
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

CE003
ESTATÍSTICA II
(Notas de Aula)

DEPARTAMENTO DE ESTATÍSTICA
UFPR

Curitiba, 27 de fevereiro de 2009

Sumário

1	Conceitos Básicos e Técnicas de Estatística Descritiva	1
1.1	Introdução	1
1.1.1	Estatística descritiva x estatística inferencial	3
1.1.2	População e amostra	5
1.1.3	Variáveis e suas classificações	8
1.2	Técnicas de estatística descritiva	9
1.2.1	Tabelas de frequências	10
1.2.2	Medidas-resumo	13
1.2.3	Gráficos	21
2	Teoria das Probabilidades	33
2.1	Introdução	33
2.2	Conceitos Básicos de Probabilidade	33
2.2.1	Definição clássica de probabilidade	34
2.2.2	Aproximação da Probabilidade pela frequência relativa	35
2.2.3	Propriedades de probabilidades	35
2.2.4	Teorema da soma	35
2.2.5	Probabilidade condicional	36
2.2.6	Teorema do produto	38
2.2.7	Teorema da probabilidade total	38
2.2.8	Teorema de Bayes	39
3	Variáveis Aleatórias	40
3.1	Introdução	40
3.2	Variável Aleatória Discreta	40
3.3	Variável Aleatória Contínua	41
3.4	Esperança Matemática	41
3.5	Variância	42
3.6	Principais Distribuições de Probabilidades	42
3.6.1	Distribuição de Bernoulli	42
3.6.2	Distribuição Binomial	42
3.6.3	Distribuição de Poisson	44

3.6.4	Distribuição Normal	44
3.6.5	Distribuição Normal Padrão	45
3.6.6	Uso da tabela da Normal Padrão	45
4	Inferência Estatística - Teoria da Estimação	48
4.1	Introdução	48
4.2	Propriedades dos Estimadores	51
4.3	Distribuições Amostrais	53
4.3.1	Introdução	53
4.3.2	Distribuição amostral de \bar{X}	53
4.3.3	Teorema central do limite (TCL)	55
4.4	Estimação da Média Populacional (μ)	57
4.5	Estimação de μ em Amostras Pequenas	60
4.6	Estimação da Diferença entre Duas Médias Populacionais (μ_1 e μ_2)	61
4.7	Estimação de $\mu_1 - \mu_2$ em Amostras Pequenas	62
4.8	Estimação de uma Proporção Populacional (p)	64
4.8.1	TCL para proporção amostral	64
4.8.2	Intervalo de Confiança para p	65
4.9	Determinação do Tamanho da Amostra (n)	66
5	Testes de Hipóteses	70
5.1	Introdução	70
5.2	Conceitos Estatísticos dos Testes de Hipóteses	71
5.2.1	Hipóteses estatísticas paramétricas	71
5.2.2	Testes	72
5.2.3	Tipos de erros cometidos ao se tomar uma decisão	72
5.2.4	Região crítica (RC) e regra de decisão (RD)	73
5.2.5	Procedimentos para realização de um teste de significância	73
5.3	Exemplos	74
5.4	Alguns Testes Paramétricos mais Utilizados.	85
5.4.1	Teste para a média (μ) com σ^2 desconhecida.	85
5.4.2	Teste para a comparação de duas médias populacionais (μ_1 e μ_2)	89
5.4.3	Teste para amostras independentes com $\sigma_1^2 = \sigma_2^2$	90
5.4.4	Teste para amostras independentes com $\sigma_1^2 \neq \sigma_2^2$	92
5.4.5	Teste para amostras dependentes	93
5.5	Teste para Proporção Populacional (p)	94
5.6	Teste para a Comparação de duas Proporções Populacionais (p_1 e p_2).	96
5.7	Testes não Paramétricos	97
5.7.1	Teste de aderência	97
5.7.2	Teste qui-quadrado para tabelas de contingência	99

6	Correlação e Regressão Linear	103
6.1	Introdução	103
6.2	Coeficiente de Correlação de Pearson	103
6.2.1	Teste de significância para ρ	105
6.3	Regressão Linear Simples	105
6.3.1	Estimação dos parâmetros por MQO	107
6.3.2	Adequação do modelo de regressão linear ajustado	108
6.3.3	Interpretação dos parâmetros do modelo	110
7	Análise de Variância	112
7.1	Introdução	112
7.2	Conceitos Básicos sobre Experimentação	112
7.2.1	Tratamento	112
7.2.2	Unidade experimental ou parcela	113
7.2.3	Repetição	113
7.2.4	Variável resposta ou variável dependente	114
7.2.5	Delineamento experimental (Design)	114
7.2.6	Modelo e análise de variância	115
7.2.7	Delineamento experimental	116
7.3	Análise de Variância	116
7.4	Teste de Tukey para Comparação de Médias	118
7.5	Teste de Kruskal-Wallis	120
8	Controle Estatístico de Qualidade	122
8.1	Introdução	122
8.1.1	Gráficos de controle	122
8.1.2	Construção do gráfico	123
8.1.3	Análise do padrão de gráficos de controle	124
8.2	Gráficos de Controle para Variáveis	127
8.2.1	Gráficos de controle para \bar{x} e R	128
8.2.2	Gráficos de controle para \bar{x} e s	129
8.2.3	Exemplos	131
	Tabelas	136

Lista de Tabelas

1.1	Resumo de técnicas de estatística descritiva	3
1.2	Resumo de técnicas de estatística inferencial	4
1.3	Frequências de estado civil em uma amostra de 385 indivíduos.	10
1.4	Tabela de frequências para a variável Idade.	11
1.5	Tabela de frequências para a variável horas semanais de atividade física . .	12
1.6	Tabela de frequências para a variável Peso	12
1.7	Tipos sanguíneos de 1000 pacientes.	16
1.8	Medidas de tendência central para as notas das turmas A e B.	17
1.9	Principais medidas de dispersão.	18
1.10	Peso de 10 nascidos vivos	19
1.11	Intenção de votos para os partidos A,B,C e D.	22
1.12	Número de crianças por família.	23
1.13	Nível de estresse em 70 funcionários de uma empresa.	25
1.14	Resumo de 5 números para o número de laranjas por caixas.	27
1.15	Alturas de crianças do sexo masculino (m) e feminino (f).	29
2.1	Gosto pela disciplina de estatística segundo sexo.	37
4.1	População de alunos.	49
4.2	Todas as possíveis amostras aleatórias simples com reposição de tamanho 2, da população de alunos.	50
4.3	Distribuição amostral da idade média.	53
5.1	Erros cometidos na tomada de decisão.	72
5.2	Algumas ocorrências, implicações e decisões após a retirada da amostra. . .	74
5.3	Distribuição de probabilidades das possíveis amostras.	75
5.4	Resumo das decisões para o Exemplo 5.1.	76
5.5	Resumo das decisões para o novo experimento.	77
5.6	Algumas tomadas de decisão e regras de decisão conforme a hipótese nula, o nível de significância e a distribuição de probabilidade.	81
5.7	Valores de $1 - \beta(\mu^*)$ para o exemplo 5.2 de acordo com os parâmetros α , σ , n e μ^*	85

5.8	Resistência (kgf) de dois tipos de concreto.	92
5.9	Pressão antes e após seis meses da administração do medicamento.	94
5.10	Tabela auxiliar.	97
5.11	Número de acidentes por dia da semana.	98
5.12	Quadro auxiliar com as frequências esperadas.	99
5.13	Renda e número de filhos por família em uma cidade.	100
5.14	Representação de duas características (A e B).	100
5.15	Número esperado para número de filhos e renda.	102
6.1	Tempo de reação ao estímulo em função da idade.	106
7.1	Tabela da análise de variância.	116
7.2	Crescimento de explantes de morangos em gramas.	117
7.3	Análise de variância do exemplo 7.1.	118
7.4	Consumo de energia elétrica de três motores durante uma hora.	121
8.1	Dados de espessura (mm) de uma peça de metal.	132
8.2	Dados de espessura (mm) de uma peça de metal avaliados após intervenção no processo.	134
1	Distribuição Normal: $P(0 \leq Z \leq z_c)$	137
2	Distribuição Normal padrão com valores de $P[-\infty \leq Z \leq Z_c]$	138
3	Distribuição Normal padrão com valores de $P[-\infty \leq Z \leq Z_c]$ (continuação).	139
4	Limites unilaterais de F ao nível de 5% de probabilidade n1=número de graus de liberdade do numerador, n2= número de graus de liberdade do denominador	140
5	Limites unilaterais de F ao nível de 1% de probabilidade n1=número de graus de liberdade do numerador, n2= número de graus de liberdade do denominador	141
6	Valores de t em níveis de 10% a 0,1% de probabilidade.	142
7	Valores da amplitude total estudentizada (q), para uso no teste de Tukey, ao nível de 5% de probabilidade. I=número de tratamentos, GLRES= número de graus de liberdade do resíduo.	143
8	Distribuição de Qui-quadrado. Valor crítico de χ^2 tal que $P(\chi_k^2 > \chi_0^2) = \alpha$	144
9	Constantes utilizadas em gráficos de controle.	145
10	Tabela de números aleatórios.	146

Lista de Figuras

1.1	Gráfico de setores para a intenção de votos nos partidos A,B,C e D.	22
1.2	Gráfico de barras para o número de filhos por família.	24
1.3	Histograma para o nível de estresse.	25
1.4	Desenho esquemático do box-plot com base no resumo de 5 números.	26
1.5	Box-plot do número de laranjas nas 20 caixas.	27
1.6	Desenho esquemático do box-plot com base nos quartis e critério para valores atípicos.	28
1.7	Altura de crianças conforme o sexo.	30
1.8	Variação mensal na Taxa Selic no período de 1995 a 2005.	31
1.9	Gráfico sequencial das vendas ao longo dos meses.	32
3.1	Densidade Normal.	45
4.1	Analogia entre as propriedades dos estimadores e o jogo de dardos.	52
4.2	Distribuição de \bar{X} quando X tem distribuição normal, para alguns tamanhos de amostra.	54
4.3	Densidades de T e Z	55
4.4	Densidade de Z e o quantil z	58
4.5	Máximo de $p(1 - p)$	66
5.1	Área hachurada relativa ao P-Valor do teste	80
5.2	Probabilidade de não rejeitar H_0 quando ela é falsa.	84
5.3	Região crítica associada à estatística t	87
5.4	Região crítica associada à estimativa da média	88
5.5	Probabilidade associada à ocorrência de estimativas da média menores do que 495 g.	88
5.6	Gráfico da distribuição χ^2	101
6.1	Gráficos de dispersão e coeficientes de correlação associados.	104
6.2	Idade <i>versus</i> tempo de reação a um estímulo.	107
6.3	Análise gráfica dos resíduos associados ao modelo ajustado.	109
6.4	QQplot dos resíduos.	110
6.5	Tempos de reação em função da idade e MRLS ajustado.	111

8.1	Gráfico de controle: idéia básica.	123
8.2	Gráfico de controle: limites de aviso.	125
8.3	Gráfico de controle: processo tendencioso.	125
8.4	Gráfico de controle: processo cíclico.	126
8.5	Efeito da média e desvio padrão em relação aos limites de especificação (LIE= limite inferior de especificação e LSE= limite superior de especificação).	127
8.6	Gráfico de controle para média - sem problemas (Tabela 8.1).	132
8.7	Gráfico de controle para amplitude - dados da Tabela 8.1	133
8.8	Gráfico de controle para o desvio padrão - dados da Tabela 8.1.	133
8.9	Gráfico de controle para média - com problemas (Tabela 8.2).	135

Capítulo 1

Conceitos Básicos e Técnicas de Estatística Descritiva

1.1 Introdução

A estatística torna-se a cada dia uma importante ferramenta de apoio à decisão. O objetivo deste capítulo é apresentar importantes conceitos de estatística e trabalhar a intuição do aluno para que este raciocine em cima de problemas que são tipicamente solucionados com o uso de técnicas estatísticas.

Para iniciar toda a discussão, primeiramente é necessário conhecer melhor o conceito de estatística. As pessoas comumente escutam falar sobre estatística na mídia. Diariamente são divulgadas informações tais como: índices de inflação, taxa de mortalidade, índice de desenvolvimento humano, proporção de eleitores, dentre outras.

O primeiro cuidado que devemos tomar é distinguir as estatísticas (valores numéricos que resumem informações) da Estatística que ganhou status de ciência. Nesta introdução, esta distinção será feita da seguinte forma, a **Estatística** como ciência terá sempre a primeira letra maiúscula, enquanto a **estatística** que transmite uma informação numérica será mencionada com a primeira letra minúscula.

No dicionário Aurélio, podemos encontrar como a primeira definição para Estatística:

[Do fr. statistique.] S. f. 1. Parte da matemática em que se investigam os processos de obtenção, organização e análise de dados sobre uma população ou sobre uma coleção de seres quaisquer, e os métodos de tirar conclusões e fazer ilações ou predições com base nesses dados.

Nesta definição, a Estatística é definida como parte da matemática. Entretanto, a Estatística já se desenvolveu o bastante para ocupar um campo no cenário científico como ciência que possui métodos e técnicas próprias. O famoso matemático John Tukey, que muito contribuiu para a Estatística, a caracterizou como:

"É uma ciência e não apenas um ramo da matemática, embora ferramentas da matemática sejam essenciais".

Em um rápido levantamento é possível encontrar várias definições para Estatística, das quais citaremos algumas interessantes.

"Ciência que utiliza métodos rígidos para lidar com incertezas".

"Ciência que procura estabelecer os limites da incerteza".

"Ciência que coleta, classifica e avalia numericamente fatos que servirão de base para inferência".

"Ciência da Incerteza".

Outras definições de conteúdo metafórico são:

"...é a arte de torturar os dados até que eles confessem a verdade".

"...nada mais é do que o bom senso expresso em números".

Embora todas as definições apresentadas contenham elementos importantes, a Estatística a ser apresentada neste material estará mais relacionada a definição a seguir:

A Estatística é um conjunto de métodos e técnicas que auxiliam a tomada de decisão sob a presença de incerteza."

Na maioria das definições apresentadas, verificamos a presença da palavra incerteza. De fato, o conceito de incerteza está vinculado à aplicação dos métodos e técnicas de análise estatística.

A incerteza

A incerteza permeia várias áreas do conhecimento: física, ciências sociais, comportamento humano, economia e ciências naturais. O tratamento quantitativo adequado a incerteza é obtido por meio do estudo de probabilidades.

A incerteza é consequência da variabilidade de um fenômeno e dificulta a tomada de decisões. Considere um simples exemplo da vida cotidiana: a ida de uma pessoa a uma agência bancária. Em torno deste fenômeno há uma série de incertezas, por exemplo: a quantidade de pessoas na fila, o número de atendentes, o tempo de atendimento, as condições do tempo, a cotação da moeda, etc.

Mesmo que um indivíduo procure informações prévias sobre todos estes elementos, sob os quais paira a incerteza, ainda assim não será possível predizer o desfecho. Podemos, por exemplo, analisar as condições do tempo, obter informações sobre o tráfego, ligar para a agência bancária e, ainda assim, não conseguimos precisar o horário em que receberemos o desejado atendimento bancário.

1.1.1 Estatística descritiva x estatística inferencial

A Estatística é conhecida, por muitas pessoas, como uma ferramenta meramente descritiva, ou seja, descreve dados por meio de percentagens, gráficos e tabelas. Apesar da estatística cumprir, também, este papel de resumir as informações, seu potencial de uso é muito mais amplo.

A tomada de decisão se apóia no uso da Estatística Inferencial. A seguir são delineadas as funções destas duas abordagens:

Estatística descritiva (Dedutiva)

O objetivo da Estatística Descritiva é resumir as principais características de um conjunto de dados por meio de tabelas, gráficos e resumos numéricos. Descrever os dados pode ser comparado ao ato de tirar uma fotografia da realidade. Caso a câmera fotográfica não seja adequada ou esteja sem foco, o resultado pode sair distorcido. Portanto, a análise estatística deve ser extremamente cuidadosa ao escolher a forma adequada de resumir os dados. Apresentamos na Tabela 1.1 um resumo dos procedimentos da Estatística Descritiva.

Tabela 1.1: Resumo de técnicas de estatística descritiva

Tabelas de frequência	Ao dispor de uma lista volumosa de dados, as tabelas de frequência servem para agrupar informações de modo que estas possam ser analisadas. As tabelas podem ser de frequência simples ou de frequência em faixa de valores.
Gráficos	<p>O objetivo da representação gráfica é dirigir a atenção do analista para alguns aspectos de um conjunto de dados.</p> <p>”Um gráfico vale mais que mil palavras”.</p> <p>Alguns exemplos de gráficos são: diagrama de barras, diagrama em setores, histograma, box-plot, ramo-e-folhas, diagrama de dispersão, gráfico sequencial.</p>
Resumos numéricos	Por meio de medidas ou resumos numéricos podemos levantar importantes informações sobre o conjunto de dados tais como: a tendência central, variabilidade, simetria, valores extremos, valores discrepantes, etc.

Estatística inferencial (Indutiva)

A Estatística Inferencial utiliza informações incompletas para tomar decisões e tirar conclusões satisfatórias. O alicerce das técnicas de estatística inferencial está no cálculo de probabilidades. Duas técnicas de estatística inferencial são as mais conhecidas: a estimação e o teste de hipóteses que são descritas na Tabela 1.2.

Tabela 1.2: Resumo de técnicas de estatística inferencial

Estimação	A técnica de estimação consiste em utilizar um conjunto de dados incompletos, ao qual iremos chamar de amostra, e nele calcular estimativas de quantidades de interesse. Estas estimativas podem ser pontuais (representadas por um único valor) ou intervalares.
Teste de Hipóteses	O fundamento do teste estatístico de hipóteses é levantar suposições acerca de uma quantidade não conhecida e utilizar, também, dados incompletos para criar uma regra de escolha.

Um exemplo tradicional do uso da estatística inferencial é apresentado a seguir.

Exemplo 1.1. Um instituto de pesquisa deseja estimar a proporção de eleitores do partido de situação no primeiro turno das eleições presidenciais. Ao coletar uma amostra de 1200 eleitores, a proporção foi estimada em 54%.

No Exemplo 1.1, a quantidade a ser estimada é a proporção de eleitores que votarão no partido de situação nas eleições presidenciais. Somente a realização das eleições revelará esta quantidade. Entretanto, estimá-la, com base em uma amostra, auxilia a tomada de decisões tais como a alteração de uma estratégia de campanha política.

Uma outra aplicação da estatística inferencial aparece no Exemplo 1.2 em que duas hipóteses são colocadas em questão. Será que uma nova droga a ser lançada aumenta, ou não, a produção de um hormônio ?

Exemplo 1.2. Um laboratório deseja verificar se uma nova droga aumenta a produção de testosterona em homens com idade acima de 35 anos. Ao aplicá-la em um grupo de 40 indivíduos, constatou-se que após um período de tempo a droga aumentou significativamente a quantidade do referido hormônio.

Exemplo 1.3. Em uma fábrica de parafusos, a peça é considerada dentro da especificação caso seu comprimento esteja no intervalo entre 4,8cm e 5,2cm. Os técnicos de controle de qualidade selecionam diariamente 100 parafusos fabricados e calculam o comprimento médio. Conhecendo a variabilidade nos tamanhos dos parafusos fabricados, caso o comprimento médio esteja abaixo de 4,99 cm ou acima de 5,01 cm, o processo será interrompido.

No Exemplo 1.3, espera-se que o comprimento médio de um conjunto de parafusos amostrados esteja dentro de um intervalo. Caso isto não ocorra, o processo de produção sofre uma interrupção. Neste caso, a estatística inferencial é utilizada para criar uma regra de decisão com base em observações de um subconjunto de 100 peças.

1.1.2 População e amostra

O uso da Estatística Inferencial oferece suporte à tomada de decisão com base em apenas uma parte das informações relevantes no problema estudado. A partir de agora, vamos utilizar os conceitos de *população* e *amostra* para representar, respectivamente, o conjunto total e o conjunto parcial destas informações.

População: é o conjunto de todas as unidades sobre as quais há o interesse de investigar uma ou mais características. O conceito de população em Estatística é bem mais amplo do que o uso comum desta palavra. A população pode ser formada por pessoas, domicílios, peças de produção, cobaias, ou qualquer outro elemento a ser investigado.

Amostra: é um subconjunto das unidades que constituem a população.

A caracterização da população é feita em função de um problema a ser estudado. Se um vendedor deseja fazer um levantamento dos potenciais clientes para o seu produto, a população será formada por todos os indivíduos com possibilidade de consumir aquele produto. Se este produto for, por exemplo, um iate, a população deve ser constituída apenas por indivíduos com renda suficiente para comprá-lo. Se o objetivo for avaliar a eficácia de tratamento contra um tipo de câncer, somente indivíduos com este problema devem compor a população.

Para que haja uma clara definição das unidades que formam a população é necessária a especificação de 3 elementos: uma característica em comum, localização temporal e localização geográfica.

Exemplo 1.4. Estudo da inadimplência de clientes em um banco multinacional com agências no Brasil.

Problema: Estudar a inadimplência dos clientes do banco HSBC.

Característica	correntista do banco HSBC
Tempo	clientes com cadastro em julho de 2007
Região	agências de Curitiba e região metropolitana

Exemplo 1.5. Estudo da obesidade em alunos do segundo grau por intermédio da medida do índice de massa corpórea.

Sem a definição dos 3 elementos, conforme os exemplos acima, torna-se difícil proceder a coleta de dados. Quando um estudo estatístico levanta informações de todas as

Problema: Estudar a obesidade em alunos do segundo grau.

Característica	alunos de 2o. grau da rede pública
Tempo	matriculados em janeiro de 2007
Região	região metropolitana de Curitiba

unidades da população, este chama-se **Censo**. Há casos em que o levantamento censitário é inviável, como por exemplo em testes destrutivos. Mesmo quando é possível realizar o Censo, o custo representa quase sempre um entrave. Nestes casos, a saída é estudar apenas parte da população para inferir sobre o todo e o levantamento é dito ser por **amostragem**.

O processo de amostragem pode ser probabilístico (aleatório) ou não-probabilístico (não-aleatório). Os métodos de inferência estatística são aplicáveis no primeiro caso pois a amostragem probabilística garante a representatividade da amostra. Como nem sempre as unidades da amostra podem ser obtidas por meio de seleção probabilística, existem alternativas como a amostragem de conveniência e amostragem por quotas.

Quando a amostragem é probabilística é necessária a existência de um *cadastro* que contenha a relação de todas as unidades na população.

Técnicas de amostragem

A Teoria de Amostragem é um ramo da estatística que estuda métodos para levantar amostras que sejam representativas da população. O princípio básico desta teoria é ter o máximo de precisão na avaliação das quantidades de interesse com o mínimo tamanho de amostra. Nem sempre é possível ponderar estas duas questões de forma a obter amostras representativas. Sendo assim, diferentes métodos de seleção de amostras foram desenvolvidos para situações específicas. Apresentamos e discutimos alguns deles.

- **Amostragem Aleatória Simples:** Consiste em selecionar aleatoriamente uma amostra de tamanho n em uma população formada por N indivíduos. A grande vantagem desta técnica é atribuir igual probabilidade de seleção a todas as possíveis amostras. Entretanto, para este tipo de amostragem, é crucial a existência de um cadastro com a relação de todas as unidades populacionais. Isto é inviável em muitas situações. Esta amostragem pode ser feita com reposição, o que garante a probabilidade $1/N$ de um elemento da população participar da amostra. Por outro lado, nessa situação um elemento pode aparecer múltiplas vezes. Quando a amostragem é feita sem reposição, a probabilidade de um elemento ser incluído na amostra se modifica durante o processo de seleção, pois a população é sequencialmente reduzida de 1 elemento.
- **Amostragem Aleatória Estratificada:** Este tipo de amostragem busca a formação de h estratos homogêneos em relação à característica estudada e, posteriormente, amostragem aleatória simples ou amostragem sistemática dentro de cada estrato. A amostragem estratificada é vantajosa quando há o conhecimento prévio de grupos

que sejam mais homogêneos internamente e heterogêneos entre si, em relação à característica investigada. Nestas situações há um ganho em relação à amostragem aleatória simples pois a seleção dentro dos estratos leva a diminuição do tamanho de amostra, mantendo a precisão das estimativas. Uma etapa importante da amostragem aleatória estratificada é a alocação da amostra pelos estratos, ou seja, quantos elementos da amostra pertencerão ao estrato 1, estrato 2, ..., estrato h . Dois tipos de alocação são comumente aplicados: alocação uniforme (mesmo número de elementos nos estratos) e a alocação proporcional (número de elementos proporcional ao tamanho do estrato).

- Amostragem por Conglomerados (Clusters): Neste método, ao invés da seleção de unidades da população, são selecionados conglomerados (clusters) destas unidades. Esta é uma alternativa para quando não existe o *cadastro*. Se a unidade de interesse, por exemplo, for um aluno, pode ser que não exista um cadastro de alunos, mas sim de escolas. Portanto, pode-se selecionar escolas e nelas investigar todos os alunos. Este tipo de amostragem induz indiretamente aleatoriedade na seleção das unidades que formarão a amostra e tem a grande vantagem de facilitar a coleta de dados.
- Amostragem Sistemática: Caso exista uma lista das unidades populacionais, a amostragem sistemática é uma técnica simples que a partir da razão $k = \frac{N}{n}$, de unidades populacionais para cada unidade amostral, sorteia-se um número inteiro no intervalo $[1, k]$ que serve como ponto de partida para a escolha do primeiro elemento a ser incluído na amostra. Descartando os $k - 1$ próximos elementos, seleciona-se o segundo e assim por diante. Tal como na amostragem aleatória simples, é necessária a existência de um cadastro, entretanto nem todas amostras são passíveis de seleção, por isto este procedimento é classificado como quasi-aleatório. Uma das grandes vantagens da amostragem sistemática, em relação à amostragem aleatória simples, é a praticidade na seleção dos elementos. Problemas com a amostragem sistemática podem surgir quando a sequência dos elementos no cadastro induz um comportamento periódico ou cíclico na principal variável a ser investigada. Considere, por exemplo, uma vila com 20 casas numeradas de 1 a 20. Se todas as casas cujos números são múltiplos de 4 estiverem mais perto da linha de trem e o intuito é medir poluição sonora, a amostragem sistemática não será adequada.
- Amostragem por Cotas: A amostragem por cotas assemelha-se é amostragem estratificada, embora dentro dos estratos não seja feita a amostragem aleatória simples. é uma alternativa para casos em que não há a existência de um cadastro, mas há informação disponível sobre o perfil desta população em relação a um fator de estratificação que pode auxiliar a representatividade da amostra (exemplo: 50% de homens e 50% de mulheres).
- Amostragem de Conveniência: Esta é uma forma de amostragem não-probabilística que leva em conta as restrições envolvidas no levantamento amostral. A unidades

amostrais são incluídas por algum tipo de conveniência, em geral ausência de tempo e recursos materiais para o levantamento dos dados. Embora não sejam feitas inferências em amostras de conveniência, estas podem ser importantes para levantar hipóteses e formular modelos.

Exemplo 1.6. Uma firma de contabilidade tem $N = 50$ clientes comerciantes. Seu proprietário pretende entrevistar uma amostra de 10 clientes para levantar possibilidades de melhora no atendimento. Escolha uma amostra aleatória simples de tamanho $n = 10$.

- Primeiro passo: atribuir a cada cliente um número entre 1 e 50.
- Segundo passo: recorrer à tabela de números aleatórios para selecionar aleatoriamente 10 números de 1 a 50. Os clientes identificados pelos números selecionados comporão a amostra.

Exemplo 1.7. Uma escola tem um arquivo com 5000 fichas de alunos e será selecionada, sistematicamente, uma amostra de 1000 alunos. Neste caso, a fração de amostragem é igual a $\frac{n}{N} = 1000/5000$ que representa $k = 5$ elementos na população para cada elemento selecionado na amostra. Na amostragem sistemática somente o ponto de partida é sorteado dentre as 5 primeiras fichas do arquivo. Admitamos que foi sorteado o número 2, então a amostra é formada pelas fichas 2, 7, 12, 17, \dots , 4992, 4997.

1.1.3 Variáveis e suas classificações

Em um levantamento de dados, censitário ou por amostragem, investiga-se uma ou mais características de interesse que supostamente variam de uma unidade para outra. Estas características serão chamadas a partir de agora de variáveis. A variável pode ser uma quantidade, sobre a qual podem ser realizadas operações aritméticas, ou pode ser um atributo como cor de pele, zona de moradia ou classe social. No primeiro caso, a variável é classificada como quantitativa e na outra situação ela é dita ser qualitativa.

A classificação da variável vai ser determinante para o tipo de análise estatística a ser conduzida. Sobre uma variável qualitativa, não podemos calcular muitos dos resumos numéricos tais como a média aritmética, a variância e o desvio padrão. Por outro lado, o gráfico de setores (ou pizza), não é adequado para representar as freqüências das temperaturas observadas durante um ano, ao menos que os valores sejam categorizados.

As variáveis quantitativas possuem uma subclassificação, elas podem ser discretas ou contínuas. O primeiro caso ocorre quando os possíveis valores da variável podem ser enumerados. Esta situação é típica de dados oriundos de contagens, como por exemplo o número diário de assaltos em um quarteirão que pode assumir valores no conjunto $\{0, 1, 2, 3, \dots\}$. A segunda subclassificação ocorre nos casos em que a variável pode assumir valores em um intervalo contínuo, por conseqüência, os possíveis valores são infinitos e não-enumeráveis. A variável idade, por exemplo, é uma variável contínua pois se for medida com bastante precisão, um indivíduo pode apresentar 32,1023 anos de idade e, dificilmente

dois indivíduos terão idades iguais. A seguir são apresentados alguns outros exemplos de variáveis quantitativas:

- **Variáveis quantitativas**

Discretas: número de filhos, número de plantas, quantidade de peças e número de assaltos.

Contínuas: as variáveis contínuas podem assumir infinitos valores (índice de preços, salário, peso, altura e pressão sistólica).

Toda variável que não é quantitativa, será classificada como qualitativa. Os valores que a variável pode assumir são chamados de níveis ou categorias. Caso estes níveis sejam ordenáveis, a variável é dita ser ordinal, caso contrário ela é classificada como nominal. É importante ressaltar que esta ordenação nos níveis (categorias) da variável é natural tal como ocorre com a variável classe social. Nesta situação, Classe A > Classe B > Classe C > Classe D. Como já foi comentado, o tipo de variável determina o tipo de análise e, para variáveis qualitativas ordinais, um resumo numérico, uma técnica gráfica ou uma tabela de frequência deve incorporar a idéia de ordenação.

- **Variáveis qualitativas (atributos)**

Ordinais (ex: classe social, cargo na empresa e classificação de um filme.)

Nominais (ex: sexo, bairro, cor de pele e canal de TV preferido.)

além das classificações mencionadas, vamos destacar uma outra situação em que a característica de interesse é investigada ao longo do tempo (espaço) constituindo o que chamamos de uma série temporal. A análise de uma variável que é medida ao longo do tempo deve considerar aspectos específicos como tendência e sazonalidade. Ao resumir estas variáveis, quando há a presença de tendência o valor médio modifica-se ao longo do tempo, enquanto a sazonalidade pode explicar variações periódicas, como o aumento de venda de televisores nos meses de novembro e dezembro.

Série temporal

Conjunto de observações ordenadas no tempo (índice mensal de inflação, temperatura máxima diária, cotação diária do dólar e número de nascimentos diários.).

1.2 Técnicas de estatística descritiva

A principal função da Estatística Descritiva é resumir as informações contidas em um conjunto de dados por meio de tabelas, gráficos e medidas características (resumos numéricos). A descrição dos dados deve ser objetiva, ter precisão de significado e simplicidade no cálculo para que outras pessoas possam compreender e, eventualmente, reproduzir os resultados. Recorremos novamente aqui à metáfora da fotografia pois realizar uma análise

descritiva é como tirar uma foto da realidade, caso a lente esteja desfocada, o resultado não será claro.

As técnicas de estatística descritiva são aplicadas a observações de uma ou mais variáveis, tomadas nas unidades de interesse. Quando apenas uma variável é resumida, a descrição é univariada, caso duas variáveis sejam resumidas conjuntamente, a descrição é bivariada. Ao conjunto de observações de uma variável chamaremos de dados brutos, lista ou rol.

1.2.1 Tabelas de frequências

A partir dos dados brutos, podemos agrupar os valores de uma variável quantitativa ou qualitativa e construir a chamada *tabela de frequências*. As tabelas de frequências podem ser simples ou por faixas de valores, dependendo da classificação da variável.

Tabelas de frequências simples

As tabelas de frequências simples são adequadas para resumir observações de uma variável qualitativa ou quantitativa discreta, desde que esta apresente um conjunto pequeno de diferentes valores.

Utilizamos os dados presentes em Magalhães & Lima (2004), referentes a um questionário aplicado a moradores de comunidades de baixa renda em São paulo, para construir a Tabela 1.3 que resume as observações da variável estado civil.

Tabela 1.3: Frequências de estado civil em uma amostra de 385 indivíduos.

Estado Civil	Frequência Absoluta	Frequência Relativa Percentual
Solteiro	165	42,86%
Casado	166	43,12%
Divorciado	10	2,6%
Viúvo	12	3,12%
Outro	32	8,31%
Total	385	100%

A variável estado civil é qualitativa nominal e no levantamento feito nos 385 indivíduos apareceram respostas que foram agrupadas em 5 níveis (categorias) para esta variável: Solteiro, Casado, Divorciado, Viúvo e Outro. A construção da tabela de frequência simples, neste caso, resume os dados brutos pela contagem de vezes (frequência absoluta) que uma determinada categoria foi observada.

A partir da Tabela 1.3, podemos rapidamente tirar conclusões a respeito dos dados como a constatação de que neste grupo de indivíduos, a quantidade de solteiros(165) e casados(166) é praticamente a mesma e há uma parcela muito pequena de divorciados(10).

Estes comentários, embora simples, tornam-se mais claros quando analisamos a coluna das frequências relativas.

- n_i : frequência do valor i
- n : frequência total
- $f_i = \frac{n_i}{n}$: frequência relativa (útil quando comparamos grupos de tamanhos diferentes)
- $f_i \times 100$: frequência relativa percentual

Para variáveis cujos valores possuem ordenação natural faz sentido incluirmos também uma coluna contendo frequências acumuladas fac . Sua construção ajuda a estabelecer pontos de corte, chamados de separatrizes ou quantis, a partir dos quais está concentrada uma determinada frequência de valores da variável.

A Tabela 1.4 exhibe as frequências das idades em uma amostra de 50 estudantes que preencheram um questionário sobre hábitos de lazer (ver Magalhães & Lima(2004)).

Tabela 1.4: Tabela de frequências para a variável Idade.

Idade	n_i	f_i	fac
17	9	0,18	0,18
18	22	0,44	0,62
19	7	0,14	0,76
20	4	0,08	0,84
21	3	0,06	0,90
22	0	0	0,90
23	2	0,04	0,94
24	1	0,02	0,96
25	2	0,04	1,00
total	n=50	1	

Tabelas de frequências em faixas de valores

Para agrupar dados de uma variável quantitativa contínua ou até mesmo uma variável quantitativa discreta com muitos valores diferentes, a tabela de frequências simples não é mais um método de resumo, pois corremos o risco de praticamente reproduzir os dados brutos.

A utilização de tabelas, nestas situações em que a variável registra diversos valores, é feita mediante a criação de faixas de valores ou intervalos de classe. Utilizando este procedimento, devemos tomar cuidado pois ao contrário da tabela de frequência simples, não é mais possível reproduzir a lista de dados a partir da organização tabular. Em outras palavras, estamos perdendo informação ao condensá-las.

Veja o exemplo na Tabela 1.5 que traz dados sobre as horas semanais de atividades físicas dos 50 estudantes que participaram do levantamento sobre hábitos de lazer.

Tabela 1.5: Tabela de frequências para a variável horas semanais de atividade física

horas semanais de atividade física	n_i	f_i	fac
0 – 2	11	0,22	0,22
2 – 4	14	0,28	0,5
4 – 6	12	0,24	0,74
6 – 8	8	0,16	0,90
8 – 10	3	0,06	0,96
10 – 12	2	0,04	1,00
total	50	1	

O resumo na Tabela 1.5 é feito mediante a construção de 6 intervalos de comprimento igual a 2 horas e posteriormente a contagem de indivíduos com valores identificados ao intervalo. Um indivíduo que gastou 6 horas semanais de exercício será contado no quarto intervalo (6|–8) que inclui o valor 6 e exclui o valor 8.

No mesmo levantamento amostral foi observado o peso dos 50 estudantes. A variável peso é classificada como quantitativa contínua e foi mensurada com uma casa decimal. Com esta precisão de medida foram observados 36 valores diferentes, o que inviabiliza a construção da tabela de frequência simples.

Novamente o recurso a ser utilizado é construir classes ou faixas de pesos e contar o número de ocorrências em cada faixa. Com 6 intervalos de peso, os dados foram agrupados conforme a Tabela 1.6.

Tabela 1.6: Tabela de frequências para a variável Peso

Peso de crianças	n_i	f_i	fac
40,0 – 50,0	8	0,16	0,16
50,0 – 60,0	22	0,44	0,60
60,0 – 70,0	8	0,16	0,76
70,0 – 80,0	6	0,12	0,88
80,0 – 90,0	5	0,10	0,98
90,0 – 100,0	1	0,02	1,00
total	100	1	

Se concordamos que a tabela em faixa de valores ajuda a resumir a quantidade de informações em um conjunto de dados, com variáveis contínuas ou discretas que assumam muitos valores, ainda fica pendente a questão de quantos intervalos serão necessários para a construção desta tabela.

Para a decepção de muitos, não há uma resposta definitiva a esta pergunta e existem várias sugestões na literatura para se chegar a este número. Esta questão será discutida posteriormente ao falarmos de uma técnica gráfica chamada de histograma, mas o bom senso indica que o número de intervalos deve estar entre 5 e 10 neste tipo de descrição.

1.2.2 Medidas-resumo

Em um processo de coleta de dados, por meio de amostragem ou censo, faz-se necessário resumir as informações contidas nas observações das variáveis utilizando as medidas adequadas. Neste capítulo, estas serão chamadas medidas-resumo. Veja o exemplo a seguir.

Exemplo 1.8. Em um ponto de ônibus, uma pessoa pergunta sobre o tempo até a passagem de uma determinada linha. Suponha que você havia registrado, na semana anterior, os tempos (em minutos) e obteve os seguintes resultados:

9; 12; 8; 10; 14; 7; 10

Ao responder: "o ônibus demora, em média, 10 minutos", você está trocando um conjunto de valores por um único número que os resume. Ao adotar este procedimento foi utilizada uma medida-resumo, neste caso a média aritmética.

Novamente, a classificação da variável vai orientar a escolha da medida-resumo mais adequada. A maior parte das medidas a serem apresentadas aplicam-se somente à variáveis quantitativas.

As medidas-resumo podem focar vários aspectos no conjunto de dados; tendência central, dispersão, ordenação ou simetria na distribuição dos valores. Aqui serão apresentadas 3 classes de medidas:

- Tendência Central
- Dispersão (Variabilidade)
- Separatrizes ¹

Tendência central

As medidas de tendência central indicam, em geral, um valor central em torno do qual os dados estão distribuídos. Este valor no Exemplo 1.8 é igual a 10 e corresponde a média aritmética. As principais medidas de tendência central na Estatística são: média, mediana e moda. Além destas, outras medidas são utilizadas com fins específicos tais como: média geométrica, média harmônica, média ponderada e trimédia.

Sejam as observações obtidas a partir da variável X , em uma população ou em uma amostra:

¹Alguns autores classificam as medidas de tendência central e separatrizes como medidas de posição.

$$x_1, x_2, \dots, x_n$$

e considere a seguinte notação para os dados ordenados:

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

em que $x_{(1)}$ é o menor valor (mínimo) no conjunto de dados e $x_{(n)}$ é o maior valor (máximo).

Com base nesta notação, apresentamos a seguir os conceitos de média, mediana e moda.

Média (Aritmética)

A média aritmética também é conhecida como ponto de equilíbrio e centro de gravidade, denominações surgidas da Física. Ela indica o valor em torno do qual há um equilíbrio na distribuição dos dados. O seu cálculo é feito conforme:

$$\bar{x}_{obs} = \frac{\sum_{i=1}^n x_i}{n}.$$

Definindo desvio da i -ésima observação, em torno da média observada, como $d_i = x_i - \bar{x}_{obs}$, a soma destes desvios sempre será igual a zero. A demonstração deste resultado é trivial. Basta observar que:

$$\sum_{i=1}^n (x_i - \bar{x}_{obs}) = \sum_{i=1}^n x_i - n\bar{x}_{obs} = \sum_{i=1}^n x_i - \sum_{i=1}^n x_i = 0.$$

A média aritmética é pouco robusta às mudanças em valores extremos no conjunto de dados observados.

Suponha um conjunto de valores ordenados de forma crescente, $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ e neles a média aritmética permanece \bar{x}_{obs} . Se um erro de anotação acrescentasse k unidades ao maior valor da amostra ($x_{(n)}$), a média inicialmente calculada \bar{x}_{obs} será acrescida de k/n unidades. O impacto da alteração na média será diretamente proporcional a magnitude de k e inversamente proporcional a quantidade de observações n .

A média só poderá ser calculada para variáveis quantitativas (discretas e contínuas). A única exceção ocorre quando a variável qualitativa é binária, ou seja, apresenta duas categorias como por exemplo: masculino e feminino. Se atribuímos os valores 0 e 1 às categorias masculino e feminino, respectivamente, ao realizar o cálculo da média o resultado indica a proporção de mulheres na amostra.

Exemplo 1.9. Uma pesquisa registrou em um grupo de 10 pessoas a satisfação em relação ao governo. Cada respondente deveria simplesmente informar se estava satisfeito ou não. Para os que estavam satisfeitos, anotou-se o valor 1 e os que estavam insatisfeitos 0. No final foi obtido o seguinte conjunto de dados:

$$x_1 = 1; x_2 = 1; x_3 = 0; x_4 = 0; x_5 = 0; x_6 = 0; x_7 = 0; x_8 = 1; x_9 = 0; x_{10} = 1.$$

A média calculada a partir dos dados acima é igual a :

$$\bar{x}_{obs} = \frac{\sum_{i=1}^n x_i}{n} = \frac{4}{10} = 0,4.$$

A interpretação deste resultado é de que 40% dos entrevistados estão satisfeitos com o governo.

Com isto, verificamos que ao calcular a proporção estamos calculando a média de uma variável qualitativa binária.

Mediana

A mediana observada md_{obs} é o valor central em um conjunto de dados ordenados. Pela mediana o conjunto de dados é dividido em duas partes iguais sendo metade dos valores abaixo da mediana e, a outra metade, acima.

Vamos denominar md_{obs} o valor da mediana observado em um conjunto de dados. Repare que para encontrar um número que divida os n dados ordenados em duas partes iguais devem ser adotados dois procedimentos:

se n é ímpar

$$md_{obs} = x_{(\frac{n+1}{2})}.$$

se n é par

$$md_{obs} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

1. Para um conjunto com um número n (ímpar) de observações, a mediana é o valor na posição $\frac{n+1}{2}$.
2. Para um conjunto com um número n (par) de observações a mediana é a média aritmética dos valores nas posições $\frac{n}{2}$ e $\frac{n}{2} + 1$.

Exemplo 1.10. Um levantamento amostral coletou e ordenou de forma crescente a renda mensal(em reais) de 8 trabalhadores da construção civil.

500	550	550	550	600	600	700	1750
-----	-----	-----	-----	-----	-----	-----	------

Neste exemplo, há $n = 8$ observações. Portanto, a mediana será obtida como:

$$md_{obs} = \frac{x_{(4)} + x_{(5)}}{2} = \frac{550 + 600}{2} = 575,$$

isto é, a média aritmética entre a quarta ($x_{(4)} = 550$) e a quinta ($x_{(5)} = 600$) observações que resulta em 575. Neste caso, note que um trabalhador com renda mediana ganha 150 reais a menos do que um trabalhador com renda média que é igual a ($\bar{x}_{obs} = 725$).

Nas situações em que a mediana é exatamente igual a média, diz-se que os dados tem distribuição simétrica, ou seja, a probabilidade de sortear um número do conjunto de dados e este estar localizado abaixo da média (mediana) é igual a 50%. Um modo direto de mensurar a simetria na distribuição dos dados é calcular a diferença entre a mediana e a média, quanto mais próximo de zero for este resultado, maior a simetria no conjunto de dados..

Moda

A moda observada (mo_{obs}) é simplesmente o valor mais freqüente em um conjunto de dados. Considere o seguinte conjunto de dados: 3; 4; 5; 7; 7; 7; 9; 9. Temos $mo_{obs} = 7$, pois o valor 7 é aquele que ocorre com a maior freqüência.

A moda é uma medida de tendência central que pode ser calculada para qualquer tipo de variável, seja ela quantitativa ou qualitativa. Veja o exemplo a seguir.

Exemplo 1.11. Em uma amostra de pacientes em um laboratório foram observados os tipos sanguíneos encontrados em 1000 exames, com os seguintes resultados mostrados na Tabela 1.7.

Tabela 1.7: Tipos sanguíneos de 1000 pacientes.

Tipo de Sangue	Freqüência Absoluta(n_i)
O	497
A	441
B	123
AB	25
Total	1000

Para estes dados, a moda observada é o sangue do tipo O, pois este é o mais freqüente. Cabe ressaltar que, para esta variável, apenas podemos calcular esta medida de tendência central.

Pode ocorrer a situação em que existam duas modas, como por exemplo para o conjunto $A = \{3, 4, 4, 5, 5, 6, 7, 8\}$. Neste caso, os valores 4 e 5 são os mais freqüentes. O conjunto é dito ser **bimodal**.

Quando o conjunto possui mais do que duas modas ele é dito ser **multimodal**.

Outra situação extrema é aquela em que a moda não existe tal como ocorre para o conjunto $B = \{3, 4, 5, 6, 7, 8, 9\}$ cujas freqüências estão distribuídas uniformemente entre os diferentes valores, ou seja, nestes dados cada valor tem freqüência igual a 1 e o conjunto é dito ser **amodal**.

Medidas de dispersão

Muito embora as medidas de tendência central sejam utilizadas como o primeiro resumo numérico de um conjunto de dados, a sua representatividade está diretamente ligada com a variabilidade. Veja o Exemplo 1.12 a seguir.

Exemplo 1.12. Ao aplicar a mesma prova em dois grupos de 4 alunos cada, foram obtidos os resultados:

Notas da Turma A

aluno	1	2	3	4
nota	5	5	5	5

Notas da Turma B

aluno	1	2	3	4
nota	10	0	10	0

Ao utilizar a média, mediana e moda para resumir as informações das duas turmas, repare que os resultados coincidem (Tabela 1.8). A nota média é 5 e, em ambas as turmas, 50% dos alunos têm nota igual ou abaixo da média.

Embora as turmas sejam iguais em relação às medidas de tendência central, a heterogeneidade da turma B é maior, ou seja, a variabilidade das notas é maior nestes alunos. Isto faz com que a média da turma B, seja menos representativa do que a média da turma A, que realmente reflete o conhecimento dos 4 alunos.

Tabela 1.8: Medidas de tendência central para as notas das turmas A e B.

	Média	Mediana	Moda
Turma A	5	5	não existe
Turma B	5	5	não existe

As medidas de dispersão servem para quantificar a variabilidade dos valores em um conjunto de dados. Uma medida de tendência central para ser melhor compreendida deve estar acompanhada de uma medida de dispersão.

Nesta seção, serão apresentadas 5 medidas de dispersão (ver Tabela 1.9) para variáveis quantitativas, sendo que 4 delas utilizam a média como referência: desvio médio absoluto, variância, desvio padrão e coeficiente de variação.

Amplitude total

Esta medida é obtida a partir da diferença entre o máximo($x_{(n)}$) e o mínimo ($x_{(1)}$) em um conjunto de dados ordenados. Esta medida possui o valor 0 como limite inferior e é altamente sensível à valores extremos.

Tabela 1.9: Principais medidas de dispersão.

Medidas	Notação
Amplitude Total	Δ_{obs}
Desvio absoluto médio	dma_{obs}
Variância	var_{obs}
Desvio padrão	dp_{obs}
Coeficiente de Variação	cv_{obs}

$$\Delta_{obs} = x_{(n)} - x_{(1)}.$$

Para o Exemplo 1.12, o valor calculado para esta medida foi $\Delta_{obs} = 0$ para a turma A e $\Delta_{obs} = 10$ para a turma B. A diferença entre as amplitudes das duas turmas é a máxima que poderia ocorrer. A turma A tem menor variabilidade possível, pela amplitude total, enquanto a turma B tem a maior variabilidade possível de ser encontrada com o uso desta medida.

A grande limitação da amplitude total é quantificar a variabilidade com apenas o uso de duas observações; máximo e mínimo. Outras medidas exploram com maior profundidade o conjunto de dados e, apresentaremos na seqüência, 4 medidas baseadas no desvio em relação à média.

Desvio médio absoluto

É simplesmente o cálculo da média dos desvios absolutos. Para o seu cálculo, primeiramente deve ser calculada a média (\bar{x}_{obs}), posteriormente os desvios d_i das observações em relação a média e, por último, a média do módulo destes desvios conforme a fórmula a seguir.

$$dma_{obs} = \sum_{i=1}^n \frac{|x_i - \bar{x}_{obs}|}{n}.$$

Exemplo 1.13. Em uma prova, os alunos obtiveram as seguintes notas: 5; 6; 9; 10; 10. Obtenha o desvio médio absoluto.

$$dma_{obs} = \frac{|5 - 8| + |6 - 8| + |9 - 8| + |10 - 8| + |10 - 8|}{5} = 2.$$

Algo importante sobre esta medida, assim como a variância, desvio padrão e o coeficiente de variação é que todas são calculadas usando como referência de tendência central a média (\bar{x}_{obs}).

Exemplo 1.14. Um estudo sobre aleitamento materno investigou o peso de 10 nascidos vivos cuja média observada foi $\bar{x}_{obs} = 3,137$. Cada um dos pesos é apresentado na Tabela 1.10.

Tabela 1.10: Peso de 10 nascidos vivos

peso	2,50	2,45	4,15	3,30	2,86	3,45	3,48	2,33	3,70	3,15
desvios	-0,63	-0,69	1,01	0,16	-0,28	0,31	0,34	-0,81	0,56	0,01

$$dma_{obs} = \frac{0,63 + 0,69 + \dots + 0,01}{10} = 0,48$$

A interpretação desta medida para o Exemplo 1.14 indica que, em média, um nascido vivo tem peso 0,48kg distante da média observada que é 3,137kg.

Variância

Esta é a mais conhecida medida de variabilidade. Como será visto mais adiante, em muitas situações o cálculo de probabilidades depende exclusivamente do conhecimento da média e variância de uma variável na população.

O cálculo da variância assemelha-se com o do dma_{obs} , pois utiliza desvios quadráticos em vez dos absolutos. Assim, a variância também é chamada de média dos desvios quadráticos.

$$var_{obs} = \sum_{i=1}^n \frac{(x_i - \bar{x}_{obs})^2}{n}.$$

Para o mesmo conjunto de dados do Exemplo 1.13, a variância observada é igual a:

$$var_{obs} = \frac{(5-8)^2 + (6-8)^2 + (9-8)^2 + (10-8)^2 + (10-8)^2}{5} = 4,4.$$

Propriedades da variância:

1. A variância de um conjunto de números iguais é sempre 0.
2. Ao multiplicar todos os valores do conjunto por uma constante, a variância fica multiplicada por esta constante ao quadrado.
3. Ao somar uma constante a todos os valores de um conjunto, a variância não se altera.

Exemplo 1.15. O conjunto de notas do Exemplo 1.13 deve ser multiplicado por 10 para que este possa ser lançado no boletim. Deste modo, o novo conjunto é: 50, 60, 90, 100, 100. Qual a variância das notas lançadas no boletim ?

Solução: Basta multiplicar a variância encontrada anteriormente por 100.

Por utilizar a média como referência, o desvio absoluto médio e a variância também são afetados por valores extremos. No caso da variância o efeito é ainda maior pois os valores estão elevados ao quadrado.

Desvio padrão

Embora seja uma medida importante, a variância carece de interpretação pois é uma medida dos valores ao quadrado. Isto é contornado com o uso do desvio padrão que é obtido pelo cálculo da raiz quadrada da variância.

$$dp_{obs} = \sqrt{var_{obs}} = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x}_{obs})^2}{n}}.$$

A grande vantagem do desvio padrão é que este possibilita analisar a variabilidade na mesma escala em que os dados foram medidos. Se a variável em questão é peso (kg), o desvio padrão é expresso em kg, ao contrário da variância que é expressa em kg^2 .

Coefficiente de Variação

O coeficiente de variação é a razão entre o desvio padrão e a média. Esta é uma medida relativa que avalia o percentual de variabilidade em relação à média observada. Uma das grandes vantagens desta medida é a possibilidade de comparar a variabilidade de conjuntos medidos em diferentes escalas.

$$cv_{obs} = 100 \times \frac{dp_{obs}}{\bar{x}_{obs}}.$$

Exemplo 1.16. Um exame físico examinou 6 indivíduos cujos pesos(kg) foram:68; 70; 86; 55; 75 e 90. No mesmo exame, foram também tomadas medidas de altura (cm), com os seguintes valores:170; 160; 164; 164; 170 e180. Os indivíduos apresentam maior variabilidade no peso ou altura?

Como as unidades de medida são diferentes, utilizaremos o coeficiente de variação como medida de comparação entre os dois conjuntos.

Resumo	Peso (Kg)	Altura (cm)
média	74	168
desvio padrão	11,65	6,43
coeficiente de variação	15,7%	3,83 %

A partir dos coeficientes de variação constatamos que os indivíduos apresentam maior variabilidade no peso.

Separatrizes

As separatrizes são valores de referência em um conjunto de valores ordenados e, portanto, são aplicadas a variáveis quantitativas e qualitativas ordinais. A mediana md_{obs} é um exemplo destas medidas, pois separa o conjunto de dados em dois subconjuntos, com as menores e maiores observações.

Se o interesse é subdividir o conjunto ordenado em 4 partes de igual tamanho, serão necessários 3 valores para estabelecer esta separação. Estes valores são chamados quartis. O primeiro quartil (Q_1) estabelece o limite entre as 25% menores observações e as 75% maiores. O segundo quartil ($Q_2 = md_{obs}$) é igual a mediana e o terceiro quartil (Q_3) separa as 75% menores observações das 25% maiores.

Em um conjunto ordenado de observações de uma variável $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, o primeiro e o terceiro quartis podem ser obtidos avaliando as quantidades $Q_1 = x_{(\frac{n}{4})}$ e $Q_3 = x_{(\frac{3n}{4})}$, respectivamente. Tais quantidades são avaliadas mediante interpolações conforme o Exemplo 1.17.

Exemplo 1.17. Diâmetros de 9 peças são medidos em milímetros, com os seguintes resultados:

$$x_1 = 3 \quad x_2 = 1,5 \quad x_3 = 2,5 \quad x_4 = 3,5 \quad x_5 = 4 \quad x_6 = 2 \quad x_7 = 3,5 \quad x_8 = 2 \quad x_9 = 1,5.$$

e a amostra ordenada é representada da seguinte maneira:

$$x_{(1)} = 1,5 \quad x_{(2)} = 1,5 \quad x_{(3)} = 2 \quad x_{(4)} = 2 \quad x_{(5)} = 2,5 \quad x_{(6)} = 3 \quad x_{(7)} = 3,5 \quad x_{(8)} = 3,5 \quad x_{(9)} = 4.$$

O primeiro e terceiro quartis são encontrados pelos números $x_{(2,25)}$ e $x_{(6,75)}$. Repare que o primeiro quartil é um valor que fica entre $x_{(2)}$ e $x_{(3)}$, entretanto mais próximo de $x_{(2)}$. Realizando uma interpolação, avaliamos esta quantidade da seguinte forma:

$$Q_1 = x_{(2,25)} = x_{(2)} + 0,25[x_{(3)} - x_{(2)}] = 1,625.$$

e, segundo o mesmo raciocínio, também avaliamos o terceiro quartil:

$$Q_3 = x_{(6,75)} = x_{(6)} + 0,75[x_{(7)} - x_{(6)}] = 3,375.$$

Caso o interesse seja dividir o conjunto de dados em 10 partes iguais, serão necessários 9 números, chamados de decis. Para calcular um decil, utilizamos a formulação:

$$x_{(\frac{in}{10})} \quad i = 1, 2, 3, \dots, 9$$

e interpolamos os valores de forma similar ao que foi mostrado para os quartis.

1.2.3 Gráficos

Muitas vezes as informações contidas em tabelas podem ser mais facilmente entendidas se visualizadas em gráficos. Graças à proliferação dos recursos gráficos, existe hoje uma infinidade de tipos de gráficos que podem ser utilizados.

No entanto, a utilização de recursos visuais deve ser feita cuidadosamente; um gráfico desproporcional em suas medidas pode conduzir a conclusões equivocadas

Vamos abordar três tipos básicos de gráficos: setores ou pizza, barras e histograma.

Gráfico de setores ou pizza

Este gráfico é adequado para representar variáveis qualitativas. Sua construção consiste em repartir um disco em setores cujos ângulos são proporcionais às frequências relativas observadas nas categorias da variável.

Exemplo 1.18. Uma pesquisa de intenção de votos para os partidos A,B,C e D, realizada com 100 eleitores resultou na Tabela 1.11.

Tabela 1.11: Intenção de votos para os partidos A,B,C e D.

Partido	Frequência Absoluta	Frequência Relativa
A	40	0,4
B	30	0,3
C	20	0,2
D	10	0,1
Total	100	1

Conforme a Figura 1.1 a maior fatia corresponde ao partido A que detem 40% das intenções de voto. Embora tal informação esteja na Tabela 1.11, a assimilação das diferenças entre as intenções de votos é mais rápida no gráfico de setores.

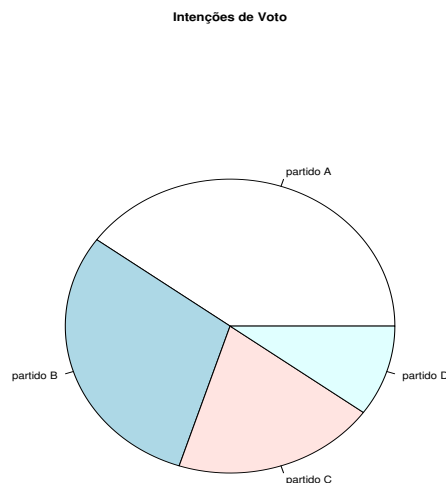


Figura 1.1: Gráfico de setores para a intenção de votos nos partidos A,B,C e D.

Gráfico de barras

Este gráfico representa a informação de uma tabela de frequências simples e, portanto, é mais adequado para variáveis discretas ou qualitativas ordinais. Utiliza o plano cartesiano com os valores da variável no eixo das abscissas e as frequências no eixo das ordenadas.

Para cada valor da variável desenha-se uma barra com altura correspondendo à sua frequência. É importante notar que este gráfico sugere uma ordenação dos valores da variável, podendo levar a erros de interpretação se aplicado à variáveis quantitativas nominais.

Exemplo 1.19. Um posto de saúde contém um cadastro das famílias regularmente atendidas em que consta o número de crianças por família. Ao resumir esta informação para todas as famílias em que há no máximo 5 crianças é obtida a Tabela 1.12.

Tabela 1.12: Número de crianças por família.

Número de crianças	Frequência Absoluta	Frequência Relativa
0	52	0,302
1	38	0,221
2	43	0,25
3	22	0,128
4	11	0,064
5	6	0,035
Total	172	1

A representação gráfica da Tabela 1.12 é apresentada na Figura 1.2. A altura de cada barra é diretamente proporcional ao número de famílias com a quantidade de filhos especificada no eixo das abcissas.

Histograma

O histograma é um gráfico que possibilita o primeiro contato com a formato da distribuição dos valores observados. Precede a sua construção a organização dos dados de uma variável quantitativa em faixas de valores.

Consiste em retângulos contíguos com base nas faixas de valores da variável e com área igual à frequência relativa da faixa. A altura de cada retângulo é denominada densidade de frequência ou simplesmente densidade definida pelo quociente da frequência relativa pela amplitude da faixa².

²Alguns autores usam a frequência absoluta ou porcentagem na construção do histograma. O uso da densidade impede que o histograma fique distorcido quando as faixas têm amplitudes diferentes.

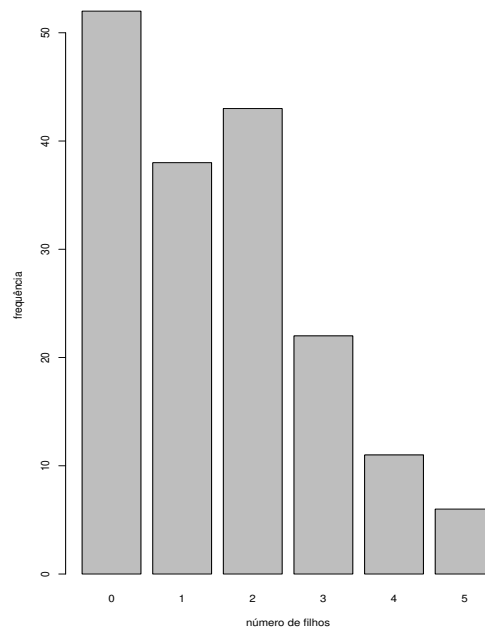


Figura 1.2: Gráfico de barras para o número de filhos por família.

Há 3 elementos que determinam a configuração da tabela de frequências em faixas de valores e do histograma:

- L - Número de faixas de valores
- h - Comprimento dos intervalos de classe
- Δ_{obs} - Amplitude total.

com a seguinte relação entre eles:

$$L = \frac{\Delta_{obs}}{h}.$$

Conforme já foi comentado, não existe uma regra definitiva para a determinação destes elementos. Entretanto, algumas formulações para L , o número de faixas de valores, são utilizadas com bastante frequência em pacotes computacionais. Dentre estas fórmulas, vamos citar duas de fácil aplicação que dependem somente de n , a quantidade de observações:

1. Fórmula de Sturges

$$L = 1 + 3,3 \log n.$$

2. Raiz quadrada de n

$$L = \sqrt{n}.$$

Assim como o gráfico de setores e o gráfico de barras são construídos a partir de uma tabela de frequências simples, o histograma é construído a partir de uma tabela de frequências em faixa de valores.

Exemplo 1.20. Um determinado teste mede o nível de estresse por uma escala de valores que varia continuamente de 0 a 13. Uma empresa aplicou o teste a 70 funcionários obtendo os seguintes resultados:

Tabela 1.13: Nível de estresse em 70 funcionários de uma empresa.

Nível de estresse	Frequência Absoluta (n_i)	Frequência Relativa (f_i)
0 –2	5	0,07
2 –4	10	0,14
4 –6	13	0,19
6 –8	16	0,23
8 –10	11	0,16
10 –12	9	0,13
12 –14	6	0,09
Total	70	1

A informações da Tabela 1.13, com $L = 7$ faixas de valores para a variável nível de estresse, são diretamente transpostas para o gráfico histograma conforme a Figura 1.3.

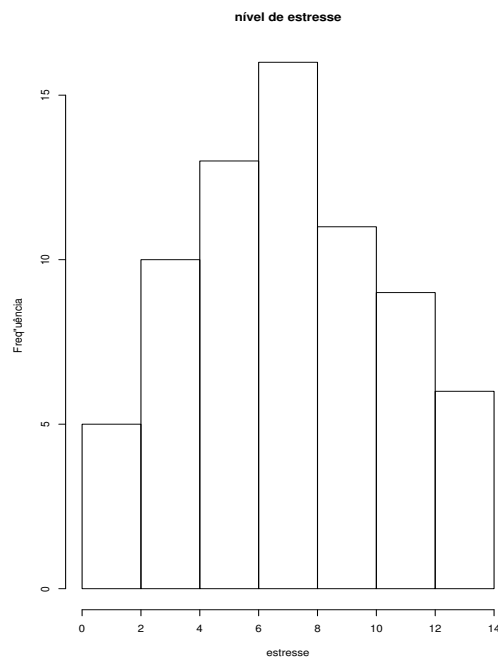


Figura 1.3: Histograma para o nível de estresse.

Box-plot

O Box-Whisker Plot, mais conhecido por box-plot, é uma ferramenta gráfica apropriada para resumir o conjunto de observações de uma variável contínua. Este gráfico revela vários aspectos dos dados, dentre eles: tendência central, variabilidade e simetria. O boxplot também possibilita visualizar valores atípicos(*outliers*).

A construção do box-plot é feita com base no chamado resumo de cinco números: o mínimo, o primeiro quartil(Q_1), a mediana (md_{obs}), o terceiro quartil (Q_3) e o máximo. Após calcular estes cinco números em um conjunto de dados observados, eles são dispostos de acordo com a Figura 1.4.

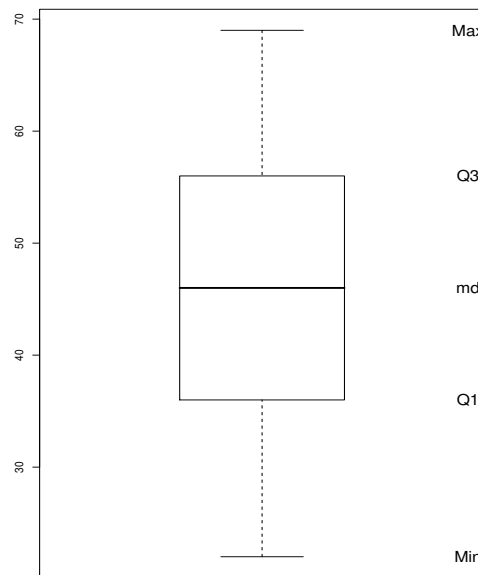


Figura 1.4: Desenho esquemático do box-plot com base no resumo de 5 números.

A parte central do gráfico é composta de uma “caixa” com o nível superior dado por Q_3 e o nível inferior por Q_1 . O tamanho da caixa é uma medida de dispersão chamada amplitude interquartílica ($AIQ = Q_3 - Q_1$).

A mediana, medida de tendência central, é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até os valores máximo e mínimo.

Exemplo 1.21. Suponha que um produtor de laranjas, que costuma guardar as frutas em caixas, está interessado em estudar o número de laranjas por caixa. Após um dia de colheita, 20 caixas foram contadas. Os resultados brutos, após a ordenação, foram:

22 29 33 35 35 37 38 43 43 44 48 48 52 53 55 57 61 62 67 69

Para esses dados temos o resumo de 5 números apresentados na Tabela 1.14.

Tabela 1.14: Resumo de 5 números para o número de laranjas por caixas.

Mediana Observada (md_{obs})	46
Primeiro Quartil (Q_1)	36, 50
Terceiro Quartil (Q_3)	55, 50
Mínimo ($x_{(1)}$)	22
Máximo ($x_{(20)}$)	69

Na Figura 1.5 é apresentado para esses dados o box-plot com base no resumo de 5 números.

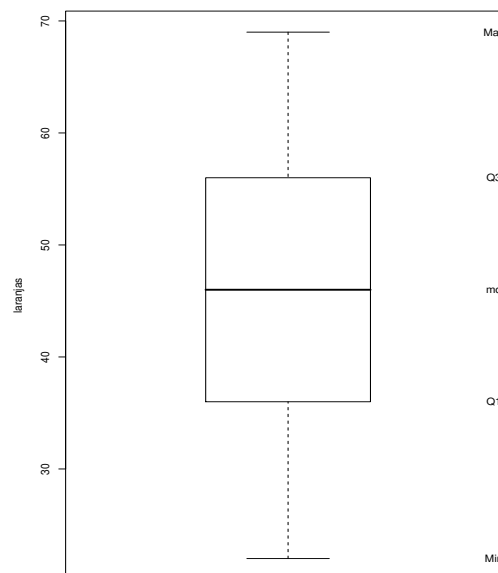


Figura 1.5: Box-plot do número de laranjas nas 20 caixas.

A representação gráfica no box-plot informa, dentre outras coisas, a variabilidade e simetria dos dados. Na Figura 1.5, a distribuição dos dados está muito próxima da perfeita simetria pois: a diferença entre a mediana(46) e a média(46,55) é pequena e a distância da mediana para os quartis é a mesma.

Outra possibilidade na construção do box-plot é utilizar nas extremidades dos traços adjacentes à caixa um critério para identificar observações atípicas. Este critério é baseado na amplitude interquartis($AIQ = Q_3 - Q_1$). A esquematização que utiliza este critério é apresentada na Figura 1.6.

No exemplo das laranjas, não há valores fora destes limites e, quando isto ocorre, os limites são representados pelo mínimo e máximo conforme a Figura 1.4.

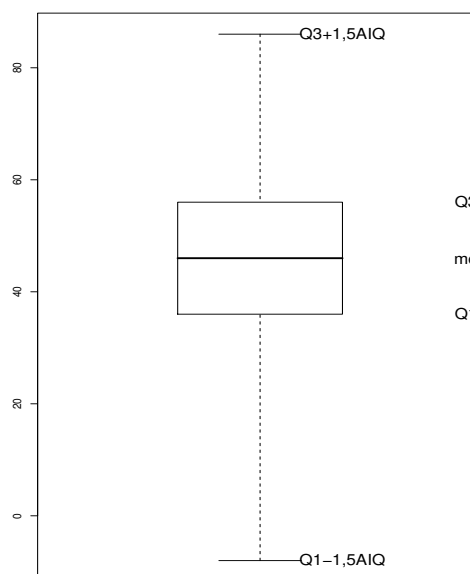


Figura 1.6: Desenho esquemático do box-plot com base nos quartis e critério para valores atípicos.

O box-plot pode também ser utilizado como ferramenta de análise bivariada. O exemplo na Figura 1.7 compara alturas de crianças dos sexos masculino e feminino. Os dados utilizados para elaboração dessa figura estão na tabela 1.15

Tabela 1.15: Alturas de crianças do sexo masculino (m) e feminino (f).

criança	altura	sexo	criança	altura	sexo
1	99.00	m	39	118.00	f
2	115.00	m	40	118.00	m
3	114.00	f	41	86.00	m
4	133.00	m	42	124.00	m
5	106.00	m	43	113.00	m
6	160.00	m	44	121.00	f
7	96.00	m	45	92.00	m
8	96.00	m	46	104.00	m
9	127.00	f	47	75.00	f
10	110.00	f	48	108.00	m
11	111.00	f	49	105.00	f
12	128.00	f	50	102.00	f
13	107.00	f	51	96.00	m
14	134.00	f	52	96.00	f
15	109.00	f	53	113.00	m
16	104.00	f	54	88.00	m
17	106.00	m	55	100.00	m
18	117.00	m	56	152.00	f
19	147.00	m	57	88.00	f
20	132.00	m	58	108.00	m
21	148.00	f	59	120.00	m
22	80.00	f	60	93.00	f
23	91.00	f	61	98.00	m
24	107.00	f	62	110.00	f
25	79.00	f	63	108.00	m
26	127.00	m	64	119.00	m
27	107.00	m	65	93.00	f
28	123.00	m	66	116.00	m
29	91.00	f	67	98.00	m
30	119.00	m	68	108.00	m
31	75.00	m	69	91.00	m
32	75.00	m	70	109.00	f
33	101.00	m	71	97.00	m
34	105.00	f	72	115.00	m
35	97.00	m	73	88.00	m
36	100.00	f	74	58.50	m
37	116.00	m	75	88.00	m
38	127.00	m	76	103.00	f

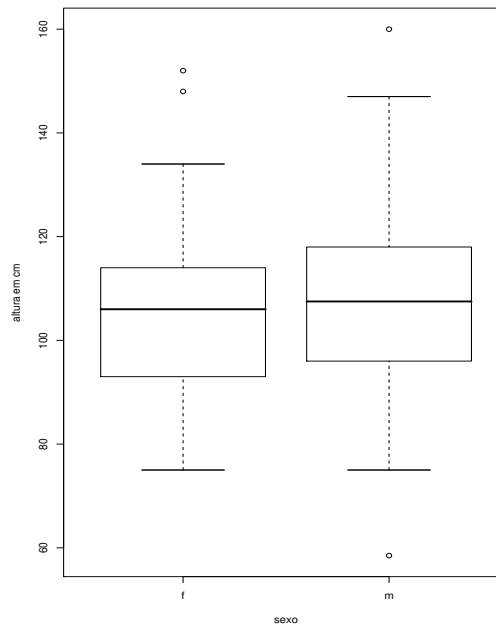


Figura 1.7: Altura de crianças conforme o sexo.

A partir do box-plot por categorias, pode-se constatar as diferenças nas tendências centrais dos grupos pelo posicionamento do traço central na caixa. Neste exemplo, a altura mediana dos meninos é ligeiramente maior do que a das meninas. Por outro lado, quando comparamos a variabilidade nos dois conjuntos, por meio dos tamanhos das caixas, não há evidência, sob o aspecto visual, de diferença entre os sexos feminino e masculino.

Outro aspecto que pode ser visto no gráfico é a existência de pontos discrepantes (*outliers*) nas alturas de meninos e meninas. É importante ressaltar que ao separar os valores de altura por sexo, podem surgir pontos discrepantes que não eram evidentes nos dados agregados, pois eles são identificados em relação a tendência central e variabilidade do grupo ao qual pertence. É bem provável que uma menina cuja altura é discrepante em relação às outras meninas, não se destaque em relação às alturas de todas as crianças pois, ao incluir meninos, a medida de tendência central é aumentada.

Gráfico seqüencial

As observações provenientes de uma série temporal em geral apresentam mudanças na média e variância ao longo do tempo que podem resultar de algum tipo de tendência no comportamento da variável. Este fato pode ser verificado descritivamente no gráfico seqüencial. A construção do gráfico consiste em plotar o par de valores (x, y) , em que x representa um índice de tempo (espaço) e y o valor observado da variável correspondente àquele índice.

A Figura 1.8 exhibe dados da variação mensal da taxa SELIC durante o período de 1995 até 2005. No gráfico podemos observar períodos de maior turbulência em que a taxa

SELIC apresenta maior variabilidade. Por este conjunto de dados apresentar aspectos típicos de uma série temporal, devemos tomar extremo cuidado ao calcular resumos numéricos pois estes podem variar em função do período de tempo em que são calculados.

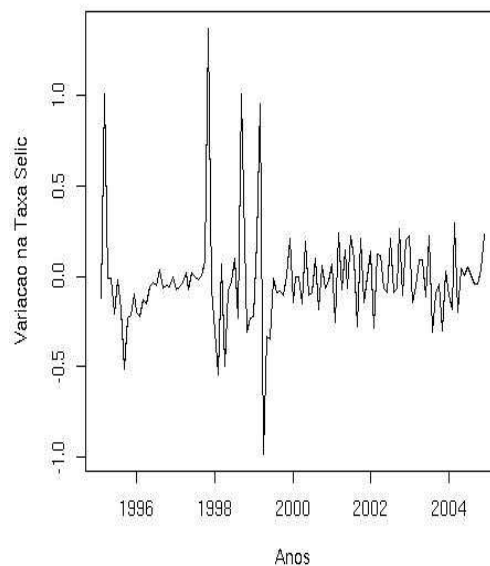


Figura 1.8: Variação mensal na Taxa Selic no período de 1995 a 2005.

Exemplo 1.22. Uma importante rede de lojas registra durante um ano a quantidade (em milhares) de eletrodomésticos vendidos.

mês	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
vendas de eletrodomésticos	25	23	17	14	11	13	9	10	11	9	20	22

A evolução das vendas ao longo do tempo é melhor entendida no gráfico seqüencial apresentado na Figura 1.9. Repare que após o mês de janeiro há uma tendência de queda que é revertida novamente nos últimos meses do ano.

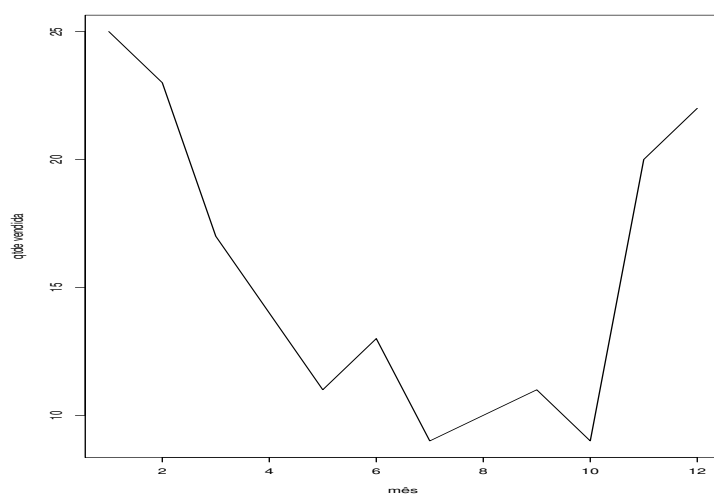


Figura 1.9: Gráfico sequencial das vendas ao longo dos meses.

Capítulo 2

Teoria das Probabilidades

2.1 Introdução

No capítulo anterior, foram mostrados alguns conceitos relacionados à estatística descritiva. Neste capítulo apresentamos a base teórica para o desenvolvimento de técnicas estatísticas a serem utilizadas nos capítulos posteriores.

Vamos considerar as seguintes questões: Como saber se um determinado produto está sendo produzido dentro dos padrões de qualidade? Como avaliar a capacidade de um determinado exame acertar o verdadeiro diagnóstico? Questões como estas envolvem algum tipo de variabilidade ou incerteza, e as decisões podem ser tomadas por meio da teoria de probabilidades que permite a quantificação da incerteza.

A seguir, veremos alguns conceitos básicos de probabilidade.

2.2 Conceitos Básicos de Probabilidade

- Fenômeno Aleatório: É um processo de coleta de dados em que os resultados possíveis são conhecidos mas não se sabe qual deles ocorrerá. Assim, um fenômeno aleatório pode ser a contagem de ausências de um funcionário em um determinado mês, o resultado do lançamento de uma moeda, verificar o resultado de um exame de sangue, entre outros.
- Espaço Amostral: O conjunto de todos os resultados possíveis do fenômeno aleatório é chamado de espaço amostral. Vamos representá-lo por Ω .

Exemplo 2.1. Lançamento de uma moeda. $\Omega = \{\text{cara, coroa}\}$.

Exemplo 2.2. Lançamento de um dado. $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Exemplo 2.3. Número de chips defeituosos em uma linha de produção durante 24 horas. $\Omega = \{0, 1, 2, 3, \dots, n\}$, sendo n o número máximo de itens defeituosos.

Exemplo 2.4. Tempo de reação de uma pomada anestésica aplicada em queimados. $\Omega = \{t \in \mathbb{R} \mid t \geq 0\}$.

- **Evento:** Qualquer subconjunto do espaço amostral Ω é chamado de evento. Serão representados por letras maiúsculas A, B, \dots . Dentre os eventos podemos considerar o evento união de A e B , denotado por $A \cup B$, que, equivale à ocorrência de A , ou de B , ou de ambos. A ocorrência simultânea dos eventos A e B , denotada por $A \cap B$ é chamada de evento interseção. Dois eventos A e B dizem-se mutuamente exclusivos ou disjuntos, quando a ocorrência de um deles impossibilita a ocorrência do outro. Os dois eventos não têm nenhum elemento em comum, isto é, $A \cap B = \emptyset$ (conjunto vazio).

Exemplo 2.5. Suponha um fenômeno aleatório conduzido com a finalidade de se conhecer a eficiência de uma terapia na cura de uma síndrome. Para tanto, dois pacientes foram tratados com a referida terapia. Vamos representar C e \overline{C} , como curado e não curado, respectivamente. O espaço amostral nesse caso é dado por:

$$\Omega = \{CC, C\overline{C}, \overline{C}C, \overline{C}\overline{C}\}.$$

Considere os seguintes eventos: A “obter uma cura” e B “obter quatro curas”: Sendo assim, temos:

$$A = \{C\overline{C}, \overline{C}C\}$$

e

$$B = \emptyset.$$

2.2.1 Definição clássica de probabilidade

Em fenômenos aleatórios tais como lançamento de uma moeda, de um dado, extração de uma carta de um baralho entre outros, temos que todos os resultados possíveis tem a mesma chance de ocorrer. Assim, por exemplo no lançamento de uma moeda a probabilidade do evento cara ou coroa ocorrer são igualmente prováveis, ou seja, a probabilidade atribuída a cada um é $1/2$.

A probabilidade de um evento A qualquer ocorrer pode ser definida por:

$$P(A) = \frac{\text{número de casos favoráveis ao evento } A}{\text{número de casos possíveis}}.$$

Exemplo 2.6. Considere o fenômeno aleatório lançamento de um dado e o evento A “sair número par”. Qual a probabilidade deste evento ocorrer?

$$P(A) = \frac{3}{6} = 0,50.$$

Na maioria das situações práticas, os resultados não têm a mesma chance de ocorrer, deste modo, a probabilidade dos eventos deve ser calculada pela frequência relativa.

2.2.2 Aproximação da Probabilidade pela frequência relativa

Quando não se tem conhecimento sobre as probabilidades dos eventos, estas podem ser atribuídas após repetidas observações do fenômeno aleatório, ou seja, a proporção de vezes que um evento A qualquer ocorre pode ser estimada como segue:

$$P(A) = \frac{\text{número de ocorrências do evento A}}{\text{número de observações}}.$$

Exemplo 2.7. Uma escola particular pretende oferecer um treinamento esportista aos seus alunos. Dos 300 alunos entrevistados, 142 optaram pelo voleibol, 123 indicaram o basquete e 35 indicaram o futebol. Selecionado aleatoriamente um desses alunos, qual a probabilidade de obter alguém que prefere o voleibol?

$$P(V) = \frac{142}{300} = 0,47333.$$

À medida que o número de observações aumenta, as aproximações tendem a ficar cada vez mais próximas da probabilidade efetiva.

2.2.3 Propriedades de probabilidades

É uma função $P(\cdot)$ que associa números reais aos elementos do espaço amostral e satisfaz as condições:

1. $0 \leq P(A) \leq 1$, para qualquer evento A;
2. $P(\Omega) = 1$;
3. $P(\emptyset) = 0$;
4. $P(\cup_{j=1}^n A_j) = \sum_{j=1}^n P(A_j)$.

Se \bar{A} for o evento complementar de A, então $P(A) = 1 - P(\bar{A})$.

2.2.4 Teorema da soma

Dado dois eventos A e B, a probabilidade de pelo menos um deles ocorrer é igual a soma das probabilidades de cada um menos a probabilidade de ambos ocorrerem simultaneamente, ou seja:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Se A e B forem mutuamente exclusivos, teremos $P(A \cap B) = 0$. Assim,

$$P(A \cup B) = P(A) + P(B).$$

Exemplo 2.8. Considere o experimento lançamento de um dado e os seguintes eventos:

$A = \{\text{sair número } 5\}$,

$B = \{\text{sair número par}\}$ e

$C = \{\text{sair número ímpar}\}$.

Determinar: Ω , $P(A)$, $P(B)$, $P(C)$, $P(A \cup B)$, $P(A \cup C)$ e $P(\overline{A})$.

$$\begin{aligned}\Omega &= \{1, 2, 3, 4, 5, 6\}. \\ P(A) &= \frac{1}{6}. \\ P(B) &= \frac{3}{6}. \\ P(C) &= \frac{3}{6}. \\ P(A \cup B) &= \frac{1}{6} + \frac{3}{6} = \frac{4}{6}. \\ P(A \cup C) &= \frac{1}{6} + \frac{3}{6} - \frac{1}{6} = \frac{3}{6}. \\ P(\overline{A}) &= 1 - \frac{1}{6} = \frac{5}{6}.\end{aligned}$$

Exemplo 2.9. Um estudo realizado por uma empresa de recursos humanos mostrou que 45% dos funcionários de uma multinacional saíram da empresa porque estavam insatisfeitos com seus salários, 28% porque consideraram que a empresa não possibilitava o crescimento profissional e 8% indicaram insatisfação tanto com o salário como com sua impossibilidade de crescimento profissional. Considere o evento S: “o funcionário sai da empresa em razão do salário” e o evento I: “o funcionário sai da empresa em razão da impossibilidade de crescimento profissional”. Qual é a probabilidade de um funcionário sair desta empresa devido a insatisfação com o salário ou insatisfação com sua impossibilidade de crescimento profissional?

$$\begin{aligned}P(S \cup I) &= P(S) + P(I) - P(S \cap I) \\ P(S \cup I) &= 0,45 + 0,28 - 0,08 = 0,65.\end{aligned}$$

2.2.5 Probabilidade condicional

Existem situações em que a chance de um particular evento acontecer depende do resultado de outro evento. A probabilidade condicional de A dado que ocorreu B pode ser determinada dividindo-se a probabilidade de ocorrência de ambos os eventos A e B pela probabilidade do evento B; como se mostra a seguir:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0.$$

Exemplo 2.10. Em uma universidade foi selecionada uma amostra de 500 alunos que cursaram a disciplina de Estatística. Entre as questões levantadas estava: Você gostou da disciplina de Estatística? De 240 homens, 140 responderam que sim. De 260 mulheres, 200 responderam que sim. Para avaliar as probabilidades podemos organizar as informações em uma tabela. maneira:

Tabela 2.1: Gosto pela disciplina de estatística segundo sexo.

Sexo	Gostou		Total
	Sim	Não	
Homem	140	100	240
Mulher	200	60	260
Total	340	160	500

Qual é a probabilidade de que um aluno escolhido aleatoriamente:

(a) H = Seja um homem?

$$P(H) = \frac{240}{500} = 0,48.$$

(b) G = Gostou da disciplina de Estatística?

$$P(G) = \frac{340}{500} = 0,68.$$

(c) M = Seja uma mulher?

$$P(M) = \frac{260}{500} = 0,52.$$

(d) NG = Não gostou da disciplina de Estatística?

$$P(NG) = \frac{160}{500} = 0,32.$$

(e) Seja uma mulher ou gostou da disciplina de Estatística.

$$P(M \cup G) = \frac{260}{500} + \frac{340}{500} - \frac{200}{500} = 0,80.$$

(f) Seja uma mulher e gostou da disciplina de Estatística.

$$P(M \cap G) = \frac{200}{500} = 0,40.$$

(g) Dado que o aluno escolhido gostou da disciplina de Estatística. Qual a probabilidade de que o aluno seja um homem?

$$P(H | G) = \frac{P(H \cap G)}{P(G)} = \frac{140}{340} = 0,41176.$$

(h) Dado que o aluno escolhido é uma mulher. Qual a probabilidade de que ela não gostou da disciplina de Estatística?

$$P(NG | M) = \frac{P(NG \cap M)}{P(M)} = \frac{60}{260} = 0,23077.$$

2.2.6 Teorema do produto

Da definição de probabilidade condicional $P(A|B) = \frac{P(A \cap B)}{P(B)}$ podemos obter o teorema do produto, que nos permite calcular a probabilidade da ocorrência simultânea de dois eventos. Sejam A e B eventos de Ω , a probabilidade de A e B ocorrerem juntos é dada por:

$$P(A \cap B) = P(A) P(B|A), \text{ com } P(A) > 0$$

ou

$$P(A \cap B) = P(B) P(A|B), \text{ com } P(B) > 0.$$

Dois eventos A e B são independentes quando a ocorrência de um não altera a probabilidade de ocorrência do outro. Desse modo,

$$P(A \cap B) = P(A) P(B).$$

Exemplo 2.11. Uma empresária sabe por experiência, que 65% das mulheres que compram em sua loja preferem sandálias plataformas. Qual é a probabilidade de as duas próximas clientes comprarem cada uma delas, uma sandália plataforma? Vamos admitir que o evento A “a primeira cliente compra uma sandália plataforma” e o evento B “a segunda cliente compra uma sandália plataforma”. Então,

$$P(A \cap B) = (0,65)(0,65) = 0,4225.$$

2.2.7 Teorema da probabilidade total

Suponha que os eventos C_1, C_2, \dots, C_n formam uma partição do espaço amostral. Os eventos não têm interseções entre si e a união destes é igual ao espaço amostral. Seja A um evento qualquer desse espaço, então a probabilidade de ocorrência desse evento será dada por:

$$P(A) = P(A \cap C_1) + P(A \cap C_2) + \dots + P(A \cap C_n)$$

e usando a definição de probabilidade condicional,

$$P(A) = P(C_1) P(A|C_1) + P(C_2) P(A|C_2) + \dots + P(C_n) P(A|C_n).$$

Exemplo 2.12. Uma caixa I contém 2 fichas verdes e 3 vermelhas. Uma segunda caixa II contém 4 fichas verdes e 3 vermelhas. Escolhe-se, ao acaso, uma caixa e dela retira-se, também ao acaso uma ficha. Qual a probabilidade de que a ficha retirada seja verde? Se denotarmos por I e II o evento caixa I e caixa II, respectivamente e V o evento a ficha é verde, temos: $P(I) = \frac{1}{2}$, $P(V|I) = \frac{2}{5}$, $P(II) = \frac{1}{2}$ e $P(V|II) = \frac{4}{7}$.

Desta forma, o evento V (“ficha verde”) pode ser escrito em termos de interseções do evento V com os eventos I e II,

$$V = (V \cap I) \cup (V \cap II)$$

$$\begin{aligned} P(V) &= (P(V \cap I)) + (P(V \cap II)) \\ &= P(I)(P(V|I)) + (P(II)P(V|II)) \\ &= \frac{1}{2} \frac{2}{5} + \frac{1}{2} \frac{4}{7} = 0,48571. \end{aligned}$$

Exemplo 2.13. Um estabilizador pode provir de três fabricantes I, II e III com probabilidades de 0,25, 0,35 e 0,40, respectivamente. As probabilidades de que durante determinado período de tempo, o estabilizador não funcione bem são, respectivamente, 0,10; 0,05 e 0,08 para cada um dos fabricantes. Qual é a probabilidade de que um estabilizador escolhido ao acaso não funcione bem durante o período de tempo especificado. Se denotarmos por A o evento “um estabilizador não funcione bem” e por C_1 , C_2 e C_3 os eventos “um estabilizador vem do fabricante I, II e III”, respectivamente. A probabilidade de que um estabilizador escolhido ao acaso não funcione bem durante o período de tempo especificado é:

$$\begin{aligned} P(A) &= P(C_1)P(A|C_1) + P(C_2)P(A|C_2) + P(C_3)P(A|C_3) \\ &= (0,25)(0,10) + (0,35)(0,05) + (0,40)(0,08) = 0,07450. \end{aligned}$$

2.2.8 Teorema de Bayes

Considere C_1, C_2, \dots, C_n eventos que formam uma partição do espaço amostral Ω , cujas probabilidades são conhecidas. Considere que para um evento A se conheçam as probabilidades condicionais, desta forma:

$$P(C_j|A) = \frac{P(C_j) P(A|C_j)}{P(C_1) P(A|C_1) + P(C_2) P(A|C_2) + \dots + P(C_n) P(A|C_n)}, \quad j = 1, 2, \dots, n.$$

Exemplo 2.14. Considere o exemplo anterior para o desenvolvimento do Teorema de Bayes. Dado que o estabilizador escolhido ao acaso não funciona bem durante o período de tempo especificado, qual a probabilidade de que tenha sido produzido pelo fabricante I, isto é, $P(C_1|A)$.

$$\begin{aligned} P(C_1|A) &= \frac{P(C_1) P(A|C_1)}{P(C_1)P(A|C_1) + P(C_2)P(A|C_2) + P(C_3)P(A|C_3)} \\ &= \frac{(0,25)(0,10)}{0,07450} = 0,33557. \end{aligned}$$

Capítulo 3

Variáveis Aleatórias

3.1 Introdução

Neste capítulo vamos dar continuidade ao estudo de probabilidades, introduzindo os conceitos de variáveis aleatórias e de distribuições de probabilidade.

Uma variável aleatória (v.a.) associa um valor numérico a cada resultado de um fenômeno aleatório e uma distribuição de probabilidade associa uma probabilidade a cada valor de uma variável aleatória. Como exemplos, podemos citar: o número de usuários que consultam um site de busca, em um certo tempo do dia, o número de atletas com lesões traumáticas no joelho, o número de ovos de codorna incubados durante um período, o tempo de uso de uma máquina agrícola, a pressão sanguínea de mulheres na menopausa, dentre outros. Uma v.a. pode ser classificada como variável aleatória discreta ou variável aleatória contínua.

3.2 Variável Aleatória Discreta

Uma variável aleatória que pode assumir um número finito de valores ou uma quantidade enumerável de valores, cujas probabilidades de ocorrência são conhecidas é denominada variável aleatória discreta.

A função que atribui a cada valor da variável aleatória sua probabilidade é denominada função discreta de probabilidade ou função de probabilidade, isto é:

$$P(X = x_i) = p(x_i) = p_i, \quad i = 1, 2, \dots$$

Uma função de probabilidade satisfaz as duas condições seguintes:

- i) $0 \leq p_i \leq 1$;
- ii) $\sum p_i = 1$.

Exemplo 3.1. Um pediatra de um hospital público constatou que das crianças internadas durante um ano, 20% não tiveram internamento por infecção da faringe, 45% tiveram um internamento por infecção da faringe, 25% tiveram dois internamentos por infecção da

faringe, 9% tiveram três internamentos por infecção da faringe e 1% tiveram quatro internamentos por infecção da faringe. Seja X uma variável aleatória discreta que representa o número de internamentos por infecção da faringe. Logo, a função de probabilidade para X é dada na tabela a seguir:

X	0	1	2	3	4
p_i	0,20	0,45	0,25	0,09	0,01

No exemplo temos: $0,20 + 0,45 + 0,25 + 0,09 + 0,01 = 1$.

3.3 Variável Aleatória Contínua

Uma variável aleatória que pode tomar um número infinito de valores, e esses valores podem ser associados a mensurações em uma escala contínua e as probabilidades necessárias ao seu estudo são calculadas como a área abaixo da curva da distribuição, é chamada de função de densidade de probabilidade. Como exemplos de variáveis aleatórias contínuas, podemos citar: o tempo de uso de um equipamento eletrônico; medidas do tórax, diâmetro de um cabo de vídeo, tempo de atendimento a clientes; altura, peso, etc.

Uma função de densidade de probabilidade $f(x)$ pode ser usada para descrever a distribuição de probabilidades de uma variável aleatória contínua X , se satisfaz:

- i) $f(x) \geq 0$, para todo $x \in (-\infty, \infty)$;
- ii) $\int_{-\infty}^{\infty} f(x)dx = 1$.

3.4 Esperança Matemática

Seja X uma variável aleatória. A esperança matemática, média ou valor esperado de X é definida por:

$$E(X) = \mu = \sum_{i=1}^{\infty} x_i P(x_i), \text{ se } X \text{ for discreta.}$$

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx, \text{ se } X \text{ for contínua.}$$

Exemplo 3.2. As probabilidades de um corretor de imóveis vender uma sala comercial com lucro de R\$ 3500,00, de R\$ 2500,00, R\$ 800,00 ou com um prejuízo de R\$ 500,00 são 0,20, 0,35, 0,22 e 0,10, respectivamente. Qual é o lucro esperado do corretor de imóveis? Sendo

$$x_1 = 3500, x_2 = 2500, x_3 = 800 \text{ e } x_4 = -500$$

e

$$p_1 = 0,20, p_2 = 0,35, p_3 = 0,22 \text{ e } p_4 = 0,10$$

tem-se $E(X) = 3500(0,20) + 2500(0,35) + 800(0,22) - 500(0,10) = 1801$

3.5 Variância

A variância de uma variável aleatória X é dada por:

$$\text{Var}(X) = \sigma^2 = E(X)^2 - [E(X)]^2.$$

3.6 Principais Distribuições de Probabilidades

A distribuição de probabilidade de uma variável descreve como as probabilidades estão distribuídas sobre os valores da variável aleatória. A seguir veremos as distribuições de probabilidade Bernoulli, Binomial e Poisson para variáveis aleatórias discretas e a distribuição Normal para uma variável aleatória contínua.

3.6.1 Distribuição de Bernoulli

Em situações em que o fenômeno aleatório é realizado uma só vez e a variável de interesse assume somente dois valores, tais como: um gestor de informação reconhece uma determinada editora ou não, um paciente sobrevive a um transplante de medula óssea ou não, um equipamento eletrônico é classificado como bom ou defeituoso. Estas situações têm alternativas dicotômicas, ou seja, podem ser representadas por respostas do tipo sucesso com probabilidade p que se atribui o valor 1 ou fracasso com probabilidade q que se atribui o valor 0. Podemos definir estes experimentos como ensaios de Bernoulli.

Uma variável X tem distribuição de Bernoulli e sua função discreta de probabilidade é dada por:

$$P(X = x) = p^x q^{1-x}, \quad x = 0, 1.$$

Exemplo 3.3. Uma caixa tem 20 bolas azuis e 30 verdes. Retira-se uma bola dessa caixa. Seja X o número de bolas verdes. Determinar $P(X)$.

Para $x = 0$ temos $q = \frac{20}{50} = 0,4$ e para $x = 1$, $p = \frac{30}{50} = 0,6$.

Logo, $P(X = x) = 0,6^x 0,4^{1-x}$.

3.6.2 Distribuição Binomial

Consideremos n tentativas independentes de ensaios de Bernoulli. Cada tentativa admite apenas dois resultados complementares: sucesso com probabilidade p ou fracasso com probabilidade q , de modo a se ter $p + q = 1$. As probabilidades de sucesso e fracasso são as mesmas para cada tentativa. A variável aleatória X que conta o número total de sucessos é denominada Binomial.

Para indicar que a variável aleatória X segue o modelo Binomial, usaremos a notação $X \sim b(n, p)$, em que n e p são denominados parâmetros dessa distribuição. A sua função de probabilidade é dada por:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n,$$

em que

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

sendo

n = número de tentativas,

x = número de sucessos,

p = probabilidade de sucesso,

q = probabilidade de fracasso e

$P(x)$ = a probabilidade de se obter exatamente x sucessos em n provas.

Para uma variável aleatória X com distribuição binomial a média e sua variância são dadas, respectivamente, por:

$$\mu = E(X) = np; \text{ e } \sigma^2 = npq.$$

Exemplo 3.4. Uma moeda é lançada 6 vezes. Qual a probabilidade de:

a) Exatamente duas caras ocorrerem?

$$P(X = 2) = \binom{6}{2} 0,5^2 0,5^{6-2} = 0,23438.$$

b) Ocorrerem pelo menos 4 caras?

$$\begin{aligned} P(X \geq 4) &= P(X = 4) + P(X = 5) + P(X = 6) \\ &= \binom{6}{4} 0,5^4 0,5^{6-4} + \binom{6}{5} 0,5^5 0,5^{6-5} + \binom{6}{6} 0,5^6 0,5^{6-6} \\ &= 0,23438 + 0,09375 + 0,01563 = 0,34375. \end{aligned}$$

c) Pelo menos 1 cara?

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ P(X \geq 1) &= \binom{6}{0} 0,5^0 0,5^{6-0} \\ P(X \geq 1) &= 1 - 0,01563 = 0,98438. \end{aligned}$$

Exemplo 3.5. Num município, há uma probabilidade de 0,70 de uma empresa de materiais recicláveis ter seguro contra incêndio; qual a probabilidade de que, dentre cinco empresas:

a) Nenhuma tenha seguro contra incêndio?

$$P(X = 0) = \binom{5}{0} 0,7^0 0,3^{5-0} = 0,00243.$$

b) Exatamente quatro tenham seguro contra incêndio?

$$P(X = 4) = \binom{5}{4} 0,7^4 0,3^{5-4} = 0,36015.$$

3.6.3 Distribuição de Poisson

Como aplicações da distribuição de Poisson podemos citar estudos de acidentes com veículos; número de mortes por derrame cerebral por ano, numa cidade, número de reclamações que chegam em uma operadora telefônica por hora, número de clientes que chegam numa loja durante uma hora de promoção relâmpago e número de usuários de computador ligados à Internet.

A distribuição de Poisson é uma distribuição discreta de probabilidade, aplicável a ocorrência de um evento em um intervalo especificado (tempo, distância, área, volume ou outra unidade análoga). A probabilidade do evento ocorrer x vezes em um intervalo é dada a seguir:

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

em que λ é a média ($\lambda = np$) ou o número esperado de ocorrências num determinado intervalo de tempo, por exemplo. Utilizaremos a seguinte notação: $X \sim \text{Po}(\lambda)$.

Exemplo 3.6. O número de mulheres que entram diariamente em uma clínica de estética para bronzamento artificial apresenta distribuição de Poisson, com média de 5 mulheres por dia. Qual é a probabilidade de que em um dia particular, o número de mulheres que entram nesta clínica de estética para bronzamento artificial, seja:

a) Igual a 2?

$$P(X = 2) = \frac{5^2 e^{-5}}{2!} = 0,08422.$$

b) Inferior ou igual a 2?

$$P(X \leq 2) = \frac{5^0 e^{-5}}{0!} + \frac{5^1 e^{-5}}{1!} + \frac{5^2 e^{-5}}{2!} = 0,12465.$$

3.6.4 Distribuição Normal

A distribuição Normal é a mais importante distribuição de probabilidade para descrever variáveis aleatórias contínuas. Isto justifica-se pelo grande número de aplicações que a utilizam tais como, altura, pressão arterial, medidas de testes psicológicos, tempo de vida útil de um dispositivo eletrônico, temperatura corporal, dentre outras. Além disso, pela sua capacidade de aproximar outras distribuições e também pela grande aplicação na inferência estatística.

A variável aleatória contínua X com distribuição Normal tem função de densidade de probabilidade dada por:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ para } -\infty < x < \infty$$

em que os parâmetros μ e σ representam a média e o desvio padrão, respectivamente.

Usaremos a notação $X \sim N(\mu, \sigma^2)$, para indicar que a variável aleatória X tem distribuição Normal com parâmetros μ e σ^2 .

Algumas características da distribuição normal são:

- a curva normal tem forma de sino, é simétrica em relação à média, como representada na Figura 3.1;
- a média, mediana e moda são valores coincidentes;
- a variável aleatória X associada a sua distribuição varia de $-\infty < x < \infty$;
- a função $f(x)$ tem ponto máximo em $x = \mu$.

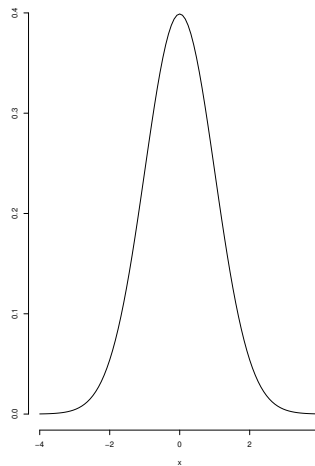


Figura 3.1: Densidade Normal.

Para o cálculo das probabilidades, surgem dois problemas: primeiro, a integração de $f(x)$, pois para a sua resolução é necessário o desenvolvimento em séries; segundo, seria a elaboração de uma tabela de probabilidades, pois $f(x)$ depende das combinações da média e variância. Para resolver esses problemas, optou-se por uma mudança de variável obtendo-se, assim, a distribuição normal padronizada que chamamos aqui de Z .

3.6.5 Distribuição Normal Padrão

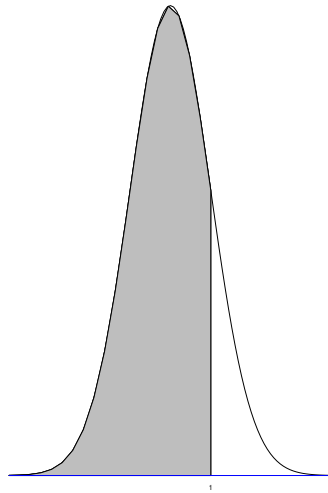
Denomina-se distribuição Normal Padrão a distribuição Normal de média zero e variância 1, usaremos a notação $Z \sim N(0, 1)$. As probabilidades associadas à distribuição normal padrão são apresentadas em tabelas. A tabela da normal padrão que iremos utilizar fornece a probabilidade de Z tomar um valor não superior a z , está apresentada no Apêndice.

3.6.6 Uso da tabela da Normal Padrão

Com o uso da tabela da distribuição Normal Padrão pode-se encontrar as probabilidades de Z . Vejamos alguns exemplos:

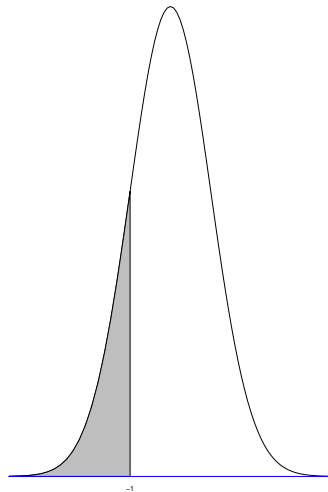
Exemplo 3.7. $P(Z \leq 1)$.

A probabilidade que está sendo pedida está representada na área sombreada sob a curva normal exibida na figura abaixo. Portanto, com auxílio da tabela chega-se ao resultado 0,84134.



Exemplo 3.8. $P(Z \leq -1)$.

Neste caso a probabilidade que está sendo pedida está a esquerda de -1. Portanto, com auxílio da tabela chega-se ao resultado 0,15865.



Exemplo 3.9. $P(Z \leq 1,72) = 0,95728$.

Exemplo 3.10. $P(Z \leq -0,53) = 0,29805$.

Exemplo 3.11. $P(-1 \leq Z \leq 1) = 0,84134 - 0,15865 = 0,68269$.

Exemplo 3.12. $P(0,7 \leq Z \leq 1,35) = 0,91149 - 0,75803 = 0,15346$.

Exemplo 3.13. $P(Z \geq 1,8) = 0,03593$.

Os problemas da vida real, entretanto, não se apresentam já na forma reduzida, ao contrário, são formulados em termos da variável normal original X , com média μ e desvio padrão σ . É preciso então, antes de passarmos à sua resolução, padronizamos ou reduzimos a v.a. Normal X , transformando-a na v.a. Z .

O resultado da padronização é a obtenção de uma escala de distribuição denominada escala reduzida, escore Z , que mede o afastamento das variáveis em relação à média em número de desvios-padrão.

$$Z = \frac{x - \mu}{\sigma},$$

em que:

Z representa o número de desvios-padrão a contar da média;

x representa um valor qualquer da variável aleatória;

μ = média da distribuição;

σ = desvio padrão da distribuição.

Exemplo 3.14. As vendas diárias de um confeitaria no centro de uma cidade têm distribuição normal, com média igual R\$ 450,00 por dia e desvio padrão igual a R\$ 95,00. Qual é a probabilidade das vendas excederem R\$ 700,00 em determinado dia?

Sendo $Z = \frac{700-450}{95} = 2,63$ e utilizando a tabela da Normal padrão, encontramos a probabilidade de 0,00426 para o valor de $Z = 2,63$.

Exemplo 3.15. Suponha que entre pacientes o nível de colesterol tenha uma distribuição aproximadamente Normal de média 105 mg por 100 ml e um desvio padrão 9 mg por 100 ml. Qual a proporção de diabéticos que tem níveis entre 90 e 125 mg por 100 ml?

Como $Z = \frac{90-105}{9} = -1,67$ e $Z = \frac{125-105}{9} = 2,22$ temos a partir da tabela da Normal padrão as seguintes probabilidades: 0,98679 para o valor de $Z = 2,22$ e 0,04745 para o valor de $Z = -1,67$. Desse modo, a proporção de diabéticos é 93,93.

Capítulo 4

Inferência Estatística - Teoria da Estimação

4.1 Introdução

Neste capítulo abordaremos situações em que o interesse está em obter informações da população a partir dos resultados de uma amostra. Como exemplo, consideremos uma indústria de produtos de cabelo que pretende investigar a aceitação, entre as mulheres, de seu novo produto tonalizante. Para tanto, selecionamos uma amostra aleatória de mulheres que utilizaram o produto e verificamos qual é o percentual de aprovação desse produto na amostra. Outro exemplo trata-se de um psiquiatra interessado em determinar o tempo de reação de um medicamento anti-depressivo em crianças. Uma amostra aleatória de crianças que utilizaram o medicamento é obtida e analisamos o tempo médio de reação. Nestes dois exemplos, deseja-se determinar o valor de um parâmetro por meio da avaliação de uma amostra.

A seguir vamos definir alguns conceitos básicos de inferência estatística.

- **Parâmetro:** é uma medida numérica, em geral desconhecida, que descreve uma característica de interesse da população. São representados, geralmente, por letras gregas tais como, μ (média populacional), σ (desvio-padrão populacional), entre outros. Neste texto, usaremos a letra p para representar a proporção populacional.
- **Estatística:** é qualquer valor calculado a partir dos dados amostrais. Por exemplo, \bar{X} (média amostral), S (desvio-padrão amostral) e \hat{p} (proporção amostral). A estatística é uma variável aleatória, por dois motivos: porque é uma quantidade incerta (antes de obter a amostra não sabemos seu valor) e porque seu valor varia de amostra para amostra. É claro que, quando uma amostra é selecionada e uma estatística é calculada, torna-se então uma constante, ou seja, é o resultado da observação de uma variável aleatória.
- **Estimador e Estimativa:** uma estatística destinada a estimar um parâmetro é cha-

mada estimador. Dada uma amostra, o valor assumido pelo estimador é chamado de estimativa ou valor estimado do parâmetro. As estimativas obtidas por meio da estatística variam de acordo com a amostra selecionada.

Os procedimentos básicos de inferência estatística compreendem duas metodologias. Uma é chamada de estimação, na qual nós usamos o resultado amostral para estimar o valor desconhecido do parâmetro; a outra é conhecida como teste de hipóteses, em que nós usamos o resultado amostral para avaliar se uma afirmação sobre o parâmetro (uma hipótese) é sustentável ou não. Teoria de estimação é o assunto principal deste capítulo e teste de hipóteses será retomado no próximo capítulo.

Para motivação, consideremos uma população formada por 5 alunos. Temos informações sobre a idade e sexo dos alunos na Tabela 4.1.

Tabela 4.1: População de alunos.

Identificação dos alunos	Idade em anos completos	Sexo feminino(f); masculino(m)
1	22	f
2	19	f
3	19	m
4	20	f
5	22	m

A população de alunos tem, em média, $\mu = 20,4$ anos com variância $\sigma^2 = 1,84$ anos² (desvio-padrão $\sigma = 1,36$ anos, aproximadamente). Verificamos que 40% dos alunos são homens, ou seja, a proporção de homens é $p = 0,40$. A idade média, o desvio-padrão de idade e a proporção de homens descrevem a população de alunos, portanto são parâmetros.

Embora tenhamos acesso a todos os dados de sexo e idade dos 5 alunos, vamos recorrer a amostragem para estimar μ por \bar{X} , a idade média amostral, e p por \hat{p} , a proporção amostral de homens.

Na Tabela 4.2 estão relacionadas todas as amostras possíveis de tamanho dois, que podem ser selecionadas da população dos 5 alunos, por amostragem aleatória simples com reposição. São $5^2 = 25$ possíveis amostras. Para cada amostra i , podemos calcular \bar{X}_i , uma estimativa para a idade média e \hat{p}_i , uma estimativa para a proporção de homens. Temos estimativas variadas para μ e p , e é o que ocorre quando tiramos várias amostras de uma população. Nenhuma das estimativas coincide com o valor do parâmetro. Por exemplo, a amostra $i = 4$, formada pelos alunos 1 e 4, apresenta uma superestimativa para a idade média, $\bar{X}_4 = 21$ anos, e subestima a proporção de homens, $\hat{p}_4 = 0,0$. Já na amostra $i = 8$, em que foram selecionados os alunos 2 e 3, a idade média é subestimada em $\bar{X}_8 = 19$ anos, e a proporção de homens é superestimada em $\hat{p}_8 = 0,5$. Nesta ilustração, estamos medindo o erro das estimativas, pois o valor do parâmetro é conhecido, situação que não acontece nos problemas reais de estimação.

Tabela 4.2: Todas as possíveis amostras aleatórias simples com reposição de tamanho 2, da população de alunos.

Amostra i	Alunos selecionados	Dados amostrais	\bar{X}_i	\hat{p}_i
1	1 e 1	22 f, 22 f	22,0	0,0
2	1 e 2	22 f, 19 f	20,5	0,0
3	1 e 3	22 f, 19 m	20,5	0,5
4	1 e 4	22 f, 20 f	21,0	0,0
5	1 e 5	22 f, 22 m	22,0	0,5
6	2 e 1	19 f, 22 f	20,5	0,0
7	2 e 2	19 f, 19 f	19,0	0,0
8	2 e 3	19 f, 19 m	19,0	0,5
9	2 e 4	19 f, 20 f	19,5	0,0
10	2 e 5	19 f, 22 m	20,5	0,5
11	3 e 1	19 m, 22 f	20,5	0,5
12	3 e 2	19 m, 19 f	19,0	0,5
13	3 e 3	19 m, 19 m	19,0	1,0
14	3 e 4	19 m, 20 f	19,5	0,5
15	3 e 5	19 m, 22 m	20,5	1,0
16	4 e 1	20 f, 22 f	21,0	0,0
17	4 e 2	20 f, 19 f	19,5	0,0
18	4 e 3	20 f, 19 m	19,5	0,5
19	4 e 4	20 f, 20 f	20,0	0,0
20	4 e 5	20 f, 22 m	21,0	0,5
21	5 e 1	22 m, 22 f	22,0	0,5
22	5 e 2	22 m, 19 f	20,5	0,5
23	5 e 3	22 m, 19 m	20,5	1,0
24	5 e 4	22 m, 20 f	21,0	0,5
25	5 e 5	22 m, 22 m	22,0	1,0

Com os resultados da Tabela 4.2 podemos calcular a média dos estimadores \bar{X} e \hat{p} .
Média de \bar{X} :

$$\frac{\sum_{i=1}^{15} \bar{X}_i}{15} = 20,4 = \mu$$

Média de \hat{p} :

$$\frac{\sum_{i=1}^{15} \hat{p}_i}{15} = 0,4 = p$$

Não foi por acaso que as médias de \bar{X} e \hat{p} coincidiram com os valores dos correspondentes parâmetros. Podemos demonstrar que a média dessas estimativas é igual ao

parâmetro que está sendo estimado.

Para as variâncias de \bar{X} e \hat{p} , temos um outro resultado interessante. Denotando tamanho da amostra por n , podemos mostrar também que a variância de \bar{X} é:

$$\frac{\sigma^2}{n},$$

e a variância de \hat{p} é:

$$\frac{p(1-p)}{n}.$$

Verifique este resultados com os dados da Tabela 4.2. Nas fórmulas para a variância dos estimadores, n aparece no denominador, isto quer dizer que quanto maior o tamanho da amostra, menos dispersas serão as estimativas.

Quando obtemos todas as amostras de tamanho da população encontramos estimativas diferentes para o mesmo parâmetro, porém, em média, são iguais ao parâmetro; e tendem a ser mais homogêneas com o aumento do tamanho da amostra. Este resultado é tão surpreendente, que torna possível o uso de uma amostra para estimar os parâmetros da população.

Podemos propor vários estimadores para um determinado parâmetro. Para estimar, por exemplo, a média populacional μ da variável X , nós poderíamos usar a média amostral \bar{X} , a mediana amostral, ou a primeira observação X_1 , entre outras possibilidades. Alguns estimadores em potencial não tem sentido como X_1 , que considera a primeira observação como estimador de μ e despreza toda a informação proveniente das outras observações na amostra. Pode ser natural usar a estatística análoga para estimar o parâmetro, ou seja, usar a média amostral para estimar μ , mas esta estratégia nem sempre leva ao melhor estimador. Basicamente, um bom estimador tem uma média igual ao parâmetro sendo estimado e desvio padrão pequeno.

4.2 Propriedades dos Estimadores

- **Vício**

Um estimador é não viciado se sua média é igual ao parâmetro. A média amostral é um estimador não viciado da média populacional. Por outro lado, um estimador viciado, em média, tende a subestimar ou sobrestimar o parâmetro. As vezes, pode ser interessante usar estimadores viciados, com vícios que tendem a desaparecer quando o tamanho da amostra aumenta.

- **Eficiência**

Uma segunda propriedade interessante para um estimador é ter um erro padrão pequeno, comparado a outros estimadores. Um estimador com essa propriedade é dito ser eficiente.

É desejável que o estimador de um parâmetro deva ser não viciado e eficiente. Vimos que a média amostral e a proporção amostral são estimadores não viciados para μ e p , respectivamente; e é possível mostrar que são estimadores eficientes. Por outro lado,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

é um estimador viciado de σ^2 e

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

é não viciado.

Uma boa maneira de visualizar as propriedades dos estimadores é fazer uma analogia com o jogo de dardos. Na Figura 4.1 estão esquematizados o desempenho de 4 jogadores, cada um com 8 dardos. Os dardos são as amostras e os jogadores representam 4 tipos de estimadores. O jogador da Figura 4.1a representa um bom estimador, pois os dardos estão em torno do alvo (não viciado) e bem concentrados (eficiente). Nas Figuras 4.1b a 4.1d os jogadores não tem um desempenho tão bom. Na Figura 4.1b está representado o estimador mais eficiente, comparando com os outros estimadores, mas tem vícios. Já o estimador caracterizado na Figura 4.1c não tem vícios porém, não é eficiente. O jogador da Figura 4.1d representa o pior dos 4 estimadores: é viciado e não pode ser considerado eficiente.

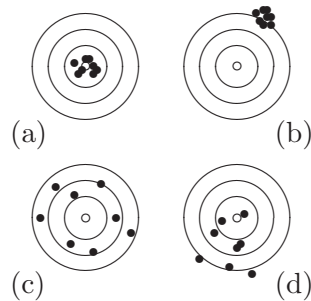


Figura 4.1: Analogia entre as propriedades dos estimadores e o jogo de dardos.

É importante estudar as propriedades do estimador para verificar se é um bom estimador para o parâmetro de interesse. Podemos também avaliar a qualidade de um estimador associando um erro máximo às estimativas. Este erro máximo é um limite que desejamos não ser ultrapassado pela estimativa. Para verificar se um estimador é bom, basta calcular a probabilidade do erro amostral não ultrapassar o máximo estipulado. Esperamos que essa probabilidade seja muito alta, perto de 100%.

Na seção anterior, vimos que as estatísticas e portanto os estimadores são variáveis aleatórias. A distribuição de probabilidades de uma estatística é conhecida como distribuição amostral e seu desvio-padrão é referido como erro padrão. Se conhecermos a distribuição amostral do estimador podemos calcular as probabilidades que precisamos para avaliar o estimador.

4.3 Distribuições Amostrais

4.3.1 Introdução

Uma forma de obter a distribuição amostral de um estimador é pensarmos em todas as amostras possíveis de tamanho n que podem ser retiradas da população, usando por exemplo, amostragem aleatória simples com reposição, como foi feito para a população de 5 alunos. Resumimos as informações sobre as estimativas da idade média na Tabela 4.2 com a distribuição de probabilidades para \bar{X} (Tabela 4.3).

Tabela 4.3: Distribuição amostral da idade média.

Possíveis valores de \bar{X}	Probabilidades
19,0	0,16
19,5	0,16
20,0	0,04
20,5	0,32
21,0	0,16
22,0	0,16
Total	1,00

Por que é tão importante conhecer a distribuição de \bar{X} ? Vamos lembrar que estamos procurando um bom estimador para μ , a média populacional. Alguém pode pensar que \bar{X} é um bom estimador de μ quando o erro amostral, $\bar{X} - \mu$, for pequeno. Mas como μ é desconhecido não é possível mensurar o erro amostral. Vamos por um limite para esse erro amostral e vamos denotá-lo por e . Se conhecermos a distribuição de \bar{X} é possível calcular a probabilidade do erro amostral ser no máximo igual a e . Já que o erro pode ser cometido para mais ou para menos, a probabilidade que devemos calcular é

$$P(-e \leq \bar{X} - \mu \leq e).$$

Se esta fosse a única forma de obter a distribuição amostral, o processo de inferência ficaria inviável para populações reais, pois é necessário obter todas as possíveis amostras de tamanho n para construir a distribuição amostral. Felizmente, pela teoria de probabilidades podemos mostrar que se uma variável X tem distribuição Normal, a média amostral, \bar{X} , também tem distribuição Normal.

4.3.2 Distribuição amostral de \bar{X}

Consideremos que uma amostra aleatória de tamanho n é selecionada da população de interesse, para observar a variável aleatória contínua, X , com distribuição Normal de média μ e desvio-padrão σ . A média amostral, \bar{X} , tem distribuição Normal com média μ

e desvio-padrão σ/\sqrt{n} . Tanto que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tem distribuição Normal Padrão.

A Figura 4.2 apresenta a forma da distribuição de X e de \bar{X} para diferentes tamanhos de amostra. Observamos que a distribuição de X na população é bem mais dispersa que a de \bar{X} ; e a medida que aumentamos o tamanho da amostra, a distribuição de \bar{X} vai ficando mais concentrada.

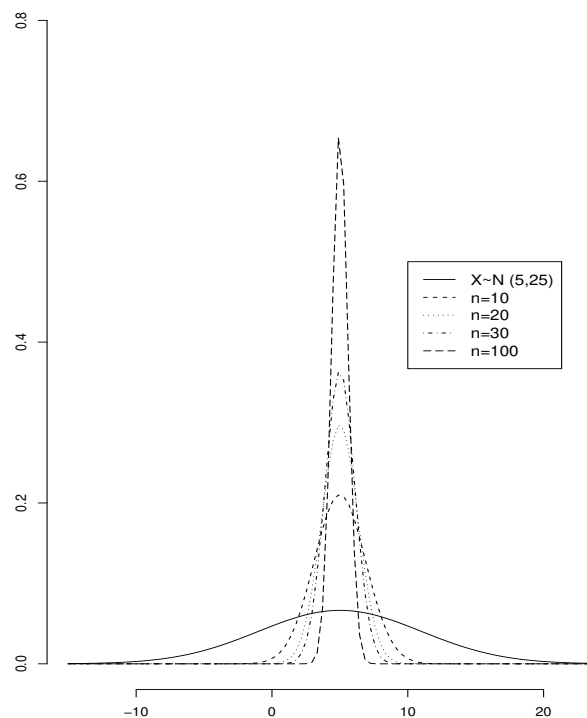


Figura 4.2: Distribuição de \bar{X} quando X tem distribuição normal, para alguns tamanhos de amostra.

Se σ é desconhecido, a teoria de probabilidades mostra que, mesmo assim, é possível obter uma distribuição amostral para \bar{X} , utilizando uma distribuição, denominada *t* de Student. Esta distribuição tem forma parecida com a da Normal Padrão, com caudas um pouco mais pesadas, ou seja, a dispersão da distribuição *t* de Student é maior. Se uma distribuição tem caudas mais pesadas, valores extremos têm maior probabilidade de ocorrerem.

Esta dispersão varia com o tamanho da amostra, sendo bastante dispersa para amostras pequenas, mas se aproximando da Normal Padrão para amostras grandes. A distribuição *t* de Student tem apenas um parâmetro, denominado graus de liberdade, *gl*.

Se uma variável aleatória X tem distribuição Normal com média μ , então

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tem distribuição t de Student com $(n - 1)$ graus de liberdade, sendo que S é o estimador de σ , o desvio-padrão de X .

Na Figura 4.3 estão representadas as densidades das distribuições de Z e T .

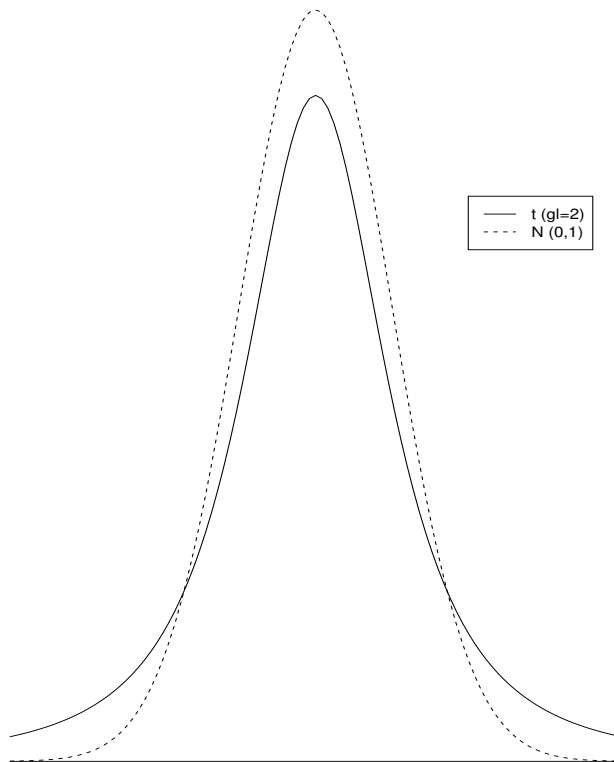


Figura 4.3: Densidades de T e Z .

Mesmo que a variável de interesse X não tenha uma distribuição Normal, ainda podemos obter uma distribuição aproximada para \bar{X} . A teoria de probabilidades fornece um teorema, que é um dos resultados mais importantes da Estatística, pois encontra uma aproximação para a distribuição amostral de \bar{X} , sem a necessidade de se conhecer muito sobre a população em estudo.

4.3.3 Teorema central do limite (TCL)

Seja uma variável aleatória contínua X com média μ e desvio-padrão σ . Uma amostra aleatória de tamanho n é selecionada da população de interesse, para observar X .

Quando n é grande o suficiente (em geral, $n \geq 30$), a média amostral, \bar{X} , tem

distribuição aproximadamente Normal com média μ e desvio-padrão σ/\sqrt{n} . Assim,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tem aproximadamente distribuição Normal Padrão.

Se σ é desconhecido, pode ser substituído por sua estimativa S , o desvio-padrão amostral, e mesmo assim, podemos então dizer que

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tem aproximadamente distribuição Normal Padrão.

Exemplo 4.1. Como ilustração do uso de distribuições amostrais, consideremos uma amostra aleatória de $n = 100$ trabalhadores imigrantes, com a informação sobre X , salário mensal dessa amostra de trabalhadores. O salário médio mensal da amostra, $\bar{X} =$, é de 1500 reais e é uma estimativa de μ , o salário médio mensal de todos os trabalhadores imigrantes. E o desvio-padrão amostral de salário, $S =$, é de 1200 reais. Os responsáveis pela pesquisa consideram que uma boa estimativa para o salário médio mensal deve ter no máximo um erro e de 200 reais.

Vamos calcular a probabilidade do erro máximo ser de 200 reais, isto é,

$$P(-200 \leq \bar{X} - \mu \leq 200).$$

Para tanto, usaremos o TCL, já que $n = 100$. Assim, dividindo todos os termos dentro da probabilidade por

$$\frac{S}{\sqrt{n}} = \frac{1200}{\sqrt{100}} = 120,$$

teremos:

$$P(-200 \leq \bar{X} - \mu \leq 200) = P\left(-1,67 \leq \frac{\bar{X} - \mu}{120} \leq 1,67\right).$$

Pelo TCL,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tem aproximadamente distribuição Normal Padrão. Podemos então escrever:

$$P\left(-1,67 \leq \frac{\bar{X} - \mu}{120} \leq 1,67\right) \cong P(-1,67 \leq Z \leq 1,67) = 0,905,$$

obtida da tabela de distribuição Normal Padrão, já que Z tem exatamente a distribuição Normal Padrão.

Temos então que:

$$P(-200 \leq \bar{X} - \mu \leq 200) \cong 0,905,$$

ou seja, a probabilidade do erro amostral ser no máximo R\$200,00 é aproximadamente igual a 90,5%. O valor 90,5% é conhecido como nível de confiança. Em outras palavras,

não podemos garantir que o erro máximo não será ultrapassado, mas temos 90,5% de confiança que ele não será maior que R\$200,00.

Para aumentar o nível de confiança, é preciso aumentar n , o tamanho da amostra; ou diminuir e , o erro máximo. O processo de encontrar n , fixados o nível de confiança e o erro máximo, é um problema de cálculo de tamanho de amostra, apresentado na Seção 4.9. O processo de encontrar e , fixados o tamanho de amostra e o nível de confiança, é um problema de estimação. Neste tipo de problema, fazemos o cálculo da probabilidade ao contrário: temos o valor 0,905 e queremos encontrar 1,67. Os detalhes desse processo inverso são vistos a partir da Seção 4.4.

Até o momento apresentamos uma estimativa do parâmetro baseado em um único valor, referido como estimativa pontual. Por exemplo, se a proporção de mulheres com osteoporose em uma comunidade for estimada em 45%, essa estimativa é pontual pois se baseia em um único valor numérico. Muitas vezes, entretanto, queremos considerar, conjuntamente, o estimador e a sua variabilidade. A forma usual de incorporar esta informação é por meio do chamado intervalo de confiança.

Com uma amostra disponível, nós podemos usar a distribuição amostral do estimador para formar um intervalo de confiança, isto é, um intervalo de valores que deve conter o verdadeiro valor do parâmetro com uma probabilidade pré-determinada, referida por nível de confiança. Na estimação por intervalo, acreditamos, com um certo nível de confiança, que o intervalo contém o valor do parâmetro. Um exemplo seria dizer que a proporção de mulheres com osteoporose está estimada entre 40% e 50% com um nível de 95% de confiança.

4.4 Estimação da Média Populacional (μ)

Consideremos uma população em que há interesse em estimar μ , a média de X , uma variável aleatória contínua. Suponhamos que uma amostra aleatória de tamanho n foi selecionada da população para obter uma estimativa de μ .

O estimador \bar{X} é aquele que apresenta as melhores propriedades para estimar média populacional μ .

Para estimar, por intervalo, o parâmetro μ , a partir de \bar{X} , podemos pensar em 3 condições diferentes.

Condição 1: Pelos resultados da Seção 4.3.2, se $X \sim N(\mu, \sigma^2)$ então

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tem distribuição Normal Padrão.

Como representado na Figura 4.4, podemos encontrar o valor de z , pela tabela da distribuição Normal Padrão, tal que

$$P(-z \leq Z \leq z) = 1 - \alpha.$$

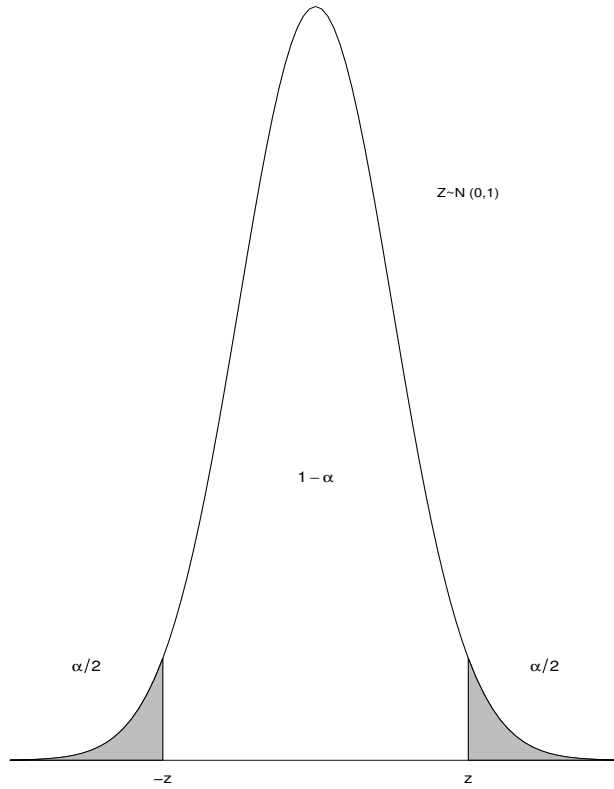


Figura 4.4: Densidade de Z e o quantil z.

Assim,

$$P\left(-z \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = 1 - \alpha.$$

Isolando μ nas desigualdades, temos que:

$$P\left(\bar{X} - z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \quad (4.1)$$

Devemos ter algum cuidado na interpretação da probabilidade em (4.1). Entre as desigualdades está o parâmetro μ , e não são calculadas probabilidades de parâmetros, pois são valores fixos. As partes aleatórias são os limites do intervalo. A interpretação para (4.1) é que com $(1 - \alpha)$ de confiança, o intervalo

$$\left(\bar{X} - z\frac{\sigma}{\sqrt{n}}; \bar{X} + z\frac{\sigma}{\sqrt{n}}\right)$$

contém o parâmetro.

O erro máximo da estimativa, e , também conhecido como margem de erro, para um nível de confiança $(1 - \alpha)$, é dado por:

$$e = z\frac{\sigma}{\sqrt{n}},$$

em que z pode ser obtido da tabela da Normal Padrão, em função do nível de confiança desejado.

Condição 2: Novamente $X \sim N(\mu, \sigma^2)$, mas é comum não conhecermos o valor de σ , o desvio-padrão de X . Neste caso, usamos

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

que tem distribuição t de Student com $(n - 1)$ graus de liberdade. O valor de t pode ser encontrado usando a tabela da distribuição t de Student, tal que

$$P(-t \leq T \leq t) = 1 - \alpha. \quad (4.2)$$

Assim, reescrevendo (4.2) temos que:

$$P\left(-t \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t\right) = 1 - \alpha$$

e isolando μ nas desigualdades, resulta em:

$$P\left(\bar{X} - t \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Um intervalo de confiança de $(1 - \alpha)$ é então:

$$\left(\bar{X} - t \frac{S}{\sqrt{n}}; \bar{X} + t \frac{S}{\sqrt{n}}\right).$$

Condição 3: Para o caso em que X não segue uma distribuição Normal, vamos supor que n seja suficientemente grande para a aplicação do TCL. Então:

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad (4.3)$$

tem aproximadamente distribuição Normal Padrão.

Podemos então escrever:

$$P\left(-z \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq z\right) \cong 1 - \alpha.$$

Isolando μ , chegamos ao intervalo:

$$\left(\bar{X} - z \frac{S}{\sqrt{n}}; \bar{X} + z \frac{S}{\sqrt{n}}\right),$$

que é o intervalo de confiança aproximado de $(1 - \alpha)$ para estimar μ .

E para um nível de confiança $(1 - \alpha)$, o erro máximo aproximado é

$$e = z \frac{S}{\sqrt{n}}$$

em que z pode ser obtido da tabela da Normal Padrão, em função do nível de confiança desejado.

Exemplo 4.2. Um centro de ortodontia deseja conhecer a estimativa do tempo médio que um membro da equipe gasta para atender a cada paciente. Suponha que uma amostra de 38 especialistas revelou que a média foi de 45 minutos com um desvio-padrão de 6 minutos. Determine um intervalo de 99% de confiança para o parâmetro.

Desejamos uma estimativa para o parâmetro desconhecido μ , o tempo médio que um membro da equipe gasta para atender um paciente. Temos que $n = 38$, $\bar{X} = 45$ e $S = 6$. Não temos muita informação sobre a distribuição de X , o tempo gasto para atender um paciente, então aplicaremos a Condição 3. Para um nível de confiança de 99%, o valor de z é 2,58. Portanto, o intervalo aproximado de 99% de confiança para μ é

$$\left(\bar{X} - z \frac{S}{\sqrt{n}}; \bar{X} + z \frac{S}{\sqrt{n}} \right) = \left(45 - 2,58 \frac{6}{\sqrt{38}}; 45 + 2,58 \frac{6}{\sqrt{38}} \right) = (42,49; 47,51).$$

Podemos dizer então que o intervalo entre 42,49 e 47,51 contém o tempo médio gasto por um membro da equipe para atender a cada paciente, com 99% de confiança.

4.5 Estimação de μ em Amostras Pequenas

Quando dispomos de uma amostra pequena ($n < 30$), não temos a garantia da aplicação do TCL, portanto a distribuição amostral da média pode ou não estar próxima da distribuição Normal. Mas se X tem distribuição Normal podemos obter as estimativas intervalares de μ , adotando a Condição 1 ou 2.

Na prática, é difícil provar que uma variável aleatória X tem distribuição Normal. Porém, se X tem apenas uma moda e é basicamente simétrica, obtemos em geral, bons resultados, com intervalos de confiança precisos. Se há forte evidência de que a população tem distribuição bastante assimétrica, então uma alternativa é utilizar métodos não-paramétricos, descritos no Capítulo 5 sobre teste de hipóteses, na Seção 5.7.

Exemplo 4.3. Uma rede de lanchonetes deseja estimar a quantia média que cada cliente gasta por lanche. Foram coletados dados de uma amostra de 22 clientes que revelou uma quantia média de R\$ 15 com um desvio-padrão de 5. Construir um intervalo de confiança de 95% para a média populacional.

O objetivo é estimar o parâmetro μ , a quantia média que cada cliente gasta por lanche, em todas as lanchonetes da rede. Assumindo que X , a quantidade gasta por lanche, tem distribuição Normal, podemos usar a Condição 2, já que σ é desconhecido. Podemos obter uma estimativa para o erro padrão da média por:

$$\frac{5}{\sqrt{22}} = 1,066.$$

Usando um nível de 95% de confiança e 21 graus de liberdade ($gl = 21$, pois $n = 22$ e $gl = n - 1$), obtemos na tabela da distribuição t de Student o valor $t = 2,08$, e assim podemos calcular o erro máximo da estimativa

$$e = t \frac{S}{\sqrt{n}} = 2,08 \cdot 1,066 = 2,217.$$

Então, temos o seguinte intervalo de 95% de confiança para o parâmetro μ :

$$\left(\bar{X} - t \frac{S}{\sqrt{n}}; \bar{X} + t \frac{S}{\sqrt{n}} \right) = (15 - 2,217; 15 + 2,217) = (12,783; 17,217).$$

O intervalo de 12,783 a 17,217 contém a quantia média que cada cliente gasta por lanche, com 95% de confiança.

Pode parecer um pouco estranho que, com uma população distribuída normalmente, venhamos eventualmente a utilizar a distribuição t de Student para achar os valores associados ao nível de confiança; mas quando σ não é conhecido, a utilização de S incorpora outra fonte de erro. Para manter o grau desejado de confiança compensamos a variabilidade adicional ampliando o intervalo de confiança por um processo que substitui o valor z por um valor maior, t . Para ilustrar esta idéia considere para o Exemplo 4.3 um intervalo de confiança de 95%, utilizando o valor z de uma distribuição Normal Padrão. Este intervalo apresenta uma amplitude menor.

$$(15 - 2,089; 15 + 2,089) = (12,911; 17,089).$$

4.6 Estimação da Diferença entre Duas Médias Populacionais (μ_1 e μ_2)

Em muitas situações há a necessidade de comparar duas populações diferentes. Como exemplos, podemos ter interesse em saber: se idosos que praticam exercícios físicos diariamente apresentam nível de colesterol menor do que idosos, com as mesmas condições, mas que não praticam exercícios físicos diariamente; se um tipo de equipamento eletrônico tem maior durabilidade do que outro; e assim por diante. A seguir vamos utilizar um método para construir um intervalo de confiança para a diferença entre duas médias, com duas amostras independentes. Para tanto, uma amostra aleatória é selecionada independentemente de cada população.

Seja X a variável aleatória a ser comparada. Suponhamos que X tenha média μ_1 e desvio-padrão σ_1 na População 1 e tenha média μ_2 e desvio-padrão σ_2 na População 2. Suponhamos também que as amostras tenham tamanhos n_1 e n_2 para as Populações 1 e 2, respectivamente.

Se n_1 e $n_2 \geq 30$, podemos estender o resultado do TCL, tanto que $(\bar{X}_1 - \bar{X}_2)$ tem distribuição aproximadamente Normal com média $\mu_1 - \mu_2$ e desvio-padrão

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Quando os parâmetros σ_1 e σ_2 são desconhecidos podem ser substituídos por S_1 e S_2 , os desvios-padrão amostrais.

Usando o mesmo processo para a construção de intervalo de confiança para uma média, o intervalo de confiança aproximado para $(\mu_1 - \mu_2)$ será dado por:

$$\left((\bar{X}_1 - \bar{X}_2) - z\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}; (\bar{X}_1 - \bar{X}_2) + z\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right),$$

em que o valor de z é encontrado na tabela da distribuição Normal Padrão, conforme o nível de confiança do intervalo.

Exemplo 4.4. Com o objetivo de comparar dois métodos de redução de gordura localizada nas coxas, foram criados dois grupos, 1 e 2, cada um com 30 pessoas que apresentam as mesmas condições, recebendo um tipo de tratamento. Antes e depois de um período de 60 dias de utilização do aparelho foi anotado a perda em mm. Obtendo-se:

$$\bar{X}_1 = 21,3; S_1 = 2,6$$

e

$$\bar{X}_2 = 13,4; S_2 = 1,9,$$

construir o intervalo de 95% de confiança para a diferença de médias.

A diferença observada na amostra é

$$\bar{X}_1 - \bar{X}_2 = 7,9$$

e o intervalo:

$$\left(7,9 - 1,96\sqrt{\frac{2,6^2}{30} + \frac{1,9^2}{30}}; 7,9 + 1,96\sqrt{\frac{2,6^2}{30} + \frac{1,9^2}{30}} \right) = (6,748; 9,0527).$$

é o intervalo de confiança de 95% para a diferença entre as reduções médias dos dois métodos.

Quando o zero pertence ao intervalo de confiança, há forte evidência de que não há diferença entre as duas médias populacionais. No Exemplo 4.4, com base no intervalo de confiança de 95%, podemos concluir que há diferença significativa entre a redução média de gordura das coxas entre os dois métodos utilizados, pois o valor zero não pertence ao intervalo de confiança.

4.7 Estimação de $\mu_1 - \mu_2$ em Amostras Pequenas

Se n_1 ou n_2 é menor que 30, formas alternativas devem ser usadas para os intervalos de confiança. A teoria usada na Seção 4.5 pode ser estendida para o caso de duas amostras independentes. Assumimos normalidade para as distribuições populacionais, como no caso

de uma amostra. Além disso, vamos assumir aqui que $\sigma_1 = \sigma_2$, isso quer dizer que as duas populações tem o mesmo desvio-padrão, digamos, σ . A variância populacional σ^2 pode ser estimada por:

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2},$$

que representa uma variância combinada, pois é uma média ponderada das duas variâncias amostrais, S_1^2 e S_2^2 .

O intervalo de confiança para $\mu_1 - \mu_2$ será dado por:

$$\left((\bar{X}_1 - \bar{X}_2) - t \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}; (\bar{X}_1 - \bar{X}_2) + t \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right),$$

em que o valor de t é encontrado na tabela da distribuição t de Student com $(n_1 + n_2 - 2)$ graus de liberdade, conforme o nível de confiança do intervalo.

Quando não podemos assumir que $\sigma_1 = \sigma_2$, não podemos combinar as variâncias e os graus de liberdade da distribuição t de Student não são tão simples de serem obtidos. A inferência sobre $\mu_1 - \mu_2$, neste caso, é visto somente para teste de hipóteses na Seção 5.4.4.

Exemplo 4.5. O tempo para realizar uma tarefa, em segundos, foi anotado para 10 homens e 11 mulheres, igualmente treinados. As médias e variâncias obtidas foram:

Homem	Mulher
$n_1 = 10$	$n_2 = 11$
$\bar{X}_1 = 45,33$	$\bar{X}_2 = 43,54$
$S_1^2 = 1,54$	$S_2^2 = 2,96$

Determine um intervalo de confiança de 99% para a diferença entre os tempos médios de homens e mulheres.

Primeiramente, vamos calcular S_p^2

$$S_p^2 = \frac{9 \times 1,54 + 10 \times 2,96}{19} = 2,29.$$

Na tabela da distribuição t de Student com 19 gl encontramos que $t = 2,86$, para 99% de confiança.

Como $\bar{X}_1 - \bar{X}_2 = 1,79$, o intervalo de confiança para $(\mu_1 - \mu_2)$ será dado por:

$$\left(1,79 - 2,86 \sqrt{2,29 \left(\frac{1}{10} + \frac{1}{11} \right)}; 1,79 + 2,86 \sqrt{2,29 \left(\frac{1}{10} + \frac{1}{11} \right)} \right) = (-0,10; 3,68).$$

Com base neste intervalo com 99% confiança, não existe diferença entre os tempos médios de homens e mulheres.

4.8 Estimação de uma Proporção Populacional (p)

Nesta seção, as variáveis são referentes a contagens, como o número de fumantes, número de unidades defeituosas em uma linha de produção, e assim por diante. Primeiramente, abordaremos a distribuição amostral de \hat{p} , em seguida a estimativa pontual do parâmetro p , a proporção populacional, e, por último, construiremos estimativas intervalares.

Consideremos o caso em que o parâmetro a ser estimado é a proporção p de indivíduos em uma população, que apresentam uma certa característica. Retira-se da população uma amostra de tamanho n . Considerando a variável $X_i = 1$, se o elemento i da população pertence a categoria de interesse, e $X_i = 0$, se não pertence a categoria, temos que $\sum_{i=1}^n X_i$ será o número de elementos da amostra que apresentam a característica em estudo.

Para uma população de N elementos podemos calcular a proporção populacional por:

$$p = \frac{\sum_{i=1}^N X_i}{N}$$

Um estimador da proporção populacional p será a proporção amostral \hat{p} :

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

Então, o cálculo de p e \hat{p} é análogo ao de uma média e, portanto, é possível aplicar o TCL para a proporção amostral.

4.8.1 TCL para proporção amostral

Seja uma amostra aleatória de tamanho n , selecionada de uma população com proporção populacional igual a p . Quando n é grande o suficiente (em geral, $n \geq 30$, se valor de p não for muito próximo de 0 ou 1), então \hat{p} , a proporção amostral, tem aproximadamente uma distribuição Normal com média p e o desvio-padrão é:

$$\sqrt{p(1-p)/n}$$

estimado por:

$$\sqrt{\hat{p}(1-\hat{p})/n}.$$

Assim,

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

tem aproximadamente distribuição Normal Padrão.

4.8.2 Intervalo de Confiança para p

Para estimar, por intervalo, o parâmetro p , a partir de \hat{p} , podemos seguir os mesmos princípios da estimação da média populacional.

Suponhamos que n seja suficientemente grande para aplicar o TCL para proporção. Temos que:

$$P(-z \leq Z \leq z) = 1 - \alpha \Rightarrow P\left(-z \leq \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} \leq z\right) \cong 1 - \alpha,$$

em que o valor de z é encontrado na tabela da distribuição Normal Padrão.

Isolando p nas desigualdades, resulta em:

$$P\left(\hat{p} - z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right) \cong 1 - \alpha. \quad (4.4)$$

A probabilidade em (4.4) pode ser interpretada como: o intervalo

$$\left(\hat{p} - z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}; \hat{p} + z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right)$$

estima o parâmetro p , com aproximadamente $(1 - \alpha)$ de confiança.

Um intervalo de confiança conservativo é:

$$\left(\hat{p} - z\sqrt{\frac{1}{4n}}; \hat{p} + z\sqrt{\frac{1}{4n}}\right).$$

A abordagem conservativa substitui o produto $p(1 - p)$ por $1/4$. Como indicado na Figura 4.5, o produto $p(1 - p)$ é, no máximo igual a $1/4$, de modo que ao usar $1/4$, obtemos um intervalo de confiança mais amplo.

Encontrando z para um nível de confiança $(1 - \alpha)$, os erros máximos de estimação na abordagem otimista e conservativa são, respectivamente, dados por:

$$e = z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

e

$$e = z\sqrt{\frac{1}{4n}}.$$

Exemplo 4.6. Um especialista em educação pretende avaliar a aceitação de um projeto educacional numa cidade. Depois de apresentá-lo às escolas do município, os responsáveis por sua execução desejam avaliar o valor aproximado do parâmetro p , a proporção de diretores favoráveis ao projeto, dentre as escolas do município. Para estimar este parâmetro, o especialista planeja observar uma amostra aleatória simples de $n = 600$ escolas. Se na amostra 420 são favoráveis, temos a seguinte estimativa pontual para o parâmetro p :

$$\hat{p} = \frac{420}{600} = 0,70.$$

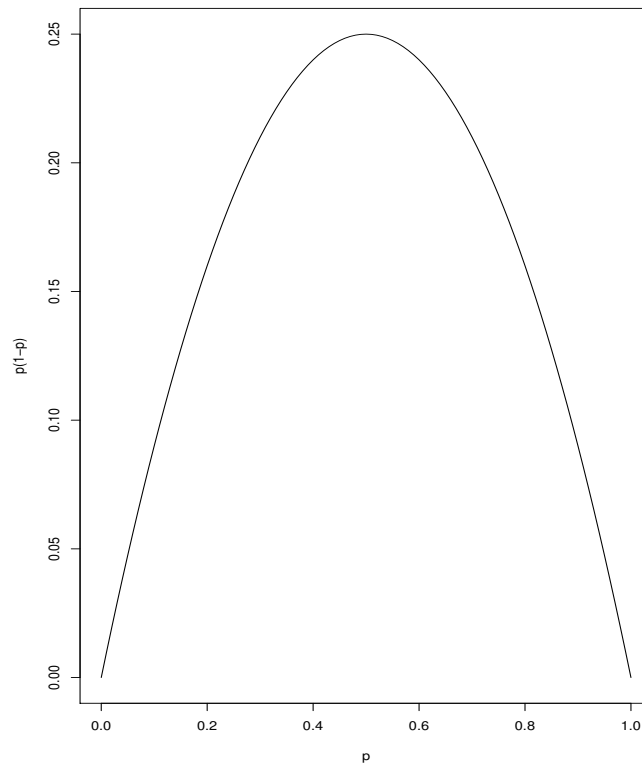


Figura 4.5: Máximo de $p(1 - p)$.

Usando um nível de 95% de confiança temos o seguinte intervalo otimista:

$$\left(0,7 - 1,96\sqrt{\frac{0,70 \cdot 0,30}{600}}; 0,70 + 1,96\sqrt{\frac{0,70 \cdot 30}{600}} \right) = (0,663; 0,737).$$

Ou seja, o intervalo de 66,3% a 73,7% contém, com 95% de confiança, a porcentagem de favoráveis ao projeto, dentre todas as escolas municipais.

4.9 Determinação do Tamanho da Amostra (n)

Supondo que há condições para aplicação do TCL, as fórmulas para o cálculo de n são derivadas, fixando o erro máximo, e e o nível de confiança $(1 - \alpha)$. A determinação de n também depende do plano amostral adotado e do parâmetro a ser estimado.

No caso da amostragem aleatória simples, a fórmula de n para a estimação de μ é encontrada isolando n em

$$e = z \frac{\sigma}{\sqrt{n}}.$$

Portanto,

$$n = \left(\frac{z\sigma}{e} \right)^2,$$

em que σ deve ser previamente estimado e z é obtido conforme o nível de confiança.

Sugestões para estimação prévia de σ :

1. usar estimativas de σ , de um estudo similar feito anteriormente ou de uma amostra piloto;
2. Em muitas situações, podemos considerar que $\sigma \approx \frac{\text{amplitude}}{4}$. O argumento teórico para o uso desta aproximação está baseado na propriedade da distribuição Normal com média μ e desvio-padrão σ , de que a área entre $\mu - 2\sigma$ e $\mu + 2\sigma$ é igual a 95,5%; portanto esta aproximação não deve ser usada se a variável em estudo for muito assimétrica.

Exemplo 4.7. Qual é o tamanho de amostra necessário para estimar a renda média mensal das famílias de uma pequena comunidade, com um erro máximo de 100 reais com 95% de confiança, usando amostragem aleatória simples? Sabe-se que a renda mensal familiar está entre 50 e 1000 reais.

Temos que $e = 100$ e para um nível de confiança igual a 95%, $z = 1,96$. Com a informação de que a renda varia entre 50 e 1000, uma aproximação para σ é:

$$\sigma \approx \frac{\text{amplitude}}{4} = \frac{1000 - 50}{4} = 237,5.$$

Assim,

$$n = \left(\frac{z\sigma}{e} \right)^2 = \left(\frac{1,96 \cdot 237,5}{100} \right)^2 = 21,67 \approx 22.$$

Portanto, cerca de 22 famílias devem ser entrevistadas.

Para a estimação de p , a fórmula para determinar n , usando amostragem aleatória simples, é encontrada isolando n em

$$e = z \sqrt{\frac{p(1-p)}{n}}.$$

Portanto,

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

em que p deve ser previamente estimado e z é obtido conforme o nível de confiança.

Sugestões para estimação prévia de p :

1. Usar estimativas de p de um estudo similar feito anteriormente ou de uma amostra piloto;
2. Substituir o produto $p(1-p)$ por 0,25. Notamos que ao substituir por 0,25, o tamanho da amostra pode ser maior que o necessário. É por isso que chamamos de abordagem conservadora, quando fazemos esta substituição nas fórmulas do intervalo de confiança.

Exemplo 4.8. Líderes estudantis de uma faculdade querem conduzir uma pesquisa para determinar a proporção p de estudantes a favor de uma mudança no horário de aulas. Como é impossível entrevistar todos os 2000 estudantes em um tempo razoável, decide-se fazer uma amostragem aleatória simples dos estudantes:

- a) Determinar o tamanho de amostra (número de estudantes a serem entrevistados) necessário para estimar p com um erro máximo de 0,05 e nível de confiança de 95%. Assumir que não há nenhuma informação a priori disponível para estimar p .

Temos que $e = 0,05$ e que $z = 1,96$. Como não há informação a priori sobre p , segue que

$$n = z^2 \frac{p(1-p)}{e^2} = 1,96^2 \frac{0,25}{0,05^2} = 384,16 \approx 385.$$

Para estimar a proporção de estudantes favoráveis a mudança de horário, com um erro máximo de 0,05 a 95% de confiança, é necessária uma amostra de 385 estudantes.

- b) Os líderes estudantis também querem estimar a proporção de p de estudantes que sentem que a representação estudantil atende adequadamente as suas necessidades. Com um erro máximo de 7% e nível de confiança de 95%, determinar o tamanho de amostra para estimar p . Utilizar a informação de uma pesquisa similar conduzida alguns anos, quando 60% dos estudantes acreditavam que estavam bem representados.

$$n = z^2 \frac{p(1-p)}{e^2} = 1,96^2 \frac{0,60(1-0,60)}{0,07^2} = 188,16 \approx 189.$$

Para estimar a proporção de estudantes que se consideram bem representados, é necessária uma amostra de 189 estudantes; considerando um erro máximo de 0,07 a 95% de confiança.

- c) Qual o tamanho de amostra adequado para atingir ambos os objetivos da pesquisa? Para atingir ambos os objetivos da pesquisa, devemos considerar a maior amostra, que é a de 385 estudantes.

Quando N (tamanho da população) é conhecido, o valor de n para estimar μ e p pode ser corrigido (n^*):

$$n^* = \frac{Nn}{N+n}.$$

Notamos que se N é muito maior que n , então n^* é aproximadamente n .

Exemplo 4.9. Determinar o tamanho de amostra necessário para estimar o volume médio de vendas de carros novos nacionais entre as concessionárias, fixando um nível de confiança de 99% para um erro de estimação igual a 1 automóvel. É conhecido que existem 200 concessionárias na região em estudo. Em uma pesquisa similar feita 5 anos antes, o desvio-padrão amostral foi igual a 2,8. Supor que foi feita uma amostragem aleatória simples.

Temos que $e = 1$ e para um nível de confiança igual a 99% temos que $z = 2,58$. Usaremos a estimativa a priori para σ , substituindo-o na fórmula por 2,8. Assim,

$$n = \left(\frac{z\sigma}{e} \right)^2 = \left(\frac{2,58 \cdot 2,8}{1} \right)^2 = 52,19 \approx 53.$$

Com a informação de que há $N = 200$, podemos corrigir o valor de n

$$n^* = \frac{Nn}{N+n} = \frac{200 \cdot 53}{200+53} = \frac{10600}{253} = 41,90 \approx 42.$$

Portanto, é necessário selecionar 42 concessionárias de automóveis, para estimar o número médio de carros vendidos, com um erro máximo de 1 automóvel a 99% de confiança.

Capítulo 5

Testes de Hipóteses

5.1 Introdução

Os testes estatísticos são regras de decisões, vinculadas a um fenômeno da população, que nos possibilitam avaliar, com o auxílio de uma amostra, se determinadas hipóteses (suposições, conjecturas, algo qualquer que um pesquisador esteja estabelecendo) podem ser rejeitadas, ou não.

No campo da Inferência Estatística, a busca por respostas acerca de certas características de uma população estudada é de fundamental importância. Apenas com base nessas características é que se devem estabelecer regras e tomar decisões sobre qualquer hipótese formulada no que se refere à população. Dessa forma, escolhida uma variável X e colhida uma amostra aleatória da população, podemos estar interessados em inferir a respeito de alguns de seus parâmetros (média, variância e proporção, por exemplo) e, também, sobre o comportamento da variável (a sua distribuição de probabilidade). A realização de testes de hipóteses nos fornece meios para que possamos, com determinado grau de certeza, concluir se os valores dos parâmetros ou mesmo a distribuição associados à população considerada, podem representá-la de forma satisfatória. Nesse contexto, temos os Testes Paramétricos, vinculados a estimação dos valores dos parâmetros e os Testes de Aderência, associados à busca da distribuição de X . Na verdade, como veremos nos exemplos do item 5.3, quando realizamos Testes Paramétricos, esses estão intimamente ligados aos Testes de Aderência. Pois, para se obter a “determinada certeza” citada, é necessário que saibamos qual a distribuição de probabilidade que melhor se ajusta às estimativas observadas por intermédio das amostras.

A maior parte das ciências se utiliza da técnica Estatística denominada Teste de Hipóteses. Podemos citar algumas suposições: o dado de certo cassino é honesto; a propaganda de um produto vinculada na televisão surtiu o efeito desejado; uma raça desenvolvida para certo animal proporcionou um ganho maior de peso do que aquela já utilizada há anos; vale a pena trocar as máquinas desta indústria; qual medicamento é mais eficaz no tratamento de certa doença; a metodologia empregada na educação infantil está associada ao aprendizado; o candidato A está com uma intenção de votos superior ao

seu adversário; o número de acidentes por dia na BR-116 segue uma distribuição de Poisson; o tempo de vida de uma marca de lâmpada pode ser representado pela distribuição Exponencial. E assim, poderíamos citar inúmeras suposições dentro de todas as áreas do conhecimento. Vejamos, agora, algumas terminologias e conceitos adotados na área dos Testes de Hipóteses Paramétricos.

5.2 Conceitos Estatísticos dos Testes de Hipóteses

5.2.1 Hipóteses estatísticas paramétricas

As hipóteses paramétricas são suposições sobre os valores dos parâmetros de certa população.

Tipos de hipóteses

1. Hipótese de Nulidade (Nula): é a hipótese que está sendo testada. Colhida uma amostra a fim de inferirmos a respeito do valor paramétrico (θ), obtemos a estimativa do parâmetro ($\hat{\theta}$) por intermédio de um Estimador. Daí, por meio do Cálculo de Probabilidades, cujos resultados são obtidos em função da Hipótese Nula (H_0), tomamos a decisão de rejeitar, ou não, H_0 . Nessa decisão, é verificada se a diferença observada entre o valor suposto na Hipótese Nula e a estimativa do parâmetro, $\theta - \hat{\theta}$, é significativa, ou não. Veja que quanto menor a diferença, maior será a probabilidade de não rejeitarmos H_0 . Assim, dizemos que $(\theta - \hat{\theta})$ não foi significativa, concluindo-se que tal diferença ocorreu por acaso. Caso contrário, devemos rejeitar H_0 e concluir que a diferença foi suficientemente grande para não ter, provavelmente, ocorrido ao acaso. Vejamos o exemplo: Será que a altura média ($\theta = \mu$) dos alunos da UFPR é de 1,71 m? Resumidamente, a Hipótese Nula poderia ser descrita desta forma:

$$H_0 : \mu = 1,71 \text{ m}$$

Para respondermos a essa questão, deveríamos colher uma amostra de tamanho n e obtermos a estimativa da média ($\hat{\theta} = \bar{X}_{obs}$), função da amostra, e, posteriormente verificarmos a diferença entre μ e \bar{x}_{obs} . Caso H_0 fosse rejeitada, concluiríamos que a diferença observada foi significativa e que não se deveu ao acaso. Logo, a média verdadeira (μ) assume um outro valor, desde que diferente de 1,71 m. E, conseqüentemente, esses possíveis valores pertenceriam à Hipótese Alternativa.

2. Hipótese Alternativa (H_a ou H_1): é uma hipótese que, necessariamente, difere de H_0 . Assim, nesse contexto, teríamos: $H_1 : \mu \neq 1,71 \text{ m}$ ou $H_1 : \mu < 1,71 \text{ m}$ ou $H_1 : \mu > 1,71 \text{ m}$.

5.2.2 Testes

São regras de decisão acerca da rejeição ou não rejeição de uma determinada Hipótese Nula.

Tipos de testes

Dependendo do interesse da pesquisa, podemos estabelecer testes específicos conforme o objetivo do pesquisador. Por exemplo:

1. Teste Bilateral (Bicaudal): $H_0 : \mu = 1,71 \text{ m}$ vs $H_1 : \mu \neq 1,71 \text{ m}$.

Note que o objetivo desse teste é decidir se a média populacional não difere de 1,71 m, não importando se μ será maior ou menor do que 1,71 m.

2. Teste Unilateral à Direita: $H_0 : p = 0,30$ vs $H_1 : p > 0,30$.

Esse teste tem por finalidade verificar se, por exemplo, a proporção verdadeira não só difere de 0,30, mas, também, necessariamente, se p é maior do que 0,30. Objetivamente, poderíamos citar uma pesquisa que visa verificar se um determinado candidato a Reitor, conseguiu aumentar sua intenção de votos após a realização de um debate com seu adversário realizado pela televisão.

3. Teste Unilateral à Esquerda $H_0 : \sigma^2 = 5$ vs $H_1 : \sigma^2 < 5$.

Nesse contexto, visamos estabelecer uma Regra de Decisão para verificarmos se a variabilidade é menor do que 5. Pois, por exemplo, se for menor do que 5, não seria recomendado investirmos num melhoramento genético.

5.2.3 Tipos de erros cometidos ao se tomar uma decisão

Ao trabalharmos com amostras para tomarmos decisões, é bem provável incorrer em erros. A esses erros chamamos de Erro Tipo I (de 1ª Espécie) e Erro Tipo II (de 2ª Espécie). E às probabilidades associadas a eles denotaremos por α e β , respectivamente. A Tabela 5.1 ilustra tais idéias:

Tabela 5.1: Erros cometidos na tomada de decisão.

Decisão	Realidade	
	H_0 é Verdadeira	H_0 é Falsa
Rejeitar H_0	Erro Tipo I	Decisão Correta
Não Rejeitar H_0	Decisão Correta	Erro tipo II

Portanto, as probabilidades ficariam:

$\alpha = P(\text{Rejeitar } H_0 | H_0 \text{ é Verdadeira})$ e $\beta = P(\text{Não Rejeitar } H_0 | H_0 \text{ é Falsa})$. Na realização de um Teste de Hipóteses levamos em consideração essas duas probabilidades.

Caso utilizarmos apenas α , como é usual, estaríamos realizando um Teste de Significância. O valor de β está associado ao poder do teste. Pois, a Função Poder do teste é dada por $1 - \beta(\mu^*)$. Sendo μ^* o valor verdadeiro, porém, desconhecido da média populacional.

Espera-se que quanto menor o valor de β , maior poder terá o teste. Se realizarmos um Teste de Hipóteses, esse valor deverá ser considerado na análise. Pois, como $\beta = P(\text{Não Rejeitar } H_0 | H_0 \text{ é Falsa})$, β dependeria do valor da média verdadeira para sua obtenção, visto que o cálculo dessa probabilidade está condicionada a H_0 ser falsa. Portanto, teríamos de obter tal probabilidade para valores de μ diferentes de 1,71 m. Detalhes desse cálculo serão vistos no Exemplo 5.2 do 5.3.

5.2.4 Região crítica (RC) e regra de decisão (RD)

Quando efetuamos um Teste de Hipóteses (ou de Significância), a probabilidade α , também denotada por nível de significância, estará intimamente ligada à Regra de Decisão do Teste. Logo, ao estabelecermos α , podemos construir uma região onde devemos rejeitar H_0 . Portanto, sob H_0 , essa região, chamada de Região Crítica (ou de Região de Rejeição), deverá conter todas as possíveis amostras raras de ocorrerem. Dessa forma, escolhido um $\alpha = 0,01$, dizemos que qualquer amostra cuja probabilidade de ocorrência for menor do que 0,01, nos levará à decisão de rejeitar H_0 . Pois, sob a referência adotada (α), tal amostra será considerada rara (probabilidade < 0,01) se admitirmos H_0 como verdadeira. Devemos notar, então, que existem amostras que podem ser coletadas, de certa população, que nos fornecem resultados bem diferentes dos esperados (sob H_0), que pertencem à Região Crítica, porém, com probabilidades bastante baixas a ponto de concluirmos que se tratam de amostras raras. Daí, a conclusão mais plausível é a de rejeitarmos a Hipótese Nula. Note, ainda, que se H_0 fosse falsa, a probabilidade dessas amostras ocorrerem poderia ser alta. Esse conceito será tratado, detalhadamente, na discussão da Tabela 5.6.

Tomando-se novamente o exemplo da altura média dos alunos da UFPR, criaríamos uma Regra de Decisão com base em $\alpha = 0,01$ e $H_1 : \mu < 1,71$ m. Assim, poderíamos estabelecer a seguinte regra: Caso coletarmos uma amostra cujo resultado (\bar{x}_{obs}) for menor do que 1,67 m, decidiremos por rejeitar H_0 , pois a probabilidade disso ocorrer é menor do que $\alpha = 0,014$. Ou seja, sob a referência ($\alpha = 0,01$), a amostra coletada deverá ser vista como rara se a Hipótese Nula for verdadeira ($H_0 : \mu = 1,71$). Consequentemente, seria mais conveniente optarmos em dizer que $\mu < 1,71$ m.

Nesse contexto, vejamos no item 5.2.5 algumas etapas de um Teste de Significância.

5.2.5 Procedimentos para realização de um teste de significância

1. Estabelecer as Hipóteses Nula e Alternativa;
2. Identificar a Distribuição Amostral associada ao Estimador e obter a Estimativa do Parâmetro;

3. Fixar um valor para o Nível de Significância (α) e obter a estatística de teste do Parâmetro por meio da Estatística do Teste;
4. Construir a Região Crítica (RC) com base na Hipótese Alternativa e no valor de α e estabelecer a Regra de Decisão (RD);
5. Concluir o Teste: Se a Estimativa do Parâmetro pertencer à Região Crítica, rejeitamos a Hipótese Nula. Caso contrário, não.

Com a finalidade de ilustramos tais conceitos, vejamos alguns exemplos.

5.3 Exemplos

Serão vistos dois exemplos com a finalidade de explicar, detalhadamente, as idéias sobre os Testes de Hipóteses. São eles:

Exemplo 5.1. Imagine que seu amigo José lhe diga o seguinte: Possuo em meu bolso do paletó 16 bolas de gude, sendo que 10 são brancas (B), 4 são verdes (V) e 2 são amarelas (A). Você acreditaria nessa afirmação? Quais seriam os meios que você disporia para verificar a verdade? Formalmente, poderíamos estabelecer, primeiramente, as seguintes hipóteses: H_0 : José não está mentindo versus H_1 : José está mentindo, bem como fixar um valor para α , por exemplo 0,10. E, posteriormente, criar algumas alternativas para uma tomada de decisão. A primeira delas seria, simplesmente, esvaziar o bolso do José e, conseqüentemente, tomar uma decisão. Como, nessa situação, foram observadas todas as bolas, a sua decisão será absolutamente correta. Ou seja, não restará dúvida (incerteza) em relação à afirmação de José, caso decidíssemos rejeitar H_0 . Porém, em Estatística, nem sempre isso é possível, ou por inacessibilidade de toda população, ou pela impossibilidade desse procedimento ou, ainda, por tal procedimento ser inviável, ou seja, não possuir um sentido prático. Como exemplo, podemos citar a verificação do tempo de vida de certa marca de lâmpada. Obviamente, não faria o menor sentido amostrarmos todas as lâmpadas de uma linha de produção, pois assim não restaria nenhuma delas para a venda. E o resultado desse procedimento seria desastroso.

Nesse contexto, uma segunda alternativa nos levaria à coleta de uma amostra de tamanho n e observaríamos o resultado e, assim, tomaríamos uma decisão. Portanto, você poderia, por exemplo, retirar 2 bolas ($n = 2$) do bolso do José, sem reposição. Imaginemos, conforme a Tabela 5.2, algumas ocorrências, implicações e possíveis decisões:

Tabela 5.2: Algumas ocorrências, implicações e decisões após a retirada da amostra.

	Ocorrência	Implicação	Decisão
I	1 Bola Branca e 1 Bola Preta	Afirmação Falsa	José mentiu
II	2 Bolas Brancas	Afirmação Possível	José, provavelmente, não mentiu
III	2 Bolas Amarelas	Afirmação Possível	José, provavelmente, mentiu

Note que, a ocorrência I é impossível, pois não existem bolas pretas. Porém, é imprescindível que notemos que a ocorrência I só se torna inconsistente se a afirmação de José for verdadeira. Afirmação essa, que transformamos em hipótese estatística (H_0 : José não está mentindo) a fim de criar uma referência para a tomada de decisões. Note, então, que a ocorrência I, na verdade, é possível, tanto é que ela ocorreu. Mas, a probabilidade dela ter ocorrido dado que H_0 é verdadeira, é igual a zero. Daí, a opção em rejeitar a hipótese. José certamente mentiu. Veja, também, que o fato de ter ocorrido I, implica em várias outras possibilidades de ocorrência caso fossem coletadas outras amostras ($n=2$). Mas, isso não nos importaria mais, visto que a decisão já foi tomada com uma certeza absoluta.

Analiseemos, agora, as ocorrências II e III:

Nunca perca de vista que, tanto a ocorrência II quanto a ocorrência III só se tornam possíveis sob a afirmação de José (sob H_0). Ou seja, se José afirma que existem 10 bolas brancas e 2 amarelas, é claramente viável que se retirem 2 bolas brancas ou 2 amarelas dentre as 16. Agora, a diferença fundamental para a ocorrência I é que nessas outras ocorrências a tomada de decisão, sobre a honestidade de José, vem carregada de um grau de incerteza. Portanto, admitindo-se que cada bola possui a mesma probabilidade de pertencer à amostra, podemos, então, obter as probabilidades de ocorrência para todos os possíveis eventos associados a todas as amostras de tamanho 2. Com base nesse cálculo de probabilidades verificamos se a hipótese testada (H_0) deve, ou não, ser rejeitada. Como podemos ver na Tabela 5.3, as probabilidades, sob a hipótese de que a afirmação de José é verdadeira, ficariam:

Tabela 5.3: Distribuição de probabilidades das possíveis amostras.

Amostra	AA	VV	AV	AB	BV	BB
Probabilidade	1/120*	6/120	8/120	20/120	40/120	45/120

$$* P(AA) = \frac{\binom{2}{2}}{\binom{16}{2}} = \frac{1}{120}$$

Observe que sob a alegação de José, todas essas amostras são possíveis e cada uma delas possui uma determinada chance de ocorrer. Daí, caso ocorra a amostra BB, seria plausível admitirmos que José esteja falando a verdade. Pois, a probabilidade associada ao evento BB sob a hipótese de que José não esteja mentindo é alta (0,375). Ou seja, se refizéssemos o experimento 100 vezes, esperaríamos que em quase 38 vezes ocorressem 2 bolas brancas. Além disso, comparativamente com o nível de significância adotado (α) esperaríamos que as amostras raras ocorressem com frequências inferiores ou iguais a 10 em 100. Logo, a amostra BB não poderia ser considerada rara. Portanto, por conveniência, com base no cálculo de probabilidades, é melhor que não rejeitemos a afirmação. Assim, H_0 não deve ser rejeitada e a ocorrência de BB deve ter sido por acaso. Agora, se ocorrer a amostra AA, a decisão deve ser oposta. Pois, a probabilidade desse evento ocorrer é 45 vezes menor do que a anterior (0,0083). Assim, se o experimento fosse feito 120 vezes, em apenas 1 vez esperaríamos uma amostra do tipo AA. Frequência essa bem abaixo da

referência adotada de 10 em 100. Conseqüentemente, concluímos que a amostra AA é rara. O que nos induz a pensar que, provavelmente, o José estaria mentindo.

Note que tivemos de tomar uma decisão com base na distribuição de probabilidades das possíveis amostras. Dessa forma, deveríamos ter pensado o seguinte: Se admitirmos que José disse a verdade, o que seria mais fácil de ocorrer quando retiramos a amostra AA? O José estar mentindo ou, de fato, termos realmente colhido tal amostra rara? Veja que não se trata de amizade e sim de possibilidades. Tome sua decisão como base naquilo que foi amostrado, no que realmente aconteceu, baseando-se na distribuição de probabilidades das possíveis amostras. Dessa forma, a Regra de Decisão adotada deve ser a seguinte: Se coletarmos uma amostra e sua probabilidade de ocorrência for menor do que 0,10, concluiremos que se trata de amostra rara e assim devemos rejeitar H_0 . Caso contrário, não. Veja um resumo dessa idéia na Tabela 5.4.

No momento em que tomamos a decisão de não rejeitar a Hipótese Nula quando retiramos a amostra BB, é evidente que podemos estar incorretos nessa decisão, uma vez que há outras probabilidades envolvidas. E todas elas menores do que 1. A esse tipo de erro, como já discutido, chamamos de Erro Tipo II. Assim, imagine que José estivesse a fim de que tomássemos uma decisão incorreta. Ele, sabendo desses conceitos, poderia ter colocado em seu bolso apenas bolas brancas. Dessa forma, ele saberia *a priori* que você só poderia coletar 2 bolas brancas e, conseqüentemente, tomar a decisão errada.

Outra tomada de decisão incorreta seria você rejeitar a hipótese quando ela é verdadeira. Assim você diria que seu amigo mentiu, quando, na realidade, ele disse a verdade. Esse erro é denotado por Erro Tipo I, que para esse exemplo estabelecemos sua probabilidade em 0,10.

A Tabela 5.4 resume as decisões adotadas em função de todas as possíveis amostras.

Tabela 5.4: Resumo das decisões para o Exemplo 5.1.

Amostra	Decisão		
	$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,01$
AA	Rejeitar H_0	Rejeitar H_0	Rejeitar H_0
AB	Não Rejeitar H_0	Não Rejeitar H_0	Não Rejeitar H_0
AV	Rejeitar H_0	Não Rejeitar H_0	Não Rejeitar H_0
BB	Não Rejeitar H_0	Não Rejeitar H_0	Não Rejeitar H_0
BV	Não Rejeitar H_0	Não Rejeitar H_0	Não Rejeitar H_0
VV	Rejeitar H_0	Rejeitar H_0	Não Rejeitar H_0

Pela análise da Tabela 5.4 podemos tirar algumas conclusões:

1. À medida que diminuimos o Nível de Significância do Teste, torna-se mais difícil rejeitar a Hipótese Nula. É óbvio que isso ocorra, pois com α menor, a amostra coletada terá uma chance cada vez menor de ser considerada rara;

2. Optamos por rejeitar H_0 para a ocorrência de VV com $\alpha = 0,05$. Porém, segundo a Tabela 3, a probabilidade de retirarmos essa amostra é igual a α . Valor esse que se encontra exatamente no limite de nossa Regra de Decisão. Então, intuitivamente, o que deveríamos fazer?

Naturalmente, a solução seria repetirmos o experimento k vezes, ou seja, em vez de apenas uma amostra, deveríamos colher k amostras. Assim, a fim de tomarmos decisões mais acertadas obteríamos uma nova Distribuição Amostral com uma gama bem maior de eventos possíveis, aumentando o poder do Teste, tornando-o melhor. Caso coletássemos outra amostra ($n = 2$), ou seja, se refizéssemos o experimento, teríamos 21 resultados possíveis (21 novas amostras) com suas respectivas probabilidades associadas. Veja alguns resultados e decisões desse novo experimento na Tabela 5.5.

Tabela 5.5: Resumo das decisões para o novo experimento.

Amostra	Probabilidade	Decisão		
		$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,01$
(AA,AA)	1/14400	Rejeitar H_0	Rejeitar H_0	Rejeitar H_0
(AA,AB)	40/14400	Rejeitar H_0	Rejeitar H_0	Rejeitar H_0
(AA,AV)	16/14400	Rejeitar H_0	Rejeitar H_0	Rejeitar H_0
...
(BB,AB)	1800/14400	Não Rejeitar H_0	Não Rejeitar H_0	Não Rejeitar H_0
(BB,AV)	720/14400	Rejeitar H_0	Rejeitar H_0	Não Rejeitar H_0
(BB,BB)	2025/14400	Não Rejeitar H_0	Não Rejeitar H_0	Não Rejeitar H_0
(BB,BV)	3600/14400	Não Rejeitar H_0	Não Rejeitar H_0	Não Rejeitar H_0
...
(VV,VV)	36/14400	Rejeitar H_0	Rejeitar H_0	Rejeitar H_0
(AV,BV)	640/14400	Rejeitar H_0	Rejeitar H_0	Não Rejeitar H_0
(BV,BV)	1600/14400	Não Rejeitar H_0	Não Rejeitar H_0	Não Rejeitar H_0

Observe na Tabela 5.5 que se a amostra AA já era considerada rara, a amostra (AA, AA) se tornou ainda mais. Pois, se refizéssemos o experimento 14400, esperaríamos em apenas 1 vez coletar simultaneamente a amostra AA. Por consequência, a decisão em rejeitar H_0 é muito mais segura do que a anterior. Além disso, o valor de $\beta = P(\text{Não Rejeitar } H_0 | H_0 \text{ é Falsa})$ se torna cada vez menor. Lembre-se que decidimos modificar o experimento porque a realização de VV nos forneceu uma insegurança na tomada de decisão. Agora, com o experimento modificado, a coleta da amostra (VV, VV) possui uma probabilidade muito pequena (0,0025), indicando, nesse caso, ser uma amostra rara. Daí, a decisão deixa de ser insegura.

Nesse contexto, com base em seus conhecimentos de Cálculo de Probabilidades e Distribuições Amostrais, complete a Tabela 5.5 e, também, construa uma outra tabela similar à Tabela 5.5 refazendo o experimento do Exemplo 5.1 três vezes retirando amostras de tamanho 2, em cada uma das vezes, sem reposição.

Finalmente, é essencial que notemos que, dependendo da forma como se realiza o experimento, a distribuição de probabilidades das possíveis amostras, sempre utilizada com intuito de se tomar uma decisão em relação a Hipótese Nula, se modifica substancialmente. Assim, no experimento inicial, com a retirada de apenas uma amostra ($n=2$), a amostra BB possuía, como já era esperado, a maior probabilidade (0,375) uma vez que tínhamos 10 bolas brancas. Já no experimento modificado com a retirada de duas amostras de tamanho 2, a probabilidade associada à amostra (BB, BB), mesmo possuindo ainda as mesmas 10 bolas brancas, diminui consideravelmente (0,14). Assim, se modificarmos ainda mais o experimento inicial, repetindo-o, por exemplo quatro vezes, a probabilidade observada de (BB, BB, BB, BB) ficaria aproximadamente em 0,0198. Portanto, rejeitaríamos H_0 para níveis de significância maiores do que 0,0198. Conseqüentemente, mudaríamos nossas decisões em relação aos experimentos anteriores.

Exemplo 5.2. Antes de enunciarmos este exemplo, gostaríamos de refletir sobre a seguinte pergunta: O valor 167 cm é menor do que 171 cm? Obviamente que muitos, talvez a maioria, diriam que sim. Porém, antes que saibamos como esses resultados foram obtidos, a melhor resposta seria: depende. Vejamos, então, as reflexões 1 e 2:

1. Se medíssemos as alturas de duas pessoas A e B, da mesma maneira e obtivéssemos, respectivamente, 167 cm e 171 cm. Concluiríamos que A é, de fato, menor do que B;
2. Se o interesse for descobrir e comparar a altura média de duas turmas (A e B) da UFPR, poderíamos obter essas alturas médias de várias maneiras, vejamos dois casos:
 - (a) com a coleta das duas populações, as médias obtidas seriam as médias verdadeiras, ou seja, os valores paramétricos (μ_A e μ_B). Assim, diríamos novamente que 167 cm é menor do que 171 cm.
 - (b) coletando-se a população de A e uma amostra de B, e obtidas as médias $\mu_A = 167$ cm e $\bar{x}_B = 171$ cm, não poderíamos afirmar com absoluta certeza que 167 cm é menor do 171 cm. Pois, recordando o conceito de Distribuição Amostral, sabemos que \bar{X} é uma variável aleatória. Portanto, a $P(X \in RC)$ é dependente de uma Distribuição de Probabilidade, conseqüentemente, apenas com base no comportamento de \bar{X} é que poderíamos decidir se, provavelmente, $\mu_A < \mu_B$. Assim, se tanto na turma A quanto na B, ou nas duas forem coletadas amostras, a resposta para a questão proposta sempre dependerá do comportamento das estimativas das possíveis amostras. Comportamento esse, representado por meio de uma Distribuição de Probabilidades e, portanto, toda decisão a respeito da questão virá acompanhada de um grau de incerteza. A Inferência Estatística, por intermédio do Teste de Hipóteses, visa responder a essa questão.

Feitas as reflexões, podemos enunciar o Exemplo 5.2: Sabemos que a variável (X) altura dos alunos da Universidade A, local A, segue uma distribuição normal com altura média de 171 cm e desvio padrão de 9 cm. Se recebermos, de uma origem desconhecida, local B, uma amostra de 27 alunos, poderíamos decidir se essa amostra foi retirada da Universidade A ou se o local B possui a mesma média do local A?

Admitamos que a população cuja amostra ($n=27$) foi retirada seja bem representada por uma distribuição normal com desvio padrão igual ao da Universidade A ($\sigma = 9\text{cm}$). Sabemos da Teoria da Estimação que se $X_A \sim N(171, 9^2)$, então, $\bar{X}_A \sim N(171, 9^2/27)$. Assim, o comportamento das estimativas das médias das possíveis amostras da Universidade A fica bem caracterizada: $\bar{X}_A \sim N(171, 3)$. Supondo que $\bar{x}_B = 167$ cm, essa estimativa pode ser vista como rara ou não? Como poderíamos, com base na estimativa da média da amostra B, obter uma Regra de Decisão para concluirmos sobre a origem dessa amostra. Enfim, qual conclusão deveríamos tomar? A solução é simples. Assim como no Exemplo 5.1, basta verificar se seria plausível coletarmos do Local A uma amostra de 27 alunos cuja estimativa da altura média fosse de 167 cm. Porém, diferentemente do Exemplo 5.1, a variável X desse exemplo é contínua. Portanto, sendo a probabilidade pontual igual a zero, devemos obter uma probabilidade intervalar. Assim, poderíamos calcular a probabilidade da estimativa da média ser menor ou igual a 167 cm sob H_0 , ou seja, supondo que $\mu = 171$ cm, $P(\bar{x}_B \leq 167 | \mu = 171)$. Assim, estaríamos contemplando no cálculo dessa probabilidade valores iguais a 167 cm, mas, também, valores com médias inferiores a 167 cm. Dessa forma, comparando-se com a Regra de Decisão adotada, podemos concluir se essas estimativas são raras. Com o auxílio da transformação da variável X na Normal Padronizada, $Z \sim N(0, 1)$, calcularíamos a probabilidade dessa forma:

$$\begin{aligned} P(\bar{x}_B \leq 167 | \mu = 171) &= P\left[Z \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right] \\ &= P\left[Z \leq \frac{167 - 171}{9/\sqrt{27}}\right] \\ &= P(Z \leq -2,31) = 0,0105. \end{aligned}$$

Graficamente, teríamos (Figura 5.1):

Com base nessa probabilidade, chamada de Nível Descritivo (P-valor) do Teste, podemos tomar uma decisão da seguinte forma: se acharmos que essa probabilidade é baixa, concluiríamos que:

- a amostra deve ser rara;
- deveríamos rejeitar a Hipótese Nula de que os 27 alunos pertencem à Universidade A;
- μ_A deve ser superior a μ_B ;
- a diferença observada $(\mu_A - \bar{x}_B) = 4$ cm, provavelmente, foi significativa;

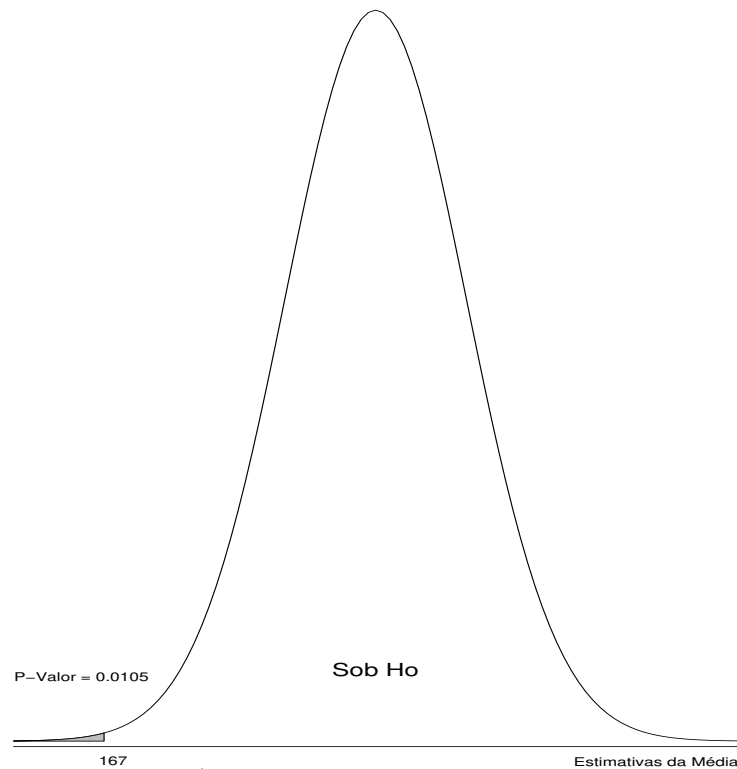


Figura 5.1: Área hachurada relativa ao P-Valor do teste

- a diferença observada não deve ter ocorrido por acaso;
- se refizéssemos o experimento inúmeras vezes, esperaríamos que na maioria delas as estimativas vindas das possíveis amostras ($n=27$) do local B, nos fornecesse valores inferiores a 171cm, indicando, portanto, uma tendência.

Formalmente, poderíamos realizar o teste segundo as etapas do item 5.2.5. Assim,

1. Estabelecer as Hipóteses Nula e Alternativa;

$$\begin{cases} H_0 & : \mu = 171 \text{ cm} \\ H_1 & : \mu < 171 \text{ cm} \end{cases}$$

2. Identificar a Distribuição Amostral associada ao Estimador e obter a Estimativa do Parâmetro;

$$\bar{X}_A \sim N(171, 3) \quad e \quad \bar{x}_B = 167 \text{ cm}$$

3. Fixar um valor para o Nível de Significância (α) e obter a estatística de teste do Parâmetro;

$$\alpha = 0,05 \quad e \quad z_{\text{calculado}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = -2,31$$

4. Construir a Região Crítica com base na Hipótese Alternativa e no valor de α ;

$$z_{teórico} = -1,6449 \quad \text{ou} \quad \bar{x}_{crítico} = 168,2 \text{ cm}$$

Regra de decisão: Se a Estimativa do Parâmetro pertencer à Região Crítica, rejeitamos a Hipótese Nula. Caso contrário, não. Ou seja, se $z_{calculado} < z_{teórico}$, ou se $\bar{x}_B < 168,2$, rejeitamos H_0 .

5. Conclusão: Como $Z_{calculado} = -2,31 < Z_{Teórico} = -1,6449$ ou $\bar{x}_B = 167 \text{ cm} < 168,2 \text{ cm}$. Decidimos por rejeitar H_0 .

A obtenção do P-valor, essencial para a tomada de decisão, está intimamente ligado ao comportamento da distribuição das estimativas obtidas com o auxílio das possíveis amostras coletadas da população. Esse comportamento pode ser descrito pela função de densidade de probabilidade (f.d.p) associada à distribuição amostral dessas estimativas. Porém, essa f.d.p. só poderá ser perfeitamente caracterizada sob a referência da Hipótese Nula. Então, os valores paramétricos que a caracterizam estão contidos em H_0 . Dessa forma, dependendo dos valores paramétricos supostos em H_0 , obteremos P-valores distintos e, em consequência, poderemos tomar decisões diferentes. Para ilustrar tal fato, tomemos o Exemplo 5.2 como referência e imaginemos algumas situações em que σ , μ_0 e n são modificados. As Decisões e as Regras de Decisão, em função de $\bar{x}_{obs} = 167$, podem ser vistas na Tabela 5.6:

Tabela 5.6: Algumas tomadas de decisão e regras de decisão conforme a hipótese nula, o nível de significância e a distribuição de probabilidade.

		Decisão($\bar{x}_{obs} = 167$)			Regra de decisão para rejeitar H_0		
Situação	P-valor	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,10$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,10$
Exemplo 5.2 ($\mu = 171; \sigma = 9$)	0,0105	NR	R	R	Se $\bar{X} < 166,97$	Se $\bar{X} < 168,2$	Se $\bar{X} < 168,8$
I $\mu = 170$ $\sigma = 9$	0,0416	NR	R	R	Se $\bar{X} < 166$	Se $\bar{X} < 167,2$	Se $\bar{X} < 167,8$
II $\mu = 170$ $\sigma = 10$	0,06	NR	NR	R	Se $\bar{X} < 165,5$	Se $\bar{X} < 166,8$	Se $\bar{X} < 167,5$
III $\mu = 170$ $\sigma = 6$	0,0047	R	R	R	Se $\bar{X} < 168,7$	Se $\bar{X} < 169,1$	Se $\bar{X} < 169,3$
IV $\mu = 168$ $\sigma = 2$	0,0047	R	R	R	Se $\bar{X} < 167,1$	Se $\bar{X} < 167,4$	Se $\bar{X} < 167,5$
V $\mu = 173$ $\sigma = 12$	0,0047	R	R	R	Se $\bar{X} < 167,6$	Se $\bar{X} < 169,2$	Se $\bar{X} < 170,0$
VI $\mu = 172$ $\sigma = 12$	0,1056	NR	NR	NR	Se $\bar{X} < 162,7$	Se $\bar{X} < 165,4$	Se $\bar{X} < 166,87$

De maneira geral, de acordo com a Hipótese Alternativa, sabemos que a Região Crítica situa-se nas caudas da distribuição de densidade. Assim, para a distribuição normal padrão, valores muito altos ou muito baixos da estatística $Z_{calculado}$ indicariam uma tendência de rejeição de $H_0 : \mu = \mu_0$. Dessa forma, pelo estudo dessa Estatística, notamos que $|\bar{x} - \mu_0|$, σ e n são os fatores que contribuem para a decisão em se rejeitar, ou não, H_0 . Logo, verificamos que:

- Um aumento na diferença observada, $\bar{x} - \mu_0$, contribuirá na tendência em se rejeitar H_0 ;
- Um aumento na variabilidade dos dados, σ , contribuirá para a não rejeição de H_0 ;
- Um aumento no tamanho da amostra, n , contribuirá para a rejeição de H_0 .

Nesse contexto, comparando-se o Exemplo 5.2 com a Situação I, da Tabela 5.6, notamos que, embora as decisões sejam as mesmas para $\bar{x} = 167$, pela Regra de Decisão adotada, o valor de \bar{x}_{obs} do Exemplo 5.2, com $\alpha = 0,01$, está praticamente no limite da Região Crítica, ou seja, quase rejeitamos H_0 , o que poderia nos causar uma dúvida maior na decisão em comparação à Situação I. Veja que o único fator modificado nesses dois casos foi a suposição em torno da média. Dessa forma, a diferença observada, $|\bar{x} - \mu_0|$, no Exemplo 5.2 é superior à diferença para a Situação I, 4 contra 3, respectivamente. Daí, é intuitivo pensar que quanto mais distante se encontrar a estimativa da média, \bar{x}_{obs} , do valor suposto em H_0 , μ_0 , maior será a tendência em se rejeitar H_0 . Agora, se compararmos a Situação I com a II, verificamos que apenas o fator variabilidade, σ , foi alterado. O aumento de σ de 9 para 10, acarretou numa mudança na decisão acerca de H_0 . Passamos a não rejeitar H_0 quando a variância aumentou. É evidente que isso poderia acontecer, pois a função de probabilidade tornou-se mais platicúrtica, portanto com caudas menos densas, possuindo, assim, uma quantidade maior de amostras de tamanho 27, cujas estimativas da média resultassem em valores menores do que 167 cm. Essa idéia fica, de fato, evidenciada pela observação das Regras de Decisão, que indicam, para um mesmo nível de significância, valores sempre inferiores de para a Situação II. Ou seja, para $\alpha = 0,05$, temos, para a Situação I, 5% das possíveis amostras com estimativas da média inferiores a 167,2 cm e para a Situação II temos 5% com valores menores do que 166,8 cm.

Nesse contexto, note que na Situação III diminuimos ainda mais o valor de σ . Logo, é de se esperar que a estimativa, $\bar{x}_{obs} = 167$ cm, seja mais representativa comparativamente com as situações anteriores, pois n e μ_0 permaneceram os mesmos. Assim, a diferença observada, $|\bar{x} - \mu_0| = 3$, também será mais representativa para a Situação III. Embora essa diferença seja a mesma para as três situações, esperamos que a diferença associada à Situação III, não tenha ocorrido por acaso, indicando, portanto, uma tendência de que, independentemente da amostra coletada, esperamos que na grande maioria das coletas, \bar{x}_{obs} seja inferior a 170 cm. Quando isso ocorre, dizemos que a diferença foi significativa. A interpretação seria análoga a essa se o tamanho da amostra fosse aumentado e os demais fatores permanecessem constantes.

Na Situação II, se coletássemos inúmeras amostras, várias delas forneceriam estimativas menores do que 170 cm e muitas outras resultariam em valores superiores a 170 cm, implicando, assim, que a diferença deve ter ocorrido por acaso, não havendo uma tendência. Dessa forma, não seria conveniente decidirmos que $\mu < 170$, e sim, concluirmos que a diferença não foi grande o suficiente (não foi significativa) a ponto de descartarmos H_0 . Logo, μ poderia ser 170 cm. Por outro lado, por se tratar de amostragem, também

não poderíamos afirmar que $\mu = 170$. Porém, deveríamos afirmar que provavelmente μ não é menor do que 170 cm.

Ao compararmos as situações III, IV e V, verificamos que os P-valores são os mesmos. Portanto as decisões são rigorosamente as mesmas. Porém, note que, pelas Regras de Decisão adotadas, os valores de $\bar{x}_{\text{crítico}}$ são bem diferentes para essas três situações. Veja que na Situação IV, quase rejeitamos H_0 para $\alpha = 0,01$, pois $\bar{x}_{\text{crítico}} = 167,1$ e $\bar{x}_{\text{obs}} = 167$, resultando numa diferença de apenas 0,1 cm. Ao contrário da Situação III, cuja diferença foi de 1,7 cm. Porém, teoricamente, essa observação está equivocada, pois, apenas poderíamos ter essa impressão porque as diferenças são aparentemente muito distintas. Mas, é imprescindível notarmos que, tanto essas diferenças quanto os intervalos associados a elas, embora bem diferentes, são equivalentes uma vez que suas distribuições são distintas. Assim, deveríamos verificar que, tanto no intervalo $167 < \bar{X}_{III} < 168,7$, 1 quanto no intervalo $167 < \bar{X}_{IV} < 167,1$, há o mesmo número de possíveis amostras ($n = 27$). Ou seja, $P(167 < \bar{X}_{III} < 168,7) = P(167 < \bar{X}_{IV} < 167,1) = 0,053$, pois $\bar{X}_{III} \sim N(170; 6^2/\sqrt{27})$ e $\bar{X}_{IV} \sim N(168; 2^2/\sqrt{27})$. Assim, se dissermos que quase rejeitamos H_0 na Situação IV, devemos dizer o mesmo para a Situação III. Além disso, note que embora a suposição para a Situação IV é de $\mu = 168$ e para a Situação V seja de $\mu = 173$, as decisões também são as mesmas. Ainda que as diferenças observadas sejam de 1 cm e de 6 cm, respectivamente, não podemos nos esquecer que o desvio padrão é também outro fator responsável pela tomada de decisão. Logo, notamos que o desvio padrão da Situação IV é seis vezes menor, implicando, portanto, em estimativas (\bar{x}_{obs}) bastante próximas de μ_0 , caso H_0 seja verdadeira. Dessa forma, é de se esperar que as diferenças observadas na Situação III estejam mais próximas de zero do que para a Situação V.

Finalmente, na Situação VI, tem-se um tamanho de amostra três vezes menor do que nas situações anteriores e, também, o maior desvio padrão dentre todas as situações. Note que, mesmo para uma média suposta de 172 cm, encontramos um P-valor que resulta em rejeição para todos os níveis de significância adotados na Tabela 5.6. Comparativamente, vê-se que a diferença observada na Situação IV é de 1 cm e na Situação VI é de 5 cm. Porém, a primeira diferença foi significativa, enquanto a segunda, não.

Com vimos, há uma probabilidade de não rejeitar H_0 quando ela é falsa, ou seja, tomamos a decisão de que $\mu = \mu_0$ quando, na verdade, não é. Obtém-se essa probabilidade calculando-se $\beta = P(\text{Não Rejeitar } H_0 | H_0 \text{ é Falsa})$. Veja que se não rejeitamos H_0 , decidimos que $\mu = \mu_0$, portanto β será calculado em função de $\mu = \mu_0$, ou seja, em função de $\bar{X} \sim N(\mu_0, \sigma^2/n)$. Porém, também para obtermos β é necessário que seja escolhido um outro valor para μ , uma vez que H_0 é falsa, por exemplo, $\mu = \mu^* = 169$ cm. Portanto, para o Exemplo 5.2, teríamos, para $\alpha = 0,05$:

$$\begin{aligned}
\beta &= P(\text{Não Rejeitar } H_0 | H_0 \text{ é Falsa}) \\
&= P(\bar{X}_{obs} > \bar{X}_{crítico} | \mu^* = 169) \\
&= P\left[Z > \frac{\bar{x} - 169}{\sigma/\sqrt{n}}\right] \\
&= P\left[Z > \frac{168,2 - 169}{9/\sqrt{27}}\right] = 0,678
\end{aligned}$$

Graficamente, teríamos (Figura 5.2):

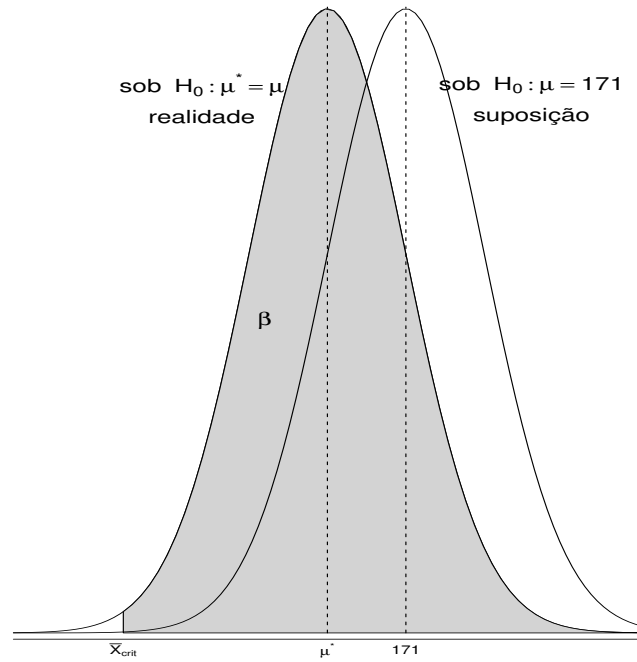


Figura 5.2: Probabilidade de não rejeitar H_0 quando ela é falsa.

Note que a expressão $\bar{x}_{obs} > \bar{x}_{crítico}$ está associada à Regra de Decisão com base em $H_0: \mu = \mu_0 = 171$. Por isso, $\bar{x}_{crítico} = 168,2$ para $\alpha = 0,05$ (veja Tabela 5.6). Mas, o valor de β , deve ser obtido por meio da probabilidade condicional que depende da média suposta (μ_0) e da média real (μ^*). Assim, ao transformarmos $\bar{x}_{crítico}$ em $z_{crítico}$, devemos primeiramente levarmos em consideração $\mu_0 = 171$, para a obtenção de $\bar{x}_{crítico}$, e posteriormente $\mu^* = 169$, para o cálculo final de β .

Vejamos, na Tabela 5.7, com base no Exemplo 5.2, o que ocorreria com o valor de β , criadas novas situações com mudanças nos valores de α , σ , n e μ^* e, conseqüentemente, com o Poder do Teste, $1 - \beta(\mu^*)$.

Tabela 5.7: Valores de $1 - \beta(\mu^*)$ para o exemplo 5.2 de acordo com os parâmetros α , σ , n e μ^* .

		$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,10$
Situação		$1 - \beta(\mu^*)$		
I	$\mu^* = 169; \sigma = 9; n = 27$	0,121	0,322	0,454
II	$\mu^* = 169; \sigma = 9; n = 200$	0,793	0,933	0,969
III	$\mu^* = 169; \sigma = 2; n = 27$	0,998	0,9998	$\cong 1$
IV	$\mu^* = 172; \sigma = 9; n = 27$	0,002	0,0013	0,032
V	$\mu^* = 166; \sigma = 9; n = 27$	0,712	0,893	0,946
VI	$\mu^* = 169; \sigma = 9; n = 9$	0,048	0,164	0,269

Notamos, claramente, que independentemente da situação, que à medida que α cresce, β diminui. Ou seja, o poder do teste aumenta e teríamos uma menor chance de errarmos na decisão de não se rejeitar H_0 quando H_0 é falsa.

Pela comparação das Situações I, II e III verificamos que, independentemente do valor de α , o poder do teste aumenta consideravelmente de I para III, passando por II. Como μ^* se manteve constante, os fatores σ e n foram responsáveis diretos para que o poder se modificasse. Além disso, note que o aumento verificado no tamanho da amostra de I para II, de quase 7,5 vezes, contribuiu menos para um aumento do poder, do que a diminuição no desvio padrão de 9 para 2, 4,5 vezes, verificada entre as situações I e III. Porém, como não podemos interferir em μ^* e σ para a obtenção do poder, resta-nos trabalharmos com n . Assim, se quisermos testes mais poderosos, devemos, em princípio, aumentar o tamanho da amostra. Nesse contexto, compare a Situação I com a V e posteriormente, reflita sobre as demais situações consideradas na Tabela 5.7.

Vejamos, agora, alguns testes mais utilizados.

5.4 Alguns Testes Paramétricos mais Utilizados.

5.4.1 Teste para a média (μ) com σ^2 desconhecida.

O objetivo desse teste é verificar a hipótese $H_0 : \mu = \mu_0$. Porém, como σ^2 é desconhecida, podemos estimá-la, obtendo-se s^2 , por meio da mesma amostra utilizada para a obtenção de \bar{X}_{obs} . Portanto, a diferença fundamental desse teste para o teste descrito na Situação II do Exemplo 5.2, em que σ^2 era conhecido, é dada pelo comportamento das estimativas de \bar{X}_{obs} vindas das possíveis amostras de tamanho n . Enquanto na Situação I \bar{X}_{obs} seguia uma distribuição Normal (apenas \bar{X}_{obs} varia, σ^2 é constante), na Situação II, tanto \bar{X}_{obs} quanto σ^2 variam, logo, outra distribuição deverá representar esse comportamento. Nesse caso, sabemos que a distribuição a t de Student, poderá se ajustar a essas estimativas.

Porém, à medida que n cresce, os testes tendem a ser equivalentes, pois os valores de Z e t se aproximam. Dessa forma, os P -valores estarão bastante próximos, para $n > 30$, e, conseqüentemente, a decisão em relação à H_0 será praticamente a mesma.

Exemplo 5.3. Uma máquina enche pacotes de café de uma marca X deve completá-los, em média, com no mínimo 500 g. Se coletássemos de uma amostra aleatória de tamanho 16, a fim de verificarmos se a máquina se encontra regulada, e obtivéssemos uma média igual a 495 g e desvio padrão de 5 g, seria plausível concluirmos que a média é menor do que 500 g, ou seja, a máquina se encontraria regulada?

Os dados observados são: 498,8; 503,1; 497,6; 491,6; 499,3; 491,3; 499,8; 492,1; 498,1; 493,2; 487,2; 489,8; 495,8; 498,2; 498,8; 485,7

Solução: Devemos proceder ao Teste Paramétrico, segundo as etapas descritas no item 5.2.5. Assim:

1. Estabelecer as Hipóteses Nula e Alternativa;

$$H_0 : \mu = 500g \text{ vs } H_1 : \mu < 500g$$

2. Identificar a Distribuição Amostral associada ao Estimador e obter a Estimativa do Parâmetro; Distribuição Amostral: t de Student com 15 g.l.. Pois, $n < 30$ e σ^2 desconhecida;

$$\text{Estimativas: } \bar{X}_{obs} = 495g \text{ e } s = 5g$$

3. Fixar um valor para o Nível de Significância (α) e obter a estatística de teste do Parâmetro por meio da Estatística do Teste;

$$\text{Nível de Significância: } \alpha = 0,01$$

Estatística de teste:

$$t_{calculado} = \frac{\bar{X}_{obs} - \mu_0}{s/\sqrt{n}} = \frac{495 - 500}{5/\sqrt{16}} = -4,0$$

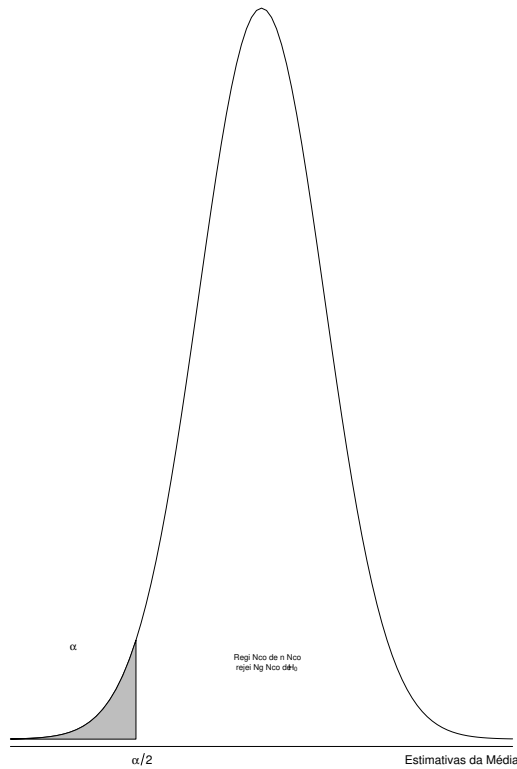
Logo: $P\text{-Valor} = P(t < -4,0) = 0,0006$.

4. Construir a Região Crítica (RC) com base na Hipótese Alternativa e no valor de α e estabelecer a Regra de Decisão (RD);

A Região Crítica é a área hachurada cuja probabilidade é igual a $\alpha = 0,01$. Observe a Figura 5.3 associada à RC.

Assim, qualquer valor de $t_{calculado}$ menor do que -2,602, pertencerá à RC. Ou seja, $RC = \{t \in \Re | t < -2,602\}$; $P(t < -2,602) = 0,01$. Agora, transformando $t = -2,602$ em \bar{X} , obtemos o $\bar{X}_{crítico}$, assim:

$$-2,602 = \frac{\bar{X}_{crítico} - 500}{5/\sqrt{16}} \Rightarrow \bar{x}_{crítico} = 496,75.$$

Figura 5.3: Região crítica associada à estatística t

. Logo, qualquer valor da média estimada, \bar{X}_{obs} , inferior a 496,75 g, pertencerá à RC. A Figura 5.4 ilustra essa idéia.

Regra de Decisão:

Se $t_{calculado} < t_{crítico} = -2,602$, ou se $\bar{X}_{obs} < \bar{X}_{crítico} = 495,75$ g, optamos em rejeitar $H_0 : \mu = 500$ g.

5. Concluir o Teste: Como a Estimativa do Parâmetro ($\bar{X}_{obs} = 495$ g) pertence à Região Crítica, rejeitamos a Hipótese Nula. Dessa forma, podemos concluir que:

- a diferença observada ($495 - 500$) foi significativa;
- a estimativa da média, 495 g, não deve ter ocorrido por acaso. Assim, se refizermos o experimento inúmeras vezes, esperaríamos que os valores obtidos para \bar{X}_{obs} fossem, na sua grande maioria inferiores a 500 g, indicando, portanto, uma tendência;
- conforme a probabilidade (P-Valor), cujo valor está associado à ocorrência de estimativas da média menores do que 495 g, que $\bar{X}_{obs} = 495$ g, sob H_0 , é raro (veja a Figura 5.5);
- Conseqüentemente, de acordo com essas conclusões, seria plausível admitirmos que μ deve ser menor do que 500 g, indicando que a máquina, provavelmente, esteja desregulada.

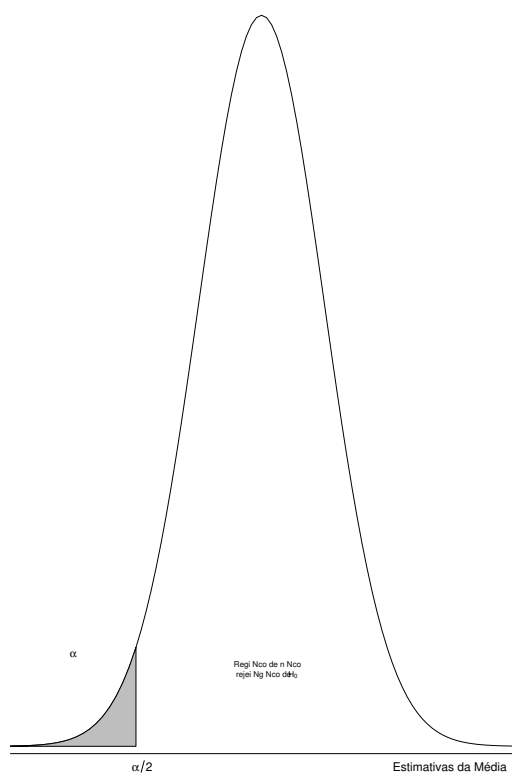


Figura 5.4: Região crítica associada à estimativa da média

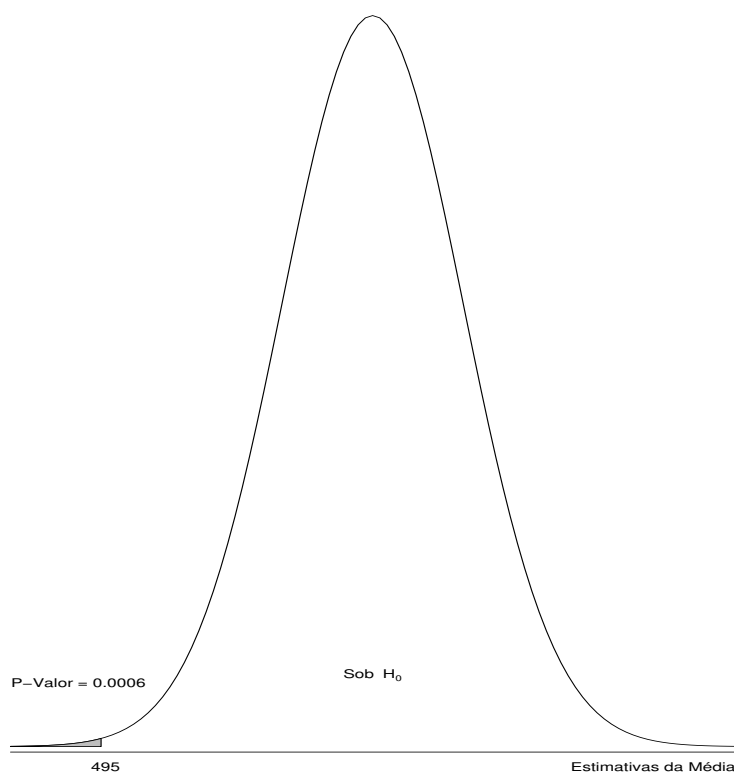


Figura 5.5: Probabilidade associada à ocorrência de estimativas da média menores do que 495 g.

5.4.2 Teste para a comparação de duas médias populacionais (μ_1 e μ_2)

O objetivo desse teste é verificar se a diferença suposta (D) em H_0 entre μ_1 e μ_2 pode ser considerada significativa. Ou seja, a Hipótese Nula será dada por:

$$H_0 : \mu_1 - \mu_2 = D.$$

A fim de procedermos ao teste corretamente, devemos observar, primeiramente, se as amostras (n_1 e n_2) provenientes das duas populações são independentes ou não. Caso sejam, torna-se primordial a realização de um teste para se verificar se as variâncias diferem. Pois, as metodologias para as realizações dos testes diferem. Logo, as decisões acerca da comparação entre μ_1 e μ_2 podem diferir. Por outro lado, se as amostras forem dependentes, um outro teste específico deverá ser realizado. Esquemáticamente, teríamos:

Comparação entre μ_1 e μ_2

$$\text{Comparação entre } \mu_1 \text{ e } \mu_2 \begin{cases} \text{amostras independentes} \begin{cases} \sigma_1^2 = \sigma_2^2 \rightarrow \text{Teste} \\ \sigma_1^2 \neq \sigma_2^2 \rightarrow \text{Teste} \end{cases} \\ \text{Amostras Dependentes} \Rightarrow \text{Teste} \end{cases}$$

Note, então, que antes de compararmos as duas médias, é fundamental que façamos primeiramente um teste para a verificação da homocedasticidade ($H_0 : \sigma_1^2 = \sigma_2^2$). Assim, poderemos efetuar o teste adequado para testar $H_0 : \mu_1 - \mu_2 = D$. Dessa forma, as etapas do Teste para compararmos as duas variâncias, fica:

1. Estabelecer as Hipóteses Nula e Alternativa;

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs } H_1 : \sigma_1^2 \neq \sigma_2^2$$

2. Distribuição Amostral: F de Snedecor, que sob H_0 , é dada por: $F_{(a,b,\alpha)}$, sendo a e b os números de graus de liberdade do numerador e denominador, respectivamente;

Estimativas: s_1^2 e s_2^2 ;

3. Nível de Significância: $\alpha = \alpha_0$

Estatística: $F_{calculado} = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$. Note que, embora $H_1 : \sigma_1^2 \neq \sigma_2^2$, como o numerador será maior do que o denominador, $F_{calculado}$ será maior do que 1. Logo, o teste adotado será o unilateral à direita;

4. Região Crítica (RC): valores de $F > F_{crítico} = F_{tabelado} = F_{(a,b,\alpha_0)}$ pertencerão à RC;

5. Concluir o Teste: Se $F_{calculado}$ pertencer à Região Crítica, rejeitamos a Hipótese Nula. Caso contrário, não.

Feitas as observações, vejamos os testes 5.4.3, 5.4.4 e 5.4.5.

5.4.3 Teste para amostras independentes com $\sigma_1^2 = \sigma_2^2$.

As amostras serão consideradas independentes (ou não correlacionadas, ou não pareadas, ou não emparelhadas) se a ocorrência de um valor específico retirado da população 1, não interferir, ou mesmo não estiver correlacionado com a observação colhida da população 2. Nesse contexto, observe os casos seguintes:

Caso 1) Qual das duas rações A e B proporcionaria um maior ganho de peso médio, em 20 dias, numa determinada raça de animal?

Note que os animais que receberão a Ração A são diferentes daqueles que receberão a Ração B. Veja, também, que o ganho de peso de um certo animal que recebeu a ração A não interferirá no ganho de peso de qualquer outro animal em que administramos a ração B. Além disso, podemos ter tamanhos de amostra diferentes para as duas rações. Logo, as amostras podem ser consideradas independentes.

Caso 2) Será que uma determinada ração A atinge um ganho de peso médio igual a μ_0 , em 20 dias, numa determinada raça de animal?

Observe que nesse caso, para sabermos se houve, de fato, um ganho de peso desejado, precisaríamos observar os pesos dos n animais da amostra antes da ração ser administrada e após os 20 dias considerados. Assim, há uma amostra inicial, associada ao início do tratamento com a ração, e uma amostra associada ao término do tratamento com a ração A. E, com base, nos pesos iniciais e finais desses n animais, podemos estabelecer o ganho de peso individual, e conseqüentemente, o ganho de peso médio por intermédio da diferença observada entre os pesos. Portanto, só obteremos o ganho de peso médio se tivermos as duas informações obtidas de um mesmo animal. Logo, fica evidente que essas informações são dependentes. Ou seja, são pareadas.

Exemplo 5.4. Vamos utilizar o Caso 1, com as seguintes informações: A Ração A foi administrada em 8 animais (n_A) e observou-se uma média de 3,12 kg (\bar{x}_A) e um desvio padrão de 0,15 kg (s_A^2). Já na Ração B, obtivemos $\bar{x}_B = 3,05$ kg e $s_B = 0,11$ kg para $n_B = 10$. Vejamos, portanto, se as rações diferem em relação ao ganho de peso. Esse problema pode ser resolvido por meio de um teste paramétrico. Porém, qual metodologia devemos utilizar? As referentes aos testes 5.4.3, 5.4.4 e 5.4.5? Como já discutido, as amostras são independentes. Portanto, devemos, antes de proceder ao teste, verificar se. Assim, as etapas do teste ficam:

Os dados utilizados são:

Ração A: 3,40; 2,99; 3,21; 3,07; 3,01; 3,27; 3,23; 3,02.

Ração B: 2,82; 3,16; 2,98; 3,04; 3,15; 3,20; 3,00; 3,01; 3,08; 3,06.

1. As Hipóteses: $H_0 : \sigma_A^2 = \sigma_B^2$ vs $H_0 : \sigma_A^2 > \sigma_B^2$
2. Distribuição Amostral: Sob H_0 , $F_{(a,b,\alpha)} = F_{\text{tabelado}} = F(7, 9, \alpha_0)$.
Estimativas: $s_A^2 = 0,15^2$ e $s_B^2 = 0,11^2$
3. Nível de Significância: $\alpha = \alpha_0 = 0,05$ (por exemplo)

Estatística: $F_{calculado} = \frac{0,15^2}{0,11^2} = 1,8595$.

4. Região Crítica (RC): valores de $\frac{s_A^2}{s_B^2} > 3,293$ pertencem à RC;

Regra de Decisão (RD): se $F_{calculado} > F_{tabelado} = 3,293$, rejeita-se H_0 .

5. Concluir o Teste: Como $F_{calculado}$ não pertence à RC, não se rejeita $H_0 : \sigma_A^2 = \sigma_B^2$. Logo, há homocedasticidade e o teste 5.4.3 deverá ser o escolhido.

Dessa forma, o teste fica:

1. As Hipóteses:

$$H_0 : \mu_A - \mu_B = 0 \text{ g vs } H_1 : \mu_A - \mu_B \neq 0 \text{ g}$$

Note que, neste caso, $D = 0$. Pois o objetivo é apenas verificar se existe diferença significativa entre o ganho de peso médio das rações. Porém, $D \in \Re$, assim, poderíamos estar interessados em investigar se $D = 0,05$. Ou seja, será que a Ração A proporciona um ganho de peso superior ao da Ração B em pelo menos 50 g?

2. Distribuição Amostral: t de Student com γ g.l..

Pois, $n_A + n_B < 30$ e as variâncias são desconhecidas.

Sendo $\gamma = n_A + n_B - 2 = 8 + 10 - 2 = 16$ g.l.;

Estimativas: $\bar{x}_A = 3,12$; $s_A = 0,15$; $\bar{x}_B = 3,05$ e $s_B = 0,11$

3. Nível de Significância: $\alpha = 0,05$

Estatística de teste:

$$t_{calculado} = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{s_c \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{3,12 - 3,05}{0,1290 \sqrt{\frac{1}{8} + \frac{1}{10}}},$$

Como as variâncias não diferem, calculamos uma variância comum dada por:

$$s_c^2 = \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A+n_B-2}$$

Logo: P-Valor = $P(t > 1,143) = 0,1348$

4. A Região Crítica é a área hachurada cuja probabilidade é igual a $\alpha = 0,05$. Observe, então, que:

$$RC = \{t \in \Re | t < -1,746\} \text{ ou } t > t_{\text{crítico}} = 1,746$$

ou

$$RC = \{(\bar{x}_A - \bar{x}_B) \in \Re | (\bar{x}_A - \bar{x}_B) < -0,11 \text{ ou } (\bar{x}_A - \bar{x}_B) > 0,11\} \text{ e}$$

$$P(t > 1,746) = 0,05,$$

Regra de Decisão: Se $t_{calculado} > 1,746$, ou se $((\bar{x}_A - \bar{x}_B)) > 0,11 \text{ kg}$, rejeitamos H_0 .

5. Concluir o Teste: como $(\bar{x}_A - \bar{x}_B) = 0,07$ não pertence à Região Crítica, não rejeitamos a Hipótese Nula. Dessa forma, podemos concluir que:

- (a) a diferença observada $(3,12 - 3,05)$ não foi significativa;
- (b) a estimativa $(\bar{x}_A - \bar{x}_B) = 0,07$ deve ter ocorrido por acaso. Assim, se refizermos o experimento inúmeras vezes, esperaríamos que os valores obtidos para $(\bar{x}_A - \bar{x}_B)$ fossem, ora maiores do que zero e ora inferiores a zero, indicando, portanto, a ausência de tendência;
- (c) conforme a probabilidade $(P - Valor = 0,1348)$, notamos que, sob H_0 , a estimativa obtida $(0,07)$ não é rara;
- (d) Conseqüentemente, de acordo com essas conclusões, seria plausível admitirmos que $\mu_A = \mu_B$. Portanto, as rações possuem, provavelmente, o mesmo efeito no ganho de peso desses animais.

5.4.4 Teste para amostras independentes com $\sigma_1^2 \neq \sigma_2^2$.

Vejamos, agora um exemplo para o caso em que as variâncias diferem.

Exemplo 5.5. Um experimento com o objetivo de verificar a resistência (kgf) de dois tipos de concreto foi realizado. Os dados são os seguintes:

Tabela 5.8: Resistência (kgf) de dois tipos de concreto.

Concreto 1	101,2	102,0	100,8	102,3	101,6
Concreto 2	100,0	102,8	101,5	99,0	102,0

Pede-se: Há evidência de que o Concreto 1 seja mais resistente do que o Concreto 2? Estabeleça $\alpha = 0,05$.

Solução: Embora $n_1 = n_2 = 5$, nota-se claramente que as amostras não são pareadas. Além disso, $F_{calculado} = 6,54 > F_{tabelado} = 6,39$, logo $\sigma_1^2 \neq \sigma_2^2$. Portanto, o teste fica:

1. As Hipóteses: $H_0 : \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_1 - \mu_2 > 0$
2. Distribuição Amostral: t de Student com γ^* g.l..

Sendo γ^* dado, por exemplo, pela fórmula de Aspin-Welch com arredondamento para menos:

$$\gamma^* = \frac{(v_1 + v_2)^2}{[v_1^2/(n_1 + 1)] + [v_2^2/(n_2 + 1)]} - 2 \quad e \quad v_1 = \frac{s_1^2}{n_1} \quad e \quad v_2 = \frac{s_2^2}{n_2}$$

Assim, $\gamma^* = 5,79 \Rightarrow \gamma^* = 5g.l.$

Portanto, $t_{tabelado} = 2,015$

3. Nível de Significância: $\alpha = 0,05$

Estatística de teste:

$$t_{\text{calculado}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{101,6 - 101,1}{\sqrt{\frac{0,362}{5} + \frac{2,368}{5}}}$$

Logo: $P - \text{Valor} = P(t > 0,67) = 0,2643$.

4. Região Crítica: $RC = \{t \in \Re | t > 2,015\}$; $P(t > 2,015) = 0,05$, ou $RC = \{(\bar{x}_1 - \bar{x}_2) \in \Re | > 1,49\}$

Regra de Decisão: Se $t_{\text{calculado}} > t_{\text{crítico}}$, ou se $(\bar{x}_1 - \bar{x}_2) > 1,49$, optamos em rejeitar H_0 .

5. Concluir o Teste: Como $t_{\text{calculado}}=0,67$ (ou $(\bar{x}_1 - \bar{x}_2) = 0,5$) não pertence à Região Crítica, não rejeitamos $H_0 : \mu_1 - \mu_2 = 0$. Logo, concluímos que não há evidência estatística, estabelecendo-se um nível de significância de 0,05, de que o Concreto 1 seja mais resistente do que o Concreto 2.

5.4.5 Teste para amostras dependentes

No caso em que duas amostras estão correlacionadas segundo algum critério (por exemplo o caso antes e depois), dizemos que aos dados estão emparelhados (veja o Caso 2 do item 5.4.3). Assim, devemos calcular as diferenças (d_i), para cada par de valores, obtendo-se uma única amostra de n diferenças ($n_1 = n_2 = n$). Então, testar a hipótese de que a diferença entre as médias das duas populações pareadas é igual a um certo valor D_0 , equivale a testar a hipótese de que a média de todas as diferenças (μ_d) seja igual a D_0 , ou seja:

$$H_0 : \mu_1 - \mu_2 = D_0 \iff H_0 : \mu_{\text{Antes}} - \mu_{\text{Depois}} = D_0 \iff H_0 : \mu_d = D_0$$

Dessa Forma, retornamos ao item 5.4.1 em que testamos $H_0 : \mu = \mu_0$. Então, de forma análoga, a Estatística do Teste é dada por:

$$t_{\text{calculado}} = \frac{\bar{d} - D_0}{s_d / \sqrt{n}}$$

em que:

\bar{d} é a média da amostra das diferenças;

D_0 é o valor suposto para a média das diferenças;

s_d é o desvio padrão da amostra das diferenças;

n é o tamanho da amostra pareada.

Vejamos um exemplo:

Exemplo 5.6. Um novo medicamento está sendo pesquisado com intuito de diminuir a pressão sistólica em indivíduos hipertensos. Dez pacientes voluntários submeteram-se ao

Tabela 5.9: Pressão antes e após seis meses da administração do medicamento.

Antes	179	200	161	170	181	190	202	220	195	165
Após	160	180	161	180	165	170	196	216	170	160
Diferença (d_i)	-19	-20	0	10	-16	-20	-8	-4	-25	-5

tratamento que consistia em medir a pressão antes e após seis meses da administração do medicamento. Os dados são os seguintes (Tabela 5.9):

Você acreditaria que o medicamento surte o efeito desejado, com $\alpha = 0,01$?

Solução: Etapas do teste:

1. As Hipóteses: $H_0 : \mu_d = D_0 = 0$ vs $H_1 : \mu_d \neq 0$
2. Distribuição Amostral: t de Student com $(n - 1)$ g.l. = 9 g.l.;
Portanto, $t_{tabelado} = -3,25$
Estimativas: $\bar{d} = -10,7$ e $s_d = 11,066$
3. Nível de Significância: $\alpha = 0,01$

Estatística de teste:

$$t_{calculado} = \frac{-10,7 - 0}{11,066/\sqrt{10}}$$

Logo: P-Valor= $P(t < -3,06) = 0,007$

4. Região Crítica (RC): $RC = \{t \in \Re | t < -3,25\} \text{ ou } \{t > 3,25\} = \{\bar{d} \in \Re | \bar{d} < -11,37 \text{ ou } \bar{d} > 11,37\}$.

Regra de Decisão: Se $t_{calculado} < -3,25$, ou se $\bar{d} < -11,37$, rejeitamos H_0 .

5. Concluir o Teste: Como $\bar{d} = -10,7 > -11,37$, podemos concluir que:

- (a) a diferença observada (- 10,7) não foi significativa;
- (b) pelo resultado do P-Valor, podemos dizer que provavelmente a amostra das diferenças não é rara;
- (c) Assim, concluímos que o medicamento não obteve um resultado desejado.

5.5 Teste para Proporção Populacional (p)

O objetivo desse teste é verificar se a proporção verdadeira (populacional) não difere de um valor suposto p_0 . Dessa forma, dada uma população, cuja variável aleatória X estuda o número de sucessos em n observações, coletamos uma amostra aleatória (n) e verificamos

o número de sucessos ocorridos (k). Assim, podemos obter a proporção (porcentagem) estimada de sucessos ($\hat{p} = \frac{k}{n}$) e concluir a respeito do valor suposto p_0 .

De acordo com o item 2.5, vejamos um exemplo.

Exemplo 5.7. Um Candidato A a Reitor da UFPR afirma que 57% (p_0) dos professores irão votar nele na próxima eleição. O Candidato B, desconfiado desse percentual, resolveu encomendar uma pesquisa de intenção de votos para verificar a autenticidade dessa afirmação. Após a coleta de uma amostra aleatória de 200 professores (n), constatou-se que 98 (k) tinham a intenção de votar no Candidato A. Segundo a pesquisa, qual a conclusão deveríamos tomar, ao nível de 0,05 de probabilidade, em relação à afirmação do Candidato A?

As Etapas:

1. As Hipóteses: $H_0 : p = p_0 = 0,57$ vs $H_1 : p < 0,57$.

Optamos por $H_1 : p < 0,57$, pois o Candidato B desconfiou da afirmação do Candidato A;

2. Distribuição Amostral: a Estatística \hat{p} segue uma distribuição aproximadamente normal, isto é,

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

Portanto, $z_{\text{tabelado}} = -1,64$.

Estimativas: $\hat{p} = \frac{k}{n} = \frac{98}{200} = 0,49$;

3. Nível de Significância: $\alpha = 0,05$

Estatística de teste: $z_{\text{calculado}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

Que, sob H_0 verdadeira, tem-se:

$$z_{\text{calculado}} = \frac{0,49 - 0,57}{\sqrt{\frac{0,57(1-0,57)}{200}}} = -2,285$$

Logo: P-Valor = $p(Z < -2,285) = 0,0112$

4. Região Crítica (RC): $RC = \{Z \in \mathbb{R} | Z < -1,645\} = \{\hat{p} \in \mathbb{R} | \hat{p} < 0,5124\}$;

Regra de Decisão: Se $z_{\text{calculado}} < -1,645$, ou se $\hat{p} < 0,5124$, rejeitamos H_0 .

5. Concluir o Teste: Como $\hat{p} < 0,5124$, podemos concluir que, provavelmente, a afirmação do Candidato A não é verdadeira.

5.6 Teste para a Comparação de duas Proporções Populacionais (p_1 e p_2).

É um teste similar ao anterior. Porém, comparamos, nesse caso, as proporções entre duas populações.

Exemplo 5.8. Os engenheiros da Indústria Romi de tornos anunciaram que um novo torno, desenvolvido por eles, produz eixos dentro das especificações com uma porcentagem maior do que seu concorrente para serem utilizados em automóveis. Uma pesquisa foi realizada para se verificar a autenticidade do anúncio. Foram retiradas duas amostras aleatórias independentes e encontrando-se: 171 eixos em 180 dentro da especificação para a ROMI, e 171 eixos em 190 para a concorrente. Qual conclusão devemos tomar, se admitirmos um nível de significância de 0,01?

Solução: As etapas do teste:

A variável associada ao teste é: X : número de eixos dentro da especificação em n_i ; i = Romi; Concorrente.

1. As Hipóteses:

$$H_0 : p_{Romi} = p_{Concorrente} \text{ vs } H_1 : p_{Romi} > p_{Concorrente}$$

Note que $H_1 : p_{Romi} > p_{Concorrente}$, pois os Engenheiros da Romi disseram que seus tornos são superiores em relação à variável X ;

2. Distribuição Amostral: a Estatística $\hat{p}_R - \hat{p}_C$ segue uma distribuição aproximadamente normal, isto é,

$$\hat{p}_R - \hat{p}_C \sim N(p_R - p_C; \frac{p_R(1-p_R)}{n_R} + \frac{p_C(1-p_C)}{n_C}).$$

Portanto, $z_{\text{tabelado}} = 2,33$.

Estimativas: $\hat{p}_R = \frac{k_R}{n_R} = \frac{171}{180} = 0,95$ e $\hat{p}_C = \frac{k_C}{n_C} = \frac{171}{190} = 0,90$

3. Nível de Significância: $\alpha = 0,01$

Estatística de teste:

$$Z_{\text{calculado}} = \frac{(\hat{p}_R - \hat{p}_C) - (p_R - p_C)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_R} + \frac{1}{n_C})}} \sim N(0,1)$$

Em que, $\hat{p} = \frac{n_R \hat{p}_R + n_C \hat{p}_C}{n_R + n_C} = \frac{k_R + k_C}{n_R + n_C}$.

Que, sob H_0 verdadeira, tem-se:

$$Z_{\text{calculado}} = \frac{(0,95 - 0,90) - (0)}{\sqrt{0,9243(1 - 0,9243)(\frac{1}{180} + \frac{1}{190})}}$$

Logo: $P - \text{Valor} = P(Z > 1,8176) = 0,0346$

4. Região Crítica (RC):

$$RC = Z \in \mathfrak{R} | Z > 2,33 = \{(\hat{p}_R - \hat{p}_C) \in \mathfrak{R} | (\hat{p}_R - \hat{p}_C) > 0,064\};$$

Regra de Decisão: Se $Z_{calculado} > 2,33$, ou se $(\hat{p}_R - \hat{p}_C) > 0,064$, rejeitamos H_0 .

5. Concluir o Teste: Como $(\hat{p}_R - \hat{p}_C) = 0,05 < 0,064$, podemos concluir que, provavelmente, o anúncio feito pelos engenheiros não procede.

5.7 Testes não Paramétricos

5.7.1 Teste de aderência

Verifica-se nesse teste a adequabilidade de um modelo probabilístico de uma variável X a um conjunto de dados observados, que serão divididos em categorias. Ou seja, verifica-se se os dados de uma amostra se ajustam, de forma satisfatória, a um modelo proposto. O princípio básico deste método é comparar proporções obtidas com auxílio das frequências obtidas dentro de cada categoria, isto é, comparam-se as possíveis divergências entre as frequências observadas e esperadas para um certo evento. Assim, as hipóteses testadas são:

1. Hipótese nula (H_0): X segue o modelo proposto. ou ainda, as frequências observadas são iguais às frequências esperadas, assim, não existe diferença entre as frequências (contagens) dos grupos.
2. Hipótese alternativa: X não segue o modelo proposto. Ou seja, existe ao menos uma frequência observada que difere de sua frequência esperada correspondente.

Nesse contexto, construiríamos uma tabela auxiliar da seguinte forma:

Tabela 5.10: Tabela auxiliar.

Categoria (k_i)	1	2	3	...	k
Freq. Observada	O_1	O_2	O_3	...	O_k
Freq. Esperada e_i	e_1	e_2	e_3	...	e_k

Karl Pearson propôs a seguinte fórmula para medir as possíveis discrepâncias entre proporções observadas e esperadas:

$$Q^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

em que,

k é o número de classes (categorias) consideradas;

o_i é a frequência observada;

e_i é a frequência esperada.

Note que as frequências observadas são obtidas diretamente dos dados das amostras, enquanto que as frequências esperadas são calculadas a partir do modelo probabilístico proposto pelo pesquisador, ou seja, em função de H_0 . além disso, como $(o_i - e_i)$ é a diferença entre a frequência observada e a esperada, quando as frequências observadas são muito próximas às esperadas, o valor de Q^2 tende a ser pequeno. Conseqüentemente, tendemos, também, a não rejeitar H_0 .

Supondo-se H_0 verdadeira, é possível demonstrar que a variável aleatória Q^2 segue uma distribuição aproximada qui-quadrado com q graus de liberdade. Assim,

$$Q^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \sim \chi_q^2$$

sendo que $q = k - 1$ representa o número de graus de liberdade.

A aproximação para o modelo Qui-Quadrado será melhor, para um n grande e se todas as frequências esperadas forem maiores do que 4 ($e_i \geq 5$). Caso isso não ocorra para alguma categoria, devemos combiná-la com outra categoria de forma que esse requisito seja satisfeito.

Com base nessa distribuição amostral, podemos, então, obter o P-Valor e, conseqüentemente, tomarmos uma decisão em relação à H_0 .

Exemplo 5.9. Deseja-se verificar se o número de acidentes em uma estrada muda conforme o dia da semana. O número de acidentes observado para cada dia de uma semana escolhida aleatoriamente foram:

Tabela 5.11: .Número de acidentes por dia da semana.

Dia da semana (k)	Domingo	Sábado	Sexta	Quinta	Quarta	Terça	Segunda
Número de acidentes	30	18	26	13	7	8	17

Solução: As etapas para realização do teste são:

1. Hipóteses a serem testadas:

H_0 : O número de acidentes não muda conforme o dia da semana;

H_1 : Pelo menos um dos dias tem número diferente dos demais.

Note que, se p_i representa a probabilidade de ocorrência de acidentes no i -ésimo dia da semana, então,

$H_0 : p_i = 1/7$ para todo $i = 1, \dots, 7$;

$H_1 : p_i \neq 1/7$ para pelo menos um valor de k .

2. Distribuição Amostral: Qui-quadrado com $k - 1$ g.l.

Portanto: $\chi^2 = 14,449$, para $\alpha = 0,05$.

Estimativa: Total de acidentes na semana: $n = 119$.

Logo, se H_0 for verdadeira, $e_i = 119 \times 1/7 = 17$; $i = 1, \dots, 7$. Ou seja, esperam-se, independentemente do dia, 17 acidentes.

Note que se trata de uma distribuição Uniforme.

3. Nível de Significância: $\alpha = 0,05$;

Tabela 5.12: Quadro auxiliar com as frequências esperadas.

Dia da semana (k)	Domingo	Sábado	Sexta	Quinta	Quarta	Terça	Segunda
Número de acidentes	30	18	26	13	7	8	17
Freq. Esperada	17	17	17	17	17	17	17
$(o_i - e_i)$	13	1	9	-4	-10	-9	0

Estatística de teste: .

$$Q^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

Assim,

$$Q^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_7 - e_7)^2}{e_7}$$

$$Q^2 = \frac{(13)^2}{17} + \frac{(1)^2}{17} + \dots + \frac{(0)^2}{17} = 26,35$$

Portanto, $\chi^2_{calculado} = 26,35$

Logo: P-Valor= $P(\chi^2 > 26,35) = 0,0002$

4. Região Crítica (RC): $RC = \{\chi^2 \in \mathbb{R}^* | \chi^2 > 14,449\}$ Regra de Decisão: Se $\chi^2_{calculado} > 14,449$ rejeita-se H_0 .
5. Concluir o Teste: Como $\chi^2_{calculado} = 26,35 > 14,449$, podemos concluir que o modelo probabilístico não é Uniforme. Assim, concluímos que o número de acidentes deve variar conforme o dia da semana.

5.7.2 Teste qui-quadrado para tabelas de contingência

O objetivo desse teste é verificar se existe independência entre duas variáveis medidas nas mesmas unidades experimentais.

Exemplo 5.10. Seja o exemplo: Deseja-se verificar se existe dependência entre a renda e o número de filhos em famílias de uma cidade. 250 famílias escolhidas ao acaso forneceram a Tabela 5.13:

Tabela 5.13: Renda e número de filhos por família em uma cidade.

Renda (R\$)	Número de filhos				
	0	1	2	+ de dois	Total
menos de 2000	15	27	50	43	135
2000 a 5000	8	13	9	10	40
5000 ou mais	25	30	12	8	75
Total	48	70	71	61	250

Tabela 5.14: Representação de duas características (A e B).

	B_1	B_2	\dots	B_s	Total
A_1	n_{11}	n_{12}	\dots	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	\dots	n_{2s}	$n_{2.}$
\dots	\dots	\dots	\dots	\dots	\dots
A_r	n_{r1}	n_{r2}	\dots	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	\dots	$n_{.s}$	$n_{..}$

Em geral, os dados referem-se a mensurações de duas características (A e B) feitas em n unidades experimentais, que são apresentadas conforme a seguinte tabela (Tabela 5.14):

As Hipóteses a serem testadas (Teste de independência) são:

$$\begin{cases} H_0 : & \text{A e B são variáveis independentes} \\ H_1 : & \text{As variáveis A e B não são independentes.} \end{cases}$$

Para testarmos H_0 , deveríamos verificar quantas observações (n_{ij}) devemos ter em cada casela. Dessa forma, se A e B forem independentes, temos que, para todos os possíveis (A_i e B_j):

$$P(A_i \cap B_j) = P(A_i) \times P(B_j) \quad \text{para } i = 1, 2, \dots, r \quad \text{e } j = 1, 2, \dots, s$$

Logo, o número esperado de observações com as características (A_i e B_j) entre as $n_{..}$ observações, sob a hipótese de independência, é dado por:

$$E_{ij} = n_{..} p_{ij} = n_{..} p_{i.} p_{.j} = n_{..} \frac{n_{i.} n_{.j}}{n_{..} n_{..}}$$

sendo p_{ij} a proporção de observações com as características (A_i e B_j). Assim,

$$E_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$$

O processo deve ser repetido para todas as caselas (i, j).

A distância entre os valores observados e os valores esperados sob a suposição de independência pode ser obtido pela Estatística do teste de independência dada por:

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^r \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

em que $O_{ij} = n_{ij}$ representa o total de observações na casela (i, j) .

Supondo H_0 verdadeira, temos que:

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^r \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_q^2.$$

sendo $q = (r - 1) \times (s - 1)$ graus de liberdade.

A Regra de decisão pode ter como base o nível descritivo P , neste caso,

$$P = P(\chi_q^2 \geq \chi_{obs}^2)$$

em que χ_{obs}^2 é o valor calculado, a partir dos dados, usando a expressão apresentada para χ^2 . Graficamente, teríamos (Figura 5.6):

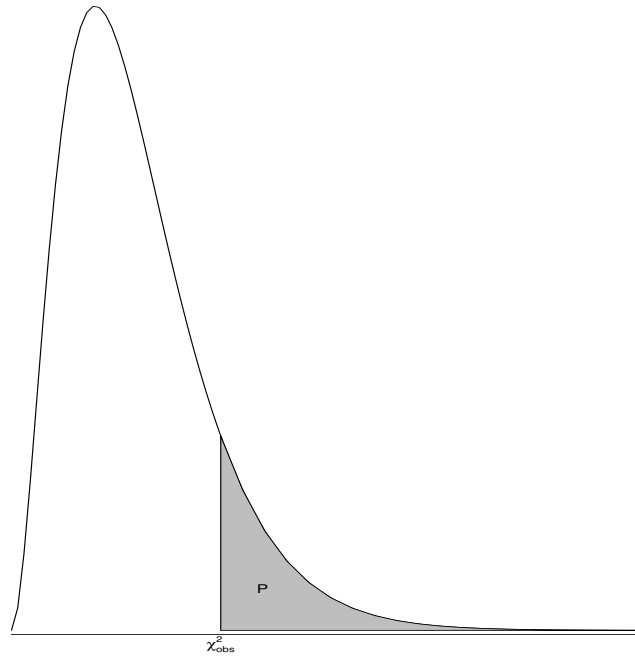


Figura 5.6: Gráfico da distribuição χ^2 .

Se, para α fixado, obtivermos $P \leq \alpha$, rejeitaremos a hipótese H_0 de independência.

Vejamos, então, a decisão que deveríamos tomar para o exemplo do estudo da dependência entre renda e o número de filhos. As hipóteses ficariam:

H_0 : O número de filhos e a renda são independentes;

H_1 : Existe dependência entre o número de filhos e a renda.

O cálculo dos valores esperados sob H_0 (independência), fica:

Número esperado de famílias sem filhos e renda menor que R\$ 2000:

$$E_{11} = \frac{48 \times 135}{250} = 25,92$$

Feito todos os cálculos, podemos construir a tabela de valores observados e esperados (entre parênteses). Assim,

Tabela 5.15: Número esperado para número de filhos e renda.

Renda (R\$)	Número de filhos				
	0	1	2	+ de dois	Total
menos de 2000	15 (25,92)	27 (37,80)	50 (38,34)	43 (32,94)	135
2000 a 5000	8 (14,40)	13 (21,00)	9 (21,30)	10 (18,30)	40
5000 ou mais	25 (7,68)	30 (11,20)	12 (11,36)	8 (9,76)	75
Total	48	70	71	61	250

Obtidos os valores esperados podemos proceder ao cálculo da estatística de qui-quadrado. Logo,

$$\begin{aligned}
 \chi_{obs}^2 &= \frac{(15 - 25,92)^2}{25,92} + \frac{(27 - 37,80)^2}{37,80} + \frac{(50 - 38,34)^2}{38,34} + \frac{(43 - 32,94)^2}{32,94} \\
 &+ \frac{(8 - 14,40)^2}{14,40} + \frac{(13 - 21,00)^2}{21,00} + \frac{(9 - 21,30)^2}{21,30} + \frac{(10 - 18,30)^2}{18,30} \\
 &+ \frac{(25 - 7,68)^2}{7,68} + \frac{(30 - 11,20)^2}{11,20} + \frac{(12 - 11,36)^2}{11,36} + \frac{(8 - 9,76)^2}{9,76} \\
 &= 36,621
 \end{aligned}$$

O número de graus de liberdade será determinado por: $q = (r-1) \times (s-1) = 2 \times 3 = 6$.

Sendo que:

Categorias de renda: $r = 3$;

Categorias de n° de filhos: $s = 4$.

Portanto, $\chi^2 \sim \chi_6^2$ e, supondo $\alpha = 0,05$, $P(\chi_6^2 \geq 36,62) = 0,0000021$

Como $P = 0,000 < \alpha = 0,05$, rejeitamos a independência entre número de filhos e renda familiar.

Capítulo 6

Correlação e Regressão Linear

6.1 Introdução

Considere a existência de uma variável quantitativa X a qual acreditamos apresentar alguma relação com uma outra variável quantitativa Y . Por exemplo: consumo de eletricidade e valor da conta de energia elétrica; idade e tempo de reação a um estímulo; temperatura e tempo de uma reação química, dentre outros.

Em situações como as citadas, a construção de um gráfico de dispersão dos valores de X *versus* os valores de Y , se constitui numa ferramenta estatística simples, porém muito útil, para investigar a existência de uma possível relação entre essas duas variáveis. Adicionalmente, podemos também fazer uso dos coeficientes de correlação, como por exemplo, o de Pearson, apresentado a seguir.

6.2 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é utilizado quando desejamos verificar a existência de associação linear entre duas variáveis quantitativas, X e Y , e é obtido dividindo-se a covariância de X e Y pelo produto dos respectivos desvios-padrão de ambas as variáveis, isto é:

$$\rho = \text{cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}. \quad (6.1)$$

Esse coeficiente resulta sempre em um valor entre -1 e 1 e sua interpretação depende do seu valor numérico e do seu sinal. Quanto mais próximo de -1 e 1 , mais forte é o grau de relação linear existente entre X e Y e, quanto mais próximo de 0 , mais fraco é o grau desta relação. Uma correlação linear negativa indica que quando o valor de uma variável aumenta, o valor da outra diminui e, uma correlação linear positiva, indica que quando o valor de uma variável aumenta, o valor da outra também aumenta.

Para uma amostra de tamanho n , em que para cada indivíduo i ($i = 1, \dots, n$) observamos os pares de valores (x_i, y_i) , o coeficiente de correlação linear entre X e Y é

calculado por:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^n (x_i - \bar{x})^2][\sum_{i=1}^n (y_i - \bar{y})^2]}}$$

$$= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{[\sum_{i=1}^n x_i^2 - n\bar{x}^2][\sum_{i=1}^n y_i^2 - n\bar{y}^2]}}$$

sendo \bar{x} e \bar{y} as médias amostrais dos x_i 's e y_i 's, respectivamente.

Os gráficos de dispersão apresentados na Figura 6.1 ilustram algumas situações com diferentes coeficientes de correlação. No gráfico (a) desta figura, podemos notar a ausência de associação entre X e Y . Já nos gráficos (b) e (c), podemos notar forte relação linear entre X e Y , pois os valores dos coeficientes de correlação de Pearson estão muito próximos de 1 e -1 , respectivamente. Na situação ilustrada no gráfico (b), à medida que os valores de uma variável crescem, os da outra também crescem, e isto ocorre de forma linear. Já na situação ilustrada no gráfico (c), à medida que os valores de uma variável crescem, os da outra decrescem, também de forma linear. Na situação mostrada no gráfico (d) podemos observar a ausência de relação linear entre X e Y . Neste caso, há a presença de uma relação quadrática, ou seja, não-linear entre elas.

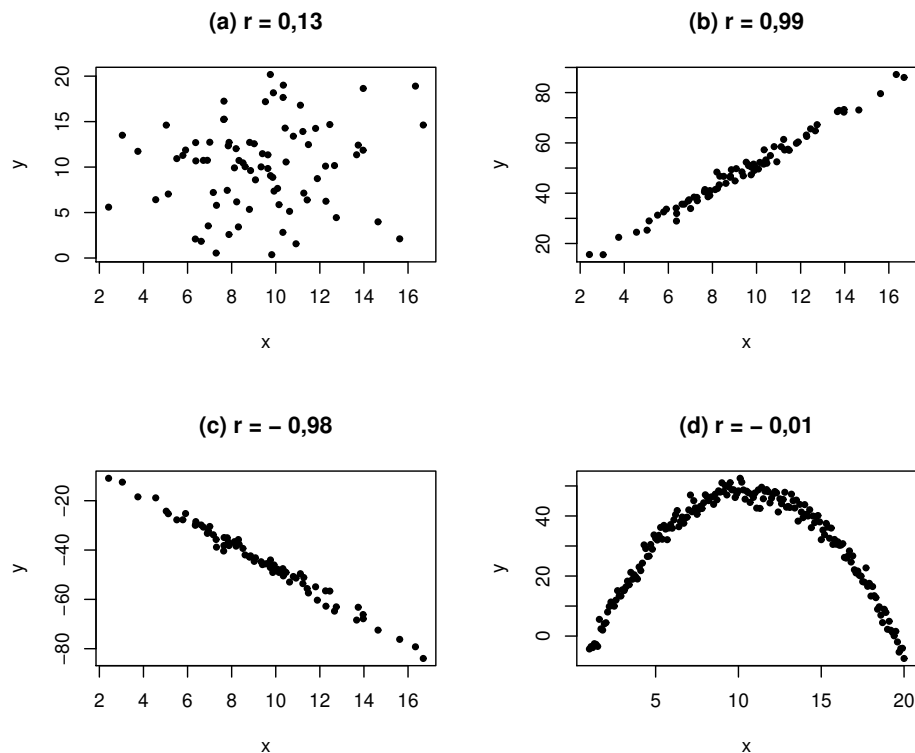


Figura 6.1: Gráficos de dispersão e coeficientes de correlação associados.

Do que foi apresentado, podemos observar que o coeficiente de correlação de Pearson é uma ferramenta útil para a investigação de relação *linear* entre duas variáveis quantitativas. A ausência de relação linear, quando indicada por este coeficiente, não implica

na ausência de relação entre elas. Outro tipo de relação pode estar presente, como, por exemplo, a não-linear.

6.2.1 Teste de significância para ρ

Na prática, desejamos testar se a associação linear entre X e Y é estatisticamente diferente de zero, bem como concluir a respeito desta associação não somente para a amostra em estudo, mas também para a população da qual a referida amostra foi extraída. Para tanto, uma estatística de teste bastante simples utilizada para testar as hipóteses:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \text{ (ou } H_1 : \rho > 0 \text{ ou } H_1 : \rho < 0) \end{cases}$$

é dada por:

$$T = r \sqrt{\frac{n-2}{1-r^2}}$$

a qual, sob H_0 , tem distribuição t -Student com $(n-2)$ graus de liberdade, sendo n o tamanho amostral e r o coeficiente de correlação linear entre X e Y . Calculado o valor dessa estatística, comparamos o mesmo com o valor tabelado, que é obtido a partir da Tabela t -Student a um nível de significância α pré-estabelecido. Se o valor calculado for maior que o tabelado, podemos rejeitar a hipótese nula ao nível de significância considerado. Rejeição da hipótese nula indica que a correlação linear observada na amostra é estatisticamente diferente de zero e que essa correlação pode ser inferida para a população da qual a mesma foi retirada.

Exemplo 6.1. Para uma amostra de tamanho $n = 80$, em que a relação entre duas variáveis quantitativas é de interesse, foi obtido para o coeficiente de correlação de Pearson o valor $r = 0,78$. Para testar se a correlação linear indicada por este coeficiente é estatisticamente diferente de zero, foi utilizado o teste de significância apresentado nesta seção. Com base nos resultados obtidos, isto é, $t_{cal} = 11,0$ e $t_{tab} = 1,99$ ao nível de 5% de significância (teste bilateral), podemos rejeitar a hipótese nula e, conseqüentemente, concluir que a correlação é estatisticamente diferente de zero, bem como que a mesma pode ser inferida para a população da qual a amostra foi extraída.

Uma vez constatada a existência de relação linear entre duas variáveis, é de usual interesse descrever essa relação por meio de uma equação linear. Na seção a seguir, discutimos esse assunto em detalhes.

6.3 Regressão Linear Simples

Se uma relação linear é válida para sumarizar a dependência observada entre duas variáveis quantitativas, então a equação que descreve esta relação é dada por:

$$Y = \beta_0 + \beta_1 X. \quad (6.2)$$

Os valores observados não se encontram, contudo, exatamente sobre esta linha reta, ou seja, existe uma diferença entre o valor observado e o valor fornecido pela equação. Esta diferença é denominada *erro* e é representada por ϵ . Este erro é assumido ser um *erro estatístico*, isto é, uma variável aleatória que quantifica a falha do modelo em ajustar-se aos dados exatamente. Tal erro pode ser devido ao efeito, dentre outros, de variáveis não consideradas e de erros de medição. Incorporando esse erro à equação (6.2) temos:

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad (6.3)$$

que é denominado modelo de regressão linear simples. Para cada indivíduo i ($i = 1, \dots, n$) na amostra, o modelo (6.3) fica representado por:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \quad (6.4)$$

A variável X em (6.3), denominada variável regressora ou independente, é considerada uma variável controlada pelo analista dos dados e medida com erro desprezível. Já Y , denominada variável resposta ou dependente, é considerada uma variável aleatória, isto é, existe uma distribuição de probabilidade para Y em cada valor possível de X . É muito freqüente, na prática, encontrarmos situações em que Y tenha distribuição Normal. Nesses casos, os erros ϵ_i (em que alguns são positivos e outros negativos) são assumidos serem normalmente distribuídos com média zero e variância constante desconhecida σ^2 , bem como independentes, isto é, o valor de um erro independe do valor de qualquer outro erro. Sendo assim, a média e a variância da variável Y serão, respectivamente:

$$\begin{aligned} E(Y | X = x) &= E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x \\ V(Y | X = x) &= V(\beta_0 + \beta_1 x + \epsilon) = \sigma^2. \end{aligned}$$

Exemplo 6.2. Um psicólogo investigando a relação entre a idade e o tempo que um indivíduo leva para reagir a um certo estímulo, obteve as informações apresentadas na Tabela 6.1

Tabela 6.1: Tempo de reação ao estímulo em função da idade.

y_i = tempo (em segundos)				x_i = idade (em anos)			
96	109	106	112	20	30	20	35
92	100	100	105	20	30	20	35
98	118	110	113	25	35	25	40
104	108	101	112	25	35	25	40
116	127	106	117	30	40	30	40

Fonte: Bussab, W. O. (1988).

A partir da representação gráfica desses dados mostrada na Figura 6.2, é possível visualizar uma relação linear positiva entre a idade e o tempo de reação. O coeficiente

de correlação de Pearson para esses dados resultou em $r = 0,768$, bem como seu respectivo teste de significância em $t_{calc} = 5,09$, que comparado ao valor tabelado $t_{tab,5\%} = 2,1$, fornece evidências de relação linear entre essas duas variáveis. Podemos, então, usar um modelo de regressão linear simples para descrever essa relação. Para isso, é necessário estimar, com base na amostra observada, os parâmetros desconhecidos β_0 e β_1 deste modelo. O método de estimação denominado Mínimos Quadrados Ordinários (MQO) é freqüentemente utilizado em regressão linear para esta finalidade e é apresentado a seguir.

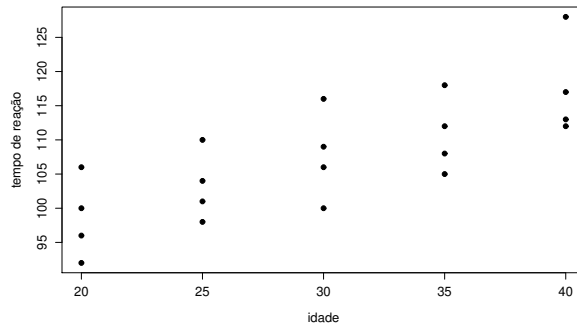


Figura 6.2: Idade *versus* tempo de reação a um estímulo.

6.3.1 Estimação dos parâmetros por MQO

Com base nos n pares de observações $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$, o método de estimação por MQO consiste em escolher β_0 e β_1 de modo que a soma dos quadrados dos erros, ϵ_i ($i = 1, \dots, n$), seja mínima. De (6.4) note que $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$. Para minimizar esta soma, que é expressa por:

$$SQ = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2, \quad (6.5)$$

devemos, inicialmente, diferenciar a expressão (6.5) com respeito a β_0 e β_1 e, em seguida, igualar a zero as expressões resultantes. Feito isso, e após algumas operações algébricas, os estimadores resultantes são:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i x_i - n \bar{y} \bar{x}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}, \end{aligned}$$

em que \bar{y} é a média amostral dos y_i 's e \bar{x} a média amostral dos x_i 's. Logo,

$$\widehat{E(Y | x)} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (6.6)$$

é o modelo de regressão linear simples ajustado, em que $\widehat{E(Y | x)}$, denotado também \hat{Y} por simplicidade, é o valor médio predito de Y para qualquer valor $X = x$ que esteja na

variação observada de X . No exemplo 6.2, a variação de X se encontra entre 20 e 40 anos e as estimativas dos parâmetros resultaram em $\hat{\beta}_0 = 80,5$ e $\hat{\beta}_1 = 0,9$.

Os desvios $e_i = y_i - \hat{y}_i$ ($i = 1 \cdots n$) são denominados *resíduos* e são considerados uma amostra aleatória dos erros. Por este fato, uma análise gráfica dos resíduos é, em geral, realizada para verificar as suposições assumidas para os erros ϵ_i .

6.3.2 Adequação do modelo de regressão linear ajustado

Após ajustar o modelo de regressão linear simples devemos, antes de adotá-lo definitivamente para fazer previsões (interpolações), verificar:

1. se o modelo se ajusta bem aos dados e,
2. se as suposições básicas se encontram satisfeitas.

Quanto a qualidade de ajuste do modelo, podemos fazer uso do coeficiente de determinação, R^2 , que nos fornece a porcentagem da variação total de Y explicada pelo modelo, ou seja, o percentual da variabilidade da variável dependente Y explicada pela variável independente X . Em regressão linear simples esse coeficiente pode ser obtido por $R^2 = r^2$, em que r é o coeficiente de correlação de Pearson amostral. O coeficiente de determinação varia de 0 a 1 (ou 0 a 100%), sendo que quanto mais próximo de 1 (100%), melhor o ajuste do modelo considerado.

Podemos, também, obter o coeficiente de determinação R^2 a partir da análise de variância da regressão, em que a variação total de Y é decomposta como mostrado no Quadro I. Fazendo-se uso da decomposição apresentada, temos que $R^2 = SQ_{Reg}/SQ_{Total}$.

Quadro I: Análise de Variância (ANOVA) da Regressão.

Fonte de Variação	g.l.	Soma de Quadrados	Quadrado Médio
Regressão	$p - 1$	$SQ_{Reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$QM_{Reg} = \frac{SQ_{Reg}}{(p-1)}$
Resíduos	$n - p$	$SQ_{Res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$QM_{Res} = \frac{SQ_{Res}}{(n-p)}$
Total	$n - 1$	$SQ_{Total} = \sum_{i=1}^n (y_i - \bar{y})^2$	

p = número de parâmetros do modelo e n = tamanho amostral.

Para testarmos a significância do parâmetro β_1 , o que, na prática, significa verificar se a covariável X influencia a resposta Y , testamos as hipóteses $H_0: \beta_1 = 0$ contra $H_0: \beta_1 \neq 0$. A estatística de teste utilizada para esta finalidade é dada por:

$$F = \frac{QM_{Reg}}{QM_{Res}},$$

em que QM_{Reg} e QM_{Res} são, respectivamente, os quadrados médios da regressão e dos resíduos apresentados no Quadro I. Sob H_0 , tal estatística tem distribuição F de Snedecor

com $p - 1$ e $n - p$ graus de liberdade. Assim, rejeitamos H_0 se o valor calculado de F for maior que o valor de F tabelado a um nível α de significância pré-estabelecido.

O Quadro II apresenta a ANOVA para os dados da Tabela 6.1. A partir desse quadro, temos que $R^2 = 0,589$, o que nos indica que 58,9% da variação total do tempo de reação está sendo explicada pela idade. Podemos também concluir pela rejeição de H_0 : $\beta_1 = 0$ ao nível de significância de 5%, visto que $F = 25,897 > F_{tab(1,18,5\%)} = 4,41$. Concluimos, assim, que a covariável idade realmente influencia o tempo de reação.

Quadro II: Análise de Variância da Regressão - Dados Tabela 6.1.

Fonte de Variação	g.l.	Soma de Quadrados	Quadrado Médio	F
Regressão	1	$SQ_{Reg} = 810$	$QM_{Reg} = 810$	25,897
Resíduos	18	$SQ_{Res} = 563$	$QM_{Res} = 31,28$	
Total	19	$SQ_{Total} = 1373$		

Quanto às suposições, devemos verificar se os erros encontram-se aleatoriamente distribuídos em torno de zero, bem como se a variância dos mesmos é constante e se são independentes. A suposição de independência está intimamente relacionada à forma como os dados foram coletados. Se o experimento foi conduzido de forma a garantir que as informações observadas em uma unidade amostral não tenham sido influenciadas pelas das outras unidades, então esta suposição é razoável. Por outro lado, o gráfico dos resíduos, e_i , *versus* os valores preditos pelo modelo, \hat{y}_i , bem como o gráfico dos resíduos *versus* os valores x_i , nos auxiliam a verificar se a média dos erros é zero e se a variância é constante. Para os dados do exemplo 6.2, a Figura 6.3 mostra ambos os gráficos.

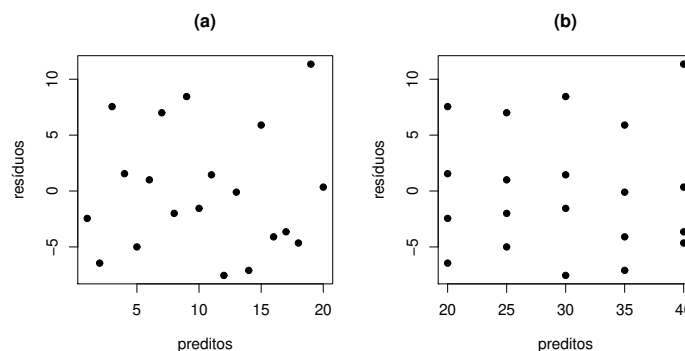


Figura 6.3: Análise gráfica dos resíduos associados ao modelo ajustado.

Note, a partir do gráfico (a) mostrado na Figura 6.3, que os resíduos encontram-se distribuídos aleatoriamente em torno de zero, indicando que a média dos mesmos se encontra próxima de zero. No gráfico (b), desta mesma figura, podemos observar que os resíduos, em $x = 20, 25, 30, 35$ e 40 , apresentam variabilidades semelhantes, indicando-nos que a variância dos erros pode ser considerada constante. Para verificar a suposição de que os erros seguem a distribuição Normal, um gráfico dos quantis teóricos *versus* os quantis amostrais dos resíduos, conhecido por QQplot, deve apresentar um comportamento próximo do linear. Para os dados do exemplo da idade e tempo de reação, obtivemos o

QQplot apresentado na Figura 6.4. A partir desta figura, notamos que a suposição de normalidade dos erros, e conseqüentemente da variável resposta Y , é razoável para os dados desse exemplo.

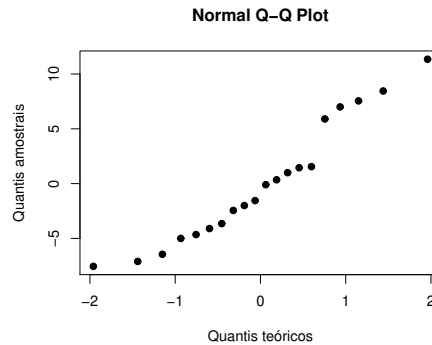


Figura 6.4: QQplot dos resíduos.

6.3.3 Interpretação dos parâmetros do modelo

Se o modelo de regressão linear simples (MRLS) for considerado adequado para descrever a relação linear entre Y e X , os coeficientes β_0 e β_1 são interpretados do seguinte modo:

- i) se a variação dos dados em X incluir $x = 0$, então o intercepto β_0 é a resposta esperada (média) em $x = 0$. Caso contrário, β_0 não apresenta interpretação prática;
- ii) o parâmetro β_1 é interpretado como a *mudança* no valor esperado de Y produzido por uma unidade de mudança em X .

Para os dados do exemplo apresentado na Tabela 6.1, o modelo de regressão linear simples ajustado é dado por:

$$\widehat{E(Y | x)} = 80,5 + 0,9 x.$$

Como a variação dos dados em X não inclui $x = 0$, não há interpretação prática do coeficiente $\hat{\beta}_0 = 80,5$. Por outro lado, $\hat{\beta}_1 = 0,9$ significa que a cada aumento de 1 ano na idade das pessoas, o tempo esperado de reação aumenta, em média, 0,9 segundos. Por exemplo, de 20 para 21 anos, estimamos um aumento no tempo de reação de, em média, 0,9 segundos.

Na Figura 6.5, apresentamos os tempos de reação registrados para as idades observadas, bem como os tempos médios de reação em cada uma dessas idades. O modelo de regressão linear simples ajustado, que podemos visualizar nesta mesma figura, apresenta um ajuste satisfatório aos tempos esperados de reação em função da idade.

De acordo com o modelo ajustado, estimamos, portanto, que o tempo de reação ao estímulo de pessoas com idade igual a 20 anos, seja de, em média, $\hat{y} = 80,5 + (0,9)(20) = 98,5$ segundos. Esse mesmo tempo para pessoas com 25 anos de idade é esperado ser, em

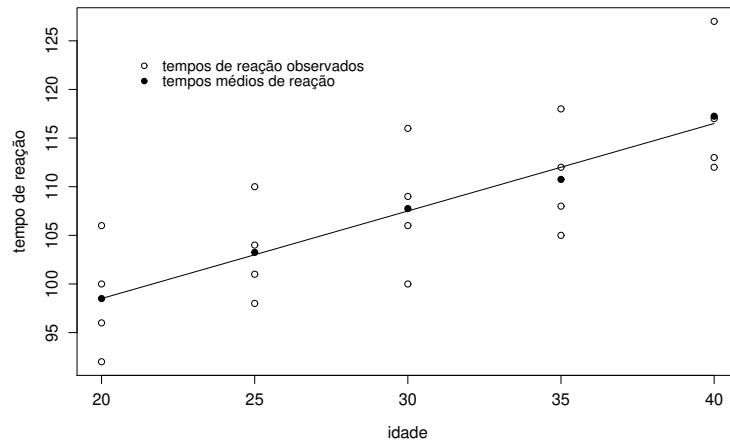


Figura 6.5: Tempos de reação em função da idade e MRLS ajustado.

média, 103 segundos. Estimativas para qualquer outra idade entre 20 e 40 anos podem ser obtidas de forma análoga.

As estimativas pontuais, \hat{y} , obtidas por meio do modelo de regressão linear simples ajustado, claramente não refletem a variação que certamente ocorre entre pessoas de uma mesma idade. Uma estimativa intervalar do tempo esperado de reação seria, desse modo, conveniente e recomendável. Este intervalo para uma idade $X = x$, a um nível de confiança de $(1 - \alpha)\%$, pode ser obtido por:

$$\hat{y} \pm t_{\alpha/2, n-2} \sqrt{\widehat{var}(\hat{y})}$$

sendo $t_{\alpha/2, n-2}$ o quantil $\alpha/2$ da distribuição t -Student com $(n - 2)$ graus de liberdade, n o tamanho amostral e $\widehat{var}(\hat{y})$, a variância de \hat{y} a qual é estimada por:

$$\widehat{var}(\hat{y}) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

em que:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{(n - 2)}.$$

Assim, se considerarmos pessoas com idade igual a $x = 28$ anos, estimamos que o tempo de reação delas seja de, em média, 105,7 segundos. Este tempo médio de reação, a um nível de confiança de 95%, pode variar entre 102,98 e 108,42 segundos.

Capítulo 7

Análise de Variância

7.1 Introdução

A Análise de Variância (ANOVA) é um procedimento utilizado para comparar três ou mais tratamentos. Existem muitas variações da ANOVA devido aos diferentes tipos de experimentos que podem ser realizados. Nesse curso será estudado apenas a análise de variância com um fator.

Inicialmente, são apresentados alguns conceitos utilizados em planejamento de experimentos e na análise de variância.

7.2 Conceitos Básicos sobre Experimentação

7.2.1 Tratamento

Um *tratamento* é uma condição imposta ou objeto que se deseja medir ou avaliar em um experimento. Normalmente, em um experimento, é utilizado mais de um tratamento. Como exemplos de tratamentos, podem-se citar: equipamentos de diferentes marcas, diferentes tamanhos de peças, doses de um nutriente em um meio de cultura, quantidade de lubrificante em uma máquina, temperatura de armazenamento de um alimento.

Os tratamentos que podem ser dispostos em uma ordem, como por exemplo, doses de nutrientes, quantidade de lubrificante, níveis de temperatura, são ditos tratamentos *quantitativos*. Já os tratamentos que não podem ser dispostos numa ordem, são ditos tratamentos *qualitativos*, por exemplo, variedades de plantas, métodos de preparação de alimento, marcas de equipamentos e outros.

Cada tipo de tratamento também pode ser chamado de um fator. Nesse texto, serão estudados somente experimentos com um fator de interesse.

O tipo de tratamento tem importância na forma como os dados serão analisados. Quando os tratamentos são quantitativos, pode-se usar, por exemplo, técnicas de análise de regressão.

Os tratamentos são chamados de *variáveis independentes*. Quando, em um experimento, estamos interessados em estudar apenas um tipo de variável independente, dizemos que possuímos apenas um fator. Em um experimento, um fator pode ter várias categoriais que são chamadas de *níveis*.

Exemplo: Um laboratório deseja estudar o efeito da composição de peças de metal sobre a dilatação.

Neste exemplo, a composição das peças é o fator (variável independente). Os diferentes tipos de composição são os níveis do fator. A dilatação das peças, medida em milímetros, por exemplo, é a variável resposta (variável dependente).

Em um experimento podem existir mais de um fator e mais de uma variável resposta.

Toda e qualquer variável que possa interferir na variável resposta ou dependente deve ser mantida constante. Quando isso não é possível, existem técnicas (estratégias) que podem ser utilizadas para reduzir ou eliminar essa interferência.

7.2.2 Unidade experimental ou parcela

Unidade experimental ou parcela é onde é feita a aplicação do tratamento. É a unidade experimental que fornece os dados para serem avaliados. Como exemplos de unidades experimentais ou parcelas pode-se citar: um motor, uma peça do motor, uma placa de Petri com meio de cultura, uma porção de algum alimento.

As unidades experimentais podem ser formadas por grupos ou indivíduos. Por exemplo, quando trabalha-se com cobaias, pode-se ter apenas uma cobaia como unidade experimental, ou seja, apenas um animal fornecerá a resposta do tratamento, ou ainda, pode-se ter um grupo de cobaias em uma gaiola fornecendo as informações. O uso de grupos ou indivíduos como unidades experimentais depende do fenômeno que se está estudando, da forma como o experimento é conduzido e dos recursos disponíveis. De modo geral, a escolha da unidade experimental deve ser feita de forma a minimizar o erro experimental.

7.2.3 Repetição

Repetição é o número de vezes que um tratamento aparece no experimento.

O número de repetições, em um experimento, vai depender também dos recursos disponíveis, do tipo de experimento (delineamento) e, também, da variabilidade do experimento ou da variável resposta. Existem várias metodologias para estimar o número satisfatório de repetições em um experimento. Mas, em função das possíveis limitações acima, a definição do número de repetições, muitas vezes, torna-se uma tarefa difícil. A experiência do pesquisador sobre o fenômeno em estudo deve ser levada em consideração. Além disso, as metodologias empregadas, para esse cálculo, pressupõem que uma estimativa do erro experimental é conhecida. Nem sempre essa informação está disponível

antes da realização de um experimento e, como cada experimento é uma nova história, em função de características intrínsecas de cada fenômeno, esse cálculo pode ser em vão.

7.2.4 Variável resposta ou variável dependente

Uma variável é qualquer característica que apresenta variação, por exemplo, a altura de pessoas, o peso de animais, o comprimento de uma peça, o número de microrganismos em um litro de leite etc.

Quando o valor de uma variável não pode ser determinado antes da realização de um experimento, tem-se então uma *variável aleatória*.

As variáveis que assumem valores enumeráveis, são denominadas variáveis aleatórias *discretas*. Por exemplo, o número de sementes germinadas, o número de microrganismos em um litro de leite.

As variáveis que assumem valores em um intervalo, são denominadas variáveis aleatórias *contínuas*. Por exemplo, o peso de animais, o teor de umidade em um alimento, o conteúdo de óleo em uma semente.

Em um experimento, podem ser medidas muitas variáveis, mas deve-se considerar somente aquelas que possam contribuir para a explicação da hipótese formulada.

É o pesquisador, em geral, quem sabe quais serão as variáveis que serão medidas em um experimento. Ele deve ser alertado, sempre, sobre as condições para a realização de tais medições, no sentido de evitar gastar recursos com variáveis que não fornecerão as informações para se testar a(s) hipótese(s). Quando o volume de dados de um experimento torna-se grande, aumentam os riscos de erros grosseiros, como de registro, de inversão de variáveis etc.

7.2.5 Delineamento experimental (Design)

Com a finalidade de reduzir o *erro experimental*, existem os chamados *delineamentos experimentais*. Um delineamento experimental é a forma como os tratamentos ou níveis de um fator são designados às unidades experimentais ou parcelas. A análise de variância (que será vista mais adiante) é baseada no delineamento experimental utilizado.

Por isso, saber como o experimento foi instalado e conduzido, é de fundamental importância. Pequenas modificações podem acarretar em grandes mudanças na forma da análise estatística. Não raro, acontecem situações em que as hipóteses formuladas, *a priori*, não podem ser testadas, ou ainda, é impossível de se realizar uma análise estatística. Por isso, deve-se dar muita importância ao planejamento experimental.

Um delineamento experimental é planejado de tal forma que a variação ao acaso seja reduzida o máximo possível. Alguns dos principais delineamentos experimentais são: delineamento completamente casualizado (DCC), delineamento em blocos casualizados (DBC) e quadrado latino.

7.2.6 Modelo e análise de variância

Em um experimento completamente casualizado, cada observação Y_{ij} pode ser decomposta conforme o modelo a seguir:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad i = 1, \dots, I \text{ e } j = 1, \dots, J \quad (7.1)$$

em que:

Y_{ij} é a observação do i -ésimo tratamento na j -ésima unidade experimental ou parcela;

μ é o efeito constante (média geral);

τ_i é o efeito do i -ésimo tratamento;

ϵ_{ij} é o erro associado ao i -ésimo tratamento na j -ésima unidade experimental ou parcela assumido como: $\epsilon_{ij} \stackrel{IID}{\sim} N(0, \sigma^2)$. Aqui, *IID* significa que os erros devem ser independentes e identicamente distribuídos.

Em um experimento, existe o interesse em testar se há diferenças entre as médias dos tratamentos, o que equivale a testar as hipóteses:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_I \\ H_1 : \mu_i \neq \mu_{i'} \text{ para pelo menos um par } (i, i'), \text{ com } i \neq i' \end{cases}$$

em que:

$$\mu_i = \mu + \tau_i \quad i = 1, 2, \dots, I.$$

De forma equivalente, podemos escrever tais hipóteses da seguinte forma:

$$\begin{cases} H_0 : \tau_1 = \tau_2 = \dots = \tau_I = 0 \\ H_1 : \tau_i \neq 0 \text{ para pelo menos um } i. \end{cases}$$

Note que, se a hipótese nula for verdadeira, todos os tratamentos terão uma média comum μ .

A análise de variância, baseia-se na decomposição da variação total da variável resposta em partes que podem ser atribuídas aos tratamentos (variância entre) e ao erro experimental (variância dentro). Essa variação pode ser medida por meio das somas de quadrados definidas para cada um dos seguintes componentes:

$$SQ_{Total} = \sum_{i=1}^I \sum_{j=1}^J y_{ij}^2 - C, \text{ em que } C = \frac{(\sum_{i=1}^I \sum_{j=1}^J y_{ij})^2}{IJ},$$

$$SQ_{Trat} = \frac{\sum_{i=1}^I y_i^2}{J} - C,$$

e a soma de quadrados dos resíduos pode ser obtida por diferença:

$$SQ_{Res} = SQ_{Total} - SQ_{Trat}.$$

A SQ_{Trat} também é chamada de variação Entre, que é a variação existente entre os diferentes tratamentos e a SQ_{Res} é chamada de variação Dentro que é função das diferenças existentes entre as repetições de um mesmo tratamento.

Essas somas de quadrados podem ser organizadas em uma tabela, denominada tabela da análise de variância, como apresentado na Tabela 7.1.

Para testar a hipótese H_0 , utiliza-se o teste F apresentado na tabela da Análise de Variância (Tabela 7.1). Convém lembrar que esse teste é válido se os pressupostos assumidos para os erros do modelo estiverem satisfeitos.

Tabela 7.1: Tabela da análise de variância.

Causas de Variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	F calculado
Tratamentos	I-1	SQTrat	QMTrat	QMTrat/QMRes
Resíduo	I(J-1)	SQRes	QMRes	
Total	IJ-1	SQTotal		

em que $QMTrat = SQTrat / (I-1)$ e $QMRes = SQRes / (I(J-1))$.

Pode-se mostrar que o quociente $QMTrat/QMRes$ tem distribuição F com $(I-1)$ e $I(J-1)$ graus de liberdade, supondo que y_{ij} sejam variáveis aleatórias independentes, todos os tratamentos têm variâncias iguais a σ^2 e $Y_{ij} \sim N(\mu_i, \sigma^2)$. Por esses motivos, os pressupostos da ANOVA devem ser testados ou avaliados em qualquer análise.

Se $F_{calculado} > F_{tabelado}$, rejeitamos a hipótese de nulidade H_0 , ou seja, existem evidências de diferença significativa entre pelo menos um par de médias de tratamentos, ao nível α de significância escolhido. Caso contrário, não rejeita-se a hipótese de nulidade H_0 , ou seja, não há evidências de diferença significativa entre tratamentos, ao nível α de significância escolhido.

Outra maneira de avaliar a significância da estatística F é utilizando o p-valor. Se o $p\text{-valor} < \alpha$, rejeitamos a hipótese de nulidade H_0 . Caso contrário, não se rejeitamos a hipótese de nulidade H_0 , ou seja, não há evidências de diferenças significativas entre os tratamentos, ao nível α de significância escolhido.

7.2.7 Delineamento experimental

Quando as unidades experimentais são homogêneas, ou seja, as parcelas são uniformes, os tratamentos podem ser sorteados nas unidades experimentais sem qualquer restrição. Nessa situação, o delineamento experimental é chamado de delineamento completamente casualizado (DCC). Neste caso, todos os tratamentos têm a mesma chance de serem aplicados em qualquer unidade experimental ou parcela. Nesse texto, abordaremos apenas esse tipo de delineamento que é o caso mais simples da ANOVA.

7.3 Análise de Variância

Exemplo 7.1. Considere o seguinte experimento que foi conduzido, considerando um delineamento inteiramente casualizado. Foram comparados 4 tratamentos (tipos de cultivo:

Ágar (A), Cássia (C), Guar (G), Leucena (L)). Mediu-se o crescimento, em gramas, de explantes de morango (Tabela 7.2).

Tabela 7.2: Crescimento de explantes de morangos em gramas.

Trat.	Repetições								Total
	I	II	III	IV	V	VI	VII	VIII	
A	0.1958	0.1301	0.1806	0.1545	0.1252	0.1882	0.2211	0.1734	1,3689
G	0.3627	0.4841	0.4119	0.4457	0.4755	0.5174	0.4173	0.4001	3,5147
L	0.1621	0.1150	0.2011	0.2123	0.1475	0.1922	0.1802	0.2248	1,4352
C	0.2841	0.3099	0.2922	0.1505	0.2345	0.1652	0.1379	0.1960	1,7703
Total									8,0891

Para este experimento, consideramos o modelo:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \text{em que} \quad \epsilon_{ij} \stackrel{IID}{\sim} N(0, \sigma^2)$$

$i = 1, 2, \dots, 4$ tratamentos;

$j = 1, 2, \dots, 8$ repetições;

y_{ij} é o peso em gramas correspondente ao i -ésimo tratamento na j -ésima unidade experimental;

τ_i é o efeito do i -ésimo tratamento;

ϵ_{ij} é o erro experimental associado ao i -ésimo tratamento e a j -ésima repetição.

As hipóteses testadas neste experimento são:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4$$

$$H_1 : \tau_i \neq \tau_{i'} \quad \text{para pelo menos um par, com } i \neq i'.$$

Cálculos para a Análise de Variância

Tem-se que:

$$\sum_{i=1}^I \sum_{j=1}^J y_{ij} = 0,1958 + 0,1301 + \dots + 0,1960 = 8,0891.$$

$$\sum_{i=1}^I \sum_{j=1}^J y_{ij}^2 = 0,1958^2 + 0,1301^2 + \dots + 0,1960^2 = 2,4952.$$

$$\text{Graus de liberdade de tratamentos} = I - 1 = 4 - 1 = 3.$$

$$\text{Graus de liberdade do resíduo} = I(J - 1) = 4(8 - 1) = 28.$$

$$\text{Graus de liberdade total} = IJ - 1 = 4 \times 8 - 1 = 31.$$

As somas de quadrados são obtidas da seguinte forma:

$$1. \text{ SQTotal} = \sum_{i=1}^I \sum_{j=1}^J y_{ij}^2 - \frac{(\sum_{i=1}^I \sum_{j=1}^J y_{ij})^2}{IJ} = 2,4952 - \frac{(8,0891)^2}{32} = 0,4504$$

Obs: A expressão $\frac{(\sum_{i=1}^I \sum_{j=1}^J y_{ij})^2}{IJ}$ é referenciada em alguns textos como fator de correção da soma de quadrados.

$$2. \text{SQTrat} = \frac{\sum_{i=1}^I y_{i.}^2}{J} - \frac{(\sum_{i=1}^I \sum_{j=1}^J y_{ij})^2}{IJ} = \frac{1,3689^2 + 1,7703^2 + 3,5147^2 + 1,4352^2}{8} - \frac{(8,0891)^2}{32} = 0,3828.$$

3. A Soma de Quadrados dos resíduos é obtida por diferença:

$$\text{SQRes} = \text{SQTotal} - \text{SQTrat} = 0,4504 - 0,3828 = 0,0676.$$

Os quadrados médios são obtidos pela divisão da soma de quadrados, pelos seus respectivos graus de Liberdade. Assim,

$$\text{QMTrat} = \text{SQTrat} / (I-1) = 0,3828 / 3 = 0,1276 \text{ e}$$

$$\text{QMRes} = \text{SQRes} / I(J-1) = 0,0676 / 28 = 0,002414.$$

O teste F é o quociente entre o QMTrat e o QMRes. Logo,

$$F_{\text{calculado}} = \text{QMTrat} / \text{QMRes} = 0,1276 / 0,002414 = 52,8583.$$

O $F_{\text{calculado}}$ é comparado com o F_{tabelado} , com 3 e 28 graus de liberdade, na tabela de F (Tabela):

F_{tabelado} a 1% = 2,95

F_{tabelado} a 5% = 4,57.

Efetuada os cálculos, podemos resumi-los na tabela da análise de variância apresentada a seguir:

Tabela 7.3: Análise de variância do exemplo 7.1.

Causas de Variação	GL	Soma de Quadrados	Quadrados Médios	F calculado
Tratamentos	4-1=3	0,3828	0,1276	52,8583**
Resíduo	4(8-1)=28	0,0676	0,002414	
Total	4×8-1=31	0,4504		

** Significativo ao nível de 1% de probabilidade

Conclusão da análise de variância: de acordo com o teste F, foram encontradas evidências de diferenças significativas, ao nível de 1% de probabilidade, entre os tratamentos, com relação ao crescimento. Rejeitamos, portanto, a hipótese de nulidade H_0 . Deve existir, pelo menos, um contraste significativo entre as médias de tratamentos, com relação ao crescimento médio.

O procedimento seguinte, quando de interesse do pesquisador, é o de comparar as médias de tratamentos utilizando algum teste de comparação de médias ou contrastes para identificar qual ou quais tratamentos é ou são diferente(s).

7.4 Teste de Tukey para Comparação de Médias

Após concluirmos que existe diferença significativa entre tratamentos, por meio do teste F, podemos estar interessados em avaliar a magnitude destas diferenças utilizando um teste de comparações múltiplas. Será utilizado o teste de Tukey.

O teste de Tukey permite testar qualquer contraste, sempre, entre duas médias de tratamentos, ou seja, não permite comparar grupos entre si.

O teste baseia-se na Diferença Mínima Significativa (DMS) Δ . A estatística do teste é dada da seguinte forma:

$$\Delta = q \sqrt{\frac{QMRes}{r}}, \quad (7.2)$$

em que q é a amplitude total studentizada, tabelada (tabela 7), $QMRes$ é o quadrado médio do resíduo, e r é o número de repetições. O valor de q depende do número de tratamentos e do número de graus de liberdade do resíduo. Também, em um teste de comparações de médias, deve-se determinar um nível de significância α para o teste. Normalmente, utiliza-se o nível de 5% ou 1% de significância.

Como o teste de Tukey é, de certa forma, independente do teste F, é possível que, mesmo sendo significativo o valor de $F_{calculado}$, não se encontrem diferenças significativas entre contrastes de médias.

Aplicando o teste de Tukey às médias dos tratamentos do exemplo 7.1, temos:

$$\Delta(5\%) = 3,85 \sqrt{\frac{0,00242}{8}} = 0,06696.$$

sendo

$q=3,85$ e $\alpha = 0,05$

Se o contraste for maior do que Δ , então as médias diferem ao nível α de significância.

Utilizar-se-á o método de letras para exemplificar o uso do teste, mas existem outras maneiras de representação como, por exemplo, o uso de tabelas ou barras.

Inicialmente, ordenamos as médias de forma crescente ou decrescente, para facilitar as comparações. Colocamos uma letra do alfabeto na primeira média (normalmente a letra 'a') e, em seguida, comparamos a diferença com as médias seguintes. Se a diferença for superior ao valor de $\Delta(5\%) = 0,06696$, a diferença entre duas médias será considerada significativa. Sendo representada pela presença de letras diferentes. O resultado final é o seguinte:

\bar{G}	0,4393	a
\bar{C}	0,2213	b
\bar{L}	0,1794	b
\bar{A}	0,1711	b

Temos que, médias de crescimento seguidas de letras iguais, não diferem significativamente entre si, pelo teste de Tukey, ao nível de 5% de probabilidade.

7.5 Teste de Kruskal-Wallis

A análise de variância exige que os erros ϵ_{ij} tenham distribuição Normal e deve haver homocedasticidade entre os tratamentos (variâncias homogêneas). Estes pressupostos nem sempre são satisfeitos em um experimento ou conjunto de dados.

Como uma alternativa para a análise de variância paramétrica para um delineamento completamente casualizado, $k \geq 3$ tratamentos, existe o teste não paramétrico de Kruskal-Wallis. Este teste pode ser utilizado para testar a hipótese $H_0 : \tau_1 = \tau_2 = \dots = \tau_k$. No lugar das medidas, utiliza-se os postos e não há suposições com relação a Normalidade e Homocedasticidade.

Uma exigência do teste de Kruskal-Wallis é que a variável em estudo seja contínua. Outra é que as observações devem ser independentes. A análise consiste em obter o posto de cada uma das observações. Adota-se que o menor valor recebe (ranking ou posto) 1, o segundo 2 e assim por diante, até que todas as observações tenham sido consideradas. Quando ocorrerem empates, atribui-se o valor médio entre as observações, ou seja, atribui-se a média das ordens que seriam atribuídas a elas se não ocorresse o empate. Se, por exemplo, as duas menores observações forem iguais há um empate. Neste caso, cada uma recebe o posto 1,5 que é a média dos valores 1 e 2.

Para testar a hipótese nula, utilizamos a estatística de teste:

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

em que:

N = número total de observações;

k = número de tratamentos;

n_j = número de observações no j -ésimo tratamento;

R_j = soma dos postos do j -ésimo tratamento.

Rejeitamos H_0 se $H \geq \chi^2$ com $k-1$ graus de liberdade ao nível α de significância.

Se ocorrerem empates, a estatística de teste H deve ser corrigida com a seguinte expressão:

$$C = 1 - \frac{\sum (t_i^3 - t_i)}{N^3 - N},$$

em que t_i é o número de observações empatadas no i -ésimo grupo.

Assim, temos a estatística corrigida:

$$H_1 = \frac{H}{C}$$

Para testar H_0 , procedemos exatamente como se não houvesse empates.

Exemplo 7.2. Em um experimento para avaliar o consumo de energia elétrica em KWh de três motores durante um hora de funcionamento, obteve-se os seguintes resultados:

Aplicando-se o teste de Kruskal-Wallis, temos que:

Tabela 7.4: Consumo de energia elétrica de três motores durante uma hora.

Motor 1	Motor 2	Motor 3
2212 (13)	2195 (12)	1770 (4)
2025 (9)	2031 (11)	1800 (5)
1989 (8)	1876 (7)	1852 (6)
2232 (14)	1750 (2)	1769 (3)
2027 (10)	1060 (1)	
$R_1 = 54$	$R_1 = 33$	$R_1 = 18$

$$H = \frac{12}{14(15)} \left[\frac{54^2}{5} + \frac{33^2}{5} + \frac{18^2}{4} \right] - 3(15) = 5,4$$

O valor χ^2 , com $k - 1 = 3 - 1 = 2$ graus de liberdade e um nível de significância de 5% é 5,99. portanto, não rejeitamos H_0 , ou seja, não há evidências de que os motores possuem um consumo diferente de energia elétrica.

Capítulo 8

Controle Estatístico de Qualidade

8.1 Introdução

A análise e interpretação de dados, voltados para a melhoria da qualidade de produtos ou serviços, é chamada de **controle estatístico de qualidade** ou **controle estatístico de processos (CEP)**.

No controle estatístico de qualidade existem vários métodos que podem ser utilizados para monitorar um processo: desde a estatística descritiva ou um simples gráfico de dispersão, até métodos mais específicos como gráficos de controle e índices de capacidade. Neste texto, abordaremos apenas alguns tipos de gráficos de controle.

8.1.1 Gráficos de controle

Basicamente, um gráfico de controle consiste no acompanhamento de um processo ao longo do tempo. Um linha média é inserida no gráfico acompanhada de uma linha superior e uma linha inferior, chamadas de limite superior e limite inferior de controle, respectivamente. Esses limites são construídos segundo critérios estatísticos. Na Figura 8.1 é mostrado um exemplo de gráfico de controle.

Diz-se que um processo está *sob controle* ou estável quando a variação observada é devida somente a causas de variação naturais do processo, ou seja, o gráfico mostra apenas flutuações aleatórias. Caso contrário, o processo é dito *fora de controle*.

O objetivo principal de um gráfico de controle é identificar se a variação existente é uma função de causas naturais de variação do processo ou de causas especiais. No caso de causas especiais, é necessário intervir no processo para reduzir a variabilidade.

Os gráficos de controle são elaborados em função do tipo de variável e da característica da amostra. Por exemplo, para atributos ou variáveis do tipo contagem e para variáveis contínuas são elaborados gráficos para cada tipo de variável. Também, dependendo do tamanho da amostra (n), metodologias específicas devem ser utilizadas.



Figura 8.1: Gráfico de controle: ideia básica.

8.1.2 Construção do gráfico

No eixo horizontal são inseridos os números das amostras de tamanho n (em geral constante), observadas no processo. No eixo vertical é inserida a unidade de medida da variável que está sendo estudada ou controlada. As amostras são chamadas também de subgrupos.

No gráfico, são representadas as médias de cada amostra que irão refletir o comportamento de variação do processo.

A linha central, dependendo das informações disponíveis, representa, em geral, a média do processo. A linha central ou o valor médio \bar{x} é obtido pela média das médias amostrais e pode, também, ser denotada por $\bar{\bar{x}}$. Se existem informações anteriores sobre o processo, pode-se utilizar um valor de referência. Por último, se for conhecida, a média da população, μ deve ser utilizada.

As duas linhas que definem os limites de controle: LSC - limite superior de controle e LIC - limite inferior de controle, são utilizadas, entre outras, para decidir quando o processo está sob controle ou não.

Não se deve confundir limites de controle com limites de especificação. Os limites de especificação são definidos pela natureza do produto. É uma questão técnica, definida pelo projeto do produto. Os limites de controle são sempre menores do que os limites de especificação. Os limites de controle, por outro lado, podem ser definidos de duas maneiras:

1. Utilizando a distribuição da variável X que mede o desempenho do processo. Nesse caso, podemos encontrar limites de controle de tal forma que,

$$P(LIC \leq X \leq LSC) \geq 1 - \alpha$$

sendo α um número arbitrário e fixo, normalmente pequeno ($\alpha = 0,01$). Esse limite é chamado de limite probabilístico. Esperamos que, por α ser pequeno, um valor além dos limites inferior e superior devam ocorrer raramente se o processo estiver sob controle. Nesse caso, um ponto fora dos limites indicará que o processo está fora de controle e que uma ou mais causas de variação especiais estão atuando sobre o processo.

2. Outra maneira é definir os limites de controle por meio de múltiplos do desvio padrão da variável X :

$$LSC = \mu_x + k\sigma_x$$

$$LIC = \mu_x - k\sigma_x,$$

em que μ_x é a média de X , σ_x é o desvio padrão de X e k é uma constante positiva. Em geral, costuma-se utilizar $k = 3$.

Se X tiver uma distribuição Normal, a probabilidade de um ponto cair fora dos limites é aproximadamente 0,003.

O tamanho da amostra e o tipo de variável influenciam na construção dos limites de controle.

É importante que os limites sejam definidos para um processo que esteja sob controle. Ou seja, como os limites dependem da variabilidade do processo (σ), se o processo não estiver sob controle, ou seja, quando a variação observada é devida somente a causas de variação naturais do processo, os limites podem não refletir o verdadeiro comportamento do processo.

Limites de aviso ou de alerta

Podemos utilizar, além dos limites superior e inferior, limites de aviso. Quando um ponto ultrapassa os limites de $\pm 3\sigma$, ações corretivas devem ser utilizadas. No entanto, um limite menor, por exemplo, 2σ pode ser utilizado como um limite de advertência (Figura 8.2).

Quando um ponto atinge os limites de aviso, podemos, por exemplo, coletar amostras com um n maior e/ou amostras com maior frequência, para obter informações sobre o processo mais rapidamente.

Limites de aviso aumentam a sensibilidade do gráfico de controle. Por outro lado, esses mesmos limites podem gerar alarmes falsos ou falsos positivos. Por isso, eles devem ser utilizados com cautela, pois podem, inclusive, aumentar custos sem necessidade.

8.1.3 Análise do padrão de gráficos de controle

Estabelecidos os limites de controle, devemos analisar e interpretar as informações fornecidas pelo gráfico.

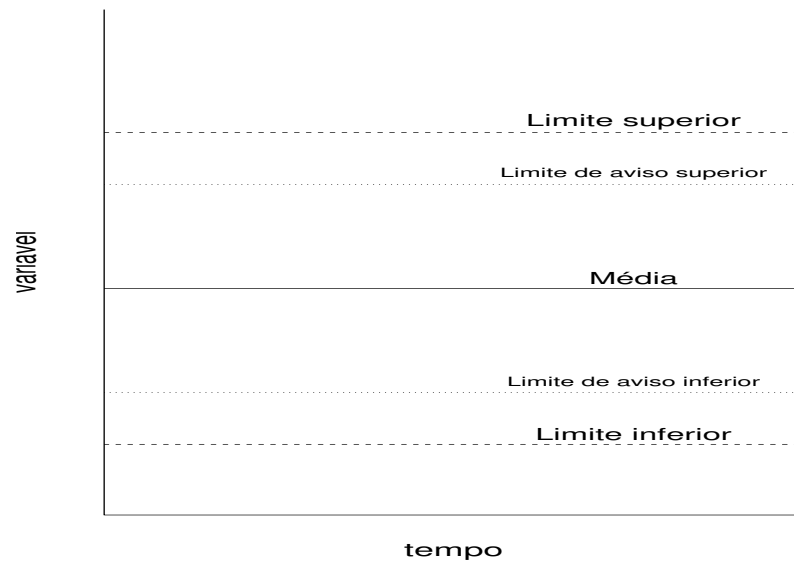


Figura 8.2: Gráfico de controle: limites de aviso.

Se todos os pontos estiverem dentro dos limites de controle, isso não indica, necessariamente, que o processo esteja sob controle. A ausência de aleatoriedade nos pontos pode indicar uma fonte de variação especial.

Por exemplo, na Figura 8.3 há somente 2 pontos abaixo da média e 8 acima. Em geral, espera-se uma distribuição proporcional dos pontos acima e abaixo da média.

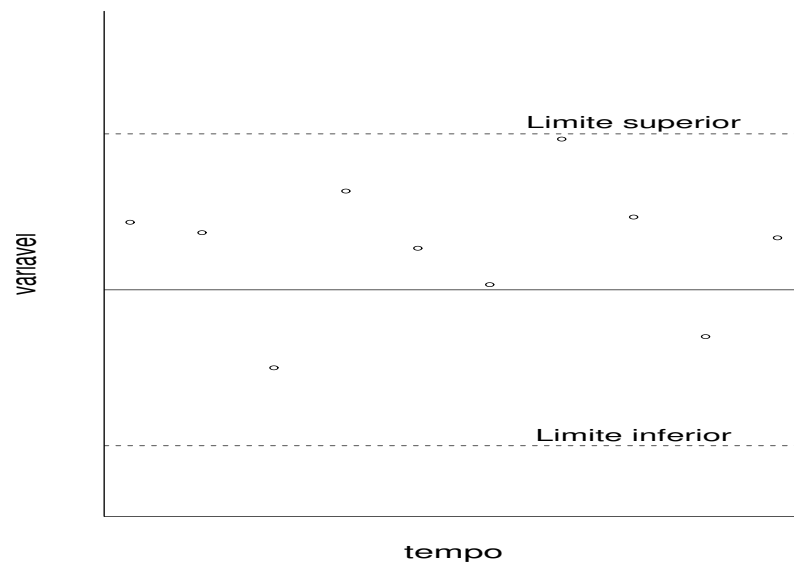


Figura 8.3: Gráfico de controle: processo tendencioso.

Na Figura 8.4 observamos um comportamento cíclico. Nesse caso, o gráfico pode indicar problemas no processo, como por exemplo, o cansaço de operadores.

Obviamente, quando um ou mais pontos ultrapassam os limites, o processo deve ser analisado. Claro que outras regras podem ser adotadas, dependendo do processo do tipo

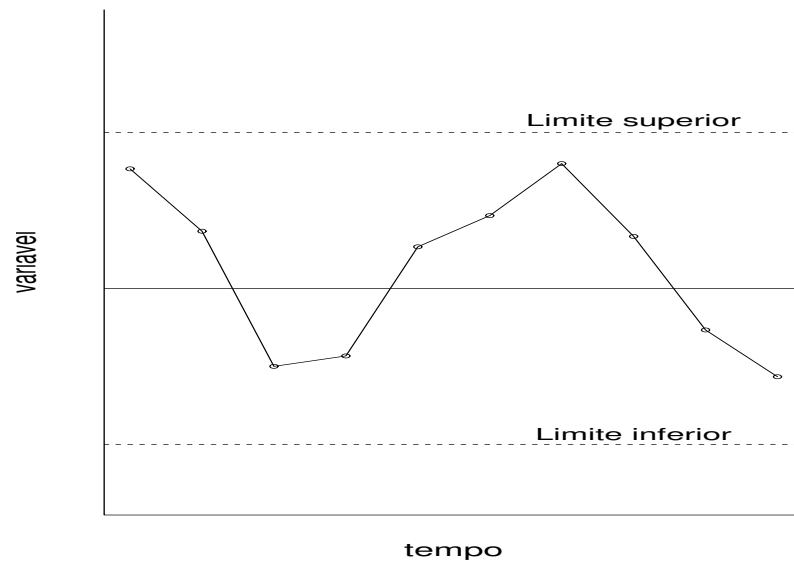


Figura 8.4: Gráfico de controle: processo cíclico.

de variável, entre outras.

Algumas regras podem ser (Montgomery, 1991):

1. um ou mais pontos fora dos limites de controle;
2. dois de três pontos consecutivos além dos limites de 2σ (mas ainda dentro dos limites de controle);
3. quatro de cinco pontos consecutivos além do limite de 1σ ;
4. oito pontos consecutivos acima ou abaixo da linha central;
5. seis pontos em uma linha crescente ou decrescente;
6. quinze pontos em linha, acima ou abaixo da média;
7. quatorze pontos alternados para cima e para baixo;
8. oito pontos em linha em ambos os lados da linha central e nenhum em até 1σ ;
9. um padrão incomum ou não aleatório;
10. um ou mais pontos próximos dos limites de controle.

Em resumo, um gráfico de controle ajuda a:

- monitorar e reduzir a variabilidade;
- monitorar um processo ao longo do tempo;
- detectar rapidamente pontos fora de controle e tendências;
- realizar correções, diminuindo o número de produtos defeituosos.

8.2 Gráficos de Controle para Variáveis

Introdução

Quando trabalhamos com variáveis do tipo comprimento, largura, diâmetro, ou seja, variáveis que possuem uma escala contínua, são utilizados gráficos de controle para variáveis, que serão apresentados nessa seção.

No caso de variáveis, existem três gráficos usuais: \bar{x} , R e s .

O gráfico \bar{x} é utilizado para monitorar a média do processo, enquanto os gráficos R e s são utilizados para monitorar a variabilidade do processo.

Nos gráficos da Figura 8.5, a seguir, pode-se perceber a importância de se monitorar a média e a variância de um processo.

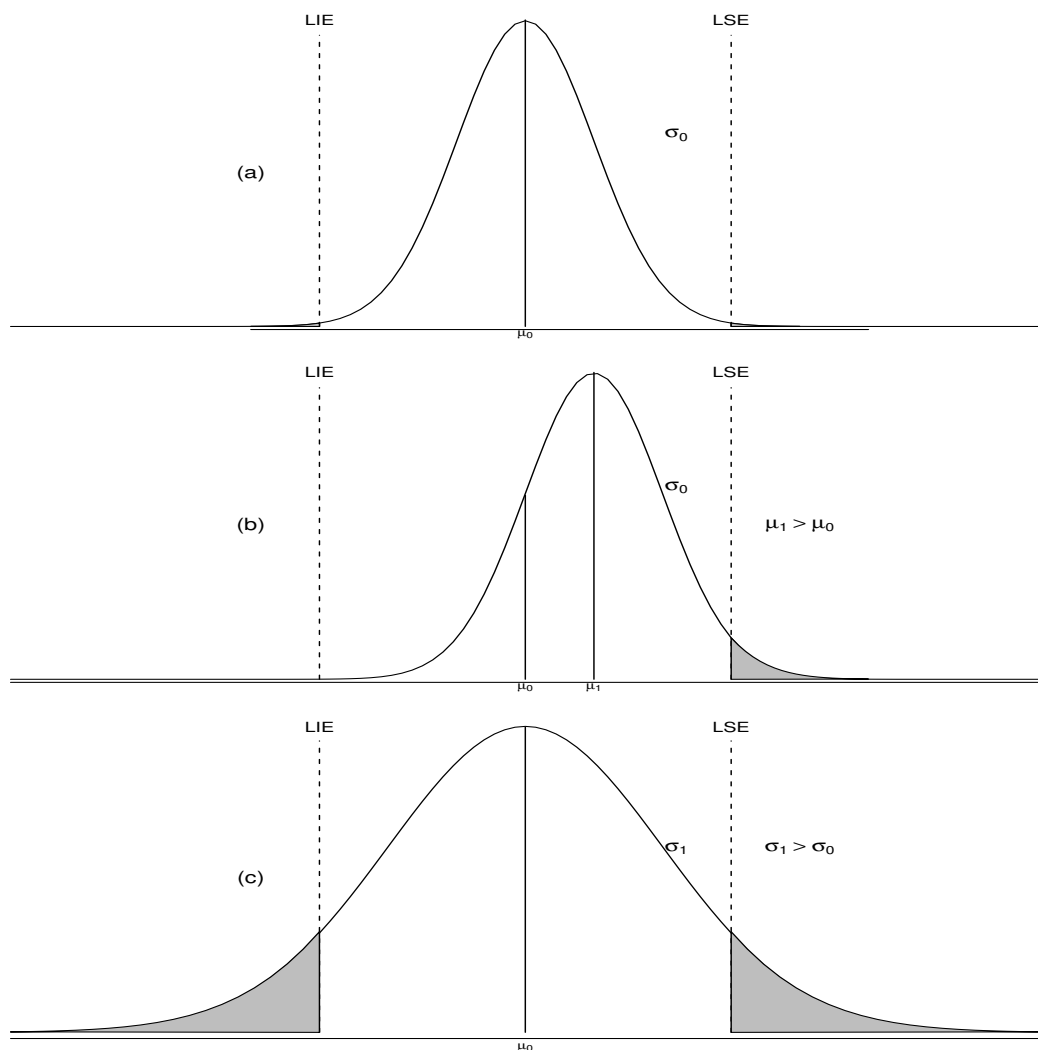


Figura 8.5: Efeito da média e desvio padrão em relação aos limites de especificação (LIE= limite inferior de especificação e LSE= limite superior de especificação).

Na Figura 8.5(a), o processo comporta-se dentro dos limites de especificação. Já, nas Figuras 8.5(b) e 8.5(c), existe uma chance maior de produtos serem defeituosos ou não-conformes.

8.2.1 Gráficos de controle para \bar{x} e R

Considere uma variável com distribuição Normal com média μ e desvio padrão σ (μ e σ desconhecidos). Se x_1, x_2, \dots, x_n é uma amostra de tamanho n , então a média da amostra é dada por:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

sendo \bar{x} normalmente distribuído com média μ e desvio padrão $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Sabemos, também, que, a probabilidade de que alguma média amostral não pertença ao intervalo $\mu \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ é $1 - \alpha$.

Como visto antes, $z_{\alpha/2}$ é geralmente substituído por 3, que significa um limite de 3 sigmas.

Em geral, para construção do gráfico de controle são necessárias $m = 20$ a $m = 25$ amostras de tamanho n . Normalmente, as amostras têm tamanho pequeno, 4 a 6 observações.

Sejam agora, $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$ as médias de cada uma das m amostras, então,

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_m}{m}$$

é o melhor estimador da média do processo. Nesse caso, $\bar{\bar{x}}$ representa a linha central do gráfico de controle.

Os limites de controle podem ser estimados a partir do desvio padrão ou da amplitude das m amostras.

Inicialmente, vamos trabalhar com a amplitude. Se x_1, x_2, \dots, x_n é uma amostra de tamanho n , a amplitude é a diferença entre o maior e o menor valor observado na amostra, isto é:

$$R = x_{max} - x_{min}.$$

Considerando R_1, R_2, \dots, R_m , a amplitude das m amostras, então,

$$\bar{R} = \frac{R_1 + R_2 + \dots + R_m}{m}$$

é definido como a amplitude média.

Os limites de controle para o gráfico \bar{x} serão dados por:

$$LSC = \bar{\bar{x}} + A_2\bar{R}$$

$$\text{Linha central} = \bar{\bar{x}}$$

$$LIC = \bar{\bar{x}} - A_2\bar{R}.$$

A_2 é uma constante tabelada para diferentes tamanhos de n . Essa constante é determinada em função da amplitude de variação e do desvio padrão das amostras, considerando uma distribuição Normal (demonstração omitida).

Comentários sobre o gráfico R

Embora o gráfico R tenha sido elaborado inicialmente para facilitar a obtenção dos limites de controle, ele é simples de ser construído e interpretado e isso é um ponto importante.

Para amostras com n pequeno (4, 5, 6) a informação fornecida pela amplitude não é muito diferente daquela fornecida pelo desvio padrão.

O gráfico R é sensível a dados discrepantes (*outliers*). Por isso, é importante analisar a origem de um outlier antes de tirar conclusões sobre o gráfico R. Claro que um processo que contém *outliers* deve ter problemas.

8.2.2 Gráficos de controle para \bar{x} e s

Além de avaliarmos a variabilidade pela amplitude, é possível também, utilizar o desvio padrão em um gráfico de controle.

Em geral, o gráfico de \bar{x} e s é utilizado quando:

- $n > 10$ observações por amostra;
- n é variável de amostra para amostra.

Para $n > 10$, é preferível adotar o gráfico s em vez do gráfico R , pois a precisão será melhor.

Gráfico s

Para construção do gráfico utilizamos métodos semelhantes aos utilizados nos gráficos de controle \bar{x} e R , com modificações nas expressões.

Para cada amostra, devemos obter a média e o desvio padrão. Uma estimativa da variância amostral pode ser obtida por:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}.$$

Mas, o desvio padrão amostral de s não é um estimador não viesado de σ . Considerando uma distribuição Normal, s estima $c_4\sigma$. Aqui, c_4 é uma constante que varia em função do tamanho amostral n . Ainda, o desvio padrão de s é dado por $\sigma\sqrt{1 - c_4^2}$.

Considerando, também, que, $E(s) = c_4\sigma$, a linha central é obtida por $c_4\sigma$. Assim, os limites de controle 3 sigmas para s são obtidos por:

$$\begin{aligned} LSC &= c_4\sigma + 3\sigma\sqrt{1 - c_4^2} \\ LIC &= c_4\sigma - 3\sigma\sqrt{1 - c_4^2}. \end{aligned}$$

Também, como nos gráficos de controle para \bar{x} e R , podemos definir constantes. Nesse caso,

$$\begin{aligned} B_5 &= c_4 - 3\sqrt{1 - c_4^2} \\ B_6 &= c_4 + 3\sqrt{1 - c_4^2} \end{aligned}$$

Logo, os limites de confiança podem ser reescritos como:

$$\begin{aligned} LSC &= B_6\sigma \\ \text{Linha central} &= c_4\sigma \\ LIC &= B_5\sigma. \end{aligned}$$

Valores de B_5 e B_6 são tabelados para vários tamanhos de n .

Se o valor de σ não é fornecido, ele pode ser estimado de dados anteriores. Considerando que se tenham m amostras, cada uma com tamanho n , e seja s_i o desvio padrão da i -ésima amostra, a média dos desvios-padrão é:

$$\bar{s} = \frac{1}{m} \sum_{i=1}^m s_i.$$

A estatística $\frac{\bar{s}}{c_4}$ é um estimador não viesado de σ . Logo, podemos reescrever os parâmetros do gráfico de controle s :

$$\begin{aligned} LSC &= \bar{s} + 3\frac{\bar{s}}{c_4}\sqrt{1 - c_4^2} \\ \text{Linha central} &= \bar{s} \\ LIC &= \bar{s} - 3\frac{\bar{s}}{c_4}\sqrt{1 - c_4^2} \end{aligned}$$

Definindo B_3 e B_4 como constantes tal que:

$$\begin{aligned} B_3 &= 1 - \frac{3}{c_4}\sqrt{1 - c_4^2} \\ B_4 &= 1 + \frac{3}{c_4}\sqrt{1 - c_4^2}, \end{aligned}$$

Pode-se reescrever as expressões do gráfico de controle por:

$$LSC = B_4 \bar{s}$$

$$\text{Linha central} = \bar{s}$$

$$LIC = B_3 \bar{s}.$$

Observe, ainda, que, $B_4 = B_6/c_4$ e $B_3 = B_5/c_4$.

Gráfico \bar{x}

Utilizando $\frac{\bar{s}}{c_4}$ como estimador de σ , pode-se definir os limites de controle para o gráfico \bar{x} como:

$$LSC = \bar{\bar{x}} + \frac{3\bar{s}}{c_4\sqrt{n}}$$

$$\text{Linha central} = \bar{\bar{x}}$$

$$LIC = \bar{\bar{x}} - \frac{3\bar{s}}{c_4\sqrt{n}}$$

Definindo-se a constante $A_3 = 3/(c_4\sqrt{n})$ tem-se

$$LSC = \bar{\bar{x}} + A_3 \bar{s}$$

$$\text{Linha central} = \bar{\bar{x}}$$

$$LIC = \bar{\bar{x}} - A_3 \bar{s}$$

8.2.3 Exemplos

Exemplo 8.1. Considere um conjunto de dados sobre espessura de uma peça de metal medida em milímetros (Tabela 8.1). Nesse exemplo, 20 amostras de tamanho $n=5$ foram obtidas. Um gráfico de controle para a média para esses dados é apresentado na Figura 8.6

Podemos observar no gráfico da Figura 8.6 que não há pontos localizados além dos limites de controle inferior e superior. Portanto, considerando as regras sugeridas na Seção 8.1.3 podemos dizer que o processo está sob controle.

Além disso, para os dados da Tabela 8.1, os gráficos de controle R e s são apresentados nas Figuras 8.7 e 8.8, respectivamente. Nesses gráficos, podemos observar que não há pontos fora dos limites de controle. Seguindo as regras sugeridas na página 8.1.3, também não há suspeitas de que o processo esteja fora de controle.

Tabela 8.1: Dados de espessura (mm) de uma peça de metal.

Amostra	Medidas				
	1	2	3	4	5
1	9.88	10.03	10.09	9.96	9.92
2	9.99	10.02	9.97	9.86	10.03
3	9.93	10.06	10.01	10.10	10.00
4	10.00	10.04	10.12	9.99	10.08
5	9.93	9.99	9.97	10.00	9.99
6	9.83	9.98	9.94	10.04	10.07
7	10.11	9.92	9.96	9.99	9.99
8	10.00	9.96	10.03	9.96	9.97
9	9.99	9.89	9.97	9.85	9.96
10	10.06	10.02	10.06	10.01	10.01
11	10.06	10.03	9.95	9.93	9.94
12	10.13	10.10	9.87	9.81	10.06
13	9.98	10.01	9.94	10.11	9.91
14	10.01	10.04	10.00	10.11	10.03
15	10.03	9.95	10.23	9.93	9.96
16	10.09	9.97	9.93	10.02	10.09
17	9.96	10.02	10.07	10.03	10.14
18	9.97	10.04	9.96	9.85	9.90
19	9.93	10.12	9.88	9.81	9.94
20	9.99	9.92	9.99	9.92	9.95

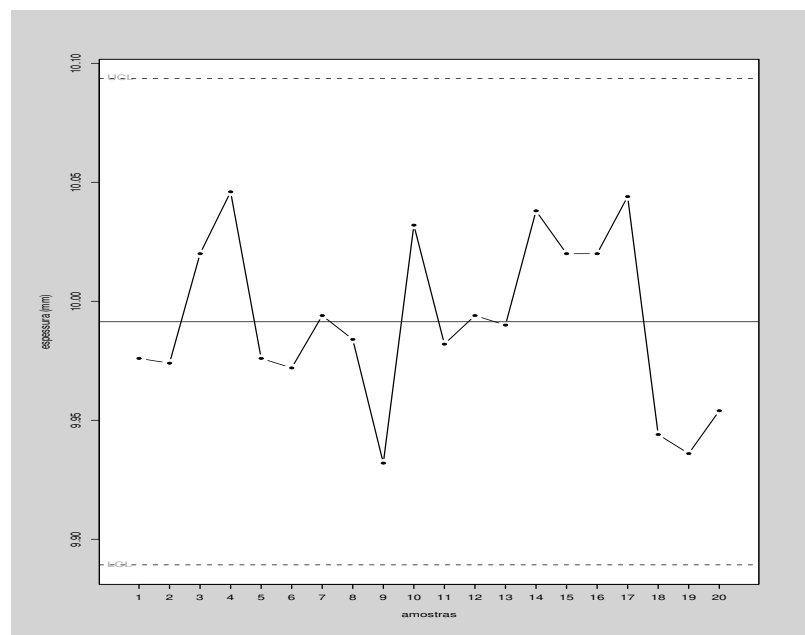


Figura 8.6: Gráfico de controle para média - sem problemas (Tabela 8.1).

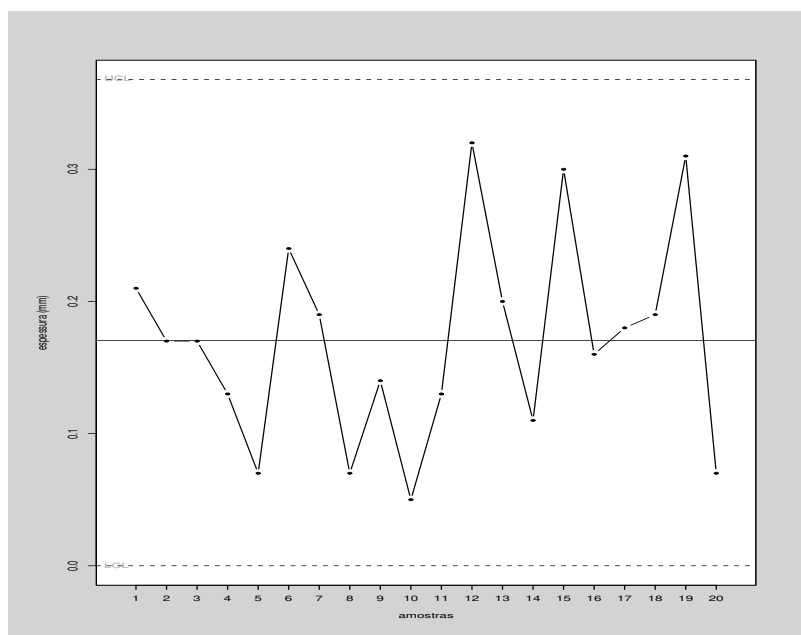


Figura 8.7: Gráfico de controle para amplitude - dados da Tabela 8.1

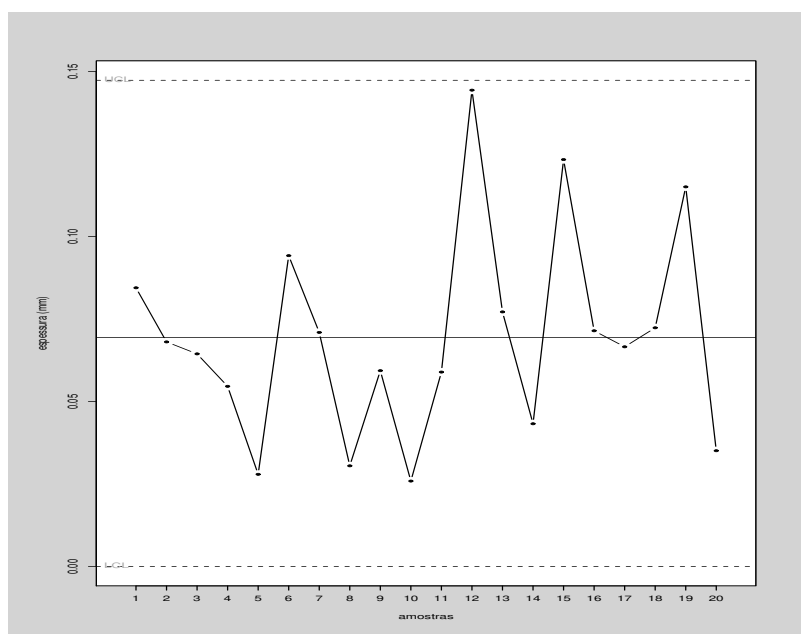


Figura 8.8: Gráfico de controle para o desvio padrão - dados da Tabela 8.1.

Exemplo 8.2. Consideremos agora, outro conjunto de dados sobre espessura de uma peça de metal, avaliados em outro momento (Tabela 8.2). Nessa situação, observe que um dos pontos amostrais ultrapassa o limite superior de controle (Figura 8.9). Em princípio, nesse momento deveria ser realizada uma intervenção no processo para descobrir a causa do problema.

Tabela 8.2: Dados de espessura (mm) de uma peça de metal avaliados após intervenção no processo.

Amostra	Medidas				
	1	2	3	4	5
1	10,18	9,90	9,94	9,97	9,92
2	9,92	10,06	9,99	10,01	10,18
3	10,03	10,22	10,18	10,03	10,15
4	10,04	9,93	9,98	10,40	10,08
5	10,07	10,06	10,10	9,89	10,10
6	10,01	10,06	10,05	9,92	9,98
7	9,93	10,06	10,01	9,99	9,97
8	9,96	10,08	9,91	9,99	10,03
9	10,10	9,94	9,98	9,90	9,97
10	9,97	10,11	10,05	10,01	10,07
11	9,98	10,00	10,03	10,14	10,06
12	10,04	9,87	10,03	10,01	9,93
13	10,01	10,01	10,04	10,06	10,03
14	10,09	10,06	10,09	10,10	10,06
15	10,03	9,93	9,95	9,93	9,87
16	9,89	9,93	9,96	10,00	10,19
17	10,03	10,00	10,00	9,97	9,94
18	9,98	9,97	9,98	9,86	10,05
19	10,13	9,95	10,05	10,02	10,07
20	9,95	10,09	9,88	9,96	9,99

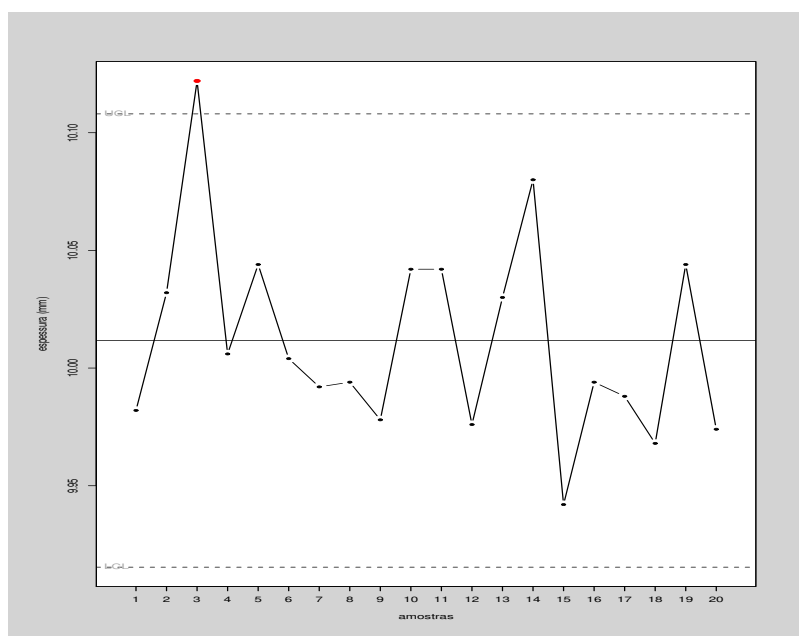


Figura 8.9: Gráfico de controle para média - com problemas (Tabela 8.2).

TABELAS

Tabela 4: Limites unilaterais de F ao nível de 5% de probabilidade

n1=número de graus de liberdade do numerador, n2= número de graus de liberdade do denominador

n2\ n1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,0	243,9	244,7	245,4	245,9	246,5	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,42	19,42	19,43	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,73	8,71	8,70	8,69	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,89	5,87	5,86	5,84	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,60	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,92	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51	3,49	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22	3,20	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01	2,99	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85	2,83	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72	2,70	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62	2,60	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53	2,51	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46	2,44	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40	2,38	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35	2,33	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31	2,29	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27	2,25	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,21	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20	2,18	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,22	2,20	2,18	2,16	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,20	2,17	2,15	2,13	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,18	2,15	2,13	2,11	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,15	2,13	2,11	2,09	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,11	2,09	2,07	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,12	2,09	2,07	2,05	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13	2,10	2,08	2,06	2,04	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,09	2,06	2,04	2,02	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,08	2,05	2,03	2,01	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,06	2,04	2,01	1,99	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,97	1,95	1,92	1,90	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,89	1,86	1,84	1,82	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,87	1,83	1,80	1,78	1,75	1,73	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,72	1,69	1,67	1,64	1,57	1,52	1,46	1,39	1,32	1,22	1,01

Tabela 5: Limites unilaterais de F ao nível de 1% de probabilidade

n1=número de graus de liberdade do numerador, n2= número de graus de liberdade do denominador

n2\n1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	20	24	30	40	60	120	∞
1	4052,2	4999,3	5403,5	5624,3	5764,0	5859,0	5928,3	5981,0	6022,4	6055,9	6083,4	6106,7	6125,8	6143,0	6157,0	6170,0	6208,7	6234,3	6260,4	6286,4	6313,0	6339,5	6365,6
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,41	99,42	99,42	99,43	99,43	99,44	99,45	99,46	99,47	99,48	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05	26,98	26,92	26,87	26,83	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,31	14,25	14,20	14,15	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,82	9,77	9,72	9,68	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,52	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,36	6,31	6,28	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,61	5,56	5,52	5,48	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,01	4,96	4,92	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,65	4,60	4,56	4,52	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,21	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,97	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,91	3,86	3,82	3,78	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,62	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,49	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,50	3,45	3,41	3,37	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,27	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,32	3,27	3,23	3,19	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,12	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,18	3,13	3,09	3,05	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,12	3,07	3,03	2,99	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	3,07	3,02	2,98	2,94	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	3,02	2,97	2,93	2,89	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,98	2,93	2,89	2,85	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,94	2,89	2,85	2,81	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,90	2,86	2,81	2,78	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99	2,93	2,87	2,82	2,78	2,75	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,96	2,90	2,84	2,79	2,75	2,72	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93	2,87	2,81	2,77	2,73	2,69	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,79	2,74	2,70	2,66	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,61	2,56	2,52	2,48	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,44	2,39	2,35	2,31	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,40	2,34	2,28	2,23	2,19	2,15	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,25	2,18	2,13	2,08	2,04	2,00	1,88	1,79	1,70	1,59	1,47	1,32	1,01

Tabela 6: Valores de t em níveis de 10% a 0,1% de probabilidade.

GL	0,1	0,05	0,02	0,01	0,001
1	6,31	12,71	31,82	63,66	636,58
2	2,92	4,30	6,96	9,92	31,60
3	2,35	3,18	4,54	5,84	12,92
4	2,13	2,78	3,75	4,60	8,61
5	2,02	2,57	3,36	4,03	6,87
6	1,94	2,45	3,14	3,71	5,96
7	1,89	2,36	3,00	3,50	5,41
8	1,86	2,31	2,90	3,36	5,04
9	1,83	2,26	2,82	3,25	4,78
10	1,81	2,23	2,76	3,17	4,59
11	1,80	2,20	2,72	3,11	4,44
12	1,78	2,18	2,68	3,05	4,32
13	1,77	2,16	2,65	3,01	4,22
14	1,76	2,14	2,62	2,98	4,14
15	1,75	2,13	2,60	2,95	4,07
16	1,75	2,12	2,58	2,92	4,01
17	1,74	2,11	2,57	2,90	3,97
18	1,73	2,10	2,55	2,88	3,92
19	1,73	2,09	2,54	2,86	3,88
20	1,72	2,09	2,53	2,85	3,85
21	1,72	2,08	2,52	2,83	3,82
22	1,72	2,07	2,51	2,82	3,79
23	1,71	2,07	2,50	2,81	3,77
24	1,71	2,06	2,49	2,80	3,75
25	1,71	2,06	2,49	2,79	3,73
26	1,71	2,06	2,48	2,78	3,71
27	1,70	2,05	2,47	2,77	3,69
28	1,70	2,05	2,47	2,76	3,67
29	1,70	2,05	2,46	2,76	3,66
30	1,70	2,04	2,46	2,75	3,65
40	1,68	2,02	2,42	2,70	3,55
60	1,67	2,00	2,39	2,66	3,46
120	1,66	1,98	2,36	2,62	3,37
∞	1,65	1,96	2,33	2,58	3,30

Tabela 7: Valores da amplitude total estudentizada (q), para uso no teste de Tukey, ao nível de 5% de probabilidade.

I=número de tratamentos, GLRES= número de graus de liberdade do resíduo.

GLRES\I	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	17,97	26,98	32,82	37,08	40,41	43,40	45,40	47,36	49,07	50,59	51,96	53,20	54,33	55,36	56,32	57,22	58,04	58,83	59,56
2	6,09	8,33	9,80	10,88	11,74	12,44	13,03	13,54	13,99	14,39	14,75	15,08	15,33	15,65	15,91	16,14	16,37	16,57	16,77
3	4,50	5,91	6,83	7,50	8,04	8,48	8,85	9,18	9,46	9,72	9,95	10,15	10,35	10,53	10,69	10,84	10,98	11,11	11,24
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83	8,03	8,21	8,37	8,53	8,66	8,79	8,91	9,03	9,13	9,23
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	7,00	7,17	7,32	7,47	7,60	7,72	7,83	7,93	8,03	8,12	8,21
6	3,46	4,34	4,90	5,31	5,63	5,90	6,12	6,32	6,49	6,65	6,79	6,92	7,03	7,14	7,24	7,34	7,43	7,51	7,59
7	3,34	4,17	4,68	5,06	5,36	5,61	5,82	6,00	6,16	6,30	6,43	6,55	6,66	6,76	6,85	6,94	7,02	7,10	7,17
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,05	6,18	6,29	6,39	6,48	6,57	6,65	6,73	6,80	6,87
9	3,20	3,95	4,42	4,76	5,02	5,24	5,43	5,60	5,74	5,87	5,98	6,09	6,19	6,28	6,36	6,44	6,51	6,58	6,64
10	3,15	3,88	4,33	4,65	4,91	5,12	5,31	5,46	5,60	5,72	5,83	5,94	6,03	6,11	6,19	6,27	6,34	6,41	6,47
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49	5,61	5,71	5,81	5,90	5,98	6,06	6,13	6,20	6,27	6,33
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,40	5,51	5,62	5,71	5,80	5,88	5,95	6,02	6,09	6,15	6,21
13	3,06	3,74	4,15	4,45	4,69	4,89	5,05	5,19	5,32	5,43	5,53	5,63	5,71	5,79	5,86	5,93	6,00	6,06	6,11
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46	5,55	5,64	5,71	5,79	5,85	5,92	5,97	6,03
15	3,01	3,67	4,08	4,37	4,60	4,78	4,94	5,08	5,20	5,31	5,40	5,49	5,57	5,65	5,72	5,79	5,85	5,90	5,96
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26	5,35	5,44	5,52	5,59	5,66	5,73	5,79	5,84	5,90
17	2,98	3,63	4,02	4,30	4,52	4,71	4,86	4,99	5,11	5,21	5,31	5,39	5,47	5,54	5,61	5,68	5,73	5,79	5,84
18	2,97	3,61	4,00	4,28	4,50	4,67	4,82	4,96	5,07	5,17	5,27	5,35	5,43	5,50	5,57	5,63	5,69	5,74	5,79
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14	5,23	5,32	5,39	5,46	5,53	5,59	5,65	5,70	5,75
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,11	5,20	5,28	5,36	5,43	5,49	5,55	5,61	5,66	5,71
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	5,01	5,10	5,18	5,25	5,32	5,38	5,44	5,49	5,55	5,59
30	2,89	3,49	3,85	4,10	4,30	4,46	4,60	4,72	4,82	4,92	5,00	5,08	5,15	5,21	5,27	5,33	5,38	5,43	5,48
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,64	4,74	4,82	4,90	4,98	5,04	5,11	5,16	5,22	5,27	5,31	5,36
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81	4,88	4,94	5,00	5,06	5,11	5,15	5,20	5,24
120	2,80	3,36	3,69	3,92	4,10	4,24	4,36	4,47	4,56	4,64	4,71	4,78	4,84	4,90	4,95	5,00	5,04	5,09	5,13
∞	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47	4,55	4,62	4,69	4,74	4,80	4,85	4,89	4,93	4,97	5,01

Tabela 8: Distribuição de Qui-quadrado. Valor crítico de χ^2 tal que $P(\chi_k^2 > \chi_0^2) = \alpha$.

GL	0,995	0,975	0,05	0,025	0,01	0,005
1	0,00	0,00	3,84	5,02	6,63	7,88
2	0,01	0,05	5,99	7,38	9,21	10,60
3	0,07	0,22	7,81	9,35	11,34	12,84
4	0,21	0,48	9,49	11,14	13,28	14,86
5	0,41	0,83	11,07	12,83	15,09	16,75
6	0,68	1,24	12,59	14,45	16,81	18,55
7	0,99	1,69	14,07	16,01	18,48	20,28
8	1,34	2,18	15,51	17,53	20,09	21,95
9	1,73	2,70	16,92	19,02	21,67	23,59
10	2,16	3,25	18,31	20,48	23,21	25,19
11	2,60	3,82	19,68	21,92	24,73	26,76
12	3,07	4,40	21,03	23,34	26,22	28,30
13	3,57	5,01	22,36	24,74	27,69	29,82
14	4,07	5,63	23,68	26,12	29,14	31,32
15	4,60	6,26	25,00	27,49	30,58	32,80
16	5,14	6,91	26,30	28,85	32,00	34,27
17	5,70	7,56	27,59	30,19	33,41	35,72
18	6,26	8,23	28,87	31,53	34,81	37,16
19	6,84	8,91	30,14	32,85	36,19	38,58
20	7,43	9,59	31,41	34,17	37,57	40,00
21	8,03	10,28	32,67	35,48	38,93	41,40
22	8,64	10,98	33,92	36,78	40,29	42,80
23	9,26	11,69	35,17	38,08	41,64	44,18
24	9,89	12,40	36,42	39,36	42,98	45,56
25	10,52	13,12	37,65	40,65	44,31	46,93
26	11,16	13,84	38,89	41,92	45,64	48,29
27	11,81	14,57	40,11	43,19	46,96	49,65
28	12,46	15,31	41,34	44,46	48,28	50,99
29	13,12	16,05	42,56	45,72	49,59	52,34
30	13,79	16,79	43,77	46,98	50,89	53,67
40	20,71	24,43	55,76	59,34	63,69	66,77
50	27,99	32,36	67,50	71,42	76,15	79,49
60	35,53	40,48	79,08	83,30	88,38	91,95
70	43,28	48,76	90,53	95,02	100,43	104,21
80	51,17	57,15	101,88	106,63	112,33	116,32
90	59,20	65,65	113,15	118,14	124,12	128,30
100	67,33	74,22	124,34	129,56	135,81	140,17

Tabela 9: Constantes utilizadas em gráficos de controle.

n	2	3	4	5	6	7	8	9	10
d_2	1,128	1,693	2,059	2,326	2,534	2,704	2,847	2,970	3,078
A_2	1,880	1,023	0,729	0,577	0,483	0,419	0,373	0,337	0,308
D_3	0	0	0	0	0	0,076	0,136	0,184	0,223
D_4	3,267	2,574	2,282	2,115	2,004	1,924	1,864	1,816	1,777

Tabela 10: Tabela de números aleatórios.

5831	3593	7697	2402	7192	9763	2608	7666	4805	8983	0329	4626
1431	2626	2218	6421	4003	0693	4081	9964	0887	4587	2648	2129
0498	9704	4756	0118	1180	1277	1498	1963	4045	1073	6264	5038
4925	2853	1290	0099	9595	6956	2372	0274	2471	3788	9312	4956
7262	4057	4845	8640	0425	4696	4774	4046	0852	5475	7236	4777
5170	0203	4461	4874	1298	2457	2775	3462	6009	5119	7337	1302
4168	7779	4144	1390	9695	6552	9329	2647	2746	6260	3101	4268
6591	1320	8902	3486	9066	5121	9471	8821	4898	6327	2711	2013
0792	5373	4664	9335	6172	3755	5232	6237	9471	1128	2456	8640
3924	1947	1923	5535	1086	6247	5881	1976	5393	9006	9362	7370
1070	7911	0222	2388	2552	4188	4593	8292	3719	1226	9038	4724
1221	6165	0037	4000	5508	9928	8988	1470	5709	0600	3585	1096
4035	5872	3871	7458	6621	7333	2129	7857	7369	4400	8369	6732
8793	8982	8134	2611	4941	6740	8781	2886	2012	4945	0264	3763
1762	7386	6202	4037	9508	5436	8916	0458	7179	6309	4185	4682
6866	9907	9743	1329	4079	3955	9463	9986	5227	1770	2769	5101
5428	7775	7116	0745	7552	1681	5522	1667	4898	6958	5210	4028
8048	9149	3589	0240	3546	4945	4353	9374	7637	3488	0494	8868
6563	6774	9651	1073	5409	8034	3418	9449	5214	6998	4539	7871
5044	3203	4891	7862	8298	9234	7204	7190	6566	1856	2168	9239
0360	7198	2739	3830	3786	8491	9299	3728	7012	1126	3983	9363
1160	3338	6426	0725	6483	3444	1453	2868	4664	2212	1538	2411
7811	8247	8095	2753	6160	7533	6438	4394	3756	8422	4025	1408
1366	7558	9937	6569	7616	2278	7790	3027	0741	8139	5109	3346
1354	4499	9764	7151	7209	9928	4897	4321	7532	6877	3750	3949
1637	6059	6343	0849	8914	3694	3972	0833	8150	2178	2447	5433
3293	8925	7496	6506	4186	9895	0069	0818	3768	8880	3248	8456
8902	5854	4263	6916	4360	7137	4826	9219	1854	4410	0628	7700