de $\overline{x}_1 = 90,0$, com um desvio padrão da amostra de $s_1 = 5,0$, enquanto 15 amostras do cimento com chumbo tiveram um teor médio percentual em peso de cálcio de $\overline{x}_2 = 87,0$, com um desvio padrão da amostra de $s_2 = 4,0$.

Considere que o teor percentual em peso de cálcio seja normalmente distribuído e que ambas as populações tenham o mesmo desvio padrão. Encontre o intervalo de confiança de 95% para as diferenças de médias.

$$s_p = \sqrt{\frac{9 \cdot (5,0)^2 + 14 \cdot (4,0)^2}{(10 + 15 - 2)}} = 4,4$$

$$\left[(90 - 87) - 2,069 \cdot 4,4\sqrt{\frac{1}{10} + \frac{1}{15}};\ \ (90 - 87) + 2,069 \cdot 4,4\sqrt{\frac{1}{10} + \frac{1}{15}}\right]$$

$$\Rightarrow [-0,72;\ \ 6,72]$$

Note que o intervalo de confiança de 95% inclui o zero; assim, nesse nível de confiança, não podemos concluir que haja uma diferença nas médias, ou seja, não há evidências de que o cimento contendo chumbo tenha afetado o percentual médio em peso de cálcio; desse modo, não podemos afirmar que a presença do chumbo afete esse aspecto do mecanismo de hidratação com um nível de 95% de confiança.

iii) Com variâncias populacionais desconhecidas e diferentes:

Se $\overline{x}_1$, $\overline{x}_2$, s_1^2, s_2^2 forem as médias e as variâncias de duas amostras aleatórias de tamanhos n_1 e n_2, respectivamente, provenientes de duas populações normais independentes, com variâncias desconhecidas e desiguais, então um intervalo de confiança de $100(1-\alpha)\%$ para a diferença nas médias $\mu_1 - \mu_2$ é:

$$\left[(\overline{x}_1 - \overline{x}_2) - t_{\frac{\alpha}{2},(\nu)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}};\ \ (\overline{x}_1 - \overline{x}_2) + t_{\frac{\alpha}{2},(\nu)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

em que

$$\nu = \frac{\left[\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

7.4 Intervalo de Confiança para a Proporção populacional

Seja X o número de elementos de uma amostra de tamanho n que apresentam a característica de interesse. Para estabelecer um I.C. para a proporção populacional p temos que usar o seguinte procedimento:

$$\left[\overline{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{p}(1-\overline{p})}{n}};\ \overline{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right]$$

que será o intervalo de $100(1-\alpha)\%$ de confiança para p.

Exemplo 7.4.1 Numa pesquisa de mercado, $n = 400$ pessoas foram entrevistas sobre determinado produto, e 60% delas preferiram a marca A. Aqui $\overline{p} = 0,60$ e um I.C. para p com 95% de confiança será:

$$\left[0,60 - 1,96 \cdot \sqrt{\frac{0,60 \cdot 0,40}{400}};\ 0,60 + 1,96 \cdot \sqrt{\frac{0,60 \cdot 0,40}{400}}\right]$$

$$\Rightarrow [0,552;\ 0,648]$$

7.5 Intervalo de Confiança para a Diferença entre duas Proporções populacionais

Se $\overline{p}_1$ e $\overline{p}_2$ forem as proporções amostrais de observação em duas amostras aleatórias e independentes, de tamanhos n_1 e n_2, que pertençam à característica de interesse, então um intervalo de confiança de $100(1-\alpha)\%$ para a diferença $p_1 - p_2$ será:

$$\left[(\overline{p}_1 - \overline{p}_2) - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{p}_1 \cdot (1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2 \cdot (1-\overline{p}_2)}{n_2}};\ (\overline{p}_1 - \overline{p}_2) + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{p}_1 \cdot (1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2 \cdot (1-\overline{p}_2)}{n_2}}\right]$$

Exemplo 7.5.1 Em um processo de fabricação de mancais para eixos de manivela foi retirada uma amostra $n_1 = 85$ e encontrou-se uma proporção de defeituosos de $\overline{p}_1 = 0,12$. Suponha que uma modificação seja feita no processo de acabamento da superfície e que, subseqüentemente, obtenha-se uma segunda amostra aleatória de 85 eixos. O número de eixos defeituosos nessa amostra é 8. Obtenha um intervalo de confiança para a diferença da proporção de mancais defeituosos produzidos pelos dois processos.

$$\left[(0,12 - 0,09) - 1,96 \cdot \sqrt{\frac{0,12 \cdot (0,88)}{85} + \frac{0,09 \cdot (0,91)}{85}};\ (0,12 - 0,09) + 1,96 \cdot \sqrt{\frac{0,12 \cdot (0,88)}{85} + \frac{0,09 \cdot (0,91)}{85}}\right]$$

$$\Rightarrow [-0,06;\ 0,12]$$

Esse intervalo de confiança inclui o zero; assim, baseado nos dados das amostras, com

95% de confiança não é possível afirmar que mudanças feitas no processo de acabamento da superfície tenham reduzido a proporção de mancais com eixos defeituosos sendo produzidos.

7.6 Intervalo de Confiança para a Variância Populacional

Vamos estabelecer uma estimativa intervalar para σ^2 utilizando a estatística:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Essa estatística tem distribuição $\chi^2_{(n-1)}$, chamada "qui-quadrado"(χ, letra grega pronunciada como "qui"), se a população é normal. O subscrito $n-1$ é o número de graus de liberdade da distribuição, quando o tamanho da amostra é n.

A distribuição qui-quadrado não é simétrica, diferentemente das distribuições normal e t de Student. À medida que número de graus de liberdade aumenta, a distribuição se torna mais simétrica se aproximando da normal. Seus valores podem ser zero ou positivos, mas nunca podem ser negativos.

Se $\chi^2_{\frac{\alpha}{2}}$ e $\chi^2_{1-\frac{\alpha}{2}}$ são valores da distribuição qui-quadrado que deixam áreas $1-\frac{\alpha}{2}$ e $\frac{\alpha}{2}$, respectivamente à esquerda, temos:

$$P\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\right] = 1-\alpha$$

Um intervalo de confiança de $100(1-\alpha)\%$ de confiança para a variância de uma população normal é:

$$\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}};\ \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\right]$$

em que s^2 é a variância de uma amostra de tamanho n e $\chi^2_{\frac{\alpha}{2}}$ e $\chi^2_{1-\frac{\alpha}{2}}$ são valores obtidos de uma tabela de distribuição qui-quadrado com $n-1$ graus de liberdade.

Exemplo 7.6.1 Têm-se os seguintes pesos, em gramas, de 10 pacotes postais remetidos por certas empresas:

46,4	46,1	45,8	47,0	46,1	45,9	45,8	46,9	45,2	46,0

Admitindo Normal a distribuição dos pesos, determine um intervalo de 95% de confiança para a variância dos pesos de todos os pacotes expedidos pela empresa:

$\bar{x} = 46,12$ $\quad s = 0,535$ $\quad s^2 = 0,286$ $\quad \alpha = 0,05$ $\quad \nu = 10–1 = 9$

$\chi^2_{0,975,(9)} = 2,70$ $\qquad\qquad$ $\chi^2_{0,025,(9)} = 19,023$

Portanto:

$$\left[\frac{9 \cdot 0,286}{19,023};\ \frac{9 \cdot 0,286}{2,70}\right] \Rightarrow [0,135;\ 0,953]$$

7.7 Intervalo de Confiança para a razão de Variâncias populacionais de duas distribuições Normais

Se s_1^2, s_2^2 forem as variâncias de amostras aleatórias de tamanhos n_1 e n_2, respectivamente, provenientes de duas populações normais independentes, com variâncias desconhecidas σ_1^2, σ_2^2, então um intervalo de confiança de $100(1-\alpha)\%$ para a razão $\dfrac{\sigma_1^2}{\sigma_2^2}$ será:

$$\left[\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\frac{\alpha}{2},(n_1-1,n_2-1)}};\ \frac{s_1^2}{s_2^2} \cdot F_{\frac{\alpha}{2},(n_2-1,n_1-1)}\right]$$

A distribuição F possui dois parâmetros: os graus de liberdade das variâncias amostrais no numerador e no denominador da estatística $\dfrac{s_1^2}{s_2^2}$. O número de graus de liberdade do numerador deve ser sempre mencionado em primeiro lugar. Como a troca de posição dos graus de liberdade modifica a distribuição, a ordem tem grande importância. Esta distribuição não é simétrica e sua forma exata depende dos dois diferentes graus de liberdade. Os valores dessa distribuição não podem ser negativos.

Exemplo 7.7.1 Uma companhia fabrica propulsores para uso em motores de turbina de avião. Uma das operações envolve esmerilhar o acabamento de uma superfície particular para um componente de liga de titânio. Dois processos diferentes para esmerilhar podem ser usados, podendo produzir peças com iguais rugosidades médias na superfície. Uma amostra aleatória de $n_1 = 11$ peças, proveniente do primeiro processo, resulta em um desvio padrão de $s_1 = 5,1$ micropolegadas. Uma amostra aleatória de $n_2 = 16$ peças, proveniente do segundo processo, resulta em um desvio padrão de $s_2 = 4,7$ micropolegadas. Encontre um intervalo de confiança de 90% para a razão de duas variâncias $\dfrac{\sigma_1^2}{\sigma_2^2}$. (Considere que os dois processos são independentes e que a rugosidade na superfície seja normalmente distribuída).

$$\left[\frac{5,1^2}{4,7^2} \cdot 0,39;\ \frac{5,1^2}{4,7^2} \cdot 2,85\right]$$

$$\Rightarrow [0,46;\ 3,56]$$

Uma vez que esse intervalo de confiança inclui a unidade, não podemos afirmar que os desvios padrões da rugosidade da superfície para os dois processos sejam diferentes com um nível de 90% de confiança.

7.8 Exercícios

7.8.1 Em um esforço para estimar a quantia gasta por cliente para jantar em um grande restaurante de Atlanta, foram coletados os dados de uma amostra de 49 clientes em um período de três semanas. Considere um desvio padrão da população de 2,50 dólares

a) Qual é o erro padrão da média?

b) Se a média da amostra é de $22,60$ dólares, qual é o intervalo de confiança de 95% pra a média da população?

7.8.2 No teste de um novo método de produção, 18 empregados foram aleatoriamente selecionados e solicitados a tentar o novo método. A taxa média de produção da amostra para os 18 empregados foi de 80 peças por hora e o desvio padrão foi de 10 peças por hora. Forneça intervalos de confiança de 90% e de 95% para a taxa média de produção da população para o novo método, considerando que a população tenha uma distribuição normal de probabilidade.

7.8.3 O Departamento de Transportes dos EUA relatou o número de quilômetro que os residentes das áreas metropolitanas percorrem de carro por dia (*1994 Information Please Environmental Almanac*). Suponha que uma amostra aleatória simples de 15 residentes de Cleveland forneceu os seguintes dados sobre os quilômetros percorridos de carro por dia.

20	16	23	18	29
20	11	16	10	19
28	17	22	22	32

Calcule a estimativa do intervalo de confiança de 95% do número médio de quilômetros da população que os residente de Cleveland percorrem de carro por dia.

7.8.4 Um fabricante produz anéis para pistões de um motor de automóveis. É sabido que o diâmetro do anel é distribuído de forma aproximadamente normal e tem um desvio padrão de $0,001$ mm. Uma amostra aleatória de 15 anéis tem um diâmetro médio de $\bar{x} = 74,036$ mm. Construa um intervalo de confiança de 99% para o diâmetro médio do anel do pistão.

7.8.5 Um fabricante de calculadoras eletrônicas está interessado em estimar a fração de unidades defeituosas produzidas. Uma amostra aleatória de 800 calculadoras contém 10 defeitos. Encontre um intervalo de confiança de 90%.

7.8.6 Um levantamento pela *Wirthlin Worldwide* coletou dados sobre as atitudes com relação à qualidade dos serviços aos clientes em lojas de atacado. O levantamento concluiu que 28% dos americanos consideram que o serviço aos clientes é melhor hoje do que era há dois anos (*USA Today*, 20 de janeiro de 1998). Se 650 adultos foram incluídos na amostra, desenvolva um intervalo de confiança de 95% para a proporção da população de adultos que sente o serviço ao cliente é melhor hoje do que era há dois anos.

7.8.7 A percentagem de titânio em uma liga usada na fundição de aeronaves é medida em 51 peças selecionadas aleatoriamente. O desvio padrão da amostra é $s = 0,37$. Construa um intervalo com 90% de confiança para σ^2.

7.8.8 Um rebite deve ser inserido em um orifício. Se o desvio padrão do diâmetro do orifício exceder $0,01$ mm, haverá uma probabilidade inaceitavelmente alta de que o rebite não se ajuste. Uma amostra aleatória de $n = 15$ peças é selecionada e o diâmetro do orifício é medido. O desvio padrão das medidas do diâmetro d orifício é $s = 0,008$ mm. Construa um intervalo de confiança de 99% para σ.

7.8.9 Estão sendo estudadas as taxas de queima de dois diferentes propelentes sólidos usados no sistema de escapamento de aeronaves. Sabe-se que ambos os propelentes têm aproximadamente o mesmo desvio padrão de taxa de queima, ou seja, $\sigma_1 = \sigma_2 = 3$ cm/s. Duas amostras aleatórias de $n_1 = 20$ e $n_2 = 20$ espécimes são testadas; as taxas médias de queima das amostras são $\bar{x}_1 = 18$ cm/s e $\bar{x}_2 = 24$ cm/s. Construa um intervalo de confiança de 95% para a diferença nas médias $\mu_1 - \mu_2$. Qual é o significado prático desse intervalo?

7.8.10 Duas formulações diferentes de um combustível oxigenado de um motor devem ser testadas com a finalidade de estudar seus números de octanagem na estrada. A variância do número de octanagem na estrada no caso da formulação 1 $\sigma_1^2 = 1,2$ e a variância da formulação dois é $1,0$. Duas amostras aleatórias de tamanho $n_1 = 15$ e $n_2 = 20$, sendo os números médios observados de octanagem dados por $\bar{x}_1 = 89,6$ e média amostral 2 igual $90,8$. Considere normalidade e construa um intervalo de confiança de 95% para a diferença nos números médios observados de octanagem na estrada.

7.8.11 Dois catalisadores podem ser usados em um processo químico em batelada. Doze bateladas foram preparadas usando catalisador 1, resultando em um rendimento médio de 86 e um desvio padrão da amostra igual a 3. Quinze bateladas foram preparadas, usando catalisador 2, resultando em um rendimento médio de 89, com um desvio padrão de 2. Considere que as medidas de rendimento sejam distribuídas aproximadamente de forma normal, com o mesmo desvio padrão. Encontre um intervalo de confiança de 95% para a diferença entre os rendimentos médios.

7.8.12 O diâmetro de bastões de aço, fabricados em duas máquinas extrusoras diferentes, está sendo investigado. Duas amostras aleatórias de tamanho $n_1 = 15$ e $n_2 = 17$ são selecionadas e as médias e as variâncias das amostras são $\bar{x}_1 = 8,73$, $s_1^2 = 0,35$, $\bar{x}_2 = 8,68$ e $s_2^2 = 0,40$, respectivamente. Suponha $\sigma_1^2 \neq \sigma_2^2$ e que os dados sejam retirados de uma população normal. Construa um intervalo de confiança de 95% para a diferença no diâmetro médio dos bastões. Interprete esse intervalo.

7.8.13 Dois tipos diferentes de máquinas de injeção-moldagem são usadas para formar peças de plásticos. Uma peça é considerada defeituosa se ela tiver excesso de encolhimento ou se for descolorida. Duas amostras aleatórias, cada uma de tamanho 300, são selecionadas e 15 peças defeituosas são encontradas na amostra da máquina 1, enquanto 8 peças defeituosas são encontradas na amostra da máquina 2. Use um intervalo de confiança com 95% e conclua se ambas as máquinas produzem ou não a mesma fração de peças defeituosas.

7.8.14 Duas companhias químicas podem fornecer uma matéria-prima, cuja concentração de um determinado elemento é importante. A concentração média para ambos os fornecedores é a mesma, porém suspeitamos de que a variabilidade na concentração pode diferir entre as duas companhias. O desvio padrão da concentração em uma amostra aleatória $n_1 = 10$ bateladas produzidas pela companhia 1 é $s_1 = 4,7$ g/l, enquanto para a companhia 2, uma amostra aleatória de $n_2 = 16$ bateladas resulta em $s_2 = 5,8$ g/l. Construindo um intervalo de 90% de confiança, há evidência suficiente para concluir que as variâncias das duas populações difiram?

7.8.15 O gerente de controle de qualidade de uma fábrica de lâmpadas precisa estimar a média aritmética da vida útil de uma grande remessa de lâmpadas. O desvio padrão do processo é conhecido como sendo igual a 100 horas. Uma amostra aleatória de 64 lâmpadas indicou uma média aritmética da vida útil da amostra igual a 350 horas.

a) Construa uma estimativa para o intervalo de confiança de 95% para a verdadeira média aritmética da população relativa à vida útil de lâmpadas nesta remessa.

b) Você acha que o fabricante tem o direito de declarar que as lâmpadas duram em média, 400 horas? Explique.

7.8.16 Para analisar a variação de cápsulas de suplemento vitamínico, foram selecionadas aleatoriamente e pesadas 14 cápsulas. Os resultados da amostra mostram que o desvio padrão amostral foi de $0,020$. Para que a variação esteja em um nível aceitável o desvio padrão populacional dos pesos deve ser menor que $0,015$ miligrama. Construa um intervalo com 90% de confiança para σ e responda se a variação está em um nível aceitável. Justifique.

7.9 Testes de Hipóteses

Com frequência o pesquisador tem idéia ou é informado sobre um possível valor do parâmetro e para dar continuação a sua pesquisa ele é obrigado a aceitar ou rejeitar tal valor. Cabe então perguntar-se: O que ele deveria fazer para verificar se a idéia ou informação sobre o parâmetro é ou não correta?

Por exemplo, na compra de uma máquina de empacotar café, o fabricante vende a máquina afirmando que o peso médio dos pacotes (μ) é de 1000 g. O que faria o responsável pelo controle de qualidade para verificar se o peso médio dos pacotes é na verdade 1000 g?

A partir da informação obtida numa amostra aleatória é estatisticamente possível tomar uma decisão quanto à aceitação ou rejeição da afirmação feita sobre o parâmetro. *A teoria de decisão preocupa-se em construir testes que permitam aceitar ou rejeitar afirmações feitas sobre o parâmetro.*

Uma afirmação feita sobre um parâmetro é chamada de HIPÓTESE ESTATÍSTICA. Logo, a teoria de decisão preocupa-se em construir testes que permitam aceitar ou rejeitar as hipóteses estatísticas, a partir da informação obtida numa amostra aleatória.

Hipótese Estatística e Erros Envolvidos num Processo de Decisão:

Uma afirmação feita, a priori, sobre um parâmetro em estudo é chamada de HIPÓTESE ESTATÍSTICA. Um TESTE DE HIPÓTESE é o critério do qual se lança mão para tomar a decisão de aceitar ou rejeitar a hipótese estatística.

A hipótese estatística divide-se em duas partes complementares:

Hipótese nula (H_0) – Aquela que será testada.

No exemplo:

$H_0 : \mu = 1000\ g$ ou $H_0 : \mu \geqslant 1000\ g$ ou $H_0 : \mu \leqslant 1000\ g$

Hipótese alternativa (H_1) - Aquela que afirma o contrário de H_0.

No exemplo:

$H_1 : \mu \neq 1000\ g$	ou	$H_1 : \mu < 1000\ g$	ou	$H_1 : \mu > 1000\ g$
Bilateral		**Unilateral à esquerda**		**Unilateral à direita**

Observação 7.9.1

i) Estabelecer H_0 e H_1 depende exclusivamente da natureza do problema em estudo.

ii) Por convenção os símbolos $=$, $\geqslant$ e $\leqslant$ estão associadas com H_0 e os símbolos $\neq$, $<$ e $>$ com H_1.

iii) A rejeição de H_0 implicará na aceitação de H_1.

Como a tomada de decisão sobre a aceitação ou rejeição de uma hipótese está baseada apenas na informação dos dados amostrais, dois tipos de erros podem ser cometidos:

i) **ERRO TIPO I**: Rejeitar H_0 quando ela é verdadeira;

ii) **ERRO TIPO II**: Não rejeitar H_0 quando ela é falsa.

A probabilidade de se cometer ERRO TIPO I é denotada por α e é chamada de *Nível de Significância* do teste. A probabilidade de ocorrência do ERRO TIPO II é denotada por β.

Para que um teste de hipótese seja considerado bom deve-se ter uma pequena probabilidade de rejeitar H_0 se esta for verdadeira, mas também, uma grande probabilidade de rejeitá-la se ela for falsa.

O quadro abaixo resume a natureza dos erros envolvidos no processo de decisão através dos testes de hipóteses:

	H_0 **Verdadeira**	H_0 **Falsa**
Rejeição de H_0	ERRO TIPO I	Decisão Correta
Não rejeição de H_0	Decisão Correta	ERRO TIPO II

Nesta disciplina serão realizados testes de hipótese em que apenas o ERRO TIPO I é controlado (chamados TESTES DE SIGNIFICÂNCIA), isto devido a que o controle do outro tipo de erro precisa de técnicas mais avançadas.

Observação 7.9.2 Em geral o valor do nível de significância (α) será considerado pequeno ou grande dependendo da precisão que o pesquisador deseja nas suas conclusões. Porém, é freqüente usar como referência $\alpha = 0,05$ ou $\alpha = 0,01$.

Mecânica Operacional de Aplicação de Testes

Passos para execução de um teste de hipótese:

1. Formular as hipóteses H_0 e H_1, segundo a natureza do problema em estudo.
2. Escolher a estatística de teste adequada.
3. Especificar o nível de significância α e estabelecer a *região crítica* (ou *região de rejeição*). A região crítica será determinada por meio da hipótese H_1 e de acordo com α.
4. Calcular o valor da estatística de teste com base em uma amostra de tamanho n extraída da população.
5. Se o valor da estatística pertencer a região crítica, rejeita-se H_0, caso contrário, não é possível rejeitar H_0 com nível de significância α.

7.10 Teste para a Média Populacional de uma População Normal Quando o Desvio Padrão da População é Conhecido

Neste caso a estatística de teste é dada por:

$$Z_c = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Exemplo 7.10.1 Um fabricante de fio de arame alega que seu produto tem uma resistência média à ruptura superior a 10 Kg, com desvio padrão de 0,5 Kg. Um consumidor resolve testar essa afirmativa. Extrai uma amostra de 50 peças de arame, a qual acusou média de 10,4 Kg. Com $\alpha = 0,05$ é possível afirmar que é válida a alegação do fabricante?

$$\begin{cases} H_0: & \mu \leqslant 10 \\ H_1: & \mu > 10 \end{cases} \quad \text{ou} \quad \begin{cases} H_0: & \mu = 10 \\ H_1: & \mu > 10 \end{cases}$$

$$Z_c = \frac{10,4 - 10}{\frac{0,5}{\sqrt{50}}} = 5,66$$

Como 5,66 é maior que 1,645 (o valor limite da região critica), devemos rejeitar H_0, ou seja, há evidência de que a resistência média seja superior a 10 Kg, como alegado pelo fabricante, com $\alpha = 0,05$.

7.11 Teste de Hipóteses para a Média Populacional de uma População Normal com Desvio Padrão Populacional Desconhecido

Neste caso a estatística de teste é dada por:

$$t_c = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Exemplo 7.11.1 Um teste de resistência à ruptura feito em uma amostra com seis cordas acusou resistência média de 3770 Kg e desvio padrão de 66 Kg. O fabricante afirma que seu produto tem resistência média superior a 3650 Kg. Pode-se justificar a alegação do fabricante, no nível de significância de 1%?

$$\begin{cases} H_0: & \mu \leqslant 3650 \\ H_1: & \mu > 3650 \end{cases} \quad \text{ou} \quad \begin{cases} H_0: & \mu = 3650 \\ H_1: & \mu > 3650 \end{cases}$$

$$t_c = \frac{3770 - 3650}{\frac{66}{\sqrt{6}}} = 4,45$$

Como o valor calculado 4,45 é superior ao valor 3,365, que é o valor t tabelado com $n-1$ graus de liberdade, então devemos rejeitar H_0, ou seja, há evidência de que a resistência média seja superior a 3650 Kg, como alegado pelo fabricante, com $\alpha = 0,01$.

7.12 Teste de Hipóteses para Diferença de Duas Médias de Populações Normais, Quando as Variâncias Populacionais são Conhecidas

Assuma que os parâmetros para duas populações são: μ_1, μ_2, σ_1 e σ_2. Quando σ_1 e σ_2 são conhecidos, a estatística de teste (Z efetivo) é:

$$Z_c = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Exemplo 7.12.1 Um idealizador de produtos está interessado em reduzir o tempo de secagem de um zarcão. Duas formulações de tinta são testadas; a formulação 1 tem uma química padrão e a formulação 2 tem um novo ingrediente para secagem, que deve reduzir o tempo de secagem. Da experiência, sabe-se que o desvio padrão do tempo de secagem é 8 minutos e essa variabilidade inerente não deve ser afetada pela adição do novo ingrediente. Dez espécimes são pintados com a formulação 1 e outros dez espécimes são pintados com a formulação 2. Os 20 espécimes são pintados em uma ordem aleatória. Os tempos médios de secagem das duas amostras são $\overline{x}_1 = 121\ \mathrm{min}$ e $\overline{x}_2 = 112\ \mathrm{min}$, respectivamente. Quais as conclusões que o idealizador de produtos pode tirar sobre a eficiência do novo ingrediente, usando $\alpha = 0,05$?

$$\begin{cases} H_0: & \mu_1 \leqslant \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases} \qquad \text{ou} \qquad \begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases}$$

$$Z_c = \frac{121 - 112}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} = 2,52$$

Como $2,52$ é maior que $1,645$ (o valor limite da região critica), devemos rejeitar H_0, ou seja, a adição do novo ingrediente à tinta reduz significativamente o tempo de secagem, com $\alpha = 0,05$.

7.13 Teste de Hipóteses para Diferença de Duas Médias de Populações Normais, Quando as Variâncias Populacionais são Desconhecidas, Porém Iguais

A estatística do teste será dada por:

$$t_c = \frac{\overline{x}_1 - \overline{x}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Sendo que:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}}$$

Exemplo 7.13.1 Duas técnicas de venda são aplicadas por dois grupos de vendedores com a expectativa de que a técnica B produza melhores resultados. A técnica A foi utilizada em uma amostra de 12 vendedores apresentando os seguintes resultados: $\overline{x}_A = 68$ vendas e variância igual a 50. A técnica B foi utilizada em uma amostra de 15 vendedores e os resultados foram: $\overline{x}_B = 76$ vendas e variância igual a 75. Teste com nível de significância de 5%, se há evidências para afirmar que em média a técnica B apresenta resultados superiores.

$$\begin{cases} H_0: & \mu_A \geqslant \mu_B \\ H_1: & \mu_A < \mu_B \end{cases} \qquad \text{ou} \qquad \begin{cases} H_0: & \mu_A = \mu_B \\ H_1: & \mu_A < \mu_B \end{cases}$$

$$s_p = \sqrt{\frac{11 \cdot 50 + 14 \cdot 75}{25}} = 8$$

$$t_c = \frac{68 - 76}{8 \cdot \sqrt{\frac{1}{12} + \frac{1}{15}}} = -2.58$$

Como $-2,58$ está abaixo do valor da região crítica $-1,7081$ (o valor limite da região critica), devemos rejeitar H_0, ou seja, existe evidência de que a técnica B produz melhores resultados do que a técnica A, com $\alpha = 0,05$.

7.14 Teste de Hipóteses para Diferença de Duas Médias de Populações Normais, Quando as Variâncias Populacionais são Desconhecidas e Diferentes

A estatística do teste será dada por:

$$t_c = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Sendo que o número de graus de liberdade do valor tabelado será dado por:

$$\nu = \frac{\left[\left(\frac{s_1^1}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

Exemplo 7.14.1 Um fabricante de unidades de vídeos está testando dois projetos de microcircuitos para determinar se eles produzem correntes médias equivalentes. A engenharia de desenvolvimento obteve os seguintes dados:

Projeto 1	$n_1 = 15$	$\bar{x}_1 = 24,2$	$s_1^2 = 10$
Projeto 2	$n_2 = 10$	$\bar{x}_2 = 23,9$	$s_2^2 = 20$

Usando $\alpha = 0,10$, determine se há qualquer diferença na corrente média entre os dois projetos, supondo que ambas as populações sejam normais, embora as variâncias populacionais são desconhecidas e diferentes.

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases}$$

$$t_c = \frac{24,2 - 23,9}{\sqrt{\frac{10}{15} + \frac{20}{10}}} = 0,18$$

$$\nu = \frac{\left[\left(\frac{10}{15}\right) + \left(\frac{20}{10}\right)\right]^2}{\frac{\left(\frac{10}{15}\right)^2}{16} + \frac{\left(\frac{20}{10}\right)^2}{11}} - 2 = 16$$

$\therefore$ O valor tabelado é $1,746$.

Como $0,18$ está entre $-1,746$ e $1,746$, não devemos rejeitar H_0, ou seja, não existe evidência de que as correntes médias sejam diferentes, com $\alpha = 0,10$.

7.15 Teste t Emparelhado: Comparando Duas Amostras Relacionadas

Os procedimentos de testes de hipóteses examinados até agora possibilitam que você compare examine diferenças entre duas populações independentes, com base em amostras que contenham dados numéricos. Nesta seção, será desenvolvido um procedimento para analisar a diferença entre as médias aritméticas de dois grupos, quando os dados numéricos selecionados são obtidos a partir de populações relacionadas, ou seja, quando os resultados do primeiro grupo não são independentes dos resultados do segundo grupo. Esta característica de dependência dos dois grupos pode ocorrer devido ao fato de os itens ou indivíduos serem colocados em pares, ou combinados, de acordo com alguma característica, ou em decorrência de as medições repetidas serem obtidas a partir de um mesmo conjunto de itens ou indivíduos. Em qualquer um desses dois casos, a variável de interesse representa a diferença entre os valores das observações, e não os valores das próprias observações.

Independentemente de serem utilizadas amostras combinadas (em pares) ou medições repetidas, o objetivo é estudar as diferenças entre as duas medições, reduzindo o efeito da variabilidade decorrente dos próprios itens ou indivíduos.

A estatística do teste é dada por:

$$t_c = \frac{\overline{D}}{\frac{s_D}{\sqrt{n}}}$$

Onde:

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} \qquad s_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}}$$

sendo que D_i representa a diferença dos pares.

A estatística do teste segue uma distribuição t, com n–1 graus de liberdade.

Exemplo 7.15.1 Para testar a diferença entre dois métodos de produção, uma amostra aleatória de seis trabalhadores é usada. Os dados sobre os tempos de término para os seis trabalhadores estão listados na tabela abaixo. Use $\alpha = 5\%$.

Trabalhador	Tempo de término para o método 1 (minutos)	Tempo de término para o método 2 (minutos)	Diferença nos tempos de término
1	6,0	5,4	0,6
2	5,0	5,2	–0,2
3	7,0	6,5	0,5
4	6,2	5,9	0,3
5	6,0	6,0	0,0
6	6,4	5,8	0,6

$$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D \neq 0 \end{cases}$$

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{1,8}{6} = 0,30$$

$$s_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}} = \sqrt{\frac{0,56}{5}} = 0,335$$

$$t_c = \frac{\overline{D}}{\frac{s_D}{\sqrt{n}}} = \frac{30}{\frac{0,335}{\sqrt{6}}} = 2,19$$

Como $2,19$ não está na região de rejeição, já que o valor tabelado é $2,571$, a hipótese nula não é rejeitada. Portanto não é possível afirmar que há diferença entre os tempos médios dos dois métodos de produção.

7.16 Teste para o Valor da Proporção Populacional

Proporção: Uma fração ou porcentagem que indica uma parte da população ou amostra que tem um particular traço de interesse.

A proporção amostral é denotada por $\overline{p}$ onde:

$$\overline{p} = \frac{\text{número de sucessos na amostra}}{\text{tamanho da amostra}}$$

Estatística de teste para testar uma proporção simples de uma população:

$$Z_c = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Exemplo 7.16.1 Um jornal alega que 25% dos seus leitores pertencem à classe A. Se em uma amostra de 740 leitores encontramos 156 de classe A, qual a sua decisão a respeito da veracidade da alegação veiculada pelo jornal?

$$\begin{cases} H_0: & p = 0,25 \\ H_1: & p \neq 0,25 \end{cases} \qquad \overline{p} = \frac{156}{740} = 0,21$$

$$Z_c = \frac{0,21 - 0,25}{\sqrt{\frac{0,25 \cdot 0,75}{740}}} = -2,52$$

Como $-2,52 < -1,96$ (valor limite da região crítica) então rejeitamos H_0, ou seja, a proporção de leitores de classe A é diferente de 25%, com $\alpha = 0,05$.

7.17 Teste Hipóteses para Comparar duas Proporções Populacionais

A Estatística de teste (Z efetivo) neste caso é:

$$Z = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\overline{p}_c(1-\overline{p}_c)}{n_1} + \frac{\overline{p}_c(1-\overline{p}_c)}{n_2}}}$$

n_1 é o tamanho da amostra da população 1.
n_2 é o tamanho da amostra da população 2.
$\overline{p}_c$ é a média ponderada das duas proporções amostrais, calculada por:

$$\overline{p}_c = \frac{\text{número total de sucessos}}{\text{tamanho total das duas amostras}} = \frac{X_1 + X_2}{n_1 + n_2}$$

X_1 é o número de sucessos em n_1.
X_2 é o número de sucessos em n_2.

Exemplo 7.17.1 Dois tipos diferentes de solução de polimento estão sendo avaliados para possível uso em uma operação de polimento na fabricação de lentes intra-oculares usadas no olho humano depois de uma operação de catarata. Trezentas lentes forma polidas usando a primeira solução de polimento e, desse número, 253 não tiveram defeitos induzidos pelo polimento. Outras 300 lentes foram polidas, usando a segunda solução de polimento, sendo 196 lentes consideradas satisfatórias. Há qualquer razão para acreditar que as duas soluções de polimento diferem? Use $\alpha = 0,01$.

$$\begin{cases} H_0: & p_1 = p_2 \\ H_1: & p_1 \neq p_2 \end{cases} \qquad \overline{p}_1 = \frac{253}{300} = 0,8433 \qquad \overline{p}_2 = \frac{196}{300} = 0,6533$$

$$\overline{p}_c = \frac{253 + 196}{300 + 300} = 0,7483$$

$$Z = \frac{0,8433 - 0,6533}{\sqrt{\frac{0,7483(1-0,7483)}{300} + \frac{0,7483(1-0,7483)}{300}}} = 5,36$$

Como $5,36 > 2,58$ (valor limite da região crítica) então rejeitamos H_0, ou seja, há evidências para confirmar a afirmação de que os dois fluídos de polimento sejam diferentes, com $\alpha = 0,01$.

7.18 Teste de Hipóteses para a Variância Populacional

A Estatística do teste será:

$$\chi^2_c = \frac{(n-1)s^2}{\sigma^2_0}$$

Regra de decisão:

H_1	**Rejeitar H_0 se:**	**Não rejeitar H_0 se:**
$\sigma^2 < \sigma^2_0$	$\chi^2_c < \chi^2_{1-\alpha}$	$\chi^2_c \geqslant \chi^2_{1-\alpha}$
$\sigma^2 > \sigma^2_0$	$\chi^2_c > \chi^2_{\alpha}$	$\chi^2_c \leqslant \chi^2_{\alpha}$
$\sigma^2 \neq \sigma^2_0$	$\chi^2_c < \chi^2_{1-\frac{\alpha}{2}}$ ou $\chi^2_c > \chi^2_{\frac{\alpha}{2}}$	$\chi^2_{1-\frac{\alpha}{2}} \leqslant \chi^2_c \leqslant \chi^2_{\frac{\alpha}{2}}$

Exemplo 7.18.1 Para avaliar certas características de segurança de um carro, um engenheiro precisa saber se o tempo de reação dos motoristas a uma determinada situação de emergência tem variância de $0,0001$, ou se é superior a $0,0001$. Se o engenheiro obtém $s = 0,014$ para uma amostra de tamanho $n = 15$, qual é a sua conclusão ao nível de $0,05$ de significância?

$$\begin{cases} H_0: & \sigma^2 \leqslant 0,010 \\ H_1: & \sigma^2 > 0,010 \end{cases} \quad \text{ou} \quad \begin{cases} H_0: & \sigma^2 = 0,010 \\ H_1: & \sigma^2 > 0,010 \end{cases}$$

$$\chi^2_c = \frac{(15-1)\cdot 0,014^2}{0,0001} = 27,44$$

Como $\chi^2_c = 27,44$ excede $23,685$ (o valor tabelado), a hipótese nula deve ser rejeitada; em outras palavras, o engenheiro pode concluir que a variância dos tempos de reação dos motoristas a determinada situação de emergência é superior a $0,0001$, com significância de 5%.

7.19 Teste de Hipóteses para Comparar Duas Variâncias Populacionais

Regra de Decisão:

H_1	**Estatística do Teste**	**Rejeitar H_0 se:**	**Não rejeitar H_0 se:**	**g.l.**
$\sigma_1^2 < \sigma_2^2$	$F = \frac{s_2^2}{s_1^2}$	$F > F_\alpha$	$F \leqslant F_\alpha$	$n_2 - 1, n_1 - 1$
$\sigma_1^2 > \sigma_2^2$	$F = \frac{s_1^2}{s_2^2}$	$F > F_\alpha$	$F \leqslant F_\alpha$	$n_1 - 1, n_2 - 1$
$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{max(s_1^2, s_2^2)}{min(s_1^2, s_2^2)}$	$F > F_{\frac{\alpha}{2}}$	$F \leqslant F_{\frac{\alpha}{2}}$	$n_2 - 1, n_1 - 1$ ou $n_1 - 1, n_2 - 1$

Exemplo 7.19.1 Deseja-se determinar se há menor variabilidade no revestimento a ouro feito pela companhia 1 do que o novo revestimento feito pela companhia 2. Se amostras independentes acusaram $s_1 = 0,033$ mil (com base em 12 elementos) e $s_2 = 0,061$ mil (com base em 10 elementos), teste hipótese de que o novo revestimento apresenta maior variância, com 5% de significância.

$$\begin{cases} H_0: & \sigma_1^2 \geqslant \sigma_2^2 \\ H_1: & \sigma_1^2 < \sigma_2^2 \end{cases} \qquad \text{ou} \qquad \begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 < \sigma_2^2 \end{cases}$$

$$F = \frac{s_2^2}{s_1^2} = \frac{0,061^2}{0,033^2} = 3,42$$

Como $F = 3,42$ excede $2,90$ (o valor tabelado), a hipótese nula será rejeitada, então existe menor variância no revestimento feito na companhia 1, com 5% de significância.

7.20 Teste χ^2 de Independência: Tabelas de Contingência

Uma outra importante aplicação da distribuição Qui-quadrado diz respeito ao uso de dados de amostra pra **testar a independência de duas variáveis**. Para testar se duas variáveis são independentes, seleciona-se uma amostra e usa-se a tabulação cruzada para resumir os dados simultaneamente para as duas variáveis. O teste de independência usa o formato da tabela de

contingência e, por essa razão, é algumas vezes chamado de *teste da tabela de contingência*. As hipóteses para este teste são:

H_0: A variável 1 é independente da variável 2;

H_1: A variável 1 não é independente da variável 2.

A estatística do teste é:

$$\chi^2_c = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

onde:

O_{ij} = freqüência observada para a categoria da tabela de contingência na linha i e coluna j;

E_{ij} = freqüência esperada para a categoria da tabela de contingência na linha i e coluna j baseada na suposição de independência, sendo que:

$$E_{ij} = \frac{(\text{Total da linha } i)(\text{Total da coluna } j)}{\text{Total geral}}$$

Com r linhas e c colunas na tabela de contingência, as estatísticas de teste têm uma distribuição de qui-quadrado com $(r-1)(c-1)$ graus de liberdade contanto que as freqüências esperadas sejam 5 ou mais para todas as categorias. Rejeitaremos a hipótese nula, ou seja, a hipótese de independência, se o valor da estatística do teste for superior ao χ^2 tabelado com $(r-1)(c-1)$ graus de liberdade.

Exemplo 7.20.1 Uma companhia tem de escolher entre três planos de pensão. O gerente deseja saber se a preferência para os planos é independente da classificação do trabalho e quer usar $\alpha = 0,05$. As opiniões de uma amostra aleatória de 500 empregados são mostradas na Tabela 7.20.1:

Tabela 7.20.1: Opiniões dos trabalhadores segundo classificação do trabalho

	Plano de Pensão			
Classificação do Trabalho	1	2	3	**Total**
Assalariados	160	140	40	340
Horistas	40	60	60	160
Total	200	200	100	500

H_0: A preferência é independente da classificação de trabalho.

H_1: A preferência não é independente da classificação de trabalho.

$$E_{11} = \frac{340 \cdot 200}{500} = 136 \qquad E_{12} = \frac{340 \cdot 200}{500} = 136 \qquad E_{13} = \frac{340 \cdot 100}{500} = 68$$

$$E_{21} = \frac{160 \cdot 200}{500} = 64 \qquad E_{22} = \frac{340 \cdot 200}{500} = 64 \qquad E_{23} = \frac{160 \cdot 100}{500} = 32$$

$$\chi_c^2 = \frac{(160-136)^2}{136} + \frac{(140-136)^2}{136} + \frac{(40-68)^2}{68} + \frac{(60-64)^2}{64} + \frac{(60-32)^2}{32} = 49,63$$

Uma vez que $\chi_c^2 = 49,63$ é superior ao $\chi_2^2 = 5,99$ (valor tabelado com $\alpha = 5\%$), rejeitamos a hipótese nula, e concluímos que a preferência para planos de pensão não é independente da classificação de trabalho.

Probabilidade de Significância (ou $p-valor$)

O método de construção de um teste de hipóteses, descrito nas seções anteriores, parte da fixação do nível de significância α. Pode-se argumentar que esse procedimento pode levar à rejeição da hipótese nula para um valor de α e à não-rejeição para um valor menor. Outra maneira de proceder consiste em apresentar a **probabilidade de significância** ou $p-valor$ do teste. Os passos são muito parecidos aos já apresentados, a principal diferença está em não construir a região crítica. O que se faz é indicar a probabilidade de ocorrer valores da estatística mais extremos do que o observado, sob a hipótese de que H_0 ser verdadeira.

7.21 Exercícios

7.21.1 O fabricante de determinado remédio alega que o mesmo acusou mais do que 90% de eficiência em aliviar a alergia por um período de 8 horas. Em uma amostra de 200 indivíduos que sofriam de alergia, o remédio deu resultado positivo em 160. Determine se a alegação do fabricante é legítima, ou não, com $0,05$ de nível de significância.

7.21.2 Os produtores de um programa de televisão pretendem modificá-lo se for assistido regularmente por menos de 25% dos possuidores de televisão. Uma pesquisa encomendada a uma empresa especializada mostrou que, de 400 famílias entrevistadas, 80 assistem ao programa regularmente. Com base nos dados, qual deve ser a decisão dos produtores, com $\alpha = 0,01$.

7.21.3 Uma linha de montagem produz peças cujos pesos, em grama, obedecem ao modelo normal com variância 30 g^2. Os equipamentos foram modernizados e, para verificar se o processos continua sob controle, foi tomada uma amostra de 23 peças, que forneceu $s^2 = 40\ g^2$. Existem evidências indicando que a variância mudou, considerando $\alpha = 10\%$?

7.21.4 Uma grande empresa do setor alimentício possui sua sede na grande São Paulo. Há apenas dois anos esta empresa abriu duas filiais, uma na região sul do país e outra na região nordeste. Após este período de funcionamento os acionistas ficaram interessados em verificar se os lucros médios das duas empresas são diferentes. Sabendo que as variâncias são de 3600 para a empresa do sul e 4900 para empresa do nordeste, foi retirada uma amostra dos lucros (ou prejuízos, caso tenha acontecido) de 12 meses para as duas empresas e verificou-se uma média de R$ 195.000,00 para a empresa do sul e R$ 201.000,00 para a empresa do nordeste. Teste a hipótese, ao nível de $0,01$ de significância.

7.21.5 Um técnico em relações públicas e consultor para a indústria aeronáutica está planejando uma estratégia para influenciar a percepção dos eleitores da regulamentação oficial das tarifas aéreas. Em uma pesquisa *New York Times/CBS*, verificou-se que 35% de uma amostra de 552 democratas acham que o governo deve regulamentar os preços de passagens aéreas, em comparação com 41% de uma amostra de 417 republicanos pesquisados. Com o nível de $0,05$ de significância, teste a afirmação de que há diferença entre as proporções de democratas e republicanos que apóiam a regulamentação oficial das tarifas aéreas.

7.21.6 Karl Pearson, que elaborou muitos conceitos importantes em estatística, coletou dados sobre crimes em 1909. Dos condenados por incêndio criminoso, 50 bebiam e 43 eram abstêmios. Dos condenados por crime de fraude, 63 bebiam e 44 eram abstêmios. Com nível de significância de $0,01$, teste a afirmação de que a proporção dos que bebem entre os incendiários é maior do que a proporção dos bebedores condenados por fraude.

7.21.7 Deseja-se investigar se uma certa moléstia que ataca o rim altera o consumo de oxigênio desse órgão. Para indivíduos sadios, admite-se que esse consumo tem distribuição Normal com média 12 cm^3/min. Os valores medidos em cinco pacientes com a moléstia foram: $14,4$; $12,9$; $15,0$; $13,7$ e $13,5$. Qual seria a conclusão, ao nível de 5% de significância?

7.21.8 Em um estudo sobre os salários de comissários de bordo, selecionaram-se aleatoriamente salários pagos por duas companhias de aviação diferentes. Para 40 comissários de bordo da *American Airlines*, a média foi de \$23.870 e para 35 comissários da *TWA*, a média foi de \$22.025. Com Nível de significância de $0,10$ e sabendo que o desvio padrão para a *American Airlines* é \$2.960 e o desvio padrão para *TWA* é \$3.065, teste a afirmação de que as duas empresas pagam salários médios diferentes.

7.21.9 Um estudo sobre pacientes com transtornos do pânico que são atendidos em um Centro de Atendimento Psicológico, de uma grande universidade, foi realizado para verificar se a idade média destes pacientes era inferior a 35 anos. Para isso, foi coletada uma amostra com 25 destes pacientes e nesta amostra foi encontrada uma média de 32 anos e uma variância de 16 $anos^2$. Utilizando um nível de significância de 1%, verifique qual a conclusão será encontrada.

7.21.10 O consumidor de um certo produto acusou o fabricante, dizendo que mais de 20% das unidades fabricadas apresentam defeito. Para confirmar sua acusação, ele usou uma amostra de tamanho 50, onde 27% das peças eram defeituosas. Utilize um nível de significância de 5% e verifique qual conclusão foi encontrada.

7.21.11 Para determinar se o clima frio contribui para uma ausência maior às aulas, selecionaram-se aleatoriamente dois grupos de alunos do primeiro grau: um de 300 alunos do Rio Grande do Sul; outro de 400 alunos de Pernambuco. No primeiro grupo, constataram-se 72 ausências de um ou mais dias em um semestre; no segundo, 70. Pode-se concluir que um clima mais frio favorece a ausência nas aulas? Use nível de significância de 5%.

7.21.12 O salário de recém-formados em Veterinária foi amostrado em duas cidades. Na cidade A, 10 profissionais foram sorteados e na cidade B, 15. Os resultados, em salários mínimos, são apresentados a seguir:

Cidade A: $7,3$; $6,6$; $6,8$; $7,4$; $8,3$; $6,5$; $7,9$; $8,7$; $8,1$; $8,5$.

Cidade B: $6,5$; $7,8$; $8,2$; $6,9$; $7,9$; $9,7$; $9,1$; $9,5$; $7,4$; $8,0$; $6,9$; $7,9$; $8,4$; $9,3$; $9,5$.

Teste com $\alpha = 5\%$ se as variâncias dos salários das duas cidades são iguais.

7.21.13 O diâmetro de bastões de aço, fabricados em duas máquinas extrusoras diferentes, está sendo investigado. Duas amostras aleatórias de tamanhos $n_1 = 15$ e $n_2 = 17$ são selecionadas e as médias e as variâncias das amostras são $\bar{x}_1 = 8,73$, $s_1^2 = 0,35$, $\bar{x}^2 = 8,68$ e $s_2^2 = 0,40$, respectivamente. Suponha $\sigma_1^2 = \sigma_2^2$, verifique se há evidência que confirme a afirmação de que as duas máquinas produzem bastões com diferentes diâmetros médios? Use $\alpha = 0,05$.

7.21.14 Dois fornecedores fabricam uma engrenagem de plástico usada em uma impressora a laser. A resistência de impacto (medida em ftlbf) dessas engrenagens é uma característica importante. Uma amostra aleatória de 10 engrenagens do fornecedor 1 resulta em $\bar{x}_1 = 290$ e $s_1 = 12$, enquanto a outra amostra aleatória de 16 engrenagens do fornecedor 2 resulta em $\bar{x}_2 = 321$ e $s_2 = 22$. Há evidência confirmando a afirmação de que o fornecedor 2 fornece engrenagens com maiores resistências médias de impacto? Use $\alpha = 0,05$ e considere que as variâncias populacionais não sejam iguais.

7.21.15 Dois diferentes testes analíticos podem ser usados para determinar o nível de impurezas em ligas de aço. Oito espécimes são testados, usando ambos os procedimentos, sendo os resultados mostrados na tabela a seguir. Há evidência suficiente para concluir que ambos os testes fornecem o mesmo nível médio de impureza? Use $\alpha = 1\%$.

Espécime	1	2	3	4	5	6	7	8
Teste 1	1,2	1,3	1,5	1,4	1,7	1,8	1,4	1,3
Teste 2	1,4	1,7	1,5	1,3	2,0	2,1	1,7	1,6

7.21.16 Uma companhia opera quatro máquinas com três mudanças todo dia. Dos registros de produção, são coletados os seguintes dados do número de interrupções:

	Máquinas			
Mudanças	**A**	**B**	**C**	**D**
1	41	20	12	16
2	31	11	9	14
3	15	17	16	10

Teste a hipótese (usando $\alpha = 5\%$) de que as interrupções independentes da mudança.

7.21.17 Um estudo está sendo feito sobre as falhas de um componente eletrônico. Há quatro tipos possíveis de falhas e duas posições de montagem do componente. Os seguintes dados forma obtidos:

	Tipo de Falha			
Posição de Montagem	**A**	**B**	**C**	**D**
1	22	46	18	9
2	4	17	6	12

Você concluiria que o tipo de falha seria independente da posição de montagem? Use $\alpha = 1\%$.

7.21.18 Uma construtora está concluindo um conjunto residencial que contém 500 apartamentos de dois dormitórios. A imobiliária encarregada das vendas, com o objetivo de definir o perfil do potencial comprador deste tipo de apartamento, amostrou aleatoriamente em seus arquivos 200 potenciais compradores casados e verificou que 25 efetivaram a compra, enquanto que em uma amostra de 120 descasados, 30 efetivaram a compra.

Se o gerente da imobiliária está propenso a acreditar que a promoção deve ser direcionada aos não casados, teste a hipótese de que não há diferença significativa na proporção de compradores dos dois grupos, ao nível de 5%.

7.21.19 Uma firma de semicondutores produz aparelhos lógicos. O contrato com o um cliente exige uma fração de defeituosos de não mais do que $0,05$. Uma amostra aleatória de 200 aparelhos resulta em seis defeituosos. Aplicando um teste de hipóteses e usando $\alpha = 0,05$, pode- se afirmar que a firma está executando o trabalho de acordo com o contrato?

7.21.20 Um vendedor de frutas recebe caixas de laranjas de dois produtores e embala as laranjas em sacos de 5 Kg para a venda posterior. Para controlar o custo da mercadoria e negociar preço com os produtores, amostra regularmente 10 caixas ao acaso de cada produtor, calculando o peso médio e o desvio padrão correspondente. Se no último levantamento obteve um desvio padrão

de $1,4$ Kg e $0,9$ Kg para as caixas dos dois produtores, supondo o peso das caixas normalmente distribuído, testar ao nível de 10% a hipótese da variância dos pesos do primeiro fornecedor ser maior que a do segundo.

7.21.21 Uma máquina é projetada para fazer esferas de aço de 1 cm de raio. Uma amostra de 10 esferas é produzida e tem o raio médio de $1,004$ cm, com $s = 0,003$. Há razões para suspeitar que a máquina esteja produzindo em média, esferas com raio maior que 1 cm, ao nível de 10%?

7.21.22 Dois fornecedores fabricam uma engrenagem de plástico usada em uma impressora a laser. A resistência de impacto (medida em ftlbf) dessas engrenagens é uma característica importante. Uma amostra aleatória de 10 engrenagens do fornecedor 1 resulta em $\overline{x}_1 = 290$ e $s_1 = 12$, enquanto a outra amostra aleatória de 16 engrenagens do fornecedor 2 resulta em $\overline{x}_2 = 321$ e $s_2 = 22$. Há evidência confirmando a afirmação de que o fornecedor 2 fornece engrenagens com maiores resistências médias de impacto? Use $\alpha = 0,05$ e considere que as variâncias populacionais sejam iguais.

8 Correlação Linear e Regressão Linear Simples

Freqüentemente procura-se verificar existe relação (ou associação) entre duas ou mais variáveis. O peso pode estar relacionado com a idade das pessoas; o consumo das famílias pode estar relacionado com sua renda; as vendas de uma empresa e os gastos promocionais podem relacionar-se, bem como a demanda de um determinado produto e seu preço. A verificação da existência e do grau de relação entre variáveis é objeto do estudo da correlação. Uma vez caracterizada, procura-se descrever uma relação sob forma matemática, através de uma função (modelo). A estimação dos parâmetros desse modelo matemático é o objeto da regressão.

8.1 Correlação Linear Simples

O estudo da correlação tem por objetivo medir e avaliar o grau de relação existente entre duas variáveis aleatórias. A correlação linear procura medir a relação entre as variáveis X e Y através da disposição dos pontos (X, Y) em torno de uma reta.

O instrumento de medida da correlação linear é dado pelo coeficiente de correlação de Pearson (no caso de uma amostra, representado por $r_{X,Y}$ e no caso de uma população representado por $\rho_{X,Y}$):

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

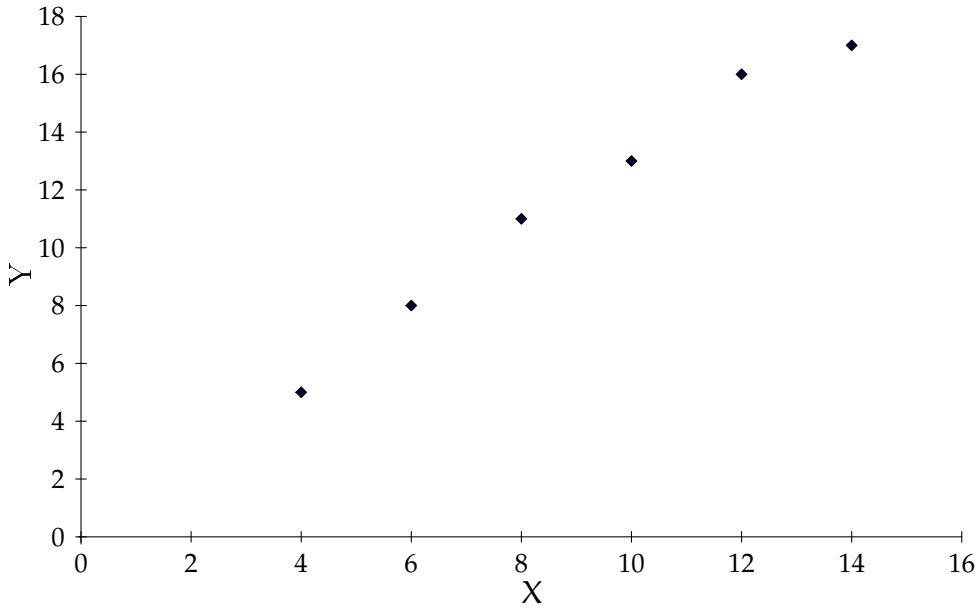
em que n é o número de observações na amostra, sendo que $-1 \leqslant r_{X,Y} \leqslant 1$.

Correlação Linear Positiva

A correlação será considerada positiva se valores crescentes de X estiverem associados a valores crescentes de Y, ou valores decrescentes de X estiverem associados a valores decrescentes da variável

Y. Um exemplo é apresentado no Diagrama de Dispersão 1 (Figura 8.1.1). Neste caso $0 < r_{X,Y} < 1$.

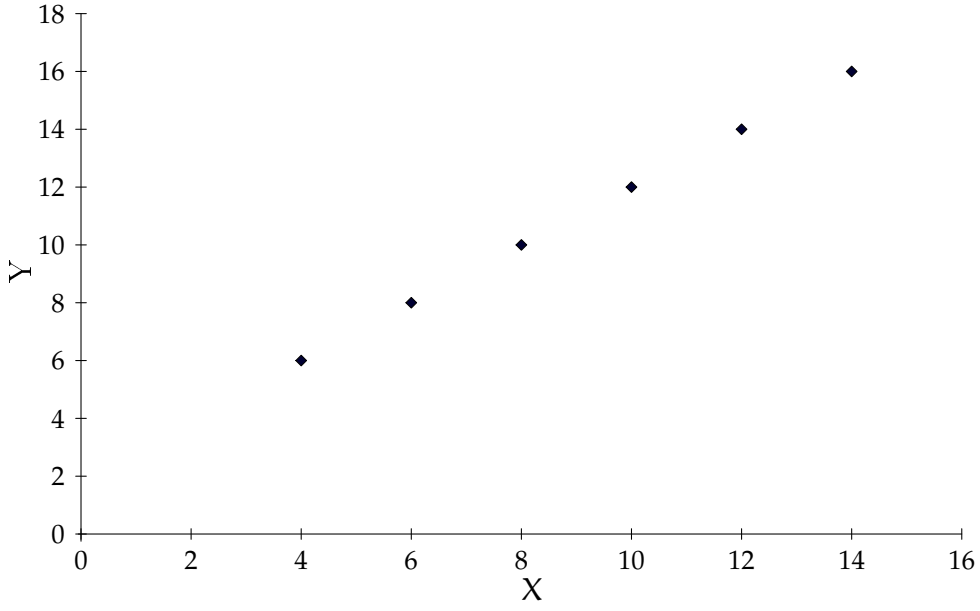
Figura 8.1.1: Diagrama de Dispersão 1



Correlação Linear Perfeita Positiva

A correlação linear perfeita positiva corresponde ao caso anterior, só que os pontos (X, Y) estão perfeitamente alinhados como apresenta o Diagrama de Dispersão 2 (Figura 8.1.2). Neste caso, $r_{X,Y} = 1$.

Figura 8.1.2: Diagrama de Dispersão 2

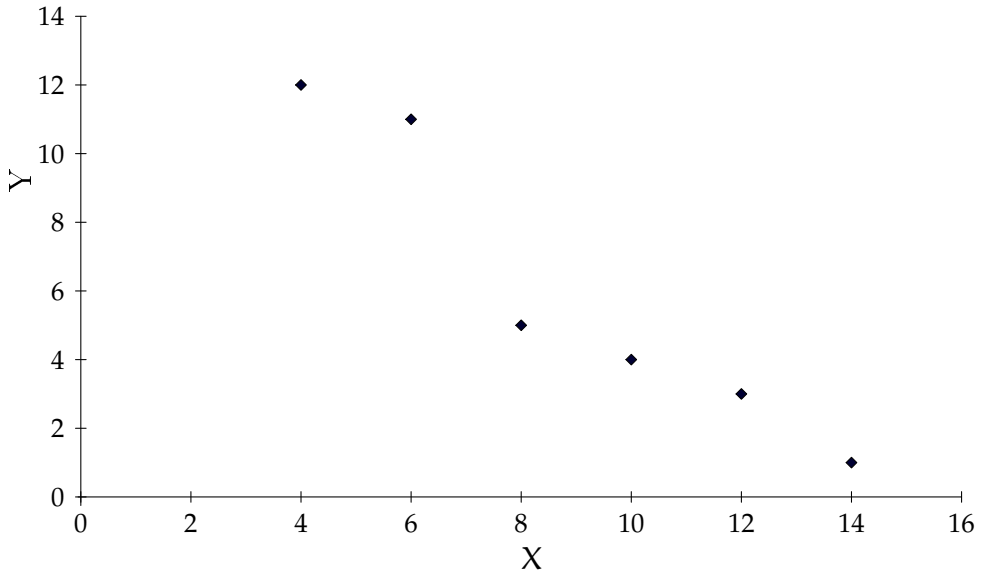


Correlação Negativa

A correlação é considerada negativa quando os valores crescentes da variável X estiverem associados a valores decrescentes da variável Y, ou valores decrescentes de X associados a valores

crescentes da variável Y. Neste caso, $-1 < r_{X,Y} < 0$. O Diagrama de Dispersão 3 (Figura 8.1.3) apresenta essa situação.

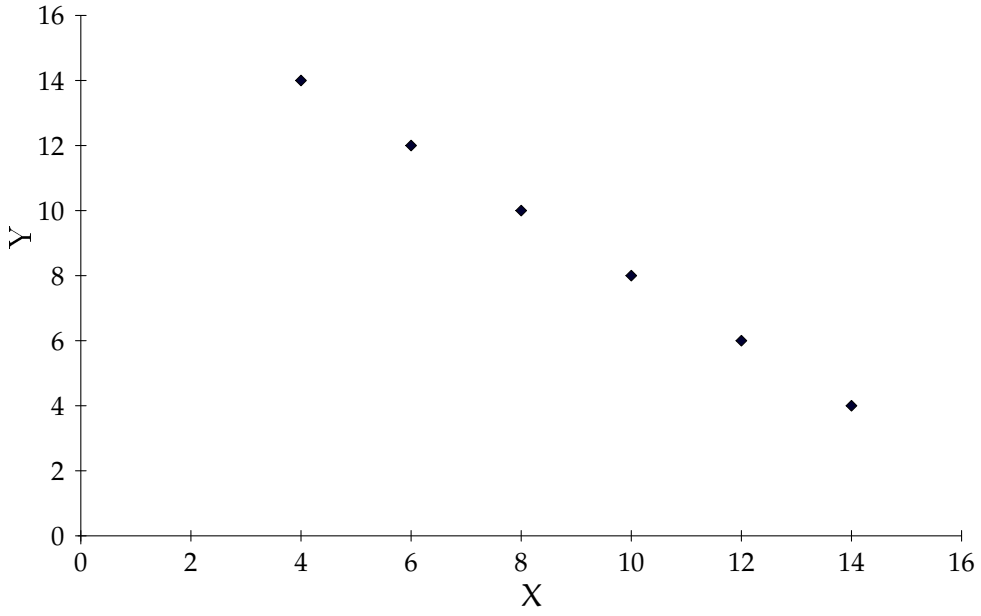
Figura 8.1.3: Diagrama de Dispersão 3



Correlação Perfeita Negativa

Quando os pontos estiverem perfeitamente alinhados, mas em sentido contrário, a correlação é denominada perfeita negativa. O Diagrama de Dispersão 4 (Figura 8.1.4) apresenta um resultado semelhante, $r_{X,Y} = -1$.

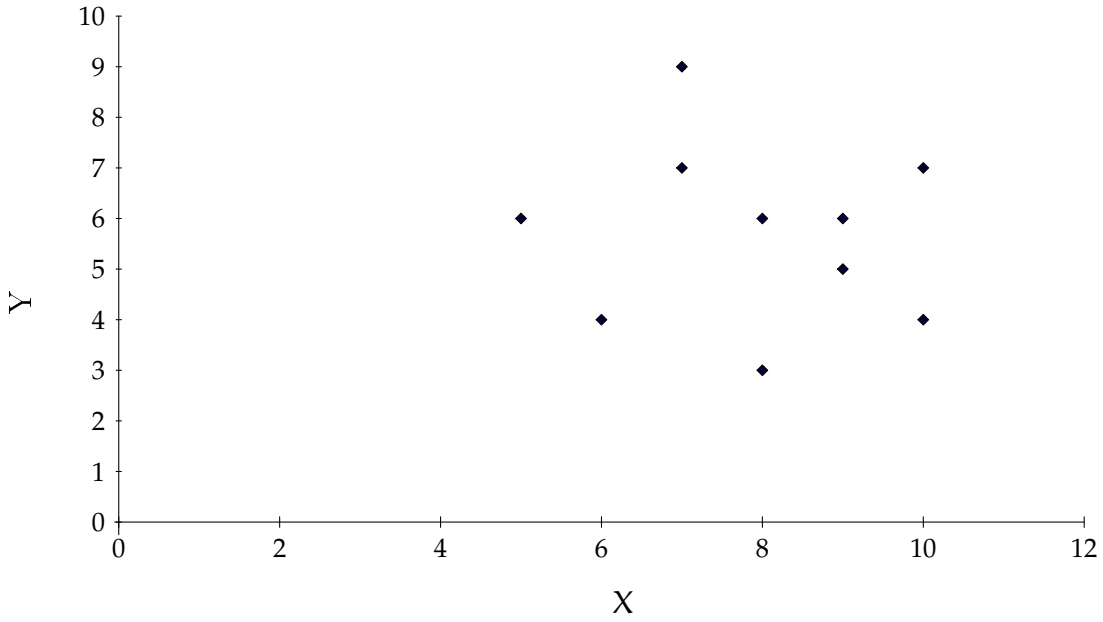
Figura 8.1.4: Diagrama de Dispersão 4



Ausência de Correlação

Quando não houver relação entre as variáveis X e Y, ou seja, quando as variações de X e Y ocorrem independentemente não existe correlação entre elas, então $r_{X,Y} = 0$. Um exemplo está no Diagrama de Dispersão 5 (Figura 8.1.5).

Figura 8.1.5: Diagrama de Dispersão 5



Correlação e Causalidade

A observação de que duas grandezas tendem simultaneamente a variar no mesmo sentido não implica a presença de um relacionamento causal entre elas. Se anotarmos o número mensal X de homicídios e o número mensal Y de Concentrações religiosas para várias cidades de grande porte, os dados provavelmente indicarão uma correlação positiva elevada. É a flutuação de uma terceira variável (a população da cidade) que faz com que X e Y variem no mesmo sentido, embora X e Y possam ser não-correlacionadas ou até mesmo correlacionadas negativamente. A terceira variável que, nesse exemplo, causa a correlação observada entre crimes e concentrações religiosas é chamada variável **intercorrente** (não-conhecida), e a falsa correlação que ela origina é chamada **correlação espúria**.

Por isso, ao se utilizar um coeficiente de correlação como medida de relacionamento, deve- se verificar a possibilidade de uma variável intercorrente estar afetando qualquer das variáveis em estudo. Isso se faz intuitivamente ou via **análise de regressão múltipla**.

8.2 Teste de Correlacionamento

Em uma amostra bivariada, é possível determinar se as duas variáveis aleatórias são realmente relacionadas ou não. Quando a população segue o modelo bivariado normal, dispomos de um teste bastante simples pra verificar se X e Y estão associadas. As hipóteses serão:

$$\begin{cases} H_0: & \rho_{X,Y} = 0 \\ H_1: & \rho_{X,Y} \neq 0 \end{cases}$$

sendo a estatística do teste dada por:

$$t_c = \frac{r_{X,Y}\sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}}$$

que tem distribuição *t de Student* com $n-2$ graus de liberdade. A hipótese nula será rejeitada se $|t_c| > t_{\frac{\alpha}{2},n-2}$ e pode-se afirmar que a correlação existe e é significativa.

8.3 Regressão Linear Simples

A análise de regressão tem por objetivo descrever através de um modelo matemático, a relação existente entre duas variáveis, a partir de n observações dessas variáveis.

Na terminologia de regressão, a variável que está sendo calculada é chamada de variável dependente. A variável ou as variáveis que estão sendo usadas para calcular a variável dependente são chamadas de variáveis independentes. Por exemplo, ao analisar os efeitos dos gastos com publicidade sobre as vendas, o desejo do gerente de marketing de prever as vendas sugeriria que se tomasse as vendas como variável dependente. Os gastos com publicidade seriam a variável independente usada para calcular as vendas. Em notação estatística, Y denota a variável dependente e X denota a variável independente.

O tipo mais simples de análise de regressão, envolvendo uma variável independente e uma variável dependente na qual a relação entre as variáveis é aproximada por uma linha reta, é chamado de regressão linear simples. A análise de regressão envolvendo duas ou mais variáveis independentes é chamada de análise de regressão múltipla.

A equação que descreve como Y está relacionado com X e com um erro é chamada de modelo de regressão. O modelo de regressão usado em uma regressão linear simples é apresentado a seguir:

$$Y = \alpha + \beta X + \varepsilon$$

No modelo de regressão linear simples, Y é uma função linear de X (a parte $\alpha + \beta X$) mais ε. α e β são denominados parâmetros do modelo, e ε é uma variável aleatória definida como o termo erro (ou resíduo). O termo erro mede a variabilidade em Y que não pode ser explicada pela relação linear entre X e Y. Uma das suposições para o modelo de regressão linear simples é que o valor esperado de ε é zero.

Geralmente os valores dos parâmetros não são conhecidos na prática e devem ser calculados usando os dados da amostra, então a equação de regressão estimada (também chamada de reta ajustada) será:

$$\widehat{Y} = a + bX$$

sendo os valores de a e b as estimativas de α e β; a é a intersecção de Y, ou seja, o ponto onde a reta ajustada corta o eixo da variável Y, b é a inclinação e $\widehat{Y}$ é o valor estimado de Y para um dado valor de X.

A reta ajustada é denominada, também, reta de mínimos quadrados, pois os valores de a e b são obtidos de tal forma que é mínima a soma dos quadrados das diferenças entre os valores observados de Y e os obtidos a partir da reta ajustada para os mesmos valores de X.

Para obter os estimadores a e b aplica-se a condição necessária de mínimo à função $\sum_{i=1}^{n}(Y-\widehat{Y})^2$. Para tanto basta derivá-la com relação a esses parâmetros e igualar as derivadas a zero. Utilizando este procedimento podemos deduzir que as estimativas de mínimos quadrados da intersecção e da inclinação no modelo de regressão linear simples são:

$$a = \overline{y} - b\overline{x} \qquad \text{e} \qquad b = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \frac{\left(\displaystyle\sum_{i=1}^{n} x_i\right)\left(\displaystyle\sum_{i=1}^{n} y_i\right)}{n}}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Observação 8.3.1

1. O método dos mínimos quadrados busca traçar a melhor reta através dos pontos, ou seja, aquela que torna mínima distância entre os pontos e a reta;
2. Sempre é possível obter a equação de uma reta que passa por um conjunto de pontos, mas isto não significa que o modelo seja adequado. Para verificar a adequação do modelo, emprega-se a metodologia da Análise de Variância (ANOVA) e também é recomendável fazer uma análise de resíduos.

8.4 Testes para a Significância da Regressão

A Tabela de Análise de Variância

Obtida a reta de regressão estimada, é necessário determinar na sua precisão, isto é, verificar a sua utilidade se ela é útil na representação da tendência dos dados observados. Um método chamado análise de variância pode ser usado para testar a significância da regressão. O procedimento divide a variância total na variável de resposta em componentes significantes como base para o teste.

Consideremos a seguinte equação:

$$y_i - \widehat{y} = (y_i - \overline{y}) - (\widehat{y}_i - \overline{y})$$

O que a identidade anterior nos diz é que o resíduo $e_i = y_i - \widehat{y}_i$ é a diferença entre o desvio do valor observado y_i em relação a sua média $\overline{y}$ e o desvio do valor estimado $\widehat{y}_i$ em relação à sua média $\overline{y}$ ($\overline{y}$ é a média tanto dos $\widehat{y}_i$ quanto dos y_i).

Reescrevamos a equação como:

$$y_i - \overline{y} = (y_i - \widehat{y}_i) - (\widehat{y}_i - \overline{y})$$

e elevemos ambos os membros ao quadrado, somando para $i = 1, 2, \ldots, n$:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

Analisando a expressão acima, vemos que a variação total das observações em torno de sua média, dada por $\displaystyle\sum_{i=1}^{n} (y_i - \overline{y})^2$, se decompõe em duas parcelas: soma dos quadrados dos resíduos $\displaystyle\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ que mede a variação em torno da reta de reta de regressão (variação "não-explicada"),

e soma dos quadrados dos desvios dos valores preditos em relação à sua média (variação "explicada"pela regressão de Y em X) medida por $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$.

Resumindo, temos:

Soma Total de Quadrados (SQTotal) = *Soma de Quadrados dos Resíduos* (SQR) + *Soma de Quadrados devida à Regressão* (SQReg)

em que:

$$SQTotal = \left(\sum_{i=1}^{n} y_i^2\right) - n\overline{y}^2$$

$$SQReg = b^2 \sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$SQR = SQT - SQReg$$

Essa decomposição é sintetizada em uma tabela, chamada Tabela de Análise de Variância (ANOVA), conforme é apresentado abaixo na Tabela 8.4.1:

Tabela 8.4.1: Tabela da Análise de Variância

Causa de Variação (CV)	**g.l.**	SQ	QM	F
Regressão	1	SQReg	$QMReg = \frac{SQReg}{g.l}$	$F = \frac{QMReg}{QMR}$
Resíduos	$n-2$	SQR	$QMR = \frac{SQR}{g.l}$	
Total	$n-1$	SQTotal		

A primeira coluna apresenta a decomposição explicada acima: há duas fontes de variação, uma associada aos resíduos, outra, à reta de regressão. A segunda coluna apresenta os graus de liberdade. Se o modelo proposto é correto, QMR estima σ^2. Por isso, é muitas vezes representado por s^2. Contudo, se o modelo proposto não é correto, s^2 superestima σ^2. Medirá não só a variação aleatória de Y (ou ε) em torno de sua media, mas também o mau ajustamento dos dados ao modelo escolhido, o que é denominado **falta de ajuste** (aderência).

Com esta tabela de Análise de Variância podemos testar a seguinte hipótese:

$$\begin{cases} H_0: & \beta = 0 \\ H_1: & \beta \neq 0 \end{cases}$$

fazendo a comparação dos valores de QMR e QMReg. Se eles forem muito afastados ou, equivalentemente, se $F = \frac{QMReg}{QMR}$ for significativamente maior que 1, rejeitaremos H_0. Resta decidir quando considerar F significativamente maior que 1.

É possível mostra que, sob o modelo de erro normal, F tem distribuição F de Fisher com 1 e $(n-2)$ graus de liberdade, e H_0 será rejeitada se $F > F_{\alpha,1,n-2}$.

Coeficiente de Determinação (R^2)

Tem por objetivo avaliar a "qualidade"do ajuste de um modelo de regressão. Seu valor fornece a proporção da variação total da variável Y explicada pela variável X através da função ajustada. Podemos expressar R^2 por:

$$R^2 = (r_{X,Y})^2 \qquad \text{ou} \qquad R^2 = \frac{SQReg}{SQTotal}$$

sendo $0 \leqslant R^2 \leqslant 1$.

Quando $R^2 = 0$, a variação explicada de Y é zero, a reta ajustada é paralela ao eixo de X. Se $R^2 = 1$, a reta ajustada explicará toda a variação de Y.

Se, por exemplo, $R^2 = 0,98$, isto significa que 98% das variações de Y são explicadas por X através da função escolhida para relacionar as duas variáveis e 2% são atribuídas a causas aleatórias.

Análise Residual: Validando Suposições do Modelo

Ajustar um modelo de regressão requer várias suposições. A estimação dos parâmetros do modelo requer a suposição de que:

i) Os valores dos erros ε sejam variáveis aleatórias não correlacionadas, ou seja, independentes;

ii) A $E(\varepsilon) = 0$;

iii) A variância de ε é a mesma para todos os valores de X, ou seja, é constante;

iv) Testes de hipóteses e estimação do intervalo requerem que o erro ε seja normalmente distribuído.

Essas suposições fornecem a base teórica para o teste t e o teste F usados para determinar se a relação entre X e Y é significante. Se as suposições sobre o erro ε parecem questionáveis, os testes de hipóteses sobre o significado da relação de regressão podem não ser válidos.

Os resíduos fornecem a melhor informação sobre ε; por isso uma análise dos resíduos é um passo importante ao determinar se as suposições para ε são apropriadas.

O resíduo para a observação i é

$$\widehat{e}_i = y_i - \widehat{y}_i,\ i = 1,\ 2,\ \ \ldots,\ \ n$$

onde:

y_i é o valor observado da variável dependente;

$\widehat{y}_i$ é o valor estimado da variável dependente.

Em outras palavras, o i-ésimo resíduo é o erro resultante do uso da equação de regressão estimada para prever o valor de Y_i.

Existem várias técnicas formais para conduzir essa análise, mas aqui iremos ressaltar basicamente métodos gráficos.

Exemplo 8.4.1 Na Tabela 8.4.2 temos os seguintes dados:

Tabela 8.4.2: Dados da população de estudantes e vendas trimestrais para 10 Armand's Pizza Parlors

Restaurante	X = População de estudantes (1.000)	Y = Vendas Trimestrais (US$1.000)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Esses dados nos levam aos resultados apresentados na Tabela 8.4.3:

Tabela 8.4.3: Dados da população de estudantes e vendas trimestrais para 10 Armand's Pizza Parlors

X = **População de estudantes** (1.000)	Y = **Vendas Trimestrais** (US$1.000)	**Vendas estimadas** ($\widehat{y}_i = 60 + 5x_1$)	**Resíduos** ($\widehat{e}_i = y_i - \widehat{y}_i$)
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

8.5 Regressão Linear Múltipla

Na discussão sobre regressão linear simples, o foco se concentra em um modelo no qual uma variável independente (ou explanatória) X é utilizada para prever o valor de uma variável Y dependente. Geralmente, pode ser desenvolvido um modelo com um melhor ajuste, caso mais de uma variável explanatória seja considerada e neste caso vamos considerar o modelo de regressão múltipla, nos quais diversas variáveis explanatórias podem ser utilizadas no sentido de prever o valor de uma variável dependente.

Modelo de regressão múltipla com k variáveis independentes:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon_1$$

onde:

α = intersecção de Y;

β_1 = inclinação de Y em relação à variável X_1, mantendo constantes as variáveis X_2, X_3, ..., X_k;

β_2 = inclinação de Y em relação à variável X_2, mantendo constantes as variáveis X_1, X_3, ..., X_k;

$\vdots$

β_k inclinação de Y em relação à variável X_k, mantendo constantes as variáveis X_1, X_2, X_3, ..., X_{k-1};

ε_i erro aleatório em Y, para a observação i.

Estimação dos Parâmetros da Regressão Múltipla

Os valores estimados da variável dependente são calculados através da equação de regressão múltipla estimada,

$$\hat{y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k$$

Para a regressão linear múltipla, o princípio é o mesmo da regressão linear simples, ou seja, o método dos mínimos quadrados, mas os detalhes são mais complicados. Em regressão múltipla, a apresentação de fórmulas para os coeficientes de regressão a, b_1, b_2, ..., b_k envolve o uso de álgebra de matrizes.

No ajuste de um modelo de regressão múltipla, é muito mais conveniente expressar as operações matemáticas usando notação matricial. Suponha que haja k regressores e n observações (x_{i1}, x_{i2}, ..., x_{ik}, y_i), $i = 1, 2, \ldots, n$, e que o modelo relacionando os regressores à resposta seja

$$y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik} + \varepsilon_1$$

Esse modelo é um sistema de n equações, que pode ser expresso na notação matricial como

$$y = X\beta + \varepsilon$$

sendo

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{e} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Em geral, y é um vetor $(nX1)$ das observações, X é uma matriz (nXk) dos níveis das variáveis independentes, β é um vetor $(kX1)$ dos coeficientes de regressão eε é um vetor $(nX1)$ dos erros aleatórios.

O que se deseja é encontrar $\widehat{\beta}$, o vetor dos estimadores de mínimos quadrados.

$$\widehat{\beta} = (X'X)^{-1}X'y$$

Na prática, os cálculos de regressão múltipla são quase sempre realizados por métodos computacionais.

Exemplo 8.5.1 A seguir temos dados da resistência ao puxamento de um fio colado, em um processo de fabricação de semicondutores. Temos também o comprimento do fio e a altura da garra. Ajustaremos, por meio de uma abordagem matricial o modelo de regressão para esses dados.

Resistência ao puxamento (Y)	Comprimento do Fio (X_1)	Altura da Garra (X_2)
9,95	2	50
24,45	8	110
31,75	11	120
35,00	10	550
25,02	8	295
16,86	4	200
14,38	2	375
9,60	2	52
24,35	9	100
27,50	8	300
17,08	4	412
37,00	11	400
41,95	12	500
11,66	2	360
21,65	4	205
17,89	4	400
69,00	20	600
10,30	1	585
34,93	10	540
46,59	15	250
44,88	15	290
54,12	16	510
56,23	17	590
22,13	6	100
21,15	5	400

A matriz $X'X$ é

$$X'X = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 2 & 8 & 11 & \cdots & 5 \\ 50 & 110 & 120 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 1 & 8 & 110 \\ 1 & 11 & 120 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 400 \end{bmatrix} = \begin{bmatrix} 25 & 206 & 8.294 \\ 206 & 2.396 & 77.177 \\ 8.294 & 77.177 & 3.531.848 \end{bmatrix}$$

e o vetor $X'y$ é

$$X'y = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 2 & 8 & 11 & \cdots & 5 \\ 50 & 110 & 120 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 9,95 \\ 24,45 \\ 31,75 \\ \vdots \\ 21,15 \end{bmatrix} = \begin{bmatrix} 725,82 \\ 8.008,37 \\ 274.811,31 \end{bmatrix}$$

As estimativas de mínimos quadrados são encontradas a partir de:

$$\widehat{\beta} = (X'X)^{-1}X'y$$

ou

$$\begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 25 & 206 & 8.294 \\ 206 & 2.396 & 77.177 \\ 8.294 & 77.177 & 3.531.848 \end{bmatrix}^{-1} \begin{bmatrix} 725,82 \\ 8.008,37 \\ 274.811,31 \end{bmatrix} =$$

$$= \begin{bmatrix} 0,214653 & -0,007491 & -0,00034 \\ -0,007491 & 0,001671 & -0,000019 \\ -0,00034 & -0,000019 & 0,0000015 \end{bmatrix} \begin{bmatrix} 725,82 \\ 8.008,37 \\ 274.811,31 \end{bmatrix} =$$

$$= \begin{bmatrix} 2,2638 \\ 2,7443 \\ 0,0125 \end{bmatrix}$$

Portanto, o modelo ajustado de regressão é dado por:

$$\widehat{y} = 2,2638 + 2,7443X_1 + 0,0125X_2$$

Interpretação dos Coeficientes

Em regressão linear simples, interpretamos b como uma estimativa da mudança em y para uma mudança de uma unidade na variável independente. Em análise de regressão múltipla, essa interpretação deve ser modificada de alguma forma. Isto é, em análise de regressão múltipla, interpretamos cada coeficiente da regressão como segue: b_1 representa uma estimativa da mudança em y correspondente a uma mudança de uma unidade em x_i quando todas as outras variáveis independentes permanecem constantes.

Exemplo 8.5.2 A maior parte dos negócios da *Butler Trucking Company* envolve entregas na região do sul da Califórnia. Para desenvolver melhores horários de trabalho, os gerentes querem estimar o tempo total de viagens diárias de seus motoristas, como mostra a Tabela

8.5.1:

Tabela 8.5.1: Dados para *Butler Trucking* com quilômetros percorridos e números de entregas como variáveis independentes

Entrega	X_1 = **Quilêmtros Percorridos**	X_2 = **Número de entregas**	Y = **Tempo de viagens (horas)**
1	100	4	9,3
2	50	3	4,8
3	100	4	8,9
4	100	2	6,5
5	50	2	4,2
6	80	2	6,2
74	75	3	7,4
8	65	4	6,0
9	90	3	7,6
10	90	2	6,1

A equação de regressão estimada é:

$$\hat{y} = -0,869 + 0,0611X_1 + 0,923X_2$$

Interpretação: Dessa forma, $b_1 = 0,0611$ horas é uma estimativa do aumento esperado no tempo de viagem correspondente a um aumento de um quilômetro na distância percorrida quando o número de entregas permanece constante. Similarmente, já que $b_2 = 0,923$, uma estimativa do aumento esperado no tempo de viagem correspondente a um aumento de uma entrega quando o número de quilômetros percorridos é mantido constante é $0,923$ horas.

Teste F para o Modelo Completo da Regressão na Regressão Múltipla

Uma vez que existe mais de uma variável explanatória, as hipóteses nula e alternativa são construídas da seguinte forma:

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (Nenhuma relação linear entre a variável dependente e as variáveis explanatórias)

H_1 : Pelo menos um $\beta_j \neq 0$ (Existe relação linear entre a variável dependente e pelo menos uma das variáveis explanatórias)

A estatística F é dada por:

$$F = \frac{QMReg}{QMR}$$

onde:

F = estatística do teste, a partir de uma distribuição F, com k e $n - k - 1$ graus de liberdade;

k = número de variáveis explanatórias no modelo de regressão.

A hipótese nula será rejeitada, no nível de significância α, se $F > F_{(k,n-k-1)}$; caso contrário, H_0 não é rejeitada. A seguir é apresentada a Tabela 8.5.2 de análise de variância para este teste:

Tabela 8.5.2: Tabela da Análise de Variância

Causa de Variação (CV)	g.l	SQ	QM	F
Regressão	k	$SQReg$	$QMReg = \frac{SQReg}{k}$	$F = \frac{QMReg}{QMR}$
Resíduos	$n–k-1$	SQR	$QMR = \frac{SQR}{n-k-1}$	
Total	$n-1$	$SQTotal$		

sendo,

$$SQTotal = y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \qquad SQReg = \widehat{\beta}'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$SQR = SQTotal - SQReg$$

Coeficientes de Determinação Múltipla

Na regressão múltipla, uma vez que existem pelo menos duas variáveis explanatórias, o coeficiente de determinação múltipla representa a proporção da variação em Y que é explicada através do conjunto de variáveis explanatórias selecionadas. É definido da seguinte forma:

$$R^2 = \frac{SQReg}{SQTotal}$$

onde:

$SQReg$ = soma dos quadrados devida à regressão;

$SQtotal$ = soma total dos quadrados.

Como no caso da regressão linear simples, temos de ter $0 \leqslant R^2 \leqslant 1$. Entretanto um valor grande de R^2 não implica necessariamente que o modelo de regressão seja bom. A adição de uma variável ao modelo sempre aumentará R^2, independente da variável adicional ser ou não estatisticamente significativa. Assim, modelos que tenham valores grandes de R^2 podem resultar em previsões pobres de novas observações ou estimativas da resposta média.

A raiz quadrada positiva de R^2 é chamada de coeficiente de correlação múltipla entre Y e o conjunto de regressores.

8.6 Exercícios

8.6.1 Presidentes e executivos dirigentes de companhias são pagos de acordo com a performance dos lucros da companhia? A tabela a seguir mostra dados associados com a mudança de porcentagem de lucros sobre um período de dois anos e a mudança de porcentagem no pagamento de presidentes e executivos dirigentes imediatamente após o período de dois anos (*Business Week, 21 de abril de 1997*):

Companhia	Mudança dos lucros em dois anos (%)	Mudança na compensação dos executivos (%)
Dow Chemical	201,3	18
Rohm & Haas	146,5	28
Morton International	76,7	10
Union Caribe	158,2	28
Praxair	–34,9	15
Air Products & Chemicals	73,2	–9
Eastman Chemical	–7,9	–20

a) Desenvolva a equação de regressão estimada com a mudança de porcentagem dos lucros como variável independente.

b) Calcule ao coeficiente de correlação e indique que tipo de correlação existe entre as duas variáveis.

c) Teste se a correlação é significativa com $\alpha = 0,10$.

8.6.2 Uma aplicação importante da análise de regressão em contabilidade é a estimativa de custo. Coletando dados de volume e custo e desenvolvendo uma equação de regressão estimada relacionando volume e custo, um contador pode estimar o custo associado a uma determinada operação de manufatura. Considere a amostra de volume de produção e o custo total para a operação de manufatura que segue:

Custo Total (em mil R$)	Volume de produção (em mil unidades)
120,0	15
150,0	17
161,4	20
169,0	26
192,3	30
210,0	33

a) Use esses dados para desenvolver uma equação de regressão estimada que poderia ser usada para prever o custo total para um dado volume de produção.

b) Use a análise de variância e teste a significância da equação encontrada. Use $\alpha = 5\%$

8.6.3 Um administrador de entrevistadores deseja desenvolver um modelo para prever o número de entrevistas em um dado dia. Ele acredita que a experiência do entrevistador (medida em semanas trabalhadas) é determinante do número de entrevistas realizadas. Uma amostra de 10 entrevistadores revelou os seguintes dados:

Semanas de experiência	15	41	58	18	37	52	28	24	45	33
Número de entrevistas realizadas	4	9	12	6	8	10	6	5	10	7

a) Determine a equação de regressão que explique essa relação;

b) Calcule o coeficiente de correlação;

c) Construa o diagrama de dispersão;

d) Determine o coeficiente de determinação.

8.6.4 São dados os custos totais e as quantidades produzidas de certo artigo:

X = **quantidades produzidas**	38	48	71	64	60	8	45	34	28	15
Y = **custos totais**	375	500	720	600	580	95	460	350	250	160

a) Determine uma equação de regressão linear simples para estudar os custos do artigo referido;

b) Use a análise de variância e teste a significância da regressão com $\alpha = 2,5\%$

8.6.5 Um artigo em *Concrete Research ("Near Surface Characteristics of Concrete: Intrinsic Permeability"* – Características do Concreto Perto da Superfície: Permeabilidade Intrínseca, Vol. 41, 1989) apresentou dados sobre resistência à compressão, x, e permeabilidade intrínseca, y, de várias mistruras e curas de concreto. Um sumário das quantidades é:

$$n = 14,\ \sum_{i=1}^{n} y_i = 572,\ \sum_{i=1}^{n} y_i^2 = 23.530,\ \sum_{i=1}^{n} x_i = 43,\ \sum_{i=1}^{n} x_i^2 = 157,42 \text{ e } \sum_{i=1}^{n} x_i y_i = 1.697,80.$$

Considere que as duas variáveis estejam relacionadas através de um modelo de regressão linear simples.

a) Estime a equação de regressão linear simples;

b) Use a equação da linha ajustada para prever que o valor de permeabilidade seria observado quando a resistência à compressão fosse $x = 4,3$;

c) Suponha que o valor observado da permeabilidade em $x = 3,7$ seja $y = 46,1$. Calcule o valor do resíduo correspondente.

8.6.6 Mostre que:

$$\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-\frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n}}{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}$$

8.6.7 Um artigo no *Tappi Journal* (março de 1986) apresentou dados sobre a concentração do licor verde de Na_2S e da produção de uma máquina de papel. Os dados (lidos a partir de um gráfico) são mostrados na seguinte tabela:

Número da observação	1	2	3	4	5	6	7	8	9	10	11	12	13
Concentração (g/l) licor verde de Na_2S	40	42	49	46	44	48	46	43	53	52	54	57	58
Produção (t/dia)	825	830	890	895	890	910	915	960	990	1010	1012	1030	1050

a) Ajuste um modelo de regresso linear simples, relacionando a concentração do licor verde de Na_2S, y, com a produção, x. Desenhe um diagrama de dispersão dos dados e o modelo resultante do ajuste pelo método dos mínimos quadrados;

b) Utilize a análise de variância para testar a regressão;

c) Faça uma análise completa de resíduo.

d) Encontre o coeficiente de correlação e o coeficiente de determinação. Explique o resultado coeficiente de determinação.

8.6.8 Um estudo foi realizado sobre o desgaste de um mancal, y, e sua relação com x_1 = viscosidade do óleo e x_2 = carga. Os seguintes dados foram obtidos:

y	193	230	172	91	113	125
x_1	1,6	15,5	22,0	43,0	33,0	40,0
x_2	851	816	1.058	1.201	1.357	1.115

a) Ajuste um modelo de regressão linear múltipla a esses dados. Interprete o resultado encontrado.

b) Monte a tabela de análise de variância e aplique o teste F;

c) Use o modelo para prever o desgaste, quando $x_1 = 25$ e $x_2 = 1.000$.

8.6.9 (USE O COMPUTADOR) A potência elétrica consumida mensalmente por uma indústria química está relacionada à temperatura média ambiente (x_1), ao número de dias no mês (x_2), à pureza média do produto (x_3) e às toneladas do produto produzido (x_4). Os dados históricos do ano passado estão disponíveis e são apresentados na seguinte tabela:

y	240	236	290	274	301	316	300	296	267	276	288	261
x_1	25	31	45	60	65	72	80	84	75	60	50	38
x_2	24	21	24	25	25	26	25	25	24	25	25	23
x_3	91	90	88	87	91	94	87	86	88	91	90	89
x_4	100	95	110	88	94	99	97	96	110	105	100	98

a) Ajuste um modelo de regressão linear múltipla a esses dados. Interprete o resultado encontrado.

b) Monte a análise de variância e verifique que conclusão é encontrada.

c) Preveja o consumo de potência para um mês em que $x_1 = 75^oF$, $x_2 = 24$ dias, $x_3 = 90\%$ e $x_4 = 98$ toneladas.

8.6.10 (USE O COMPUTADOR) A resistência ao puxamento de um fio colado é uma característica importante. A tabela a seguir fornece informação sobre a resistência ao puxamento (y), a altura da garra (x_1), altura da coluna (x_2), a altura da alça (x_3), comprimento do fio (x_4), largura da cola no molde (x_5) e da largura da cola na coluna (x_6).

y	8,0	8,3	8,5	8,8	9,0	9,3	9,3	9,3	9,5	9,8	10,0	10,3	10,5	10,8	11,0
x_1	5,2	5,2	5,8	6,4	5,8	5,2	5,6	6,0	5,2	5,8	6,4	6,0	6,2	6,2	6,2
x_2	19,6	19,8	19,6	19,4	18,6	18,8	20,4	19,0	20,8	19,9	18,0	20,6	20,2	19,2	17,0
x_3	29,6	32,4	31,0	32,4	28,6	30,6	32,4	32,6	32,2	31,8	32,6	33,4	31,8	32,4	31,4
x_4	94,9	89,7	96,2	95,6	86,5	84,5	88,8	85,7	93,6	86,0	87,1	93,1	83,4	94,5	83,4
x_5	2,1	2,1	2,0	2,2	2,0	2,1	2,2	2,1	2,3	2,1	2,0	2,1	2,2	2,1	1,9
x_6	2,3	1,8	2,0	2,1	1,8	2,1	1,9	1,9	2,1	1,8	1,6	2,1	2,1	1,9	1,8

a) Ajuste o modelo de regressão linear múltipla;

b) Ajuste o modelo de regressão linear múltipla usando como variáveis independentes apenas x_2, x_4 e x_6.

9 Introdução ao Planejamento e Análise de Experimentos

Experimentos são uma parte natural dos processos de tomada de decisão em qualquer área da ciência. Suponha, por exemplo, que um engenheiro civil esteja investigando os efeitos de diferentes métodos de cura sobre a resistência compressiva do concreto. O experimento poderia consistir em fabricar vários corpos de prova de concreto usando cada um dos métodos propostos de cura e então testar a resistência compressiva de cada espécime. Os dados desse experimento poderiam ser usados para determinar qual método de cura deveria ser usado para fornecer a máxima resistência compressiva média.

Se houver somente dois métodos de cura que sejam de interesse, esse experimento poderia ser planejado e analisado usando os métodos de hipóteses estatísticas para duas amostras discutidos anteriormente. Nesse caso o pesquisador tem um único *fator* de interesse (métodos de cura) e há somente dois *níveis* do fator.

Muitos experimentos com único fator requerem que mais de dois níveis do fator sejam considerados. Por exemplo, o engenheiro civil pode querer investigar cinco métodos diferentes de cura, e nessa situação poderá ser utilizado o método de **análise de variância** (ANOVA) para comparar médias, quando houver mais de dois níveis de um único fator.

Técnicas de planejamento de experimentos, baseadas estatisticamente, são particularmente úteis no mundo de engenharia, a fim de melhorar o desempenho de um processo de fabricação. Elas têm também aplicação extensiva no desenvolvimento de novos processos. A maioria dos processos pode ser descrita em termos de muitas variáveis controláveis, tais como temperatura, pressão e taxa de alimentação. Usando planejamento de experimentos, os engenheiros podem determinar que subconjunto das variáveis de processo tem maior influência no desempenho do processo. Os resultados de tal experimento podem conduzir a um melhor rendimento do processo, redução na variabilidade do processo, redução nos tempos de projeto e desenvolvimento, redução nos custos de operação, entre outros.

Métodos de planejamento de experimentos são úteis também em atividades de projeto de engenharia, em que novos produtos sejam desenvolvidos e produtos já existentes seja melhorados. Algumas aplicações típicas de experimentos planejados estatisticamente em projeto de engenharia incluem entre outros a avaliação e comparação de configurações básicas de projeto, avaliação de materiais diferentes e determinação dos parâmetros de projeto dos produtos chaves que causem impacto no desempenho do produto.

O uso de planejamento de experimentos no projeto de engenharia pode resultar em produtos que sejam mais fáceis de fabricar, em produtos que tenham melhores desempenhos no campo e melhor confiabilidade do que seus competidores e em produtos que possam ser projetados,

desenvolvidos e produzidos em menos tempo.

Princípios Básicos da Experimentação

São três os princípios básicos da experimentação, a saber:

i) **Repetição:** Consiste em se terem várias parcelas com o mesmo tratamento. Estaríamos com isso, procurando confirmar a resposta que o indivíduo dá a um determinado tratamento.

ii) **Casualização:** Consiste em se distribuírem os tratamentos pelas parcelas através de sorteio. Com isso, estaremos oferecendo a mesma chance a todos os tratamentos de ocuparem uma determinada posição na área experimental. Elimina-se com isso, a intuição ou desejo involuntário de proteger determinados tratamentos. Estes dois princípios, casualização e repetição são obrigatórios em todos os experimentos.

iii) **Controle Local:** É usado quando a área experimental é heterogênea. Nestes casos, ela é subdividida em áreas menores e homogêneas chamadas de blocos. Em cada uma devem-se colocar todos os tratamentos, de preferência em igual número. Como exemplo, podemos citar os terrenos em declive onde se espera que haja uma grande fertilidade, ou seja, que as partes mais baixas do terrenos ejam mais férteis que as partes mais altas.

Quando a área experimental for homogênea, por exemplo, uma área plana, dispensa-se o controle local; todos os tratamentos com todas as suas repetições são dispostos por sorteio nessa área de modo que todos têm a mesma chance de ocupar qualquer posição. Estes são chamados de **Experimento Inteiramente ao Acaso**.

Experimento Inteiramente ao Acaso (DIC)

Consideremos testes de hipóteses de que três ou mais médias populacionais sejam iguais, como em:

$H_0 : \mu_1 = \mu_2 = \mu_3$ ou Não existem diferenças entre os efeitos dos tratamentos
$H_1 :$ Existe pelo menos, uma diferença entre os efeitos dos tratamentos

O nome deste tipo de experimento se deve ao fato de que é utilizada uma única propriedade, ou característica, para categorizar as populações. Essa característica é, algumas vezes, chamada de **tratamento** ou **fator**.

Suposições:

i) As populações têm distribuições que são aproximadamente normais;

ii) As populações têm a mesma variância;

iii) As amostras são amostras aleatórias simples;

iv) As amostras são independentes umas das outras;

v) As diferentes amostras são de populações que são categorizadas de apenas uma maneira.

Tabela da Análise de Variância

Tabela 9.0.1: Tabela da Análise de Variância

Causa de Variação (CV)	g.l	SQ	QM	F
Tratamentos	$a-1$	SQTrat	$QMTrat = \frac{SQTrat}{a-1}$	$F = \frac{QMTrat}{QMR}$
Resíduo	$a(n-1)$	SQR	$QMR = \frac{SQR}{a(n-1)}$	
Total	$an-1$	SQTotal		

As fórmulas de cálculo das somas quadráticas, para análise de variância com tamanhos iguais de amostra em cada tratamento, são:

$$SQTotal = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y^2}{N} \qquad SQTrat = \sum_{i=1}^{a} \frac{y_i^2}{n} - \frac{y^2}{N} \qquad SQR = SQTotal - SQTrat$$

Em alguns experimentos com único fator, o número de observações sujeitas a cada tratamento pode ser diferente. Dizemos, então que o planejamento está *desbalanceado*. A análise descrita anteriormente é ainda válida, porém leves modificações têm de ser feitas nas fórmulas das somas quadráticas.

As fórmulas de cálculo das somas quadráticas, para análise de variância com tamanhos diferentes em cada tratamento, são:

$$SQTotal = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y^2}{N} \qquad SQTrat = \sum_{i=1}^{a} \frac{y_i^2}{n_i} - \frac{y^2}{N} \qquad SQR = SQTotal - SQTrat$$

E sua respectiva tabela de Análise de Variância:

Tabela 9.0.2: Tabela da Análise de Variância

Causa de Variação (CV)	g.l	SQ	QM	F
Tratamentos	$a-1$	SQTrat	$QMTrat = \frac{SQTrat}{a-1}$	$F = \frac{QMTrat}{QMR}$
Resíduo	$N-a$	SQR	$QMR = \frac{SQR}{N-a}$	
Total	$N-1$	SQTotal		

Observação 9.0.1 Escolher um planejamento balanceado tem duas vantagens importantes. Primeira, o procedimento de teste será relativamente insensível a pequenos desvios da suposição de igualdade de variâncias se as amostras tiverem o mesmo tamanho. Esse não é o caso para amostras de tamanhos diferentes. Segunda, a potência do teste será maximizada se as amostras tiverem o mesmo tamanho.

A regra de decisão para o teste F é:

i) Se o valor calculado de F for maior que o valor de F tabelado, ao nível α de significância e com $a-1$ e $a(n-1)$ graus de liberdade, rejeita-se H_0. O teste é considerado significativo ao nível de α% de probabilidade e admite-se que, ao nível de α% de probabilidade, existe pelo menos uma diferença entre os efeitos dos tratamentos.

ii) Caso o valor calculado de F seja menor ou igual ao valor de F ao nível de α%, não existem evidencias para rejeitar H_0. O teste não é significativo ao nível de α% e, ao nível de α% de probabilidade, não existem diferenças entre os efeitos dos tratamentos.

Exemplo 9.0.1 Um fabricante de papel usado para fabricar sacos de papel pardo está interessado em melhorar a resistência do produto à tensão. A engenharia de produto pensa que a resistência à tensão seja uma função da concentração de madeira de lei na polpa e que a faixa prática de interesse das concentrações de madeira de lei esteja entre 5 e 20%. Um time de engenheiros responsáveis pelo estudo decide investigar quatro níveis de concentração de madeira de lei: 5%, 10%, 15%, e 20%. Eles decidem fabricar seis corpos de prova, para cada nível de concentração, usando uma planta piloto. Todos os 24 corpos de prova são testados, em uma ordem aleatória, em um equipamento de teste de laboratório. Os dados desse experimento são mostrados a seguir:

Concentração de Madeira de Lei (%)	1	2	3	4	5	6	**Totais**	**Médias**
5	7	8	15	11	9	10	60	10,00
10	12	17	13	18	19	15	94	15,67
15	14	18	19	17	16	18	102	17,00
20	19	25	22	23	18	20	127	21,17
							383	15,96

Use a análise de variância pra testar a hipótese de que diferentes concentrações de madeira de lei não afetam a resistência média do papel à tensão. (Use $\alpha = 1\%$).

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ ou Não existem diferenças entre os efeitos das Concentrações de Madeira de Lei

H_1 : Existe pelo menos, uma diferença entre os efeitos das Concentrações de Madeira de Lei

$$SQTotal = \sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}^2 - \frac{y^2}{N} = 7^2 + 8^2 + \cdots + 20^2 - \frac{(383)^2}{24} = 512,96$$

$$SQTrat = \sum_{i=1}^{4} \frac{y_i^2}{n} - \frac{y^2}{N} = \frac{60^2 + 94^2 + 102^2 + 127^2}{6} - \frac{(383)^2}{24} = 382,79$$

$$SQR = 512,96 - 382,79 = 130,17$$

Tabela 9.0.3: Tabela da Análise de Variância para os dados de resistência à tensão

Causa de Variação (CV)	g.l	SQ	QM	F
Tratamentos = Concentração de madeira de lei	3	382,79	127,60	19,60
Resíduo	20	130,17	6,51	
Total	23	512,96		

Como o F calculado $19,60$ é superior ao F tabelado $= 4,94$, rejeitamos H_0 e concluímos que a concentração de madeira de lei na polpa afeta significativamente a resistência do papel ao nível de 1% de significância.

Comparação Múltipla

Suponha que a análise de variância indique que a hipótese nula deva ser rejeitada. Isso implica diferenças entre as médias dos tratamentos; mas que médias exatamente são diferentes não é especificado. Procedimentos para comparar as médias individuais dos tratamentos são chamados de **comparação múltipla**.

Teste de Tukey

É um teste de comparação múltipla entre todas as médias dos tratamentos tomadas duas a duas. O roteiro para a aplicação do Teste Tukey é:

1. Escolha o nível de significância α;
2. Calcule a Diferença Mínima Significativa (DMS), dada por:

   i) No caso de o número de repetições dos tratamentos serem iguais:

$$DMS = q_{(i,\nu)}\sqrt{\frac{QMR}{j}}$$

   sendo:

   $q_{(i,\nu)}$ é a amplitude total studentizada, com i sendo o número de tratamentos e ν o número de graus de liberdade do resíduo. "Estudentizar"(em inglês *to studentize*) é dividir uma variável pelo seu respectivo desvio padrão;

j é o número de repetições dos tratamentos.

ii) No caso de o número de repetições dos tratamentos serem diferentes:

$$DMS = q_{(i,\nu)}\sqrt{\frac{1}{2}\left(\frac{1}{r_i}+\frac{1}{r_k}\right)\cdot QMR}$$

sendo:

r_i é o número de repetições do tratamento i;

r_k é o número de repetições do tratamento k.

3. Cálculo do módulo dos valores de todos os contrastes entre as médias 2 a 2;

4. Comparação do valor de cada contraste com DMS. Se o valor do módulo do contraste for maior que a DMS, o teste é significativo e então as duas médias são consideradas diferentes.

9.1 Experimento Aleatorizado com Blocos Completos

Em muitos problemas de planejamento de experimentos, é necessário planejar o experimento de modo que a variabilidade aparecendo de um fator perturbador possa ser controlada. No delineamento em blocos casualizados, o material experimental é dividido em grupos homogêneos, cada grupo constituindo uma repetição. Cada repetição ou bloco deve conter uma vez cada tratamento, no caso de blocos completos. O objetivo em todas as etapas do experimento é manter o erro, dentro de cada bloco, tão pequeno quanto seja possível na prática. Na condução do ensaio deve ser empregada uma técnica uniforme para todas as parcelas de um mesmo bloco. Quaisquer alterações na técnica de condução ou em outras condições que possam afetar os resultados devem ser feitas entre os blocos.

O procedimento geral para um planejamento aleatorizado com blocos completos consiste em selecionar b blocos e correr uma réplica completa do experimento em cada bloco. Haverá a observações (uma por nível do fator) em cada bloco e a ordem em que essas observações são corridas é designada aleatoriamente dentro do bloco.

A Análise de Variância para um experimento aleatorizado com blocos completos está logo abaixo:

Tabela 9.1.1: Tabela da Análise de Variância

Causa de Variação (CV)	g.l	SQ	QM	F
Tratamentos	a–1	SQTrat	$QMTrat = \frac{SQTrat}{a-1}$	$F = \frac{QMTrat}{QMR}$
Blocos	$b-1$	SQBlocos	$QMBlocos = \frac{SQBlocos}{b-1}$	
Resíduo	$(a-1)(b-1)$	SQR	$QMR = \frac{SQR}{(a-1)(b-1)}$	
Total	$(ab)-1$	SQTotal		

As fórmulas de cálculo para as somas quadráticas na análise de variância para um planejamento aleatorizado com blocos completos são:

$$SQTotal = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{y^2}{ab} \qquad SQTrat = \frac{1}{b}\sum_{i=1}^{a} y_i^2 - \frac{y^2}{ab} \qquad SQBlocos = \frac{1}{a}\sum_{j=1}^{b} y_j^2 - \frac{y^2}{ab}$$

$$SQR = SQTotal – SQTrat – SQBlocos$$

Rejeitaremos a hipótese nula, com um nível de significância α, se o valor calculado de F for maior que o valor de F tabelado com $a–1$ e $(a–1)(b–1)$ graus de liberdade.

As principais características e vantagens em relação ao experimento inteiramente casualizado são:

i) Permite o controle da influência de uma fonte de variação além do efeito de tratamentos, pelo agrupamento hábil das parcelas (controle local);

ii) Dentro de cada bloco (repetição), as condições ambientais devem ser homogêneas, podendo variar de bloco para bloco;

iii) As repetições podem ser distribuídas por uma área maior permitindo conclusões mais gerais.

Exemplo 9.1.1 Os dados na Tabela 9.1.2 referem-se a um ensaio sobre influência da dosagem de K_2O aplicada ao solo, nas propriedades do algodão utilizado na fabricação de fibras. A característica selecionada para análise foi o índice de resistência Pressley. Este índice foi determinado medindo-se a tensão de ruptura de um atado de fibras com uma secção transversal pré-fixada. O ensaio foi conduzido segundo o experimento blocos casualizados com três repetições. As repetições foram constituídas por três máquinas de teste de resistência, cada uma operada por um técnico diferente. Use $\alpha = 5\%$.

Tabela 9.1.2: Índices de resistência "Pressley"de fibras de algodão em um ensaio com cinco dosagens de K_2O

Níveis de K_2O(Kg/ha)	I	II	III	Totais
40	7,62	8,00	7,93	23,55
60	8,14	8,15	7,87	24,16
80	7,76	7,73	7,74	23,23
100	7,17	7,57	7,80	22,54
120	7,46	7,68	7,21	22,35
Totais	38,15	39,13	38,55	115,83

$H_0 : \mu_1 = \mu_2 = \mu_3$ ou Não existem diferenças entre os efeitos da dosagem de K_2O
H_1 : Existe pelo menos, uma diferença entre os efeitos da dosagem de K_2O

$$SQTotal = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{y^2}{ab} = [(7,62)^2 + (8,00)^2 + \cdots + (7,21)^2] - \frac{(115,83)^2}{15} = 1,179$$

$$SQTrat = \frac{1}{b}\sum_{i=1}^{a} y_i^2 - \frac{y^2}{ab} = \frac{1}{3}[(23,55)^2 + \cdots + (22,35)^2] - \frac{(115,83)^2}{15} = 0,7324$$

$$SQBlocos = \frac{1}{a}\sum_{j=1}^{b} y_j^2 - \frac{y^2}{ab} = \frac{1}{5}[(38,15)^2 + (39,13)^2 + (38,55)^2] - \frac{(115,83)^2}{15} = 0,0971$$

$$SQR = 1,179 - 0,7324 - 0,0971 = 0,3495$$

Tabela 9.1.3: Análise de Variância para os dados da tensão de ruptura de fibras de algodão

Causa de Variação (CV)	g.l	SQ	QM	F
Tratamentos	4	0,0971	0,0485	4,19
Blocos	2	0,7324	0,1813	
Resíduo	8	0,3495	0,0437	
Total	14	1,179		

Como F calculado 4,19 é maior que o F tabelado 3,84, rejeita-se a hipótese nula, portanto há pelo menos uma diferença entre os efeitos da dosagem de K_2O.

9.2 Experimento Fatorial

Em muitos experimentos é interessante estudar os efeitos de 2 ou mais fatores conjuntamente. Nestes casos, os Planejamentos Fatoriais são, geralmente, os mais eficientes e mais utilizados.

Um Planejamento Fatorial é um experimento que em cada repetição do experimento, todas as possíveis combinações dos níveis dos fatores são investigadas.

Exemplo 9.2.1 Uma companhia tem interesse em investigar o efeito do preço e do tipo de campanha publicitária nas vendas de um de seus produtos. Para isto, ela vai realizar um experimento considerando 3 preços do produto (R$ 100,00, R$ 110,00, R$ 120,00) e dois tipos de campanha publicitária (anúncio em rádio e anúncio em jornal).

Temos, neste exemplo, 2 fatores: preço do produto (fator A) e tipo de campanha publicitária (fator B) com 3 e 2 níveis respectivamente.

Combinando cada nível de A com um nível de B, obtemos 6 tratamentos em comparação

como mostra o esquema dado em seguida:

	Preços do Produto (Fator A)		
Tipo de Campanha Publicitária (Fator B)	100 (A_1)	110 (A_2)	120 (A_3)
em rádio (nível B_1)	A_1B_1	A_2B_1	A_3B_1
em jornal (nível B_2)	A_1B_2	A_2B_2	A_1B_2

Notemos que os 3 preços considerados são os mesmos para cada tipo de veiculação publicitária. Portanto, num Planejamento Fatorial:

i) Cada nível de um fator está combinado como todos os níveis do outro fator. Dizemos que os fatores obedecem a uma classificação Cruzada;

ii) Combinando cada um dos a níveis de A com cada um dos b níveis de B obtemos ab tratamentos.

9.3 Exercícios

9.3.1 Um trabalho no periódico *Journal of the Association of Asphalt Paving Technologists* (Vol. 59, 1990) descreve um experimento com o objetivo de determinar o efeito de bolhas de ar sobre a percentagem da resistência residual do asfalto. Para finalidades do experimento, bolhas de ar são controladas em três níveis: baixo $(2-4\%)$, médio $(4-6\%)$ e alto $(6-8\%)$. Os dados mostrados na seguinte tabela:

Bolhas de Ar	**Resistência Residual**							
Baixa	106	90	103	90	79	88	92	95
Média	80	69	94	91	70	83	87	83
Alta	78	80	62	69	76	85	69	85

a) Os diferentes níveis de bolhas de ar afetam significativamente a resistência média retida? Use $\alpha = 1\%$.

b) Encontre o $p-valor$ para a estatística F calculada no item a).

9.3.2 Um engenheiro eletrônico está interessado no efeito, na condutividade do tubo, de cinco tipos diferentes de recobrimento de tubos de raios catódicos em uma tela de um sistema de telecomunicações. Os seguintes dados de condutividade são obtidos:

Tipo de Recobrimento	**Condutividade**			
1	143	141	150	146
2	152	149	137	143
3	134	133	132	127
4	129	127	132	129
5	147	148	144	142

a) Há qualquer diferença na condutividade devido ao tipo de recobrimento? Use $\alpha = 5\%$.

b) Caso tenha encontrado diferença significativa, faça o teste de Tukey e diga que tipos de recobrimento resultam em diferentes médias.

9.3.3 A resistência de blocos de cimento está sendo estudada. Quatro diferentes técnicas de produção podem ser usadas economicamente. Os seguintes dados foram coletados:

Técnicas	**Resistência** (lb/in^2)			
1	3129	3000	2865	2890
2	3200	3300	2975	3150
3	2800	2900	2985	3050
4	2600	2700	2600	2765

Testar a hipótese de que as técnicas afetam a resistência do cimento. Use 10% de significância.

9.3.4 Um experimento foi feito para determinar se quatro temperaturas específicas de queima afetam a densidade de um certo tipo de tijolo. O experimento conduziu aos seguintes dados:

Temperatura (°F)	**Densidade**						
100	21,8	21,9	21,7	21,6	21,7	21,5	21,8
125	21,7	21,4	21,5	21,5			
150	21,9	21,8	21,8	21,6	21,5		
175	21,9	21,7	21,8	21,7	21,6	21,8	

A temperatura de queima afeta a densidade dos tijolos? Use $\alpha = 5\%$.

9.3.5 Usando cada um dos três tipos de gasolina em cada um dos quatro carros diferentes, obtemos os resultados apresentados na tabela a seguir:

	BLOCO			
Tipo de Combustível	**Carro 1**	**Carro 2**	**Carro 3**	**Carro 4**
Regular	9,3	9,4	9,6	10,0
Extra	9,4	9,3	9,8	9,9
Premium	9,2	9,4	9,5	9,7

a) Aplique a Análise de Variância correta para testar a afirmação de que o tipo de combustível afetam a milhagem. Use $\alpha = 1\%$.

b) Caso tenha encontrado diferença significativa, faça o teste de Tukey e diga que tipos de combustíveis resultam em diferentes médias

9.3.6 Suponhamos que desejamos determinar se 4 diferentes ponteiras produzem ou não diferentes leituras numa máquina de teste de durabilidade. A máquina opera prensando a ponteira de metal e, da depressão resultante, a durabilidade da placa pode ser determinada. Há 4 ponteiras e 4 espécies de metal avaliados. Cada ponteira é testada uma vez em cada espécie resultando num planejamento Bloco Aleatorizado. Os dados obtidos são mostrados na tabela abaixo:

	Espécie de Metal			
Tipo de Ponteira	1	2	3	4
1	9,3	9,4	9,6	10,0
2	9,4	9,3	9,8	9,9
3	9,2	9,4	9,5	9,7
4	9,7	9,6	10,0	10,2

O tipo de ponteira afeta a durabilidade média? Use 5% de significância.

9.3.7 Os resultados seguintes foram obtidos de um experimento para verificar se operadores diferentes obtiveram resultados médios diferentes numa análise de solo para nitrogênio.

	Operador			
Dia	**A**	**B**	**C**	**D**
Terça	509	512	532	506
Quarta	505	507	542	520
Quinta	465	472	498	483

Em cada um dos 3 dias, uma amostra de solo foi selecionada e dividida em 4 partes. Aleatoriamente, essas 4 partes foram distribuídas para os operadores para fazerem a análise.

Faça uma análise de variância e conclua de acordo com os resultados encontrados. Use 5% de significância.

10 Anexos

10.1 Tabela da Distribuição Normal

Tabela 10.1.1: Distribuição Normal - Valores de $P(0 \leqslant Z \leqslant Z_0)$

Z_0	0	1	2	3	4	5	6	7	8	9
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	0,0675	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	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3138	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	0,2790	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	0,4935	0,4946	0,4948	0,4949	0,4951	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	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4967	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

10.2 Tabela da Distribuição t de *Student*

Tabela 10.2.1: Distribuição t de *Student*

gl	Teste Unilateral								
	15%	10%	5%	2,5%	2%	1%	0,5%	0,1%	0,05%
	Teste Bilateral								
	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%
1	1,9626	3,0777	6,3137	12,7062	15,8945	31,8210	63,6559	318,2888	636,5776
2	1,3862	1,8856	2,9200	4,3027	4,8487	6,9645	9,9250	22,3285	31,5998
3	1,2498	1,6377	2,3534	3,1824	3,4819	4,5407	5,8408	10,2143	12,9244
4	1,1896	1,5332	2,1318	2,7765	2,9985	3,7469	4,6041	7,1729	8,6101
5	1,1558	1,4759	2,0150	2,5706	2,7565	3,3649	4,0321	5,8935	6,8685
6	1,1342	1,4398	1,9432	2,4469	2,6122	3,1427	3,7074	5,2075	5,9587
7	1,1192	1,4149	1,8946	2,3646	2,5168	2,9979	3,4995	4,7853	5,4081
8	1,1081	1,3968	1,8595	2,3060	2,4490	2,8965	3,3554	4,5008	5,0414
9	1,0997	1,3830	1,8331	2,2622	2,3984	2,8214	3,2498	4,2969	4,7809
10	1,0931	1,3722	1,8125	2,2281	2,3593	2,7638	3,1693	4,1437	4,5868
11	1,0877	1,3634	1,7959	2,2010	2,3281	2,7181	3,1058	4,0248	4,4369
12	1,0832	1,3562	1,7823	2,1788	2,3027	2,6810	3,0545	3,9296	4,3178
13	1,0795	1,3502	1,7709	2,1604	2,2816	2,6503	3,0123	3,8520	4,2209
14	1,0763	1,3450	1,7613	2,1448	2,2638	2,6245	2,9768	3,7874	4,1403
15	1,0735	1,3406	1,7531	2,1315	2,2485	2,6025	2,9467	3,7329	4,0728
16	1,0711	1,3368	1,7459	2,1199	2,2354	2,5835	2,9208	3,6861	4,0149
17	1,0690	1,3334	1,7396	2,1098	2,2238	2,5669	2,8982	3,6458	3,9651
18	1,0672	1,3304	1,7341	2,1009	2,2137	2,5524	2,8784	3,6105	3,9217
19	1,0655	1,3277	1,7291	2,0930	2,2047	2,5395	2,8609	3,5793	3,8833
20	1,0640	1,3253	1,7247	2,0860	2,1967	2,5280	2,8453	3,5518	3,8496
21	1,0627	1,3232	1,7207	2,0796	2,1894	2,5176	2,8314	3,5271	3,8193
22	1,0614	1,3212	1,7171	2,0739	2,1829	2,5083	2,8188	3,5050	3,7922
23	1,0603	1,3195	1,7139	2,0687	2,1770	2,4999	2,8073	3,4850	3,7676
24	1,0593	1,3178	1,7109	2,0639	2,1715	2,4922	2,7970	3,4668	3,7454
25	1,0584	1,3163	1,7081	2,0595	2,1666	2,4851	2,7874	3,4502	3,7251
26	1,0575	1,3150	1,7056	2,0555	2,1620	2,4786	2,7787	3,4350	3,7067
27	1,0567	1,3137	1,7033	2,0518	2,1578	2,4727	2,7707	3,4210	3,6895
28	1,0560	1,3125	1,7011	2,0484	2,1539	2,4671	2,7633	3,4082	3,6739
29	1,0553	1,3114	1,6991	2,0452	2,1503	2,4620	2,7564	3,3963	3,6595
30	1,0547	1,3104	1,6973	2,0423	2,1470	2,4573	2,7500	3,3852	3,6460
35	1,0520	1,3062	1,6896	2,0301	2,1332	2,4377	2,7238	3,3400	3,5911
40	1,0500	1,3031	1,6839	2,0211	2,1229	2,4233	2,7045	3,3069	3,5510
50	1,0473	1,2987	1,6759	2,0086	2,1087	2,4033	2,6778	3,2614	3,4960
60	1,0455	1,2958	1,6706	2,0003	2,0994	2,3901	2,6603	3,2317	3,4602
120	1,0409	1,2886	1,6576	1,9799	2,0763	2,3578	2,6174	3,1595	3,3734
$+\infty$	1,0364	1,2816	1,6449	1,9600	2,0537	2,3264	2,5758	3,0902	3,2905

10.3 Tabela da Distribuição Qui-quadrado

$$P(\chi^2 \text{ com } n \text{ graus de liberdade} \geqslant \text{valor tabelado}) = \alpha$$

Tabela 10.3.1: Valores críticos (unilaterais à esquerda) da distribuição Qui-quadrado

	0,995	0,99	0,975	0,95	0,9	0,1	0,05	0,025	0,01	0,005
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	5,308	16,928	18,939	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	21,434	41,422	44,985	48,232	52,191	55,002
32	15,134	16,362	18,291	20,072	22,271	42,585	46,194	49,480	53,486	56,328
33	15,815	17,073	19,047	20,867	23,110	43,745	47,400	50,725	54,775	57,648
34	16,501	17,789	19,806	21,664	23,952	44,903	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	24,797	46,059	49,802	53,203	57,342	60,275
36	17,887	19,233	21,336	23,269	25,643	47,212	50,998	54,437	58,619	61,581
37	18,586	19,960	22,106	24,075	26,492	48,363	52,192	55,668	59,893	62,883
38	19,289	20,691	22,878	24,884	27,343	49,513	53,384	56,895	61,162	64,181
39	19,996	21,426	23,654	25,695	28,196	50,660	54,572	58,120	62,428	65,475
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	3,691	66,766
41	21,421	22,906	25,215	27,326	29,907	52,949	56,942	60,561	64,950	68,053
42	22,138	23,650	25,999	28,144	30,765	54,090	58,124	61,777	66,206	69,336
43	22,860	24,398	26,785	28,965	31,625	55,230	59,304	62,990	67,459	70,616
44	23,584	25,148	27,575	29,787	32,487	56,369	60,481	64,201	68,710	71,892
45	24,311	25,901	28,366	30,612	33,350	57,505	61,656	65,410	69,957	73,166
46	25,041	26,657	29,160	31,439	34,215	58,641	62,830	66,616	71,201	74,437
47	25,775	27,416	29,956	32,268	35,081	59,774	64,001	67,821	72,443	75,704
48	26,511	28,177	30,754	33,098	35,949	60,907	65,171	69,023	73,683	76,969
49	27,249	28,941	31,555	33,930	36,818	62,038	66,339	70,222	74,919	78,231
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490

Observação 10.3.1 Para graus de liberdade que não estão na tabela, isto é acima de 50, use a aproximação:

$$\chi^2_p = \frac{1}{2}\left(Z_p + \sqrt{2k-1}\right)^2$$

onde Z_p é o valor correspondente na normal padrão.

10.4 Tabela da Distribuição F − Snedecor

$$X \sim F_{m,n}$$

$$P(X > F_{m,n,\alpha}) = \alpha$$

m - graus de liberdade do numerador
n - graus de liberdade do denominador

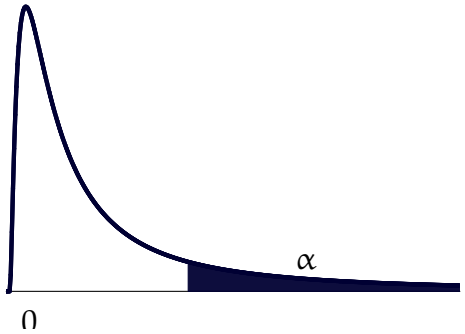


Tabela 10.4.1: Distribuição F − Snedecor

		m																	
n	α	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	0,100	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06
	0,050	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	243,90	245,95	248,02	249,05	250,10	251,14	252,20	253,25
	0,025	647,79	799,48	864,15	899,60	921,83	937,11	948,20	956,64	963,28	968,63	976,72	984,87	993,08	997,27	1001,40	1005,60	1009,79	1014,04
	0,010	4052,18	4999,34	5403,53	5624,26	5763,96	5858,95	5928,33	5980,95	6022,40	6055,93	6106,68	6156,97	6208,66	6234,27	6260,35	6286,43	6312,97	6339,51
2	0,100	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48
	0,050	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49
	0,025	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49
	0,010	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,48	99,48	99,49
3	0,100	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14
	0,050	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55
	0,025	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95
	0,010	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22
4	0,100	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78
	0,050	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66
	0,025	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31
	0,010	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56
5	0,100	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12
	0,050	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40
	0,025	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07
	0,010	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11
6	0,100	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74
	0,050	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70
	0,025	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90
	0,010	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97

		m																	
n	α	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
7	0,100	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49
	0,050	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27
	0,025	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,41	4,36	4,31	4,25	4,20
	0,010	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74
8	0,100	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32
	0,050	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97
	0,025	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73
	0,010	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95
9	0,100	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18
	0,050	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75
	0,025	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39
	0,010	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40
10	0,100	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08
	0,050	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58
	0,025	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14
	0,010	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00
11	0,100	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00
	0,050	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45
	0,025	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94
	0,010	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69
12	0,100	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93
	0,050	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34
	0,025	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79
	0,010	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45
13	0,100	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88
	0,050	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25
	0,025	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66
	0,010	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25
14	0,100	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83
	0,050	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18
	0,025	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55
	0,010	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09
15	0,100	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79
	0,050	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11
	0,025	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46
	0,010	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96

		m																
n	α	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
16	0,100	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75
	0,050	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06
	0,025	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38
	0,010	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84
17	0,100	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72
	0,050	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01
	0,025	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32
	0,010	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75
18	0,100	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69
	0,050	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97
	0,025	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26
	0,010	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66
19	0,100	2,	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67
	0,050	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93
	0,025	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20
	0,010	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58
20	0,100	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64
	0,050	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90
	0,025	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16
	0,010	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52
22	0,100	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60
	0,050	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84
	0,025	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08
	0,010	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40
24	0,100	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57
	0,050	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79
	0,025	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01
	0,010	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31
26	0,100	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54
	0,050	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75
	0,025	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95
	0,010	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23
26	0,100	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52
	0,050	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71
	0,025	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91
	0,010	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17

		m																	
n	α	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
30	0,100	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50
	0,050	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68
	0,025	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87
	0,010	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11
40	0,100	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42
	0,050	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58
	0,025	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72
	0,010	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92
60	0,100	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35
	0,050	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47
	0,025	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58
	0,010	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73
120	0,100	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26
	0,050	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35
	0,025	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43
	0,010	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53