Ciências Contábeis
UNIVERSIDADE FEDERAL DA BAHIA

José Sérgio Casé de Oliveira

MATF10

Estatística Aplicada às Ciências Sociais Aplicadas II

ESTATÍSTICA APLICADA ÀS CIÊNCIAS SOCIAIS APLICADAS II

UNIVERSIDADE FEDERAL DA BAHIA
FACULDADE DE CIÊNCIAS CONTÁBEIS
BACHARELADO EM CIÊNCIAS CONTÁBEIS

ESTATÍSTICA APLICADA ÀS CIÊNCIAS SOCIAIS APLICADAS II

José Sérgio Casé de Oliveira

Salvador, 2018

UNIVERSIDADE FEDERAL DA BAHIA
Reitor: João Carlos Salles Pires da Silva
Vice-Reitor: Paulo César Miguez de Oliveira

Pró-Reitoria de Ensino de Graduação
Pró-Reitor: Penildon Silva Filho

Faculdade de Ciências Contábeis
Diretor: Prof. Joséilton Silveira da Rocha

Superintendência de Educação a Distância -SEAD
Superintendente: Márcia Tereza Rebouças Rangel

Coordenação de Tecnologias Educacionais
CTE-SEAD
Haenz Gutierrez Quintana

Coordenação de Design Educacional
CDE-SEAD
Lanara Souza

Coordenadora Adjunta UAB
Andréa Leitão
UAB -UFBA

Bacharelado em Ciências Contábeis EaD
Coordenadora:
Profª Inês Teresa Lyra Gaspar da Costa

Produção de Material Didático
Coordenação de Tecnologias Educacionais
CTE-SEAD

Núcleo de Estudos de Linguagens & Tecnologias - NELT/UFBA

Coordenação
Prof. Haenz Gutierrez Quintana

Projeto gráfico
Prof. Haenz Gutierrez Quintana
Projeto da Capa: Prof. Alessandro Faria

Arte da Capa: Alessandro Faria

Foto de capa: Designed by mindandi / Freepik

Revisão:
Júlio Pereira Neves

Equipe Design - Supervisão
Alessandro Faria
Editoração / Ilustração
Moema Baião
Ariana Santana
Marcos Nascimento
Design de Interfaces
Raissa Bomtempo
Equipe Audiovisual
Direção:
Prof. Haenz Gutierrez Quintana
Produção:
Letícia Moreira de Oliveira
Ana Paula Borges
Câmera
Maria Christina Souza
Edição:
Victor Foseca
Imagens de cobertura:
Marcone Silva | Freepik
Animação e videografismos:
Dominique de Andrade Santos
Edição de áudio
Greice Mara
Trilha Sonora:
José Balbino

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFBA

O48

Oliveira, José Sérgio Casé de.
Estatística aplicada às ciências sociais aplicadas II / José Sergio Casé de Oliveira. - Salvador: UFBA, Faculdade de Ciências Contábeis; Superintendência de Educação a Distância, 2018.
112 p. : il.

Esta obra é um Componente Curricular do Curso de Bacharelado em Ciências Contábeis na modalidade EaD da UFBA/SEAD/UAB.

ISBN: 978-85-8292-162-3

1. Estatística matemática. 2. Amostragem (Estatística). 3. Análise multivariada. 4. Ciências sociais - Métodos estatísticos. I. Universidade Federal da Bahia. Faculdade de Ciências Contábeis. II. Universidade Federal da Bahia. Superintendência de Educação a Distância. III. Título.

CDU: 519.2

SUMÁRIO

CARTA DE APRESENTAÇÃO DA DISCIPLINA

Prezado(a) estudante,

Vivemos a era da informação, nunca se coletaram e analisaram tantos dados como atualmente. E isso segue como uma tendência crescente, espera-se que com o passar do tempo e com os avanços tecnológicos, cada vez seja possível coletar e analisar mais e mais informação. Nesse contexto, é de fundamental importância para qualquer profissional ser capaz de utilizar toda essa informação a seu favor. Para que isso seja feito de forma eficiente, é imprescindível o conhecimento de estatística, mais precisamente, métodos estatísticos capazes guiar a tomada de decisão.

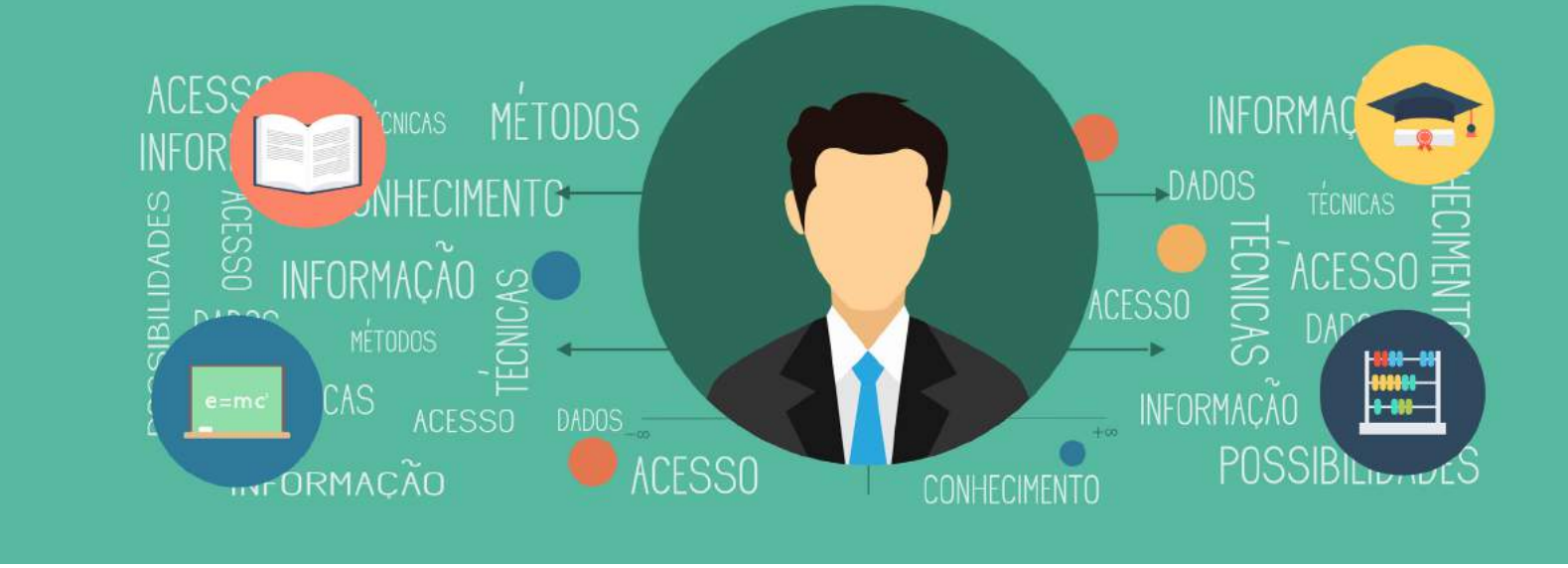
Pensando nisso, esse módulo foi confeccionado a fim de possibilitar um acesso suave a algumas das principais técnicas para extração de informação relevante a partir de dados. O objetivo principal aqui é apresentar uma série de ferramentas estatísticas que podem ser utilizadas em diferentes contextos e que são capazes de fornecer respostas sobre questões normalmente complexas de se avaliar.

Aqui serão apresentadas algumas etapas fundamentais para se conseguir uma informação relevante e confiável a partir de dados. Serão discutidas formas de coleta de dados e a importância dessa etapa. Serão também apresentadas técnicas para extração de informação considerando um ou mais bancos de dados.

O modelo está organizado em cinco unidades. A primeira unidade introduz noções gerais sobre amostragem e sua importância. A segunda unidade apresenta técnicas de estimação, com ênfase em estimação intervalar. A terceira unidade traz conceitos básicos sobre testes de hipóteses, discutindo diversos testes bem como suas aplicações. A quarta unidade introduz uma visão geral sobre análise de regressão simples. E por fim, a quinta unidade apresenta noções sobre estatística multivariada.

Bons estudos!

José Sérgio Casé de Oliveira



MINI CURRICULUM DO PROFESSOR

O Prof. José Sérgio Casé de Oliveira é bacharel em ciências econômicas pela Universidade Federal de Pernambuco, mestre em estatística, também pela Universidade Federal de Pernambuco, e doutor em economia pela Universidade Federal da Paraíba. Atualmente é professor da Faculdade de Ciências Contábeis da Universidade Federal da Bahia. Tem experiência com pesquisas em econometria, macroeconomia, distribuições de probabilidade e estatística computacional.

UNIDADE 1 - NOÇÕES GERAIS SOBRE AMOSTRAGEM

Os estudos que se utilizam de estatística normalmente seguem uma série de etapas para sua realização. Essas etapas, por vezes, são chamadas de método estatístico, e podem ser apresentadas como:

1. Delimitação do problema
2. Planejamento
3. Coleta de dados
4. Organização e apresentação dos dados
5. Análise e Interpretação dos resultados

Nesse capítulo, estamos especialmente interessados na etapa 3. Do método estatístico. A coleta de dados nada mais é do que o passo por meio do qual se obtém a informação relevante sobre o objeto de estudo. Discutiremos aqui a importância desse passo para o bom funcionamento da investigação por meio do método estatístico, e veremos algumas das principais técnicas para coleta de dados, as quais são comumente chamadas de amostragem.

1.1 Alguns conceitos importantes

Antes de mais nada é importante se estabelecer alguns conceitos preliminares.

Será denominada de **População** o conjunto de todos os elementos a serem estudados. Por exemplo, imagine que se deseja saber a idade média de pessoas fumantes do estado da Bahia, logo, a população de interesse para o estudo são todos os indivíduos fumantes residentes do estado da Bahia.

Chamaremos de **Amostra** um subconjunto população. Considerando o exemplo anterior, uma amostra dos indivíduos fumantes do estado da Bahia seria um subconjunto do total de indivíduos fumantes do estado da Bahia.

Chamaremos de **Amostragem** o procedimento para obtenção da amostra a partir de uma dada população.

Por fim, será chamado de **Parâmetro** uma medida numérica que descreve uma determinada característica da população. Por exemplo, considere a variável "nota final dos alunos de Estatística II". O número de alunos que também será o número de notas final é um parâmetro desta população. A média das notas final também é um parâmetro desta população.

Admita que a quantidade de elementos que compõe a população é N, e que a quantidade de elementos que compõe a amostra é n. Posto que a amostra é um subconjunto da população, ela sempre terá um número de indivíduos menor que o da população, tal que $N > n$.

Ex. Exemplos

Exemplo1.1: Suponha que se deseja descobrir a nota média dos alunos do 6º período do curso de ciências contábeis da UFBA no semestre 2017.2.

Delimitamos aqui nossa população:

População: Todos os alunos do 6º período do curso de ciências contábeis da UFBA no semestre 2017.2.

Suponha ainda que todos os alunos matriculados no 6º período do curso de ciências contábeis da UFBA no semestre 2017.2 somaram 20 indivíduos. Logo, tem-se que $N = 20$.

Por fim, admita que são selecionados 4 alunos dessa população para realização do cálculo da média.

Delimitamos aqui nossa amostra:

Amostra: 4 dos 20 alunos do 6º período do curso de ciências contábeis da UFBA no semestre 2017.2.

Note que isso implica $n = 4$. Note ainda que $N > n$ e que, como os 4 alunos selecionados na amostra, foram retirados do conjunto população, esta amostra é um subconjunto da população.

A forma como é feita a escolha dos 4 elementos da população para compor a amostra é a amostragem.

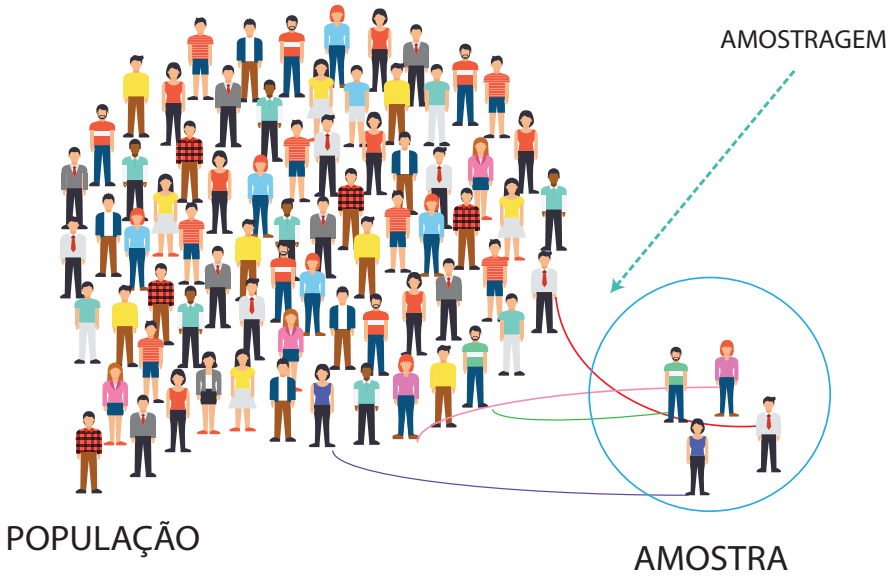


Figura 1: Relação entre população, amostragem e amostra.

Ilustração: Ariana Santana

Como a população é o conjunto de todos os elementos a serem estudados, obviamente ela contém toda informação necessária de interesse. Posto isso, por que não utilizar a população ao invés da amostra?

A motivação para utilização de amostragem é bastante óbvia, claramente a coleta de informação sobre toda uma população demanda muito mais tempo e custos para sua realização. Além disso, por vezes é inviável a obtenção de informações sobre cada unidade da população. Assim a possibilidade de conseguir informações relevantes sobre a população a partir de uma amostra é bastante atraente.

Existem algumas características que são importantes para que se tenhamos uma amostragem de boa qualidade.

Primeiramente, a amostragem deve ser capaz de fornecer uma amostra **representativa**. Mas o que vem a ser uma amostra representativa? É uma amostra que contém as características relevantes da população.

Ex. Exemplos

Exemplo1.2: Suponha que a idade média dos cidadãos baianos seja igual a 36 anos. Uma amostra representativa para população dos cidadãos baianos deve ter idade média aproximadamente igual a 36 anos.

Outra característica importante é a imparcialidade da amostragem, ou seja, o procedimento de amostragem deve ser conduzido de tal forma que todos os elementos da população devem ter igual oportunidade de compor a amostra.

Ex. Exemplos

Exemplo1.3: Considere que se deseja obter informações sobre a média de chuvas no mês de janeiro no estado da Bahia. Cidades com maior intensidade de chuvas devem ter as mesmas chances de compor a amostra que cidades com menos intensidade de chuvas, caso contrário, a amostra não será capaz de representar bem essa característica da população.

Outro conceito importante é o **Erro** associado ao processo de utilização da amostra. Como se deseja obter informações a partir de um subconjunto da população, é natural que este, por mais representativo que seja, não seja capaz de carregar consigo todas as características inerentes à população como um todo. Na prática, quando se colhe uma amostra, por mais rigoroso que seja o procedimento adotado na coleta, esta pode não ser perfeitamente representativa para população. Podemos subdividir esse erro em duas categorias, o **erro não amostral** e o **erro amostral**.

O erro não amostral, é o erro associado ao procedimento utilizado para obtenção da amostra. Normalmente está associado a não imparcialidade da amostra. Esse erro pode ser minimizado utilizando técnicas de amostragem imparciais.

O erro amostral está associado a incapacidade da amostra de conseguir representar de forma precisa as características intrínsecas a população. Ou seja, por melhor que seja o procedimento de amostragem, na prática é muito improvável que a estimativa obtida a partir da amostra seja idêntica ao parâmetro real da população. Se, por exemplo, for coletada informação referente à altura de todos os homens do Brasil, e à média proveniente dessa população for igual a 1,74 m, dificilmente a média obtida a partir de uma amostra da altura dos homens brasileiros será exatamente igual a 1,74 m. Além disso, diferentes

amostras sobre uma mesma população podem gerar resultados diferentes, mesmo utilizando técnicas similares para obtenção da amostra.

Matematicamente podemos definir o erro amostral como a diferença entre a estimativa obtida a partir da amostra e o parâmetro populacional ().

ERRO AMOSTRAL = ESTIMATIVA - θ.

Ex. Exemplos

Exemplo1.4: Suponha que determinada cidade tem 10 empresas (população), as quais tiveram os seguintes lucros (em milhões) no ano de 2016.

A = -1,70; B = -1,45; C = -0,42; D = 0,04; E = 0,39;

F = 0,48; G = 1,31; H = 1,63; I = 1,70; J = 2,00.

Note que a média de lucros das empresas dessa cidade é aproximadamente igual a 0,40.

Se para compor minha amostra, eu escolho de forma parcial apenas as 5 primeiras empresas que tiveram desempenhos abaixo da média, A, B, C, D e E, vou ter como resultado que o lucro médio da cidade foi aproximadamente igual a -0.63. Note que isso é muito distante da realidade que é um lucro médio de 0,40. Este erro está associado ao erro não amostral, posto que se deve a forma como minha amostra foi colhida.

Note que se eu escolher, mais uma vez de forma parcial, as 5 empresas com melhor desempenho (F, G, H, I e J) será obtida também uma média de lucros incompatível com a realidade.

Por outro lado, eu posso ser imparcial na escolha da amostra (controlando assim o erro não amostral), utilizando por exemplo, um gerador aleatório de números para selecionar as empresas. Admita que o gerador sorteia as empresas G, F, B, J e C, que passam a compor nossa amostra. Note que a média de lucros da cidade, considerando essas empresas, é aproximadamente igual a 0,38, um resultado similar a média de lucros real. De forma que o erro amostral pode ser estimado em

ERRO AMOSTRAL=ESTIMATIVA- θ

ERRO AMOSTRAL=0,38 – 0,4 = -0,02.

Admita uma segunda amostra obtida também a partir de um gerador aleatório de números composta por C, A, H, E e I. De forma que o lucro médio das empresas da cidade considerando essa amostra é igual a 0,32. Logo,

$$\text{ERRO AMOSTRAL}=\text{ESTIMATIVA}-\theta$$

$$\text{ERRO AMOSTRAL}=0{,}32-0{,}4=-0{,}08.$$

Note que mesmo se utilizando do mesmo procedimento de amostragem obtivemos diferentes erros amostrais.

1.2 Métodos de Amostragem

Normalmente, os métodos de amostragem são divididos em dois grupos. A amostragem é denominada **probabilística** se todos os elementos da população possuem probabilidade conhecida e diferente de zero, de pertencer à amostra. Caso contrário, a amostragem é chamada **não probabilística**.

Os métodos de amostragem não probabilística se utilizam de algum critério não probabilístico para seleção da amostra, por exemplo, o pesquisador pode optar por compor sua amostra a partir de indivíduos de fácil acesso, a fim de se obter ganhos em termos de tempo e custos, prezando pela conveniência.

Entretanto, utilizando-se uma amostra não probabilística não é possível generalizar os resultados da pesquisa para a população, uma vez que amostra não é imparcial e/ou representativa.

Para realizar inferências ou induções sobre a população com base em uma determinada amostra, é necessário que esta tenha sido obtida a partir de um processo de amostragem probabilística. Esse método embasa a posterior utilização de técnicas estatísticas sobre a amostra. Por ser a forma de amostragem mais indicada, daremos enfoque aqui a amostragem probabilística.

1.2.1 Amostragem aleatória simples (AAS)

Este é o procedimento mais elementar, neste tipo de amostragem, deve-se garantir que todos os elementos de uma população de tamanho N tenham a mesma probabilidade de serem selecionados. O procedimento consiste basicamente em:

i. Rotula-se todos os elementos da população;
ii. Sorteia-se aleatoriamente e sem reposição um elemento dessa população;
iii. Repete-se o sorteio até que se obtenha n elementos para compor a amostra.

Ex. Exemplos

Exemplo1.5: Considere uma população composta por 6 indivíduos, a saber, João, Marilia, Luiz, Fernanda, Maria e Camila.

Note que N=6. Imagine que se deseja obter uma amostra de tamanho 3 (n=3) com base nessa população. Rotulando os elementos da população, temos como possibilidade

João=1, Marilia=2, Luiz=3, Fernanda=4, Maria=5 e Camila=6.

Agora utilizando um gerador de números aleatórios, sorteamos 3 números no intervalo contínuo de 1 a 6. Suponha que o resultado do sorteio é 2,3 e 6 (Esse sorteio pode ser feito ainda utilizando um dado não viciado, por exemplo, lançando seguidas vezes até se obter 3 números distintos). Por fim, temos que nossa amostra será composta por

Marilia, Luiz e Camila.

É possível também que o sorteio dos elementos da população seja feito com reposição (de forma que o mesmo indivíduo possa aparecer mais que uma vez na amostra), entretanto, essa abordagem não é comum na prática.

Vale ressaltar ainda que quando o tamanho da população é grande, as duas formas de sorteio (com e sem reposição) são equivalentes, posto que a chance de um elemento da população aparecer mais que uma vez na amostra passar a ser irrisória.

Note que na AAS, a probabilidade de um dos elementos que compõe a população estar presente em uma amostra é dada por n/N. Note ainda que a medida que N se aproxima de infinito, essa probabilidade se aproxima de 0, ou ainda, para N grande essa probabilidade tende a 0.

Podemos citar aqui algumas possíveis formas para realização do sorteio aleatório:

a) Utilização de algoritmos computacionais para geração de números aleatórios;

b) Sorteio de bolinhas enumeradas, como as que são utilizadas em bingos, por exemplo;

c) Utilização de dados;

Uma limitação da ASS que merece destaque é a necessidade de que a população seja homogênea.

1.2.2 Amostragem aleatória estratificada (AAE)

Quando a população tem características observáveis que variam muito de indivíduo para indivíduo, ou seja, a população é muito heterogênea, recomenda-se a divisão da população em subgrupos homogêneos. Esses subgrupos são chamados de estratos. O procedimento consiste basicamente em:

i) Divide-se a população em *k* estratos;

ii) Em cada um dos *k* estratos é realizada uma AAS, de forma a se obter *k* subamostras a partir dos *k* estratos;

iii) Junta-se as *k* subamostras para se obter uma única amostra estratificada.

Vale ressaltar aqui que a AAE é capaz de produzir amostras de melhor qualidade do que a AAS considerando o mesmo tamanho de amostra. Entretanto, é mais cara e demorada, dada a segmentação da população.

Podemos elencar aqui como exemplos algumas das características mais comuns para definição dos estratos:

a) Sexo;

b) Faixas de idade;

c) Classe social;

d) Naturalidade;

e) Profissão.

Suponhamos que para o procedimento de AAE tomamos a decisão de subdividir a população em 2 grupos conforme o sexo do indivíduo, de tal forma que a população terá: 2 estratos, os indivíduos que compõe a população cujo sexo é masculino; e os indivíduos que compõe a população cujo sexo é feminino, de tal forma que k=2. Com isso, tem-se que a população total com N indivíduos será subdividida em dois subconjuntos de tamanhos respectivamente iguais a N_1 e N_2, tal que $N_1 + N_2 = N$.Seguindo o procedimento de AAE, para cada um desses dois subconjuntos teremos uma AAS, de forma a termos 2 subamostras de tamanhos respectivamente iguais a n_1 e n_2, tais que $n_1 + n_2 = n$.

A essa altura surge uma dúvida pertinente, dado que o tamanho desejado para a amostra é n, como decidir os tamanhos de n_1 e n_2? Ou ainda, como decidir qual o tamanho da representatividade de cada estrato na amostra final?

Trataremos aqui de duas formas mais usuais para definir o tamanho de cada estrato, são elas AAE por igual ou uniforme e a AAE proporcional.

1.2.2.1 Amostragem Aleatória Estratificada Uniforme

Na AAE uniforme (AAEu), é atribuído a cada subamostra de cada estrato o mesmo número de indivíduos. Assim, dada uma amostra de tamanho n e um procedimento de AAEu que considere k estratos, tem-se que cada uma das subamostras dos estratos será composta por n/k indivíduos.

Ex. Exemplos

Exemplo1.6: Suponha que se deseja obter uma amostra composta por 1000 indivíduos (assim, n = 1000) a partir de uma determinada população. Suponha ainda que esta população pode ser desmembrada em 4 diferentes estratos (logo, k=4), por exemplo

i) Indivíduos desempregados;

ii) Indivíduos que ganham até 3 salários mínimos;

iii) Indivíduos que ganham mais que 3 e menos que 10 salários mínimos;

iv) Indivíduos que ganham 10 ou mais salários mínimos;

Assim, utilizando a AAEu optaremos por admitir que cada subamostra de cada estrato terá indivíduos,

$$\frac{n}{k} = \frac{1000}{4} = 500$$

de tal forma que

$$n_1 = n_2 = n_3 = n_4 = 500.$$

Note ainda que.

$$n_1 + n_2 + n_3 + n_4 + 1000.$$

1.2.2.2 Amostragem Aleatória Estratificada Proporcional

Na AAE proporcional (AAEp) a representatividade de cada estrato é reproduzida na amostra. Ou seja, a proporção de cada estrato com relação à população é a mesma proporção de cada subamostra com relação a amostra total. Ou ainda

$$\frac{N_1}{N} = \frac{n_1}{n}, \quad \frac{N_2}{N} = \frac{n_2}{n}, \quad \ldots, \quad \frac{N_k}{N} = \frac{n_k}{n}.$$

A partir disto, podemos facilmente deduzir que

$$n_1 = \frac{nN_1}{N}, \quad n_2 = \frac{nN_2}{N}, \quad \ldots, \quad n_k = \frac{nN_k}{N}.$$

Ex. Exemplos

Exemplo1.7: Suponha que se deseja obter uma amostra composta por 200 indivíduos (assim, n =200) a partir de uma determinada população. Suponha ainda que esta população pode ser desagregada em 2 diferentes estratos (logo, k=2), indivíduos do sexo masculino ou feminino, por exemplo. Admita que a população é comporta por 10000 indivíduos (N=10000), em que 6000 são do sexo feminino (N_1=6000) e 4000 são do sexo masculino (N_2=4000).

Assim, utilizando a AAEp optaremos por admitir que a subamostra referente aos indivíduos do sexo feminino será composta por

$$n_1 = \frac{nN_1}{N} = \frac{200 * 6000}{10000} = 120$$

indivíduos. Enquanto que a subamostra referente aos indivíduos do sexo masculino será composta por

$$n_2 = \frac{nN_2}{N} = \frac{200 * 4000}{10000} = 80$$

indivíduos.

Note que

$$\frac{N_1}{N} = \frac{6000}{10000} = 0{,}6 = \frac{120}{200} = \frac{n_1}{n}$$

e que

$$\frac{N_2}{N} = \frac{4000}{10000} = 0{,}4 = \frac{80}{200} = \frac{n_2}{n}.$$

Garantindo assim a proporcionalidade.

1.3 O tamanho da amostra

Visto todo o conteúdo até o momento, um leitor mais atento facilmente percebeu que em nossas abordagens e exemplos sempre partimos do pressuposto do conhecimento do tamanho da amostra. Entretanto, uma dúvida bastante pertinente é, como escolher o tamanho da amostra? Esse é um assunto bastante complexo por envolver diversas metodologias adequadas para as mais diferentes perguntas que se deseja responder a partir de uma amostra. Aqui será apresentado um caso bastante simples, porém usual dadas as limitações de profundidade teórica. Considere que o objeto de desejo é estimar a média populacional a partir da média amostral.

A escolha do tamanho ideal da amostra deve levar em consideração o chamado **erro aceitável** e a probabilidade de ocorrer o erro aceitável (1-α). O erro aceitável é a margem de erro máxima para o erro amostral que o pesquisador está disposto a aceitar.

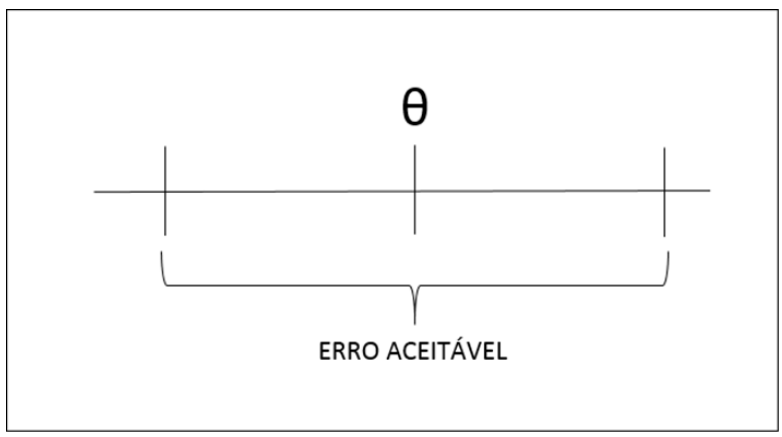


Figura 2: Erro aceitável. / Fonte: Elaboração do autor

Ou seja, a estimativa do parâmetro deve pertencer ao intervalo compreendido pelo erro aceitável. Se, por exemplo θ=2, , e o erro amostral que estou disposto a aceitar é 0,5, então o erro aceitável compreende o intervalo de 1,5 até 2,5. Dessa forma, deseja-se uma amostra que seja capaz de gerar uma precisão tal que minha estimativa de θ esteja entre 1,5 e 2,5.

Como o processo é aleatório, não é possível garantir que para qualquer amostra gerada a partir de uma população se possa de fato obter estimativas dentro do intervalo do erro aceitável, dessa forma, recorre-se a uma probabilidade tolerável de que isso venha a acontecer. Note que como a probabilidade de ocorrer o erro aceitável é 1-α, por complementariedade, a probabilidade de que não ocorra o erro aceitável será igual a α, dessa forma, definimos α como a probabilidade de que a estimativa esteja fora do intervalo do erro aceitável. Por exemplo, podemos estar satisfeitos em conseguir que a estimativa esteja dentro do intervalo do erro aceitável em 95% das amostras obtidas a partir da população. Assim tem-se que 1-α=0.95 .

A probabilidade de ocorrer o erro aceitável por vezes é chamada também de nível de confiança. Podemos descrever isso estatisticamente como

$$P(|\text{ESTIMATIVA}-\theta|<\text{ERRO AMOSTRAL})=1-\alpha$$

Assumindo que os elementos da população seguem distribuição Normal, e considerando um determinado nível de confiança, podemos determinar o tamanho da amostra utilizando a seguinte fórmula

$$n = \frac{\sigma^2 z_{\alpha}^2}{\varepsilon^2}$$

em que é a variância populacional da variável a ser estudada, e é o valor crítico correspondente ao nível de confiança desejado. Os níveis de confiança mais usuais, bem como seus respectivos valores críticos podem ser observados na tabela a seguir.

Tabela 1: Valores críticos da distribuição Normal padrão.

Nível de confiança	α	z
90%	0,10	1,645
95%	0,05	1,960
99%	0,01	2,575

Ilustração: Ariana Santana

Ex. Exemplos

Exemplo1.8: Suponha que nosso interesse reside em estimar a renda média dos contadores em 2016. Admita ainda que desejamos um nível de confiança de 90% de que a média amostral esteja a menos de R$ 200,00 da média real (populacional). Admita ainda que conhecemos a informação sobre o desvio padrão dessas rendas, tal que . Qual o tamanho necessário para amostra?

$$n = \frac{\sigma^2 z_\alpha^2}{\varepsilon^2} = \frac{(7000)^2 * (1{,}645)^2}{200^2} \approx 3314{,}88 \approx 3315,$$

de forma que necessitamos de uma amostra composta por 3315 observações.

Na prática, dificilmente conhecemos de fato a variância populacional da variável de interesse, dessa forma, uma possibilidade mais realista para a escolha do tamanho da amostra é considerar primeiramente uma pesquisa preliminar, a partir da qual se estima a variância amostral e a utiliza como aproximação para σ^2 . Ou seja:

i) Coleta-se uma pequena amostra preliminar de tamanho n=20 , por exemplo;

ii) Calcula-se a variância dessa amostra (S^2);

iii)Aplica a fórmula anterior substituindo σ^2 por S^2 .

UNIDADE 2 - ESTIMAÇÃO

Vimos até aqui formas adequadas para coleta de informações que sejam capazes de carregar consigo característica intrínsecas de determinada população. Coletada a amostra de forma pertinente surge a necessidade de estimar a partir da amostra determinadas características (ou parâmetros) populacionais, tais como média e variância, por exemplo. Essa parte do processo é conhecida como estimação pontual, posto que a partir dela se obtém um único valor (ou ponto). Por exemplo, a estimação da média de uma amostra é dada por um único valor (o mesmo vale para a variância).

Vimos ainda no capítulo anterior que dependendo da forma como a amostra é coletada, o valor da estimativa pode variar (ver **Exemplo 1.4**), mesmo que os procedimentos de coleta da amostra estejam corretos. De forma que a estimação pontual não possibilita ter uma ideia do erro cometido ao se proceder a estimação. Assim, é interessante se construir um intervalo em torno da estimação pontual para que se tenha mais confiança e se tenho uma ideia do tamanho do erro que o estimador utilizado pode causar. Dessa forma, seria interessante que ao invés de termos apenas um valor para estimativa de determinado parâmetro (como ocorre na estimação pontual), nós tivéssemos um intervalo de valores que melhor reflita o parâmetro populacional.

Esse tipo de estimação denominamos **estimação por intervalo** ou **estimação intervalar**. Por exemplo: Suponha que a média de determinada população é igual a 100. São coletadas 3 amostras, cujas médias são respectivamente iguais a 99, 101 e 102. Se as 3 amostras foram coletadas corretamente, temos que os três valores são estimativas válidas para média populacional, logo, estas têm equivalência estatística. Assim, o objetivo aqui é obter um intervalo de valores ao qual o verdadeiro valor do parâmetro populacional deve ter grandes chances de pertencer. Esse intervalo é conhecido como **Intervalo de Confiança** (IC).

2.1 Estimação intervalar

Trataremos aqui de algumas técnicas para estimação intervalar considerando seis casos específicos de interesse, a saber, estimativa intervalar para média populacional quando a variância populacional é conhecida e quando é desconhecida; estimativa intervalar para proporção populacional; estimativa intervalar para diferença entre duas médias populacionais considerando tanto variância populacional conhecida quanto desconhecida; e diferença entre duas proporções populacionais.

Assumiremos aqui que *n* é grande. Essa hipótese é importante porque a partir dela podemos presumir (dado o Teorema Central do Limite) que a média amostral segue distribuição Normal. Isso facilita bastante a estimação, posto que já existe todo um arcabouço teórico construído, de forma que focaremos aqui em como utilizar essas técnicas.

2.1.1 Estimação intervalar para média populacional quando a variância é conhecida

Temos interesse aqui em estimar os limites inferiores e superiores do IC, tal que exista uma probabilidade pré-definida (*1-α*) de que a média populacional esteja contida nesse IC. Assim determinaremos duas estatísticas, a saber:

Limite inferior - L_{inf}; e

Limite superior -L_{sup} ;

denotando nossa pretensão matematicamente, desejamos obter a

$$P\,(L_{inf} < \mu < L_{sup}\,) = 1\text{-}\alpha,$$

lembrando que (1-α) é o nível de confiança.

Para isso será necessário recorrer a alguns fundamentos teóricos. Do Teorema Central do Limite (TCL) temos que $\bar{X} - \mu \sim N(0, \sigma_{\bar{X}}^2)$

em que μ é a média populacional e $\sim N(0, \sigma_{\bar{X}}^2)$ denota "segue distribuição Normal com média 0 e variância $\sigma_{\bar{X}}^2$". Tem-se ainda que que

$$Var(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n},$$

partindo disso, podemos estabelecer que

$$P(|\bar{X} - \mu| < Z_\alpha \sigma_{\bar{X}}) = 1 - \alpha\,,$$

em que Z_α é determinado com base em 1-α como descrito na **Tabela 1**. Após algumas manipulações podemos reescrever a fórmula anterior como

$$P(\bar{X} - Z_\alpha \sigma_{\bar{X}} < \mu < \bar{X} + Z_\alpha \sigma_{\bar{X}}) = 1 - \alpha,$$

ou seja, estabelecemos que o verdadeiro valor da média populacional (μ) está contido em um intervalo com

$$L_{inf} = \bar{X} - Z_\alpha \sigma_{\bar{X}} \quad \text{e} \quad L_{sup} = \bar{X} + Z_\alpha \sigma_{\bar{X}}$$

com probabilidade de [(1-α)*100] %, como desejávamos.

Em termos práticos, e assumindo que $\sigma^2_{\bar{X}}$ é conhecido, podemos obter a intervalo de confiança para média populacional utilizando a seguinte fórmula:

$$[\bar{X} - Z_\alpha \sigma_{\bar{X}} \; ; \; \bar{X} + Z_\alpha \sigma_{\bar{X}}].$$

Ex. Exemplos

Exemplo 2.1: Assuma uma população que possui desvio padrão igual 5 e média desconhecida. Um estudo consegue a informação de que , a partir de uma amostra de tamanho 100. Deseja-se determinar o intervalo de 95% de confiança para a média populacional.

Note que

- n = 100;
- $1 - \alpha = 95\%$;
- $\bar{X} = 25$;
- $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{5^2}{100} = \frac{25}{100} = 0{,}25$;
- $\sigma_{\bar{X}} = \sqrt{\sigma^2_{\bar{X}}} = \sqrt{0{,}25} = 0{,}5$.

Utilizando essas informações podemos escrever o intervalo de confiança como

$$[\bar{X} - Z_\alpha \sigma_{\bar{X}} \; ; \; \bar{X} + Z_\alpha \sigma_{\bar{X}}]$$
$$[25 - Z_\alpha * 0{,}5 \; ; \; 25 + Z_\alpha * 0{,}5],$$

como 1-α=95%, da **Tabela 1** temos que $Z_a=1,96$, logo

$$[25 - 1,96 * 0,5 \; ; \; 25 + 1,96 * 0,5]$$
$$[24,02 \; ; \; 25,98] \, .$$

Ou seja, com 95% de confiança, acredita-se que μ está contida no intervalo[24,02 ; 25,98].

2.1.2 Estimação intervalar para média populacional quando a variância é desconhecida

Aqui será tratado um caso mais realista, no qual admitimos que desconhecemos o valor da variância populacional (σ^2). Nesse caso, devemos nos contentar em utilizar uma aproximação desse valor, que pode ser obtido pela variância da amostra (S^2). Essa não é a única inconveniência resultante de não se ter essa informação, agora não podemos mais utilizar os valores de referência disponíveis na **Tabela 1**. Caso a amostra seja grande (algo em torno de 100), podemos ainda assim utilizar a fórmula apresentada para estimação de IC com variância conhecida, para tamanhos amostrais menores, a distribuição não é mais adequada e é necessário utilizar os quantis da distribuição *t* de Student.

Diferentemente da distribuição Normal padrão, em que observando a tabela, é necessário apenas o conhecimento do nível de confiança de interesse para encontrar o valor crítico correspondente, na distribuição *t* de Student deve-se levar em consideração o nível de confiança de interesse é o **grau de liberdade** (ν) correspondente. O grau de liberdade é dado simplesmente por $n-1$, ou seja, tem-se uma amostra de 25 observações, $\nu=n-1=24$. Denotaremos aqui o valor crítico correspondente ao nível de confiança$(1-\alpha)$ com ν graus de liberdade, da distribuição *t* de Student por $T_{\nu,a}$.

Ex. Exemplos

Exemplo 2.2: Considere que se deseja descobrir o valor crítico da distribuição *t* de Student referente a um nível de confiança de 95% e com 10 graus de liberdade.

No **Anexo A,** está disponível uma tabela com os principais níveis de confiança e os respectivos valores críticos dados diferentes graus de liberdade. Para construção de IC deve-se observar os níveis de confiança bilaterais.

Grau de Liberdade	α unilateral / bilateral	0,005 / 0,010	0,025 / 0,050	0,05 / 0,100
1		63,657	12,706	6,314
2		9,925	4,303	2,920
3		5,841	3,182	2,353
4		4,604	2,776	2,132
5		4,032	2,571	2,015
6		3,707	2,447	1,943
7		3,499	2,365	1,895
8		3,355	2,306	1,860
9		3,250	2,262	1,833
10		3,169	2,228	1,812
11		3,106	2,201	1,796
12		3,055	2,179	1,782

Figura 3: Obtenção de valores críticos a partir da tabela *t* de Student. / Ilustração: Ariana Santana

Assim, como , então , e se tem interesse em , como indicado na figura acima, tem-se que .

Nesse caso a fórmula para obtenção do intervalo de confiança não tem mudanças drásticas, e pode ser expressa por

em que é obtido da tabela de valores críticos da distribuição *t* de Student e a variância do estimador é dada por

Ex. Exemplos

Exemplo 2.3: Assuma uma população que possui desvio padrão e média desconhecidos. Um estudo consegue as informações de que $\bar{X}$=25, S=10, e que a partir de uma amostra de tamanho 21. Deseja-se determinar o intervalo de 95% de confiança para a média populacional.

Note que

- $n = 21$;
- $v = 21 - 1 = 20$;
- $1 - \alpha = 95\%$;
- $\bar{X} = 25$;
- $\widehat{\sigma_{\bar{X}}^2} = \frac{S^2}{n} = \frac{10^2}{21} = \frac{100}{21} = 4{,}762$;
- $\sigma_{\bar{X}} = \sqrt{\widehat{\sigma_{\bar{X}}^2}} = \sqrt{4{,}762} = 2{,}182$.

Utilizando essas informações podemos escrever o intervalo de confiança como

$$\left[\bar{X} - T_{v,\alpha}\widehat{\sigma_{\bar{X}}}\,;\ \bar{X} + T_{v,\alpha}\widehat{\sigma_{\bar{X}}}\right],$$
$$\left[25 - T_{v,\alpha} * 2{,}182\,;\ 25 + T_{v,\alpha} * 2{,}182\right],$$

como 1-α=95% , da de valores críticos da distribuição *t* de Student temos que $T_{v,a}$=2,086, logo

$$[25 - 2{,}086 * 2{,}182\,;\ 25 + 2{,}086 * 2{,}182]$$
$$[20{,}448\,;\ 29{,}552]\,.$$

Ou seja, com 95% de confiança, acredita-se que está μ contida no intervalo [20,448 ; 29,552].

2.1.3 Estimação intervalar para proporção populacional

O terceiro caso a ser abordado aqui será o da estimação intervalar para proporção populacional (p). Como se sabe, a proporção populacional é representa pela percentagem de indivíduos de determinada população que têm determinada característica. Para a estimativa da proporção amostral podemos utilizar uma representação binária para indicar a existência ou não de determinada característica, atribuindo 1 aos indivíduos que possuem a característica de interesse, e 0 aos que não possuem. Posteriormente, calcula-se a média simples dessa amostra, assim, podemos expressar matematicamente a proporção amostral como

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n},$$

onde é o valor correspondente a cada elemento da população.

Ex. Exemplos

Exemplo 2.4: Considere uma população composta por 60% de mulheres e 40% de homens, dessa população é obtida uma amostra aleatória composta por 5 indivíduos, a saber,

João, Marilia, Luiz, Fernanda, Maria e Camila.

Deseja-se saber a proporção amostral de mulheres.

Como a característica de interesse é "ser mulher", podemos representar a amostra como

0, 1, 0, 1, 1 e 1.

Aplicando a fórmula para proporção temos

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{0+1+0+1+1+1}{6} \approx 0{,}67\,.$$

Partindo mais uma vez do TCL, podemos supor que $\hat{p}$ segue distribuição Normal. Assim, a construção dos intervalos de confiança é similar às formas apresentadas até o momento. Podemos representar o intervalo de confiança para proporção populacional como

$$\left[\hat{p} - Z_\alpha \sigma_p \,;\, \hat{p} + Z_\alpha \sigma_p\right],$$

em que

$$Var(p) = \sigma_p^2 = \frac{p(1-p)}{n},$$

surge aqui um problema, a estimação do IC depende de algo que pode não ser conhecido, p . Existem duas abordagens para lidar com esse problema, a primeira é a chamada abordagem **conservadora**, a qual parte do fato de que

$$p(1-p) \leq \frac{1}{4},$$

posto que o valor máximo de *p(1-p)* se dá quando *p=1/2*. Nesse caso, assumimos que

$$Var(p) = \sigma_p^2 = \frac{p(1-p)}{n} = \frac{1}{4n},$$

e o IC é dado por

$$\left[\hat{p} - Z_\alpha \sigma_p \,;\, \hat{p} + Z_\alpha \sigma_p\right]$$
$$\left[\hat{p} - Z_\alpha \sqrt{\frac{1}{4n}} \,;\, \hat{p} + Z_\alpha \sqrt{\frac{1}{4n}}\right].$$

A segunda abordagem consiste em simplesmente substituir p por $\hat{p}$, de forma que

$$Var(\hat{p}) = \widehat{\sigma^2_{\hat{p}}} = \frac{\hat{p}(1-\hat{p})}{n},$$

e consequentemente o IC é dado por

$$\left[\hat{p} - Z_\alpha \widehat{\sigma_{\hat{p}}} \; ; \; \hat{p} + Z_\alpha \widehat{\sigma_{\hat{p}}}\right]$$

$$\left[\hat{p} - Z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \; ; \; \hat{p} + Z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Vale ressaltar aqui que o IC não conservador tende a fornecer IC com mais precisão do que o IC conservador. Caso a proporção populacional esteja muito próxima de 0 ou de 1, o IC conservador tende a não funcionar de forma satisfatória.

Ex. Exemplos

Exemplo 2.5: Um fabricante de celulares deseja estimar a fração de p aparelhos defeituosos produzidos em sua linha de montagem. Para isso colheu-se uma amostra aleatória composta por 200 unidades, na qual verificou-se que 20 destas eram defeituosas. Com base nessas informações, vamos construir o intervalo de 95% de confiança para p, considerando as duas técnicas vistas até aqui. Começando pelo IC conservador.

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{20}{200} = 0{,}1.$$

Da Tabela 1, considerando o nível de confiança de 95%, temos que Z_a=1,96. O IC conservador é dado por

$$\left[\hat{p} - Z_\alpha \sqrt{\frac{1}{4n}} \; ; \; \hat{p} + Z_\alpha \sqrt{\frac{1}{4n}}\right],$$

logo

$$\left[0{,}1 - 1{,}96\sqrt{\frac{1}{4*200}}\ ;\ 0{,}1 + 1{,}96\sqrt{\frac{1}{4*200}}\right]$$
$$[0{,}1 - 0{,}069\ ;\ 0{,}1 + 0{,}069]$$
$$[0{,}031\ ;\ 0{,}169]\,.$$

Ou seja, temos como resultado que a proporção populacional é algo entre 3,1% e 16,9%.

Utilizando agora a fórmula do IC não conservador, temos que

$$\left[\hat{p} - Z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ;\ \hat{p} + Z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$\left[0{,}1 - 1{,}96\sqrt{\frac{0{,}1(1-0{,}1)}{200}}\ ;\ 0{,}1 + 1{,}96\sqrt{\frac{0{,}1(1-0{,}1)}{200}}\right]$$
$$[0{,}1 - 0{,}041\ ;\ 0{,}1 + 0{,}041]$$
$$[0{,}058\ ;\ 0{,}141]\,.$$

Ou seja, aqui temos que a proporção populacional está entre 5,8% e 14,1%.

Note que o IC não conservador gerou um IC de amplitude menor que o IC conservador.

Daqui em diante consideraremos o caso em que temos interesse em extrair informações de mais de uma população, utilizando mais de uma amostra. Basicamente, vamos analisar a diferença entre médias e proporções de diferentes populações considerando mais de uma amostra aleatória. Os casos a seguir são especialmente interessantes para testar se populações diferentes têm ou não característica similares.

2.1.4 Estimação intervalar para diferença entre duas médias com variância conhecida

Nesse tópico vamos tratar de um caso particularmente diferente dos anteriores, principalmente porque agora serão utilizadas duas amostras ao invés de apenas uma. Suponha agora que consideramos duas populações, e que temos duas amostras correspondentes, denotadas pelo conjunto X composto por n_x indivíduos, e pelo conjunto Y composto por

n_y indivíduos. Chamemos as médias amostrais de X e Y de $\bar{X}$ e $\bar{Y}$, respectivamente. A ideia aqui reside em estimar IC para diferença entre as médias populacionais dessas duas amostras, em que podemos representar matematicamente essa diferença por $\mu_x - \mu_y$.

Para construção desse IC, vamos admitir que as variâncias populacionais correspondentes a cada uma das amostras são conhecidas e são denotadas por σ_X^2 e σ_Y^2;

A derivação do estimador do IC para diferença de médias é similar a apresentada para o caso de uma só amostra apresentada no início desse capítulo. Entretanto, ao invés de trabalhar com o estimador $\bar{X}$, vamos trabalhar com o estimador $\bar{X} - \bar{Y}$, cuja variância é dada por

$$Var(\bar{X} - \bar{Y}) = \sigma^2_{\bar{X}-\bar{Y}} = \frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_Y}.$$

Assim, temos que o IC para diferença de médias é dado por

$$\left[(\bar{X} - \bar{Y}) - Z_\alpha \sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_Y}} ; (\bar{X} - \bar{Y}) + Z_\alpha \sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_Y}}\right].$$

Ex. Exemplos

Exemplo 2.6: Assuma que existem duas empresas concorrentes no mercado de notebooks, a empresa X e a empresa Y . Suponha que se deseja saber a diferença entre a média de idade dos consumidores da marca X($\bar{X}$) e a média de idade dos consumidores da marca Y($\bar{Y}$). Sabe-se, por meio de um estudo preliminar que $\sigma_x=9$ e $\sigma_y=10$. São coletadas duas amostras, uma para cada consumidor dos produtos de cada empresa, de forma que se obteve a partir de uma amostra de tamanho $n_x=36$, enquanto que foi obtido $\bar{Y}=35$ de uma amostra de tamanho $n_y=49$. Vamos então calcular o IC para $\mu_x - \mu_y$ considerando um nível de confiança de 95%.

$$\left[(\bar{X} - \bar{Y}) - Z_\alpha \sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_Y}} ; (\bar{X} - \bar{Y}) + Z_\alpha \sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_Y}}\right]$$

$$\left[(40 - 35) - 1{,}96 * \sqrt{\frac{81}{36} + \frac{100}{49}} ; (40 - 35) + 1{,}96 * \sqrt{\frac{81}{36} + \frac{100}{49}}\right]$$

$$[5 - 4{,}06 ; 5 + 4{,}06]$$

$$[0{,}94 ; 9{,}06]$$

De forma que a diferença entre as idades médias dos consumidores das duas marcas está entre 0,94 anos e 9,06 anos. Ou seja, podemos afirmar com 95% de confiança que existe de fato diferença de idade média entre os consumidores de cada marca.

2.1.5 Estimação intervalar para diferença entre duas médias com variância desconhecida

Aqui admitimos que as variâncias populacionais são desconhecidas, entretanto, elas podem se apresentar de duas formas, onde cada uma tem seu IC correspondente. Por isso, nesse tópico serão tratados dois casos, a saber

I. A variância populacional correspondente a cada uma das amostras é desconhecida, entretanto, sabe-se que são iguais. Nesse caso teremos que trabalhar com suas estimativas, denotadas por S_X^2 e S_Y^2, e , em que $S_X^2 = S_Y^2 = S^2$;

II. A variância populacional correspondente a cada uma das amostras é desconhecida e são diferentes. Nesse caso, teremos que trabalhar com suas estimativas, denotadas por S_X^2 e S_Y^2.

Caso I – variâncias desconhecidas, porém iguais

O formato do IC para este caso é bastante similar ao caso anterior, cujas variâncias populacionais são conhecidas. As únicas diferenças aqui dizem respeito à variância do estimador para diferença de médias, como as variâncias populacionais são agora desconhecidas temos que utilizar a variância amostral como aproximação, de forma que, agora, a variância do estimador é dada por

$$Var(\bar{X} - \bar{Y}) = S^2_{\bar{X}-\bar{Y}} = S^2\left(\frac{1}{n_x} + \frac{1}{n_Y}\right),$$

em que

$$S^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2},$$

onde S_X^2 é a variância amostral da amostra X e S_Y^2 é a variância amostral da amostra Y.

Além disso, não podemos mais utilizar valores críticos da distribuição Normal padrão, agora utilizaremos os valores críticos da distribuição t de Student com $n_x = n_y - 2$ graus de liberdade.

Assim, nesse caso, o IC é dado por

$$\left[(\bar{X}-\bar{Y}) - T_{v,\alpha}S_{\bar{X}+\bar{Y}}\,;(\bar{X}-\bar{Y}) + T_{v,\alpha}S_{\bar{X}+\bar{Y}}\right].$$

Ex. Exemplos

Exemplo 2.7: Suponha que nosso interesse é investigar a influência do uso de tabaco por mulheres gravidas no peso das crianças ao nascer. Assim, temos duas populações, as mulheres grávidas fumantes e as mulheres grávidas não-fumantes. Com base em uma pesquisa, são disponibilizadas as seguintes informações:

Mulheres não-fumantes: $n_x = 25;\ \bar{X} = 3{,}6\ Kg;\ S_x = 0{,}7\ Kg$

Mulheres fumantes: $n_y = 15;\ \bar{Y} = 3{,}2\ Kg;\ S_y = 1\ Kg$

Vamos agora calcular o IC para diferença de médias considerando um nível de confiança de 95%.

Antes de mais nada, precisamos calcular o valor de $S_{\bar{X}+\bar{Y}}$. Sabe-se que

$$S^2_{\bar{X}-\bar{Y}} = \left(\frac{(n_X-1)S^2_X + (n_Y-1)S^2_Y}{n_X+n_Y-2}\right) * \left(\frac{1}{n_x}+\frac{1}{n_Y}\right),$$

logo, substituindo as informações disponíveis, temos que

$$S^2_{\bar{X}-\bar{Y}} = \left(\frac{(25-1)*0{,}7^2 + (15-1)*1^2}{25+15-2}\right) * \left(\frac{1}{25}+\frac{1}{15}\right)$$

$$S^2_{\bar{X}-\bar{Y}} = \left(\frac{24*0{,}49+14*1}{38}\right) * (0{,}107)$$

$$S^2_{\bar{X}-\bar{Y}} = (0{,}678) * (0{,}107) \approx 0{,}072\,.$$

Dado que conhecemos $S^2_{\bar{X}-\bar{Y}}$ conhecemos também $S_{\bar{X}+\bar{Y}}$, e da tabela dos valores críticos da distribuição *t* de Student tem-se que $T_{38,0,05}$ é igual a 2,024. Finalmente podemos montar o IC, o qual é dado por

$$\left[(\bar{X}-\bar{Y}) - T_{v,\alpha}S_{\bar{X}+\bar{Y}}\,;(\bar{X}-\bar{Y}) + T_{v,\alpha}S_{\bar{X}+\bar{Y}}\right]$$

$$[(3{,}6-3{,}2) - 2{,}024*0{,}268\ ;(3{,}6-3{,}2) + 2{,}024*0{,}268\ ]$$

$$[0{,}4-0{,}542\ ;0{,}4+0{,}542\ ]$$

$$[-0{,}142\ ;0{,}942\ ].$$

Ou seja, estamos 95% confiantes de que a verdadeira diferença entre as médias de pesos entre as crianças nascidas de mães fumantes e não fumantes está contida no intervalo [-0,142 ; 0,942].

Note que o intervalo contém o valor zero, o que indica que não existem evidências estatísticas de que exista diferença entre as médias.

Caso II – variâncias desconhecidas e diferentes

Consideraremos aqui que além das variâncias populacionais serem desconhecidas, elas também são diferentes (diferentemente do Caso I). Nesse caso, nós desconhecemos os valores de σ_X^2 e σ_Y^2;mas sabemos que $\sigma_X^2 \neq \sigma_Y^2$. Este é um cenário uma pouco mais realista. Normalmente como a variância populacional é desconhecida, utilizaremos as variâncias amostrais S_X^2 e S_Y^2,. Aqui a variância do estimador será dada por

$$Var(\bar{X} - \bar{Y}) = S^2_{\bar{X}-\bar{Y}} = \frac{S_X^2}{n_x} + \frac{S_Y^2}{n_Y},$$

e assim como no Caso I, os valores críticos são obtidos da distribuição *t* de Student, entretanto, agora os graus de liberdade são dados por

$$v = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}.$$

Assim o IC nesse caso pode ser obtido por

$$\left[(\bar{X} - \bar{Y}) - T_{v,\alpha}\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\,; (\bar{X} - \bar{Y}) + T_{v,\alpha}\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\right].$$

Ex. Exemplos

Exemplo 2.8: Considere duas amostras retiradas de duas populações como variâncias diferentes e desconhecidas.

Amostra 1: $n_X = 28; \bar{X} = 102{,}5; S_X = 15;$
Amostra 2: $n_Y = 22; \bar{Y} = 91; S_Y = 12{,}5.$

Com base nessas informações, vamos construir o IC com nível de confiança de 95%. Primeiramente precisamos definir os graus de liberdade de $T_{v,a}$, a fim de que se possa obter o valor crítico tabelado.

$$v = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}$$

$$v = \frac{\left(\frac{15^2}{28} + \frac{12{,}5^2}{22}\right)^2}{\frac{\left(\frac{15^2}{28}\right)^2}{28 - 1} + \frac{\left(\frac{12{,}5^2}{22}\right)^2}{22 - 1}}$$

$$v = \frac{(8{,}036 + 7{,}102)^2}{\frac{(8{,}036)^2}{27} + \frac{(7{,}102)^2}{21}}$$

$$v = \frac{229{,}159}{4{,}793} \approx 47{,}811 \approx 48$$

Da tabela da distribuição *t* de Student temos que $T_{48,005} = 2{,}011$. Agora aplicando a fórmula do IC, temos

$$\left[(\bar{X} - \bar{Y}) - T_{v,\alpha}\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\,;(\bar{X} - \bar{Y}) + T_{v,\alpha}\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\right]$$

$$\left[(102{,}5 - 91) - 2{,}011 * \sqrt{\frac{15^2}{28} + \frac{12{,}5^2}{22}}\,;(102{,}5 - 91) + 2{,}011 * \sqrt{\frac{15^2}{28} + \frac{12{,}5^2}{22}}\right]$$

$$\left[11{,}5 - 2{,}011 * \sqrt{8{,}036 + 7{,}102}\,;11{,}5 + 2{,}011 * \sqrt{8{,}036 + 7{,}102}\right]$$

$$[11{,}5 - 7{,}824\,;11{,}5 + 7{,}824\,]$$

$$[3{,}676\,;19{,}324\,].$$

Assim, podemos afirmar com 95% de confiança que o real valor da diferença entre as médias populacionais está entre 3,676 e 19,324 .

2.1.6 Estimação intervalar para diferença entre duas proporções

Vimos anteriormente sobre estimação intervalar para proporção quando se trabalha com uma única amostra, agora vamos considerar o caso em que temos duas populações independentes, e deseja-se estimar o IC para diferença de suas proporções populacionais. Denotaremos a proporção da população X por p_x e a proporção da população Y por p_y, de forma que temos interesse em $(p_x - p_y)$.

A variância do estimador para diferença da proporção populacional $(\widehat{p_X} - \widehat{p_Y})$ é dada por

$$Var(\widehat{p_X} - \widehat{p_Y}) = \widehat{\sigma^2_{\widehat{p_X}-\widehat{p_Y}}} = \frac{\widehat{p_X}(1-\widehat{p_X})}{n_X} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{n_Y},$$

de forma que o IC pode ser obtido por meio da seguinte fórmula

$$\left[(\widehat{p_X} - \widehat{p_Y}) - Z_\alpha \sqrt{\frac{\widehat{p_X}(1-\widehat{p_X})}{n_X} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{n_Y}} \; ; \; (\widehat{p_X} - \widehat{p_Y}) + Z_\alpha \sqrt{\frac{\widehat{p_X}(1-\widehat{p_X})}{n_X} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{n_Y}}\right].$$

Ex. Exemplos

Exemplo 2.9: Assuma que temos interesse em estudar a população de determinada espécie de peixe, considerando duas lagoas. Da lagoa X, é retirada uma amostra de 116 peixes, e observa-se que apenas 84 são da espécie que temos interesse. Da lagoa Y é retirada uma amostra de 80 peixes, dos quais 45 são da espécie que temos interesse. Deseja-se estimar o IC de 90% de confiança para diferença entre as proporções.

$$\widehat{p_X} = \frac{84}{116} = 0{,}724\,,$$

$$\widehat{p_Y} = \frac{45}{80} = 0{,}562\,.$$

Aplicando a fórmula do IC, temos

$$\left[(\widehat{p_X} - \widehat{p_Y}) - Z_\alpha \sqrt{\frac{\widehat{p_X}(1-\widehat{p_X})}{n_X} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{n_Y}}\ ;\ (\widehat{p_X} - \widehat{p_Y}) + Z_\alpha \sqrt{\frac{\widehat{p_X}(1-\widehat{p_X})}{n_X} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{n_Y}}\right]$$

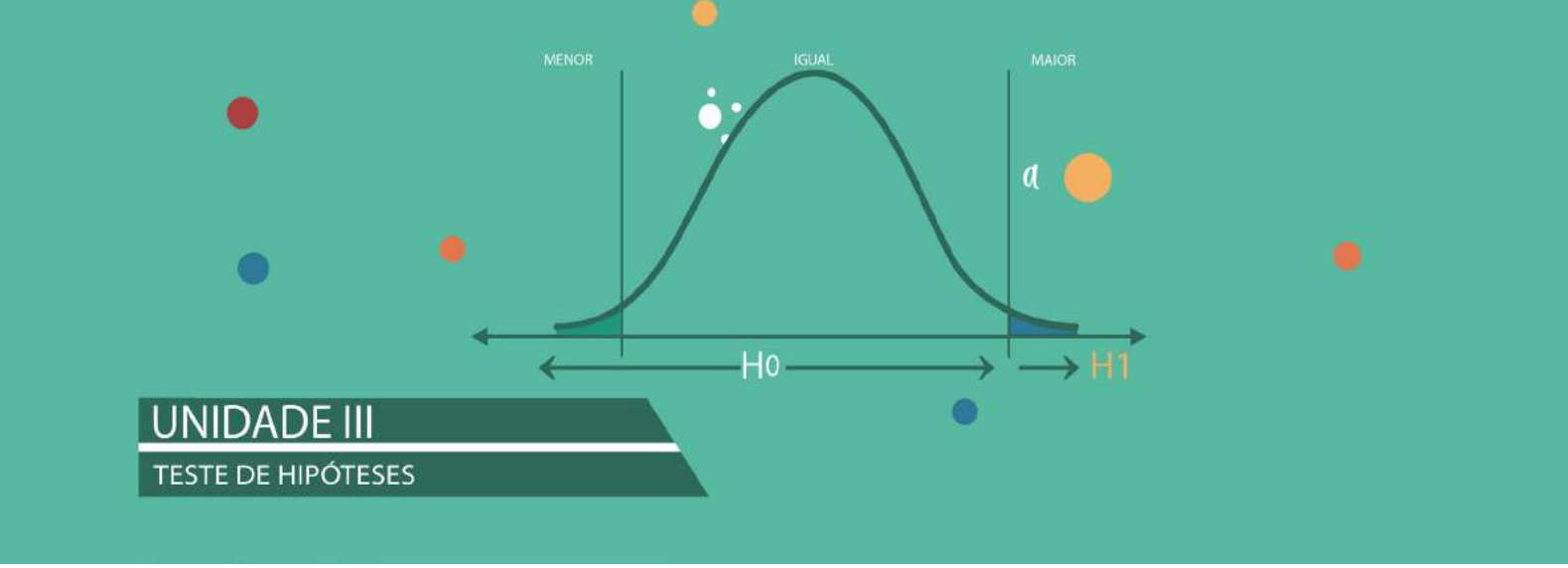
$$\left[(0{,}724 - 0{,}562) - 1{,}645 * \sqrt{\frac{0{,}724(1-0{,}724)}{116} + \frac{0{,}562(1-0{,}562)}{80}}\ ;\ (0{,}724 - 0{,}562) + 1{,}645 * \sqrt{\frac{0{,}724(1-0{,}724)}{116} + \frac{0{,}562(1-0{,}562)}{80}}\right]$$

$$\left[0{,}162 - 1{,}645 * \sqrt{0{,}002 + 0{,}003}\ ;\ 0{,}162 + 1{,}645 * \sqrt{0{,}002 + 0{,}003}\right]$$

$$[0{,}162 - 1{,}645 * 0{,}071\ ;\ 0{,}162 + 1{,}645 * 0{,}071\ ]$$

$$[0{,}045\ ;\ 0{,}279\ ].$$

Ou seja, a verdadeira diferença entre as proporções populacionais está entre 4,5% e 27,9%, com 95% de confiança.



UNIDADE 3 - TESTE DE HIPÓTESES

Nesse capítulo vamos estudar uma poderosa ferramenta estatística com a qual é possível testar suposições inicialmente feitas sobre os parâmetros de determinada população de interesse, levando em consideração apenas as informações disponíveis em uma amostra aleatória.

Na prática, é bastante comum que o pesquisador esteja mais interessado nas afirmações que podem ser feitas a respeito dos parâmetros populacionais do que simplesmente sua estimação em si. Por exemplo, um pesquisador deseja investigar a temperatura corporal média de ursos polares. Ela acredita que esta temperatura é igual a 98ºC. Dessa forma, para validar essa hipótese será necessário obter uma amostra de temperaturas corporais de ursos polares, para a partir das informações contidas na amostra, testar sua hipótese inicial.

3.1 Conceitos básicos

O chamado **teste de hipóteses** é uma ferramenta para **rejeitar ou não** determinada hipótese estatística levando em consideração as informações contidas na amostra.

Vamos definir aqui os conceitos principais para se entender o teste de hipóteses.

I. Hipóteses estatísticas

É aquilo que se tem interesse em validar utilizando testes estatísticos. As hipóteses são basicamente afirmações a respeito do parâmetro de interesse. Normalmente o teste de hipóteses leva em consideração duas hipóteses, são elas,

a) Hipótese Nula

É a hipótese principal do teste de hipóteses, nela é feita a afirmação de maior interesse ao pesquisador. A hipótese nula é comumente denotada por H_0. Exemplos de hipóteses nulas considerando afirmações sobre determinada média populacional

$$H_0: \mu=\mu_0; \qquad H_0: \mu\geq\mu_0; \qquad H_0: \mu\leq\mu_0$$

b) Hipótese Alternativa

É a afirmação que deve ser verdadeira caso a hipótese nula seja rejeitada, ou seja, caso H_0 seja rejeitada, o parâmetro deve ter características não condizentes com H_0. A hipótese alternativa representa a afirmação sobre o parâmetro de interesse que se acredita ser verdadeira quando a hipótese nula não o for. A hipótese alternativa é denotada por H_1. Exemplos de hipóteses alternativas considerando afirmações sobre determinada média populacional

$$H_1: \mu\neq\mu_0; \qquad H_1: \mu<\mu_0; \qquad H_1: \mu>\mu_1$$

A escolha das hipóteses nula e alternativa pelo pesquisador determina o procedimento de teste de hipóteses a ser utilizado. A depender dessa escolha podemos ter que utilizar testes **unilaterais** ou **bilaterais**. Por exemplo, suponha que as hipóteses de interesse são

$$H_0: \mu=\mu_0 \quad vs \quad H_1: \mu<\mu_0 \ ,$$

Nesse caso, é necessário utilizar um teste unilateral. Para melhor entender a ideia de teste unilateral e bilateral, considere a reta dos reais na figura abaixo. Caso a hipótese alternativa seja $H_1: \mu<\mu_0$ ou $H_1: \mu>\mu_0$, então o teste será unilateral por que caso se rejeite H_0, a hipótese alternativa indica que o valor de μ estará à direita **ou** à esquerda de μ_0, na reta dos reais.

Já no caso em que a hipótese alternativa é algo como $H_1: \mu\neq\mu_0$, tem-se que essa supõe que μ pode assumir qualquer valor a direita e a esquerda de μ_0 , desde que não seja exatamente igual a μ_0. Nesse caso, utilizamos testes bilaterais.

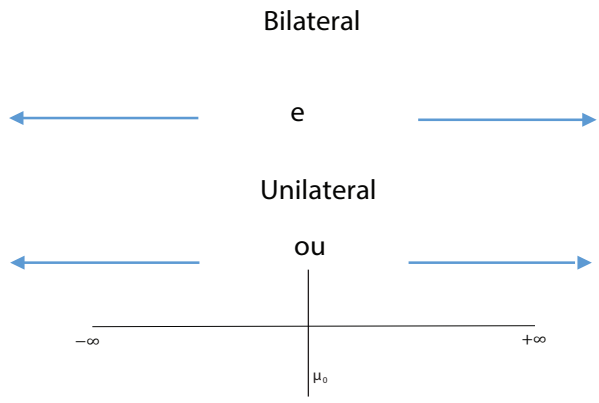


Figura 4: Testes unilaterais e bilaterais. / Ilustração: Ariana Santana

A escolha sobre as hipóteses nula e alternativa é bastante importante, de forma que é indispensável que esta seja feita de forma correta. Em teoria a hipótese nula deve refletir o cenário mais importante para pesquisa a ser desenvolvida. Por exemplo, se já existe

um conhecimento preestabelecido ou uma afirmação alheia sobre determinada variável, pode-se utilizar esse conhecimento para formular a hipótese nula.

A hipótese alternativa deve refletir um cenário que dê suporte a uma argumentação alternativa ou que uma afirmação que confronte o que foi estabelecido na hipótese nula.

II. Erros

A ideia central do teste de hipóteses é chegar a uma decisão sobre as hipóteses nula e alternativa, a rigor, decidimos se rejeitamos ou não a hipótese nula. Entretanto, ao tomar essa decisão (rejeitar ou não H_0) podemos incorrer em erro. Mais precisamente, podemos cometer dois tipos de erros, podemos rejeitar H_0 quando H_0 é verdadeira e podemos não rejeitar H_0 quando H_0 é falsa, esses dois tipos de erro são denominados **Erro tipo I** e **Erro tipo II**, respectivamente.

		Realidade	
		H_0 é verdadeira	H_0 é falsa
Decisão tomada	Rejeitar H_0	**ERRO TIPO I** (rejeitar H_0 quando H_0 é verdadeira)	Decisão correta
	Não rejeitar H_0	Decisão correta	**ERRO TIPO II** (não rejeitar H_0 quando H_0 é falsa)

Figura 5: Tipos de erros.

Ilustração: Ariana Santana

III. Nível de significância

Normalmente se estabelece o Erro tipo I como o mais danoso ou inconveniente para a pesquisa.

Exemplo clássico: Considere um júri popular, onde este tem que absolver ou condenar o réu, e o réu pode ser culpado ou inocente. Neste caso, temos que

a) Erro tipo I: condenar o réu quando o mesmo é inocente;

b) Erro tipo II: absolver o réu quando o mesmo é culpado.

Dessa forma, é tido como mais danoso cometer o erro tipo I.

Assim, é importante que se controle a probabilidade de se cometer o erro tipo I, de forma que ela seja a menor possível. A essa probabilidade, damos o nome de **nível de significância** e a denotamos por α .

$$P(\text{Erro tipo I}) = P(\text{rejeitar } H_0 \mid H_0 \text{ é verdadeira}) = \alpha .$$

Normalmente α é estabelecido como 1% ou 5%.

Além disso, denotaremos aqui a probabilidade de se cometer o erro tipo II por β , ou seja,

$$P(\text{Erro tipo II}) = P(\text{não rejeitar } H_0 \mid H_0 \text{ é falsa}) = \beta .$$

IV. Estatística de teste

A estatística de teste é a forma como condessaremos as informações contidas na amostra para ser capazes de tomar a decisão de rejeitar ou não a hipótese nula. Está pode estar associada a alguma distribuição de probabilidade ou não. Caso esteja associada a alguma distribuição de probabilidade, dizemos que o teste é **paramétrico**. Caso não seja necessário assumir que a estatística de teste segue alguma distribuição dizemos que o teste é **não paramétrico**.

V. Região de rejeição (RR)

Também chamada de **região crítica**, a **região de rejeição** compreende os valores para estatística de teste que se mostram demasiadamente destoantes do que foi assumido na hipótese nula.

Por outro lado, a região que compreende os valores da estatística de teste que são estatisticamente equivalentes ao valor assumido na hipótese nula é chamada **região de não rejeição** (RNR). O conceito de RNR em muito se assemelha ao conceito de intervalo de confiança.

O valor que delimita a fronteira entre as regiões de rejeição e não rejeição é chamado **valor crítico**. Para uma melhor compreensão, considere o seguinte exemplo, onde deseja-se testar

$$H_0: \mu = \mu_0 \quad vs \quad H_1: \mu \neq \mu_0 ,$$

lembrando que um teste para essas hipóteses é do tipo bilateral, por isso teremos dois valores críticos. Fixado o valor de α , suponha os valores críticos são dados por Z_α e $-Z_\alpha$, de forma que a RR e a RNR podem ser observadas na reta real abaixo.

Note que neste caso, para testes bilaterais, caso a estatística de teste calculada esteja entre $-Z_{\alpha}$ e Z_{α}, isso implica que a estatística calculada pertence a RNR, consequentemente, temos evidencias que indicam que não devemos rejeitar H_0. Por outro lado, caso a estatística calculada seja maior que Z_{α} ou menor que $-Z_{\alpha}$, as evidências apontam para rejeição de H_0.

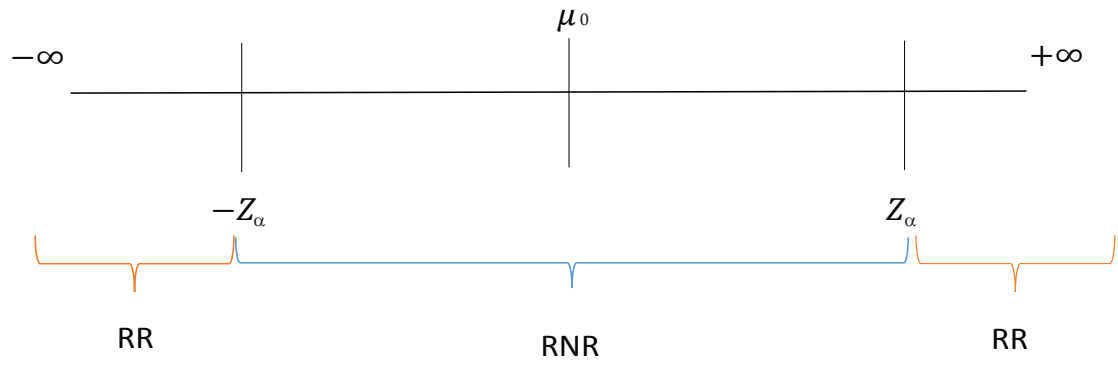


Figura 6: Regiões de rejeição e região de não rejeição. / Ilustração: Ariana Santana.

Considere agora que nosso desejo é testar as seguintes hipóteses

$$H_0: \mu \leq \mu_0 \quad vs \quad H_1: \mu > \mu_0 \ ,$$

Note que aqui, um teste para essas hipóteses é do tipo unilateral, por isso teremos apenas um valor crítico. Fixado o valor de α, suponha o valor crítico é dado por Z_{α} de forma que a RR e a RNR podem ser observadas na reta real abaixo.

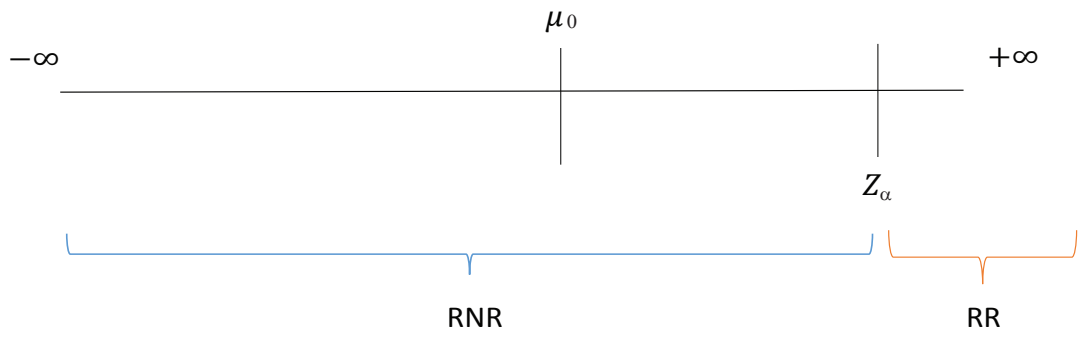


Figura 7: Região de rejeição e região de não rejeição. / Ilustração: Ariana Santana.

Nesse caso, se a estatística de teste calculada for maior que Z_{α}, estamos RR, logo rejeitamos H_0 .

Para o caso em que se deseja testar hipóteses como

$$H_0: \mu \geq \mu_0 \quad vs \quad H_1: \mu < \mu_0,$$

a RR é similar a apresentada na figura acima, entretanto, agora ela fica à esquerda da RNR. E a regra para decisão passa a ser, se a estatística de teste calculada for menor que Z_α, estamos RR, logo rejeitamos H_0.

Visto isso, em resumo temos que um teste de hipóteses estatístico é realizado a fim de se determinar se a amostra aleatória contém informação suficiente para rejeitar H_0, levando-se a concluir que H_1 é verdadeira.

VI. *p*-valor

O chamado *p*-valor (denotado por p) é uma medida de probabilidade de significância do teste de hipóteses. Ele é dado por

$$p = P(\text{estatistica de teste} > |\text{valor crítico}|),$$

ou seja, é a probabilidade de que o valor da estatística de teste seja maior que o valor crítico em módulo, ou ainda, é a probabilidade de que o valor da estatística de teste esteja na RR.

O *p*-valor serve como referência para rejeitar ou não H_0, de forma que se for maior que p o nível de significância (α) então rejeitamos H_0. E em caso contrário, não rejeitamos H_0. É importante conhecer o que é e para que serve o *p*-valor, entretanto, devido à complexidade que envolve seu cálculo, aqui não calcularemos o *p*-valor.

Ex. Exemplos

Exemplo 3.1: Assuma que uma montadora de carros compre de uma fabricante pinos de aço cuja especificação técnica para sua resistência média a ruptura é igual a 30 kgf. Para fazer um controle de qualidade, um lote desses pinos é escolhido ao acaso para verificar se este atende a especificação fornecida pelo fabricante.

Nesse caso, temos

H_0: O lote testado atende as especificações;

H_1: O lote testado não atende as especificações.

ou matematicamente

$$H_0: \mu = 30 \qquad vs \qquad H_1: \mu \neq 30.$$

Note que nesse caso, é necessário um teste bilateral.

Em resumo, para proceder um teste de hipóteses estatístico, precisamos de três informações básicas, são elas

1) Hipóteses estatísticas;

2) Estatística de teste; e

3) Valores críticos.

3.2 Teste para diferença de duas médias populacionais

Suponha agora que temos interesse em duas populações, e que temos duas amostras correspondentes, denotadas pelo conjunto X composto n_x por indivíduos, e pelo conjunto Y composto por n_y indivíduos. Chamemos as médias amostrais de X e Y de $\bar{X}$ e $\bar{Y}$, respectivamente.

A ideia aqui é bastante similar ao que foi visto sobre IC para diferença entre as médias populacionais. Entretanto, agora temos interesse em testar hipóteses sobre duas médias populacionais, mais especificamente, estaremos aqui interessados em testar se as médias populacionais de duas diferentes populações independentes são estatisticamente iguais ou não, com base em suas respectivas amostras. Ou ainda,

$$H_0: \mu_x=\mu_y \quad vs \quad H_1: \mu_x\neq\mu_y \ ,$$

que pode ainda ser expresso por,

$$H_0: \mu_x-\mu_y=0 \quad vs \quad H_1: \mu_x-\mu_y\neq 0 \ .$$

Os procedimentos para o teste de hipóteses, nesse caso, variam de acordo com as informações disponíveis sobre as variâncias populacionais (assim como na estimação de IC). Dessa forma, vamos tratar aqui três casos diferentes, em que as variâncias são conhecidas; as variâncias são desconhecidas, porém iguais; e as variâncias são desconhecidas e diferentes.

3.2.1 Variâncias conhecidas

Este é o caso mais simples, em que a estatística de teste será dada por

$$Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}},$$

onde σ_X^2 e σ_Y^2; são conhecidos.

Nesse caso, os valores críticos Z_α e $-Z_\alpha$ são obtidos da distribuição Normal padrão.

Ex. Exemplos

Exemplo 3.2: Uma máquina enche latas de refrigerante com base no peso. Duas amostras são retiradas ao acaso em turnos diferentes de produção. A primeira amostra tem 10 latas e a segunda tem 20 latas, com respectivos pesos médios e desvios padrão iguais a 174,6 g e 5 g; e 178,9 g e 6 g. Deseja-se saber se a máquina está bem regulada considerando um nível de significância de 1%.

Nosso interesse aqui é testar

$$H_0: \mu_x = \mu_y \quad vs \quad H_1: \mu_x \neq \mu_y \ .$$

Primeiramente vamos obter o valor da estatística de teste

Conhecido o valor da estatística de teste, precisamos agora conhecer o valor do nível

$$Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

$$Z = \frac{(174{,}6 - 178{,}9)}{\sqrt{\frac{5}{10} + \frac{6}{10}}}$$

$$Z = \frac{-4{,}6}{1{,}049} \approx -1{,}049 \ .$$

crítico que pode ser obtido da tabela da distribuição Normal padrão. Lembrando que consideramos um nível de significância de 1%, então

$$Z_\alpha = 2{,}575.$$

Como o valor calculado da estatística de teste está contido na RNR (-2,575<-1,049 < 2,575), então temos evidencias para a não rejeição de H_0.

3.2.2 Variâncias desconhecidas, porém, iguais

Vamos considerar agora que temos a informação de que as variâncias populacionais são iguais, ou seja, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Entretanto, as variâncias populacionais são desconhecidas, de forma que teremos que recorrer a suas estimativas dadas pela variância de cada amostra. Assim, a estatística de teste será dada por

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{S^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}},$$

em que

$$S^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2},$$

onde S_X^2 é a variância amostral da amostra X e S_Y^2 é a variância amostral da amostra Y.

De forma similar a metodologia para estimação do IC, aqui será utilizada a distribuição t de Student com graus de liberdade (v) dados por

$$v = n_X + n_Y - 2,$$

para obtenção dos valores críticos $T_{v,a}$ e $-T_{v,a}$.

Obs.: Assim como em IC, para obtenção dos valores críticos a partir da distribuição t de Studet, usar como referência valores **bilaterais**.

Ex. Exemplos

Exemplo 3.3: Deseja-se investigar a quantidade de nicotina em duas marcas de cigarros, coletadas as amostras tem-se que

Cigarro X: $n_X = 4;\ \bar{X} = 20;\ \text{e}\ S_X^2 = 6;$

Cigarro Y: $n_Y = 5;\ \bar{Y} = 21;\ \text{e}\ S_Y^2 = 5.$

Assumindo que as variâncias populacionais são desconhecidas, porém iguais, desejamos testar a hipótese de que

$$H_0: \mu_x = \mu_y \quad vs \quad H_1: \mu_x \neq \mu_y ,$$

considerando um nível de significância de 5%.

Começaremos pelo cálculo de S^2 .

$$S^2 = \frac{(4-1)*6+(5-1)*5}{4+5-2} = \frac{38}{7} \approx 5{,}4$$

Seguindo agora para obtenção do valor da estatística de teste temos

$$T = \frac{(20-21)}{\sqrt{5{,}428 * \left(\frac{1}{4}+\frac{1}{5}\right)}} = \frac{-1}{1{,}563} \approx -0{,}640$$

O próximo passo agora é obter o valor crítico, lembrando que agora ele será obtido da distribuição *t* de Student. Considerando que

$$v=n_X+n_Y-2=4+5-2=7\,,$$

e considerando ainda um nível de confiança de 5%, temos que $T_{v,\alpha}=2{,}365$.

Assim, como o valor calculado da estatística de teste está contido na RNR (-2,365<-0,64 < 2,365), então temos evidencias para a não rejeição de H_0 .

3.2.3 Variâncias desconhecidas e diferentes

Diferentemente do caso anterior, aqui consideramos que além de não conhecermos as variâncias populacionais, elas ainda são diferentes. Vamos novamente recorrer as variâncias das amostras como aproximação para as variâncias populacionais. Nesse caso a estatística de teste é dada por

$$T = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{S_X^2}{n_x}+\frac{S_Y^2}{n_Y}}}.$$

Assim como caso anterior, para obtenção dos valores críticos ($T_{v,\alpha}$ e $-T_{v,\alpha}$) aqui será utilizada a distribuição *t* de Student com graus de liberdade (v) dados por

$$v = \frac{\left(\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X-1}+\frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y-1}}.$$

Obs.: Assim como no caso anterior, para obtenção dos valores críticos a partir da distribuição *t* de Studet, usar como referência valores **bilaterais**.

Ex. Exemplos

Exemplo 3.4: Deseja-se investigar o desemprenho médio de duas turmas de matemática financeira, uma diurna (X) e uma noturna (Y). Para isso foram colhidas duas amostras aleatórias, das quais se observa as seguintes informações:

Turma *X*: $n_X = 4; \bar{X} = 8{,}1;$ e $S_X^2 = 2;$

Turma *Y*: $n_Y = 6; \bar{Y} = 7{,}7;$ e $S_Y^2 = 1{,}4.$

Assuma ainda que as variâncias populacionais são desconhecidas e diferentes, e suponha um nível de confiança de 5%.

Nosso interesse mais uma vez reside em testar

$$H_0: \mu_x = \mu_y \quad vs \quad H_1: \mu_x \neq \mu_y \ .$$

Comecemos pela estatística de teste

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} = \frac{8{,}1 - 7{,}7}{\sqrt{\frac{2}{4} + \frac{1{,}4}{6}}} = \frac{0{,}4}{0{,}856} \approx 0{,}467 \ .$$

Seguindo, agora precisamos definir os valores críticos, para isso precisamos primeiro obter o número de graus de liberdade da distribuição *t* de Student.

$$v = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}} = \frac{\left(\frac{2}{4} + \frac{1{,}4}{6}\right)^2}{\frac{\left(\frac{2}{4}\right)^2}{4-1} + \frac{\left(\frac{1{,}4}{6}\right)^2}{6-1}} = \frac{0{,}538}{0{,}094} \approx 5{,}723 \approx 6 \ .$$

Observando a tabela da distribuição *t* de Student, considerando 6 graus de liberdade e um nível de confiança de 5%, temos que os valores críticos são dados por $T_{v,\alpha} = 2{,}447$.

Assim, como o valor calculado da estatística de teste está contido na RNR (-2,447<0,467 < 2,447), então temos evidencias para a não rejeição de H_0 .

3.3 Teste para diferença de médias populacionais em amostras pareadas

Até aqui tratamos de amostras independentes, vamos considerar agora a existência de dependências entre as amostras. Nesse caso, as observações aparecem aos pares, por exemplo, cada indivíduo é observado duas vezes ao longo do tempo, de forma que teremos duas amostras considerando os mesmos indivíduos, porém, em contextos diferentes.

Ex. Exemplos

Exemplo 3.5: Seleciona-se uma amostra de funcionários de uma empresa. Cada funcionário realiza determinada função utilizando um método tradicional chamado método 1 para o fazer, e o tempo que ele leva para realizar a função é cadastrado na amostra X , em seguida, os mesmos funcionários devem realizar a mesma função utilizando um método alternativo chamado método 2, e o tempo que cada um leva para realizar a função é cadastrado na amostra Y .

Agora estaremos interessados em uma terceira amostra gerada a partir das duas amostras iniciais. Essa terceira amostra é denotada por d e é obtida fazendo a diferença entre x_i e y_i , para $i=1, \ldots, n$, em que n é o tamanho das amostras, e x_i e y_i representam os elementos que compõe as amostras X e Y . Assim temos que

$$d_i = x_i - y_i \ .$$

Note que agora temos uma única amostra (d) de tamanho n, e que é preciso que $n_X=n_Y=n$.

A ideia consiste em testar se a as médias populacionais diferem significativamente ou não, ou seja

$$H_0: \mu_x=\mu_y \quad vs \quad H_1: \mu_x \neq \mu_y \ ,$$

ou ainda, queremos testar se a média populacional da diferença entre as médias populacionais é estatisticamente igual a zero ou não, de forma que temos interesse em testar se

$$H_0: \mu_d=0 \quad vs \quad H_1: \mu_d \neq 0 \ .$$

Para testar essas hipóteses utiliza-se a seguinte estatística de teste

$$T = \frac{\bar{d}}{S_d/\sqrt{n}},$$

em que $\bar{d}$ e S_d são respectivamente a média simples e o desvio padrão da amostra d.

Os valores críticos são obtidos da distribuição t de Student com n-1 graus de liberdade.

Ex. Exemplos

Exemplo 3.6: Seis cobaias foram submetidas a dietas com determinadas rações durando uma semana cada dieta com cada ração. Ao término de cada semana são coletadas informações sobre os pesos (em gramas) das cobaias, as informações estão disponíveis abaixo

Cobaia	Dieta X	Dieta Y	d
1	635	640	-5
2	704	712	-8
3	662	661	1
4	560	558	2
5	603	610	-7
6	745	740	5

Ilustração: Ariana Santana

Com base nessas informações, temos que $\bar{d}$ = -2 e S_d = 5,366 . Assim a estatística de teste é igual a

$$T = \frac{\bar{d}}{S_d/\sqrt{n}} = \frac{-2}{5{,}366/\sqrt{6}} = \frac{-2}{2{,}191} = -0{,}913\,.$$

Considerando um nível de significância de 1%, e n-1=5 graus de liberdade, temos que $T_{\nu,\alpha}$=4,032.

Como o valor calculado da estatística de teste está contido na RNR (-4,032<-0,913 < 4,032), então temos evidências para a não rejeição de H_0.

3.4 Teste para diferença de duas proporções populacionais

A ideia aqui é bastante similar ao que foi visto em teste para diferença de duas médias populacionais. Mais uma vez, suponha que temos interesse em duas populações, e que temos duas amostras correspondentes, denotadas pelo conjunto X composto n_x por indivíduos, e pelo conjunto Y composto n_y por indivíduos. Chamemos as proporções de X e Y de $\widehat{p_X}$ e $\widehat{p_Y}$ respectivamente.

Nosso interesse reside em testar hipóteses sobre duas proporções populacionais, mais especificamente, estaremos aqui interessados em testar se as proporções populacionais de duas diferentes populações independentes são estatisticamente iguais ou não, com base em suas respectivas amostras. Ou ainda,

$$H_0: p_x = p_y \quad vs \quad H_1: p_x \neq p_y \ ,$$

que pode ainda ser expresso por

$$H_0: p_x - p_y = 0 \quad vs \quad H_1: p_x - p_y \neq 0 \ .$$

Para os procedimentos de teste, utilizaremos a seguinte estatística de teste

$$Z = \frac{\widehat{p_X} - \widehat{p_Y}}{\sqrt{\hat{\bar{p}}(1-\hat{\bar{p}})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}},$$

em que

$$\hat{\bar{p}} = \frac{n_X \widehat{p_X} + n_y \widehat{p_Y}}{n_X + n_Y} = \frac{P_X + P_Y}{n_X + n_Y},$$

com

P_x: número de indivíduos da amostra que possuem a característica de interesse;

P_y: número de indivíduos da amostra que possuem a característica de interesse.

Neste caso, os valores críticos são obtidos da distribuição Normal padrão.

Ex. Exemplos

Exemplo 3.7: Um candidato ao governo do estado deseja saber se sua intenção de votos é a mesma em duas cidades, para assim nortear os próximos passos de sua campanha política. Colhidas duas amostras aleatórias, obteve-se os seguintes resultados

Cidade X : n_x=500; e 116 se declaram eleitores do candidato;

Cidade Y : n_y=600 ; e 105 se declaram eleitores do candidato.

Podemos afirmar que as duas cidades têm a mesma proporção de eleitores do candidato considerando um nível de confiança de 5%?

A hipótese a ser testada nesse caso é

$$H_0: p_x\text{-}p_y=0 \quad vs \quad H_1: p_x\text{-}p_y\neq 0$$

Seguindo, calculemos a estatística de teste

$$\hat{\bar{p}} = \frac{P_X + P_Y}{n_X + n_Y} = \frac{116 + 105}{500 + 600} = 0{,}201\,.$$

$$Z = \frac{\widehat{p_X} - \widehat{p_Y}}{\sqrt{\hat{\bar{p}}(1-\hat{\bar{p}})\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} = \frac{\frac{116}{500} - \frac{105}{600}}{\sqrt{0{,}201 * (1-0{,}201) * \left(\frac{1}{500}+\frac{1}{600}\right)}}$$

$$Z = \frac{0{,}057}{0{,}024} \approx 2{,}375\,.$$

Considerando um nível de significância de 5%, temos que Z_α=1,96.

Como o valor calculado da estatística de teste está contido na RR (2,375> 1,96), então temos evidencias para a rejeição de H_0.

3.5 Comparando três ou mais médias

Até aqui, abordamos metodologias para comparação de duas médias populacionais. Entretanto, por vezes, é de interesse comparar três médias populacionais ou mais. Tendo isso em vista, nessa seção, serão apresentadas metodologias adequadas a esse cenário.

3.5.1 Teste *F*

Também conhecido como ANOVA (análise de variância) o teste *F* é recomendado para comparação as médias de vários grupos descritos por variáveis quantitativas que descreva um fator característico de cada grupo.

A ideia consiste em subdividir a **variabilidade total** (VT) das observações em duas fontes de variação distintas.

> **Variabilidade entre**(VE): diz respeito à variabilidade entre os grupos, ou seja, a variabilidade associada as diferenças entre os grupos;
>
> **Variabilidade dentro**(ND): diz respeito à variabilidade dentro de cada grupo, ou seja, a variabilidade associada às diferenças dentro de cada grupo.

Assim, temos que

$$VT=VE+VD\,.$$

Essa metodologia pode ser utilizada para *k* grupos distintos, e a ideia consiste em testar

$$H_0: \mu_1=\mu_2=\ldots=\mu_k \quad vs \quad H_1: \textit{ao menos uma igualdade não é atendida}\,.$$

A estatística de teste é dada por

$$F=\frac{variação\ entre\ grupos}{variação\ dentro\ dos\ grupos}=\frac{\sum_{i=1}^{k}\frac{n_i(\bar{\bar{x}}-\bar{x}_i)^2}{k-1}}{\sum_{i=1}^{k}\frac{(n_i-1)Sa_i^2}{n-k}},$$

em que $\bar{x}_1$ é a média simples de cada grupo (logo teremos $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$), $\bar{\bar{x}}$ a média simples da médias de cada grupo, *n* o tamanho total da amostra (considerando todos os grupos), n_i é número de observações em cada grupo *n* (logo, teremos $n_1,n_2,\ldots,n_k$), é número de grupos e Sa_i^2 é a variância amostral de cada grupo, qual é dada por

$$Sa_i^2=\sum_{j=1}^{n_i}\frac{(\bar{x}_i-x_{ji})^2}{n_i-1}\,.$$

O valor crítico ($F_{(k-1;n-1),\alpha}$) é obtido da distribuição *F* de Snedecor, também conhecida simplesmente por distribuição *F*, com *(k-1; n-1)* graus de liberdade. Note que para esse teste tomamos como referência apenas um valor crítico. Se o valor calculado da estatística de teste for maior que o quantil tabelado, rejeitamos H_0.

Existem algumas limitações ao uso desse procedimento de teste, é necessário que haja independência entre as variáveis, as observações sigam distribuição Normal e tenham mesma variância populacional.

Ex. Exemplos

Exemplo 3.8: Vinte e um ratos foram divididos em três grupos, em que cada grupo recebe uma dieta rica em vitaminas A, B e C, respectivamente, por uma semana. Após esse período, mediu-se o ganho de peso (em gramas) dos animais, e esses valores estão disposto na tabela a seguir

A	B	C
5,1	4,2	4,7
4,4	5,4	5,2
3,7	4,3	4,0
4,1	4,6	3,6
5,0	4,7	4,9
3,3	4,7	3,8
3,7	3,8	4,6

Ilustração: Ariana Santana

Neste caso, temos interesse em testar

$$H_0: \mu_1=\mu_2=\mu_3 \quad vs \quad H_1: \text{ao menos uma igualdade não é atendida}.$$

Da tabela acima, temos as seguintes informações

$$\begin{aligned}&\text{A: } \bar{x}_1 = 4{,}186;\ Sa_1^2 = 0{,}468;\ n_1 = 7;\\ &\text{B: } \bar{x}_2 = 4{,}443;\ Sa_2^2 = 0{,}27;\ n_2 = 7;\\ &\text{C: } \bar{x}_3 = 4{,}4;\ Sa_3^2 = 0{,}363;\ n_3 = 7.\end{aligned}$$

Lembrando ainda que $n=n_1+n_2+n_3=21$, e $k=3$.

Para calcular o valor da estatística de teste é necessário primeiro obter

$$\bar{\bar{x}} = \frac{4{,}186 + 4{,}443 + 4{,}4}{3} = 4{,}343\,.$$

Passemos agora para estatística de teste

$$F = \frac{\sum_{i=1}^{k} \frac{n_i(\bar{\bar{x}}-\bar{x}_i)^2}{k-1}}{\sum_{i=1}^{k} \frac{(n_i-1)Sa_i^2}{n-k}} =$$

$$F = \frac{\frac{7*(4{,}343-4{,}186)^2+7*(4{,}343-4{,}443)^2+7*(4{,}343-4{,}4)^2}{2-1}}{\frac{(7-1)*0{,}468+(7-1)*0{,}27+(7-1)*0{,}363}{21-3}}$$

$$F = \frac{0{,}133}{0{,}367} = 0{,}362\,.$$

Considerando um nível de confiança de 5%, e observado os valores na tabela referentes a distribuição *F* com (2;18) graus de liberdade, temos que $F_{(k-1;n-1),\alpha} = 3{,}55$.

Como o valor calculado da estatística de teste está contido na RR (0,362< 3,55), então temos evidencias para a não rejeição de H_0.

3.5.2 Teste de Tukey

O teste *F* é apropriado para comparação de várias médias simultaneamente, entretanto, caso a hipótese nula seja rejeitada, ou seja, ao menos uma média difere das demais, este teste não possibilita saber qual dos grupos destoa com relação aos demais.

O teste de Tukey promove a comparação de médias duas a duas, no caso de o teste *F* rejeitar a hipótese nula. Dessa forma, nossas hipóteses de interesse agora são

$$H_0: \mu_i \neq \mu_k,\ \forall\, i \neq k \qquad vs \qquad H_1: \textit{ao menos um par de médias é desigual}\,.$$

O procedimento de Tukey usa a distribuição da estatística de variação "studentizada"

$$q = \frac{\bar{x}_{max} - \bar{x}_{min}}{\sqrt{QM_d/n}},$$

em que $\bar{x}_{max}$ e $\bar{x}_{min}$ são a maior e a menor médias entre os grupos, e QM_d é o denominador da estatística de teste do teste *F*, ou seja,

$$QM_d = \sum_{i=1}^{k} \frac{(n_i - 1)Sa_i^2}{n-k},$$

caso o valor absoluto de q exceda

$$T_\alpha = q_\alpha(k, n-k)\sqrt{\frac{QM_d}{n_i}},$$

o teste indica que as médias dos grupos são significativamente diferentes.

Ex. Exemplos

Exemplo 3.9: Crianças foram separadas em três grupos aos quais foram dados diferentes níveis de motivação para estudar matemática (baixa, média e alta). Aplicou-se um exame com as crianças e tabulou-se as notas que segue

A	B	C
4	12	1
5	8	3
4	10	4
3	5	6
6	7	8
10	9	5
1	14	3
8	9	2
9	4	2

Ilustração: Ariana Santana

A princípio foi aplicado o teste F para avaliar se as médias são iguais. O resultado por $F=7,82 > F_{2;24;0,05} = 3,403$, logo rejeita-se H_0.

Sabendo que $\bar{x}_A = 5,11,\ \bar{x}_B = 8,67,\ \bar{x}_C = 3,78\ QM_d = 7,35$, utilize o teste de Tukey para investigar quais médias são diferentes.

Vamos testar as médias de A e B, A e C, e B e C.

$$q_{A\,e\,B} = \frac{\bar{x}_{max} - \bar{x}_{min}}{\sqrt{QM_d/n}} = \frac{8,67 - 5,11}{\sqrt{7,35/9}} = 3.939\,,$$

$$q_{A\,e\,C} = \frac{\bar{x}_{max} - \bar{x}_{min}}{\sqrt{QM_d/n}} = \frac{5,11 - 3,78}{\sqrt{7,35/9}} = 1,472\,,$$

$$q_{B\,e\,C} = \frac{\bar{x}_{max} - \bar{x}_{min}}{\sqrt{QM_d/n}} = \frac{8,67 - 3,78}{\sqrt{7,35/9}} = 5,411\,.$$

Considerando um nível de significância de 5%, observando a tabela de Tukey, temos que $q_\alpha(k,n-k)=3,61$. De forma que

$$T_\alpha = q_\alpha(k, n-k)\sqrt{\frac{QM_d}{n}} = 3,61 * \sqrt{\frac{7,35}{9}} = 3,262\,.$$

Note que o teste aponta que apenas as médias dos grupos A e C são iguais.

3.6 Noções de testes não paramétricos

Todos os procedimentos de teste de hipóteses abordados até aqui eram considerados paramétricos, ou seja, era necessário se fazer suposições sobre a distribuição da variável de interesse. Por vezes, essa suposição pode não se mostrar razoável. Uma alternativa aos testes paramétricos são os chamados testes não paramétricos. Aqui apresentaremos dois desses testes para comparação de médias.

3.6.1 Teste de Mann-Whitney

Também conhecido como teste *U* de Mann-Whitney, esse teste é correspondente não paramétrico indicado quando se tem interesse em testar se as médias de duas populações são iguais a partir de duas amostras aleatórias e independentes. Entretanto, a interpretação das hipóteses muda um pouco, de forma que se estabelece que o interesse aqui é testar

H_0: *os tratamentos tem efeios similares* *vs* H_1: *os tratamentos não tem efeitos similares,*

em que o tratamento se refere a característica de interesse quantificada para descrever o valor quantitativo de cada observação.

Os testes não paramétricos são baseados em posto, que é um sistema de classificação das observações disponíveis, onde se atribui um número natural para indicar a posição de cada observação. Assim, para proceder a realização do teste de Mann-Whitney precisamos primeiro fazer a ordenação do posto dos dados. Para as observações empatadas, deve-se atribuir a média dos postos correspondentes.

As estatísticas de teste são dadas por

$$U = n_X n_Y + \frac{n_X(n_X+1)}{2} - R_X, \qquad U = n_X n_Y + \frac{n_Y(n_Y+1)}{2} - R_Y\,,$$

onde R_x é a soma dos postos do grupo X e R_y é a soma dos postos do grupo Y.

Deve-se escolher o menor U obtido e comparar com o valor crítico. Se o menor valor calculado for menor ou igual ao valor crítico, rejeita-se H_0.

Se $n = n_1 + n_2 \leq 20$, utilizamos os valores críticos da tabela de Mann-Whitney.

Caso contrário, calculamos

$$z = \frac{U - \mu_R}{\sigma_R}, \qquad \mu_R = \frac{n_X n_Y}{2}, \qquad \sigma_R = \sqrt{\frac{n_X n_Y (n_X + n_Y + 1)}{12}},$$

e comparamos o valor de z com os valores críticos da distribuição Normal padrão.

Ex. Exemplos

Exemplo 3.10: Deseja-se comparar a eficácia da publicidade de dois produtos. Assim, ofereceu-se cada um dos produtos a um consumidor e pediu que lhe atribuísse uma nota, o resultado foi

Marca X	Marca Y
3	9
4	7
2	5
6	10
2	6
5	8

Ilustração: Ariana Santana

As marcas têm notas médias equivalentes?

Aqui estamos interessados em testar

H_0: *os tratamentos tem efeios similares* *vs* H_1: *os tratamentos não tem efeitos similares* .

Primeiro precisamos definir o posto das observações

Marca X		Marca Y	
Nota	Posto	Nota	Posto
2	1,5	5	5,5
2	1,5	6	7,5
3	3	7	9
4	4	8	10
5	5,5	9	11
6	7,5	10	12
Soma	**23**	**Soma**	**55**

Ilustração: Ariana Santana

Assim, temos que $R_x = 23$ e $R_y = 55$. Seguindo agora, vamos calcular U.

$$U = n_X n_Y + \frac{n_X(n_X + 1)}{2} - R_X = 6 * 6 + \frac{6 * (6 + 1)}{2} - 23 = 34,$$

$$U = n_X n_Y + \frac{n_Y(n_Y + 1)}{2} - R_Y = 6 * 6 + \frac{6 * (6 + 1)}{2} - 55 = 2.$$

O menor valor de U é 2.

Da tabela de Maan-Whitney, temos que o valor crítico considerando um nível crítico de 5%, é igual a 5. Como o valor calculado de U é menor que o valor tabelado ($2 < 5$), então o teste indica que devemos rejeitar H_0.

3.6.2 Teste de Kruskal Wallis

O teste de Kruskal Wallis é o equivalente não paramétrico para o teste F. Aqui, assim como no caso anterior, é necessário que a variável de interesse esteja em escala ordinal. É necessário que existam ao menos 5 indivíduos e ao menos 3 grupos para compor a amostra. Assim, de forma similar ao teste F, deseja-se testar algo como

H_0: *os tratamentos tem efeios similares* *vs* H_1: *os tratamentos não tem efeitos similares.*

A estatística de teste é dada por

$$H = \left(\frac{12}{n(n+1)}\right)\left(\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right) - 3 * (n+1).$$

Se o valor de H calculado for maior que o valor crítico correspondente da tabela da distribuição qui-quadrado, com k-1 graus de liberdade, então rejeitamos H_0.

Caso a amostra seja muito pequena, por exemplo, $k=3$ e a amostra de cada grupo contém cinco ou menos elementos, a aproximação pela distribuição qui-quadrado não é boa.

Caso um ou mais valores observados apareçam mais que uma vez, é necessário que se proceda uma correção no valor de H, de forma que a estatística de teste será dada por

$$H = \frac{\left(\frac{12}{n(n+1)}\right)\left(\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right) - 3*(n+1)}{1 - \frac{\sum_{l=1}^{g} t_l^3 - t_l}{n^3 - n}},$$

em que t_l é tamanho do grupo de elementos repetidos l, e g é a quantidade de grupos. Os elementos que não se repetem correspondem cada um a um grupo de tamanho 1.

Ex. Exemplos

Exemplo 3.11: Toma-se aleatoriamente três amostras em três capitais diferentes, onde pergunta-se a cada indivíduo quantas vezes ele foi ao shopping no mês anterior, o resultado está disposto na tabela a seguir

Grupo 1	Grupo 2	Grupo 3
20	12	8
4	21	22
7	9	10
2	0	5
17	14	6
3	1	19

Ilustração: Ariana Santana

Com base nessas informações, podemos admitir que os tratamentos são similares?

Temos interesse aqui em testar se

H_0: os tratamentos tem efeios similares vs H_1: os tratamentos não tem efeitos similares.

Primeiramente, vamos atribuir os postos.

Grupo 1		Grupo 2		Grupo 3	
Quantidade	Posto	Quantidade	Posto	Quantidade	Posto
20	16	12	12	8	9
4	5	21	17	22	18
7	8	9	10	10	11
2	3	0	1	5	6
17	14	14	13	6	7
3	4	1	2	19	15
Soma	**50**	**Soma**	**55**	**Soma**	**66**

Ilustração: Ariana Santana

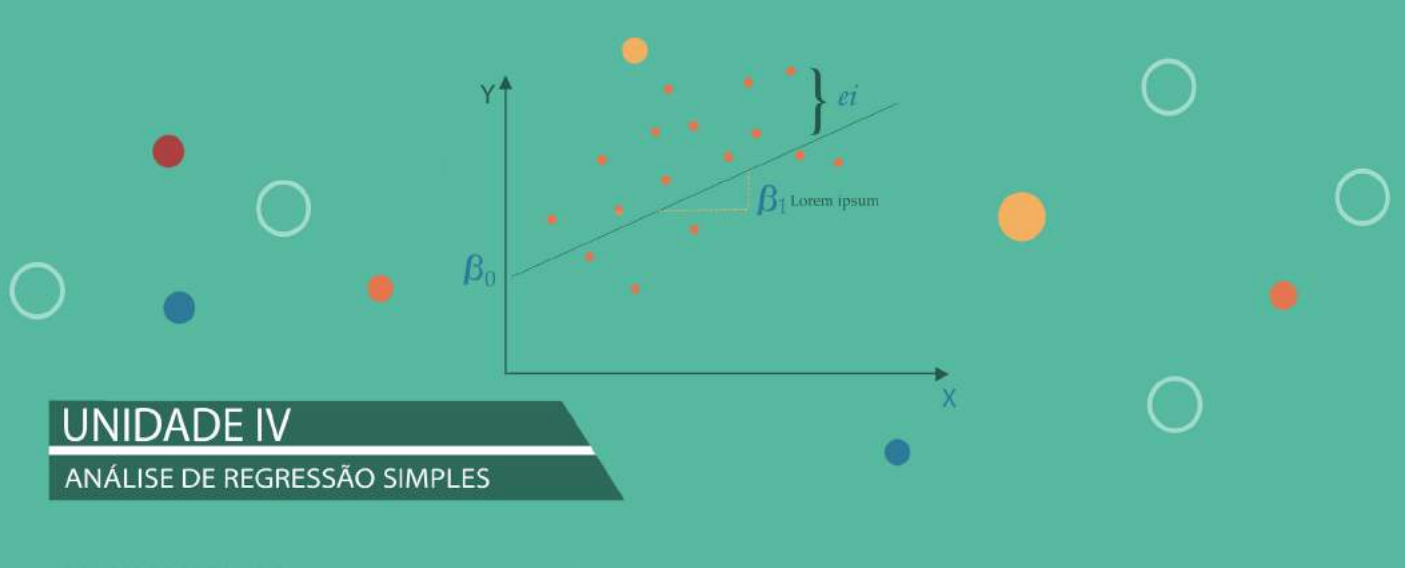
Vamos agora obter o valor da estatística de teste

$$H = \left(\frac{12}{n(n+1)}\right)\left(\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right) - 3*(n+1)$$

$$H = \left(\frac{12}{18*(18+1)}\right)\left(\frac{(50)^2}{6}+\frac{(55)^2}{6}+\frac{(66)^2}{6}\right) - 3*(18+1)$$

$$H = \left(\frac{12}{342}\right)\left(\frac{9881}{6}\right) - 57 = 0{,}784\,.$$

Considerando um nível de confiança de 5%, da tabela qui-quadrado temos que o valor crítico é dado por 5,991. Como o valor calculado de H é menor que o valor tabelado (0,784<5,991), então o teste indica que não devemos rejeitar H_0.



UNIDADE 4 –ANÁLISE DE REGRESSÃO SIMPLES

Aqui veremos a técnica estatística provavelmente mais utilizada para trabalhos científicos. Por vezes, o pesquisador acredita que uma variável pode influenciar fortemente o comportamento de outra variável, entretanto, o contrário não seria válido. Por exemplo, as variáveis consumo e renda. É fácil ver que para se poder consumir mais produtos é necessário dispor de mais renda. Entretanto, a renda do indivíduo não varia caso ele consuma mais ou menos (lembrando que renda é fluxo de dinheiro e não estoque).

Note que nesse caso temos duas variáveis de interesse em que uma exerce influência sobre a outra, entretanto, sem reciprocidade. Para avaliar a direção o grau e a relevância dessa influência podemos utilizar a análise de regressão simples.

4.1 Introdução

Regressão é um termo introduzido por Francis Galton (1889) ao analisar dados sobre alturas de pais e seus filhos. Embora houvesse uma tendência de pais altos terem filhos altos, e de pais baixos terem filhos baixos, a altura média dos filhos de pais de uma dada altura tendia a se deslocar ou "regredir" até a altura média da população como um todo.

Hoje em dia, regressão ocupa-se do estudo da dependência de uma variável (chamada variável endógena, resposta ou dependente), em relação a uma ou mais variáveis, chamadas variáveis explicativas (ou exógenas). E tem como objetivo estimar a média (da população) ou valor médio da variável dependente em termos dos valores conhecidos (ou fixos) das explicativas.

Aqui estamos interessados em estudar a chamada **regressão linear simples**, em que se analisa a influência de uma única variável explicativa sobre uma variável dependente, assumindo uma função linear para expressar essa relação de influência. O caso em que existe mais que uma variável explicativa chamamos **regressão linear** múltipla.

É importante aqui fazer algumas distinções para melhor compreender o que é uma regressão. A primeira delas é que apesar de regressão lidar com a dependência de uma variável em relação a outras variáveis, ela não implica necessariamente em causa. Relações estatísticas não estabelecem relações causais, por mais fortes que essas possam ser.

Ex. Exemplos

Exemplo 4.1: Um estudo com regressão pode indicar que fumantes estão mais propensos a ter câncer de pulmão que não fumantes, mas não pode estabelecer que fumar causa câncer de pulmão, isso fica a critério da medicina.

Outra distinção importante diz respeito a diferença entre regressão e correlação. Análise de correlação tem como objetivo medir a intensidade ou o grau de associação linear entre duas variáveis, sob um contexto de influência mútua. Na análise de regressão tentamos estimar ou prever o valor médio de uma variável com base nos valores fixados de outras variáveis. Note que correlação é uma medida de influência bilateral, enquanto que em regressão, assume-se que uma variável influência a outra.

4.2 O modelo

A ideia consiste em estabelecer uma forma funcional e linear nos parâmetros para a variável que se deseja modelar (Y), com relação as informações disponíveis (X). Por exemplo,

$$y_i=\beta_0+\beta_1 x_i+e_i, \qquad i=1,\dots,n$$

em que y_i e x_i são respectivamente, os i-ésimos elementos das minhas variáveis dependente e independente. β_0 é o chamado intercepto. β_1 é chamado coeficiente angular. Aqui temos que β_0 e β_1 são nossos parâmetros de interesse. E e_i é o i-ésimo erro não observável.

Note que estamos supondo uma relação de linearidade entre as variáveis, e estamos supondo ainda que uma variável é perfeitamente explicada por uma segunda variável, por isso é imprescindível a utilização do termo de erro, uma vez que muito dificilmente no mundo real essas duas suposições serão satisfatórias. De forma que, desde o princípio, já temos em mente que o modelo não é perfeito e poderá cometer erros.

Para minimizar esse problema, podemos escolher os valores de β_0 e β_1 que tornam o e_i menor possível. Ou seja, temos interesse em estimativas de β_0 e β_1 (denotadas por $\widehat{\beta_0}$ e $\widehat{\beta_1}$)

que me ofereçam o menor erro possível. O estimador capaz de me oferecer isso é o chamado **Estimador de Mínimos Quadrados Ordinários**, ou simplesmente **MQO**.

Na prática e_i é não observável, por isso, o método de estimação leva em consideração uma aproximação do erro, a qual chamaremos de resíduo, e será denotado por $\widehat{e_i}$. Assim, temos que os resíduos são dados por

$$\widehat{e_i} = y_i - \widehat{y_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i .$$

A ideia do modelo consiste em traçar uma reta média capaz de caracterizar a relação de influência de X sobre Y de forma que a distância entre cada observação e a reta média seja a menor possível.

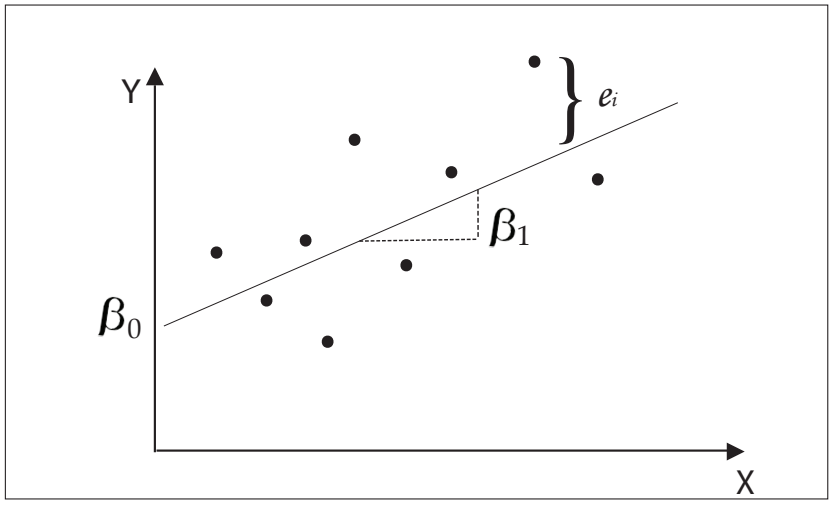


Figura 8: Reta de regressão.

Ilustração: Ariana Santana

Para se obter o estimador de MQO é necessária a suposição de algumas hipóteses básicas. São elas

I. O modelo está corretamente especificado;

II. X é não estocástico;

III. Os erros têm média zero;

IV. Os erros têm variância constante igual a σ^2;

V. Os erros são independentes;

VI. Os erros têm distribuição Normal.

A primeira hipótese nos diz que o modelo postulado é o modelo correto, ou seja, Y pode realmente ser explicado por X. A segunda hipótese supõe que X é não aleatório, ou seja, seus valores são fixos. A terceira hipóteses supõe que os erros são em média iguais a zero, ou seja, às vezes meu erro será positivo, outras vezes negativo, de forma que em média, esses erros se anulam.

A quarta hipótese, é comumente chamada de homocedasticidade. E supõe que todos os erros têm variância igual. A quinta suposição assume que os erros são independentes entre si, ou seja, um erro com relação a uma observação não influencia em outro erro com relação a outra observação. Por fim, a sexta suposição supõe que os erros seguem distribuição Normal. Note que podemos resumir as hipóteses III, IV, V e VI simplesmente em são independente e identicamente distribuídos (iid) tais que

$$e_i \sim N(0,\sigma^2).$$

Sob essas hipóteses básicas, podemos seguir para derivação do modelo. O Método de MQO, consiste basicamente é obter os $\widehat{\beta_0}$ e $\widehat{\beta_1}$ que minimizam a soma dos quadrados dos resíduos, ou seja, deseja-se obter $\widehat{\beta_0}$ e $\widehat{\beta_1}$, tais que minimizam

$$\sum_{i=1}^{n} \widehat{e_i}^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2 .$$

A obtenção dos estimadores envolve cálculo diferencial para solução do problema de minimização. Os estimadores que MQO para β_0 e β_1 são dados por

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X},$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})},$$

em que

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ e } \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Podemos ainda obter as estimativas das variâncias dos estimadores por

$$Var(\widehat{\beta_0}) = \frac{\sum_{i=1}^{n} x_i^2}{n * \sum_{i=1}^{n}(x_i - \bar{X})^2}\hat{\sigma}^2, \qquad Var(\widehat{\beta_1}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}, \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}.$$

Ex. Exemplos

Exemplo 4.2: Considerando os dados disposto na tabela a abaixo, estime os parâmetros do seguinte modelo de regressão, bem como seus erros padrão.

$$Y_i=\beta_0+\beta_1 X_i+e_i.$$

Observação	1	2	3	4		Média
Y	20	35	28	40		30,75
X	3	7	6	10		6,5

Note que a fórmula para obtenção de $\widehat{\beta_0}$ depende do conhecimento prévio de $\widehat{\beta_1}$, assim, começaremos por calcular este

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})}$$

$$\widehat{\beta_1} = \frac{(3-6{,}5)*(20-30{,}75)+(7-6{,}5)*(35-30{,}75)+}{(3-6{,}5)+(7-6{,}5)+(6-6{,}5)+(10-6{,}5)}$$

$$\frac{+(6-6{,}5)*(28-30{,}75)+(10-6{,}5)*(40-30{,}75)}{(3-6{,}5)+(7-6{,}5)+(6-6{,}5)+(10-6{,}5)}$$

$$\widehat{\beta_1} \approx 2{,}94\,.$$

Agora podemos obter $\widehat{\beta_0}$,

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X} = 30{,}75 - 2{,}94 * 6{,}5 \approx 11{,}64\,.$$

Assim o modelo de regressão estimado é

$$Y_i = 11{,}64+2{,}94 * X_i\,.$$

Seguindo agora para o cálculo das variâncias dos estimadores, precisamos primeiro definir $\hat{e}_i$, que é dado por

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \widehat{\beta_0} - \widehat{\beta_1}X_i = Y_i - 11{,}64 + 2{,}94 * X_i\,,$$

aplicando a essa fórmula os dados da tabela, temos que

$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_3$	$\hat{e}_4$
-0,46	2,78	-1,28	-1,04

Ilustração: Ariana Santana

logo, aplicando a fórmula para estimação da variância dos resíduos, tem-se

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2} = \frac{(-0{,}46)^2 + (2{,}78)^2 + (-1{,}28)^2 + (-1{,}04)^2}{4-2} = 5{,}33\,,$$

passando então para o cálculo das variâncias, temos

$$Var(\widehat{\beta_0}) = \frac{\sum_{i=1}^{n} x_i^2}{n * \sum_{i=1}^{n} (x_i - \bar{X})^2} \hat{\sigma}^2$$

$$Var(\widehat{\beta_0}) = \frac{3^2 + 7^2 + 6^2 + 10^2}{4 * ((3-6{,}5)^2 + (7-6{,}5)^2 + (6-6{,}5)^2 + (10-6{,}5)^2)} * 5{,}33$$

$$Var(\widehat{\beta_0}) = 1{,}94 * 5{,}33 \approx 10{,}34\,.$$

E

$$Var(\widehat{\beta_1}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2} = \frac{5{,}33}{(3-6{,}5)^2 + (7-6{,}5)^2 + (6-6{,}5)^2 + (10-6{,}5)^2}$$

$$Var(\widehat{\beta_1}) \approx 0{,}21\,.$$

Conhecida a forma dos estimadores, é importante agora saber os interpretar. $\widehat{\beta_0}$ é o intercepto, como vimos anteriormente, ele pode ser interpretado como o valor médio de Y quando X é nulo. Já $\widehat{\beta_1}$ (o coeficiente angular) interpretado como o impacto que é causado em Y dado uma variação marginal em X, ou ainda, é a magnitude (ou grau de influência) de X sobre Y.

4.3 Validação do modelo

Feita a estimação do modelo, devemos seguir o passo seguinte que é a validação do mesmo. Nesta etapa, procedemos testes de hipóteses a fim de se garantir a validade das hipóteses postuladas.

Como aqui abordaremos apenas o caso de regressão simples, iremos nos ater a testar apenas as hipóteses IV, V e VI, que são referentes ao erro do modelo. Existem diversos testes para avaliar cada uma dessas hipóteses, a maior parte deles exige um pouco mais de profundidade teórica para um melhor entendimento, de forma que aqui vamos apresentar alguns testes simples para cada caso. Em caso de os testes serem baseados em conteúdos mais avançados, apenas mencionaremos o teste. Lembrando que na prática todos

esses testes já estão programados em diversos *softwares*. Não sendo imprescindível o profundo conhecimento da teoria por trás do teste para sua utilização.

i. Testando se os erros têm variância constante

Essa hipótese é de fundamental para se garantir as boas propriedades dos estimadores. Caso ela seja violada, dizemos que os erros têm problemas de heteroscedasticidade. Um dos testes mais comuns é o teste de Goldfeld-Quandt, o qual tem como hipóteses de teste

$$H_0: \text{variância dos erros constante} \quad vs \quad H_1: \text{não } H_0$$

A ideia do teste consiste ordenar as n observações de forma crescente com respeito a variável explicativa. Em seguida, divide-se a amostra em três partes, de forma que a parte central contenha 25% das observações. Em seguida, estima-se duas regressões para as partes 1 e 3. A estatística de teste é dada por

$$F_{GQ} = \frac{\frac{SQR_{reg2}}{n_3 - p + 1}}{\frac{SQR_{reg1}}{n_1 - p + 1}},$$

em que n_3 e n_1 são os números de observações das partes 1 e 3. é o número de observações centrais que forma omitidas. SQR_{reg1} é a soma dos quadrados dos resíduos da regressão feita com a primeira parte dos dados. E SQR_{reg2} é a soma dos quadrados dos resíduos da regressão feita com a terceira parte dos dados.

Sob homocedasticidade, F_{GQ} deve ser próximo de 1. Os valores críticos são obtidos da distribuição F, com $(n_3\text{-}p+1, n_1\text{-}p+1)$ graus de liberdade.

ii. Testando se os erros têm correlação

Essa hipótese normalmente é testada quando se deseja modelar uma série de tempo, ou seja, as observações são obtidas a partir de um indivíduo que é acompanhado ao longo do tempo. Os testes mais comuns nesse caso são os testes de Durbin-Watson, *h* de Durbin e Breuch-Goldfrey. Esses testes são baseados em modelo de séries temporais, e avaliam se há ou não correlação entre os erros com base nos resíduos.

iii. Testando se os erros seguem distribuição Normal

Essa hipótese é especialmente importante na etapa posterior a estimação do modelo, quando se deseja validar as informações obtidas a partir do mesmo. Isso porque a maior parte dos testes de hipóteses utilizados parte do princípio que os erros seguem distribuição Normal para derivar a distribuição de sua estatística de teste.

Os testes mais comuns para indicar a normalidade dos erros são os testes de Bera-Jarque e Kolmogorov-Smirnov. Ambos se baseiam nos resíduos do modelo de regressão.

Além das hipóteses do modelo, convém também testar a validade dos valores estimados para β_0 e β_1. A ideia do teste consiste em avaliar se o valor estimado é estatisticamente diferente de zero ou não. Assim testa-se

$$H_0:\beta_i=0 \qquad vs \qquad H_1:\beta_i\neq 0\,.$$

Para isso utiliza-se a seguinte estatística de teste

$$t=\frac{\widehat{\beta_i}}{\sqrt{Var(\widehat{\beta_i})}},$$

e a região crítica é obtida da tabela da distribuição *t* de Student com *n - 2* graus de liberdade.

Esse teste é especialmente importante por ser capaz de identificar se X tem ou não influencia sobre Y. Caso H_0 seja rejeitada, temos indícios de que X influencia Y.

Ex. Exemplos

Exemplo 4.3: Considere os dados do Exemplo 4.2 e teste as hipóteses de que os coeficientes estimados são iguais a zero.

Temos que $\widehat{\beta_0}=11{,}64$, $\widehat{\beta_1}=2{,}94$, $Var(\widehat{\beta_0})=10{,}34$ e $Var(\widehat{\beta_1})=0{,}21$

Primeiro testaremos

$$H_0:\beta_0=0 \qquad vs \qquad H_1:\beta_0\neq 0\,.$$

A estatística de teste é dada por

$$t=\frac{\widehat{\beta_0}}{\sqrt{Var(\widehat{\beta_0})}}=\frac{11{,}64}{\sqrt{10{,}34}}\approx 3{,}62\,.$$

Da tabela *t* de Student, considerando 2 graus de liberdade e um nível de significância de 5%, temos que $t_{v,\alpha}=4{,}303$. Como 3,62 está na RNR, não podemos rejeitar a hipótese de que $\beta_0=0$.

Testando agora

$$H_0:\beta_1=0 \qquad vs \qquad H_1:\beta_1\neq 0\,.$$

A estatística de teste de dada por

$$t=\frac{\widehat{\beta_1}}{\sqrt{Var\left(\widehat{\beta_1}\right)}}=\frac{2{,}94}{\sqrt{0{,}21}}\approx 6{,}41\,.$$

Como $t_{v,\alpha}=4{,}303$, temos que a estatística de teste calculada está contida na RR, logo rejeitamos a hipótese de que $\beta_1=0$.

4.4 Observando os resíduos

A análise gráfica dos resíduos pode fornecer informações valiosas sobre o comportamento do termo de erro. Ressaltando que a análise gráfica por si só não é suficiente para validação do modelo por ser muitas vezes subjetiva. Ela é importante para fornecer indícios, entretanto, é sempre necessário proceder os testes de hipóteses.

Comecemos pelo **gráfico de dispersão dos resíduos**. Este gráfico é útil para avaliar se os erros têm média zero e se a variância dos erros é constante. O ideal é que os dados se concentrem em torno de zero, tenham uma dispersão constante e estejam dispersos aleatoriamente (sem padrão definido), como no exemplo que segue.

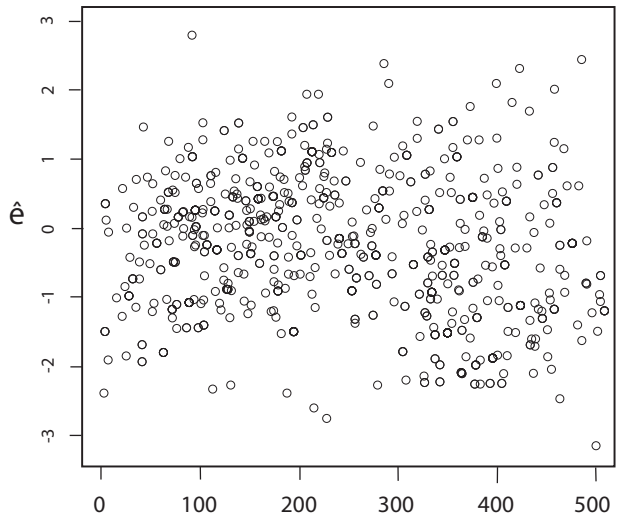


Figura 9: Gráfico de dispersão dos resíduos. / Ilustração: Ariana Santana

Caso os resíduos não se apresentem em torno de zero, tem-se indícios de que o erro não tem média zero. Caso os resíduos não apresentem uma dispersão constante, tem-se indícios de que os erros são heteroscedasticos. E caso a dispersão não seja aleatória, há indícios de relação não linear entre as variáveis. Interpretação similar tem o **gráfico de resíduos contra valor ajustado**. Segue exemplo.

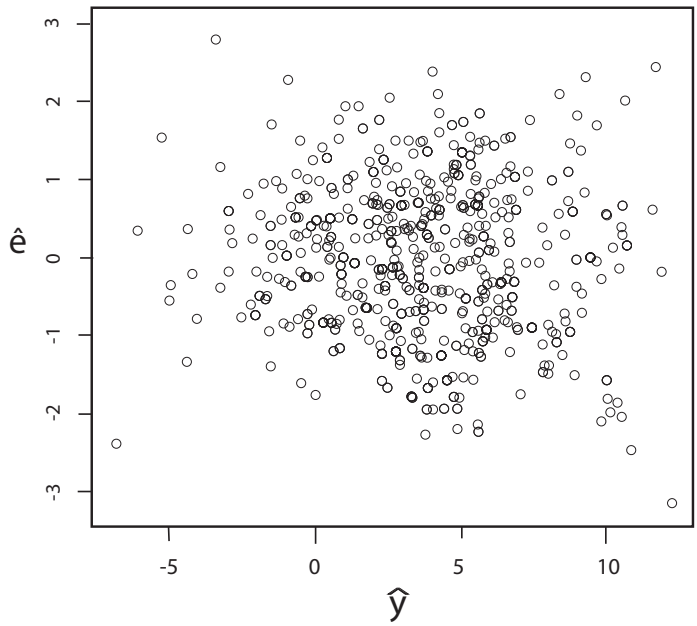


Figura 10: Gráfico de resíduos contra valor ajustado. / Ilustração: Ariana Santana.

Outro gráfico bastante útil para avaliar resíduos é o **histograma dos resíduos**. Esse gráfico serve para se ter uma noção da distribuição empírica dos dados. Como a função de densidade de probabilidade da distribuição Normal tem formato de sino, espera-se que o histograma dos resíduos também o tenha. Abaixo segue exemplo de histograma de resíduos normais.

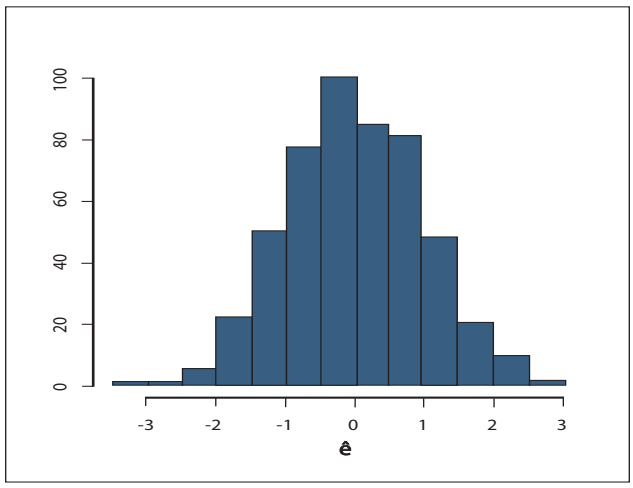


Figura 11: Histograma dos resíduos. / Ilustração: Ariana Santana.

4.5 Coeficiente de determinação

O coeficiente de determinação, denotado por R^2, é uma métrica que aponta a qualidade do ajuste da reta de regressão aos dados. É uma medida que se encontra entre zero e um, ou seja, $0 \leq R^2 \leq 1$, em que quanto mais próximo de 1, melhor o ajuste do modelo, e quanto mais próximo de 0, pior o ajuste do modelo. Podemos obter o R^2 por meio da seguinte fórmula

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2} = \frac{SQE}{SQT},$$

em que SQE é a soma dos quadrados explicados pelo modelo de regressão e SQT é a soma dos quadrados totais.

Assumindo que podemos decompor a SQT de tal forma que, SQT = SQE + SQR, em que SQR é a soma dos quadrados dos resíduos. Podemos representar alternativamente o por R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{e}_i^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2} = 1 - \frac{SQR}{SQT}.$$

O R^2 pode ainda ser interpretado como a proporção da variação total de Y explicada pelo modelo de regressão.

Ex. Exemplos

Exemplo 4.4: Considerando novamente o Exemplo 4.2, calcule o R^2 do modelo ajustado.

Lembrando que

$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_3$	$\hat{e}_4$
-0,46	2,78	-1,28	-1,04

Ilustração: Ariana Santana

Então,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sum_{i=1}^{n} (y_i - \bar{Y})^2}$$

$$R^2 = 1 - \frac{(-0{,}46)^2 + (2{,}78)^2 + (-1{,}28)^2 + (-1{,}04)^2}{(20 - 30{,}75)^2 + (35 - 30{,}75)^2 + (28 - 30{,}75)^2 + (40 - 30{,}75)^2}$$

$$R^2 \approx 0{,}953\,.$$

4.6 Aplicação prática com auxílio de software

O procedimento de regressão é deveras trabalhoso para ser realizado manualmente, principalmente em um contexto em que temos muitas observações. Para facilitar as estimações, existem uma variedade grande de *softwares* que possuem algoritmos capazes de realizar todos os cálculos para grandes bases de dados (R, STATA, SPSS, Excel, entre outros). Veremos aqui como proceder uma regressão linear simples utilizando o *software* Excel da Microsoft.

Em nossa aplicação, teremos interesse em explicar a **taxa de mortalidade infantil** nos municípios baianos (Y) por meio da **renda média das pessoas ocupadas** de cada cidade (X). De forma que o modelo a ser estimado pode ser expresso por

$$Y_i = \beta_0 + \beta_1 X_i + e\,,$$

para $i=1,\ldots,n$.

Os dados para aplicação são provenientes do Atlas do Desenvolvimento Humano, compreendem 418 municípios que foram observados no ano de 2010.

Antes de iniciar as estimações, note que é de se esperar que a variável renda média tenha impacto negativo sobre a variável taxa de mortalidade, posto que se a renda média é elevada, a cidade tem habitantes de maior poder aquisitivo, que podem pagar mais por remédios e tratamentos.

Dado que as informações já estão na planilha do Excel, para o procedimento de regressão, seleciona-se o menu **DADOS** e em seguida, na faixa de opções correspondente, seleciona-se a função **Análise de Dados**. De forma que surgirá uma segunda janela denominada Análise de Dados. Nela, deve-se selecionar a opção **Regressão** e posteriormente selecionar **OK**.

Feito isso, surgirá uma terceira janela denominada **Regressão**, nela devem ser informadas as variáveis dependente (*Y*) e independente (*X*), respectivamente, nas estradas correspondentes a **Intervalo Y de entrada** e **Intervalo X de entrada**.

Ainda na janela **Regressão**, existem outras funções interessantes a se explorar. Na seção **Resíduos**, podemos selecionar **Plotar revsíduos** e **Plotar ajuste de linha**.

Por fim, na janela **Regressão**, selecionando **OK**, é gerada uma saída de informações sobre as estimativas do modelo de regressão, além dos dois gráficos que também foram solicitados. As etapas até aqui descritas por ser melhor compreendidas com auxílio das imagens a seguir.

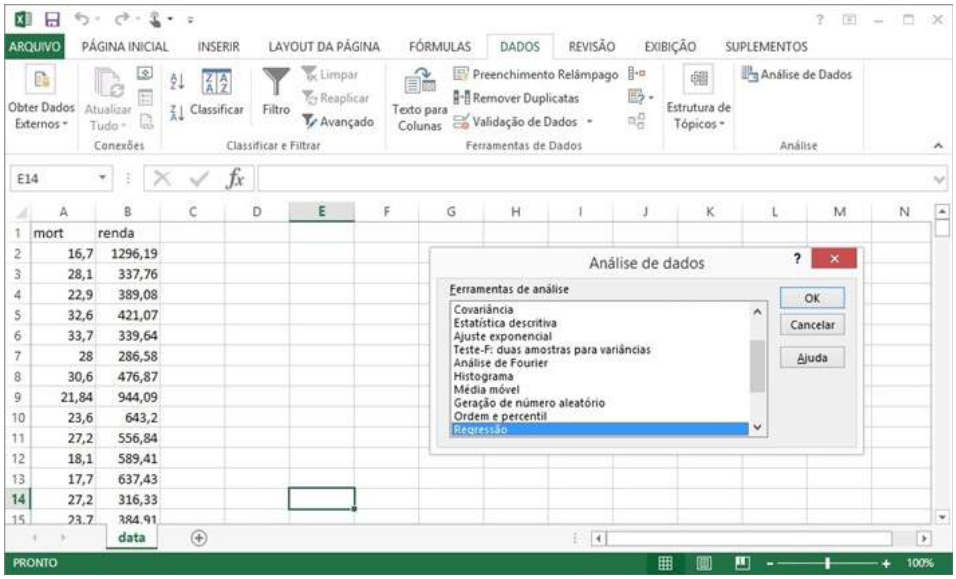


Figura 12: Regressão em planilha eletrônica - 1. / Fonte: Elaboração do autor.

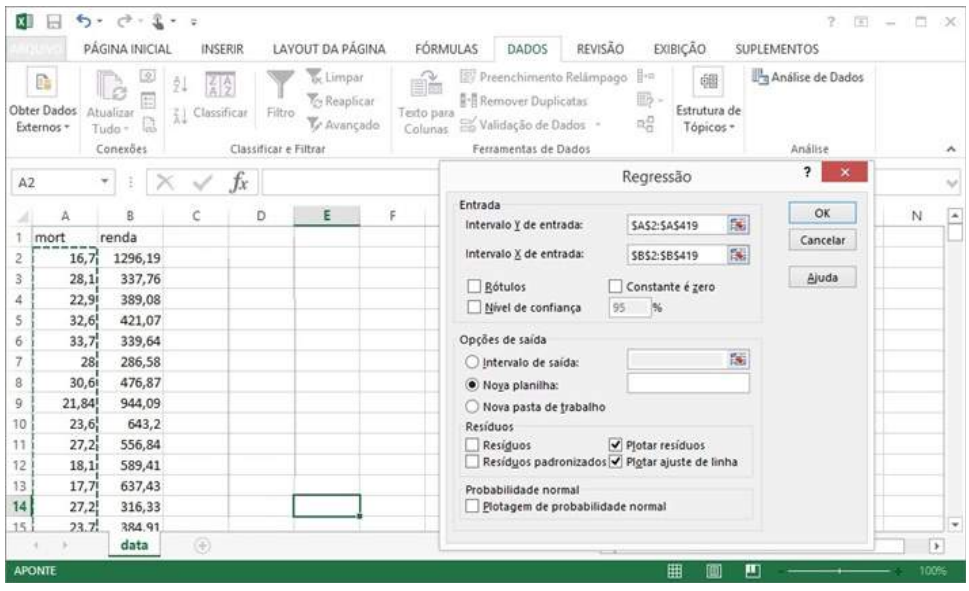


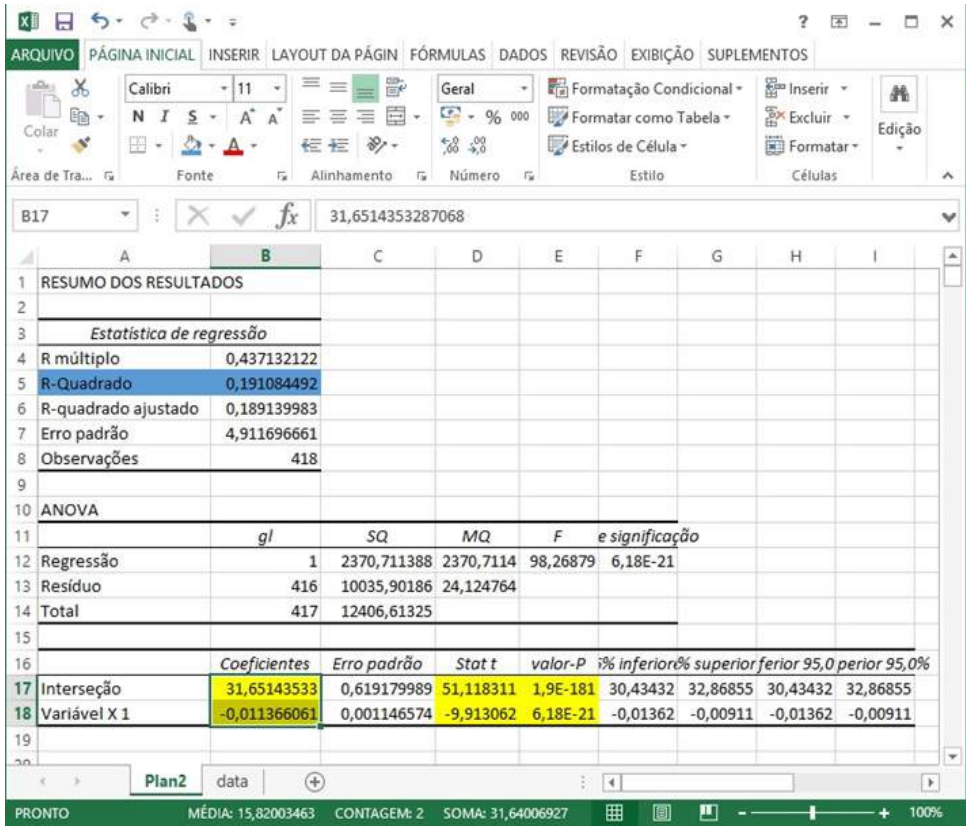
Figura 13: Regressão em planilha eletrônica - 2. / Fonte: Elaboração o autor.

Observando a saída do modelo de regressão, note que são calculadas diversas estatísticas, dentre elas, podemos destacar o R^2, grifado em azul na figura a seguir, que foi estimado em aproximadamente 0,19. Ou seja, tem-se que cerca de 19% das variações em são explicadas pelo modelo.

Além disso, destacado em amarelo, temos os **Coeficientes** estimados, tais que são aproximadamente

$$\beta_0=31,65 \qquad e \qquad \beta_1=-0,01.$$

Note que como era de se esperar, a influência de X sobre Y é negativa (β_1).



	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	*Estatística de regressão*								
4	R múltiplo	0,437132122							
5	R-Quadrado	0,191084492							
6	R-quadrado ajustado	0,189139983							
7	Erro padrão	4,911696661							
8	Observações	418							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*e significação*			
12	Regressão	1	2370,711388	2370,7114	98,26879	6,18E-21			
13	Resíduo	416	10035,90186	24,124764					
14	Total	417	12406,61325						
15									
16		*Coeficientes*	*Erro padrão*	*Stat t*	*valor-P*	*i% inferior*	*% superior*	*ferior 95,0*	*perior 95,0%*
17	Interseção	31,65143533	0,619179989	51,118311	1,9E-181	30,43432	32,86855	30,43432	32,86855
18	Variável X 1	-0,011366061	0,001146574	-9,913062	6,18E-21	-0,01362	-0,00911	-0,01362	-0,00911
19									

Figura 14: Regressão em planilha eletrônica - 3.

Fonte: Elaboração do autor.

Note ainda que o *software* calcula a estatística de teste para avaliar a hipótese de que o coeficiente estimado seja igual a zero. Esta é dada por **Stat t**, também destacada em amarelo. Na prática, não é necessário comparar o valor calculado com os valores críticos da tabela *t* de Student porque o *software* já faz isso. A saída valor-P é o *p*-valor visto no capítulo sobre Teste de hipótese, que é uma métrica para tomada de decisão acerca do teste de hipóteses. De forma que se o *p*-valor for maior que o nível de significância (α), rejeitamos H_0. Como os p-valores calculados são demasiadamente pequenos, rejeitamos

as hipóteses de que os parâmetros estimados são iguais a zero, considerando um nível de significância igual a 5%. Assim, temos que o modelo estimado é dado por

$$Y_i = 31{,}65 - 0{,}01 * X_i .$$

Além das estimativas, temos também acesso ao gráfico dos resíduos. Note que a grande maioria dos pontos aparecem dispersos de forma similar e aleatoriamente em torno de zero, com exceção de algumas poucas cidades com renda média acima de 1000 reais. Isso indica que não existe problemas de heteroscedasticidade.

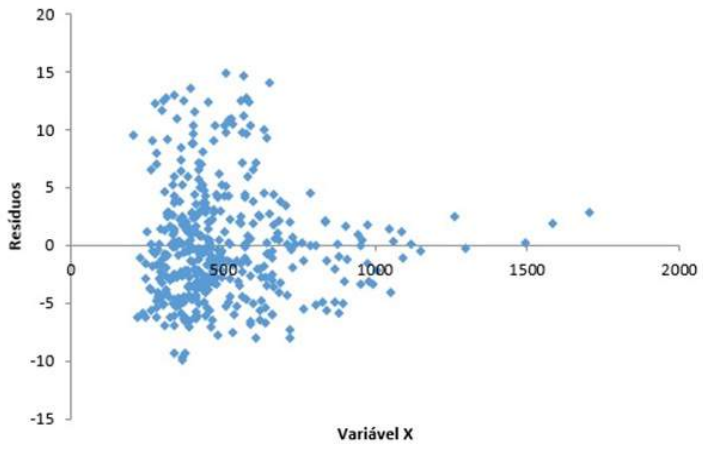


Figura 15: Gráfico de dispersão dos resíduos. / Fonte: Elaboração do autor.

Temos ainda o gráfico que apresenta a reta de regressão ajustada. Em que é notório que a mesma segue a tendência dos dados, que como já era de se esperar, apresenta relação de influência negativa.

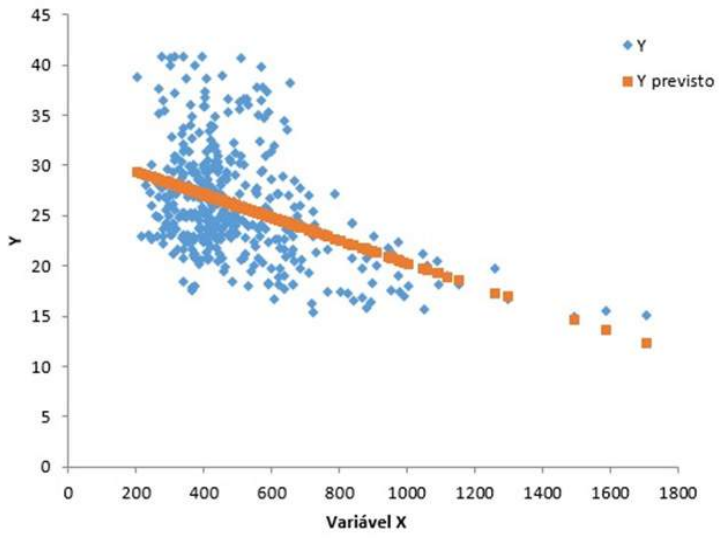
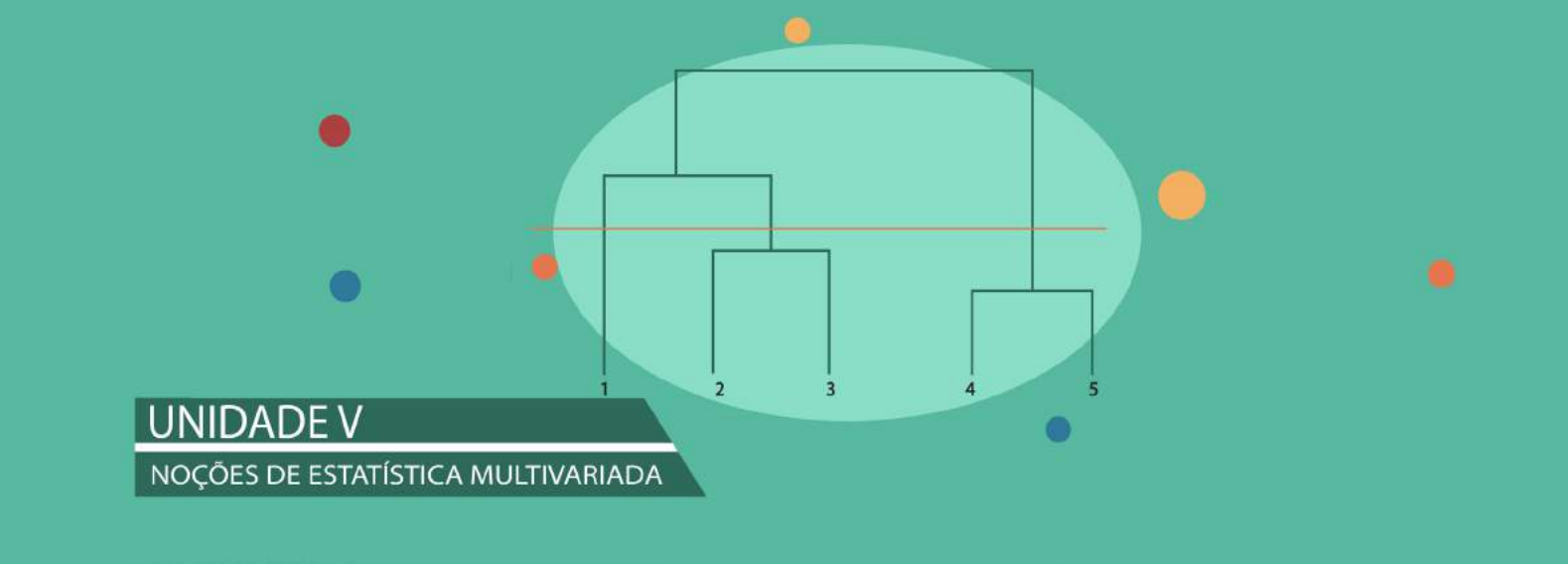


Figura 16: Gráfico da reta de regressão ajustada. / Fonte: Elaboração do autor.



UNIDADE 5 - NOÇÕES DE ESTATÍSTICA MULTIVARIADA

Todas as técnicas que vimos até aqui tratam da utilização de um, ou no máximo dois conjuntos de dados que são utilizados para inferir informações sobre a amostra. Na prática, temos acesso a diversos conjuntos de dados que podem estar relacionados entre si, e é de suma importância que se utilize o máximo de informação possível para uma tomada de decisão precisa e correta.

Antes de tomar qualquer decisão, sempre pensamos e ponderamos com respeito a qual melhor alternativa escolher. Esse processo envolve uma série de fatores que são analisados e ponderados. Por exemplo, se se deseja comprar um relógio entre diversas opções existentes, analisa-se a durabilidade, a precisão, a aparência, entre tantos outros fatores. Normalmente esses fatores tem pesos diferentes na decisão.

Nesse capítulo, veremos uma série de técnicas estatísticas para se extrair informações relevantes de um conjunto de variáveis de interesse. Veremos que é possível extrair diferentes tipos de informações dessas variáveis e que para cada problemática, existe uma técnica apropriada para análise dos dados.

5.1 Conceitos introdutórios

Antes de mais nada é importante fazer a distinção do que é uma **análise univariada** e uma **análise multivariada**.

> UNIVARIADA – diz respeito a análise feita sobre uma única variável. Exemplos dessa abordagem são a estatística descritiva, teste de média, análise variância (ANOVA), entre outras.

Existe ainda uma categoria entre análise univariada e multivariada, denominada, **análise bivariada**.

BIVARIADA – este é o caso intermediário entre a análise univariada e a multivariada. Na análise bivariada trabalhamos com **duas variáveis**. Por exemplo, teste para diferença de média, regressão simples, entre outras.

MULIVARIADA – diz respeito a análise estatística feita sobre **três ou mais variáveis**.

A nomenclatura **análise multivariada** diz respeito a divversas técnicas estatísticas que utilizam simultaneamente mais que duas variáveis para construção de informação relevante. A técnica utilizada depende diretamente da pergunta que se deseja responder com base nos dados disponíveis, de forma que cada técnica de análise multivariada é diferente da outra, cada qual com suas condições particulares de uso.

Por exemplo, podemos ter interesse em identificar o quanto amostras se relacionam segundo alguns critérios, podemos utilizar análise de agrupamento hierárquico ou análise de componentes principais. Podemos ainda ter interesse em como uma série de variáveis influenciam uma única variável, assim podemos utilizar regressão múltipla.

O procedimento de escolha da metodologia adequada para análise dos dados passa pela categorização dos dados disponíveis. Os dados podem ser **qualitativos** ou **quantitativos**.

A variável qualitativa expressa determinada característica do indivíduo, a ausência dela (nesse caso temos uma escala nominal), ou sua intensidade (nesse caso temos uma escala ordinal). É possível converter a informação em um padrão numérico, entretanto, esse valor não tem um significado quantitativo.

Ex. Exemplos

Exemplo 5.1: Quando se pergunta o indivíduo nasceu no Brasil ou não (escala nominal), este só tem duas opções viáveis, sim ou não. É possível converter a resposta para uma escala numérica atribuindo 1 a sim e 0 a não.

Exemplo 5.2: Pode-se ter interesse em saber a satisfação do indivíduo com determinado serviço prestado (escala ordinal). Para isso, elabora-se uma escala de graus de satisfação, por exemplo, não satisfeito, indiferente, satisfeito. É possível aqui converter a resposta para uma escala numérica atribuindo 0 para não satisfeito, 1 para indiferente e 2 para satisfeito.

A variável quantitativa permite uma métrica real da característica que se deseja analisar. As medidas de uma variável quantitativa podem ter escala de intervalo, (como temperatura em graus Celsius, por exemplo) e podem ter escala de razão (por exemplo, peso, altura).

Existem diversas técnicas de análise multivariada, e surgem novas técnicas com frequência, por isso daremos enfoque aqui as técnicas mais usuais, as quais serão classificadas em 2 grupos, as técnicas de modelagem via regressão e a técnicas baseadas em correlação.

5.2 Modelagem via Regressão

A ideia aqui é bastante similar ao que foi visto anteriormente sobre regressão simples. Em regressão simples, assumimos que uma variável pode ser explicada por outra. Aqui passaremos a assumir que uma variável de interesse (dependente) pode ser explicada por duas ou mais outras variáveis (independentes), o que é uma conjuntura bem mais realista.

5.2.1 Regressão Múltipla

Quando temos uma única variável dependente quantitativa e desejamos relacionar as influências de outras variáveis independentes, sejam elas quantitativas ou qualitativas, é recomendável o uso de regressão múltipla. O objetivo central do modelo de regressão múltipla é conseguir prever satisfatoriamente a variável dependente com base nas informações disponível sobre as variáveis dependentes. Para isso são estimados os pesos de cada variável independente na composição da variável dependente.

O modelo de regressão múltipla pode ser ainda utilizado para investigar se determinado grupo de variáveis independentes afetam ou não o valor médio de uma variável dependente.

Ex. Exemplos

Exemplo 5.3: Deseja-se saber os fatores que influenciam o peso do recém-nascido segundo os hábitos da mãe durante a gravidez. Existem uma série de fatores que podem influenciar, como por exemplo, se a mãe fumou ou não durante a gravidez, o ganho de peso da mãe durante a gravidez, se a mãe tomou remédios controlados ou não durante a gravidez, entre outros.

5.2.2 Análise Discriminante

Esta técnica é utilizada para discriminar e classificar indivíduos. A ideia consiste em classificar os indivíduos em duas ou mais populações, de forma a minimizar a probabilidade de que um indivíduo seja erroneamente classificado em determinada população. Assim, podemos dizer que o objetivo da análise descriminante é identificar as variáveis que discriminam os grupos, para assim poder elaborar previsões sobre uma possível nova observação, identificando qual grupo melhor lhe representa.

Para isso é levado em consideração as características mais presentes em indivíduos de cada população. Tecnicamente isso é feito por meio de uma função discriminante. Essa função estabelece um critério numérico (escores de corte) baseado nas variáveis disponíveis para alocar cada indivíduo em uma população de indivíduos homogêneos.

Esta técnica é indicada quando temos interesse em definir quais variáveis explicativas quantitativas tem maior poder de discriminar determinada variável qualitativa (categórica) em diferentes populações.

Ex. Exemplos

Exemplo 5.4: Deseja-se estudar o desempenho de alunos do ensino médio. Até o momento para cada amostra de aluno foi atribuído um índice correspondente as suas notas (nota/10), de forma que a amostra foi classificada (categorizada) a partir desse índice, logo, se o aluno tem nota 7, seu índice correspondente é 0,7. A classificação foi feita da seguinte forma.

i. Se o aluno tem índice igual ou superior a 0,7 ele é considerado bom;

ii. Se o índice está entre 0,69 e 0,5 ele é considerado regular; e por fim,

iii. Se ele tem índice abaixo de 0,49 é considerado ruim.

Levando em consideração variáveis como horas de estudo, horas de lazer e distância entre escola e casa, por exemplo, deseja-se elencar quais variáveis são mais importantes para discriminar os alunos entre esses 3 grupos. Neste caso, deve-se utilizar análise discriminante.

5.2.3 Modelos Lineares de probabilidade

Essa técnica é popularmente conhecida como **modelo de regressão logit**, e pode ser considerada uma mistura de regressão múltipla e análise discriminante. Ela se assemelha a análise de regressão múltipla porque é também de interesse se estimar pesos para as variáveis independentes a fim de melhor prever o comportamento da variável independente. Entretanto, diferentemente da regressão múltipla, a variável dependente é qualitativa e não quantitativa, e nesse aspecto, o modelo é semelhando a análise discriminante.

Neste caso, considerando que a variável dependente seja qualitativa em escala nominal, pode-se utilizar o modelo de regressão logit (também conhecido como regressão logística). Em caso de a variável dependente seja qualitativa em escala ordinal, pode ser utilizado o modelo de **regressão logit ordenado**.

Uma característica que merece destaque nesse modelo é a sua interpretação, pela qual é possível determinar a probabilidade de que determinado evento de interesse ocorra, de acordo com as características observadas desse evento.

Exemplo 5.5: Suponha que se deseja saber a probabilidade de que uma pessoa venha a ter câncer de pulmão. Para isso, tem-se uma amostra de pessoas que desenvolveram a doença (a estes se atribui o valor 1) e pessoas que não desenvolveram a doença (a estes se atribui o valor 0). Com base em características comportamentais (como hábito de fumar, praticar exercícios físicos, entre outros) e física (peso, altura, características genéticas), se pode estabelecer quais características são influentes para que o indivíduo tenha uma maior probabilidade de desenvolver a doença ou não.

5.3 Técnicas Baseadas em Correlação

As técnicas apresentadas até aqui têm por característica principal se utilizarem de variáveis dependentes e independentes. Entretanto, nem sempre é possível previamente definir que determinada variável é dependente e determinada variável não o é. Por vezes, nosso interesse reside simplesmente em investigar a interdependência de todas as variáveis conjuntamente, nesse caso existem técnicas multivariadas apropriadas. Note que isso

implica uma análise de caráter mais explanatório dos dados. Aqui serão apresentadas as principais técnicas utilizadas para esta finalidade.

5.3.1 Análise de fatores

Por vezes, nos deparamos com um grande banco de dados e não sabemos exatamente como trabalhar tanta informação simultaneamente. Surge a necessidade de traçar perfis aos dados. Observar quais variáveis tem maior relação umas com as outras. É possível que se possa condensar a informação contida em muitas variáveis em uma única variável não observada diretamente, esta variável pode ser chamada de **fator**. Quando nos deparamos com interesses nessa linha, é recomendável o uso de **análise de fatores**.

Por vezes se utiliza análise fatorial com o intuito de

i. Reduzir grandes conjuntos de variáveis para conjuntos menores e mais significativos, e

ii. Identificar grupos de variáveis inter-relacionadas e distingui-los de outros grupos de variáveis inter-relacionadas.

Essa técnica é baseada na matriz de correlações das variáveis. A ideia por trás do método é que é possível condensar a informação contida em um determinado número de variáveis com uma perda mínima de informação. Ou seja, com base na matriz de correlação das variáveis, busca-se determinar grupos, tais que, as variáveis dentro de cada grupo possuam alta correlação entre si, enquanto possuírem correlação relativamente pequena com variáveis de outro grupo.

Uma outra forma de pensar é se duas ou mais variáveis estão altamente correlacionadas, quer dizer que, se elas estão "dizendo a mesma coisa", assim será possível "passar essa mensagem" utilizando uma variável apenas.

Ex. Exemplos

Exemplo 5.6: Suponha um estudo de mercado sobre automóveis. Neste estudo, levantaram-se diversas variáveis levadas em consideração pelo consumidor ao escolher um carro, por exemplo.

i. Espaço interno;

ii. Custo de manutenção;

iii. Design;

iv. Baixo consumo de combustível;

v. Preço de revenda;

vi. Fácil manuseio de instrumentos;

vii. Robustez do motor;

viii. Robustez do cambio;

ix. Porta malas;

x. Conforto interno;

xi. Variedade de cores;

xii. Preço de seguro;

xiii. Distância entre eixos;

xiv. Altura em relação ao solo.

Note que elencamos ao todo 14 variáveis. Para o departamento de marketing, seria problemático ter que desenvolver uma campanha publicitária levando em consideração cada uma das 14 variáveis analisadas. Em vez disso, seria ideal se fosse possível saber como o consumidor pensa em termos de variáveis mais geral, ou **fatores**.

Utilizando análise fatorial, pode-se delimitar três categorias mais gerais como

I. Custo-benefício;

II. Conforto; e

III. Segurança.

De forma que podemos obter características referentes aos fatores a partir do conhecimento das variáveis disponíveis.

5.3.2 Análise de *clusters*

Esta técnica tem por objetivo simplesmente alocar os indivíduos em grupos de similaridade com base em suas características observadas. Note que esta técnica propõe algo similar ao que foi visto de análise discriminante. Entretanto, aqui não precisamos definir uma variável dependente, analisamos todas as características de cada indivíduo simultaneamente, e a partir disto, definimos a qual grupo o indivíduo pertence.

Também conhecida como análise de agrupamento, a análise de *clusters* dividi os elementos da amostra em grupos de similaridade. Para isso, é definida uma **medida de similaridade** para os indivíduos. Duas medidas bastante comuns são as **medidas de correlação** e a **distância euclidiana** entre os indivíduos. A medida de similaridade quantifica a proximidade entre cada indivíduo de acordo com suas características observáveis.

Estabelecidas as similaridades entre os indivíduos, o passo seguinte é a construção dos *clusters*. Essa construção pode ser feita de duas formas, basicamente, utilizando

i. Métodos hierárquicos – é utilizado quando não se sabe previamente a quantidade de grupos que se deseja classificar as observações. Essa quantidade é fornecida pelo método;

ii. Métodos não hierárquicos – é utilizado quando o pesquisador já tem predeterminada a quantidade de grupos de interesse.

Existe um gráfico bastante interessante na análise de *clusters*, o chamado **Dendrograma**. Ele sumariza as informações relacionadas as distâncias calculadas e a formação dos *clusters*. Quanto maior a exigência do nível de similaridade entre os indivíduos pertencentes a cada grupo, maior será a quantidade de grupos.

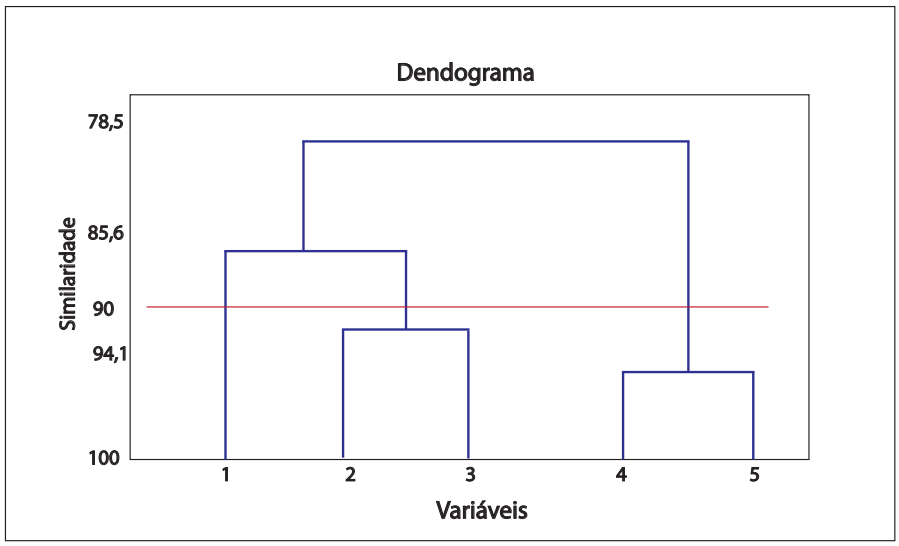


Figura 17: Dendrograma.

Ilustração: Ariana Santana.

Com base no dendrograma acima, tem-se que considerando um nível de similaridade entre os indivíduos dentro do mesmo grupo de 90, são deduzidos 3 *clusters*, são eles, o *cluster* 1, o *cluster* [2,3] e o *cluster* [4,5]. Entretanto, considerando o máximo de similaridade possível (100), obtemos 5 *clusters*.

É claro que quanto maior for a similaridade, melhor a qualidade de classificação dos indivíduos, entretanto, normalmente, para se obter altos nível de similaridade é necessário um número demasiadamente grande de *clusters*, tornando a análise de dados ineficiente. Por isso, é interessante que se escolha um número de interesse de clusters considerando o maior nível de similaridade possível.

Ex. Exemplos

Exemplo 5.7: Suponha que temos dados de 100 indivíduos, e desejamos agrupá-los de acordo com suas similaridades. Para cada um dos indivíduos temos as informações sobre Renda, Idade e Peso. Se nosso desejo é classificar esses 100 indivíduos em grupos similares (mesmo não sabendo delimitar a princípio a quantidade de grupos), de acordo com suas características observadas, é indicada a análise de ***clusters*****.**

ANEXO A

Tabela A: Valores críticos da distribuição *t* de Student.

G.L.	α unilateral	0,005	0,025	0,050	G.L.	α unilateral	0,005	0,025	0,050
	bilateral	0,010	0,050	0,100		bilateral	0,010	0,050	0,100
1		63,657	12,706	6,314	41		2,701	2,020	1,683
2		9,925	4,303	2,920	42		2,698	2,018	1,682
3		5,841	3,182	2,353	43		2,695	2,017	1,681
4		4,604	2,776	2,132	44		2,692	2,015	1,680
5		4,032	2,571	2,015	45		2,690	2,014	1,679
6		3,707	2,447	1,943	46		2,687	2,013	1,679
7		3,499	2,365	1,895	47		2,685	2,012	1,678
8		3,355	2,306	1,860	48		2,682	2,011	1,677
9		3,250	2,262	1,833	49		2,680	2,010	1,677
10		3,169	2,228	1,812	50		2,678	2,009	1,676
11		3,106	2,201	1,796	51		2,676	2,008	1,675
12		3,055	2,179	1,782	52		2,674	2,007	1,675
13		3,012	2,160	1,771	53		2,672	2,006	1,674
14		2,977	2,145	1,761	54		2,670	2,005	1,674
15		2,947	2,131	1,753	55		2,668	2,004	1,673
16		2,921	2,120	1,746	56		2,667	2,003	1,673
17		2,898	2,110	1,740	57		2,665	2,002	1,672
18		2,878	2,101	1,734	58		2,663	2,002	1,672
19		2,861	2,093	1,729	59		2,662	2,001	1,671
20		2,845	2,086	1,725	60		2,660	2,000	1,671
21		2,831	2,080	1,721	61		2,659	2,000	1,670
22		2,819	2,074	1,717	62		2,657	1,999	1,670
23		2,807	2,069	1,714	63		2,656	1,998	1,669
24		2,797	2,064	1,711	64		2,655	1,998	1,669
25		2,787	2,060	1,708	65		2,654	1,997	1,669
26		2,779	2,056	1,706	66		2,652	1,997	1,668
27		2,771	2,052	1,703	67		2,651	1,996	1,668
28		2,763	2,048	1,701	68		2,650	1,995	1,668
29		2,756	2,045	1,699	69		2,649	1,995	1,667
30		2,750	2,042	1,697	70		2,648	1,994	1,667
31		2,744	2,040	1,696	71		2,647	1,994	1,667
32		2,738	2,037	1,694	72		2,646	1,993	1,666
33		2,733	2,035	1,692	73		2,645	1,993	1,666
34		2,728	2,032	1,691	74		2,644	1,993	1,666
35		2,724	2,030	1,690	75		2,643	1,992	1,665
36		2,719	2,028	1,688	76		2,642	1,992	1,665
37		2,715	2,026	1,687	77		2,641	1,991	1,665
38		2,712	2,024	1,686	78		2,640	1,991	1,665
39		2,708	2,023	1,685	79		2,640	1,990	1,664
40		2,704	2,021	1,684	80		2,639	1,990	1,664

Fonte: Elaboração do autor.

ANEXO B

Tabela B: Valores críticos do teste de Tukey.

GL (n-k)	α	k níveis								
		2	3	4	5	6	7	8	9	10
5	0,05	3,64	4,6	5,22	5,67	6,03	6,33	6,58	6,8	6,99
	0,01	5,7	6,98	7,8	8,42	8,91	9,32	9,67	9,97	10,24
6	0,05	3,46	4,34	4,9	5,3	5,63	5,9	6,12	6,32	6,49
	0,01	5,24	6,33	7,03	7,56	7,97	8,32	8,61	8,87	9,1
7	0,05	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6	6,16
	0,01	4,95	5,92	6,54	7,01	7,37	7,68	7,94	8,17	8,37
8	0,05	3,26	4,04	4,53	4,89	5,17	5,4	5,6	5,77	5,92
	0,01	4,75	5,64	6,2	6,62	6,96	7,24	7,47	7,68	7,86
9	0,05	3,2	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74
	0,01	4,6	5,43	5,96	6,35	6,66	6,91	7,13	7,33	7,49
10	0,05	3,15	3,88	4,33	4,65	4,91	5,12	5,3	5,46	5,6
	0,01	4,48	5,27	5,77	6,14	6,43	6,67	6,87	7,05	7,21
11	0,05	3,11	3,82	4,26	4,57	4,82	5,03	5,2	5,35	5,49
	0,01	4,39	5,15	5,62	5,97	6,25	6,48	6,67	6,84	6,99
12	0,05	3,08	3,77	4,2	4,51	4,75	4,95	5,12	5,27	5,39
	0,01	4,32	5,05	5,5	5,84	6,1	6,32	6,51	6,67	6,81
13	0,05	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32
	0,01	4,26	4,96	5,4	5,73	5,98	6,19	6,37	6,53	6,67
14	0,05	3,03	3,7	4,11	4,41	4,64	4,83	4,99	5,13	5,25
	0,01	4,21	4,89	5,32	5,63	5,88	6,08	6,26	6,41	6,54
15	0,05	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,2
	0,01	4,17	4,84	5,25	5,56	5,8	5,99	6,16	6,31	6,44
16	0,05	3	3,65	4,05	4,33	4,56	4,74	4,9	5,03	5,15
	0,01	4,13	4,79	5,19	5,49	5,72	5,92	6,08	6,22	6,35
17	0,05	2,98	3,63	4,02	4,3	4,52	4,7	4,86	4,99	5,11
	0,01	4,1	4,74	5,14	5,43	5,66	5,85	6,01	6,15	6,27
18	0,05	2,97	3,61	4	4,28	4,49	4,67	4,82	4,96	5,07
	0,01	4,07	4,7	5,09	5,38	5,6	5,79	5,94	6,08	6,2
19	0,05	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04
	0,01	4,05	4,67	5,05	5,33	5,55	5,73	5,89	6,02	6,14
20	0,05	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,9	5,01
	0,01	4,02	4,64	5,02	5,29	5,51	5,69	5,84	5,97	6,09
24	0,05	2,92	3,53	3,9	4,17	4,37	4,54	4,68	4,81	4,92
	0,01	3,96	4,55	4,91	5,17	5,37	5,54	5,69	5,81	5,92
30	0,05	2,89	3,49	3,85	4,1	4,3	4,46	4,6	4,72	4,82
	0,01	3,89	4,45	4,8	5,05	5,24	5,4	5,54	5,65	5,76
40	0,05	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73
	0,01	3,82	4,37	4,7	4,93	5,11	5,26	5,39	5,5	5,6
60	0,05	2,83	3,4	3,74	3,98	4,16	4,31	4,44	4,55	4,65
	0,01	3,76	4,28	4,59	4,82	4,99	5,13	5,25	5,36	5,45
120	0,05	2,8	3,36	3,68	3,92	4,1	4,24	4,36	4,47	4,56
	0,01	3,7	4,2	4,5	4,71	4,87	5,01	5,12	5,21	5,3

Fonte: Elaboração do autor.

ANEXO C

Tabela C1: Valores críticos da distribuição *F* (5% de significância).

n-1 \ k-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	242,98	243,91	244,69	245,36	245,95	246,46	246,92	247,32	247,69	248,01
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,42	19,42	19,43	19,43	19,44	19,44	19,44	19,45
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,73	8,71	8,70	8,69	8,68	8,67	8,67	8,66
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,89	5,87	5,86	5,84	5,83	5,82	5,81	5,80
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,60	4,59	4,58	4,57	4,56
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,92	3,91	3,90	3,88	3,87
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51	3,49	3,48	3,47	3,46	3,44
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22	3,20	3,19	3,17	3,16	3,15
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01	2,99	2,97	2,96	2,95	2,94
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85	2,83	2,81	2,80	2,79	2,77
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72	2,70	2,69	2,67	2,66	2,65
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62	2,60	2,58	2,57	2,56	2,54
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53	2,51	2,50	2,48	2,47	2,46
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46	2,44	2,43	2,41	2,40	2,39
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40	2,38	2,37	2,35	2,34	2,33
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35	2,33	2,32	2,30	2,29	2,28
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31	2,29	2,27	2,26	2,24	2,23
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27	2,25	2,23	2,22	2,20	2,19
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,21	2,20	2,18	2,17	2,16
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20	2,18	2,17	2,15	2,14	2,12

Fonte: Elaboração do autor.

Observação: A tabela C1 foi rotacionada 90º em sentido anti-horário para melhor visualização.

Tabela C2: Valores críticos da distribuição *F* (1% de significância).

n-1 \ k-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	4052,2	4999,5	5403,4	5624,6	5763,7	5859,0	5928,4	5981,1	6022,5	6055,9	6083,3	6106,3	6125,9	6142,7	6157,3	6170,1	6181,4	6191,5	6200,6	6208,7
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,41	99,42	99,42	99,43	99,43	99,44	99,44	99,44	99,45	99,45
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,13	27,05	26,98	26,92	26,87	26,83	26,79	26,75	26,72	26,69
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,31	14,25	14,20	14,15	14,11	14,08	14,05	14,02
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,82	9,77	9,72	9,68	9,64	9,61	9,58	9,55
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,52	7,48	7,45	7,42	7,40
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,36	6,31	6,28	6,24	6,21	6,18	6,16
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,61	5,56	5,52	5,48	5,44	5,41	5,38	5,36
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,01	4,96	4,92	4,89	4,86	4,83	4,81
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,65	4,60	4,56	4,52	4,49	4,46	4,43	4,41
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,21	4,18	4,15	4,12	4,10
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,97	3,94	3,91	3,88	3,86
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,91	3,86	3,82	3,78	3,75	3,72	3,69	3,66
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,62	3,59	3,56	3,53	3,51
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,49	3,45	3,42	3,40	3,37
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,50	3,45	3,41	3,37	3,34	3,31	3,28	3,26
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,27	3,24	3,21	3,19	3,16
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,32	3,27	3,23	3,19	3,16	3,13	3,10	3,08
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,12	3,08	3,05	3,03	3,00
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,18	3,13	3,09	3,05	3,02	2,99	2,96	2,94

Fonte: Elaboração do autor.

Observação: A tabela C2 foi rotacionada 90º em sentido anti-horário para melhor visualização.

Tabela C3: Valores críticos da distribuição *F* (10% de significância).

		k-1																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n-1	1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,47	60,71	60,90	61,07	61,22	61,35	61,46	61,57	61,66	61,74
	2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41	9,41	9,42	9,42	9,43	9,43	9,44	9,44	9,44
	3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22	5,21	5,20	5,20	5,20	5,19	5,19	5,19	5,18
	4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91	3,90	3,89	3,88	3,87	3,86	3,86	3,85	3,85	3,84
	5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28	3,27	3,26	3,25	3,24	3,23	3,22	3,22	3,21	3,21
	6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92	2,90	2,89	2,88	2,87	2,86	2,85	2,85	2,84	2,84
	7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68	2,67	2,65	2,64	2,63	2,62	2,61	2,61	2,60	2,59
	8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52	2,50	2,49	2,48	2,46	2,45	2,45	2,44	2,43	2,42
	9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40	2,38	2,36	2,35	2,34	2,33	2,32	2,31	2,30	2,30
	10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30	2,28	2,27	2,26	2,24	2,23	2,22	2,22	2,21	2,20
	11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23	2,21	2,19	2,18	2,17	2,16	2,15	2,14	2,13	2,12
	12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17	2,15	2,13	2,12	2,10	2,09	2,08	2,08	2,07	2,06
	13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,12	2,10	2,08	2,07	2,05	2,04	2,03	2,02	2,01	2,01
	14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,07	2,05	2,04	2,02	2,01	2,00	1,99	1,98	1,97	1,96
	15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,04	2,02	2,00	1,99	1,97	1,96	1,95	1,94	1,93	1,92
	16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	2,01	1,99	1,97	1,95	1,94	1,93	1,92	1,91	1,90	1,89
	17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,98	1,96	1,94	1,93	1,91	1,90	1,89	1,88	1,87	1,86
	18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,95	1,93	1,92	1,90	1,89	1,87	1,86	1,85	1,84	1,84
	19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,93	1,91	1,89	1,88	1,86	1,85	1,84	1,83	1,82	1,81
	20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,91	1,89	1,87	1,86	1,84	1,83	1,82	1,81	1,80	1,79

Fonte: Elaboração do autor.

Observação: A tabela C3 foi rotacionada 90º em sentido anti-horário para melhor visualização.

ANEXO D

Tabela D: Valores críticos da distribuição qui-quadrado com k-1 graus de liberdade.

	α		
k-1	0,01	0,05	0,1
1	6,635	3,841	2,706
2	9,210	5,991	4,605
3	11,345	7,815	6,251
4	13,277	9,488	7,779
5	15,086	11,070	9,236
6	16,812	12,592	10,645
7	18,475	14,067	12,017
8	20,090	15,507	13,362
9	21,666	16,919	14,684
10	23,209	18,307	15,987
11	24,725	19,675	17,275
12	26,217	21,026	18,549
13	27,688	22,362	19,812
14	29,141	23,685	21,064
15	30,578	24,996	22,307
16	32,000	26,296	23,542
17	33,409	27,587	24,769
18	34,805	28,869	25,989
19	36,191	30,144	27,204
20	37,566	31,410	28,412
21	38,932	32,671	29,615
22	40,289	33,924	30,813
23	41,638	35,172	32,007
24	42,980	36,415	33,196
25	44,314	37,652	34,382

Fonte: Elaboração do autor.

ANEXO E

Tabela E1: Valores críticos para teste bilateral de Mann Whitney (5% de significância).

N_2	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
N_1																
2	*	*	*	0	0	0	0	1	1	1	1	1	2	2	2	2
3	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	0	1	2	3	4	4	5	6	7	9	10	11	11	12	13	14
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	*	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	*	*	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	*	*	*	13	15	17	19	22	24	26	29	31	34	36	38	41
9	*	*	*	*	17	20	23	26	28	31	34	37	39	42	45	48
10	*	*	*	*	*	23	26	29	33	36	39	42	45	48	53	55
11	*	*	*	*	*	*	30	33	37	40	44	47	51	55	58	62
12	*	*	*	*	*	*	*	37	41	45	49	53	57	61	65	69
13	*	*	*	*	*	*	*	*	45	50	54	59	63	67	72	76
14	*	*	*	*	*	*	*	*	*	55	59	64	69	74	78	83
15	*	*	*	*	*	*	*	*	*	*	64	70	75	80	85	90
16	*	*	*	*	*	*	*	*	*	*	*	75	81	86	92	98
17	*	*	*	*	*	*	*	*	*	*	*	*	87	93	99	105
18	*	*	*	*	*	*	*	*	*	*	*	*	*	99	106	112
19	*	*	*	*	*	*	*	*	*	*	*	*	*	*	113	119
20	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	127

Fonte: Elaboração do autor.

Tabela E2: Valores críticos para teste bilateral de Mann Whitney (1% de significância).

N_2	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
N_1																
2	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0	0
3	*	*	*	*	0	0	0	1	1	1	2	2	2	2	3	3
4	*	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	*	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	*	*	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	*	*	*	7	9	11	13	15	17	18	20	22	24	26	28	30
9	*	*	*	*	11	13	16	18	20	22	24	27	29	31	33	36
10	*	*	*	*	*	16	18	21	24	26	29	31	34	37	39	42
11	*	*	*	*	*	*	21	24	27	30	33	36	39	42	45	48
12	*	*	*	*	*	*	*	27	31	34	37	41	44	47	51	54
13	*	*	*	*	*	*	*	*	34	38	42	45	49	53	57	60
14	*	*	*	*	*	*	*	*	*	42	46	50	54	58	63	67
15	*	*	*	*	*	*	*	*	*	*	51	55	60	64	69	73
16	*	*	*	*	*	*	*	*	*	*	*	60	65	70	74	79
17	*	*	*	*	*	*	*	*	*	*	*	*	70	75	81	86
18	*	*	*	*	*	*	*	*	*	*	*	*	*	81	87	92
19	*	*	*	*	*	*	*	*	*	*	*	*	*	*	93	99
20	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	105

Fonte: Elaboração do autor.

REFERENCIAS

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A.; PAIVA, L. S. C. (Trad.). Estatística aplicada à administração e economia. 2.ed São Paulo: Cengage Learning, 2009.

FREUND, J. E.; SIMON, G. A. Estatística aplicada: economia, administração e contabilidade. 11.ed. Porto Alegre: Bookman, 2006.

GALTON, F. Natural Inheritance. London: Macmillan, 1889.

GUJARATI, D. N. Econometria básica. 4.ed. Rio de Janeiro: Editora Campus, 2006.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 4.ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999.

LEVIN, Jack. Estatística Aplicada a Ciências Humanas. 2.ed. São Paulo: Editora Harbra Ltda, 1987.

MASON, R. D., DOUGLAS, L. A. *Statistical Techniques in Business And Economics*. IRWIN,Boston, 1990.

MORETTIN, P. A.; BUSSAB, W. O. Estatística básica. 6.ed. São Paulo: Saraiva, 2010.

SPIEGEL, M. R. Estatística. 4.ed. Porto Alegre: Bookman, 2009.

STEVENSON, W. J. Estatística aplicada a administração. São Paulo: Harper & Row do Brasil, 1981.

WONNACOTT, T. H.; WONNACOTT, R. J. Estatística aplicada a economia e a administração. Rio de Janeiro: Livros Técnicos e Científicos, 1981.

TRIOLA, M. F. Introdução a estatística. 7.ed. Rio de Janeiro: Livros Técnicos e Científicos, 1999.

ANOTAÇÕES

ANOTAÇÕES

ANOTAÇÕES

Universidade Federal da Bahia

Estatística Aplicada às Ciências Sociais Aplicadas II

Vivemos a era da informação, nunca se coletaram e analisaram tantos dados como atualmente. E isso segue como uma tendência crescente, espera-se que com o passar do tempo e com os avanços tecnológicos, cada vez seja possível coletar e analisar mais e mais informação. Nesse contexto, é de fundamental importância para qualquer profissional ser capaz de utilizar toda essa informação a seu favor. Para que isso seja feito de forma eficiente, é imprescindível o conhecimento de estatística, mais precisamente, métodos estatísticos capazes guiar a tomada de decisão.

Pensando nisso, esse módulo foi confeccionado a fim de possibilitar um acesso suave a algumas das principais técnicas para extração de informação relevante a partir de dados.

O objetivo principal aqui é apresentar uma série de ferramentas estatísticas que podem ser utilizadas em diferentes contextos e que são capazes de fornecer respostas sobre questões complexas de se avaliar.

PROGRAD
PRÓ-REITORIA DE GRADUAÇÃO

Ciências Contábeis
UNIVERSIDADE FEDERAL DA BAHIA

SEAD
Superintendência de Educação a Distância | UFBA

UNIVERSIDADE ABERTA DO BRASIL

NELT NÚCLEO DE ESTUDOS DE Linguagens & Tecnologias