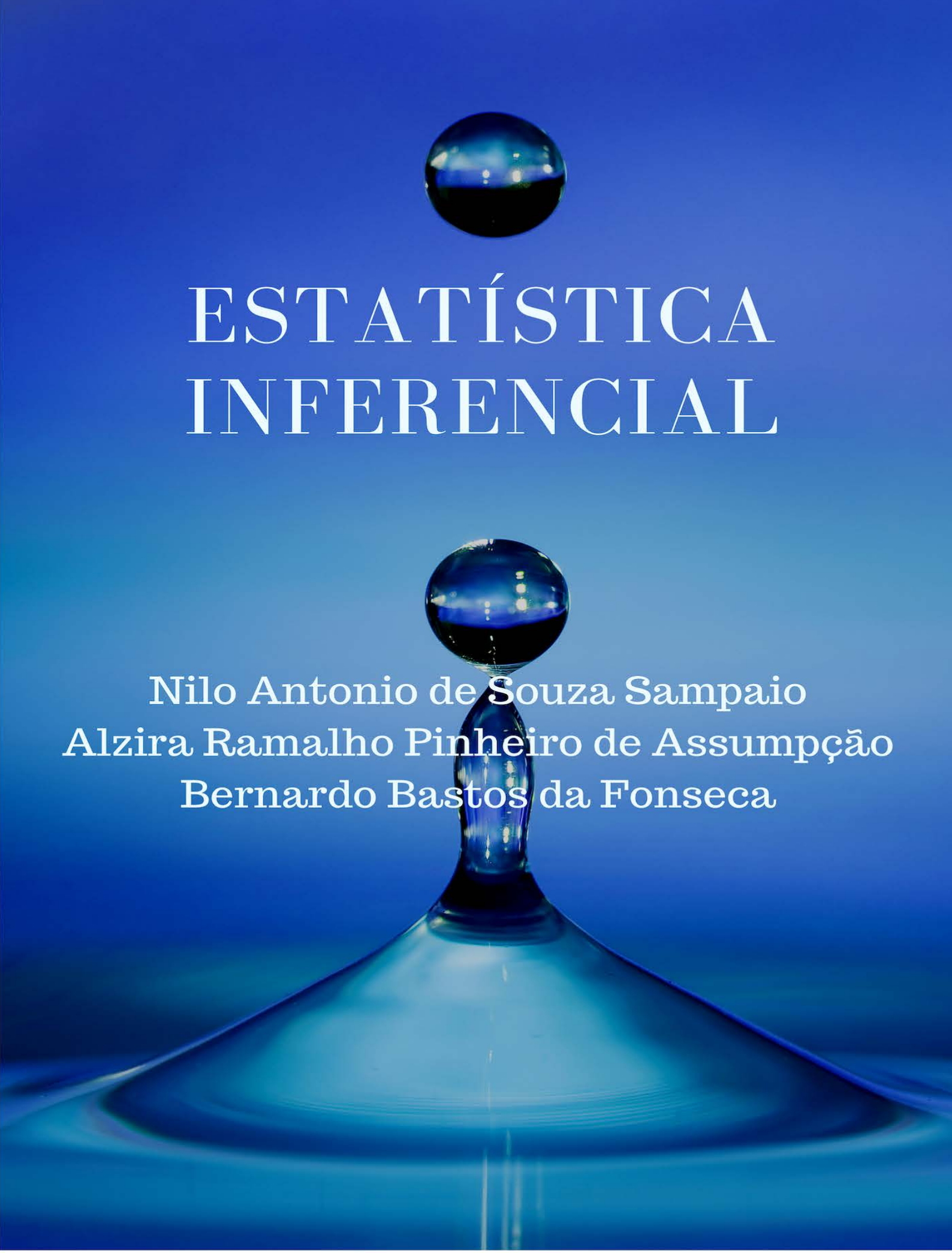


ESTATÍSTICA INFERENCIAL

Nilo Antonio de Souza Sampaio
Alzira Ramalho Pinheiro de Assumpção
Bernardo Bastos da Fonseca

Editora Poisson

Nilo Antônio de Souza Sampaio
Alzira Ramalho Pinheiro de Assumpção
Bernardo Bastos da Fonseca

Estatística Inferencial

1ª Edição

Belo Horizonte
Poisson
2018

Dados Internacionais de Catalogação na Publicação (CIP)

S192
SAMPAIO. Nilo Antônio de Souza; ASSUMPÇÃO, Alzira Ramalho Pinheiro de; FONSECA, Bernardo Bastos da – Estatística Inferencial. Belo Horizonte, Editora Poisson, 2018.
70p.

Formato: PDF
ISBN: 978-85-7042-028-2
DOI: 10.5935/978-85-7042-028-2.2018B001

Modo de acesso: World Wide Web
Inclui bibliografia

1. **Estatística 2. Métodos Quantitativos**
I. Título

CDD-519.5

www.poisson.com.br

contato@poisson.com.br

Nilo Antônio de Souza Sampaio

- Doutor em Engenharia Mecânica pela UNESP-FEG - 2011.
- Professor Adjunto de Probabilidade e Estatística UERJ-FAT, Professor Associação Educacional Dom Bosco (AEDB) e Professor Faculdade Sul Fluminense (FASF).
- Trabalha com Matemática Aplicada, Probabilidade e Estatística, trabalhando nas seguintes áreas: Pesquisa sobre Aplicações da Estatística e da Matemática em todas as Ciências e Planejamento de Experimentos

Alzira Ramalho Pinheiro de Assumpção

- Doutora em Engenharia de Produção pelo Instituto Luiz Alberto Coimbra de Pós Graduação e Pesquisa de Engenharia da Universidade Federal do Rio de Janeiro – PEP/ COPPE/UFRJ – 1996.
- Professora Associada UERJ-FAT.
- Trabalha com Pesquisa Operacional, Raciocínio Lógico, Negociação, Metodologia do Ensino Superior.

Bernardo Bastos da Fonseca

- Doutor em Engenharia de Produção pelo Instituto Luiz Alberto Coimbra de Pós-Graduação e Pesquisa de Engenharia da Universidade Federal do Rio de Janeiro - PEP/COPPE/UFRJ - 2014.
- Professor Adjunto UERJ-FAT.
- Trabalha com pesquisa na área de segurança, meio ambiente e saúde.

PREFÁCIO

Estamos colocando a disposição dos colegas professores e aos interessados em estatística de modo geral este livro de Estatística Inferencial. O conteúdo deste livro apresenta os conceitos básicos de Inferência Estatística que possibilita aos estudantes de graduação adquirir os conhecimentos essenciais sobre esta disciplina. A abordagem mais uma vez é essencialmente acadêmica, os conteúdos são acompanhados de exercícios resolvidos e propostos para ajudar na consolidação da parte conceitual.

AGRADECIMENTOS

Gostaria de Agradecer primeiramente a Deus por me possibilitar realizar este trabalho, sem seu direcionamento e proteção eu jamais teria conseguido e também a minha família que me apóia em todos os momentos da minha vida. Gostaria ainda externar a minha gratidão a Prof (a) Alzira Ramalho Pinheiro de Assumpção e ao Prof Bernardo Bastos da Fonseca pelo apoio e incentivo na realização deste trabalho.

Nilo Sampaio

SUMÁRIO

Capítulo 1

Inferência Estatística e Amostragem

1.1 DEFINIÇÃO DE INFERÊNCIA ESTATÍSTICA:

Inferência estatística é uma área da Estatística cujo objetivo é fazer afirmações a partir de um conjunto de valores representativos (amostra) sobre um universo e se assume que a amostra é muito maior do que o conjunto de dados observados. Esta afirmação deve sempre vir acompanhada de uma medida de precisão sobre sua veracidade. Para realizar este trabalho, o estatístico coleta informações de dois tipos: experimentais (as amostras) e aquelas que obtém na literatura. As duas principais escolas de inferência são a inferência frequentista (ou clássica) e a inferência bayesiana.

1.2 DEFINIÇÕES BÁSICAS

Abaixo, algumas definições utilizadas em Inferência Estatística são apresentadas:

Variável Aleatória:

- Característico numérico do resultado de um experimento.
- É a Função que associa a cada elemento do espaço amostral um número real.

População e Amostra:

- População é o conjunto de todos os elementos ou resultados de um problema que está sendo estudado.
- Amostra é qualquer subconjunto da população que contém os elementos que podem ser observados e é onde as quantidades de interesse podem ser medidas.

Parâmetros:

- Característica numérica (desconhecida) da distribuição dos elementos da população.

Estimador:

- É a Função da amostra, construída com a finalidade de representar, ou estimar um parâmetro de interesse na população.

Estimativa:

- Valor numérico que um estimador assume

Exemplo:

A distribuição da altura da população brasileira adulta pode ser representada por um modelo normal (embora as alturas não possam assumir valores negativos). Neste caso, temos como interesse estimar os parâmetros média e variância dessa distribuição.

- **Solução 1:** Medir a altura de todos os brasileiros adultos.
- **Solução 2:** Selecionar de forma aleatória algumas pessoas (amostra), analisá-las e inferir propriedades para toda a população.

1.3 TÉCNICAS DE AMOSTRAGEM

As Técnicas de Amostragem atuam no estudo de um pequeno grupo de elementos retirado de uma população que se pretende conhecer. Esses pequenos grupos retirados da população são chamados de Amostras.

Veremos a seguir as principais técnicas de amostragem, divididas em probabilísticas e não-probabilísticas:

1.3.1 TÉCNICAS PROBABILÍSTICAS (ALEATÓRIAS)

As técnicas probabilísticas garantem a possibilidade de realizar afirmações sobre a população com base nas amostras. Normalmente, todos os elementos da população possuem a mesma

probabilidade de serem selecionados. Assim, considerando N como o tamanho da população, a probabilidade de cada elemento ser selecionado será 1/N. Estas técnicas garantem o acaso na escolha.

São técnicas probabilísticas:

- Amostragem Aleatória Simples

Amostragem Aleatória Simples é o processo mais elementar e freqüentemente utilizado. Ela pode ser realizada a partir da numeração dos elementos da população de 1 a n e sorteando, por meio de um dispositivo aleatório qualquer, X números dessa seqüência, que corresponderão aos elementos pertencente à amostra.

Exemplo

Obter uma amostra representativa de 10% de uma população de 200 alunos de uma escola.

1º) Numerar os alunos de 1 a 200;

2º) Escrever os números de 1 a 200 em pedaços de papel e colocá-los em uma urna;

3º) Retirar da urna 20 pedaços de papel, um a um, formando a amostra da população.

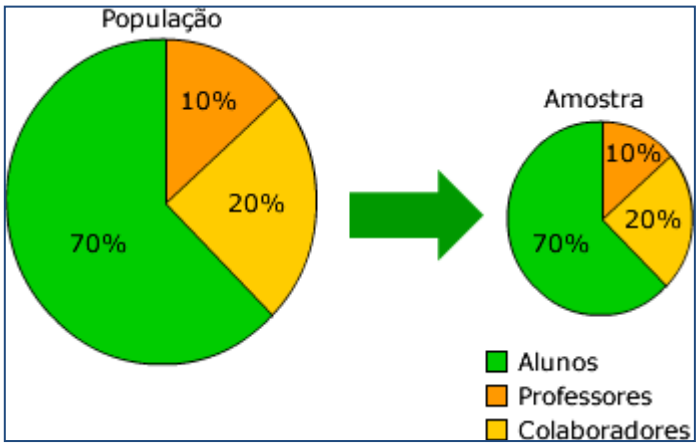
Nesta técnica de amostragem, todos os elementos da população têm a mesma probabilidade de serem selecionados: 1/N, onde N é o número de elementos da população.

- Amostragem Estratificada

Quando a população possui características que permitem a criação de subconjuntos, as amostras extraídas por amostragem simples são menos representativas. Nesse caso, a amostragem estratificada é utilizada.

Como a população se divide em subconjuntos, convém que o sorteio dos elementos leve em consideração tais divisões para que os elementos da amostra sejam proporcionais ao número de elementos desses subconjuntos. Observe a figura abaixo:

Figura 01: Relação entre população e amostra



Exemplo

Em uma população de 400 alunos, há 240 meninos e 160 meninas. Extraia uma amostra representativa de 10% dessa população.

Nesse exemplo, há uma característica que permite identificar 2 subconjuntos, a característica Sexo. Considerando essa divisão, vamos extrair a amostra da população.

Tabela-1: Relação entre População e Amostra.

SEXO	POPULAÇÃO	AMOSTRA (10%)
Masculino	240	24
Feminino	160	16
Total	400	40

Portanto, a amostra deve conter 24 alunos do sexo masculino e 16 do sexo feminino, totalizando 40 alunos, que correspondem a 10% da população.

Para selecionar os elementos da população com o objetivo de formar a amostra, podemos executar os seguintes passos:

1º) Numerar os alunos de 1 a 400, sendo os meninos numerados de 1 a 240 e as meninas, de 241 a 400;

2º) Escrever os números de 1 a 240 em pedaços de papel e colocá-los em uma urna A;

3º) Escrever os números de 241 a 400 em pedaços de papel e colocá-los em uma urna B;

4º) Retirar da urna A 24 pedaços de papel, um a um, e 16 da urna B, formando a amostra da população.

São exemplos desta técnica de amostragem as pesquisas eleitorais por região, cidades pequenas e grandes, área urbana e área rural, sexo, faixa etária, faixa de renda, etc.

- **Amostragem Sistemática**

Esta técnica de amostragem é aplicada em populações que possuem os elementos ordenados em que não há a necessidade de construir um sistema de referência. Nesta técnica, a seleção dos elementos que comporão a amostra pode ser feita por um sistema criado pelo pesquisador.

Exemplo

Obter uma amostra de 80 casas de uma rua que contém 2000 casas. Nesta técnica de amostragem, podemos realizar o seguinte procedimento:

1º) Como 2000 dividido por 80 é igual a 25, escolhemos por um método aleatório qualquer um número entre 1 e 25, o que indica o primeiro elemento selecionado para a amostra.

2º) Consideramos os demais elementos, periodicamente, de 25 em 25.

Se o número sorteado entre 1 e 25 for o número 8, a amostra será formada pelas casas: 8ª, 33ª, 58ª, 83ª, 108ª, etc.

Apesar de esta técnica ser de fácil execução, há a possibilidade de haver ciclos de variação, o que tornariam a amostra não-representativa da população.

- **Amostragem por Conglomerados**

Esta técnica é usada quando a identificação dos elementos da população é extremamente difícil. Todavia, pode ser relativamente fácil dividir a população em conglomerados (subgrupos) heterogêneos representativos da população global.

A seguir, é descrito o procedimento de execução desta técnica:

1º) Seleciona uma amostra aleatória simples dos conglomerados existentes;

2º) Realizar o estudo sobre todos os elementos do conglomerado selecionado.

São exemplos de conglomerados: quarteirões, famílias, organizações, agências, edifícios, etc.

Exemplo

Estudar a população de uma cidade, dispondo apenas do mapa dos quarteirões da cidade.

Neste caso, não temos a relação dos moradores da cidade, restando o uso dos subgrupos heterogêneos (conglomerados). Para realizar o estudo estatístico sobre a cidade, realizaremos os seguintes procedimentos:

1º) Numerar os quarteirões de 1 a n;

2º) Escrever os números de 1 a n em pedaços de papel e colocá-los em uma urna;

3º) Retirar um pedaço de papel da urna e realizar o estudo sobre os elementos do conglomerado selecionado.

1.3.2 TÉCNICAS NÃO-PROBABILÍSTICAS (NÃO-ALEATÓRIAS)

São técnicas em que há uma escolha deliberada dos elementos da população onde não permite generalizar os resultados das pesquisas para a população, pois amostras não garantem a representatividade desta.

São técnicas não-probabilísticas:

- **Amostragem Acidental**

Trata-se da formação de amostras por aqueles elementos que vão aparecendo. Este método é utilizado, geralmente, em pesquisas de opinião em que os entrevistados são acidentalmente escolhidos.

Exemplo

Pesquisas de opinião em shoppings, praças e locais públicos de grandes cidades, etc.

- **Amostragem Intencional**

De acordo com determinado critério, é escolhido intencionalmente um grupo de elementos que comporão a amostra. O pesquisador se dirige intencionalmente a grupos de elementos dos quais deseja saber a opinião.

Exemplo

Em uma pesquisa sobre preferência por determinada cerveja, o pesquisador entrevista os freqüentadores dos bares de uma cidade.

Agora que já conhecemos as principais técnicas de amostragem, vamos aprender a calcular o tamanho das amostras dos estudos estatísticos.

Antes de prosseguir, vamos definir alguns termos:

Parâmetro: Característica da população.

Estatística: Característica descritiva de elementos de uma amostra.

Estimativa: valor acusado por uma estatística que estima o valor de um parâmetro.

O cálculo do tamanho da amostra está diretamente ligado ao erro amostral tolerável.

Mas o que é erro amostral?

É a diferença entre o valor que a estatística pode acusar e o verdadeiro valor do parâmetro que se deseja estimar.

O erro amostral tolerável é a margem de erro aceitável em um estudo estatístico. Para esclarecer melhor, é quando o apresentador do telejornal, em ano de eleições, anuncia:

"O candidato A tem 42% das intenções de voto, 2 para mais, 2 para menos."

Quando o apresentador cita "2 para mais, 2 para menos", ele se refere ao erro amostral tolerável para aquela pesquisa de intenções de voto.

Tamanho da Amostra

Obs.: um passo importante antes de iniciar o cálculo do tamanho da amostra é definir qual o erro amostral tolerável para o estudo que será realizado.

Observe a seguinte fórmula:

Onde:

- n_0 é a primeira aproximação do tamanho da amostra
- E_0 é o erro amostral tolerável (Ex.: 2% = 0,02)

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

, onde:

- N é o número de elementos da população
- n é o tamanho da amostra

Observe o seguinte exemplo para compreender melhor:

Exemplo

Em uma empresa que contém 2000 colaboradores, deseja-se fazer uma pesquisa de satisfação. Quantos colaboradores devem ser entrevistados para tal estudo?

Resolução

$N = 2000$

Definindo o erro amostral tolerável em 2%

$E_0 = 0{,}02$

$$n_0 = 1/(E_0)^2$$

$$n_0 = 1/(0{,}02)^2$$

$$n_0 = 2500$$

$$n = (N\ .\ n_0)/(N + n_0)$$

$$n = (2000\ .\ 2500)/(2000 + 2500)$$

$$n = 1111 \text{ colaboradores}$$

Com o erro amostral tolerável em 2%, 1111 colaboradores devem ser entrevistados para a pesquisa.

Vamos repetir os cálculos, definindo o erro amostral tolerável em 4%.

$N = 2000$

$E_0 = 0{,}04$

$$n_0 = 1/(E_0)^2$$
$$n_0 = 1/(0{,}04)^2$$
$$n_0 = 625$$
$$n = (N \,.\, n_0)/(N + n_0)$$
$$n = (2000 \,.\, 625)/(2000 + 625)$$
$$n = 476 \text{ colaboradores}$$

Através deste segundo cálculo, é possível observar que, quando aumentamos a margem de erro, o tamanho da amostra reduz.

E se houvesse 300.000 colaboradores na empresa?

$N = 300000$

$E_0 = 0{,}04$

$$n_0 = 1/(E_0)^2$$
$$n_0 = 1/(0{,}04)^2$$
$$n_0 = 625$$
$$n = (N \,.\, n_0)/(N + n_0)$$
$$n = (300000 \,.\, 625)/(300000 + 625)$$
$$n = 623 \text{ colaboradores}$$

Observe que a diferença entre *n* e no, neste último cálculo, é muito pequena.

Portanto: se o número de elementos da população (N) é muito grande, a primeira aproximação do tamanho da amostra já é suficiente.

Observe ainda:

$N = 2000$

$E_0 = 0{,}04$

$n = 476$ colaboradores = 23,8% da população

$N = 300.000$

$E_0 = 0{,}04$

$n = 623$ colaboradores = 0,2% da população

EXERCÍCIOS – CAPÍTULO 1 - INFERÊNCIA ESTATÍSTICA E AMOSTRAGEM.

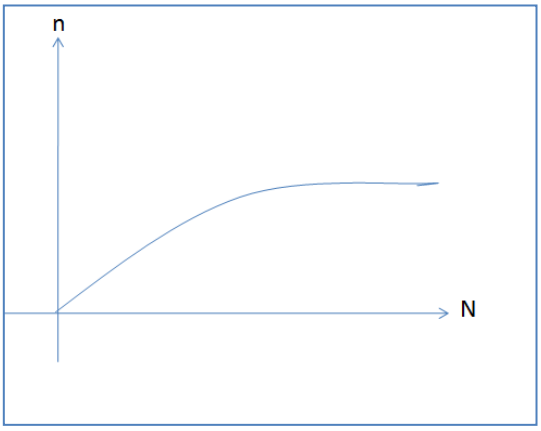
1) Exemplo: calcule o tamanho da amostra: $N = 200$ famílias $E_0 =$ erro amostral tolerável = 4% ($E_0 = 0{,}04$) $n_0 = 1/(0{,}04)\ 2 = 625$ famílias n (tamanho da amostra corrigido) = $n = 200\text{x}625/200{+}625 = 125000/825 = 152$ famílias

E se a população fosse de 200.000 famílias? $n = (200.000)\ .\ 625/\ (200.000 + 625) = 623$ famílias.

Obs.: Observe que se N é muito grande, não é necessário considerar o tamanho exato N da população. Nesse caso, o cálculo da primeira aproximação já é suficiente para o cálculo.

Tamanho da amostra: Observe que N = 200 famílias, E_0 = 4% n = 152 famílias → 76% da população. Observe que N = 200.000 famílias, E_0 = 4% n = 623 famílias -→ 0,3% da população. Logo, é errôneo pensar que o tamanho da amostra n deve ser tomado como um percentual do tamanho da população para ser representativa.

Figura-2: Tamanho da Amostra em Relação ao tamanho da População.



2) Numa pesquisa para uma eleição presidencial, qual deve ser o tamanho de uma amostra aleatória simples, se deseja garantir um erro amostral não superior a 2%?

Sol: $n = n_0 = 1/(0{,}02)2 = 1/0{,}0004 = 2500$ eleitores

3) Numa empresa com 1000 funcionários, deseja-se estimar a percentagem dos favoráveis a certo treinamento. Qual deve ser o tamanho da amostra aleatória simples que garanta um erro amostral não superior a 5%?

$N = 1000$ empregados E_0 = erro amostral tolerável = 5% ($E_0 = 0{,}05$) $n_0 = 1/(0{,}05)\ 2 = 400$ empregados $n = 1000\ .\ 400\ /\ (1000 + 400) = 286$ empregados

CAPÍTULO 2

Intervalo de Confiança

2.1 INTERVALO DE CONFIANÇA

Em estatística, o **intervalo de confiança (IC)** é um tipo de estimativa por intervalo de um parâmetro populacional desconhecido. Introduzido na estatística por Jerzy Neyman em 1937, é um intervalo observado (calculado a partir de observações) que pode variar de amostra para amostra e que, com dada frequência (nível de confiança), inclui o parâmetro de interesse real não observável.

2.2 INTERVALO DE CONFIANÇA PARA A MÉDIA

Quando queremos estimar a média de uma população através de uma amostra temos dois casos distintos a considerar: quando a variância da população é conhecida e quando ela é desconhecida. A seguir, temos os dois casos.

♦ Variância Conhecida

Consideremos uma amostra aleatória simples $X_1,\ldots,X_n$ obtida de uma população com distribuição normal, com média μ e variância σ^2 conhecida. Desta forma, a distribuição amostral da média também é Normal com média μ e variância $\frac{\sigma^2}{n}$, ou seja:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Assim, temos que a variável Z tem distribuição normal padronizada.

Os valores mais comuns para a variável Z são:

$$Z_{90\%} = 1{,}64 \text{ - } Z_{94\%} = 1{,}88 \text{ - } Z_{95\%} = 1{,}96 \text{ - } Z_{98\%} = 2{,}33 \text{ - } Z_{99\%} = 2{,}58$$

Com isso, o intervalo de confiança da média é dado por:

$$IC(\mu, 1-\alpha) = \left(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; X + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Exemplo:

O projetista de uma indústria tomou uma amostra de 36 funcionários para verificar o tempo médio gasto para montar um determinado brinquedo. Lembrando que foi verificado que $\overline{x}$ =19,9 e $\sigma = 5{,}73$, construir um intervalo de confiança de nível 95% para μ.

Solução:

Na tabela da distribuição normal padronizada, obtemos que $Z_{0{,}025} = 1{,}96$.

Substituindo $\overline{x} = 19{,}9, n = 36, \sigma = 5{,}73$ e $Z_{0{,}025} = 1{,}96$ na fórmula para o intervalo de confiança, temos

$$19{,}9 - 1{,}96\frac{5{,}73}{\sqrt{36}} \leq \mu \leq 19{,}9 + 1{,}96\frac{5{,}73}{\sqrt{36}}$$

e, portanto,

$$IC(\mu, 0,95) = (18,02; 21,77)$$

Uma das principais interpretações do intervalo de confiança consiste em avaliar a incerteza que temos a respeito de estimarmos o parâmetro populacional μ a partir de uma amostra aleatória de tamanho n.

A raiz quadrada do fator abaixo é utilizada para correção do intervalo de confiança quando a população é finita, isto é, quando se conhece a população N.

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

- Variância Desconhecida:

Tendo os conceitos básicos sobre intervalos de confiança, vamos agora tratar uma situação mais realista: quando a variância σ^2 da população é desconhecida.

Consideremos uma amostra aleatória simples $X_1, X_2, \ldots, X_n$, obtida de uma população com distribuição normal, com média μ e variância σ^2 desconhecidas. Como neste caso a variância é desconhecida, utilizaremos a variância amostral S^2 no lugar de σ^2. Assim, temos que:

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

Isto representa que a variável T tem distribuição t de Student com $n-1$ graus de liberdade.

Analogamente ao caso anterior, obtemos que

$$\mathbb{P}\left(-t_{((n-1),\alpha/2)} \leq \frac{\overline{X} - \mu}{s/\sqrt{n}} \leq t_{((n-1),\alpha/2)}\right) = 1 - \alpha$$

Exemplo-1:

Consideremos que o projetista de uma indústria tomou uma amostra de 36 funcionários para verificar o tempo médio gasto para montar um determinado brinquedo. Os tempos estão colocados na Tabela a seguir. Dado que o projetista não tem conhecimento da variabilidade da população, construir um intervalo de confiança com $(1-\alpha) = 0{,}95$ para a média μ.

Tabela de dados			
17,1000	16,8930	14,6004	13,0053
29,6292	19,2500	17,7504	24,6337
29,3567	25,0798	16,7914	29,4087
23,8807	15,2133	19,1536	30,3199
13,0050	24,6795	29,3308	20,7309
16,4541	26,2017	21,7857	19,7393
24,6042	18,6442	21,2594	26,9123
16,9896	32,8977	21,3627	15,4958
18,3113	23,6931	19,5429	16,3855

Solução:

A partir da análise do conjunto de dados temos que $\overline{x} = 21,39$ e $s = 5,38$ Substituindo esses valores na fórmula do intervalo de confiança temos que

$$21,39 - 2,03\frac{5,38}{\sqrt{36}} \leq \mu \leq 21,39 + 2,03\frac{5,38}{\sqrt{36}}$$

Portanto,

$IC(\mu, 0,95) = (19{,}56; 23,21)$

Exemplo-2:

Foram realizados testes glicêmicos em 25 pacientes após um jejum de 8 horas. Os resultados são apresentados na tabela abaixo. Encontrar um intervalo de confiança de nível 95% para a média μ.

Teste glicêmico (mg/dL)				
80	118	100	90	83
117	95	84	102	80
112	78	102	121	82
77	88	73	104	88
132	91	103	140	101

Solução:

Inicialmente, calculamos a média amostral $\overline{X}$ e o desvio padrão amostral s, que são dados por:

$$\overline{X} = \frac{1}{25}\sum_{i=1}^{25} X_1 = 97,64 \qquad s = \sqrt{\frac{1}{24}\sum_{i=1}^{25}(X_1 - \overline{X})^2} = 17,82.$$

Como a confiança é de 95%, segue $t_{0,025,24} = 2,06$ e então, substituindo esses valores na fórmula do intervalo de confiança, temos que

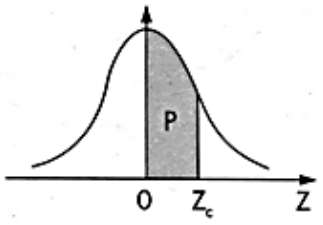
$$IC = (\mu; 0,95) = \left[97,64 - 2,06\frac{17,82}{\sqrt{25}}; 97,64 + 2,06\frac{17,82}{\sqrt{25}}\right] = [90,28; 105].$$

Tabela-2: - Distribuição Normal Padrão

Distribuição Normal Padrão

$Z \sim N(0, 1)$

Corpo da tabela dá a probabilidade p, tal que $p = P(0 < Z < Z_c)$



parte inteira e primeira decimal de Z_c	Segunda decimal de Z_c										parte inteira e primeira decimal de Z_c
	0	1	2	3	4	5	6	7	8	9	
	p = 0										
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574	2,1
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899	2,2
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158	2,3
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361	2,4
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520	2,5
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643	2,6
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736	2,7
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807	2,8
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861	2,9
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900	3,0
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929	3,1
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950	3,2
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965	3,3
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976	3,4
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983	3,5
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989	3,6
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992	3,7
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995	3,8
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997	3,9
4,0	49997	49997	49997	49997	49997	49997	49998	49998	49998	49998	4,0
4,5	49999	50000	50000	50000	50000	50000	50000	50000	50000	50000	4,5

Tabela 3 – Tabela t Student

Unilateral α	0,25	0,10	0,05	0,025	0,01	0,005
Bilateral α	0,50	0,20	0,10	0,05	0,02	0,01
c	0,50	0,80	0,90	0,95	0,98	0,99
G.L						
1	1,000	3,078	6,314	12,706	31,821	63,657
2	0,816	1,886	2,920	4,303	6,965	9,925
3	0,765	1,638	2,353	3,182	4,541	5,841
4	0,741	1,533	2,132	2,776	3,747	4,604
5	0,727	1,476	2,015	2,571	3,365	4,032
6	0,718	1,440	1,943	2,447	3,143	3,707
7	0,711	1,415	1,895	2,365	2,998	3,499
8	0,706	1,397	1,860	2,306	2,896	3,355
9	0,703	1,383	1,833	2,262	2,821	3,250
10	0,700	1,372	1,812	2,228	2,764	3,169
11	0,697	1,363	1,796	2,201	2,718	3,106
12	0,695	1,356	1,782	2,179	2,681	3,055
13	0,694	1,350	1,771	2,160	2,650	3,012
14	0,692	1,345	1,761	2,145	2,624	2,977
15	0,691	1,341	1,753	2,131	2,602	2,947
16	0,690	1,337	1,746	2,120	2,583	2,921
17	0,689	1,333	1,740	2,110	2,567	2,898
18	0,688	1,330	1,734	2,101	2,552	2,878
19	0,688	1,328	1,729	2,093	2,539	2,861
20	0,687	1,325	1,725	2,086	2,528	2,845
21	0,686	1,323	1,721	2,080	2,518	2,831
22	0,686	1,321	1,717	2,074	2,508	2,819
23	0,685	1,319	1,714	2,069	2,500	2,807
24	0,685	1,318	1,711	2,064	2,492	2,797
25	0,684	1,316	1,708	2,060	2,485	2,787
26	0,684	1,315	1,706	2,056	2,479	2,779
27	0,684	1,314	1,703	2,052	2,473	2,771
28	0,683	1,313	1,701	2,048	2,467	2,763
29	0,683	1,311	1,699	2,045	2,462	2,756
∞	0,674	1,282	1,645	1,960	2,326	2,576

EXERCÍCIOS PARA TREINAMENTO

Questão 1

Um dos principais produtos de uma indústria siderúrgica é a folha de flandres. Havia uma preocupação com a possibilidade de haver um número de folhas fora da faixa de especificação de dureza (LIE = 58,0 HR e LSE = 64,0 HR). A partir desta informação a empresa decidiu estimar a dureza média das folhas de flandres (μ) coletando uma amostra aleatória de 49 folhas.

Medidas de dureza (HR) das folhas-de-flandres fabricadas pela siderúrgica						
61,0	60,2	60,3	60,3	60,0	61,0	60,3
60,0	60,0	60,9	61,0	61,2	59,2	60,9
60,0	60,5	59,8	59,3	61,0	59,6	59,8
59,6	60,1	58,0	59,8	58,9	57,6	58,0
60,5	60,1	61,6	61,1	59,7	58,3	61,6
59,5	59,0	60,3	58,7	59,6	54,2	60,3
61,0	59,7	59,9	59,9	60,0	58,6	59,9

Para um grau de confiança de 95%, determine a margem de erro (E) e o intervalo de confiança para média populacional (μ).

a) [60,04; 60,38]HR

b) [80,04; 60,38]HR

c) [60,04; 100,38]HR

d) [40,04; 60,38]HR

e) nda

Questão 2

A altura dos alunos de uma academia apresenta uma distribuição aproximadamente normal. Para estimar a altura média dessa população, foi observada a altura de 30 alunos, obtendo-se x =175 cm e s=15 cm. Determine um intervalo de confiança de 99% para a média populacional.

a) 187,95 <μ<182,05

b) 167,95 <μ<182,05

c) 167,95 <μ<192,05

d) 467,95 <μ<782,05

e) 267,95 <μ<782,05

Questão 3

Sabe-se que uma amostra possui 25 elementos, média 150 e desvio padrão igual a 10. Represente um intervalo de confiança em nível de 90%.

a) 146,57; 153,42

b) 176,57; 193,42

c) 126,57; 143,42

d) 146,57; 253,42

e) 156,57; 353,42

RESOLUÇÕES:

Resposta Questão 1

$\overline{X} = 60{,}21$

s=0,61

n=49

Grau de confiança de 95% implica em: 1 – α = 95%, logo $\alpha = 5\% = 0{,}05$ e $\alpha/2 = 0{,}025$.

$Z_{\alpha/2} = Z_{0{,}025} = 1{,}96$

$$E = Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$E = 1{,}96 . \frac{0{,}61}{\sqrt{49}} = 0{,}1708 = 0{,}17$$

Intervalo de confiança:

$$\overline{x} - E < \mu < x + E$$

$$60{,}21 - 0{,}17 < \mu < 60{,}21 + 0{,}17$$

Se fôssemos selecionar muitas amostras de 49 elementos da produção de folhas e construíssemos um intervalo de 95% de confiança para cada amostra, 95% desses intervalos conteriam a média populacional μ

[60,04 ; 60,38]HR

Gabarito: **LetraA.**

Resposta Questão 2

Para encontrarmos o erro, utilizamos a fórmula: $E = Z_C \frac{S}{\sqrt{n}}$ pois $n \geq 30$ e $\sigma \cong s$

C=99%, então $Z_C = 2{,}575$ vide (Tabela 1)

n=30

s=15 cm

$$E = 2{,}575 . \frac{15}{\sqrt{30}} = 7{,}05$$

O intervalo de confiança é dado por: $\overline{X} - E < \mu < \overline{X} + E$

175 - 7,05 <μ< 175 + 7,05

167,95 <μ<182,05 .

Portanto, com 99% de confiança, podemos dizer que a média populacional está entre 167,95 cm e 182,05 cm.

Gabarito:Letra B.

Resposta Questão 3

$$\overline{x} \pm t\frac{s}{\sqrt{n}}$$

$$150 \pm 1{,}7109\frac{10}{\sqrt{25}}$$

$$150 \pm 3{,}4218$$

$$P(146{,}5782 \leq \overline{x} \leq 153{,}4218) = 0{,}90$$

Gabarito: Letra A.

2.3 INTERVALO DE CONFIANÇA PARA A PROPORÇÃO:

Consideremos X a variável aleatória que representa a presença (ou não) de determinada característica de uma população. Assim, temos que X tem distribuição de Bernoulli com parâmetro *P*, no qual *P* representa a probabilidade de um determinado elemento da amostra ter a característica de interesse. Retiramos uma amostra aleatória $X_1, \dots, X_n$ desta população. Cada $X_i, i = 1, \dots, n$ tem distribuição de Bernoulli com parâmetro *P*, isto é,

$$X_1, X_2, \dots, X_n \sim Bernoulli(p)$$

com média $\mu = p$ e variância $\sigma^2 = p(1-p)$

Neste caso, o estimador de máxima verossimilhança($\hat{p}$) para o parâmetro populacional *p* é dado por:

$$\hat{p} = \frac{\text{Nordm; de elementos da amostra com a característica}}{\text{Total de elementos da amostra}} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

Utilizaremos três métodos diferentes para encontrar o intervalo de confiança para a proporção: Aproximação Normal, Aproximação Normal com Correção de Continuidade e Binomial Exata.

2.3.1 APROXIMAÇÃO NORMAL:

Vejamos como construir intervalos de confiança para a proporção p, utilizando a aproximação Normal. Consideremos $\hat{p}$ a proporção amostral. Pelo Teorema Central do Limite temos que para um tamanho de amostra grande, podemos considerar a proporção amostral $\hat{p}$ como tendo aproximadamente distribuição normal com média p e variância *p(1-p)/n*. Partindo-se destas premissas pode-se afirmar que

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Observamos que a variância de $\hat{p}$ depende do parâmetro desconhecido P. No entanto, pelo fato de n ser grande, podemos substituir P por $\hat{p}$ Com isso temos que:

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1).$$

Considerando o mesmo procedimento de montagem do intervalo para a média, construímos o intervalo com $100(1-\alpha)\%$ de confiança para a proporção p:

$$IC(p,1-\ \alpha) = \left(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Exemplo-1:

Numa amostra aleatória de tamanho $\mathrm{n}=700$ foram encontrados 68 elementos defeituosos. Achar um intervalo de confiança de nível 95% para a proporção p de defeituosos.

Temos que $\hat{p} = 68/700 \ = \ 0{,}0971$. Para $\alpha = 0{,}05$, temos pela tabela da distribuição normal que $Z_{0{,}025} = 1{,}96$. Então, o intervalo de confiança é dado por

$$\left(0{,}0971 - 1{,}96\sqrt{\frac{0{,}0971(0{,}9028)}{700}}; 0{,}0971 + 1{,}96\sqrt{\frac{0{,}0971(0.9028)}{700}}\right) = (0{,}0752; 0{,}119).$$

2.3.2 APROXIMAÇÃO NORMAL COM CORREÇÃO DE CONTINUIDADE

Uma outra maneira de obtermos um intervalo de confiança para proporção é através da aproximação normal com correção de continuidade. Considerando o processo anterior, a única diferença é que aqui não consideraremos simplesmente a proporção amostral $\hat{p}$, mas sim uma correção dela. Assim, para determinar o intervalo de confiança consideramos uma modificação da proporção $\hat{p}$, dada por:

$$\hat{p} = \begin{cases} \hat{p} + \dfrac{1}{2n} \ se\ \hat{p} < 0{,}5 \\ \hat{p} - \dfrac{1}{2n} \ se\ \hat{p} > 0{,}5 \end{cases}$$

Assim, o intervalo de confiança para proporção $p\,\hat{p}$ com correção de continuidade, é dado por

$$IC(p,1-\alpha) = \left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n}}\right), \hat{p}_c + Z_{\alpha/2}\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n}}$$

O fator de continuidade é utilizado para melhorar a aproximação de uma variável aleatória discreta $\hat{P}$ pela distribuição normal que é contínua.

Exemplo-2:

Consideremos novamente o Exemplo-1. Vamos agora encontrar o intervalo de confiança com correção de continuidade.

Temos que $\hat{p} = 68/700 = 0.0971$. Assim, $\hat{p} < 0{,}5$. Então $\hat{p}_c = 0{,}0971 + 1/1400 = 0{,}0978$. Para $\alpha = 0{,}05$, temos pela tabela da distribuição normal que $Z_{0{,}025} = 1{,}96$. Então o intervalo de confiança é dado por:

$$IC(p, 0{,}95) = \left(0{,}0978 - 1{,}96\sqrt{\frac{0{,}0978(1-0{,}0978)}{700}}; 0{,}0978 + 1{,}96\sqrt{\frac{0{,}0978(1-0.0978)}{700}} \right)$$

$$IC(p, 0{,}95) = (0{,}07579; 0{,}1198)$$

2.4 INTERVALO DE CONFIANÇA PARA A VARIÂNCIA:

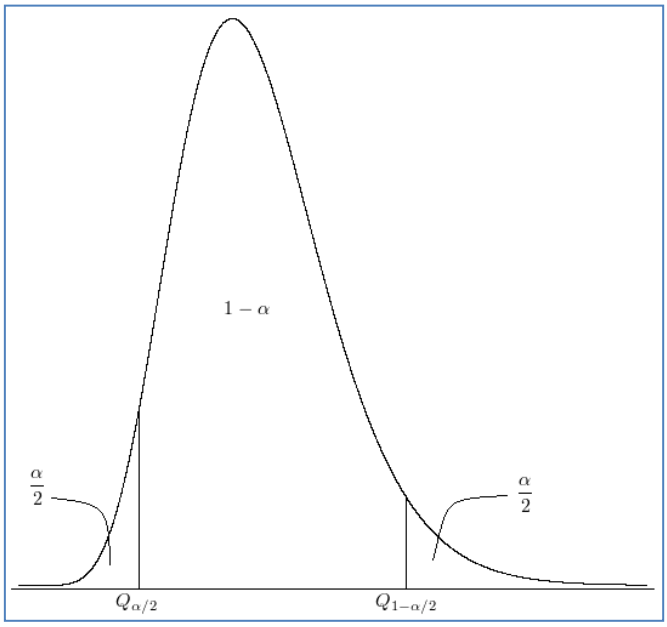
Consideremos uma amostra aleatória $X_1, \ldots, X_n$ de tamanho n de uma população com distribuição normal com média μ e variância σ^2. Um estimador para σ^2 é a variância amostral s^2. Assim, sabemos que a quantidade pivô é:

$$Q = \frac{(n-1)s^2}{\sigma^2} \sim X^2_{n-1}$$

Seja $1 - \sigma$ a probabilidade da variável Q, com $n - 1$ graus de liberdade, tomar valores entre $Q_{\alpha/2}$ e $Q_{1-\alpha/2}$, valores obtidos na tabela da distribuição qui-quadrado tais que

$\mathbb{P}[Q < Q_{\alpha/2}] = P[Q > Q_{1-\alpha/2}] = \alpha/2$.

Figura 3 – Distribuição Qui-Quadrado



Observando a equação

$$Q_{\alpha/2} \leq Q \leq Q_{1-\alpha/2}$$

vemos que podemos substituir Q pela expressão acima e então obtemos

$$Q_{\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq Q_{1-\alpha/2}$$

Reescrevendo esta desigualdade, obtemos o intervalo de confiança para a variância,

$$\frac{(n-1)s^2}{Q_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{Q_{\alpha/2}}$$

Assim,

$$\mathbb{P}\left(\frac{(n-1)s^2}{Q_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{Q_{\alpha}}\right) = 1-\alpha$$

Logo, o intervalo com nível $100(1-\alpha)\%$ de confiança para σ^2 será dado por

$$IC(\sigma^2, 1-\alpha) = \left(\frac{(n-1)s^2}{Q_{1-\alpha/2}}, \frac{(n-1)s^2}{Q_{\alpha/2}}\right)$$

Exemplo-3:

O peso de componentes mecânicos produzidos por uma determinada empresa é uma variável aleatória que se supõe ter distribuição normal. Pretende-se estudar a variabilidade do peso dos referidos componentes. Para isso, uma amostra de tamanho 11 foi obtida, cujos valores em grama são:

98 97 102 100 98 101 102 105 95 102 100

Construa um intervalo de confiança para a variância do peso, com um grau de confiança igual a 95%.

Temos que $n = 11, \overline{x} = 100$ e,

$$s^2 = \sum_{i=1}^{1} 1\frac{(x_1 - \overline{x})^2}{10} = \frac{4 + 9 + \cdots + 25 + 4 + 0}{10} = 8$$

Pela Tabela da distribuição qui-quadrado com 10 graus de liberdade, temos que $Q_{0,025} = 3,25$ e$Q_{0,975} = 20,48$. Assim,

$$IC(\sigma^2, 1-\alpha) = \left(\frac{10\,.\,8}{20,48}, \frac{10\,.\,8}{3,25}\right) = (3,90; 24,61)$$

2.5 INTERVALO DE CONFIANÇA PARA A RAZÃO ENTRE DUAS VARIÂNCIAS:

Vejamos como construir um intervalo de confiança para a razão entre duas variâncias de populações normais independentes. Para isso retiramos uma amostra aleatória $X_1, X_2, \dots, X_{n_1}$ da população 1, com distribuição $N(\mu_1, \sigma_1^2)$, e uma amostra $Y_1, Y_2, \dots, Y_{n_2}$ da população 2, com distribuição $N(\mu_2, \sigma_2^2)$. Como

$$Q_1 = \frac{(n_1 - 1)}{\sigma_1^2} s_1^2 \sim X_{n_1}^2 - 1 \quad (Qui - quadrado\ com\ n_1 - 1\ graus\ de\ liberdade)$$

$$Q_2 = \frac{(n_2 - 1)}{\sigma_2^2} s_2^2 \sim X_{n_2}^2 - 1 \quad (Qui - quadrado\ com\ n_2 - 1\ graus\ de\ liberdade)$$

em que s_1^2 é a variância amostral da população 1 e s_2^2 a variância amostral da população 2. Neste caso, a expressão de F é definida por

$$F = \frac{\frac{Q_1}{N_1 - 1}}{\frac{Q_2}{n_2}} = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = \frac{s_1^2}{s_2^2}\frac{\sigma_2^2}{\sigma_1^2}$$

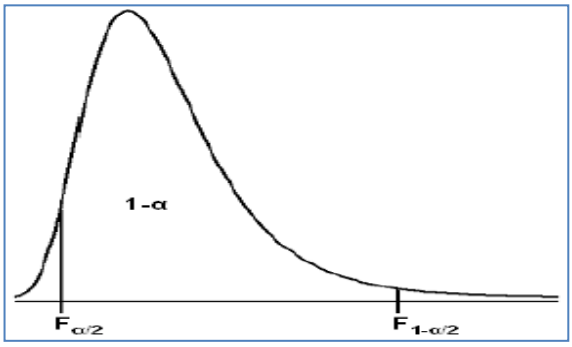
tem distribuição F (Fisher-Snedecor) de com $n_1 - 1$ graus de liberdade no numerador e $n_2 - 1$ graus de liberdade no denominador e denotamos por $F_{(n_{1-1};n_2-1)}$.

Consideremos que a probabilidade da variável F tomar valores entre $F_{(\frac{\sigma}{2};\, n_1-1\,;\, n_2-1)}$ e

e $F_{(1-\frac{\sigma}{2};\, n_1-1\,;\, n_2-1)}$ é $1 - \sigma$. Esses valores são obtidos na Tabela da distribuição de Fisher-Snedecor referente ao valor de α e aos graus de liberdade do numerador e do denominador,

$n_1 - 1$ e $n_2 - 1$, respectivamente. Veja a figura a seguir.

Figura-4: Distribuição F



Observando a equação

$$F_{\sigma/2} < F < F_{(1-\sigma/2)}$$

vemos que podemos substituir F pela expressão acima e assim temos:

$$F_{\alpha/2} < \frac{s_1^2}{s_2^2}\frac{\sigma_2^2}{\sigma_1^2} < F_{(1-\sigma/2)}$$

Reescrevendo esta equação obtemos:

$$\frac{1}{F_{(1-\sigma/2)}}\frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{\sigma/2}}\frac{s_1^2}{s_2^2}$$

Assim,

$$P\left(\frac{1}{F_{\left(1-\frac{\sigma}{2}\right)}}\frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{\frac{\sigma}{2}}}\frac{s_1^2}{s_2^2}\right) = 1-\sigma$$

Observe que $F_{(1-\alpha/2,n_1-1,n_2-1)} = \frac{1}{F_{(\alpha/2,\ n_2-1,n_1-1)}}$ e $F_{(\alpha/2,n_1-1,n_2-1)} = \frac{1}{F_{(1-\alpha/2,\ n_2-1,n_1-1)}}$

Logo, o intervalo de confiança com nível $100(1-\alpha)\%$ para a razão entre duas variâncias será dado por

$$IC(\sigma_1^2/\sigma_2^2, 1-\alpha) = \left(\frac{1}{F_{(1-\alpha/2)}}\frac{s_1^2}{s_2^2}; \frac{1}{F_{(\alpha/2)}}\frac{s_1^2}{s_2^2}\right)$$

2.6 INTERVALO DE CONFIANÇA PARA A DIFERENÇA DE MÉDIAS:

2.6.1 VARIÂNCIAS CONHECIDAS:

Consideremos duas amostras aleatórias, $X_1, X_2, \ldots, X_{n_1}$ de tamanho n_1 e $Y_1, Y_2, \ldots, Y_{n_2}$ de tamanho n_2, ambas com distribuição normal, médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2, respectivamente. Assim, a Média Amostral é aproximadamente Normal.

$$\overline{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ e } \overline{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Daí, temos que,

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

o que implica em

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Consideremos que a probabilidade da variável Z tomar valores entre $-Z_{\alpha/2}$ e $Z_{\alpha/2}$ é $1-\alpha$. Observando a equação

$$-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$$

vemos que podemos substituir Z pela expressão acima e assim obtemos

$$-Z_{\alpha/2} \leq \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leq Z_{\alpha/2}$$

Reescrevendo esta desigualdade, obtemos o intervalo de confiança para a diferença das médias $\mu_1-\mu_2$

$$IC(\mu_1-\mu_2, 1-\alpha) = \left(\left(\overline{X}-\overline{Y}\right) - Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right); \left(\overline{X}-\overline{Y}\right) + Z_{\alpha/2}\left(\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)\right)$$

e podemos afirmar que se pudéssemos construir uma quantidade grande de intervalos $IC(\mu_1-\mu_2, 1-\alpha)$, todos baseados em amostras de tamanho n_1 e n_2, em torno de $100(1-\alpha)\%$ deles conteriam o valor verdadeiro da média populacional.

2.6.2 VARIÂNCIAS DESCONHECIDAS - PORÉM IGUAIS:

Considerando agora duas amostras aleatórias, $X_1, X_2, \ldots, X_{n_1}$ de tamanho n_1 e $Y_1, Y_2, \ldots, Y_{n_2}$ de tamanho n_2, com apenas uma diferença do caso anterior: as variâncias são desconhecidas, porém iguais. Isto é, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Como

$$\frac{(n_1-1)s_1^2}{\sigma^2} \sim X^2_{n_{1-1}} \text{ e } \frac{(n_2-1)s_2^2}{\sigma^2} \sim X^2_{n_{2-1}}$$

onde s_1^2 é a variância amostral da população 1 e s_2^2 é a variância amostral da população 2, temos que:

$$T = \frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_{1+n_2}-2}$$

Onde:

$$S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

Daí, utilizando a tabela da distribuição t de Student com $a = n_1 + n_2 - 2$ graus de liberdade, obtemos o valor de $t_{(a,\alpha/2)}$ de forma que

$$-t_{(a,\alpha/2)} < \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{(a,\alpha/2)}$$

Reescrevendo esta desigualdade, obtemos o intervalo de confiança para a diferença das médias $\mu_1 - \mu_2$ quando as variâncias são desconhecidas, porém iguais,

$$(\overline{X} - \overline{Y}) - t_{(a,\alpha/2)}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{X} - \overline{Y}) + t_{(a,\alpha/2)}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

ou

$$IC(\mu_1 - \mu_2, 1-\alpha) = \left((\overline{X} - \overline{Y}) - t_{\left(a,\frac{\alpha}{2}\right)}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; (\overline{X} - \overline{Y}) + t_{\left(a,\frac{\alpha}{2}\right)}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

e podemos afirmar que se pudéssemos construir uma grande quantidade de intervalos $IC(\mu_1 - \mu_2, 1-\alpha)$, todos baseados em amostras de tamanho n_1 e n_2 em torno de $100(1-\alpha)\%$ deles conteriam a verdadeira diferença das médias populacionais.

2.6.3 VARIÂNCIAS DESCONHECIDAS E DIFERENTES:

Consideremos duas amostras aleatórias, $X_1, X_2, \ldots, X_{n_1}$ de tamanho n_1 e $Y_1, Y_2, \ldots, Y_{n_2}$ de tamanho n_2, com distribuições normais, mas agora com variâncias desconhecidas e diferentes, isto é, $\sigma_1^2 \neq \sigma_2^2$. Como as variâncias populacionais são desconhecidas, usaremos as variâncias amostrais s_1^2 e s_2^2 em seus lugares. Considerando a variável T tal que

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_v$$

ou seja, a variável T dada pela equação acima tem distribuição t de Student com v graus de liberdade, onde

$$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Fazendo uma construção análoga a do caso anterior, obtemos o intervalo de confiança para a diferença de duas médias com variâncias desconhecidas e desiguais:

$$IC(\mu_1-\mu_2,1-\alpha)=\left((\overline{X}-\overline{Y})-\ t_{\left(v,\frac{\alpha}{2}\right)}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}};(\overline{X}-\overline{Y})+\ t_{\left(v,\frac{\alpha}{2}\right)}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$$

Exemplo-4:

Os dados a seguir correspondem a teores de um elemento indicador da qualidade de um certo produto. Foram coletadas 2 amostras referentes a 2 métodos de produção. Construa um intervalo de confiança para a diferença das médias dos dois métodos.

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

A média referente ao método 1 é $\overline{X}_1=3,63$ e do método 2 é $\overline{X}_2=3,96$. Calculando as variâncias amostrais, obtemos:

$$s_1^2=\sum_{i=1}^{10}\frac{(x_{1i}-\overline{x}_1)^2}{9}=8,29 \qquad s_2^2=\sum_{i=1}^{10}\frac{(x_{2i}-\overline{x}_2)^2}{9}=2,53$$

em que x_{1i} são os teores referentes ao método 1 e x_{2i} ao método 2, $i=1,\dots,10$. Os graus de liberdade são dados por:

$$v=\frac{(\frac{8,29}{10}+\frac{2,53}{10})^2}{\frac{\left(\frac{8,20}{10}\right)^2}{9}+\frac{\left(\frac{2,35}{10}\right)^2}{9}}=14,028$$

Assim, da Tabela da distribuição t de Student obtemos que $t_{14,0,025}=2,145$ e, então, temos que:

$$IC(\mu_1-\mu_2,1-\alpha)=\left((3,63-3,96)(-2,145)\sqrt{\frac{8,29}{10}+\frac{2,53}{10}};(3,63-3,96)(2,145)\sqrt{\frac{8,29}{10}+\frac{2,53}{10}}\right)$$

ou seja, $\text{IC}(\mu_1-\mu_2,1-\alpha)=(-2,56;1,90)$.

EXERCÍCIOS PARA TREINAMENTO

a. Considere uma amostra aleatória n=25 que possui uma média amostral de 51,3 e um desvio padrão populacional de σ=2. Construa o intervalo com 95% de confiança para a média populacional μ.

b. Sabe-se que a vida em horas de um bulbo de lâmpada de 75 W é distribuída de forma aproximadamente normal com desvio padrão de σ=25. Uma amostra aleatória de 20 bulbos tem uma vida média de 1.014 horas. Construa um intervalo de confiança de 95% para a vida média.

c. Qual deve ser o tamanho da amostra para que o intervalo com 99,5% de confiança para a média populacional tenha uma semi amplitude não superior a 1,5? Sabe-se que a variância populacional é de 23.

d. Calcular o intervalo de confiança de 95% para a seguinte amostra, com variância populacional desconhecida:

19,8 18,5 17,6 16,7 15,8 15,4 14,1 13,6 11,9 11,4 11,4 8,8 7,5 15,4 15,4 19,5 14,9 12,7 11,9 11,4 10,1 7,9

e. Uma marca particular de margarina diet foi analisada para determinar o nível em porcentagem de ácidos graxos insaturados. Uma amostra de seis pacotes resultou nos seguintes dados: 16,8; 17,2; 17,4; 16,9; 16,5 e 17,1. Encontre o intervalo de confiança de 99% para a amostra.

f. Uma amostra piloto com 12 elementos traça uma média de 6,7 e desvio padrão de 1,7. Qual deve ser o tamanho da amostra para que a semi amplitude do intervalo de 99,5% de confiança da média populacional não seja superior a 0,8?

g. O conteúdo de açúcar na calda de pêssegos em lata é normalmente distribuído. É extraída uma amostra de n=10 latas que resulta em um desvio padrão amostral de s=4,8. Encontre o intervalo de confiança para de 95% para a variância populacional σ2.

h. Se uma amostra de tamanho n=20, a média e o desvio padrão são X=1,25 e s=0,25, respectivamente. Construa um intervalo de confiança de 99% para σ 2.

i. Em uma amostra aleatória de 85 mancais de eixos de manivelas de motores de automóveis, 10 têm um acabamento superficial mais rugoso do que as especificações permitidas. Calcule um intervalo de confiança para o 95% da proporção.

j. De 1.000 casos selecionados de aleatoriamente de câncer de pulmão, 823 resultaram em morte. Construa um intervalo de confiança de 95% para a taxa de morte de câncer de pulmão.

GABARITO:

1. I.C. = 51,3 ± 0,78
2. I.C. = 1014 ± 11
3. $2{,}81 \cdot 4{,}8 / \sqrt{n} < 1{,}5 \Rightarrow n = 80{,}85 \approx 81$ elementos
4. I.C. = 13,71 ± 1,57
5. I.C. = 16,98 ± 0,53
6. ≈ 56 elementos.
7. = [10,9; 76,8]
8. = [0,03; 0,17]
9. [0,05; 0,19]
10. [0,799; 0,847]

Capítulo 3

Teste de Hipóteses

3.1 TESTE DE HIPOTESES

Teste de Hipóteses é um procedimento que permite tomar uma decisão (aceitar ou rejeitar a hipótese nula) entre duas ou mais hipóteses (hipótese nula ou hipótese alternativa), utilizando os dados observados de um determinado experimento. Há diversos métodos para realizar o teste de hipóteses, dos quais se destacam o método de Fisher (teste de significância), o método de Neyman–Pearson e o método de Bayes.

São dois os tipos de erros que podemos cometer na realização de um teste de hipóteses:

1. Rejeitar a hipótese H_0, quando ela é verdadeira.
2. Não rejeitar a hipótese H_0, quando ela é falsa.

A Tabela a seguir resume as situações acima.

Tabela 4: Tipos de Erros

	Aceitar H_0	Rejeitar H_0
H_0 Verdadeira	Decisão Correta	Erro Tipo I
H_0 Falsa	Erro Tipo II	Decisão Correta

Se a hipótese H_0 for verdadeira e não rejeitada ou falsa e rejeitada, a decisão estará correta. No entanto, se a hipótese H_0 for rejeitada sendo verdadeira ou se não for rejeitada sendo falsa, a decisão estará errada. O primeiro destes erros é chamado de Erro do Tipo I e a probabilidade de cometê-lo é denotada pela letra grega α (alfa); o segundo é chamado de Erro do Tipo II e a probabilidade de cometê-lo é denotada pela letra grega β (beta). Assim temos,

$$\alpha = \mathbb{P}(\text{Erro do tipo I}) = P(\text{rejeitar } H_0 \text{ dado } H_0 \text{ verdadeira});$$

$$\beta = \mathbb{P}(\text{Erro do tipo II}) = P(\text{aceitar } H_0 \text{ dado } H_0 \text{ falsa});$$

Considere um teste unilateral dado pelas hipóteses:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 = \mu < \mu_0 \end{cases}$$

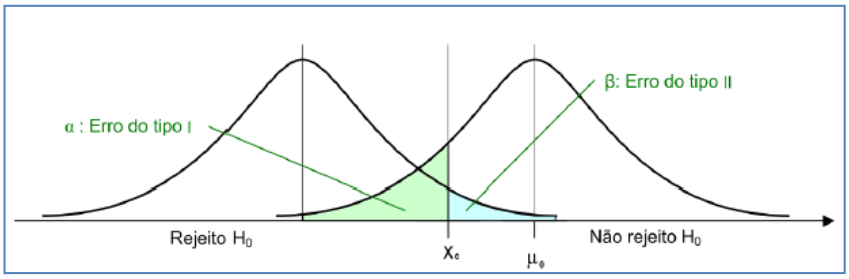
Neste caso, a região de rejeição é determinada por $\{\overline{X} < X_C\}$ e a interpretação dos erros pode ser vista como:

$$\alpha = \mathbb{P}\left(\overline{X} < X_C\right) | \mu = \mu_0$$

$$\beta = \mathbb{P}\left(\overline{X} > X_C\right) | \mu < \mu_0$$

A situação ideal é aquela em que ambas as probabilidades, α e β, são próximas de zero. No entanto, é fácil ver que a medida que diminuímos α, β aumenta. A Figura a seguir apresenta esta relação.

Figura-5: Testes de Hipótese.



Para um teste de hipóteses do tipo acima, onde estamos interessados em testar a média de uma população, utilizamos a expressão

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

que é a estatística do teste de hipóteses. A partir do Teorema Central do Limite, sabemos que, desde que tenhamos um tamanho amostral suficientemente grande, esta estatística tem distribuição normal padrão, isto é,

$$Z \sim N(0,1)$$

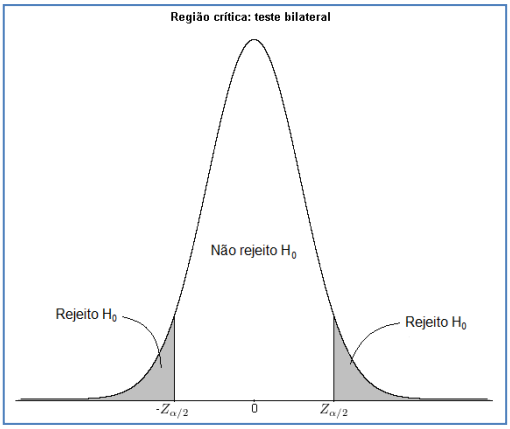
A partir dos valores de Z e da especificação do erro cometido, podemos definir a região crítica do teste.

Vamos considerar que o erro mais importante a ser evitado seja o Erro do Tipo I. A probabilidade de ocorrer o erro do tipo I σ é denominada nível de significância do teste. O complementar do nível de significância $(1 - a)$ é denominado nível de confiança. Supondo que o nível de significância α seja conhecido, temos condições de determinar o(s) valor(es) crítico(s). Se considerarmos o teste bilateral

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 = \mu \neq \mu_0 \end{cases}$$

a figura a seguir representa a região de rejeição para um valor fixo de α.

Figura-6: Teste Bilateral.

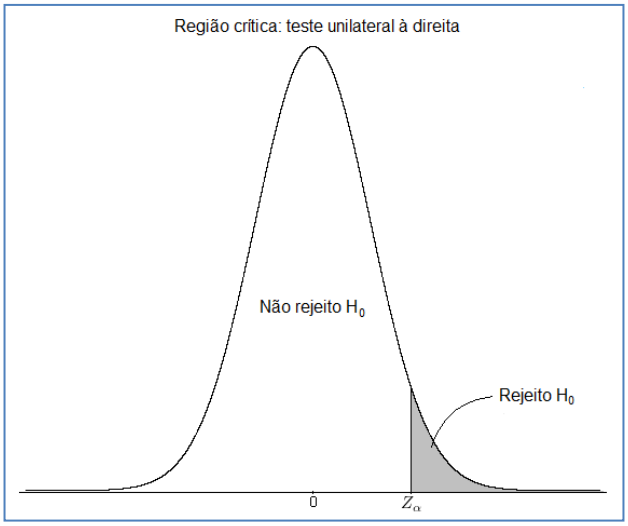


Se considerarmos o teste unilateral à direita

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 = \mu > \mu_0 \end{cases}$$

a região crítica é representada segundo a figura abaixo.

Figura-7: Teste Unilateral à Direita.

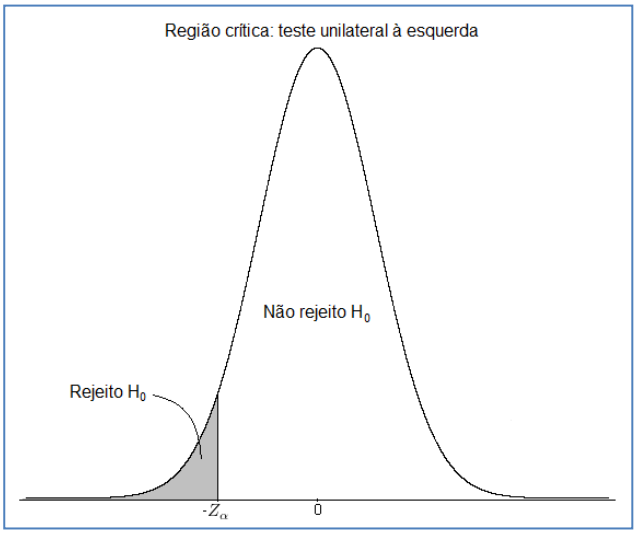


E, se considerarmos o teste unilateral à esquerda

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 = \mu < \mu_0 \end{cases}$$

a região crítica é representada segundo a figura abaixo.

Figura-8: Teste Unilateral à Esquerda.



Os valores $-Z_a$ e Z_a nas duas últimas figuras são tais que as áreas à esquerda e à direita, respectivamente, sob a curva Normal padrão, valem α. Agora, os valores $-Z_{a/2}$ e $Z_{a/2}$ na primeira figura, são tais que as áreas à esquerda e à direita, respectivamente, sob a curva Normal padrão, valem $\alpha/2$

Como foi dito inicialmente, o objetivo do teste de hipótese é determinar, através de uma estatística, se a hipótese nula é aceitável ou não. Essa decisão é tomada considerando a região de rejeição ou região crítica (RC). Caso o valor observado da estatística pertença à região de rejeição, rejeitamosH_0; caso contrário, não rejeitamos H_0. Analogamente, definimos a região de aceitação (complementar da região de rejeição): caso o valor observado pertença à região de aceitação, não rejeitamos H_0; se não pertencer, rejeitamos.

Se o nível de significância é 0,05, os valores críticos são $-1,645$ ou $1,645$ para as alternativas unilaterais e$-1,96$ e $1,96$ para a alternativa bilateral; se o nível de significância é $0,01$, os valores críticos são $-2,33$ ou $2,33$ para as alternativas unilaterais e $-2,575$ e $2,575$ para a alternativa bilateral (valores obtidos na Tabela da distribuição normal). A tabela a seguir apresenta alguns critérios para o teste de hipótese.

Tabela 5: Tipos de Hipóteses

Hipótese Alternativa	Rejeita H_0 se	Aceita H_0 se
u < uo	$Z < -Z_\alpha$	$Z \geq -Z_\alpha$
u > uo	$Z > Z_\alpha$	$Z \leq Z_\alpha$
u ≠ uo	$Z < -Z_{\alpha/2}$ ou $Z > Z_{\alpha/2}$	$-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$

Exemplo-5

Um supervisor da qualidade quer testar, com base numa amostra aleatória de tamanho n = 35 e para um nível de significância $\alpha = 0{,}05$, se a profundidade média de um furo numa determinada peça é 72,4mm. O que podemos dizer se ele obteve $\overline{X} = 73,2$ mm e se sabe, de informações anteriores, que $\sigma = 2{,}1$ mm?

1. Primeiro vamos estabelecer as hipóteses:

$$\begin{cases} H_0 : \mu = 72,4 \\ H_1 = \mu \neq 72,4 \end{cases}$$

2. Como $\alpha = 0{,}05$, temos que $Z_{\alpha/2} = Z_{0{,}025} = 1,96$.

3. Critério: rejeitar H_0 se $Z_{obs} < -1,96$ ou se $Z_{obs} > 1,96$ em que

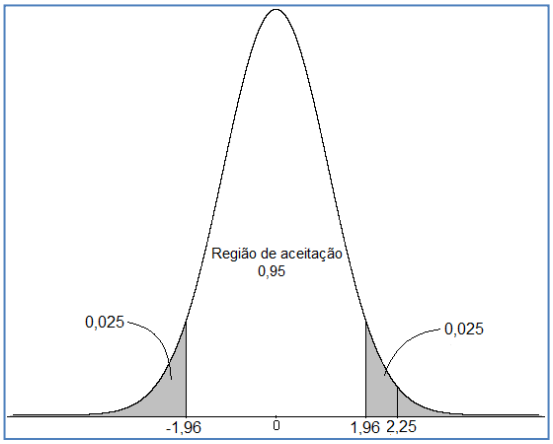
$$Z_{obs} = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

4. Substituindo $\mu_0 = 72,4$, $\sigma = 2,1$, $n = 35$, $\overline{x} = 73$ na equação acima, obtemos

$$Z_{obs} = \frac{73,2 - 72,4}{\frac{2,1}{\sqrt{35}}} = 2,25$$

5. Conclusão: Como $Z_{obs} = 2{,}25 > 1{,}96$, a hipótese nula deve ser rejeitada. Em outras palavras, não podemos assumir que a média populacional μ seja igual a 72,4, isto é, a diferença entre 73,2 e 72,4 é significativa. Veja a figura abaixo

Figura-9: Teste Bilateral



P-valor

O p-valor, também denominado nível descritivo do teste, é a probabilidade de que a estatística do teste (como variável aleatória) tenha valor extremo em relação ao valor observado (estatística) quando a hipótese H_0 é verdadeira.

Para exemplificar a definição de p-valor, considere um teste de hipóteses para a média no qual o valor da estatística é dado por Z_{obs}, ver Exemplo-5. As figuras a seguir representam, respectivamente, o p-valor nos casos em que temos um teste de hipóteses bilateral com rejeição da hipótese nula e sem rejeição da hipótese nula.

Figura-10: p Valor

A seguir, temos a figura de um teste de hipóteses unilateral para média. Na primeira das figuras, rejeitamos a hipótese nula e na segunda não rejeitamos.

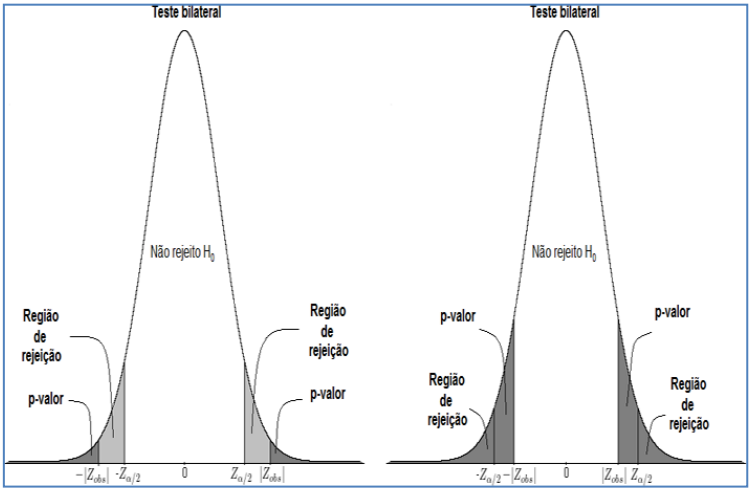
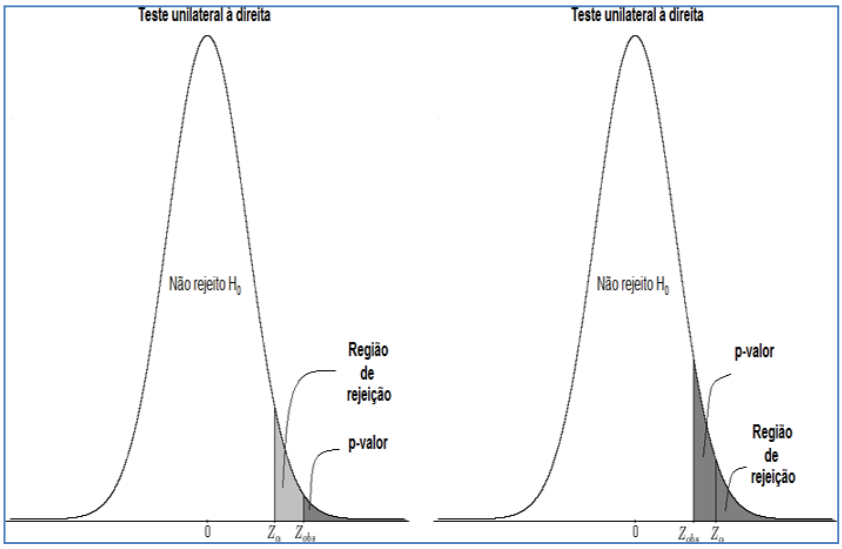


Figura-11: Teste de Hipóteses Unilateral para média



Observe que se o p-valor é menor que o nível de significância proposto α. Então, Z_{obs} está na região crítica e, portanto, rejeitamos a hipótese nula H_0. Por outro lado, se o p-valor é maior que o nível de significância, não rejeitamos a hipótese nula (ver figura acima). Além disso, quanto menor for o p-valor, mais "distante" estamos da hipótese nula H_0. Portanto, o p-valor tem mais informações sobre a evidência contra H_0 e, assim, o experimentador tem mais informações para decidir sobre H_0 com o nível de significância apropriado.

Também podemos interpretar o p-valor como o menor valor do nível de significância para o qual rejeitamos H_0. Desta forma, se o nível de significância (α) proposto para o teste for menor que o p-valor, não rejeitamos a hipótese H_0.

Em muitas situações, a região de rejeição de um teste de hipótese com nível de significância α apresenta seguinte forma:

Rejeitamos H_0 se e somente se $W(X) \geq c_a$,

em que $W(X)$ é a estatística do teste apropriada para o problema e a constante c_a é escolhida de modo que o teste tenha nível de significância α. Neste caso, o p-valor para o ponto amostral x é definido matematicamente como

$$p(x) = \sup_{\theta \epsilon \theta_0} P\theta[W(X) \geq W(x)]$$

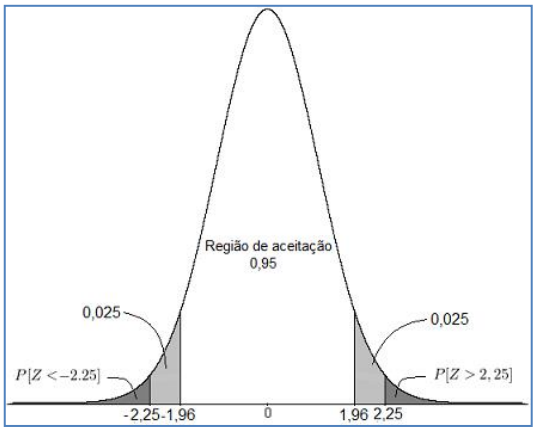
em que θ é um parâmetro pertencente ao espaço paramétrico θ sob a hipótese nula (H_0).

Voltando ao Exemplo-5, vamos calcular o p-valor do teste de médias. No decorrer deste módulo calculamos o p-valor para todos os testes estatísticos clássicos.

Neste caso, como temos um teste bilateral, segue que o p-valor é dado por

$$\mathrm{P-valor} = \mathbb{P}[\mathrm{Z} > |Z_{obs}| + \mathbb{P}[Z < -|Z_{obs}|] = \mathbb{P}[Z > 2{,}25] + \mathbb{P}[Z < -2.25] = 0{,}0122 = 0{,}0244$$

Figura-12: Gráfico do P-Valor



Portanto, podemos concluir que, para qualquer nível de significância maior que 0,0244, temos evidências para rejeitar a hipótese nula.

Análise do p-valor

Consideremos um teste de hipóteses no qual R_α é a região de rejeição com nível de significância α. Suponha que, para diferentes valores de α, essas regiões podem ser encaixadas no sentido que

$$R_\alpha \subset R_\alpha, \qquad \text{para qualquer } \alpha < \alpha. \qquad (5.1.2.1)$$

Sob essa situação, além de conseguirmos saber se a hipótese é rejeitada ou não, conseguimos ainda determinar o p-valor, que aqui é definido por

$$p = p(X) = \inf\{\alpha : X \in R_\alpha\}$$

no qual X representa a amostra.

O p-valor nos fornece uma ideia de quanto os dados contradizem a hipótese nula. Além disso, ele permite que diferentes experimentadores utilizem seus respectivos níveis de significância para avaliar os resultados do teste de hipóteses.

Exemplo-6:

Considere uma amostra de tamanho um de uma população X com distribuição $N(\mu, \sigma^2)$, com σ^2 conhecido. Consideremos sob $H_0, \mu = 0$, e sob $H_1, \mu = \mu_1$, para algum $\mu_1 > 0$. Seja Φ a função de distribuição acumulada da normal padrão e $z_{1-\alpha}$ o quantil $1-\alpha$ da distribuição normal padrão. Então, a região de rejeição pode ser denotada como

$$R_\alpha = \{X : X > \sigma z_{1-\alpha}\} = \left\{X : \Phi\left(\frac{X}{\sigma}\right) > 1-\alpha\right\} = \left\{X : 1 - \Phi\left(\frac{X}{\sigma}\right) < \alpha\right\}$$

Dessa maneira, para um valor observado de X dado, o ínfimo sobre todos α em que a última desigualdade se mantém é

$$p = 1 - \Phi\left(\frac{X}{\sigma}\right)$$

Alternativamente, podemos escrever que o p-valor é $\mathbb{P}_0[X \geq x]$, em que x é o valor observado de X. Notemos ainda que sob a hipótese nula, $\mu = 0$, a distribuição de p é dada da seguinte maneira

$$\mathbb{P}_0[p \leq u] = \mathbb{P}_0\left[1 - \Phi\left(\frac{X}{\sigma}\right) \leq u\right] = \mathbb{P}_0\left[\Phi\left(\frac{X}{\sigma}\right) \geq 1 - u\right] = u$$

pois $\Phi(X/\sigma)$ é uniformemente distribuído sobre (0,1), portanto p é uniformemente distribuído em (0,1). Esse resultado segue da transformação integral de probabilidade (*probability integral transformation),* que garante que:

Se X tem uma função de distribuição contínua F, então $F(X)$ é uniformemente distribuído sobre (0,1).

O Lema a seguir traz uma propriedade geral do p-valor.

Lema:

Suponhamos que X tem distribuição de probabilidade $\mathbb{P}_0$, para algum $\theta \epsilon \Theta$. Consideremos $\theta \in \Theta_0$, em que Θ_0 representa o espaço paramétrico sob a hipótese nula H_0. Assumimos ainda que as regiões de rejeição satisfazem (5.1.2.1)

i) Se

$$\sup_{\theta \epsilon \Theta_0} P\theta[X \in R_\alpha] \leq \alpha \quad para\ todo\ 0 < \alpha < 1, \quad (5.1.2.2)$$

então a distribuição de p sobre $\theta \in \Theta_0$ satisfaz

$$\mathbb{P}_\theta[p \leq u] \leq u \quad para\ todo\ 0 < u < 1.$$

Prova:

Se $\theta \in \Theta_0$, pela definição do p-valor, $p = p(X) = \inf\{\alpha: X \in R_\alpha\}$ e, temos que, para todo $v > u, [p \leq u] \subset [X \in \square_v]$, o que implica em $\mathbb{P}_\theta[p \leq u] \leq \mathbb{P}_\theta[X \in R_v]$ Assim, escrevendo

$$\lim_{v \to u^+} \mathbb{P}_\theta\,[p \leq u] \leq \lim_{v \to u^+} \mathbb{P}_\theta[X \in R_v]$$

como (5.1.2.2) é válido, segue que $\mathbb{P}_\theta[p \leq u] \leq u$.

ii) Se, para $\theta \in \Theta_0$,

$$\mathbb{P}_\theta[X \in R_\alpha] = \alpha \quad para\ todo\ 0 < \alpha < 1, \quad (5.1.2.3)$$

então

$$\mathbb{P}_\theta[p \leq u] = u \quad para\ todo\ 0 \leq u \leq 1,$$

ou seja, p é uniformemente distribuído sobre (0,1).

Prova:

Novamente pela definição do p-valor, temos que se$[X \in R_u]$ então $[p \leq u]$. Dessa forma, segue que

$$\mathbb{P}_\theta[p \leq u] \geq \mathbb{P}_\theta[X \in R_u]$$

Assim, por (5.1.2.3) temos que $\mathbb{P}_\theta[p \leq u] \geq u$. Do resultado obtido em (i), concluímos que $\mathbb{P}_\theta[p \leq u] \geq u$, ou seja, p tem distribuição uniforme em (0,1).

Passos para realização do teste de hipóteses

- Estabelecer as hipóteses;
- Determinar o nível de significância do teste (α);
- Determinar a região de rejeição;
- Calcular o p-valor

A seguir, vamos aplicar os conceitos discutidos acima para tratar diversos exemplos de testes de hipóteses.

3.2 TESTE DE HIPOTESES PARA A MÉDIA.

Considere uma população da qual retiramos uma amostra $X_1, X_2, \ldots, X_n$. Estamos interessados em realizar inferência sobre a média populacional μ.

Se não conhecemos o valor do desvio padrão populacional σ e a amostra é pequena, $n < 30$, devemos substituir a expressão

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

pela expressão

$$T = \frac{\overline{X} - \mu_0}{\frac{8}{\sqrt{n}}}$$

onde T tem distribuição t de Student com $n-1$ graus de liberdade. Para facilitar a execução do teste, podemos seguir os passos:

1. Estabelecer as hipóteses:

Fixamos $H_0: \mu = \mu_0$. Dependendo da informação que fornece o problema que estamos estudando, a hipótese alternativa pode ter uma das três formas abaixo:

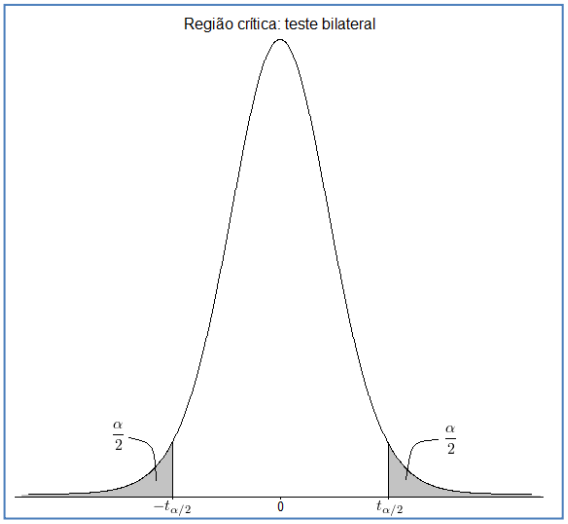
- $H_1: \mu \neq \mu_0$ (teste bilateral);
- $H_1: \mu > \mu_0$ (teste unilateral à direita);
- $H_1: \mu < \mu_0$ (teste unilateral à esquerda).

2. Fixar o nível de significância α.

3. Determinar a região crítica.

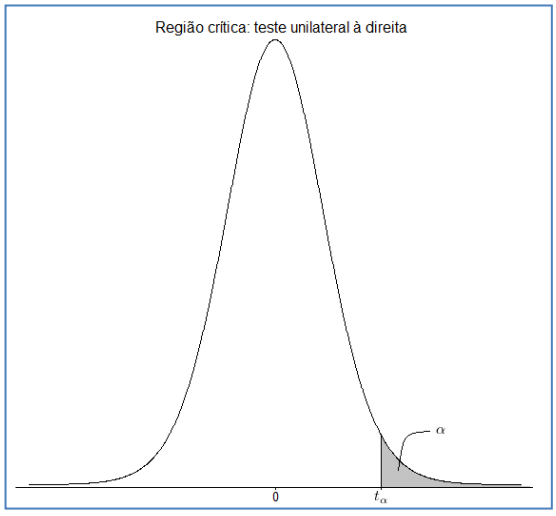
- Se o teste é bilateral, determinamos os pontos críticos $-t_{\alpha/2}$ e $t_{\alpha/2}$ tais que
- $\mathbb{P}[T > t_{\alpha/2}] = \mathbb{P}[T < -t_{\alpha/2}]$ a partir da distribuição t de Student com $n-1$ graus de liberdade.

Figura-13: Teste Bilateral



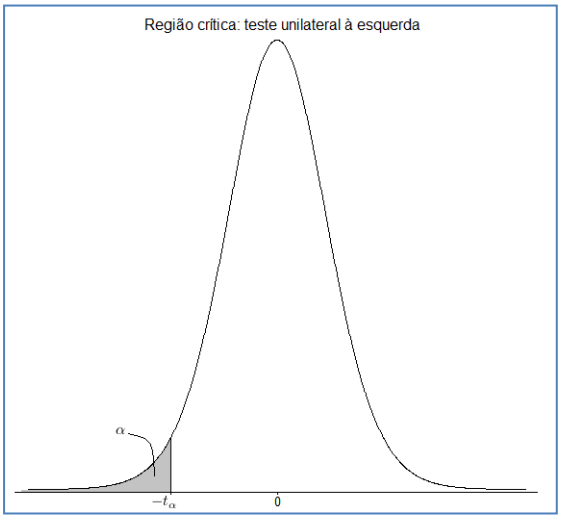
- Se o teste é unilateral, determinamos o ponto crítico t_α tal que $\mathbb{P}[T > t_\alpha] = \alpha$.

Figura-14: Teste Unilateral à Direita.



- Se o teste é unilateral à esquerda, determinamos o ponto $-t_\alpha$ tal que $\mathbb{P}[T < -t_\alpha] = \alpha$

Figura-15: Teste Unilateral à Esquerda.



4. Calcular, sob a hipótese nula, o valor:

$$T_{obs} = \frac{\overline{X} - \mu_0}{\frac{8}{\sqrt{n}}}$$

onde

- $\overline{X}$: valor da média amostral.
- μ_0: valor da média populacional sob a hipótese nula.
- s: valor do desvio padrão amostral.
- n: tamanho da amostra.

5. Critério:

- Teste bilateral: se $T_{obs} > t_{\alpha/2}$ ou se $T_{obs} > t_{\alpha/2}$ ou se $T_{obs} < -t_{-\alpha/2}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Teste unilateral à direita: se $T_{obs} > t_{\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Teste unilateral à esquerda: se $T_{obs} < -t_{\alpha/2}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.

6. O p-valor no teste bilateral é dado por

$$p - valor = \mathbb{P}[|t| > |T_{obs}||H_0 = 2\mathbb{P}[T > |T_{obs}||H_0]$$

Se o teste é unilateral à direita, o p-valor é dado por

$$p - valor = \mathbb{P}[T > T_{obs}|H_0]$$

e, se o teste é unilateral à esquerda, o p-valor é dado por

$$p - valor = \mathbb{P}[T < T_{obs}|H_0]$$

7. Como vimos anteriormente o intervalo de confiança é dado por

$$IC(\mu, 1-\alpha) = \left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

se o teste é bilateral. Se o teste é unilateral à direita, então o intervalo de confiança para o parâmetro μ é dado por

$$IC(\mu, 1-\alpha) = \left(\overline{X} - t_{\alpha}\frac{s}{\sqrt{n}}; \infty\right)$$

e, se o teste é unilateral à esquerda, então o intervalo de confiança para o parâmetro μ é dado por

$$IC(\mu, 1-\alpha) = \left(-\infty; \overline{X} + t_{\alpha}\frac{s}{\sqrt{n}}\right)$$

Exemplo-7:

Uma firma está convertendo as máquinas que aluga para uma versão mais moderna. Até agora foram convertidas 40 máquinas. O tempo médio de conversão foi de 24 horas, com desvio padrão de 3 horas.

a) Determine um intervalo de 98% de confiança para o tempo médio de conversão.

R. [22,895; 25,105]

b) O fabricante das novas máquinas afirma que a conversão em média dura no máximo 25 horas. Com base nas conversões feitas até o momento, e exigindo uma confiança de 99%, a afirmação do fabricante é verdadeira? R. Sim. Z = -2,1082

3.3 TESTE DE HIPOTESES PARA A PROPORÇÃO.

Pelo teorema central do limite, $\overline{X}$ terá distribuição aproximadamente normal, com média p e variância $\frac{p(1-p)}{n}$, ou seja,

$$\overline{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Observamos que $\overline{X}$ é um estimador de máxima verossimilhança para p, a proporção populacional, e, desse modo, para n suficientemente grande podemos considerar a distribuição amostral de

$\hat{p} = \overline{X}$ como aproximadamente normal:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Daí, temos que

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

Vejamos os passos para a construção do teste para proporção.

1. Estabelecer as hipóteses

$$\begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases} \quad \begin{cases} H_0: p = p_0 \\ H_1: p < p_0 \end{cases} \quad \begin{cases} H_0: p = p_0 \\ H_1: p > p_0 \end{cases}$$

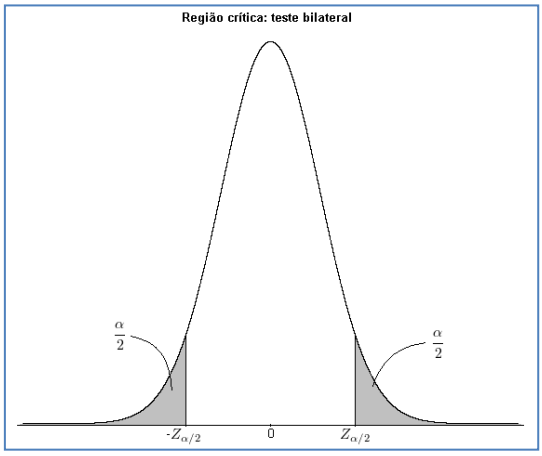
se o teste é bilateral, unilateral à esquerda ou unilateral à direita, respectivamente.

2. Fixar o nível de significância α.

3. Determinar a região crítica.

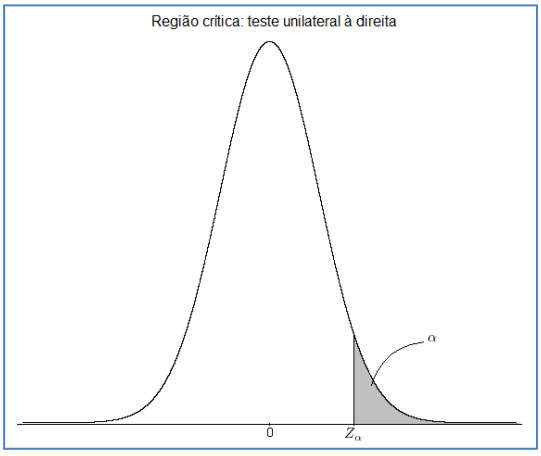
- Se o teste é bilateral, determinamos os pontos $-Z_{\alpha/2}$ e $Z_{\alpha/2}$ usando a tabela da distribuição normal, tais que $\mathbb{P}[Z > Z_{\alpha/2}] = \mathbb{P}[Z < -Z_{\alpha/2}] = \alpha/2$.

Figura-16: Teste Bilateral.



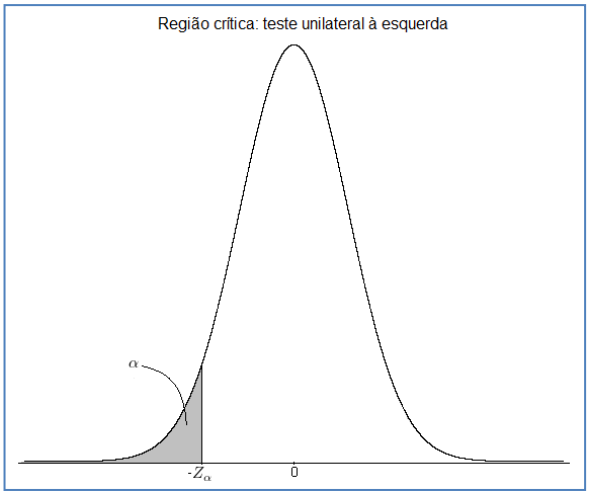
- Se o teste é unilateral à direita, determinamos o ponto crítico Z_α tal que $\mathbb{P}[Z > Z_\alpha] = \alpha$.

Figura-17: Teste Unilateral à Direita.



- Se o teste é unilateral à esquerda, determinamos o ponto crítico $-Z_\alpha$ tal que $\mathbb{P}[Z < -Z_\alpha] = \alpha$

Figura-18: Teste Unilateral à Esquerda.



4. Calcular, sob a hipótese nula, o valor

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

5. Critério:

- Se o teste é bilateral e $Z_{obs} > Z_{\alpha/2}$ ou $Z_{obs} < -Z_{\alpha/2}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Se o teste é unilateral à direita e $Z_{obs} > Z_\alpha$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Se o teste é unilateral à esquerda e $Z_{obs} < -Z_\alpha$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.

6. O p-valor é determinado por

$$p - valor = \mathbb{P}[|Z| > |Z_{obs}||H_0 = 2\mathbb{P}[Z > |Z_{obs}||H_0]$$

no teste bilateral. Se o teste é unilateral à direita, o p-valor é determinado por

$$p - valor = \mathbb{P}[Z| > |Z_{obs}|H_0]$$

e, se o teste é unilateral à esquerda

$$p - valor = \mathbb{P}[Z < |Z_{obs}|H_0]$$

7. Como foi visto anteriormente, o intervalo de confiança é dado por

$$IC(p, 1-\alpha) = \left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

se o teste é bilateral. Observamos aqui que o limite inferior do intervalo de confiança não pode ser inferior a zero e o limite superior não deve ser superior a um, uma vez que estamos calculando o intervalo de confiança para uma proporção e não faz sentido considerar uma proporção negativa ou maior do que um neste caso. No caso em que o teste é unilateral à direita, o intervalo de confiança para o parâmetro p é dado por

$$IC(p, 1-\alpha) = \left(\hat{p} - Z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; 1 \right)$$

e, se o teste é unilateral à esquerda, o intervalo de confiança para o parâmetro p é dado por

$$IC(p, 1-\alpha) = \left(0; \hat{p} + Z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Exemplo-8:

Um fabricante garante que 90% das peças que fornece à linha de produção de uma determinada fábrica estão de acordo com as especificações exigidas. A análise de uma amostra de 200 peças revelou 25 defeituosas. A um nível de 5%, podemos dizer que é verdadeira a afirmação do fabricante?

1. Estabelecemos as hipóteses

$$\begin{cases} H_0: p = 0{,}9 \\ H_1: p < 0{,}9 \end{cases}$$

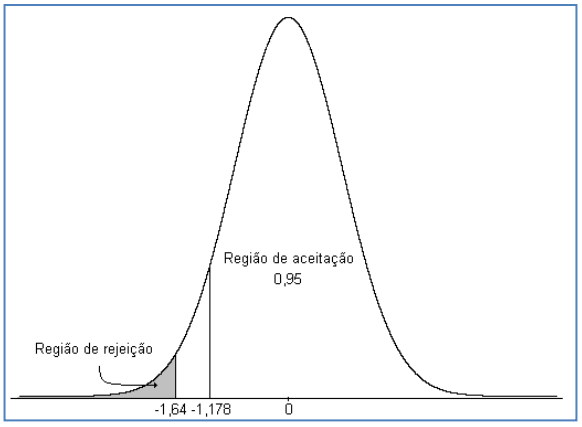
2. Fixemos o nível de significância $\alpha = 0{,}05$.

3. Como $\alpha = 0{,}05$, $-Z_{\alpha} = -1{,}64$.

4. Temos que $\hat{p} = 0{,}875$ e, sob a hipótese nula, $p_0 = 0{,}9$. Assim,

$$Z_{obs} = \frac{0{,}875 - 0{,}9}{\sqrt{(0{,}9)(0{,}1)/200}} = -1{,}178$$

Figura-19: Região Crítica e Região de Aceitação.



5. Conclusão: como $-1{,}64 = -Z_\alpha < Z_{obs} = -1{,}178$, não rejeitamos H_0. Portanto, temos evidências de que a afirmação do fabricante é verdadeira.

6. Vamos agora calcular o P-valor:

$$P - valor = \mathbb{P}[Z < Z_{obs}|H_0] = \mathbb{P}[Z < -1{,}178|H_0] = 0{,}1192$$

7. Como $n = 200$, $\hat{p} = 0{,}875$, $-Z_\alpha = -1{,}64$, temos que o intervalo de confiança é

$$\left(0; 0{,}875 + 1{,}64\sqrt{\frac{0{,}875(1-0{,}875}{200}}\right) = (0; 0{,}9134)$$

3.4 TESTE DE HIPOTESES PARA A VARIÂNCIA.

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de tamanho n retirada de uma população normal $N(\mu, \sigma^2)$. Suponha que desejamos testar uma hipótese sobre a variância σ^2 desta população. Sabemos que a estatística

$$Q = \frac{(n-1)s^2}{\sigma^2}$$

tem distribuição qui-quadrado com $n-1$ graus de liberdade. Denotamos $Q \sim X^2_{(n-1)}$. Para executar este tipo de teste, podemos seguir os passos:

1. Estabelecer uma das hipóteses (bilateral, unilateral à direita ou unilateral à esquerda)

$$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases} \quad \begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 > \sigma_0^2 \end{cases} \quad ou \quad \begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$$

OBS: As hipóteses H_0 podem ser substituídas por

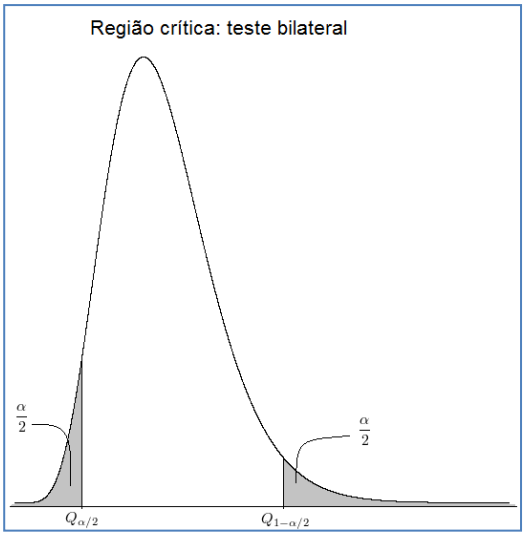
$H_0: \sigma^2 \geq \sigma_0^2$, $H_0: \sigma^2 \leq \sigma_0^2$, $H_0: \sigma^2 > \sigma_0^2$ ou $H_0: \sigma^2 < \sigma_0^2$.

2. Fixar o nível de significância α.

3. Determinar a região crítica.

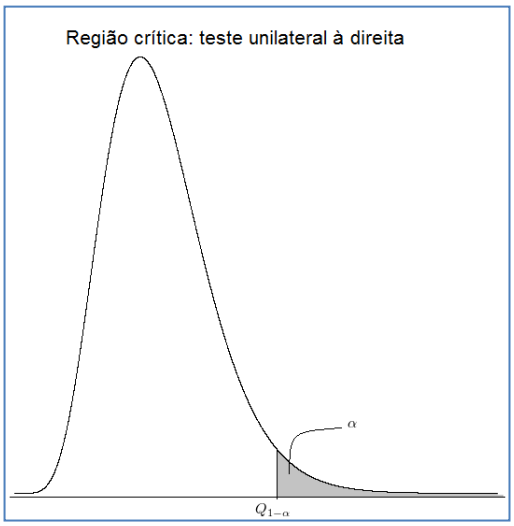
- Se o teste é bilateral, devemos determinar os pontos críticos $Q_{\alpha/2}$ e $Q_{1-\alpha/2}$ tais que
- $\mathbb{P}[Q < Q_{\alpha/2}] = \alpha/2$ e $\mathbb{P}[Q > Q_{1-\alpha/2}] = \alpha/2$ utilizando a tabela da distribuição qui-quadrado com $n-1$ graus de liberdade.

Figura-20: Teste Bilateral.



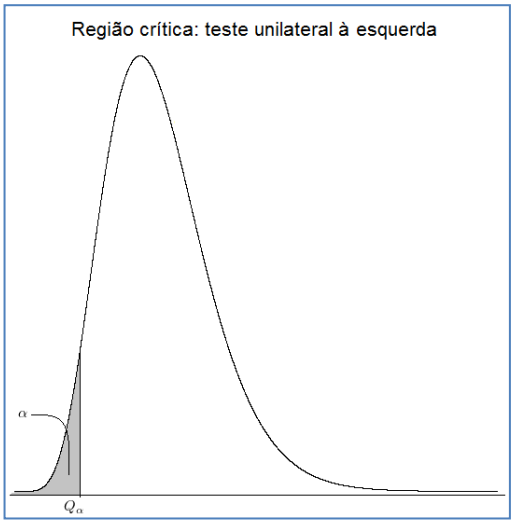
- Se o teste é unilateral à direita, devemos determinar o ponto crítico $Q_{1-\alpha}$ tal que $\mathbb{P}[Q > Q_{1-\alpha}] = \alpha$.

Figura-21: Teste Unilateral à Direita.



- Se o teste é unilateral à esquerda, devemos determinar o ponto crítico Q_{α} tal que $\mathbb{P}[Q < Q_{\alpha}] = \alpha$.

Figura-22: Teste Unilateral à Esquerda.



4. Calcular, sob a hipótese nula, o valor

$$Q_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

5. Critério:

(a) Teste bilateral: Se $Q_{obs} > Q_{\alpha/2}$ ou se $Q_{obs} < Q_{1-\alpha/2}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.

(b) Teste unilateral à direita: se $Q_{obs} > Q_{1-\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.

(c) Teste unilateral à esquerda: se $Q_{obs} < Q_{\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.

6. O p-valor é dado por

$$p - valor = 2\min(\mathbb{P}[Q > Q_{obs}|H_0], \mathbb{P}[Q < Q_{obs}|H_0])$$

no caso bilateral.

No caso unilateral à direita, o p-valor é dado por

$$p - valor = \mathbb{P}[Q > Q_{obs}|H_0]$$

e, no caso unilateral à esquerda, o p-valor é dado por

$$p - valor = \mathbb{P}[Q < Q_{obs}|H_0]$$

7. Como vimos na anteriormente, o intervalo de confiança para a variância populacional σ^2 é dado por

$$IC(\sigma^2, 1-\alpha) = \left(\frac{(n-1)s^2}{Q_{1-\alpha/2}}; \frac{(n-1)s^2}{Q_{\alpha/2}}\right)$$

se o teste é bilateral. Se o teste é unilateral à direita, o intervalo de confiança é dado por

$$IC(\sigma^2, 1-\alpha) = \left(\frac{(n-1)s^2}{Q_{1-\alpha}}; \infty\right)$$

e se o teste é unilateral à esquerda, o intervalo de confiança é dado por

$$IC(\sigma^2, 1-\alpha) = \left(0; \frac{(n-1)s^2}{Q_{\alpha}}\right)$$

Exemplo-9:

Uma máquina de preenchimento automático é utilizada para encher garrafas com detergente líquido. Uma amostra aleatória de 20 garrafas resulta em uma variância da amostra do volume de enchimento de $s^2nbsp; = 0{,}0153\ onça\ fluída^2$. Se a variância do volume de enchimento exceder $0{,}01\ onças\ fluídas^2$, existirá uma proporção inaceitável de garrafas cujo enchimento não foi completo ou foi em demasia. Há evidência nos dados da amostra sugerindo que o fabricante tenha um problema com garrafas com falta ou excesso de detergente? Use $\alpha = 0{,}05$ e considere que o volume de enchimentos tem distribuição normal.

O parâmetro de interesse é a variância da população

1. Primeiro vamos estabelecer as hipóteses:

$$\begin{cases} H_0: \sigma^2 = 0{,}01 \\ H_1: \sigma^2 > 0{,}01 \end{cases}$$

2. Como $\alpha = 0{,}05$ temos que $Q_{0{,}95} = 30{,}14$.

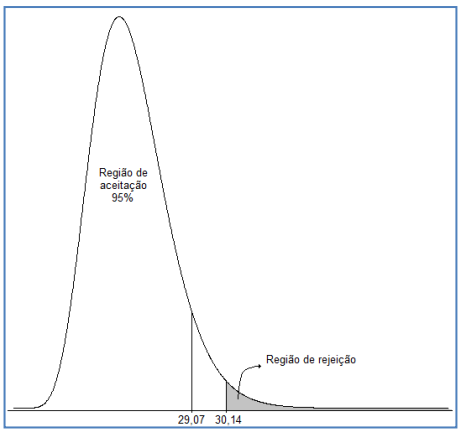
3. Critério: Rejeitar H_0 se $Q_{obs} > 30{,}14$

4. Calcular Q_{obs}, dado por

$$Q_{obs} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19(0{,}0153)}{0{,}01} = 29{,}07$$

5. Conclusão: como $Q_{obs} = 29{,}07 < 30{,}14$, a hipótese nula não deve ser rejeitada. Ou seja, não há evidências de que a variância do volume de enchimento exceda $0{,}01\ onça\ fluída^2$.

Figura-23: Teste Unilateral à Direita 95% de Confiança..



6. Vamos agora calcular o p-valor:

$$p - valor = \mathbb{P}[Q > Q_{obs}] = \mathbb{P}[Q > 29{,}07] = 0{,}064892$$

7. Como $n = 20$, $s^2 = 0{,}0153$ e $Q_{0,95} = 30{,}14$, segue que o intervalo de confiança para σ^2 com 95% de confiança é dado por

$$IC(\sigma^2, 95\%) = \left(\frac{(n-1)s^2}{Q_{0,95}}; \infty\right) = (0{,}00964, \infty)$$

EXERCÍCIOS PARA TREINAMENTO

Questão 1

Uma empresa produz saquinhos de salgadinhos de 500g. Para verificar se a máquina de empacotar está trabalhando corretamente o controle de qualidade tomou uma amostra de 50 saquinhos, que apresentou uma média amostral de 475g e desvio padrão amostral de 30g. Os dados obtidos proporcionam evidências suficientes para concluir que a máquina de empacotar não está trabalhando adequadamente (ou seja, a máquina empacota com pesos diferentes do proposto)? Realize o teste com α = 0,01. Observando o problema acima assinale a alternativa que representa a hipótese nula e a hipótese alternativa.

A) Ho: μ=475g e Ha: $\mu \neq$475g.

B) Ho: μ=475g e Ha: $\mu >$475g.

C) Ho: μ=475g e Ha: $\mu <$475g.

D) Ho: μ=500g e Ha: $\mu \neq$ 500g.

E) Ho: μ=500g e Ha: $\mu <$475g

Questão 2

Uma empresa produz saquinhos de salgadinhos de 500g. Para verificar se a máquina de empacotar está trabalhando corretamente o controle de qualidade tomou uma amostra de 50 saquinhos, que apresentou uma média amostral de 475g e desvio padrão amostral de 30g. Os dados obtidos proporcionam evidências suficientes para concluir que a máquina de empacotar não está trabalhando adequadamente (ou seja, a máquina empacota com pesos diferentes do proposto)? Realize o teste com α = 0,01. Após a realização do teste o que podemos concluir?

A) Rejeitamos a hipótese nula. A máquina não está trabalhando adequadamente.

B) Não rejeitamos a hipótese nula. A máquina não está trabalhando adequadamente.

C) Não rejeitamos a hipótese nula. A máquina está trabalhando adequadamente.

D) Rejeitamos a hipótese nula. A máquina está trabalhando adequadamente.

E) Nada podemos concluir.

Questão 3

Pesquisadores de uma clínica de emagrecimento desejam comparar a eficácia de uma dieta com exercícios contra uma dieta sem exercícios. Oitenta pacientes foram aleatoriamente selecionados e divididos em dois grupos. O primeiro grupo, com 35 pacientes foi colocado no programa de dieta com exercícios. O segundo grupo, com 45 pacientes, foi colocado no programa com dieta sem exercícios. Os resultados com a perda de peso, em quilogramas, após 4 meses, foram: Grupo 1:

média amostral de 8kg e desvio padrão amostral de 1,5 kg. Grupo 2: média amostral de 8,2 kg e desvio padrão amostral de 1,8kg. Determine com o nível de significância de 0,05, se existe diferença entre os dois tratamentos. Observe o problema acima e assinale a alternativa que representa a hipótese nula e a hipótese alternativa.

A) Ha: $\mu 1=\mu 2$ e Ho: $\mu 1 <\mu 2$

B) Ho: $\mu 1=\mu 2$ e Ha: $\mu 1>\mu 2$

C) Ho: $\mu 1=\mu 2$ e Ha: $\mu 1<\mu 2$

D) Ha: $\mu 1=\mu 2$ e Ho: $\mu 1 \neq \mu 2$

E) Ho: $\mu 1=\mu 2$ e Ha: $\mu 1 \neq \mu 2$

RESOLUÇÕES:

Resposta Questão 1

Gabarito:Letra D.

Resposta Questão 2

Gabarito:Letra A.

Resposta Questão 3

Gabarito:Letra E.

Capítulo 4

Correlação e Regressão Linear

4.1 DEFINIÇÕES:

- REGRESSÃO LINEAR

Em estatística ou econometria, **regressão linear** é uma equação para se estimar a condicional (valor esperado) de uma variável y, dados os valores de algumas outras variáveis x.

- CORRELAÇÃO LINEAR

Em probabilidade e estatística, **correlação**, dependência ou associação é qualquer relação estatística (causal ou não causal) entre duas variáveis e correlação é qualquer relação dentro de uma ampla classe de relações estatísticas que envolva dependência entre duas variáveis.

Uma medida do grau e do sinal da correlação é dada pela covariância entre as duas variáveis aleatórias X e Y que é uma medida numérica de associação linear existente entre elas, e definida por:

4.2 PARÂMETROS IMPORTANTES:

4.2.1 COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON:

Esse coeficiente serve para detectar padrões lineares (não vale para os padrões não lineares).

$$r = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2 \sqrt{n\sum y^2 - (\sum y)^2}}$$

O valor de r está sempre entre 1 e -1, ou seja $-1 \leq r \leq 1$.

Se r está próximo de 1, há uma forte correlação positiva.

Se r está próximo de -1, há uma forte correlação negativa.

Se r está próximo de 0, não há correlação linear.

Exercício 01

É esperado que a massa muscular de uma pessoa diminua com a idade. Para estudar essa relação, uma nutricionista selecionou 18 mulheres, com idade entre 40 e 79 anos, e observou em cada uma delas a idade (X) e a massa muscular (Y).

Massa muscular (Y)	Idade (X)
82.0	71.0
91.0	64.0
100.0	43.0
68.0	67.0
87.0	56.0
73.0	73.0
78.0	68.0
80.0	56.0
65.0	76.0
84.0	65.0
116.0	45.0
76.0	58.0
97.0	45.0
100.0	53.0
105.0	49.0
77.0	78.0
73.0	73.0
78.0	68.0

Calcule o coeficiente de correlação linear entre X e Y.

(Denotamos as variáveis: Y = Massa Muscular e X = Idade n=18)

$$\overline{X} = 61.556\ \overline{Y} = 85\ \sum_{i=1}^{18} X_i^2 = 70362\ \sum_{i=1}^{18} Y_i^2 = 133300\ \sum_{i=1}^{18} Y_i X_i = 91064$$

$$S_{XX} = \sum_{i=1}^{18} X_i^2 - 18\left(\overline{X}\right)^2 = 70362 - 18(61{,}556)^2 = 2157{,}460$$

$$S_{YY} = \sum_{i=1}^{18} Y_i^2 - 18\left(\overline{Y}\right)^2 = 133300 - 18(85)^2 = 3251$$

$$r = \frac{\sum_{i=1}^{18}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum_{i=1}^{18} X_i Y_i - 18\overline{X}\overline{Y}}{\sqrt{S_{XX}S_{YY}}} = \frac{91964 - 18(85)(61{,}556)}{\sqrt{(2157{,}460)(3250)}} = -0{,}837$$

O resultado demonstra que existe uma forte correlação negativa.

Exercício 02

Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 25 famílias.

Renda Familiar (X)	Gasto com Alimentação (Y)
3	1,5
5	2,0
10	6,0
10	7,0
20	10,0
20	12,0
20	15,0
30	8,0
40	10,0
50	20,0
60	20,0
70	25,0
70	30,0
80	25,0
100	40,0
100	35,0
100	40,0
120	30,0
120	40,0
140	40,0
150	50,0
180	40,0
180	50,0
200	60,0
200	50,0

(a)Calcular o coeficiente de correlação entre essas variáveis.

Denotamos as variáveis: Y = Gasto com Alimentação e X = Renda familiar

$$\overline{X} = 83{,}120 \quad \overline{Y} = 26{,}660 \quad \sum_{i=1}^{25} X_i^2 = 271934 \quad \sum_{i=1}^{25} Y_i^2 = 24899{,}250 \quad \sum_{i=1}^{25} Y_i X_i = 80774{,}500$$

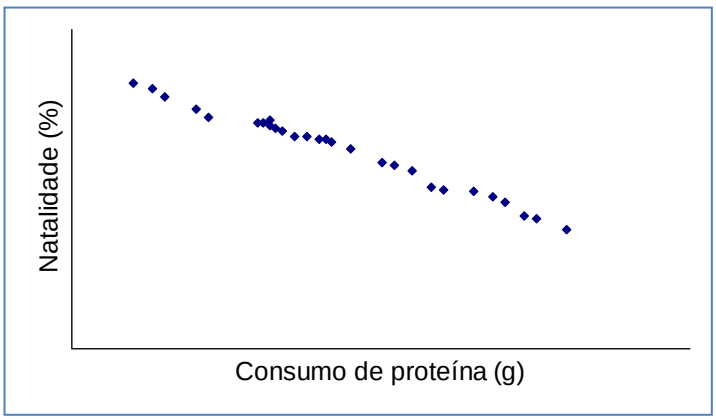
$$r = \frac{S_{XY}}{\sqrt{S_X S_Y}} = \frac{\sum_{i=1}^{25} X_i Y_i - 25\overline{X}\overline{Y}}{\sqrt{S_X S_Y}} = 0{,}954$$

O resultado demonstra que existe uma forte correlação positiva.

4.2.2 CORRELAÇÃO E CAUSA:

É importante salientar que o coeficiente de correlação define apenas o sentido da variação conjunta das variáveis. A observação que duas variáveis tendem variar simultaneamente em uma direção ou em direções contrárias, onde os dados provavelmente indicariam uma correlação positiva ou negativa, alta, não implicaria necessariamente na presença de uma relação de causa e efeito entre elas. Assim, na Figura 9 nota-se que existe uma correlação negativa entre o consumo de proteínas e o coeficiente de natalidade. Entretanto, isto não implica em afirmar que um aumento no consumo de proteínas determina redução da fertilidade. Portanto, uma correlação observada pode ser falsa (**correlação espúria**), isto é, pode ser devido a uma terceira e desconhecida variável causal.

Figura-24: Diagrama de dispersão para o consumo individual diário de proteínas de origem animal e a natalidade, em 28 países



4.2.3 REGRESSÃO LINEAR: ESTIMAÇÃO DE PARÂMETROS

Em experimentos que procuram determinar a relação existente entre duas variáveis, por exemplo, a dose de uma droga e a reação, concentração e densidade ótica, peso e altura, idade da vaca e a produção de leite, etc., dois tipos de situações podem ocorrer:

(a) uma variável (X) pode ser medida acuradamente e seu valor escolhido pelo experimentador. Por exemplo, a dose de uma droga a ser ministrada no animal. Esta variável é a **variável independente**. A outra variável (Y), dita **variável dependente ou resposta**, está sujeita a erro experimental e seu valor depende do valor escolhido para a variável independente. Assim, a resposta (reação, Y) é uma variável dependente da variável independente dose (X). Este é o caso da **Regressão**.

(b) as duas variáveis quando medidas estão sujeitas a erros experimentais, isto é, erros de natureza aleatória inerentes ao experimento. Por exemplo, produção de leite e produção de gordura medidas

em vacas em lactação, peso do pai e peso do filho, comprimento e a largura do crânio de animais, etc. Este tipo de associação entre duas variáveis constitui o problema da **Correlação**.

Atualmente, se dá à técnica de correlação uma importância menor do que a da regressão. Se duas variáveis estão correlacionadas, é muito mais útil estudar as posições de uma ou de ambas por meio de curvas de regressão, as quais permitem, por exemplo, a predição de uma variável em função de outra, do que estudá-las por meio de um simples coeficiente de correlação.

- Regressão linear simples

O termo regressão é usado para designar a expressão de uma variável dependente (Y) em função de outra (X), considerada independente. Diz-se regressão de Y em (sobre) X. Se a relação funcional entre elas é expressa por uma equação do 1º grau, cuja representação geométrica é uma linha reta, a regressão é dita linear.

Para introduzir a idéia de regressão linear simples, consideremos o seguinte exemplo:

Tabela 6: Tempo, em minutos, e quantidade de procaina[1] hidrolizada, em 10^{-5} moles/litro, no plasma canino.

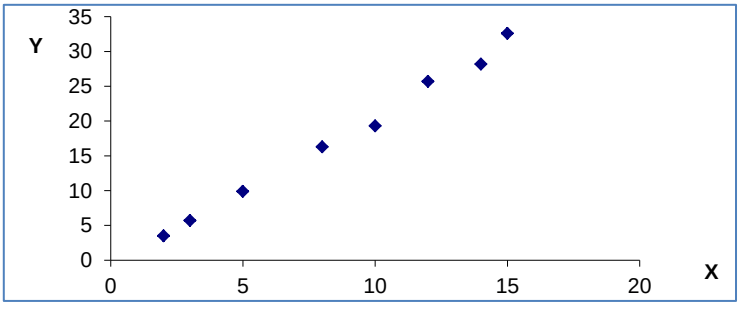
Tempo (X)		Quantidade hidrolizada (Y)	X .Y	X^2	Y^2
	2	3,5	7,0	4,0	12,3
	3	5,7	17,1	9,0	32,5
	5	9,9	49,5	25,0	98,0
	8	16,3	130,4	64,0	265,7
	10	19,3	193,0	100,0	372,5
	12	25,7	308,4	144,0	660,5
	14	28,2	394,8	196,0	795,2
	15	32,6	489,0	225,0	1062,8
Total	69	141,2	1589,2	767,0	3299,5

[1] anestésico local

A simples observação dos dados apresentados na Tabela 5, mostra que no intervalo estudado a quantidade de procaina hidrolizada varia em função do tempo.

Na resolução de problemas de regressão, o primeiro passo é traçar o **diagrama de dispersão** correspondente, marcando em um sistema cartesiano bidimensional os diversos pares de valores observados (x_i , y_i). Os dados da Tabela 1 estão apresentados na Figura 1.

Figura 24. Diagrama de dispersão dos dados da Tabela 6.



É fácil ver observando essa figura, que os pontos relativos aos dados de tempo e quantidade de procaina hidrolizada estão praticamente sobre uma reta. Parece então razoável estabelecer que a variação da quantidade de procaina hidrolizada (Y) pode ser considerada como uma função linear do tempo (X).

Postulada a existência de uma relação linear entre duas variáveis, pode-se representar o conjunto de pontos (x_i, y_i) pela equação da reta:

$$y = \alpha + \beta x + \varepsilon$$

que expressa o valor de Y como função do valor de X, onde ε, conhecido como *erro* ou *resíduo*, é a distância que um resultado y em particular se encontra da linha de regressão da população, representada pela equação:

$$E(y/x) = \alpha + \beta x,$$

em que **α** indica o intercepto da linha com o eixo do Y e **β** o coeficiente angular ou inclinação da reta.

Se ε [y – E(y/x)] é positivo, y é maior do que E(y/x); se é negativo, y é menor do que E(y/x); e a soma dos $\varepsilon_i's$ é igual a zero ($\sum \varepsilon_i = 0$). Logo, a média dos erros é nula, isto é, $E(\varepsilon_i) = 0$.

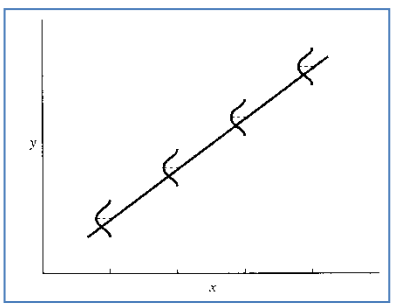
Como veremos a seguir, os parâmetros α e β da linha de regressão da população são estimados a partir da amostra aleatória de observações (x_i, y_i).

Considerando, então, que observações $x_1, x_2, ..., x_k$ sejam obtidas sobre a variável independente x, tal que $y_1, y_2, ..., y_k$ sejam as observações feitas sobre a variável dependente y, todas sujeitas a erros experimentais, pode-se querer saber como é que y varia, em média, para um dado x. Ou seja, como os y_s' variam aleatoriamente, deseja-se conhecer a distribuição do y quando x é conhecido. Isto é feito por meio da esperança condicionada de y dado x, simbolizada por E(y/x), que depende em geral de x. E(y/x) é também chamada de função de regressão de y em x.

A Figura 2 mostra as distribuições de y dados certos valores de x, supondo a função de regressão de y em x linear.

Modelo. A reta da Figura 2 é simbolizada por $E(y/x) = \alpha + \beta x$, onde α e β são os parâmetros a serem estimados.

Figura-25: Normalidade dos resultados y para determinado valor de x



A partir de agora, se o modelo acima for desenvolvido num contexto paramétrico, uma hipótese simplificadora e muito simples deve ser feita, a saber: a distribuição da variável aleatória y, para um dado x, é normal. Mais especificamente, fixado um x_i (X não é uma variável aleatória), os y_s' constituem variáveis independentes normais $N(\alpha + \beta x_i, \sigma^2)$; o que equivale dizer que as médias das distribuições de y/x estão sobre a verdadeira reta $\alpha + \beta x$ ou seja, $E(y_i) = E(\alpha) + E(\beta x_i) + E(\varepsilon_i) = \alpha + \beta x_i$, onde $E(\varepsilon_i) = 0$, e que para um dado valor de x, a variância do erro é sempre σ^2, denominada variância residual, isto é, $E[y_i - E(y_i/x_i)]^2 = E(\varepsilon_i)^2 = \sigma^2$ (propriedade homocedástica). Estes conceitos estão ilustrados na Figura 2. À parte do fato que σ^2 é desconhecido, a reta na qual as médias estão localizadas é também desconhecida. Assim, um objetivo importante da análise estatística é estimar os parâmetros α e β para que se conheça totalmente a função de regressão E(y/x). A teoria mostra que a melhor maneira de estimá-los é por meio do **método dos quadrados mínimos**, que consiste em minimizar a soma dos quadrados das distâncias $y_i - \hat{y}_i$, onde $\hat{y}_i = a + bx_i$ representa a equação de regressão estimada, tal que $a = \hat{\alpha}$ e $b = \hat{\beta}$ são os estimadores de α e β, respectivamente.

Sendo, então, $y_i - \hat{y}_i$ a diferença entre o valor observado e o estimado pela equação de regressão para cada observação, a qual é rotulada por e_i, procura-se estimar α e β, de modo que $\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$ seja o menor possível. As diferenças $e_i = y_i - \hat{y}_i$ são chamadas "desvios da regressão" ou "erros de estimativas". Se todos os desvios (e_i) são iguais a zero, implica que cada ponto (x_i, y_i) se encontra diretamente sobre a linha ajustada; os pontos estão tão próximos quanto possíveis da linha.

Estimadores. Dado um conjunto de n pares de observações (x_1, y_1), (x_2, y_2), ... , (x_n, y_n), pode-se mostrar usando métodos de cálculo infinitesimal não utilizado aqui, que os estimadores de quadrados mínimos são:

$$b = \hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad a = \hat{\alpha} = \bar{y} - b\bar{x}$$

Dividindo-se o numerador e o denominador de b por $(n - 1)$, vê-se que

b é denominado coeficiente de regressão de Y em X; simboliza-se por $b_{Y.X}$

$$b = \frac{Cov(X,Y)}{s_X^2} = \frac{[\sum (x_i - \bar{x})(y_i - \bar{y})]/n - 1}{\sum (x_i - \bar{x})^2]/n - 1}$$

Fórmulas de cálculo:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Note-se que, além da suposição da normalidade do y, outras hipóteses usadas pelo método de mínimos quadrados são:

- para qualquer valor específico de x, $\sigma_{y/x}$, o desvio padrão dos resultados y, não se modifica.

Esta hipótese de variabilidade constante em todos os valores de x é conhecida como homoscedasticidade, e

(b) a relação (verdadeira) entre y e x é suposta linear; mais claramente, E(y/x) = α + βx.

$$b_{YX} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

A fórmula de cálculo acima pode ser melhor trabalhada e ficaria expressa como:

$$b = \frac{n\sum x.y - (\sum x).(\sum y)}{n\sum x^2 - (\sum x)^2}$$

e $a = \bar{y} - b\bar{x}$ sendo a equação de regressão: $\hat{Y}$ = a + b.x

Para traçar a reta de regressão, basta dar valores quaisquer para X dentro do intervalo estudado e calcular os respectivos valores de $\hat{Y}$ (Figura 3). Os valores calculados de $\hat{Y}$ não coincidem necessariamente com os valores observados de Y. A curva resultante é denominada de regressão de Y para X, visto que Y é avaliado a partir de X. O mais importante objetivo de um estudo de regressão é usar o modelo linear desenvolvido para estimar a resposta esperada correspondente a um valor futuro.

- Coeficiente de Determinação: O coeficiente de determinação, também chamado de R^2, é uma medida de ajustamento de um modelo estatístico linear generalizado, como a regressão linear, em relação aos valores observados. O R^2 varia entre 0 e 1, indicando, em percentagem, o quanto o modelo consegue explicar os valores observados. Quanto maior o R^2, mais explicativo é o modelo, melhor ele se ajusta à amostra. Por exemplo, se o R^2 de um modelo é 0,8234, isto significa que 82,34% da variável dependente consegue ser explicada pelos regressores presentes no modelo.

Sendo:

SQE = Soma dos quadrados dos Erros.

SQT = Soma dos Quadrados Total Corrigida.

SQR = Soma dos Quadrados da Regressão.

Sendo SQT = SQR + SQE ------→ SQR = SQT – SQE -----→

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} = \frac{\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

ou seja, é a razão entre a soma de quadrados da regressão e a soma de quadrados total. No modelo com intercepto, podemos escrever

$$R^2 = \frac{\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i \sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{(\sum_{i=1}^{n}(x_1 - \overline{x})Y_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Notemos que

$$0 \leq R^2 \leq 1$$

O R^2 é, portanto, uma medida descritiva da qualidade do ajuste obtido. Em geral, referimo-nos ao R^2 como a quantidade de variabilidade nos dados que é explicada pelo modelo de regressão ajustado. Entretanto, o valor do coeficiente de determinação depende do número de observações *(n)*, tendendo a crescer quando n diminui. Se $n = 2$, tem-se sempre $R^2 = 1$

O R^2deve ser usado com precaução, pois é sempre possível torná-lo maior pela adição de um número suficiente de termos ao modelo. Assim, se, por exemplo, não há dados repetidos (mais do que um valor y para um mesmo x) um polinômio de grau $(n - 1)$ dará um ajuste perfeito

$R^2 = 1$ para n dados. Quando há valores repetidos, o R^2 não será nunca igual a 1, pois o modelo não poderá explicar a variabilidade devido ao erro puro.

Obs.: O Coeficiente de Determinação pode ser calculado simplesmente elevando o Coeficiente de Correlação Linear de Pearson ao quadrado.

- Coeficiente de Determinação Ajustado

Para evitar dificuldades na interpretação de R^2, alguns estatísticos preferem usar o R_a^2 (R^2 ajustado), definido para uma equação com 2 coeficientes como:

$$R_a^2 = 1 - (\frac{n-1}{n-2})(1 - R^2)$$

Assim como o Coeficiente de Determinação R^2, quanto maior R_a^2, mais a variável resposta é explicada pela regressora X.

Exercício 03

Um motorista deseja prever seus gastos com seu automóvel em função dos quilômetros que roda por mês.

3203	400
3203	400
2603	340
3105	400
1305	150
804	100
1604	200
2706	300
805	100
1903	200
3203	400
3702	450
3203	400
3203	400
803	100
803	100
1102	130
3202	400
1604	150
1603	200
3203	400
3702	450
3403	440

Estatística de Regressão	
R múltiplo	0,9931
R-Quadrado	0,9862
R-quadrado ajustado	0,9855
Erro padrão	127,51
Observações	23

Observando a tabela acima, percebe-se uma forte correlação entre as variáveis, onde R está muito próximo de 1. Quilômetros rodados explica 98% da variância de gastos.

Exercício 04 – Retornando aos dados do Exercício 02

Os dados a seguir correspondem à variável renda familiar e gasto com alimentação (em unidades monetárias) para uma amostra de 25 famílias.

Renda Familiar (X)	Gasto com Alimentação (Y)
3	1,5
5	2,0
10	6,0
10	7,0
20	10,0
20	12,0
20	15,0
30	8,0
40	10,0
50	20,0
60	20,0
70	25,0
70	30,0
80	25,0
100	40,0
100	35,0
100	40,0
120	30,0
120	40,0
140	40,0
150	50,0
180	40,0
180	50,0
200	60,0
200	50,0

Denotamos as variáveis: Y = Gasto com Alimentação e X = Renda familiar

$$\overline{X} = 83{,}120 \quad \overline{Y} = 26{,}660 \quad \sum_{i=1}^{25} X_i^2 = 271934 \quad \sum_{i=1}^{25} Y_i^2 = 24899{,}250 \quad \sum_{i=1}^{25} Y_i X_i = 80774{,}500$$

Obtenha a equação de regressão do gasto com alimentação em função da renda familiar.

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{25} X_i Y_i - 25\overline{X}\,\overline{Y}}{S_{XX}} = \frac{80774{,}5 - 25(83{,}12)(26{,}66)}{271934 - 25(83{,}12)^2} = 0{,}256$$

e

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 26{,}66 - 0{,}256(83{,}120) = 5{,}380$$

A reta de regressão estimada da variável Gasto de alimentação (Y) em função da Renda familiar (X) é

$$\hat{Y} = 5{,}380 + 0{,}256X$$

Qual o significado prático do valor da inclinação da reta de regressão do item (c)?

O valor $\hat{\beta}_1$ =0,256 significa que estima-se que para cada aumento de uma unidade monetária da renda familiar ocorre um acréscimo em média de 0,256 unidades no gasto

EXERCÍCIOS PARA TREINAMENTO

Questão 1

Uma agência de turismo estudou a demanda de passagem sem relação à variação do preço de venda e obteve os valores da tabela a seguir:

Preço de Venda (x)	33	25	24	18	12	10	8	4
Demanda de Passagens (y)	300	400	500	600	700	800	900	1000

Preço de Venda (x)	33	25	24	18	12	10	8	4
Demanda de Passagens (y)	300	400	500	600	70	800	900	1000

Preencha a tabela a seguir e calcule o coeficiente de correlação linear por meio da fórmula do coeficiente de correlação de Pearson.

x	y	$x.y$	x^2	y^2
Total				

R = -0,62

Questão 2

Uma agência de viagens realizou um estudo sobre as passagens de avião que vendeu nos últimos meses e a soma de horas trabalhadas por todos seus funcionários (lembre que o número de

funcionários é variável). Calcule o coeficiente de correlação linear pelo coeficiente de correlação de Pearson.

Meses	Horas Trabalhadas	Passagens
	x	y
Janeiro	1378	154
Fevereiro	1292	146
Março	1146	110
Abril	854	98
Maio	973	105
Junho	996	118
Julho	1241	143
Agosto	1208	105
Setembro	1045	112
Total		

R=0,8227

Questão 3

Considere os valores da tabela a seguir e calcule o coeficiente de correlação linear por meio da fórmula do coeficiente de correlação de Pearson.

x	3	5	8	13	16	17	20	22
y	6	17	27	20	45	28	34	53

R= 0,8464

Questão 4

Como resultado de um experimento foram obtidos os seguintes valores para a função f(x)

x	-1	0	1	2	3	4	5	6
F(x)	10	9	7	5	4	3	0	-1

Determinar qual é a melhor reta g(x)=ax+b, que ajusta esses pontos através do método da Regressão Linear

Lembre que: $$\begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_1 \\ \sum_{i=1}^{n} x_1 & \sum_{i=1}^{n} x_1^2 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} f(x_i) \\ \sum_{i=1}^{n} f(x_i)\, x_i \end{bmatrix}$$

r: y = -1,6071 + 8,6427

Questão 5

Dada a tabela de pontos experimentais:

X	1	2	3	4	5
F(x)	2,2	3,3	4,2	5,1	6,3

Obtenha a reta que melhor ajusta os pontos através do método da Regressão Linear.

R: y = x + 1,22

Referências

REFERÊNCIAS

[1] DEVORE, J. L. Probabilidade e Estatística para Engenharia e Ciências. Ed. Thomson, 2006.

[2] COSTA NETO, PEDRO. L. O. Estatística. Ed. Edgard Blücher, LTDA. 2002.

[3] GOODE, WILLIAM J.& HATT, PAUL K. Métodos em Pesquisa Social. Ed. Companhia Editora Nacional, 1979.

[4] HINES, W. & MONTGOMERY, D. C. Probability and Statistics in Engineering and Management Science. Ed. Wiley, 1990.

[5] MONTGOMERY, D. C. & RUNGER, G. C. Estatística Aplicada e Probabilidade para Engenheiros. LTC, 2009

[6] MOOD A. M, GRAYBILL F., BOES, D. C. Introduction to the Theory of Statistics. Editora McGraw-Hill, 1974.

[7] SIEGEL, SIDNEY. Estatística Não – Paramétrica. Para Ciências do Comportamento. Ed.McGraw-Hill, 1979.

[8] WALPOLE, R. ; MYERS, R.; MYERS, S & YE, K. Probabilidade e Estatística para Engenharia e Ciências. Ed. Pearson, 2009.

[9] BARBETTA, Pedro Alberto. Estatística Aplicada às Ciências Sociais, Ed. UFSC, 5ª Edição, 2002.

[10] http://www.portalaction.com.br

Agência Brasileira do ISBN
ISBN 978-85-7042-028-2
9 788570 420282