Estatística Básica

para os cursos de Ciências Exatas e Tecnológicas

Augustus Caeser Franke Portella
Ildon Rodrigues Do Nascimento
Anatércia Ferreira Alves
Gessiel Newton Scheidt

Estatística Básica

para os cursos de Ciências Exatas e Tecnológicas

EDUFT
UNIVERSIDADE FEDERAL DO TOCANTINS

Palmas-TO
2015

Projeto Gráfico, Revisão de Texto & Impressão
ICQ Editora Gráfica e Pré-Impressão Ltda.

Designer Responsável
Gisele Skroch

Impresso no Brasil
Printed in Brazil

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins – SISBIB

E79 Estatística básica para os cursos de ciências exatas e tecnológicas / Augustus Caeser Franke Portella; Ildon Rodrigues do Nascimento; Anatércia Ferreira Alves; Gessiel Newton Scheidt. – Palmas: Universidade Federal do Tocantins / EDUFT, 2015.

168 p.:il.

ISBN: 978-85-63526-93-9

1. Estatística Básica. 2. Ciências Exatas. 3. Ciências Tecnológicas. I. Portella, Augustus Caeser Franke. II. Nascimento, Ildon Rodrigues do. III. Alves, Anatércia Ferreira. III. Scheidt, Gessiel Newton. IV. Titulo

CDD 519.5

Prefácio

O objeto desta obra é o resultado de um esforço conjunto dos autores no sentido de oferecer um material didático sobre Estatística básica. Foi escrito, em especial, para servir de livro texto em cursos de graduação em ciências exatas com exemplos direcionados a seus interesses. Porém, utilizando uma bibliografia auxiliar, também pode ser utilizado tanto em cursos de pós-graduação, quanto, tomando as devidas precauções, em cursos para alunos e profissionais de outras áreas.

Dos oito capítulos que compõem o livro, os dois primeiros são dedicados à apresentação de conceitos básicos utilizados na estatística descritiva e teoria das probabilidades que têm crescido em nossas universidades em vários currículos da graduação, exigindo um curso introdutório de Probabilidade e Estatística Descritiva. Nos capítulos 3 e 4 é apresentada a importância das Técnicas de Amostragem e a formulação das hipóteses na teoria do Teste de Hipóteses. De forma a levar o aluno a uma utilização mais profissional, a partir do capítulo 5 é contemplada a Análise de Variância, Correlação e Regressão. No último capítulo é visto um conceito sobre o Controle de Qualidade com aplicações no âmbito das engenharias.

Uma vez que o livro foi escrito para estudantes que têm somente conhecimento elementar de cálculo em matemática, grande parte da teoria estatística deve ser aceita tal como os postulados.

Vários alunos de graduação e de pós-graduação tiveram acesso a este material, ou parte dele, em disciplinas ministradas

pelos autores nas Universidades Federais do Tocantins e do Paraná. Ficam registrados os agradecimentos a todos que contribuíram, de forma direta ou indireta, na produção deste material.

A versão atual deste livro não se encontra livre de erros e imperfeições. Desse modo, comentários, críticas e sugestões dos leitores são bem-vindos.

Os Autores

SUMÁRIO

"Posso todas as coisas naquele que me fortalece."
(FIL.: 4.13)

"Amo a História, se não a amasse não seria historiador. Fazer a vida em duas: consagrar à profissão, cumprida sem amor; reservar a outra à satisfação das necessidades profundas – algo de abominável quando a profissão que se escolheu é uma profissão de inteligência. Amo a história – e é por isso que estou feliz por falar daquilo que amo."
(FEBVRE, 1985, p. 28)

Apresentação

A Estatística é a ciência que lida com a coleta, o processamento e a disposição de dados (informação), atuando como ferramenta fundamental nos processos de soluções de problemas. Em outras palavras, a Estatística trata da coleta de dados informativos e da interpretação desses dados, facilitando o estabelecimento de conclusões confiáveis sobre algum fenômeno que esteja sendo estudado.

A Estatística Descritiva vem a ser a utilização dos dados para o conhecimento de processos e produtos de interesse econômico ou social. Dessa maneira, ela abrange diferentes campos do conhecimento e por isso é considerada uma área interdisciplinar, pois a utilização de suas técnicas fornecem parâmetros aos especialistas de diversas áreas.

Falando de forma mais específica, é também importante destacar que as técnicas estatísticas são muito úteis para o controle de qualidade de bens e serviços, e por esse motivo o conhecimento destes métodos está se tornando cada vez mais importante para engenheiros e demais profissionais.

A partir desta apresentação fica claro que um profissional treinado em Estatística terá maior facilidade em identificar um problema em sua área de atuação, determinar os tipos de dados que irão contribuir para sua análise, coletar esses dados e a seguir estabelecer conclusões e traçar um plano de ação para a solução dos problemas detectados.

Introdução

A importância de que se revestem os métodos que visam exprimir a informação contida numa grande massa de dados através de um número muito menor de valores ou medidas características é tal que a Estatística se ocupa em estudar os métodos que o permitam.

Deste modo, podemos definir Estatística como sendo a ciência que se preocupa com a coleta, organização, apresentação, análise e interpretação de dados. Didaticamente, podemos dividir a estatística em duas partes: a estatística descritiva e a inferência estatística. A estatística descritiva se refere à maneira de apresentar um conjunto de dados em tabelas e gráficos, e ao modo de resumir as informações contidas nestes dados a algumas medidas. Já a inferência estatística baseia-se na teoria das probabilidades para estabelecer conclusões sobre todo um grupo (chamado população), quando se observou apenas uma parte (amostra) desta população.

O campo de aplicação da Estatística estende-se a muitas áreas do conhecimento humano. Ledo engano ao pensarmos que nos dias atuais em função da facilidade que o advento dos computadores nos proporciona, na resolução de cálculos avançados e aplicações mirabolantes de processos sofisticados com razoável eficiência e rapidez, muitos pesquisadores consideram-se aptos a fazerem análises e inferências estatísticas sem um conhecimento mais aprofundado dos conceitos e teorias.

Interpretações equivocadas e muitas vezes errôneas são cometidas em nome da facilidade. Em sua essência, a Estatística é a ciência que apresenta processos próprios para coletar, apresentar e

interpretar adequadamente conjuntos de dados de qualquer natureza, sejam eles numéricos ou não, para que se tenha maior compreensão dos fatos que os mesmos representam e para ela ser bem usada é necessário conhecer os seus fundamentos e princípios, e acima de tudo que o pesquisador desenvolva um espírito crítico e jamais deixe de pensar.

Na área de Engenharia a aplicação é muito vasta, estando presente principalmente no estudo do controle de qualidade industrial, onde a técnica tem evoluído e proporcionado resultados importantes.

O estudo que será desenvolvido pode ser dividido em quatro partes: Estatística Descritiva, Probabilidades, Amostragem e Estatística Inferencial. A estatística descritiva trata da organização dos dados e descreve um conjunto de observações. A amostragem vai possibilitar o conhecimento das principais técnicas de obtenção de amostras. O estudo das Probabilidades ajuda no desenvolvimento dos métodos utilizados na Estatística Inferencial. E por fim, a Inferência estatística vai possibilitar a tomada de decisões acerca de populações.

1
Estatística Descritiva

A Estatística é a ciência que engloba conceitos de organização. Popularmente, a estatística está relacionada com tabelas e gráficos nos quais podemos representar os resultados, porém, ela tem assumido papel bem mais abrangente, nas últimas décadas, com importância cada vez maior no campo das ciências biológicas e agrárias.

Ela não pode ser vista apenas como mais uma disciplina, pois se trata principalmente de uma ferramenta auxiliar no raciocínio e análise dos resultados obtidos. É útil em pesquisas que exigem planejamento prévio para obter indicações de qualidade, através de dados que auxiliam na interpretação e conclusões sobre o fenômeno em questão.

Como uma grande parte do aprendizado vem da leitura, o estudante interessado em se informar a partir da literatura moderna, particularmente na área tecnológica, certamente se deparará com símbolos, termos e raciocínios estatísticos. Além disso, para cursos práticos, com práticas de laboratório ou campo, existem técnicas na obtenção de resultados que envolvem informações estatísticas, assim como o planejamento de experimentos, a publicação e o treino profissional.

Ocorre que pelo próprio conceito de ciência, a atividade de pesquisa científica deve começar a partir de um problema sobre o fenômeno, e não apenas de observações ou coleta de dados, embora o problema possa surgir a partir de observações, mensurações ou vivência sobre o assunto.

Evidentemente, tanto a parte de organização e descrição dos dados no que diz respeito à sua análise e interpretação são importantes. É razoável supor que para poder fazer a análise e interpretação dos dados observados, deva-se primeiramente proceder à sua organização e descrição.

Assim sendo, podemos dividir a ciência Estatística em duas partes: a Estatística Descritiva, que trabalha com a organização e descrição dos dados experimentais e a Estatística Indutiva ou Inferencial, que cuida de sua análise e interpretação.

A finalidade da Estatística Indutiva, cujas técnicas serão objetos deste trabalho, utilizam dois conceitos fundamentais: o de população, ou universo, e o de amostra.

Uma população ou universo é um conjunto de elementos com pelo menos uma característica comum. Essa característica comum deve delimitar inequivocamente quais os elementos que pertencem à população e quais os que não pertencem.

Em qualquer estudo estatístico queremos sempre pesquisar uma ou mais características dos elementos de alguma população. Os dados que observaremos, na tentativa de tirar conclusões sobre o fenômeno que nos interessa, serão referentes a elementos desta população.

Uma vez caracterizada perfeitamente a população, o passo posterior é o levantamento de dados acerca da característica (ou características) de interesse no estudo em questão.

Na maior parte das vezes, não é conveniente, ou mesmo nem é possível, realizar o levantamento dos dados referentes a todos os elementos da população. Devemos, então, limitar nossas observações a uma parte dela, isto é, a uma amostra proveniente desta população.

Uma amostra é, pois, um subconjunto de uma população, necessariamente finito, pois todos os seus elementos serão examinados para efeito da realização do estudo estatístico desejado.

Em suma, um estudo estatístico completo que recorra às técnicas da Estatística Indutiva irá envolver também, direta ou indiretamente, tópicos de Estatística Descritiva, Cálculo de Probabilidades e Estatística Inferencial.

1.1 Tipos de variáveis estatísticas

É necessário, inicialmente, que se defina qual(is) a(s) característica(s) dos elementos. Ou seja, não se trabalha estatisticamente com os elementos existentes, mas com alguma(s) característica(s) desses elementos. Por exemplo, os elementos a serem estudados podem ser a população de determinado micro-organismo, a densidade de uma floresta ou a produtividade de uma espécie vegetal, mas estaremos interessados em alguma característica específica, tal como a produção de metabólitos, cinética de crescimento, biomassa, produtividade, etc.

Trabalha-se, portanto, com os valores de uma variável (que é a característica de interesse), e não com os elementos originalmente considerados. A escolha da variável (ou variáveis) de interesse dependerá dos objetivos do estudo estatístico em questão. Essa característica (variável) poderá ser qualitativa ou quantitativa.

A variável será qualitativa quando resultar de uma classificação por tipos ou atributos, como nos seguintes exemplos:

População: cepa *Lactobacillus* em placas com determinado meio de cultura.

- Variável: característica da colônia (grande, pequena, aeróbias, anaeróbias).
- População: cepas submetidas à coloração de Gram.
- Variável: coloração (violeta ou vermelha, positiva ou negativa).
- População: Espécies de árvores.
- Variável: diâmetro do tronco.
- População: Cultivares de oleaginosas.
- Variável: Determinação de Cu, Fe, Mn, Zn, Ca, K e Mg visando a produção de óleo vegetal e biodiesel.
- População: Efluentes industriais.
- Variável: teor de ferro.

A variável será quantitativa quando seus valores forem expressos em números. Pode ser subdivida em:

1. Quantitativa discreta: pode assumir apenas valores pertencentes a um conjunto enumerável;
2. Quantitativa contínua: pode assumir qualquer valor em um intervalo de variação.

Por exemplo, o diâmetro de um halo de inibição medido em milímetros pode estar contido em um intervalo que depende do nível de precisão e do critério utilizado ao medir. Nesse caso, ao se medir o halo como sendo 2,68 mm deveremos considerar que o valor exato deste diâmetro será algo entre 2,675 e 2,685 mm.

Para atingir os objetivos da Estatística Descritiva os dados observados são muitas vezes sintetizados e apresentados em formas de tabelas ou gráficos, os quais irão fornecer informações rápidas e seguras a respeito das variáveis.

Uma das tabelas mais utilizadas é a distribuição de frequências. Os gráficos associados a ela são o gráfico de frequências (denominado histograma, para o caso de variáveis quantitativas contínuas), o polígono de frequências, o gráfico de frequência acumulada e o polígono de frequência acumulada.

Exemplos de variáveis quantitativas discretas:

- População: cultura de micro-organismos em determinado meio de cultura.
  Variável: número de colônias formadas (UFC).
- População: bactéria *Vibrio fischeri.*
  Variável: fator de toxidade.
- População: pessoas susceptíveis a determinada doença.
  Variável: número de pessoas.
- População: cultura de determinada hortaliça.
  Variável: número de parcelas com determinada característica.
- População: Densidade de uma floresta.
  Variável: número de árvores.

As variáveis quantitativas contínuas, geralmente, se utilizam de algum instrumento para medição. Temos os exemplos que se seguem:

- População: colônias de microrganismos submetidas a testes antibiogramas.
- Váriavel: medida do halo de inibição.
- População: crescimento microbiológico.
- Variável: tempo de crescimento.
- População: Tronco de teixo.
- Variável: diâmetro do cerne.
- População: Parcela de uma variedade.
- Variável: produção por hectare.

1.2 Distribuições de frequência

Distribuição de frequência é um método de se agrupar dados em classes, de modo a fornecer a quantidade (e/ou a percentagem) de dados em cada classe.

Para se obter informações de interesse sobre o fenômeno em estudo, deve-se agrupar as observações em tabelas ou gráficos convenientemente construídos. Com isso, podemos resumir e visualizar um conjunto de dados sem precisar levar em conta os valores individuais. Uma distribuição de frequência (absoluta ou relativa) pode ser apresentada em tabelas ou gráficos.

Muitas vezes os gráficos são elaborados utilizando-se as frequências dos valores da variável. Para tal, necessitamos definir alguns conceitos importantes.

Considere a variável discreta x, representando as idades dos alunos de uma determinada classe escolar. Foram entrevistados 25 alunos fornecendo os seguintes valores para x:

23 – 21 – 22 – 25 – 21
22 – 26 – 27 – 28 – 24
21 – 22 – 24 – 25 – 25
23 – 27 – 28 – 31 – 31
30 – 23 – 31 – 30 – 23

1.2.1 Dados brutos

É o conjunto de valores obtidos após a crítica dos dados coletados. A crítica dos dados consiste na observação em busca de falhas e imperfeições, visando eliminar erros grosseiros atribuídos a fatores que podem afetar o resultado, tais como medidas, cálculos, escala e instrumento descalibrado.

1.2.2 Rol

É o arranjo dos dados brutos em determinada ordem (crescente ou decrescente).

21 – 21 – 21 – 22 – 22
22 – 23 – 23 – 23 – 23
24 – 24 – 25 – 25 – 25
26 – 27 – 27– 28 – 28
30 – 30 – 31 – 31 – 31

Tamanho da amostra n = 25

1.2.3 Amplitude total

$$AT = x_{max} - x_{min} \quad (1.1)$$

X_{max} = valor máximo da variável x.

X_{min} = valor mínimo da variável x.

AT = 31 – 21
AT = 10

1.2.4 Número de classes (K)

$$k = \sqrt{n} \tag{1.2}$$

n = número total dos elementos observados.

Ou usando a fórmula de Sturges

$$K = 1 + 3{,}3.\log n \tag{1.3}$$

Exemplo:

$$K = \sqrt{25} \Rightarrow K = 5$$

ou

$$K = 1 + 3{,}3.1{,}4 = 5{,}6 \Rightarrow 6$$

Obs.: Trabalhar com menos de 5 linhas em uma tabela afeta a frequência.

1.2.5 Amplitude dos dados

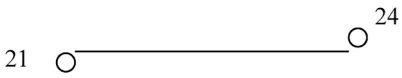
$$h = \frac{AT}{K} \tag{1.4}$$

h = amplitude da classe, AT = amplitude total, K = n° de classes.

$$h = \frac{15}{5} :. \quad h = 3$$

O arredondamento para cima ou parcial (0,5 ou 1,0) 2,2 ≅ 2,5

1.2.6 Limite das Classes



De 21 inclusive a 24 exclusive

1.2.7 Pontos Médios de classe (x_i)

$$x_i = \frac{l_i + l_s}{2} \qquad (1.5)$$

l_i= limite inferior da classe
l_s = limite superior da classe

1.2.8 Frequências absolutas e frequências relativas

Definimos frequência de um valor de uma variável (qualitativa ou quantitativa) como sendo o número de vezes que aquele valor se repete no conjunto de dados experimentais. Usaremos a notação f_i para representar a frequência do *i*-ésimo valor observado.

$$\sum_{i=1}^{k} fi = n \tag{1.6}$$

k = número de diferentes valores existentes na variável.

Do mesmo modo, podemos definir frequência relativa de um valor observado como sendo a relação:

$$p_i = Fr = \frac{f_i}{n} \tag{1.7}$$

Sendo p_i a frequência relativa ou proporção do i-ésimo elemento observado.

Verifica-se que:

$$\sum_{i=0}^{k} p_i = n \tag{1.8}$$

Onde k é o número de diferentes valores existentes na variável.

Exemplo:

Seja o conjunto de dados contido na tabela 1.1 que representa o número de Unidades Formadora de Colônias (UFC/mL) de um microrganismo específico obtidas em 50 amostras extraídas em uma lagoa de resíduos industriais durante 5 dias, diluídas e inoculadas em placas de Petri com meio seletivo.

Tabela 1.1 – Distribuição de Frequência

Tempo (dias)	Frequência (UFC/mL) f_i	Frequência relativa *fr*
0	15	0,30
1	10	0,20
2	13	0,26
3	06	0,12
4	03	0,06
5	03	0,06
Total	50	1,00

As frequências são:

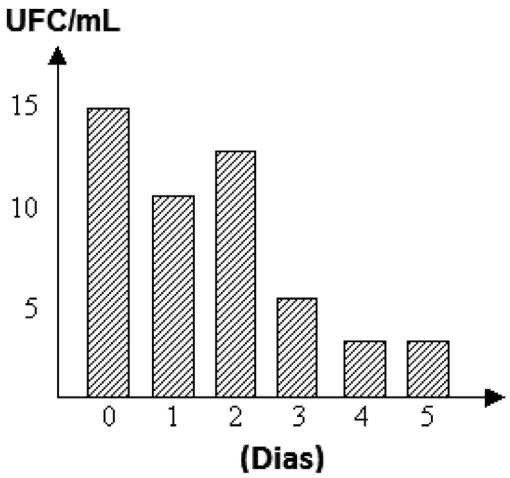
- f_0 = 15 (corresponde ao valor 0)
- f_1 = 10 (corresponde ao valor 1)
- f_2 = 13 (corresponde ao valor 2)
- f_3 = 06 (corresponde ao valor 3)
- f_4 = 03 (corresponde ao valor 4)
- f_5 = 03 (corresponde ao valor 5)

Chamamos de distribuição de frequência à associação das frequências aos respectivos valores observados. Portanto, a representação acima caracteriza uma distribuição de frequência.

1.3 Distribuição de frequência para variáveis quantitativas discretas

A Figura 1.1 representa a distribuição de frequência para a variável discreta "Unidade Formadora de Colônias (UFC)". A representação gráfica de uma distribuição de frequência de uma variável quantitativa discreta é chamada de histograma de frequência. Utilizando o exemplo, temos o seguinte histograma de frequência:

Figura 1.1 – Histograma de frequência



Outra representação utilizada é o histograma das frequências acumuladas e frequências relativas acumuladas. Tomando-se os dados do exemplo anterior podemos calcular a frequência, frequência acumulada e frequência relativa acumulada dos diversos valores. Esse cálculo está ilustrado na Tabela 1.2.

Tabela 1.2 – Frequência absoluta, frequência relativa e frequência relativa acumulada

Tempo (dias)	Freq. (UFC/mL) Fi	Freq. Relativa Fr	Freq. Relativa Acumulada
0	15	0,30	0,30
1	10	0,20	0,50
2	13	0,26	0,76
3	06	0,12	0,88
4	03	0,06	0,94
5	03	0,06	1,00
Total	**50**	**1,00**	**-**

Com os dados da Tabela 1.2 podemos construir o gráfico de frequência relativa (Figura 1.2) e frequência relativa acumulada (Figura 1.3).

Figura 1.2 – Histograma de frequência relativa

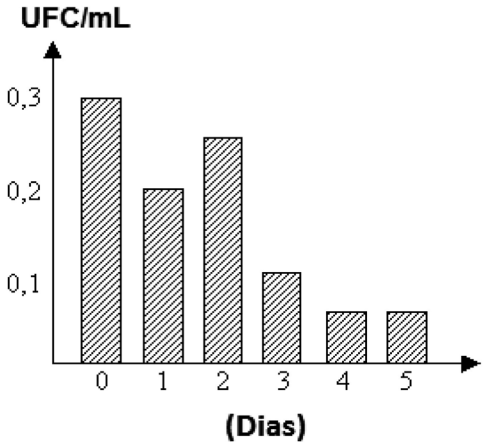
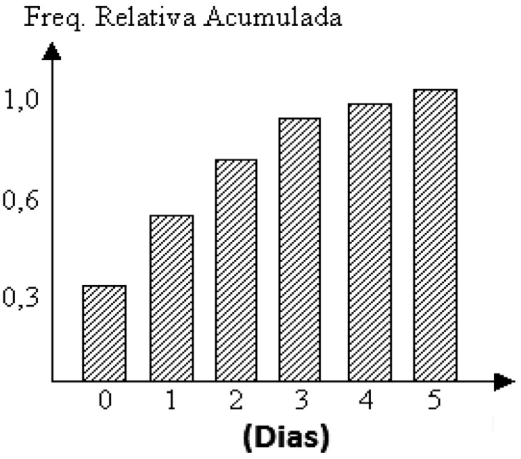


Figura 1.3 – Histograma de frequência relativa acumulada



1.4 Distribuição de frequência para variáveis quantitativas contínuas

As variáveis quantitativas contínuas diferem um pouco das discretas na sua forma de representação gráfica. Para entender essa diferença, temos que lembrar que as variáveis contínuas, por definição, têm os seus valores definidos num intervalo contínuo dos números reais.

Devemos utilizar a distribuição de frequência contínua na representação de uma série de valores, quando o número de elementos distintos da série for grande. Neste caso, os dados serão agrupados por faixas de valores (intervalos).

Portanto, não tem sentido falar em frequência de repetição de um determinado valor, pois os valores raramente se repetem.

Algumas indicações na construção de distribuição de frequências são:

- Na medida do possível, as classes deverão ter amplitudes iguais.
- Escolher os limites dos intervalos entre duas possíveis observações.
- Escolher limites que facilitem o agrupamento.
- Marcar os pontos médios dos intervalos.

A Tabela 1.3 representa uma distribuição de frequência para a variável "Diâmetro do halo de inibição".

Tabela 1.3 – Diâmetro (mm) dos halos de inibição de 50 cepas submetidas à bacteriocina nisina inoculadas em placas de Petri

Diâmetro (mm)	Frequência fi	x_i	Freq. Relativa Fr	Freq. Acumulada fac
151 ├ 159	2	155	0,04	2
159 ├ 167	11	163	0,22	13
167 ├ 175	18	171	0,36	31
175 ├ 183	10	179	0,20	41
183 ├ 191	8	187	0,16	49
191 ├ 199	1	195	0,02	50
Total	**50**		**1,00**	

Intervalos de classes: É representado pelo maior e o menor valor da classe e o símbolo (├) define o intervalo de uma classe.

Exemplo: 151 ├ 159.
Inclui o limite inferior (151) e exclui o limite superior (159);
O símbolo (|-|) inclui ambos os limites, superior (159) e inferior (151);
O símbolo (-|) só inclui o limite superior (159).

Limites da classe: Representa os números extremos de cada intervalo.
Exemplo: o limite inferior da 1ª classe é 151 e o limite superior da 1ª classe é 159.

Ponto médio de classe: ponto intermediário do intervalo de classe.
Exemplo: Ponto médio da 1ª classe é 155.

$$x_i = \frac{151+159}{2}$$

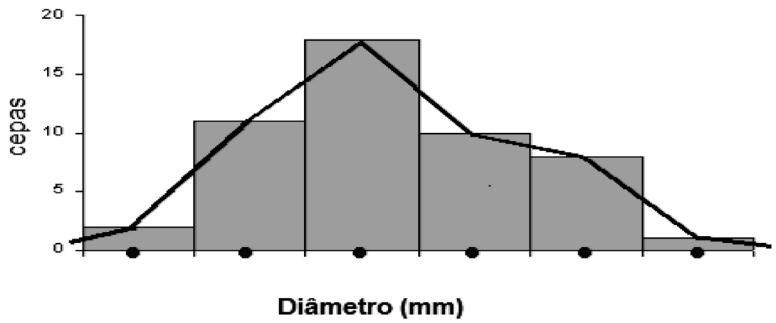
Amplitude do intervalo de classe: é a diferença entre os limites reais, superiores e inferiores de cada classe.

Exemplo: amplitude da 1ª classe é 8.

$$h_1 = 159 - 151$$

Com os dados da Tabela 1.3 podemos construir o gráfico de frequências do mesmo modo que fizemos para as variáveis discretas. A diferença mais importante é que, agora, as frequências são associadas a intervalos de valores (classes de frequências) e não mais a valores individuais da variável em estudo (Figura 1.4). O gráfico consiste em um conjunto de retângulos, com centro no ponto médio e larguras iguais aos intervalos das classes, e áreas proporcionais às frequências das classes.

Figura 1.4 – Histograma e polígono de frequência



A seguir está mostrado o histograma correspondente aos dados do exemplo acima, e o polígono de frequência, que é o gráfico obtido unindo-se os pontos médios dos patamares do histograma (Figura 1.4).

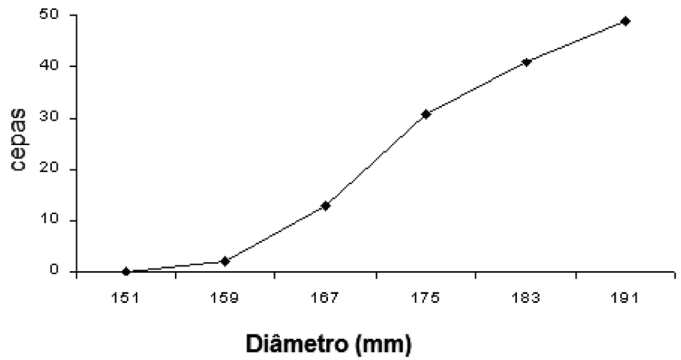
O próximo gráfico é o polígono de frequência acumulada. Ele é construído unindo-se a frequência acumulada ao final de cada classe de frequência (Tabela 1.4). Pode ser construído também com a frequência relativa acumulada e, neste caso, ele se chamará polígono de frequência relativa acumulada. O primeiro está mostrado na Figura 1.5.

Tabela 1.4 – Diâmetro (mm) dos halos de inibição de 50 cepas submetidas à bacteriocina nisina inoculadas em placas de Petri

Diâmetro (mm)	Cepas
<151	0
<159	2
<167	13
<175	31
<183	41
<191	50

Nota: < menor que

Figura 1.5 – Polígono de frequência acumulada



1.5 Medidas de tendência central

As medidas de tendência central são estimadores utilizados para definir o centro de equilíbrio de uma distribuição de frequência de uma variável.

1.5.1 Média

A média de um conjunto de *n* números x_1; x_2; ...; x_n é representada por $\bar{x}$ e definida por:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{1.9}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n} \tag{1.10}$$

De modo geral, para dados agrupados, se x_1, x_2, ..., x_n ocorrerem com as frequências f_1, f_2, ..., f_n, respectivamente, a média aritimética será dada por:

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum f_x}{\sum f} = \frac{\sum f_x}{n} \tag{1.11}$$

Tabela 1.5 – Cálculo da Média

Classes	Frequência f	x	f_x
151 ├ 159	2	155	310
159 ├ 167	11	163	1793
167 ├ 175	18	171	3078
175 ├ 183	10	179	1790
183 ├ 191	8	187	1496
191 ├ 199	1	195	195
Total	50	-	8662

$$\overline{X} = \frac{\sum f_x}{n} = \frac{8662}{50} = 173{,}24$$

Para dados distribuídos em classes, os valores x_1, x_2, ..., x_n corresponderão aos pontos médios das n classes que podem ser definidos como a média aritimética entre os limites inferior (l_i) e superior (l_s) da classe i considerada, ou seja,

$$x_i = \frac{l_i + l_s}{2} \qquad (1.12)$$

As propriedades principais da média aritimética são:

a) A soma algébrica dos desvios de um conjunto de números em relação a média aritmética deste conjunto é zero.

b) A soma dos quadrados dos desvios de um conjunto de números, em relação a um número qualquer a, é um mínimo se e somente se, $a = \bar{x}$.

1.5.2 Mediana

A mediana de um conjunto de números, ordenados em ordem de grandeza, é o valor médio (*n* impar) ou a média aritmética dos dois valores centrais (*n* par). A mediana é útil quando o conjunto de dados é muito influenciado pelos extremos, refletindo com mais fidelidade à tendência central.

Exemplos:

- {3; 4; 4; 5; 6; 8; 8; 8; 10} O elemento de ordem (n+1)/2 é igual a 6;

- {5; 6; 7; 9; 11; 12; 13; 17} A mediana é dada pela média aritimética dos dois valores de ordem n/2 e (n/2)+1 que é igual a 10.

$$Me = \frac{9+11}{2} = 10$$

No caso de dados agrupados em classes de frequências, a mediana Me pode ser calculada pela expressão (deduzida a partir do histograma de frequências).

$$Me = l_i + \frac{P - f_a}{f_{Me}} h \qquad 1.13)$$

Onde: l_i é o limite inferior da classe mediana (em uma distribuição de frequências chama-se classe mediana a classe que contém a mediana).

- P = n/2 é a posição da classe mediana;
- F_a é a frequência acumulada da classe vizinha anterior à classe mediana;
- F_{Me} é a frequência da classe mediana;
- h é a amplitude do intervalo da classe mediana.

1.5.3 Moda

A moda é o valor que ocorre com mais frequência. A moda pode não existir e, mesmo que exista, pode não ser única.

Exemplos:
{1; 1; 3; 3; 5; 7; 7; 7; 11; 13} tem moda 7
{3; 5; 8; 11; 13; 18} não tem moda (amodal)
{3; 5; 5; 5; 6; 6; 7; 7; 7; 11; 12} tem duas modas 5 e 7 (bimodal)

No caso de dados agrupados em classes de frequências, a moda (Mo) pode ser calculada pela expressão:

$$Mo = l_i + \frac{f_p}{fa + fp} h \qquad (1.14)$$

- Onde: l_i é o limite inferior da classe modal (ou seja, a classe de maior frequência);
- F_p é a frequência de classe imediatamente posterior à classe modal;
- F_a é a frequência de classe imediatamente inferior à classe modal;
- h é a amplitude de intervalo da classe modal.

Tabela 1.6 – Cálculo da Mediana e da Moda

Classes	**Frequência** f	fa
151 ├ 159	2	2
159 ├ 167	11	13
167 ├ 175	18	31
175 ├ 183	10	41
183 ├ 191	8	49
191 ├ 199	1	50
Total	**50**	_

Cálculo da mediana (Me) pela expressão (1.13)

$$Me = 167 + \frac{25-13}{18}8 \quad = 172{,}33$$

Cálculo da Moda (Mo) pela expressão (1.14)

$$Mo = 167 + \frac{11}{10+11}8 = 171{,}19$$

1.6 Medidas de dispersão

O grau aos quais os dados numéricos tendem a dispersar-se em torno de um valor médio chama-se variação ou dispersão dos dados.

Considere os tempos de reação de três substâncias para executar certa titulação. Foram tomados os tempos (em segundos) de cinco operações para cada substância, fornecendo os resultados:

Substância A: 5, 5, 5, 5, 5;
Substância B: 5, 3, 9, 5, 3;
Substância C: 3, 4, 5, 8, 5;

Calculando a média aritmética para cada titulação, obtém-se:

$$\overline{X}_A = \overline{X}_B = \overline{X}_C = 5s$$

Ou seja, o tempo médio para executar a operação é o mesmo para as três substâncias. Mas observando mais detalhadamente, os tempos se distribuem diferentemente em relação ao tempo médio (5s).

Para uma análise quantitativa dessa maior ou menor variação (ou dispersão) do conjunto de valores em torno do valor médio, devemos estudar as medidas de dispersão.

As medidas mais comuns são: amplitude total, desvio médio, desvio padrão e variância.

1.6.1 Amplitude Total

É a diferença entre o maior e o menor valor (expressão 1.1). Exemplo: a amplitude total de {4; 7; 9; 11; 11; 15; 20} é 16 (ou seja, 20 – 4).

1.6.2 Variância (s^2)

A variância (s^2) de um conjunto de n valores x_1, x_2, ..., x_n é dada pela média aritmética dos quadrados dos desvios desses valores em relação à sua média aritmética, ou seja,

$$\delta^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n} \qquad (1.15)$$

Para dados agrupados temos:

$$\delta^2 = \frac{\sum_{i=1}^{k} f_i \left(x_i - \bar{x} \right)^2}{n} \qquad (1.16)$$

Tabela 1.6 – Cálculo da Variância

Classes	Frequência f	x	$f\ (x-x)^2$
151 ├ 159	2	155	332,70
159 ├ 167	11	163	104,86
167 ├ 175	18	171	5,02
175 ├ 183	10	179	33,18
183 ├ 191	8	187	189,34
191 ├ 199	1	195	473,50
Total	**50**	**-**	**1138,59**

A média calculada utilizando a tabela 1.5 forneceu 173,24, e considerando os dados como uma população obtém-se:

$$\delta^2 = \frac{1138,59}{50} = 22,77$$

Obs.: Quando a variância corresponde aos dados de uma amostra, é em geral calculado com o divisor (n-1) ao invés de n, para que se tornem estimadores não tendenciosos. Neste caso geralmente utiliza-se as letras s e s^2 para representar o desvio padrão e a variância, respectivamente.

A variância para os dados da substância B do exemplo 1.8 será dada por:

$$s_B^2 = \frac{(5-5)^2 + (3-5)^2 + (9-5)^2 + (5-5)^2 + (3-5)^2}{5-1} = 6$$

1.6.3 Desvio Padrão (s)

Como a unidade da variância é expressa pelo quadrado da variável em estudo, é inconveniente o uso prático da variância. Para contornar o problema da unidade, define-se o desvio padrão. O desvio padrão (s) é definido como a raiz quadrada positiva da variância.

O desvio padrão de X_1; X_2; ...; X_N é dado por:

$$\delta = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{n}} \qquad (1.17)$$

No caso do exemplo anterior, o desvio padrão será:

$$\delta = \sqrt{22{,}77} = 4{,}77$$

1.6.4 Coeficiente de Variação

O efeito da variação ou dispersão em relação à média é medido pela dispersão relativa, que é definida por:

$$DR = \frac{Disp.\ absoluta}{Média} \qquad (1.18)$$

Se a dispersão absoluta for o desvio padrão, a dispersão relativa é denominada Coeficiente de Variação (*CV*), que pode ser representado por:

$$CV = \frac{s}{\overline{X}} \qquad (1.19)$$

Obs: O Coeficiente de variação deixa de ser útil quando o $\overline{X}$ está próximo de zero.

O coeficiente de variação pode ser expresso em porcentagem e é uma medida adimensional de dispersão, sendo definida como o coeficiente entre o desvio padrão (s) e a média ($\overline{X}$). Assim, quando se deseja comparar dois conjuntos de dados com médias

diferentes, deve-se utilizar o coeficiente de variação, pois o mesmo leva em consideração a ordem de grandeza dos mesmos.

Supondo que um conjunto de dados tem média $\overline{X}_1 = 20$ mm e desvio padrão s = 2 mm, enquanto um segundo conjunto tem média $\overline{X}_2 = 50$ mm e desvio padrão s = 4 mm.

$$CV_1 = \frac{2}{20} = 0{,}10 \text{ ou } 10\%$$

$$CV_2 = \frac{4}{50} = 0{,}08 \text{ ou } 8\%$$

Nota-se que em termos absolutos a dispersão do primeiro conjunto é menor que a do segundo, mas em termos relativos, o primeiro conjunto apresenta uma dispersão maior, ou seja, $CV_1 > CV_2$.

1.7 Curvas de frequência

As curvas de frequência determinam o grau de afastamento ou de desvio de uma distribuição em torno da média (assimetria). Quantitativamente, o grau de desvio ou afastamento pode ser determinado pelas medidas denominadas de coeficientes do momento de assimetria e coeficeinte de assimetria de Pearson.

O coeficiente do momento de assimetria (a_3) é uma medida adimensional definida como o quociente entre o terceiro momento centrado na média (m_3) e o cubo do desvio padrão, ou seja:

$$a_3 = \frac{m_3}{s^3} \qquad (1.20)$$

O momento de ordem r (m_r) centrado na média de um conjunto de n valores x_1, x_2, ..., x_n é definido pela quantidade:

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n} \tag{1.21}$$

Para o caso de dados agrupados em classes de frequências, a expressão (1.21) fica sendo:

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n} \tag{1.22}$$

Para r = 1 (momento de primeira ordem) verifica-se que:

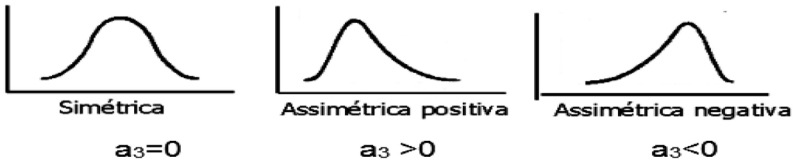
$$m_1 = 0,$$

enquanto que para r = 2:

$$m_2 = s^2$$

Para $a_3 = 0$ tem-se uma distribuição simétrica, caso contrário, a distribuição é dita assimétrica. Quando $a_3 < 0$, a distribuição é dita alongada à esquerda, sendo denominada de assimétrica negativa, enquanto que, para $a_3 > 0$, a distribuição é alongada à direita, sendo denominada assimétrica positiva. Na figura 1.6 podemos verificar os três casos.

Figura 1.6 – Curvas de frequência.



Como exemplo, considere os dados da tabela 1.7, cujo desvio padrão resultou em s = 4,77 e o terceiro momento centrado na média (m_3) será calculado a seguir.

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n} \tag{1.23}$$

O coeficiente do momento de assimetria será

$$a_3 = \frac{m_3}{S^3} \tag{1.24}$$

Tabela 1.7 – Cálculo do terceiro momento

Classes	**Frequência** *f*	***x***	$f\,(x - \bar{x})^3$
151 ├159	2	155	-6068,40
159 ├167	11	163	-1073,74
167 ├175	18	171	-11,24
175 ├183	10	179	191,10
183 ├191	8	187	2605,29
191 ├199	1	195	10303,31
Total	**50**	**-**	**5946,31**

$$m_3 = \frac{5946,31}{50} = 118,93$$

O coeficiente do momento de assimetria será:

$$a_3 = \frac{118,93}{108,53} = 1,09$$

O coeficiente de assimetria de Pearson (A) é outra medida adimensional de assimetria, sendo definida pela expressão:

$$A = \frac{\bar{x} - m_O}{s} \qquad (1.25)$$

No caso dos dados da tabela 1.7, onde:

$\bar{x}$=173,24, Mo=171,19 e s= 4,77

Temos pela equação 1.14:

$$A = \frac{173,24 - 171,19}{4,77} = 0,43$$

Como os valores da a_3 e A estão acima de zero, a distribuição da Tabela 1.7 possui uma assimetria positiva.

1.8 Curtose

A curtose é definida como o grau de achatamento de uma distribuição, considerado usualmente em relação à distribuição normal (objeto de estudo do capítulo 2). Com relação ao achatamento, a distribuição normal é dita mesocúrtica. As distribuições mais achatadas que o normal são ditas platicúrticas, enquanto que as menos achatadas são ditas leptocúrticas.

O coeficiente do momento de curtose (a_4) é definido pelo quociente entre o quarto momento centrado na média e o quadrado da variância, ou seja,

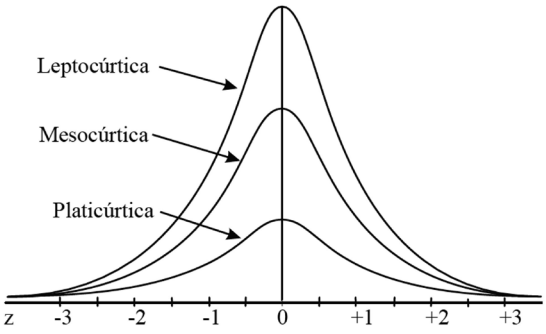
$$a_4 = \frac{m_4}{(s^2)^2} = \frac{m_4}{s^4} \tag{1.26}$$

O coeficiente do momento de curtose é uma medida adimensional de curtose, sendo $a_4 = 3$ para a distribuição normal, $a_4 < 3$ para as distribuições platicúrticas e $a_4 > 3$ para as distribuições leptocúrticas.

Na prática, a curtose só se aplica para distribuições simétricas ou aproximadamente simétricas.

A figura 1.7 mostra os três casos de curtose, utilizando a representação pelas curvas de frequências.

Figura 1.7 – Distribuições quanto à curtose



Exercícios

1) Em uma fermentação por batelada a volume constante, foram obtidos os seguintes dados experimentais (g/L)

6,0 – 0 – 2,0 – 6,5 – 5,0 – 3,5 – 4,0 – 7,0 – 8,0 – 7,0

4,5 – 0 – 6,5 – 6,0 – 2,0 – 5,0 – 5,5 – 5,0 – 7,0 – 1,5

5,0 – 5,5 – 4,0 – 4,5 – 4,0 – 1,0 – 5,5 – 3,5 – 2,5 – 4,5

Determinar:

a) As distribuições de frequências para dados agrupados em classes de igual amplitude, calcular a AT, K e h.
b) O maior e o menor grau (valores) da amostra.
c) Qual a porcentagem dos volumes que tiveram produções inferiores ao limite superior da terceira classe.
d) Qual o limite superior da segunda classe.
e) Qual o ponto médio da quarta classe.
f) Qual a frequência relativa da terceira classe.
g) Os gráficos.

2) Dada à amostra:

3 – 4 – 4 – 5 – 7 – 6 – 6 – 7 – 7 – 4 – 5 – 5 – 6 – 6 – 7 – 5– 8 – 5 – 6 – 6.

Pede-se:

a) Construir a distribuição de frequência para dados agrupados, mas não em classes.
b) Determinar as frequências relativas.
c) Determinar as frequências acumuladas.

d) Qual a porcentagem de elementos maiores que 5.

e) Construir o gráfico das frequências absolutas.

3) Considere os dados obtidos pela leitura de absorbância do crescimento de 60 cepas **(dados em nanômetros (nn)**

0,151-0,152-0,154-0,155-0,158-0,159-0,159-0,160-0,161-0,161-0,166-0,166-0,166-0,167-0,167-0,167-0,167-0,167-0,168-0,168-0,169-0,169-0,169-0,169-0,169-0,170-0,170-0,170-0,170-0,170-0,173-0,173-0,174-0,174-0,174-0,175-0,175-0,175-0,175-0,176-0,176-0,176-0,176-0,177-0,177-0,180-0,181-0,181-0,181-0,182-0,182-0,182-0,183-0,185-0,185-0,186-0,187-0,188-0,190-0,190.

Pede-se:

a) Amplitude total.

b) O número de classes.

c) A amplitude das classes.

d) A distribuição de frequência para dados agrupados em classes conforme valores obtidos nos itens a, b e c (a primeira classe deve começar com o menor valor da amostra).

e) As frequências relativas, frequências acumuladas e os pontos médios.

f) Os gráficos (Histograma, Polígono de frequência, polígono de frequência acumulada).

4) O número de colônias de um determinado microrganismo em uma placa de Peters, de acordo com o seu ciclo de vida, está representado na tabela a seguir:

Classe	Frequência
0 ├ 6	280
6 ├ 12	140
12 ├ 18	60
18 ├ 24	15
24 ├ 30	5

a) Qual a duração média e a mediana do ciclo de vida?

b) Encontre a variância e o desvio padrão amostral da duração do ciclo de vida.

5) Em três reatores de batelada, testa-se a produção de cada cepa tomando-se uma amostra de 100 ml e determinando-se o pH necessário para a sobrevivência de cada cepa. Os resultados de cada teste são os seguintes:

A	B	C
7,0	6,0	5,0
1,8	1,5	1,0

a) Qual reator apresenta a menor variação absoluta no pH?

b) E a maior variação no pH?

6) O quadro abaixo apresenta os resultados do inventário das áreas de manejo florestal em assentamentos no Brasil.

Projeto Assentamento	Inventário Florestal				Área de Manejo	Produção Anual	
	N° de Parcelas	Estoque Total (st/ha)	Estoque explorável (st/ha)	N° de espécies	(ha)	Lenha (st)	Carvão (Sacos)
Brejinho	13	145,22	120,16	40	200	1602	4806
Pipoca	12	138,40	127,40	18	100,8	856	2569

Projeto Assentamento	Inventário Florestal				Área de Manejo	Produção Anual	
	N° de Parcelas	Estoque Total (st/ha)	Estoque explorável (st/ha)	N° de espécies	(ha)	Lenha (st)	Carvão (Sacos)
Sítio do Meio	15	144,50	132,30	29	120	1058	3174
Barra Nova	13	169,90	107,30	22	45,2	333	999
Catolé	18	117,20	115,40	36	213	1639	4917

st/ha = metro de lenha empilhada por hectare

a) Qual o total de hectares de manejo dos cinco assentamentos com sua produção em metros esteres de lenha ou sacas de carvão?

b) Qual o possível rendimento bruto baseado no preço de comercialização praticado nas respectivas regiões, que gira em torno de R$ 4,00 por saca de carvão de 25 kg (2006-2007)?

c) Calcule as médias e os coeficientes de variação dos itens constantes no inventário florestal.

d) Faça o histograma.

e) Pode-se concluir então que o manejo florestal sustentado da caatinga representa uma alternativa viável para ser desenvolvida em Projetos de Assentamento rurais no semiárido nordestino?

2
Probabilidade

A teoria moderna das probabilidades constitui a base de um dos ramos de maior aplicação nas ciências, a Estatística. É conveniente dispormos de uma medida que exprima a incerteza presente em afirmações, tais como: "É possível que chova amanhã", ou "Não há chance de vitória", em termos de uma escala numérica que varie do impossível ao certo. Essa medida é a probabilidade.

A teoria de probabilidades lida com a realização de experimentos, naturais ou planejados pelo homem, cujos resultados não podem ser previstos com exatidão. As primeiras aplicações do cálculo das probabilidades ocorreram em função de jogos de azar, no século XVI. As pessoas se utilizavam do conhecimento da teoria das probabilidades para planejar estratégias de apostas.

Embora os resultados de um experimento, realizado sob condições uniformes e não tendenciosas, não possam ser antecipados com exatidão, é possível estabelecer o conjunto que contém todos os resultados possíveis ou esperados de tal experimento.

Para se obter informações de uma amostra de dados que sejam úteis à tomada de decisões no planejamento e projeto de sistemas especialistas é necessário estabelecer um modelo matemático que contenha os principais elementos do processo que determinou a ocorrência daquelas observações. Tal modelo deve ser probabilístico pela impossibilidade de se sintetizar em um conjunto de equações a lei que descreve rigorosamente a variação de um certo fenômeno.

Um modelo probabilístico, embora seja incapaz de prever com exatidão a data e a magnitude de um fenômeno climático, por

exemplo, revela-se muito útil no estudo do regime local de chuvas, especificando com que probabilidade uma determinada precipitação irá ser igualada ou superada, em um ano qualquer.

O presente capítulo tem por objetivo estabelecer os princípios da teoria de probabilidades, necessários à construção de modelos probabilísticos.

2.1 Experimentos determinísticos e aleatórios determinísticos

2.1.1 Experimentos determinísticos

São aqueles que repetidos em idênticas condições apresentam os mesmos resultados.

Exemplo: Ao nível do mar, a água entra em ebulição a 100 °C.

2.1.2 Experimentos Aleatórios

São aqueles que repetidos em idênticas condições podem produzir resultados diferentes. Embora não saibamos qual o resultado que irá ocorrer num experimento, em geral, conseguimos descrever o conjunto de todos os resultados possíveis.

Exemplos:

1. Contagem do número de células por mL da suspensão utilizando câmara de Neubauer.
2. Avaliação do grau de contaminação ambiental.
3. Determinação do ciclo de vida de um determinado microrganismo.
4. Incidência de pragas em determinadas culturas.

A teoria das probabilidades estuda a forma de estabelecermos as possibilidades de ocorrência num experimento aleatório.

Uma aplicação particularmente importante é quando um pesquisador conduz um experimento, a fim de comparar os efeitos de diferentes tratamentos (variações de um fator a ser estudado).

Para se estimar os efeitos dos tratamentos e também para executar os testes estatísticos é necessário o uso de repetições (aplicações do mesmo tratamento em diversas unidades experimentais e que formará a amostra de estudo), por meio das quais vamos ter a possibilidade de calcular a variabilidade dos dados, ou seja, a variância.

2.2 Espaço amostral e evento

2.2.1 Espaço amostral

Chamamos de espaço amostral, e indicamos por S ou Ω, um conjunto formado por todos os possíveis resultados de um experimento aleatório.

2.2.2 Evento

Em termos de conjunto, um evento é um subconjunto de resultados do experimento, ou seja, é um subconjunto de *{S}*. Os eventos são denotados por letras maiúsculas (A, B, C...).

Exemplos:

(1) Lançamento de duas moedas: S = {kk, kc, ck, cc}.
(cara, cara), (cara, coroa), (coroa, cara), (coroa, coroa).

(2) Sexo de acordo com a ordem de nascimentos de dois filhos de um casal:
S = {mm, mf, fm, ff}.
(masculino, masculino), (masculino, feminino), (feminino, masculino),
(feminino, feminino).

Do ponto de vista prático, os eventos são as sentenças que podemos formular sobre nosso experimento. Assim, desejamos definir formas de manipular, ou seja, de operar essas sentenças. As três operações básicas são:

a) ***União* (∪):** A união de dois conjuntos quaisquer E e F conterá todos os elementos de E e de F, incluindo os elementos que sejam comum aos dois ou não.

b) ***Intersecção* (∩)**: A intersecção de dois conjuntos quaisquer E e F conterá os elementos comuns a E e F.

c) ***Complementar* (A^c)**: O evento complementar ao evento A é o conjunto dos elementos do espaço amostral que não pertencem a A.

d) Na terminologia da teoria de conjuntos, o conjunto vazio é o conjunto composto por nenhum elemento, que denotaremos por ∅. Esse conjunto está contido em qualquer outro evento do espaço amostral.

A probabilidade é uma forma de atribuirmos "pesos" relativos à ocorrência dos eventos.

Se os elementos de um espaço amostral $S = \{e_1, e_2, ..., e_n\}$ (finito) são equiprováveis, isto é, todos os elementos do espaço amostral têm o mesmo "peso" (probabilidade) de ocorrer, temos que:

$$P(e_i) = \frac{1}{n} \qquad (2.1)$$

onde n é o número total de elementos equiprováveis.

2.3 Definição clássica de Probabilidades

Dado um experimento aleatório, sendo S o seu espaço amostral, vamos admitir que todos os elementos de S tenham a mesma chance de acontecer, ou seja, que S é um conjunto equiprovável.

a) Definimos probabilidade de um evento A ($A \subset S$) (A está contido em S) ao número real $P(A)$, tal que:

$$P(A) = \frac{n^o \text{ de resultados favoráveis a } A}{n^o \text{ de resultados possíveis}} = \frac{n(A)}{n(S)} \qquad (2.2)$$

Exemplo: Considerando o lançamento de um dado, pede-se:

a) A probabilidade do evento A "obter um número par na face superior".

   Temos:
   $S = \{1; 2; 3; 4; 5; 6\} \Rightarrow n(S) = 6$
   $A = \{2; 4; 6\} \Rightarrow n(A) = 3$:
   Logo, $P(A) = 3/6 = 1/2$

b) A probabilidade do evento B "obter um número menor ou igual a 6 na face superior".

   Temos:
   $S = \{1; 2; 3; 4; 5; 6\} \Rightarrow n(S) = 6$
   $B = \{1; 2; 3; 4; 5; 6\} \Rightarrow n(B) = 6$
   Logo, $P(B) = 6/6 = 1$

c) A probabilidade do evento C "obter um número 4 na face superior".

   Temos:
   $S = \{1; 2; 3; 4; 5; 6\} \Rightarrow n(S) = 6$
   $C = \{4\} \Rightarrow n(C) = 1$
   Logo, $P(C) = 1/6$

d) A probabilidade do evento D "obter um número maior que 6 na face superior".

Temos:
$S = \{1; 2; 3; 4; 5; 6\} \Rightarrow n(S) = 6$
$D = \varnothing \Rightarrow n(D) = 0$: Logo, $P(D) = 0/6 = 0$
(Evento vazio ou impossível)

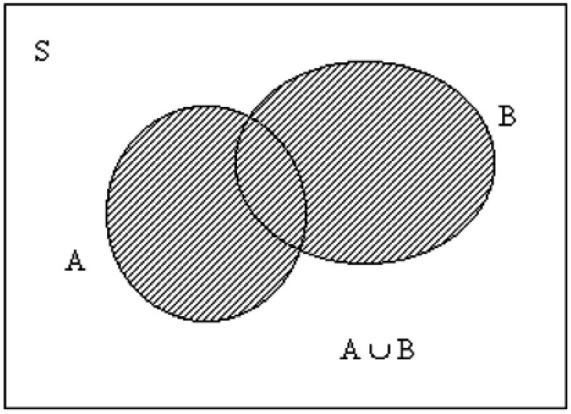
2.4 Operações com eventos aleatórios – Teoria dos Conjuntos

Consideremos um espaço amostral finito $S = \{e_1, e_2, e_3, ..., em\}$. Sejam A e B dois eventos de $F(S)$. As seguintes operações são definidas:

2.4.1 União de conjuntos

Definição: $A \cup B = \{e_i \in S = / e_i \in A$ ou $e_i \in B\}$; $i = 1,..., n$. Portanto, o evento união é formado pelos pontos amostrais que pertençam a pelo menos um dos conjuntos. A união pode ser vista na Figura 2.1.

Figura 2.1 – União dos conjuntos *A* e *B*



Observação

a) $A \cup B = B \cup A$

b) $A \cup A = A$

c) $A \cup \varnothing = A$

d) Se $A \subset B$, $A \cup B = B$ (em particular $A \cup S = S$).

A representação da união de A_i eventos A1, A2, ..., An (A1 ∪ A2 ∪ ... ∪ An) é dada por:

$$\bigcup_{i=1}^{n} A_i$$

Exemplo:

Os microrganismos causadores de enfermidades transmitidas por alimentos podem ser:

Liberadores de toxina: *S. aureus, Clostridium perfringens, C. botulinum, Vibrio cholerae, Bacillus cereus,* fungos filamentosos.

Causadores de infecções: *Salmonella sp, E. coli, Shigella sp, Vibrio parahaemoliticus, Campilobacter sp, Listeria monocytogenes, Yersinia sp.*

Considerando a seguinte distribuição nos alimentos:

Microrganismos	**Maioneses (*UFC/g*)**	**Verduras *(UFC/g)***	**Total *(UFC/g)***
Staphylococcus aureus	*10*	*113*	*123*
Salmonella	*107*	*28*	*135*
Escherichia coli	*106*	*102*	*208*

Qual a probabilidade de um dos alimentos servidos, escolhido ao acaso:

a) Estar contaminado por *Staphylococcus aureus* ou *Salmonella?*

$$P(S.\ aureus \cup\ Salmonella) = P(S.\ aureus) + P(Salmonella)$$

$$P\ (S.\ aureus \cup\ \ Salmonella) \quad = \frac{123}{466} + \frac{135}{466} = 0{,}55$$

b) Ser verdura ou estar contaminada por *Escherichia coli*:

$$P\ (verdura \cup\ Escherichia\ coli) = P(verdura) + P(Escherichia\ coli)$$

$$= \frac{243}{466} + \frac{208}{466} - \frac{102}{466} = 0{,}75$$

c) Não ser *Staphylococcus aureus:*

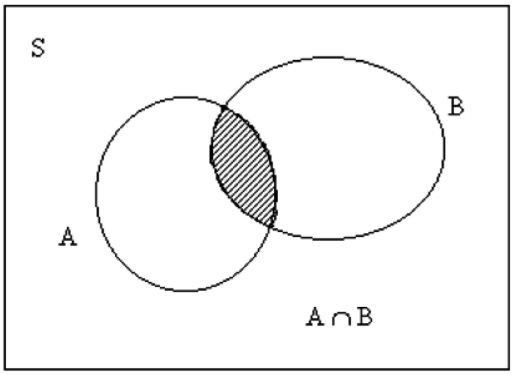
$$P(S.\overline{aureus}) = 1 - P(S.aureus)$$

$$= 1 - \frac{123}{466} = 0{,}73$$

2.4.2 Intersecção de conjuntos

Definição: $A \cap B = \{e_i \in S = e_i \in A\ e\ e_i \in B\}$; $i = 1, \ldots, n$. Portanto, o evento intersecção é formado pelos pontos amostrais que pertençam simultaneamente aos eventos *A* e *B*. A intersecção pode ser vista na Figura 2.2.

Figura 2.2 – Intersecção dos conjuntos *A* e *B*.



Observação

a) $A \cap B = B \cap A$

b) $A \cap A = A$

c) $A \cap \acute{A} = \acute{A}$

d) Se $A \subset B$ $A \cap B = A$ (em particular $A \cap S = A$)

e) $(A \cap B) \cap X = A \cap B \cap X)$.

A representação da intersecção de A_i eventos A1, A2,,..., An(A1 ∩ A2 ∩ ... ∩∪ An) é dada por:

$$\bigcap_{i=1}^{n} A_i$$

2.4.3 Probabilidade condicional

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

$$P(A/B) = \frac{P(A \bigcap B)}{P(B)} \tag{2.3}$$

$$P(A \bigcap B) = P(B)P(A/B) \tag{2.4}$$

$$P(B/A) = \frac{P(A \bigcap B)}{P(A)} \tag{2.5}$$

$$P(B \bigcap A) = P(A)P(B/A) \tag{2.6}$$

Exemplo:

Em um processo fermentativo utilizando 12 tubos de ensaios como meios de cultura, verificou-se que 4 estavam contaminados, duas amostras são retiradas ao acaso sem reposição. Qual a probabilidade de que ambas estejam sem contaminação?

- O evento A seria a probabilidade de retirar um tubo sem contaminação;
- O evento B seria a probabilidade de retirar um tubo contaminado;
- Na primeira amostra eu posso tirar 8 tubos não contaminados, dado que 4 dos 12 estavam contaminados;
- Na segunda amostra eu posso retirar 7 tubos não contaminados dos 11 restantes, uma vez que não houve reposição.

$$P(A \bigcap B) = \frac{8x7}{12x11} = 0{,}424 \text{ ou } 42{,}4\%$$

2.4.4 Independência Estatística

Um evento A é considerado independente de um evento B, se a probabilidade de A é igual à probabilidade de A dado B.

$$P(A) = P(A/B) \text{ ou } P(B) = P(B/A) \qquad (2.7)$$

Exemplo de Probabilidade Condicional.

Microrganismos	Maioneses (UFC/g)	Verduras (UFC/g)	Total (UFC/g)
Staphylococcus aureus	*10*	*113*	*123*
Salmonella	*107*	*28*	*135*
Escherichia coli	*106*	*102*	*208*

a) Qual a probabilidade de conter contaminação de *Salmonella* nas verduras?

$$P(S/V) = \frac{18}{466} = 0{,}039 \text{ ou } 4\%$$

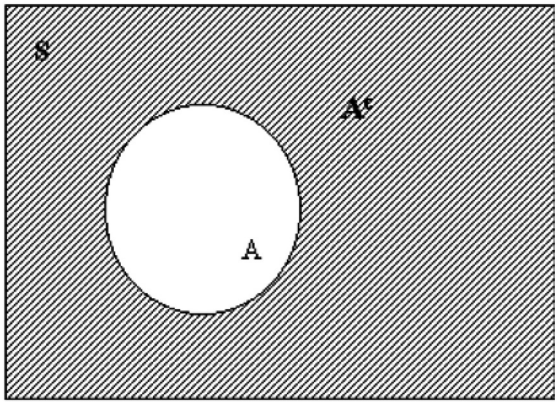
b) Qual a probabilidade de ser maionese a causadora da infecção, dado que está contaminada com *S.aureus*?

$$P(M/Sa) = \frac{10}{123} = 0{,}081 \text{ ou } 8\%$$

2.4.5 Complemento

Definição: $S - A = A^c = \{e_i \in S = e_i \notin A\}$; $i = 1,\ldots,n$. O complemento de um evento *A* é, portanto, o evento contendo todos os resultados no espaço amostral *S* que não pertençam a *A*. O complemento de *A* pode ser visto na Figura 2.3.

Figura 2.3 – Complemento do conjunto A



Observação

a) $(A^c)^c = A$

b) $A \cup A^c = S$

c) $\varnothing^c = S$

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

e) $S^c = \varnothing$.

2.4.6 Eventos mutuamente exclusivos ou disjuntos

Definição: Dois eventos são ditos mutuamente exclusivos ou disjuntos se A e B não puderem ocorrer juntos, ou seja, a realização de um exclui a realização do outro. Segue que A e B são disjuntos se $A \cap B = \varnothing$.

2.5 Definição axiomática de Probabilidade

Para um dado experimento é necessário atribuir para cada evento A no espaço amostral S um número $P(A)$ que indica a

probabilidade de *A* ocorrer. Para satisfazer a definição matemática de probabilidade, este número *P*(*A*) deve satisfazer três axiomas específicos:

- Axioma 1: Para qualquer evento A, P(A) ≥ 0;
- Axioma 2: P(S) = 1;
- Axioma 3: Para qualquer sequência infinita de eventos disjuntos $P(A_1 \cup A_2)=P(A_1)+P(A_2)$

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right) \tag{2.8}$$

A distribuição de probabilidade, ou simplesmente a probabilidade no espaço amostral *S* é uma especificação de números *P*(*A*) que satisfazem os axiomas 1, 2 e 3.

Teorema 1: $P(\varnothing) = 0$;
Teorema 2: Para qualquer sequência finita de eventos disjuntos $A_1, A_2, \ldots, A_N$

$$p\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right) \tag{2.9}$$

Teorema 3: Para qualquer evento A, $P(A^c) = 1 - P(A)$
Teorema 4: Para qualquer evento A, $0 \leq P(A) \geq 1$
Teorema 5: Se $A \subset B$, então $P(A) \leq P(B)$
Teorema 6: Para qualquer dois eventos *A* e *B*

$$P(A \cup B)=P(A)+P(B)-P(A \cap B) \tag{2.10}$$

Exemplo:

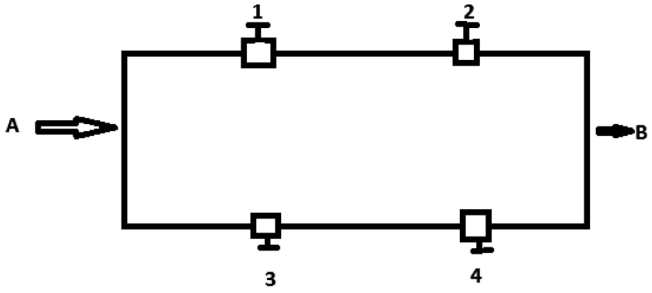
Em um fluxo de tanques para decantação, a probabilidade de que cada registro esteja fechado é de 0,6. Supondo que cada registro seja aberto ou fechado independentemente um do outro, calcular a probabilidade de que o líquido passe de A para B.

Sejam os eventos:

R_1= o registro 1 está fechado.
R_2= o registro 2 está fechado.
R_3=o registro 3 está fechado.
R_4=o registro 4 está fechado.

$P(R_1)=P(R_2)=P(R_3)=P(R_4) = 0,6$

Figura 2.4 – Fluxo tanques de decantação



O fluxo passa de A para B se estiverem fechados os registros 1 e 2 ou 3 e 4, portanto, podemos calcular $P[(R_1 \cap R_2) \cup (R_3 \cap R_4)]$ onde os eventos $(R_1 \cap R_2)$ e $(R_3 \cap R_4)$ podem ocorrer simultaneamente (se os quatro registros estiverem fechados).

Logo,

$P[(R_1 \cap R_2) \cup (R_3 \cap R_4)]=P(R_1 \cap R_2) + P(R_3 \cap R_4)-P(R_1 \cap R_2) \cap (R_3 \cap R_4)$
$=P(R_1).P(R_2)+P(R_3).P(R_4)-P(R_1).P(R_2).P(R_3).P(R_4)$
$=(0,6.0,6) + (0,6.0,6 – (0,6.0,6.0,6.0,6))=0,60$ ou 60%

Teorema 7: Teorema de Bayes - Se $E_1, E_2, \ldots, E_n$ são eventos dois a dois mutuamente exclusivos e exauram o conjunto S dos eventos elementares, então se $P(E_i)>0$, (i=1,2, ...,n), tem-se

$$p(E_i / B) = \frac{P(E_i)P(B/E_i)}{P(E_1)P(B/E_1) + P(E_2)P(B/E_2) + \ldots + P(E_n)P(B/E_n)} \qquad (2.11)$$

Onde B é um evento que só pode ocorrer como efeito de uma das causas mutuamente exclusiva E_1. O teorema de Bayes fornece a probabilidade de que o evento E_1 tenha ocorrido na hipótese de que o evento B tenha sido observado.

Exemplo:

Uma indústria produz quatro tipos de válvulas eletrônicas: A, B, C, D. A probabilidade de uma válvula do tipo A falhar é 1%, do tipo B é 0,5%, do tipo C é 2% e do tipo D é 0,2%. Em um depósito existem 1000 válvulas do tipo A, 500 do tipo B, 300 do tipo C e 200 do tipo D. Uma válvula é retirada ao acaso do depósito e verifica-se que é defeituosa. Qual a probabilidade de que a válvula retirada seja do tipo D?

Sejam os eventos:
A= a válvula é do tipo A.
B= a válvula é do tipo B.
C= a válvula é do tipo C.
D= a válvula é do tipo D.
E= a válvula é defeituosa.

O evento válvula defeituosa (E) ocorreu, portanto, a probabilidade de que seja do tipo D (sendo A, B, C e D mutuamente exclusivos) é dada por:

$$p(D/E) = \frac{P(D)P(E/D)}{P(A)P(E/A) + P(B)P(E/B) + P(C)P(E/C) + P(D)P(E/D)}$$

Onde:

$$p(A)=\frac{1000}{2000}=0{,}50;\ p(B)=\frac{500}{2000}=0{,}25;\ p(C)=\frac{300}{2000}=0{,}15;\ p(D)=\frac{200}{2000}=0{,}10$$

$$p(E/A)=0{,}010;\ p(E/B)=0{,}005;\ p(E/C)=0{,}020;\ p(E/D)=0{,}002$$

Portanto:

$$p(D/E)=\frac{0{,}10\times 0{,}002}{0{,}50\times 0{,}010+0{,}25\times 0{,}005+0{,}15\times 0{,}020+0{,}10\times 0{,}002}=0{,}021$$

Ou 2,1%

Problemas Propostos

1) Uma indústria química produz três tipos de produtos altamente tóxicos. Durante as reações, na obtenção desses produtos, cuidados especiais são tomados para evitar o vazamento de gases. De acordo com o setor de segurança dessa indústria, a probabilidade de vazamento do gás tipo 1 é de 0,001, do tipo 2 é 0,002 e do tipo 3 é de 0,015. Em um dia qualquer, durante as reações acima citadas, qual a probabilidade de haver simultaneamente, vazamento de:

a) Dois tipos de gases?

b) Três tipos de gases?

2) Em uma indústria, cinco máquinas (A, B, C, D e E) produzem os mesmos tipos de peças, que serão utilizadas na montagem de equipamentos elétricos. Sabe-se que a produção diária da máquina A é o dobro da produção da máquina D, que a produção das máquinas B e C são iguais e que a máquina E produz 20 peças a mais que a máquina A. De acordo com

o setor de controle de qualidade dessa indústria, são defeituosas, respectivamente, 1%, 2%, 5%, 1% e 3% das peças produzidas pelas máquinas A, B, C, D e E. Uma peça foi tomada aleatoriamente e verificou-se que ela é defeituosa. Calcular a probabilidade de que essa máquina tenha sido fabricada pela máquina E, sabendo-se que as máquinas A e B produzem, respectivamente, 200 e 150 peças.

2.6 Distribuições de Probabilidades

Até o momento construímos a distribuição de probabilidade de uma variável discreta (nº de faces no lançamento de dois dados), empregando nosso conhecimento para o cálculo das probabilidades envolvidas.

Veremos adiante alguns modelos probabilísticos padrões que nos auxiliarão em diversas situações práticas. Nosso problema passa a ser determinar qual modelo é o mais adequado para a situação em estudo.

2.7 Principais distribuições discretas

Para identificarmos uma variável aleatória discreta temos de conhecer quais resultados podem ocorrer e quais são as probabilidades associadas a tais resultados.

2.7.1 Distribuição de Bernoulli

Caracteriza o tipo mais simples de experimento, quando queremos apenas observar a presença ou não de alguma característica, cujos dois únicos resultados denominamos de sucesso ou fracasso.

Neste contexto, sucesso não significa algo bom ou excepcional, mas apenas um resultado no qual temos interesse, enquanto fracasso significa exatamente o outro resultado possível.

- Seja X a variável aleatória que admite apenas os valores x_1= 1 (sucesso) e x_2= 0 (fracasso).

- Seja P(X) a função de distribuição de X, tal que $p(x_1)$ = p e $p(x_2)$= q , onde p + q = 1.

Definimos a seguinte variável aleatória discreta X, número de sucessos em uma única tentativa do experimento.

X assume os valores:

$$X = \begin{cases} 0, & fracasso \\ & com \quad P(x=0)=q \quad e \quad P(x=1)=p \\ 1, & sucesso \end{cases}$$

Nessas condições a variável aleatória X tem distribuição de Bernoulli e sua função de probabilidade é dada por:

$$P(X = x) = p^x.q^{1-x} \qquad (2.12)$$

Uma variável aleatória, assim definida, tem uma distribuição de Bernoulli e suas principais características são:

X	P(X)	X.P(X)	X^2.P(X)
0	q	0	0
1	p	p	p
	1	p	p

Média μ_x = E[X] = p e **Variância** σ^2_x = V [X] = p - p^2 = p(1-p) = p.q.

A distribuição fica completamente especificada ao estabelecermos um valor para p.

Exemplo:

Uma urna tem 30 bolas brancas e 20 verdes. Retira-se uma bola dessa urna. Seja X: nº de bolas verdes. Calcular E(X), Var(X) e determinar P(X).

$$X = \begin{cases} 0 \Rightarrow q = 30/50 = 3/5 \\ 1 \Rightarrow p = 20/50 = 2/5 \end{cases} \quad com \quad P(x = x) = (2/5)^x.(3/5)^{1-x}$$

E(X) = p = 2/5 Var(X) = p.q = (2/5).(3/5) = 6/25

E(X) = p = 2/5 Var(X) = p.q = (2/5).(3/5) = 6/25

2.7.2 Distribuição Binomial

Trata-se de uma distribuição de probabilidade adequada aos experimentos que apresentam apenas dois resultados possíveis: sucesso ou fracasso.

Por exemplo:

a) Lançar uma moeda 5 vezes e observar o número de caras.

b) Numa linha de produção, observar 10 itens tomados ao acaso e verificar o número de defeituosos.

c) Verificar o número de bits que não estão afetados por ruído num pacote com n bits.

Define-se a Variável Binomial X como o número de sucessos em n repetições do experimento. A expressão geral da Distribuição Binomial é:

$$P(X = x) = \binom{n}{x} p^x . q^{n-x} \qquad (2.13)$$

A Esperança, Variância e Desvio Padrão da variável aleatória do tipo Binomial são calculadas respectivamente por:

$$E(Y) = n \,.p \qquad (2.14)$$

$$V(Y) = n.\ p.\ Q \qquad (2.15)$$

$$DP(Y) = \sqrt{V(Y)} \qquad (2.16)$$

Exemplo:

Uma moeda não viciada é lançada 5 vezes. Encontre a probabilidade de:

a) Dar exatamente 3 caras.

b) Pelo menos uma cara.

c) No máximo 2 caras.

d) Calcular o valor esperado e o desvio padrão.

Seja X a variável Binomial com os parâmetros: n=5, p=1/2 e q=1/2

a) Desejamos *P(X = 3)* por (2.13)

$$\Rightarrow P(X = 3) = \frac{5!}{3!\,2!} \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^2$$

Onde,

$$P(X = 3) = 10.(1/2)^5 = 10/32 = 31{,}25\%$$

b) Desejamos *P(X ≥ 1)* que é o mesmo que $1 - P(X < 1)$ equivalente a

$$1. - P(X = 0) = 1 - 0{,}03125 = 96{,}88\%$$

c) Desejamos *P(X ≤ 2)* que equivale a $P(X = 0) + P(X = 1) + P(X = 2) = 50\%$

$E[X] = np$

Logo,

$E[X] = 2{,}5$ caras, e $V[X] = npq = 5/4 = 1{,}25$. Logo $DP[X] = 1{,}12$ caras.

2.7.3 Distribuição Poisson

Considere as situações em que se avalia o número de ocorrências de um determinado evento por unidade de tempo, de comprimento, de área ou de volume (genericamente denominados de área de oportunidade).

Em muitos casos conhece-se o número de sucessos, mas às vezes é muito difícil, ou até mesmo impossível, determinar o número de fracassos. Imagine o número de automóveis que passam por uma esquina: pode-se anotar o número de veículos que passaram num determinado intervalo de tempo, mas não se pode determinar quantos deixaram de passar.

A distribuição de Poisson é aplicada nos tipos de situações em que nos interessa o número de vezes em que um evento pode ocorrer durante um intervalo de tempo ou em determinado ambiente físico (área de oportunidade).

Tomando como referência o número de ocorrências em determinado intervalo de tempo, em um processo de Poisson, podem ser observados eventos discretos num intervalo de tempo, de tal forma que, reduzindo suficientemente este intervalo, tenhamos:

A função de probabilidade da distribuição de Poisson é:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad (2.14)$$

onde:

e é uma constante (base do logarítmo neperiano), valendo aproximadamente 2,718...
λ é o número esperado de sucessos no intervalo considerado;
x é o número de sucessos (x = 0, 1, 2, ...,∞.);

Principais Características da Distribuição de Poisson

Média $\mu_x = E[X] = \lambda$ e Variância $\sigma^2_x = V[X] = \lambda$.

Exemplo:

As consultas a um banco de dados ocorrem de forma independente e aleatória, à base de 3 consultas por minuto. Calcule as probabilidades:

a) No próximo minuto ocorrerem exatamente 3 consultas.

b) No próximo minuto ocorrerem menos de 3 consultas.

c) Nos próximos dois minutos ocorrerem mais do que 5 consultas.

Seja X a variável Poisson com ocorrência média de 3 consultas por minuto (λ=3):

- Desejamos $P(X = 3) = [e^{-3}. 3^3]/3! = 22,4\%$
- Desejamos $P(X < 3) = P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) = 42,32\%$

Observe que a unidade de tempo alterou de 1 para 2 minutos. Como a taxa média é de 3 por minuto, então em dois minutos teremos $\lambda = 6$. Desejamos assim:

$$P(X > 5) = 1 - P(X \leq 5) = 1 - 0{,}42358 = 57{,}64\%$$

2.8 Principais distribuições contínuas

Quando os valores que a variável aleatória pode assumir pertencem ao conjunto dos números reais, a variável aleatória é denominada contínua. Como não é possível registrar todos os valores de uma variável aleatória contínua numa lista ou tabela, a distribuição de probabilidade deste tipo de variável aleatória é definida por uma curva contínua e não por pontos discretos de uma tabela.

2.8.1 Função densidade de probabilidade

De forma análoga à Distribuição de Probabilidade de uma variável aleatória discreta para variável aleatória contínua define-se a função Densidade de Probabilidade [f(x)] com as seguintes características:

a) A probabilidade da variável aleatória X é sempre definida em um intervalo de valores de X_i, por exemplo, (x_1, x_2), e sempre temos que $f(x) \geq 0$ para todo $x \in S$.

b) A probabilidade da variável aleatória X é medida pela área sob a curva da função densidade em um determinado intervalo.

$$P(a < X < b) = \int_a^b f(x)dx \qquad (2.15)$$

c) A área total sob a curva de densidade é igual a 1.

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \qquad (2.16)$$

Observa-se que a função densidade f(x) não mede a probabilidade no ponto x da variável aleatória X.

Pela própria característica mencionada no item b acima, P(x≤X≤ x)=0, onde utilizamos o seguinte artifício para representarmos (X = x) ≡ (x ≤ X ≤ x).

Por considerarmos a probabilidade de um ponto como igual a zero, decorre imediatamente que:

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) \qquad (2.17)$$

Função de distribuição acumulada

De forma análoga às variáveis aleatórias discretas, pode-se definir também uma Função de Distribuição Acumulada para variáveis aleatórias contínuas como sendo:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx \qquad (2.18)$$

Qualquer probabilidade pode ser obtida através da Função de Distribuição Acumulada. Para a < b sempre teremos:

- P(X < a) = P(X ≤ a) = F(a)
- P(X > b) = 1 - P(X ≤ b) = 1 - F(b)
- P(a < X < b) = F(b) - F(a)

2.8.3 Valor Esperado de uma Variável Aleatória Contínua

Define-se Esperança Matemática ou Média de uma variável aleatória contínua como:

$$\mu_x = E(X) = \int_{-\infty}^{+\infty} xf(x)dx \qquad (2.19)$$

2.8.4 Variância e Desvio Padrão de uma Variável Aleatória Contínua

Define-se Variância para uma variável aleatória contínua como:

$$\delta^2{}_x = V(X) = E(X - \mu_x)^2 = \int_{-\infty}^{+\infty} (x - \mu_x)^2 f(x)dx \quad (2.20)$$

Por definição, o desvio padrão é sempre a Raiz Quadrada da Variância.

Alternativamente, podemos calcular a Variância com o uso da fórmula:

$$V(X) = \delta^2 x = E(X^2) - \mu^2{}_x \quad (2.21)$$

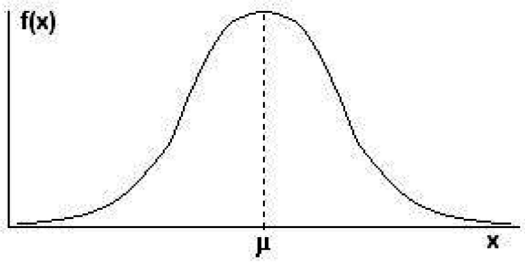
2.9 Principais distribuições contínuas

2.9.1 Distribuição Normal

É considerada a distribuição de probabilidades mais importante, pois permite modelar uma infinidade de fenômenos naturais e, além disso, possibilita realizar aproximações para calcular probabilidades de muitas variáveis aleatórias que têm outras distribuições, tais como a Binomial quando n é grande e p não muito grande nem muito pequeno.

É também conhecida como distribuição de Gauss, Laplace ou Laplace-Gauss, e é muito importante também na inferência estatística.

A distribuição Normal é caracterizada por uma Função de Densidade de Probabilidade, cujo gráfico descreve uma curva em forma de sino, que evidencia maior probabilidade de a variável aleatória assumir valores próximos aos valores centrais.



Uma Variável aleatória terá Distribuição Normal se sua Função Densidade de Probabilidade for da forma

$$f(x) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)^2} \qquad (2.22)$$

onde:
μ = média da distribuição
σ = desvio padrão da distribuição
π *e e* são constantes (3,1416..... *e* 2,718.....)

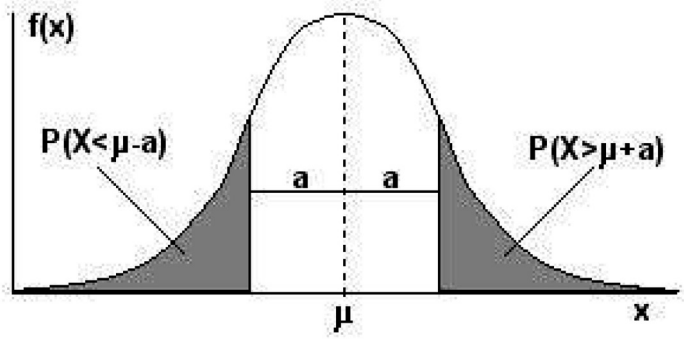
A Distribuição Normal tem como parâmetros a Média ou Valor Esperado $\mu_x = E[X] = \mu$ *e* Variância $\sigma^2_x = V[X] = \sigma^2$ *e* suas principais características são denotadas por:

1. Teoricamente, a curva prolonga-se de **−∞** a **+∞**, sendo que limite de f(x) =0 para x tendendo a ±∞.

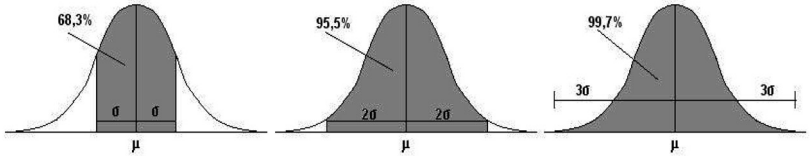
2. A área total sob a curva é igual a 1, ou seja:

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

3. A curva é simétrica em torno de μ, o que faz com que média = mediana = moda. Adicionalmente, temos também que $P(X < \mu - a) = P(X > \mu + a)$.



4. A curva tem dois pontos de inflexão, respectivamente em μ-σ e μ+σ. Cerca de 68% dos valores recaem no intervalo de um desvio padrão de cada lado da média, 95% recaem no intervalo média ± 2 desvios e 99,7% recaem no intervalo média ± 3 desvios.

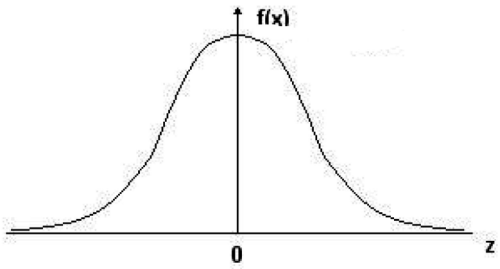


Considerando a enorme dificuldade de calcularmos probabilidades pela integração da Função de Densidade de Probabilidade (FDP) para as infinitas combinações de valores de μ e σ, utiliza-se a Distribuição Normal Padrão ou Reduzida, definida conforme a seguir.

Seja Z a variável com distribuição normal com média = 0 e variância = 1, geralmente denotada por N (0;1). Neste caso (lembrando que desvio-padrão = variância = 1) a FDP de Z será:

$$z = \frac{x - \mu}{\sigma} \qquad 2.23)$$

com a forma:



Observe-se a conveniência de termos a média igual a zero e o desvio padrão igual a 1, fazendo com que essa distribuição passe a representar os valores de z_i como número de desvios em relação à média (origem). Assim, essa distribuição nos permite trabalhar com valores *relativos* de desvios em relação à média.

Qualquer distribuição normal com média μ e desvio padrão σ pode ser transformada, para efeito de cálculo de probabilidades, na distribuição normal padrão, através de uma mudança de variável.

Há vários tipos de tabelas que fornecem as áreas (probabilidades) sob a curva normal padrão. O tipo mais comum é a tabela de faixa central. Esse tipo de tabela fornece a área sob a curva normal padrão entre z= 0 e qualquer valor positivo de z. A simetria em torno de z= 0 permite-nos obter a área entre quaisquer valores de z, sejam positivos ou negativos, não sem razoável esforço na identificação correta de intervalos.

Exemplos de transformações:

a) Calcule P(z < 0,85)

A área solicitada é exatamente a área fornecida pela tabela. Basta procurar a linha que contenha o valor 0,8 e sua interseção com a coluna que contenha o valor 0,05. (lembrando que 0,85 = 0,8 + 0,05).
Logo, P(z < 0,85) = 0,8023 (ou 80,23%).

b) Calcule P(0 < z < 1,25)

O valor procurado corresponde a P(z<1,25) – P(z<0). Da tabela, tiramos que P(z<1,25)=0,8944 e P(z<0)=0,5.
Logo, P(0<z<1,25) = 0,8944 -0,5000 = 0,3944 (ou 39,44%).

c) Calcule P(z>2,39)

Observe que o valor tabelado é P(z<2,39). Como a área total sob a curva vale 1, então P(z>2,39)=1-P(z<2,39).
Logo, P(z>2,39) = 1 – 0,9916 = 0,0084 ou 0,84%

d) Calcule P(z=1)

Considerando que a probabilidade é medida pela área sob a curva definida por um intervalo, P(z= 1) pode ser escrita como P(1 ≤ z ≤ 1). Isso reduz o intervalo a um só ponto e, portanto, a área é zero. Outra forma de se obter esse resultado é pela utilização do conceito da Função de Distribuição Acumulada, pois P(1 ≤ z ≤ 1) = F(1) – F(1) = 0.

e) Calcule P(-2,55<z<1,2)

P(-2,55<z<1,2) = P(z<1,2) - P(z<-2,55) = 0,8849 - 0,0054 = 0,8795 ou 87,95%

f) O diâmetro do halo de inibição formado por um bactericida ao inibir o crescimento de bacilos germinativos é normalmente distribuída com média 1,60mm e desvio

padrão 0,30mm. Calcule a probabilidade de um halo medir entre 1,50mm e 1,80mm. Seja X a variável aleatória $N(1,60; 0,30^2)$. Deseja-se a probabilidade $P(1,50 < x < 1,80)$

Precisamos primeiro transformar os limites do intervalo da VA X para a VA Z (Normal reduzida ou Normal padrão), para que possamos, pela tabela, calcular $P(z_1 < z < z_2)$. Assim procedendo teremos:

$$z_1 = (1,50 - 1,60)/0,30 = -0,10/0,30 = - 0,33$$

$$z_2 = (1,80 - 1,60)/0,30 = \ 0,20/0,30 = \ 0,67$$

Assim,

$P(- 0,33 < z < 0,67) = P(z<0,67) - P(z <- 0,33) = 0,7486 - 0,3707 = 0,3779$ ou 37,79%

Nos exemplos anteriores foram fornecidos os valores do intervalo para que fossem calculadas as probabilidades associadas ao intervalo. Existem aplicações em que devemos determinar os valores de z a partir do conhecimento das probabilidades associadas a esses valores.

O software Excel disponibiliza as seguintes funções para cálculos com a Distribuição Normal:

Função DIST.NORMP(z), onde
z: valor da VA Normal Padrão ou Reduzida.

Esta função retorna a probabilidade $P(-\infty < Z < z) = P(Z < z)$, para qualquer valor de z, da mesma forma que a tabela apresentada no final deste capítulo.

Para um intervalo genérico P(a<z<b) pode-se aplicar F(b) – F(a) diretamente na forma:

P(a<z<b) = DIST.NORMP(b) – DIST.NORMP(a)

Para P(z>a), usa-se 1 – P(z<a) e, portanto, P(z>a) = 1 – DIST.NORMP(a). Aplicável ao exemplo "c" acima.

- Função DIST.NORM(x ; média ; desv_padrão ; cumulativo), onde x: valor da VA Normal
- Média: média da variável aleatória X.
- Desv_padrão: desvio padrão da variável aleatória X.
- Cumulativo: um valor lógico que define o tipo de distribuição.
- VERDADEIRO: retorna o valor da função de distribuição acumulada (FDA) $F(x) = P(X \leq x)$
- FALSO: retorna o valor da função densidade de probabilidade (FDP) no ponto x: f(x)

É a função mais completa para tratamento de distribuição normal. Observe que no caso dos parâmetros média= 0, desvio= 1 e cumulativo= 1 ou verdadeiro, esta função retorna o mesmo valor da **DIST.NORMP**.

Função INV. NORMP (probabilidade)

Retorna o valor z da VA Normal Padrão, abaixo do qual se tem a probabilidade informada. É o inverso da função DIST. NORMP (z)

No caso do exemplo "g", a função inversa registrada como INV.NORMP (0,3015) retorna exatamente - 0,520091.

Para o caso da sugestão apresentada,

INV.NORMP (0,30)= - 0,524401

Função INV.NORM (probabilidade; média; desv_padrão)

Como no caso acima, é o inverso da função geral DIST. NORMP(z), aplicável a qualquer variável aleatória Normal X, desde que conhecidos sua média e desvio padrão.

Função PADRONIZAR (x; média; desv_padrão)

Retorna o desvio padrão normalizado z, considerando os argumentos x, média e desvio padrão, utilizando a fórmula já apresentada (2.24):

2.9.2 Distribuição de Qui-Quadrado (χ^2)

É um modelo de distribuição contínua muito importante para a teoria da inferência estatística.

Considere x_1, x_2, x_3, ..., x_p, "n" variáveis aleatórias independentes, normalmente distribuídas com média zero e variância 1, ou seja, "n" variáveis tipo normal padrão.

Define-se a variável aleatória com distribuição Qui-Quadrado como:

$$\chi_n^2 = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2 \text{ ou}$$

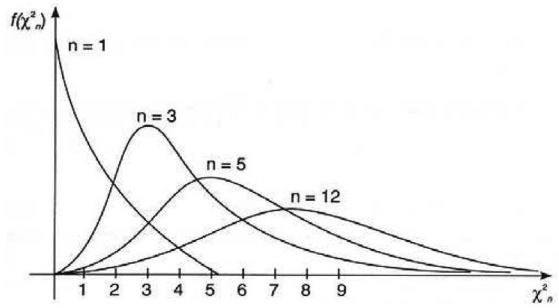
$$\chi_n^2 = \sum_{i=1}^{n} z_i^2 \qquad (2.24)$$

Onde "n" é um parâmetro da função densidade de probabilidade denominado grau de liberdade e geralmente denotado pela letra grega φ (lê-se fi), ou eventualmente por gl.

As principais características da distribuição Qui-Quadrado são:

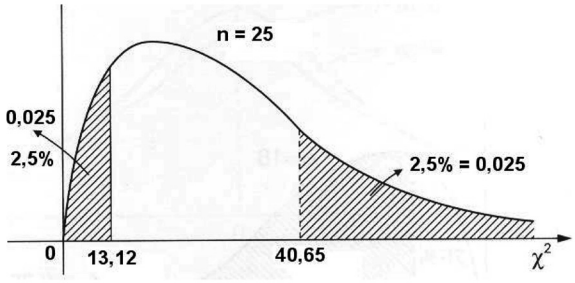
- $\chi_n^2 \geq 0$
- Média = n
- Variância = 2n

A função densidade de probabilidade está representada graficamente para alguns valores de n:



Observe que na medida em que n cresce, a função de densidade de probabilidade tende à forma da Função Normal.

A tabela do Qui-Quadrado em função do grau de liberdade n apresenta o valor numérico da VA que deixa à sua direita determinada área α, ou seja, $\alpha = P(X \geq x)$.



Para cálculo da probabilidade $P(X \leq x)$, ou seja, área na cauda esquerda da distribuição utiliza-se a propriedade $P(X \leq x) = 1 - P(X \geq x) = 1 - \alpha$, conforme ilustrado abaixo.

1. O valor à direita, chamado qui-quadrado superior, é obtido na tabela com:
   n= 25 e $\alpha = 0,025$.
   Logo, $x^2 = 40,65$

2. O valor da abscissa à esquerda, chamado qui-quadrado inferior, é obtido da tabela com n = 25 e $\alpha = 1 - 0{,}025$, portanto $\alpha = 0{,}975$.

   Logo, $x^2 = 13{,}12$

O Excel disponibiliza as seguintes funções para cálculos com a Distribuição Qui-Quadrado:

Função DIST.QUI (x ; graus_liberdade), onde x: valor da variável aleatória Qui-Quadrado, e graus_liberdade = n
Retorna a probabilidade P(X > x), ou seja, se n= 25, *DIST.QUI(18 ; 25)* = P(X>18) = 84,24%.

Função INV.QUI (probabilidade ; graus_liberdade)
Retorna o inverso da probabilidade uni-caudal da distribuição qui-quadrada.
Se probabilidade = *DIST.QUI(x ; n)*, então *INV.QUI(probabilidade; n) = x*.
No caso do exemplo anterior, *INV.QUI(* 0,8424 ; 25*)* =18.

2.9.3 Distribuição t de Student

É um modelo de distribuição contínua que se assemelha à distribuição normal padrão N(0 ; 1). É utilizada para inferências estatísticas, quando se tem amostras com tamanhos inferiores a 30 elementos. A distribuição t de Student, com *n* graus de liberdade é dada por:

$$t = \frac{z}{\sqrt{\frac{x^2 n}{n}}} \qquad (2.25)$$

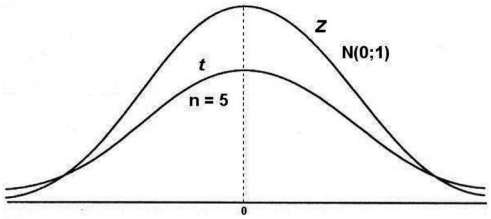
Principais características da distribuição *t* de Student.

1. Média = 0

2. Variância = $\frac{n}{n-2}$

3. A distribuição é simétrica em relação à média.

4. A comparação entre *t* e z é mostrada no gráfico



Para valores de n < 30, a distribuição apresenta maior dispersão que z N(0;1). À medida que n aumenta, *t* se aproxima cada vez mais de z.

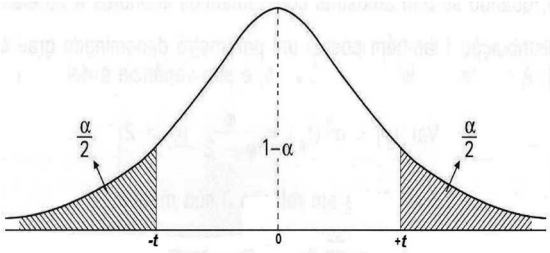
A distribuição t também está tabelada. No final deste livro é apresentada uma tabela que fornece as abscissas da distribuição para diversas áreas (probabilidades) nas caudas. Trata-se de uma tabela bicaudal, conforme ilustrado a seguir.

Exemplo:

Seja n = 9 e α = 5%. Consultando a tabela com estes valores, encontramos o valor t = 2,2622. Observe que a tabela mostra os valores de x tais que $P(-x \leq X \leq x) = 1 - \alpha$.

O Excel disponibiliza as seguintes funções para cálculos com esta distribuição:

Função DISTT (x ; graus_liberdade ; caudas), onde x: valor da abscissa, graus_liberdade = n, e caudas igual a 1 ou 2, conforme se deseje 1 ou 2 caudas.
Retorna a probabilidade $P(-x \leq X \leq x)$ para n grau de liberdade.



No exemplo acima, n= 9 e t= 2,2622, a função DISTT(2,2622 ; 9 ; 2) retorna 0,0499965 ou 5%. Alterando o número de caudas, DISTT(2,2622 ; 9 ; 1) retorna 0,02499825 ou 2,5%.

Exercícios:

1. X é uma variável aleatória contínua, tal que X = N (12;25). Qual a probabilidade de uma observação ao acaso:

   a) ser menor do que -3.
   b) cair entre -1 e 15.

2. Suponha que o diâmetro médio de um halo de inibição ocasionado pela aplicação de um antibiótico contra bactérias Gram-positivas é de 0,25mm, e o desvio padrão 0,02mm. O efeito é considerado inibidor se o diâmetro do halo formado é maior que 0,28mm ou sem efeito significativo se menor que 0,20mm.

   a) Encontre a porcentagem de halos considerados com inibição;

   b) Qual deve ser a medida mínima para que tenhamos no máximo 12% de inibição?

3. A fase exponencial de certo microrganismo em um meio nutritivo tem em média 8 dias e desvio padrão 4 dias. Calcular a probabilidade de esse microrganismo crescer:

a) Entre 7 e 10 dias.

b) Mais que 8 dias.

c) Menos que 7 dias.

d) Exatamente 10 dias.

e) Qual deve ser o número de dias necessários para que tenhamos de repor no máximo 5% dos nutrientes?

4. A produção de nisina por 6 espécies de *Lactococcus lactis* são normalmente distribuídos com média 6,5 g/L e desvio padrão 0,5 g/L. Encontre o número de microrganismos que produzem:

   a) Entre 6 e 7 g/L.

   b) Mais que 6,2 g/L.

5. Certo produto tem peso médio de 10 g e desvio padrão 0,5 g. É embalado em caixas de 120 unidades que pesam em média 150 g e desvio padrão 8 g. Qual a probabilidade de que uma caixa cheia pese mais de 157 g?

6. Suponha que a duração de vida de dois equipamentos E1 e E2 tenham respectivamente distribuições N(45;9) e N(40;36). Se o equipamento tiver que ser usado por um período de 45 horas, qual deles deve ser preferido?

7. Certa máquina de empacotar determinado produto oferece variações de peso com desvio padrão de 20 g. Em quanto deve ser regulado o peso médio do pacote para que apenas 10% tenham menos que 400 g? Calcule a probabilidade de um pacote sair com mais de 450 g.

8. Num laticínio, a temperatura do pasteurizador deve ser de 75°C. Se a temperatura ficar inferior a 70°C, o leite poderá ficar com bactérias maléficas ao organismo humano. Observações do processo mostram que valores da temperatura seguem uma distribuição normal com média 75,4°C e desvio padrão 2,2°C.

a) Qual é a probabilidade da temperatura ficar inferior a 70°C?

b) Qual é a probabilidade de que em 500 utilizações do pasteurizador, em mais do que cinco vezes a temperatura não atinja 70°C?

3
Técnicas de amostragem

Devido à impossibilidade de obtermos todas as informações disponíveis, podendo até mesmo ser desnecessários os levantamentos de dados por amostragem, quando realizados seguindo rigidamente conceitos científicos podem fornecer resultados preciosos a custos desprezíveis quando comparados aos levantamentos que tenham por alvo toda a população da pesquisa, além disso, em alguns casos são mais confiáveis que um censo.

A amostragem estuda técnicas de planejamento de pesquisa para possibilitar inferências sobre um universo a partir do estudo de uma pequena parte de seus componentes, uma amostra. O processo de amostragem tem a finalidade de definir o tamanho de determinada amostra, baseado em informações fornecidas por uma variável base, que é uma característica medida, controlada ou manipulada numa pesquisa. Basicamente, o tamanho da amostra dependerá da variância, da variável base e dos níveis de precisão exigidos.

A amplitude das conclusões de um estudo estatístico está limitada pela qualidade do processo de amostragem. Se a amostra for representativa as conclusões que podemos tirar aplicam-se a toda população, sendo possível calcular as incertezas. Já se a amostra não for representativa as conclusões devem limitar-se à própria amostra.

Uma amostra será representativa de uma população, em relação a um caráter variável, se não houver qualquer razão para pensar que o valor desse caráter possa diferir da amostra para a população, sendo preciso também que todos os elementos da população tenham a mesma probabilidade de serem selecionados.

Já quando a amostra é extraída da população segundo algum método de seleção, seja, por exemplo, por razões de comodidade do experimentador, esta não é representativa da população e não podemos extrair dela quaisquer conclusões relativas à população, podendo, no máximo, fazer indicações acerca da população.

Basicamente, existem dois tipos de amostragem: a amostragem probabilística, quando todos os elementos da população tem a mesma chance de pertencer à amostra, e a amostragem não probabilística, quando os elementos da população não possuem probabilidade conhecida de pertencer à amostra.

A vantagem do uso da amostragem probabilística é a determinação do erro amostral, o que não é verificado na amostragem não probabilística.

3.1 Técnicas de amostragem probabilística

As principais técnicas são:

3.1.1 Amostragem por conglomerado

A população é dividida em diferentes conglomerados (grupos), extraindo-se de cada grupo uma amostra dos conglomerados selecionados, e não de toda a população. O ideal seria que cada conglomerado representasse tanto quanto possível o total da população. Na prática, selecionam-se os conglomerados geograficamente.

Quando a população pode ser dividida em grupos homogêneos, ou seja, subgrupos que consistem, todos eles, em indivíduos bastante semelhantes entre si, pode-se obter uma amostra aleatória desta população bastante precisa.

3.1.2 Amostragem aleatória simples

A amostragem aleatória simples (AAS) é a maneira mais fácil para selecionarmos uma amostra probabilística de uma população. Comecemos introduzindo o conceito de AAS de uma

população finita, para a qual temos uma listagem de todas as unidades elementares.

Podemos obter uma amostra nessas condições escrevemos cada elemento num cartão, misturando-os numa urna e sorteando tantos cartões quantos desejarmos na amostra. Esse procedimento torna-se inviável quando a população é muito grande. Nesse caso, usa-se um processo alternativo no qual os elementos são numerados e em seguida sorteados por meio de uma tabela de números aleatórios.

Utilizando-se um procedimento aleatório, sorteia-se um elemento da população, sendo que todos os elementos têm a mesma probabilidade de serem selecionados. Repete-se o procedimento até que sejam sorteadas as unidades da amostra.

Podemos ter uma AAS com reposição se for permitido que uma unidade possa ser sorteada mais de uma vez e sem reposição se a unidade sorteada for removida da população.

Do ponto de vista da quantidade de informação contida na amostra, amostrar sem reposição é mais adequado. Contudo, a amostragem com reposição conduz a um tratamento teórico mais simples, pois ela implica que tenhamos independência entre as unidades selecionadas. Essa independência facilita o desenvolvimento das propriedades dos estimadores que serão considerados.

Se a população for infinita então as retiradas com e sem reposição serão equivalentes, isto é, se a população for infinita (ou então muito grande) o fato de se recolocar o elemento retirado de volta na população não vai afetar em nada a probabilidade de extração do elemento seguinte.

Se, no entanto, a população for finita (e pequena) será necessário fazer uma distinção entre os dois procedimentos, pois na extração com reposição as diversas retiradas serão independentes, mas no processo sem reposição haverá dependência entre as retiradas, isto é, o fato de não recolocar o elemento retirado afeta a probabilidade de o elemento seguinte ser retirado.

A amostragem sem reposição é mais eficiente que a amostragem com reposição e reduz a variabilidade, uma vez que não é possível retirar elementos extremos mais de uma vez.

3.1.3 Amostragem sistemática

Quando os elementos da população se apresentam ordenados e a retirada dos elementos da amostra é feita periodicamente temos uma amostragem sistemática. Assim, por exemplo, em uma linha de produção podemos a cada dez itens produzidos retirar um, para pertencer a uma amostra da produção diária.

Amostras não probabilísticas são também, muitas vezes, empregadas em trabalhos estatísticos, por simplicidade ou por impossibilidade de se obterem amostras probabilísticas, como seria desejável. No entanto, processos não probabilísticos de amostragem têm também sua importância e sua utilização deve ser feita com cuidado. Apresentamos a seguir algumas técnicas de amostragem não probabilística.

3.1.4 Inacessibilidade a toda população

Esta situação ocorre com muita frequência na prática. Por exemplo, seja a população que nos interessa constituída de todas as peças produzidas por certa máquina. Mesmo estando a máquina em funcionamento normal, existe uma parte da população que é formada pelas peças que ainda irão ser produzidas.

Ou então se nos interessar a população de todos os portadores de febre tifoide, estaremos diante de um caso semelhante. Deve-se notar que, em geral, estudos realizados com base nos elementos da população amostrada terão, na verdade, seu interesse de aplicação voltado para os elementos restantes da população.

Este caso de amostragem não probabilística pode ocorrer também quando, embora se tenha a possibilidade de atingir toda a população, retiramos a amostra de uma parte que seja prontamente acessível. Assim, se fôssemos recolher uma amostra de um monte de minério poderíamos por simplificação retirar a amostra de uma camada próxima da superfície do monte, pois o acesso às porções interiores seria problemático.

3.1.5 Amostragem a esmo

É a amostragem em que o amostrador, para simplificar o processo, procura ser aleatório sem, no entanto, realizar propriamente o sorteio usando algum dispositivo aleatório confiável.

Por exemplo, se desejarmos retirar uma amostra de 100 parafusos de uma caixa contendo 10.000, evidentemente não faremos uma AAS, pois seria muito trabalhosa, mas retiramos simplesmente a esmo.

Os resultados da amostragem a esmo são, em geral, equivalentes aos da amostragem probabilística se a população é homogênea e se não existe a possibilidade de o amostrador ser inconscientemente influenciado por alguma característica dos elementos da população.

3.1.6 Amostragens intencionais

Enquadram-se aqui os diversos casos em que o amostrador deliberadamente escolhe certos elementos para pertencer à amostra por julgar tais elementos bem representativos. O perigo desse tipo de amostragem é grande pois o amostrador pode facilmente se enganar em seu pré-julgamento.

3.1.7 Amostragem por voluntários

Ocorre, por exemplo, no caso da aplicação experimental de uma nova droga em pacientes quando a ética obriga que haja concordância dos escolhidos.

3.2 Distribuições amostrais

O conceito de distribuição de probabilidade de uma variável aleatória será agora utilizado para caracterizar a distribuição dos diversos valores de uma variável em uma população.

Considere todas as amostras possíveis de tamanho ***n*** que podem ser retiradas (com ou sem reposição) de uma certa popu-

lação. Ao retirar uma amostra aleatória de uma população estaremos considerando cada valor da amostra como um valor de uma variável aleatória cuja distribuição de probabilidade é a mesma da população no instante da retirada desse elemento para a amostra.

Para cada amostra podemos calcular uma grandeza estatística (média, desvio padrão, variância, etc.) que varia de amostra para amostra obtendo um conjunto de valores da grandeza estatística calculada, denominada distribuição amostral. Se essas grandezas forem calculadas para a média obtém-se a distribuição amostral das médias, se for o desvio padrão obtém-se a distribuição amostral dos desvios padrões, etc.

Em consequência do fato de os valores da amostra serem aleatórios, qualquer quantidade calculada em função dos elementos da amostra também será uma variável aleatória.

As grandezas estatísticas calculadas para cada amostra são denominadas simplesmente de estatísticas e as grandezas calculadas para a população são denominadas parâmetros.

Para o estudo das distribuições amostrais devemos diferenciar as populações finitas das infinitas. Populações suficientemente grandes podem ser consideradas como infinitas, uma vez que não conseguimos abranger toda a população, independente do número de amostras que possamos extrair.

3.2.1 Distribuição amostral das médias

Sabemos que a média aritmética amostral é um estimador da média aritmética populacional. Como a média amostral é uma variável aleatória, busca-se conhecer sua distribuição de probabilidade.

Propriedades ou teoremas

A média da Distribuição Amostral das Médias, denotada por μ, é igual à média populacional μ.

$$E[\bar{x}] = \mu(\bar{x}) = \mu \qquad (3.1)$$

Se a população é infinita (ou muito grande) ou se a amostragem é com reposição, a Variância Amostral das Médias é igual à razão da variância populacional pelo tamanho da amostra, ou seja, a variância da média amostral é menor que a variância da população:

$$E[(\bar{x}-\mu)^2] = \sigma^2(\bar{x}) = \frac{\sigma^2}{n} \tag{3.2}$$

Se a população é finita (N < 20n ou n>5% de N) ou se a amostragem é sem reposição, então a variância da distribuição amostral das médias é dada por:

$$\sigma^2(\bar{x}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \tag{3.3}$$

Observação: ao termo (N-n)/(N-1) denomina-se Fator de Correção para População Finita (FCPF).

Teorema Central do Limite: se o tamanho da amostra for razoavelmente grande (n ≥ 30), então a DISTRIBUIÇÃO AMOSTRAL DA MÉDIA pode ser aproximada pela DISTRIBUIÇÃO NORMAL.

Se a população tem ou não Distribuição Normal com média μ e variância σ2, então a Distribuição das Médias Amostrais será normalmente distribuída com média μ e variância dada por:

Para População Infinita:

$$\frac{\sigma^2}{n} \tag{3.4}$$

Para População Finita:

$$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \tag{3.5}$$

3.2.2 Distribuição amostral das Frequências Relativas

Queremos determinar qual é a distribuição amostral da frequência relativa ou proporção. Seja X uma população infinita, e seja p a probabilidade (ou proporção) para um certo evento de X. Assim, q = 1 – p é a probabilidade do evento não ocorrer.

Seja (x_1,x_2,x_3,..., x_n) uma amostra aleatória de n elementos dessa população, e seja x o número de sucessos nesta amostra. Identifica-se facilmente que x é uma variável aleatória com Distribuição Binomial, tendo média= nxp e variância= nxpxq.

A Distribuição Amostral da Frequência Relativa $\hat{p} = f = \frac{x}{n}$ terá por parâmetros:

$$Média = E[f] = E\left[\frac{x}{n}\right] = \frac{np}{n} = p \tag{3.6}$$

$$Variância = V[f] = V\left[\frac{x}{n}\right] = \frac{npq}{n^2} = \frac{pq}{n} \tag{3.7}$$

Para n ≥ 30 a Distribuição Amostral da Frequência Relativa f será Normal com parâmetros:

$$f \overset{d}{\equiv} N\left(p\,;\,\frac{pq}{n}\right) \tag{3.8}$$

3.2.3 Distribuição Amostral de Variâncias

Seja a Variância Populacional designada por σ2 e a Variância Amostral designada por s^2. Logo, s^2 é o estimador de σ^2. Pode ser demonstrado que a Distribuição de s^2 tem parâmetros:

$$Média = E[s^2] = \sigma^2 \tag{3.9}$$

$$\text{Variância} = V\left[s^2\right] = \frac{2\sigma^4}{n-1} \tag{3.10}$$

Prova-se também que s^2 tem Distribuição Qui-Quadrado com (n-1) graus de liberdade, ou seja:

$$\frac{(n-1)\,S^2}{\sigma^2} \overset{d}{\equiv} \chi^2_{n-1} \tag{3.11}$$

Assim, a relação entre s^2 e σ^2 é dada por uma Distribuição Qui-Quadrado.

3.2.4 Distribuição Amostral da Soma ou Diferença de Duas Médias

Desejamos identificar a distribuição amostral do estimador $(\bar{x}_1 \pm \bar{x}_2)$. Sabe-se que a distribuição amostral da média é Normal com média = μ e variância = σ^2/n.

A soma ou diferença de duas médias terá também Distribuição Normal, com média igual à soma ou diferença das médias populacionais e variância igual à soma das variâncias populacionais.

$$(x_1 \pm x_2) \overset{d}{\equiv} N\left(\mu_1 \pm \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \tag{3.12}$$

3.2.5 Distribuição amostral da Soma ou Diferença de Duas Frequências Relativas

Desejamos identificar a distribuição amostral do estimador $(\bar{f}_1 \pm \bar{f}_2)$. Sabe-se que a distribuição amostral da frequência relativa, considerando-se $n \geq 30$, é Normal com média = p e variância = pq/n.

Considerando-se amostras independentes de duas populações, a soma ou diferença de duas proporções terá distribuição Normal, com média igual à soma ou diferença das proporções populacionais e variância igual à soma das variâncias populacionais.

$$f_1 \pm f_2) \overset{d}{\equiv} N\left(p_1 \pm p_2; \frac{p_1.q_1}{n_1} + \frac{p_2.q_2}{n_2} \right) \tag{3.13}$$

3.2.6 Distribuição Amostral das Médias quando a Variância da População é Desconhecida

Sabe-se que a distribuição amostral da média é Normal com média = μ e variância = σ^2/n, o que implica em sua distribuição normal padronizada ser representada por:

$$Z_i = \frac{\bar{x}_i - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{3.14}$$

Como não se conhece o valor da variância populacional σ^2, e portanto não se conhece também o valor do desvio padrão populacional σ, uma possibilidade é substituir o desvio padrão populacional pelo seu estimador, o desvio padrão amostral. Neste caso passamos a ter a estatística T.

$$T_i = \frac{\bar{x}_i - \mu}{\frac{s}{\sqrt{n}}} \tag{3.15}$$

Que possui Distribuição de Student com (n-1) graus de liberdade e portanto:

$$t_{n-1} = \frac{\bar{x}_i - \mu}{\frac{s}{\sqrt{n}}} \tag{3.16}$$

3.3 Estimação

A inferência estatística tem por objetivo fazer generalizações sobre uma população com base nos dados de amostra. Um dos itens básicos nesse processo é a estimação de parâmetros. A estimação pode ser por ponto ou por intervalo.

- *Estimativa por Ponto*: é a estimativa de um parâmetro populacional por um único valor.
- *Estimativa por Intervalo*: consiste em um intervalo em torno da estimativa por ponto de tal forma que ele possua probabilidade conhecida (nível de confiança **(1-α)**) de conter o verdadeiro valor do parâmetro. Este intervalo é conhecido por intervalo de confiança **(IC)**.

3.3.1 Propriedades de um Estimador

Por se tratar de uma variável aleatória, um estimador pode assumir valores segundo uma distribuição de probabilidades. A principal característica que um estimador deve apresentar é a de que, em média, ele seja igual ao parâmetro populacional que se deseja estimar.

3.3.2 Estimador Não Tendencioso

Seja T um estimador do parâmetro θ. O estimador T é não tendencioso (ou não viesado) se E[T] = θ.

Na prática, normalmente retiramos apenas uma amostra da população e produzimos através dela um único valor para o Estimador: uma Estimativa. Ainda que nosso estimador seja não tendencioso o valor da estimativa pode ser diferente do valor do parâmetro populacional. É desejável, portanto, que nosso estimador tenha variância pequena para reduzir a chance de nossa estimativa se afastar muito do valor do parâmetro.

3.3.3 Eficiência do Estimador

Sejam T1 e T2 dois estimadores não tendenciosos de um parâmetro, sendo V[T1] < V[T2]. Neste caso, T1 é dito mais eficiente que T2 e a eficiência relativa de T1 em relação a T2 é dada por:

$$ef(T_1, T_2) = \frac{V[T_2]}{V[T_1]} \qquad (3.17)$$

3.4 Erro amostral

Usando as Distribuições Amostrais é possível avaliar probabilisticamente o erro que se está cometendo por se usar uma amostra e não toda a população. Conforme anteriormente mencionado, a este erro dá-se o nome de Erro Amostral ou Erro de Estimativa e seu cálculo fica evidenciado na estimativa em forma de Intervalo de Confiança.

3.4.1 Intervalo de confiança para a média μ de uma população

Os intervalos de confiança para a média são tipicamente construídos com o estimador $\overline{X}$ no centro do intervalo quando conhecemos o Desvio Padrão da população (σ).

Quando o uso da distribuição normal está garantido o intervalo de confiança para a média é determinado por:

z	(1-α)
1,65	0,90
1,96	0,95
2,58	0,99

$$IC = \left(\bar{x} - z \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + z \frac{\sigma}{\sqrt{n}} \right) \qquad (3.18)$$

ou

$$IC = \left(\bar{x} - z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \; ; \; \bar{x} + z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right) \qquad (3.19)$$

No caso de população finita de tamanho N e amostragem sem reposição.

Exemplo:

A duração da vida de uma peça tem desvio padrão σ = 5 horas. Foram amostradas 100 peças observando-se a média de 500 horas. Construir o intervalo de confiança para a verdadeira duração da peça, com um nível de confiabilidade de 95%.

Temos que:

$\sigma = 5$; $n=100$; $\bar{x} = 500$; $(1-\alpha).100 = 95\%$

Da tabela da distribuição Normal (apêndice A) retiramos o valor da abscissa Zα/2 como sendo 1,96 (para 97,5%) e substituindo os valores na fórmula (3.20) para população infinita obtemos a inequação:

$$P\left(500 - 1{,}96.\frac{5}{\sqrt{100}} \leq \mu \leq 500 + 1{,}96.\frac{5}{\sqrt{100}}\right) = 95\%$$

Cujo cálculo resulta em uma margem de erro (ou erro de estimativa) de 0,98 horas.

Assim, o intervalo 500±0,98, ou [499,02 ; 500,98] contém a duração média da peça com 95% de confiança, significando com isso que permanece uma chance de 5% de a real duração da peça não pertencer a este intervalo.

Os intervalos de confiança mais frequentemente utilizados são os de 90%, 95% e 99%.

Quando σ é desconhecido

Quando o desvio padrão da população não é conhecido usa-se o desvio padrão da amostra como estimativa, substituindo-se σ por **s** nas equações para intervalo de confiança (Distribuição da população normal).

A distribuição "t de Student" é utilizada quando o desvio padrão da população é desconhecido. A forma da distribuição t é muito semelhante com a normal, sendo a principal diferença entre as duas distribuições o fato de que a distribuição t apresenta maior área nas caudas.

Para calcularmos t necessitamos conhecer o nível de confiança desejado e o número de graus de liberdade (gl= n-1).

O intervalo de confiança para a média é determinado por:

$$IC = \left(\bar{x} - t\,\frac{s}{\sqrt{n}}\ ;\ \bar{x} + t\,\frac{s}{\sqrt{n}}\right) \qquad (3.20)$$

ou

$$IC = (\bar{x} - t \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} ; \bar{x} + t \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}) \qquad (3.21)$$

No caso de população finita de tamanho N e amostragem sem reposição.

Exemplo:

A amostra: 9; 8; 12; 7; 9; 6; 11; 6; 10; 9 foi extraída de uma população normal. Construir o intervalo de confiança para a média ao nível de 95%.

Solução: calculando a média aritmética e o desvio padrão da amostra obtemos os seguintes resultados:

$$\bar{x} = 8{,}7 \qquad e \qquad s = 2$$

Considerando que (1-α) = 95% e g.l.= 9 (graus de liberdade= n-1) da tabela da distribuição t (Apêndice A) retiramos o valor 2,26 para a abscissa tα/2. Com tais valores o erro de estimativa (ou margem de erro) é 1,43 e o intervalo de confiança 8,7±1,43 torna-se [7,27 ; 10,13], o qual contém a média da população com 95% de confiança.

3.4.2 Intervalo de confiança para a proporção π de uma população

A distribuição amostral da proporção é aproximadamente normal para n ≥ 30, pode-se então usar a distribuição normal para estabelecer o intervalo de confiança:

$$IC=(p - z \sqrt{\frac{p(1-p)}{n}} ; p + z \sqrt{\frac{p(1-p)}{n}}) \qquad (3.22)$$

ou

$$IC=(p - z \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}} ; p + z \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}) \qquad (3.23)$$

No caso de população finita de tamanho N e amostragem sem reposição.

Exercício:

1. Seja a população formada pelos valores obtidos no lançamento de dois dados, obter a distribuição amostral das médias, calculando a média e o desvio padrão dessa distribuição.

   As amostras possíveis são:

   {1,1}, {1,2}, {1,3}, {1,4}, {1,5}, {1,6}
   {2,1}, {2,2}, {2,3}, {2,4}, {2,5}, {2,6}
   {3,1}, {3,2}, {3,3}, {3,4}, {3,5}, {3,6}
   {4,1}, {4,2}, {4,3}, {4,4}, {4,5}, {4,6}
   {5,1}, {5,2}, {5,3}, {5,4}, {5,5}, {5,6}
   {6,1}, {6,2}, {6,3}, {6,4}, {6,5}, {6,6}

4
Teste de hipóteses

Trata-se de uma regra de decisão para aceitar ou rejeitar uma hipótese estatística (acerca de parâmetros populacionais) com base nos elementos da amostra (estimadores).

A teoria dos Testes de Hipóteses exige a formulação de duas hipóteses mutuamente excludentes: a hipótese nula e a hipótese alternativa.

4.1 Hipótese Nula - H_0

É a hipótese estatística a ser testada que é aceita como verdadeira até prova estatística em contrário. Constitui o ponto de partida para a análise dos dados, e em geral é formulada em termos de igualdade (=) entre dois parâmetros ou entre um parâmetro e um valor constante. Geralmente representa o contrário do que desejamos provar, ou seja, é planejada para que se obtenham evidências para sua rejeição, acarretando com isto a aceitação de uma hipótese alternativa.

Exemplos de representação matemática para os casos apresentados:

a) $\mathbf{H_0: \mu_A = \mu_B}$ (processador A e processador B)

b) $\mathbf{H_0: \mu_A = \mu_D}$ (Antes e Depois)

c) $\mathbf{H_0: p_A = p_D}$ (Antes e Depois)

4.2 Hipótese Alternativa – Ha

É normalmente formulada em termos de uma desigualdade ($\neq$, $>$, $<$), representando aquilo que se deseja provar, e que será aceita sempre que se possa rejeitar a hipótese nula.

Aceitar a hipótese alternativa é uma posição mais forte do que aceitar a hipótese nula, pois neste caso é necessário que se obtenha as evidências necessárias enquanto a hipótese nula é aceita por falta de evidências.

Exemplos de representação matemática para os casos apresentados:

a) $\mathbf{H_a: \mu_A \neq \mu_B}$ (processador A e processador B). Teste bicaudal ou bilateral.

b) $\mathbf{H_a: \mu_A < \mu_D}$ (Antes e Depois). Teste unicaudal ou unilateral à direita.

c) $\mathbf{H_a: p_A > p_D}$ (Antes e Depois). Teste unicaudal ou unilateral à esquerda.

4.3 Aceitação da Hipótese Nula - H_0

A hipótese nula deve ser aceita e a hipótese alternativa deve ser rejeitada. Aceitar a hipótese nula significa que não há evidências suficientes para rejeitá-la e, portanto, ela deve ser verdadeira. (Observe que não é afirmado que a hipótese nula seja verdadeira)

A média da amostra não é significativamente diferente da média μ_0 (a média estabelecida na H_0). É razoável aceitar que a diferença entre a média da amostra e a média da população seja somente devida à aleatoriedade da amostra escolhida (ou acaso). O resultado não é estatisticamente significante.

4.4 Rejeição da Hipótese Nula - H_0

A hipótese nula deve ser rejeitada e a hipótese alternativa deve ser aceita. Aceitar a hipótese alternativa significa que há evidências de que a hipótese nula seja falsa.

A média da amostra é significativamente diferente da média μ_0 (a média estabelecida na H_0). Não é razoável aceitar que a diferença entre a média da amostra e a média da população seja somente devida à aleatoriedade da amostra escolhida (ou acaso).

O resultado é estatisticamente significante. Isto significa que as evidências contra a hipótese nula alcançaram o erro tolerado ou nível de significância do teste.

4.5 Tipos de erro

No teste de uma hipótese estatística há dois tipos possíveis de erro:

4.5.1 Erro Tipo I

Rejeitar a hipótese nula quando ela é de fato verdadeira. A probabilidade de ocorrência deste erro é denotada por α. É denominada de nível de significância do teste. Só ocorre quando rejeitamos H_0.

4.5.2 Erro Tipo II

Aceitar a hipótese nula quando ela é de fato falsa. A probabilidade de ocorrência deste erro é denotada por β. Só ocorre quando aceitamos H_0.

		Realidade	
		H_0 Verdadeira	**H_0 Falsa**
Decisão	**Aceitar H_0**	Decisão Correta **p = 1 - α**	**Erro tipo II** **β**
	Rejeitar H_0	Erro tipo I α	Decisão Correta **p = 1 - β**

Desejamos evidentemente reduzir ao mínimo as probabilidades dos dois tipos de erro. Isto é muito difícil porque para uma amostra de determinado tamanho à medida que um tipo de erro diminui o outro aumenta e vice-versa. A redução simultânea dos dois tipos de erro pode ser alcançada pelo aumento do tamanho da amostra. O Teste de Hipóteses compreende o controle dos dois tipos de erro.

Nas aplicações práticas observa-se que o erro Tipo I é socialmente mais importante que o erro Tipo II. Observe por exemplo o caso de um julgamento: condenar um inocente *versus* inocentar um culpado.

4.6 Testes de Significância

Testes realizados com níveis de significância $\alpha \leq 5\%$ são considerados altamente significativos. Testes com níveis de significância $5\% < \alpha < 10\%$ são considerados provavelmente significativos.

Testes com níveis de significância $\alpha \geq 10\%$ são considerados pouco significativos.

4.6.1 Região crítica

É a região onde os valores da estatística dos testes levam à rejeição da hipótese nula. A sua área é igual ao nível de significância e sua direção é a mesma da hipótese alternativa.

Unilateral à esquerda: H0: $\mu = 50$

Ha: $\mu < 50$

Unilateral à direita: H0: $\mu = 50$

Ha: $\mu > 50$

Bilateral: H0: $\mu = 50$

Ha: $\mu \neq 50$

Etapas da realização de um Teste

A metodologia de testes compreende as seguintes etapas:

1) Identificar Ho;
2) Identificar Ha. Aqui se define o tipo de teste a ser aplicado: unilateral ou bilateral;
3) Identificar o modelo de distribuição de probabilidades de referência para o parâmetro em teste;
4) Fixar o nível de significância do teste (α);
5) Construir a Região Crítica para o tipo de teste escolhido, que é a região de rejeição da hipótese nula. Isto automaticamente define a Região de Aceitação de Ho, bem como determina o Valor Crítico do parâmetro que vai balizar a comparação com o estimador a ser utilizado no teste;

6) Calcular o estimador segundo o modelo de referência para obter o Valor de Teste e verificar em que região ele se situa por comparação com o Valor Crítico;

7) Formular a conclusão do teste.

Exemplo:

O peso médio de litros de leite de embalagens cheias em uma linha de produção está sendo estudado. O padrão prevê um conteúdo médio de 1000 ml por embalagem. Sabe-se que o desvio padrão é de 10 ml e que a variável tem distribuição normal.

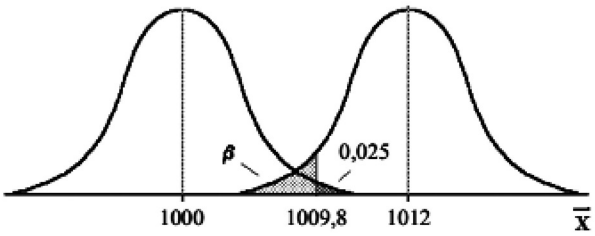
Para encontrar a probabilidade de erro tipo II ao testamos se a média é diferente de 1000 ml sendo que o real conteúdo médio da embalagem é 1012 ml ao nível de 5% de significância, utilizamos 4 unidades amostrais escolhidas ao acaso, e, temos:

H0: $\mu = 1000$
H1: $\mu \neq 1000$

P (erro tipo II) = P (aceitar H0/ H0 é falsa) = ?

$$Z\alpha/2 = Z=0,025 = 1,96$$

$$1,96 = P(\frac{\bar{x}-1000}{\frac{10}{\sqrt{4}}}) \Leftrightarrow \bar{x} = 1009,8$$



Erros tipo I e tipo II

P (aceitar H0/ H0 é falsa) = P (X < 1009,8 / μ = 1012)

$$P(\frac{\bar{x}-\mu}{\frac{\delta}{\sqrt{n}}}) < \frac{1000-1012\mu}{\frac{10}{4}}$$

$$= P (Z < -0,44) = 0,33$$

Ou seja, a probabilidade de não rejeitarmos Ho, quando a média real da embalagem é de 1012 ml é de 0,33. A partir dessa informação podemos obter o poder do teste que é de 1-β=1-0,33=0,67.

Esta é a abordagem clássica do teste de significância. Com o advento de computadores e *softwares* especialistas surgiu a abordagem do *p-value* (ou valor p) para auxiliar na tomada de decisão do teste.

O *p-value* (ou probabilidade de significância) é definido como a probabilidade da estatística do teste acusar um resultado tão ou mais distante do esperado, considerando o resultado observado na amostra em teste, sem rejeitar a hipótese nula. Na prática, o *p-value* representa o valor da probabilidade do modelo de referência calculada para o Valor de Teste, considerando Ho como verdadeira.

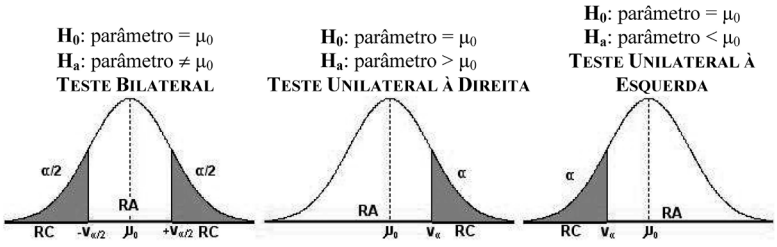
Encarando o *p-value* como o limiar para rejeição de Ho, o julgamento do teste passa a ser como se segue:

1. Se o *p-value* é maior do que o nível de significância estabelecido para o teste, aceita-se Ho.

2. Se o *p-value* for menor ou igual ao nível de significância, rejeita-se Ho. Quanto menor o *p-value*, mais evidências existem de que Ho deve ser rejeitada.

4.6.2 Região de aceitação

Em função do tipo de teste escolhido a posição da região crítica e da região de aceitação de Ho fica definida conforme ilustrado abaixo.



Exemplo:

Suponhamos que uma indústria compre de certo fabricante parafusos cuja carga média de ruptura por tração é especificada em 50 Kg, o desvio-padrão das cargas de ruptura é supostamente igual a 4 Kg. O comprador deseja verificar se um grande lote de parafusos recebidos deve ser considerado satisfatório, no entanto existe alguma razão para se temer que a carga média de ruptura seja eventualmente inferior a 50 Kg. Se for superior não preocupa o comprador, pois neste caso os parafusos seriam de melhor qualidade que o especificado. Neste exemplo, a hipótese do comprador é que a carga média da ruptura é inferior a 50 Kg.

O comprador pode ter o seguinte critério para decidir se compra ou não o lote: Resolve tomar uma amostra aleatória simples de 25 parafusos e submetê-los ao ensaio de ruptura. Se a carga média de ruptura observada nesta amostra for maior que 48 Kg ele comprará o lote, caso contrário se recusará a comprar.

- Hipótese Nula (H0): É um valor suposto para um parâmetro. No exemplo acima H0:μ=50.
- Hipótese Alternativa (Ha): É uma hipótese que contraria a hipótese nula, complementar de H0, no exemplo Ha: μ <50.

Ou seja, no exemplo
Ho: $\mu = 50$
Ha: $\mu < 50$

Supondo H0 verdadeira, X da amostra aleatória de 25 valores será uma variável aleatória com média também de 50 Kg e desvio padrão $\frac{\sigma}{\sqrt{n}}$.

No exemplo

$$\sigma = \frac{4}{\sqrt{25}} = 0{,}8$$

Sabemos que X é aproximadamente normal, então podemos calcular a probabilidade de obtermos um valor inferior a 48.

$$P(X < 48) = P(x - < \mu - P(\frac{x - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{48 - 50}{0{,}8}) = P(Z < 2{,}5) = 0{,}0062$$

Existe uma probabilidade de 0,0062 de que, mesmo sendo a hipótese H0 verdadeira, X assuma um valor na região que leva à rejeição de H0, conforme critério adotado anteriormente.

Exercícios propostos

1) Uma amostra de 25 elementos resultou média 13,5 com desvio padrão de 4,4. Efetuar o teste ao nível de 1% para a hipótese que a média seja inferior a 16.

2) As estaturas de 20 recém nascidos foram tomadas no Departamento de Pediatria da FMRP, cujos resultados são, em cm:

41 50 52 49 49 54 50 47 52 49
50 52 50 47 49 51 46 50 49 50

a) Suponha inicialmente que a população das estaturas é normalmente distribuída com variância 2 cm^2; Teste a hipótese de que a média seja diferente de 50cm ($\propto$=0,05).

b) Faça o mesmo teste para a média, mas agora desconhecendo a variância ($\propto$=0,05).

3) Um processo deveria produzir mesas com 0,85m de altura. O engenheiro desconfia que as mesas que estão sendo produzidas são menores que o especificado. Uma amostra de 8 mesas foi coletada e indicou média 0,847m. Sabendo que o desvio padrão é σ= 0,010m, teste a hipótese do engenheiro usando um nível de significância de 3%.

4) As condições de mortalidade de uma região são tais que a proporção de nascidos que sobrevivem até 60 anos é de 0,6. Testar essa hipótese ao nível de 5% se em 1000 nascimentos amostrados aleatoriamente, verificou-se 530 sobreviventes até 60 anos.

5
Análise de Variância

Para se fazer uma análise de variância seria necessária a verificação de atendimento às pressuposições que sustentam o modelo, as quais serão discutidas adiante. Por ora será introduzida apenas a metodologia que envolve a análise de variância, com o intuito de familiarização com os cálculos e o seu significado.

A ideia por trás da análise de variância é comparar a variação devida aos tratamentos (as variedades no caso) com a variação devida ao acaso ou resíduo, como normalmente é designado esse tipo de variação. O modelo exige a aplicação de muitas fórmulas e também o conhecimento da notação empregada.

A Tabela 5.1 simboliza um experimento com k tratamentos, sendo que cada tratamento tem r repetições.

Tabela 5.1 – Experimento Inteiramente ao Acaso

Discriminação	**Tratamento**					**Total**
	1	**2**	**3**	**...**	**k**	
	y_{11}	y_{21}	y_{31}		y_{k1}	
	y_{12}	y_{22}	y_{32}		y_{k2}	
	y_{13}	y_{23}	y_{33}		y_{k3}	
	:	:	:		:	
	y_{1r}	y_{2r}	y_{3r}		y_{kr}	
Total	T_1	T_2	T_3		T_k	$\Sigma T=\Sigma y$
Nº repetições	r	r	r		r	n=k.r
Média	$\overline{y}_1$	$\overline{y}_2$	$\overline{y}_3$		$\overline{y}_k$	

A soma dos resultados das r repetições de um mesmo tratamento constitui o total desse tratamento e as médias dos tratamentos são designadas por $\overline{y}_{1,}\ \overline{y}_{2,}\ \overline{y}_{3,\ldots}\ \overline{y}_{k}$. O total geral é dado pela soma dos totais de tratamentos.

Para se proceder à analise de variância de um experimento inteiramente ao acaso é preciso calcular as seguintes quantidades:

a) os graus de liberdade:
de tratamentos: **k – 1**
do total: **n – 1**, com n= k.r
do resíduo: (n – 1) – (k – 1) = **n - k**

b) o valor C, conhecido pela designação de correção, calculado como a razão do total geral elevado ao quadrado pelo número de observações:

$$C = \frac{(\sum y)^2}{n} \qquad (5.1)$$

c) a soma de quadrados total:

$$SQT = \Sigma y^2 - C \qquad (5.2)$$

d) a soma de quadrados de tratamentos:

$$SQTr = \Sigma T^2/r - C \qquad (5.3)$$

e) a soma de quadrados de resíduo:

$$SQR = SQT - SQTr \qquad (5.4)$$

f) o quadrado médio de tratamentos:

$$QMTr = SQTr/(k - 1) \tag{5.5}$$

g) o quadrado médio de resíduos:

$$QMR = SQR/(n - k) \tag{5.6}$$

h) o valor da estatística F:

$$F = QMTr/QMR \tag{5.7}$$

Observe que os quadrados médios são obtidos pela divisão das somas dos quadrados correspondentes pelos respectivos graus de liberdade. Estes quadrados médios representam, na verdade, as Variâncias dos Tratamentos e dos Resíduos, respectivamente.

Assim, o valor F representa então a razão da Variância Explicada pela variação dos tratamentos pela Variância dos Resíduos, cuja variação não é explicada.

Estas quantidades são calculadas e são apresentadas em uma tabela de análise de variância, cuja forma de apresentação é padrão e é mostrada na Tabela 5.2 a seguir.

Tabela 5.2 – ANOVA de um experimento inteiramente ao acaso

Causas de variação	GL	SQ	QM	F
Tratamentos	k – 1	SQTr	QMTr	F
Resíduos	n – k	SQR	QMR	
Total	n – 1	SQT		

5.1 Exemplo de aplicação

Tomando os dados do exemplo apresentado na Tabela podemos construir a Tabela correspondente à Tabela 5.2 em conformidade com a estrutura ali apresentada para obter a tabela 5.3 a seguir:

Tabela 5.3 – Exemplo de Experimento Inteiramente ao Acaso

Operações	**Tratamento**				
	A	**B**	**C**	**D**	**Total**
	25	31	22	33	
	26	25	26	29	
	20	28	28	31	
	23	27	25	34	
	21	24	29	28	
T = Total Tratamentos	115	135	130	155	535
Σy^2 = Soma Quadrados	2671	3675	3410	4831	14587
T^2 = Quadrados Tratamentos	13225	18225	16900	24025	72375
R = n° de repetições	5	5	5	5	20
Média	23	27	26	31	

Para proceder à análise de variância desse experimento precisamos calcular:

a) os graus de liberdade:
de tratamentos: **k – 1 =** (4 – 1) = 3
do total: **n – 1 =** (4x5 – 1) = 19
do resíduo: (n – 1) – (k – 1) = **n – k =** (20 – 4) = 16

b) o valor C (correção) calculado como:

$$C = \frac{(\sum y)^2}{n} = \frac{(535)^2}{20} = 14311{,}25$$

c) a soma de quadrados total:

$$SQT = \sum y^2 - C = 14587 - 14311{,}25 = 275{,}75$$

d) a soma de quadrados de tratamentos:

$$SQTr = \frac{\sum T^2}{r} - C = \frac{72375}{5} - 14311{,}25 = 163{,}75$$

e) a soma de quadrados de resíduo:

SQR = SQT – SQTr = 275,75 – 163,75 = **112,00**

f) o quadrado médio de tratamentos:

$$QMTr = \frac{SQTr}{(k-1)} = \frac{163{,}75}{3} = 54{,}58$$

g) o quadrado médio de resíduos:

$$QMR = \frac{SQR}{(n-k)} = \frac{112{,}0}{16} = 7{,}0$$

h) o valor da estatística:

$$F = \frac{QMTr}{QMR} = \frac{54{,}58}{7{,}0} = 7{,}8$$

Com esses dados a Tabela da ANOVA fica:

Tabela 4.4 – ANOVA para o Exemplo

Causas de variação	GL	SQ	QM	F	P
Tratamentos	3	163,75	54,58	7,80	0,0020
Resíduos	16	112,00	7,0		
Total	19	275,75			

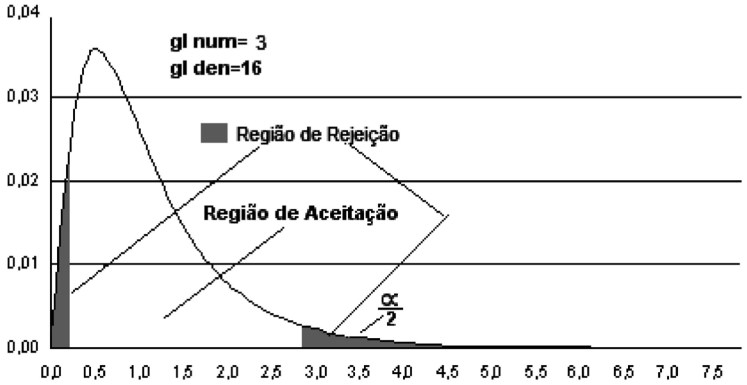
O *p-value* **P=** 0,0020 foi obtido através da chamada função **DISTF**(x; graus_liberdade1; graus_liberdade2) do Excel, com os parâmetros x= 7,80, gl_numerador= 3 e gl_denominador= 16. Como 0,0020 (0,2%) é bem menor que α**=** 5%, a esse nível de significância (e 95% de confiança) rejeita-se Ho, que no caso é a hipótese de que as médias de variedades são iguais.

5.1.1 Interpretação do valor de F

Primeiramente convém aprender como se obtém o correspondente valor crítico de F. Esse valor é retirado da Tabela da Distribuição F de Snedecor (Anexo C), para o caso presente de 5% de nível de significância temos graus de liberdade do numerador igual a 3 e graus de liberdade do denominador igual a 16 e a consulta à tabela com esses parâmetros fornece F=3,24.

Continuando com o caso de exemplo, observa-se então que o F crítico, aquele que se obtém do modelo probabilístico, atingiu o valor 3,24. O F de teste, obtido por cálculos a partir dos valores observados, obteve o valor 7,80.

À luz dos conceitos de testes de significância do capítulo anterior constata-se que o valor de teste é superior ao valor crítico, o que nos coloca na Região de Rejeição da Hipótese Nula, conforme se pode notar pela ilustração a seguir.



Seja pela abordagem clássica, conforme mostrado neste item, seja pela abordagem do *p-value*, conforme mostrado no item anterior, ao nível de significância de 5% o pesquisador deve rejeitar a hipótese de que as médias sejam iguais, o que permite concluir, portanto, que as variedades A, B, C e D não têm médias de produção iguais.

É importante notar que o teste não comprova nenhuma das hipóteses. Havendo significância no resultado, o teste indica que há evidências contra a hipótese da nulidade. O pesquisador deve então rejeitar a hipótese de igualdade das médias e ao fazer isso o pesquisador corre um risco de 5% (significância) de estar cometendo um erro ao tomar essa decisão. Alternativamente, ao tomar essa decisão o pesquisador tem uma confiança de 95% de não estar cometendo um erro.

É importante também destacar que o pesquisador pôde concluir pela não igualdade das produções das variedades porque ele procedeu à casualização quando do delineamento, evitando qualquer tendenciosidade ou favoritismo.

A análise de variância aqui mostrada é indicada para experimentos feitos de acordo com as normas técnicas. É essencial que as unidades experimentais utilizadas no experimento sejam, de início, similares e que os tratamentos sejam designados às unidades experimentais através de processo aleatório.

5.1.2 Ferramenta ANOVA do Excel

A ferramenta do Excel **ANOVA:** fator único chamada com os dados do exemplo deste capítulo fornece os seguintes resultados, onde se procedeu à formatação dos dados para exibição do mesmo número de casas decimais para facilidade de comparação com os resultados já obtidos na tabela 4.4.

Ferramenta **ANOVA fator único** do Excel.

Resumo

Grupo	Contagem	Soma	Média	Variância
A	5	115	23	6,5
B	5	135	27	7,5
C	5	130	26	7,5
D	5	155	31	6,5

ANOVA

Fonte da variação	SQ	*gl*	MQ	*F*	valor-P	*F* crítico
Entre grupos	163,75	3	54,58	0,807	0,0020	3,24
Dentro dos grupos	112,00	16	7,00			
Total	275,75	19				

Observe que os resultados da ferramenta incorporam também uma análise descritiva dos tratamentos. A tabela ANOVA apresenta elementos para julgamento do Teste F, tanto pela abordagem clássica (comparação de F calculado com F crítico) quanto pela abordagem do *p-value* (comparação do valor-p com o alfa de significância).

Para facilidade de comparação dos resultados emitidos pela ferramenta ANOVA do Excel, repetimos aqui a tabela ANOVA elaborada manualmente conforme tabela 4.4.

Tabela 4.5 – ANOVA para o Exemplo

Causas de variação	GL	SQ	QM	F	P
Tratamentos	3	163,75	54,58	7,80	0,0020
Resíduos	16	112,00	7,0		
Total	**19**	**275,75**			

Exercícios:

1) Explicar como proceder para designar 5 tratamentos (A, B, C, D e E) para 25 unidades experimentais similares.

2) Os dados obtidos num experimento inteiramente ao acaso estão apresentados na tabela abaixo. Calcule as médias, faça um gráfico e proceda à análise da variância interpretando o resultado.

Operações	Tratamento				
	A	B	C	D	E
	12	11	8	15	16
	13	8	11	17	17
	10	7	13	17	19
	13	9	12	17	16
	13	9	12	14	16
	11	10	10	16	18

3) Num laboratório são usados quatro voltímetros diferentes. Para verificar se os quatro voltímetros estão igualmente calibrados mediu-se a mesma tensão constante de 110 volts, cinco vezes com cada voltímetro. Os dados obtidos estão na tabela abaixo. Faça uma análise de variância e interprete o resultado.

Operações	Voltímetros			
	A	B	C	D
	117	115	118	125
	120	110	123	121
	114	116	119	123
	119	115	122	118
	115	114	118	118

4) Os dados abaixo representam a produção, por hectare, de três safras de milho plantadas com 4 tipos de fertilizantes. Utilizando a ANOVA, teste se existe diferença significativa na produção por hectare. Usar o nível de significância de 5%.

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

6
Correlação

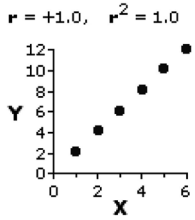
O problema da correlação está ligado ao grau de relação entre duas ou mais variáveis quantitativas. Alguns métodos estatísticos visam estudar a associação entre duas ou mais variáveis aleatórias.

A teoria de Correlação e Regressão ocupa um lugar de destaque por seu uso mais difundido.

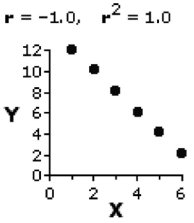
Nesses estudos, o primeiro objetivo é o de analisar o comportamento simultâneo das variáveis, tomadas duas a duas, verificando se a variação positiva ou negativa de uma delas está associada a uma variação positiva ou negativa da outra, ou em outras palavras, se não há nenhuma forma de dependência entre elas.

Em uma primeira análise exploratória podemos ter um diagrama cartesiano bidimensional. Tal diagrama chama-se diagrama de dispersão. O diagrama de dispersão permite visualizar o grau de associação entre as variáveis e a tendência de variação conjunta que apresentam.

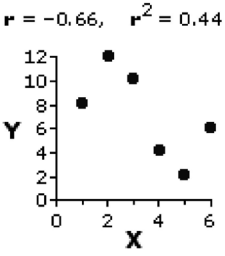
Em outras palavras, um diagrama de dispersão pode nos dar ideia se a correlação é:



a) **Linear positiva: se os pontos do diagrama têm como "imagem" uma reta ascendente.**



b) **Linear negativa: se os pontos têm como "imagem" uma reta descendente.**



c) **Não há relação: se os pontos apresentam-se dispersos, não oferecendo ideia de uma "imagem" definida.**

Uma medida utilizada em correlação linear é conhecida como coeficiente de correlação linear de Pearson, definido por:

$$r = \frac{Cov(X, Y)}{s_x . s_y} \qquad (6.1)$$

Sendo Sx e Sy os desvios padrões amostrais de X e Y, respectivamente.

A expressão (6.1) pode ser colocada na forma

$$r = \frac{\sum_{i=1}^{n} X_1 Y_1 - \frac{\sum_{i=1}^{n} X_1 \sum_{i=1}^{n} Y_1}{n}}{\sqrt{\left[\sum_{i=1}^{n} X_1^2 - \frac{\left(\sum_{i=1}^{n} X_1\right)^2}{n}\right]\left[n\sum_{i=1}^{n} Y_1^2 - \frac{\left(\sum_{i=1}^{n} Y_1\right)^2}{n}\right]}} \qquad (6.2)$$

O coeficiente de correlação linear r é adimensional e demonstra que pode variar de -1 a 1, ou seja $-1 \leq r \geq 1$. Quando r = -1 tem-se a correlação linear negativa perfeita, enquanto que para r = 1 a correlação linear é positiva perfeita. Para r = 0 não existe correlação linear entre as variáveis, podendo existir relação de outro tipo.

Quanto mais o valor de r aproxima-se de -1 ou 1 melhor é o grau de correlação entre as variáveis. A interpretação do valor de r depende dos objetivos de sua utilização e as razões para os quais este valor é calculado. O valor de r pode ser qualitativamente avaliado da seguinte forma:

- Se $0 < | r | < 0,3$ - existe fraca correlação linear;
- Se $0,30 < | r | < 0,60$ - existe moderada correlação linear;
- Se $0,60 < | r | < 0,90$ - existe forte correlação linear;
- Se $0,90 < | r | < 1,00$ - existe correlação linear muito forte.

Exemplo:

A tabela seguinte fornece valores das variáveis X (Volume por recipiente (ml)) e Y (Tempo de autoclavagem (min.)). **Período mínimo recomendado para esterilização de meio para cultura de tecidos de plantas a** 121°C e 1,05kg/cm/cm2 = 105kPa **(extraído do Catálogo da Sigma). Calcular o coeficiente de correlação linear de Pearson e construir o diagrama de dispersão.**

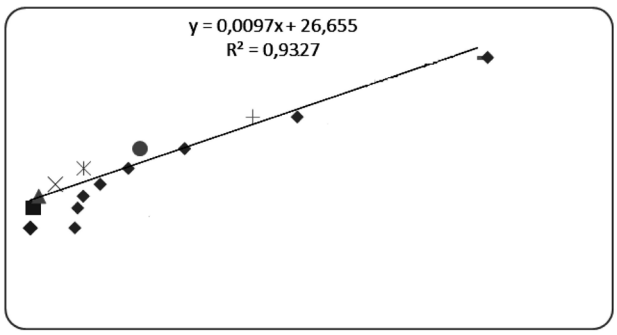
Volume por recipiente (ml)	Tempo de autoclavagem (min.)
5	20
50	25
100	28
250	31
500	35
1000	40
2000	48
4000	63

Como $\sum X=7925$, $\sum Y=290$, $\sum XY=417800$, $\sum X^2=21325625$, $\sum Y^2=11868$

Então por 6.2,

$$r = \frac{\sum_{i=1}^{n} X_1 Y_1 - \frac{\sum_{i=1}^{n} X_1 \sum_{i=1}^{n} Y_1}{n}}{\sqrt{\left[\sum_{i=1}^{n} X_1^2 - \frac{\left(\sum_{i=1}^{n} X_1\right)^2}{n}\right]\left[n\sum_{i=1}^{n} Y_1^2 - \frac{\left(\sum_{i=1}^{n} Y_1\right)^2}{n}\right]}} \quad r = \frac{16324,8}{19307,0} = 0,8450$$

Figura 6.4 – Exemplo de diagrama de dispersão



Como ***r*** não está muito próximo de 1 temos uma forte correlação linear positiva. Isto significa que quanto maior o volume do recipiente maior será o tempo de autoclavagem.

O R^2 denominado coeficiente de determinação nos diz que 93% da variabilidade no tempo de utilização do autoclave pode ser explicada em função da capacidade do recepiente.

Problemas propostos

1. A seguinte amostra de tamanho 7 foi obtida da variável aleatória bidimensional (X,Y). Utilizando estes valores calcule o coeficiente do correlação linear.

X	1	2	3	4	5	6	7
Y	9	7	6	6	5	4	2

2. O alongamento (X) de uma mola foi medido em função de 5 valores (Y) da carga aplicada. Calcular o coeficiente de correlação de linear de Pearson e construir o diagrama de dispersão.

Carga (kg)	4	5	6	7	8
Alogamento (cm)	7,3	8,5	9,0	9,5	9,9

3. As importações de uma determinada matéria prima (em toneladas) no período de 1980 a 1986 estão na tabela seguinte:

Ano (X)	1980	1981	1982	1983	1984	1985	1986
Importações (Y)	97	86	74	64	58	43	39

Pede-se:

a) Calcular o coeficiente linear de Pearson e interpretar o resultado;

b) Construir o diagrama de dispersão.

7
Regressão

Muitas vezes estamos interessados em estabelecer uma relação funcional entre duas variáveis para estudarmos o fenômeno pela qual ela é regida. A regressão e a correlação são técnicas utilizadas para estimar uma relação que possa existir na população, enquanto as técnicas anteriormente estudadas (Medidas de Tendência Central e de Dispersão: Média, Desvio Padrão, Variância, etc.) servem para estimar um único parâmetro populacional.

A análise de correlação e regressão compreende a análise de dados amostrais para saber como duas ou mais variáveis estão relacionada uma com a outra numa população. A correlação mede a força ou grau de relacionamento entre duas variáveis; a regressão dá a equação que descreve o relacionamento em termos matemáticos.

Os dados para análise de regressão e correlação provêm de observações de variáveis emparelhadas. Na regressão pressupõe-se alguma relação de causa e efeito, de explanação do comportamento entre as variáveis. Ex: a idade e a altura de cada indivíduo; a alíquota de imposto e a arrecadação; preço e quantidade.

Análise de regressão é uma técnica de modelagem utilizada para analisar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes (X_1, X_2, X_3,..., X_n). O objetivo dessa técnica é identificar (estimar) uma função que descreve, o mais próximo possível, a relação entre essas variáveis e assim podermos predizer o valor que a variável dependente (Y) irá assumir para um determinado valor da variável independente (X).

Exemplos de relação entre variáveis são o consumo em relação à taxa de inflação; a produção de leite e temperatura ambiente; a resistência de um material e sua composição química; o número de peças com defeitos e a experiência; receita e gasto com publicidade e etc.

O modelo de regressão poderá ser escrito genericamente como:

$$Y = f(X_1, X_2, X_3, ..., X_N) + \varepsilon \,, \qquad (7.1)$$

onde o termo ε representa uma perturbação aleatória na função, ou o erro da aproximação. O número de variáveis independentes varia de uma aplicação para outra. Quando se tem apenas uma variável independente chama-se Modelo de Regressão Simples e quando se tem mais de uma variável independente chama-se Modelo de Regressão Múltipla.

A forma da função f (.) também varia, podendo ser representada por um modelo linear, polinomial ou até mesmo uma função não linear.

7.1 Regressão Linear Simples

Este modelo é utilizado quando existe uma relação linear entre a variável independente e a variável dependente (neste caso apenas uma). A função que expressa esse modelo será dada pela forma abaixo

:

$$Y_i = b_0 + b_1 X_i + \varepsilon \qquad (7.2)$$

Uma vez escolhido o modelo de regressão devem-se estimar seu parâmetro, neste caso os coeficientes da equação da reta b_0, b_1. Isso pode ser feito a partir da aplicação do Método dos Mínimos Quadrados:

$$\hat{b}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{7.3}$$

E o estimador $\hat{b}_0$ pode ser calculado a partir de:

$$\hat{b}_o = \overline{Y} - \hat{b}_1 \overline{X} \tag{7.4}$$

Sendo que a equação de estimativa será dada por:

$$\hat{Y} = \hat{b}_o + \hat{b}_1 \overline{X} \tag{7.5}$$

Exemplo:

Em uma determinada região do país foram coletados os índices pluviométricos e a produção de leite do tipo C. Sabendo-se que existe uma previsão para o próximo ano de um índice pluviométrico de 24mm, determine então a produção de leite dessa região.

Anos	Produção de Leite C (1.000.000 litros)	Índice pluviométrico (mm)
1970	26	23
1971	25	21
1972	31	28
1973	29	27
1974	27	23
1975	31	28
1976	32	27
1977	28	22

Anos	Produção de Leite C (1.000.000 litros)	Índice pluviométrico (mm)
1978	30	26
1979	30	25

Resolução:

	Y	X	y	x	y^2	x^2	xy
1970	26	23	-2.9	-2	8.41	4	5.8
1971	25	21	-3.9	-4	15.21	16	15.6
1972	31	28	2.1	3	4.41	9	6.3
1973	29	27	0.1	2	0.01	4	0.2
1974	27	23	-1.9	-2	3.61	4	3.8
1975	31	28	2.1	3	4.41	9	6.3
1976	32	27	3.1	2	9.61	4	6.2
1977	28	22	-0.9	-3	0.81	9	2.7
1978	30	26	1.1	1	1.21	1	1.1
1979	30	25	1.1	0	1.21	0	0
Soma	**289**	**250**	**0**	**0**	**48.9**	**60**	**48**
Média	**28.9**	**25**	**0**	**0**	**4.89**	**6**	**4.8**

$$\hat{b}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \text{ , assim } \hat{b}_1 = \frac{48}{60} = 0.8$$

e

$$\hat{b}_o = \overline{Y} - \hat{b}_1 \overline{X} \text{ ,}$$

que

$$\hat{b}_o = 28{,}9 - 0.8.25 = 8{,}9$$

Assim, a equação pode ser escrita como:

$$\hat{Y} = 8.9 + 0.8\overline{X}$$

Uma maneira de avaliar a bondade do modelo é através da diferença entre os valores amostrais reais (Y) e os valores estimados ($\hat{Y}$). A essa diferença damos o nome de resíduo.

Continuando o exemplo,

	Y	X	y	X	y^2	x^2	xy	$\hat{Y}$	Y-$\hat{Y}$	(Y-$\hat{Y}$)²
1970	26	23	-2.9	-2	8.41	4	5.8	27.3	-1.3	1.69
1971	25	21	-3.9	-4	15.21	16	15.6	25.7	-0.7	0.49
1972	31	28	2.1	3	4.41	9	6.3	31.3	-0.3	0.09
1973	29	27	0.1	2	0.01	4	0.2	30.5	-1.5	2.25
1974	27	23	-1.9	-2	3.61	4	3.8	27.3	-0.3	0.09
1975	31	28	2.1	3	4.41	9	6.3	31.3	-0.3	0.09
1976	32	27	3.1	2	9.61	4	6.2	30.5	1.5	2.25
1977	28	22	-0.9	-3	0.81	9	2.7	26.5	1.5	2.25
1978	30	26	1.1	1	1.21	1	1.1	29.7	0.3	0.09
1979	30	25	1.1	0	1.21	0	0	28.9	1.1	1.21
Soma	**289**	**250**	**0**	**0**	**48.9**	**60**	**48**	**289**	**0**	**11**
Média	**28.9**	**25**	**0**	**0**	**4.89**	**6**	**4.8**	**28.9**	**0**	**1**

Podemos perceber que as diferenças (Y-$\hat{Y}$) são relativamente pequenas. Uma análise mais cuidadosa pode ser feita através da aplicação de testes estatísticos, nesse caso ANOVA (teste de variância) e teste *t-Student*.

Tabela **ANOVA**

Soma dos Quadrados	Graus de Liberdade (g.l.)	Quadrados Médios (QM)	Teste F
SQE= $\hat{b}_1^{\,2}\sum x_i^2$	1	SQE/g.l.	SQEmed/SQRmed
SQR= $\sum\left(Y-\hat{Y}\right)^2$	n-2	SQR/g.l.	
SQT= $\sum y_i^2$	n-1	SQE/g.l + SQR/g.l.	

Obs: O grau de liberdade em relação ao SQE é devido a termos apenas uma variável independente. Em relação a SQT, os graus devem ser iguais à variância amostral, ou seja, n-1 (onde n é o número da elementos da amostra). E o grau de liberdade para SQR seria dado pela diferença entre este, ou seja, n-2.

Onde:

Soma dos quadrados dos totais de y centrado:

$$SQT = \sum y_i^2 \tag{7.6}$$

Soma dos quadrados explicados:

$$SQE = \sum \hat{Y}_i^{\,2} = \sum \hat{b}_1^{\,2} x_i^2 = \hat{b}_1^{\,2} \sum x_i^2 \tag{7.7}$$

Soma dos quadrados dos resíduos:

$$SQR = \sum\left(Y-\hat{Y}\right)^2 \tag{7.8}$$

Outro parâmetro utilizado constantemente é o coeficiente de determinação, R^2, que explica percentualmente a relação entre as variáveis do problema.

$$R^2 = \frac{SQE}{SQT} \tag{7.9}$$

Retornando ao exemplo,

Tabela **ANOVA**

Soma dos Quadrados	Graus de Liberdade (g.l.)	Quadrados Médios (QM)	Teste F
SQE=38.4 SQR=11.0	1 8	38.4 1.38	27.83
SQT=49.4	7	7.06	

Agora que já temos o valor de F precisamos testar a hipótese nula que as variâncias são diferentes, ou seja:

$Ho = m_{1 \neq} m_2$

Adotaremos um nível de significância (a) de 5%. Com esse valor e os números de graus de liberdade acha-se na tabela um valor crítico de 5.32.

Como o F calculado é maior que o F crítico então rejeita-se a hipótese Ho, o que também quer dizer que as variâncias são iguais e consequentemente o modelo de regressão é válido.

7.2 Regressão Linear Múltipla

Em algumas situações mais do que uma variável independente ($X_1, X_2, \ldots, X_n$) pode ser necessária para predizer o valor da variável independente (Y). O modelo matemático para esse caso é dado abaixo:

$$Y_i = b_0 + b_1 X_i + b_2 X_{2i} + \ldots + b_k X_{ki} + \varepsilon_i \qquad (7.10)$$

Que para as n observações poderá ser escrito da seguinte forma:

$$\begin{aligned}
Y_1 &= b_0 + b_1 X_1 + b_2 X_{21} + \ldots + b_k X_{k1} + \varepsilon_1 \\
Y_2 &= b_0 + b_1 X_2 + b_2 X_{22} + \ldots + b_k X_{k2} + \varepsilon_2 \\
\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\
Y_n &= b_0 + b_1 X_n + b_2 X_{2n} + \ldots + b_k X_{kn} + \varepsilon_n
\end{aligned}$$

Que forma um sistema linear que podemos escrever na forma de matriz como:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & X_1 & X_{21} & X_{k1} \\ 1 & X_2 & X_{22} & X_{k2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_n & X_{2n} & X_{kn} \end{bmatrix} . \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_k \end{bmatrix}$$

Escrevendo em outra forma mais compacta teremos:

$$Y = bX + \varepsilon \qquad (7.11)$$

O estimador para *b* será dado por:

$$\hat{b} = \left(X'X\right)^{-1}\left(X'Y\right) \qquad (7.12)$$

Pela equação acima há necessidade que o produto *X'X* tenha uma matriz inversa, o que implica na condição obrigatória que nenhuma coluna da matriz *X* seja combinação linear das outras.

Exemplo:

Uma agroindústria quer saber o custo de manutenção de seus caminhões durante o corrente ano. Para tanto foram coletadas informações de quilometragem e tempo do caminhão.

A tabela abaixo nos mostra esses valores.

Custo da manutenção	Quilometragem (1000Km)	Tempo da manutenção (Anos)
832	6	8
73	7	7
647	9	6
553	11	5
467	13	4
373	15	3
283	17	2
189	18	1
96	19	0

Nesse caso será feita diretamente a análise, sem plotar o gráfico. O procedimento no software Excel é: Ferramenta Þ Análise de Dados Þ Regressão. No campo Intervalo X de Entrada deve ser preenchida com a faixa de valores das variáveis independentes, que nesse caso são a quilometragem e o tempo do caminhão.

Da planilha de resultados destacam-se os seguintes valores:

Na estatística padrão: *R-quadadro* = 0.99
Erro padrão: 2.106
Na Anova: *gl* ⇨ *total* =*8 F*=56501.23
Interseção ⇨ *17.73 Variável X1*⇨ *4.06 e X2*⇨ 98.507

Assim, a equação do modelo poderá ser escrita como:

$$\hat{Y} = 17.73 + 4.06X_{1i} + 98.507X_2$$

Assim, para um caminhão com 5 anos de uso, com quilometragem de 10.000 milhas, o custo de manutenção será de $550.89.

Problemas Propostos:

1. Suponha que a análise de certo combustível apresentou para Y (poder calorífico) e para X (% de cinzas) os resultados:

(X)	13100	11200	10200	9600	8800
(Y)	18,3	27,5	36,4	48,5	57,8

Determinar a equação de regressão linear e estimar o poder calorífico para X= 30%. Construir o diagrama de dispersão e traçar a reta ajustada.

2. Um pesquisador realizou certa experiência relacionando pressão Y com temperatura X, obtendo os resultados:

(X) Em °C	30°	40°	50°	60°	70°	80°
(Y) Em atm	1,3	1,9	2,5	3,0	3,7	4,1

Pede-se:

a) A equação de regressão linear de mínimos quadrados;
b) A estimativa da pressão para a temperatura de 45°;
c) O diagrama de dispersão e a correspondente reta;
d) O coeficente de correlação linear.

3. A tabela seguinte relaciona o consumo (em toneladas) de matéria prima para uma indústria produzir dois tipos de produtos: A e B, sendo X1 e X2 as quantidades produzidas de A e B, respectivamente.

Consumo (Y)	3,5	4,0	5,4	6,1	7,0	7,5	8,0
X1	10	12	15	17	20	23	25
X2	8	9	11	13	15	16	18

Determinar a equação da regressão linear múltipla.

8
Controle Estatístico de Processo

O método preventivo de se comparar continuamente os resultados de um processo com os padrões/especificações identificados a partir de dados estatísticos é definido como Controle Estatístico de Processo. Este método engloba as avaliações das tendências para determinadas variações significativas no processo produtivo, a fim de eliminar/controlar essas variações e reduzi-las cada vez mais.

O controle de qualidade é constituído de um conjunto amplo de operações que envolvem todos os setores de uma empresa, visando a obtenção de produtos em níveis econômicos que satisfaçam seus consumidores.

O acompanhamento estatístico da qualidade pode ser feito de duas maneiras: por gráficos de controle ou por inspeção de amostragem. No primeiro caso temos um controle preventivo, já que o mesmo procura impedir a produção de itens defeituosos, enquanto que no segundo temos um controle curativo, ou seja, apenas separamos os produtos defeituosos dos perfeitos.

8.1 Gráficos de controle

A aplicação de técnicas estatísticas tem por principal objetivo oferecer aos responsáveis pela tomadas de decisões referências relativas ao grau de confiabilidade dos resultados gerados pelos controles e aos riscos envolvidos nas decisões tomadas. A

sistematização dos dados de controle que normalmente é feita sob a forma de "gráficos de controle" tem por objetivo facilitar a "visualização" dos resultados.

São três os principais tipos de gráficos usados em controle da qualidade, a saber:

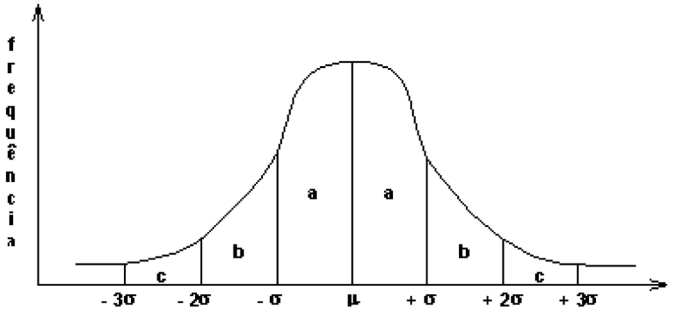
- Gráficos de controle por média;
- Gráficos de controle por amplitude;
- Gráficos de controle para frações defeituosas.

Os controles por média e amplitude são feitos com base na teoria estatística da distribuição normal. Já o controle de frações defeituosas é mais frequentemente fundamentado na distribuição de Poisson. Para alguns casos de controle de frações defeituosas a aplicação de teoria estatística da distribuição binomial pode ser vantajosa.

Quando os valores de uma determinada variável estão distribuídos normalmente (simetricamente) em torno da média ou quando estes obedecem a uma curva de distribuição, como a da Figura 8.1, é representada pela equação.

$$y = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{1}{2}\right)x^2} \qquad (8.1)$$

Figura 8.1 - Curva de distribuição normal



Para as variáveis que se distribuem segundo uma distribuição normal podemos dizer que:

a) 68% dos valores encontrados caem no intervalo $\mu \pm \sigma$ (região a);

b) 95% dos valores encontrados caem no intervalo $\mu \pm 2\sigma$ (regiões a e b);

c) 99,7% dos valores encontrados caem no intervalo $\mu \pm 3\sigma$ (regiões a, b e c).

Onde:

μ é a média da população;

σ é o desvio-padrão da população, ou sua melhor estimativa, quando se trabalha com uma amostra da população. Nesse caso, usa-se S como símbolo do desvio-padrão ao invés de "δ"

Sendo:

$$\mu = \frac{\sum x}{n} \tag{8.2}$$

e

$$\delta = \sqrt{\frac{\sum_{i=n}^{i=1} (x_i - \mu)^2}{n}} \tag{8.2}$$

Para quando se está trabalhando com a população, ou:

$$s = \sqrt{\frac{\sum_{i=n}^{i=1} (x_i - \bar{x})^2}{n-1}} \tag{8.3}$$

Para quando se está trabalhando com uma amostra da população. Neste caso, S é apenas a melhor estimativa do desvio-padrão da população, onde:

x são valores individuais;
x é a média dos valores individuais de uma amostra;
n é o número de itens que compõem a amostra.

Pelo exposto, pode-se afirmar que estatisticamente espera-se que para cada:

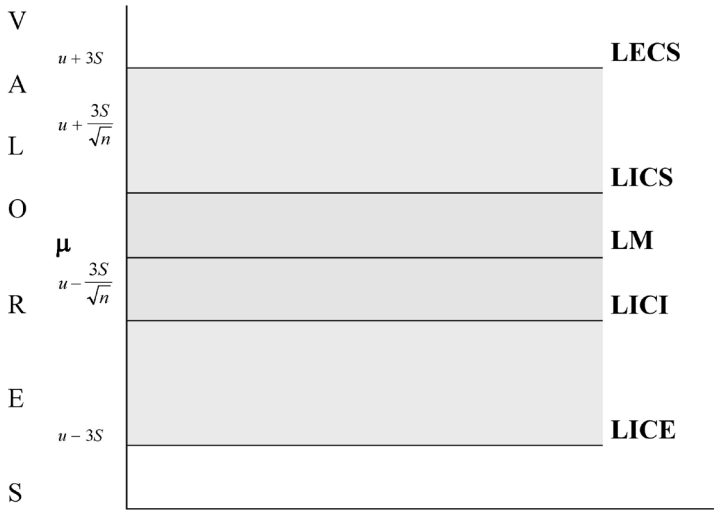
a) 100 amostras analisadas, 32 devem apresentar resultados fora dos limites $\mu \pm \delta$;
b) 100 amostras analisadas, 5 devem apresentar resultados fora dos limites $\mu \pm 2\delta$;
c) 1.000 amostras analisadas, apenas 3 devem apresentar resultados fora dos limites $\mu \pm 3\delta$.

8.2 Gráficos de controle por média

Os gráficos de controle por média são os mais usados. Eles são construídos com base na teoria da distribuição normal, apresentada no Capítulo 2. Conhecidos os valores μ e S, a sua construção é simples e pode assumir duas configurações:

A primeira (Figura 8.2) adota como linhas dos limites superior e inferior do controle interno as posições $\mu + 2s$ e $\mu - 2s$, respectivamente (Sistema inglês).

Figura 8.2 – Gráfico de controle por média (Sistema Inglês)



Onde temos:

LSCE	-	Limite superior de controle externo;
LSCI	-	Limite superior de controle interno;
LM	-	Linha da média da população ou da amostra;
LICI	-	Limite inferior de controle interno;
LICE	-	Limite inferior de controle externo.

Como as linhas LSCI e LICI situam-se a +2s e -2s da média, respectivamente, espera-se que apenas 5 em cada 100 amostras (ou 1 em cada 20) venham a se posicionar fora das mesmas. Se isto ocorrer diz-se que o processo está sob controle.

Se mais de 1 amostra em cada 20 avaliadas apresentar resultados fora dos limites estabelecidos pelas linhas LSCI e LICI diz- se que o processo está fora de controle.

Do mesmo modo como as linhas LSCE e LICE situam-se a + 3s e - 3s da média, respectivamente, espera-se que apenas 3 em cada 1000 (ou 1 em cada 333) amostras analisadas estejam fora desses limites.

A vantagem de se trabalhar com duas linhas de controle (controle interno e controle externo) reside no fato de que quando mais de 1 amostra em 20 analisadas, no caso da Figura 8.2, apresentar resultados fora das linhas LSCI e LICI já se pode tomar decisões relativas ao seu ajuste.

Como estatisticamente espera-se que, neste caso, mais de 3 amostras em 1000 (ou 1 em 333) venham a apresentar resultados fora dos limites estabelecidos pelas linhas LSCE e LICE não será necessário esperar pelas próximas 313 avaliações para fazer os devidos ajustes no processo, evitando assim que o mesmo seja conduzido em condições fora de controle

A outra maneira de se construir um gráfico de controle por média é posicionando as linhas de controle interno a $\mu \pm \frac{3S}{\sqrt{n}}$ (Figura 8.3; Sistema americano). Como é o desvio-padrão das médias e os controles internos foram definidos como:

$$\mathbf{LSCI} = \mu + \frac{3S}{\sqrt{n}} \quad \mathbf{e}$$

$$\mathbf{LICI} = \mu - \frac{3S}{\sqrt{n}}$$

Pode- se concluir que 99,7% das médias das amostras analisadas deverão situar-se na região compreendida entre LSCI e LICI, enquanto que a região compreendida entre os limites LSCE e LICE constitui o intervalo onde 99,7% dos resultados individuais estarão localizados sempre que o processo estiver sob controle.

Assim, espera-se que no máximo 1 amostra em 333 apresente médias fora dos limites estabelecidos pelas linhas LSCI e LICI e no máximo 1 amostra em 333 apresente valores individuais fora dos limites estabelecidos pelas linhas LSCE e LICE.

O gráfico da Figura 8.2 só se aplica quando o controle é feito mediante análise de amostras com mais de uma unidade amostral, o que torna o processo de controle mais oneroso. No Sistema Americano as linhas de controle interno (LICI e LSCI) definem os limites do lugar geométrico das médias, enquanto que as linhas de

controles externos (LSCE e LICE) estabelecem os limites do lugar geométrico dos valores individuais.

Portanto, no Sistema Americano deve ser lançado tanto os valores encontrados para a média como os valores individuais, o que pode gerar confusões na sua interpretação. Assim, para o acompanhamento de processos, o gráfico da Figura 8.2 não só é mais prático como também é mais barato, uma vez que as amostras são constituídas de uma única unidade amostral.

Quando o produto que está sendo controlado deve obedecer a normas metrológicas ou é especificado com limites de tolerância definidos, o gráfico de controle da qualidade deve ser elaborado a partir do conhecimento destes limites.

Exemplo:

Para atender a Portaria INMETRO 74 de 25 de maio de 1995, xampu veterinário envasado em frascos de 1000 ml deve atender ao limite de tolerância para média dado pela equação u ≥ Qn – kS, onde:

u é a média da amostra;
Qn é o valor nominal (1000 ml, no caso);
S é o desvio-padrão da amostra; e
k é um fator que depende do tamanho da amostra (para amostras com 20 itens k = 0,64).

Este é um caso típico de produtos especificados apenas pelo limite de tolerância inferior. Neste caso, a linha LICE passará a ser definida pelo limite de tolerância inferior (LTI):

$$LICE = LTI$$

E a linha LICI deverá se situar a:

$$LICI = LTI + (3S - \frac{3S}{\sqrt{n}})$$

A linha LSCI deverá se situar a:

$$LSCI = LTI + (3S + \frac{3S}{\sqrt{n}})$$

Isto implica em que a máquina de envasar deve ser ajustada para produzir itens com peso médio igual a:

μ = LTI + 3 S

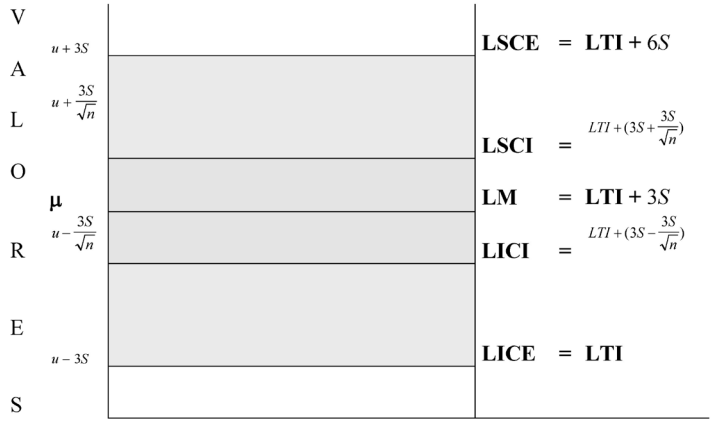
Nesse caso a linha de controle superior externo, LSCE, será dada por:

LSCE = LTI + 2 (3 S)

O gráfico da Figura 8.3 se aplica para controle por média quando mais de um item é analisado e, neste caso, a região compreendida entre as linhas LSCI e LICI definem o intervalo onde se espera que 99,7% das amostras venham situar-se as suas médias.

Como nos exemplos anteriores, as linhas LSCE e LICE limitam a região onde se espera que 99,7% dos itens individuais das amostras venham se posicionar. Caso isso não ocorra o processo estará fora de controle.

Figura 8.3 – Gráfico de controle para produtos com limite de tolerância inferior especificado



8.3 Diagramas de Ishikawa e análise de causa raiz

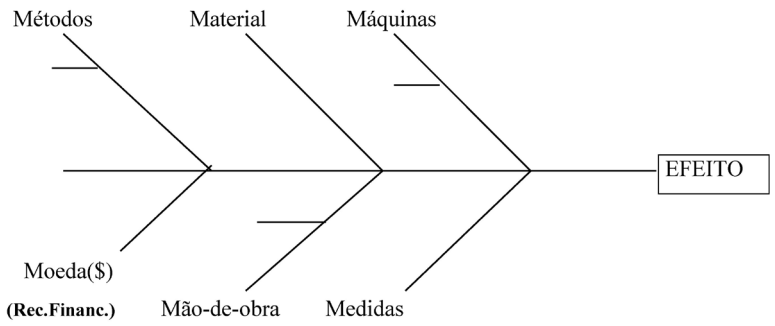
Outra ferramenta de qualidade que visa a identificação de defeitos e a avaliação de seu impacto no processo global, com vistas a sua otimização, seria o diagrama de causa-efeito de Ishikawa.

Este diagrama, originalmente proposto por Kaoru Ishikawa na década de 60, já foi bastante utilizado em ambientes industriais para a localização de causa de dispersão de qualidade no produto e no processo de produção.

Ele é uma ferramenta gráfica utilizada para explorar e representar opiniões a respeito de fontes de variações em qualidade de processo, que em termos teóricos faz parte da atividade de validação de processos e obviamente no programa de Análise de Perigos e Pontos Críticos de Controle (APPCC). Especificamente ele é utilizado na etapa referente à identificação de perigos e pontos críticos de controle, ou na linguagem farmacêutica, de pontos críticos.

Sua maior função seria focada para a identificação de direcionadores, ou drivers, que potencialmente levam ao efeito indesejável. Ele é uma ferramenta analítica que, utilizada por um grupo de projeto, parte de um "problema de interesse" e possibilita a execução de um brainstorm no sentido de identificar as causas possíveis para o problema. De uma forma geral, pode-se exemplificar sua aplicação como segue abaixo:

Figura 8.4 - Diagrama de causa-e-efeito de Ishikawa



As principais categorias de causas, denominadas de "primárias", são divididas em seis categorias, porém, dependendo da situação, podem ser elencados outros componentes. As causas primárias podem ainda ser subdivididas em causas secundárias e assim por diante.

Roteiro para o método:

1. Identificar o EFEITO: deve ser identificado com clareza o efeito do problema ou não conformidade a ser corrigida. O EFEITO pode também ser uma meta / objetivo a ser atingido ao invés de um problema específico;
2. Geração dos dados: através de uma seção de *brainstorming* serão coletadas informações que corresponderão às causas secundárias;
3. Identificar as causas secundárias: cada causa pode gerar inúmeras subcausas. Quanto maior o volume de informações advindas das pessoas que têm relação direta e indireta com o problema / efeito / objetivo, mais fáceis podem ser as alternativas de equacionamento da questão;
4. Análise: preenchido o diagrama fica fácil a análise de todas as causas e a identificação daquelas que efetivamente estão produzindo o efeito. Isto posto devem ser desenvolvidas ações imediatas e diretas junto às causas selecionadas, de forma a resolver a questão.

Mesmo com sua grande aplicabilidade o diagrama de Ishikawa conduz à identificação de causas de desvios, sem estabelecer exatamente quais as raízes do problema. O diagrama apresenta como principais vantagens os seguintes aspectos:

- É uma boa ferramenta de levantamento de tendências;
- É uma boa ferramenta de comunicação;
- Estabelece a relação entre o efeito e suas causas;
- Possibilita um detalhamento das causas.

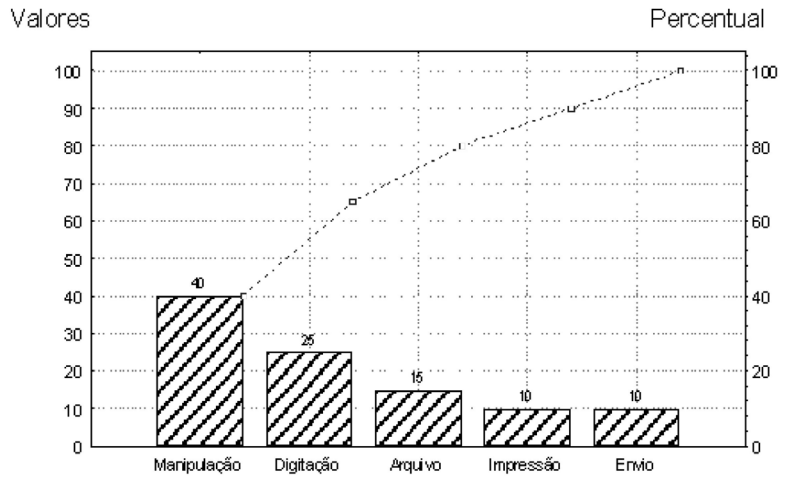
8.4 Diagrama de Pareto

Com a construção do gráfico de Pareto pode-se evidenciar as causas de maior impacto, onde deverão, a princípio, despender mais esforços. Uma aplicação é seu uso para a visualização das barreiras priorizadas pela aplicação de uma matriz de decisão.

Etapas de Construção:

1. Compare as medições de cada categoria, estabelecendo uma lista em ordem decrescente de importância;
2. Os itens de menor importância podem ser agrupados na categoria outros. Como regra pode-se utilizar este recurso para os itens com valor inferior a 1% do total;
3. Trace o eixo horizontal e estabeleça uma largura para cada barra de forma que a soma das larguras de todas as barras possa ficar contida no espaço deste eixo;
4. Trace o eixo vertical. Verifique na etapa de construção 4 qual o maior valor (topo da lista) e faça a escala do eixo de forma a poder comportar todos os valores.
5. Desenhe as barras seguindo a ordem decrescente da lista da etapa 4. A categoria outros, mesmo que não seja a menor de todas, é colocada como última barra (no extremo direito). O limite superior da barra representa seu valor. O valor numérico correspondente deve ser escrito acima da barra.

Figura 7.7 – Diagrama de Pareto



BIBLIOGRAFIA

BARROS NETO, B.; SCARMÍNIO, I. S.; BRUNS, R. E. **Planejamento e Otimização de Experimentos**. Campinas, SP: Editora da Unicamp, 1995.

BEVERIDGE, G. S. G.; SCHECHTER, R. S. **Optimization**: Theory and Practice. New York: McGraw-Hill, 1970.

BOX, G. P. G.; HUNTER, W. G.; HUNTER, J. S. **Statistics for Experimenter**: An Introduction To Design, Data Analysis And Model Building. New York: John Wiley & Sons, 1978.

BRERETON, R. G. Chemometrics in Analytical Chemistry - **A Review**, Analyst, 1987, 112, p. 1635 - 1657.

BRITO, E. S.; PINTO, Gustavo A. S.; BRUNO, Laura M.; AZEREDO, Henriette M. C. de **A metodologia de superfície de resposta (MSR) na otimização de processos biológicos**: A determinação de valores de pH e temperatura ótimos para a atividade enzimática, Acesso: em maio de 2008.

BULISANI, A. E. **Feijão**: fatores de produção e qualidade. Campinas: Fundação Cargill, 1987. 326 p.

BURTON, K. W. C.; NICKLESS, G. **Optimization via Simplex**. Part I. Background, definitions and a simple application. Chemometrics Intel. Lab. Syst., 1987. 1, p. 135- 149.

CHEMKEYS. http://www.chemkeys.com/bra/mdpf_3/mdpf_3.htm. Acesso em: abril de 2008.

COSTA NETO, P.L.O. **Estatística**. São Paulo: Edgard Blucher, 1977. 264p.

DEMING, S. N.; MORGAN, S. L., **Simplex Optimization of Variables in Analytical Chemistry**, Anal. Chem., 1973, 45, p. 278 A-283 A.

EIRAS, S. P.; CUELBAS, C. J.; DE ANDRADE, J. C.. **Um Estudo Comparativo sobre a Eficiência de Estratégias Quimiométricas de Otimização**. Química Nova, 1994, 16, p. 216 - 219.

FERRARI, C. K. B. Oxidação lipídica em alimentos e sistemas biológicos: mecanismos gerais e implicações nutricionais e patológicas. **Rev. Nutr**., n. 11, v. 1, p. 3-14, 1998.

FISHER, R. A. **The Design of Experiments**. Oliver & Boyd, Edinburg, 1935.

LEGRET, M.; DIVET, L. **Application of Factorial** - designs in Optimization of the Determination of Lead by Electrotermal atomization. Analusis, 1988, 16, p. 97 - 106.

MARQUES, J. M; Marques, M. A. M. **Estatística básica para os cursos de engenharia**. Domínio do saber, 2005. 272p.

MONTGOMERY, D. G. **Design and analysis of experiments**. New York: Johr Wiley & Sons, 2001.

NELDER, J. A.; MEAD, D. R. **A simplex method for function minimization**. Computer J. 1965, 7, p. 308-12.

RODRIGUES, M. I. IEMA, A. F. **Planejamento de experimentos e otimização de processo**: uma estratégia seqüencial de planejamentos. Campinas, SP: Casa do Pão Editora, 2005.

ROUTH, M. W.; SWARTZ, P. A e DENTON, M. B. **Performance of the super modified simplex**. Anal. Chem. 1977. 49, p. 1422-1428.

SPLENDLEY, W.; HEXT, G. R.; HIMSWORTH, F. R. **Sequential application of simplex designs in optimization and evolutionary operation**. Technometrics, 1962, 4, p. 441-461.

VOET, J.; VOET, D. **Biochemistry.** Johr Wiley & Sons, New York, 2001.

Apêndices

Apêndice – Tabela da Distribuição Normal Padrão *P(Z<z)*

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

P(Z<z)

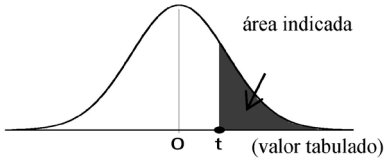
z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
-3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,7	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,8	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

TABELA DA DISTRIBUIÇÃO QUI-QUADRADO

n \ α	0,995	0,990	0,975	0,950	0,900	0,750	0,500	0,250	0,100	0,050	0,025	0,010	0,005
1	0,0000	0,0002	0,0010	0,0039	0,0158	0,1015	0,4549	1,3233	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201	0,0506	0,1026	0,2107	0,5754	1,3863	2,7726	4,6052	5,9915	7,3778	9,2103	10,5966
3	0,0717	0,1148	0,2158	0,3518	0,5844	1,2125	2,3660	4,1083	6,2514	7,8147	9,3484	11,3449	12,8382
4	0,2070	0,2971	0,4844	0,7107	1,0636	1,9226	3,3567	5,3853	7,7794	9,4877	11,1433	13,2767	14,8603
5	0,4117	0,5543	0,8312	1,1455	1,6103	2,6746	4,3515	6,6257	9,2364	11,0705	12,8325	15,0863	16,7496
6	0,6757	0,8721	1,2373	1,6354	2,2041	3,4546	5,3481	7,8408	10,6446	12,5916	14,4494	16,8119	18,5476
7	0,9893	1,2390	1,6899	2,1673	2,8331	4,2549	6,3458	9,0371	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,3444	1,6465	2,1797	2,7326	3,4895	5,0706	7,3441	10,2189	13,3616	15,5073	17,5345	20,0902	21,9550
9	1,7349	2,0879	2,7004	3,3251	4,1682	5,8988	8,3428	11,3888	14,6837	16,9190	19,0228	21,6660	23,5894
10	2,1559	2,5582	3,2470	3,9403	4,8652	6,7372	9,3418	12,5489	15,9872	18,3070	20,4832	23,2093	25,1882
11	2,6032	3,0535	3,8157	4,5748	5,5778	7,5841	10,3410	13,7007	17,2750	19,6751	21,9200	24,7250	26,7568
12	3,0738	3,5706	4,4038	5,2260	6,3038	8,4384	11,3403	14,8454	18,5493	21,0261	23,3367	26,2170	28,2995
13	3,5650	4,1069	5,0088	5,8919	7,0415	9,2991	12,3398	15,9839	19,8119	22,3620	24,7356	27,6882	29,8195
14	4,0747	4,6604	5,6287	6,5706	7,7895	10,1653	13,3393	17,1169	21,0641	23,6848	26,1189	29,1412	31,3193
15	4,6009	5,2293	6,2621	7,2609	8,5468	11,0365	14,3389	18,2451	22,3071	24,9958	27,4884	30,5779	32,8013
16	5,1422	5,8122	6,9077	7,9616	9,3122	11,9122	15,3385	19,3689	23,5418	26,2962	28,8454	31,9999	34,2672
17	5,6972	6,4078	7,5642	8,6718	10,0852	12,7919	16,3382	20,4887	24,7690	27,5871	30,1910	33,4087	35,7185
18	6,2648	7,0149	8,2307	9,3905	10,8649	13,6753	17,3379	21,6049	25,9894	28,8693	31,5264	34,8053	37,1565
19	6,8440	7,6327	8,9065	10,1170	11,6509	14,5620	18,3377	22,7178	27,2036	30,1435	32,8523	36,1909	38,5823
20	7,4338	8,2604	9,5908	10,8508	12,4426	15,4518	19,3374	23,8277	28,4120	31,4104	34,1696	37,5662	39,9968
21	8,0337	8,8972	10,2829	11,5913	13,2396	16,3444	20,3372	24,9348	29,6151	32,6706	35,4789	38,9322	41,4011
22	8,6427	9,5425	10,9823	12,3380	14,0415	17,2396	21,3370	26,0393	30,8133	33,9244	36,7807	40,2894	42,7957
23	9,2604	10,1957	11,6886	13,0905	14,8480	18,1373	22,3369	27,1413	32,0069	35,1725	38,0756	41,6384	44,1813
24	9,8862	10,8564	12,4012	13,8484	15,6587	19,0373	23,3367	28,2412	33,1962	36,4150	39,3641	42,9798	45,5585
25	10,5197	11,5240	13,1197	14,6114	16,4734	19,9393	24,3366	29,3389	34,3816	37,6525	40,6465	44,3141	46,9279
26	11,1602	12,1981	13,8439	15,3792	17,2919	20,8434	25,3365	30,4346	35,5632	38,8851	41,9232	45,6417	48,2899
27	11,8076	12,8785	14,5734	16,1514	18,1139	21,7494	26,3363	31,5284	36,7412	40,1133	43,1945	46,9629	49,6449
28	12,4613	13,5647	15,3079	16,9279	18,9392	22,6572	27,3362	32,6205	37,9159	41,3371	44,4608	48,2782	50,9934
29	13,1211	14,2565	16,0471	17,7084	19,7677	23,5666	28,3361	33,7109	39,0875	42,5570	45,7223	49,5879	52,3356
30	13,7867	14,9535	16,7908	18,4927	20,5992	24,4776	29,3360	34,7997	40,2560	43,7730	46,9792	50,8922	53,6720
31	14,4578	15,6555	17,5387	19,2806	21,4336	25,3901	30,3359	35,8871	41,4217	44,9853	48,2319	52,1914	55,0027
32	15,1340	16,3622	18,2908	20,0719	22,2706	26,3041	31,3359	36,9730	42,5847	46,1943	49,4804	53,4858	56,3281
33	15,8153	17,0735	19,0467	20,8665	23,1102	27,2194	32,3358	38,0575	43,7452	47,3999	50,7251	54,7755	57,6484
34	16,5013	17,7891	19,8063	21,6643	23,9523	28,1361	33,3357	39,1408	44,9032	48,6024	51,9660	56,0609	58,9639
35	17,1918	18,5089	20,5694	22,4650	24,7967	29,0540	34,3356	40,2228	46,0588	49,8018	53,2033	57,3421	60,2748
36	17,8867	19,2327	21,3359	23,2686	25,6433	29,9730	35,3356	41,3036	47,2122	50,9985	54,4373	58,6192	61,5812
37	18,5858	19,9602	22,1056	24,0749	26,4921	30,8933	36,3355	42,3833	48,3634	52,1923	55,6680	59,8925	62,8833
38	19,2889	20,6914	22,8785	24,8839	27,3430	31,8146	37,3355	43,4619	49,5126	53,3835	56,8955	61,1621	64,1814
39	19,9959	21,4262	23,6543	25,6954	28,1958	32,7369	38,3354	44,5395	50,6598	54,5722	58,1201	62,4281	65,4756
40	20,7065	22,1643	24,4330	26,5093	29,0505	33,6603	39,3353	45,6160	51,8051	55,7585	59,3417	63,6907	66,7660
41	21,4208	22,9056	25,2145	27,3256	29,9071	34,5846	40,3353	46,6916	52,9485	56,9424	60,5606	64,9501	68,0527
42	22,1385	23,6501	25,9987	28,1440	30,7654	35,5099	41,3352	47,7663	54,0902	58,1240	61,7768	66,2062	69,3360
43	22,8595	24,3976	26,7854	28,9647	31,6255	36,4361	42,3352	48,8400	55,2302	59,3035	62,9904	67,4593	70,6159
44	23,5837	25,1480	27,5746	29,7875	32,4871	37,3631	43,3352	49,9129	56,3685	60,4809	64,2015	68,7095	71,8926
45	24,3110	25,9013	28,3662	30,6123	33,3504	38,2910	44,3351	50,9849	57,5053	61,6562	65,4102	69,9568	73,1661
46	25,0413	26,6572	29,1601	31,4390	34,2152	39,2197	45,3351	52,0562	58,6405	62,8296	66,6165	71,2014	74,4365
47	25,7746	27,4158	29,9562	32,2676	35,0814	40,1492	46,3350	53,1267	59,7743	64,0011	67,8206	72,4433	75,7041
48	26,5106	28,1770	30,7545	33,0981	35,9491	41,0794	47,3350	54,1964	60,9066	65,1708	69,0226	73,6826	76,9688
49	27,2493	28,9406	31,5549	33,9303	36,8182	42,0104	48,3350	55,2653	62,0375	66,3386	70,2224	74,9195	78,2307
50	27,9907	29,7067	32,3574	34,7643	37,6886	42,9421	49,3349	56,3336	63,1671	67,5048	71,4202	76,1539	79,4900

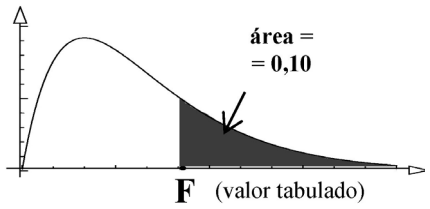
Distribuição *t* de *Student*



	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

	Área na cauda superior								
gl	**0,25**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**	**0,0025**	**0,001**	**0,0005**
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

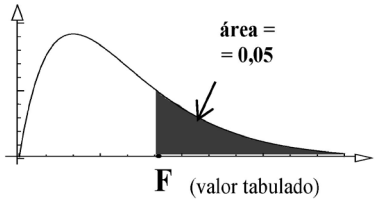
Distribuição
F de Snedecor
α = 0,10



gl	**graus de liberdade no numerador**									
denom.	**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

gl	graus de liberdade no numerador									
denom.	1	2	3	4	5	6	7	8	9	10
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
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

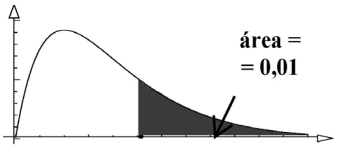
Distribuição
F de Snedecor
α = 0,05



gl	graus de liberdade no numerador									
denom.	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

gl	graus de liberdade no numerador									
denom.	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

Distribuição
F de Snedecor
α = 0,01



F (valor tabulado)

gl	graus de liberdade no numerador									
denom.	1	2	3	4	5	6	7	8	9	10
1	4052,2	4999,3	5403,5	5624,3	5764,0	5859,0	5928,3	5981,0	6022,4	6055,9
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85

gl	graus de liberdade no numerador									
denom.	1	2	3	4	5	6	7	8	9	10
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
35	7,42	5,27	4,40	3,91	3,59	3,37	3,20	3,07	2,96	2,88
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
45	7,23	5,11	4,25	3,77	3,45	3,23	3,07	2,94	2,83	2,74
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50