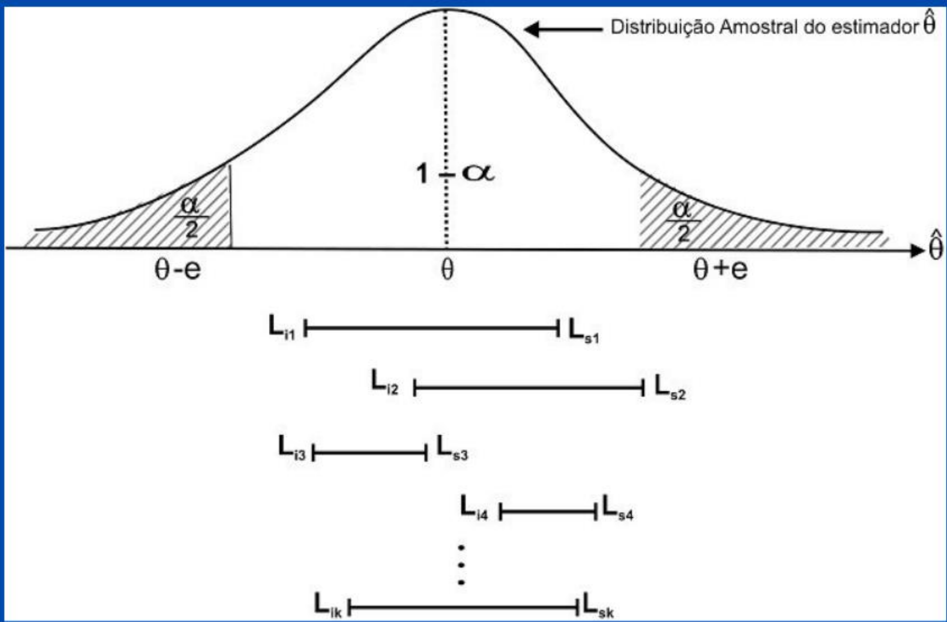
ESTIMAÇÃO ESTATÍSTICA

JANILSON PINHEIRO DE ASSIS
ROBERTO PEQUENO DE SOUSA
PAULO CÉSAR FERREIRA LINHARES
BEN DÊIVIDE DE OLIVEIRA BATISTA
EUDES DE ALMEIDA CARDOSO



Pantanal Editora

2021

JANILSON PINHEIRO DE ASSIS
ROBERTO PEQUENO DE SOUSA
PAULO CÉSAR FERREIRA LINHARES
BEN DÊIVIDE DE OLIVEIRA BATISTA
EUDES DE ALMEIDA CARDOSO

ESTIMAÇÃO ESTATÍSTICA

Pantanal Editora

2021

Copyright© Pantanal Editora
Editor Chefe: Prof. Dr. Alan Mario Zuffo
Editores Executivos: Prof. Dr. Jorge González Aguilera e Prof. Dr. Bruno Rodrigues de Oliveira

Diagramação: A editora. **Diagramação e Arte:** A editora. Imagens de capa e contra-capa: Canva.com. **Revisão:** O(s) autor(es), organizador(es) e a editora.

Ficha Catalográfica

Dados Internacionais de Catalogação na Publicação (CIP)
(eDOC BRASIL, Belo Horizonte/MG)

E81 Estimação estatística [livro eletrônico] / Janilson Pinheiro de Assis... [et al.]. – Nova Xavantina, MT: Pantanal Editora, 2021. 130p.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
ISBN 978-65-88319-63-5
DOI https://doi.org/10.46420/9786588319635

1. Estatística. 2. Probabilidade. 3. Matemática. I. Assis, Janilson Pinheiro de. II. Sousa, Roberto Pequeno de. III. Linhares, Paulo César Ferreira. IV. Batista, Ben Dêivide de Oliveira. V. Cardoso, Eudes de Almeida.
CDD 519.5

Elaborado por Maurício Amormino Júnior – CRB6/2422

Pantanal Editora

Rua Abaete, 83, Sala B, Centro. CEP: 78690-000.
Nova Xavantina – Mato Grosso – Brasil.
Telefone (66) 99682-4165 (Whatsapp).
https://www.editorapantanal.com.br
contato@editorapantanal.com.br

APRESENTAÇÃO

Este livro apresenta uma importante ferramenta da inferência estatística, utilizada pelos pesquisadores em geral, e tendo como público alvo estudantes de curso de graduação que desenvolvem atividades principalmente na área biológica, mas sem deixar de ser uma importante obra de apoio na área das ciências físicas, engenharias e sociais. No entanto nada impede que outros pesquisadores ou técnicos de órgãos públicos e privados, de agências governamentais, bancos, escolas, etc. que utilizam com menor ou maior frequência a estimação intervalar nas suas atividades do dia a dia, não possa se beneficiar desta obra.

Esta obra descreve de forma minuciosamente os conteúdos de forma clara e objetiva de todos os conceitos, definições e propriedades dos estimadores, os procedimentos para a obtenção de estimadores amostrais justo, não viesado, não viciado ou não tendencioso, consistente ou coerente, eficiente e suficiente, além de técnicas para o dimensionamento amostral. Também aborda a teoria para proporcionar o manuseio dos fatores que influenciam a amplitude e confiança dos intervalos fiduciais e para a construção e interpretação de intervalos de confiança paramétricos para grandezas populacionais referentes a média, diferença entre médias, variância, razão entre variâncias, desvio padrão, proporção e diferença entre proporções, com diversos exercícios de aplicação práticos e formulários, bem como o uso de tabelas para facilitar o aprendizado e a compreensão deste assunto. O leitor terá facilidade na assimilação do conteúdo, pois os capítulos seguem uma ordem e sequência lógica e bem didática.

No final do livro são expostas diversas situações práticas de pesquisa com dados experimentais, tabelas auxiliares, o alfabeto grego com todas as letras e pronuncia em português, formulários para dimensionamento amostral bem como dos principais tipos de intervalos de confiança numa tabela resumo, além de uma lista de referências bibliográficas para o leitor interessado em se aprofundar mais no estudo da estatística indutiva ou analítica frequentista.

Os autores,

Mossoró, RN, Brasil, Agosto de 2021

PREFÁCIO

Este livro surgiu das notas de aulas e apostilas utilizadas nas disciplinas de estatística e estatística experimental ministradas nos últimos trinta anos na Escola Superior de Agricultura de Mossoró (ESAM), bem como na Universidade Federal Rural do Semiárido (UFERSA) localizada na cidade de Mossoró, Estado do Rio Grande do Norte, Brasil, além de ser fruto da larga experiência adquirida em diversos atividades de docência, ministrando cursos, participação em cursos e em eventos técnicos científicos, orientação de bolsistas e na elaboração e publicação de livros, apostilas, notas de aulas bem como de artigos científicos em diversos periódicos nacionais e internacionais por parte dos autores.

O estudo da inferência estatística surgiu da necessidade e do desafio do pesquisador lidar com populações infinitas tanto estimando parâmetros como na tomada de decisão baseada em hipóteses sobre características destes universos e comparando fenômenos semelhantes que é o objetivo principal da ciência estatística frequentista, e assim capacitar o pesquisador a poder realizar generalizações acerca de características de interesse da pesquisa dessas populações infinitas. Este fato permite ao investigador conhecer a natureza da distribuição de probabilidade desses universos permitindo ao mesmo obter conclusões válidas e racionais sobre a população partindo-se dos resultados estatísticos através de estimadores ou estatísticos de amostras principalmente aleatórias o que garante sua representatividade do universo de onde foi selecionada. O estudo dos intervalos de confiança denominados ainda de estimação intervalar faz desta ferramenta estatística uma das mais importantes áreas da teoria estatística.

Intervalos fiduciais são técnicas ou processos com objetivos de estimar ou "adivinhar" valores de parâmetros ou medidas características da população no que se refere a obter um valor pontual e intervalar com o objetivo de permitir a comparação de igualdade ou desigualdade entre duas ou mais medidas características das populações infinitas, entre valores esperados ou previstos e valores ocorridos, ou entre estatísticas de dois ou mais conjuntos separados no tempo e no espaço, pode-se estimar, por exemplo, as intenções da população de eleitores da cidade de Natal no Estado do Rio Grande do Norte sobre suas opiniões quanto ao pleito eleitoral para prefeito da capital do Estado do Rio Grande do Norte no ano de 2020, com uma determinada probabilidade de confiança dentro de uma margem máxima de erro de amostragem ou de estimação no qual os pesquisadores esperam estar o valor do parâmetro estimado ou no caso a verdadeira porcentagem de eleitores que deverá votar neste escrutínio eleitoral, bem como permitindo se obter qual é o tamanho mínimo que deve ter a amostra de eleitores para que os resultados obtidos reflitam a realidade da população de interesse no estudo com a máxima economia possível. Tudo isso tendo como suporte o cálculo de probabilidades que é outra área da estatística de grande importância e excelente suporte na pesquisa científica para que assim proporcione ou dê ao pesquisador o poder de realização de inferências confiáveis e precisas.

Nesse sentido o texto apresenta uma importante ferramenta de inferência estatística denominada de intervalos de confiança ou fiduciais paramétricos para estimação dos parâmetros populacionais referentes a média, diferença entre médias, dados emparelhados, variância, desvio padrão, razão entre variâncias, proporção relativa, diferença entre proporções percentuais, bem como um capítulo precedido da teoria dos métodos de estimação e do dimensionamento amostral. Também é apresentado um apêndice com os principais intervalos de confiança construídos e usados na estatística experimental e em estudos de regressão e correlação, e finalmente é apresentado um apêndice com planilhas, tabelas, distribuições de frequências bem como um formulário de apoio computacional para o auxílio da determinação das principais medidas ou estimadores pontuais amostrais, grandezas estatísticas amostrais, para serem utilizadas na construção dos intervalos de confiança. O texto é apresentado de forma clara e objetiva com diversos exercícios de aplicação e formulários, bem como o uso de tabelas para facilitar o aprendizado e a compreensão deste assunto. O leitor terá facilidade na assimilação do assunto, pois os capítulos seguem uma ordem e sequência lógica e bem didática.

São mostrados no final do livro diversos exemplos, tabelas auxiliares, o alfabeto grego com todas as letras, formulários dos principais tipos de dimensionamento amostral e construção dos intervalos de confiança, bem como uma lista atual da literatura especializada para o leitor interessado em se aprofundar mais no assunto.

Esta obra tem como principal objetivo atingir o público dos alunos de graduação dos cursos da área biológica, sem, contudo deixar de servir como um texto de apoio apropriado para pessoas que trabalham ou estudam nas áreas das ciências exatas, engenharias, administração, ciências contábeis, agronomia, zootecnia, engenharia de pesca, biologia, ecologia, medicina veterinária, medicina, bem como os demais cursos que tenham alguma abordagem quantitativa dos seus conteúdos programáticos. No entanto pesquisadores, técnicos, dirigentes de empresas, profissionais que utilizam a estatística em suas atividades profissionais, bem como alunos de pós-graduação podem também se nortear e devem se beneficiar com a leitura deste livro.

É evidente que uma obra desta natureza apresente falhas e imperfeições agravadas pelas limitações dos autores. Esperamos dos colegas e especialistas na área de estatística, valiosas críticas e sugestões que nos permita aprimorar esta valiosa obra de trabalho estatístico inferencial.

Mossoró, RN, Brasil, Agosto de 2021

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Os avanços no campo da probabilidade levaram ao desenvolvimento da teoria estatística frequentista permitindo aos pesquisadores das mais diversas áreas do conhecimento humano que eles façam generalizações científicas para as populações a partir de informações incompletas obtidas nas amostras. Enquanto a estatística descritiva trata da coleta, organização e apresentação dos dados, a inferência estatística trata de generalizações da parte para o todo, ou seja, da amostra para a população estudada. Na estatística descritiva medidas de posição ou tendência central tais como as médias, medidas de dispersão, medidas de assimetria, etc., representam fins em si mesmos, ao passo que na inferência estatística são meios no processo de investigação.

A inferência estatística tem como objetivo estudar generalizações sobre uma população infinita através de evidências fornecidas por uma amostra retirada de forma conveniente desta população. A amostra contém os elementos que podem ser observados e é onde as quantidades de interesse podem ser medidas. Estatística envolve métodos para o planejamento e condução de um estudo, descrição dos dados coletados e para tomada de decisões, predições ou inferências sobre os fenômenos representados pelos dados, ou seja, procedimentos para fazer generalizações sobre as características de uma população a partir da informação contida na amostra onde é possível a tomada de decisões e/ou a validação científica de uma conclusão. A inferência estatística consiste em fazer afirmações probabilísticas sobre as características do modelo probabilístico, que se supõe representar uma população infinita, a partir dos dados de uma amostra aleatória (probabilística) desta mesma população.

No cotidiano costuma-se fazer estimativas, como por exemplo, a Fundação IBGE estima a safra de grão do Brasil geralmente um ano antes da colheita, um estudante estima suas possibilidades de ser aprovado na disciplina de estatística são de 80%, um professor estima que 30% dos seus alunos são de nível superior, ou um Agrônomo estima que a produtividade de melão no Estado do Rio Grande do Norte aumentará 25%. Independente da forma como foram efetuadas, são estimativas pontuais dos verdadeiros parâmetros da população de interesse em estudo, mas desconhecidos valores para respectivamente, a probabilidade da previsão da safra de grãos, do estudante passar na disciplina de estatística, da proporção de alunos de nível superior, e da porcentagem de aumento da produtividade de melão.

A estimação estatística de características numéricas de uma população inclui situações muito diversas, como por exemplo, estimar o consumo médio de leite *In Natura* para o fornecedor prever as necessidades dos clientes, estimar as intenções de votos dos eleitores para os partidos políticos alterarem, se necessário, a forma como deverão se dirigir aos eleitores para fazer passar a sua mensagem, ou, ainda, estimar a inflação média para um governo poder corrigir eventualmente a sua política econômica.

Com o surgimento da computação eletrônica na primeira metade do século XX, o aumento da população humana, o avanço da medicina, o surgimento, de doenças, pragas, etc. a demanda cada vez maior por água, alimentos e energia, fez com que os cientistas se esforçassem em descobrir novas técnicas de análise de dados ampliando o campo de atuação da estatística indutiva, com o objetivo de beneficiar a sociedade com suas descobertas, principalmente devido ao desafio de se estudar populações grandes e infinitas, diante da limitação de tempo, pessoal, recuso e material o que muitas vezes torna esta tarefa difícil, cara, demorada, trabalhosa ou mesmo impossível. Restou a escolha de tratar as populações mediante as condições de incerteza mediante o uso de amostras oriundas destes universos de interesse, e sendo assim com o desafio das condições de incerteza foi feito o uso da teoria de probabilidades para medir o risco de suas conclusões de forma, econômica, precisa e com elevada exatidão.

E assim surgiu a importante ferramenta da estimação intervalar que é um dos componentes da chamada inferência estatística, que proporcionou a ciência um grande avanço e apoio nas suas descobertas, norteando os pesquisadores nas suas comparações, e atualmente continua se expandindo de forma espetacular principalmente com a utilização de softwares e pacotes estatísticos cada vez mais poderosos, baratos, de livre acesso e de código aberto e de largo alcance, o que produziu uma imensa gama de artigos científicos sobre este assunto, bem como um grande volume de publicações de excelentes livros principalmente editados na língua inglesa com menor ou maior profundidade da teoria.

Nesta unidade, vamos estudar como, a partir de estatísticas baseadas numa amostra aleatória, podemos fazer inferências ou generalizações acerca do valor de parâmetros de uma distribuição de probabilidade representativa de um particular universo estatístico. Embora existam dois métodos distintos para se fazer inferências, os quais são o método clássico e o método de Bayes, vamos considerar apenas o método clássico de estimação. Neste método, as inferências são baseadas apenas na informação contida numa amostra aleatória, enquanto no método de Bayes, além da informação contida numa amostra aleatória, também se usa o conhecimento prévio e subjetivo acerca da distribuição de probabilidade de parâmetros desconhecidos.

Os intervalos de confiança foram descobertos por Jerzy Neyman, Estatístico Polonês, que nasceu em 16 de abril de 1894 em Bendery na Moldavia e faleceu em 05 agosto de 1981 em Oaklandna Califórnia nos Estados Unidos. Seu nome original era Splawa-Neyman, mas ele desistiu da primeira parte do seu nome aos 30 anos. Ele lecionou matemática e estatística em Varsóviana Polônia e doutorou-se em 1924. Recebeu uma bolsa de estudos para trabalhar com o Estatístico Inglês Karl Pearson que nasceu em Londres em 27 de março de 1857 e faleceu em 27 de abril de 1936, em Londres, mas ficou desapontado ao descobrir que ele ignorava a moderna matemática.

A generalização para uma população estatística denominada de inferências estatísticas referente ao valor de um parâmetro θ de uma distribuição teórica de probabilidade de uma variável aleatória X ou população, são sempre baseadas em informação incompleta baseada numa amostra. O início de qualquer processo de estimação que é implementado é uma amostra aleatória simples de tamanho n independente e identicamente distribuída (i.i.d.), $x_1, x_2, \ldots, x_n$, extraída de uma população objeto de estudo, a partir da qual o pesquisador tenta obter respostas para as seguintes perguntas: i) Existe alguma estatística amostral $\hat{\theta}$ obtida em função dos valores $x_1, x_2, \ldots, x_n$ da amostra aleatória que possa ser usada como estimador ou estatística amostral do parâmetro populacional desconhecido θ ? ii) Se existir mais do que um estimador $\hat{\theta}$ de θ , como deve ser o processo de escolha deste, procurando qual é o estimador de melhor qualidade estatística, ou seja , verificar as propriedades desejáveis deste estatísticas amostral.

Frequentemente os pesquisadores nunca conhece o verdadeiro valor do parâmetro populacional θ, o qual ele pretende estimar ou estudar. No entanto este pesquisador pode definir critérios que permita mensurar a qualidade dos estimadores.

A inferência estatística ou a estatística indutiva pode ser visualizada de forma compreensível através do seguinte gráfico ou esquema de estudo. Onde pode-se visualizar diversas amostras aleatórias simples de tamanho "n" sendo extraídas de uma população alvo, e para cada uma dessas amostras o pesquisador obtém o estimador que lhe interessa no seu estudo (Figura 1).

Sendo assim um dos grandes objetivos da estatística é o conhecimento de grandezas desconhecidas da população que são os parâmetros, através de grandezas conhecidas nas amostras que sejam fiéis representantes dessa população. Estes parâmetros desconhecidos podem ser médias, desvios padrões e outros.

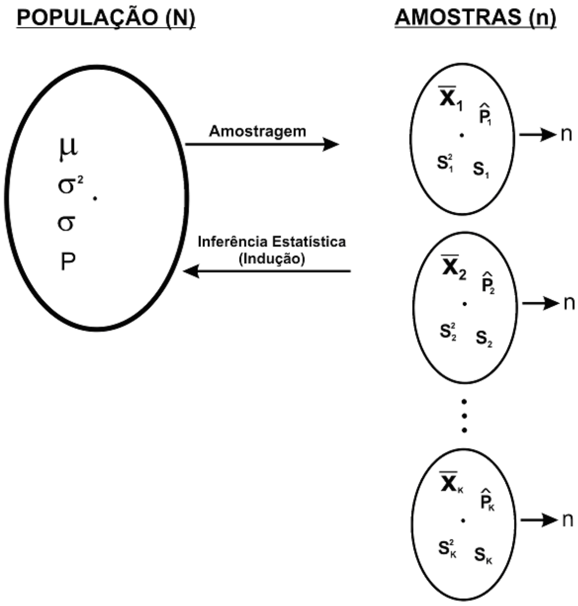


Figura 1. Diagrama representativo do processo de inferência estatística.

Como exemplo podemos citar a média μ e o desvio padrão σ de uma população são parâmetros populacionais (constantes fixas e desconhecidas em geral). A média $\overline{X}$ e o desvio padrão S de uma amostra são estimadores dos parâmetros. O valor numérico do estimador numa determinada amostra é chamado estimativa (ou estatística) amostral.

É importante destacar a diferença entre Dedução e Indução. A diferença entre o raciocínio dedutivo e o raciocínio indutivo. A dedução é o mecanismo que envolve a argumentação do geral para o específico, isto é da população para a amostra. A indução é o inverso, ou seja, é o raciocínio do especifico para o geral, isto é, da amostra para a população. Para tornar este processo melhor compreendido, vale salientar que a população estatística é o ponto de referência, o qual o pesquisador se norteia. O prefixo De significa A partir de. Assim dedução significa conclusões a partir da população, ou com base na população. O prefixo In indica o termo Para ou Em direção a. Assim, a dedução é a extrapolação em direção à população. Finalmente, a inferência estatística se baseia na indução, por isso é denominada de estatística indutiva.

2. CONCEITOS BÁSICOS

É importante para todo pesquisador das mais diferentes áreas do conhecimento humano conhecer antes do levantamento amostral ou de experimento conceitos, definições básicas próprios usados na inferência estatística ou termos próprios e de uso corriqueiro e frequente adotados na construção e interpretação desta importante ferramenta denominada de intervalo de confiança ou intervalo fiducial. Antes, porém vale apenas lembrar que um estimador ou uma estatística amostral, ou simplesmente uma estatística é uma função real das variáveis aleatórias independentes que são os elementos componentes da amostra aleatória simples usada na pesquisa em questão. Esses termos são os seguintes:

2.1. POPULAÇÃO

É o conjunto de dados, observações ou variáveis aleatórias X que apresenta pelo menos uma característica comum, cujo comportamento interessa analisar ou inferir.

Exemplo: A produção anual de melão em toneladas por hectare no Nordeste do Brasil.

2.2. AMOSTRA

É o subconjunto necessariamente finito convenientemente selecionado da população, o qual deve ser representativo desta, ou seja, que contenha suas características básicas, objeto de estudo ou de inferência.

Exemplo: A produção de melão em toneladas por hectare no Estado do Rio Grande do Norte (RN) em 2004.

2.3. PARÂMETRO θ

São as quantidades, medidas ou grandezas da população, em geral desconhecidas e constantes (fixas) e sobre as quais o pesquisador está interesse em estudar, e são usualmente representadas por letras gregas, tais como θ, μ e σ, entre outras.

Exemplo: A média da população μ.

2.4. ESTIMAÇÃO

É o processo através do qual se utiliza valores de estimadores ou estatísticos amostrais, para representar parâmetros populacionais desconhecidos.

Exemplo: método dos mínimos quadrados, método da máxima verossimilhança e método dos momentos.

2.5. ESTIMADOR $\hat{\theta}$ OU ESTATÍSTICA

É a grandeza estatística obtida através da combinação dos elementos da amostra, a qual é construída com a finalidade de representar ou estimar um parâmetro desconhecido de interesse na população, como por exemplo a média aritmética da amostra. Em geral, destaca-se por símbolos com o acento circunflexo: $\hat{m}$, $\hat{\theta}$, $\hat{\sigma}$, etc.

Exemplo: A média aritmética da amostra de tamanho n, a qual é dada pela seguinte equação:

$$\hat{m} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

2.6. ESTIMATIVA AMOSTRAL

São os valores numéricos assumidos pelos estimadores numa particular amostra.

Exemplo: O valor de 20 quilogramas assumidos pela média aritmética de uma amostra aleatória simples, $\overline{X} = 20$ kg.

2.7. INFERÊNCIA ESTATÍSTICA

É o estudo de técnicas que possibilitam a extrapolação ou generalização para uma população objeto de estudo das informações e conclusões obtidas a partir de amostras representativas dessa população.

Exemplo: Obter as preferências eleitorais da população de 450.000 eleitores de Natal, RN, numa eleição para prefeito dessa cidade, com uma margem de erro de 3 pontos percentuais para mais ou para menos, e uma nível de confiança de 95% de probabilidade.

2.8 TIPOS DE ESTIMAÇÃO

A estimação de parâmetros pode ser feita através de duas maneiras.

2.9. ESTIMAÇÃO POR PONTO

Uma estimativa por ponto constitui-se de um número obtido de computações sobre os valores da amostra que serve como uma aproximação do parâmetro estimado. Ou seja, é uma estatística $\hat{\theta}$ usada para estimar o valor do parâmetro θ.

Exemplo: Uma estatística $\hat{\theta}$ é um estimador pontual, se o seu objetivo for adivinhar o valor de θ. Por exemplo, a média amostral $\overline{X}$ é uma estimativa por ponto da média da população μ.

A tabela a seguir apresenta os estimadores por pontos (Tabela 1).

Tabela 1. Parâmetros populacionais e seus respectivos estimadores.

Parâmetro Populacional $\boldsymbol{\theta}$	**Estimador $\hat{\boldsymbol{\theta}}$**
μ	$\overline{X}$
$(\mu_1 - \mu_2)$	$(\overline{X}_1 - \overline{X}_2)$
P	$\hat{P}$
$(P_1 - P_2)$	$(\hat{P}_1 - \hat{P}_2)$
σ^2	S^2
$\frac{\sigma_1^2}{\sigma_2^2}$	$\frac{S_1^2}{S_2^2}$

Por exemplo, podemos afirmar que em uma amostra de 10 propriedades agrícolas produtores da cultura de melão no Rio Grande do Norte têm como estimador ($\hat{\theta}$) da verdadeira produtividade média (μ) de todas as propriedades desse Estado (população) a média da amostra $\overline{X}$, cujo valor pode ser $\overline{X} = 3$ toneladas/hectare.

2.10. ESTIMAÇÃO POR INTERVALOS OU INTERVALAR (INTERVALOS DE CONFIANÇA)

A estimação pontual não fornece ideia da margem de erro que é cometida ao se estimar um determinado parâmetro. A estimação por intervalo procura corrigir essa lacuna a partir da criação de um intervalo que garanta uma alta probabilidade de conter o verdadeiro valor do parâmetro desconhecido. Assim, para ilustrar esse processo, na estimação da produtividade média de uma nova cultiva de feijão pode-se apresentar o intervalo: [8,0 t/ha; 9,2 t/ha]. Isso significa que o verdadeiro valor da produtividade média, o qual é desconhecido, tem uma elevada probabilidade de ser um único valor compreendido pelo intervalo apresentado.

2.10.1. Apresentação

Quando a partir da amostra procuramos construir um intervalo do tipo $(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2)$, determinado por dois números obtidos de computação sobre os valores da amostra no qual se espera que contenha o valor do parâmetro no seu interior, temos o que se denomina de intervalo de confiança.

A estimativa por intervalo é feita de tal maneira que a probabilidade de o intervalo conter o parâmetro possa ser especificada, o qual é o coeficiente de confiança do intervalo dado por $[1 - \alpha]$. Ou seja,

$$P\,(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$$

2.10.2. Vantagem da estimativa por intervalo

É que ela mostra quão precisamente está sendo estimado o parâmetro.

2.10.3. Erro de estimação

Num intervalo de confiança o erro diz respeito ao desvio (diferença) entre o estimador $\hat{\theta}$ como, por exemplo, a média amostral e o verdadeiro valor do parâmetro populacional θ como, por exemplo, a verdadeira média da população. Como o intervalo de confiança tem centro na média amostral, o erro máximo provável é igual à metade da amplitude do intervalo, por exemplo,

$$\overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \frac{\sigma_{(x)}}{\sqrt{n}} \therefore \overline{X} \pm e$$

onde o erro é dado por $e = \frac{Z_{\left(\frac{\alpha}{2}\right)} \sigma_{(x)}}{\sqrt{n}}$.

$$e = \frac{1}{2} \times \text{(amplitude do intervalo de confiança)}$$

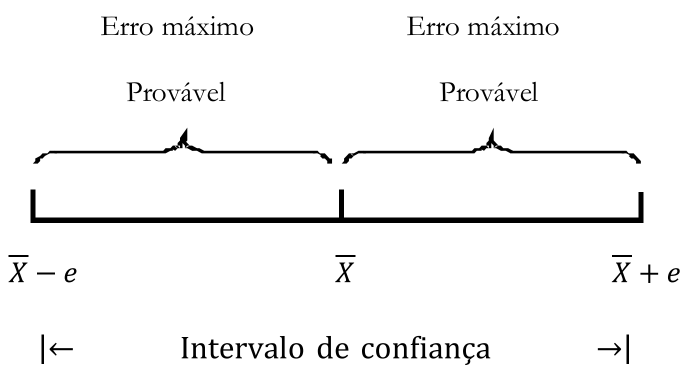


Figura 2. Gráfico representativo do erro de estimação num intervalo de confiança.

2.10.4. Fatores que influenciam na amplitude de um intervalo de confiança

Os fatores que podem determinar o comprimento ou amplitude de um intervalo de confiança são os seguintes:

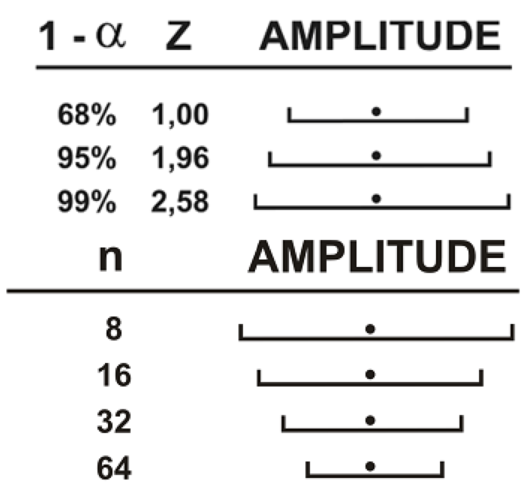
i) Coeficiente de confiança

Quanto maior é o valor do coeficiente de confiança, maior é a amplitude do intervalo de confiança, como pode ser visto no diagrama abaixo.

Sendo assim, quanto maior o intervalo de confiança, mais confiante o pesquisador estará de que realmente o intervalo de confiança calculado contenha o verdadeiro valor do parâmetro θ. Por outro lado, quanto maior o intervalo, menos informação teremos sobre o verdadeiro valor do parâmetro θ. A situação ideal seria aquela em que obtemos um intervalo de confiança relativamente curto com alta confiança.

É sabido que o comprimento do intervalo de confiança mede a precisão da estimação, e que podemos observar que a precisão é inversamente relacionada com o nível de confiança. O desejável seria obter um intervalo de confiança que fosse curto o suficiente para o propósito de tomada de decisão, e que também tivesse uma confiança adequada. Uma forma de conseguir isso seria através da escolha do tamanho da amostra n para ser grande o suficiente a fim de obtermos um intervalo de confiança de um determinado comprimento com uma confiança definida.

Exemplos:

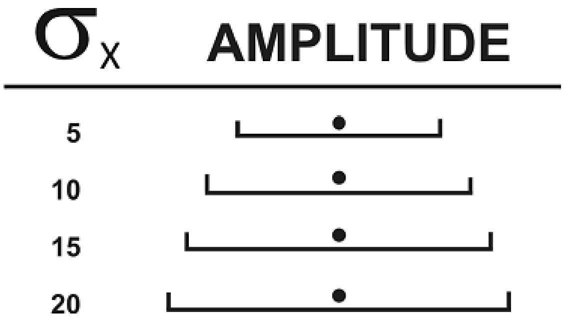


ii) Tamanho da amostra

Quanto maior for o tamanho da amostra "n", menor é a amplitude do intervalo de confiança, como é mostrado no diagrama abaixo:

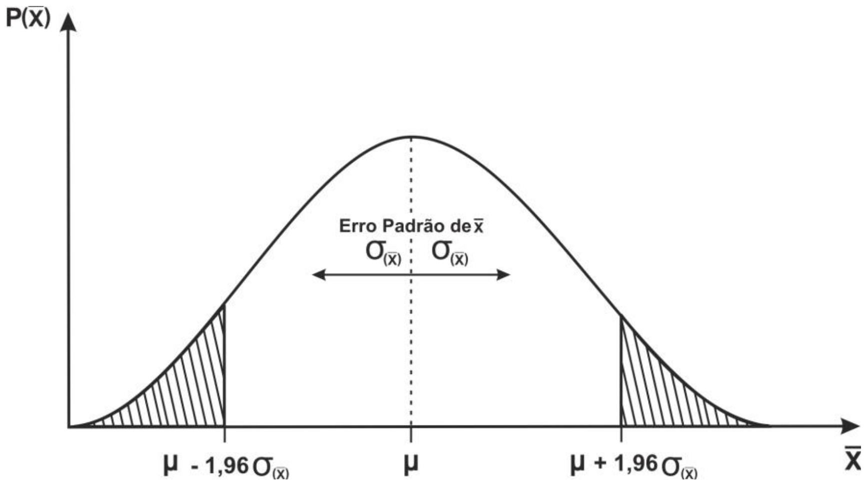
iii) Dispersão da população

Quanto maior for à dispersão da população medida pelo seu desvio padrão σ_x, maior será a amplitude do intervalo de confiança obtido como se observa a seguir:



Se o interesse for, por exemplo, a estimação da média populacional "μ" devemos construir um intervalo de confiança da forma $\mu = \overline{X} \pm$ e, ou seja, $\mu = \overline{X} \pm$ um erro de amostragem.

A questão crucial é saber qual deve ser a amplitude da tolerância para este erro amostral. A resposta, obviamente, dependerá de quanto o estimador $\hat{\theta}$ como, por exemplo, a média $\overline{X}$ flutuar isto é, da distribuição amostral de $\overline{X}$.



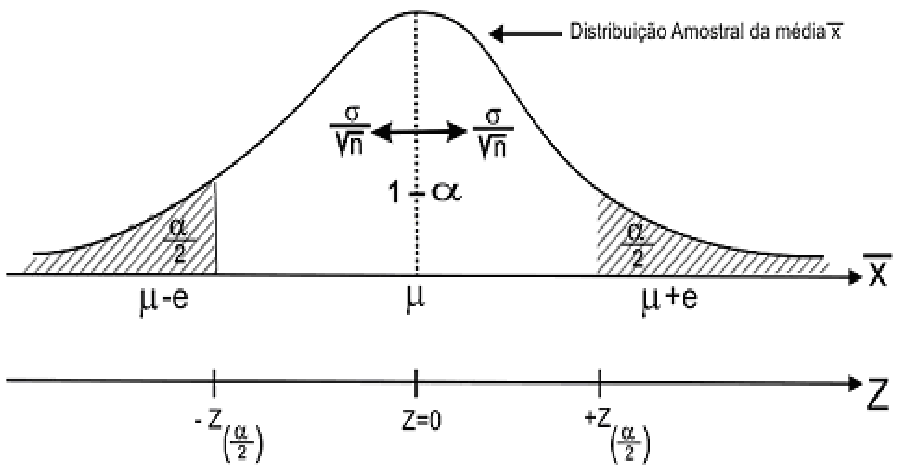


Figura 3. Gráficos representativos da distribuição amostral da média $\overline{X}$ para uma região crítica de 5 % de probabilidade e para o caso em geral.

Primeiro, devemos decidir quanto ao grau de confiança em nosso intervalo que deve efetivamente englobar μ é comum escolher um nível de confiança de 95%. Em outras palavras, utilizarem uma técnica que, a longo alcance nos dar um intervalo correto 19 vezes em cada 20.

O fato de o intervalo de $-1{,}96$ a $+1{,}96$ ser simétrico implica que é o mais curto de todos os intervalos que contém 95% da área isto, por sua vez, significa que o intervalo de confiança de 95% resultante é mais curto que qualquer outro intervalo do mesmo nível de confiança.

Os níveis ou coeficientes de confiança podem ser, por exemplo.

Nível de confiança $1 - \alpha = \frac{19}{20} = 0{,}95 = 95\%$

Nível de confiança $1 - \alpha = \frac{18}{20} = 0{,}90 = 90\%$

Nível de confiança $1 - \alpha = \frac{19{,}8}{20} = 0{,}99 = 99\%$

Isto significa, por exemplo, que ao adotarmos um nível de confiança de 95% de probabilidade, estaremos utilizando uma técnica que nos dará em repetidas amostragens da população de interesse, dezenove intervalos em cada 20 contendo o verdadeiro valor do parâmetro populacional desconhecido.

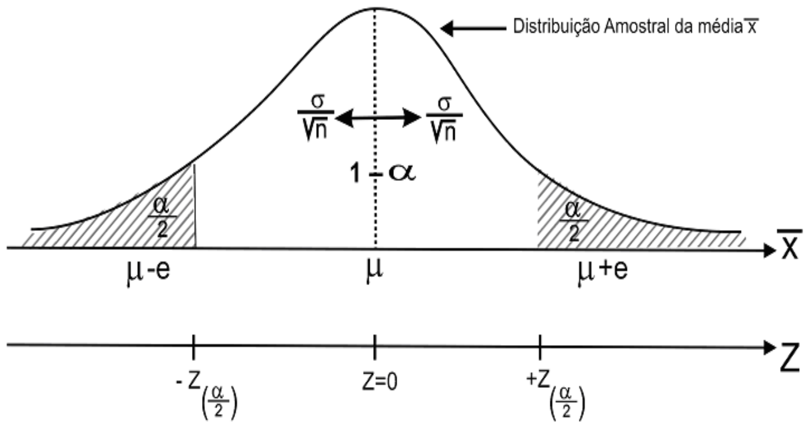


Figura 4. Distribuição amostral da média $\overline{X}$ representada pela distribuição normal padrão.

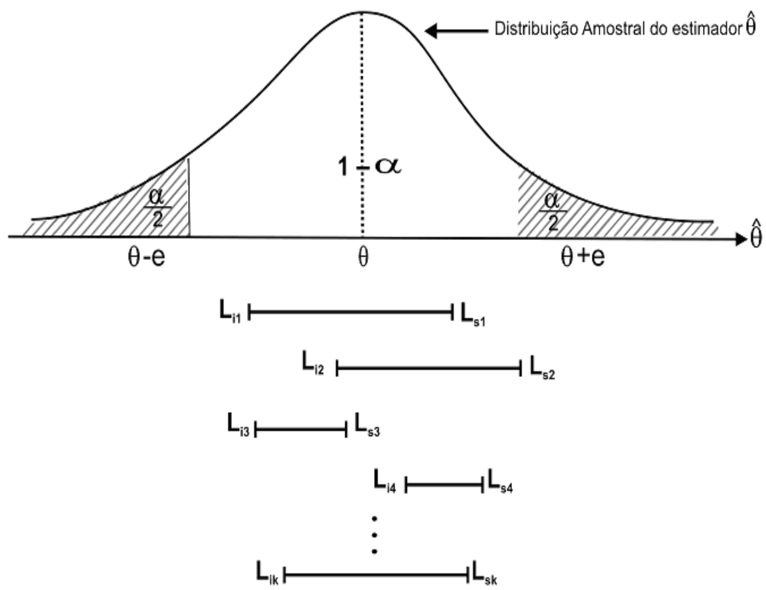


Figura 5. Gráfico representativo da distribuição amostral de um estimador $\hat{\theta}$, mostrando a construção dos diversos intervalos de confiança.

Podemos ainda afirma que na estimação por intervalos, em vez de se propor apenas um valor concreto para certo parâmetro da população, constrói-se um intervalo de valores (t_1, t_2) que, com certo grau de certeza, previamente estipulado, contenha o verdadeiro valor do parâmetro. Em muitos casos o intervalo é da forma a $(\hat{\theta} - e, \hat{\theta} + e)$ e "e" é considerado uma medida de precisão ou medida do erro inerente a estimativa $\hat{\theta}$. Desta forma, este método de estimação privilegia a confiança que se pode atribuir às estimativas.

Sendo assim, podemos definir o intervalo de confiança da seguinte maneira.

Seja θ um parâmetro desconhecido. Suponha-se que com base na informação amostral, pode-se encontrar as variáveis aleatórias T_1 e T_2, sendo $T_1 < T_2$, tais que:

$$P(T_1 < \theta < T_2) = 1 - \alpha, \text{ com } 0 < \alpha < 1$$

Sejam t_1 e t_2 realizações amostrais específicas de T_1 e T_2. Então o intervalo (t_1, t_2) e chamado o intervalo de confiança (I. C.) a $100(1 - \alpha)\%$ para θ, designando-se $(1 - \alpha)$ o grau de confiança do intervalo e α o nível de significância.

Se forem retiradas repetidamente um elevado número de amostras da população, o parâmetro θ deve estar contido em $100\,(1 - \alpha)\%$ dos intervalos calculados da forma atrás descrita. Portanto α representa o risco de que o parâmetro procurado não esteja no intervalo de confiança calculado.

A semiamplitude do intervalo de confiança também designada por erro de estimativa corresponde ao erro máximo que, com a confiança especificada, se pode cometer na estimativa de θ.

A amplitude do intervalo de confiança, como já comentado pode ser reduzida se:

- Se aumentar a dimensão ou tamanho da amostra "n";
- Mantendo a dimensão da amostra, se diminuir o grau de confiança $(1 - \alpha)$.

2.10.5. Método da variável pivotal

A especificação de um intervalo de confiança [I.C.] para um dado parâmetro θ implica o conhecimento simultâneo de:

1. Um estimador para o parâmetro em causa;
2. A sua distribuição amostral;
3. Uma estimativa pontual do parâmetro em causa.

2.10.5.1. Definição

Seja $X_1, X_2, \dots, X_n$, uma amostra aleatória de uma dada população com função (densidade) de probabilidade $f(x; \theta)$. Diz-se que a função das observações e o valor do parâmetro θ, $T(x_1, x_2, \dots, x_n; \theta)$, é uma variável pivotal se a respectiva função de densidade, $f_T(t)$, é independente de θ.

2.10.5.2. Obtenção do intervalo de confiança

i) Escolher a variável pivotal adequada para estimar o parâmetro pretendido;

ii) Fixar o grau de confiança $(1 - \alpha)$;

iii) Procurar dois números no domínio de T, $t_{1\left(\frac{\alpha}{2}\right)}$ e $t_{2\left(\frac{\alpha}{2}\right)}$, tais que:

i) $P\left(T < t_{1_{\left(\frac{\alpha}{2}\right)}}\right) = \frac{\alpha}{2}$

$$\Rightarrow P\left(t_{1_{\left(\frac{\alpha}{2}\right)}} < T < t_{2_{\left(\frac{\alpha}{2}\right)}}\right) = 1 - \alpha$$

ii) $P\left(T > t_{2_{\left(\frac{\alpha}{2}\right)}}\right) = \frac{\alpha}{2}$

Vale salientar que as condições $i)$ e $ii)$ permitem obter intervalos de amplitude mínima quando a distribuição de T é simétrica. Nos restantes casos as obtenções dos intervalos de confiança se tornam mais simples e conduz a intervalos com amplitude próxima da mínima.

iv) Resolver a desigualdade $t_{1_{\left(\frac{\alpha}{2}\right)}} < T < t_{2_{\left(\frac{\alpha}{2}\right)}}$ de forma a obter, $T_1 < \theta < T_2$;

v) Com base numa amostra concreta obter as realizações t_1e t_2 de T_1 e T_2;

vi) (t_1, t_2) e o I.C. para θ com $100\ (1 - \alpha)\%$ de confiança.

3. PROPRIEDADES DOS ESTIMADORES

Diversas propriedades dos estimadores são esperadas pelo pesquisador no processo de inferência estatística. A escolha de um estimador $\hat{\theta}$ de um parâmetro θ qualquer em detrimento de outro parâmetro θ_i depende de uma criteriosa avaliação dessas propriedades.

As seguintes propriedades conferem a um estimador a qualidade de um bom estimador, e são buscadas por qualquer pesquisador.

3.1. JUSTO, NÃO VIESADO, NÃO VICIADO OU NÃO TENDENCIOSO

Um estimador justo é uma estatística amostral cujo valor esperado em repetidas amostragens aleatório converge para o valor do parâmetro desconhecido que está sendo estimado.

Seja X uma variável aleatória discreta ou contínua (população) que dependa do parâmetro θ, e seja $\hat{\theta}$ um estimador ou estatística amostral de θ, sendo assim $\hat{\theta}$ é considerado justo se a sua média, esperança matemática ou valor esperado a longo prazo for igual ao valor do parâmetro populacional desconhecido, ou seja, $E[\hat{\theta}] = \theta$. Um estimador é não viesado ou não viciado se, em repetidas amostragens aleatórias da população, o estimador (estatística amostral) correspondente da distribuição amostral teórica desta, for igual ao parâmetro desconhecido da população θ o qual que quer estimar.

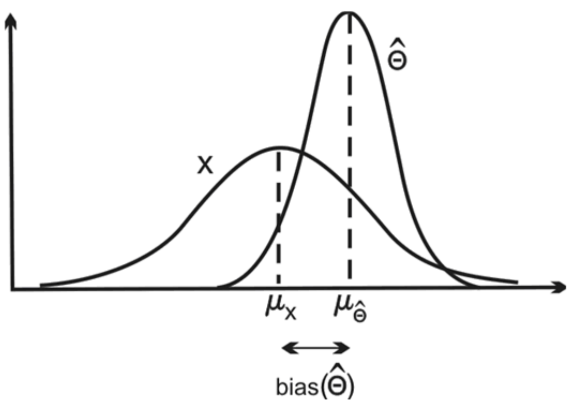


Figura 6. Curvas da distribuição amostral do estimador $\hat{\theta}$ e da população X, mostrando o enviesamento do estimador $\hat{\theta}$.

3.1.1. Exemplo 1

A média aritmética de uma amostra aleatória $\overline{X}$é um estimador justo da média da população μ, pois $E[\overline{X}] \ = \ \mu$, ou seja,

$$E[\overline{X}] = \left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n}E[X_1 + X_2 + \cdots X_n] = \frac{1}{n}E[X_1] + E[X_2] + \cdots + E[X_n] = \frac{1}{n}\{\mu + \mu + \ldots + \mu\}$$

$$= \frac{1}{n}n\mu = \mu$$

portanto $E[\overline{X}] = \mu$.

3.1.2. Exemplo 2

A variância amostral $S^2 = \frac{\sum(X-\overline{X})^2}{n}$, não é um estimador justo da variância populacional σ^2, pois $E(S^2) \ < \ \sigma^2$. Para se tornar justa temos que multiplicar $\frac{\sum(x-\bar{x})^2}{n}$ por $\frac{n}{n-1}$

$E\left(\frac{n}{n-1}S^2\right) = \sigma^2$, é uma estimativa justa de σ^2 é $\frac{S^2}{n} = \frac{\sum(x-\bar{x})^2}{n-1}$.

Aliás, a variância σ^2 é por definição $\sigma^2 = \frac{\sum(x-\mu)^2}{n}$ e como parâmetro μ, é desconhecido, é substituída pela estimativa $\overline{X}$, correspondendo a isso a perda de um grau de liberdade.

3.2. Consistente ou coerente

Um estimador amostral $\hat{\theta}$ de um parâmetro populacional θ é considerado como consistente se ele for justo ou não viciado e sua variância diminui tendendo para zero quando o tamanho da amostra adotada aumentar indefinidamente, sendo assim, o limite da probabilidade do desvio ou da diferença entre o parâmetro populacional e o estimador amostral ser maior que um determinado número positivo ε tende para zero quando o tamanho da amostra tende para infinito, ou de outra forma o limite da probabilidade do desvio ou da diferença entre o parâmetro populacional e o estimador amostral ser no máximo um determinado valor positivo ε tende para um quando o tamanho da amostra tende para infinito, ou seja, satisfazendo as seguintes condições:

$\lim_{n\to\infty} Prob\{|\theta - \hat{\theta}| > \varepsilon\} = 0$, para todo $\varepsilon > 0$, ou ainda,

$\lim_{n\to\infty} Prob\{|\theta - \hat{\theta}| < \varepsilon\} = 1$, onde ε é um valor positivo ($\varepsilon > 0$).

Na pesquisa prática a consistência de um estimador $\hat{\theta}$ não é muito simples de ser verificada. Sendo assim utilizam-se os seguintes critérios ou regras:

i) Se $E\,(\hat{\theta}) = \theta$, isto é, se o estimador $\hat{\theta}$ for justo, não viciado ou não viesado.

ii) Se $\lim_{n \to \infty} Var(\hat{\theta}) = 1$

Sendo assim $\hat{\theta}$ é um estimador consistente do parâmetro populacional desconhecido θ, ou ainda um estimador $\hat{\theta}$ de um parâmetro θ de uma população é consistente quando sua distribuição amostral se torna mais concentrada à medida que o tamanho da amostra n cresce indefinidamente. Um estimador $\hat{\theta}$ com consistência simples é aquele que converge em probabilidade para o parâmetro θ à medida que o tamanho da amostra aleatória n tende a infinito, ou seja, além de não ser viesado, sua variância converge para zero com o aumento de n, ou seja, $\lim_{n \to \infty} \sigma_{\hat{\theta}}^2 = 0$.

3.2.1. Exemplo

Por exemplo, para os estimadores da média populacional μ (média aritmética $\bar{X}$, e mediana M_d), de uma população normal pode ser verificada a consistência simples e construir um gráfico comparativo de ambas as situações em função do tamanho n da amostra aleatória adotada.

Para a média $\overline{X}$, temos que: $\lim_{n \to \infty} \sigma_{(\overline{X})}^2 = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0$

Para a mediana M_d, temos que: $\lim_{n \to \infty} \sigma_{(Md)}^2 = \lim_{n \to \infty} \left(\frac{\sigma^2}{n} \cdot \frac{\pi}{2}\right) = 0$

Neste caso, a média aritmética e a mediana são estimadores consistentes da média populacional μ, se $\overline{X}$ tem um erro padrão dado por $\frac{\sigma}{\sqrt{n}}$, e quando n cresce indefinidamente, isto é, $n \to \infty$, então a variância da média tende para zero, ou seja, $\frac{\sigma}{\sqrt{n}} \to 0$, e assim a distribuição amostral da média $\overline{X}$ torna-se mais centrada em torno da média populacional μ ou da distribuição de probabilidade da população X.

3.3. EFICIÊNCIA

Se dois estimadores amostrais $\hat{\theta}_1$ e $\hat{\theta}_2$ são centrados e estimam um mesmo parâmetro θ de uma determinada população estatística; baseados no mesmo número de observação n, então $\hat{\theta}_1$ serão mais eficientes que $\hat{\theta}_2$ quando e somente quando a variância do primeiro for menor que a variância do segundo, ou seja, $Var[\hat{\theta}_1] < Var[\hat{\theta}_2]$.

A eficiência relativa do primeiro estimador amostral relativamente ao segundo estimador pode ser determinada através da seguinte equação:

Eficiência Relativa $[\text{ER}] = \frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$

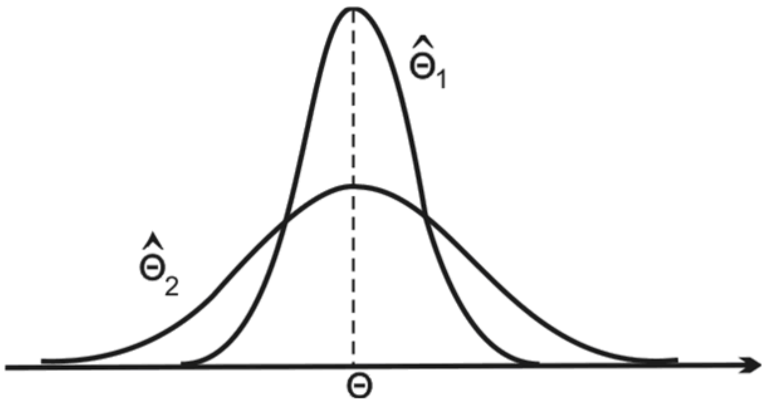


Figura 7. Curvas das distribuições amostrais dos estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, mostrando a eficiência de estimadores não enviesados $[Var(\hat{\theta}_1) < Var(\hat{\theta}_2)]$.

3.3.1. Exemplo

Considere dois estimadores médios aritméticos $\overline{X}$ e mediana M_d de um mesmo parâmetro populacional desconhecido neste caso à média de uma população normal (μ) obtidos em uma amostra aleatória de tamanho n. Verifique qual destes estimadores é o mais eficiente. É sabido que devido a simetria da distribuição normal ou Gaussiana, tem-se que a esperança matemática da mediana e a da média e igual a média populacional μ, ou seja,

$$E(Md) = E(\overline{X}) = \mu$$

Para determinar a eficiência relativa dos estimadores é necessário conhecer suas variâncias. A variância da média aritmética $\overline{X}$ é conhecida e dada por $\sigma^2_{(\overline{X})} = \frac{\sigma^2}{n}$ mas a variância da mediana M_d não é conhecida de forma exata, tendo apenas um valor determinado com uma boa aproximação. Essas variâncias estão apresentadas a seguir:

$$\sigma^2_{(\overline{X})} = \frac{\sigma^2}{n} \ e \ \sigma^2_{(Md)} \cong \frac{\sigma^2}{n} \cdot \frac{\pi}{2}$$

Através da razão, quociente ou da divisão entre as duas variâncias $\sigma^2_{(\overline{X})}$ e $\sigma^2_{(Md)}$ obtém-se a eficiência relativa (ER). Sendo assim tem-se que,

$$ER = \frac{\sigma^2_{(\overline{X})}}{\sigma^2_{(Md)}} = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} \cdot \frac{\pi}{2}} = \frac{2}{\pi} \cong 0{,}64 = 64\%$$

A eficiência relativa da mediana M_d em comparação com à média aritmética $\overline{X}$ é de apenas 64%. Isso pode ser interpretado em função do tamanho amostral, que seria necessário para se estimar a média da população μ usando a mediana, em detrimento daquele que seria necessário, usando a média aritmética como estimador amostral. Se o tamanho da amostra fosse igual a n para se estimar a média populacional μ através da média aritmética amostral $\overline{X}$, então o tamanho de uma amostra para estimá-la, usando a mediana M_d como estimador seria $1{,}57n$, para se alcançar a mesma precisão. Isso significa que uma estimativa da média da população "μ" usando 64 observações, para obter a média da amostra $\overline{X}$ dá a mesma informação que a de 100 observações usando-se a mediana amostral.

Se, entretanto, a população estivesse contaminada, ou seja, for muito heterogênica com presença de valores que dá a ela uma condição de heterogeneidade elevada, então a média aritmética da amostra ($\overline{X}$) teria uma maior possibilidade de apresentar maiores desvios da média populacional μ. Por outro lado a mediana M_d, por ser menos afetada por esses elementos discrepantes da amostra, poderia ser um estimador mais adequado.

3.4. SUFICIENTE

Seja θ um parâmetro populacional de interesse em uma determinada pesquisa científica. O princípio da suficiência estabelece que uma estatística amostral ou estimador amostral $\hat{\theta}$ é dito suficiente para o parâmetro · se ela captura toda a informação sobre θ contida na amostra, ou seja, se $X_1, X_2, \dots, X_n$ é uma amostra aleatória com variáveis aleatórias independente s e identicamente distribuídas (i.i.d.) retirada da população X, $\hat{\theta}$ é uma estatística amostral ou estimador suficiente para o parâmetro θ se qualquer inferência sobre este parâmetro θ depende na amostra n somente do valor $\hat{\theta}$, ou seja, não depende de um outro estimador $\hat{\theta}$. Se n_1 e n_2 são duas amostras aleatórias tais que $\hat{\theta}_{n_1} = \hat{\theta}_{n_2}$, então a inferência sobre o parâmetro θ é a mesma independente da observação n_1 e n_2 .

Sendo assim um estimador amostral $\hat{\theta}$ é suficiente se para uma dada amostra ele fornece toda a informação possível ou relevante a respeito do parâmetro θ, independentemente do auxílio de qualquer outra estimativa amostral. Vale salientar que os estimadores suficientes são os mais esperados pelos estatísticos e pelos pesquisadores em geral, em função de suas características e propriedades.

Um estimador $\hat{\theta}$ diz-se suficiente para um parâmetro populacional θ se a distribuição condicional da amostra $(X_1, X_2, \dots, X_n)$ dado o valor observado $\hat{\theta} = t$, não depende de θ.

Um estimador $\hat{\theta}$ diz-se suficiente se extrai da amostra toda a informação que esta contém sobre o parâmetro θ de tal maneira que, dado o valor observado $\hat{\theta} = t$, o conhecimento dos valores observados para os elementos da amostra nada acrescente sobre θ. Os estimadores suficientes gozam da propriedade de retirar da amostra toda a informação relevante sobre o parâmetro (Murteira et al., 2001).

3.4.1. Exemplo

Um estimador de máxima verossimilhança é um tipo de estimador suficiente, conforme exposto a seguir.

O princípio de máxima verossimilhança é um dos métodos utilizados na estatística usados para se obter estimadores. O estimador deste tipo lida com a estimação baseado nos resultados obtidos pela amostra aleatória e deve-se determinar qual a distribuição, dentre todas aquelas definidas pelos possíveis valores de seus parâmetros, com maior possibilidade de ter gerado tal amostra. Por exemplo, se for considerada uma população e uma variável aleatória X, relacionada a essa população, com função de probabilidade (se X é uma variável aleatória discreta) ou função densidade de probabilidade (se X é uma variável aleatória contínua) $f(x;\theta)$, sendo θ o parâmetro desconhecido.

Seja $P = (P_\theta : \theta \in \Theta)_{n \geq 1}$, sequência de modelos, com espaço paramétrico $\Theta \in \mathbb{R}^P$. Desta forma, seleciona-se uma amostra aleatória simples de X, de tamanho n, $X_1, X_2, \dots, X_n$, e sejam $X_1, X_2, \dots, X_n$, os valores efetivamente observados. A função de verossimilhança L é definida através da seguinte equação:

$$L(\theta; x_1, \dots, x_n) = f(x_1;\theta) \times \dots \times f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Se X é uma variável aleatória discreta com função de distribuição $p(x,\theta)$, a função de verossimilhança é dada por

$$L(\theta; x_1, \dots, x_n) = p(x_1;\theta) \times \dots \times p(x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta)$$

que deve ser interpretada como uma função de θ. Com isso, afirma-se que $\hat{\theta}$ é um estimador amostral de máxima verossimilhança (EMV) para $\theta = \theta_0$, se $f(x,\hat{\theta}) \in P$, e para algum $f(x,\theta_0) \in P$ obtém-se desde que qualquer parametrização seja identificável. Em outras palavras:

$$\prod_{i=1}^{n} f(\boldsymbol{x};\hat{\theta}) \geq \prod_{i=1}^{n} f(\boldsymbol{x};\theta_0)$$

$$L(\hat{\theta};\boldsymbol{x}) = \sup_{\theta_0 \in \Theta} L(\theta_0;\boldsymbol{x})$$

4. MÉTODOS ESTATÍSTICOS DE ESTIMAÇÃO UTILIZADOS PARA SE OBTER ESTIMADORES

Quando, em um estudo científico, se faz referência ao processo de estimar um parâmetro, se está buscando um método que permitirá conhecer um valor para a quantidade fixa θ que é desconhecida. Para se realizar esse processo são usadas expressões, as quais são funções das observações amostrais denominados de estimadores ou estatísticas amostrais. A seguir são apresentados alguns métodos para a obtenção de tais expressões. Além disso, procura-se mostrar ao leitor que tais estimadores são frutos de métodos cientificamente desenvolvidos. Muitos são os métodos existentes para obtê-las. Entretanto, são apresentados somente aqueles mais importantes.

4.1. TIPOS DE MÉTODOS DE ESTIMAÇÃO

Dentre os métodos de estimações pontuais mais usuais na inferência estatística destacam-se os seguintes:

4.1.1. Método dos momentos

Os estimadores obtêm-se por substituição dos momentos da amostra nas expressões que representam os momentos na população;

4.1.2. Método dos mínimos quadrados

São usualmente utilizados no âmbito da regressão linear.

4.1.3. Método da máxima verossimilhança

É provavelmente o método mais importante. Geralmente, os estimadores de máxima verossimilhança gozam das propriedades desejáveis num bom estimador: são os mais eficientes e consistentes. Embora, usualmente, não sejam centrados costumam ser assintoticamente não enviesados.

Uma forma de obter uma estimativa pontual de um parâmetro da população (por exemplo: μ, σ, σ^2 e p) é retirar uma amostra aleatória representativa dessa população e calcular o valor da estatística correspondente (por exemplo: $\bar{x},\ s,\ s^2,\ \bar{p}$).

4.1.4. Descrição dos métodos

4.1.4.1. Método dos momentos

Este método de estimação é um dos mais simples e mais antigos para obter estimadores de um ou mais parâmetros de uma distribuição. A ideia base é utilizar os momentos da amostra para estimar os correspondentes momentos da população, e, a partir daí, estimar os parâmetros de interesse (Murteira *et al.*, 2001). Seja $X_1, X_2, \dots, X_n$ uma amostra aleatória de uma dada população com função (densidade) de probabilidade f.d.p., $f(x;\theta_1,\theta_2,\dots,\theta_k)$ que depende de k parâmetros. Admitindo que existam os momentos ordinários, μ_r' da população X, estes são funções dos k parâmetros,

$$\mu_r' = E[X^r] = \begin{cases} \sum_{i=1}^{N} x^r f(x;\theta_1,\theta_1\dots,\theta_k), & \text{para distribuições discretas} \\ \int_{-\infty}^{\infty} x^r f(x;\theta_1,\theta_1\dots,\theta_k), & \text{para distribuições contínuas} \end{cases}$$

Os correspondentes momentos amostrais são dados por $m_r' = \frac{1}{n}\sum_{i=1}^{n} x_i^r$.

O método dos momentos consiste em considerar que os estimadores dos momentos ordinários são dados pelos momentos ordinários amostrais, ou seja, $\hat{\mu}_r' = m_r', r = 1,\dots,k.$

4.1.4.2. Método da máxima verossimilhança

O método da máxima verossimilhança foi introduzido por Ronald Aylmer Fisher em 1922. Sua introdução, em muitos aspectos, determinou o começo da teoria estatística moderna. Para apresentar o seu conceito, considere que $X_1, X_2, \dots, X_n$ seja uma amostra aleatória de uma população com densidade $f(x)$, determinada pelos parâmetros $\theta_i, i = 1,2,\dots,k$. Inicialmente, é considerada a situação específica de apenas um parâmetro $\theta\,(k = 1)$, por facilitar o entendimento do conceito a ser apresentado. Para uma amostra particular $X_1, X_2, \dots, X_n$, o estimador de máxima verossimilhança $(\hat{\theta})$ do parâmetro θ é aquele que maximiza a densidade conjunta de $X_1, X_2, \dots, X_n$.

Em razão do fato de os valores amostrais $X_1, X_2, \dots, X_n$ serem independentes é possível definir a densidade conjunta ou função de verossimilhança (L) pelo produtório das densidades de cada $X_i\,(i = 1,2,\dots,n)$. Assim, a função de verossimilhança, L, é definida pela seguinte equação: $L = f(x_1)f(x_2)f(x_3)\dots f(x_n) = \prod_{i=1}^{n} f(x_i)$.

O estimador de máxima verossimilhança é aquele que maximiza o valor de L na equação anterior.

Para obter o estimador de máxima verossimilhança ($\hat{\theta}$), basta tomar a primeira derivada de L em relação ao parâmetro θ, igualar a zero e resolver para θ. A solução é o estimador de máxima verossimilhança. Nem sempre uma solução explícita existe e métodos numéricos são utilizados para se obterem as estimativas. Quando se tem mais de um parâmetro, tomam-se as derivadas parciais de L com respeito a cada um deles. Iguala-se cada derivada a zero e resolve-se o sistema formado obtendo-se os estimadores de máxima verossimilhança dos parâmetros.

Algumas propriedades matemáticas da função L garantem a possibilidade de usar a função $S = ln(L)$ em seu lugar, uma vez que apresentam o máximo para o mesmo valor de θ. Isso é feito para tornar mais fácil a obtenção do máximo, uma vez que o produtório se transforma em somatório. Essa função é denominada de função suporte.

Na estimação paramétrica, este método só pode ser aplicado se a distribuição da população for conhecida.

A definição desse método pode ser explicada da seguinte maneira:

Seja $X_1, X_2, \dots, X_n$, uma amostra aleatória de uma dada população com função (densidade) de probabilidade, f.d.p., $f(x;\theta)$. Então a f.d.p. conjunta das variáveis que constituem a amostra é dada por:

$$f = (x_1, x_2, \dots, x_n; \theta) = f(x_1;\theta) f(x_2;\theta) \dots f(x_{ni};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Considerando uma amostra em concreto, designa-se por função de verossimilhança a função de θ e da amostra tal que: $L(x_1, x_2, \dots, x_n; \theta) = L(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$.

O Método da Máxima Verossimilhança pode ainda ser descrito da seguinte forma:

Consiste em encontrar o estimador $\hat{\theta}$ que maximiza o valor a função de verossimilhança para uma determinada amostra, ou seja, o valor de $\hat{\theta}$ que torna aquela amostra concreta mais provável, isto é, mais verossímil.

Para se obter o estimador de máxima verossimilhança devem-se fazer as seguintes operações:

i) Determinar a função de verossimilhança $L(x;\ \theta)$;

ii) Se necessário, aplicar a transformação logarítmica à função de verossimilhança $ln\, L(x;\ \theta)$. Geralmente, esta transformação torna o problema da maximização mais simples.

iii) Determinar os pontos onde a 1ª derivada da função ($L(x;\ \theta)$ ou $ln\ L(x;\ \theta)$) em relação a θ se anula (condição de primeira ordem):

$\frac{\partial L(x;\theta)}{\partial \theta} = 0$ ou $\frac{\partial InL(x;\theta)}{\partial \theta} = 0$

iv) Verificar se a 2ª derivada da função em relação a θ é negativa (condição de segunda ordem):

$$\frac{\partial^2 L(x;\theta)}{\partial \theta^2} < 0$$

ou

$$\frac{\partial^2 ln\ L(x;\theta)}{\partial \theta^2} < 0$$

O método da máxima verossimilhança pode ser utilizado para estimar mais do que um parâmetro em simultâneo.

4.1.4.3. Método dos quadrados mínimos

Outro importante método de estimação que aparece na teoria estatística da inferência é o dos quadrados mínimos ou mínimos quadrados em inglês *least squares method*. Grande parte da teoria em que se fundamenta a estatística experimental e baseada nesse método. Para apresentar o conceito e ilustrar o método, considerar o modelo linear a seguir para cada observação da amostra aleatória $X_1, X_2, \ldots, X_n$.

A observação amostral é modelada como resultante da soma de dois componentes básicos: um componente fixo (constante) μ e outro de natureza aleatória e_i.

$$X_i = \mu + e_i$$

em geral, o componente e_i é suposto normal com média 0 e variância constante σ^2 para todo valor de $i, i = 1,2,\ldots,n$.

Pela observação do modelo linear apresentado, verifica-se que μ e σ^2 são os parâmetros desconhecidos desse modelo. O parâmetro σ^2 não é, entretanto, mencionado explicitamente nesse modelo.

O método de estimação de quadrados mínimos baseia-se na minimização da soma de quadrados da variável aleatória denominada erro ou resíduo, e_i. Assim, isolando-se e_i no modelo anterior resulta em:

$$e_i = X_i - \mu$$

Tomando-se a soma dos quadrados (SQ) de e_i para as n observações amostrais tem-se:

$$SQ = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (X_i - \mu)^2$$

Para se obter o estimador de quadrados mínimos e necessário minimizar o valor dessa soma de quadrados (SQ). Para se obter esse mínimo, é necessário derivar em relação a cada parâmetro, igualar as derivadas a zero e resolver o sistema de equações formado. Nesse modelo, em que se está exemplificando o método, a soma de quadrados SQ só depende de um parâmetro (μ), então,

$$\frac{dSQ}{d\mu} = -2 \sum_{i=1}^{n} (X_i - \mu)$$

Igualando a zero a derivada primeira de SQ em relação ao parâmetro μ tem-se:

$$-2 \sum_{i=1}^{n} (X_i - \hat{\mu}) = 0$$

$$-\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \hat{\mu} = 0$$

$$n\hat{\mu} = \sum_{i=1}^{n} X_i \quad \therefore \quad \hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

A segunda derivada de SQ em relação à μ é:

$$\frac{dSQ}{d\mu d\mu} = 2n > 0$$

Esse resultado indica que $\mu = \bar{X}$ representa um ponto de mínimo da função SQ. Consequentemente, o estimador de quadrados mínimos da média $\mu = \bar{X}$. O estimador de momentos da variância α^2, não viesado, é: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \hat{X}_1\right)^2$. Em que, $\hat{X}_1 = \hat{\mu} = \overline{X}$ é o preditor de quadrados mínimos de X_i logo, $S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$.

4.1.4.3.1. Exemplos

1. Suponha que uma variável aleatória X representa o número de avarias ou defeitos de um motor usado em irrigação por aspersão durante um período de tempo de 3 meses e que obedece a uma lei de Poisson de parâmetro λ desconhecido. Para este parâmetro foram sugeridos dois estimadores:

$$\hat{\lambda} = \frac{X_1 + \cdots + X_n}{n} \quad \text{e} \quad \hat{\lambda} = \frac{X_1 + X_n}{n}$$

i) Compare-os quanto ao enviesamento.
ii) Deduza a variância para cada um deles.
iii) Qual dos dois estimadores é mais eficiente? Justifique a sua escolha.
v) Estude os dois estimadores quanto à consistência.

A resolução se dá da seguinte maneira:

i) $E[\hat{\lambda}] = E\left[\frac{X_1+\cdots+X_n}{n}\right] = \frac{1}{n}\{E[X_1] + E[X_2] + \cdots + E[X_n]\}$

$$= \frac{1}{n}[\mu + \mu + \cdots + \mu] = \frac{1}{n}n\mu = \mu$$

$$E[\hat{\lambda}] = E\left[\frac{X_1+X_2}{n}\right] = \frac{1}{2}\{E[X_1] + E[X_2]\} = \frac{1}{2}[\mu + \mu] = \mu$$

ii) Portanto, ambos os estimadores são centrados.

$Var[\hat{\lambda}] = Var\left[\frac{X_1+\cdots+X_n}{n}\right] \underbrace{=}_{X_i\ independentes} \frac{1}{n^2}\{Var[X_1] + Var[X_2] + \cdots + Var[X_n]\} =$

$\frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$

iii) $Var[\hat{\lambda}] = Var\left[\frac{X_1++X_n}{n}\right] \underbrace{=}_{X_i\ independentes} \frac{1}{2^2}\{Var[X_1] + Var[X_n]\} = \frac{1}{4}[\sigma^2 + \sigma^2] = \frac{\sigma^2}{2}$

iv) $Var[\hat{\lambda}] < Var[\hat{\lambda}]$, para n > 2. Portanto $\hat{\lambda}$ é mais eficiente.

v) $\hat{\lambda}$ é um estimador consistente, pois $E[\hat{\lambda}] = \mu$ e $Var[\hat{\lambda}] = \frac{o^2}{n} \to 0$, quando $n \to \infty$.

vi) $\hat{\lambda}$ não é um estimador consistente, pois $E[\hat{\lambda}] = \mu$ mas $Var[\hat{\lambda}] = \frac{o^2}{n}$, seja qual for o valor de n.

2. Seja $(X_1, X_2, \dots, X_n)$ uma amostra aleatória de uma distribuição Normal, $X \cap N\ (\mu, \sigma)$. Estime os parâmetros μ e pelo método:

i) Dos momentos.

ii) Da Máxima verossimilhança.

A resolução se dá da seguinte maneira:

i) Sabe-se que:

$$\mu_1' = E[X] = \mu$$

$$\mu_2' = E[X^2] = Var[X] + E^2[X] = \sigma^2 + \mu^2.$$

Para obter os estimadores $\hat{\mu}$ e $\hat{\sigma}^2$ pelos métodos dos momentos, é preciso resolver o seguinte sistema de equações:

$$\begin{cases} \mu_1' = m_1' = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i \\ \mu_2' = m_2' = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 \end{cases}$$

$$\begin{cases} \mu = \frac{1}{n}\sum_{i=1}^{n} X_i \\ o^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases}$$

$$\begin{cases} \mu = \bar{X} \\ o^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \end{cases}$$

Portanto, os estimadores obtidos foram: $\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \end{cases}$

ii) Função densidade de probabilidade (f.d.p.):

$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ para, $\sigma > 0$

Função de verossimilhança:

$$L(\mu, \sigma^2, x) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma^2}\right)^2}$$

$$= \frac{1}{\left[\sqrt{2\pi\sigma^2}\right]} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Logaritmo da Função de Verossimilhança:

$$ln\, L(\mu, \sigma^2; x) = -\frac{n}{2}(\ln 2 + \ln \pi + \ln \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Condição de 1ª ordem:

$$\begin{cases} \frac{\partial \ln L(\mu, \sigma^2; x)}{\partial \mu} = 0 \\ \frac{\partial \ln L(\mu, \sigma^2; x)}{\partial \sigma^2} = 0 \end{cases} \therefore \begin{cases} -\frac{1}{2\sigma^2}\left(-2\sum_{i-1}^{n} x_i + 2n\mu = 0\right) \\ -\frac{n}{2}\frac{1}{\sigma^2} + \sum_{i=1}^{n}(x_i - \mu)^2 \frac{2}{4\sigma^4} = 0 \end{cases} \therefore \begin{cases} \sum_{i=1}^{n} x_i - n\mu = 0 \\ -n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases} \therefore$$

$$\begin{cases} \mu = \sum_{i=1}^{n} \frac{x_i}{n} \\ \sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} \end{cases} \therefore \begin{cases} \mu = \bar{x} \\ \sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{n} \end{cases} \therefore \begin{cases} \mu = \bar{x} \\ \sigma^2 = S^2 \end{cases}$$

Condição de 2ª ordem:

$$\begin{cases} \dfrac{\partial^2 \ln L(\mu,\sigma^2;x)}{\partial \mu^2} = -\dfrac{1}{2\sigma^2} 2n < 0 \\ \dfrac{\partial^2 \ln L(\mu,\sigma^2;x)}{\partial \sigma^4} = \dfrac{n}{2}\dfrac{1}{\sigma^4} - \sum_{i=1}^{n} \dfrac{(x_i-\mu)^2}{2}\dfrac{2\sigma^2}{\sigma^8} < 0 \end{cases}$$

Portanto, os estimadores de máxima verossimilhança obtidos foram:

$$\begin{cases} \hat{\mu} = \bar{X} = \sum_{i=1}^{n} \dfrac{X_i}{n} \\ \hat{\sigma}^2 = \sum_{i=1}^{n} \dfrac{(X_i-\bar{X})^2}{n} = \dfrac{n}{n-1} S^2 \end{cases}$$

3. Considere uma população com distribuição de Bernoulli, com parâmetro p $(0 \leq p \leq 1)$.

i) Derive o estimador de máxima verossimilhança para o parâmetro p, sabendo-se que foi obtida uma amostra de dimensão $n = 3$, cujos valores observados foram $(1,1,0)$.

ii) Esboce o gráfico da função de verossimilhança e interprete-o.

iii) Forneça uma estimativa para p com base no método da máxima verossimilhança.

A resolução é processada da seguinte forma:

i) Função de probabilidade (f.p.): $f(x,p) = p^x(1-p)^{1-x}$, para $x = 0, 1$ e $0 \leq p \leq 1$.

Função de verossimilhança:

$$L(p;x) = \prod_{i=1}^{n} f(X_i,p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}$$

Logaritmo da Função de Verossimilhança: $ln\, L(p;x) = \sum_{i=1}^{n} X_i \, ln\, p + (n - \sum_{i=1}^{n} X_i)\, ln\,(1-p)$

Condição de 1ª ordem: $\frac{\partial\, ln\, L(p;X)}{\partial p} = 0 \therefore \sum_{i=1}^{n} X_i \frac{1}{p} + (n - \sum_{i=1}^{n} X_i)\frac{-1}{1-p} = 0 \therefore$

$$\therefore (1-p)\sum_{i=1}^{n} X_i - p\left(n - \sum_{i=1}^{n} X_i\right) = 0 \Leftrightarrow \sum_{i=1}^{n} X_1 - pn = 0 \Leftrightarrow p = \sum_{i=1}^{n} \frac{X_i}{n}$$

Condição de 2ª ordem:

$$\frac{\partial^2 lnL(p;x)}{\partial p^2} = -\frac{\sum_{i=1}^{n} X_i}{p^2} + \frac{(n - \sum_{i=1}^{n} X_i)(-1)}{(1-p)^2} = -\frac{\sum_{i=1}^{n} X_i}{p^2} - \frac{(n - \sum_{i=1}^{n} X_i)}{(1-p)^2} < 0$$

Pois $X_i \geq 0, p^2 \geq 0, n > 0, (1-p)^2 \geq 0$ e $n \geq \sum_{i=1}^{n} X_i$ pois $X_i = 0$ ou 1.

Portanto, o estimador de máxima verossimilhança é $\hat{p} = \sum_{i=1}^{n} \frac{X_i}{n}$

ii)

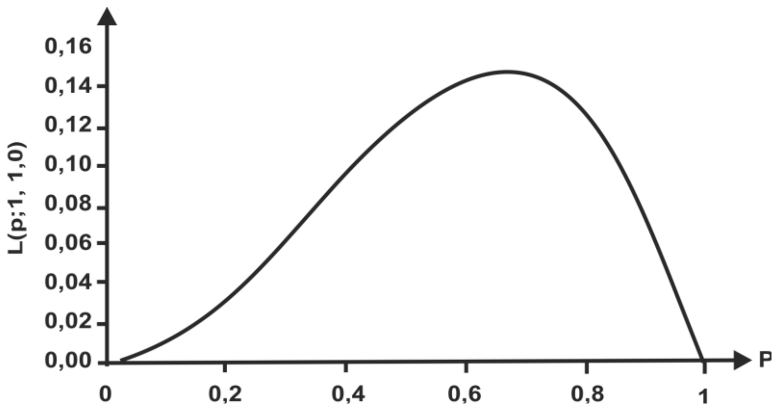


Figura 8. Gráfico representativo do estimador de máxima verossimilhança da proporção $\hat{p}$, com o tamanho da amostra $n = 3$.

iii) $\hat{p} = \frac{2}{3} = 0{,}6667$

iv) Obter o estimador de máxima verossimilhança do parâmetro λ da distribuição exponencial, a partir de uma amostra aleatória de tamanho n.

A função de verossimilhança da distribuição exponencial é:

$$L = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i,} = \lambda^n e^{-\lambda \sum_{i=l}^{n} X_i}$$

Tomando-se o logaritmo neperiano de L: $ln\, L = n \ln(\lambda) - \lambda \sum_{i=1}^{n} X_i$

Deriva-se $ln\, L$ em relação ao parâmetro λ, iguala-se a zero, resolve-se a equação formada e obtém-se o estimador de máxima verossimilhança:

$$\frac{d\, ln\, L}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i$$

$$\frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} X_i = 0 \Rightarrow \frac{n}{\hat{\lambda}} = \sum_{i=1}^{n} X_i$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$$

Logo, $\hat{\lambda} = \frac{1}{\overline{X}}$, é o estimador de máxima verossimilhança do parâmetro λ da distribuição exponencial.

5. ESTIMAÇÃO POR INTERVALO

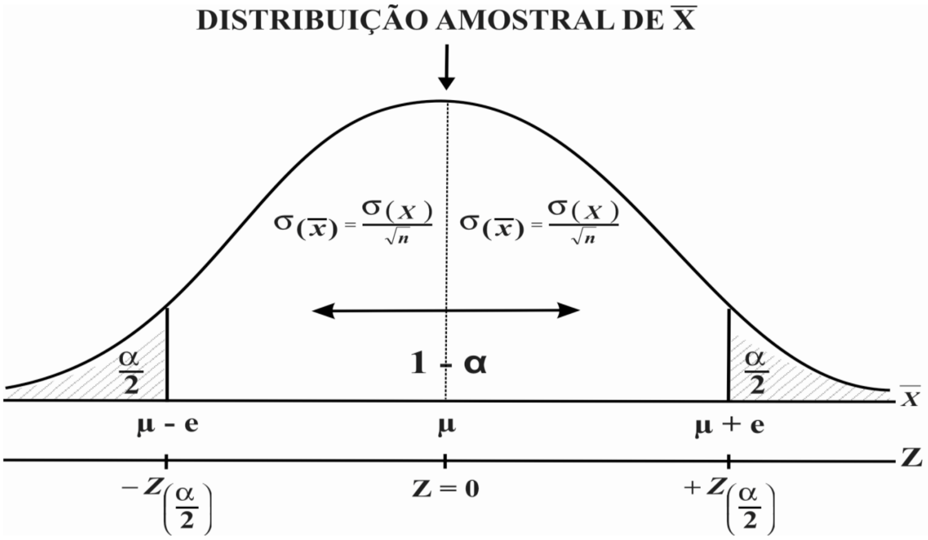
5.1 TIPOS DE INTERVALOS DE CONFIANÇA PARA MÉDIA (μ), DIFERENÇA ENTRE MÉDIAS $(\mu_1 - \mu_2)$, VARIÂNCIA (σ^2), DESVIO PADRÃO (σ), RAZÃO ENTRE VARIÂNCIAS $\left[\frac{\sigma_1^2}{\sigma_2^2}\right]$, PROPORÇÃO (P) E DIFERENÇA ENTRE PROPORÇÕES $(P_1 - P_2)$

Os tipos de intervalos de confiança mostrados aqui foram construídos através do método da função pivô ou da quantidade pivotal, onde demonstraremos o primeiro caso, sendo que os demais intervalos são deduzidos de forma semelhante, portanto temos que:

5.2. INTERVALO DE CONFIANÇA PARA A MÉDIA POPULACIONAL (μ) DE UMA POPULAÇÃO INFINITA OU FINITA E AMOSTRAGEM COM REPOSIÇÃO, NORMAL COM VARIÂNCIA (S^2) CONHECIDA OU NÃO, E COM A UTILIZAÇÃO DE AMOSTRA GRANDE $[N \geq 30]$.

5.2.1. O intervalo de confiança

O intervalo de confiança é construído baseado na distribuição amostral do estimador em questão que neste caso é a média aritmética de amostras aleatórias simples obtidas ou selecionadas em infinitas escolhas casuais, tendo uma distribuição de frequência ou de probabilidade Normal Padrão ou gaussiana, sendo assim com o uso da variável aleatória normal reduzida da variável aleatória média aritmética o tem-se assim obtido o erro de amostragem, erro de estimação máximo tolerável ou precisão estatística do intervalo a ser construído.



$\mu \leq \overline{X} + e$ e $\overline{X} - e \leq \mu$

Figura 9. Gráfico representativo da distribuição amostral da média $\overline{X}$ utilizado na construção do intervalo de confiança para a média populacional μ, quando se utiliza grandes amostras ($n \geq 30$).

Ou seja, temos que: $\overline{X} - e \leq \mu \leq \overline{X} + e$, e assim o intervalo é dado por:

$$P\left[\overline{X} - e \leq \mu \leq \overline{X} + e\right] = 1 - \alpha$$

$$Z_{\left(\frac{\alpha}{2}\right)} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \therefore \quad Z_{\left(\frac{\alpha}{2}\right)} = \frac{(\mu + e) - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \therefore \quad Z_{\left(\frac{\alpha}{2}\right)} = \frac{+e}{\frac{\sigma}{\sqrt{n}}} \quad \therefore$$

$$e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\overline{X} - e \leq \mu \leq \overline{X} + e$$

$$\overline{X} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}$$

$$P\left[\underbrace{\overline{X} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{LIC}} \leq \underbrace{\mu}_{\text{PARÂMETRO}} \leq \underbrace{\overline{X} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{LSC}}\right] = \underbrace{1 - \alpha}_{\text{COEFICIENTE DE CONFIANÇA}}$$

I. C.: $\overline{X} \pm e \therefore \quad \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}$ OU $\overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}$

Sendo assim o intervalo de confiança é determinado conforme a expressão seguinte:

$$P\left[\overline{X} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

onde $\overline{X}$ é a média de uma amostra aleatória, isto é, $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ ou $\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{n}$, onde, $n = \sum_{i=1}^{n} fi$ e X_i é o ponto médio de um intervalo de classe. σ o desvio padrão de uma população com Distribuição Normal e variância conhecida, e assim o intervalo de confiança de $100(1 - \alpha)\%$ para a média populacional μ é dado pela equação acima, onde $Z_{\left(\frac{\alpha}{2}\right)}$ é o valor crítico, tabelado ou teórico sob a curva da distribuição teórica de probabilidade normal padrão, reduzida ou gaussiana $f(Z)$. Caso não se conheça na prática o verdadeiro valor da variância populacional σ^2, pode-se estima-la através das variâncias amostrais utilizando-se as seguintes equações a seguir.

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1},\ S^2 = \frac{\sum X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1};$$

ou então,

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{n-1}, S^2 = \frac{\sum_{i=1}^{n} X_i^2 f_i - \frac{\left(\sum_{i=1}^{n} X_i f_i\right)^2}{n}}{n-1}$$

Por outro lado se o pesquisador desejar obter um intervalo de confiança para estimar o total T da população este é obtido multiplicando-se os limites obtidos no intervalo para a média populacional pelo número total de observações desta população o N.

$$N \cdot \left(\bar{X} - Z_{\left[\frac{\alpha}{2}\right]} \frac{\sigma}{\sqrt{n}} < T < \bar{X} + Z_{\left[\frac{\alpha}{2}\right]} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha(100\%)$$

Se a população for finita, isto é, $n \geq 5\%N$, e a amostragem for realizada sem reposição (ASR) então o erro padrão do estimador amostral deve ser multiplicado pelo fator de correção para população finita (FCPF) obtido pela equação $\frac{N-n}{N-1}$. Sendo assim o intervalo para a média populacional passa a ser este dado pela seguinte equação, com a mesma interpretação do intervalo de confiança (IC) no caso de população infinita (PI) ($n < 5\%N$) ou população finita (PF) ($n \geq 5\%N$) e amostragem com reposição (ACR).

$$\left(\bar{X} - Z_{\left[\frac{\alpha}{2}\right]} \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + Z_{\left[\frac{\alpha}{2}\right]} \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha(100\%)$$

5.2.2. Exercício de aplicação

Sabe-se de experimentos anteriores que a variância da produção leiteira de um rebanho bovino da raça holandês preto e branco criado na região de Natal, RN, é de 4,0 $(\text{Kg})^2$. Para avaliar a produção leiteira média leiteira atual, um criador(comprador) retirou uma amostra aleatória simples de $n = 36$ vacas, obtendo uma estimativa por ponto $\overline{X} = 5{,}4$.

a) Determine estimativas por intervalos para a produção leiteira média populacional atual (μ), aos níveis de confiança de (a.1) 0,95 e (a.2) 0,99.

Resolução do item (a.1):

$\sigma^2 = 4{,}0Kg^2$, ou seja, a variância populacional σ^2 é conhecida, sendo assim a população é normal. $n = 36, \overline{X} = 5{,}4$ e $\sigma = \sqrt{4} = 2Kg, 1 - \alpha = 0{,}95, \alpha = 0{,}05$. Pela tabela da normal temos que o valor crítico ou tabelado sob a curva normal padrão é dado por: $Z_{\left(\frac{0{,}05}{2}\right)} = \pm\ 1{,}96$. Sendo assim o intervalo de confiança é obtido como mostrado abaixo.

$$\text{I. C. }[\,\mu]0{,}95 \rightarrow \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \therefore 5{,}4 \pm 1{,}96\frac{2}{\sqrt{36}} \therefore 5{,}4 \pm 1{,}96\frac{2}{6} \therefore 5{,}4 \pm 0{,}6533$$

$$4{,}7467 \leq \mu \leq 6{,}0533 \therefore P[4{,}75 \leq \mu \leq 6{,}05] = 0{,}95$$

Interpretação: O intervalo de confiança de limites $4{,}75$ e $6{,}05$, deve conter a verdadeira produção leiteira média populacional (μ) atual expressa em quilogramas, do rebanho bovino de vacas leiteiras da raça holandês preto e branco, com 95% de probabilidade de confiança.

Resolução do item (a.2):

Neste caso agora o nível de confiança é 99% de probabilidade assim temos que: $1 - \alpha = 0{,}99, \alpha = 0{,}01$. Pela tabela da normal temos que o valor crítico ou tabelado sob a curva normal padrão é dado por: $Z_{\left(\frac{0{,}01}{2}\right)} = \pm\ 2{,}58$. Sendo assim o intervalo de confiança é obtido como mostrado abaixo:

$$\text{I. C. }[\,\mu]0{,}99 \rightarrow \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \therefore 5{,}4 \pm 2{,}58\frac{2}{\sqrt{36}} \therefore 5{,}4 \pm 2{,}58 \cdot \frac{2}{6} \therefore 5{,}4 \pm 0{,}86$$

$$4{,}54 \leq \mu \leq 6{,}26 \therefore P[4{,}54 \leq \mu \leq 6{,}26] = 0{,}99$$

Interpretação: O intervalo de confiança de limites $4{,}54$ e $6{,}26$, deve conter a verdadeira produção leiteira média populacional (μ) atual expressa em quilogramas, do rebanho bovino de vacas leiteiras da raça holandês preto e branco, com 99% de probabilidade de confiança.

b) Qual é o tamanho da amostra necessária para que se obtenha um intervalo de confiança com precisão $e = 0{,}5\ Kg$, ao nível $\alpha = 0{,}05$?

$n = ?;\ e = 0{,}5;\ 1 - \alpha = 0{,}95;\ \alpha = 0{,}05$ e assim o valor tabelado de Z é dado por: $Z_{\left(\frac{0{,}05}{2}\right)} = \pm\ 1{,}96$

$$\text{I. C. }[\,\mu\,] \qquad \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \quad \therefore \quad \overline{X} \pm\ e$$

$$e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \quad \therefore\ \sqrt{n} \cdot e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}\sqrt{n}\ \therefore \sqrt{n} \cdot e = Z_{\left(\frac{\alpha}{2}\right)}\sigma \therefore \frac{\sqrt{n} \cdot e}{e} = \frac{Z_{\left(\frac{\alpha}{2}\right)}\sigma}{e}$$

$$\sqrt{n} = \frac{Z_{\left(\frac{\alpha}{2}\right)}\sigma}{e} \therefore \left(\sqrt{n}\right)^2 = \left(\frac{Z_{\left(\frac{\alpha}{2}\right)}\sigma}{e}\right)^2 \therefore n = \frac{\left[Z_{\left(\frac{\alpha}{2}\right)}\right]^2 \cdot \sigma^2}{e^2}$$

$$n = \frac{[1{,}96]^2 \cdot 2^2}{[0{,}5]^2} = \frac{15{,}3664}{0{,}25} = 61{,}47 \cong 62 \quad vacas$$

Portanto o tamanho da amostra "n" deve ser de aproximadamente 62 animais. Ou seja, esse é o menor tamanho "n" que deve ter a amostra de vacas para ser usada na estimação da produção leiteira média populacional atual "μ" em quilogramas, em uma população infinita, normal, com variância (σ^2) conhecida de $4Kg^2$, e com amostra grande ($n = 36$), com um erro de estimação ou de amostragem "e" de no máximo 0,5 kg, e uma confiança de 95% de probabilidade.

5.3. INTERVALO DE CONFIANÇA PARA A MÉDIA (μ) DE UMA POPULAÇÃO INFINITA OU FINITA E AMOSTRAGEM COM REPOSIÇÃO, APROXIMADAMENTE NORMAL COM VARIÂNCIA (σ^2) DESCONHECIDA E AMOSTRA PEQUENA (N < 30)

5.3.1. O intervalo de confiança

$$P\left[\overline{X} - t_{\left(v;\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\left(v;\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

lembrado que a variável "t" possui $v = n - 1$ graus de liberdade, onde $\overline{X}$ e S são, respectivamente a média e o desvio padrão de uma amostra aleatória, Isto é, $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$; $\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{n}$, onde, $n = \sum_{i=1}^{n} fi$ e X_i é o ponto médio de um intervalo de classe.

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1},\ S^2 = \frac{\sum X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1};$$

ou ainda

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{n-1}, S^2 = \frac{\sum_{i=1}^{n} X_i^2 f_i - \frac{\left(\sum_{i=1}^{n} X_i f_i\right)^2}{n}}{n-1}$$

amostra esta extraída de uma população com distribuição normal e variância desconhecida, e assim o intervalo de confiança de $100(1 - \alpha)\%$ para a média populacional μ é dado pela equação acima, onde $t_{\left(v;\frac{\alpha}{2}\right)}$ é o valor crítico, tabelado ou teórico sob a curva da distribuição teórica de probabilidade t de Student com $v = n - 1$ graus de liberdade.

Por outro lado se o pesquisador desejar obter um intervalo de confiança para estimar o total T da população ele é construído multiplicando-se os limites obtidos no intervalo para média pelo número total de observações desta população o N.

$$N \cdot \left(\bar{X} - t_{\left(v;\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}} < T < \bar{X} + t_{\left(v;\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha(100\%)$$

Se a população for finita, isto é, $n \geq 5\%N$, e a amostragem for realizada sem reposição (ASR) então o erro padrão do estimador amostral deve ser multiplicado pelo fator de correção para população finita (FCPF) obtido pela equação $\frac{N-n}{N-1}$. Sendo assim o intervalo para a média populacional μ passa a ser este dado pela seguinte equação, com a mesma interpretação do intervalo de confiança (IC) no caso de população infinita (PI) ($n < 5\%N$) ou população finita (PF) ($n \geq 5\%N$) e amostragem com reposição (ACR)

$$\left(\bar{X} - t_{\left(v;\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + t_{\left(v;\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha(100\%).$$

5.3.2. Exercício de aplicação

Seja X a variável aleatória que representa a taxa normal de colesterol no plasma sanguíneo de coelhos da raça Norfolk. Suponhamos que com base em uma amostra casual simples de 25 animais normais, um pesquisador obteve a média $\overline{X} = 198\ mg/100\ ml$ de plasma e o desvio padrão $S = 30mg/100\ ml$ de plasma.

a) Obtenha com base nessa amostra, o intervalo de confiança com 90% de probabilidade para μ, isto é para a taxa média populacional de colesterol dos coelhos da raça Norfolk.

b) Para um erro de amostragem $e = 9{,}00$, a amostra satisfaz?

5.4. INTERVALO DE CONFIANÇA PARA A MÉDIA DAS DIFERENÇAS (μ_D) DE DUAS POPULAÇÕES INFINITAS OU FINITA E AMOSTRAGEM COM REPOSIÇÃO, QUE NÃO SÃO INDEPENDENTES, ISTO É, AS VARIÁVEIS SÃO EMPARELHADAS (DADOS EMPARELHADOS) E COM A AMOSTRA DE DIFERENÇAS (N) PEQUENA (N < 30)

$$P\left[\overline{D} - t_{\left(v;\frac{\alpha}{2}\right)} \frac{S_D}{\sqrt{n}} \leq \mu_D \leq \overline{D} + t_{\left(v;\frac{\alpha}{2}\right)} \frac{S_D}{\sqrt{n}}\right] = 1 - \alpha$$

onde, $v = n - 1$; n é o número de elementos da amostra de n diferenças; $\overline{D}$ é a média da amostra de n diferenças; SD é o desvio padrão da amostra de n diferenças.

5.4.1. Dados emparelhados

Os resultados das duas amostras constituem dados emparelhados, quando estão relacionados dois a dois segundo algum critério que introduz uma influência marcante entre os diversos pares, que supomos, porém, influir igualmente sobre os valores de cada par.

Ora, se os dados das duas amostras estão emparelhados tem sentido calcularmos as diferenças (d_i) correspondentes a cada par de valores, reduzindo assim os dados a uma única amostra de n diferenças.

Os exemplos 1 e 2 a seguir ilustram situações em que os dados obtidos de suas amostras são correlacionados (dados emparelhados).

Exemplo 1. Quando certo caráter é medido no mesmo indivíduo, em épocas diferentes, os valores obtidos nas duas mensurações tendem a ser mais parecidos entre si do que se houvessem sido obtidos de indivíduos diferentes. Como por exemplo, a medição de taxas de crescimento de plantas de uma determinada cultura, antes e depois de se aplicar uma substância inibidora da fotossíntese.

Exemplo 2. As eficiências de duas rações podem ser comparadas utilizando-se vários pares de animais irmãos de uma mesma leitegada (suínos, por exemplo), ou de uma mesma ninhada (camundongos, por exemplo).

O experimento consiste em alimentar cada membro, de cada par, com uma das rações alocada ao acaso. Indivíduos deste tipo (irmãos-germanos) pelas suas semelhanças genéticas tendem a apresentar respostas correlacionadas aos estímulos a que são submetidos.

5.5. INTERVALO DE CONFIANÇA PARA A DIFERENÇA ENTRE AS MÉDIAS (μ_1 - μ_2) DE DUAS POPULAÇÕES INFINITAS OU FINITAS E AMOSTRAGEM COM REPOSIÇÃO, COM VARIÂNCIAS (σ_1^2 E σ_2^2) CONHECIDAS [AMOSTRAS GRANDES ($N_1 \geq 30$ E $N_2 \geq 30$) OU PEQUENAS ($N_1 < 30$ E $N_2 < 30$)], OU COM VARIÂNCIAS DESCONHECIDAS [AMOSTRAS GRANDES ($N_1 \geq 30$ E $N_2 \geq 30$)].

$$P\left[\left(\overline{X}_1 - \overline{X}_2\right) - Z_{\left(\frac{\alpha}{2}\right)}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) + Z_{\left(\frac{\alpha}{2}\right)}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$

5.6. INTERVALO DE CONFIANÇA PARA A DIFERENÇA ENTRE AS MÉDIAS ($\mu_1 - \mu_2$) DE DUAS POPULAÇÕES INFINITAS OU FINITAS E AMOSTRAGEM COM REPOSIÇÃO, COM VARIÂNCIAS (σ_1^2 E σ_2^2) DESCONHECIDAS, E ESTATISTICAMENTE IGUAIS ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), E AMOSTRAS PEQUENAS ($N_1 < 30$ E $N_2 < 30$).

$$P\left[\left(\overline{X}_1 - \overline{X}_2\right) - t_{\left(v;\frac{\alpha}{2}\right)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) + t_{\left(v;\frac{\alpha}{2}\right)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = 1 - \alpha$$

Em que os graus de liberdade da variável t são dados por $v = n_1 + n_2 - 2$. Como não conhecemos σ^2, deveremos estimá-lo assim:

$$S_p^2 = \frac{(n_1 - 1)\ S_1^2 + (n_2 - 1)\ S_2^2}{n_1 + n_2 - 2} \text{ onde } S_p = \sqrt{S_p^2}.$$

5.7. INTERVALO DE CONFIANÇA PARA A DIFERENÇA ENTRE AS MÉDIAS ($\mu_1 - \mu_2$) DE DUAS POPULAÇÕES INFINITAS OU FINITAS E AMOSTRAGEM COM REPOSIÇÃO, APROXIMADAMENTE NORMAIS, COM VARIÂNCIAS (σ_1^2 E σ_2^2) DESCONHECIDAS E ESTATISTICAMENTE DESIGUAIS ($\sigma_1^2 \neq \sigma_2^2$) E COM AMOSTRAS PEQUENAS ($N_1 < 30$ E $N_2 < 30$).

$$P\left[\left(\overline{X}_1 - \overline{X}_2\right) - t_{\left(v;\frac{\alpha}{2}\right)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) + t_{\left(v;\frac{\alpha}{2}\right)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right] = 1 - \alpha$$

em que

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

onde $w_1 = \frac{S_1^2}{n_1}$ e $w_2 = \frac{S_2^2}{n_2}$.

5.8. INTERVALO DE CONFIANÇA PARA A VARIÂNCIA (σ^2) DE UMA POPULAÇÃO NORMAL

$$P\left[\frac{(n-1)S^2}{\chi^2_{\text{sup}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{inf}}\right] = 1 - \alpha$$

Observação: $v = n - 1$ dá os graus de liberdade da distribuição Qui-quadrado (χ^2), onde n é o número de elementos da amostra e S^2 é a variância da amostra

$$\chi_{inf}^{2^2_{\left(v;1-\frac{\alpha}{2}\right)_{sup}^{2^2_{\left(v;\frac{\alpha}{2}\right)}}}}$$

5.9. INTERVALO DE CONFIANÇA PARA O DESVIO PADRÃO (σ) DE UMA POPULAÇÃO NORMAL

$$P\left[\frac{(n-1)S^2}{\chi^2_{\text{sup}}} \leq \sigma \leq \frac{(n-1)S^2}{\chi^2_{inf}}\right] = 1-\alpha$$

Observação: $v = n - 1$ dá os graus de liberdade da distribuição Qui-quadrado (χ^2), onde n é o número de elementos da amostra e S^2 é a variância da amostra

$$\chi_{inf}^{2^2_{\left(v;1-\frac{\alpha}{2}\right)_{sup}^{2^2_{\left(v;\frac{\alpha}{2}\right)}}}}$$

5.10. INTERVALO DE CONFIANÇA PARA A RAZÃO $\left[\frac{\sigma_1^2}{\sigma_2^2}\right]$ ENTRE AS VARIÂNCIAS (σ_1^2 E σ_2^2), DE DUAS POPULAÇÕES NORMAIS

$$P\left[\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{\left(v_1,v_2;\frac{\alpha}{2}\right)}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2}\cdot F_{\left(v_1,v_2;\frac{\alpha}{2}\right)}\right] = 1-\alpha$$

onde: $V_1 = n_1 - 1$ e $V_2 = n_2 - 1$.

5.11. INTERVALO DE CONFIANÇA PARA A PROPORÇÃO (P) DE UMA POPULAÇÃO INFINITA OU FINITA E AMOSTRAGEM COM REPOSIÇÃO, NORMAL, E COM AMOSTRA GRANDE (N > 30)

5.11.1. O intervalo de confiança

$$P\left[f - Z_{\left(\frac{\alpha}{2}\right)}\cdot\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + Z_{\left(\frac{\alpha}{2}\right)}\cdot\sqrt{\frac{f(1-f)}{n}}\right] = 1-\alpha$$

onde $f = \hat{p} = \frac{x}{n}$, $\hat{p}$ é a proporção (frequência relativa) de eventos sucesso na amostra; x é o número (frequência absoluta) de eventos sucesso na amostra e n é o número de elementos da amostra (tamanho da amostra).

Para construir o intervalo de confiança para a proporção populacional p, é necessário conhecer o valor da verdadeira proporção p. Como não se conhece p na prática, pode-se proceder de duas maneiras:

i) Substituir $\mathrm{p}(1-\mathrm{p})$ por $\hat{\mathrm{p}}(1-\hat{\mathrm{p}})$ no erro padrão do estimador da proporção, ou seja,

$$P\left[\hat{p} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = 1 - \alpha(100\%)$$

ii) Usar a condição de que $\mathrm{p}(1-\mathrm{p}) \leq 1/4$, de modo que,

$$\sqrt{\frac{\mathrm{p}(1-\mathrm{p})}{n}} \leq \sqrt{\frac{1}{4n}}.$$

Desta forma o intervalo de confiança é dado por:

$$P\left[\hat{p} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{1}{4n}} \leq p \leq \hat{p} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{1}{4n}}\right] = 1 - \alpha(100\%)$$

Este intervalo é chamado de intervalo conservativo.

Por outro lado se o pesquisador desejar obter um intervalo de confiança para estimar o total T da população é obtido multiplicando-se os limites obtidos no intervalo para proporção pelo número total de observações desta população, N.

$$N \cdot \left[\hat{p} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq T \leq \hat{p} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = 1 - \alpha(100\%)$$

5.11.2. Exercício de aplicação

Para testar uma vacina contra a febre aftosa, um médico veterinário da UFERZAM em Mossoró, RN, selecionou ao acaso uma amostra aleatória (ou casual) simples de 80 bovinos da raça Gir, e, observou após um determinado período de 20 dias, que 60 desses animais foram imunizados.

a) Calcular a estimativa intervalar para a verdadeira proporção populacional "P" de bovinos da raça Gir imunizados com a vacina. Use um coeficiente de confiança de 95% de probabilidade.

b) Se a precisão for aumentada em 5%, qual é o número de animais da raça Gir que deve ser selecionado?

5.12. INTERVALO DE CONFIANÇA PARA A DIFERENÇA ENTRE AS PROPORÇÕES (P_1 – P_2), DE DUAS POPULAÇÕES INFINITAS OU FINITAS E AMOSTRAGEM COM REPOSIÇÃO, NORMAIS, COM AMOSTRAS GRANDES ($N_1 \geq 30$ E $N_2 \geq 30$).

$$P\left[(\hat{p}_1-\hat{p}_2)-Z_{\left(\frac{\alpha}{2}\right)}\cdot\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\leq \hat{p}_1-\hat{p}_2\leq(\hat{p}_1-\hat{p}_2)+Z_{\left(\frac{\alpha}{2}\right)}\cdot\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\right]=1-\alpha$$

onde, $\hat{p}_1=\frac{x_1}{n_1}$, $\hat{p}_2=\frac{x_2}{n_2}$, $\hat{q}_1=1-\hat{p}_1$, $\hat{q}_2=1-\hat{p}_2$.

É importante lembrar, através da seguinte observação, que é necessário usar o fator de correção para população finita dada por (FCPF $=\frac{N-n}{N-1}$), quando necessário. Por exemplo

$$t_{\left(v;\frac{\alpha}{2}\right)}\cdot\frac{S}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$

ou seja, para multiplicar o valor do erro padrão do estimador que está sendo usado na construção do intervalo de confiança.

Vale lembrar que a população infinita é aquela em que $n < 0{,}05.N$ ou $\frac{n}{N} < 0{,}05(ou\ 5\%)$ e a população finita é aquela em que $n \geq 0{,}05.N$ ou $\frac{n}{N} \geq 0{,}05(ou\ 5\%)$.

6. DIMENSIONAMENTO DE AMOSTRAS (N)

6.1. INTRODUÇÃO

O dimensionamento amostral através de cálculos do tamanho da amostra ou explicações para o tamanho da amostra é um dos primeiros itens na elaboração de um projeto de pesquisa experimental ou em um levantamento em qualquer área do conhecimento. O tamanho da amostra é o número de animais, plantas, ou outro elemento ou unidades investigadas que serão incluídos em um estudo e necessários para responder à hipótese da pesquisa no estudo. O objetivo principal do cálculo do tamanho da amostra é determinar o número suficiente de unidades necessárias para detectar diferenças entre tratamentos ou os parâmetros clínicos em um estudo médico desconhecidos ou os efeitos do tratamento ou a associação após a coleta de dados.

Fatores que determinam o tamanho da amostra: i) Hipótese de pesquisa a qual resume os elementos do estudo: a amostra, o tamanho da amostra, o desenho, o preditor e as variáveis de resultado; ii) Tamanho do efeito: Diferença mínima esperada; iii) Erro tipo I, α a probabilidade de fazer erro tipo isto é, rejeitar a hipótese nula quando ela é realmente verdadeira também chamada de nível de significância estatística ou α; iv) Erro Tipo II β que é a aceitação da hipótese nula quando ela for falsa. O erro tipo I é muito grave em comparação com o erro Tipo II. e v) Poder do teste de hipótese $1 - \beta$. A probabilidade de cometer um erro de tipo II é dada pelo valor beta (β), enquanto a probabilidade de evitar tal erro é denominada poder estatístico do estudo.

Power é a quantidade de $1 - \beta$. É a possibilidade de se observar uma associação de um determinado tamanho ou maior em uma amostra, se uma associação estiver realmente presente na população. Ao dimensionar uma amostra, necessita-se do conhecimento prévio da variância da população e do grau de precisão desejado, mas quando não se dispõem de informações sobre a variabilidade da população a ser amostrada, deve-se realizar uma pré-amostragem, em pequena escala, a fim de que se possa obter estimativas dos parâmetros populacionais (média e variância), que serão usados na obtenção do melhor tamanho da amostra (Silveira et al., 1980).

Conforme Bolfarine e Bussab (2005, p.70):

> Para a determinação do tamanho da amostra, é preciso fixar o erro máximo desejado (B), com algum grau de confiança $1 - \alpha$ (traduzido pelo valor tabelado Z_{α}) e possuir algum conhecimento a priori da variabilidade da população (σ^2). Os dois primeiros são fixados pelo pesquisador e, quanto ao terceiro, a resposta exige mais trabalho. O uso de pesquisas passadas, "adivinhações" estatísticas, ou amostras piloto são os critérios mais usados. Em

> muitos casos, uma amostra piloto pode fornecer informação suficiente sobre a população, de tal forma que se pode obter um estimador inicial razoável para σ^2. Em outros casos, pesquisas amostrais efetuadas anteriormente sobre a população também podem fornecer estimativas iniciais bastante satisfatórias para σ^2. Um outro procedimento, talvez menos dispendioso, seria considerar um intervalo onde aproximadamente 95% dos indivíduos da população estariam concentrados, e aí, igualar ao comprimento deste intervalo a quantidade 4σ. Teríamos então um valor aproximado para σ^2. Tal procedimento é baseado no fato de que no intervalo compreendido entre a média menos dois desvios padrões e a média mais dois desvios padrões (média $\pm$ 2DP), tem-se, em populações (aproximadamente) simétricas, aproximadamente 95% da população.

Se o tamanho da amostra for muito pequeno, o investigador pode não ser capaz de responder à questão do estudo. Por outro lado, o número de elementos, ou unidades amostrais ou ainda de pacientes em muitos estudos clínicos médicos é limitado devido a aspectos práticos como custo, inconveniência do paciente, decisões de não prosseguir com uma investigação ou um tempo de estudo prolongado. Os pesquisadores devem calcular o tamanho ideal da amostra antes da coleta de dados para evitar os erros por causa do tamanho da amostra muito pequeno e também desperdiçar dinheiro, trabalho e tempo, devido ao tamanho da amostra muito grande.

Além disso, os cálculos do tamanho da amostra para projetos de pesquisa são parte essencial de um protocolo de estudo para submissão a órgãos de financiamento de pesquisa, para comitês de ética ou para alguns periódicos científicos de revisão por pares. É muito importante determinar o tamanho da amostra de acordo com o desenho do estudo e os objetivos do estudo. Cometer erros no cálculo do tamanho da amostra pode levar a resultados incorretos ou insignificantes, ou não detectar efeito entre tratamentos na estatística experimental.

Existem dois tipos de erros que devem ser levados em conta ao projetar um estudo. Um erro do tipo I é o erro de rejeitar erroneamente a hipótese nula quando ela é verdadeira. O nível de significância é definido como a probabilidade de gerar um erro do tipo I e é denotado por α. Para proteger contra erros do tipo I, geralmente é definido como valores pequenos, como 0,05. Um erro do tipo II é o erro de aceitar incorretamente a hipótese nula quando ela é falsa. A probabilidade de fazer um erro do tipo II é denotada por β. O poder de um teste de hipótese é igual a $1 - \beta$ e é frequentemente expresso como uma porcentagem, em vez de uma proporção. Na pesquisa médica, por exemplo, é frequentemente definido pelo menos em 80%.

Amostras pequenas reduzem o poder de um estudo; no entanto, amostras grandes em cada grupo praticamente garantirão significância estatística entre os dois grupos. Portanto, um pesquisador precisa decidir com antecedência qual diferença entre os dois grupos seria de importância prática ou clínica no caso de estudos médicos.

A determinação do tamanho da amostra é uma parte importante do delineamento de ambos os estudos científicos sejam eles analíticos ou descritivos. O tamanho da amostra é uma estimativa do número de indivíduos ou elementos requeridos para detectar uma associação de um determinado tamanho e variabilidade do efeito, com uma probabilidade especificada de se cometer os erros do Tipo I (falso-positivo) e do Tipo II (falso negativo) por parte do pesquisador. A probabilidade máxima de fazer um erro de Tipo 1 é chamada α, e de fazer um erro de tipo 11, β. A quantidade $(1-\beta)$ é o poder do teste, a possibilidade de observar uma associação de um determinado tamanho ou maior em uma amostra, se uma estiver realmente presente na população. Esses estudos, que concluem sem resultados significativos, poderiam na verdade ser um exemplo de estudo sem poder adequado.

Para alcançar o objetivo desejado em estudos de pesquisa preocupados em estabelecer uma diferença entre grupos ou naqueles conduzidos para estimar uma quantidade, O planejamento adequado do tamanho da amostra é obrigatório. O cálculo do tamanho da amostra é um requisito fundamental de qualquer estudo controlado randomizado. A falha em realizar um cálculo do tamanho da amostra geralmente pode levar a uma tentativa que é fraca e pode perder uma diferença importante. Embora se diga que o inverso é verdade que um julgamento pode ser muito grande, na prática isso raramente ocorre, se é que ocorre.

Antes de se realizar qualquer cálculo, deve-se decidir o que constitui uma diferença de significado prático em qualquer área do conhecimento humano. A decisão sobre o que constitui uma diferença importante é geralmente arbitrária. Uma noção mais baseada em evidências sobre diferenças prováveis é observar estudos anteriores semelhantes que sejam úteis, mesmo que eles usem medidas de resultados diferentes. No entanto, geralmente, em nossa experiência, novos tratamentos, especialmente quando eles são testados contra comparações eficazes, tendem a produzir efeitos relativamente pequenos. Para estimar nosso tamanho de amostra, primeiro é necessário calcular um tamanho de efeito "padronizado". Isto é feito simplesmente tomando a diferença entre dois meios e dividindo pelo desvio padrão, que pode ser uma estimativa agrupada ou, mais simplesmente, o desvio padrão do grupo de controle.

É muito importante que a amostra retirada forneça uma representação precisa da população da qual ela é selecionada, a única maneira de se conseguir isso é selecionando a amostra através de um dispositivo aleatório qualquer mediante um mecanismo de sorteio o que garante igual probabilidade de seleção para os membros dos elementos populacionais que vão fazer parte da amostra. Do contrário, as conclusões sobre a população podem ser distorcidas ou viciadas. Todavia, na utilização da amostragem os resultados estão sujeitos a certo grau de incerteza, pois os dados mensurados em amostras podem conduzir a uma variação aleatória relativa ao método de medição, ao próprio material e também por considerar apenas uma parte da população (Heath, 1981).

Uma opção para se obter amostras de tamanho eficiente em estimar parâmetros populacionais desejáveis é a utilização da técnica de reamostragem de subamostras de tamanho reduzido ao de uma amostra de referência. O método da reamostragem permite uma comparação eficiente dos efeitos do tamanho da amostra na estimação de parâmetros genéticos e fenotípicos. Neste sentido, a técnica da simulação de subamostras permite obter, na maioria das vezes, um número inferior de dados da amostra referencial que seja capaz de fornecer as mesmas estimativas desta última e com a mesma exatidão.

Em uma amostragem não probabilística o tamanho ideal da amostra, é estabelecido sem nenhuma base de sustentação técnica. Comumente corresponde a 10% ou 15% da população alvo de tamanho N.

Já, em uma amostragem probabilística, o tamanho da amostra é função: do(s) parâmetro(s) a estimar como, por exemplo, da média, do nível de confiança desejável, do erro tolerável, de estimação ou de amostragem, índice de precisão escolhidos, do grau de dispersão da população medida pelo desvio padrão, pode, ainda, depender do tamanho da população N e de outros parâmetros específicos, tais como tempo, recursos financeiros, tipo de pesquisa, objetivo do estudo, tipo de material, etc..

Basicamente, o tamanho da amostra depende da precisão desejada o qual está em função da decisão do pesquisador. Assim, é intuitivo perceber que o tamanho depende do erro aleatório mencionado acima.

Há uma relação inversa entre o erro e o tamanho da amostra. Amostras grandes estão associadas a erros pequenos e amostras pequenas a erros grandes. Assim, deve-se procurar uma compatibilidade entre o tamanho amostral e o erro que máximo tolerável que se assume cometer em um estudo ou pesquisa.

A precisão do intervalo de confiança que é dada por $(1 - \alpha) \times 100\ \%$ para média, por exemplo, é metade da sua amplitude (semiamplitude do intervalo de confiança), ou seja, $Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}$, $t_{\left(v;\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}$ ou $Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}$.Assim, antes de efetuar a amostragem, pode-se estimar, com um grau de confiança de $(1 - \alpha) \times 100\ \%$ dado, o tamanho "n" da amostra que garante que o erro máximo cometido (precisão) o qual é dado por: $Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}$, $t_{\left(v;\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}$ ou $Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}$, não ultrapasse um valor e desejado. Para isso conforme o caso procura-se solucionar a seguinte inequação.

$$Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \leq e,\ t_{\left(v;\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}} \leq e \ ou\ Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}} \leq e$$

em relação à "n", obtendo-se, respectivamente,

$$n \geq \left(\frac{Z_{\left(\frac{\alpha}{2}\right)}\sigma}{e}\right)^2, \; n \geq \left(\frac{t_{\left(v;\frac{\alpha}{2}\right)}S}{e}\right)^2 \text{ ou } n \geq \left(\frac{Z_{\left(\frac{\alpha}{2}\right)}S}{e}\right)^2$$

e, portanto basta tomar para "n" o menor inteiro que satisfaz a desigualdade. Sendo assim, podemos concluir de imediato que o tamanho da amostra é inversamente proporcional ao quadrado do erro amostral e. Por exemplo, para o erro diminuir 10 vezes, o tamanho da amostra tem de aumentar 100 vezes.

É claro que, na maioria das situações, a variância populacional σ^2 é desconhecida. Nestas circunstâncias, antes de se determinar a ordem de grandeza de n através das seguintes fórmulas,

$$n \geq \left(\frac{t_{\left(v;\frac{\alpha}{2}\right)}S}{e}\right)^2 \text{ ou } n \geq \left(\frac{Z_{\left(\frac{\alpha}{2}\right)}S}{e}\right)^2$$

recorre-se a uma amostra preliminar ou amostra piloto de tamanho $n \geq 30$, para se determinar o desvio padrão amostral S.

Sendo assim é de fundamental importância o dimensionamento amostral para o processo de estimação ou de inferência estatística em geral, pois a determinação do tamanho da amostra ou o cálculo do tamanho amostral, ou seja, o seu tamanho n é um fator que influencia todo o processo, tais como determina a amplitude do intervalo de confiança, ou seja, a precisão da estimação, os custos, a distribuição amostral do estimador, e finalmente a validade das conclusões e generalizações sobre a população de interesse obtida através do processo de indução estatística. É mostrado a seguir então alguns exemplos de aplicação sobre dimensionamento amostral em diversas situações.

6.1.1. Exemplo 1

Um experimento consiste em estimar a resistência média de fibras de algodão mocó. Qual deve ser o tamanho da amostra, para um erro de estimação inferior a 6 kgf e um nível de confiança de 95%, sabendo-se que uma amostra piloto forneceu um valor para o desvio padrão $S = 24\ kgf$.

O tamanho da amostra é dado por $n \geq \left(\frac{Z_{\left(\frac{\alpha}{2}\right)}S}{e}\right)^2$. Neste caso, $Z_{\left(\frac{0,05}{2}\right)} = 1{,}960,\ S = 24$ e o erro $e = 6$. Portanto,

$$n \geq \left(\frac{1{,}90.24}{6}\right)^2 = \frac{(1{,}960)^2.(24)^2}{(6)^2} = \frac{3{,}8416.576}{36} = 61{,}4656 \text{ fibras}$$

6.1.2. Exemplo 2

Um experimento consiste em estimar a resistência média de fibras de algodão mocó. Qual deve ser o tamanho da amostra, para um erro de estimação inferior a 3 kgf e um nível de confiança de 95 %, sabendo-se que uma amostra piloto forneceu um valor para o desvio padrão $S = 24$ kgf.

O tamanho da amostra é dado por $n \geq \left(\frac{Z_{\left(\frac{\alpha}{2}\right)} S}{e}\right)^2$. Neste caso, $Z_{\left(\frac{0{,}05}{2}\right)} = 1{,}960,\ S = 24$ e o erro $e = 3$. Portanto,

$$n \geq \left(\frac{1{,}90.24}{3}\right)^2 = \frac{(1{,}960)^2.(24)^2}{(3)^2} = \frac{3{,}8416.576}{9}$$

$$n = 246 \text{ fibras}$$

Como se pode verificar, para reduzir o erro pela metade e manter o mesmo nível de confiança, o tamanho da amostra é multiplicado por 4 (que é o quadrado de 2).

6.2. TAMANHO DA AMOSTRA PARA ESTIMAR A MÉDIA DA POPULAÇÃO "M" DE UMA POPULAÇÃO NORMAL INFINITA OU FINITA E AMOSTRAGEM COM REPOSIÇÃO E COM VARIÂNCIA CONHECIDA (σ^2) OU DESCONHECIDA (S^2), OU COM AMOSTRA GRANDE (N ≥ 30), COM UMA CONFIANÇA ($1 - \alpha$) E UM ERRO DE AMOSTRAGEM DE NO MÁXIMO IGUAL A "E"

6.2.1. Para população infinita ou população finita e amostragem com reposição

Partindo da fórmula do erro de estimação ou de amostragem "e" temos que:

$$e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma_{(X)}}{\sqrt{n}} \quad \therefore e \cdot \sqrt{n} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma_{(X)}}{\sqrt{n}} \cdot \sqrt{n} \therefore e \cdot \sqrt{n} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma_{(X)} \therefore \frac{e \cdot \sqrt{n}}{e} = \frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma_{(X)}}{e}$$

$$\sqrt{n} = \frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma_{(X)}}{e} \therefore \left(\sqrt{n}\right)^2 = \left(\frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma_{(X)}}{e}\right)^2 \therefore \left(\sqrt{n}\right)^2 = \left(\frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma_{(X)}}{e}\right)^2$$

$$n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)} \cdot \sigma^2_{(X)}}{e^2} = \frac{Z^2_{\left(\frac{\alpha}{2}\right)} \cdot S^2_{(X)}}{e^2}$$

Portanto esse é o menor tamanho que deve ter a amostra "n" para se estimar a média "μ" de uma população normal, com variância (σ^2) conhecida ou não, usando grandes amostras ($n \geq$

30), com um coeficiente de confiança $(1-\alpha)\%$, e um erro de amostragem ou de estimação de no máximo igual a "e".

O valor $Z_{\left(\frac{\alpha}{2}\right)}$ corresponde ao valor crítico ou tabelado sob a curva normal ao nível de confiança $(1-\alpha)$ desejado, $\sigma_{(X)} = S^2_{(X)}$ é uma estimativa da variabilidade da população e "e" representa o erro de amostragem ou de estimação máximo permitido, o qual mede a precisão do processo de estimação.

6.2.1.1. Exemplo

Qual é o tamanho da amostra necessária para se estimar a produtividade média de algodão em kg/ha de uma população de plantios considerada infinita, cujo desvio padrão é igual a 4t/ha, com 98% de confiança e uma precisão de 0,50.

$$1 - \alpha = 98\,\%, \alpha = 2\,\% = 0{,}02$$

$$Z_{\left(\frac{\alpha}{2}\right)} = 2{,}33 e = 0{,}50$$

$$n = \left(\frac{(2{,}33)^2 4^2}{0{,}50^2}\right) = \left(\frac{5{,}4289.16}{0{,}25}\right) = \left(\frac{86{,}8624}{0{,}25}\right) = 347{,}4496 = 348 \text{ plantios}$$

6.2.2. Para população finita e amostragem sem reposição

Se a população for finita ou se amostragem for feita sem reposição a expressão do erro de amostragem ou de estimação deve ser multiplicado pelo fator de correção de população finita, o que nos leva a seguinte fórmula para "n".

$e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma_{(X)}}{\sqrt{n}} \cdot \sqrt{\frac{(N-n)}{(N-1)}}$, onde N é o tamanho da população que estamos estudando.

Resolvendo esta expressão para n', obtemos:

$$\sqrt{n}\sqrt{\frac{(N-n)}{(N-1)}} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma_{(X)}}{e}$$

Elevando ambos os membros da equação ao quadrado, vem o seguinte:

$$\left[\sqrt{n}\sqrt{\frac{(N-n)}{(N-1)}}\right]^2 = \left[Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma_{(X)}}{e}\right]^2$$

$$n \cdot \frac{(N-n)}{(N-1)} = Z^2{}_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma^2{}_{(X)}}{e^2}$$

Substituindo o segundo termo por n_0, fica o seguinte:

$$n \cdot \frac{(N-n)}{(N-1)} = n_0$$

$$n \cdot N - n = n_0 N - n_0 n$$

$$n(N + n_0 - 1) = n_0 N$$

$$n = \frac{n_0 N}{N + n_0 - 1}$$

Dividindo ambos os termos por "N", temos que:

$$n' = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

onde o "n_0" é o "n" mostrado anteriormente.

6.3. TAMANHO DA AMOSTRA PARA ESTIMAR A MÉDIA "M" DE UMA POPULAÇÃO INFINITA OU FINITA E AMOSTRAGEM COM REPOSIÇÃO DE UMA POPULAÇÃO APROXIMADAMENTE NORMAL COM VARIÂNCIA DESCONHECIDA (σ^2), OU COM AMOSTRA PEQUENA (N < 30), COM UMA CONFIANÇA ($1 - \alpha$) E UM ERRO DE AMOSTRAGEM DE NO MÁXIMO IGUAL A "E"

6.3.1. Para população infinita ou população finita e amostragem com reposição

Partindo da fórmula do erro de estimação ou de amostragem "e" temos que:

$$e = t_{\left(v;\,\frac{\alpha}{2}\right)} \cdot \frac{S_{(X)}}{\sqrt{n}} \quad \therefore e \cdot \sqrt{n} = t_{\left(v;\,\frac{\alpha}{2}\right)} \cdot \frac{S_{(X)}}{\sqrt{n}} \cdot \sqrt{n} \therefore e \cdot \sqrt{n} = t_{\left(v;\,\frac{\alpha}{2}\right)} \cdot S_{(X)} \therefore \frac{e \cdot \sqrt{n}}{e} = \frac{t_{\left(v;\,\frac{\alpha}{2}\right)} \cdot S_{(X)}}{e}$$

$$\sqrt{n} = \frac{t_{\left(v;\,\frac{\alpha}{2}\right)} \cdot S_{(X)}}{e} \therefore \left(\sqrt{n}\right)^2 = \left(\frac{t_{\left(v;\,\frac{\alpha}{2}\right)} \cdot S_{(X)}}{e}\right)^2 \therefore \left(\sqrt{n}\right)^2 = \left(\frac{t_{\left(v;\,\frac{\alpha}{2}\right)} \cdot S_{(X)}}{e}\right)^2$$

$$n = \frac{t_{\left(v;\,\frac{\alpha}{2}\right)} \cdot S^2_{(X)}}{e^2}$$

Portanto esse é o menor tamanho que deve ter a amostra "n" para se estimar a média "μ" de uma população aproximadamente normal, com variância (σ^2) desconhecida, usando pequenas amostras ($n < 30$), com um coeficiente de confiança $(1 - \alpha)\%$, e um erro de amostragem ou de estimação de no máximo igual a "e".

O valor $t_{\left(v;\, \frac{\alpha}{2}\right)}$ corresponde ao valor crítico ou tabelado sob a curva da distribuição t de "Student" ao nível de confiança $(1 - \alpha)$ desejado, $S^2_{(X)}$ é uma estimativa da variabilidade da população e "e" representa o erro de amostragem ou de estimação máximo permitido, o qual mede a precisão do processo de estimação.

Não conhecendo o desvio padrão da população, deveríamos substituí-lo por sua estimativa "S" e usar "t" de Student na expressão do tamanho da amostra "n". Ocorre, porém, que, não tendo ainda sido retirada a amostra, não dispomos, em geral, do valor de "S". Se não conhecemos nem ao menos uma limitação superior para "σ", a única solução será, então, colher uma amostra piloto de "n" elementos para, com base nela, obtermos uma estimativa para o desvio padrão que é o "S", empregando, a seguir, a expressão vista anteriormente $n = \frac{t_{\left(v;\, \frac{\alpha}{2}\right)} \cdot S^2_{(X)}}{e^2}$. No entanto se $n \leq n'$, a amostra piloto já terá sido suficiente para a estimação. Caso contrário deveremos retirar, ainda, da população, os elementos necessários à complementação do tamanho mínimo da amostra "n".

6.3.1.1. Exemplo

Qual é o tamanho da amostra necessária para se estimar a produtividade média diária de leite em litros por dia, de uma população de um rebanho caprino da raça morada nova considerada infinita, cuja amostra piloto de 10 animais forneceu um desvio padrão amostral "S" igual a 6 litros/dia, com 90% de confiança e uma precisão de 0,15 litros.

$$1 - \alpha = 0{,}90;\ \alpha = 0,10, v = n - 1 = 10 - 1 = 9;$$

$$t_{\left(9;\, \frac{0{,}10}{2}\right)} = 1{,}833;\ S = 6, e = 0{,}15.$$

$$n = \frac{(1{,}833)^2 .6^2}{(0{,}15)^2} = \frac{3{,}359889.36}{0{,}0225} = \frac{120{,}956004}{0{,}0225} = 5375{,}8224 = 5376 \text{ caprinos.}$$

6.3.1.2. Exemplo

Qual é o tamanho da amostra necessária para se estimar o consumo médio de Biodiesel em litros/hora de trabalho (aração por hectare), de tratores agrícolas da marca AGRIBRASIL, sabendo-se que a população é considerada infinita, e que uma amostra de 20 tratores forneceu uma média de 30 litros por hora e cujo desvio padrão S é igual a 8,0 l/h, com 95% de confiança e uma precisão de 0,09.

$$1 - \alpha = 0{,}95;\ \alpha = 0{,}05, v = n - 1 = 20 - 1 = 19;$$

$$t_{\left(19;\ \frac{0{,}05}{2}\right)} = 2{,}093;\ S = 8, e = 0{,}09.$$

$$n = \frac{(2{,}093)^2.8^2}{(0{,}09)^2} = \frac{4{,}380649.64}{0{,}0081} = \frac{280{,}361536}{0{,}0081} = 34612{,}53531 = 34613 \text{ tratores}$$

6.3.2. Para população finita e amostragem sem reposição

Se a população for finita ou se a amostragem usada for sem reposição, a expressão do erro deve ser multiplicada pelo fator de correção da população finita. Nestas condições o tamanho da amostra "n" será dado por:

$$n' = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

6.4. TAMANHO DA AMOSTRA PARA ESTIMAR A PROPORÇÃO RELATIVA "P" DE UMA POPULAÇÃO NORMAL INFINITA OU FINITA E AMOSTRAGEM COM REPOSIÇÃO E COM AMOSTRA GRANDE (N ≥ 30), COM UMA CONFIANÇA $(1 - \alpha)$ E UM ERRO DE AMOSTRAGEM DE NO MÁXIMO IGUAL A "E"

6.4.1. O dimensionamento amostral

$$e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{\hat{P}\left(1 - \hat{P}\right)}{n}} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{\hat{P}(\hat{q})}{n}} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sqrt{\hat{P}.\hat{q}}}{\sqrt{n}}$$

$$e\sqrt{n} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sqrt{\hat{P}.\hat{q}}}{\sqrt{n}}\sqrt{n} \therefore e\sqrt{n} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{P} \cdot \hat{q}} \therefore \frac{e\sqrt{n}}{e} = \frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{P} \cdot \hat{q}}}{e}$$

$$\sqrt{n} = \frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{P} \cdot \hat{q}}}{e} \therefore \left(\sqrt{n}\right)^2 = \left(\frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{P} \cdot \hat{q}}}{e}\right)^2 \therefore n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)} \cdot \hat{P}\hat{q}}{e^2}$$

$$n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)} \cdot \hat{P}\hat{q}}{e^2} = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{e^2} \cdot P(1 - P)$$

Portanto esse é o menor tamanho que deve ter a amostra "n" para se estimar a proporção relativa de sucessos "P" de uma população normal, usando grandes amostras ($n \geq 30$), com um coeficiente de confiança $(1 - \alpha)\%$, e um erro de amostragem ou de estimação de no máximo igual a "e".

O valor $Z_{\left(\frac{\alpha}{2}\right)}$ corresponde ao valor crítico ou tabelado sob a curva normal ao nível de confiança $(1 - \alpha)$ desejado, $\hat{P}$ é uma estimativa da proporção relativa de sucessos na população e "e" representa o erro de amostragem ou de estimação máximo permitido, o qual mede a precisão do processo de estimação.

A limitação à determinação do tamanho da amostra por meio da expressão de "n" mostrada anteriormente está em desconhecermos P e tampouco dispormos de sua estimativa P, pois a amostra ainda não foi retirada. Essa dificuldade pode ser resolvida através de uma mostra piloto, de forma semelhante ao caso descrito para a estimação da média de uma população "μ". Ou ainda podemos resolver o problema analisando-se o comportamento do fator $P(1-P)$ para $0 \leq P \leq 1$. Verifica-se facilmente que $P(1-P)$ é a expressão de uma parábola cujo ponto de máximo é $P = \frac{1}{2}$.

Se agora substituirmos, na expressão $n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{e^2} \cdot P(1-P)$, $P(1-P)$ pelo seu valor máximo, que é $\frac{1}{4}$, seguramente o tamanho da amostra obtido será suficiente para a estimação, qualquer que seja P. Isso equivale a considerar o seguinte:

$$n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{e^2} \cdot \frac{1}{4} = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{(2e)^2} = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{4e^2}$$

Pelo mesmo raciocínio, se sabemos que seguramente$P \leq P^{\prime} \leq 0{,}50$ ou $P \geq P^{`} \geq 0{,}50$, podemos usar o limitante P' ao invés de P, na expressão $n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{e^2} \cdot P(1-P)$, obtendo um tamanho de amostra suficiente, pois teremos então, $P(1-P) \leq P'(1-P')$.

Evidentemente, que usando-se a expressão $n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{e^2} \cdot \frac{1}{4} = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{(2e)^2} = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{4e^2}$, corre-se o risco de dimensionar uma amostra bem maior do que a realmente necessária. Isso ocorrerá se P' for, na realidade, próximo de 0 ou de 1. Se o custo envolvido for elevado e proporcional ao tamanho da amostra, será desejável evitar que tal fato ocorra, sendo mais prudente a tomada de uma amostra piloto. Inversamente, em muitos casos, é preferível, por simplificação, proceder mostrado antes, com base em uma limitação superior para o fator $P(1-P)$.

O erro relativo de estimação é definido como o erro absoluto dividido pelo valor do parâmetro a ser estimado.

Nestas circunstâncias se desejarmos estimar uma proporção populacional "P" com o erro relativo fixado (e_r) a expressão $n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{e^2} \cdot P(1-P)$ passa a ser $n = \frac{Z^2_{\left(\frac{\alpha}{2}\right)}}{e_r^2} \cdot \frac{(1-P)}{P}$, onde $e_r = \frac{e}{P}$ é o erro relativo.

6.4.2. Exemplo

Qual é o tamanho da amostra suficiente para estimarmos a proporção de frutos de melão considerados fora dos padrões de especificações para exportação e comercialização no mercado internacional, com precisão de $0{,}02$ e 95% de confiança, sabendo que essa proporção seguramente não é superior a $0{,}20$.

Sendo que

$$1 - \alpha = 0{,}95;\ \alpha = 0{,}05;\ P = 0,20, Z_{\left(\frac{0{,}05}{2}\right)} = 1{,}96$$

$$n = \frac{(1{,}96)^2}{(0{,}02)^2} 0{,}20(1-0{,}20) = \frac{(3{,}8416)}{(0{,}0004)} 0{,}16 = 1536{,}64 = 1537 \text{ frutos de melão}$$

7. EXERCÍCIOS DE APLICAÇÃO SOBRE TEORIA DA ESTIMAÇÃO

7.1. DEFINA E EXEMPLIFIQUE

i) Inferência estatística, estatística indutiva ou estatística analítica.

ii) Estatística descritiva, estatística dedutiva ou análise exploratória de dados (AED).

iii) Parâmetro.

iv) Estimador ou estatística amostral ou estatística.

v) Estimativa.

vi) Estimação por ponto.

vii) Estimação intervalar ou por intervalo.

viii) População ou universo estatísticos.

ix) Amostra.

x) População finita.

xi) População infinita.

xii) População real.

xiii) População hipotética.

xiv) População objeto de estudo.

xv) População amostrada.

7.2. QUAIS SÃO AS PROPRIEDADES DOS ESTIMADORES OU ESTATÍSTICAS AMOSTRAIS. DESCREVA E DÊ EXEMPLOS.

7.3. SEJA A SEGUINTE DISTRIBUIÇÃO DE FREQUÊNCIA, REPRESENTATIVA DOS DADOS DE UMA AMOSTRA DE CINQUENTA ELEMENTOS.

Tabela 2. Distribuição de frequências de uma variável quantitativa contínua com uma amostra de 50 valores.

Classes	f_i
10 – 20	3
20 – 30	9
30 – 40	15
40 – 50	10
50 – 60	8
60 – 70	5
Soma	50

i) Calcule a média e o desvio padrão.

ii) Construa um intervalo de confiança com 90% de probabilidade para a média de população.

iii) Interprete o intervalo encontrado.

7.4. UMA AMOSTRA DE TAMANHO COM 10 ELEMENTOS, OU SEJA, $n = 10$ TEM $\overline{X} = 110$ E $S = 10$, DETERMINAR OS INTERVALOS DE CONFIANÇA PARA A MÉDIA POPULACIONAL AO NÍVEL DE 90% E 95%.

7.5. QUAL O INTERVALO DE CONFIANÇA QUE CONTARÁ COM 90% A VERDADEIRA MÉDIA DE UMA POPULAÇÃO NORMAL QUE RESULTOU $\sum x_i = 700,8$ E $\sum x_i^2 = 23436,8$ DE UMA AMOSTRA DE 30 ELEMENTOS.

7.6. UMA CENTENA DE COMPONENTES FOI ENSAIADA E 93 DELES FUNCIONARAM MAIS DE 500 HORAS. DETERMINAR UM INTERVALO DE CONFIANÇA DE 95% DE CONFIANÇA PARA A PROPORÇÃO.

7.7. O FABRICANTE DE UM INSTRUMENTO DE MEDIDA DE PRECISÃO GARANTE QUE ELE INDICA MEDIDAS CORRETAS COM PRECISÃO TOLERÁVEL DE 2 UNIDADES. UM OBJETO FOI MEDIDO QUATRO VEZES COM ESTE INSTRUMENTO ENCONTRANDO-SE OS SEGUINTES RESULTADOS: 353, 351, 352, 355. DETERMINA UM INTERVALO DE CONFIANÇA COM 90% DE PROBABILIDADE PARA A MÉDIA POPULACIONAL.

7.8. O QUE SIGNIFICA?

i) Inferência estatística. Qual é sua função e importância para a sua área de conhecimento.

ii) Dedução

iii) Indução

7.9. UMA AMOSTRA AO ACASO DE BATATAS MOSTROU OS SEGUINTES PESOS (G/BATATA).

70	75	72	80	85	65	70	72
75	79	82	86	68	70	72	82
65	66	72	71	76	71	73	74
78							

i) Calcular a média da amostra.

ii) Calcular a variância e o desvio padrão da amostra.

iii) Calcular o coeficiente de variação.

iv) Calcular os limites de confiança da média da população, com 90% de probabilidade.

7.10. UMA AMOSTRA CASUAL DE 625 DONAS DE CASA REVELA QUE 70% DELAS PREFEREM A MARCA X DE DETERGENTE. CONSTRUIR UM INTERVALO DE CONFIANÇA PARA P, A PROPORÇÃO DA POPULAÇÃO DE DONAS DE CASA QUE PREFEREM A MARCA X, COM COEFICIENTE DE CONFIANÇA DE 95%.

7.11. UM FARMACOLOGISTA DESEJA SABER, EM MÉDIA, QUAL O TEMPO DE REAÇÃO A DETERMINADO ANALGÉSICO EM PACIENTES PORTADORES DE DETERMINADAS CARACTERÍSTICAS. SUPONHAMOS QUE EXPERIÊNCIAS PRÉVIAS LHE GARANTAM QUE O TEMPO, EM PROBLEMAS DESTE TIPO, TEM DISTRIBUIÇÃO NORMAL COM VARIÂNCIA IGUAL A 16 MIN^2. VAMOS INDICAR POR X O TEMPO DE REAÇÃO.

i) Suponhamos que aplicando o analgésico a 36 pacientes, tenha observado que a média de tempo de reação, em minutos, tenha sido $\overline{X} = 28$. Construir um intervalo de confiança de 90% para a verdadeira média da distribuição.

ii) Suponhamos que o farmacologista deseje que a probabilidade de que o valor a ser observado para X_n não difira de μ por mais do que 2 minutos seja 0,98. Para isto, quantos pacientes devem ele observar, isto é, qual deverá ser o tamanho n da amostra?

7.12. UM CONJUNTO DE 12 ANIMAIS DE EXPERIÊNCIA FOI ALIMENTADO COM UMA DIETA ESPECIAL DURANTE 3 SEMANAS E PRODUZIU-SE OS SEGUINTES AUMENTOS DE PESO.

30	22	32	26	24	40
34	36	32	33	28	30

i) Determinar o intervalo de confiança para a média ($\alpha = 1\%$ e $\alpha = 5\%$).

ii) Determinar o tamanho da amostra para aumentar a precisão de estimativa em 25%, para os 2 níveis de confiança, considerando a população como infinita.

7.13. AS MEDIDAS DOS DIÂMETROS DE UMA AMOSTRA ALEATÓRIA DE 200 ROLAMENTOS ESFÉRICOS PRODUZIDOS POR CERTA MÁQUINA DURANTE UMA SEMANA APRESENTAM A MÉDIA DE 0,824 POLEGADAS O DESVIO PADRÃO DE 0,042 POLEGADAS. DETERMINAR OS LIMITES DE CONFIANÇA DE:

i) 95%

ii) 99%

Para ao diâmetro médio de todos os rolamentos esféricos.

7.14. UMA AMOSTRA CONSTITUÍDA DE 12 MEDIDAS DA TENSÃO DE RUPTURA DE UM FIO DE ALGODÃO APRESENTOU A MÉDIA DE 7,38 KG E O DESVIO PADRÃO DE 1,24 KG. PEDE-SE:

i) Os limites de confiança de 95% para se estimar a média populacional.

ii) O novo intervalo de confiança para a tensão de ruptura real, diminuindo-se o risco da afirmativa em 5 %.

iii) O tamanho da amostra para que a precisão da estimativa seja aumentada de 30% da anterior, no nível de 5% de probabilidade. Considere a população como infinita.

7.15. SUPONHA QUE O DESVIO PADRÃO DA VIDA ÚTIL DE UMA DETERMINADA MARCA DE TUBO DE IMAGEM DE TV É CONHECIDA, E É IGUAL A $\sigma = 500$, MAS QUE A MÉDIA DA VIDA ÚTIL É DESCONHECIDA. SUPÕE-SE QUE A VIDA ÚTIL DOS TUBOS DE IMAGEM TEM UMA DISTRIBUIÇÃO APROXIMADAMENTE NORMAL. PARA UMA AMOSTRA DE $n = 15$, A MÉDIA DE VIDA ÚTIL É $\overline{X} = 8.900$ HORAS DE OPERAÇÃO. CONSTRUIR UM INTERVALO DE CONFIANÇA DE 95%, PARA A VERDADEIRA MÉDIA DE VIDA ÚTIL.

7.16. ENTRE 500 PESSOAS INQUIRIDAS A RESPEITO DE SUAS PREFERÊNCIAS ELEITORAIS, 260 MOSTRAM-SE FAVORÁVEIS AO CANDIDATO "HONESTO DA SILVA". CALCULAR UM INTERVALO DE CONFIANÇA AO NÍVEL DE 95% PARA A PERCENTAGEM DOS ELEITORES FAVORÁVEIS A ESSE CANDIDATO.

7.17. UMA AMOSTRA DE TAMANHO 10, DE UMA VARIÁVEL X NORMALMENTE DISTRIBUÍDA, FORNECEU UMA MÉDIA $\bar{X} = 20$ E UM DESVIO PADRÃO $S = 4$. ENCONTRAR OS LIMITES DE CONFIANÇA PARA A MÉDIA POPULACIONAL, AOS NÍVEIS DE CONFIANÇA DE 90 %; 95% E 99% DE PROBABILIDADE.

7.18. ESTIMAR, NO NÍVEL DE 95% DE PROBABILIDADE, O QI (QUOCIENTE DE INTELIGÊNCIA) DE UMA COMUNIDADE, ONDE UMA AMOSTRA DE 50 PESSOAS FORNECEU $\bar{X} = 100,8$ E $S = 12,3$.

7.19. ANOTARAM-SE OS PESOS DE TARTARUGAS CAPTURADAS EM UM LAGO COM O OBJETIVO DE SE ESTUDAR AS CONSEQUÊNCIAS DO TIPO DE MEIO AMBIENTE SOBRE O CRESCIMENTO DAS TARTARUGAS. TODOS OS ANIMAIS USADOS ERAM DA MESMA IDADE E FORAM MARCADOS ANTES DE SEREM COLOCADOS DE VOLTA NO LAGO. OS PESOS DAS 20 TARTARUGAS ($n = 20$) DO LAGO ESTÃO NA TABELA A SEGUIR:

Tabela 3. Distribuição de dos pesos em quilos de 20 tartarugas.

Peso – Tartarugas (kg)									
14,1	15,2	13,9	14,5	14,7	13,8	14,0	16,1	12,7	15,1
12,2	13,0	14,1	13,6	12,4	11,9	12,5	13,8	13,4	14,8

Use os dados para construir um intervalo de confiança para "μ" com 95% de probabilidade, e interprete o resultado.

7.20. PARA SE ESTIMAR A PRODUÇÃO TOTAL DE ARROZ DE 600 PROPRIEDADES RURAIS, ESCOLHEU-SE UMA AMOSTRA ALEATÓRIA SIMPLES DE TAMANHO 90, DA QUAL SE EXTRAIU UMA MÉDIA DE 180 TONELADAS E UM DESVIO PADRÃO DE 30 TONELADAS. DETERMINE OS LIMITES DE 95% DE CONFIANÇA PARA O TOTAL DA PRODUÇÃO DE 600 PROPRIEDADES.

7.21. UM CONJUNTO DE 12 COELHOS DA RAÇA NOLFOX, EM EXPERIÊNCIA, FOI ALIMENTADO COM CERTA DIETA ESPECIAL DURANTE TRÊS SEMANAS, VISANDO AUMENTAREM SEUS PESOS. APÓS A EXPERIÊNCIA VERIFICOU-SE OS SEGUINTES AUMENTOS DE PESOS, EM GRAMAS: 300, 220, 320, 260, 240, 490,

350, 350, 320, 330, 280 E 300. ENCONTRE OS LIMITES DE CONFIANÇA DE 95% PARA μ (MÉDIA POPULACIONAL) E INTERPRETE O RESULTADO.

APÊNDICE 1

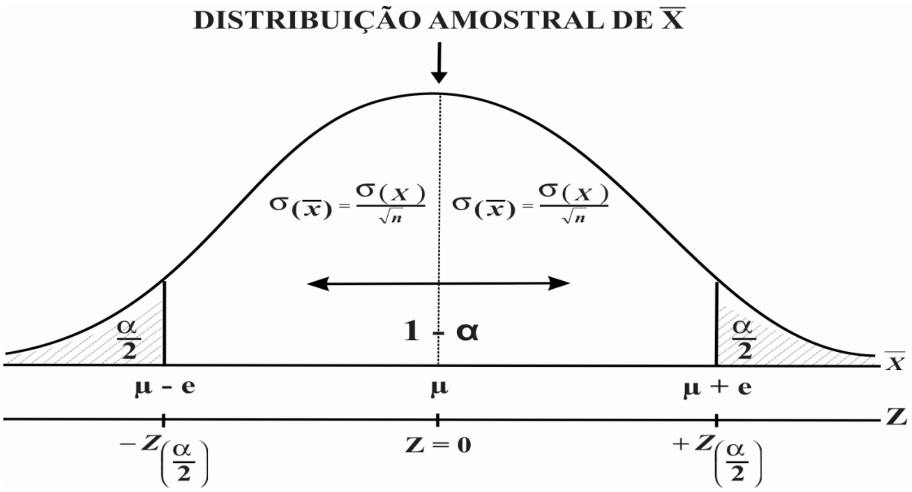
EXERCÍCIOS RESOLVIDOS. ESTIMAÇÃO POR INTERVALO DE CONFIANÇA

ESTIMAÇÃO POR INTERVALO: TIPOS DE INTERVALOS DE CONFIANÇA PARA A MÉDIA (μ), PARA A PROPORÇÃO (P), EM POPULAÇÕES INFINITAS (N/N<0,05) OU FINITAS (N/N≤0,05) E AMOSTRAGEM COM REPOSIÇÃO E TAMBÉM EM POPULAÇÕES FINITAS (N/N≤0,05) E AMOSTRGAEM SEM REPOSIÇÃO

INTERVALO DE CONFIANÇA PARA A MÉDIA POPULACIONAL (μ) DE UMA POPULAÇÃO INFINITA, NORMAL COM VARIÂNCIA (Σ^2) CONHECIDA OU NÃO, E COM A UTILIZAÇÃO DE AMOSTRA GRANDE [N ≥ 30]

O intervalo de confiança

Figura 10. Gráfico representativo da distribuição amostral da média $\overline{X}$ utilizado na construção do intervalo de confiança para a média populacional μ, quando se utiliza grandes amostras ($n \geq 30$).



$\mu - e \leq \overline{X}$ e $\overline{X} \leq \mu + e$

$\mu \leq \overline{X} + e$ e $\overline{X} - e \leq \mu$

Ou seja, temos que:

$$\overline{X} - e \leq \mu \leq \overline{X} + e$$

e assim o intervalo é dado por:

$$P\left[\overline{X} - e \leq \mu \leq \overline{X} + e\right] = 1 - \alpha$$

$$Z_{\left(\frac{\alpha}{2}\right)} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \therefore Z_{\left(\frac{\alpha}{2}\right)} = \frac{(\mu - e) - \mu}{\frac{\sigma}{\sqrt{n}}} \therefore Z_{\left(\frac{\alpha}{2}\right)} = \frac{e}{\frac{\sigma}{\sqrt{n}}} \therefore e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\overline{X} - e \leq \mu \leq \overline{X} + e$$

$$\overline{X} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}$$

$$P\left[\underbrace{\overline{X} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{LIC}} \leq \underbrace{\mu}_{\text{PARÂMETRO}} \leq \underbrace{\overline{X} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{LSC}}\right] = \underbrace{1 - \alpha}_{\text{COEFICIENTE DE CONFIANÇA}}$$

$$\text{I.C.: } \overline{X} \pm e \therefore \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \text{ ou } \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}$$

Sendo assim o intervalo de confiança é determinado conforme a expressão abaixo.

$$P\left[\overline{X} - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

onde $\overline{X}$ e σ são, respectivamente a média da amostra e o desvio padrão da população com distribuição normal e variância conhecida, e assim o intervalo de confiança de $100(1 - \alpha)\%$ para a media populacional μ é dada pela equação acima, onde $Z_{\left(\frac{\alpha}{2}\right)}$ é o valor critico ou tabelado sob a curva da distribuição normal padrão.

EXERCÍCIO DE APLICAÇÃO

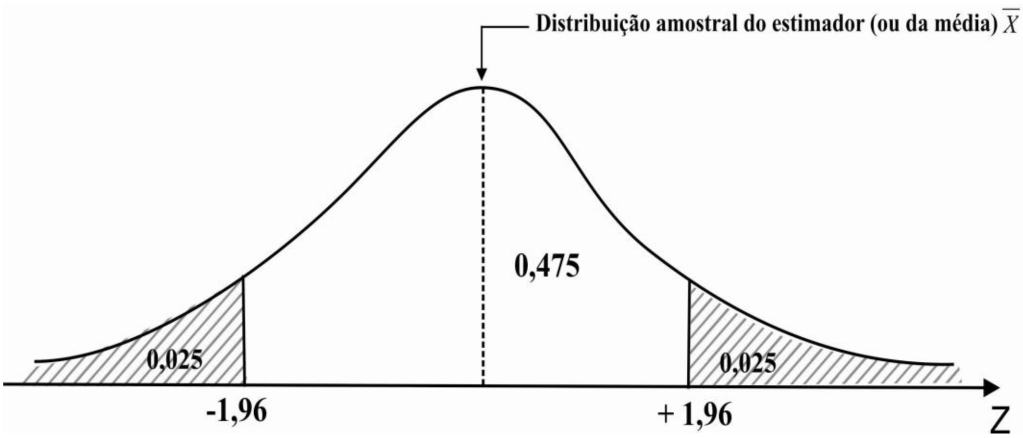
Sabe-se de experimentos anteriores que a variância da produção leiteira de um rebanho bovino da raça holandês preto e branco criado na região de Natal-RN, durante o período de 1970 a 2012, é de $4{,}0\text{Kg}^2$. Para avaliar a produção leiteira média atual, um criador(comprador) retirou uma amostra aleatória simples de $n = 36$ vacas, obtendo uma estimativa por ponto $\overline{X} = 5{,}4$.

a) Determine estimativas por intervalos para a produção leiteira média populacional atual (μ), aos níveis de confiança de:

(a.1) 0,95

(a.2) 0,99.

Resolução do item a.1:



$\sigma^2 = 4{,}0\text{Kg}^2$, ou seja, a variância populacional σ^2 é conhecida, sendo assim a população é normal. $n = 36$, $\overline{X} = 5{,}4$ e $\sigma = \sqrt{4} = 2$ Kg, $1 - \alpha = 0{,}95$, $\alpha = 0{,}05$. Pela tabela da normal temos que o valor crítico ou tabelado sob a curva normal padrão é dado por: $Z_{\left(\frac{0,05}{2}\right)} = \pm 1{,}96$. Sendo assim o intervalo de confiança é obtido como mostrado abaixo.

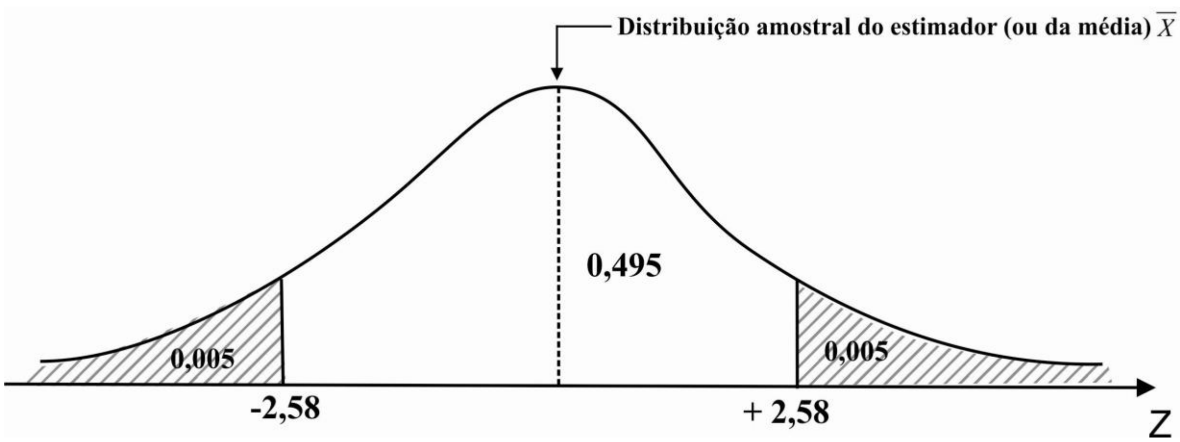
$$I.C.[\mu]_{0,95} \to \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \therefore\ 5{,}4 \pm 1{,}96 \frac{2}{\sqrt{36}} \therefore 5{,}4 \pm 1{,}96 \cdot \frac{2}{6} \therefore\ 5{,}4 \pm 0{,}6533 \therefore 4{,}7467 \leq \mu \leq 6{,}0533\ \therefore\ P[4{,}75 \leq \mu \leq 6{,}05] = 0{,}95$$

Interpretação: O intervalo de confiança de limites 4,75 e 6,05 deve conter a verdadeira produção leiteira média populacional (μ) atual expressa em quilogramas, do rebanho bovino de vacas leiteiras da raça holandês preto e branco, com 95% de probabilidade de confiança e uma precisão ou erro de estimação ou de amostragem de no máximo 0,6533 kg.

Resolução do item a.2:

Neste caso agora o nível de confiança é 99% de probabilidade assim temos que:

$1 - \alpha = 0{,}99, \alpha = 0{,}01$. Pela tabela da distribuição normal temos que o valor crítico ou tabelado sob a curva normal padrão é dado por: $Z_{\left(\frac{0,01}{2}\right)} = \pm 2{,}58$.



Sendo assim o intervalo de confiança é obtido como mostrado a seguir:

$$\text{I. C. } [\mu]_{0,99} \rightarrow \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \therefore\ 5,4 \pm 2,58 \frac{2}{\sqrt{36}} \therefore 5,4 \pm 1,58 \cdot \frac{2}{6} \therefore\ 5,4 \pm 0,86 \therefore 4,54 \leq \mu \leq 6,26 \ \therefore\ P[4,54 \leq \mu \leq 6,26] = 0,99$$

Interpretação: O intervalo de confiança de limites 4,54 e 6,26, deve conter a verdadeira produção leiteira média populacional (μ) atual expressa em quilogramas, do rebanho bovino de vacas leiteiras da raça holandês preto e branco, com 99% de probabilidade de confiança. Em outras palavras, se adotarmos a mesma metodologia, e construirmos 100 intervalos deste tipo para a variável produção leiteira média populacional (μ) atual expressa em quilogramas, do rebanho bovino de vacas leiteiras da raça holandês preto e branco, em Média 99 intervalos conterão o valor desta produção média populacional (μ), e apenas 1 deles não conterão este valor desconhecido do parâmetro μ.

b) Dimensionamento amostral: Qual é o tamanho da amostra necessária para que se obtenha um intervalo de confiança com precisão $e = 0,5$ Kg, ao nível $\alpha = 0,05$?

$n = ?$; $e = 0,5$; $1 - \alpha = 0,95$; $\alpha = 0,05$ e assim o valor tabelado de Z é dado por: $Z_{\left(\frac{0,05}{2}\right)} = \pm 1,96$.

$$\text{I. C. } [\mu]_{0,95} \rightarrow \overline{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \therefore \overline{X} \pm e$$

$$e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \ \therefore\ e\sqrt{n} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}}\sqrt{n} \ \therefore\ e\sqrt{n} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma \ \therefore\ \frac{e\sqrt{n}}{e} = \frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma}{e}$$

$$\sqrt{n} = \frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma}{e} \ \therefore\ \sqrt{n}^2 = \left(\frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma}{e}\right)^2 \ \therefore\ n = \left(\frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sigma}{e}\right)^2$$

$$n = \frac{[1,96]^2 . 2^2}{[0,5]^2} = \frac{15,3664}{0,25} = 61,47 \cong 62 \text{ vacas}$$

Portanto o tamanho da amostra "n" deve ser de aproximadamente 62 animais. Ou seja, esse é o menor tamanho "n" que deve ter a amostra de vacas para ser usada na estimação da produção leiteira média populacional atual "μ" em quilogramas, em uma população infinita, normal, com variância (σ^2) conhecida de 4Kg^2), com um erro de estimação ou de amostragem "e" de no máximo $0,5\text{kg}$, e uma confiança de 95% de probabilidade.

INTERVALO DE CONFIANÇA PARA A MÉDIA (μ) DE UMA POPULAÇÃO INFINITA, APROXIMADAMENTE NORMAL COM VARIÂNCIA (σ^2) DESCONHECIDA E AMOSTRA PEQUENA (N < 30)

O intervalo de confiança

$$P\left[\overline{X} - t_{\left(v,\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\left(v,\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

lembrado que a variável "t" possui $v = n - 1$ graus de liberdade.

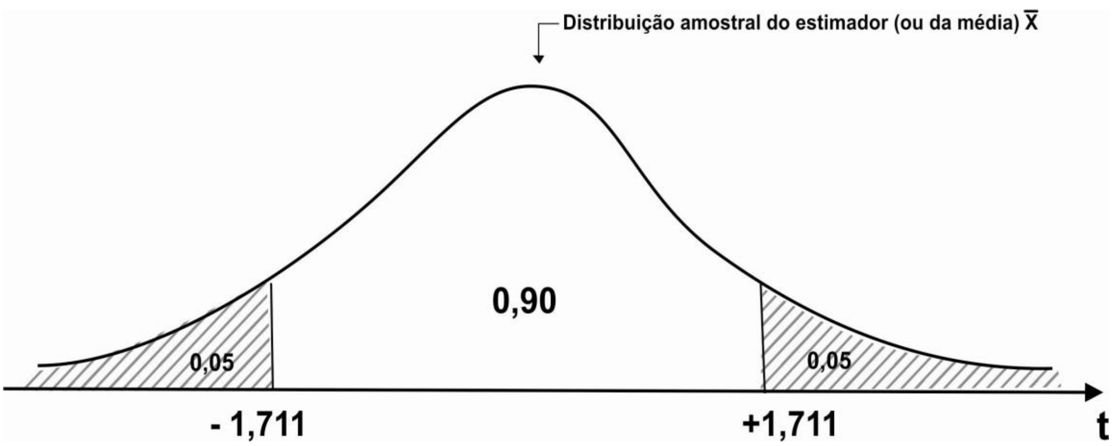
Onde $\overline{X}$ e S são, respectivamente a média e o desvio padrão de uma amostra aleatória de uma população com distribuição normal e variância desconhecida, e assim o intervalo de confiança de $100(1-\alpha)\%$ para a média populacional μ é dada pela equação acima, onde $t_{\left(v,\frac{\alpha}{2}\right)}$ é o valor critico ou tabelado sob a curva da distribuição t de Student, com "v" graus de liberdade.

Exercício de aplicação

Seja "X" a variável aleatória que representa a taxa normal de colesterol no plasma sanguíneo de coelhos da raça Norfolk, criados em uma granja no município de Mossoró – RN, durante o ano de 2012. Suponhamos que com base em uma amostra casual simples de 25 animais normais, um pesquisador obteve a média $\overline{X} = 198mg/100ml$ de plasma e o desvio padrão $S = 30\text{mg}/100\text{ml}$ de plasma.

Obtenha com base nessa amostra, o intervalo de confiança com 90% de probabilidade para μ, isto é para a taxa média populacional de colesterol dos coelhos da raça Norfolk.

Solução:



$$v = n - 1 = 25 - 1 = 24 \text{ g.l.}$$

$$\alpha = 10\% = 0{,}10$$

$$\frac{\alpha}{2} = \frac{10\%}{2} = 0{,}05 = 5\%$$

$$t_{\left(24,\frac{0,10}{2}\right)} = \pm 1{,}711$$

$$198 - 1{,}711 \cdot \frac{30}{\sqrt{25}} \leq \mu \leq 198 + 1{,}711 \cdot \frac{30}{\sqrt{25}}$$

$$P\left[198 - 1{,}711 \cdot \frac{30}{\sqrt{25}} \leq \mu \leq 198 + 1{,}711 \cdot \frac{30}{\sqrt{25}}\right] = 90\%$$

$$P[198 - 10{,}26 \leq \mu \leq 198 + 10{,}26] = 90\%$$

$$P[187{,}74 \leq \mu \leq 208{,}26] = 90\%$$

Interpretação: Este é o intervalo de 90% de confiança para a média populacional "μ", obtido com base na amostra. Então, 90% dos intervalos calculados dessa forma, para amostras casuais simples de 25 indivíduos, conterão o parâmetro que é a taxa normal média populacional (μ) de colesterol em $\mathrm{mg/100ml}$ de plasma no plasma sanguíneo de coelhos da raça Norfolk. De outra forma: O intervalo de confiança de limites $187{,}74$ e $208{,}26$, contém a verdadeira taxa média populacional (μ) de colesterol em $\mathrm{mg/100ml}$ de plasma dos coelhos da raça Norfolk com 90% de probabilidade de confiança, e uma precisão ou um erro de amostragem ou de estimação de no máximo $10{,}26\mathrm{mg/100ml}$.

b) Dimensionamento amostral: Para um erro de amostragem $e = 9{,}00$, a amostra satisfaz?

$$\mathrm{I.C.}[\mu] \quad \rightarrow \overline{X} \pm t_{\left(v,\frac{\alpha}{2}\right)} \cdot \frac{S}{\sqrt{n}} \therefore \overline{X} \pm e$$

$$t_{\left(24,\frac{0,10}{2}\right)} = \pm 1{,}711$$

$$e = t_{\left(24,\frac{0,10}{2}\right)} \cdot \frac{S}{\sqrt{n}} \therefore e\sqrt{n} = t_{\left(24,\frac{0,10}{2}\right)} \cdot \frac{S}{\sqrt{n}}\sqrt{n} \therefore e\sqrt{n} = t_{\left(24,\frac{0,10}{2}\right)} \cdot S \therefore \frac{e\sqrt{n}}{e} = \frac{t_{\left(24,\frac{0,10}{2}\right)} \cdot S}{e}$$

$$\sqrt{n} = \frac{t_{\left(24, \frac{0,10}{2}\right)} \cdot S}{e} \;\therefore\; \sqrt{n}^2 = \left(\frac{t_{\left(24, \frac{0,10}{2}\right)} \cdot S}{e}\right)^2 \;\therefore\; n = \left(\frac{t_{\left(24, \frac{0,10}{2}\right)} \cdot S}{e}\right)^2$$

$$n = \left(\frac{1{,}711 \cdot 30}{9}\right)^2 = 32{,}49 \cong 32{,}50 \cong 33 \text{ indivíduos}$$

Conclusão: Para um erro de amostragem de $9{,}00\ \mathrm{mg/100ml}$ de plasma será necessária uma amostra de 33 indivíduos, portanto a amostra com 25 indivíduos não satisfaz, ou seja, são necessários mais 8 elementos.

Portanto o tamanho da amostra "n" deve ser de aproximadamente 33 coelhos. Ou seja, esse é o menor tamanho "n" que deve ter a amostra de coelhos para ser usada na estimação da taxa normal média populacional "μ" em $\mathrm{mg/100ml}$ de plasma, de colesterol em uma população infinita, com 90% de confiança e um erro de amostragem ou de estimação de no máximo $9{,}00\ \mathrm{mg/100ml}$ de plasma.

INTERVALO DE CONFIANÇA PARA A PROPORÇÃO (P) DE UMA POPULAÇÃO INFINITA, NORMAL, E COM AMOSTRA GRANDE (N > 30)

O intervalo de confiança

$$P\left[f - Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{f(1-f)}{n}} \leq p \leq f + Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{f(1-f)}{n}}\right] = 1 - \alpha$$

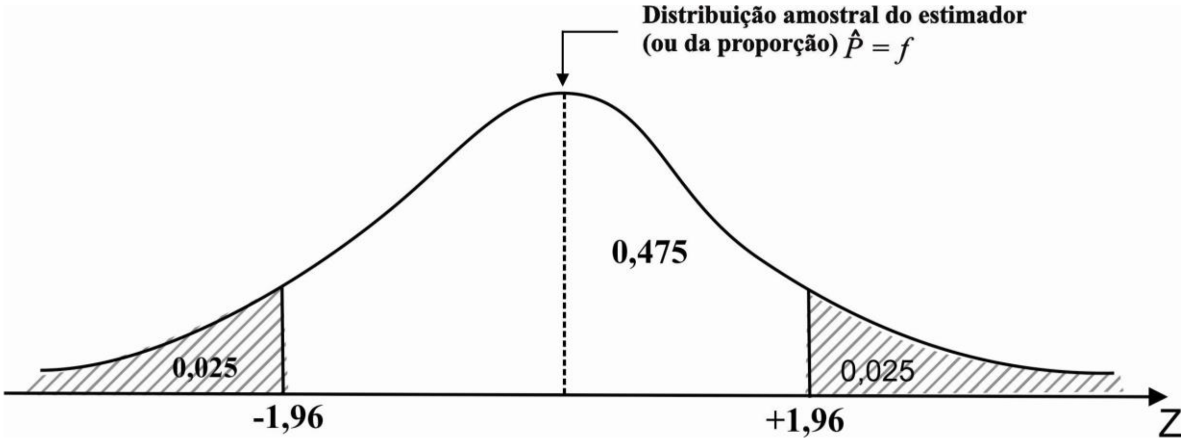
onde $f = \hat{p} = \frac{x}{n}$ é a proporção (frequência relativa) de eventos sucesso na amostra; x é o número (frequência absoluta) de eventos sucesso na amostra; e n é o número de elementos da amostra (tamanho da amostra).

Exercício de aplicação

Para testar uma vacina contra a febre aftosa, um médico veterinário da UFERZAM em Mossoró-RN, selecionou ao acaso em 2012, uma amostra aleatória (ou casual) simples de 80 bovinos da raça Gir, e, observou após um determinado período de 20 dias, que 60 desses animais foram imunizados.

Calcular a estimativa intervalar para a verdadeira proporção populacional "P" de bovinos da raça GIR imunizados com a vacina. Use um coeficiente de confiança de 95% de probabilidade.

Solução:



$$n = 80; X = 60\text{, onde,}$$

$$f = \hat{p} = \frac{x}{n} = \frac{60}{80} = 0{,}75 = 75\% \text{ e}$$

$$\hat{q} = 1 - \hat{p} = 1 - 0{,}75 = 0{,}25 = 25\%.$$

$$\alpha = 0{,}05 = 5\%$$

$$1 - \alpha = 1 - 0{,}05 = 0{,}95$$

$$Z_{\left(\frac{0,05}{2}\right)} = \pm 1{,}96$$

$$0{,}75 \pm 1{,}96 \cdot \sqrt{\frac{0{,}75(1 - 0{,}75)}{80}} = 0{,}75 \pm 0{,}095$$

$$P[0{,}655 \leq p \leq 0{,}845] = 0{,}95$$

$$P[65{,}5 \leq p \leq 84{,}5] = 95\%$$

Interpretação: O intervalo de confiança de limites $65{,}5\%$ e $84{,}5\%$ contém a verdadeira proporção populacional de bovinos da Raça GIR, imunizados através da vacina contra a febre aftosa, com 95% de confiança. Isto é se forem construídos 100 intervalos deste tipo usando o mesmo tamanho amostral ($n = 80$) para a mesma variável resposta e adotando a mesma metodologia, espera-se que em média 95 destes intervalos contenham a verdadeira proporção populacional (parâmetro) de bovinos da raça Gir imunizados contra a febre aftosa.

b) Dimensionamento amostral: Se a precisão for aumentada em 5%, qual é o número de animais da raça Gir que deve ser selecionado?

O erro amostral ou de estimação é dado por: $e = 0{,}095$;

O novo erro amostral ou de estimação passa a ser: $e' = 0{,}095 - (0{,}05) \cdot (0{,}095) = 0{,}09025$.

$$e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \therefore e = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \therefore e\sqrt{n} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\sqrt{n}$$

$$e\sqrt{n} = Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{p}(1-\hat{p})} \therefore \frac{e\sqrt{n}}{e} = \frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{p}(1-\hat{p})}}{e} \therefore \sqrt{n} = \frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{p}(1-\hat{p})}}{e}$$

$$\sqrt{n}^2 = \left(\frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{p}(1-\hat{p})}}{e}\right)^2 \therefore n = \left(\frac{Z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\hat{p}(1-\hat{p})}}{e}\right)^2$$

$$n = \left(\frac{1{,}96 \cdot \sqrt{0{,}75(1-0{,}75)}}{0{,}09025}\right)^2 \cong 88{,}43 \cong 89 \text{ animais}$$

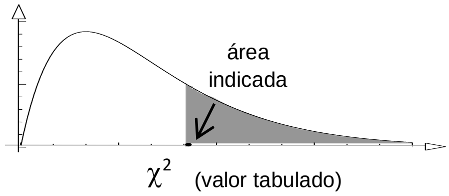
Portanto se a precisão for aumentada em 5% serão necessários 89 bovinos para a pesquisa, ou seja, este é o menor tamanho que deve ter a amostra de animais, para se estimar a verdadeira proporção populacional de bovinos imunizados contra a febre aftosa, com 95% de confiança ou de probabilidade, e um erro de amostragem ou de estimação de no máximo $0{,}09025$, ou 9% aproximadamente.

APÊNDICE 2

Tabela 4. Áreas ou probabilidades sob a curva normal padrão ente $Z = 0,00$ e um valor positivo de Z. Para os valores das probabilidades entre os valores negativos de Z e $Z = 0,00$, as áreas são obtidas por simetria.

Z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4965	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4983	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,49	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500

Tabela 5. Distribuição de qui-quadrado para diversos níveis de significância.



	Área na cauda superior							
gl	0,999	0,9975	0,995	0,99	0,975	0,95	0,9	0,75
1	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,10
2	0,00	0,01	0,01	0,02	0,05	0,10	0,21	0,58
3	0,02	0,04	0,07	0,11	0,22	0,35	0,58	1,21
4	0,09	0,14	0,21	0,30	0,48	0,71	1,06	1,92
5	0,21	0,31	0,41	0,55	0,83	1,15	1,61	2,67
6	0,38	0,53	0,68	0,87	1,24	1,64	2,20	3,45
7	0,60	0,79	0,99	1,24	1,69	2,17	2,83	4,25
8	0,86	1,10	1,34	1,65	2,18	2,73	3,49	5,07
9	1,15	1,45	1,73	2,09	2,70	3,33	4,17	5,90
10	1,48	1,83	2,16	2,56	3,25	3,94	4,87	6,74
11	1,83	2,23	2,60	3,05	3,82	4,57	5,58	7,58
12	2,21	2,66	3,07	3,57	4,40	5,23	6,30	8,44
13	2,62	3,11	3,57	4,11	5,01	5,89	7,04	9,30
14	3,04	3,58	4,07	4,66	5,63	6,57	7,79	10,17
15	3,48	4,07	4,60	5,23	6,26	7,26	8,55	11,04
16	3,94	4,57	5,14	5,81	6,91	7,96	9,31	11,91
17	4,42	5,09	5,70	6,41	7,56	8,67	10,09	12,79
18	4,90	5,62	6,26	7,01	8,23	9,39	10,86	13,68
19	5,41	6,17	6,84	7,63	8,91	10,12	11,65	14,56
20	5,92	6,72	7,43	8,26	9,59	10,85	12,44	15,45
21	6,45	7,29	8,03	8,90	10,28	11,59	13,24	16,34
22	6,98	7,86	8,64	9,54	10,98	12,34	14,04	17,24
23	7,53	8,45	9,26	10,20	11,69	13,09	14,85	18,14
24	8,08	9,04	9,89	10,86	12,40	13,85	15,66	19,04
25	8,65	9,65	10,52	11,52	13,12	14,61	16,47	19,94
26	9,22	10,26	11,16	12,20	13,84	15,38	17,29	20,84
27	9,80	10,87	11,81	12,88	14,57	16,15	18,11	21,75
28	10,39	11,50	12,46	13,56	15,31	16,93	18,94	22,66
29	10,99	12,13	13,12	14,26	16,05	17,71	19,77	23,57
30	11,59	12,76	13,79	14,95	16,79	18,49	20,60	24,48
35	14,69	16,03	17,19	18,51	20,57	22,47	24,80	29,05
40	17,92	19,42	20,71	22,16	24,43	26,51	29,05	33,66
45	21,25	22,90	24,31	25,90	28,37	30,61	33,35	38,29
50	24,67	26,46	27,99	29,71	32,36	34,76	37,69	42,94
100	61,92	64,86	67,33	70,06	74,22	77,93	82,36	90,13

Tabela 6. Distribuição de qui-quadrado para diversos níveis de significância.

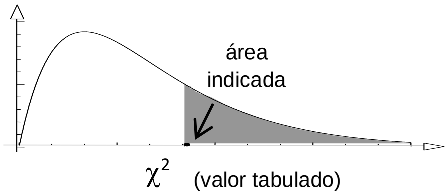
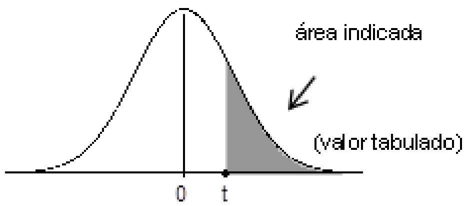


Tabela 7. Tabela da distribuição t de "Student" com valores críticos ou tabelados unilaterais para diversos níveis de significância.

	Área na cauda superior							
gl	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001
1	1,32	2,71	3,84	5,02	6,63	7,88	9,14	10,83
2	2,77	4,61	5,99	7,38	9,21	10,60	11,98	13,82
3	4,11	6,25	7,81	9,35	11,34	12,84	14,32	16,27
4	5,39	7,78	9,49	11,14	13,28	14,86	16,42	18,47
5	6,63	9,24	11,07	12,83	15,09	16,75	18,39	20,51
6	7,84	10,64	12,59	14,45	16,81	18,55	20,25	22,46
7	9,04	12,02	14,07	16,01	18,48	20,28	22,04	24,32
8	10,22	13,36	15,51	17,53	20,09	21,95	23,77	26,12
9	11,39	14,68	16,92	19,02	21,67	23,59	25,46	27,88
10	12,55	15,99	18,31	20,48	23,21	25,19	27,11	29,59
11	13,70	17,28	19,68	21,92	24,73	26,76	28,73	31,26
12	14,85	18,55	21,03	23,34	26,22	28,30	30,32	32,91
13	15,98	19,81	22,36	24,74	27,69	29,82	31,88	34,53
14	17,12	21,06	23,68	26,12	29,14	31,32	33,43	36,12
15	18,25	22,31	25,00	27,49	30,58	32,80	34,95	37,70
16	19,37	23,54	26,30	28,85	32,00	34,27	36,46	39,25
17	20,49	24,77	27,59	30,19	33,41	35,72	37,95	40,79
18	21,60	25,99	28,87	31,53	34,81	37,16	39,42	42,31
19	22,72	27,20	30,14	32,85	36,19	38,58	40,88	43,82
20	23,83	28,41	31,41	34,17	37,57	40,00	42,34	45,31
21	24,93	29,62	32,67	35,48	38,93	41,40	43,77	46,80
22	26,04	30,81	33,92	36,78	40,29	42,80	45,20	48,27
23	27,14	32,01	35,17	38,08	41,64	44,18	46,62	49,73
24	28,24	33,20	36,42	39,36	42,98	45,56	48,03	51,18
25	29,34	34,38	37,65	40,65	44,31	46,93	49,44	52,62
26	30,43	35,56	38,89	41,92	45,64	48,29	50,83	54,05
27	31,53	36,74	40,11	43,19	46,96	49,65	52,22	55,48
28	32,62	37,92	41,34	44,46	48,28	50,99	53,59	56,89
29	33,71	39,09	42,56	45,72	49,59	52,34	54,97	58,30
30	34,80	40,26	43,77	46,98	50,89	53,67	56,33	59,70
35	40,22	46,06	49,80	53,20	57,34	60,27	63,08	66,62
40	45,62	51,81	55,76	59,34	63,69	66,77	69,70	73,40
45	50,98	57,51	61,66	65,41	69,96	73,17	76,22	80,08
50	56,33	63,17	67,50	71,42	76,15	79,49	82,66	86,66
100	109,1	118,5	124,3	129,6	135,8	140,2	144,3	149,4

Tabela 8. Valores críticos ou tabelados bilaterais sob a distribuição de t de Student.



	Área na cauda superior								
gl	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,660
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
35	0,682	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
45	0,680	1,301	1,679	2,014	2,412	2,690	2,952	3,281	3,520
50	0,679	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
z	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

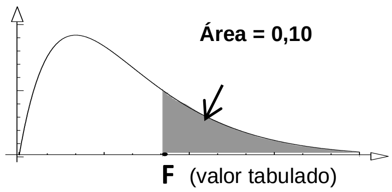
Tabela 9. Valores da distribuição t de “Student” com valores críticos ou tabelados bilaterais para diferentes números de graus de liberdade e diversos níveis de significância.

g.l. P	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
01	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
02	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
03	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,541	12,924
04	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
05	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
06	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
07	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,365	3,499	5,408
08	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
09	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,726
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,856	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,856	1,056	1,313	1,701	2,048	2,467	2,763	3,674

Continua na próxima página

g.l. P	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Tabela 10. Distribuição "F" de Snedecor-Fisher, mostrando os valores críticos ou tabelados para o nível de significância de 0,10.



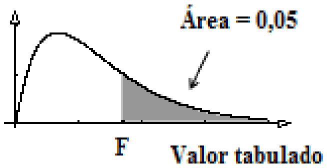
g.l. den.	graus de liberdade no numerador									
	1	2	3	4	5	6	7	8	9	10
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86

Continua na próxima página

g.l. den.	graus de liberdade no numerador									
	1	2	3	4	5	6	7	8	9	10
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82
35	2,85	2,46	2,25	2,11	2,02	1,95	1,90	1,85	1,82	1,79
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76
45	2,82	2,42	2,21	2,07	1,98	1,91	1,85	1,81	1,77	1,74
50	2,81	2,41	2,20	2,06	1,97	1,90	1,84	1,80	1,76	1,73
100	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,69	1,66

g.l. den. = Grau de liberdade do denominador.

Tabela 11. Distribuição "F" de Snedecor-Fisher, mostrando valores os críticos ou tabelados para o nível de significância de 0,05.



g.l. den.	Graus de liberdade no numerador									
	1	2	3	4	5	6	7	8	9	10
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24

Continua na próxima página

g.l. den.	Graus de liberdade no numerador									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
45	4,06	3,20	2,81	2,58	2,42	2,31	2,22	2,15	2,10	2,05
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93

g.l. den. = Grau de liberdade do denominador.

A tabela seguinte, que fornece os valores críticos, teóricos ou tabelados de "Z" para ambos os testes unilateral e bilateral e também para intervalos de confiança unilateral e bilateral em vários níveis de significância α, pode ser muito útil como referência a todos os leitores interessados em realizar estudos de inferência estatística. Os valores críticos de "Z" para outros níveis de significância são determinados através do uso das tabelas de áreas sob a curva normal ou gaussiana padrão ou reduzida.

Tabela 12. Valores críticos de "Z" para alguns níveis de significância (α) mais comuns na construção de intervalos de confiança e aplicação de testes de hipóteses.

Nível de significância α	0,10	0,05	0,01	0,005	0,002
Valores críticos (tabelados) de Z para testes de hipóteses e intervalos de confiança unilaterais	-1,280 OU 1,280	-1,645 OU 1,645	-2,330 OU 2,330	-2,580 OU 2,580	-2,880 OU 2,880
Valores críticos (tabelados) de Z para testes de hipóteses e intervalos de confiança bilaterais	-1,645 E 1,645	-1,960 E 1,960	-2,580 E 2,580	-2,810 E 2,810	-3,080 E 3,080

APÊNDICE 3

$P_1 - P_2$ $\binom{n_1 \geq 30\ e}{n_2 \geq 30}$	———	BERNOULLI	$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}} \cap N(0,1)$	$\left](\hat{P}_1 - \hat{P}_2) - Z_{\left(\frac{\alpha}{2}\right)}\sigma_{(\hat{P}_1-\hat{P}_2)}; (\hat{P}_1 - \hat{P}_2) + Z_{\left(\frac{\alpha}{2}\right)}\sigma_{(\hat{P}_1-\hat{P}_2)}\right[$ onde $\sigma_{(\hat{P}_1-\hat{P}_2)} = \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$
$\frac{\sigma_1^2}{\sigma_2^2}$	———	NORMAIS	$F = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \cap F_{(v_1=n_1-1;v_2=n_2-1)}$	$\left]\frac{S_1^2}{S_2^2}\frac{1}{F_{\left(v_1=n_1-1;v_2=n_2-1;\frac{\alpha}{2}\right)}}; \frac{S_1^2}{S_2^2}F_{\left(v_2=n_2-1;v_1=n_1-1;\frac{\alpha}{2}\right)}\right[$

PARÂMETRO	$\boldsymbol{\sigma^2}$ CONHECIDO	TIPO DE POPULAÇÃO	VARIÁVEL PIVOTAL	INTERVALO DE CONFIANÇA $(\mathbf{1-\alpha})$ 100%
μ	SIM	NORMAL	$Z = \dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \cap N(0{,}1)$	$\left]\overline{X} - Z_{\left(\frac{\alpha}{2}\right)}\dfrac{\sigma}{\sqrt{n}}; \overline{X} + Z_{\left(\frac{\alpha}{2}\right)}\dfrac{\sigma}{\sqrt{n}}\right[$
	NÃO	NORMAL	$t = \dfrac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \cap t_{\left(v=n-1;\frac{\alpha}{2}\right)}$	$\left]\overline{X} - t_{\left(v,\frac{\alpha}{2}\right)}\dfrac{S}{\sqrt{n}}; \overline{X} + t_{\left(v,\frac{\alpha}{2}\right)}\dfrac{S}{\sqrt{n}}\right[$
P (n ≥ 30)	______	BERNOULLI	$Z = \dfrac{\hat{P}-P}{\sqrt{\frac{P(1-P)}{n}}} \cap N(0{,}1)$	$\left]\hat{P} - Z_{\left(\frac{\alpha}{2}\right)}\sqrt{\dfrac{\hat{P}(1-P)}{n}}; \hat{P} + Z_{\left(\frac{\alpha}{2}\right)}\sqrt{\dfrac{\hat{P}(1-P)}{n}}\right[$
σ^2	______	NORMAL	$\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} \cap \chi_{(v=n-1)}$	$\left]\dfrac{(n-1)S^2}{\chi^2_{\left(v=n-1;1-\frac{\alpha}{2}\right)}}; \dfrac{(n-1)S^2}{\chi^2_{\left(v=n-1;\frac{\alpha}{2}\right)}}\right[$
PARÂMETRO	$\boldsymbol{\sigma_1^2\ e\ \sigma_2^2}$ **CONHECIDOS?**	**TIPO DE POPULAÇÃO**	**VARIÁVEL PIVOTAL**	**INTERVALO DE CONFIANÇA** $(\mathbf{1-\alpha})$ **100%**
$(\mu_1-\mu_2)$	SIM	NORMAIS	$Z = \dfrac{\left(\overline{X}_1-\overline{X}_2\right)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \cap N(0{,}1)$	$\left]\left(\overline{X}_1-\overline{X}_2\right) - Z_{\left(\frac{\alpha}{2}\right)}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}; \left(\overline{X}_1-\overline{X}_2\right) + Z_{\left(\frac{\alpha}{2}\right)}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}\right[$

	NÃO	NORMAIS	$t = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \cap t_{(v = n_1 + n_2 - 2)}$	$\left]\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2) - t_{\left(v = n_1 + n_2 - 2; \frac{\alpha}{2}\right)} S_P; \left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2) + t_{\left(v = n_1 + n_2 - 2; \frac{\alpha}{2}\right)} S_P; \right[$ onde $S_P = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
	NÃO	NORMAIS	$t = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \cap t_{\left(\left[v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}\right]\right)}$	$\left]\left(\overline{X}_1 - \overline{X}_2\right) - t_{\left(v; \frac{\alpha}{2}\right)}\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}; \left(\overline{X}_1 - \overline{X}_2\right) + t_{\left(v; \frac{\alpha}{2}\right)}\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}\right[$

APÊNDICE 4

ALFABETO GREGO

NOME DA LETRA	SÍMBOLOS	
	MAIÚSCULA	MINÚSCULA
Alfa	Α	α
Beta	Β	β
Gama	Γ	γ
Delta	Δ	δ
Épsilon	Ε	ε
Zeta	Ζ	ζ
Eta	Η	η
Téta	Θ	θ
Iota	Ι	ι
Capa	Κ	ϰ
Lambda	Λ	λ
Mu(mi)	Μ	μ
Nu(ni)	Ν	ν
Csi	Ξ	ξ
Omicron	Ο	ο
Pi	Π	π
Ró	Ρ	ϱ
Sigma	Σ	σ
Tau	Τ	τ
Upsilon(ipsilon)	ϒ	υ
Fi	Φ	φ
Chi(qui)	Χ	χ
Psi	Ψ	ψ
Omega	Ω	ω

APÊNDICE 5

INTERVALOS DE CONFIANÇA CONSTRUÍDOS E APLICADOS NA ANÁLISE ESTATÍSTICA INFERENCIAL FREQUENTISTA EM ENSAIOS PLANEJADOS E CONDUZIDOS NA ESTATÍSTICA EXPERIMENTAL

As informações aqui prestadas aos pesquisadores das diversas áreas do conhecimento, tais como, biologia, ecologia, medicina, medicina veterinária, zootecnia, administração, ciências contábeis, agronomia, ciências físicas, ciências sociais, engenharias dentre outras, apresenta-se como um conjunto de intervalos de confiança para estimação de comparações de contrastes de médias e de totais de tratamentos ou fatores experimentais e estimação de parâmetros populacionais desconhecidos, que podem levar o investigador a tomar decisões importantes na sua pesquisa científica. As comparações de médias referem-se a estudos experimentais, cujo objetivo principal ´e verificar se existem diferenças significativas entre os tratamentos utilizados na pesquisa experimental. Utiliza-se o teste F de Fischer-Snedecor no processo da análise de variância (ANOVA), para testar a hipótese $H_0: \mu_1 = \cdots = \mu_I = \mu$, isto é, a hipótese de nulidade dos efeitos das médias dos I tratamentos, contra a hipótese alternativa $H_1: \mu_i \neq \mu_j$, para algum par i, j $(i, j = 1, 2, \cdots, I;\ i \neq j)$. Sempre que o resultado do teste F for significativo, existem fortes indícios de que pelo menos um dos contrastes dos I tratamentos seja significativo ao nível de significância α adotado. Os dados obtidos nos experimentos podem ser submetidos a vários tipos de análise estatística, sendo que uma das mais importantes é a análise de variância.

Após a execução da análise de variância (ANOVA), e constatado que houve efeito significativo nos tratamentos, aplicam-se os testes de comparações de médias, e os intervalos de confiança para contrastes entre médias a fim de verificar quais médias diferem entre si e em quanto oscila essa diferença. Apresenta-se a seguir então, os deferentes tipos de intervalos de confiança usados para realizar inferência estatística: LSD-Diferenãa mínima significativa; Método de Bonferroni, onde utiliza-se a desigualdade de Bonferroni; Método de Tukey, método que compara as médias, duas a duas; Método de Duncan, realizado de forma sequencial e depende do número de médias que se comparam; Newman Keuls, Dunnet, SNK e por último o m´método de Scheffé, que se propõe a construir para comparação quaisquer contrastes entre médias de tratamentos.

Método baseado na distribuição teórica de probabilidade t de "Student" (teste t de Student).

O Teste t de Student é também utilizado para testar ou verificar hipótese referentes a contrastes de médias tais como:

$$H_0: y(m) = 0 \textit{ versus } H_1: y(m) \neq 0$$

Em que, $m = 1,2,\ldots,(I-1)$ contrastes dados por: $y(m) = \sum_{i=1}^{I} r_i c_{i,k} \mu_i$, sendo que para ser contraste a equação anterior tem que assumir a seguinte restrição: $y(m) = \sum_{i=1}^{I} r_i c_{i,k} = 0$

O pesquisador pode montar ou estabelecer a priori na fase de planejamento experimental tantos contrastes quanto são os graus de liberdade para tratamentos $I-1$.

Para o teste t de Student a ortogonalidade dos contrastes também deve ser verificada a priori, ou seja, a soma dos produtos dos coeficientes das médias correspondentes dos contrastes deve ser nula. A ortogonalidade se traduz na independência entre os contrastes, isto é, a variação de um deles é independente da variação no outro, portanto a covariância entre eles deve ser nula. Os contrastes podem ser estabelecidos com os totais de tratamentos ou com as médias de tratamentos, ou seja, $cov[y(m), y(m')] = 0$ para $m \neq m'$.

Sendo assim o intervalo de confiança para o contraste usando médias de tratamentos é determinado conforme mostrado a seguir:

$$\hat{y}(m) = \sum_{i=1}^{I} c_{i,k}\, \bar{y}_i \pm \sqrt{\sum_{i=1}^{I} c_{i,k}^2 \frac{QM_{erro}}{J}}$$

onde o quadrado médio do erro (QM_{erro}) é a estimativa da variância do erro experimental ou resíduo e J o número de repetições do ensaio, no caso particular se forem todas com o mesmo número de repetições então $r_i = J$.

LSD - CONTRASTE DA DIFERENÇA MÍNIMA SIGNIFICATIVA (FISHER, 1935)

Os testes estatísticos para comparações de médias mais utilizados baseiam-se na distribuição teórica de probabilidade t de Student. Cujo objetivo é o de comparar os efeitos das médias de I tratamentos e neste caso o pesquisador tem interesse em contrastar quaisquer hipóteses da forma:

$$H_0: \mu_i = \mu_j \textit{ versus } H_1: \mu_i \neq \mu_j$$
$$\forall i = j;\ i,j = 1,2,\ldots,I$$

A técnica clássica e mais comum para efetuar as comparações múltiplas é o procedimento LSD (*Least Significant Difference*), que deve ser aplicado quando a hipótese de igualdade de todas as médias dos tratamentos tiver sido rejeitada, ou melhor o método LSD de Fisher é usado na análise de variância (ANOVA) para construir intervalos de confiança de todas as diferenças de pares entre médias de níveis de fatores ou tratamentos, controlando a taxa de erro individual para um nível de significância especificado.

O Intervalo de confiança, por exemplo, para estimar a diferença entre médias de contrastes é construído conforme a equação a seguir.

$$\mathrm{IC}_{[(1-\alpha)100\%]}(\mu_i - \mu_j): \left[(\bar{y}_i - \bar{y}_j) - t_{\left(v,\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}\hat{S}_R^2};\ (\bar{y}_i - \bar{y}_j) + t_{\left(v,\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}\hat{S}_R^2}\right]$$

ou,

$$\mathrm{IC}_{[(1-\alpha)100\%]}(\mu_i - \mu_j): [(\bar{y}_i - \bar{y}_j) - LSD;\ (\bar{y}_i - \bar{y}_j) + LSD]$$

onde, a quantidade LSD, é denominada diferença mínima significativa, $\hat{S}_R^2$ é o quadrado médio do resíduo obtido a partir da análise de variância, n_i e n_j é o número de observações de cada média e v é o número de gruas de liberdade associado a $\hat{S}_R^2$, sendo assim os intervalos com $(1-\alpha)100\%$ de confiança para $\mu_i - \mu_j$ serão obtidos com base na equação apresentada a seguir: sendo, n_i e n_j o número de observações correspondente a cada média observada no experimento;$t_{\left(v,\frac{\alpha}{2}\right)}$ o valor crítico ou tabelado da distribuição t-Student com v graus de liberdade, cuja área a sua direita corresponde a $\frac{\alpha}{2}$. Caso o experimento seja balanceado, isto é, se cada média dos tratamentos for calculada com o mesmo número de unidades experimentais n, então tem-se que: $\mathrm{LSD} = t_{\left(v,\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}\hat{S}_R^2}$.

O procedimento LSD é simples de utilizar. Pode ser aplicado tanto em modelos equilibrados como em modelos não equilibrados. Além disso permite a construção de intervalos de confiança para diferença entre médias, estes intervalos são obtidos conforme a seguinte equação:

$$\mathrm{IC}_{[(1-\alpha)100\%]}(\bar{y}_i - \bar{y}_j): [(\bar{y}_i - \bar{y}_j) - LSD;\ (\bar{y}_i - \bar{y}_j) + LSD]$$

MÉTODO DE BONFERRONI

Bonferroni propôs um ajuste para o nível de significância do teste LSD de Fisher, garantindo assim um nível de significância para experimento (*experimentwise Type I error rate* - EER) abaixo do escolhido. O procedimento é denominado de teste de Bonferroni.

Neste método se fixa um nível de significância α que se divide entre cada uma das comparações consideradas e se utiliza a desigualdade de Bonferroni:

$$P\left(\bigcup_{m=1}^{M} A_m\right) \leq \sum_{m=1}^{M} P(A_m)$$

$$P(A_1 \cup A_2 \cup \ldots A_M) \leq P(A_1) + P(A_2) + \ldots + P(A_M)$$

Constroem-se estimação por intervalo para as $M = \binom{I}{2}$ possíveis comparações, cada uma ao nível de significância $\alpha^* = \frac{\alpha}{M}$, isso dá origem a M intervalos de confiança contendo cada uma das possíveis diferenças $\mu_i - \mu_j$ com probabilidade $1 - \alpha^*$ nomeando C_m o m-ésimo intervalo, tem-se que: $P(\mu_{1m} - \mu_{2m} \in C_m) = 1 - \alpha^*$, $m = 1, 2, \dots, M$, sendo μ_{1m} e μ_{2m} a primeira e a segunda média da correspondente comparação, onde supõem-se que $1 \leq 1_m < 2_m \leq I$. Aplicando a desigualdade de Bonferroni $\mathrm{P}(\bigcup_{m=1}^{M} A_m) \leq \sum_{m=1}^{M} P(A_m)$ tem-se que:

$$P\left(\bigcap_{m=1}^{M} C_m\right) = 1 - \mathrm{P}\left(\bigcup_{m=1}^{M} C_m\right) \geq 1 - \sum_{m=1}^{M} P(\bar{C}_m) = 1 - \sum_{m=1}^{M} \alpha^*$$

onde $\bar{C}_m$ é o símbolo para o complemento do intervalo α.

Com base nestes resultados e tendo como objetivo principal garantir um nível de significância α para um conjunto de M comparações por pares, ou um nível de confiança $1 - \alpha$ para o conjunto de intervalos, basta tomar: $\alpha^* = \frac{\alpha}{M}$. Sendo assim, é obvio que a probabilidade de que todos os intervalos de confiança $\bar{C}_m$ contenham as correspondentes diferenças entre médias, será de pelo menos $1 - \alpha$. Sendo assim os intervalos de confianças serão construídos conforme a equação a seguir:

$$(\bar{y}_{1m} - \bar{y}_{2m}) \pm t_{\left(v, \frac{\alpha}{2M}\right)} \cdot \sqrt{\hat{S}_R^2 \frac{1}{n_{1m}} + \frac{1}{n_{2m}}}$$

onde, $\bar{y}_{1m}, \bar{y}_{2m}$ e n_{1m}, n_{2m}, são as médias e os tamanhos amostrais, correspondentes a m-ésima comparação.

MÉTODO DE TUKEY (HONESTLY-SIGNIFICNAT-DIFFERENCE). DIFERENÇA HONESTAMENTE SIGNIFICANTE (TUKEY, 1953)

Este método que compara todo e qualquer contraste entre duas médias de tratamentos, isto é, duas a duas, permite ao pesquisador construir intervalos conjuntos com $1 - \alpha$ de confiança tanto para modelos balanceados como não balanceado para todas as possíveis comparações entre duas médias associadas aos I níveis dos tratamentos, isto é, o número de possíveis comparações entre duas médias é dado por $C_2^I = \binom{I}{2} = \frac{I!}{(I-2)! \cdot 2!}$. O coeficiente de confiança conjunto $1 - \alpha$ indica que de cada 100 amostras em $100(1 - \alpha)\%$ delas, cada um dos intervalos contém a sua correspondente diferença entre médias. Portanto, o coeficiente de confiança de cada um dos intervalos será de pelo menos $100(1 - \alpha)\%$.

Para construir estes intervalos considera-se os desvios, $(\bar{y}_1 - \mu_1)$, $(\bar{y}_2 - \mu_2)$, ..., $(\bar{y}_I - \mu_I)$, esses desvios são variáveis aleatórias independentes com média 0, e variância $\frac{\sigma^2}{n}$. Além disso $\frac{\hat{S}_R^2}{n}$ é um estimador de $\frac{\sigma^2}{n}$, que é independente de tais desvios. Sendo assim o intervalo de confiança para a diferença entre médias é construído conforme a equação a seguir:

$$\left[(\bar{y}_i - \bar{y}_j) - \text{HSD} \leq (\mu_i - \mu_j) \leq (\bar{y}_i - \bar{y}_j) + \text{HSD}\right]$$

onde $\text{HSD} = q_{[I,v,\alpha]} \cdot \sqrt{\frac{\hat{S}_R^2}{n}}$, sendo q a amplitude total estudentizada, cujo valor é encontrado nas tabelas, v o número de graus de liberdade associados ao erro experimental, I é o número de tratamentos, n o número de observações ou repetições associada a cada tratamento, α o nível de significância do teste e $\hat{S}_R^2$ o estimador do erro ou variância experimental.

No caso do modelo não ser balanceado aplica-se o ajuste de Tukey-Cramer, que consiste em trocar n pela média harmônica, n_h do tamanho dos grupos, ou seja, $n_h = 2/\sum_{i=1}^{2} n_i$.

Se o intervalo de confiança acima não contém o valor zero, conclui-se que as médias μ_i e μ_j diferem significativamente ou estatisticamente entre si.

MÉTODO DE DUNCAN

Este método é realizado de forma sequencial e depende do número de médias de tratamentos que se comparam no contraste. Utiliza a mesma estatística ou a distribuição da amplitude estudentizada HSD que é a amplitude estudentizada do método de Tukey, ou seja, como já foi exposto a aplicação do teste de hipótese de Duncan é sequencial, isto é, não se utiliza o mesmo valor crítico para todas as diferenças entre médias como é feito no teste de Tukey, mas um valor crítico que dependa do número de médias ordenadas compreendidas entre as médias que se comparam, sendo estas colocadas previamente em ordem crescente. O método tanto serve para modelos balanceados como não balanceados (DUNCAN, 1955).

O intervalo de confiança para diferença entre as médias é construído conforme a equação a seguir:

$$\left[(\bar{y}_i - \bar{y}_j) - \text{R}_\text{p} \leq (\mu_i - \mu_j) \leq (\bar{y}_i - \bar{y}_j) + \text{R}_\text{p}\right]$$

$$\text{R}_\text{p} = q_{[\alpha_p,p,v]} \cdot \sqrt{\frac{\hat{S}_R^2}{n}}$$

onde $p = 2, \dots, I$ ou então para modelos desbalanceados tem-se que,

$$\text{R}_\text{p} = q_{[\alpha_p,p,v]} \cdot \sqrt{\frac{\hat{S}_R^2}{n_h}}$$

No caso do modelos não ser balanceado aplica-se o ajuste de Tukey-Cramer, que consiste em trocar n pela média harmônica, n_h do tamanho dos grupos, ou seja, $n_h = 2/\sum_{i=1}^{2} n_i$, onde $q_{[\alpha_p,p,v]}$ é o valor crítico da amplitude estudentizada, encontrado nas tabelas, baseado na comparação da média maior, e a média menor de p médias; $\hat{S}_R^2$ é o quadrado médio do resíduo ou erro experimental com v graus de liberdade; α_p é o nível de significância conjunto relativo as p médias consecutivas, quer dizer é a probabilidade de rejeitar erroneamente pelo menos uma das $p-1$ comparações independentes associadas as médias consideradas. Tal nível de significância está associado ao nível de significância α de uma comparação individual através da equação $\alpha_p = 1-(1-\alpha)^{p-1}$.

MÉTODO DE SCHEFFÉ

Este procedimento é usado para comparar qualquer contraste entre médias. O teste de Scheffé pode ser usado quando as comparações são selecionadas depois de olhar para os dados e incluem os contrastes, que nem todos são aos pares. Scheffé provou que a probabilidade do erro do tipo I para cada um dos testes não ultrapassa o valor α. É um teste semelhante ao teste de Tukey, mas difere daquele por utilizar os valores das tabelas da distribuição de F-Fisher-Snedecor, e não os valores da amplitude estudentizda (q). Mas é um teste superior ao teste de Tukey para comparar mais de duas médias, e inferior neste caso de duas médias apenas como é o procedimento de Tukey (SHEFFÉ,1935).

Este método propõe realizar qualquer contraste entre duas ou mais médias de tratamentos. Este procedimento não requer que o modelo seja balanceado. Senão veja.

Seja uma família de contrastes da seguinte forma: $\psi = \sum_i l_i\mu_i$. O objetivo deste procedimento é capacitar o pesquisador a tomar a decisão para cada um destes contrastes entre as hipóteses $H_0\colon \psi = 0$ versus $H_0\colon \psi \neq 0$. O método de Scheffé está baseado na construção de intervalos de confiança para todos os possíveis contrastes da forma $H_0\colon \psi = 0$ versus $H_0\colon \psi \neq 0$.

Esses intervalos tem um nível de confiança simultâneo $1-\alpha$, isto é, a probabilidade de que todos os intervalos de confiança sejam iguais simultaneamente é igual a $1-\alpha$. Scheffé mostrou que esses intervalos de confiança tem a seguinte expressão:

$$\psi \pm S(\hat{\psi}) \cdot \sqrt{(I-1)F_{[\alpha,I-1,v]}}$$

onde $\psi = \sum_i l_i\mu_i$ é um estimador de ψ, μ_i é a média observada no tratamento i, l_i são constantes reais, tais que $\sum_i l_i = 0$ e $S(\hat{\psi}) = \sqrt{\hat{S}_R^2 \frac{\sum_i l_i^2}{n_i}}$, $\hat{S}_R^2$ sendo a variação residual com v graus de liberdade, ou seja, a estimação do erro experimental como regra de decisão tem-se que rejeitar a hipótese de nulidade H_0 sobre um contraste ψ se o intervalo de confiança para ψ,

$$\left[\hat{\psi} - S(\hat{\psi}) \cdot \sqrt{(I-1)F_{[\alpha, I-1, v]}}\ ;\ \hat{\psi} + S(\hat{\psi}) \cdot \sqrt{(I-1)F_{[\alpha, I-1, v]}}\right]$$

não incluir o zero, isto é, se $|\hat{\psi}| \geq S(\hat{\psi}) \cdot \sqrt{(I-1)F_{[\alpha, I-1, v]}}$.

Na aplicação prática das pesquisas científicas, procede-se da seguinte maneira: calcula-se $S = S(\hat{\psi}) \cdot \sqrt{(I-1)F_{[\alpha, I-1, v]}}$, onde I é o número de tratamentos ou fatores experimentais comparados no experimento e F é o valor tabelado ou crítico sob a distribuição teórica de probabilidade F de Fisher – Snedecor para um número de graus de liberdade de tratamentos, e do resíduo ou erro experimental, com nível de significância alfa (α).

STUDENT-NEWMAN-KEULS OU SNK

Neste método ou teste de hipótese a posteriori, o objetivo é contornar as limitações do teste t de Student, quando mais de dois tratamentos estão envolvidos no experimento. O teste SNK (Student-Newman-Keuls) procura ajustar o valor de t de acordo com as distâncias entre as médias ordenadas dos tratamentos ou fatores experimentais. Em uma relação decrescente com I médias, duas delas $\bar{X}_1$ e $\bar{X}_2$ apresentarão significância se o valor calculado em módulo para t_{snk} for maior ou igual ao valor tabelado para o nível de significância α com v graus de liberdade (gl) para o resíduo ou erro experimental e uma distância i entre as médias dado por $i = p + 2$ (sendo p o número de médias existente entre as duas médias comparadas na relação decrescente). O teste SNK é mais rigoroso, apresenta menor diferença entre as médias do que o teste t de Student.

Newman (1939) e Keuls (1952) criaram um procedimento ou teste de hipótese que consiste em ajustar um valor do teste t para mais de dois tratamentos. O SNK é muito semelhante ao teste de Tukey, exceto que ele analisa as diferenças em termos de camadas. Para uma camada, o teste dá o mesmo resultado que seria obtido através do teste de Tukey. Também se pode afirmar que este teste de poder intermediário entre os testes de Tukey e Duncan. Como no teste de Duncan, o teste SNK tem por objetivo fazer aplicação de múltiplas amplitudes na comparação das médias. Sendo que a diferença básica entre estes testes é o uso da amplitude q que é um valor tabelado utilizada no teste de Tukey, em substituição ao valor Z também tabelado nas mesmas condições do teste de Duncan. Ou seja, utiliza a tabela do teste de Tukey onde n passa a ser o número de médias abrangidas pelo contraste ou comparação.

Procedimentos para aplicar o Teste de Student-Neuman-Keuls (SNK): i) colocar as médias dos I tratamentos em ordem decrescente; ii) calcular a estimativa do contraste da forma $\hat{Y}_I = \bar{Y}_{maior} - \bar{Y}_{menor}$, em que este contraste abrange todos os I tratamentos; iii) calcular o diferença mínima significativa (DMS) dada pela seguinte equação, $DMS = q\sqrt{\frac{QM\ Erro\ Experimental}{r}}$ em que, q é o valor tabelado em função do número de tratamentos I envolvidos no contraste e do número de graus de liberdade do erro experimental ou resíduo; r é o número de repetições de todos os tratamentos; iv) comparar $\hat{Y}_1$ com a DMS: se $\hat{Y}_1 < DMS$, o teste não é significativo, indicando que as duas médias que entraram no contraste $\hat{Y}_1$ não diferem, então, une-se as médias abrangidas pelo contraste por uma barra contínua e não pode-se comparar médias dentro da barra. Se $\hat{Y}_1 \geq DMS$, o teste é significativo, indicando que as duas médias que entraram no contraste $\hat{Y}_1$ diferem. Então, se passa a testar contrastes que abrangem um número imediatamente inferior de médias $I - 1$; v) voltar ao passo ii obtendo $\hat{Y}_2$ e vi) fazer isso até não ser mais necessário realizar comparações. Os valores tabelados da amplitude estudentizada q_i diminuem com o aumento do número de graus de liberdade, mas aumentam com a distância entre as médias, corrigindo os excessos de erro tipo I.

Sendo assim pode-se construir o intervalo de confiança conforme se mostra a seguir:

$$\hat{Y}_I = \bar{Y}_{maior} - \bar{Y}_{menor} \pm q\sqrt{\frac{QM\ Erro\ Experimental}{r}}$$

DUNNETT

Dunnett (1955) propôs um procedimento para comparações múltiplas onde o interesse é o de comparar um grupo particular (muitas vezes o chamado grupo de controle, testemunha ou placebo) com cada um dos grupos restantes. A significância deste teste implicará apenas na conclusão de que os grupos tratados apresentam diferença com o grupo controle. Isto é, em situações frequentes o interesse do pesquisador não está baseado em descobrir diferenças entre tratamentos, mas em descobrir se existem tratamentos estatisticamente diferentes, ou seja, melhores ou piores, que o tratamento considerado testemunha, padrão, controle ou placebo. Este conjunto de comparações não forma um conjunto independente, o que fez com que Dunnett (1955) apresentasse uma alternativa a este procedimento que requer um único valor para julgar a significância ou não das diferenças existentes. O procedimento pode ser aplicado como um teste unilateral, ou seja, melhor ou pior que a testemunha, ou bilateral diferente do controle. A significância deste teste implicará apenas na conclusão de que os grupos tratados apresentam diferenças do grupo testemunha.

Os procedimentos para aplicar o Teste de Dunnett são mostrados a seguir:

calcular a diferença mínima significativa (DMS), dada por:

$$DMS = D\sqrt{2\frac{QM\ Erro\ Experimental}{r}}$$

em que, D ou $t_{[\alpha,p,n]}$ é o valor tabelado da amplitude estudentizada, obtida na tabela a um nível de α% de significância, para ou em função de p médias de tratamentos envolvidos e do número de $v = n$ graus de liberdade do erro experimental ou resíduo, r ou J é o número de repetições respectivamente de todos os tratamentos;

calcular as estimativas dos contrastes, dados por:

$$\hat{Y}_1 = \hat{m}_1 - \hat{m}_{controle}$$
$$\hat{Y}_2 = \hat{m}_2 - \hat{m}_{controle}$$
$$\vdots$$
$$\hat{Y}_i = \hat{m}_i - \hat{m}_{controle}$$

comparar o valor absoluto de cada estimativa do controle como valor da DMS.

Se $\hat{Y} \geq DMS$ o teste é significativo indicando que o grupo tratado difere do grupo controle, e se $\hat{Y} < DMS$ então o grupo controle não difere do grupo tratado.

No caso de tratamentos com número diferentes de repetições ou modelos desbalanceados, a fórmula do cálculo do teste de Dunnett deve ser modificada, e tem-se então que o valor do erro padrão da diferença entre a média do tratamento testemunha com r_0 e qualquer outro tratamento com r repetições é dado por:

$$S_d = \sqrt{QM\ Erro\ Experimental\left(\frac{1}{r} + \frac{1}{r_0}\right)}$$

Sendo assim o intervalo de confiança para um contraste particular é dado por:

$$\hat{Y} \pm D\sqrt{2\frac{QM\ Erro\ Experimental}{r}}$$

ou então

$$\hat{Y} \pm \sqrt{QM\ Erro\ Experimental\left(\frac{1}{r} + \frac{1}{r_0}\right)}$$

MÉTODO DE SCOTT KNOTT

Este teste ou método foi proposto por Scott e Knott (1974), e é diferente aos outros testes apresentados neste trabalho, o qual é um procedimento que compara as médias dos tratamentos ou tem sua base teórica por conglomerados e sua significância é analisada por meio da distribuição teórica de probabilidade de Qui-quadrado (χ^2) . A maior vantagem de sua aplicação é decorrente do fato de que nenhuma média pode pertencer a mais de um agrupamento, como ocorre nos outros testes de comparações múltiplas, ou seja, o teste determina a constituição de grupos disjuntos, sempre que for encontrada significância na aplicação do teste F da análise de variância (ANOVA). Este procedimento não apresenta uma fórmula básica para obtenção de um ou mais valores para comparações das médias. Na verdade antes de mais nada estabelece os grupos de médias em função da variabilidade entre os grupos dessas médias.

O teste Scott e Knott (1974) utiliza a razão de verossimilhança para atestar a significância de que I tratamentos podem ser divididos em grupos que maximizem a soma de quadrados entre grupos.

Se por exemplo o ensaio tiver 4 tratamentos, A, B, C e D, o processo consiste em determinar uma partição, em três grupos, que maximize a soma de quadrados. Nesse caso são possíveis $2^I - 1$ grupos. Para minimizar a formação dos grupos, basta ordenar as médias dos tratamentos. Nesse caso o número de partições possíveis passa a ser obtido por $n - 1$. Uma vez que o pesquisador tenha ordenado as médias de tratamentos, procede-se da seguinte maneira, fazendo inicialmente o número de tratamentos envolvidos no grupo de médias considerado (g) igual ao o número total de tratamentos (I).

Sendo assim os procedimentos para se aplicar o Teste de Scott-Knott, se dá conforme o seguinte algoritmo:

i) ordenar as I médias;

ii) fazer o número de médias no grupo g igual ao número total de I tratamentos, criando $I - 1$ partições de grupos de médias da seguinte maneira:

Partição 1: grupo do tratamento 1 e o grupo formado pelos tratamentos $2,3,\ldots,I$.

Partição 2: grupo dos tratamentos 1 e 2 e o grupo dos tratamentos $3,4,\ldots,I$, e assim sucessivamente até a

Partição I-1: grupo formado pelos tratamentos $1,2,,\ldots,I-1$ e o grupo do tratamento I.

iii) determinar $g - 1$ partições com dois grupos;

iv) para cada partição calcular a soma de quadrado entre grupos de cada partição, usando a seguinte equação: $SQ_{grupos} = \frac{(T_1)^2}{i_1} + \frac{(T_2)^2}{i_2} - \frac{(T_1+T_2)^2}{i}$, onde T_1 e T_2 são os totais das médias componentes do grupo

1 e do grupo 2 e $(T_1 + T_2)$ o total geral do grupo, i_1 e i_2 o número de médias de tratamento em cada grupo na partição. O máximo valor da soma de quadrados de grupos é chamado de β_0^2;

v) calcular o estimador de máxima verossimilhança de $\sigma_{\bar{y}}^2$

$$\sigma_0^2 = \frac{1}{g+v}\left[\sum_{i=1}^{k_1}(\bar{y}_i - \bar{y})^2 + vS_{\bar{y}}^2\right] = \frac{SQ_{médias} + v\left(\frac{QME}{r}\right)}{i+v}$$

onde r é o número de repetições de cada tratamento, v é o número de graus de liberdade do erro ou resíduo experimental, $\bar{y}$ é a média geral de todos os tratamentos envolvidos na comparação e $S_{\bar{y}}^2 = \frac{QMErro}{r}$;

vi) calcular o valor da estatística λ para a partição que maximiza a soma de quadrados, dado pela seguinte equação: $\lambda = \frac{\pi}{2(\pi-2)} \cdot \frac{\beta_0^2}{\sigma_0^2}$, com λ tendo uma distribuição teórica de probabilidade de qui quadrado χ^2 com w graus de liberdade, para $w = \frac{k}{(\pi-2)}$.

Se $\lambda \geq \chi^2_{\left[\alpha, \frac{g}{(\pi-2)}\right]}$ rejeita-se a hipótese de que os dois grupos são idênticos em favor da hipótese alternativa de que os dois grupos diferem estatisticamente. Por outro lado caso se rejeite esta hipótese, os dois subgrupos formados serão independentemente submetidos aos passos iii) a vi), fazendo, respectivamente, $g = i_1$ e $g = i_2$. E assim o processo em cada subgrupo se encerra ao se aceitar a hipótese H_0 no passo vi) ou se cada subgrupo contiver apenas uma média.

Sendo assim o intervalo de confiança neste caso é obtido usando na sua construção a estatística $\lambda = \frac{\pi}{2(\pi-2)} \cdot \frac{\beta_0^2}{\sigma_0^2}$, sendo usado o critério $\lambda \geq \chi^2_{\left[\alpha, \frac{g}{(\pi-2)}\right]}$.

INTERVALO DE CONFIANÇA PARA O COEFICIENTE LINEAR ALFA (A) DO MODELO DE REGRESSÃO LINEAR SIMPLES

Segundo Assis (2013), neste caso o intervalo de confiança é construído através da seguinte equação:

$$P\left(\hat{a} - t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{a}) \leq \alpha \leq \hat{a} + t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{a})\right) = 1 - \alpha$$

onde $\hat{a}$ é o valor do coeficiente linear da equação de regressão, $S(\hat{a})$ é o erro padrão do coeficiente linear da equação de regressão, a é o valor do coeficiente linear populacional a ser estimado, $t_{\left(v,\frac{\alpha}{2}\right)}$ é o valor crítico ou tabelado sob a curva do teste ou distribuição amostral do estimador $\hat{a}$, a qual é do tipo t de Student e $1 - \alpha$ é o coeficiente de confiança do intervalo. A interpretação deste intervalo é feito da seguinte forma: O intervalo de confiança de limites fiduciais dados por $\hat{a} - t_{\left(v,\frac{\alpha}{2}\right)}$ e $\hat{a} + t_{\left(v,\frac{\alpha}{2}\right)}$ deve conter o verdadeiro valor do coeficiente linear populacional a (parâmetro populacional) da equação com um grau de $(1-\alpha)\%$ de probabilidade de confiança, ou seja, se for construído 100 intervalos do tipo $P\left(\hat{a}-t_{\left(v,\frac{\alpha}{2}\right)}\cdot S(\hat{a})\leq \alpha\leq \hat{a}+t_{\left(v,\frac{\alpha}{2}\right)}\cdot S(\hat{a})\right)=1-\alpha$ usando a mesma metodologia, então em média $100(1-\alpha)\%$ destes deverão incluir o verdadeiro valor populacional do coeficiente linear α com uma probabilidade de confiança de $100(1-\alpha)\%$. O comprimento ou amplitude do intervalo de confiança depende do valor do erro padrão de $\hat{a}$, $S(\hat{a})$, o qual é determinado conforme a equação a seguir:

$$S(\hat{a})=S\cdot\sqrt{\frac{\sum_{i=1}^{n}X_i^2}{n\left[\sum_{i=1}^{n}X_i^2-\frac{(\sum_{i=1}^{n}X_i)^2}{n}\right]}}$$

De antemão é sabido que o valor do coeficiente linear da equação do modelo de regressão é dado por:

$$\hat{a}=\frac{\sum_{i=1}^{n}Y_i}{n}-\hat{b}\frac{\sum_{i=1}^{n}X_i}{n}=\frac{1}{n}\sum_{i=1}^{n}Y_i-X_i\hat{b}$$

então a $\mathrm{Var}(\hat{a})=\frac{1}{n}+\hat{b}\frac{\bar{X}^2}{\mathrm{SQ}X}=\frac{\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}(X_i-\bar{X})^2}$, ou então do seguinte modo: se $\hat{a}=\bar{Y}-\hat{b}\bar{X}$, a

$$\mathrm{Var}(\hat{a})=\frac{1}{n}\mathrm{Var}(\bar{Y})+\mathrm{Var}(\hat{b}\bar{X})-2\mathrm{Cov}(\bar{Y},\hat{b}\bar{X})=\mathrm{Var}(\bar{Y})+\bar{X}^2\mathrm{Var}(\hat{b})-2\bar{X}\mathrm{Cov}(\bar{Y},\hat{b})$$

Considerando que a estimativa do valor da covariância entre a média de Y e o coeficiente angular é nula, ou seja, $Cov(\bar{Y},\hat{b})=0$, fica então assim:

$$\mathrm{Var}(\hat{a})=\mathrm{Var}(\bar{Y})+\bar{X}^2\mathrm{Var}(\hat{b})=\frac{S^2}{n}+\bar{X}^2\frac{S^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)S^2$$

$$\mathrm{Var}(\hat{a})=\left[\frac{\sum_{i=1}^{n}X_i^2}{n\left(\sum_{i=1}^{n}X_i^2-\frac{(\sum_{i=1}^{n}X_i)^2}{n}\right)}\right]S^2$$

ou seja,

$$\mathrm{Var}(\hat{a})=\mathrm{Var}(\bar{Y}-\bar{X}\hat{b})=V(\hat{a})=\left[\frac{\sum_{i=1}^{n}X_i^2}{n\left(\sum_{i=1}^{n}X_i^2-\frac{(\sum_{i=1}^{n}X_i)^2}{n}\right)}\right]S^2$$

e o erro padrão de $\hat{a}$ é dado por:

$$S(\hat{a}) = S \cdot \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)}}$$

Intervalo de confiança para o coeficiente de regressão ou de inclinação beta (β) do modelo de regressão linear simples.

Para este coeficiente o intervalo de confiança é construído conforme a equação a seguir (ASSIS, 2013): $P\left(\hat{b} - t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{b}) \leq \beta \leq \hat{b} + t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{b})\right) = 1 - \alpha$, onde $\hat{b}$, é o valor do coeficiente de regressão da equação de regressão, $S(\hat{b})$ é o erro padrão do coeficiente de regressão da equação de regressão, β é o valor do coeficiente de regressão populacional a ser estimado, $t_{\left(v,\frac{\alpha}{2}\right)}$, é o valor crítico ou tabelado sob a curva do teste ou distribuição amostral do estimador $\hat{b}$, a qual é do tipo t de Student e $1 - \alpha$ é o coeficiente de confiança do intervalo. A interpretação deste intervalo é feito da seguinte forma: O intervalo de confiança de limites fiduciais dados por $\hat{b} - t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{b})$ e $\hat{b} + t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{b})$ deve conter o verdadeiro valor do coeficiente de regressão ou angular populacional β (parâmetro populacional) da equação com um grau de $(1 - \alpha)\%$ de probabilidade de confiança, ou seja, se for construído 100 intervalos do tipo $P\left(\hat{b} - t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{b}) \leq \beta \leq \hat{b} + t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{b})\right) = 1 - \alpha$ usando a mesma metodologia, então em média $(1 - \alpha)\%$ destes deverão incluir o verdadeiro valor populacional do coeficiente de regressão β com uma probabilidade de confiança de $(1 - \alpha)\%$. O comprimento ou amplitude do intervalo de confiança depende do valor do erro padrão de $\hat{b}$, $S(\hat{b})$, dado como mostrado a seguir: sendo o coeficiente de regressão linear simples dado por:

$$\hat{b} = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j - (\sum_{i=1}^{n} X_i)(\sum_{j=1}^{n} Y_i)}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \sum_{i=1}^{n} Y_i \left[\frac{nX - \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\right]$$

portanto tem-se que,

$$\mathrm{Var}(\hat{b}) = S^2 \cdot \sum_{i=1}^{n}\left[\frac{nX - \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\right]^2 = S^2 \cdot \sum_{i=1}^{n}\left[\frac{X - \bar{X}}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\right]^2 = \frac{S^2}{\sum_{i=1}^{n} x_i^2}$$

Esta última fórmula é válida se for utilizada o valor da variável centrada $x_i = X_i - \bar{X}$, ou melhor, ainda, a variância e o erro padrão do coeficiente de regressão amostral são dados respectivamente pelas seguintes fórmulas:

$$\mathrm{V}(\hat{b}) = \sum_{i=1}^{n} \left[\frac{X - \bar{X}}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \right]^2 = \frac{S^2}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

e

$$S(\hat{b}) = \frac{S}{\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}}$$

Intervalo de confiança ou intervalo de predição ou previsão para a resposta média, dado um particular valor da variável independente $f(x)$ ou $\mathrm{E}[Y/x_i]$.

Neste caso o intervalo de confiança é dado pela equação a seguir (ASSIS, 2013):

$$P\left(\hat{Y}_i - t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{Y}_i) \leq Y_i \leq \hat{Y}_i + t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{Y}_i)\right) = 1 - \alpha$$

onde $t_{\left(v=n-2,\frac{\alpha}{2}\right)}$ é o $\left(1 - \frac{\alpha}{2}\right) \cdot 100$ percentil da distribuição t-Student e $S(\hat{Y}_i)$ é o erro padrão de $\hat{Y}_i$, dado por:

$$\mathrm{Var}(\hat{Y}_i) = S^2 \cdot \left[\frac{1}{n} + \frac{X_i - \bar{X}}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\right]$$

e

$$S(\hat{Y}_i) = S \cdot \sqrt{\frac{1}{n} + \frac{X_i - \bar{X}}{SQX}}$$

Intervalo de confiança ou de previsão para um valor individual de Y, Y_i, dado um particular valor da variável X_i.

Conforme ASSIS (2013) para um valor individual da variável X modelo de regressão, por exemplo X_0, o valor previsto pela equação de regressão ajustada para uma nova observação Y_0 é dado por $\hat{Y}_0 = b_0 + b_1 X_0$. Demonstra-se que a média de $\hat{Y}_0$ é $\mathrm{E}[\hat{Y}_0] = \beta_0 + \beta_1 X_0$. Observe que, apesar da mesma designação usada no item anterior, aqui $\hat{Y}_0$ tem uma conotação diferente. A variância de $\hat{Y}_0$ tem dois componentes: um deles é a variância de $\hat{Y}_0$ (predição da média) e outro é a variância de uma dada observação em torno de sua média $\mathrm{E}[Y]$, assumida ser σ^2. Assim, a variância de uma nova observação estimada através da equação de regressão ajustada:

$$\sigma_{\hat{Y}_0}^2 = \sigma^2 \cdot \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right]$$

A estimativa da variância de $\hat{Y}_0$ é obtida substituindo σ^2 por seu estimador, isto é,

$$S_{\hat{Y}_0}^2 = S^2 \cdot \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

e um intervalo a $(1 - \alpha) \cdot 100\%$ de confiança para Y_h é dado por:

$$P\left(\hat{Y}_i - t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{Y}_0) \leq Y_i \leq \hat{Y}_i + t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(\hat{Y}_0)\right) = 1 - \alpha$$

sendo $t_{\left(n-2,\frac{\alpha}{2}\right)}$ o $(1 - \frac{\alpha}{2}) \cdot 100$ percentil da distribuição t-Student. Esse intervalo de previsão representa um intervalo que tem uma probabilidade $(1 - \alpha)$ de conter, não um parâmetro populacional (caso dos intervalos de confiança), mas sim um novo valor y_0 da variável aleatória Y_0. Ressalte-se ainda o fato dos intervalos de previsão (resposta futura) terem uma amplitude maior do que os correspondentes intervalos de confiança (resposta média). No entanto, a amplitude dos intervalos de previsão é afetada pelos mesmos fatores que a amplitude dos intervalos de confiança, isto é, aumenta com a variância S^2; aumenta com $|x_0 - \bar{x}|$ e para uma dada amostra, é mínima para $x_0 = \bar{x}$; diminui quando a soma de quadrados de X (SQX) aumenta e diminui com o aumento do tamanho da amostra (n).

Segundo FERREIRA (2005), outra importante maneira de predição refere-se à necessidade de predizer o valor da variável independente X_i que seria esperado na população se o valor da variável resposta (dependente) Y_i fosse observado. Esse mecanismo é chamado de predição inversa. O valor pontual dessa predição é obtido de forma simples realizando algumas manipulações algébricas na equação de regressão linear ajustada. O estimador pontual para a predição inversa é determinado através da seguinte equação: $\hat{X}_i = \frac{Y_i - \hat{a}}{\hat{b}}$.

O intervalo de confiança para a predição inversa não é derivado de forma trivial (ZAR, 1996, citado por FERREIRA, 2005). Os limites dos intervalos de confiança são simétricos ao redor do valor de $\hat{Y}_i$, mas os limites para a predição de $\hat{X}_i$ não são simétricos. O intervalo de confiança para o valor populacional X_i com $(1 - \alpha)$ de probabilidade para um dado valor observado de Y_i pode ser construído através da seguinte fórmula:

$$\mathrm{IC}_{(1-\alpha)}[X_i]: \bar{X} + \frac{\hat{b}(Y_i - \bar{Y})}{K} \pm \frac{t_{\left(v,\frac{\alpha}{2}\right)}}{K}\sqrt{S^2\left[\frac{(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + K(1 + \frac{1}{n})\right]}$$

em que o valor de K é dado por: $K = \hat{b}^2 - t_{\left(v,\frac{\alpha}{2}\right)} S_{\hat{b}}^2$ e $v = n - 2$ graus de liberdade.

Intervalo de confiança ou fiducial para o coeficiente de correlação populacional linear simples de Pearson (ρ).

A distribuição amostral do coeficiente de correlação linear simples de Pearson r, em geral não é simétrica e não possui distribuição teórica de probabilidade Normal ou gaussiana, pois depende não só do sinal mas também da magnitude do coeficiente, exceto no caso particular em que $\rho = 0$. Quando o coeficiente se afasta de zero a distribuição é muito assimétrica, positiva ou negativa conforme o sinal do coeficiente. Quando o coeficiente se aproxima de zero a distribuição é simétrica e para grandes amostras ($n \geq 30$) a distribuição é aproximadamente Normal. Nessas condições o intervalo de confiança para o coeficiente de correlação linear simples populacional ρ, será dado conforme a equação seguinte (ASSIS, 2013):

$$P\left(r - t_{\left(v,\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{1-r^2}{n-2}} \leq \rho \leq r + t_{\left(v,\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{1-r^2}{n-2}}\right) = 1 - \alpha$$

$$P\left(r - t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(r) \leq \rho \leq r + t_{\left(v,\frac{\alpha}{2}\right)} \cdot S(r)\right) = 1 - \alpha$$

Essa expressão não pode ser usada para construir o intervalo de confiança, pois a distribuição amostral do coeficiente r, só é aproximadamente normal sob a hipótese $H_0: \rho = 0$, e assim só serve para aplicar testes de hipótese para a situação de distribuição normal bivariada e com $\rho = 0$, pois não existe uma forma universal para a distribuição do coeficiente r, pois depende da distribuição da população bivariada amostrada. Para contornar o problema e ser possível construir intervalos de confiança usando uma variável transformada Z_r, Fisher (1915) demonstrou que essa transformação do coeficiente de correlação r em Z_r produziria uma variável normalmente distribuída com média zero. Sendo assim esse autor derivou uma aproximação normal de "r" em "Z_r", cuja distribuição amostral possui distribuição normal aproximada ou assintótica cuja média e desvio padrão são mostrados a seguir: sendo assim, $Z_r = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$, possui distribuição normal assintótica com média $\mu_{Z_r} = \frac{1}{2}\left[\ln\left(\frac{1+\rho}{1-\rho}\right) + \frac{\rho}{n-\rho}\right]$, variância dada por $\sigma^2_{Z_r} \cong \frac{1}{n-3}$, onde o desvio padrão de Z_r é dado por $S_{Z_r} = \frac{1}{\sqrt{n-3}}$, onde n é a dimensão da amostra. Então a variável pivotal é dada por $Z = \frac{Z_r - Z_\rho}{\frac{1}{\sqrt{n-3}}} \cap N(0,1)$.

A variância de Z não envolve o verdadeiro valor do parâmetro e foi essa propriedade de estabilização da variância que conduziu Fisher a propor a transformação da tangente hiperbólica inversa dada por $Z_r = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$,. Essa transformação não é válida para $\rho = -1$ ou $\rho = 1$ e para $\rho = 0$, a distribuição exata de t deve ser preferida. Dessa forma, a estatística $Z = Z_c = \frac{Z_r - \mu_Z}{\sigma_Z}$ deve ser usada para realizar testes de hipóteses e obter intervalos de confiança de ρ.

Existem tabelas prontas para Z_r em função de vários valores de r. Portanto o intervalo de confiança para Z_R com $(1 - \alpha)100\%$ de confiança é dado por:

$$P\left(Z_r - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}} \leq Z_R \leq Z_r + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right) = 1 - \alpha$$

Como $Z_r = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$, assim o intervalo de confiança será dado por:

$$P\left(\frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}} \leq Z_R \leq \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right) = 1 - \alpha$$

Agora para transformar o intervalo obtido num intervalo de confiança para o coeficiente de correlação populacional ρ utiliza-se novamente a tabela que relaciona Z_r com r, ou então calculando os limites do intervalo através da expressão $r = \frac{e^{2Z_r}-1}{e^{2Z_r}+1}$. Vale salientar que o intervalo de confiança para Z_R está centrado em torno de Z_r, enquanto que o intervalo de confiança resultante para o coeficiente ρ não está centrado em torno de r. Portanto o intervalo de confiança para o coeficiente ρ é o seguinte:

$$P\left(\frac{e^{2Z_{r_{\text{inf}}}}-1}{e^{2Z_{r_{\text{inf}}}}+1} \leq \rho \leq \frac{e^{2Z_{r_{sup}}}-1}{e^{2Z_{r_{sup}}}+1}\right) = 1 - \alpha$$

onde, $Z_{r_{\text{inf}}} = Z_r - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}$, $Z_{r_{\text{sup}}} = Z_r + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}$ e $Z_r = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$. Sendo assim tem-se que:

$$P\left(\frac{e^{2\left[\frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right]} - 1}{e^{2\left[\frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] - Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right]} + 1} \leq \rho \leq \frac{e^{2\left[\frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right]} - 1}{e^{2\left[\frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] + Z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{1}{\sqrt{n-3}}\right]} + 1}\right) = 1 - \alpha$$

É possível também obter os valores limites para um intervalo que contenha o verdadeiro valor do coeficiente de correlação populacional ρ utilizando-se gráficos, que se referem aos coeficientes de confiança. A utilização desses gráficos é feita da seguinte maneira: para um coeficiente de confiança $(1 - \alpha)100\%$ fixado, toma-se o valor observado de $r = r_0$ na amostra de tamanho n em estudo e localiza-se este valor no eixo das abscissas; a partir dele levanta-se uma paralela ao eixo das ordenadas até encontrar as duas curvas referentes ao n em questão. De cada um dos pontos de encontro traça-se uma reta paralela ao eixo das abscissas até encontrar o eixo dos valores de ρ, obtendo-se assim os limites inferior e superior do intervalo de confiança para o parâmetro populacional ϱ o qual é o verdadeiro coeficiente de correlação linear simples populacional.

APÊNDICE 6

PLANILHAS, TABELAS E FORMULÁRIOS DE APOIO

Planilhas, tabelas, distribuições de frequências e formulários usados como suporte na construção de intervalos de confiança para serem utilizados na estimação dos mais diferentes tipos de parâmetros populacionais, tais como a média, a variância, o desvio padrão, a proporção, a diferença entre médias, a diferença entre proporções, a razão entre variâncias, o coeficiente de correlação, o coeficiente linear da reta de regressão, o coeficiente angular ou de regressão do modelo de regressão, para valores de variáveis dependentes em modelos de regressão, para contrastes entre médias, para contrastes de grupos de médias, dentre outros parâmetros estatísticos:

Tabela ou distribuição de frequências ou seriação para variáveis quantitativas discretas.

Número Mensal de Ovos de Galinhas Poedeiras (X_i)	Contagem	Frequência do Número de Ovos f_i	$f_{ac}\downarrow$	$f_{ac}\uparrow$	f_r	$f_{rac}\downarrow$	$f_{rac}\uparrow$	$f_{\%} = f_p$	$f_{pac}\downarrow$	$f_{pac}\uparrow$
SOMA	-		-	-		-	-		-	-

Tabela ou distribuição de frequências ou seriação para variáveis quantitativas discretas.

Número Mensal de Ovos de Galinhas Poedeiras (X_i)	Contagem	Frequência do Número de Ovos f_i	$f_{ac}\downarrow$	$X_i f_i$	X_i^2	$X_i^2 f_i$

TOTAL		---------	----------

Tabela ou distribuição de frequências ou seriação para variáveis quantitativas discretas ou contínuas.

CLASSES (Altura de Plantas de Milho em Centímetros)	CONTAGEM	f_i	X_i	$f_{ac} \downarrow$	$f_{ac} \uparrow$	f_r	$f_{rac} \downarrow$	$f_{rac} \uparrow$	$f_{\%} = f_p$	$f_{pac} \downarrow$	$f_{pac} \uparrow$
[-----------)											
[-----------)											
[-----------)											
[------------)											
[-----------)											
[-----------)											
TOTAL	-------------		-	-----	-----		-----	-----		-----	-----

Tabela ou distribuição de frequências ou seriação para variáveis quantitativas discretas ou contínuas.

Classes (Altura de Plantas de Milho em Centímetros)	Contagem	f_i	$f_{ac} \downarrow$	X_i	$X_i f_i$	X_i^2	$X_i^2 f_i$
[-----------------)							
[-----------------)							
[-----------------)							
[-----------------)							
[-----------------)							
[-----------------)							
[-----------------)							
Total			--------	--------		--------	

Planilha de apoio computacional.

	X_i	Y_i	X_iY_i	X_i^2	Y_i^2	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
	X_1	Y_1							
	X_2	Y_2							
	$\vdots$	$\vdots$							
	X_n	Y_n							
TOTAL									

FORMULÁRIO DE APOIO

A soma de n termos pode ser simbolicamente representada por:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$$

O produto de n fatores pode ser simbolicamente representado por:

$$\prod_{i=1}^{n} X_i = X_1 \cdot X_2 \cdot \ldots \cdot X_n$$

As diversas medidas podem ser determinadas de acordo com este formulário a seguir:

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\text{Total} = t = \sum_{i=1}^{n} x_i = N\bar{X}$$

$$\text{Total} = t = \sum_{i=1}^{n} x_i = n\bar{X}$$

$\text{Total} = t = \sum_{i=1}^{n} x_i = N$ (Limites do Intervalo de Confiança)

$$\text{TOTAL} = T = \sum_{i=1}^{n} X_i = N\mu$$

$$\text{CV} = \frac{S}{\bar{X}}100\%$$

$$\bar{X} = \frac{S}{CV}100\%$$

$$\bar{X}CV = S$$

$$S_{[\bar{X}]} = \frac{S}{\sqrt{n}}$$

$$S_{[\bar{X}]}\sqrt{n} = S$$

$$n = \frac{S^2}{\left(S_{[\bar{X}]}\right)^2}$$

O erro padrão pode ser fornecido como uma porcentagem do valor da média aritmética, como por exemplo, $S_{[\bar{X}]} = (4\%)(\bar{X})$, ou seja, $S_{[\bar{X}]} = (0{,}04)(20{,}5 \text{ litros}) = 0{,}82 \text{ litros}$

$$S^2 = \frac{\sum_{i=1}^{n} f_i X_i - \frac{\sum_{i=1}^{n} f_i X_i}{n}}{n-1} = \frac{\sum_{i=1}^{n} f_i \,(X_i - \bar{X})^2}{n-1}$$

$$S = \sqrt{S^2}$$

Em distribuições de frequências simétricas a média = moda = mediana, ou seja, $\bar{X} = Md = Mo$. Em particular em séries estatísticas unimodal, moderadamente assimétrica e em que o número de observações é suficientemente grande ($n \geq 30$) e se for verificado também uma pequena escala de unidades que divide a distribuição, então observa-se a seguinte relação entre as medidas descritivas média, mediana e moda, $\bar{X} - Mo = 3(\bar{X} - Md)$, sendo assim a partir desta relação observada o pesquisador pode determinar uma das três medidas de posição ou tendência central em função das outras duas medidas, conforme mostrado a seguir: $\bar{X} = \frac{3Md\,-\,Mo}{2}$, $\text{Mo} = 3\text{Md} - 2\bar{X}$ e $\text{Md} = \frac{2\bar{X}+Mo}{3}$.

Estas expressões indicam geometricamente que a mediana se situa sempre entre a média e a moda, sendo sua distância para a Moda o dobro de sua distância para a média aritmética, Ou seja.

A frequência relativa ou $\hat{p}$, f de fada, determinado pela seguinte expressão $f = \hat{p} = \frac{x}{n}$ mostra a $f = \hat{p}$ proporção (frequência relativa) de eventos sucesso na amostra; onde x é o número (frequência absoluta) de eventos sucesso na amostra; e n é o número de elementos da amostra (tamanho da amostra).

Por outro lado em muitas análises estatísticas, o investigador pode necessitar de quantidades importantes pra as suas análises tais como: somas de quadrados, somas de produtos ou soma de produtos cruzados e demais fórmulas utilizadas, por exemplo, em análise de regressão e de correlação dentre outros tipos de análises.

$$SQX = \sum_{i=1}^{n} X_i^2 \; - \frac{(\sum_{i=1}^{n} X_i)^2}{n}$$

$$SQY = \sum_{i=1}^{n} Y_i^2 \; - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

$$SPXY = \sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j \; - \frac{\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{Y} = \frac{\sum_{j=1}^{n} Y_j}{n}$$

$$Y_i = \hat{a} + \hat{b}X_i$$

$$\hat{a} = Y_i - \hat{b}X_i$$

$$\hat{b} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j - \frac{\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j}{n}}{\sum_{i=1}^{n} X_i^2 \; - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

$t_{\text{teste}} = \frac{\hat{b}-\beta}{S(\hat{b})}$ com $v = n-2$ graus de liberdade onde $S(\hat{b}) = \frac{S}{\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}}$, sendo $S^2 = \frac{SQD}{n-2}$ e

$\mathrm{SQD} = \mathrm{SQ}Y - \frac{\mathrm{SP}XY^2}{\mathrm{SQ}X}$, $S = \sqrt{S^2}$, $r^2 = \frac{\hat{b}\,SPXY}{SQY}100\% = \frac{\hat{b}^2 SQX}{SQY}100\%$.

$$r \;=\; \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j - \frac{\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j}{n}}{\sqrt{\left[\sum_{i=1}^{n} X_i^2 \; - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right]\left[\sum_{i=1}^{n} Y_i^2 \; - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}\right]}} = \frac{SPXY}{\sqrt{SQX \cdot SQY}}$$

$t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}$, ou $t_{\text{teste}} = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$, onde a variável aleatória t de Sudent se distribui com $v = n-2$ graus de liberdade.

Outra opção é achar os valores críticos do coeficiente de correlação linear simples de Pearson (r) em tabela s, resolvendo-se a seguinte equação $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ em relação a r, obtendo-se $r = \frac{t}{\sqrt{t^2+n-2}}$.

Onde o valor de t é dado pelas tabelas estatísticas supondo-se um caso de teste de hipótese ou de significância bilateral ou intervalo de confiança bilateral com um número de graus de liberdade determinado por $v = n - 2$. Existem tabelas que fornecem os resultados do coeficiente de correlação linear simples r de Pearson para valores selecionados do tamanho da amostra n para determinado nível de significância alfa (α).

Onde os termos ou elementos das fórmulas significam: $\text{SQ}X = S_{XX}$ é a soma de quadrados de X; $\text{SQ}Y = S_{YY}$ é a soma de quadrados de Y; $\text{SP}XY = S_{XY}$ é a soma de produtos cruzados ou soma de produtos de XY.

APÊNDICE 7

ROTINAS EM R PARA A TEORIA DA ESTIMAÇÃO

Introdução ao R

O R é uma linguagem de programação desenvolvido principalmente para área da estatística, isto é, análise de dados. Foi inicialmente desenvolvida por George Ross Ihaka e Robert Clifford Gentelman, Nova Zelândia, em 1993. A inspiração dessa linguagem foi baseada na linguagem S, criada por John Chambers. A base inicial da linguagem foi escrita na linguagem C, FORTRAN e na própria linguagem R.

Usamos muitas vezes o artigo "o" ou "a" para o R, porque também é considerado um ambiente de software livre, de código aberto, do inglês: *"open source"*. Apesar do R teruma interface por linha de comando, diversas GUIs[1] são disponíveis como o RStudio, Jupyter, dentre outros.

Para instalarmos o R, basta acessarmos o CRAN e escolhermos o sistema operacional para a instalação. Sempre quando vamos instalar, fazemos a instalação base do pacote, isto é, a instalação dos pacotes básicos mínimos para a utilização do software. Por exemplo, no caso do SO Windows basta acessar https://cran.r- project.org/bin/windows/base.

O R é uma linguagem orientada a objetos. Isto significa que não precisamos definira natureza de um objeto, para posteriormente criá-lo, assim como fazemos com a linguagens C ou FORTRAN, por exemplo. De forma conotativa, supondo que fôssemos reconhecer os móveis de nossa casa pela linguagem R, não precisaremos caracterizar um sofá para que a linguagem entenda que a classe desse objeto é um sofá. As suas características já o identificam como sofá. Em termos de linguagem, não precisamos definir se um objeto tem natureza numérica ou de caractere. Os elementos ao serem atribuídos a esse objeto se forem numéricos, por exemplo, já determinam a classe numérica do objeto.

A ideia de um objeto é uma forma de aproximar a linguagem as coisas do mundo real, de modo a representar qualquer coisa. Representamos um objeto por um nome ou letrano R, por exemplo,

```
x <- 5
```

O símbolo <- significa atribuição, isto é, o objeto x recebe o número 5. De outro modo, o número 5 é atribuído ao objeto x. Observe que não precisei identificar a classe do objeto x. O fato de ele receber um número, já o identifica como de natureza numérica, veja:

```
class(x)
## [1] "numeric"
```

[1] Do Inglês, *Graphical User Interface.*

Podemos destacar algumas classes de objetos: Numérico, Caractere, Vetor, Lista, Matriz, Função, Lógica, dentre outras.

Por fim, uma outra estrutura básica no R são os pacotes. Os pacotes são umconjunto de objetos, em sua grande maioria da classe de funções, que se destina solução de um problema. Assim, ao invés de criar funções para um determinado problema, e toda vida que precisar dessas funções ter que executar cada uma delas, podemos fazer um pacote, eao carregá-lo e anexá-lo, teremos todas as funções disponíveis em nosso ambiente de trabalho. Isso ajuda e muito o dia-a-dia de um programador ou pesquisador, em suasanálises.

Existem pacotes que já estão disponíveis no ato da instalação do R. São estes:

```
library(lib.loc = .Library)
# Packages in library ‘C:/PROGRA~1/R/R-36~1.1/library’:#
# base            The R Base Package
# boot            Bootstrap Functions
#                 (Originally by Angelo Canty
#                 for S)
# class           Functions for Classification#
cluster           "Finding Groups in Data":
#                 Cluster Analysis Extended
#                 Rousseeuw et al.
# codetools       Code Analysis Tools for R#
compiler          The R Compiler Package
# datasets        The R Datasets Package
# foreign         Read Data Stored by 'Minitab',#
                  'S', 'SAS', 'SPSS', 'Stata',
#                 'Systat', 'Weka', 'dBase', ...#
graphics          The R Graphics Package
# grDevices       The R Graphics Devices and
#                 Support for Colours and Fonts
# grid            The Grid Graphics Package
# KernSmooth      Functions for Kernel Smoothing#
                  Supporting Wand & Jones (1995)
# lattice         Trellis Graphics for R
```

```
# MASS            Support Functions and Datasets
```

```
#                         for Venables and Ripley's MASS
#  Matrix                 Sparse and Dense Matrix
#                         Classes and Methods
#  methods                Formal Methods and Classes
#  mgcv                   Mixed GAM Computation Vehicle
#                         with Automatic Smoothness
#                         Estimation
#  nlme                   Linear and Nonlinear Mixed
#                         Effects Models
#  nnet                   Feed-Forward Neural Networks
#                         and Multinomial Log-Linear
#                         Models
#  parallel               Support for Parallel
#                         computation in R
#  rpart                  Recursive Partitioning and
#                         Regression Trees
#  spatial                Functions for Kriging and
#                         Point Pattern Analysis
#  splines                Regression Spline Functions
#                         and Classes
#  stats                  The R Stats Package
#  stats4                 Statistical Functions using S4
#                         Classes
#  survival               Survival Analysis
#  tcltk                  Tcl/Tk Interface
#  tools                  Tools for Package Development
#  translations           The R Translations Package
#  utils                  The R Utils Package
```

Contudo, qualquer usuário poderá desenvolver pacotes R. Para mais detalhes, acessar: https://bendeivide.github.io/meupacoter. Para instalar um pacote, primeiro devemos saber o nome do pacote (nome_pacote), e posteriormente usar a função:

```
install.packages("nome_pacote")
```

Isso significa que foi realizado o *download* do pacote e instalado na biblioteca R do sistema operacional. Posteriormente, para os objetos do referido pacote estarem disponíveis no ambiente de trabalho, precisamos carregá-lo e anexá-lo. Para isso, usamos a linha de comando:

```
library("nome_pacote")
```

Conceitos básicos

Vimos anteriormente como criarmos escalares. Agora, apresentamos a forma vetorial de um objeto, usando a função concatenar, c(), isto é: vetor <- c(5, 10, 20, 30, 40, 50)

Criamos um objeto, vetor, que recebe cinco elementos numéricos. Para a criaçãode um objeto da classe função, usamos function(). Como exemplo, temos:

```
# Função para calcular a média
media <- function(x) {
  resultado <- sum(x) / length(x)
  return(resultado)
}
# Vetor
vetor <- c(5, 10, 20, 30, 40, 50)
## [1] 25.83333
```

O objeto media recebeu uma função que calcula a média aritmética. Dizemos que oelemento x é o argumento da função, que no nosso caso recebeu o objeto vetor. Ao executar a função pelo nome do objeto, media, o resultado é a média aritmética.

Escrevemos as linhas de comando em um *script*, isto é, um arquivo com a extensão .R. A execução do comando é realizada no *console*, que geralmente apresenta o símbolo ">". Se não desejar guarda o código desenvolvido, as linhas de comando podem serexecutadas todas no console.

Ao invés de criarmos uma função chamada media, poderíamos buscar algum pacote que já executasse essa função. No nosso caso, temos o pacote base, que já vem nainstalação do R, e a função mean() calcula a média aritmética. Dessa forma, como esse pacote já vem anexado ao abri o compilador do R, basta executar as linhas de comando:

```
# Vetor
vetor <- c(5, 10, 20, 30, 40, 50)

# Cálculo da média
mean(vetor)

## [1] 25.83333
```

Para obter o mesmo resultado. Uma outra coisa interessante no código, é quequando desejamos comentar uma linha, iniciamos com o símbolo #, isto significa, que esta linha não será executada.

Aplicações na estimação estatística

Estimadores intervalares e teste de hipóteses para média(a) de uma ou duas populações normais

Iremos utilizar a função z.test do pacote TeachingDemos quando a variância populacional for conhecida. A instalação e o anexamento do pacote seguem as linhas de comando:

```
install.packages("TeachingDemos")library ("TeachigDemos")
```

Por fim, segue o exemplo.

```
amostra <- stats::rnorm(n = 25, # Tamanho da amostra n = 25
                  mean = 100, # Média igual a 100
                  sd = 5) # Desvio padrão igual a 5

# Aplicação do IC e do teste de hipóteses
TeachingDemos::z.test(x = amostra, # Amostra
                      mu = 99, # Hipótese a ser testada
                      stdev = 5, # Desvio padrão conhecido
                      alternative = "two.sided", # Teste bilateral
                      conf.level = 0.95) # Nível de confiança

##
##   One Sample z-test##
## data:   amostra
```

Se houver dúvidas sobre essa função, digite no console?Z.test, que um documento de ajuda será aberto para auxiliá-lo.

Uma aplicação dentro de aplicações reais, geralmente não temos o conhecimentodos parâmetros populacionais. Dessa forma, usaremos o teste de hipóteses baseado na distribuição *t*-Student, tanto para uma quanto para duas populações. Seguem as linhas de comando:

```r
stats::t.test(x = rnorm(25), # Amostra do Grupo 1
         y = rnorm(27, 5), # Amostra do Grupo 2
         alternative = "two.sided", # Teste bilateral
         mu = 0, # Hipótese a ser testada
         paired = FALSE, # Dados pareados, sim (TRUE) ou não (FALSE)
         var.equal = TRUE, # Variâncias iguais (TRUE) ou diferentes
(FALSE)
         conf.level = 0.95) # Nível de confiança
```

```
##
##  Two Sample t-test##
## data:  rnorm(25) and rnorm(27, 5)
## t = -21.222, df = 50, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0## 95
percent confidence interval:
##  -5.767870 -4.770455
## sample estimates:
```

Nesse exemplo muitos problemas podem ser abordados, como dados emparelhados (argumento paired), variâncias homocedásticas e heterocedásticas (argumento var,equal), ou até mesmo para o caso de uma população, bastando assumir y = NULL, desconsiderando os demais argumentos referentes a dois grupos. Para mais detalhes, use no console ?t.test.

Estimadores intervalares e teste de hipóteses para proporções populacionais

Para a utilização de um teste de hipóteses ou intervalo de confiança exato deproporções em um experimento de Bernoulli, usamos as linhas de comando:

```r
stats::binom.test(x = 40, # Número de sucessos
           n = 100, # Tamanho da amostra
           p = 0.5, # Hipótese a ser testada
           alternative = "two.sided", # Teste bilateral
           conf.level = 0.95) # Nível de confiança

##
##  Exact binomial test##
## data:  40 and 100
## number of successes = 40, number of trials = 100, p-value = 0.05689##
alternative hypothesis: true probability of success is not equal to0.5
## 95 percent confidence interval: ##
0.3032948 0.5027908
## sample estimates:
## probability of
##                    0.4
```

Detalhes sobre a função, insira no console ?binom.test. Para o caso aproximado, use:

```r
stats::prop.test(x = 40, # Número de sucessos
           n = 100, # Tamanho da amostra
           p = 0.5, # Hipótese a ser testada
           alternative = "two.sided", # Teste bilateral
           conf.level = 0.95, # Nível de confiança
           correct = TRUE) # Correção de continuidade

##
##  1-sample proportions test with continuity correction##
## data:  40 out of 100, null probability 0.5 ## X-
squared = 3.61, df = 1, p-value = 0.05743
## alternative hypothesis: true p is not equal to 0.5## 95
percent confidence interval:
##  0.3047801 0.5029964
## sample estimates:##  p
```

Com esta última função é possível a realização para o caso de duas populações. Paramais detalhes, use ?prop.test.

Estimadores intervalares e teste de hipóteses para comparar variâncias de populações normais

Usando a distribuição F para comparar a igualdade entre variâncias de populaçõesnormais, seguem as linhas de comando:

```
x <- rnorm(35, mean = 0, sd = 2) # População normal com sd = 2y <-
rnorm(30,  mean  =  1,  sd  =  1)  #  População  normal  com  sd  =  1
stats::var.test(x, y)

##

##  F test to compare two variances##

## data:  x and y

## F = 8.739, num df = 34, denom df = 29, p-value = 6.056e-08

## alternative hypothesis: true ratio of variances is not equal to 1## 95
percent confidence interval:

##   4.237599 17.658770
```

REFERÊNCIAS

Azania AAPM et al. (2003). Métodos de superação de dormência em sementes de *Ipomoea* e *Merremia*. *Planta daninha*, 21(2): 203-209.

Andrade DF, Ogliari PJ (2017). Estatística para as ciências agrárias e biológicas: com noções de experimentação. 3. ed. Florianópolis: Editora da UFSC, 475 p.

Assis JP (2013). Regressão e correlação linear simples e múltipla. Mossoró, RN: edufersa, 310 p.

Barbetta P (2005). Estatística aplicada às ciências sociais. 5. ed. Florianópolis: Editora da UFSC, 340 p.

Bolfarine H, Bussab WO (2005). Elementos de amostragem. São Paulo: São Paulo: Edgard Blücher, 274p.

Bussab WO, Morettin PA (2003). Estatística básica. São Paulo: Editora Saraiva, 5. ed. 526P.

Costa Neto PLO (2000). Estatística. São Paulo: Edgard Blücher, 264p.

Curi PR (1998). Metodologia e análise da pesquisa em ciências biológicas. 2. ed. Botucatu, SP. 263 p.

Duncan DB (1955). Multiple-Range and Multiple-F Tests, Biometrics, 11, 1- 42.

Dunnett CW (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50, 1096-1121.

Ferreira DF (2005). Estatística básica. Editora UFLA. Lavras, MG. 625 p.

Fisher RA (1915). Frequency distributions of the values of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, London, v. 10, n. 4, p. 507 – 521.

Fisher RA (1922). On the mathematical foundations of theoretical statistics philosophical transactions royal society. Série A, v. 222. p. 309 – 368.

Fisher RA (1935). The design of experiments California-EUA: Oliver and Boyd. 252 P.

Heath OVS (1981). A estatística na pesquisa cientifica. São Paulo: USP, 95 p.

Keuls M (1952). The use of the Studentized Range in Connection with an analysis of variance, Euphytica, 1, 112-122.

Memória JM (1973). Curso de estatística aplicada à pesquisa científica. Viçosa: Imprensa Universitária, 304p. (apostila).

Murteira B et al. (2001). Introdução à estatística. McGraw-Hill. Portugal. 688 p.

Newman D (1939). The Distribution of the Range in Samples from a normal population, Expressed in Terms of an independent Estimate of the Standard Deviation, Biometrika, 31, 20- 30.

Nunes RP (1998). Métodos para a pesquisa agronômica. Centro de Ciências Agrárias da Universidade Federal do Ceará, Fortaleza, 564 p.

Nunes C, Afonso A (2005). Apontamentos de introdução às probabilidades e à estatística – Manuais da Universidade de Évora: Volume II. Área departamental de ciências exatas. Évora, Portugal, 234 p.

Pedrosa AC, Gama SMA (2004). Introdução computacional à probabilidade e estatística. Porto Editora: Porto, Portugal, 607 p.

Sampaio IBM (1998). Estatística aplicada à experimentação animal. Belo horizonte. Fundação de ensino e pesquisa em medicina veterinária e zootecnia. 221 p.

Scheffé H (1953). A method for Judging all contrasts in the analysis of variance. Biometrika, 40, 87-104.

Scott A, Knott M (1974). Cluster-analysis method for grouping means in analysis of variance. *Biometrics, Washington D.C.,* 30(3): 507-512.

Silveira Júnior P et al. (1980). Estatística geral: inferência estatística. Pelotas: UFPEL: DME, 156 p.

Spiegel MR (2000). Estatística. São Paulo: Macron Books, 4. ed. 580p.

Stevenson WJ (1986). Estatística aplicada à administração. São Paulo: Harper & Row do Brasil. 498 p.

Tukey JW (1953). The problem of multiple comparisons. Unpublished manuscript, Princeton University. 1953. 396p.

Zar JH (1996). Biostatistical analysis. 3 ed. Upper Saddle River. *Prentice Hall, New York*. 662 p.

Zar JH (2010). Biostatistical analysis. 5. ed. Upper Saddle River, New York: Prentice Hall. 944 p.

ÍNDICE REMISSIVO

Q

S

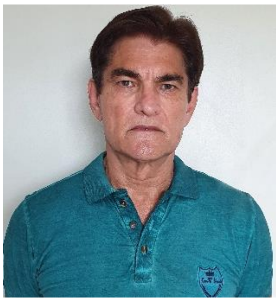
T

V

SOBRE OS AUTORES

Janilson Pinheiro de Assis

Engenheiro Agrônomo graduado em Engenharia Agronômica (1987) na Escola Superior de Agricultura de Mossoró (ESAM). Mestre (1990) em Engenharia Agronômica (Fitotecnia) na Universidade Federal do Ceará (UFC). Doutor (2014) em Produção Vegetal - Fitotecnia na Universidade de São Paulo (USP). Atualmente, é Professor Titular da Universidade Federal Rural do Semiárido (UFERSA), leciona a disciplina de Estatística, possui quatro livros publicados, 25 artigos completos publicados/aceitos em revistas nacionais e internacionais, 20 resumos simples/expandido. É revisor de dez revistas nacionais e internacionais. Contato: (85)999826636.



Roberto Pequeno de Sousa

Engenheiro Agrícola, graduado em Engenharia Agrícola (1981) na Universidade Federal da Paraíba (UFPB). Mestre (1985) em Engenharia Civil (Recursos Hídricos - Irrigação) na Universidade Federal da Paraíba (UFPB). Doutor (2013) em Agronomia - Fitotecnia na Universidade Federal Rural do Semiárido (UFERSA). Atualmente, é Professor Associado IV da Universidade Federal Rural do Semiárido (UFERSA), leciona a disciplina de Estatística Experimental, possui quatro livros publicados, 60 artigos completos publicados/aceitos em revistas nacionais e internacionais, 45 resumos simples/expandido. É revisor de cinco revistas nacionais e internacionais. Contato: (84)994115032.

Paulo César Ferreira Linhares

Engenheiro Agrônomo, graduado em Engenharia Agronômica (2002) na Escola Superior de Agricultura de Mossoró (ESAM). Mestre em Fitotecnia (2007) e Doutorado em Fitotecnia (2009) pela Universidade Federal Rural do Semiárido (UFERSA). Atualmente é Pesquisador na área de Produção Orgânica de Hortaliças da Universidade Federal Rural do Semiárido (UFERSA), possui um livro publicado, 110 artigos publicados em revistas nacionais e internacionais. 100 resumos simples/expandido. 32 orientações de trabalho de conclusão do curso de Agronomia. 22

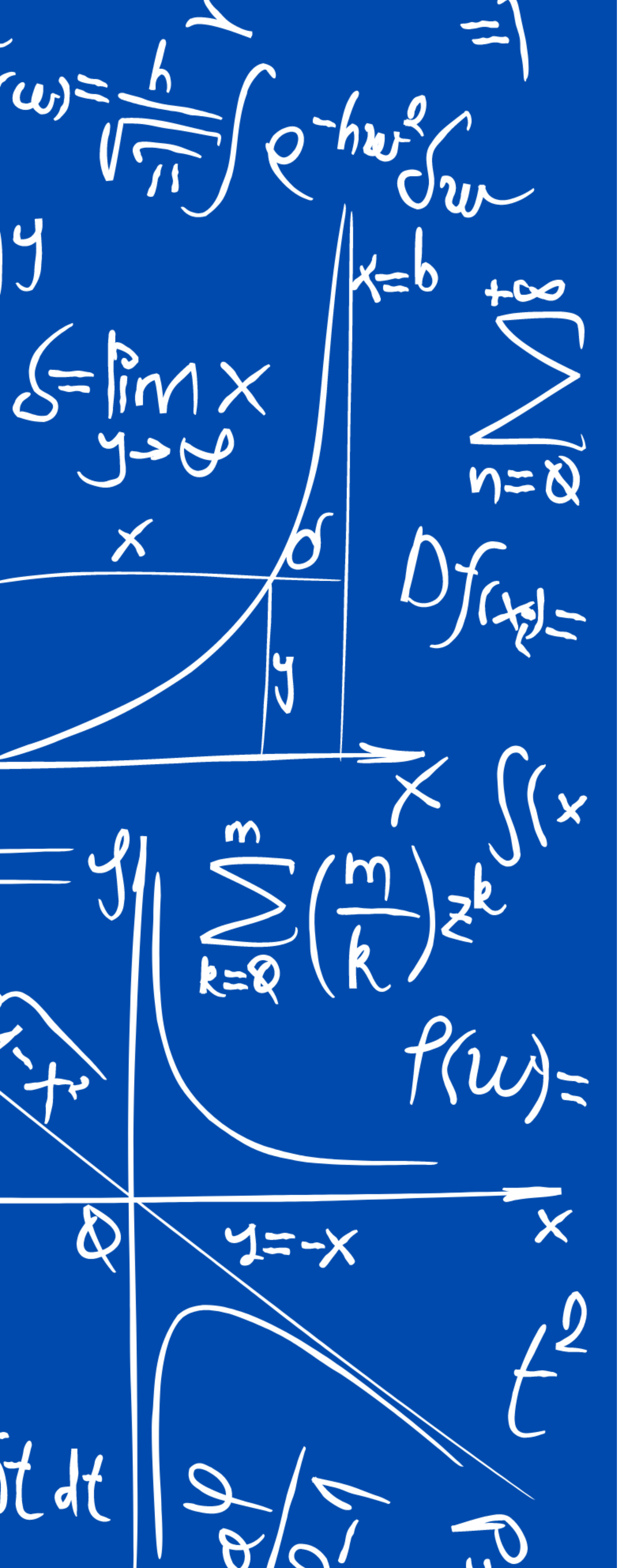
orientações de Dissertação de Mestrado. 01 coorientação de Doutorado. 07 participações em bancas de dissertação de mestrado. 03 participações em tese de Doutorado. 24 participações em trabalhos de conclusão do curso de Agronomia. Contato: (84) 996595423.

Ben Dêivide de Oliveira Batista

Possui graduação em Engenharia Agronômica pela Universidade Federal Rural do Semiárido (2010), mestrado, doutorado e pós-doutorado em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (2012, 2016 e 2019, respectivamente). Atualmente é professor Adjunto - I, Classe C, pela Universidade Federal de São João Del-Rei. Áreas de pesquisa: Estatística computacional, Estatística experimental, Probabilidade e Estatística. 12 artigos publicados nacional/internacional, 16 trabalhos de congresso nacional/internacional, 5 pacotes R publicados. Contato: (35)991776992.

Eudes de Almeida Cardoso

Engenheiro Agrônomo, graduado em Engenharia Agronômica (1988) na Universidade Federal da Paraíba (UFPB). Mestre (1992) em Agronomia (Produção Vegetal) na Universidade Federal da Paraíba (UFPB). Doutor (2005) em Agronomia - Fitotecnia na Universidade Federal do Ceará (UFC). Atualmente, é Professor Titular do curso de Agronomia da Universidade Federal Rural do Semiárido (UFERSA). Leciona a disciplina de Horticultura na graduação e a disciplina Fruticultura Tropical na Pós-Graduação, possui dois capítulos de livros publicados, 48 artigos completos publicados/aceitos em revistas nacionais e internacionais e 33 resumo expandidos publicados . É responsável principal no Brasil pela seleção de uma nova cultivar porta-enxerto para o maracujazeiro amarelo tolerante à doenças do solo, em especial a fusariose, que está em fase de registro e proteção no Ministério da Agricultura. Contato: (84)988285000



ISBN 978-658831963-5

9 786588 319635

Pantanal Editora
Rua Abaete, 83, Sala B, Centro.
CEP: 78690-000
Nova Xavantina – Mato Grosso – Brasil
Telefone (66) 99682-4165 (Whatsapp)
https://www.editorapantanal.com.br
contato@editorapantanal.com.br