

statística

NOV

**CARLA SILVA DA SILVA
SUZI SAMÁ PINTO**

**Prefácio por
TABAJARA LUCAS DE ALMEIDA**

Estadística

Volume II



UNIVERSIDADE FEDERAL DO RIO GRANDE-FURG

Reitor

DANILO GIROLDO

Vice-Reitor

RENATO DURO DIAS

Chefe de Gabinete do Reitor

JACIRA CRISTIANE PRADO DA SILVA

Pró-Reitor de Extensão e Cultura

DANIEL PORCIUNCULA PRADO

Pró-Reitor de Planejamento e Administração

DIEGO D'ÁVILA DA ROSA

Pró-Reitor de Infraestrutura

RAFAEL GONZALES ROCHA

Pró-Reitora de Graduação

SIBELE DA ROCHA MARTINS

Pró-Reitora de Assuntos Estudantis

DAIANE TEIXEIRA GAUTÉRIO

Pró-Reitora de Gestão e Desenvolvimento de Pessoas

LÚCIA DE FÁTIMA SOCOOWSKI DE ANELLO

Pró-Reitor de Pesquisa e Pós-Graduação

EDUARDO RESENDE SECCHI

Pró-Reitora de Inovação e Tecnologia da Informação

DANÚBIA BUENO ESPÍNDOLA

EDITORA DA FURG

Coordenadora

CLEUSA MARIA LUCAS DE OLIVEIRA

COMITÊ EDITORIAL

Presidente

DANIEL PORCIUNCULA PRADO

Titulares

ANDERSON ORESTES CAVALCANTE LOBATO

ANDRE ANDRADE LONGARAY

ANGELICA CONCEIÇÃO DIAS MIRANDA

CARLA AMORIM NEVES GONÇALVES

CLEUSA MARIA LUCAS DE OLIVEIRA

EDUARDO RESENDE SECCHI

ELIANA BADIALE FURLONG

GIONARA TAUCHEN

LUIZ EDUARDO MAIA NERY

MARCELO GONÇALVES MONTES D'OCA

MARCIA CARVALHO RODRIGUES

RAÚL ANDRÉS MENDOZA SASSI

Editora da FURG

Campus Carreiros

CEP 96203 900 – Rio Grande – RS – Brasil

editora@furg.br

Integrante do PIDL

Editora Associada à



Carla Silva da Silva e Suzi Samá

Estadística

Volume II



Rio Grande
2021

© **Carla Silva da Silva e Suzi Samá**

2021

Capa: Anderson Mendonça

Diagramação da capa: Anael Macedo

Formatação e diagramação:

João Balansin

Gilmar Torchelsen

Cinthia Pereira

Revisão: João Reguffe

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pela Bibliotecária Marcia Carvalho
Rodrigues, CRB 10/1411.

S586e Silva, Carla Silva da
Estatística [recurso eletrônico] : volume II / Carla
Silva da Silva e Suzi Samá. – Dados eletrônicos. –
Rio Grande: Ed. da FURG, 2021.

Modo de acesso: <<http://repositorio.furg.br/>
Edições anteriores em formato de livro impresso.
ISBN: 978-65-5754-047-3 (eletrônico)

1. Estatística matemática. 2. Probabilidade. I.
Pinto, Suzi Samá. II. Título.

CDU, 2. ed.: 519.2

Índice para o catálogo sistemático:

1. Estatística matemática	519.2
2. Probabilidade	519.2

SUMÁRIO

Prefácio	10
1 Introdução à Estatística Inferencial	13
1.1 Conceitos básicos	14
2 Métodos de amostragem	17
2.1 Métodos de amostragem probabilísticos	17
2.2 Métodos de amostragem não-probabilísticos	22
Exercícios complementares	25
Respostas	26
3 Distribuição amostral	28
3.1 Distribuição amostral de médias	29
3.2 Teorema Central do Limite	39
3.3 Determinando probabilidades a partir da distribuição amostral da média	41
Exercícios resolvidos	45
Exercícios complementares	49
Respostas	52

4 Intervalos de confiança	54
4.1 Intervalo de confiança para a média populacional – σ conhecido	56
4.2 Determinação do tamanho da amostra – σ conhecido .	63
4.3 Determinação do tamanho da amostra – σ desconhecido	63
4.4 Intervalo de confiança para a média populacional – σ desconhecido	65
4.4.1 Uso da tabela da distribuição t de Student	68
4.5 Intervalo de confiança para a proporção populacional (p)	72
4.6 Tamanho da amostra para estimar a proporção da população com p conhecido	76
4.7 Tamanho da amostra para estimar a proporção da população com p desconhecido	77
Exercícios resolvidos	78
Exercícios complementares	82
Respostas	86
5 Testes de Hipóteses	88
5.1 Tipos de hipóteses	89
5.2 Erro do tipo I e erro do tipo II	91
5.3 Nível de significância	93
5.4 Testes unilaterais e bilaterais	93
5.5 Estatística de teste	95

5.6 Teste de hipótese para a média	97
5.7 Teste de hipótese para a média com σ conhecido	97
5.8 Teste de hipótese para média com σ desconhecido	104
5.9 Fluxograma	108
5.10 Testes de hipóteses para proporção	109
Exercícios resolvidos	115
Exercícios complementares	118
Respostas	123
6 Inferência estatística para duas amostras	125
6.1 Intervalo de confiança para a diferença de duas médias populacionais com σ_1 e σ_2 conhecidos	126
6.2 Intervalo de confiança para a diferença de duas médias populacionais com σ_1 e σ_2 desconhecidos	128
6.3 Fluxograma	133
6.4 Intervalo de confiança para amostras dependentes	134
6.5 Intervalo de confiança para a diferença de duas proporções populacionais	139
Exercícios resolvidos	141
Exercícios complementares	144
Respostas	151
7 Testes de Hipóteses para duas amostras	153
7.1 Teste de Hipóteses para a diferença de duas médias populacionais com σ_1 e σ_2 conhecidos	153

7.2 Teste de Hipóteses para a diferença de duas médias populacionais com σ_1 e σ_2 desconhecidos	156
7.3 Fluxograma	161
7.4 Teste de Hipóteses para amostras dependentes	162
7.5 Teste de Hipóteses para a diferença de duas proporções populacionais	167
Exercícios resolvidos	172
Exercícios complementares	174
Respostas	182
8 Estudo da variância e do desvio-padrão	184
8.1 A distribuição qui-quadrado (χ^2)	184
8.1.1 Uso da tabela da distribuição qui-quadrado (χ^2)	186
8.2 Intervalo de confiança para variância σ^2 ou desvio-padrão σ	187
8.3 Teste de hipótese para variância e desvio-padrão	192
8.4 Teste qui-quadrado não-paramétrico	196
8.4.1 Tabelas de contingência	196
Exercícios resolvidos	201
Exercícios complementares	206
Respostas	210
9 Análise de regressão	212
9.1 Análise de regressão linear simples	213
9.2 Método dos mínimos quadrados	215

9.2.1 Critério dos mínimos quadrados	216
9.3 Análise de correlação	221
9.3.1 Interpretação do coeficiente de correlação	222
9.4 Coeficiente de determinação	226
Exercícios resolvidos	228
Exercícios complementares	233
Respostas	237
Anexo I – Uso da Tabela da Distribuição Normal Padronizada Z	238
Anexo II – Tabela da Distribuição Normal Padrão	241
Anexo III – Distribuição “ t ” de Student	243
Anexo IV – Distribuição Qui-quadrado χ^2	244

Prefácio

Este é um livro em que eu gostaria de ter estudado. Quando cursei minhas disciplinas de Estatística, passei por professores que não se comunicavam comigo, que não me ensinavam a relevância dos temas em estudo. Os livros que eles adotavam, na época, eram crivados de fórmulas e suas deduções. Eu era muito bom em Matemática, estudava Cálculo com prazer verdadeiro, por isso não me era difícil ser aprovado nas disciplinas de Estatística. Mas não era um estudo prazeroso, pois eu percebia que havia diferença entre a Matemática, que era uma linguagem, e a Estatística, que era uma ferramenta – sobretudo para um engenheiro. Eu queria usar a ferramenta quando fosse profissional, queria saber interpretar os resultados das análises. Mas sabia mesmo era deduzir as fórmulas. Com este histórico pessoal, cabe então explicar por que motivos eu acabei me tornando um professor da área. Mas os motivos não são nobres: fui convidado para lecionar nesta área e tive que aceitar por sobrevivência. No entanto, fiquei com muita vontade de lecionar de modo diferente e fui investigar.

As autoras me redimem. Elas construíram um livro leve, desprezioso, sem ranço intelectual acadêmico, diminuindo a distância entre autor e leitor. Correm o risco de ter construído um material didático tão acessível, que os estudantes poderão até faltar às aulas, como brincou um deles.

Já de início, cuidam de recuperar conceitos básicos pertencentes a um volume anterior, coisa que, por pura preguiça, à vezes se deixa de lado. Os Exercícios Complementares vêm com Respostas. Boa ideia. As Distribuições Amostrais são apresentadas com pouca matemática e, quando isso acontece, logo estampam um exemplo numérico. Assim, as fórmulas assustam menos quem não é amigo da Matemática. Ao final de cada tema há um Quadro Resumo. Boa ideia. As figuras ilustrativas são bastante presentes – deixando o texto menos sólido. Os Intervalos de Confiança são apresentados, logo a seguir, tendo o cuidado de conectá-los com um problema real: para isso eles existem – para resolver problemas reais. E aproveitam para ajudar a responder uma famosa e recorrente pergunta de clientes das assessorias estatísticas: “Que tamanho de amostra eu devo usar?”. Ao final dos exemplos há um destaque para a interpretação. Boa ideia. Aí, as autoras entram na Teoria da Decisão, onde “deitam e rolam”, para usar uma linguagem pouco acadêmica, mas dentro do espírito de comunicação do livro. Esse capítulo é importante. Elas sabem que usamos Estatística para tomar decisões. Afinal, a palavra vem de Estado – o Estado fazia censos e estatísticas para tomar decisões administrativas mais eficientes. Elas vão seguindo o programa estabelecido pela Academia, mas sem descuidar de ligá-lo à realidade. E mostram exemplos didaticamente apresentados. Quase como mastigados. Depois, são estudados os temas de duas amostras, desvio-padrão e variância, como

complementos necessários do conteúdo anterior. As técnicas seguintes são a Análise de Regressão e a de Correlação, ferramentas básicas indispensáveis. A didática é sempre preponderante.

É como se o recado geral fosse: Estudantes: percam o medo da Estatística! A ferramenta é muito útil para vocês.

Definitivamente, eu gostaria de ter estudado neste livro. Talvez eu tivesse começado a pensar estatisticamente mais cedo... E compreendido a vida melhor...

Dr. Tabajara Lucas de Almeida

1 Introdução à Estatística Inferencial

A Estatística Inferencial é um conjunto de técnicas que obtém informações sobre uma população, a partir de resultados observados numa amostra. Como as amostras são formadas por apenas alguns elementos da população, há incertezas com relação às conclusões a que podemos chegar sobre as características de uma população através da análise da amostra. Por exemplo, se cinco estudantes de uma turma obtêm as seguintes notas numa avaliação: 8, 9, 6, 7, 9, é possível, através da estatística descritiva, resumir esta informação. Portanto, podemos afirmar que a média das notas destes cinco estudantes é 7,8, mas não podemos concluir que a média de toda a turma é 7,8. A análise de amostras com a finalidade de inferir sobre a população exige generalizações que vão além dos dados coletados.

Para que possamos inferir sobre a população com base na amostra, são necessários cuidados especiais na seleção desta. Além disso, quando coletamos uma amostra, as medidas estatísticas obtidas não são exatamente iguais aos parâmetros populacionais, devido à variabilidade amostral. Portanto, a probabilidade e a amostragem estão estritamente relacionadas e juntas formam os fundamentos da teoria da inferência.

Antes de iniciarmos o estudo da estatística inferencial, vamos rever alguns conceitos básicos e os principais tipos de amostragem.

1.1 Conceitos básicos

População: é um conjunto de indivíduos ou objetos que apresentam pelo menos uma característica em comum.

Exemplo: Uma pesquisa pretende analisar a opinião dos estudantes de uma universidade com relação à mudança do sistema de avaliação da aprendizagem. Neste caso, pertencem à população todos os estudantes matriculados na instituição.

Uma população pode ser *infinita*, como o número de vezes que se pode lançar uma moeda, ou *finita*, como o número de alunos de uma escola. Em algumas situações a população é tão grande que pode ser tratada como infinita, como por exemplo, número de peixes de determinada espécie, alunos da rede pública de ensino básico.

Amostra: parte da população com as mesmas características. O objetivo de selecionar uma amostra é obter conclusões que possam ser generalizadas para a população, isto é, possam ser inferidas.

Exemplo: para uma pesquisa sobre a opinião dos estudantes da universidade com relação à mudança no processo de aprendizagem, foram selecionados 200 estudantes para compor a amostra.

Parâmetro: é uma descrição numérica de uma característica da população.

Exemplo: um professor calcula a média das notas obtidas pelos estudantes da turma A. Se considerarmos a turma A como

a população a ser considerada, a média obtida é um parâmetro populacional.

Estimador ou estatística amostral: é uma descrição numérica de uma característica da amostra, que será usada no processo de estimação de um parâmetro populacional.

Exemplo: um professor seleciona 10 estudantes entre os 50 estudantes da Turma A e calcula a média das notas destes estudantes. Como essa medida foi calculada com base em uma amostra, é considerada um estimador ou estatística amostral.

Erro amostral: é o erro que ocorre justamente pelo uso da amostra. Ele representa a diferença entre o resultado amostral e o verdadeiro resultado da população. O erro amostral ocorre devido às variações amostrais.

Exemplo: a média da turma A é 8,5 e a média da amostra com os dez estudantes é 7,8. O erro amostral é a diferença entre o parâmetro populacional e a média amostral ($8,5 - 7,8$), ou seja, 0,7.

Amostragem: é o processo que estuda as características de uma população através de uma amostra. A finalidade da amostragem é obter uma indicação do valor de um ou mais parâmetros de uma população, tais como a média, o desvio-padrão e proporção. Os estimadores que correspondem a esses parâmetros populacionais são usados para aproximar os valores desconhecidos daqueles parâmetros. Assim é que a média amostral é usada para estimar a média populacional, o desvio-padrão amostral é usado para estimar o desvio-padrão

populacional...

A partir de agora passaremos a fazer distinção entre parâmetros populacionais e estimadores. Para tal adotaremos a notação do Quadro 1.1.

Quadro 1.1 – Notação para os parâmetros populacionais e estimadores

Medidas estatísticas	Parâmetros populacionais	Estimadores
Média	μ	\bar{X}
Desvio-padrão	σ	s
Variância	σ^2	s^2
Proporções	p	\bar{p}

2 Métodos de amostragem

Na maioria dos problemas de inferência estatística, é impossível ou impraticável observar a população inteira. Logo, dependemos de uma amostra de observações da população para ajudar a tomar decisões acerca da população. Para que nossas inferências sejam válidas, a amostra tem que ser representativa da população.

Existem dois tipos de métodos de amostragem¹: os métodos não-probabilísticos e os probabilísticos. Nos métodos não-probabilísticos, os elementos da população não possuem a mesma probabilidade de compor a amostra. Nesse caso, os resultados obtidos na amostra não podem ser generalizados para a população e muitos métodos estatísticos não podem ser aplicados. Os métodos probabilísticos exigem que cada elemento ou indivíduo da população possua uma probabilidade conhecida (não-nula) de ser selecionado para compor a amostra. Dessa forma, os resultados obtidos na amostra podem ser inferidos para a população.

2.1. Métodos de amostragem probabilísticos

Dentre **os métodos de amostragem probabilísticos**, podemos citar a amostragem aleatória simples, amostragem

¹ Ver vídeo sobre este e outros conceitos do livro no canal de Suzi Samá no Youtube.

sistemática, amostragem estratificada e amostragem por agrupamento ou conglomerados.

Na **amostragem aleatória simples** (AAS), todos os elementos da população têm a mesma probabilidade de serem selecionados para compor a amostra. Ela pode ser selecionada por sorteio, ou quando a população for muito grande os elementos podem ser numerados e em seguida sorteados através de uma tabela de números aleatórios. Atualmente é possível tomar uma amostra com números aleatórios gerados por meio de calculadoras estatísticas ou computadores.

Exemplo 1: O professor deseja selecionar cinco estudantes para apresentar o trabalho realizado. Escreve o nome dos 30 estudantes em pedaços de papel e seleciona os estudantes por sorteio.

Exemplo 2: O dono de uma rede de academias de ginástica deseja fazer uma pesquisa de satisfação entre seus clientes. A amostra será composta por 50 clientes. Como a rede de academias tem 3000 clientes, fica inviável colocar o nome de cada cliente em um pedaço de papel e fazer o sorteio. Como os clientes estão organizados por números, o proprietário utiliza um computador e gera 50 números aleatoriamente de 1 a 3000. Os clientes com o número gerado pelo computador irão compor a amostra.

A **amostragem sistemática** é utilizada quando a população apresenta-se organizada segundo algum critério, de modo tal que cada um de seus elementos possa ser unicamente

identificado pela posição (p. ex., fichas, lista telefônica). Nesse método de amostragem supõe-se que a distribuição dos elementos da população, em uma lista, é aleatória. Nesse caso, a amostragem é realizada por intervalos fixos. Seleciona-se, aleatoriamente, o primeiro elemento que deve estar entre 1 e o fator de sistematização, depois escolhem-se os membros da amostra a intervalos regulares. O fator de sistematização é obtido através da divisão do número de elementos da população (N) pelo número de elementos da amostra (n):

$$\text{fator de sistematização} = \frac{N}{n}$$

O fator de sistematização é arredondado para o número inteiro mais próximo.

Exemplo: se uma amostra sistemática com 30 elementos for selecionada de uma população de 600 indivíduos, o fator de sistematização será de $600/30 = 20$. Um número entre 1 e 20 será escolhido aleatoriamente entre os primeiros indivíduos da população. Suponha que tenha sido escolhido o número 7. O sétimo elemento será o primeiro elemento da amostra; as seleções subsequentes serão 27, 47, 67, 87, ... , 567 e 587.

A **amostragem estratificada** é indicada quando a população encontra-se dividida em grupos distintos (população heterogênea). Dependendo dos objetivos do estudo, a população será dividida em dois ou mais subgrupos, denominados estratos, que compartilham uma característica comum, como sexo, grau de instrução e classe social. Depois

que uma população é dividida em estratos apropriados, podemos fazer uma amostra aleatória simples em cada estrato. Os resultados da amostragem podem então ser ponderados e combinados em estimativas apropriadas da população. Com a estratificação obtemos estratos homogêneos internamente e heterogêneos em relação aos outros estratos. Nessa situação, a estratificação gera amostras mais representativas da população.

O número de elementos de cada estrato que constituirão a amostra é calculado com base em duas informações: (1) o tamanho que deve ter a amostra total e (2) como a amostra total deve ser alocada entre os estratos. As amostras dentro de cada estrato podem ser proporcionais ou desproporcionais ao tamanho do estrato em relação à população.

Exemplo: uma comunidade universitária é formada por 8000 indivíduos, entre professores, estudantes e funcionários. Na Tabela 2.1 é apresentado o número de indivíduos em cada um destes estratos, proporcional ao seu número na população, considerando uma amostra com 5% dos elementos da população.

Tabela 2.1 – Amostragem estratificada

Estratos	População	Amostra
Professores	800	40
Funcionários	1200	60
Estudantes	6000	300
Total	8000	400

Na **amostragem por agrupamento ou conglomerado**, os elementos da população são divididos em grupos, de forma que cada grupo seja representativo da população total. Uma amostra aleatória simples dos grupos é então obtida, e todos os elementos dentro de cada grupo são analisados. Podemos citar como agrupamentos agências, bairros, edifícios ou bairros. A amostragem por agrupamentos resulta em economia de custo, particularmente se a população estiver dispersa por uma extensa área geográfica, pois em um agrupamento muitas observações da amostra podem ser obtidas em tempo relativamente curto, o que possibilita obter um tamanho de amostra maior com um custo total significativamente mais baixo.

Exemplo: Um grupo de nutricionistas investiga a desnutrição dos estudantes nas escolas públicas de um município. Para delinear a amostragem foi feito um levantamento de todas as escolas públicas do município, onde cada escola é considerada um grupo que não difere entre si em relação ao que se pretende medir, neste caso a desnutrição. Na sequência foram sorteadas aleatoriamente algumas escolas (conglomerado) e todos os estudantes de cada escola participaram da pesquisa.

Quadro 2.1 - Resumo dos métodos de amostragem probabilística

Tipo	Descrição
Aleatória simples	A seleção pode ser feita por uma lista aleatória de elementos, por sorteio ou por números gerados por um computador.
Sistemática	População organizada sob algum critério. Começa com um início aleatório e depois a amostragem é realizada por intervalos fixos.
Estratificada	A população é dividida em estratos homogêneos e são selecionadas amostras aleatórias de cada estrato.
Agrupamento	A população é dividida em seções ou grupos e uma amostra aleatória dos grupos é obtida. Todos os elementos de cada grupo são analisados.

2.2. Métodos de amostragem não-probabilísticos

Dentre **os métodos de amostragem não-probabilísticos**, podemos citar a amostragem por conveniência, por julgamento ou por quotas.

Na **amostragem por conveniência**, os elementos ou indivíduos são selecionados com base na sua semelhança presumida com a população e na sua disponibilidade imediata. Frequentemente são entrevistados clientes de estabelecimentos comerciais, possibilitando ao pesquisador fazer contato com grande número de pessoas em curtos períodos de tempo e a baixo custo. Portanto, a amostragem por conveniência tem a vantagem de ser rápida e barata pela fácil seleção da amostra e coleta dos dados, no entanto é difícil avaliar quão representativa

da população é essa amostragem.

Exemplo 1: Um programa de televisão libera um número de telefone para que os telespectadores possam ligar e dar sua opinião sobre determinado assunto.

Exemplo 2: Um repórter de TV faz entrevistas na rua.

Na **amostragem por julgamento**, a pessoa mais conhecedora do assunto a ser pesquisado escolhe intencionalmente os indivíduos ou elementos que ela considera representativos da população para comporem a amostra. Com frequência este é um modo relativamente fácil de selecionar uma amostra. No entanto, a qualidade dos resultados da amostra depende do julgamento da pessoa que faz a seleção.

Exemplo 1: Em estudos sobre o assédio sexual no trabalho, o pesquisador pode entrevistar apenas aqueles que sofreram assédio sexual no trabalho e/ou pessoas que trabalham e desenvolvem pesquisas sobre este assunto.

Exemplo 2: Antes de lançar um novo produto no mercado, algumas empresas o testam entre seus funcionários. Isso porque acredita-se que os funcionários terão reações mais favoráveis em relação ao novo produto do que o público. Dessa forma, se o produto não passar por esse grupo, não tem perspectiva de sucesso no mercado em geral.

Na **amostragem por quotas**, o pesquisador procura obter uma amostra que seja similar à população sob determinado(s) aspecto(s) ou dimensão(ões) considerando as características da população, como sexo, idade, classe social,

entre outras. A amostra deve possuir proporções similares de pessoas com as mesmas características na população. Se acreditarmos que a resposta a uma pergunta pode variar dependendo do sexo da pessoa, então devemos buscar respostas proporcionais de homens e mulheres. Podemos achar também que as pessoas da classe média têm opinião diferente das pessoas da classe baixa sobre determinado assunto, então isso seria um outro aspecto a ser considerado na coleta da amostra. Portanto, podemos pedir ao entrevistador para encontrar pessoas da classe média, sexo feminino e de determinada faixa etária.

Exemplo: Pesquisas de opinião e de marketing.

Quadro 2.2 – Resumo dos métodos de amostragem não-probabilística

Tipo	Descrição
Conveniência	Os elementos são selecionados com base na sua semelhança presumida com a população e na sua disponibilidade imediata.
Julgamento	Pesquisador usa o seu julgamento para escolher intencionalmente os indivíduos ou elementos que ele considera representativos da população.
Quotas	O pesquisador entrevista um número predefinido de pessoas segundo determinados aspectos.

Exercícios complementares

1) Faça a distinção entre parâmetro populacional e estatística amostral.

2) Defina população e amostra.

3) Para cada uma das situações a seguir, explique se a amostra selecionada é representativa da população indicada:

a) Foi selecionada, aleatoriamente, uma amostra de estudantes entre os que estavam na biblioteca durante a tarde de segunda-feira. A pesquisa tinha como finalidade verificar a opinião dos estudantes da universidade com relação à qualidade dos serviços prestados por sua biblioteca.

b) Com o objetivo de avaliar o serviço de coleta do lixo residencial de determinada cidade, pesquisadores entrevistaram seus moradores. Como a cidade está dividida em bairros, foram selecionados aleatoriamente alguns quarteirões de cada bairro e os moradores de todas as casas de cada quarteirão participaram da pesquisa.

4) Indique o tipo de amostragem mais adequado em cada caso:

a) Tendo em vista a impossibilidade de todos os estudantes apresentarem seus trabalhos para a turma, o professor decidiu selecionar cinco estudantes para fazerem a apresentação.

b) O proprietário de uma rede de academias de ginástica deseja fazer uma pesquisa de satisfação com seus clientes. A rede tem 3000 clientes cadastrados em fichas numeradas em ordem crescente.

c) Uma empresa de TV a cabo pretende averiguar o interesse dos moradores de uma cidade em adquirir seus serviços. Tendo em vista a impossibilidade de atender, inicialmente, toda a cidade, a empresa pretende levantar informações para verificar quais os bairros em que o investimento inicial será recuperado mais rapidamente.

5) Elabore um exemplo para um tipo de amostragem probabilística. Primeiro, descreva o objetivo da pesquisa. Com base nesta informação, defina a população. Escolha o tipo de amostragem (estratificada, conglomerados, julgamento,...). Justifique sua escolha.

Respostas

1. Parâmetro populacional é uma descrição numérica de uma característica da população. Estatística amostral é uma descrição numérica de uma característica da amostra.

2. População é todo conjunto de indivíduos, objetos ou produtos que apresentam, pelo menos, uma característica em comum. Amostra é um subconjunto da população que preserva as mesmas características observadas na população.

3. a) Não, pois foram entrevistados apenas os estudantes que frequentaram a biblioteca na segunda-feira à tarde.

b) Sim.

4.a) amostragem aleatória simples;

b) amostragem sistemática;

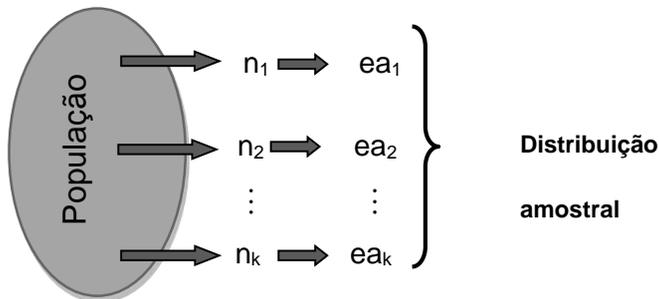
c) amostragem por conglomerados.

5. Sugestão de resposta: Objetivo da pesquisa: verificar a opinião dos estudantes de graduação de uma universidade com relação ao processo de avaliação da aprendizagem. População: estudantes regularmente matriculados em cursos de graduação da instituição. Tipo de amostragem: Tendo em vista que a universidade possui um cadastro de todos os estudantes organizados ordenadamente por número de matrícula e curso, sugerimos amostragem estratificada, sendo os cursos os estratos.

3 Distribuição amostral

Na Estatística Descritiva estudamos medidas estatísticas como média e desvio-padrão que caracterizam uma amostra e cujos valores variam de uma amostra a outra. No estudo de probabilidade estudamos os principais modelos de distribuição de probabilidade, como binomial e normal. Neste capítulo, juntam-se as medidas estatísticas e as distribuições de probabilidade, dando origem às distribuições amostrais. O conhecimento dessas distribuições é a base para aplicar as técnicas de inferência estatística que posteriormente estudaremos.

Se a seleção de todas as amostras possíveis de uma população fosse efetivamente realizada, a distribuição destas estatísticas seria chamada de distribuição amostral. Considere todas as amostras possíveis de tamanho n , retiradas de uma população de tamanho N . Se para cada amostra, calcularmos uma estatística amostral (ea), como, por exemplo, a média ou desvio-padrão, obteremos uma distribuição desses resultados denominada distribuição amostral.



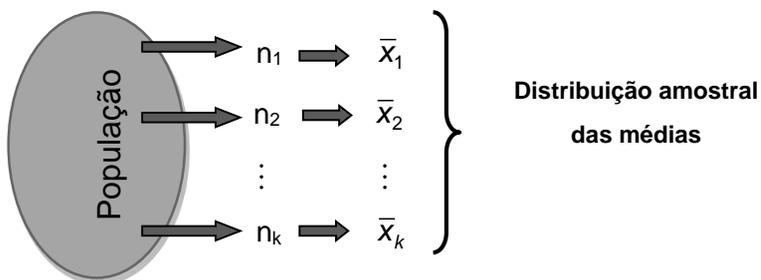
Na prática, uma única amostra é selecionada, aleatoriamente, a partir da população. No entanto, é importante conhecer as características da distribuição amostral desta estatística para poder utilizá-la na estimação do parâmetro da população. Conhecendo a distribuição amostral de determinada estatística, pode-se fazer inferências sobre a população.

Uma **distribuição amostral de uma estatística** é a distribuição de todos os valores da estatística quando todas as amostras possíveis de mesmo tamanho n são extraídas da mesma população.

Para auxiliar na compreensão do conceito de distribuição amostral, apresentamos, a seguir, a distribuição amostral de médias.

3.1 Distribuição amostral de médias

Distribuição amostral de médias é a distribuição de probabilidade de todos os valores possíveis da média amostral.



Exemplo: Dada uma população de tamanho $N = 4$, cujos elementos são os números 3, 7, 11 e 15. A Figura 3.1 apresenta a distribuição dessa população.

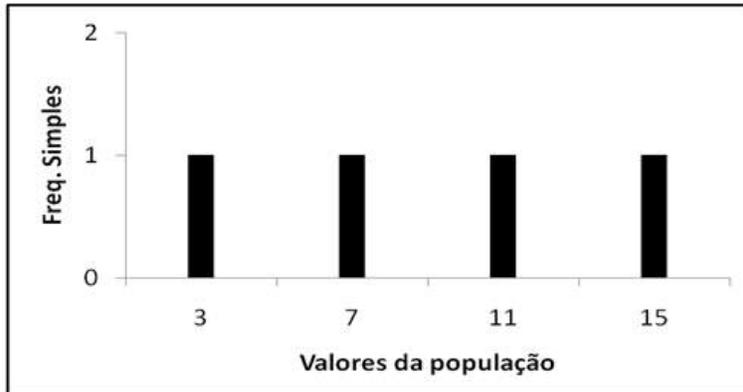


Figura 3.1 – Histograma da população

A média populacional e o desvio-padrão populacional são determinados pelas seguintes fórmulas:

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

No nosso exemplo:

$$\mu = \frac{3 + 7 + 11 + 15}{4} = 9$$

$$\sigma = \sqrt{\frac{(3-9)^2 + (7-9)^2 + (11-9)^2 + (15-9)^2}{4}} = 4,47$$

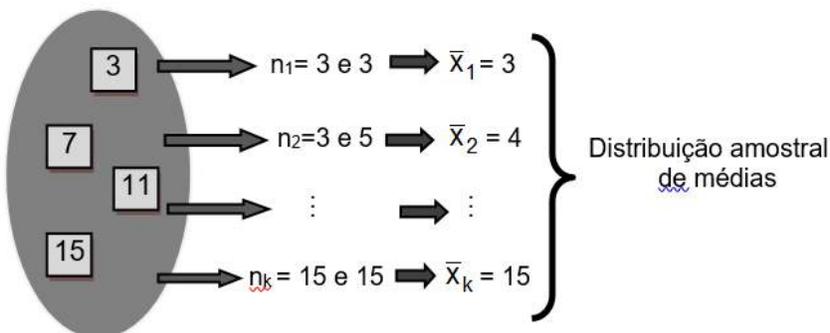
Portanto, esta população apresenta média populacional $\mu = 9$ e desvio-padrão populacional $\sigma = 4,47$.

Vamos retirar todas as amostras possíveis de tamanho 2 desta população para ilustrar que amostras diferentes são possíveis e que estas geram uma variedade de valores para as estatísticas da amostra, neste caso a média amostral.

Primeiramente, construiremos uma distribuição amostral de médias de todas as amostras possíveis de tamanho $n = 2$, **com reposição**, isto é, cada valor selecionado é recolocado antes que a próxima seleção seja feita.

Escreva os números da população em quatro pedaços de papel. Coloque-os em uma caixa. Para selecionar a primeira amostra, escolha ao acaso um pedaço de papel, anote o número e o devolva para a caixa. A seguir selecione outro pedaço de papel. Estes dois números irão compor a primeira amostra. Repita este procedimento sucessivamente até obter todas as amostras possíveis.

Representação da população e sua distribuição amostral de médias:



O Quadro 3.1 apresenta todas as amostras possíveis de tamanho $n = 2$, com reposição, retiradas da população de tamanho $N = 4$ formada pelos elementos 3, 7, 11 e 15.

Quadro 3.1 – Amostras com reposição.

Amostra	Elementos da amostra	Média da amostra \bar{x}
1	3,3	3
2	3,7	5
3	3,11	7
4	3,15	9
5	7,3	5
6	7,7	7
7	7,11	9
8	7,15	11
9	11,3	7
10	11,7	9
11	11,11	11
12	11,15	13
13	15,3	9
14	15,7	11
15	15,11	13
16	15,15	15

O **número de amostras possíveis** de tamanho n de uma população de tamanho N , **com reposição**, pode ser obtido por: N^n Neste exemplo temos $4^2 = 16$ amostras possíveis.

A seguir apresenta-se a distribuição de frequência das médias amostrais e o histograma de frequência dessa distribuição – Figura 3.2.

Distribuição de frequência da distribuição amostral de médias

Médias	Frequência simples
3	1
5	2
7	3
9	4
11	3
13	2
15	1
Total	16

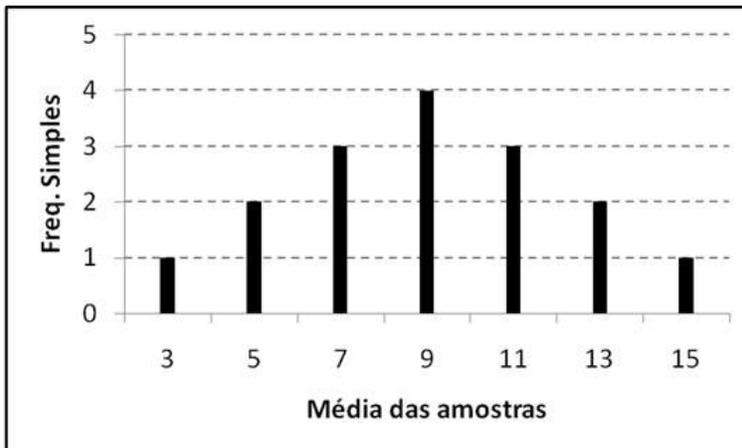


Figura 3.2 – Histograma de frequência simples para amostras com reposição.

Como diferentes amostras aleatórias resultam em uma variedade de valores para a média da amostra, estamos em geral interessados na média de todas as médias amostrais que podem ser geradas pelas várias amostras aleatórias simples. Esta média é denominada de média das médias amostrais e simbolizada por $\mu_{\bar{x}}$.

A média das médias do exemplo acima é:

$$\mu_{\bar{x}} = \frac{3 \times 1 + 5 \times 2 + 7 \times 3 + 9 \times 4 + 11 \times 3 + 13 \times 2 + 15 \times 1}{16} = \frac{144}{16} = 9$$

A média da distribuição amostral de médias, $\mu_{\bar{x}} = 9$, é igual à média da população. A média amostral, em geral, varia de uma amostra para outra, dependendo dos elementos da população que compõem a amostra. Esta medida de variabilidade da média, de amostra para amostra, é expressa pelo desvio-padrão das médias amostrais.

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^N (\bar{x}_i - \mu_{\bar{x}})^2}{\text{número de amostras}}}$$

$$\sigma_{\bar{x}} = \sqrt{\frac{(3-9)^2 + (5-9)^2 + \dots + (13-9)^2 + (15-9)^2}{16}} = 3,16$$

O desvio-padrão das médias amostrais é menor que o desvio-padrão populacional. Isto se justifica pelo fato de que as médias amostrais são menos variáveis que os elementos da população. Uma população pode ter valores muito grandes e muito pequenos. Quando se calcula a média de uma amostra, mesmo que esta contenha algum valor extremo, este será diluído entre os valores dos outros elementos da amostra, o que reduz sua influência no cálculo da média. Quanto maior o número de elementos da amostra, menor será a variabilidade das médias amostrais, pois mais diluídos estarão os valores extremos da população. Portanto, **a quantidade de dispersão**

na distribuição amostral depende da dispersão da população e do tamanho da amostra.

Na prática, é inviável extrair todas as amostras possíveis de uma população. Para determinar o desvio-padrão das médias amostrais usamos a seguinte relação entre o desvio-padrão populacional e o tamanho da amostra:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

O **desvio-padrão da distribuição amostral de médias**, em **amostras com reposição**, é igual ao desvio-padrão populacional dividido pela raiz quadrada do tamanho da amostra:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

No exemplo temos:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4,47}{\sqrt{2}} = 3,16$$

Obs.: O desvio-padrão da distribuição amostral de médias das amostras é frequentemente chamado de **erro padrão da média**.

Muitas pesquisas envolvem **amostragem sem reposição**, de modo que a probabilidade de os elementos restantes na população pertencerem à amostra é alterada. Se o tamanho da amostra é pequeno em relação ao da população, a não-reposição do elemento examinado terá efeito desprezível

nas probabilidades dos elementos restantes, e a amostragem sem reposição não causará grandes dificuldades. Por outro lado, se a amostra é relativamente grande, a probabilidade de os elementos restantes na população pertencerem à amostra é alterada. Dessa forma, a questão da reposição do elemento examinado na população, antes de se proceder à observação seguinte, surge apenas em relação às populações finitas.

Para estudar o efeito da amostragem sem reposição, vamos repetir nosso exemplo com amostras de tamanho $n = 2$, **sem reposição**. O Quadro 3.2 apresenta todas as amostras possíveis de tamanho $n = 2$, sem reposição, retiradas da população de tamanho $N = 4$ formada pelos elementos 3, 7, 11 e 15.

Neste caso, o número de amostras possíveis é dado por:

$$C_{N,n} = \frac{N!}{n!(N-n)!} = \frac{4!}{2!(4-2)!} = 6$$

Quadro 3.2 – Amostras sem reposição.

Amostra	Elementos da amostra	Média da amostra \bar{x}
1	3,7	5
2	3,11	7
3	3,15	9
4	7,11	9
5	7,15	11
6	11,15	13

A seguir são apresenta-se a distribuição de frequência das médias amostrais e o histograma de frequência dessa distribuição – Figura 3.3.

Distribuição de frequência da distribuição amostral de médias	
Médias	Frequência simples
5	1
7	1
9	2
11	1
13	1
Total	6

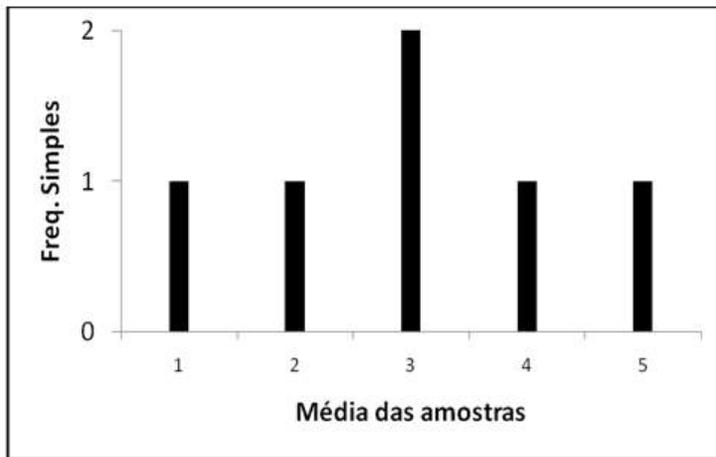


Figura 3.3 – Distribuição de frequência e histograma da distribuição amostral de médias, amostragem sem reposição.

Analisando a distribuição de frequência das médias e o histograma, podemos observar que a distribuição de médias das amostras com e sem reposição tem mesmo centro, ou seja, $\mu_{\bar{x}} = 9$, mas não apresentam a mesma variabilidade.

$$\mu_{\bar{x}} = \frac{5 \times 1 + 7 \times 1 + 9 \times 2 + 11 \times 1 + 13 \times 1}{6} = \frac{54}{6} = 9$$

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^N (\bar{x}_i - \mu_{\bar{x}})^2}{\text{número de amostras}}}$$

$$\sigma_{\bar{x}} = \sqrt{\frac{(5-9)^2 + (7-9)^2 + (9-9)^2 + (9-9)^2 + (11-9)^2 + (13-9)^2}{6}} = 2,58$$

O desvio-padrão da distribuição amostral de médias em amostragem sem reposição é menor do que em amostragem com reposição. Como as amostras são realizadas sem reposição, os valores mais extremos da população não se repetem, o que gera médias amostrais menos variáveis, resultando em menor dispersão. Isso pode ser verificado observando a tabela 3.1, amostras com reposição, em que a menor média é 3 e a maior é 15; já na amostragem sem reposição, tabela 3.2, a menor média observada é 5 e a maior é 13.

Na prática, para determinar o erro-padrão da média utilizando o desvio-padrão da população e o tamanho da amostra, em amostragens sem reposição, é necessário fazer um ajuste no desvio-padrão da distribuição amostral de médias, denominado **fator de correção finita**, e dado pela expressão:

$$\sqrt{\frac{N-n}{N-1}}$$

O **desvio-padrão da distribuição amostral de médias** em amostragens **sem reposição** e populações finitas é dado por:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

O fator de correção finita é usado quando a amostra é sem reposição e a população é finita. Quando o número de elementos da amostra for muito menor que o número de elementos da população (isto é, $n < 5\%N$), o fator de correção finita é insignificante e pode ser omitido, pois seu valor será muito próximo de 1 (um).

No nosso exemplo temos:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{4,47}{\sqrt{2}} \cdot \sqrt{\frac{4-2}{4-1}} = 2,58$$

Quadro 3.3 - Resumo da média e desvio-padrão amostral

- A média das médias amostrais, para um tamanho amostral fixo, é igual à média da população: $\mu_{\bar{x}} = \mu$
 - O desvio-padrão das médias amostrais, em amostras **com** reposição, é dado por: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
 - O desvio-padrão das médias amostrais, em amostras **sem** reposição, é dado por: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$
- Obs.: Quando $n < 5\%N$, o fator de correção finita é insignificante.

3.2 Teorema Central do Limite

Além de saber como determinar a média e o desvio-padrão de uma distribuição amostral de médias, é necessário conhecer a forma da distribuição amostral de médias. O

Teorema Central do Limite envolve duas distribuições diferentes: a distribuição da população original e a distribuição das médias amostrais.

Consideramos dois casos: um no qual se sabe que a população é distribuída normalmente e outro em que a distribuição da população é desconhecida ou não apresenta distribuição normal.

- Se a população original tem distribuição normal, a distribuição amostral das médias extraídas da população também apresentará distribuição normal, para qualquer tamanho de amostra.

- Se a população é desconhecida ou não apresenta distribuição normal, a distribuição amostral de médias pode apresentar distribuição normal ou aproximadamente normal à medida que o tamanho da amostra se torna maior.

Portanto, nem sempre é necessário conhecer a distribuição de uma população para podermos fazer inferência sobre ela a partir de dados amostrais. A única restrição é que o tamanho da amostra seja grande ($n \geq 30$).

A Figura 3.4 apresenta três populações originais diferentes. De cada uma destas populações foram extraídas amostras de diferentes tamanhos. Observe que à medida que o tamanho da amostra vai aumentando, as distribuições amostrais de médias começam a tomar a forma de sino, sendo que as amostras de tamanho 30 apresentam distribuições aproximadamente normais para as três distribuições amostrais.

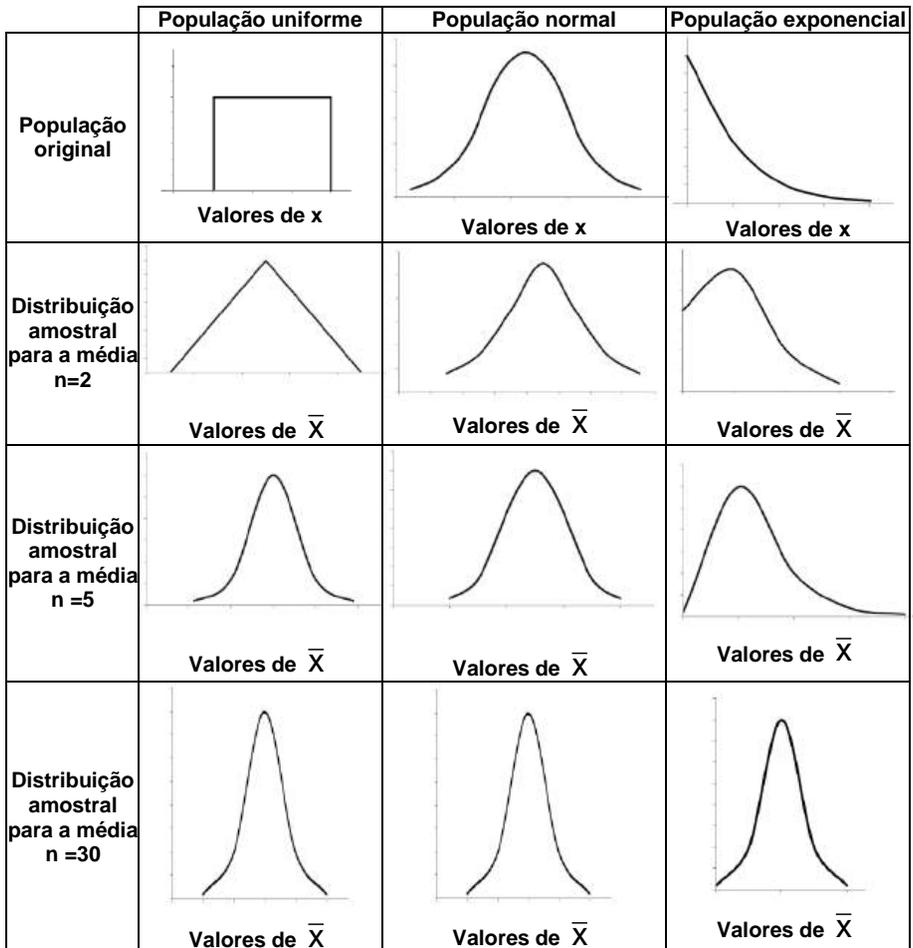


Figura 3.4 – Ilustração do Teorema Central do Limite relativa a diferentes populações, para amostras de tamanho $n=2$, $n=5$ e $n=30$

3.3 Determinando probabilidades a partir da distribuição amostral da média

Muitos problemas podem ser resolvidos com o Teorema Central do Limite. Lembre-se de que, se o tamanho da amostra for igual ou maior que 30, ou se a população for normalmente distribuída, pode-se tratar a distribuição das médias amostrais

como se fosse uma distribuição normal com média $\mu_{\bar{x}}$ e desvio-padrão $\sigma_{\bar{x}}$.

No estudo de distribuições de probabilidade contínuas, vimos como determinar a probabilidade de que uma variável aleatória, de uma população normalmente distribuída com média μ e desvio-padrão σ , esteja em um dado intervalo de valores populacionais. Para tal, transformamos a variável aleatória x na variável padronizada ou em um escore z pela seguinte fórmula:

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

Da mesma forma, pode-se obter a probabilidade de que uma média amostral esteja em um dado intervalo, transformando a média amostral na variável padronizada, z , pela seguinte fórmula:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{ou} \quad z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}^*}, \quad \text{onde} \quad \sigma_{\bar{x}}^* = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Para compreender melhor a diferença entre o cálculo de probabilidade da ocorrência de um valor individual de uma população normalmente distribuída e a média de uma distribuição amostral, vamos desenvolver o seguinte exemplo:

Exemplo 1: Uma máquina de empacotamento que abastece pacotes de feijão apresenta distribuição normal com média de 500g e desvio-padrão de 22g. De acordo com as normas de defesa do consumidor, os pacotes de feijão não

podem ter peso inferior a 2% do estabelecido na embalagem.

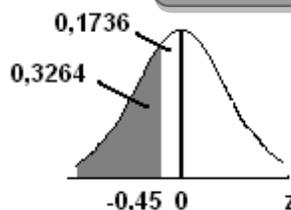
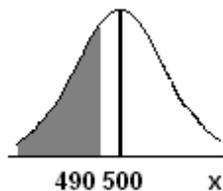
a) Determine a probabilidade de um pacote selecionado aleatoriamente ter peso inferior a 490g.

b) Determine a probabilidade de 20 pacotes selecionadas aleatoriamente, com reposição, terem peso médio inferior a 490g.

c) Considerando que, de acordo com as normas de defesa do consumidor, os pacotes de feijão não podem ter peso inferior a 2% do estabelecido na embalagem, como você interpreta os resultados dos itens anteriores? O que é mais indicado, selecionar um pacote ou uma amostra?

Solução do item a): Estamos com um valor individual, $x=490\text{g}$, de uma população normalmente distribuída com média populacional $\mu=500\text{g}$ e desvio-padrão populacional $\sigma=22\text{g}$. Transformando o peso de 490g para a variável padronizada, z , correspondente, temos:

$$z = \frac{x - \mu}{\sigma} = \frac{490 - 500}{22} = -0,45$$



Uso da tabela z está no Anexo I

Na tabela da distribuição normal padrão (Anexo II), a área entre o centro da curva e o valor de $z = -0,45$ é igual a 0,1736. Como queremos determinar a $P(x < 490\text{g})$, basta subtrair a área de 0,1736 de 0,5 (área correspondente à metade da curva). Portanto, a probabilidade de selecionarmos aleatoriamente um

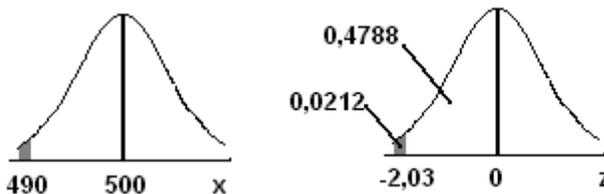
pacote nessa população e este ter menos que 490g é de 32,64%, ou seja, $P(x < 490g) = 32,64\%$.

Solução do item b: Como neste caso estamos lidando com a média de uma amostra de 20 pacotes, usaremos o teorema central do limite: como a população original tem distribuição normal, o teorema central do limite nos garante que a distribuição das médias amostrais também apresentará distribuição normal. Nesse caso, usaremos os parâmetros $\mu_{\bar{x}}$ e $\sigma_{\bar{x}}$, que são calculados como:

$$\mu_{\bar{x}} = \mu = 500 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{22}{\sqrt{20}} = 4,92$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{490 - 500}{4,92} = -2,03$$

Na tabela da distribuição normal padrão, a área entre o centro da curva e o valor de $z = -2,03$ é igual a 0,4788. Como queremos determinar a $P(\bar{x} < 490g)$, basta subtrair 0,4788 de 0,5. Portanto, a probabilidade de selecionarmos aleatoriamente uma amostra de 20 pacotes nesta população e ela ter média menor que 490g é de 2,12%.



$$P(\bar{x} < 490g) = 2,12\%$$

Solução do item c: Há uma probabilidade de 32,64% de um único pacote pesar menos que 490g e há uma probabilidade de 2,12% de 20 pacotes terem peso médio menor que 490g. Portanto, a probabilidade de que a média de uma amostra esteja distante da média populacional é menor do que a probabilidade de que um único valor individual venha a estar distante da média populacional. Isto se deve ao fato de que as médias amostrais apresentam menor variabilidade que os elementos da população. A seleção da amostra é mais indicada para fazer o controle do que a seleção de um único pacote. Como a $P(\bar{x} < 490g) = 2,12\%$, muito baixa, não há motivos de preocupação, pois é pouco provável que o peso das embalagens esteja fora do estabelecido pelas normas de defesa do consumidor.

Exercícios resolvidos

1) Os 400 empregados de uma prestadora de serviços recebem em média R\$800,00, com desvio-padrão de R\$300,00. Os salários apresentam uma distribuição normal. Foi levantada uma amostra com 35 empregados, sem reposição. Determinar:

a) a média e o desvio-padrão da distribuição amostral das médias;

b) qual a probabilidade de uma amostra de 35 empregados apresentar salário médio entre R\$720,00 e R\$850,00;

c) qual a probabilidade de uma amostra apresentar salário médio maior ou igual a R\$870,00.

Dados do exercício:

A variável “salário” apresenta distribuição normal, $\mu =$ R\$800,00 e $\sigma =$ R\$300,00

$N = 400$, $n = 35$, sem reposição.

Solução do item a:

$$\mu_{\bar{x}} = \mu = \text{R}\$800,00$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{300}{\sqrt{35}} \cdot \sqrt{\frac{400-35}{400-1}} = 48,4984$$

$$\sigma_{\bar{x}} = 48,4984$$

Solução do item b: $P(720 \leq \bar{x} \leq 850) = ?$

Padronizando a média amostral, $\bar{x}_1 = 720$, temos:

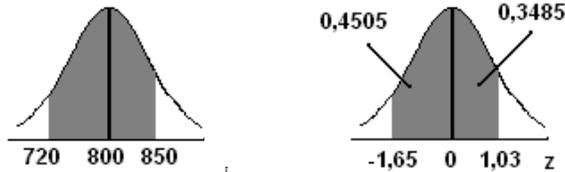
$$z_1 = \frac{\bar{x}_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{720 - 800}{48,4984} = -1,65$$

Logo, para $\bar{x}_1 = 720$ o valor de z_1 correspondente é -1,65

Padronizando a média amostral, $\bar{x}_2 = 850$, temos:

$$z_2 = \frac{\bar{x}_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{850 - 800}{48,4984} = 1,03$$

Logo, para $\bar{x}_2 = 850$ o valor de z_2 correspondente é 1,03



$$P(720 \leq \bar{x} \leq 850) = P(-1,65 \leq z \leq 1,03) = 0,4505 + 0,3485$$

$$P(720 \leq \bar{x} \leq 850) = 0,799$$

Interpretando os resultados: A probabilidade de selecionar uma amostra de 35 empregados, sem reposição, com salário médio entre R\$720,00 e R\$850,00, é de 79,9%.

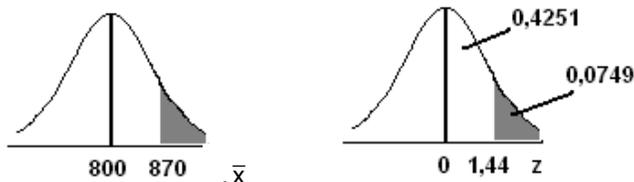
Solução do item c: $P(\bar{x} \geq 870) = ?$

$$P(\bar{x} \geq 870) = ?$$

Padronizando a média amostral, $\bar{x} = 870$, temos:

$$z_1 = \frac{\bar{x}_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{870 - 800}{48,4984} = 1,44$$

Para $\bar{x} = 870$ o valor de z correspondente é 1,44.



$$P(\bar{x} \geq 870) = P(z \geq 1,44) = 0,5 - 0,4251$$

$$P(\bar{x} \geq 870) = 0,0749$$

Interpretando os resultados: A probabilidade de selecionar uma amostra de 35 empregados, sem reposição, com salário médio maior do que R\$ 870,00, é de 7,49%.

2) Uma pesquisa realizada com a finalidade de analisar o número de horas por semana que os 12.000 estudantes universitários dedicam ao estudo acusou média de 7,3 horas e desvio-padrão de 4,2 horas. O tempo de estudo não apresenta distribuição normal. Selecionados aleatoriamente 45 estudantes, determine:

a) a probabilidade de o tempo médio de estudo exceder 8 horas.

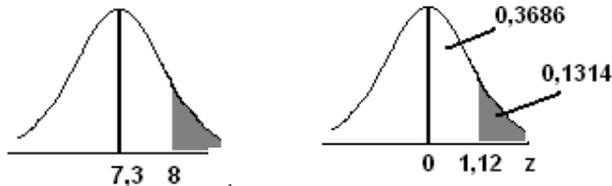
b) a probabilidade de o tempo médio de estudo ser igual ou superior a 7 horas.

Solução do item a: $P(\bar{x} \geq 8) = ?$

Padronizando a média amostral, $\bar{x} = 8$, temos:

$$z_1 = \frac{\bar{x}_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{8 - 7,3}{0,6261} = 1,12$$

Para $\bar{x} = 8$ o valor de z correspondente é 1,12.



$$P(\bar{x} \geq 8) = P(z \geq 1,12) = 0,5 - 0,3686$$

$$P(\bar{x} \geq 8) = 0,1314$$

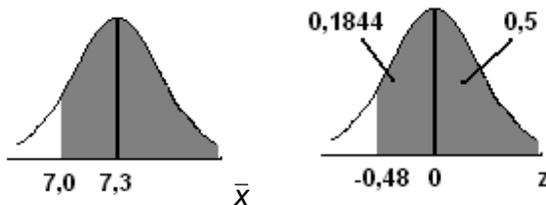
Interpretando o resultado: A probabilidade de selecionar uma amostra de 45 estudantes e o tempo médio de estudo ser maior ou igual a 8 horas é de 13,14%.

Solução do item b: $P(\bar{x} \geq 7) = ?$

Padronizando a média amostral, $\bar{x} = 7$, temos:

$$z_1 = \frac{\bar{x}_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{7 - 7,3}{0,6261} = -0,48$$

Para $\bar{x} = 7$ o valor de z correspondente é $-0,48$.



$$P(\bar{x} \geq 7) = P(z \geq -0,48) = 0,1844 + 0,5 = 0,6844$$

$$P(\bar{x} \geq 7) = 0,6844$$

Interpretando o resultado: A probabilidade de selecionar uma amostra de 45 estudantes e o tempo médio de estudo ser maior ou igual a 7 horas é de 68,44%.

Exercícios complementares

1) A vida média de baterias de uma certa marca é de 920 horas, com desvio-padrão de 55 horas. O tempo de duração das

baterias apresenta distribuição normal. Determinar a probabilidade de uma amostra aleatória de 20 baterias acusar vida média:

Obs.: Desenhe as curvas.

a) entre 910 e 930 horas

b) menos de 905 horas

c) mais de 940 horas

d) entre 900 e 940 horas

2) Refaça o problema anterior para uma amostra aleatória de 60 baterias. Compare as respostas e explique a diferença.

3) Uma linha de montagem produz peças cujos pesos, em gramas, apresentam uma distribuição normal com peso médio de 22,4 g e desvio-padrão de 3,4g. Foram selecionadas 60 amostras com 36 peças cada uma.

a) Qual a probabilidade de que uma amostra apresente peso médio acima de 21,2g?

b) Qual a probabilidade de que uma amostra apresente peso médio abaixo de 23,8g?

c) Em quantas das 60 amostras pode-se esperar que a média encontre-se entre 21,5g e 23,4g?

4) Uma montadora produz automóveis com consumo médio de combustível de 15km/l e desvio-padrão de 3km/l. O consumo apresenta uma distribuição normal. Qual a

probabilidade de uma amostra de 32 automóveis apresentar consumo médio:

a) entre 13,8km/l e 16,1km/l?

b) inferior a 13,8km/l?

c) acima de 18 km/l?

5) Faça a distinção entre parâmetro populacional e estatística amostral.

6) Com base em pesquisas passadas, sabe-se que os estudantes universitários dedicam aos estudos em média 7,2 horas com desvio-padrão de 1,2 horas. Foram selecionados aleatoriamente 35 estudantes:

a) qual a probabilidade de a média amostral ser maior ou igual a 7,8 horas?

b) qual a probabilidade de a média amostral ser menor ou igual a 7,0?

c) qual a probabilidade de a média amostral estar entre 6,9 e 7,5?

7) A altura de 3000 alunos de uma escola tem distribuição normal com média igual a 178cm e desvio-padrão de 11cm. Calcule a probabilidade de uma amostra aleatória, sem reposição, de 200 alunos apresentar média superior a 180cm.

8) Quando devemos usar o fator de correção? Onde?

9) Com base nos resultados dos exercícios 1 e 2 da lista de exercícios complementares, explique o efeito do tamanho da amostra sobre a variabilidade (dispersão) de uma distribuição amostral de médias.

Respostas

1. a) $P(910 \leq \bar{x} \leq 930) = 58,2\%$

b) $P(\bar{x} < 905) = 11,12\%$

c) $P(\bar{x} > 940) = 5,16\%$

d) $P(900 \leq \bar{x} \leq 940) = 89,68\%$

2. a) $P(910 \leq \bar{x} \leq 930) = 84,14\%$

b) $P(\bar{x} < 905) = 1,74\%$

c) $P(\bar{x} > 940) = 0,24\%$

d) $P(900 \leq \bar{x} \leq 940) = 99,52\%$

3. a) $P(\bar{x} > 21,2) = 98,3\%$

b) $P(\bar{x} < 23,8) = 99,32\%$

c) $n = 54$

4. a) $P(13,8 \leq \bar{x} \leq 16,1) = 96,93\%$

b) $P(\bar{x} < 13,8) = 1,19\%$

c) $P(\bar{x} > 18) = 0\%$

5. Parâmetro populacional é uma descrição numérica de uma característica da população. Estatística amostral é uma descrição numérica de uma característica da amostra.

6. a) $P(\bar{x} \geq 7,8) = 0,15\%$

b) $P(\bar{x} \leq 7) = 16,11\%$

c) $P(6,9 \leq \bar{x} \leq 7,5) = 86,12\%$

7. $P(\bar{x} > 180) = 0,39\%$

8. O fator de correção deve ser utilizado quando a amostragem for sem reposição e a população finita. Deve ser aplicado no cálculo do desvio-padrão das médias ou erro-padrão da média.

9. Quanto maior o tamanho da amostra, menor o desvio-padrão das médias, aumentando a probabilidade de encontrarmos valores mais próximos da média das médias ($\mu_{\bar{x}}$) e diminuindo a probabilidade de encontrarmos valores muito afastados da média das médias ($\mu_{\bar{x}}$).

4 Intervalos de confiança

A capacidade de se estimar parâmetros populacionais por meio de amostras está diretamente ligada ao conhecimento da distribuição amostral da estatística que está sendo usada na estimação do parâmetro. A estimação de parâmetros tem inúmeras aplicações. As fábricas, por exemplo, devem continuamente estimar a porcentagem de peças defeituosas em um lote, o índice de gordura no leite, o desvio-padrão da salinidade na água, entre outros.

Existem dois tipos principais de estimativas: estimativa pontual e estimativa intervalar¹. Quando selecionamos uma amostra e calculamos sua média, obtemos uma estimativa pontual da média da população. Como já observado no estudo das distribuições amostrais, as médias amostrais apresentam certa variabilidade, uma vez que dependem dos elementos que são selecionados para compor a amostra. Dificilmente a média obtida na amostra será exatamente igual à média da população, mas provavelmente estará próxima desse valor. Surge então a estimativa intervalar, que consiste em um intervalo de valores em vez de um único valor, também denominado intervalo de confiança. A estimativa pontual é usada como centro do intervalo, depois adiciona-se ou subtrai-se desse valor o erro máximo de estimativa.

¹ Vídeos sobre intervalos de confiança, gravados pelas autoras deste livro, você pode assistir no Youtube no Canal da Profa Suzi Samá.

O erro máximo de estimativa (e) é calculado levando em consideração o nível de confiança ou probabilidade de estar estimando corretamente o verdadeiro valor do parâmetro da população, bem como a dispersão da estatística que está sendo usada para estimar o parâmetro populacional.

O nível de confiança é expresso como uma probabilidade $1-\alpha$ (letra grega alfa), que representa a área sob a curva normal padrão entre dois pontos (Figura 4.1). O nível de confiança fornece a probabilidade de que o intervalo estimado contenha o verdadeiro parâmetro populacional que se quer estimar.

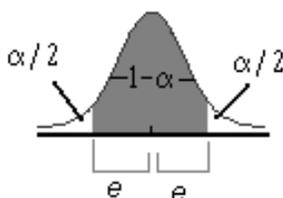


Figura 4.1 – Nível de confiança $1-\alpha$ e o erro máximo de estimativa (e)

Para se construir o intervalo de confiança em torno de uma estatística amostral, é necessário determinar o **erro máximo da estimativa (e)**, o qual é calculado com base no nível de confiança e no desvio-padrão da estatística amostral. O nível de confiança determinará o valor crítico correspondente. Na Figura 4.2 é apresentado um exemplo de intervalo de confiança para a média.

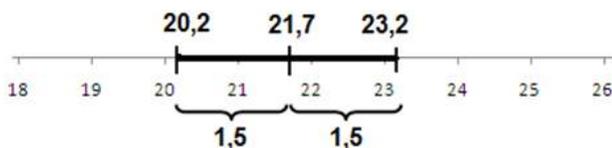
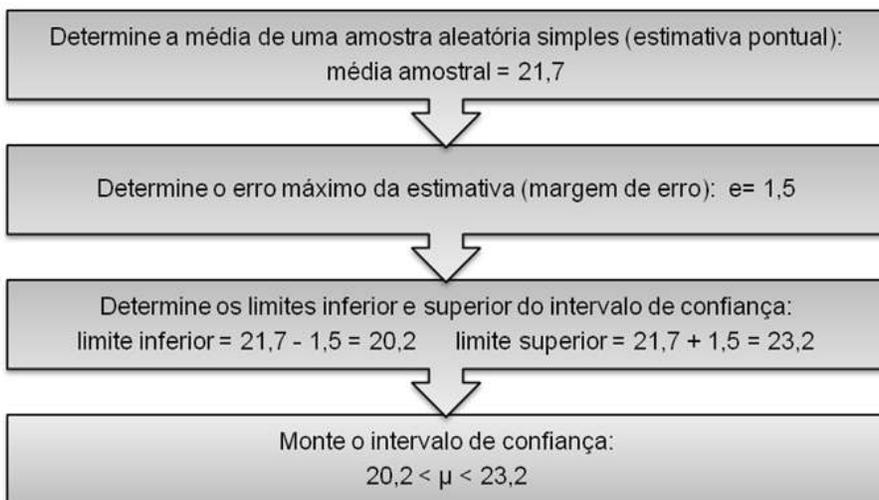


Figura 4.2 – Montagem de um intervalo de confiança usando a média da amostra como exemplo.

4.1 Intervalo de confiança para a média populacional – σ conhecido

Na estimação do intervalo de confiança para a média da população, desta seção, partiremos das seguintes suposições:

- A amostra é uma amostra aleatória simples.
- O valor do desvio-padrão populacional σ é conhecido.
- A população apresenta distribuição normal ou a amostra é maior do que 30 (Teorema Central do Limite).

O **erro máximo da estimativa** pode ser obtido multiplicando-se o valor crítico pelo desvio-padrão da distribuição amostral de médias:

$$e = z_c \cdot \sigma_{\bar{x}}, \quad \text{onde} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

O **intervalo de confiança** para a média é dado por:

$$P(\bar{x} - e \leq \mu \leq \bar{x} + e) = P(-1,65 \leq z \leq 1,03) = 0,4505 + 0,3485$$

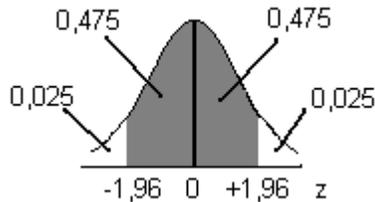
$$P(\bar{x} - e \leq \mu \leq \bar{x} + e) = 1 - \alpha \quad \rightarrow \quad \bar{x} \pm z_c \cdot \frac{\sigma}{\sqrt{n}}$$

O limite inferior do intervalo de confiança é dado por $\bar{x} - e$; o limite superior do intervalo de confiança é dado por $\bar{x} + e$.

Exemplo 1: Uma empresa que fabrica computadores deseja estimar o tempo médio de horas semanais que as pessoas utilizam o computador em casa com uma confiança de 95%. Uma amostra aleatória de 25 pessoas apresentou tempo médio de uso do computador de 22,4 horas. Com base em estudos anteriores, a empresa assume que $\sigma = 5,2$ horas e que os tempos estão normalmente distribuídos.

Solução: Primeiro precisamos verificar se as suposições exigidas são aceitas: a amostragem é aleatória, o desvio-padrão populacional é conhecido e pelo Teorema Central do Limite podemos considerar que a distribuição de médias do grau de satisfação dos clientes é uma distribuição normal, já que a população é normalmente distribuída.

1º. Com base no nível de confiança (1- α), podemos determinar o valor de $Z_{crítico}$.



Para construir um intervalo de confiança de 95%, temos como valores críticos (z_c) -1,96 e +1,96.

2º. Montando o intervalo de confiança:

$$\bar{x} \pm z_c \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow 22,4 \pm 1,96 \frac{5,2}{\sqrt{25}} \Rightarrow 22,4 \pm 2,04$$

$$22,4 - 2,04 \leq \mu \leq 22,4 + 2,04$$

$$20,36 \leq \mu \leq 24,44$$

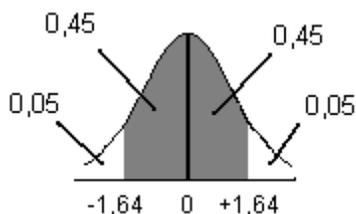
3º. Interpretando o intervalo de confiança: Com 95% de confiança podemos afirmar que o intervalo de 20,36 a 24,44 contém o verdadeiro tempo médio de uso de computadores por adolescentes em casa.

Exemplo 2: Uma empresa de entregas faz pesquisas mensais para verificar o grau de satisfação de seus clientes quanto à rapidez no serviço prestado. Os clientes são selecionados através de uma amostra aleatória simples realizada entre os clientes atendidos durante o mês. Cada

cliente amostrado é questionado sobre seu grau de satisfação com relação ao serviço prestado pela empresa. É solicitado que ele dê uma nota entre 0 (muito insatisfeito) e 10 (muito satisfeito) ao serviço prestado. A média de satisfação da amostra é calculada e usada como estimativa pontual da média de satisfação para toda a população de clientes. Apesar de a média das notas variar de um mês a outro, o valor do desvio-padrão tende a se estabilizar em torno de 1,8. Por isso, assumiremos que o desvio-padrão da população é $\sigma = 1,8$. O levantamento mais recente de satisfação dos clientes forneceu média amostral de $\bar{x} = 8,3$ entre os 100 clientes amostrados. Monte um intervalo de 90% de confiança para o grau médio populacional.

Solução: Primeiro precisamos verificar se as suposições exigidas são aceitas: a amostragem é aleatória simples, o desvio-padrão populacional é conhecido e pelo Teorema Central do Limite podemos considerar que a distribuição de médias do grau de satisfação dos clientes é uma distribuição normal, já que a amostra é maior do que 30.

1º. Com base no nível de confiança (1- α), podemos determinar o valor de $Z_{crítico}$.



Para construir um intervalo de confiança de 90%, temos como valores críticos (z_c) $-1,64$ e $+1,64$.

2º. Montando o intervalo de confiança:

$$\bar{x} \pm z_c \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow 8,3 \pm 1,64 \frac{1,8}{\sqrt{100}} \Rightarrow 8,3 \pm 0,3$$

$$22,4 - 2,04 \leq \mu \leq 22,4 + 2,04$$

$$8,0 \leq \mu \leq 8,6$$

3º. Interpretando o intervalo de confiança: Com 90% de confiança podemos afirmar que o intervalo estimado realmente contém o verdadeiro grau médio de satisfação dos clientes desta empresa.

Exemplo 3: Com os dados do exemplo anterior:

a) Estime um intervalo de 99% de confiança para o grau de satisfação dos clientes quanto à rapidez no serviço prestado, utilizando os dados do exemplo anterior.

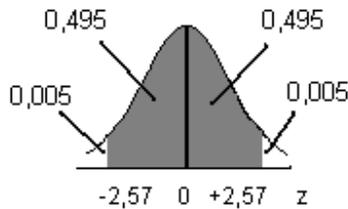
b) Compare o intervalo de confiança do item a com o intervalo de confiança do exemplo 2. Qual possui a maior amplitude entre o limite inferior e o superior? Como o aumento do nível de confiança afeta o intervalo de confiança estimado?

c) Estime um intervalo de 99% de confiança para o grau de satisfação dos clientes quanto à rapidez no serviço prestado para uma amostra de 50 clientes. Compare esse intervalo com o intervalo estimado no item a. O que você pode concluir?

d) Como o tamanho da amostra afeta o intervalo de confiança estimado?

Solução do item a:

1º Com base no nível de confiança $(1 - \alpha)$ podemos determinar o valor de $Z_{crítico}$.



2º Montando o intervalo de confiança:

$$\bar{x} \pm z_c \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow 8,3 \pm 2,57 \frac{1,8}{\sqrt{100}} \Rightarrow 8,3 \pm 0,5$$
$$8,3 - 0,5 \leq \mu \leq 8,3 + 0,5$$

$$7,8 \leq \mu \leq 8,8$$

3º Interpretando o intervalo de confiança: Com 99% de confiança podemos afirmar que o intervalo de 7,8 a 8,8 contém o verdadeiro grau médio de satisfação dos clientes.

Solução do item b: O intervalo de 99% de confiança possui a maior amplitude entre o limite inferior e o superior. Quando aumentamos a confiança no intervalo estimado, aumentamos a amplitude do intervalo.

Solução do item c: Como o tamanho da amostra é de 50 clientes, o que muda no cálculo do intervalo de confiança é o erro máximo de estimativa.

1º. Calculando o erro máximo da estimativa:

$$e = z_c \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow e = 2,57 \cdot \frac{1,8}{\sqrt{50}} = 0,65$$

2º. Montando o intervalo de confiança:

$$8,3 \pm 0,65$$

$$8,3 - 0,65 \leq \mu \leq 8,3 + 0,65$$

$$7,65 \leq \mu \leq 8,95$$

Solução do item d: Diminuindo o tamanho da amostra aumenta a amplitude do intervalo de confiança. Portanto, é necessário um intervalo mais preciso para aumentar o tamanho da amostra.

Concluindo: Do exemplo 3, item c, podemos concluir que, quando aumentamos a confiança no intervalo estimado, aumentamos também a amplitude do intervalo, passando a ter um intervalo de confiança maior, o que diminui a precisão da estimação da verdadeira média populacional. Pelo exemplo 3, item d, concluímos que, quanto maior o número de elementos na amostra, menor a amplitude do intervalo de confiança. Portanto, uma forma de aumentar a precisão da estimação da verdadeira média populacional sem a redução do nível de confiança é aumentar o tamanho da amostra.

4.2 Determinação do tamanho da amostra – σ conhecido

Fixando o erro máximo da estimativa (e) e o nível de confiança ($1 - \alpha$), podemos obter o tamanho da amostra:

$$n = \left(\frac{z_c \cdot \sigma}{e} \right)^2$$

Exemplo: Qual o tamanho de amostra necessária para se estimar a média de uma população com 95% de confiança, erro máximo de estimativa de 0,6 e desvio-padrão populacional igual a 4,3,?

Solução: $n = \left(\frac{z_c \cdot \sigma}{e} \right)^2 = \left(\frac{1,96 \cdot 4,3}{0,6} \right)^2 = 197,31 \quad n = 198$

Obs.: Ao determinar o tamanho da amostra, arredonde sempre para o **inteiro maior** mais próximo.

4.3 Determinação do tamanho da amostra – σ desconhecido

Algumas formas de estimar o valor de σ quando este for desconhecido:

- Realize um estudo-piloto com no mínimo 31 elementos na amostra, calcule o desvio-padrão da amostra, s , e use-o no lugar do desvio-padrão populacional σ .

▪ Use a regra empírica dos 68, 95 e 99,7. Nas distribuições normais, 95% das observações estão num intervalo de 2σ à direita e 2σ à esquerda, como na Figura 4.3.

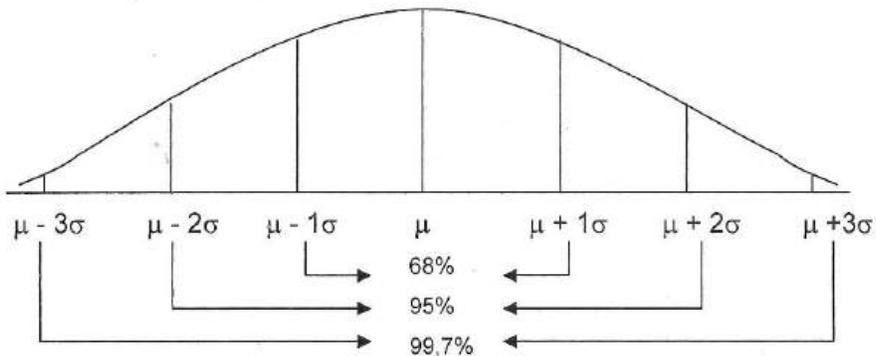


Figura 4.3 – Regra empírica

Considerando esta regra, podemos afirmar que 95% dos dados estão num intervalo de 2σ à esquerda e 2σ à direita da média, logo a amplitude total é aproximadamente 4σ . Portanto,

$$\sigma \approx \frac{\text{amplitude}}{4}$$

Exemplo: Os responsáveis pela livraria de uma universidade resolveram fazer um levantamento do preço médio de venda de livros. Quantos exemplares eles devem selecionar, para que tenham 90% de confiança de que a média amostral esteja a menos de R\$4,00 da média populacional? Como não foram realizados estudos anteriores, o desvio-padrão populacional é desconhecido. Use a informação de que o livro mais barato custa R\$10,00 e o mais caro R\$90,00.

Solução: Aplicando a regra empírica:

$$\sigma \approx \frac{\text{amplitude}}{4} = \frac{90 - 10}{4} = 20$$

$$n = \left(\frac{z_c \cdot \sigma}{e} \right)^2 = \left(\frac{1,64 \cdot 20}{4} \right)^2 = 67,24$$

$$n = 68$$

Obs.: Quando a população for finita e amostragem sem reposição, devemos aplicar o fator de correção finita no cálculo do erro-padrão da média, portanto a expressão para o cálculo do tamanho da amostra será:

$$n = \frac{z_c \cdot \sigma^2 \cdot N}{z_c \cdot \sigma^2 + e^2(N-1)}$$

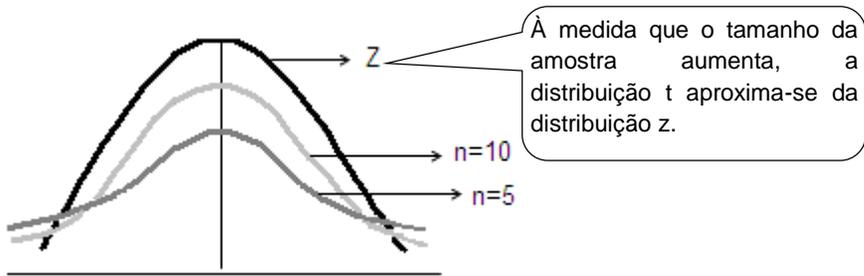
4.4 Intervalo de confiança para a média populacional – σ desconhecido

Quando o desvio-padrão populacional é desconhecido, não podemos usar a distribuição normal padrão. Neste caso usamos a **distribuição t de Student**, desde que a população original apresente distribuição normal. Isso se deve ao fato de que a estimativa intervalar da média populacional se baseia na hipótese de que a distribuição amostral das médias é normal.

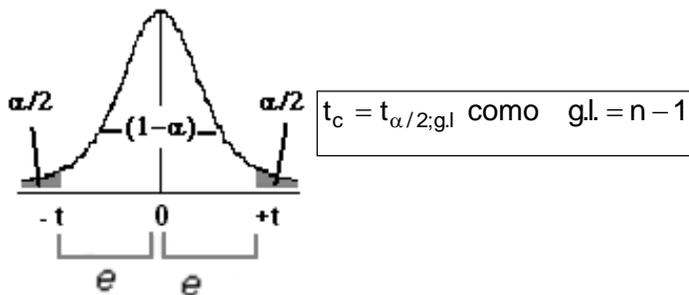
Para grandes amostras isso não apresenta dificuldade especial, pois se aplica o Teorema Central do Limite. No entanto, para amostras pequenas ($n < 30$) é importante saber que a população submetida à amostragem tem distribuição normal, ou ao menos aproximadamente normal. De outra forma essas técnicas não podem ser utilizadas.

A princípio, a distribuição t é muito parecida com a distribuição normal. Ambas as distribuições têm curvas em formato de sino e são simétricas. Entretanto, a distribuição t tem maior área nas caudas e menor área no centro do que a distribuição normal. Isso ocorre porque o desvio-padrão populacional σ é desconhecido e estamos usando o desvio-padrão amostral s para fazer sua estimativa. À medida que se aumenta o tamanho da amostra, a distribuição t aproxima-se da forma da distribuição normal.

Como no processo de inferência cada valor da média é convertido para um valor normal padronizado, sendo σ desconhecido, a fórmula de conversão $(\bar{x} - \mu)/s_{\bar{x}}$ inclui no denominador uma variável diferente para cada média de amostra. O resultado é que a inclusão da variável $s_{\bar{x}}$ em lugar da constante $\sigma_{\bar{x}}$ no denominador gera valores convertidos que não se distribuem como valores Z . Em vez disso, os valores são distribuídos de acordo com a distribuição t de Student.



A distribuição t de Student é uma família de distribuições de probabilidades similares, em que uma específica distribuição t depende de um parâmetro conhecido como graus de liberdade (g.l.). Os graus de liberdade para a estatística t de uma amostra decorrem do desvio-padrão amostral s no denominador da fórmula da distribuição t , pois s tem $n - 1$ graus de liberdade.

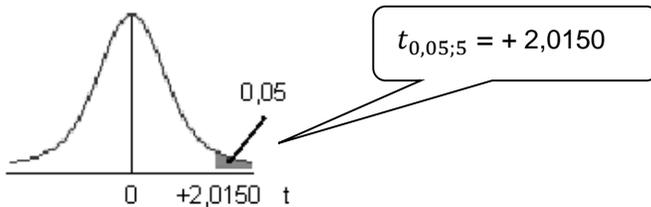


Uma distribuição t é apropriada para inferência sobre a média quando:

- o desvio-padrão populacional σ for desconhecido;
- a população for normalmente distribuída;
- a amostra é uma amostra aleatória simples.

4.4.1 Uso da tabela da distribuição t de *Student*

Para usar a tabela da distribuição t de *Student* (Anexo III), deve-se observar o fato de que a curva é simétrica e centrada na média. O corpo da tabela é constituído das probabilidades (área sob a curva entre t e $+\infty$; t e $-\infty$). Esta tabela trabalha com dois parâmetros: graus de liberdade (g.l) e a área da extremidade da curva. Na primeira linha da tabela está a área da extremidade da curva e na primeira coluna estão os valores dos graus de liberdade. Por exemplo, o valor de $t_{0,05;5}$ (onde 0,05 é a área da extremidade da curva e 5 os graus de liberdade) é obtido pela intersecção da linha que contém o valor 0,05 e a coluna que contém g.l. = 5, como apresentado no Quadro 4.1.



Quadro 4.1 – Uso da Tabela da distribuição *t* de Student

g.l.	Área da extremidade da curva				
	0,1	0,05	0,025	0,01	0,005
1	3,077	0,313	12,70	31,82	63,65
2	1,885	2,920	4,302	6,964	9,925
3	1,637	2,353	3,182	4,540	5,840
4	1,533	2,131	2,776	3,746	4,604
5	1,475	2,015	2,571	3,478	4,291
6	1,439	1,943	2,447	3,249	4,033
7	1,414	1,894	2,365	3,097	3,858
8	1,396	1,859	2,306	2,896	3,355
9	1,383	1,833	2,262	2,845	3,249
10	1,372	1,812	2,228	2,814	3,177
11	1,363	1,795	2,201	2,794	3,129
12	1,356	1,782	2,178	2,681	3,054
13	1,350	1,770	2,160	2,650	3,012
14	1,345	1,761	2,144	2,624	2,976
15	1,340	1,753	2,131	2,602	2,946
16	1,336	1,745	2,119	2,583	2,920
17	1,333	1,739	2,109	2,566	2,898
18	1,330	1,734	2,100	2,552	2,878
19	1,327	1,729	2,093	2,539	2,860
20	1,325	1,724	2,086	2,528	2,845

Área da extremidade da curva

Graus de liberdade:
g.l. = n - 1 = 6 - 1

O erro máximo da estimativa é obtido por:

$$e = t_c \cdot s_{\bar{x}} \quad \text{onde} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$\text{com } g.l. = n - 1 \quad \text{e} \quad t_c = t_{\alpha/2; g.l.}$$

O intervalo de confiança para média quando σ é desconhecido é obtido por:

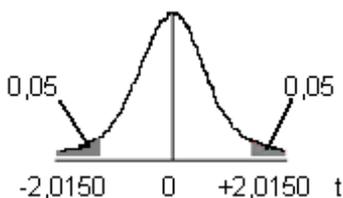
$$P(\bar{x} - e \leq \mu \leq \bar{x} + e) = 1 - \alpha$$

$$\text{ou} \quad \bar{x} \pm t_c \cdot \frac{s}{\sqrt{n}}$$

Exemplo 1: Uma amostra de seis elementos extraída aleatoriamente de uma população com distribuição normal forneceu média $\bar{x} = 10,2$ e desvio-padrão $s = 0,7$. Construa um intervalo de 90% de confiança para a média dessa população.

Solução: Primeiro precisamos verificar se as suposições exigidas são aceitas: a amostra é uma amostra aleatória simples, o desvio-padrão populacional é desconhecido e a população é normalmente distribuída.

1º Com base no nível de confiança $(1 - \alpha)$ e nos graus de liberdade podemos determinar o valor de $t_{crítico}$.



$$g.l. = n - 1 = 6 - 1 = 5$$

$$t_c = t_{\alpha/2; g.l.} = t_{0,05; 5} = 2,0150$$

2º Montando o intervalo de confiança:

$$\bar{x} \pm t_c \cdot \frac{s}{\sqrt{n}} \Rightarrow 10,2 \pm 2,0150 \cdot \frac{0,7}{\sqrt{6}} \Rightarrow 10,2 \pm 0,58$$

$$10,2 - 0,58 \leq \mu \leq 10,2 + 0,58$$

$$9,62 \leq \mu \leq 10,78$$

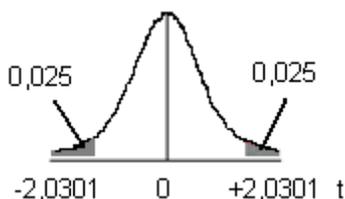
3º Interpretando o intervalo de confiança: Com 90% de confiança podemos afirmar que o intervalo de 9,62 a 10,78 contém a média populacional.

Exemplo 2: O salário médio anual de professores adjuntos de uma determinada universidade segue uma distribuição

normal. Uma amostra aleatória simples de 36 professores apresentou média de R\$86.500,00 e desvio-padrão de R\$6.500,00. Construa um intervalo de 95% de confiança para o salário médio populacional.

Solução: Verificar se as suposições exigidas são aceitas: a amostra é uma amostra aleatória simples, o desvio-padrão populacional é desconhecido e a população é normalmente distribuída.

1º Com base no nível de confiança (1- α) e nos graus de liberdade podemos determinar o valor de $t_{crítico}$.



$g.l. = n - 1 = 36 - 1 = 35$ $t_c = t_{\alpha/2; g.l.} = t_{0,025; 35} = 2,0301$
--

2º Montando o intervalo de confiança:

$$\bar{x} \pm t_c \cdot \frac{s}{\sqrt{n}} \Rightarrow 86.500 \pm 2,0301 \cdot \frac{6.500}{\sqrt{36}} \Rightarrow 86.500 \pm 2.199,28$$

$$84.300,72 \leq \mu \leq 88.699,28$$

3º Interpretando o intervalo de confiança: Com 95% de confiança podemos afirmar que o intervalo de 84.300,72 a 88.699,28 contém o verdadeiro salário médio anual de professores adjuntos.

4.5 Intervalo de confiança para a proporção populacional (p)

Se de uma população do tipo binomial com parâmetros p e $1 - p$ retirarmos todas as amostras possíveis de tamanho n e calcularmos a estatística \bar{p} , o conjunto dessas proporções será dito distribuição amostral das proporções.

Na estimação do intervalo de confiança para a proporção da população, partiremos das seguintes suposições:

- A amostra é uma amostra aleatória simples.
- As condições de uma distribuição binomial são satisfeitas: as n tentativas de um mesmo experimento são independentes; cada tentativa admite apenas dois resultados – sucesso ou fracasso; a probabilidade de sucesso (p) em cada tentativa é constante.
- A distribuição normal pode ser usada para aproximar a distribuição de proporções amostrais, desde que $np \geq 5$ e $nq \geq 5$. Quando p e q são desconhecidos, usa-se a proporção amostral para estimar seus valores, \bar{p} e $(1 - \bar{p})$.

Neste caso, o **erro máximo da estimativa** pode ser obtido multiplicando-se o valor crítico pelo desvio-padrão da estatística

amostral: $e = z_c \cdot s_{\bar{p}}$, onde $s_{\bar{p}} = \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}}$ é o erro-padrão das

proporções para amostras **com reposição** e

$s_{\bar{p}} = \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} \sqrt{\frac{N - n}{N - 1}}$ o erro-padrão das proporções para

amostras **sem reposição**.

O **intervalo de confiança** para a proporção é dado por:

$$P(\bar{p} - e \leq \mu \leq \bar{p} + e) = 1 - \alpha \quad \text{ou} \quad \bar{p} \pm z_c \cdot \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}}$$

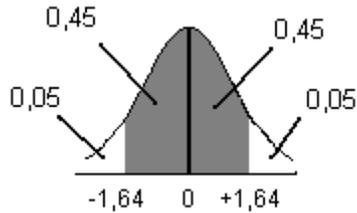
onde o limite inferior do intervalo de confiança é dado por $\bar{p} - e$; e o limite superior do intervalo de confiança é dado por $\bar{p} + e$.

A estatística \bar{p} representa a proporção de sucesso da amostra e $1 - \bar{p}$ a proporção de fracasso da amostra.

Exemplo 1: Uma pesquisa de mercado revela que 247 residências possuem um automóvel, dentre 1.016 residências selecionadas aleatoriamente. Com base nesses resultados, construa um intervalo de 90% de confiança para a proporção de todas as residências que possuem um automóvel.

Solução: Primeiro precisamos verificar se as suposições exigidas são aceitas: a amostragem é aleatória simples, as condições de uma distribuição binomial são satisfeitas e a distribuição normal pode ser usada para aproximar a distribuição de proporções amostrais, desde que $np \geq 5$ ($n \cdot \bar{p} = 246,89$) e $nq \geq 5$ ($n \cdot \bar{q} = 769,11$).

1º Com base no nível de confiança ($1 - \alpha$) podemos determinar o valor de $Z_{crítico}$.



2º Montando o intervalo de confiança:

$$\bar{p} = \frac{x}{n} = \frac{247}{1.016} = 0,243$$

$$\bar{p} \pm z_c \cdot \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} \Rightarrow 0,243 \pm 1,64 \cdot \sqrt{\frac{0,243(1 - 0,243)}{1.016}} \Rightarrow$$

$$0,243 \pm 0,0221$$

$$0,243 - 0,022 \leq p \leq 0,243 + 0,022$$

$$0,221 \leq p \leq 0,265$$

3º Interpretando o intervalo de confiança: Com 90% de confiança podemos afirmar que o intervalo de 0,221 a 0,265 contém a proporção populacional de todas as residências que possuem um automóvel.

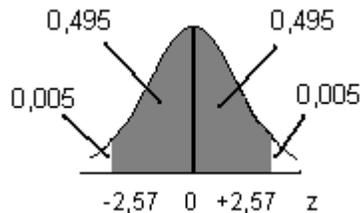
Exemplo 2: Uma amostra de 400 peças extraída aleatoriamente, sem reposição, de uma população de 2000 peças forneceu uma proporção de peças defeituosas de 10%. Determine um intervalo de confiança de 99% para a proporção populacional de peças defeituosas na população.

Solução: Primeiro precisamos verificar se as suposições exigidas são aceitas: a amostragem é aleatória simples, as condições de uma distribuição binomial são satisfeitas e a

distribuição normal pode ser usada para aproximar a distribuição de proporções amostrais, desde que $np \geq 5$ ($n \cdot \bar{p} = 40$) e $nq \geq 5$ ($n \cdot \bar{q} = 360$).

Dados do exemplo: $N = 2000$ $n = 400$ (sem reposição)
 $\bar{p} = 0,1$ $1 - \alpha = 0,99$

1º Com base no nível de confiança ($1 - \alpha$) podemos determinar o valor de $Z_{crítico}$.



2º Montando o intervalo de confiança:

Como a amostragem é sem reposição, calculamos o erro-padrão por:

$$s_{\bar{p}} = \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} \cdot \sqrt{\frac{N - n}{N - 1}}$$

$$s_{\bar{p}} = \sqrt{\frac{0,1 \cdot (1 - 0,1)}{400}} \cdot \sqrt{\frac{2000 - 400}{2000 - 1}} = 0,015 \cdot 0,8947 = 0,0134$$

$$\bar{p} \pm z_c \cdot \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} \cdot \sqrt{\frac{N - n}{N - 1}} \quad \Leftrightarrow \quad 0,1 \pm 2,57 \cdot 0,0134 \quad \Leftrightarrow$$

$$0,1 \pm 0,0344$$

$$0,0656 \leq p \leq 0,1344$$

3º Interpretando o intervalo de confiança: Com 99% de confiança podemos afirmar que o intervalo de 0,0656 a 0,1344 contém a proporção populacional de peças defeituosas.

4.6 Tamanho da amostra para estimar a proporção da população com p conhecido

Fixando o erro máximo da estimativa (e), o nível de confiança ($1 - \alpha$), e conhecendo a proporção de sucesso na população (p), podemos obter o tamanho da amostra:

$$n = \left(\frac{z_c}{e} \right)^2 \cdot p \cdot (1 - p)$$

Exemplo: Um fabricante de peças acredita que aproximadamente 10% de seus produtos são defeituosos. Se ele deseja estimar a verdadeira proporção de peças defeituosas, com um erro máximo de estimativa de 3% e nível de confiança de 90%, qual o tamanho da amostra a ser tomada?

Solução:

$$n = \left(\frac{z_c}{e} \right)^2 \cdot p \cdot (1 - p) = \left(\frac{1,64}{0,03} \right)^2 \cdot 0,1 \cdot (1 - 0,1) = 268,96 \Rightarrow n = 269$$

4.7 Tamanho da amostra para estimar a proporção da população com p desconhecido

Quando a proporção de sucesso na população é desconhecida, usa-se $p = 0,5$ e $p = (1-p) = 0,5$. Observe o Quadro 4.2.

Quadro 4.2 – Comparando os valores de $p = (1-p)$.

p	$(1-p)$	$p.(1-p)$
0,1	0,9	0,09
0,3	0,7	0,21
0,5	0,5	0,25
0,6	0,4	0,24
0,8	0,2	0,16

Como podemos observar, à medida que o valor de p aumenta, o valor de $p.(1-p)$ também aumenta até o valor de $p = 0,5$; para valores de p maiores do que 0,5 o valor de $p.(1-p)$ diminui. Portanto, o maior valor para $p.(1-p)$ é quando $p = 0,5$ e $p.(1-p) = 0,25$. Dessa forma, encontraremos o maior tamanho de amostra possível. Como desconhecemos a proporção de sucesso na população, devemos selecionar a maior amostra possível.

$$n = \left(\frac{z_c}{e} \right)^2 \cdot 0,5 \cdot (1 - 0,5)$$

Exemplo: Suponha que o fabricante do exemplo anterior não tenha ideia da proporção de produtos defeituosos que

fabrica. Se ele deseja estimar a verdadeira proporção de peças defeituosas, com um erro máximo de estimativa de 3% e nível de confiança de 90%, qual o tamanho da amostra a ser tomada?

Solução:

$$n = \left(\frac{z_c}{e}\right)^2 \cdot p(1-p) = \left(\frac{1,64}{0,03}\right)^2 \cdot 0,5(1-0,5) = 747,11 \Rightarrow n = 747$$

Portanto, podemos concluir que, por não ter informação sobre a proporção da população, o fabricante terá que obter a maior amostra possível. Esse tamanho de amostra é determinado quando se usa $p = 0,5$ e $1 - p = 0,5$.

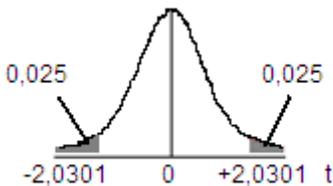
Obs.: Quando a população for finita e a amostragem sem reposição, devemos aplicar o fator de correção finita no cálculo do erro-padrão da média, portanto a expressão para o cálculo do tamanho da amostra será:

$$n = \frac{z^2 \cdot p \cdot (1-p) \cdot N}{(N-1) \cdot e^2 + z^2 \cdot p \cdot (1-p)}$$

Exercícios resolvidos

1) Uma amostra de 36 clientes aleatoriamente selecionados, da classe C, de uma empresa de cartões de crédito apresentou um gasto médio mensal de R\$350,00, com desvio-padrão amostral de R\$34,80. Estime um intervalo de 95% de confiança para o gasto médio mensal dos clientes da classe C dessa empresa. Suponha distribuição normal.

Solução: Primeiro precisamos verificar se as suposições exigidas são aceitas: a amostra é uma amostra aleatória simples, o desvio-padrão populacional é desconhecido e a população é normalmente distribuída.



$$g.l. = n - 1 = 36 - 1 = 35$$

$$t_c = t_{\alpha/2; g.l.} = t_{0,025; 35} = 2,0301$$

Montando o intervalo de confiança:

$$\bar{x} \pm t_c \cdot \frac{s}{\sqrt{n}} \Rightarrow 350 \pm 2,0301 \cdot \frac{34,8}{\sqrt{36}} \Rightarrow 350 \pm 11,77$$

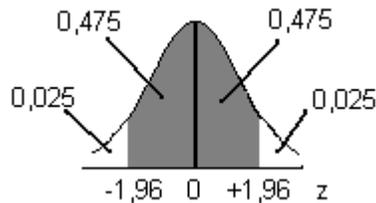
$$350 - 11,77 \leq \mu \leq 350 + 11,77$$

$$338,23 \leq \mu \leq 361,77$$

Interpretação do intervalo de confiança: Com 90% de confiança podemos afirmar que o intervalo de 9,62 a 10,78 contém a média populacional.

2) Uma amostra de dez elementos, aleatoriamente selecionados, apresentou média igual a 20. A população apresenta distribuição normal com desvio-padrão populacional igual a 3. Estime um intervalo de confiança de 95% para a média populacional.

Solução: Primeiro precisamos verificar se as suposições exigidas são aceitas: a amostragem é aleatória, o desvio-padrão populacional é desconhecido e a população apresenta distribuição normal.



Para construir um intervalo de confiança de 95%, temos como valores críticos (z_c) -1,96 e +1,96

Montando o intervalo de confiança:

$$\bar{x} \pm z_c \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow 20 \pm 1,96 \cdot \frac{3}{\sqrt{20}} \Rightarrow 20 \pm 1,86$$

$$20 - 1,86 \leq \mu \leq 20 + 1,86$$

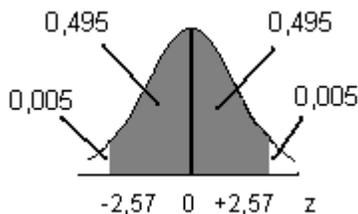
$$18,14 \leq \mu \leq 21,86$$

Interpretação do intervalo de confiança: Com 95% de confiança podemos afirmar que o intervalo de 18,14 a 21,86 contém a média populacional.

3) A fim de estimar a proporção de votos de seu candidato nas próximas eleições, os membros do partido selecionaram aleatoriamente 850 eleitores, obtendo proporção amostral de 20%. Monte um intervalo de confiança de 99% para a proporção populacional.

Solução: Primeiro precisamos verificar se as suposições exigidas são aceitas: a amostragem é aleatória simples, as condições de uma distribuição binomial são satisfeitas e a distribuição normal pode ser usada para aproximar a distribuição de proporções amostrais, desde que $np \geq 5$ ($n \cdot \bar{p} = 170$) e $nq \geq 5$ ($n \cdot \bar{q} = 680$).

1º Com base no nível de confiança $(1 - \alpha)$ podemos determinar o valor de $Z_{crítico}$.



Montando o intervalo de confiança:

$$\bar{p} \pm z_c \cdot \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n}} \Rightarrow 0,2 \pm 2,57 \cdot \sqrt{\frac{0,2(1 - 0,2)}{850}} \Rightarrow 0,2 \pm 0,0352$$

$$0,165 \leq p \leq 0,235$$

Interpretação do intervalo de confiança: Com 99% de confiança podemos afirmar que o intervalo de 0,165 a 0,235 contém a proporção populacional.

4) Uma empresa pretende coletar uma amostra para verificar se a proporção de peças defeituosas em sua produção sofreu alterações. A proporção populacional de peças defeituosas é de 5%. Qual o tamanho da amostra necessária para uma confiança de 99% e erro máximo de estimativa de 4%?

Solução:

$$n = \left(\frac{z_c}{e} \right)^2 \cdot p \cdot (1 - p) = \left(\frac{2,57}{0,04} \right)^2 \cdot 0,05 \cdot (1 - 0,05) = 196,08 \Rightarrow n = 197$$

Exercícios complementares

1) Um pesquisador está interessado em estimar a idade média em que os jovens começam a beber. Com base em pesquisas anteriores, sabe-se que a idade com que o jovem começa a beber apresenta distribuição normal com desvio-padrão populacional de 1,5 ano. Uma amostra de 25 jovens, aleatoriamente selecionados, forneceu idade média de 15,6 anos.

a) Construa um intervalo de 90% de confiança para estimar a média em que os jovens começam a beber.

b) Construa um intervalo de 99% de confiança para estimar a média em que os jovens começam a beber.

c) Compare e interprete os resultados. O que você pode afirmar quanto à influência do aumento do nível de confiança na amplitude do intervalo?

2) Uma pesquisa foi realizada com a finalidade de verificar o número médio de horas que os estudantes do curso de Administração de uma faculdade dormem a cada noite. O quadro a seguir apresenta as horas de sono por noite de cada um dos 24 estudantes, aleatoriamente selecionados.

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

a) Qual é a estimativa pontual do número de horas de sono por noite da população de estudantes?

b) Assumindo que a população tenha distribuição normal, estime um intervalo de confiança de 95% para o número médio de horas de sono a cada noite para a população.

c) Quais as suposições sobre a população (forma da distribuição) e sobre o planejamento do estudo (tipo de amostragem) que são exigidas pelo processo utilizado no item (b)? Qual dessas suposições é mais importante para a validade do processo neste caso?

3) Um estudo pretende estimar a renda média familiar dos empregados de uma empresa. Com base em informações passadas, admite-se que o desvio-padrão da renda familiar é de R\$240,00. Qual deve ser o tamanho da amostra, a fim de que o erro de estimativa da renda média populacional seja no máximo de R\$30,00, com nível de confiança de 95%? Suponha distribuição normal.

4) Um pesquisador deseja estimar o salário médio dos professores do Ensino Médio da rede privada de uma cidade. Quantos professores devem ser selecionados, para termos 90% de confiança de que a média amostral esteja a menos de R\$30,00 da média populacional? (Admitindo que os salários variem entre R\$800,00 e R\$1.200,00).

5) Qual o tamanho de amostra necessário para estimarmos a proporção de peças defeituosas produzidas por uma máquina, com 95% de confiança e erro máximo de 2%, sabendo que a proporção de peças defeituosas historicamente produzidas pela máquina não é superior a 10%?

6) Uma pesquisa de consumo de combustível realizada entre 600 clientes de um posto indicou que 245 tinham carros a álcool e 355 carros a gasolina.

a) Qual a estimativa pontual da proporção de carros a gasolina entre os clientes desse posto?

b) Monte um intervalo de 99% de confiança para a proporção populacional de carros a gasolina desse posto.

7) Com a finalidade de conhecer a proporção de pessoas vacinadas contra a febre amarela em uma cidade, foi selecionada uma amostra de 500 habitantes, dos quais 350 estavam vacinados. Monte um intervalo de 95% de confiança para a proporção de habitantes vacinados contra a febre amarela nessa cidade.

8) Um instituto de pesquisas deseja saber quantas pessoas deve entrevistar, de modo que a proporção de pessoas favoráveis ao uso de células-tronco na amostra difira de menos de 2% da proporção de pessoas favoráveis na população, com confiança de 95%.

9) As pilhas produzidas por uma fábrica apresentam desvio-padrão populacional de 3,8 horas. Uma amostra de 35 pilhas apresentou vida média de 55 horas. Determine um intervalo de confiança de 96% para a vida média populacional.

10) Uma pesquisa de consumo de combustível realizada entre 61 clientes de um posto de combustível indicou um consumo médio de 11,2km/l, com desvio-padrão de 2,8km/l. Determine um intervalo de confiança de 90% para o consumo médio populacional dos clientes desse posto. Suponha distribuição normal para o consumo médio.

11) Os ouvintes de uma rádio reclamam que esta toca mais músicas estrangeiras do que nacionais. Em uma amostra aleatória de 150 músicas, 70 eram nacionais. Monte um intervalo de confiança de 95% para a proporção de músicas nacionais tocadas pela rádio. O que podemos concluir?

12) O serviço de atendimento ao cliente de uma empresa de telefonia pretende realizar uma pesquisa de satisfação entre seus clientes a fim de verificar a proporção de clientes que obtêm a solução de problemas no primeiro contato telefônico. Qual o tamanho da amostra necessária para uma confiança de 95% e erro máximo de estimativa de 10%?

13) Uma amostra de 30 clientes de uma agência bancária indicou o tempo médio de atendimento de 22 minutos. Sabe-se por pesquisas anteriores que o tempo médio de atendimento apresenta uma distribuição normal com desvio-padrão populacional de 6 minutos. Monte um intervalo de 99% de confiança para o tempo médio de atendimento populacional.

14) A vida média das pilhas produzidas por uma fábrica apresenta distribuição normal com desvio-padrão populacional de 3,8 horas. Qual deve ser o tamanho da amostra, a fim de que o erro de estimativa do tempo médio populacional seja no máximo de 1,5 horas com nível de confiança de 99%?

Respostas

1. a) $15,11 \leq \mu \leq 16,09$

b) $14,83 \leq \mu \leq 16,37$

c) pessoal

2. a) $\bar{x} = 6,89$

b) $6,56 \leq \mu \leq 7,22$

c) pessoal

3. $n = 246$

4. $n = 30$

5. $n = 865$

6. a) $\bar{p} = 0,5917$

b) $22,4 - 54,03\% \leq p \leq 64,31\%$

7. $66\% \leq p \leq 74\%$

8. $n = 2.401$

9. $53,68 \leq \mu \leq 56,32$

10. $10,6 \leq \mu \leq 11,8$

11. $0,39 \leq p \leq 0,55$ ou $39\% \leq p \leq 55\%$. Nada podemos concluir, pois o intervalo de confiança para a proporção populacional indica que a proporção de músicas nacionais pode ser superior a 50% (maioria) ou inferior a 50% (minoria).

12. $n = 97$ clientes

13. $19,18 \leq \mu \leq 24,82$

14. $n = 43$ pilhas

5 Testes de Hipóteses

No mercado de trabalho, inúmeras vezes devemos tomar decisões que norteiem o caminho a ser seguido. Para tais decisões, devemos seguir métodos estatísticos, que assegurem a validade da resolução. A Teoria da Decisão é um método que permite testar hipóteses sobre a população com base em informações de dados amostrais. Essas decisões são chamadas de decisões estatísticas.

Neste capítulo, veremos os **testes de hipóteses**¹, procedimentos que permitem testar uma hipótese sobre uma população por meio de dados amostrais. Se os resultados obtidos a partir da amostra não são viáveis, rejeitamos a hipótese sobre a população. Se são plausíveis, mantemos a hipótese e atribuímos os desvios entre a estatística amostral e o parâmetro populacional em estudo, ao erro amostral.

Antes de iniciar o estudo dos testes de hipótese, apresentaremos separadamente cada um de seus componentes, como: hipótese nula, hipótese alternativa, tipos de erro, nível de significância, testes unilaterais ou bilaterais, valor crítico, região crítica e estatística de teste.

¹ Vídeos sobre Testes de Hipóteses, gravados pelas autoras deste livro, você pode assistir no Youtube no Canal da Profa Suzi Samá.

5.1 Tipos de hipóteses

Há dois tipos de hipóteses: a **hipótese nula**, simbolizada por H_0 , que contém uma afirmativa de igualdade tal como $=$, \leq ou \geq , e a **hipótese alternativa**, simbolizada por H_a , que é o complemento da hipótese nula e contém uma afirmativa de desigualdade como \neq , $>$ ou $<$.

Devemos tomar cuidado quando formulamos as hipóteses, pois em alguns casos a sua formulação pode não ser clara. A afirmativa ou alegação a ser testada pode estar na hipótese nula ou na hipótese alternativa, portanto devemos estar seguros de que as hipóteses estejam montadas de maneira adequada, para que a conclusão do teste de hipótese forneça uma decisão confiável.

Vejamos alguns exemplos de como montar as hipóteses em diferentes tipos de problemas.

Exemplo 1: Suponha que o fabricante de um parafuso afirme que o comprimento médio do parafuso fabricado é igual a 5cm.

A afirmação do fabricante é de que o comprimento médio dos parafusos é igual a 5cm. Como esta afirmação contém a igualdade, ela será representada pela hipótese nula. Como a hipótese alternativa é um complemento da hipótese nula, ela terá o sinal de diferente:

$$\begin{aligned}H_0 : \mu &= 5 \\H_a : \mu &\neq 5\end{aligned}$$

Exemplo 2: Um grupo de pesquisadores da área de saúde afirma que o medicamento formulado por eles consegue curar uma determinada doença em mais de 80% dos pacientes testados.

Lembre-se que a hipótese nula sempre contém o sinal de igualdade. Como os pesquisadores afirmam que “MAIS de” 80% dos pacientes se curam, esta alegação será representada pela hipótese alternativa, pois é uma afirmativa de desigualdade, conseqüentemente na hipótese nula teremos o sinal de menor e igual:

$$H_0 : p \leq 0,8$$
$$H_a : p > 0,8$$

Exemplo 3: Um fabricante de suco afirma que em média sua embalagem contém pelo menos 1,1 litro de suco.

O fabricante afirma que sua embalagem contém “pelo menos” 1,1 litro de suco. A expressão “pelo menos” significa que contém 1,1 “ou mais”. Dessa forma, a alegação do fabricante será representada pela hipótese nula, conseqüentemente a hipótese alternativa conterá o sinal de menor (<).

$$H_0 : \mu \geq 1,1$$
$$H_a : \mu < 1,1$$

Na aplicação de um teste de hipótese, podemos ter dois tipos de resultados: **não rejeitar a hipótese nula** ou **rejeitar a**

hipótese nula. Como esse resultado baseia-se em dados amostrais, corremos o risco de tomar a decisão errada, pois podemos não rejeitar uma hipótese nula que na realidade é falsa ou rejeitar uma hipótese nula que na realidade é verdadeira. Estes são os dois tipos de erros que podemos cometer, os quais serão discutidos a seguir.

5.2 Erro do tipo I e erro do tipo II

Quando realizamos um teste de hipótese podemos cometer dois tipos de erro:

Erro do tipo I – significa rejeitar uma hipótese nula verdadeira. A probabilidade de se cometer esse erro é dada por α .

Erro do tipo II – significa não rejeitar uma hipótese nula falsa. A probabilidade de se cometer esse erro é dada por β .

O Quadro a seguir mostra os quatro possíveis resultados de um teste de hipótese.

Decisão	Na realidade	
	H_0 é verdadeira	H_0 é falsa
Não rejeitar H_0	<i>Decisão correta</i>	<i>Erro do tipo II</i>
Rejeitar H_0	<i>Erro do tipo I</i>	<i>Decisão correta</i>

Exemplo: A meteorologia marca a possibilidade de chuva. Portanto, antes de sair de casa precisamos tomar uma decisão – levar ou não o guarda-chuva. O quadro a seguir apresenta as duas decisões corretas que podemos tomar e os dois tipos de erro que podemos cometer:

Decisão	Na realidade	
	Chove	Não chove
Levar o guarda-chuva	<i>Decisão correta</i>	<i>Erro do tipo II</i>
Não levar o guarda-chuva	<i>Erro do tipo I</i>	<i>Decisão correta</i>

Neste exemplo prático, somos capazes de analisar os erros que podemos cometer em um teste de hipótese. Se levarmos o guarda-chuva e chover, significa que tomamos a decisão correta; se não chover, estaremos cometendo um erro do tipo II, o que implica carregarmos o guarda-chuva que não usaremos. Se não levarmos o guarda-chuva e não chover, também estaremos tomando uma decisão correta; no entanto, se chover estaremos cometendo um erro do tipo I, o que implicaria tomarmos um banho de chuva. Neste contexto, não levar o guarda-chuva quando na realidade chove é o erro que não gostaríamos de cometer, o que torna o erro do tipo I mais importante que o erro do tipo II.

Assim, na realização de um teste para verificar a eficácia de uma nova vacina, uma opção conservadora, mas prudente, consiste em dar preferência a errar ao dizer que a nova vacina não apresenta resultados melhores do que a anterior, quando na verdade apresenta, do que errar ao dizer que uma nova vacina é melhor do que a já utilizada, quando na verdade não é. Por este motivo, neste livro trabalharemos com o erro do tipo I, ou seja, rejeitar uma hipótese nula verdadeira, onde a probabilidade de se cometer esse erro é dada pelo nível de significância (α).

5.3 Nível de significância

O **nível de significância** (α) é a probabilidade de ocorrer um erro do tipo I. Antes de iniciar o teste o pesquisador determina o nível de risco (α) que pode ser tolerado ao se rejeitar a hipótese nula quando ela for verdadeira (erro do tipo I). Portanto, o risco de se cometer o erro do tipo I está diretamente sob o controle do pesquisador. Se escolhermos aplicar um teste com um nível de significância de 5% ou 0,05, teremos cerca de 5 chances em 100 de rejeitarmos a hipótese nula e ela ser, na realidade, verdadeira.

5.4 Testes unilaterais e bilaterais

Nos testes de hipóteses podemos ter dois tipos de situações, na tentativa de detectar desvios significativos de um determinado parâmetro. Esses desvios podem ser apenas em uma direção, **teste unilateral** (Figura 5.1), ou em ambas as direções, **teste bilateral**, (Figura 5.2). Identificamos essas características quando montamos a hipótese alternativa do problema.

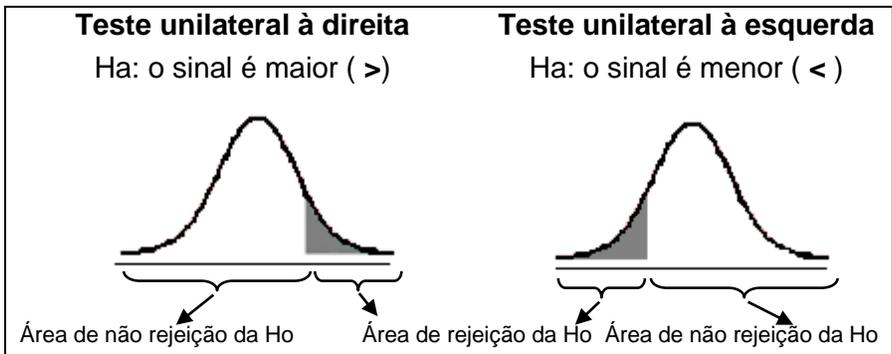


Figura 5.1 – Teste unilateral

O teste da cauda à esquerda é útil para verificar se determinado padrão mínimo foi atingido – por exemplo, conteúdo mínimo de gordura no leite, peso líquido de pacotes de determinado produto, resistência de correias à tensão. Já um teste de cauda à direita é útil para testar se determinado padrão máximo não foi excedido – por exemplo, teor máximo de gordura permitido em determinado alimento, número de unidades defeituosas numa remessa de certa mercadoria, quantidade de poluição atmosférica ocasionada por uma fábrica.

Na prática, usam-se os testes bilaterais sempre que a divergência crítica é em ambas as direções, tal como ocorreria na fabricação de roupas, em que as camisas muito grandes ou muito pequenas não correspondem a determinado padrão.

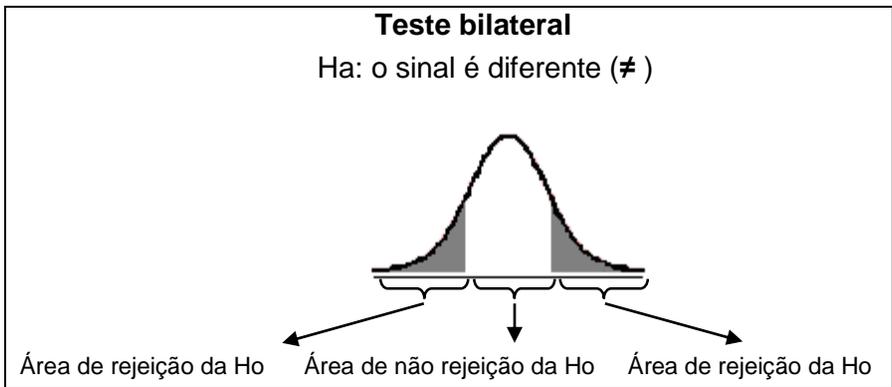


Figura 5.2 – Teste bilateral

A **região crítica** (Figura 5.3) de um teste de hipótese é a área de rejeição da hipótese nula. O **valor crítico** (Figura 5.3) é o valor que divide a área de não rejeição da área de rejeição da H_0 .

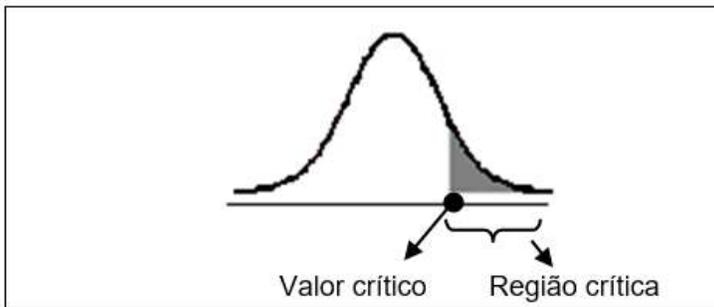


Figura 5.3 – Valor crítico e região crítica

A decisão de não rejeitar ou rejeitar a hipótese nula depende também do resultado obtido pela estatística de teste.

5.5 Estatística de teste

A estatística de teste é calculada com base em uma distribuição estatística conhecida, como por exemplo, a distribuição normal, a distribuição t de Student ou a distribuição

qui-quadrado. Se o valor da estatística de teste cair na área de rejeição, rejeitamos H_0 ; se cair na área de não rejeição, não rejeitamos H_0 .

Neste capítulo aplicaremos o teste de hipótese para dois parâmetros populacionais: a média (μ) e a proporção (p). No caso das proporções, a estatística de teste será calculada com base na distribuição normal. No caso da média, precisamos considerar se conhecemos ou não o desvio-padrão populacional (σ). Quando σ é conhecido, a estatística teste é calculada com base na Distribuição Normal; quando σ é desconhecido, usamos o desvio-padrão da amostra (s), e a distribuição usada é a Distribuição t de Student. Ambas fazem o uso de tabelas que estão em anexo.

Estatística de teste para a proporção:
$$Z_{\text{teste}} = \frac{\bar{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

Estatística de teste para a média:

$$Z_{\text{teste}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{ou} \quad t_{\text{teste}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Para tomarmos a decisão de não rejeitar ou rejeitar a H_0 , o valor da estatística de teste será comparado com o valor crítico. Este último é obtido diretamente da tabela da respectiva distribuição estatística com base no valor do nível de significância.

5.6 Teste de hipótese para a média

Antes de iniciar qualquer procedimento de teste de hipótese, devemos primeiro verificar se as suposições exigidas são satisfeitas para o conjunto de dados que está sendo usado. No caso da média, da mesma forma que nos intervalos de confiança, precisamos verificar se a amostra é aleatória simples, a população apresenta distribuição normal ou a amostra é maior do que 30 (Teorema Central do Limite). Além disso, se o desvio-padrão populacional, σ for conhecido, usaremos a distribuição normal; caso contrário, se o desvio-padrão populacional, σ , é desconhecido, usamos a distribuição t de Student.

5.7 Teste de hipótese para a média com σ conhecido

Quando o valor do desvio-padrão populacional σ é conhecido, a população original for normal ou a amostra for maior do que 30, com base no teorema central do limite podemos afirmar que a distribuição amostral da média segue uma distribuição normal e a estatística de teste é dada pela seguinte expressão:

$$Z_{\text{teste}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Exemplo 1: Um fabricante utiliza uma máquina para encher as embalagens de café. A máquina está funcionando

adequadamente se colocar 700g de café em pó em cada embalagem. A fim de verificar a calibragem da máquina a empresa coletou uma amostra aleatória de 40 embalagens. Esta amostra apresentou uma média de 698g. Sabe-se que o desvio-padrão da população é de 10g. Teste a hipótese de que o peso médio das embalagens na população é de 700g, a um nível de significância (α) de 5%.

Solução:

1ª Retire os dados do problema:

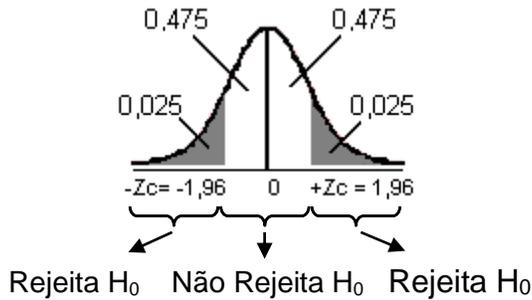
$$\mu = 700\text{g} \quad n = 40 \quad \sigma = 10\text{g} \quad \bar{x} = 698\text{g} \quad \alpha = 5\%$$

2ª Formule as hipóteses H_0 e H_a (lembre-se que a hipótese nula é aquela que contém o sinal de igual)

$$H_0 : \mu = 700$$

$$H_a : \mu \neq 700$$

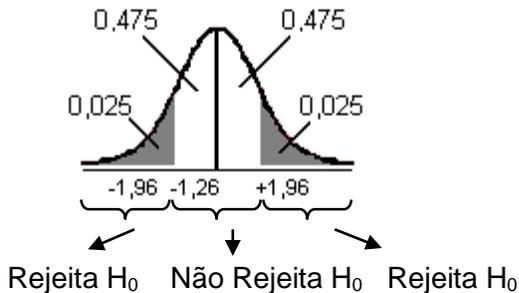
3ª Determine o valor crítico: Com base no sinal da hipótese alternativa (\neq) podemos verificar que o teste é bilateral, portanto o valor do nível de significância fica dividido por dois, obtendo assim duas áreas nas extremidades da curva de 0,025 cada uma. Com o auxílio da tabela da normal vamos descobrir quanto vale o valor crítico, que neste caso chamamos de $Z_{\text{crítico}}$, pois estamos usando a distribuição normal. Para uma área de 0,475 o valor de $Z_{\text{crítico}}$ é 1,96. Lembre-se, o $Z_{\text{crítico}}$ é o ponto que divide a área de rejeição da área de não rejeição da H_0 .



4ª Calcule a estatística de teste:

$$Z_{\text{teste}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{698 - 700}{10/\sqrt{40}} = -1,26$$

5ª Tome a decisão (não rejeitar ou rejeitar H_0): Podemos verificar que o $Z_{\text{teste}} = -1,26$ está dentro da área de não rejeição da H_0 , pois está entre $-1,96$ e $+1,96$.



6ª Interprete o resultado: Não se rejeita a hipótese nula, ao nível de significância de 5%, portanto podemos concluir que o peso médio em cada embalagem é de 700g, não havendo necessidade de parar a linha de produção para calibrar a máquina.

Exemplo 2: O rótulo em um recipiente de suco de laranja indica que o produto contém 1g de vitamina C. É selecionada uma amostra aleatória de 37 recipientes de suco, apresentando média amostral de 0,99g de vitamina C. Sabe-se que o desvio-padrão da quantidade de vitamina C nas embalagens é de $\sigma = 0,06g$.

a) Use os dados amostrais, a um nível de significância de 2%, para testar a afirmação de um gerente de produção de que a quantidade média de vitamina C no suco de laranja é menor do que 1g.

b) Com base no resultado do item a, determine qual o tipo de erro que poderemos estar cometendo.

Solução do item a):

1ª Retire os dados do problema:

$$\mu = 1g \quad n = 37 \quad \sigma = 0,06g \quad \bar{x} = 0,99g \quad \alpha = 2\%$$

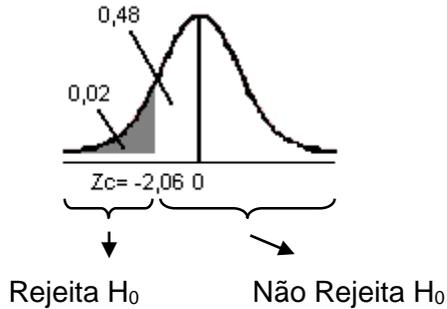
2ª Formule as hipóteses H_0 e H_a (lembre-se que a hipótese nula é aquela que contém o sinal de igual)

$$H_0 : \mu \geq 1$$

$$H_a : \mu < 1$$

3ª Determine o valor crítico: Com base no sinal da hipótese alternativa ($<$), podemos verificar que o teste é unilateral à esquerda, portanto a área da extremidade esquerda da curva é de 0,02. Na tabela da distribuição normal vamos procurar a área 0,48; verificamos que não há 0,48 exatos na

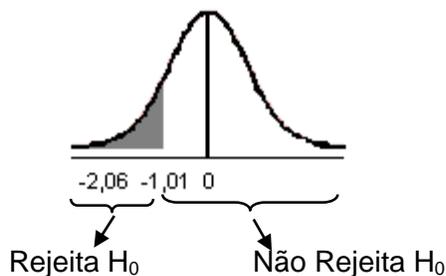
tabela, neste caso usamos o valor mais próximo, 0,4803, que equivale a um $Z_{\text{crítico}} = -2,06$ (negativo, pois o $Z_{\text{crítico}}$ está à esquerda de zero).



4ª Calcule a estatística de teste:

$$Z_{\text{teste}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{0,99 - 1}{0,06 / \sqrt{37}} = -1,01$$

5ª Tome a decisão (não rejeitar ou rejeitar H_0):
Podemos verificar que o $Z_{\text{teste}} = -1,01$ está dentro da área de não rejeição da H_0 , pois se encontra à direita de $Z_{\text{crítico}} = -2,06$.



6ª Interprete o resultado: Não se rejeita a H_0 ao nível de significância de 2%, ou seja, o suco de laranja contém em média

1g de vitamina C, o que indica que o rótulo do produto contém a informação correta.

Solução do item b): Podemos estar cometendo um erro do tipo II, que é não rejeitar a hipótese nula e ela ser falsa, ou seja, não rejeitar que o produto contém em média 1g de vitamina C quando na realidade contém uma quantidade menor.

Exemplo 3: O gerente de um *resort* afirma que os hóspedes gastam, em média, mais de R\$500,00 durante um fim de semana. Para testar a afirmação do gerente, foi selecionada aleatoriamente uma amostra de 30 hóspedes, obtendo-se uma média de R\$530,00. Sabe-se que o desvio-padrão populacional dos gastos dos hóspedes é de R\$45,00. Use $\alpha = 5\%$.

Solução:

1ª Retire os dados do problema:

$$\mu = \text{R}\$500,00 \quad n = 30 \quad \sigma = \text{R}\$45,00 \quad \bar{x} = \text{R}\$530,00 \quad \alpha = 5\%$$

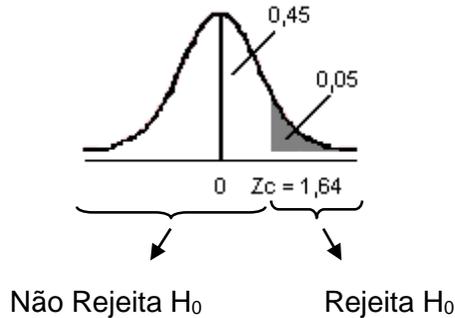
2ª Formule as hipóteses H_0 e H_a (lembre-se que a hipótese nula é aquela que contém o sinal de igual)

$$H_0 : \mu \leq 500$$

$$H_a : \mu > 500$$

3ª Determine o valor crítico: Com base no sinal da hipótese alternativa ($>$), podemos verificar que o teste é unilateral à direita, portanto a área da extremidade direita da curva é de 0,05. Na tabela da distribuição normal vamos procurar a área 0,45. Verificamos que não há 0,45 exatos na tabela; neste

caso há dois valores próximos de 0,45, que são 0,4495 e 0,4505. Como os valores apresentam a mesma diferença, sugerimos usar o primeiro valor. Portanto, $Z_{\text{crítico}} = 1,64$ (positivo, pois o $Z_{\text{crítico}}$ está à direita de zero).

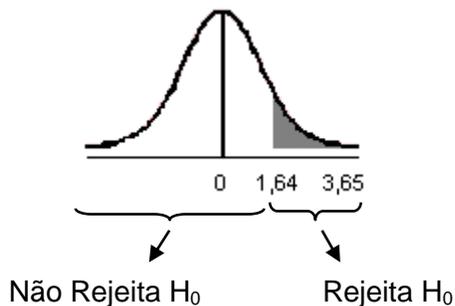


4ª Calcule a estatística de teste:

$$Z_{\text{teste}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{530 - 500}{45/\sqrt{30}} = 3,65$$

5ª Tome uma decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $Z_{\text{teste}} = 3,65$ está fora da área de não rejeição da H_0 , pois se encontra à direita de $Z_{\text{crítico}} = 1,64$.



6ª Interpretação do resultado: Rejeita-se a H_0 ao nível de significância de 5%, ou seja, o gasto médio dos hóspedes é maior do que R\$500,00, como afirma o gerente.

5.8 Teste de hipótese para média com σ desconhecido

Para amostras cujo desvio-padrão da população é desconhecido, a distribuição a ser usada no cálculo de testes de hipóteses é a distribuição t de Student, desde que possamos considerar que a população tenha distribuição normal ou aproximadamente normal. Nesse caso, a estatística de teste será dada por:

$$t_{\text{teste}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \text{com g.l.} = n - 1$$

Exemplo 1: Há doze anos o número médio de horas gasto assistindo à TV por família da população foi relatado com sendo de no máximo 6,8 horas por dia. Uma emissora de televisão alega que o tempo médio gasto assistindo à TV aumentou nos últimos anos. Uma amostra aleatória de 22 famílias forneceu um tempo médio gasto diante da televisão de 7,5 horas por dia, com desvio-padrão amostral de 1,8 horas por dia. Teste a alegação da emissora de televisão. Use um nível de significância de 1%.

Solução:

1ª Retire os dados do problema:

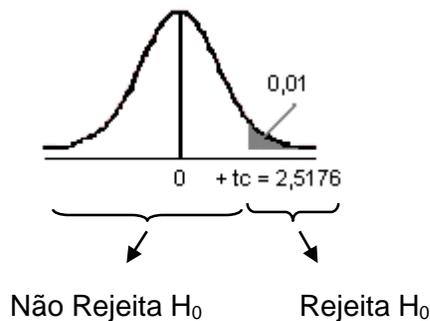
$$\mu = 6,8 \text{ h/dia} \quad n = 22 \quad \bar{x} = 7,5 \text{ h/dia} \quad s = 1,8 \text{ h/dia} \quad \alpha = 1\%$$

2ª Formule as hipóteses H_0 e H_a

$$H_0 : \mu \leq 6,8$$

$$H_a : \mu > 6,8$$

3ª Determine o valor crítico: Com base no sinal da hipótese alternativa ($>$) podemos verificar que o teste é unilateral à direita, portanto a área da extremidade à direita da curva é de 0,01. Na tabela da distribuição t de Student vamos procurar o valor de g.l. = 21 e área 0,01, portanto $t_{\text{crítico}} = 2,5176$ (positivo, pois o $t_{\text{crítico}}$ está à direita de zero).



4ª Calcule a estatística de teste:

$$t_{\text{teste}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{7,5 - 6,8}{1,8/\sqrt{22}} = 1,82$$

5ª Tome uma decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $t_{\text{teste}} = 1,82$ está dentro da área de não rejeição da H_0 , pois se encontra à esquerda de $t_{\text{crítico}} = 2,5176$, logo não rejeitamos a H_0 .

6ª Interprete o resultado: Não se rejeita a H_0 ao nível de significância de 1%, ou seja, o tempo médio gasto assistindo à TV não aumentou nos últimos anos.

Exemplo 2: Na avaliação de seus serviços, a nota máxima que um aeroporto pode alcançar é 10. São classificados como aeroportos classe A os que têm avaliação média maior que 7. A administração de um determinado aeroporto alega que seu aeroporto é de classe A. Com o objetivo de testar essa alegação, foi selecionada uma amostra aleatória de 13 viajantes. Foram obtidas as seguintes notas: 7 – 8 – 10 – 8 – 6 – 9 – 6 – 7 – 7 – 8 – 9 – 8 – 7. Considere que as notas apresentam distribuição normal. A um nível de significância de 5%, teste a alegação da administração do aeroporto.

Solução:

1ª Retire os dados do problema:

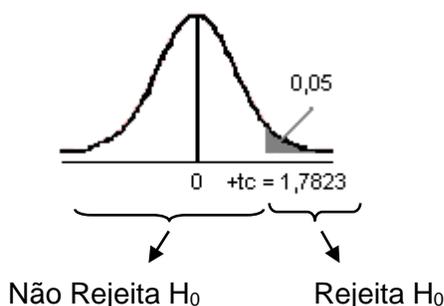
$$\mu = 7 \quad n = 13 \quad \bar{x} = 7,69 \quad s = 1,18 \quad \alpha = 5\%$$

2ª Formule as hipóteses H_0 e H_a

$$H_0 : \mu \leq 7,0$$

$$H_a : \mu > 7,0$$

3ª Determine o valor crítico: Com base no sinal da hipótese alternativa ($>$), podemos verificar que o teste é unilateral à direita, portanto a área da extremidade à direita da curva é de 0,05. Na tabela da distribuição t de Student, vamos procurar o valor de g.l. = 12 e área 0,05, portanto $t_{\text{crítico}} = 1,7823$ (positivo, pois o $t_{\text{crítico}}$ está à direita de zero).



4ª Calcule a estatística de teste:

Para calcular a estatística de teste, precisamos da média e do desvio-padrão da amostra.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{100}{13} = 7,69$$

$$s = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]} = \sqrt{\frac{1}{13-1} \left[786 - \frac{100^2}{13} \right]} = 1,18$$

$$t_{\text{teste}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{7,69 - 7}{1,18/\sqrt{13}} = 2,11$$

5ª Tome uma decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $t_{\text{teste}} = 2,11$ está fora da área de não rejeição da H_0 , pois se encontra à direita de $t_{\text{crítico}} = 1,7823$. Logo, rejeita-se H_0 .

6ª Interprete o resultado: Rejeita-se H_0 , ao nível de significância de 5%, ou seja, há evidências suficientes para apoiar a alegação da administração de que o aeroporto pertence à classe A.

5.9 Fluxograma

O fluxograma a seguir auxilia na escolha da distribuição a ser usada nos testes de hipóteses. A primeira pergunta é se o desvio-padrão populacional é conhecido; há duas saídas, sim ou não. Se a resposta for sim, podemos usar a Distribuição Normal; se for não, respondemos a outra pergunta. A distribuição é aproximadamente normal? Se a resposta for sim, usaremos o desvio-padrão amostral e usaremos a distribuição t de Student. Se a resposta for não, devemos aplicar testes não-paramétricos ou aumentar o tamanho da amostra para garantir que a distribuição seja aproximadamente normal como diz o teorema central.



5.10 Testes de hipóteses para proporção

A proporção da população, simbolizada por p , é parâmetro populacional, já estudado anteriormente. Para a proporção podemos ter três formas para um teste de hipótese, são eles:

$$\begin{array}{lll}
 H_0 : p \geq p_0 & H_0 : p \leq p_0 & H_0 : p = p_0 \\
 H_a : p < p_0 & H_a : p > p_0 & H_a : p \neq p_0
 \end{array}$$

O p_0 é um valor hipotético para a proporção da população. Podemos ter testes bilaterais ou unilaterais, que vão depender da montagem das hipóteses. Já vimos que a distribuição amostral de \bar{p} pode ser aproximada por uma distribuição de probabilidade normal desde que $np \geq 5$ e $nq \geq 5$.

O método usado para conduzir os testes é análogo ao usado para os testes de hipóteses da média populacional. A única diferença é que usamos a proporção da amostra \bar{p} e seu desvio-padrão $\sigma_{\bar{p}}$ no desenvolvimento da estatística do teste.

Estatística de teste para a proporção:

$$Z_{\text{teste}} = \frac{\bar{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}}, \text{ onde } \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Começamos formulando a hipótese nula e hipótese alternativa para o valor da proporção da população. Então, usando o valor da proporção da amostra \bar{p} e seu desvio-padrão $\sigma_{\bar{p}}$, calculamos um valor para a estatística do teste z. Comparar o valor da estatística do teste com o valor crítico nos possibilita determinar se a hipótese nula deve ser rejeitada ou não.

Exemplo 1: Num clube de pesca, 20% dos pescadores são mulheres. Em um esforço para aumentar a proporção de pescadoras, o clube utilizou uma promoção especial. Após um certo período foi selecionada uma amostra aleatória de 500

pescadores, dos quais 150 eram mulheres. Teste a hipótese de que a proporção de pescadoras aumentou com a nova promoção. Use nível de significância de 10%.

Solução: Já vimos que a distribuição amostral de \bar{p} pode ser aproximada por uma distribuição de probabilidade normal desde que $np \geq 5$ e $nq \geq 5$.

Neste exemplo, $np \geq 500 \cdot 0,2 = 100$ e $nq \geq 500 \cdot 0,8 = 400$

Assim, a aproximação da distribuição de probabilidade normal é apropriada.

1ª Retire os dados do problema:

$$p = 0,2 \quad n = 500 \quad \bar{p} = \frac{150}{500} = 0,3 \quad \alpha = 10\%$$

2ª Formule as hipóteses H_0 e H_a

$$H_0 : p \leq 0,20$$

$$H_a : p > 0,20$$

3ª Determine o valor crítico: Com base no sinal da hipótese alternativa ($>$) podemos verificar que o teste é unilateral à direita, portanto a área da extremidade à direita da curva é de 0,1. Na tabela da distribuição normal vamos procurar a área 0,40; verificamos que não há 0,40 exatos na tabela. Neste caso há dois valores próximos: 0,3997 e 0,4015; consideramos o que apresenta a menor diferença a 0,40, que é 0,3997, portanto, $Z_{\text{crítico}} = 1,28$ (positivo, pois o $Z_{\text{crítico}}$ está à direita de zero).

4ª Calcule a estatística de teste:

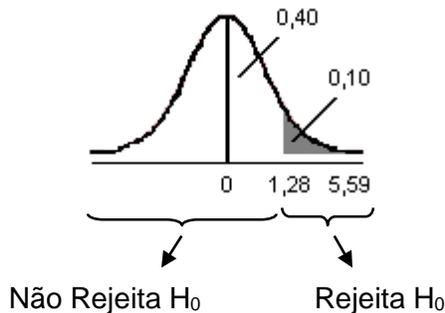
Antes de calcular a estatística de teste é necessário calcular o desvio-padrão das proporções $\sigma_{\bar{p}}$:

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0,20(1-0,20)}{500}} = 0,0179$$

$$Z_{\text{teste}} = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0,30 - 0,20}{0,0179} = 5,59$$

5ª Tome a decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $Z_{\text{teste}} = 5,59$ está fora da área de não rejeição da H_0 , pois se encontra à direita de $Z_{\text{crítico}} = 1,28$.



6ª Interprete o resultado: Rejeita-se a H_0 ao nível de significância de 10%, ou seja, a promoção especial aumentou significativamente a proporção de pescadoras no clube de pesca.

Exemplo 2: Os editores de um jornal afirmam que 25% de seus leitores são mulheres. Uma amostra aleatória de 150 leitores mostrou que 30 eram mulheres. Teste a afirmação dos editores do jornal ao nível de significância de 2%.

Solução: Já vimos que a distribuição amostral de \bar{p} pode ser aproximada por uma distribuição de probabilidade normal, desde que $np \geq 5$ e $nq \geq 5$. Neste exemplo, temos $np \geq 150 \cdot 0,25 = 37,5$ e $nq \geq 150 \cdot 0,75 = 112,5$.

Assim, a aproximação da distribuição de probabilidade normal é apropriada.

1ª Retire os dados do problema:

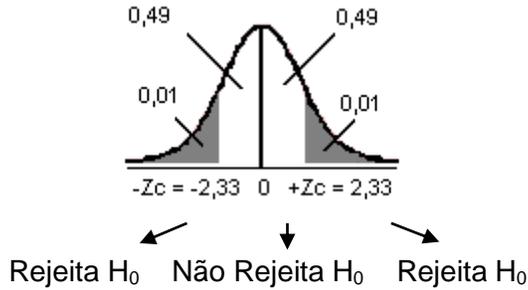
$$p = 0,25 \quad n = 150 \quad \bar{p} = \frac{30}{150} = 0,2 \quad \alpha = 2\%$$

2ª Formule as hipóteses H_0 e H_a

$$H_0 : p = 0,25$$

$$H_a : p \neq 0,25$$

3ª Determine o valor crítico: Com base no sinal da hipótese alternativa (\neq) podemos verificar que o teste é bilateral, portanto o valor do nível de significância fica dividido por dois, obtendo assim duas áreas nas extremidades da curva de 0,01 cada uma. Com o auxílio da tabela da normal vamos descobrir qual é o valor crítico. Para uma área de 0,49 verificamos que não há 0,49 exatos na tabela. Neste caso, há dois valores próximos: 0,4898 e 0,4901; consideramos o que apresenta a menor diferença, que será 0,4901, portanto, $Z_{\text{crítico}} = 2,33$.



4ª Calcule a estatística de teste:

Antes de calcular a estatística de teste é necessário calcular o desvio-padrão das proporções $\sigma_{\bar{p}}$:

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0,25(1-0,25)}{150}} = 0,0354$$

$$Z_{\text{teste}} = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0,2 - 0,25}{0,0354} = -1,41$$

5ª Tome a decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $Z_{\text{teste}} = -1,41$ está dentro da área de não rejeição da H_0 , pois se encontra entre $-2,33$ e $+2,33$.

6ª Interprete o resultado: Não se rejeita a H_0 ao nível de significância de 2%, ou seja, há evidências suficientes para apoiar a afirmação dos editores do jornal de que a proporção de leitoras é de 25%.

Exercícios resolvidos

1) De acordo com uma revista especializada em carros, o preço médio de carros usados no Brasil é de R\$15.680,00 ou mais. O gerente de uma distribuidora de carros usados de Rio Grande reviu uma amostra de 101 vendas recentes de carros usados na distribuidora, obtendo um preço médio de R\$14.880,00 e desvio-padrão R\$3.200,00. Com um nível de significância de 5%, teste a alegação da revista. Qual é a conclusão do teste?

Hipóteses: $H_0: \mu \geq 15.680$

$H_a: \mu < 15.680$

Tamanho da amostra (n) = 101

Média amostral (\bar{x}) = 14.880

Desvio-padrão amostral (s) = 3.200

Nível de significância (α) = 0,05

Estatística de teste = -2,5

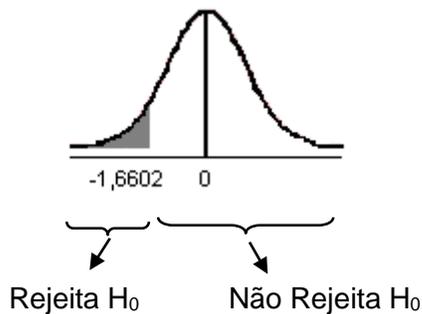
Solução:

As hipóteses foram enunciadas. Precisamos verificar qual distribuição devemos usar. Como o desvio-padrão da população não é conhecido, usaremos o desvio-padrão amostral (s) e a distribuição t de Student.

Como podemos observar, o teste é unilateral à esquerda, com um nível de significância de 0,05.

Devemos calcular os graus liberdade $g.l. = 101 - 1 = 100$. Na tabela t de Student retiraremos o $t_{\text{crítico}} = 1,6602$, que neste caso é negativo, pois está à esquerda de zero.

Podemos verificar que a estatística de teste igual a $-2,5$ está dentro da área de rejeição. Portanto, temos evidências suficientes para rejeitar a hipótese nula. Podemos concluir que a média de preços de vendas dos carros no Brasil é menor que R\$15.680.



2) Um novo programa de dieta afirma que os participantes perderão em média oito quilos durante a primeira semana do programa. Uma amostra aleatória de 41 participantes do programa mostrou uma perda de peso médio de sete quilos, com desvio-padrão amostral de 3,2 quilos, com um nível de significância de 0,05, Qual a conclusão sobre a afirmação feita pelo programa de dieta?

Tamanho da amostra (n) = 41

Média amostral (\bar{x}) = 7

Desvio-padrão amostral (s) = 3,2

Nível de significância (α) = 0,05

Estatística de teste = -1,98

Solução:

Neste exercício as hipóteses não foram enunciadas. O primeiro passo é descobrir quem são as hipóteses nula e alternativa.

$$H_0: \mu = 8$$

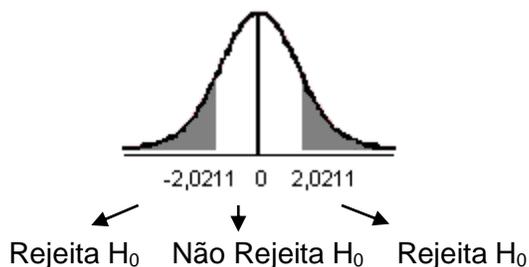
$$H_a: \mu \neq 8$$

Precisamos verificar qual distribuição devemos usar. Como o desvio-padrão da população não é conhecido, usaremos o desvio-padrão amostral (s) e a distribuição t de Student.

Como podemos observar, o teste é bilateral, com um nível de significância de 0,05. Temos duas áreas de 0,025.

Devemos calcular os graus de liberdade $g.l. = 41 - 1 = 40$. Na tabela t de Student retiraremos o $t_{crítico} = 2,0211$.

Podemos verificar que a estatística de teste igual a -1,98 está dentro da área de não rejeição. Portanto, temos evidências suficientes para não rejeitar a hipótese nula. Podemos concluir que em média os participantes da dieta perdem 8kg durante a primeira semana de dieta.



Exercícios complementares

1) Uma rede de televisão possui um canal que apresenta notícias, reportagens e anúncios direcionados para os indivíduos que esperam nas filas dos caixas de supermercados. Os programas de televisão foram concebidos com um ciclo de oito minutos, com base na teoria de que este é o tempo médio da população de compradores na fila de um caixa de supermercado. Uma amostra aleatoriamente selecionada de 101 compradores em uma grande rede de supermercados mostrou que o tempo médio de espera é de 7,5 minutos, com um desvio-padrão de 3,2 minutos. Teste, ao nível de significância de 0,05, se o tempo médio de espera no caixa do supermercado difere do tempo previsto.

Hipóteses: $H_0: \mu = 8$

$H_a: \mu \neq 8$

Tamanho da amostra (n) = 101

Média amostral (\bar{x}) = 7,5

Desvio-padrão amostral (s) = 3,2

Nível de significância (α) = 0,05

Estatística de teste = -1,57

2) Um novo programa de televisão precisa provar que tem mais que 25% de audiência de telespectadores depois das 15 primeiras semanas de exibição para ser julgado bem-sucedido. Considere uma amostra aleatória de 400 famílias; destas, 112 estavam vendo o novo programa. Com um nível de significância de 0,1, o programa pode ser considerado bem-sucedido com base na informação da amostra?

Hipóteses: $H_0: p \leq 0,25$

$H_a: p > 0,25$

Tamanho da amostra (n) = 400

Proporção amostral (\bar{p}) = $\frac{112}{400} = 0,28$

Nível de significância (α) = 0,1

Estatística de teste = 1,385

3) O gerente de um posto de gasolina está considerando um novo plano de bônus criado para aumentar as vendas (ex.: a cada mês será sorteada uma TV para os clientes do posto, ou a

cada 20 litros vendidos o cliente ganhará um refrigerante...). O volume médio atual de vendas é de até 10.000 litros mensais. O gerente aplica durante seis meses um plano de bônus e registra os seguintes volumes de vendas: 11.000 litros, 12.500 litros, 9.500 litros, 9.800 litros, 11.500 litros e 10.600 litros. Teste a hipótese de que o plano foi bem-sucedido ao um nível de significância de 0,1. Qual a sua conclusão? Suponha a população normalmente distribuída.

4) Uma indústria calçadista opera com o custo médio de fabricação de R\$300,00 ou mais por hora de produção e desvio-padrão populacional de R\$15,00. O diretor de fabricação propõe um novo método de fabricação que visa a diminuir o custo por hora de produção. Para testar o novo método é selecionado aleatoriamente um período de 41 horas de produção ao longo de um mês, verificando-se que o custo médio de produção é de R\$285,00. Com um nível de significância de 0,05, teste a hipótese de que o método proposto pelo diretor reduz o custo médio de produção. Qual sua conclusão?

5) Uma universidade federal afirma que mais de 80% de seus estudantes completam a graduação em 4 anos. Para testar a afirmação foram selecionados 400 graduados para o estudo; 300 graduados responderam que terminaram a graduação em 4 anos. Teste a afirmação da Universidade com um nível de significância de 0,1. Qual a sua conclusão?

6) Para uma população de varas de aço com desvio-padrão populacional de 21cm, desejamos testar a hipótese nula $H_0: \mu = 255$ cm contra a hipótese alternativa $H_a: \mu \neq 255$ cm, com base em uma amostra de 36 elementos e uma média amostral do comprimento das varas de 248cm, com um nível de significância de 0,1. Qual a conclusão e o erro que podemos estar cometendo?

Hipóteses: $H_0: \mu = 255$

$H_a: \mu \neq 255$

Tamanho da amostra (n) = 36

Média amostral (\bar{x}) = 248

Desvio-padrão populacional (σ) = 21

Nível de significância (α) = 0,1

Estatística de teste = -2,0

7) Um contador acredita que os problemas financeiros pelos quais a sua empresa está passando são resultado direto do atraso do pagamento das contas a receber. O contador afirma que pelo menos 70% das atuais contas a receber tem mais de dois meses de atraso. Uma amostra de 120 contas a receber mostrou que 78 têm mais de dois meses em atraso. Teste a afirmação do contador a um nível de significância de 0,05. Qual a sua conclusão?

Hipóteses: $H_0: p \geq 0,70$

$H_a: p < 0,70$

Tamanho da amostra (n) = 120

Proporção amostral (\bar{p}) = $\frac{78}{120} = 0,65$

Nível de significância (α) = 0,05

Estatística de teste = - 1,195

8) Para testar a alegação de que em média uma dona de casa com marido e dois filhos trabalhe 50 horas ou menos por semana em serviços domésticos (ex.: lavar a roupa, cozinhar, limpar a casa etc.), seleciona-se aleatoriamente uma amostra de oito donas de casa. As horas de trabalho durante uma semana para uma amostra de oito donas de casa são: 55, 60, 48, 45, 52, 54, 45 e 51. Teste a afirmação a um nível de significância de 0,05. Qual sua conclusão? Suponha população normalmente distribuída.

9) O fabricante de uma determinada marca de pneus afirma que seus pneus podem suportar uma quilometragem média de mais de 64.000km. Foi selecionada uma amostra de 60 pneus, com média de 61.000km, sabendo-se que a população tem desvio-padrão de 5.500km. Teste a afirmação do fabricante a um nível de significância de 0,05. Qual sua conclusão?

10) A empresa INSET vende um repelente de insetos que alega ser eficiente pelo prazo médio de 360 horas no mínimo. Uma análise de nove itens escolhidos aleatoriamente acusou uma média de 340 horas e desvio-padrão de 9 horas. Ao nível de significância de 0,01, teste a afirmação da empresa. Qual sua conclusão? Suponha distribuição da população aproximadamente normal.

Respostas

1. $t_{\text{teste}} = -1,57$, $t_{\text{crítico}} = \pm 1,9840$, não se rejeita a H_0
2. $Z_{\text{teste}} = 1,385$, $Z_{\text{crítico}} = 1,28$, rejeita H_0
3. $t_{\text{teste}} = 1,803$, $t_{\text{crítico}} = 1,4759$, média amostral = 10.816,67, $s = 1.108,9$, rejeita H_0
4. $Z_{\text{teste}} = -6,40$, $Z_{\text{crítico}} = -1,64$, rejeita H_0
5. $Z_{\text{teste}} = -2,5$, $Z_{\text{crítico}} = 1,28$, não se rejeita a H_0
6. $Z_{\text{teste}} = -2$, $Z_{\text{crítico}} = \pm 1,64$, rejeita H_0
7. $Z_{\text{teste}} = -1,195$, $Z_{\text{crítico}} = -1,64$, não se rejeita a H_0

8. Média amostral= 51,25, desvio-padrão amostral = 5,18, $t_{\text{teste}}= 0,683$, $t_{\text{crítico}}= 1,8946$, não se rejeita a H_0

9. $Z_{\text{teste}}= -4,23$, $Z_{\text{crítico}}= 1,64$, não se rejeita a H_0

10. $t_{\text{teste}}= - 6,66$, $t_{\text{crítico}}= -2,8965$, rejeita H_0

6 Intervalo de confiança para duas amostras

No capítulo 4, apresentamos como montar um intervalo de confiança para um único parâmetro populacional, como média (μ) e proporção (p). Neste capítulo, estenderemos estes conceitos para a comparação de dois parâmetros e consequentemente duas amostras.

É importante ressaltar a distinção entre o estudo realizado com **amostras independentes** e **amostras dependentes**¹. Duas amostras são independentes se os valores amostrais de uma população não estão relacionados ou emparelhados com os valores amostrais selecionados da outra população. Duas amostras são dependentes se cada elemento de uma amostra corresponder a um elemento de outra amostra.

Por exemplo, a fim de avaliar a influência de uma determinada campanha publicitária sobre a intenção de compra do consumidor, um pesquisador entrevista os consumidores antes e depois do lançamento da campanha. Se a amostra obtida depois que a campanha publicitária tiver sido lançada for composta por indivíduos diferentes daqueles entrevistados antes do lançamento da campanha, as amostras são independentes. Caso contrário, se a amostra obtida depois de lançada a campanha publicitária for composta pelos mesmos

¹ Vídeos sobre intervalos de confiança para duas amostras, gravados pelas autoras deste livro, você pode assistir no Youtube no Canal da Profa Suzi Samá.

indivíduos entrevistados antes do lançamento da campanha, então as amostras são consideradas dependentes, pois cada entrevistado estará sendo comparado consigo mesmo em outro instante de tempo.

6.1 Intervalo de confiança para a diferença de duas médias populacionais com σ_1 e σ_2 conhecidos

A inferência estatística para diferença de duas médias populacionais μ_1 e μ_2 de duas populações com desvios-padrões conhecidos segue as seguintes suposições:

- Ambas as amostras são aleatórias.
- As duas amostras são independentes.
- Ambas as amostras provêm de populações com distribuição normal ou ambas as amostras são maiores que 30 (Teorema Central do Limite).

Na estimação de um intervalo de confiança para duas médias populacionais o **erro máximo da estimativa** é obtido usando-se a distribuição normal e multiplicando o valor crítico ($Z_{\text{crítico}}$) pelo desvio-padrão das diferenças de médias:

$$e = z_c \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

O **intervalo de confiança** para a diferença de duas médias populacionais com σ_1 e σ_2 conhecidos é dado por:

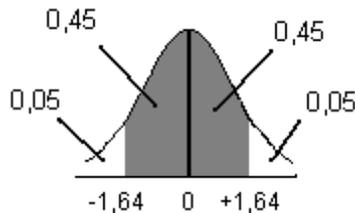
$$(\bar{x}_1 - \bar{x}_2) \pm z_c \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$P((\bar{x}_1 - \bar{x}_2) - e \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + e) = 1 - \alpha$$

Exemplo de intervalo de confiança: Um fabricante de automóveis utiliza dois tipos diferentes de pneus nos veículos que fabrica. Uma amostra aleatória de 50 pneus do tipo A apresentou duração média de 38.700km. Uma amostra aleatória de 40 pneus do tipo B apresentou durabilidade média de 41.800km. Por pesquisas anteriores, sabe-se que o pneu tipo A apresenta desvio-padrão populacional de 4.023km, e os pneus tipo B, desvio-padrão populacional de 4.827km. Construa o intervalo de 90% de confiança para diferença de duração média entre os dois tipos de pneus.

Solução:

1º Com base no nível de confiança (1- α), determine o valor de $Z_{\text{crítico}}$.



2º Determine o intervalo de confiança:

$$(\bar{x}_1 - \bar{x}_2) \pm z_c \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(38.700 - 41.800) \pm 1,64 \cdot \sqrt{\frac{4.023^2}{50} + \frac{4.827^2}{40}}$$

$$(38.700 - 41.800) \pm 1.561,18 \quad \Leftrightarrow \quad -3.100 \pm 1.561,18$$

$$-4.611,18 \leq \mu_A - \mu_B \leq -1.538,82$$

3º Interprete o intervalo de confiança: Com 90% de confiança podemos afirmar que o intervalo de -4.661,18 a -1.538,82 contém a diferença de duração média populacional. Como os dois limites do intervalo de confiança são negativos, podemos concluir que os pneus tipo A têm duração média menor que os pneus tipo B ($\mu_A < \mu_B$).

6.2 Intervalo de confiança para a diferença de duas médias populacionais com σ_1 e σ_2 desconhecidos

Nesta seção, a inferência estatística para a diferença de duas médias populacionais μ_1 e μ_2 segue as mesmas suposições do item 6.1, no entanto apresentam-se dois casos distintos:

1º Caso: Os desvios-padrões populacionais, σ_1 e σ_2 , são desconhecidos e não se faz qualquer suposição sobre a

igualdade das variâncias populacionais, $\sigma_1^2 \neq \sigma_2^2$. Neste caso, usamos as variâncias amostrais s_1^2 e s_2^2 no cálculo do desvio-padrão da diferença de médias, usando assim a distribuição t de Student.

O **intervalo de confiança** para a diferença de duas médias populacionais supondo $\sigma_1^2 \neq \sigma_2^2$ é dado por:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\text{crítico}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Como partimos da suposição de que as variâncias populacionais são diferentes $\sigma_1^2 \neq \sigma_2^2$, calcula-se o número de graus de liberdade pela seguinte expressão:

$$\text{g.l.} = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}}$$

$$\text{Sendo: } w_1 = \frac{s_1^2}{n_1} \quad \text{e} \quad w_2 = \frac{s_2^2}{n_2}$$

Exemplo de intervalo de confiança: Uma pesquisa analisou a eficácia de dois tipos de treinamento, com a finalidade de diminuir o tempo médio da realização de uma determinada tarefa. Foram selecionadas duas amostras aleatórias de populações com distribuição normal. Considere $\sigma_1 \neq \sigma_2$. Os dados da pesquisa estão no quadro a seguir. Determine um intervalo de 95% de confiança para a diferença no tempo médio

da realização da atividade. O que podemos concluir?

Treinamento 1	$n_1 = 15$	$\bar{x}_1 = 24,2\text{min}$	$s_1 = 3,16\text{ min}$
Treinamento 2	$n_2 = 10$	$\bar{x}_2 = 23,9\text{min}$	$s_2 = 4,47\text{ min}$

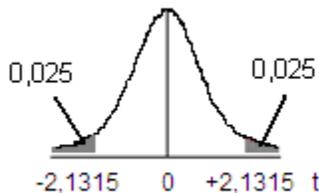
Solução:

1º Para determinar o intervalo de confiança, primeiro calcule o número de graus de liberdade para encontrar o valor de $t_{\text{crítico}}$.

$$w_1 = \frac{s_1^2}{n_1} = \frac{3,16^2}{15} = 0,666 \qquad w_2 = \frac{s_2^2}{n_2} = \frac{4,47^2}{10} = 1,98$$

$$\text{g.l.} = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}} = \frac{(0,666 + 1,998)^2}{\frac{0,666^2}{15 - 1} + \frac{1,998^2}{10 - 1}} = \frac{7,0969}{0,0317 + 0,4436} = 14,93 \cong 15$$

2º Com base no número de graus de liberdade (g.l.) e no nível de confiança $(1 - \alpha)$, determine o valor de $t_{\text{crítico}}$.



3º Determine o intervalo de confiança:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\text{crítico}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(24,2 - 23,9) \pm 2,1315 \cdot \sqrt{\frac{3,16^2}{15} + \frac{4,47^2}{10}}$$

$$0,3 \pm 2,1315.1,6321$$

$$0,3 \pm 3,479$$

$$-3,179 \leq \mu_1 - \mu_2 \leq 3,779$$

4º Interprete o intervalo de confiança: Com 95% de confiança podemos afirmar que o intervalo de -3,179 a 3,779 contém a diferença de tempo médio populacional na realização da tarefa. Como o limite inferior é negativo e o limite superior do intervalo de confiança é positivo, nada podemos concluir, pois o intervalo de confiança contém o zero, logo o tempo médio entre os dois treinamentos pode ser igual ($\mu_A = \mu_B$), ou o treinamento A ter maior tempo médio ($\mu_A > \mu_B$), ou o treinamento A ter menor tempo médio ($\mu_A < \mu_B$).

2º caso: Os desvios-padrões populacionais, σ_1 e σ_2 , são desconhecidos, mas é razoável supor que as variâncias populacionais são iguais, $\sigma_1^2 = \sigma_2^2$. Neste caso, calculamos a média aritmética ponderada das variâncias s_1^2 e s_2^2 para obter uma estimativa da variância populacional comum, denotada por \hat{s}^2 , usando assim a distribuição t de Student, a qual tem $n_1 + n_2 - 2$ graus de liberdade.

O **intervalo de confiança** supondo $\sigma_1^2 = \sigma_2^2$ é dado por:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\text{crítico}} \cdot \sqrt{\frac{\hat{s}^2}{n_1} + \frac{\hat{s}^2}{n_2}}$$

com $g.l. = n_1 + n_2 - 2$ onde $\hat{s}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Exemplo de intervalo de confiança: Em uma avaliação de estatística, foi selecionada uma amostra aleatória de 12 estudantes da turma A, resultando numa nota média igual a 7,9 com desvio-padrão 0,6. Na turma B foram selecionados aleatoriamente 15 estudantes, que tiraram nota média 6,7 com desvio-padrão 0,8. Construa um intervalo de confiança de 99% para a diferença de médias, supondo a variância das duas populações iguais $\sigma_1^2 = \sigma_2^2$. As notas apresentam distribuição normal.

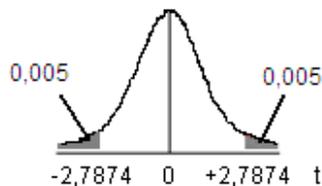
Solução:

1º Retire os dados do exemplo:

$$\bar{x}_1 = 7,9 \quad n_1 = 12 \quad s_1 = 0,6 \quad 1 - \alpha = 0,99$$

$$\bar{x}_2 = 6,7 \quad n_2 = 15 \quad s_2 = 0,8 \quad g.l. = 12 + 15 - 2 = 25$$

2º Com base no número de graus de liberdade (g.l.) e no nível de confiança (1- α), determine o valor de $t_{crítico}$.



3º Determine o intervalo de confiança: Para determinar o intervalo de confiança precisamos calcular \hat{s}^2

$$\hat{s}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1)0,6^2 + (15 - 1)0,8^2}{12 + 15 - 2} = \frac{3,96 + 8,96}{25} = 0,52$$

$$(7,9 - 6,7) \pm 2,7874 \cdot \sqrt{\frac{0,52}{12} + \frac{0,52}{15}}$$

$$1,2 \pm 2,7874 \cdot 0,2793 \quad \rightarrow \quad 1,2 \pm 0,78$$

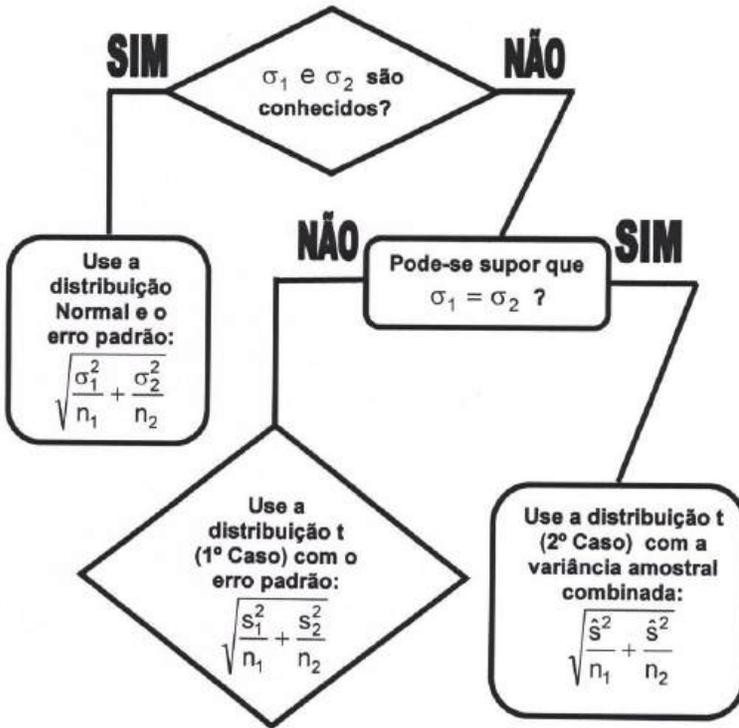
$$0,42 \leq \mu_1 - \mu_2 \leq 1,98$$

Interprete o intervalo de confiança: Com 99% de confiança podemos afirmar que o intervalo de 0,42 a 1,98 contém a diferença das notas médias populacionais. Como os limites inferior e superior do intervalo de confiança são positivos, podemos concluir que a média das notas dos alunos da turma A é maior do que a média das notas dos alunos da turma B ($\mu_A > \mu_B$).

6.3 Fluxograma

O fluxograma a seguir auxilia na escolha da distribuição a ser usada nos intervalos de confiança para diferença de médias de amostras independentes. A primeira pergunta é se o desvio-padrão populacional é conhecido, há duas saídas, sim ou não. Se a resposta for sim, podemos usar a distribuição normal; se for não, respondemos outra pergunta: podemos supor que $\sigma_1 = \sigma_2$? Se a resposta for sim, usaremos a distribuição t com os passos do 2º

caso. Se for não, devemos usar a distribuição t de Student com os passos do 1º caso.



6.4 Intervalo de confiança para amostras dependentes

O teste t de Student para amostras dependentes analisa dois grupos de observações baseados na mesma amostra de objetos ou indivíduos, a fim de verificar se o processo ao qual os indivíduos foram submetidos produziu alguma alteração. Voltemos ao exemplo do início desta unidade, em que um pesquisador está interessado em avaliar a influência de uma

determinada campanha publicitária sobre a intenção de compra do consumidor. Se ele entrevistar os mesmos consumidores antes e depois de ser lançada a campanha, as amostras serão dependentes. Neste caso, em vez de analisar cada grupo separadamente, observamos somente a diferença (d_i) entre as duas medidas em cada indivíduo, $d_i = x_i - y_i$.

No estudo de amostras dependentes, obtemos um quadro semelhante a este:

Antes (x_i)	Depois (y_i)
x_1	y_1
x_2	y_2
\vdots	\vdots
x_n	y_n

O **intervalo de confiança** para amostras dependentes é dado por:

$$\bar{d} \pm t_{\text{crítico}} \cdot \frac{s_d}{\sqrt{n}}$$

ou

$$P(\bar{d} - e \leq \mu_d \leq \bar{d} + e) = 1 - \alpha$$

sendo:

- Média das diferenças $\bar{d} = \frac{\sum d_i}{n}$
- Desvio-padrão das diferenças entre os pares:

- $s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}}$
- Graus de liberdade: g.l. = n – 1
- Número de pares observados: n
- Erro máximo da estimativa: e

Exemplo para intervalo de confiança: O centro de estudos de violência no trânsito deseja determinar se o novo programa de direção defensiva altera o comportamento dos motoristas. Foram selecionados dez motoristas, os quais responderam um questionário a fim de verificar o grau de agressividade no trânsito antes e depois de eles terem participado do programa. Os escores obtidos estão no quadro a seguir. Monte um intervalo de 90% de confiança para os escores de agressividade no trânsito. O que você pode concluir?

<i>Motorista</i>	<i>Antes</i>	<i>Depois</i>
1	4	1
2	3	3
3	10	8
4	5	1
5	8	7
6	9	8
7	5	1
8	7	5
9	1	2
10	7	6

Solução:

1º Calcule a diferença (d_i)

<i>Motorista</i>	<i>Antes (x_i)</i>	<i>Depois (y_i)</i>	$d_i = x_i - y_i$
1	4	1	3
2	3	3	0
3	10	8	2
4	5	1	4
5	8	7	1
6	9	8	1
7	5	1	4
8	7	5	2
9	1	2	-1
10	7	6	1

2º Calcule a média das diferenças:

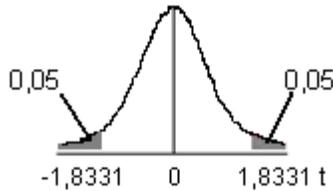
$$\bar{d} = \frac{\sum d_i}{n} = \frac{3+0+2+4+1+1+4+2+(-1)+1}{10} = \frac{17}{10} = 1,7$$

3º Calcule o desvio-padrão das diferenças entre os pares :

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{53 - 10 \cdot 1,7^2}{10-1}} = \sqrt{\frac{53 - 28,9}{9}} = 1,6364$$

4º Número de graus de liberdade: Com base no número de graus de liberdade (g.l.) e no nível de confiança $(1 - \alpha)$ determine o valor de $t_{\text{crítico}}$.

$$\text{g.l.} = n - 1 = 10 - 1 = 9$$



5º Determine o intervalo de confiança:

$$\bar{d} \pm t_{\text{crítico}} \cdot \frac{s_d}{\sqrt{n}} \rightarrow 1,7 \pm 1,8311 \cdot \frac{1,6364}{\sqrt{10}} \rightarrow 1,7 \pm 1,8311 \cdot 0,5175$$

$$1,7 \pm 0,95$$

$$0,75 \leq \mu_d \leq 2,65$$

6º Interprete o intervalo de confiança: Com 90% de confiança podemos afirmar que o intervalo de 0,75 a 2,65 contém a diferença dos escores médios populacionais. Como os limites inferior e superior do intervalo de confiança são ambos positivos, podemos concluir que a média dos escores de agressividade antes da participação do motorista no programa é maior do que a média dos escores de agressividade após a participação no programa ($\mu_A > \mu_D$). Portanto, o programa surtiu o efeito desejado.

6.5 Intervalo de confiança para a diferença de duas proporções populacionais

Na inferência estatística para diferença de duas proporções populacionais p_1 e p_2 , com base em amostras independentes é realizada através da distribuição normal.

O **intervalo de confiança** para diferença de duas proporções populacionais é dado por:

$$(\bar{p}_1 - \bar{p}_2) \pm z_c \cdot \sqrt{\frac{\bar{p}_1 \cdot (1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2 \cdot (1 - \bar{p}_2)}{n_2}}$$

$$P((\bar{p}_1 - \bar{p}_2) - e \leq p_M - p_F \leq (\bar{p}_1 - \bar{p}_2) + e) = 1 - \alpha$$

Exemplo: Uma pesquisa entre estudantes universitários, constatou que de 140 estudantes do sexo masculino, aleatoriamente selecionados, 26 usam a Internet apenas para os estudos; e 15 estudantes do sexo feminino entre as 120, aleatoriamente selecionadas, também utilizam a Internet apenas nos estudos. Determinar, ao nível de 99% de confiança, o intervalo para a diferença entre as proporções de estudantes de ambos os sexos, que utilizam a Internet apenas para estudo.

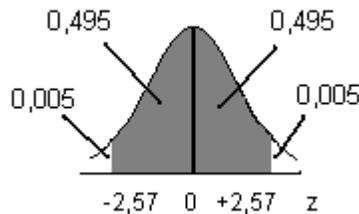
Solução:

1º Retire os dados do exemplo:

$$x_M = 26 \quad n_M = 140 \quad \bar{p}_M = \frac{x}{n} = \frac{26}{140} = 0,186$$

$$x_F = 15 \quad n_F = 120 \quad \bar{p}_F = \frac{x}{n} = \frac{15}{120} = 0,125$$

2º Com base no nível de confiança (1- α), determine o valor de $Z_{\text{crítico}}$.



3º Determine o intervalo de confiança:

$$(\bar{p}_M - \bar{p}_F) \pm z_c \cdot \sqrt{\frac{\bar{p}_M \cdot (1 - \bar{p}_M)}{n_M} + \frac{\bar{p}_F \cdot (1 - \bar{p}_F)}{n_F}}$$

$$(0,186 - 0,125) \pm 2,57 \cdot \sqrt{\frac{0,186 \cdot (1 - 0,186)}{140} + \frac{0,125 \cdot (1 - 0,125)}{120}}$$

$$0,061 \pm 2,57 \cdot \sqrt{\frac{0,1514}{140} + \frac{0,1094}{120}}$$

$$0,061 \pm 2,57 \cdot 0,0446 \rightarrow 0,061 \pm 0,1146$$

$$-0,054 \leq p_M - p_F \leq 0,176$$

$$\text{ou } -5,4\% \leq p_M - p_F \leq 17,6\%$$

4º Interprete o intervalo de confiança: Com 99% de confiança podemos afirmar que o intervalo de - 0,054 a 0,176

contém a diferença entre as proporções de estudantes, de ambos os sexos, que utilizam a Internet apenas para estudo. Como o intervalo contém o zero, nada podemos concluir em relação a proporção de estudantes do sexo masculino e feminino que utiliza a Internet apenas para estudo.

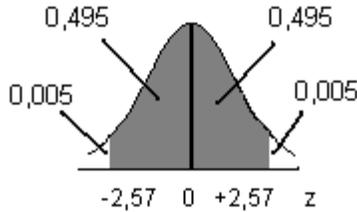
Exercícios resolvidos

1) Um pesquisador está estudando a idade média em que jovens de ambos os sexos começam a beber. Para tal, selecionou uma amostra aleatória de 26 jovens do sexo masculino, obtendo idade média de 12,3 anos. Uma amostra aleatória de 22 jovens do sexo feminino forneceu idade média de 14,2 anos. Com base em pesquisas anteriores, sabe-se que a idade em que jovens do sexo masculino e feminino começam a beber apresenta uma distribuição normal com desvio-padrão populacional igual a 0,8 e 1,3 ano, respectivamente. Monte um intervalo de confiança de 99% para a diferença de médias. O intervalo contém o valor ZERO? Interprete este resultado.

$$\sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}} = 0,3184$$

Solução:

1º Com base no nível de confiança (1- α), determine o valor de $Z_{\text{crítico}}$.



2º Determine o intervalo de confiança:

$$(\bar{x}_M - \bar{x}_F) \pm Z_c \cdot \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}$$

$$(-1,9) \pm 2,57 \cdot 0,3184 \rightarrow (-1,9) \pm 0,818$$

$$-2,718 \leq \mu_M - \mu_F \leq -1,082$$

3º Interprete o intervalo de confiança: Com 99% de confiança podemos afirmar que o intervalo de -2,718 a -1,082 contém a diferença de idade média populacional. O intervalo não contém o valor zero. Como os dois limites do intervalo de confiança são negativos, podemos concluir que a idade média populacional em que os jovens do sexo masculino começam a beber é menor do que a idade média populacional em que as jovens começam a beber ($\mu_M < \mu_F$).

2) Uma empresa realizou um estudo para avaliar o tempo médio de adaptação de funcionários do sexo masculino e feminino. Foram selecionados aleatoriamente 46 homens, obtendo-se tempo médio de adaptação de 2,8 anos com desvio-padrão de 0,6 ano, e 46 mulheres, também aleatoriamente, obtendo-se tempo médio de adaptação de 2,3 anos com desvio-

padrão de 0,7 ano. Por estudos anteriores, sabe-se que a população apresenta distribuição normal. Monte um intervalo de confiança de 95% para a diferença de médias.

$$\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} = 0,1356$$

Solução:

Retire os dados do exemplo:

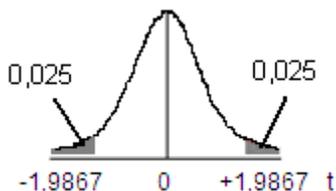
$$\bar{x}_1 = 2,8 \quad n_1 = 46 \quad s_1 = 0,6$$

$$\bar{x}_2 = 2,3 \quad n_2 = 46 \quad s_2 = 0,7$$

$$1 - \alpha = 0,95$$

Graus de liberdade: g.l. = 46 + 46 = 90

Com base no número de graus de liberdade (g.l.) e no nível de confiança (1 - α), determine o valor de $t_{\text{crítico}}$.



Determine o intervalo de confiança

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\text{crítico}} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}$$

$$(2,8 - 2,3) \pm 1,9867 \cdot 0,1356 \quad \rightarrow \quad 0,5 \pm 1,9867 \cdot 0,2694$$

$$0,231 \leq \mu_1 - \mu_2 \leq 0,769$$

Interprete o intervalo de confiança: Com 95% de confiança, podemos afirmar que o intervalo de 0,231 a 0,769 contém a diferença entre as médias populacionais. Como os limites inferior e superior do intervalo de confiança são positivos, podemos concluir que o tempo médio de adaptação dos homens é maior do que o tempo médio de adaptação das mulheres ($\mu_A > \mu_B$).

Exercícios complementares

1) A vida média das lâmpadas produzidas pelas fábricas A e B apresenta distribuição normal com desvio-padrão populacional de 200 e 150 horas, respectivamente. Uma amostra de 40 lâmpadas produzidas pela fábrica A apresentou vida média de 4.000 horas. Uma amostra de 60 lâmpadas produzidas pela fábrica B apresentou vida média de 3.600 horas. Monte um intervalo de confiança de 90% para diferença de médias. O que você pode concluir?

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 37,081 \quad 1 - \alpha = 90$$

2) Um fabricante produz dois tipos de pneus. A durabilidade dos dois tipos de pneus apresenta distribuição normal. Os pneus do tipo A apresentam desvio-padrão populacional de 4.020 km e os do tipo B, desvio-padrão populacional de 4.900 km. Uma amostra aleatória de 70 pneus do tipo A apresentou vida média de 38.700 km e uma amostra aleatória de 55 pneus do tipo B, apresentou vida média de 41.900 km. Construa um intervalo de confiança de 99% para a diferença na vida média dos dois tipos de pneus. O intervalo de confiança estimado inclui ou não incluiu o zero? Interprete este resultado.

3) Uma amostra de 52 trabalhadores de uma fábrica demora, em média, 12 minutos para completar uma tarefa, com desvio-padrão de 2 minutos. Uma amostra aleatória de 50 trabalhadores de outra fábrica demora, em média, 16 minutos para completar a mesma tarefa, com desvio-padrão de 3 minutos. As duas populações apresentam distribuição normal. Construa um intervalo de confiança de 95% para a diferença entre as duas médias populacionais. Com base no intervalo de confiança estimado, o que você pode concluir sobre o tempo médio na realização da tarefa entre as duas fábricas? Suponha $\sigma_1^2 = \sigma_2^2$.

$$\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} = 0,503$$

4) Uma escola está testando dois métodos de ensino a fim de verificar qual método é mais eficaz. Uma amostra de 32 estudantes, aleatoriamente selecionados, foi submetida ao método I e apresentou nota média de 74, com desvio-padrão igual a 5,8. Outra amostra com 40 estudantes foi submetida ao método II e apresentou nota média de 77, com desvio-padrão igual a 7,6. Construa um intervalo de confiança de 95% para a nota média entre os dois tipos de método. Suponha $\sigma_1^2 = \sigma_2^2$.

5) Para avaliar o nível de tensão ocasionada por exames escolares, oito alunos foram escolhidos e sua pulsação medida antes e depois do exame. Monte um intervalo de confiança de 90% para o nível de tensão médio.

Instante da Medição	Estudante							
	1	2	3	4	5	6	7	8
Antes	92	83	87	96	79	83	87	79
Depois	86	87	84	92	78	85	79	74

$$\bar{d} = 2,625 \quad s_d = 4,0686$$

6) Realizou-se um estudo para investigar os efeitos de uma nova dieta para emagrecer. Para tal, os participantes do estudo foram pesados antes e depois da dieta, conforme quadro a seguir. Monte um intervalo de confiança de 95%. O que você

pode concluir quanto ao efeito da dieta sobre o peso?

Antes	99	57	62	69	74	77	59	92	70	85
Depois	94	57	64	73	66	74	58	88	70	82

7) Suponha que você faz parte do comitê editorial de uma grande revista. Para enfrentar a crise financeira pela qual a revista vem passando, o comitê resolve fazer mudanças no enfoque da revista. Para tal, realiza uma pesquisa a fim de verificar o perfil de seus leitores. Dentre 200 leitores do sexo masculino, aleatoriamente selecionados, 70 têm menos de 25 anos. Das 200 leitoras, aleatoriamente selecionadas, 50 têm menos de 25 anos. Monte um intervalo de 99% de confiança para a diferença entre a proporção de leitores do sexo masculino e leitoras do sexo feminino com menos de 25 anos nesta população. O que você pode concluir?

$$\sqrt{\frac{\bar{p}_1 \cdot (1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2 \cdot (1 - \bar{p}_2)}{n_2}} = 0,0456$$

8) Defina amostras dependentes e independentes. Dê um exemplo de cada.

9) Duas empresas realizam, há alguns anos, campanhas de marketing para aumentar a quantia gasta por seus clientes com cartões de crédito. A fim de comparar as duas campanhas, foi selecionada uma amostra de 80 clientes da empresa A sendo

constatada uma média de gastos de R\$1.245,00. Uma amostra aleatória de 120 clientes da empresa B, apresentou média de gastos igual a R\$1.638,00. Com base em pesquisas anteriores, sabe-se que o desvio-padrão populacional da empresa A é de R\$253,00 e da empresa B de R\$352,00. Monte um intervalo de confiança de 96% para a média de gastos entre as duas campanhas.

10) Foi realizado um experimento para estudar o efeito do álcool sobre as pessoas. Foram selecionadas aleatoriamente duas amostras, com 36 integrantes em cada uma, sendo que em apenas uma das amostras houve consumo de álcool. Os participantes do experimento foram submetidos a um teste de habilidades motoras e visuais, sendo o número de erros cometidos anotados. Na amostra em que não houve consumo de álcool, a média de erros foi de 1,93 com desvio-padrão de 0,88. Na amostra em que houve consumo de álcool, a média de erros foi de 4,72 e desvio-padrão de 2,23. Suponha as variâncias populacionais iguais.

a) Construa um intervalo de 95% de confiança para a diferença entre as duas médias populacionais.

b) O intervalo contém o ZERO? O que você pode concluir?

11) Uma empresa oferece cursos de língua inglesa a seus funcionários, com o objetivo de melhorar a compreensão deste idioma. O quadro a seguir fornece os escores obtidos por dez de

seus funcionários antes e após a realização do curso. Quanto maior o escore, maior a capacidade de compreensão do idioma.

Funcionário	1	2	3	4	5	6	7	8	9	10
Pré-teste	28	25	31	31	30	15	20	20	34	26
Pós-teste	30	29	32	34	25	18	16	27	31	28

Construa um intervalo de confiança de 90%. O que você pode concluir?

12) De 600 moradores, aleatoriamente selecionados, de uma cidade em que a base da economia é a indústria, 450 são favoráveis a um projeto de lei, e de uma amostra de 240 moradores, aleatoriamente selecionados, de uma cidade cuja principal atividade é a agricultura, 150 são favoráveis. Monte um intervalo de confiança de 99% para a diferença entre as duas proporções populacionais.

13) Uma agência de empregados domésticos deseja verificar o grau de satisfação de seus clientes. Os clientes estão classificados em duas categorias, de acordo com a renda familiar – A (maior renda) e B (menor renda). Foi selecionada uma amostra aleatória de cada categoria e solicitado que atribuíssem uma nota de 0 (totalmente insatisfeito) a 10 (totalmente satisfeito). Os resultados estão no quadro a seguir:

Categoria A	$\bar{x}_A = 6,5$	$s_A = 0,9$	$n_A = 54$
Categoria B	$\bar{x}_B = 6,9$	$s_B = 1,2$	$n_B = 48$

Monte um intervalo de confiança de 99% para a diferença entre a satisfação média dos clientes das duas populações. Suponha variâncias populacionais iguais.

14) Uma amostra aleatória de 12 formandos da turma A de um curso de datilografia obteve a média de 80,09 palavras por minuto, com desvio-padrão de 2,6551. Uma outra amostra aleatória com 12 alunos da turma B do mesmo curso de datilografia obteve em média 73,55 palavras por minuto, com desvio-padrão de 8,908. Construa um intervalo de confiança de 95% para a diferença entre a média populacional de palavras digitadas por minuto entre as duas turmas. Considere população normal com variâncias iguais.

15) Em virtude dos protestos quanto à degradação do meio ambiente, algumas empresas têm adotado materiais totalmente recicláveis em seus produtos. Numa pesquisa, foi perguntado a 270 consumidoras (M) e 290 consumidores (H), aleatoriamente selecionados, se comprariam um produto que não tivesse a etiqueta informativa de que os materiais utilizados na embalagem são totalmente recicláveis. Dos entrevistados, 90 consumidoras e 125 consumidores responderam que não comprariam o produto. Determine um intervalo de 90% de confiança para a diferença de proporção populacional de consumidores e consumidoras que não comprariam o produto.

16) Uma associação beneficente adotou dois tipos de pedido de auxílio. No pedido feito por telefone (T), dos 1.600 realizados, obteve 310 adesões. Outro tipo de pedido de auxílio (O), mais dispendioso, teve 415 adesões dentre os 2500 contatos realizados. Construa um intervalo 95% de confiança para a diferença de proporção populacional entre os contatos realizados.

Respostas

1. $339,19 \leq \mu_A - \mu_B \leq 460,81$. Conclusão pessoal.

2. $-5.299,56 \leq \mu_A - \mu_B \leq -1.100,44$. Não inclui o zero.
Interpretação ver o exercício resolvido 1.

3. $-5 \text{ min} \leq \mu_1 - \mu_2 \leq -3 \text{ min}$. Conclusão pessoal.

4. $-6,2 \leq \mu_1 - \mu_2 \leq 0,25$

5. $-0,1 \leq \mu_d \leq 5,35$

6. $-0,7192 \leq \mu_d \leq 4,0622$

7. $-1,72\% \leq p_1 - p_2 \leq 21,72\%$. Conclusão individual.

8. Resposta pessoal

9. $-480,54 \leq \mu_A - \mu_B \leq -305,46$

10. $-3,5868 \leq \mu_1 - \mu_2 \leq -1,9932$

O intervalo de confiança não contém o zero. Como os dois limites do intervalo de confiança são negativos, podemos concluir que a média de erros entre os que consumiram álcool é maior do que a média de erros entre os que não consumiram álcool ($\mu_2 > \mu_1$).

11. $-3,2142 \leq \mu_d \leq 1,2142$

12. $0,0327 \leq p_A - p_B \leq 0,2173$

13. $-0,9478 \leq \mu_A - \mu_B \leq 0,1478$

14. $0,9751 \leq \mu_A - \mu_B \leq 12,1049$

15. $-0,1649 \leq p_M - p_H \leq 0,0311$

16. $0,0037 \leq p_T - p_O \leq 0,0523$

7 Testes de Hipóteses para duas amostras

No capítulo 5, apresentamos como realizar um teste de hipótese para um único parâmetro populacional, como média (μ) e proporção (p). Neste capítulo, estenderemos estes conceitos para a comparação de dois parâmetros e conseqüentemente duas amostras.

Da mesma forma que na estimação de um intervalo de confiança para duas médias populacionais, nos testes de hipóteses¹ também é necessário considerar a distinção entre o estudo realizado com **amostras independentes** e **amostras dependentes**. Duas amostras são independentes se os valores amostrais de uma população não estão relacionados ou emparelhados com os valores amostrais selecionados da outra população. Duas amostras são dependentes se cada elemento de uma amostra corresponder a um elemento de outra amostra.

7.1 Teste de hipótese para a diferença de duas médias populacionais com σ_1 e σ_2 conhecidos

No teste de hipóteses para diferença das médias μ_1 e μ_2 de duas populações distintas com desvios-padrões conhecidos segue as seguintes suposições:

¹ Vídeos sobre Testes de Hipóteses para duas amostras, gravados pelas autoras deste livro, você pode assistir no Youtube no Canal da Profa Suzi Samá.

- Ambas as amostras são aleatórias.
- As duas amostras são independentes.
- Ambas as amostras provêm de populações com distribuição normal ou ambas as amostras são maiores que 30 (Teorema Central do Limite).

A **estatística de teste** para o **teste de hipótese** da diferença de duas médias populacionais com σ_1 e σ_2 conhecidos é dada por:

$$Z_{\text{teste}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Exemplo de teste de hipótese: Uma transportadora de valores tem duas possibilidades de trajeto para realizar entregas em um determinado banco. O gerente de logística desconfia não haver diferença significativa entre o tempo médio de cada trajeto. Foram selecionadas, aleatoriamente, 45 entregas realizadas no primeiro trajeto, sendo anotado o tempo do percurso de cada uma delas, resultando em uma média amostral de 57 minutos. No segundo trajeto foram selecionadas aleatoriamente 30 entregas, em que foi obtido tempo médio de 54 minutos. O desvio-padrão populacional do primeiro trajeto é de $\sigma_1 = 8$ min e do segundo trajeto, $\sigma_2 = 6$ min. Teste a hipótese de que não existe diferença significativa entre o tempo médio dos dois trajetos. Use $\alpha = 1\%$.

Solução:

1º Retire os dados do problema:

$$\bar{x}_1 = 57 \text{ min}; \quad n_1 = 45; \quad \sigma_1 = 8 \text{ min}$$

$$\bar{x}_2 = 54 \text{ min}; \quad n_2 = 30; \quad \sigma_2 = 6 \text{ min}$$

2º Formule as hipóteses H_0 e H_a :

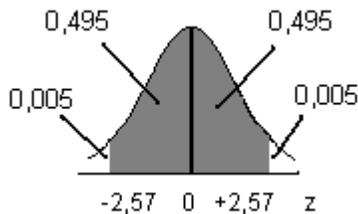
$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

3º Especifique o nível de significância:

$$\alpha = 1\%$$

4º Determine o valor crítico, monte a curva e determine a região de rejeição e não rejeição da hipótese nula:



5º Calcule a estatística de teste:

$$Z_{\text{teste}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{57 - 54}{\sqrt{\frac{8^2}{45} + \frac{6^2}{30}}} = \frac{3}{1,6193} = 1,85$$

6º Tome a decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $Z_{\text{teste}} = 1,85$ está dentro da área de não rejeição da H_0 , pois se encontra entre $-2,57$ e $+2,57$. Logo, não se rejeita a H_0 .

7º Interprete o resultado: Não se rejeita a hipótese nula, ao nível de significância de 1%, portanto podemos concluir que não existe diferença significativa no tempo médio de entrega entre os dois trajetos.

7.2 Teste de hipótese para a diferença de duas médias populacionais com σ_1 e σ_2 desconhecidos

Nesta seção, a inferência estatística para a diferença de duas médias populacionais μ_1 e μ_2 segue as mesmas suposições do item 7.1, no entanto apresentam-se dois casos distintos:

1º Caso: Os desvios-padrões populacionais, σ_1 e σ_2 , são desconhecidos e não se faz qualquer suposição sobre a igualdade das variâncias populacionais, $\sigma_1^2 \neq \sigma_2^2$. Neste caso, usamos as variâncias amostrais s_1^2 e s_2^2 no cálculo do desvio-padrão da diferença de médias, usando assim a distribuição t de Student.

A **estatística de teste** para o **teste de hipótese** da diferença de duas médias populacionais supondo $\sigma_1^2 \neq \sigma_2^2$ é dada por:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Com graus de liberdade (g.l.):

$$g.l. = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1} + \frac{w_2^2}{n_2 - 1}}$$

$$\text{Sendo: } w_1 = \frac{s_1^2}{n_1} \quad \text{e} \quad w_2 = \frac{s_2^2}{n_2}$$

Exemplo de teste de hipótese: Com os dados do exemplo anterior, teste a hipótese de que o tempo médio para a realização da tarefa é igual para os dois tipos de treinamento. Use nível de significância de 5%.

Solução:

1º Retire os dados do exemplo:

Treinamento 1	$n_1 = 15$	$\bar{x}_1 = 24,2\text{min}$	$s_1 = 3,16\text{ min}$
Treinamento 2	$n_2 = 10$	$\bar{x}_2 = 23,9\text{min}$	$s_2 = 4,47\text{ min}$

2º Formule as hipóteses nula e alternativa:

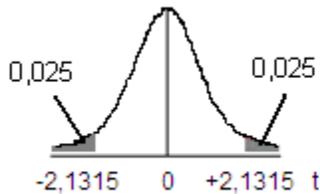
$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

3º Especifique o nível de significância:

$$\alpha = 5\%$$

4º Determine o valor crítico, monte a curva e determine a região de rejeição e não rejeição da hipótese nula: Pelo exercício anterior conhecemos o valor aproximado do número de graus de liberdade $g.l. \cong 15$ e o valor de $t_{\text{crítico}} = 2,1315$.



4º Calcule a estatística de teste:

$$t_{\text{teste}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{24,2 - 23,9}{1,6321} = 0,1838$$

6º Tome a decisão (não rejeitar ou rejeitar a H_0):

Podemos verificar que o $t_{\text{teste}} = 0,1838$ está dentro da área de **não rejeição** da H_0 , pois se encontra entre $-2,1315$ e $+2,1315$. Logo, não se rejeita a H_0 .

7º Interprete o resultado: Não se rejeita a hipótese nula, ao nível de significância de 5%, portanto podemos concluir que não existe diferença significativa entre os dois tipos de treinamento no que diz respeito ao tempo médio para a realização da tarefa.

2º caso: Os desvios-padrões populacionais, σ_1 e σ_2 , são desconhecidos, mas é razoável supor que as variâncias populacionais são iguais, $\sigma_1^2 = \sigma_2^2$. Neste caso, calculamos a média aritmética ponderada das variâncias s_1^2 e s_2^2 para obter uma estimativa da variância populacional comum, denotada por \hat{s}^2 , usando assim a distribuição t de Student, a qual tem $n_1 + n_2 - 2$ graus de liberdade.

A **estatística de teste** para o **teste de hipótese** da diferença de duas médias populacionais supondo $\sigma_1^2 = \sigma_2^2$ é dada por:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}^2}{n_1} + \frac{\hat{s}^2}{n_2}}}$$

Com graus de liberdade dado por:

$$\text{g.l.} = n_1 + n_2 - 2$$

$$\text{onde } \hat{s}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Exemplo de teste de hipótese: Com os dados do exemplo anterior, teste a hipótese de que a turma A tem média maior do que a turma B. Use nível de significância de 1%.

Solução:

1º Retire os dados do exemplo:

$$\bar{x}_1 = 7,9 \quad n_1 = 12 \quad s_1 = 0,6$$

$$\bar{x}_2 = 6,7 \quad n_2 = 15 \quad s_2 = 0,8$$

2º Formule as hipóteses nula e alternativa:

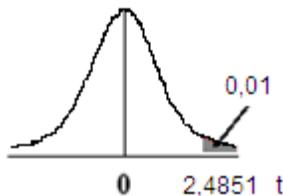
$$H_0 : \mu_A \leq \mu_B$$

$$H_a : \mu_A > \mu_B$$

3º Especifique o nível de significância:

$$\alpha = 1\%$$

4º Determine o valor crítico, monte a curva e determine a região de rejeição e não rejeição da hipótese nula: Pelos cálculos do exercício anterior, sabemos o número de graus de liberdade $g.l. = 25$, o valor crítico $t_{\text{crítico}} = 2,4851$.



5º Calcule a estatística de teste:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}^2}{n_1} + \frac{\hat{s}^2}{n_2}}} = \frac{7,9 - 6,7}{0,2793} = 4,2964$$

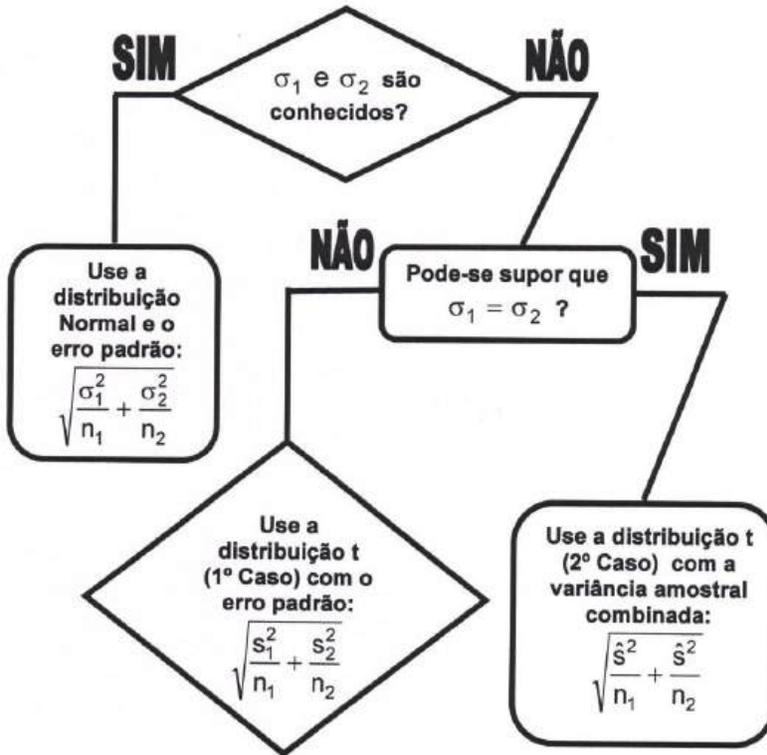
6º Tome a decisão (não rejeitar ou rejeitar a H_0):

Podemos verificar que o $t_{\text{teste}} = 4,2964$ está fora da área de **não rejeição** da H_0 , pois se encontra à direita de $+2,4851$. Logo, rejeita-se a H_0 .

7º Interprete o resultado: Rejeita-se a hipótese nula, ao nível de significância de 1%, ou seja, podemos concluir que a nota média da turma A é significativamente maior do que a nota média da turma B.

7.3 Fluxograma

O fluxograma a seguir auxilia na escolha da distribuição a ser usada nos testes de hipóteses para diferença de médias de amostras independentes. A primeira pergunta é se o desvio-padrão populacional é conhecido, há duas saídas, sim ou não. Se a resposta for sim, podemos usar a distribuição normal; se for não, respondemos outra pergunta: podemos supor que $\sigma_1 = \sigma_2$? Se a resposta for sim, usaremos a distribuição t com os passos do 2º caso. Se for não, devemos usar a distribuição t de Student com os passos do 1º caso.



7.4 Teste de hipótese para amostras dependentes

O teste t de Student para amostras dependentes analisa dois grupos de observações baseados na mesma amostra de objetos ou indivíduos, a fim de verificar se o processo ao qual os indivíduos foram submetidos produziu alguma alteração. Voltemos ao exemplo do início desta unidade, em que um pesquisador está interessado em avaliar a influência de uma determinada campanha publicitária sobre a intenção de compra

do consumidor. Se ele entrevistar os mesmos consumidores antes e depois de ser lançada a campanha, as amostras serão dependentes. Neste caso, em vez de analisar cada grupo separadamente, observamos somente a diferença (d_i) entre as duas medidas em cada indivíduo, $d_i = x_i - y_i$.

No estudo de amostras dependentes, obtemos um quadro semelhante a este:

Antes (x_i)	Depois (y_i)
x_1	y_1
x_2	y_2
\vdots	\vdots
x_n	y_n

O **intervalo de confiança** para amostras dependentes é dado por:

$$P(\bar{d} - e \leq \mu_d \leq \bar{d} + e) = 1 - \alpha \quad \rightarrow \quad \bar{d} \pm t_{\text{crítico}} \cdot \frac{s_d}{\sqrt{n}}$$

sendo:

- Média das diferenças $\bar{d} = \frac{\sum d_i}{n}$
- Desvio-padrão das diferenças entre os pares:
- $s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n - 1}}$
- Graus de liberdade: gl. = $n - 1$
- Número de pares observados: n

A **estatística de teste** para o teste de hipótese de amostras dependentes é:

$$t_{\text{teste}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

Exemplo: Sete trabalhadores foram selecionados a fim de determinar a eficiência de certo treinamento para a realização de uma tarefa. O quadro a seguir apresenta os resultados observados quanto ao tempo de execução da tarefa, em minutos, antes de os trabalhadores serem submetidos ao treinamento e depois do treinamento. Para que o treinamento seja considerado eficaz, é necessário que o tempo de realização da tarefa depois do treinamento seja significativamente menor do que o tempo de realização da tarefa antes do treinamento. Ao nível de 5% de significância, podemos concluir que o tempo da realização da tarefa é menor depois do treinamento?

Trabalhadores	Treinamento	
	Antes	Depois
1	13	10
2	08	08
3	10	07
4	11	06
5	07	09
6	14	09
7	12	11

Solução:

1º Formule as hipóteses nula e alternativa: Se o tempo médio é menor depois do treinamento, significa que o tempo médio antes é maior. Na montagem da hipótese alternativa fica ($\mu_A > \mu_D$), pois no quadro a diferença (d_i) foi calculada considerando $d_i = x_i - y_i$.

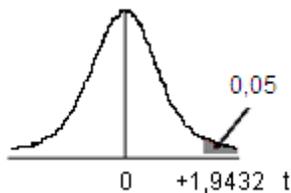
$$H_0 : \mu_A \leq \mu_D$$

$$H_a : \mu_A > \mu_D$$

2º Especifique o nível de significância:

$$\alpha = 5\%$$

3º Determine o valor crítico, monte a curva e determine a região de rejeição e não rejeição da hipótese nula:



4º Calcule a diferença (d_i)

Trabalhadores	Treinamento		
	Antes (x_i)	Depois (y_i)	$d_i = x_i$
1	13	10	3
2	08	08	0
3	10	07	3
4	11	06	5
5	07	09	-2
6	14	09	5
7	12	11	1

5º Calcule a média das diferenças e o desvio-padrão das diferenças:

$$\bar{d} = \frac{\sum d_i}{n} = \frac{3+0+3+5+(-2)+5+1}{7} = \frac{15}{7} = 2,1428$$

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{73 - 7 \cdot 2,1428^2}{7-1}} = \sqrt{\frac{40,8588}{6}} = 2,6096$$

6º Calcule a estatística de teste:

$$t_{\text{teste}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{2,1428}{\frac{2,6096}{\sqrt{7}}} = 2,1725$$

7º Tome a decisão (não rejeitar ou rejeitar a H_0):

Podemos verificar que o $t_{\text{teste}} = 2,1725$ está dentro da área de rejeição da H_0 , pois se encontra à direita de $+1,9432$. Logo, rejeita-se a H_0 .

8º Interprete o resultado: Rejeita-se a hipótese nula, ao nível de significância de 5%, ou seja, o tempo de execução da tarefa é menor depois do treinamento, portanto o treinamento foi eficaz.

7.5 Teste de hipótese para a diferença de duas proporções populacionais

Os valores das proporções populacionais, p_1 e p_2 , são desconhecidos. Devemos estimá-las através das proporções amostrais, \bar{p}_1 e \bar{p}_2 . Mas, pela hipótese nula, temos que $p_1 = p_2$, então suas estimativas também devem ser iguais. Portanto, usamos como estimativas de p_1 e p_2 a proporção amostral combinada \hat{p} :

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

onde x_1 é o número de sucessos da amostra 1 e x_2 é o número de sucessos da amostra 2.

A **estatística de teste** para a diferença de proporções é dada por:

$$Z_{\text{teste}} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Exemplo 1: Em um estudo datado de 2003, compreendendo 2.870 motoristas, 1.210 afirmaram terem

ingerido bebida alcoólica antes de dirigir. Depois de sancionada a “Lei Seca”, foi realizado outro estudo entre 2.200 motoristas, dos quais 725 afirmaram ter ingerido bebida alcoólica antes de dirigir. Usando um nível de significância de 10%, é possível não rejeitar a alegação das autoridades de que a proporção de motoristas que ingerem bebidas alcoólicas antes de dirigir diminuiu significativamente após entrar em vigor a “Lei Seca”?

Solução:

1º Retire os dados do exemplo:

$$x_1 = 1210 \quad n_1 = 2870 \quad \bar{p}_1 = \frac{x}{n} = \frac{1210}{2870} = 0,42$$

$$x_2 = 725 \quad n_2 = 2200 \quad \bar{p}_2 = \frac{x}{n} = \frac{725}{2200} = 0,33$$

2º Formule as hipóteses H_0 e H_a

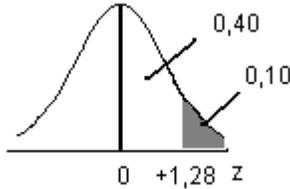
$$H_0 : p_1 \leq p_2$$

$$H_a : p_1 > p_2$$

3º Especifique o nível de significância:

$$\alpha = 10\%$$

4º Determine o valor crítico, monte a curva e determine a região de rejeição e não rejeição da hipótese nula:



5º Determine a proporção amostral combinada:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{1210 + 725}{2870 + 2200} = 0,38$$

6º Calcule a estatística de teste:

$$Z_{\text{teste}} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,42 - 0,33}{\sqrt{0,38(1-0,38)\left(\frac{1}{2870} + \frac{1}{2200}\right)}} = \frac{0,09}{0,0138} = 6,52$$

7º Tome a decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $Z_{\text{teste}} = 6,52$ está fora da área de **não rejeição** da H_0 , pois se encontra à direita de $+1,28$. Logo, rejeita-se a H_0 .

8º Interprete o resultado: Rejeita-se a hipótese nula, ao nível de significância de 10%, ou seja, há evidência suficiente, nos dados amostrais, para apoiar a alegação das autoridades de que a proporção de motoristas que ingerem bebidas alcoólicas antes de dirigir diminui significativamente após sancionada a “Lei Seca”.

Exemplo 2: Há alguns anos, um levantamento entre 80 estudantes revelou que 7,5% pretendiam prestar vestibular para

Engenharia. Em um levantamento mais recente compreendendo 75 estudantes, 12% deles pretendiam prestar vestibular para Engenharia. Sendo $\alpha = 1\%$, você pode concluir que a proporção de estudantes que pretendem estudar Engenharia aumentou?

Solução:

1º Retire os dados do exemplo:

$$x_1 = ? \quad n_1 = 80 \quad \bar{p}_1 = 0,075$$

$$x_2 = ? \quad n_2 = 75 \quad \bar{p}_2 = 0,12$$

2º Determine os valores de x_1 e x_2 :

$$0,075 = \frac{x_1}{80} \quad \therefore x_1 = 6$$

$$0,12 = \frac{x_2}{75} \quad \therefore x_2 = 9$$

3º Formule as hipóteses H_0 e H_a

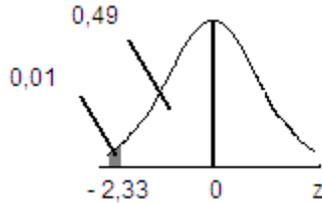
$$H_0 : p_1 \geq p_2$$

$$H_a : p_1 < p_2$$

4º Especifique o nível de significância:

$$\alpha = 1\%$$

5º Determine o valor crítico, monte a curva e determine a região de rejeição e não rejeição da hipótese nula:



6º Determine a proporção amostral combinada:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{6 + 9}{80 + 75} = 0,097$$

7º Calcule a estatística de teste:

$$Z_{\text{teste}} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,075 - 0,12}{\sqrt{0,097(1-0,097)\left(\frac{1}{75} + \frac{1}{80}\right)}} = -\frac{0,045}{0,0476} = -0,95$$

8º Tome a decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $Z_{\text{teste}} = -0,95$ está dentro da área de não rejeição da H_0 , pois encontra-se à direita de $-2,33$. Logo, não se rejeita a H_0 .

9º Interprete o resultado: Não se rejeita a hipótese nula, ao nível de significância de 1%, ou seja, a proporção de estudantes que pretendem prestar vestibular para Engenharia não aumentou com o tempo.

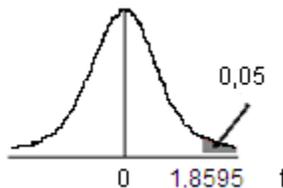
Exercícios resolvidos

1) Num estudo da eficácia de exercícios físicos na redução de peso, dez pessoas de uma amostra aleatória seguiram o programa de exercícios físicos durante um período de três meses. O peso de cada uma das dez pessoas, antes e depois do programa, está no quadro a seguir. A um nível de significância de 5%, podemos afirmar que o programa é eficaz na redução do peso?

Indivíduo	1	2	3	4	5	6	7	8	9	10
Antes	80	86	98	92	84	78	92	91	93	78
Depois	78	77	88	93	77	78	85	80	89	81

Solução:

$$\begin{aligned} H_0 &: \mu_A \leq \mu_D \\ H_a &: \mu_A > \mu_D \\ \alpha &= 5\% \end{aligned}$$



$$\text{g.l.} = n - 1 = 10 - 1 = 9$$

Calcule as diferenças:

d_i	2	9	10	-1	7	0	7	11	4	-3
-------	---	---	----	----	---	---	---	----	---	----

Calcule a média das diferenças e o desvio-padrão das diferenças:

$$\bar{d} = \frac{\sum d_i}{n} = \frac{46}{10} = 4,6$$

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{430 - 10 \cdot 4,6^2}{10-1}} = 4,9261$$

Calcule a estatística de teste:

$$t_{\text{teste}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{4,6}{\frac{4,9261}{\sqrt{10}}} = \frac{4,6}{1,5578} = 2,9529$$

Tome a decisão: Podemos verificar que o $t_{\text{teste}} = 2,9529$ está dentro da área de rejeição da H_0 , pois se encontra à direita de $+1,8595$. Logo, rejeita-se a H_0 .

Interprete o resultado: rejeita-se a hipótese nula, ao nível de significância de 5%, ou seja, o programa é eficaz na redução de peso.

2) Uma pesquisa revelou que, entre 860 mulheres entrevistadas, 74 não usam cinto de segurança ao dirigir. Na mesma pesquisa, entre 940 motoristas do sexo masculino 95 não usam o cinto de segurança. Teste a afirmativa de que não existe diferença significativa no uso do cinto de segurança entre ambos os sexos. Use $\alpha = 5\%$.

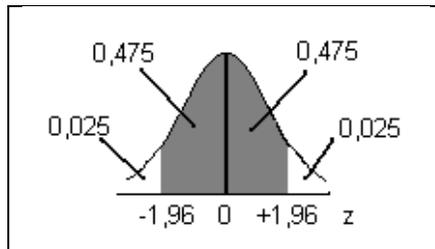
Estadística de teste = -1,03

Solução: Retirando os dados do exemplo:

$$x_1 = 74 \quad n_1 = 860 \quad \bar{p}_1 = 0,086$$

$$x_2 = 95 \quad n_2 = 940 \quad \bar{p}_2 = 0,101$$

$$\begin{aligned} H_0 : p_1 &= p_2 \\ H_a : p_1 &\neq p_2 \\ \alpha &= 5\% \end{aligned}$$



Tome a decisão: Podemos verificar que o $Z_{\text{teste}} = -1,03$ está dentro da área de **não rejeição** da H_0 , pois se encontra entre -1,96 e +1,96. Logo, não se rejeita a H_0 .

Interprete o resultado: Não se rejeita a hipótese nula, ao nível de significância de 5%, ou seja, não existe diferença significativa entre a proporção populacional do uso do cinto de segurança entre ambos os sexos.

Exercícios complementares

1) A vida média das lâmpadas produzidas pelas fábricas A e B apresenta distribuição normal com desvio-padrão populacional de 200 horas e 150 horas, respectivamente. Uma amostra de 40 lâmpadas produzidas pela fábrica A apresentou

vida média de 4.000 horas. Uma amostra de 60 lâmpadas produzidas pela fábrica B apresentou vida média de 3.600 horas. Teste a hipótese de que a vida média das lâmpadas produzidas pela fábrica A é maior do que a vida média das lâmpadas produzidas pela fábrica B ao nível de significância de 1%. O que você pode concluir?

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 37,081 \quad 1 - \alpha = 90$$

2) Um fabricante produz dois tipos de pneus. A durabilidade dos dois tipos de pneus apresenta distribuição normal. Os pneus do tipo A apresentam desvio-padrão populacional de 4.020 km e os do tipo B, desvio-padrão populacional de 4.900 km. Uma amostra aleatória de 70 pneus do tipo A apresentou vida média de 38.700 km e uma amostra aleatória de 55 pneus do tipo B, apresentou vida média de 41.900 km. Ao nível de significância de 5%, teste a hipótese de que não existe diferença significativa na vida média dos dois tipos de pneus.

$$\text{Estatística de teste} = -3,92$$

- Determine a H_0 e H_a
- Monte a curva e determine a área de não rejeição e rejeição da H_0
- Identifique a estatística de teste (Z ou t)
- Tome uma decisão

3) Uma amostra de 52 trabalhadores da fábrica A demora, em média, 12 minutos para completar uma tarefa, com desvio-padrão de 2 minutos. Uma amostra aleatória de 50 trabalhadores da fábrica B demora, em média, 16 minutos para completar a mesma tarefa, com desvio-padrão de 3 minutos. As duas populações apresentam distribuição normal. Teste a hipótese de que o tempo médio dos trabalhadores da fábrica A é menor do que o tempo médio dos trabalhadores da fábrica B ao nível de significância de 10%. O que você pode concluir? Suponha $\sigma_1^2 = \sigma_2^2$.

$$\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} = 0,503$$

4) Uma escola está testando dois métodos de ensino a fim de verificar qual método é mais eficaz. Uma amostra de 32 estudantes, aleatoriamente selecionados, foi submetida ao método I e apresentou nota média de 74, com desvio-padrão igual a 5,8. Outra amostra com 40 estudantes foi submetida ao método II e apresentou nota média de 77, com desvio-padrão igual a 7,6. Teste a hipótese de que o método II é mais eficaz do que o método I, ao nível de significância de 1%. Suponha $\sigma_1^2 = \sigma_2^2$.

$$\text{Estatística de teste} = -1,84$$

5) Para avaliar o nível de tensão ocasionada por exames escolares, oito alunos foram escolhidos e sua pulsação medida antes e depois do exame. Teste a hipótese de que não existe diferença significativa entre a pulsação dos estudantes antes e depois do exame. Use nível de significância de 5%.

Instante da medição	Estudante							
	1	2	3	4	5	6	7	8
Antes	92	83	87	96	79	83	87	79
Depois	86	87	84	92	78	85	79	74

$$\bar{d} = 2,625 \quad s_d = 4,0686$$

6) Realizou-se um estudo para investigar os efeitos de uma nova dieta para emagrecer. Ao nível de significância de 5%, teste a afirmação de que o peso médio das pessoas depois a dieta é menor do que o peso médio antes da dieta. O que se pode concluir quanto ao efeito da dieta sobre o peso?

Antes	99	57	62	69	74	77	59	92	70	85
Depois	94	57	64	73	66	74	58	88	70	82

$$\text{Estatística de teste} = 1,616$$

7) Suponha que você faz parte do comitê editorial de uma grande revista. Para enfrentar a crise financeira pela qual a revista vem passando, o comitê resolve fazer mudanças no enfoque da revista. Para tal, realiza uma pesquisa a fim de verificar o perfil de seus leitores. Dentre 200 leitores do sexo masculino, aleatoriamente selecionados, 70 têm menos de 25

anos. Das 200 leitoras, aleatoriamente selecionadas, 50 têm menos de 25 anos. Teste a hipótese de que a proporção de leitores do sexo masculino é maior do que a proporção de leitores do sexo feminino, ao nível de 10%.

$$\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} = 0,0456$$

8) Duas empresas realizam, há alguns anos, campanhas de marketing para aumentar a quantia gasta por seus clientes com cartões de crédito. A fim de comparar as duas campanhas, foi selecionada uma amostra de 80 clientes da empresa A sendo constatada uma média de gastos de R\$1.245,00. Uma amostra aleatória de 120 clientes da empresa B, apresentou média de gastos igual a R\$1.638,00. Com base em pesquisas anteriores, sabe-se que o desvio-padrão populacional da empresa A é de R\$253,00 e da empresa B de R\$352,00. Teste a hipótese de que a campanha realizada pela empresa B apresenta resultados mais eficazes do que a campanha realizada pela empresa A. Use nível de significância de 1%.

$$\text{Estatística de teste} = -9,18$$

- a) determine a H_0 e H_a
- b) monte a curva e determine a área de não rejeição da H_0
- c) identifique a estatística de teste (z ou t)
- d) tome uma decisão

9) Foi realizado um experimento para estudar o efeito do álcool sobre as pessoas. Foram selecionadas aleatoriamente duas amostras, com 36 integrantes em cada uma, sendo que em apenas uma das amostras houve consumo de álcool. Os participantes do experimento foram submetidos a um teste de habilidades motoras e visuais, sendo anotado o número de erros cometidos. Na amostra em que não houve consumo de álcool, a média de erros foi de 1,93 com desvio-padrão de 0,88. Na amostra em que houve consumo de álcool, a média de erros foi de 4,72 e desvio-padrão de 2,23. Teste a hipótese de que não existe diferença significativa entre a média de erros nos dois grupos, ao nível de significância de 5%. Suponha as variâncias populacionais iguais.

10) Uma empresa oferece cursos de língua inglesa a seus funcionários, com o objetivo de melhorar a compreensão deste idioma. O quadro a seguir fornece os escores obtidos por dez de seus funcionários antes e após a realização do curso. Quanto maior o escore, maior a capacidade de compreensão do idioma.

Funcionário	1	2	3	4	5	6	7	8	9	10
Pré-teste	28	25	31	31	30	15	20	20	34	26
Pós-teste	30	29	32	34	25	18	16	27	31	28

Teste a hipótese de que o curso proporciona aos funcionários melhoria na capacidade de compreensão da língua inglesa ao nível de significância de 10%.

11) De 600 moradores, aleatoriamente selecionados, de

uma cidade em que a base da economia é a indústria, 450 são favoráveis a um projeto de lei, e de uma amostra de 240 moradores, aleatoriamente selecionados, de uma cidade cuja principal atividade é a agricultura, 150 são favoráveis. Teste a hipótese de que a proporção populacional de moradores favoráveis é maior na cidade industrial do que na cidade em que a base da economia é a agricultura. Use nível de significância de 5%.

Estatística de teste = 3,62

- a) determine a H_0 e H_a
- b) monte a curva e determine a área de não rejeição e rejeição da H_0
- c) identifique a estatística de teste (z ou t)
- d) Tome uma decisão

12) Uma agência de empregados domésticos deseja verificar o grau de satisfação de seus clientes. Os clientes estão classificados em duas categorias, de acordo com a renda familiar – A (maior renda) e B (menor renda). Foi selecionada uma amostra aleatória de cada categoria e solicitado que atribuíssem uma nota de 0 (totalmente insatisfeito) a 10 (totalmente satisfeito). Teste, ao nível de significância de 1%, a hipótese de que não existe diferença significativa entre o grau médio populacional de satisfação dos clientes da categoria A e B. Suponha variâncias populacionais iguais. Os resultados estão

no quadro a seguir:

Categoria A	$\bar{x}_A = 6,5$	$s_A = 0,9$	$n_A = 54$
Categoria B	$\bar{x}_B = 6,9$	$s_B = 1,2$	$n_B = 48$

13) Uma amostra aleatória de 12 formandos da turma A de um curso de datilografia obteve a média de 80,09 palavras por minuto, com desvio-padrão de 2,6551. Uma outra amostra aleatória com 12 alunos da turma B do mesmo curso de datilografia obteve em média 73,55 palavras por minuto, com desvio-padrão de 8,908. Teste a hipótese, ao nível de significância de 5%, de que a média de palavras por minuto para os formandos da turma A é significativamente maior do que para a turma B. Considere população normal com variâncias diferentes.

Estadística de teste = 2,4485

14) Em virtude dos protestos quanto à degradação do meio ambiente, algumas empresas têm adotado materiais totalmente recicláveis em seus produtos. Em uma pesquisa, foi perguntado a 270 consumidoras e 290 consumidores, aleatoriamente selecionados, se comprariam um produto que não tivesse a etiqueta informativa de que os materiais utilizados na embalagem são totalmente recicláveis. Dos entrevistados, 90 consumidoras e 125 consumidores responderam que não

comprariam o produto. Teste a hipótese de que a diferença de proporção populacional de consumidores e consumidoras que não comprariam o produto não será significativa ao $\alpha = 5\%$.

15) Uma associação beneficente adotou dois tipos de pedido de auxílio. No pedido feito por telefone, dos 1.600 realizados, obteve 310 adesões. Outro tipo de pedido de auxílio, mais dispendioso, teve 415 adesões dentre os 2500 contatos realizados. Um dos diretores da associação afirma que o pedido por telefone é mais eficaz. Teste esta afirmação ao nível de significância de 10%.

Respostas

1. $Z_{\text{teste}} = 10,787$. Rejeita-se H_0 .
2. $Z_{\text{crítico}} = \pm 1,96$. Rejeita-se H_0 .
3. $t_{\text{teste}} = -7,952$. Rejeita-se H_0 .
4. $t_{\text{crítico}} = -2,3808$. Não se rejeita H_0 .
5. $t_{\text{crítico}} = 1,8248$. Não se rejeita H_0 .
6. $t_{\text{crítico}} = 1,8331$. Não se rejeita H_0 .

7. $Z_{\text{teste}} = 2,193$; $Z_{\text{crítico}} = 1,28$. Rejeita-se H_0 .
8. $Z_{\text{teste}} = -9,18$; $Z_{\text{crítico}} = -2,33$. Rejeita-se H_0 .
9. $t_{\text{teste}} = -6,982$; $t_{\text{crítico}} = -/+1,9944$. Rejeita-se H_0 .
10. $t_{\text{teste}} = -0,8257$; $t_{\text{crítico}} = -1,3830$. Não se rejeita H_0 .
11. $Z_{\text{teste}} = 3,62$; $Z_{\text{crítico}} = 1,64$. Rejeita-se H_0 .
12. $t_{\text{teste}} = -1,9175$; $t_{\text{crítico}} = \pm 2,6259$. Não se rejeita H_0 .
13. $t_{\text{teste}} = 2,4485$; $t_{\text{crítico}} = 1,7709$. Rejeita-se H_0 .
14. $Z_{\text{teste}} = -2,402$; $Z_{\text{crítico}} = -/+1,96$. Rejeita-se H_0 .
15. $Z_{\text{teste}} = 2,295$; $Z_{\text{crítico}} = 1,28$. Rejeita-se H_0 .

8 Estudo da variância e do desvio-padrão

O estudo da variância e do desvio-padrão de uma distribuição é importante para que se possa controlar a variação em um determinado processo, produto ou operação. Por exemplo, na produção de peças é importante que elas sejam uniformes, ou seja, variem muito pouco ou quase nada do padrão estabelecido; na embalagem de produtos o peso não deve variar muito do estipulado, pois acima do peso trazem prejuízo ao fabricante, e abaixo do peso, prejuízo ao consumidor. Portanto, para manter o controle de qualidade dos produtos, processos ou operações, a variabilidade também deve ser controlada.

No estudo da média e das proporções, usamos a distribuição normal e a distribuição t de Student, que são distribuições simétricas. No estudo do desvio-padrão e da variância usamos a distribuição qui-quadrado (χ^2).

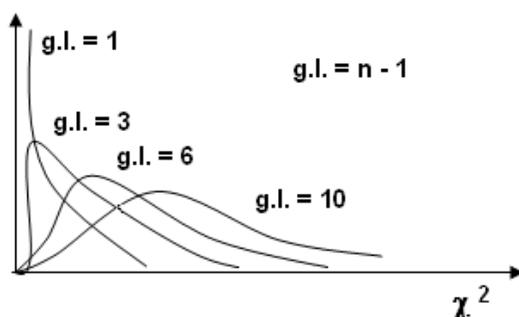
8.1 A distribuição qui-quadrado (χ^2)

Na distribuição qui-quadrado, a suposição de que a população seja distribuída normalmente é mais crítica do que nas outras distribuições, pois qualquer afastamento da normalidade acarretará erros grosseiros.

Se de uma população normalmente distribuída selecionamos aleatoriamente amostras independentes de tamanho n e calculamos a variância amostral s^2 para cada amostra, a estatística amostral $\chi^2 = (n-1)s^2/\sigma^2$ formará uma distribuição qui-quadrado (χ^2). A distribuição qui-quadrado é uma família de curvas, cada uma delas determinada pelos graus de liberdade (g.l.) semelhante a distribuição t ; a diferença é que a distribuição qui-quadrado não é simétrica.

Características da distribuição qui-quadrado (χ^2)

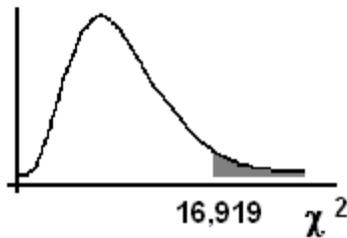
- Quando n tende ao infinito (∞), a distribuição χ^2 tende a distribuição normal.
- A curva é assimétrica.
- A área total sob a curva é 1.
- Todos os valores de χ^2 são maiores ou iguais a zero.
- Para $g.l. \geq 3$ tem origem em zero e é assintótica ao eixo horizontal à direita.



8.1.1 Uso da tabela da distribuição qui-quadrado (χ^2)

Para usar a tabela da distribuição qui-quadrado, deve-se atentar para o fato de que a curva não é simétrica. O corpo da tabela é constituído das probabilidades (área sob a curva). Esta tabela trabalha com dois parâmetros: graus de liberdade (g.l) e a área da extremidade da curva à direita. Na primeira coluna da tabela estão os valores dos graus de liberdade e na primeira linha está a área da extremidade da curva à direita.

Por exemplo, o valor de $\chi^2_{0,05;9}$ (onde 0,05 é a área da extremidade da curva à direita e 9 os graus de liberdade) é obtido pela intersecção da linha que contém o valor 0,05 e a coluna que contém g.l. = 9:



	Área à direita da curva						
<u>g.l.</u>	0,995	0,99	...	0,05	0,025	0,01	0,005
1	0,000	0,0002	...	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201		5,9915	7,3778	9,2104	10,596
3	0,0717	0,1148		7,8147	9,3484	11,345	12,838
4	0,2070	0,2971		9,4877	11,143	13,277	14,860
Graus de liberdade: <u>g.l.</u> = n - 1 <u>g.l.</u> = 9	0,4118	0,5543	...	11,071	12,832	15,086	16,750
	0,6757	0,8721		12,592	14,450	16,812	18,548
	0,9893	1,2390		14,06			
8	1,3444	1,6465	...	15,507		20,090	21,955
9	1,7349	2,0879		16,919	19,023	21,666	23,589
10	2,1558	2,5582		18,307	20,483	23,209	25,188
11	2,6032	3,0535	...	19,675	21,920	24,725	26,757
⋮	⋮	⋮	...	⋮	⋮	⋮	⋮

Área à direita da curva

$\chi^2_{0,05;9} = 16,919$

8.2 Intervalo de confiança para variância σ^2 ou desvio-padrão σ

Na estimação do intervalo de confiança para a variância ou desvio-padrão da população¹, partiremos das seguintes suposições:

- A amostra é uma amostra aleatória simples.

¹ Vídeos sobre intervalos de confiança e testes de hipóteses com a distribuição qui-quadrado, gravados pelas autoras deste livro, você pode assistir no Youtube no Canal da Prof^a Suzi Samá.

▪ A população deve apresentar distribuição normal mesmo para amostras maiores que 30.

As distribuições qui-quadrado não são simétricas. Por conseguinte, um intervalo de confiança para um desvio-padrão envolve o uso de dois valores χ^2 diferentes, ao contrário do sistema “mais e menos” usado com os intervalos de confiança, baseados na distribuição normal e distribuição t .

O intervalo de confiança para o desvio-padrão populacional é dado por:

$$\sqrt{\frac{(n-1).s^2}{\chi_{\text{direito},g.l}^2}} \leq \sigma \leq \sqrt{\frac{(n-1).s^2}{\chi_{\text{esquerdo},g.l}^2}}$$

O intervalo de confiança para variância da população é dado por:

$$\frac{(n-1).s^2}{\chi_{\text{direito},g.l}^2} \leq \sigma^2 \leq \frac{(n-1).s^2}{\chi_{\text{esquerdo},g.l}^2}$$

A tabela indica as proporções de área sob as distribuições qui-quadrado, de acordo com vários graus de liberdade (g.l.). Na fórmula geral acima, os subscritos “direito” e “esquerdo” indicam as duas áreas obtidas na construção do intervalo de confiança.

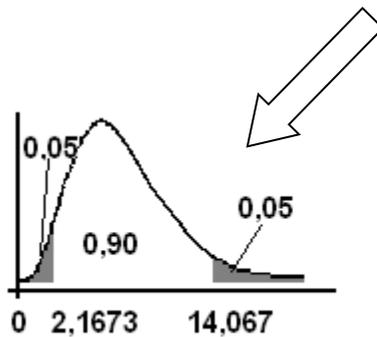
Exemplo 1: As notas dos alunos de uma escola são normalmente distribuídas. O desvio-padrão das notas de oito

estudantes escolhidos aleatoriamente de uma escola é 2,4. Determinar um intervalo de 90% de confiança para o desvio-padrão populacional.

Solução: como a amostra é aleatória e a população normalmente distribuída podemos usar a distribuição qui-quadrado para estimar o desvio-padrão populacional.

Dados do exemplo: $s = 2,4$ $1 - \alpha = 0,90$ $n = 8$

Os valores de χ^2_{direito} e χ^2_{esquerdo} são :



Monte o intervalo de confiança:

$$\sqrt{\frac{(n-1)s^2}{\chi_{\text{direito},g,l}^2}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi_{\text{esquerdo},g,l}^2}}$$

$$\sqrt{\frac{(8-1)(2,4)^2}{14,067}} \leq \sigma \leq \sqrt{\frac{(8-1)(2,4)^2}{2,1673}}$$

$$1,693 \leq \sigma \leq 4,313$$

$$P(1,693 \leq \sigma \leq 4,313) = 90\%$$

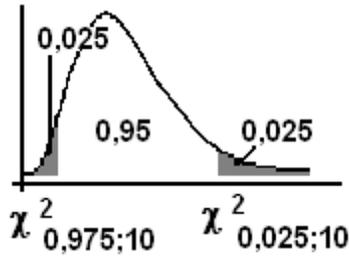
Interprete o intervalo de confiança: Com 90% de confiança podemos afirmar que o intervalo de 1,693 a 4,313 contém o desvio-padrão populacional das notas dos estudantes.

Exemplo 2: Uma amostra aleatória de onze elementos, extraída de uma população com distribuição normal, forneceu variância $s^2 = 7,08$. Construir um intervalo de 95% de confiança para a variância dessa população.

Solução: como a amostra é aleatória e a população normalmente distribuída, podemos usar a distribuição qui-quadrado para estimar a variância populacional.

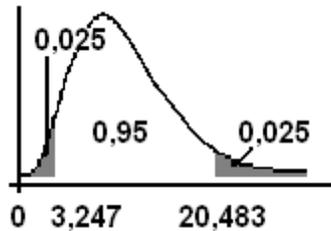
Dados do exemplo: $s^2 = 7,08$ $1 - \alpha = 0,95$ $n = 11$

Os valores de χ_{direito}^2 e χ_{esquerdo}^2 são:



$$\chi^2_{\text{direito};g.l.} = \chi^2_{0,025;10} = 20,483$$

$$\chi^2_{\text{esquerdo};g.l.} = \chi^2_{0,975;10} = 3,2470$$



Monte o intervalo de confiança:

$$\frac{(n-1)s^2}{\chi^2_{\text{direito},g.l.}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\text{esquerdo},g.l.}}$$

$$\frac{(11-1)7,08}{20,483} \leq \sigma^2 \leq \frac{(11-1)7,08}{3,2470}$$

$$3,457 \leq \sigma^2 \leq 21,805$$

$$P(3,457 \leq \sigma^2 \leq 21,805) = 95\%$$

Interprete o intervalo de confiança: Com 95% de confiança podemos afirmar que o intervalo de 3,457 a 21,805 contém a variância populacional.

8.3 Teste de hipótese para variância e desvio-padrão

Nos testes de hipótese para a variância ou desvio-padrão da população, partiremos das seguintes suposições:

- A amostra é uma amostra aleatória simples.
- A população deve apresentar distribuição normal mesmo para amostras maiores do que 30.

A estatística de teste a ser usada no teste de hipótese para a variância ou desvio-padrão da população é dada por:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad \text{com g.l.} = n - 1$$

Exemplo de teste de hipótese para variância: Um órgão de defesa do consumidor testa regularmente o peso dos produtos vendidos em estabelecimentos comerciais. Uma empresa alega que seus produtos apresentam, no máximo, variância de $50g^2$. A inspeção de uma amostra aleatória de 15 produtos acusou variância amostral de $45g^2$. Considerando que os pesos desses produtos apresentam distribuição normal, teste a alegação da empresa ao nível de significância de 10%.

Solução:

1ª Retire os dados do exemplo:

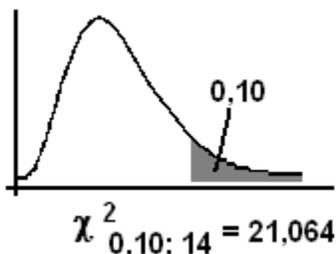
$$\sigma^2 = 50g^2 \quad n = 15 \quad s^2 = 45g^2 \quad \alpha = 10\%$$

2ª Formule as hipóteses H_0 e H_a (lembre-se que a hipótese nula é aquela que contém o sinal de igual):

$$H_0 : \sigma^2 \leq 50$$

$$H_a : \sigma^2 > 50$$

3ª Determine o valor crítico: Com base no sinal da hipótese alternativa ($>$), podemos verificar que o teste é unilateral à direita, portanto a área da extremidade direita da curva é de 0,10. Na tabela da distribuição qui-quadrado vamos procurar o valor de $\chi_{0,10;14}^2$.



4ª Calcule a estatística de teste:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(15-1)45}{50} = 12,6$$

5ª Tome a decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $\chi_{\text{teste}}^2 = 12,6$ está dentro da área de não rejeição da H_0 , pois se encontra à esquerda de 21,064.

6ª Interprete o resultado: Não se rejeita a hipótese nula, ao nível de significância de 10%, portanto podemos concluir que há evidências suficientes para apoiar a alegação da empresa de que a variância do peso é de no máximo $50g^2$.

Exemplo de teste de hipótese para o desvio-padrão:
Certo produto embalado por uma empresa apresenta distribuição normal com desvio-padrão de 8g. A máquina de embalar passou por uma revisão e o técnico afirma que a dispersão diminuiu após a calibração. Com a finalidade de testar a afirmação do técnico, ao nível de significância de 1%, foi selecionada uma amostra aleatória de 12 observações. Foram obtidos os seguintes resultados: 500g – 492g 490g – 502g – 505g – 493g – 500g – 498g – 494g – 509g – 491g 497g.

Solução:

1ª Para calcular a estatística de teste, precisamos do desvio-padrão da amostra.

$$s = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]} = \sqrt{\frac{1}{12-1} \left[2.971.453 - \frac{5.971^2}{12} \right]} = 5,9g$$

2ª Retire os dados do problema:

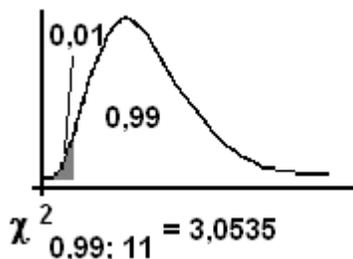
$$s = 5,9g \quad n = 12 \quad \sigma = 8g \quad \alpha = 1\%$$

3ª Formule as hipóteses H_0 e H_a (lembre-se que a hipótese nula é aquela que contém o sinal de igual):

$$H_0 : \sigma \geq 8$$

$$H_a : \sigma < 8$$

4ª Determine o valor crítico: Com base no sinal da hipótese alternativa (<), podemos verificar que o teste é unilateral à esquerda, portanto a área da extremidade esquerda da curva é de 0,1. Na tabela da distribuição qui-quadrado procuramos o valor de $\chi^2_{0,99;11}$, pois a tabela fornece o valor do qui-quadrado da direita para a esquerda.



5ª Calcule a estatística de teste:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(12-1)5,9^2}{8^2} = 5,98$$

6ª Tome a decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $\chi^2_{\text{teste}} = 5,98$ está dentro da área de não rejeição da H_0 , pois se encontra à direita de 3,0535.

7ª Interprete o resultado: Não se rejeita a hipótese nula, ao nível de significância de 1%, portanto podemos concluir que não há evidências suficientes para apoiar a afirmação do técnico de que a dispersão na máquina de embalar diminui significativamente após a calibração.

8.4 Teste qui-quadrado não-paramétrico

O teste qui-quadrado não-paramétrico é recomendado para verificar se as variáveis são independentes ou relacionadas, e também para o tratamento estatístico de dados oriundos de tabelas de dupla entrada (tabelas de contingência). Esse teste é adaptável a estudos que envolvem variáveis com níveis de mensuração nominal e ordinal. Para aplicá-los não é necessário fazer suposições quanto ao modelo de distribuição de probabilidade da população.

8.4.1 Tabelas de contingência

No estudo da relação ou independência entre duas variáveis categóricas, aplicamos o teste do qui-quadrado não-paramétrico. A representação das frequências é dada por uma tabela de dupla entrada ou tabela de contingência, em que uma variável é usada para categorizar linhas e a segunda variável é usada para categorizar colunas.

Exemplo de tabela de contingência: A tabela apresenta os resultados de um levantamento, em que foram selecionados aleatoriamente estudantes de uma universidade, que responderam a uma pesquisa que tinha o objetivo de verificar se há independência entre a variável sexo e a variável opção pela área do curso.

Sexo	Opção			Total
	Ciências Sociais	Ciências Humanas	Ciências Exatas	
Masculino	80	55	70	205
Feminino	65	85	45	195
Total	145	140	115	400

A variável linha tem duas categorias: masculino e feminino. A variável coluna também tem três categorias: ciências sociais, ciências humanas e ciências exatas. As variáveis linha e coluna podem ter mais do que duas categorias.

Nos testes de independência, na hipótese nula afirmamos que as variáveis são independentes; conseqüentemente, na hipótese alternativa afirmamos que as variáveis são dependentes:

A estatística a ser usada é dada por:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad \text{com g.l.} = (r - 1) \cdot (c - 1)$$

onde O representa as frequências observadas em um experimento, E as frequências esperadas do experimento, r o número de linhas e c o número de colunas.

A frequência esperada (E) de uma tabela de contingência é obtida pelo produto entre a probabilidade (p) de se estar na célula e o número total observado (n):

$$E = p.n$$

onde p é obtido pela aplicação do teorema do produto para eventos independentes, ou seja, $P(A \text{ e } B) = P(A).P(B)$. Isso se deve ao fato de que estamos supondo (hipótese nula) que os eventos são independentes.

Dessa forma, a frequência esperada da primeira coluna e primeira linha da tabela é dada por:

$$E = p.n = P(A \text{ e } B) . n$$

como $P(A \text{ e } B) = P(A) . P(B)$ substituindo, temos:

$$E = P(A) . P(B) . n$$

onde A corresponde ao evento “ser do sexo masculino” e B o evento “ter feito a opção por ciências sociais”.

$$E = \frac{n(A)}{n(S)} \cdot \frac{n(B)}{n(S)} \cdot n = \frac{205}{400} \cdot \frac{145}{400} \cdot 400$$

$$E = \frac{205 \cdot 145}{400} = 74,31$$

Portanto, podemos dizer que, apesar de 80 estudantes do sexo masculino terem feito a opção por ciências sociais, esperava-se que 74,31 estudantes do sexo masculino tivessem feito essa opção, se realmente as variáveis fossem independentes.

Como o valor 205 é a soma da linha, 145 a soma da coluna e 400 o total observado, em geral usa-se uma fórmula mais prática para obter o valor da frequência esperada:

$$E = \frac{(\text{soma da linha} \times \text{soma da coluna})}{\text{total observado}}$$

Obs.: 1º Todas as frequências esperadas têm que ser igual ou maior do que 5.

2º A região crítica se localiza apenas na cauda direita da curva.

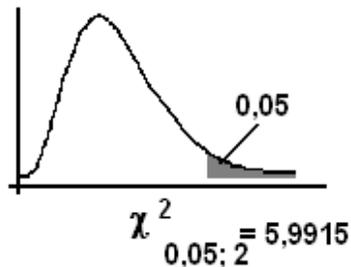
Exemplo: Para a tabela de contingência apresentada no exemplo acima, teste a hipótese de que a variável sexo e a variável opção pela área do curso universitário são independentes ao nível de significância de 5%.

Solução:

H_0 : as variáveis são independentes

H_a : as variáveis são dependentes

g.l. = $(r - 1) \cdot (c - 1) = (2 - 1) \cdot (3 - 1) = 2$



Pela fórmula mais prática para calcular a frequência esperada da célula da primeira linha e segunda coluna, basta multiplicar o total da linha (205) pelo total da coluna (140) e dividir pelo total observado (400).

$$E = \frac{(\text{soma da linha} \times \text{soma da coluna})}{\text{total observado}} = \frac{205 \cdot 140}{400} = 71,75$$

Continuando o cálculo para as demais células:

$$E = \frac{205 \cdot 115}{400} = 58,94$$

$$E = \frac{195 \cdot 145}{400} = 70,69$$

$$E = \frac{195 \cdot 140}{400} = 68,25$$

$$E = \frac{195 \cdot 115}{400} = 56,06$$

Como todas as frequências esperadas são iguais ou maiores do que 5, podemos dar continuidade ao teste.

Calculando a estatística de teste:

$$\chi_{\text{teste}}^2 = \sum \frac{(O - E)^2}{E}$$

$$\begin{aligned} \chi_{\text{teste}}^2 &= \frac{(80 - 74,31)^2}{74,31} + \frac{(55 - 71,75)^2}{71,75} + \frac{(70 - 58,94)^2}{58,94} + \frac{(65 - 70,69)^2}{70,69} + \\ &\quad + \frac{(85 - 68,25)^2}{68,25} + \frac{(45 - 56,06)^2}{56,06} = \end{aligned}$$

$$\chi_{\text{teste}}^2 = 0,4357 + 3,9103 + 2,0754 + 0,458 + 4,1108 + 2,182$$

$$\chi_{\text{teste}}^2 = 13,1722$$

5ª Tome a decisão (não rejeitar ou rejeitar H_0):

Podemos verificar que o $\chi_{\text{teste}}^2 = 13,1722$ está dentro da área de rejeição da H_0 , pois se encontra à direita do $\chi_{\text{crítico}}^2 = 5,9915$.

6ª Interprete o resultado: Rejeita-se a hipótese nula, ao nível de significância de 5%, ou seja, podemos concluir que as variáveis não são independentes.

Exercícios resolvidos

1) Uma máquina automática de enchimento de garrafas com refrigerante apresenta distribuição normal. O setor de qualidade da empresa considera que a variância do volume de enchimento não deve exceder a 100 ml^2 , evitando, dessa forma, que as garrafas de refrigerante apresentem enchimento em demasia ou incompleto. Uma amostra aleatória de 18 garrafas resulta em variância do volume de enchimento de $s^2 = 225 \text{ ml}^2$. Teste a hipótese de que a variância está dentro das especificações determinadas pelo setor de qualidade da empresa ao nível de significância de 5%.

Solução:

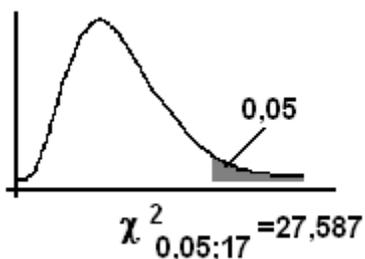
Dados do problema:

$$\sigma^2 = 100 \text{ ml}^2 \quad n = 18 \quad s^2 = 225 \text{ ml}^2 \quad \alpha = 5\%$$

$$H_0 : \sigma^2 \leq 100$$

$$H_a : \sigma^2 > 100$$

Com base no sinal da hipótese alternativa ($>$), podemos verificar que o teste é unilateral à direita, portanto a área da extremidade direita da curva é de 0,05. Na tabela da distribuição qui-quadrado vamos procurar o valor de $\chi^2_{0,05;17}$.



$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(18-1) \cdot 225}{100} = 36$$

Podemos verificar que o $\chi^2_{\text{teste}} = 36$ está fora da área de não rejeição da H_0 , pois se encontra à direita de 27,587.

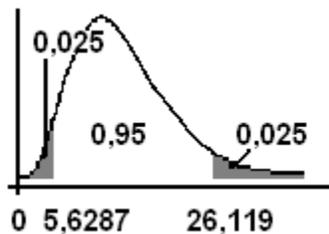
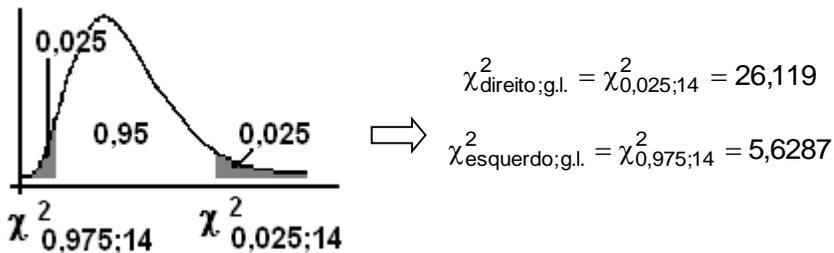
Interpretação do resultado: Rejeita-se a hipótese nula, ao nível de significância de 5%, portanto podemos concluir que a variância do volume de enchimento está fora das especificações do setor de qualidade da empresa.

2) Um fabricante de pneus está testando a vida útil de um pneu fabricado com um novo componente. As pesquisas demonstram que a vida útil do pneu apresenta distribuição normal. Uma amostra aleatória de 15 pneus apresentou desvio-

padrão de 2.978km. Monte um intervalo de 95% de confiança para o desvio-padrão da vida útil do pneu.

Solução: como a amostra é aleatória e a população normalmente distribuída, podemos usar a distribuição qui-quadrado para estimar o desvio-padrão populacional.

Dados do problema: $s = 2.978$ $1 - \alpha = 0,95$ $n = 15$



Montando o intervalo de confiança:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{direito,g.l.}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{esquerdo,g.l.}}}$$

$$\sqrt{\frac{(15-1)(2.978)^2}{26,119}} \leq \sigma \leq \sqrt{\frac{(15-1)(2.978)^2}{5,6287}}$$

$$P(2.180,27 \leq \sigma \leq 4.696,61) = 95\%$$

Interpretando o intervalo de confiança: Com 95% de confiança, podemos afirmar que o intervalo de 2.180,27 a 4.696,61 contém o desvio-padrão populacional da vida útil dos pneus.

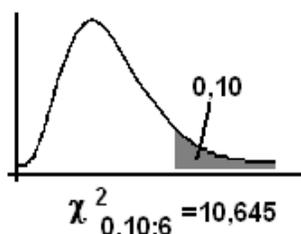
3) Uma empresa realizou uma pesquisa entre seus funcionários para verificar se há relação entre a satisfação no trabalho e a renda familiar. Teste a hipótese de que a satisfação no trabalho é independente da renda familiar ao nível de significância de 10%.

Renda familiar	Satisfação no trabalho			Total
	Satisfeito	Regular	Insatisfeito	
Menos de R\$1.200,00	73	64	39	176
R\$1.200,00 a R\$2.400,00	114	92	52	258
R\$2.401,00 a R\$3.600,00	105	71	38	214
Mais de R\$3.600,00	87	45	17	149
Total	379	272	146	797

Solução:

H₀ : as variáveis são independentes

H_a : as variáveis são dependentes



$$g.l. = (r - 1).(c - 1) = (4 - 1). (3 - 1) = 6$$

As frequências esperadas estão calculadas na tabela a seguir:

$$E = \frac{(\text{soma da linha} \times \text{soma da coluna})}{\text{total observado}}$$

Renda familiar	Satisfação no trabalho					
	Satisfeito		Regular		Insatisfeito	
	O _i	E _i	O _i	E _i	O _i	E _i
Menos de R\$1.200,00	73	83,69	64	60,07	39	32,24
R\$1.200,00 a R\$2.400,00	114	122,69	92	88,05	52	47,26
R\$2.401,00 a R\$3.600,00	105	101,76	71	73,03	38	39,20
Mais de R\$3.600,00	87	70,85	45	50,85	17	27,29
Total	379		272		146	

Calculando a estatística de teste:

$$\chi_{\text{teste}}^2 = \sum \frac{(O - E)^2}{E}$$

$$\begin{aligned} \chi_{\text{teste}}^2 &= \frac{(73 - 83,69)^2}{83,69} + \frac{(114 - 122,69)^2}{122,69} + \frac{(105 - 101,76)^2}{101,76} + \frac{(87 - 70,85)^2}{70,85} + \\ &+ \frac{(64 - 60,07)^2}{60,07} + \frac{(92 - 88,05)^2}{88,05} + \frac{(71 - 73,03)^2}{73,03} + \frac{(45 - 50,85)^2}{50,85} + \\ &+ \frac{(39 - 32,24)^2}{32,24} + \frac{(52 - 47,26)^2}{47,26} + \frac{(38 - 39,2)^2}{39,2} + \frac{(17 - 27,29)^2}{27,29} = \end{aligned}$$

$$\begin{aligned} \chi_{\text{teste}}^2 &= 1,365 + 0,616 + 0,103 + 3,681 + 0,257 + 0,177 + 0,056 + \\ &+ 0,673 + 1,417 + 0,475 + 0,037 + 3,880 \end{aligned}$$

$$\chi_{\text{teste}}^2 = 12,739$$

Podemos verificar que o $\chi_{\text{teste}}^2 = 12,739$ está fora da área de não rejeição da H_0 , pois se encontra à direita do $\chi_{\text{crítico}}^2 = 10,645$.

Interpretação do resultado: Rejeita-se a hipótese nula, ao nível de significância de 10%, ou seja, podemos concluir que as variáveis não são independentes.

Exercícios complementares

1) O gerente de qualidade de uma empresa afirma que a variabilidade da máquina de encher pacotes de café está dentro dos padrões. De acordo com as especificações do controle de qualidade, o desvio-padrão deve ser de no máximo 10g. Uma amostra aleatoriamente selecionada de 16 pacotes foi analisada e observou-se um desvio-padrão de 14,5g. Teste a afirmação do gerente ao nível de significância de 5%. Suponha população normalmente distribuída.

2) Um professor afirma que a variabilidade das notas nos exames finais dos alunos na disciplina de estatística diminuiu. Historicamente a variância das notas nos exames finais é de no mínimo 210,8. Uma amostra aleatória de 13 alunos é extraída, obtendo-se variância igual a 36,2. Teste a afirmação do professor ao nível de significância de 5%. Suponha população normalmente distribuída.

3) O gerente de uma empresa de crédito deseja estimar o desvio-padrão da renda mensal de clientes com cartão de crédito. Um auditor seleciona uma amostra aleatória de 41

contas e encontra desvio-padrão de R\$230,65. Monte o intervalo de confiança de 90% para o desvio-padrão da renda mensal dos clientes com cartão de crédito.

4) O tempo de espera (em minutos) de clientes em um banco, onde os clientes entram em uma fila única que é atendida por três guichês, apresenta distribuição normal. Uma amostra aleatória foi obtida com a finalidade de estimar o desvio-padrão populacional do tempo de espera dos clientes. A partir dos dados obtidos na amostra, construa um intervalo de 95% de confiança para o desvio-padrão populacional do tempo de espera dos clientes na fila deste banco.

Dados da amostra em minutos:

6,5 6,6 6,7 6,8 7,1 7,3 7,4 7,7 7,7 7,7

5) A variância de uma amostra escolhida aleatoriamente de 10 lâmpadas elétricas produzidas em uma determinada empresa é de 120 horas² de duração. Supondo que as horas de duração desse produto apresentam distribuição normal, construa um intervalo de confiança para a variância de todas as lâmpadas da empresa com uma confiança de 99%.

6) Um estudo é realizado para determinar se a preferência por diferentes esportes está relacionada à faixa etária de homens. Uma amostra aleatória de 300 homens é selecionada aleatoriamente, e solicita-se a cada indivíduo que indique seu

esporte preferido. Teste a hipótese de que as variáveis opção pelo esporte e faixa etária são independentes. Use nível de significância de 5%.

Faixa etária	Opção			
	Futebol	Volêi	Basquete	Total
Menos de 20	30	20	25	75
20 a 40	35	25	20	80
Mais de 40	45	45	55	145
Total	110	90	100	300

7) De acordo com as normas de controle de qualidade de uma empresa, as embalagens dos detergentes que fabrica apresentam desvio-padrão de 40g. Uma amostra aleatória de 36 caixas revelou um desvio-padrão de 48g. Teste a hipótese de que o desvio-padrão populacional é igual a 40g ao nível de significância de 1%. Suponha população normalmente distribuída.

8) Uma universidade implantou um novo sistema de avaliação de aprendizagem. Foi realizada uma pesquisa com 500 alunos aleatoriamente escolhidos entre três cursos, a fim de verificar se existe relação entre a opinião do estudante sobre a mudança na avaliação e o curso do estudante. Teste a hipótese de que as variáveis são independentes ao nível de significância de 10%.

Curso	Opinião do estudante		
	Favorável	Contrário	Indiferente
Administração	60	85	15
Direito	80	55	25
Economia	95	65	20

9) De uma população com distribuição normal, foi selecionada uma amostra aleatória de 18 estudantes do sexo feminino, obtendo-se desvio-padrão de 1,8cm para as estaturas dos estudantes. Determinar um intervalo de 95% de confiança para o desvio-padrão populacional.

10) Uma amostra, aleatoriamente selecionada, de 15 barras de ferro forneceu uma variância de $6,88\text{cm}^2$ para o comprimento das barras. Construa um intervalo de 90% de confiança para a variância populacional. Suponha população normal.

11) Para realizar um estudo sobre a variabilidade do peso das embalagens de café solúvel de uma determinada marca, selecionou-se uma amostra aleatória de 18 pacotes, encontrando-se variância $11,7\text{g}^2$. Teste a hipótese de que a variância populacional é inferior a 28g^2 , ao nível de significância de 10%. Qual sua conclusão? Suponha que a população apresenta distribuição normal.

12) Um pesquisador afirma que o desvio-padrão da renda familiar anual de um determinado bairro é maior do que R\$450,00. Uma amostra aleatoriamente selecionada de 20 famílias resultou em um desvio-padrão de R\$350,00. Por pesquisas anteriores sabe-se que a renda familiar dessa população apresenta distribuição normal. Teste a afirmação do pesquisador ao nível de significância de 5%. Qual sua conclusão?

13) Uma empresa aérea está fazendo uma pesquisa entre seus usuários. Uma das questões levantadas é se existe relação entre o tipo de passagem e o tipo de voo. Usando um nível de significância de 10%, teste a independência entre o tipo de passagem e o tipo de voo. Qual é a sua conclusão?

Tipo de passagem	Tipo de voo	
	Vôos domésticos	Vôos internacionais
Primeira classe	42	36
Classe executiva	112	135
Classe econômica	420	155

Respostas

1. $\chi^2_{\text{teste}} = 31,5375; \chi^2_{\text{crítico}} = 24,996$; Rejeita-se a H_0 .

2. $\chi^2_{\text{teste}} = 2,0607; \chi^2_{\text{crítico}} = 5,226$; Rejeita-se a H_0 .

3. $P(195,36 \leq \sigma \leq 283,33) = 90\%$

4. $P(0,3279 \leq \sigma \leq 0,8703) = 95\%$

5. $P(45,784 \leq \sigma^2 \leq 622,514) = 99\%$

6. $\chi_{\text{teste}}^2 = 5,539; \chi_{\text{crítico}}^2 = 9,4877; \text{Não se rejeita a } H_0.$

7. $\chi_{0,005;35}^2 = 60,275; \chi_{0,995;35}^2 = 17,192 ; \chi^2 = 50,4; \text{Não se rejeita a } H_0.$

8. $\chi_{\text{teste}}^2 = 15,9459; \chi_{\text{crítico}}^2 = 7,7794; \text{Rejeita-se a } H_0.$

9. $P(1,351 \leq \sigma \leq 2,698) = 95\%$

10. $P(4,067 \leq \sigma^2 \leq 14,659) = 90\%$

11. $\chi_{\text{teste}}^2 = 7,1036; \chi_{\text{crítico}}^2 = 10,0850; \text{Não se rejeita a } H_0.$

12. $\chi_{\text{teste}}^2 = 11,4938; \chi_{\text{crítico}}^2 = 30,1450; \text{Não se rejeita a } H_0.$

13. $\chi_{\text{teste}}^2 = 61,03; \chi_{\text{crítico}}^2 = 4,6052; \text{Rejeita-se a } H_0.$

9 Análise de regressão e correlação

A análise de regressão¹ é uma técnica estatística que analisa as relações existentes entre uma única variável dependente e uma ou mais variáveis independentes, permitindo estimar o valor de uma variável a partir do valor de outra(s) variável(eis), com o objetivo de estudar as relações entre elas a partir de um modelo matemático. Por sua vez, a análise de correlação linear permite determinar o grau de relação entre duas variáveis ou entre uma variável e um conjunto de outras variáveis.

Quando o problema envolve uma única variável independente, a técnica estatística é chamada análise de regressão simples; quando envolve duas ou mais variáveis independentes, denomina-se análise de regressão múltipla. O mesmo na análise de correlação.

O problema da análise de regressão consiste em definir a forma de relação existente entre as variáveis. Por exemplo, na análise de regressão simples, as variáveis, x e y , podem apresentar uma relação linear, ou uma relação exponencial ou ainda uma relação polinomial, entre outras.

$$y = b + ax \quad (\text{Relação Linear})$$

$$y = ax^2 \quad (\text{Relação Exponencial})$$

$$y = ax^2 + bx + c \quad (\text{Relação Polinomial})$$

¹ Vídeo sobre análise de regressão, gravado pelas autoras deste livro, você pode assistir no Youtube no Canal da Prof^a Suzi Samá.

A variável dependente é y , aquela que será predita a partir da variável independente x .

Numa análise de regressão linear múltipla, a relação entre as variáveis independentes e a variável dependente é dada por:

$$y = b + a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_kx_k$$

Neste livro, a seguir, estudaremos a análise de regressão linear simples e a análise de correlação linear.

9.1 Análise de regressão linear simples

Nesta subseção, veremos a análise de regressão linear para duas variáveis. Para obter a equação que relacione essas variáveis, devemos seguir os seguintes passos:

a) Coletar dados sobre a variável considerada. Exemplo: suponha que x e y representem, respectivamente, a altura e o peso de adultos. Então uma amostra com n indivíduos apresentaria as alturas (x_1, x_2, \dots, x_n) e os pesos (y_1, y_2, \dots, y_n) .

b) Colocar os pontos $(x_1 ; y_1)$, $(x_2 ; y_2)$..., $(x_n ; y_n)$ em um sistema de coordenadas cartesianas. A esse conjunto de pontos denominamos de diagrama de dispersão.

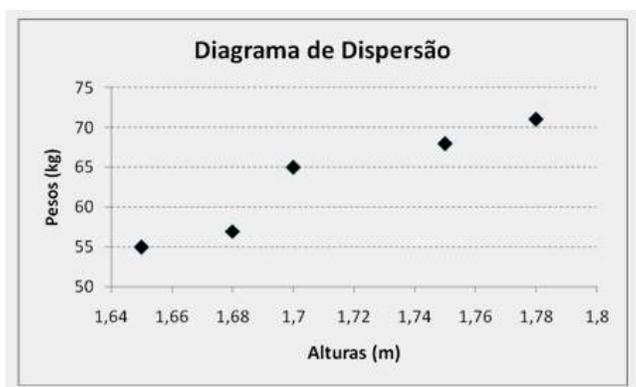
Em um diagrama de dispersão, é possível visualizar uma curva regular que se aproxime dos dados, a qual denominamos curva de ajustamento. Se os dados parecem estar próximos de uma linha reta, neste caso diremos que existe uma relação linear entre as variáveis.

Exemplo:

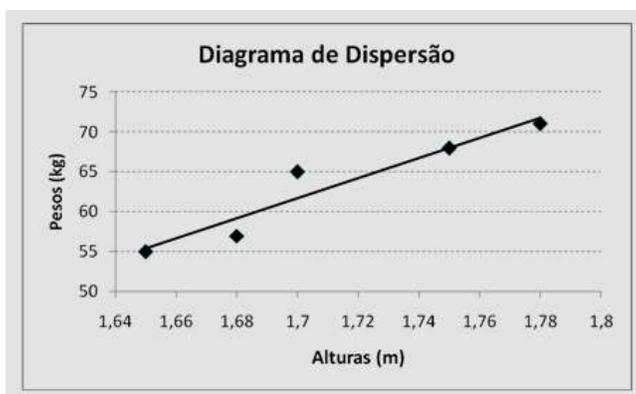
Dados a altura e o peso de um grupo de nadadoras, a variável peso é a variável dependente, pois o peso depende da altura.

x (altura m)	1,65	1,68	1,70	1,75	1,78
y (peso kg)	55	57	65	68	71

Observe o diagrama de dispersão.



Observando o diagrama de dispersão, podemos verificar que é possível ajustar uma reta aos pontos do diagrama.



O problema consiste em estimar os parâmetros, β_0 e β_1 da equação de regressão. Para todos os pontos possíveis (x, y) existe uma reta da regressão dada pela expressão:

$$y = \beta_0 + \beta_1 x$$

Como através de uma amostra obteremos uma estimativa da verdadeira equação de regressão, denominamos:

$$\hat{y} = b_0 + b_1 x$$

onde $\hat{y} = y$ estimado; $b_0 =$ estimativa de β_0 ; $b_1 =$ estimativa de β_1

Com os parâmetros estimados, a equação de regressão permite fazer previsões sobre a variável y para dados valores de x. Entretanto, não se recomenda estimar y para valores de x muito afastados do intervalo dos x_i observados na amostra.

9.2 Método dos mínimos quadrados

Um dos métodos mais simples para o cálculo das estimativas dos parâmetros β_0 e β_1 é o **método dos mínimos quadrados**.

A cada valor x_i temos um valor y_i , que é o valor observado na amostra, e um valor \hat{y}_i , que é o valor estimado pela reta de regressão. A diferença entre o valor observado (y_i) e o valor estimado (\hat{y}_i) para certo valor x_i é denominado desvio (d_i). Logo $d_i = y_i - \hat{y}_i$.

O método dos mínimos quadrados visa a estimar os parâmetros da equação de regressão, de forma que a soma dos

quadrados dos desvios (d_i) seja a menor possível.

Dessa forma,

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

9.2.1 Critério dos mínimos quadrados

O critério de mínimos quadrados corresponde, portanto, a determinar as estimativas b_0 e b_1 que minimizam a diferença ($y_i - \hat{y}_i$).

$$\text{mín} \sum (y_i - \hat{y}_i)$$

**Critério dos mínimos
quadrados**

O valor de: $\sum_{i=1}^n d_i^2$, assumirá um mínimo quando as derivadas parciais em relação a b_0 e b_1 forem nulas. Segundo esse método, poderemos avaliar as estimativas b_0 e b_1 pela aplicação das seguintes equações:

$$b_1 = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sum x_i^2 - [(\sum x_i)^2]/n}$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad \bar{x} = \frac{\sum x_i}{n} \quad \bar{y} = \frac{\sum y_i}{n}$$

Exemplo 1:

1) A tabela a seguir relaciona as distâncias percorridas por carros (km) e seus consumos de combustível em rodovias

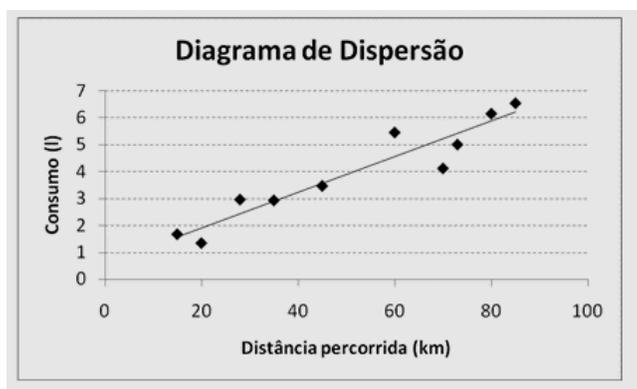
brasileiras (l), para uma amostra de carros de passeio novos. Com base nos resultados:

- a) Construa um diagrama de dispersão para esses dados.
- b) Tente aproximar a relação entre x e y traçando uma linha reta através dos pontos.
- c) Determine uma equação de regressão estimada calculando os valores de b_0 e b_1 .
- d) Pode-se concluir que para percursos mais longos há maior consumo de combustível?

Distância percorrida (km)	20	60	15	45	35	80	70	73	28	85
Consumo (l)	1,33	5,45	1,66	3,46	2,92	6,15	4,11	5,00	2,95	6,54

Resolução:

- a) Plote os pontos da tabela em um gráfico x, y.
- b) Após, com auxílio de uma regra, tente ajustar uma reta entre seus pontos.



c) Para calcular as estimativas b_0 e b_1 , montaremos uma tabela para facilitar os cálculos:

Percurso (km) - x_i	Consumo (l) - y_i	$x_i y_i$	x_i^2
20	1,33	26,6	400
60	5,45	327	3.600
15	1,66	24,9	225
45	3,46	155,7	2.025
35	2,92	102,2	1.225
80	6,15	492	6.400
70	4,11	287,7	4.900
73	5	365	5.329
28	2,95	82,6	784
85	6,54	555,9	7.225
$\Sigma 511$	$\Sigma 39,57$	$\Sigma 2.419,60$	$\Sigma 32.113$

Σx_i

Σy_i

$\Sigma x_i y_i$

Σx_i^2

$$b_1 = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sum x_i^2 - [(\sum x_i)^2]/n} = \frac{2.419,60 - [(511 \times 39,57)]/10}{32.113 - [(511)^2]/10} = 0,066$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{511}{10} = 51,1 \quad \bar{y} = \frac{\sum y_i}{n} = \frac{39,57}{10} = 3,96$$

$$b_0 = \bar{y} - b_1 \bar{x} = 3,96 - 0,0066 \times 51,1 = 0,577$$

$$\hat{y} = b_0 + b_1 x = 0,577 + 0,066 x$$

Equação de regressão $\longrightarrow \hat{y} = 0,577 + 0,066 x$

d) Quando analisamos a equação de regressão, verificamos que o consumo de combustível (l), representado por \hat{y} , depende da distância percorrida pelo carro, representada pela variável x. Observamos que à medida que a distância percorrida aumenta, o consumo de combustível (l) também aumenta. A equação linear faz uma previsão dos valores de consumo em função da distância percorrida pelo carro. Com base nessa previsão, temos evidências suficientes para dizer que com o aumento da distância percorrida, aumenta também o consumo do carro.

Exemplo 2

2) A tabela a seguir relaciona os pesos de carros (t) e as taxas de consumo de combustível em rodovias brasileiras (km/l), para uma amostra de carros de passeio novos. Com base nos resultados:

- a)** Construa um diagrama de dispersão para esses dados.
- b)** Tente aproximar a relação entre x e y traçando uma linha reta através dos pontos.
- c)** Determine uma equação de regressão estimada calculando os valores de b_0 e b_1 .
- d)** O que você pode concluir a respeito da taxa de consumo?

Obs.: taxa de consumo é expressa em quilômetros por litro (km/l).

Peso (t)	1,32	1,59	1,27	1,99	1,13	1,54	1,36	1,50	1,27	1,09
Consumo (Km/l)	13,18	11,45	12,33	10,63	13,18	12,33	11,90	11,90	11,90	14,00

Resolução:

a) Plote os pontos da tabela em um gráfico x, y.

b) Após, com auxílio de uma regra, tente ajustar uma reta entre seus pontos.



c) Para calcular as estimativas b_0 e b_1 , montaremos uma tabela para facilitar os cálculos:

Pesos (t) - x_i	Consumo (km/l) - y_i	$x_i y_i$	x_i^2
1,32	13,18	17,40	1,74
1,59	11,45	18,21	2,53
1,27	12,33	15,66	1,61
1,99	10,63	21,15	3,96
1,13	13,18	14,89	1,28
1,54	12,33	18,99	2,37
1,36	11,90	16,18	1,85
1,50	11,90	17,85	2,25
1,27	11,90	15,11	1,61
1,09	14,00	14,26	1,18
$\Sigma 14,06$	$\Sigma 122,80$	$\Sigma 170,70$	$\Sigma 20,39$

Σx_i

Σy_i

$\Sigma x_i y_i$

Σx_i^2

$$b_1 = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sum x_i^2 - [(\sum x_i)^2]/n} = \frac{170,70 - [(14,06 \times 122,8)]/10}{20,39 - [(14,06)^2]/10} = -3,15$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{14,06}{10} = 1,41 \quad \bar{y} = \frac{\sum y_i}{n} = \frac{122,8}{10} = 12,28$$

$$b_0 = \bar{y} - b_1 \bar{x} = 12,28 - (-3,15) \times 1,41 = 16,72$$

$$\hat{y} = b_0 + b_1 x = 16,72 - 3,15x$$

Equação de regressão $\Rightarrow \hat{y} = 16,72 - 3,15x$

d) Quando analisamos a equação de regressão, verificamos que a taxa de consumo de combustível, representada por \hat{y} , depende do peso do carro, representado pela variável x . Observamos que à medida que o peso do carro aumenta, a taxa de consumo de combustível em km/l diminui. A equação linear faz uma previsão dos valores de taxas de consumo em função do peso do carro. Com base nessa previsão, podemos afirmar que com o aumento do peso do carro a taxa de consumo em km/l diminui.

9.3 Análise de correlação linear

A análise de correlação linear² permite determinar o grau de relação entre duas variáveis. Assim, poderíamos determinar o grau de relacionamento entre o peso e a altura de um grupo de pessoas;

² Vídeo sobre análise de correlação, gravado pelas autoras deste livro, você pode assistir no Youtube no Canal da Prof^a Suzi Samá.

entre o tabagismo e doenças do coração. Neste livro, a seguir, estudaremos o coeficiente de correlação linear.

Para avaliar o grau de correlação linear entre duas variáveis, ou seja, medir o grau de ajustamento dos valores em torno de uma reta, usaremos o coeficiente de correlação de Pearson, que é dado por:

$$r_{xy} = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \cdot \sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}}$$

onde **n** é o número de observações.

9.3.1 Interpretação do coeficiente de correlação

O valor de r_{xy} , que sempre pertencerá ao intervalo $[-1, 1]$, representa uma medida de intensidade do inter-relacionamento de duas variáveis. Se $r_{xy} = 1$, há uma perfeita correlação positiva entre as variáveis, isto é, se os valores de uma variável aumentam (ou diminuem), em correspondência os valores da outra variável também aumentam (ou diminuem) na mesma proporção (Figura 1a). A relação entre peso e altura de uma pessoa pode ser um exemplo de correlação positiva entre as duas variáveis, pois quanto mais alta uma pessoa, maior seu peso. No entanto esta correlação não é perfeita, pois existem pessoas com estatura mais baixa e com sobrepeso, assim como

há pessoas altas e com baixo peso.

Se, por outro lado, $r_{xy} = -1$, há uma perfeita correlação negativa entre as variáveis, ou seja, os valores de uma variável variam em proporção inversa aos valores de outra variável (Figura 1b). Como exemplo podemos citar a relação entre a pressão atmosférica e a temperatura do ar: quanto maior a pressão, menor a temperatura. Se, entretanto, $r_{xy} = 0$, não há correlação entre as variáveis. Neste caso, o comportamento de uma variável não tem relação com o comportamento da outra variável, o que acaba gerando uma nuvem de pontos aleatório, como podemos observar na Figura 1c.

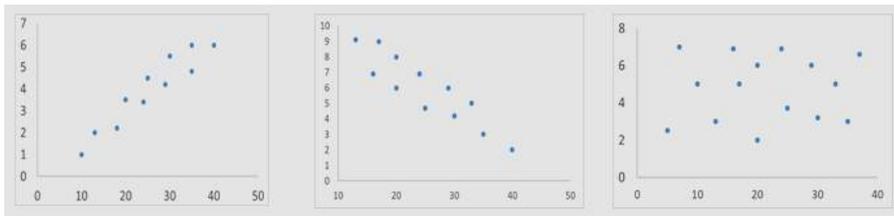


Figura 1a- Gráfico de dispersão

Figura 1b- Gráfico de dispersão

Figura 1c- Gráfico de dispersão

Como exemplo de correlação positiva podemos citar o peso e a altura de uma pessoa e de correlação negativa a pressão atmosférica e a temperatura do ar.

Quando se constata correlações entre as variáveis, podem ocorrer resultados interessantes e úteis. Não podemos afirmar a existência de correlações entre variáveis sem antes fazermos um estudo. Como r_{xy} é calculado com base em dados amostrais, é uma estatística amostral usada para medir o grau

de correlação linear entre x e y . Se tivéssemos todos os pares (x,y) , o r_{xy} seria um parâmetro populacional e representado pela letra grega ρ .

Exemplo 3:

Usando os dados do exemplo 1, vamos calcular o coeficiente de correlação. Ainda precisaremos calcular para completar os dados o Σy_i^2 , o qual não foi necessário calcular no exemplo 1.

y^2
1,77
29,70
2,76
11,97
8,53
37,82
16,89
25,00
8,70
42,77
Σ 185,91

Dados:

$\Sigma x_i =$	$\Sigma y_i =$	$\Sigma x_i y_i =$	$\Sigma x_i^2 =$	$\Sigma y_i^2 =$
511	39,57	2.419,60	32.113	185,91

$$r_{xy} = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \cdot \sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}} =$$

$$r_{xy} = \frac{2.419,60 - [(511 \times 39,57)]/10}{\sqrt{32.113 - \frac{(511)^2}{10}} \sqrt{185,91 - \frac{(39,57)^2}{10}}} =$$

$$r_{xy} = 0,948$$

O valor de r_{xy} deve estar sempre entre -1 e 1, como mostra o cálculo.

Interpretação:

A correlação é positiva, o que significa que quanto maior a variável x, maior a variável y, ou seja, quanto maior a distância percorrida pelo carro, maior o seu consumo.

Exemplo 4:

Usando os dados do exemplo 2 vamos calcular o coeficiente de correlação. Ainda precisaremos calcular para completar os dados o $\sum y_i^2$, o qual não foi necessário calcular no exemplo 2.

y^2
400
3.600
225
2.025
1.225
6.400
4.900
5.329
784
7.225
Σ 32.113

Σy_i^2

Dados:

$\Sigma x_i = 14,06$	$\Sigma y_i = 122,80$	$\Sigma x_i y_i = 170,70$	$\Sigma x_i^2 = 20,39$	$\Sigma y_i^2 = 1.516,41$
----------------------	-----------------------	---------------------------	------------------------	---------------------------

$$r_{xy} = \frac{\Sigma x_i y_i - [(\Sigma x_i \Sigma y_i)]/n}{\sqrt{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} \sqrt{\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}}} =$$
$$r_{xy} = \frac{170,70 - [(14,06 \times 122,80)]/10}{\sqrt{20,39 - \frac{(14,06)^2}{10}} \sqrt{1.516,41 - \frac{(122,80)^2}{10}}} =$$
$$r_{xy} = -0,855$$

O valor de r_{xy} deve estar sempre entre -1 e 1, como mostra o cálculo.

Interpretação:

A correlação é negativa, o que significa que quanto maior a variável x, menor a variável y, ou seja, quanto maior o peso do carro, menor a sua taxa de consumo, ou seja, menor a quilometragem percorrida por litro.

9.4 Coeficiente de determinação

O **coeficiente de determinação ou r quadrado (r^2)** é o quadrado do coeficiente de correlação, o qual nos dá a porcentagem da variação de y que pode ser explicada pela variação da variável independente x.

$$r \text{ Quadrado} = (r_{xy})^2$$

Exemplo 5:

Usando o coeficiente de correlação calculado no exemplo 3, vamos calcular o coeficiente de determinação para as distâncias percorridas por carros novos (km) e o consumo de combustível em rodovias brasileiras (l).

$$r \text{ Quadrado} = (0,948)^2 = 0,899$$

Interpretação:

O r quadrado é igual a 0,899, o que quer dizer que 89% da variação de y pode ser explicada pela variação da variável independente x. Neste caso, 89,9% da variação do consumo pode ser explicada pela variação da distância percorrida pelos carros.

Exemplo 6:

Usando o coeficiente de correlação calculado no exemplo 4, vamos calcular o coeficiente de determinação para os pesos de carros (t) e as taxas de consumo de combustível em rodovias brasileiras (km/l).

$$r \text{ Quadrado} = (0,855)^2 = 0,731$$

Interpretação:

O r quadrado é igual a 0,731, o que quer dizer que 73% da variação de y pode ser explicada pela variação da variável independente x. Neste caso, 73,1% da variação da taxa de

consumo (km/l) pode ser explicada pela variação no peso dos carros (t).

Exercícios resolvidos

1) Para divulgar a sua imagem, a empresa TCNY S.A. (nome fictício) investe em publicidade nos veículos de comunicação como TVs e jornais. Num determinado período o departamento financeiro analisa os relatórios sobre despesas com publicidade e receita. Considerando os dados fornecidos a seguir, plote o diagrama de dispersão, determine a relação funcional entre receitas e despesas com publicidade (determinar a equação de regressão), calcule o coeficiente de correlação e o r quadrado. Interprete os resultados.

Receita (R\$)	13,5	15,8	11	19	9	10,5	15	10
Despesas com publicidade (R\$1000,00)	3	5	2	6	1,5	2,5	4	1

Solução

Despesas c/ publicidade (R\$ 1000,00)	Receita(R\$)	$x_i y_i$	x_i^2	y_i^2
3	13,50	40,50	9,00	182,25
5	15,80	79,00	25,00	249,64
2	11,00	22,00	4,00	121,00
6	19,00	114,00	36,00	361,00
1,5	9,00	13,50	2,25	81,00
2,5	10,50	26,25	6,25	110,25
4	15,00	60,00	16,00	225,00
1	10,00	10,00	1,00	100,00
$\Sigma 25$	$\Sigma 103,80$	$\Sigma 365,25$	$\Sigma 99,50$	$\Sigma 1.430,14$



$$b_1 = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sum x_i^2 - [(\sum x_i)^2]/n} = \frac{365,25 - [(25 \times 103,80)]/8}{99,50 - [(25)^2]/8} = 1,91$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{25}{8} = 3,125 \quad \bar{y} = \frac{\sum y_i}{n} = \frac{103,8}{8} = 12,97$$

$$b_0 = \bar{y} - b_1 \bar{x} = 12,97 - (1,91) \times 3,125 = 7,006 \cong 7$$

$$\hat{y} = b_0 + b_1 x = 7 + 1,91x$$

Equação de regressão \Rightarrow $\hat{y} = 7 + 1,91x$

$$r_{xy} = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \cdot \sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}} =$$

$$r_{xy} = \frac{365,25 - [(25 \times 103,80)]/8}{\sqrt{99,50 - \frac{(25)^2}{8}} \cdot \sqrt{1.430,14 - \frac{(103,80)^2}{8}}} =$$

$$r_{xy} = 0,968$$

$$r \text{ Quadrado} = (0,968)^2 = 0,937$$

Interpretação:

A correlação é positiva, o que significa que quanto maior o investimento em publicidade, maior a receita.

O r quadrado é igual a 0,937, o que quer dizer que 93% da variação de y pode ser explicada pela variação da variável independente x .

2) Os dados a seguir mostram a velocidade de corte em m/min (metros por minuto) e a vida útil da ferramenta em minutos, em um determinado processo de usinagem, para uma amostra escolhida aleatoriamente de 16 ferramentas.

Velocidade de corte (m/min)	Vida útil (min)
5	41
6	43
8	35
7	32
10	22
8,5	35
8	29
15	18
13	21
20	13
17	18
19	20
18,5	15
25	11
30	06
24	10

a) Qual é a variável dependente?

b) Construa um diagrama de dispersão para esses dados.

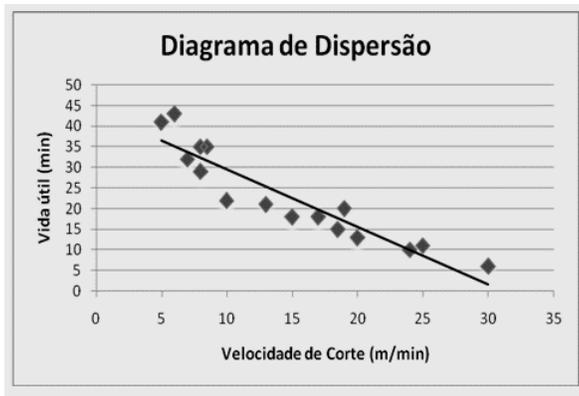
c) Tente aproximar a relação entre x e y traçando uma linha reta através dos pontos.

d) Determine uma equação de regressão estimada calculando os valores de b_0 e b_1 .

e) Calcule o coeficiente de correlação e o coeficiente de determinação.

Resolução:

- a) A variável dependente é vida útil.
- b) Plote os pontos da tabela em um gráfico x, y.
- c) Após, com auxílio de uma regra, tente ajustar uma reta entre seus pontos.



d) Calculando as estimativas b_0 e b_1 (para facilitar os cálculos você pode montar uma tabela como apresentado na próxima página).

$$b_1 = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sum x_i^2 - [(\sum x_i)^2]/n} = \frac{4.178 - [(234 \times 369)]/16}{4.297,5 - [(234)^2]/16} = -1,39$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{234}{16} = 14,625 \quad \bar{y} = \frac{\sum y_i}{n} = \frac{369}{16} = 23,06$$

$$b_0 = \bar{y} - b_1 \bar{x} = 23,06 - (-1,39) \times 14,625 = 43,42$$

$$\hat{y} = b_0 + b_1 x = 43,42 - 1,39x$$

Equação de regressão $\Rightarrow \hat{y} = 43,42 - 1,39x$

x	y	xy	x ²	y ²
5	41	205	25	1.681
6	43	258	36	1.849
8	35	280	64	1.225
7	32	224	49	1.024
10	22	220	100	484
8,5	35	297,5	72,25	1.225
8	29	232	64	841
15	18	270	225	324
13	21	273	169	441
20	13	260	400	169
17	18	306	289	324
19	20	380	361	400
18,5	15	277,5	342,25	225
25	11	275	625	121
30	06	180	900	36
24	10	240	576	100
234	369	4.178	4.297,5	10.469

Σx_i

Σy_i

Σx_iy_i

Σx_i²

Σy_i²

$$r_{xy} = \frac{\sum x_i y_i - [(\sum x_i \sum y_i)]/n}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \cdot \sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}} =$$

$$r_{xy} = \frac{4.178 - [(234 \times 369)]/16}{\sqrt{4.297,5 - \frac{(234)^2}{16}} \cdot \sqrt{10.469 - \frac{(369)^2}{16}}} =$$

$$r_{xy} = -0,930$$

O valor de r_{xy} deve estar sempre entre -1 e 1, como mostra o cálculo.

e) $r \text{ Quadrado} = (-0,930)^2 = 0,866$

Exercícios complementares

1) Uma pesquisa realizada com 26 homens, com idades entre 30 e 50 anos, anotou os pesos em kg e as leituras de pressão arterial de cada indivíduo – os dados estão na tabela a seguir. Qual é a variável dependente? Determine a equação de regressão linear que descreve a relação entre as variáveis peso e pressão arterial. Calcule o coeficiente de correlação e de determinação. Usando um nível de significância de 5%, escreva um texto sucinto relatando sua conclusão a respeito da relação entre as variáveis. Obs: Pressão arterial sistólica é o maior valor verificado durante a aferição da pressão arterial.

Indivíduo	Peso (kg)	Pressão arterial sistólica
1	74,844	130
2	75,751	133
3	81,648	150
4	70,308	128
5	96,163	151
6	79,38	146
7	86,184	150
8	95,256	140
9	90,72	148
10	67,586	125
11	71,669	133
12	76,658	135
13	77,112	150
14	78,019	153
15	72,122	128
16	76,205	132
17	78,926	149
18	83,009	158
19	97,524	150
20	88,452	163
21	81,648	156
22	64,865	124
23	108,864	170
24	106,596	165
25	87,091	160
26	84,823	159

2) Um preparador físico quer estudar a relação entre as variáveis peso e altura de jogadores de basquete do clube no qual trabalha. Para tanto, anotou o peso e a altura de cinco jogadores. Os dados estão na tabela a seguir.

Altura (cm)	Peso (kg)
185,72	79,9
175,56	75,0
170,48	72,2
179,10	77,7
182,64	78,6

- Qual é a variável dependente?
- Determine uma equação de regressão estimada calculando os valores de b_0 e b_1 .
- Calcule o coeficiente de correlação e determinação.
- O que o preparador físico pode concluir?

3) A tabela a seguir mostra idade de motociclistas e média de número de acidentes ocorridos por faixa etária, no ano de 2007, obtidos em seis meses de pesquisas. Foram entrevistados 1000 motociclistas entre 22 e 34 anos. Obs: São contados desde pequenos tombos até acidentes mais graves.

Idades	Número de acidentes
22	6
24	6
26	5
28	4
30	4
32	3
34	2

- a) Qual é a variável dependente?
- b) Construa o diagrama de dispersão para esses dados.
- c) Tente aproximar a relação entre x e y traçando uma linha reta através dos pontos.
- d) Determine uma equação de regressão estimada calculando os valores de b_0 e b_1 .
- e) Calcule o coeficiente de correlação e o coeficiente de determinação.
- f) O que você pode concluir a respeito da relação de dependência entre as duas variáveis?

4) Uma pesquisa realizada com seis famílias de quatro pessoas, cuja renda varia entre R\$1.500,00 e R\$3.000,00, mostrou os gastos médios mensais com produtos de higiene e limpeza (dados fictícios).

Salário (R\$)	1.860,00	2.540,00	1.650,00	2.200,00	2.900,00	1.700,00
Gastos (R\$)	180,00	225,00	135,00	200,00	250,00	150,00

- a) Qual a variável dependente?
- b) Construa o diagrama de dispersão para esses dados.
- c) Determine a equação de regressão.
- d) Calcule o coeficiente de correlação e o coeficiente de determinação.

5) Determine a equação de regressão linear que descreva a relação entre a frequência de acidentes em uma determinada empresa que atua na orla portuária e o número de horas/aula

preventivas (educacionais) ministradas aos trabalhadores da empresa. Determine os coeficientes de correlação e de determinação. Qual sua conclusão? Obs: São contabilizados acidentes leves, médios e graves (dados fictícios).

Número de acidentes por/ano	24	22	19	15	12
Número de horas/aula por/ano	3	5	7	8	10

6) Como resultado do crescente desenvolvimento de uma cidade, surgiram vários estabelecimentos comerciais na periferia, e com isso vários estabelecimentos comerciais do centro da cidade estão sofrendo financeiramente. O setor de propaganda de um desses estabelecimentos acha que o aumento de publicidade poderia ajudar a atrair mais compradores. Para estudar o efeito da publicidade nas vendas, um comerciante do centro registrou os gastos com propaganda nos meses de maio a setembro e os valores de vendas no mesmo período (dados fictícios).

Volume de vendas (R\$)	60.000,00	73.000,00	76.000,00	90.000,00	105.000,00
Despesas de publicidade (R\$)	5.000,00	7.000,00	10.500,00	13.000,00	18.000,00

- a)** Construa um diagrama de dispersão para esses dados.
- b)** Tente aproximar a relação entre x e y traçando uma linha reta através dos pontos.
- c)** Com os dados a seguir fornecidos, monte a equação de regressão. Qual a sua conclusão?

$$\sum x_i = 53.500,00$$

$$\sum y_i = 404.000,00$$

$$\sum x_i y_i = 4.669.000.000,00$$

$$\sum x^2 = 677.250.000,00$$

$$\sum y^2 = 33.830.000.000,00$$

$$r_{xy} = 0,982$$

Respostas

1. Variável dependente pressão arterial;
 $\hat{y} = 69,104 - 0,9246x$; $r_{xy} = 0,773$, r Quadrado = 0,597

2. Variável dependente peso; $\hat{y} = 14,53 + 0,51x$;
 $r_{xy} = 0,988$; r Quadrado = 0,976

3. Variável dependente acidentes; $\hat{y} = 13,78 - 0,339x$;
 $r_{xy} = -0,979$; r Quadrado = 0,958

4. Gastos $\hat{y} = 6,24 + 0,0858x$; $r_{xy} = 0,978$;
r Quadrado = 0,956

5. $\hat{y} = 16,48 - 0,537x$; $r_{xy} = -0,979$; r Quadrado = 0,958

6. $\hat{y} = 45.453,24 + 3,303x$

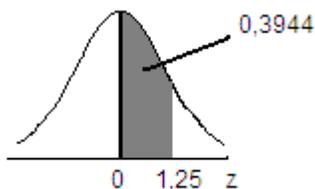
Anexo I – Uso da Tabela da Distribuição Normal Padronizada Z

Para usar a tabela z, da distribuição normal padronizada, deve-se usar o fato de que a curva é simétrica e centrada na média. O corpo da tabela é constituído das probabilidades (área sob a curva entre os limites de zero a z). Os valores de z estão nas margens da tabela, na primeira coluna está o valor inteiro e a primeira casa decimal, na primeira linha está a segunda casa decimal. Por exemplo, o valor de $z=1,25$ é obtido pela intersecção da linha que contém o valor 1,2 e a coluna que contém a segunda casa decimal, 5, do valor de z procurado.

Segunda casa decimal

z	0	1	2	3	4	5	6	7
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157
0,6	0,2258	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486
0,7				0,2673	0,2704	0,2734	0,2764	0,2794
0,8	Valor inteiro e a primeira casa decimal			0,2967	Probabilidade entre a média zero e o valor $z = 1,25$			0,3078
0,9	0,3078	0,3186	0,3212	0,3238				0,3340
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418

A área abaixo da curva entre zero e o valor de z igual a 1,25 é de 0,3944. Como mostrado na tabela acima:



Exemplo de uso da tabela:

Dados os valores de z , determine a probabilidade, ou seja, a área abaixo da curva:

a) $P(0 \leq z \leq 2)$

Estamos procurando a probabilidade de um valor z estar entre zero e dois. A tabela fornece a área entre a média de z (zero) e o valor de z procurado. Obtemos esse valor, procurando na primeira coluna da tabela o valor de $z = 2,0$ e na primeira linha o valor "0" que é a segunda casa decimal. O valor da área é 0,4772.

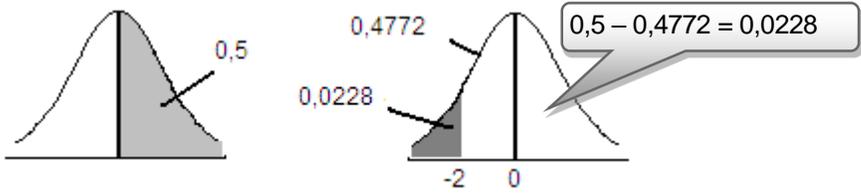


$$P(0 \leq z \leq 2) = 0,4772$$

b) $P(z \geq 2)$

Como a curva é simétrica e a área abaixo da curva é igual a um, cada metade vale 0,5. Estamos procurando a probabilidade de um valor z ser maior ou igual a 2. A tabela fornece a área entre

a média de z (zero) e o valor de z procurado. Portanto, é necessário subtrair o valor encontrado na tabela de 0,5.



$$P(z \geq 2) = 0,5 - 0,4772 = 0,0228$$

c) $P(z \leq -2)$

Como a curva é simétrica à área entre $z=0$ e $z=2$ é a mesma entre $z= -2$ e $z=0$. Com raciocínio análogo ao item anterior, temos:

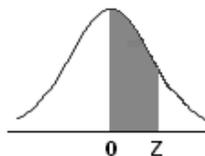
$$P(z \leq -2) = 0,5 - 0,4772$$

$$= 0,0228$$

$$P(z \leq -2) = \mathbf{0,0228}$$



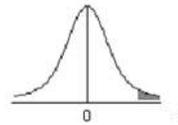
Anexo II – Tabela da Distribuição Normal Padrão



z	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	0,0636	0,0675	0,0714	0,0754
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	0,2258	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0,7	0,2580	0,2612	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2996	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974

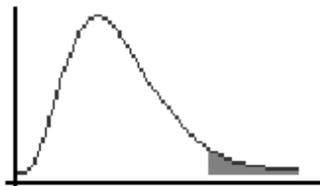
z	0	1	2	3	4	5	6	7	8	9
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4995	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

Anexo III – Distribuição “ t ” de Student



g.l.	Área da extremidade da curva				
	0,1	0,05	0,025	0,01	0,005
1	3,0777	6,3137	12,7062	31,8210	63,6559
2	1,8856	2,9200	4,3027	6,9645	9,9250
3	1,6377	2,3534	3,1824	4,5407	5,8408
4	1,5332	2,1318	2,7765	3,7469	4,6041
5	1,4759	2,0150	2,5706	3,3649	4,0321
6	1,4398	1,9432	2,4469	3,1427	3,7074
7	1,4149	1,8946	2,3646	2,9979	3,4995
8	1,3968	1,8595	2,3060	2,8965	3,3554
9	1,3830	1,8331	2,2622	2,8214	3,2498
10	1,3722	1,8125	2,2281	2,7638	3,1693
11	1,3634	1,7959	2,2010	2,7181	3,1058
12	1,3562	1,7823	2,1788	2,6810	3,0545
13	1,3502	1,7709	2,1604	2,6503	3,0123
14	1,3450	1,7613	2,1448	2,6245	2,9768
15	1,3406	1,7531	2,1315	2,6025	2,9467
16	1,3368	1,7459	2,1199	2,5835	2,9208
17	1,3334	1,7396	2,1098	2,5669	2,8982
18	1,3304	1,7341	2,1009	2,5524	2,8784
19	1,3277	1,7291	2,0930	2,5395	2,8609
20	1,3253	1,7247	2,0860	2,5280	2,8453
21	1,3232	1,7207	2,0796	2,5176	2,8314
22	1,3212	1,7171	2,0739	2,5083	2,8188
23	1,3195	1,7139	2,0687	2,4999	2,8073
24	1,3178	1,7109	2,0639	2,4922	2,7970
25	1,3163	1,7081	2,0595	2,4851	2,7874
26	1,3150	1,7056	2,0555	2,4786	2,7787
27	1,3137	1,7033	2,0518	2,4727	2,7707
28	1,3125	1,7011	2,0484	2,4671	2,7633
29	1,3114	1,6991	2,0452	2,4620	2,7564
30	1,3104	1,6973	2,0423	2,4573	2,7500
35	1,3062	1,6896	2,0301	2,4377	2,7238
40	1,3031	1,6839	2,0211	2,4233	2,7045
45	1,3007	1,6794	2,0141	2,4121	2,6896
50	1,2987	1,6759	2,0086	2,4033	2,6778
60	1,2958	1,6706	2,0003	2,3901	2,6603
70	1,2938	1,6669	1,9944	2,3808	2,6479
80	1,2922	1,6641	1,9901	2,3739	2,6387
90	1,2910	1,6620	1,9867	2,3685	2,6316
100	1,2901	1,6602	1,9840	2,3642	2,6259
1000	1,2824	1,6464	1,9623	2,3301	2,5807

Anexo IV – Distribuição Qui-quadrado χ^2



Área à direita da curva														
g.l.	0,995	0,99	0,975	0,95	0,9	0,8	0,75	0,25	0,2	0,1	0,05	0,025	0,01	0,005
1	0,000	0,0002	0,0010	0,0039	0,0158	0,0642	0,1015	1,3233	1,6424	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201	0,0506	0,1026	0,2107	0,4463	0,5754	2,7726	3,2189	4,6052	5,9915	7,3778	9,2104	10,596
3	0,0717	0,1148	0,2158	0,3518	0,5844	1,0052	1,2125	4,1083	4,6416	6,2514	7,8147	9,3484	11,345	12,838
4	0,2070	0,2971	0,4844	0,7107	1,0636	1,6488	1,9226	5,3853	5,9886	7,7794	9,4877	11,143	13,277	14,860
5	0,4118	0,5543	0,8312	1,1455	1,6103	2,3425	2,6746	6,6257	7,2893	9,2363	11,071	12,832	15,086	16,750
6	0,6757	0,8721	1,2373	1,6354	2,2041	3,0701	3,4546	7,8408	8,5581	10,645	12,592	14,450	16,812	18,548
7	0,9893	1,2390	1,6899	2,1673	2,8331	3,8223	4,2549	9,0371	9,8032	12,017	14,067	16,013	18,475	20,278
8	1,3444	1,6465	2,1797	2,7326	3,4895	4,5936	5,0706	10,219	11,030	13,362	15,507	17,534	20,090	21,955
9	1,7349	2,0879	2,7004	3,3251	4,1682	5,3801	5,8988	11,389	12,242	14,684	16,919	19,023	21,666	23,589
10	2,1558	2,5582	3,2470	3,9403	4,8652	6,1791	6,7372	12,549	13,442	15,987	18,307	20,483	23,209	25,188
11	2,6032	3,0535	3,8157	4,5748	5,5778	6,9887	7,5841	13,701	14,631	17,275	19,675	21,920	24,725	26,757
12	3,0738	3,5706	4,4038	5,2260	6,3038	7,8073	8,4384	14,845	15,812	18,549	21,026	23,337	26,217	28,299
13	3,5650	4,1069	5,0087	5,8919	7,0415	8,6339	9,2991	15,984	16,985	19,812	22,362	24,736	27,688	29,819
14	4,0747	4,6604	5,6287	6,5706	7,7895	9,4673	10,165	17,117	18,151	21,064	23,685	26,119	29,141	31,319
15	4,6009	5,2294	6,2621	7,2609	8,5468	10,307	11,036	18,245	19,311	22,307	24,996	27,488	30,578	32,801
16	5,1422	5,8122	6,9077	7,9616	9,3122	11,152	11,912	19,369	20,465	23,542	26,296	28,845	31,999	34,267
17	5,6973	6,408	7,5642	8,6718	10,085	12,002	12,792	20,489	21,615	24,769	27,587	30,191	33,409	35,718
18	6,2648	7,0149	8,2307	9,3904	10,865	12,857	13,675	21,605	22,60	25,989	28,869	31,526	34,805	37,156
19	6,8439	7,6327	8,9065	10,117	11,651	13,716	14,562	22,718	23,900	27,204	30,143	32,852	36,191	38,582
20	7,4338	8,2604	9,5908	10,851	12,443	14,578	15,452	23,828	25,037	28,412	31,410	34,170	37,566	39,997
21	8,0336	8,8972	10,283	11,591	13,240	15,445	16,344	24,935	26,171	29,616	32,671	35,479	38,932	41,401
22	8,6427	9,5425	10,982	12,338	14,041	16,314	17,240	26,039	27,301	30,813	33,924	36,781	40,289	42,796
23	9,2604	10,196	11,689	13,091	14,848	17,187	18,137	27,141	28,429	32,007	35,173	38,076	41,638	44,181
24	9,8862	10,856	12,401	13,848	15,659	18,062	19,037	28,241	29,553	33,196	36,415	39,364	42,980	45,558
25	10,519	11,524	13,119	14,611	16,473	18,940	19,939	29,339	30,675	34,382	37,653	40,647	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	19,820	20,843	30,435	31,795	35,563	38,885	41,923	45,642	48,289
27	11,808	12,879	14,573	16,151	18,114	20,703	21,749	31,528	32,912	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	21,588	22,657	32,621	34,027	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	22,475	23,566	33,711	35,139	39,087	42,557	45,722	49,588	52,335
30	13,787	14,954	16,791	18,493	20,599	23,364	24,478	34,799	36,250	40,256	43,773	46,979	50,892	53,672
35	17,192	18,509	20,569	22,465	24,797	27,836	29,054	40,223	41,778	46,059	49,802	53,203	57,342	60,275
40	20,707	22,164	24,433	26,509	29,051	32,345	33,660	45,616	47,269	51,805	55,759	59,342	63,691	66,766

EDITORA E GRÁFICA DA FURG
CAMPUS CARREIROS
CEP 96203 900
editora@furg.br

Carla Silva da Silva é professora da Escola de Engenharia da Universidade Federal do Rio Grande (EE/FURG). Bacharel em Engenharia Civil, mestre em Engenharia Oceânica pela FURG e doutora em Engenharia Civil pela UNICAMP. Suas áreas de interesse são Ensino de Estatística, Hidráulica; Mecânica, Gestão de Águas e Educação a Distância. Atuou como professora pesquisadora e formadora na modalidade a distância junto a UAB/FURG. Atualmente trabalha com os cursos de Graduação em Engenharia e Especialização em Gestão Ambiental em Municípios.

carlasilva@furg.br

Suzi Samá Pinto é professora do Instituto de Matemática, Estatística e Física da Universidade Federal do Rio Grande (IMEF/FURG); licenciada em Matemática, mestre em Engenharia Oceânica e doutoranda em Educação em Ciência pela FURG. Membro do grupo de pesquisa em Educação Estatística e do grupo de pesquisa em Educação a distância e tecnologia. Atuou como professora pesquisadora e tutora a distância junto a Secretaria de Educação a Distância da FURG.

suzisama@furg.br

ISBN 978-65-5754-047-3



9 786557 540473